Detection and Estimation in Gaussian Random Fields: Minimax Theory and Efficient Algorithms

by

Hossein Keshavarz Shenastaghi

A dissertation submitted in partial fulfilment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in the University of Michigan
2017

Doctoral Committee:

Associate Professor XuanLong Nguyen, Co-Chair
Associate Professor Clayton D. Scott, Co-Chair
Associate Professor Veronica Berrocal
Professor Xuming He


THE UNIVERSITY OF MICHIGAN
ARTES · SCIENTIA · VERITAS
1817

hksh@umich.edu

ORCID iD: 0000 – 0003 – 3809 – 7875

This dissertation is dedicated to my family and friends without whom I could not successfully finish my thesis work.

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my thesis advisors Xuan-Long Nguyen and Clayton Scott for their continuous guidance and support throughout my PhD studies. Their several thoughtful and meticulous suggestions immensely assisted me to augment my knowledge about various topics and have broaden my perspective of research. Without their sincere support and encouragement, it was extremely difficult to successfully accomplish this journey.

I am very thankful to my committee members Xuming He and Veronica Berrocal for their invaluable support and motivating comments. My sincere thanks also goes to Michael L Stein at University of Chicago. I have genuinely benefited from his broad insight and astute comments. It is also imperative to thank STATMOS research network for giving me excellent opportunity to attend many conferences, workshops and meet many experienced and exceptional researchers in US.

I have indebted to my dear mother, who has always stood by me like a pillar in times of need and to whom I owe my life for her constant love. I would like to specially thank my loving family for standing by me in all situations. Finally I am truly grateful to my great friends in Ann Arbor for their affection and kindness throughout these years.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AUC Area Under Curve

CUSUM Cumulative Sum

GLRT Generalized Likelihood Ratio Test

GP Gaussian Process

IF Inversion Free

LIF Local-Inversion Free

MLE Maximum Likelihood Estimator

PGLRT Plug-in Generalized Likelihood Ratio Test

RMSE Root Mean-Squared Error

ROC receiver operating characteristic

ABSTRACT

The strong dependence between samples in large spatial data sets is the primary challenge of designing statistically consistent and computationally efficient inference algorithms. Gaussian processes provide a powerful tool for modelling the spatial dependence patterns and play a crucial role in numerous tractable inference algorithms.

This thesis addresses two important problems on high-dimensional Gaussian spatial processes. We first focus on scalable estimation of covariance parameters. Evaluating the log-likelihood function of Gaussian process data can be computationally intractable, particularly for large and irregularly spaced observations. We build a broad family of surrogate loss functions based on local moment-matching and a block diagonal approximation of the covariance matrix. This class of algorithms provides a versatile balance between the estimation accuracy and the computational cost. The fixed domain asymptotic behaviour of the proposed method is thoroughly studied for the isotropic Matern processes observed on a multi-dimensional irregular lattice.

In the second part, the main emphasis is on minimax optimal detection of abrupt changes in the mean of a one-dimensional Gaussian process. Our main contribution is to show that in the fixed-domain asymptotic regime, neglecting the dependence structures leads to suboptimal performance. We first show that plugging the estimated covariance matrix into the Generalized Likelihood Ratio Test (GLRT) provides a test with near minimax asymptotic optimality. On the other hand, the suboptimality of the cumulative sum test, which ignores the dependence structure of data, is substantiated for a vast range of covariance functions.

CHAPTER 1

Introduction

With the advent of data mining, spatial-temporal data are more ubiquitous and richer ever than before. Substantial amount of high-dimensional spatial-temporal are continuously collected in environmental, biological and social science. The increasing availability of such large data sets bring forward new challenges and opportunities for researchers in wide variety of disciplines. Particularly in the past two decades, there has been voluminous research in the data science community on designing efficient inference algorithms for massive spatial-temporal datasets.

The field of *spatial statistics* encapsulates a broad array of methodology for analyzing spatial-temporal processes. The first law of geography states that "everything is related to everything else but nearby things are more related than distant things". In other words the independence assumption, which provides a very convenient framework for developing tractable inference algorithms and is crucial for the theoretical understanding of large data sets, is overly strong and unrealistic in the field of spatial statistics Consequently more diverse range of statistical, probabilistic and numerical tools are required for modelling the dependence structure and developing tractable estimation, detection and prediction algorithms for spatial-temporal data.

Using Gaussian Process (GP) for spatial modelling has became a common practice in the geostatistical literature (see e.g., [Cre15, GDFG10]). The versatile correlation structure in the GP models can conveniently capture a wide range of spatial behaviours. Furthermore using GPs for modelling dependencies in large data sets is essential for developing computationally tractable inference algorithms. The covariance function is typically specified up to a finite number of parameters, to guarantee the positive definiteness of its estimate.

For designing efficient inference algorithms for massive spatial GP, one is confronted with multiple modelling, computational and theoretical challenges, which we briefly mention here.

(a) Strong dependence between the near by observations requires to be intelligently in-

corporated into the inference procedure. In other words neglecting the dependence structure of the process leads to suboptimal result.

(b) From the computational perspective, the covariance matrix of the observations is commonly near ill-conditioned, particularly for large sample size. AS a result, an accurate evaluation of the likelihood function, which is pivotal for many estimation and detection algorithms, is almost infeasible.

(c) Due to the absence of independence assumption, the asymptotic analysis is typically far more difficult and requires numerous statistical and probabilistic tools.

For conducting a thorough asymptotic analysis of inference algorithms for spatial and temporal GPs , one is confronted with two fundamentally different regimes, the *increasing domain asymptotics* and *fixed domain (infill) asymptotics*. The former arises naturally in time series analysis, which is distinguished by the constraint that the distance between consecutive sampling *time* points are bounded away from zero. Fixed domain asymptotics, one the other hand, is a more suitable setting when the index set of sampling points is bounded, so that the observations get denser in a bounded region as the sample size increases. This is the case for spatially distributed data [Ste12], where the domain of the index set is typically of one, two or three dimensions. This approach is also appropriate in the context of change detection for non-stationary processes [Ada98, D$^+$97, LS08].

1.1 Preliminaries

This section provides the necessary background on GPs for easier understanding the subsequent chapters on covariance estimation and change-point detection. We mainly focus on weakly stationary processes and the associated spectral theory. Let $\mathcal{D}$ be an arbitrary subset of $\mathbb{R}^d$. Consider a stochastic process $G := \{G(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{D}\}$ on $\mathcal{D}$ (collection of random variables indexed by $\mathcal{D}$). G is called a GP on $\mathcal{D}$, if for any finite sub-collection $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\} \subset \mathcal{D}$, the random column vector $[G(\boldsymbol{s}_1), \ldots, G(\boldsymbol{s}_n)]^\top$ has a multivariate Gaussian distribution.

Definition 1.1 (*The mean and covariance functions*)**.** The real valued functions $m : \mathcal{D} \mapsto \mathbb{R}$ and $K : \mathcal{D} \times \mathcal{D} \mapsto \mathbb{R}$ respectively denote the mean and covariance functions of G, if for any finite subset $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\} \subset \mathcal{D}$,

$$\begin{bmatrix} G(\boldsymbol{s}_1) \\ \vdots \\ G(\boldsymbol{s}_n) \end{bmatrix} = \mathcal{N}\left(\begin{bmatrix} m(\boldsymbol{s}_1) \\ \vdots \\ m(\boldsymbol{s}_n) \end{bmatrix}, \begin{bmatrix} K(\boldsymbol{s}_1, \boldsymbol{s}_1) & \ldots & K(\boldsymbol{s}_1, \boldsymbol{s}_n) \\ \vdots & \ddots & \vdots \\ K(\boldsymbol{s}_n, \boldsymbol{s}_1) & \ldots & K(\boldsymbol{s}_n, \boldsymbol{s}_n) \end{bmatrix} \right).$$

The mean and covariance function are aptly named as for any pair of points $\boldsymbol{s}, \boldsymbol{s}' \in \mathcal{D}$,

$$m(\boldsymbol{s}) = \mathbb{E}G(\boldsymbol{s}), \quad K(\boldsymbol{s}, \boldsymbol{s}') = \mathbb{E}\left[(G(\boldsymbol{s}) - m(\boldsymbol{s}))(G(\boldsymbol{s}') - m(\boldsymbol{s}'))\right].$$

The covariance function must trivially satisfy the following properties:

(a) (Boundedness) $K(\boldsymbol{s}, \boldsymbol{s})$ should be bounded for any $\boldsymbol{s} \in \mathcal{D}$, i.e. $\operatorname{var} G(\boldsymbol{s}) = K(\boldsymbol{s}, \boldsymbol{s}) < \infty$.

(b) (Symmetry) For any $\boldsymbol{s}, \boldsymbol{s}' \in \mathcal{D}$, $K(\boldsymbol{s}, \boldsymbol{s}') = K(\boldsymbol{s}', \boldsymbol{s})$.

(c) (Positive semi-definiteness) The following inequality holds for any finite n, an arbitrary set of real coefficients $\{c_1, \ldots, c_n\}$ and any $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\} \subset \mathcal{D}$.

$$\operatorname{var}\left[\sum_{i=1}^{n} c_i G(\boldsymbol{s}_i)\right] = \sum_{i,j=1}^{n} c_i c_j K\left(\boldsymbol{s}_i, \boldsymbol{s}_j\right) \geq 0.$$

As a common simplifying assumption on the stochastic process, we assume that the probabilistic structure of G looks similar in the different regions of $\mathcal{D}$. The following Definition rigorously expresses such assumption.

Definition 1.2. G is said to be *strictly stationary* stochastic process on $\mathcal{D}$ if for all finite n, all real numbers $t_1, \ldots, t_n$, and arbitrary points $\boldsymbol{t}, \boldsymbol{s}_1, \ldots, \boldsymbol{s}_n \in \mathcal{D}$

$$\mathbb{P}(G(\boldsymbol{s}_1 + \boldsymbol{t}) \leq t_1, \ldots, G(\boldsymbol{s}_n + \boldsymbol{t}) \leq t_n) = \mathbb{P}(G(\boldsymbol{s}_1) \leq t_1, \ldots, G(\boldsymbol{s}_n) \leq t_n).$$

Simply put, a strictly stationary process is invariant to translations in the input space.

The strict stationarity is very restrictive and almost infeasible to validate in practical scenarios. A weaker form of stationarity typically employed in statistics and machine learning is known as the weak stationarity, which is defined in terms of the first two moments of G. In particular, G is called *weakly stationary* if

- There exists a scalar $m_0 \in \mathbb{R}$ such that $m(\boldsymbol{s}) = m_0$ for any $\boldsymbol{s} \in \mathcal{D}$. Namely the mean function is constant on $\mathcal{D}$.
- The covariance function $K(\boldsymbol{s}, \boldsymbol{s}')$ depend only on $\boldsymbol{s}' - \boldsymbol{s}$. In other words, there is a symmetric, bounded and positive semi-definite function K such that $\operatorname{cov}\left[G(\boldsymbol{s}), G(\boldsymbol{s}')\right] = K(\boldsymbol{s}' - \boldsymbol{s})$.

Note that the notions of weak and strict stationarity are identical for GPs. Thus we solely focus on weakly stationary (or in short stationary) processes in this thesis. We conclude this section by introducing a commonly used classes of stationary covariance functions.

Definition 1.3 (*Geometric anisotropic* covariance functions)**.** Let $A \in \mathbb{R}^d$ be a symmetric positive definite matrix. The real-valued stationary process G is called geometric anisotropic on $\mathcal{D}$ if the covariance function K satisfies

$$\operatorname{cov}[G(\boldsymbol{s}), G(\boldsymbol{s}')] = K\left(\sqrt{(\boldsymbol{s}'-\boldsymbol{s})^\top A\,(\boldsymbol{s}'-\boldsymbol{s})}\right),$$

for any pair of points $\boldsymbol{s}, \boldsymbol{s}' \in \mathcal{D}$. Let $B = A^{1/2}$ stands for the symmetric square-root of A. Then the covariance function can be equivalently rewritten as

$$\operatorname{cov}[G(\boldsymbol{s}), G(\boldsymbol{s}')] = K\left(\left\|B(\boldsymbol{s}'-\boldsymbol{s})\right\|_{\ell_2}\right), \quad \forall\, \boldsymbol{s}, \boldsymbol{s}' \in \mathcal{D}.$$

1.1.1 Spectral representation of the covariance function

Spectral methods are powerful tools for studying spatial processes. Expressing an integral representation of the covariance function K lies at the heart of the spectral analysis of GPs. Consider a non-negative measure F on $\mathbb{R}^d$ and construct the complex-valued function $K : \mathbb{R}^d \mapsto \mathbb{C}$ by

$$K(\boldsymbol{x}) = \int_{\mathbb{R}^d} \exp\left(j\boldsymbol{\omega}^\top \boldsymbol{x}\right) F(d\boldsymbol{\omega}), \quad \forall\, \boldsymbol{x} \in \mathbb{R}^d. \tag{1.1}$$

Here $j = \sqrt{-1}$ stands for the imaginary unit. It is easy to see that the function K defined in Eq. (1.1) is positive semi-definite. In other words, K represents the covariance function of a complex-valued stochastic process. The classical *Bochner's Theorem* ([Ste12], p. 24) guarantees the existence of such an integral representation similar to Eq. (1.1) for any covariance function K.

Theorem 1.1 (Bochner's Theorem)**.** A complex-valued function K on $\mathbb{R}^d$ is the covariance function of a stationary mean-square continuous complex-valued stochastic process G on $\mathbb{R}^d$ if and only if it can be expressed as in Eq. (1.1) with F a non-negative finite measure F.

F is typically called the associated spectral measure for G. If F has a well-defined density $\hat{K}$ (with respect to the Lebesgue measure), the spectral representation of G can be rewritten as

$$K(\boldsymbol{x}) = \int_{\mathbb{R}^d} \exp\left(j\boldsymbol{\omega}^\top \boldsymbol{x}\right) \hat{K}(\boldsymbol{\omega})\, d\boldsymbol{\omega}, \quad \forall\, \boldsymbol{x} \in \mathbb{R}^d.$$

In the subsequent chapters, $\hat{K}$ is called the *spectral density* of G. The duality property of the Fourier transform gives a spectral representation for $\hat{K}$ in terms of K. Specifically,

$$\hat{K}(\boldsymbol{\omega}) = (2\pi)^{-d} \int_{\mathbb{R}^d} \exp\left(-j\boldsymbol{\omega}^\top \boldsymbol{x}\right) K(\boldsymbol{x})\, d\boldsymbol{x}, \quad \forall\, \boldsymbol{\omega} \in \mathbb{R}^d.$$

1.1.2 Mean-square differentiability

The covariance function of a stationary GP can indirectly characterize the mean-square properties of the process. Let G be a zero mean GP on $\mathcal{D} \subset \mathbb{R}^d$. Then G is mean-square continuous at $\boldsymbol{s} \in \mathcal{D}$ if

$$\lim_{\boldsymbol{t} \to \boldsymbol{s}} \operatorname{var}[G(\boldsymbol{t}) - G(\boldsymbol{s})] = 2 \lim_{\boldsymbol{t} \to \boldsymbol{s}} [K(\mathbf{0}) - K(\boldsymbol{t} - \boldsymbol{s})] = 0. \tag{1.2}$$

According to Eq. (1.2), a stationary process is mean-square continuous at any $\boldsymbol{s} \in \mathcal{D}$ if and only if K is continuous at the origin.

Now we define the mean-square derivative of a process in $\mathcal{D}$. We use $\boldsymbol{e}_p,\ p = 1, \ldots, d$ to denote the unit vector in the pth direction. Furthermore, for any mean-square continuous stochastic process G and an arbitrary scalar δ define

$$G_{p,\delta}(\boldsymbol{s}) = \frac{G\left(\boldsymbol{s} + \delta \boldsymbol{e}_p\right) - G(\boldsymbol{s})}{\delta}, \quad p = 1, \ldots, d.$$

G is called a mean-square differentiable at $\boldsymbol{s}$ if for any real-valued vanishing sequence δ_n, the d-dimensional random vector $[G_{1,\delta_n}(\boldsymbol{s}), \ldots, G_{d,\delta_n}(\boldsymbol{s})]^\top$ converges in $\mathbb{L}^2$ to a limit independent of δ_n. The limit, which is represented by $G^{(1)}(\boldsymbol{s}) \coloneqq \left[G_1^{(1)}(\boldsymbol{s}), \ldots, G_d^{(1)}(\boldsymbol{s})\right]^\top$, refers to the mean-square derivative of G at $\boldsymbol{s}$. The covariance matrix of is $[G_{1,\delta_n}(\boldsymbol{s}), \ldots, G_{d,\delta_n}(\boldsymbol{s})]^\top$ can be easily obtained for stationary processes. Particularly as $n \to \infty$

$$\operatorname{cov}\{[G_{1,\delta_n}(\boldsymbol{s}), \ldots, G_{d,\delta_n}(\boldsymbol{s})]\} \to \left[-\frac{\partial^2 K}{\partial x_p \partial x_q}(\mathbf{0})\right]_{p,q=1}^{d}.$$

In other words, G is mean-square differentiable at any point $\boldsymbol{s} \in \mathcal{D}$, when $\nabla^2 K(\mathbf{0})$ is well-defined. In this case $\operatorname{cov}\left(G_1^{(1)}(\boldsymbol{s})\right) = -\nabla^2 K(\mathbf{0})$. In other words, G is mean-square differentiable whenever K is twice differentiable at the origin.

In summary the mean-square smoothness of the sample paths of stationary processes is closely related to the smoothness of the covariance function at $\mathbf{0}$. It is known that the near origin smoothness behaviour of K is reflected in the tail behaviour of the spectral density $\hat{K}$. We refer the interested reader to *Abelian and Tauberian Theorems* (see e.g., [Fel68, BGT87]) for further details.

1.2 Contributions of This Thesis

This thesis addresses two specific aspects of large GP data. First, we focus on a class of computationally efficient covariance estimation algorithms for high-dimensional stationary GPs. Designing Minimax optimal change-point detection procedure for an one-dimensional GP observed in a bounded domain is the second problem covered in this thesis.

In Chapter 2 we examine the increasing domain asymptotic properties of a computation and memory efficient covariance estimation algorithm introduced in Anitescu et al. [ACS16] (which will be called the inversion-free). This approach in is based upon maximizing a loss function which is independent of the precision matrix of the observations and so can be less challenging to evaluate than the Maximum Likelihood Estimator (MLE), particularly for irregularly spaced samples. However it has been claimed in that the inversion-free algorithm is statistically comparable to the MLE, only in the case that the covariance matrix has a bounded condition number. Indeed, a proper preprocessing of the samples is imperatively needed for reducing the condition number of the covariance matrix, particularly for large sample size. The consistency and asymptotic normality of the global and local maximizers of the inversion-free loss are studied in Chapter 2, for GPs observed on a d-dimensional randomly perturbed regular lattice.

In Chapter 3 we propose a novel class of covariance estimation algorithms (which will be referred to as the local inversion-free), built upon the sparse block diagonal approximation of the covariance matrix and the inversion-free approach [ACS16]. The introduced algorithm offers a rich spectrum of scalable, memory efficient and statistically consistent estimators which can be computed in a parallel fashion. We also provide sharp fixed domain asymptotic analysis of our algorithm for isotropic Matern covariance functions. Note that the technical details require a careful handling of the covariance matrix of the preprocessed samples. Our analysis surprisingly refutes the necessity of controlling the condition number of the covariance matrix for having $\sqrt{n}$-consistency and asymptotic normality properties. The performance of the proposed algorithm is also evaluated on moderate ($n = 10^4$) and large datasets ($n = 2.5 \times 10^5$) generated from isotropic and anisotropic Matern GP models.

In Chapter 4 we investigate the fixed-domain asymptotic behaviour of detecting abrupt changes in the mean of a single dimensional GP observed in $[0, 1]$. Motivated by the piecewise locally stationary time series models, developing fixed-domain asymptotics for change-point detection problem has become increasingly popular (see e.g., [Ada98,LS08]). Our main contribution is to show that in this asymptotic regime, neglecting the dependence structures leads to suboptimal performance. The objective of Chapter 4 is two fold: we first show that plugging the estimated covariance matrix into the GLRT provides a near

minimax asymptotic optimality. On the other hand, the suboptimality of the cumulative sum test is substantiated for a vast range of covariance functions, as a result of ignoring the dependence structure of data. This observation is corroborated by the simulation studies, which exhibit a wide gap between the rate of two detection algorithms.

CHAPTER 2

Inversion Free (IF) Covariance Estimation

2.1 Introduction

In the last two decades, there has been extensive research regarding the statistical and computational facets of the estimation of the GPs' covariance parameters. MLE was the earliest favored algorithm in the geostatistics community, e.g., Mardia et al. [MM84] and Ying [Yin91]. However, solving systems of linear equations is inevitable to evaluate the Gaussian likelihood. Notwithstanding the recent advances toward scalable and efficient solution of the system of linear equation (e.g., iterative *Krylov* subspace method or block preconditioned conjugate gradient algorithm [O'L80]) which moderately reduces the computational and memory costs of the direct evaluation of the precision matrix, obtaining the MLE of unknown covariance parameters using such linear systems solvers is still a strenuous task, especially for a generic Gaussian spatial process observed at numerous and possibly irregularly spaced locations. Vecchia [Vec88] proposed to approximate the likelihood function by neglecting the conditional correlation of distant sites given their nearest neighbors. Stein et al. [SCW04] extended this approximation method by considering more flexible choices of conditioning sets. Roughly speaking, sparse approximation of the inverse covariance matrix lies at the heart of the Vecchia's algorithm. Approximating the likelihood function by tapering the covariance matrix is another class of algorithms aiming to reduce the numerical burden of MLE (see Kaufman et al. [KSN08]). Tapering technique takes advantage of the sparsity of the approximated covariance matrix to accelerate linear solvers using the *Krylov subspace iteration method*, Furrer et al. [FGN06]. Recent studies [DZM$^+$09, KSN08, WL$^+$11] demonstrate the consistency and asymptotic normality of this algorithm under some mild conditions on the taper function. In more recent advances toward scalable evaluation of the MLE [ACW12, SCA$^+$13], the computational cost of each optimization iteration is reduced by considering an unbiased stochastic approximation of the score function. The proposed algorithms in [ACW12, SCA$^+$13] are statistically compa-

rable to MLE, if the condition number of the covariance matrix has a uniform upper bound (independent of the sample size).

Because of the obstacles of solving system of linear equations for massive data, which is necessary for tapered and exact MLE, it is of great interest to develop estimation techniques without requiring such extensive computations. Such class of algorithms, which will be referred to as *Inversion Free (IF)*, are based upon optimizing a loss function whose form (and its derivatives) is independent of the precision matrix of data. The first attempt toward such a goal has been done by Anitescu, Chen and Stein [ACS16]. Their proposed procedure is faster than likelihood based algorithms. In [ACS16], the covariance parameters are estimated by computing the global maximizer of a non-concave program. Simulation studies verify the efficiency of IF approach in the case that the covariance matrix has a bounded condition number. The main purpose of this chapter is to appraise the asymptotic properties of the IF algorithm such as consistency, minimax optimality and asymptotic normality. The developed theory in this chapter shows that IF algorithm has the same asymptotic rate of convergence as the MLE. In practice, the solution of IF optimization problem may also serve as the starting point (initial guess) of a likelihood maximization procedure.

Zhang [Zha04] showed that not all the covariance parameters are consistently estimable in the fixed domain regime. Strictly speaking, there is no asymptotically consistent algorithm for estimating the *non-micro ergodic* covariance parameters, which do not asymptotically affect the interpolation mean square error (see [Ste12] for a precise definition). On the other hand, it is known in the literature that subject to some mild regularity conditions, maximizing the likelihood provides a strongly consistent and asymptotically normal estimate for all the covariance parameters in the increasing domain setting [Bac14, MM84].

Increasing domain asymptotic analysis of covariance estimation has two significant benefits. First, unlike the infill asymptotic setting, the geometry of the spatial sampling has a crucial impact on the asymptotic distribution of the parameter estimate. Thus, this regime is a natural asymptotic framework for assessing the role of irregularity of spatial sampling on the covariance parameter estimation [Bac14]. This claim can be verified by a deeper look at the asymptotic distribution of the microergodic parameter estimates in the fixed domain (see e.g., [Ste12, Yin91] for MLE and [DZM$^+$09, KSN08, WL$^+$11] for tapered MLE). Another significant characteristic of increasing domain regime is that the covariance matrix has a universally bounded condition number as n grows under some mild regularity conditions. This feature of the covariance matrix plays a major role in our asymptotic analysis. Although in many geostatistical applications in a fixed bounded domain the condition number of the covariance matrix increases at least linearly with respect to n, preconditioning filters is commonly used to uniformly control the condition number independent of

n [Che13, Ste12]. Therefore, we believe that our developed increasing domain asymptotics can be useful for the fixed domain analysis of preconditioned inversion-free algorithms.

Outline of main results. In this chapter we study the increasing domain asymptotic behaviour of IF estimation algorithm introduced in [ACS16]. Specifically, suppose that G is a zero mean stationary GP in $\mathbb{R}^d$ with covariance function $\mathrm{cov}(G(s), G(s')) = R(s - s', \eta)$ in which $\eta \in \Omega$ denotes the vector of unknown covariance parameters. One realization of G has been observed on a *$d-$dimensional perturbed regular lattice* of $n = N^d$ points, which will be formally defined in Section 2.2. The specific contributions of this work are given as follow:

(a) Assuming the polynomial decay of $R(s, \eta)$ and its gradient (with respect to η) in terms of the Euclidean norm of s, and under some mild identifiability condition on R, we prove that the global maximizer of IF method consistently estimates η. Furthermore, the estimation error is of order $\sqrt{n^{-1} \ln n}$ which is shown to be minimax optimal up to some $\sqrt{\ln n}$ term.

(b) As the proposed loss function in Anitescu et al. [ACS16] is not jointly concave in η, finding its global maximizer is challenging. For a large enough sample size and under an additional condition regarding the polynomial decay of the second derivative of $R(s, \eta)$ with respect to η, we show that any *stationary point* of this non-concave program is concentrated around the true η with radius of order $\sqrt{n^{-1} \ln n}$.

(c) The asymptotic normality of the stationary points of the aforementioned algorithm will be substantiated under some mild restriction on the third derivative of R with respect to η.

Plan of the chapter. In Section 2.2, we formulate IF estimation method and precisely introduce the geometry of the sampling points. Section 2.3 expresses the necessary assumptions and studies the asymptotic properties of the estimation algorithm. Section 2.3.1 presents the convergence rate of the global and local maximizers of the optimization problem introduced in Section 2.2. We investigate the minimax optimality and the asymptotic normality of the local maximizers in Section 2.3.2. The objective of Section 2.4 is to assess the performance of IF algorithm and verify the developed theory using simulation studies on synthetic data. Section 2.5 serves as the conclusion and discusses the future directions. Section 2.6 presents the proof of the main results. Finally, the Section 2.7 contains some auxiliary technicalities on the nonasymptotic behaviour of the quadratic forms of GPs and

of large covariance matrices with polynomially decaying off-diagonal entries, which are essential in Section 2.6.

Notation. For any $m \in \mathbb{N}$, I_m and $\mathbf{0}_m$ respectively denote the m by m identity matrix and all zeros column vector of length m. Moreover, $\wedge$ and $\vee$ stand for the minimum and maximum operators. For two matrices of the same size M and M', $\langle M, M' \rangle := \sum_{i,j} M_{ij} M'_{ij}$ denotes their usual inner product. We use the following matrix norms on $M \in \mathbb{R}^{m \times n}$. For any $1 \leq p < \infty$, $\|M\|_{\ell_p} := \left(\sum_{i,j} \left|M_{ij}\right|^p\right)^{1/p}$ stands for the element-wise $p-$norm of M. $\|M\|_{2\to 2}$ represents the usual operator norm (largest singular value of M). Associated to any finite set $\mathcal{D} \subset \mathbb{R}^d$ and $s \in \mathcal{D}$, we define $\mathcal{D} - s := \{s' - s : \ s' \in \mathcal{D}\}$. We also write $\mathcal{D}(s,r) := \left\{s' \in \mathcal{D} : \ \|s' - s\|_{\ell_2} \leq r\right\}$ and $\mathcal{D}^c(s,r) = \mathcal{D} \setminus \mathcal{D}(s,r)$, for any non-negative r. $\mathcal{S}^m$ stands for the $m-$dimensional unit sphere with respect to the Euclidean norm, i.e., $\mathcal{S}^m := \left\{v \in \mathbb{R}^{m+1} : \ \|v\|_{\ell_2} = 1\right\}$. For a random sequence x_n and a deterministic positive sequence a_n, we write $x_n = \mathcal{O}_{\mathbb{P}}(a_n)$ when x_n is bounded below by a_n asymptotically, i.e., $\lim_{n\to\infty} \Pr(|x_n| \geq C a_n) = 0$ for some $C > 0$. For two sets $\Omega_1, \Omega_2 \subset \mathbb{R}^m$, $\operatorname{dist}(\Omega_1, \Omega_2) := \inf_{\omega_i \in \Omega_i,\, i=1,2} \|\omega_1 - \omega_2\|_{\ell_2}$ represents their mutual distance with respect to the Euclidean norm. Moreover, for $\mathcal{A} \subset \mathbb{R}^m$ and $r > 0$, $\mathcal{N}_r(\mathcal{A})$ denotes a subset of $\mathcal{A}$ (of minimal size) such that for any $a \in \mathcal{A}$, $\operatorname{dist}(\{a\}, \mathcal{N}_r(\mathcal{A})) \leq r$. The cardinality of such set is called the *covering number* of $\mathcal{A}$. Given spatial points $\{s_1, \ldots, s_n\} \in \mathbb{R}^d$ and the covariance function $R(\cdot, \eta)$ parametrized by $\eta = (\eta_1, \ldots, \eta_m)$, the associated covariance matrix and its derivatives are defined as

$$R_n(\eta) = \left[R\left(s_i - s_j, \eta\right)\right]_{i,j=1}^n, \quad \frac{\partial}{\partial \eta_r} R_n(\eta) = \left[\frac{\partial}{\partial \eta_r} R\left(s_i - s_j, \eta\right)\right]_{i,j=1}^n, \ \forall\, r = 1, \ldots, m.$$

The higher order derivatives can be defined in an analogous way. For two random vectors v_1 and v_2, the expression $v_1 \overset{d}{=} v_2$ means that they have the same distribution. Lastly, $D(\mathbb{P}_1 \,||\, \mathbb{P}_2)$ indicates the *Kullback-Leibler* divergence of two distributions $\mathbb{P}_i$, $i = 1, 2$.

2.2 Problem Set up and the IF Estimation Algorithm

Consider a mean zero and *stationary* (real valued) GP $G : \mathbb{R}^d \mapsto \mathbb{R}$ whose covariance function belongs to a parametric family $\mathcal{C}_{R,\Omega} := \{R(\cdot, \eta) : \eta \subset \Omega\}$. In other words, there exists $\eta_0 \in \Omega$ for which

$$\mathrm{E} G(s) G(s') = R(s - s', \eta_0), \quad \forall\, s, s' \in \mathbb{R}^d. \tag{2.1}$$

Moreover, there is $m \in \mathbb{N}$ such that Ω is a *compact* $(m+1)$ dimensional subset of $\mathbb{R}^{m+1}$ with respect to the Euclidean topology. Thus, $\mathcal{C}_{R,\Omega}$ is assumed to be a *finite dimensional* class. For analytical convenience, we consider an alternative formulation for the unknown parameters of the covariance function given as

$$\eta_0 = (\phi_0, \theta_0), \quad \phi_0 \in \mathcal{I},\ \theta_0 \in \Theta.$$

In this new representation, ϕ_0 is a strictly positive scalar denoting the variance of G and the $m-$dimensional vector θ_0 stands for the other parameters of R. Moreover, $\mathcal{I} \subset (0, \infty)$ is a bounded interval and $\Theta \subset \mathbb{R}^m$ is compact. For instance in isotropic Matern or powered exponential classes, θ_0 is a positive vector representing the *range parameter* and *fractal index*. Finally, (2.1) can be rewritten as $R(s - s', \eta_0) = \phi_0 K(s - s', \theta_0)$, in which $K(\cdot, \theta_0)$ indicates the correlation function parametrized by θ_0.

The objective is to estimate η_0 observing one realization of G at a deterministic set of spatial locations $\mathcal{D}_n = \{s_1, \ldots, s_n\} \subset \mathbb{R}^d$. It is beneficial to emphasize that our asymptotic analysis lies in the increasing domain regime in which the diameter of $\mathcal{D}_n$ tends to infinity as $n \to \infty$. The collected samples form a zero mean Gaussian column vector $Y = [G(s_1), \ldots, G(s_n)]^\top$ of length n. Before proceeding further, let us precisely introduce the geometric structure of $\mathcal{D}_n$.

Assumption 2.1. Suppose that there is $N \in \mathbb{N}$ such that $n = N^d$. There exists $\delta \in [0, 1/2)$ for which $\mathcal{D}_n$ is a *$d-$dimensional $\delta-$perturbed regular lattice* (with unit grid size). Namely,

$$\mathcal{D}_n = \left\{ v_i + \delta p_i :\ v_i \in \mathcal{V}_{N,d},\ p_i \in [-1, 1]^d \right\}_{i=1}^n,$$

in which $\mathcal{V}_{N,d} \coloneqq \{v_1, \ldots, v_n\} = \{1, \ldots, N\}^d$ denotes the $d-$dimensional regular lattice.

The condition $\delta \in [0, 1/2)$ guarantees the existence of a strictly positive minimum distance $(1 - 2\delta)$ between the distinct points in $\mathcal{D}_n$. The scalar δ quantifies the amount of irregularity in $\mathcal{D}_n$. In the case of $\delta = 0$, $\mathcal{D}_n$ forms a regular lattice and the irregularity can be more apparent as δ increases. Although the absence of randomness in Assumption 2.1 may appear problematic at first sight, our theoretical contributions are not restricted to any further set of strong conditions on p_i's. For instance, the presented results in the next section hold almost surely if p_i's are independent (or even dependent) draws of a distribution supported on $[-1, 1]^d$ which is absolute continuous with respect to the Lebesgue measure..

Now we present IF estimation algorithm introduced in [ACS16]. Define,

$$\hat{\eta}_n = \underset{\eta \in \Omega}{\arg\max}\, F_n(Y, \eta), \ \text{ where } \ F_n(Y, \eta) \coloneqq \frac{1}{n} \left\{ Y^\top R_n(\eta)\, Y - \frac{1}{2} \|R_n(\eta)\|_{\ell_2}^2 \right\}. \tag{2.2}$$

Note that $F_n(Y,\eta)$ does not depend on the Cholesky factorization of $R_n(\eta)$ and regardless of the choice of covariance function, it can be evaluated in $\mathcal{O}\left(n^2\right)$ operations, even for irregularly spaced samples which is an improvement over the conventional likelihood function. The optimization algorithm in (2.2) can be reformulated as

$$\begin{aligned}\left(\hat{\phi}_n, \hat{\theta}_n\right) &= \arg\max_{(\phi,\theta)\in\mathcal{I}\times\Theta} F_n(Y,\phi,\theta)\,,\\ \text{where}\;\; F_n(Y,\phi,\theta) &\coloneqq \tfrac{1}{n}\left\{\phi Y^\top K_n(\theta)\, Y - \tfrac{\phi^2}{2}\,\|K_n(\theta)\|_{\ell_2}^2\right\}.\end{aligned} \tag{2.3}$$

Despite the fact that $F_n(Y,\phi,\theta)$ has a simple quadratic (concave) form of ϕ, its dependence to θ is fairly complicated . For instance F_n is not a concave function of θ even for the classic case of isotropic exponential covariance. So, accurate approximation of its global maximizer can be computationally expensive.

Remark 2.1. We conclude this section mentioning two characteristics of $F_n(Y,\phi,\theta)$ that can provide a theoretical clue for generalizing IF loss function to a broader class of inversion-free losses. The first property is also critical for the theoretical analysis in the next section.

1. As stated in [ACS16], the true parameter η_0 is a stationary point of the expected value of $F_n(Y,\eta)$. That is,

$$\mathrm{E}\left\{\frac{\partial}{\partial\eta_j} F_n(Y,\eta)\,|_{\eta=\eta_0}\right\} = 0, \quad \forall\; j = 1,\ldots,(m+1)\,.$$

   Roughly speaking, η_0 is located in a small neighborhood of a stationary point of (2.2) if the gradient of $F_n(Y,\eta)$ is smooth enough and concentrated around its expected value. The next fact, which has not been stated in [ACS16], reveals a profound connection of (2.2) to MLE.

2. Define $H_n(Y,\eta) \coloneqq F_n(Y,\eta) - F_n(Y,\eta_0)$ and $L_n(\eta,\eta_0) \coloneqq R_n^{1/2}(\eta_0)\, R_n^{-1}(\eta)\, R_n^{1/2}(\eta_0)$. Also, let $\tilde{\eta}_n$ denotes the MLE of η. Obvious calculations lead to

$$\begin{aligned}\hat{\eta}_n &= \arg\max_{\eta\in\Omega} H_n(Y,\eta)\,,\\ \tilde{\eta}_n &= \arg\max_{\eta\in\Omega} H'_n(Y,\eta)\,,\\ \text{where}\;\; H'_n(Y,\eta) &\coloneqq \frac{1}{n}\Big\{-\log\det L_n(\eta,\eta_0) - n + Y^\top R_n^{-1}(\eta)\, Y\Big\}.\end{aligned}$$

   Notice that $\mathrm{E}H_n(Y,\eta_0) = \mathrm{E}H'_n(Y,\eta_0) = 0$. Under Assumption 2.1 and using similar tech-

niques as Section 2.7, one can guarantee the existence of a scalar $C \in (0, \infty)$ such that

$$\mathrm{E}H_n(Y,\eta) \le C\,\mathrm{E}H'_n(Y,\eta), \quad \forall \eta \in \Omega. \tag{2.4}$$

Namely, in the increasing domain regime, the objective function proposed in [ACS16] can be viewed as an approximate *minorizing surrogate* of the likelihood function in the expected value sense (it forms a perfect minorizer whenever $C = 1$ in (2.4)).

2.3 Main Results

We establish the asymptotic characteristics of the estimation algorithm in (2.2). Section 2.3.1 examines the consistency of the global maximizer and the stationary points of (2.2) under some sufficient conditions on Ω and the correlation function $K(\cdot,\theta)$. The near minimax optimality and the asymptotic normality of the stationary points will be covered in Section 2.3.2

2.3.1 Consistency and the Convergence Rate

The following assumptions are assumed on the parameter space $\Omega = \mathcal{I} \times \Theta$ and the correlation function $K(\cdot,\theta)$ for studying the asymptotic behaviour of the global maximizer of (2.2). Similar but slightly stronger conditions have been used in [Bac14] for the increasing domain asymptotic analysis of MLE.

Assumption 2.2. The following conditions are satisfied by Ω and K.

(A1) Θ and $\mathcal{I}$ are *compact connected* subsets of $\mathbb{R}^m$ and $(0,\infty)$, respectively.

(A2) There are bounded scalars $M > 0$ and $r_1 > 1$ such that for any $s \in \mathcal{D}_n$,

$$\max_{s' \in \mathcal{D}_n(s,r_1)} \left|K(s'-s,\theta_2) - K(s'-s,\theta_1)\right| \ge M\|\theta_2 - \theta_1\|_{\ell_2}, \quad \forall\, \theta_1,\theta_2 \in \Theta. \tag{2.5}$$

(A3) For some $q \in \{1,2,3\}$, there exists a positive scalar $C_{K,\Theta}$ such that

$$\max_{\theta\in\Theta}\left(|K(s,\theta)| \vee \left|\frac{\partial}{\partial\theta_{j_1}} \dots \frac{\partial}{\partial\theta_{j_q}} K(s,\theta)\right|\right) \le \frac{C_{K,\Theta}}{1+\|s\|_{\ell_2}^{d+1}}, \quad \forall\, s \in \mathbb{R}^m,$$

for any $j_1,\dots,j_q \in \{1,\dots,m\}$.

Condition $(A2)$, assuring the *identifiability* of θ from the K, holds for the standard class of correlation functions such as Matern, powered exponential and rational quadratic. A

detailed look at $(A2)$ is postponed to the end of this section. Before commenting on $(A3)$, let us define the family of geometric anisotropic covariance functions.

Definition 2.1. Let $G : \mathbb{R}^d \mapsto \mathbb{R}$ be a zero mean stationary GP in $\mathbb{R}^d$. Then G is called *geometric anisotropic* if

$$R(s-s',\eta_0) \coloneqq \mathrm{E}G(s)G(s') = \phi_0 K\left(\sqrt{(s-s')^\top A_0 (s-s')}\right), \quad \forall\ s,s' \in \mathbb{R}^d, \tag{2.6}$$

for $\phi_0 > 0$, *symmetric positive definite* matrix $A_0 \in \mathbb{R}^{d\times d}$, $\eta_0 = (\phi_0, A_0)$ and a correlation function K. Specifically if $A_0 = \theta_0^{-1} I_d$ for some strictly positive θ_0, then G is said to be an *isotropic* GP.

For geometric anisotropic processes, K is either assumed to be a fully known function (in this case $\eta_0 = (\phi_0, A_0)$ in which $\phi_0 \in \mathcal{I}$ and $A_0 \in \Theta$, denotes the unknown parameters of covariance function) or known up to some strictly positive scalar ν_0, usually refers to as the *fractal index*. In the latter case, $\eta_0 = \{\phi_0, \theta_0 = (A_0, \nu_0)\}$. Now, we mention some commonly used class of covariance functions, with unknown fractal index, satisfying $(A3)$ with $q = 1$ (appearing in the statement of the first main result in this section). It is supposed in the following Remark that

$$\Lambda_{\min,\Theta} \le \min_{A_0\in\Theta} \frac{1}{\left\|A_0^{-1}\right\|_{2\to 2}} \le \max_{A_0\in\Theta} \|A_0\|_{2\to 2} \le \Lambda_{\max,\Theta}, \tag{2.7}$$

for strictly positive and bounded scalars $\Lambda_{\min,\Theta}$ and $\Lambda_{\max,\Theta}$. Namely, all eigenvalues of A_0 are universally bounded away from zero and infinity.

Remark 2.2. Any compactly supported correlation function, such as *spherical or Wendland family* [Wen95] on $\mathbb{R}^d$ trivially admits $(A3)$. Assumption $(A3)$ with $q = 1$ also holds for some classical families of geometric anisotropic covariances such as:

(a) *Matern:* The GP G has Matern covariance function if it fulfills (2.6) with

$$K(r) = \frac{2^{1-\nu_0}}{\Gamma(\nu_0)} r^{\nu_0} \mathcal{K}_{\nu_0}(r), \tag{2.8}$$

in which ν_0 is an unknown, strictly positive scalar lies in a compact space. Moreover, $\Gamma(\cdot)$ and $\mathcal{K}_{\nu_0}(\cdot)$ represent the Gamma function and the modified Bessel function of the second kind, respectively. The parametric Matern family satisfies $(A3)$ provided condition (2.7).

(b) *Powered exponential:* A covariance function in this class satisfies (2.6) with $K(r) = e^{-r^{\nu_0}}$ and $\nu_0 \in (0,2)$. Like Matern class, assuming (2.7), any member of powered exponential family fulfills $(A3)$ with $q = 1$.

(c) *Rational quadratic:* The elements of this class are of the form (2.6) with $K(r) = \left(1+r^2\right)^{-\left(\frac{d}{2}+\nu_0\right)}$ and $\nu_0 > 0$. For the case of known fractal index, $(A3)$ with $q = 1$ is valid, if (2.7) holds. Note that for unknown ν_0 the exact same statement is satisfied under a slightly stronger condition of $\nu_0 > 1/2$

Parts (b) and (c) of Remark 2.2 are verifiable by straightforward algebra and differentiation rules. In order to demonstrate part (a), see [AS$^+$66] for the derivative properties of the Bessel function of the second kind (with respect to the entries of A_0) and see Lemma 2.5 for the asymptotic behaviour of the partial derivatives of the Matern covariance with respect to ν_0. Now, we state the first significant result of this section regarding the consistency of the global maximizer of (2.3) under Assumption 2.2 and perturbed regular lattice sampling.

Theorem 2.1. Suppose that Assumptions 2.1 and 2.2 with $q = 1$ hold for $\mathcal{D}_n$, Ω and K. Then the maximizer of (2.3) satisfies

$$\Pr\left(\left\|\hat{\theta}_n - \theta_0\right\|_{\ell_2} \vee \left|\frac{\hat{\phi}_n}{\phi_0} - 1\right| \geq C\sqrt{\frac{\ln n}{n}}\right) \to 0, \quad \text{as } n \to \infty, \tag{2.9}$$

for some constant C (which depends on $\mathcal{D}_n$, Ω and K).

Remark 2.3. Let $\phi_{\min}$ and $\phi_{\max}$ denote the smallest and largest element in $\mathcal{I}$. Obviously, $\phi_{\min}$ and $\phi_{\max}$ are well defined and finite due to $(A1)$. Moreover,

$$\begin{aligned}(1 \wedge \phi_{\min})\left(\left\|\hat{\theta}_n - \theta_0\right\|_{\ell_2} \vee \left|\frac{\hat{\phi}_n}{\phi_0} - 1\right|\right) &\leq \|\hat{\eta}_n - \eta_0\|_{\ell_2} \\ &\leq \sqrt{1+\phi_{\max}^2}\left(\left\|\hat{\theta}_n - \theta_0\right\|_{\ell_2} \vee \left|\frac{\hat{\phi}_n}{\phi_0} - 1\right|\right).\end{aligned}$$

Thus (2.9) is a stronger statement than $\|\hat{\eta}_n - \eta_0\|_{\ell_2} = \mathcal{O}_{\mathbb{P}}\left(\sqrt{n^{-1}\ln n}\right)$ and they are equivalent when $\phi_{\min} > 0$ (which is true under $A1$).

An analogous consistency result has been proved recently by Bachoc [Bac14] for the MLE and cross validation estimator. Based upon Theorem 2.1, the asymptotic rate of IF algorithm has not been sacrificed for increasing the speed and memory efficiency comparing to the MLE.

Finally, we concisely address the role of the identifiability condition $(A2)$ in Theorem 2.1. Actually, $(A2)$ plays a decisive role in translating consistent estimation of the correla-

tion matrix (in the relative sense) to η_0. Strictly speaking, $(A2)$ is required to deduce (2.9) from the probabilistic statement

$$\frac{1}{\sqrt{n}}\left\|K_n(\hat{\theta}_n)-K_n(\theta_0)\right\|_{\ell_2}=\mathcal{O}_{\mathbb{P}}\left(\sqrt{n^{-1}\ln n}\right).$$

The rest of this section is devoted to the analysis of the stationary points of (2.3). Solving the unique root of the derivative of $F_n(Y,\phi,\theta)$ with respect to ϕ, yields a closed form formula for $\hat{\phi}_n$ in terms of $\hat{\theta}_n$, Y and the correlation function, namely

$$\hat{\phi}_n=\frac{Y^\top K_n(\hat{\theta}_n)Y}{\left\|K_n(\hat{\theta}_n)\right\|_{\ell_2}^2}. \tag{2.10}$$

Moreover, $\hat{\theta}_n$ can be obtained using

$$\hat{\theta}_n=\underset{\theta\in\Theta}{\arg\max}\ G_n(Y,\theta),\quad \text{where}\quad G_n(Y,\theta)=\frac{Y^\top K_n(\theta)Y}{\|K_n(\theta)\|_{\ell_2}}. \tag{2.11}$$

Note that for large n, computing the global maximizer of (2.11) can be less intensive than (2.2) due to searching over a smaller space Θ. We first visually assess the key properties of F_n in some simple scenarios. In Figure 2.1, $G_n(Y,\theta)$ (which is a univariate function of scalar θ) has been plotted versus θ for the two dimensional ($d=2$) isotropic Matern covariance function in two different scenarios. In the left panel the isotropic GP G has been generated with the parameters $(\phi_0,\theta_0,\nu_0)=(1,4,0.5)$ and has been sampled in a randomly perturbed regular lattice with $\delta=0.2$ and of size $N=100$. In the right panel, the covariance parameters are given by $(\phi_0,\theta_0,\nu_0)=(1,6,1.5)$ and the GP is sampled at a randomly generated perturbed regular lattice with $N=100$ and $\delta=0.2$. As is apparent from Figure 2.1, for these two parsimonious scenarios $G_n(Y,\theta)$ is not a concave function of θ and has a single *inflection point*. However, $G_n(Y,\theta)$ has only one stationary point which coincides with its global maximizer. In the following, we rigorously study the large sample behaviour of the stationary points of $G_n(Y,\theta)$ (as well as F_n) in a generic case. We initiate our analysis by stating the sufficient conditions on K and Ω.

Assumption 2.3. $(A1)$ holds for Ω, and K fulfills $(A2)$ and $(A3)$ with $q=2$ in Assumption 2.2.

Remark 2.4. The analysis of the stationary points of (2.3) requires a slightly stronger conditions than that of the global maximizer in Assumption 2.2. The main distinction is the polynomial decay of the second order derivative of K with respect to θ. Note that the new condition on the second derivative of K is not too restrictive. For instance the same

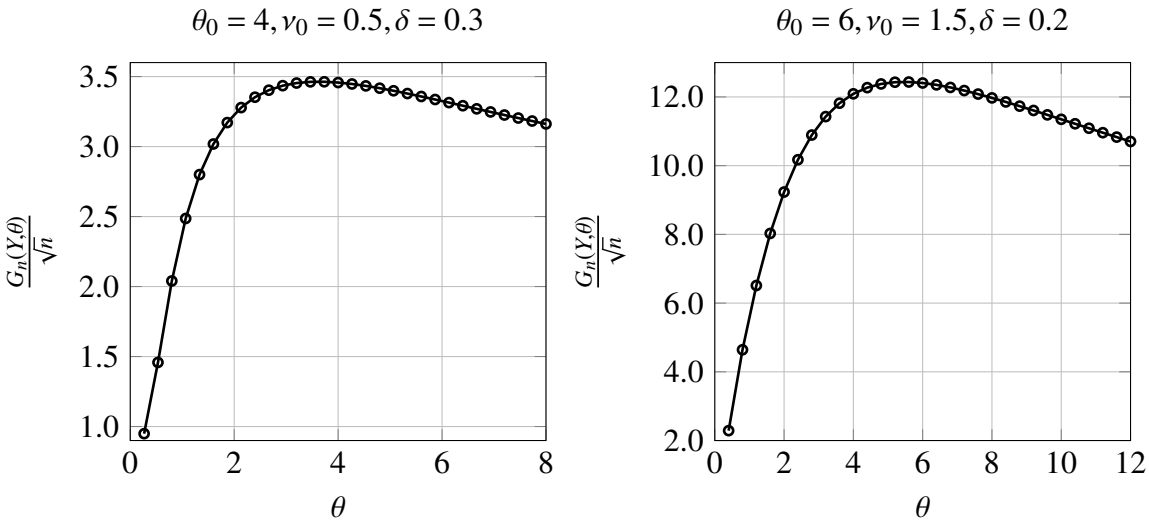

Figure 2.1: The above figures exhibit $n^{-1/2}G_n(Y,\theta)$ for the isotropic Matern covariance function (with known ν_0). In the left panel, $\theta_0 = 4$, $\nu_0 = 0.5$ and the spatial samples form a two dimensional randomly perturbed regular lattice of size $N = 100$ with $\delta = 0.3$. In the right panel, $\theta_0 = 6$, $\nu_0 = 1.5$ and $\mathcal{D}_n$ is a randomly chosen two dimensional perturbed regular lattice with $N = 100$ and $\delta = 0.3$.

analysis as Remark 2.2 validates this condition for all covariance families introduced in Remark 2.2 (with a larger constant $C_{K,\Theta}$).

Theorem 2.2. Suppose that $\mathcal{D}_n$ admits Assumption 2.1 and Assumption 2.3 holds for Ω and K. Then any stationary point of the optimization problem (2.3) satisfies

$$\lim_{n\to\infty} \Pr\left(\left\| \hat{\theta}_n - \theta_0 \right\|_{\ell_2} \vee \left| \frac{\hat{\phi}_n}{\phi_0} - 1 \right| \geq C\sqrt{\frac{\ln n}{n}} \right) = 0, \tag{2.12}$$

for an appropriately chosen constant $C > 0$ depending on $\mathcal{D}_n$, Ω and K.

Theorem 2.2 shows that any stationary point of F_n is concentrated in a small neighborhood of (ϕ_0, θ_0), with high probability. In other words, $F_n(Y,\phi,\theta)$ shows a similar behaviour as Figure 3.1 in the general case. In addition, the comparison between (2.9) and (2.12) reveals that stationary points converge to (ϕ_0, θ_0) with the same rate as the global maximizer.

We conclude this section by illustrating how restrictive the identifiability assumption $(A2)$ can be for the frequently used classes of the covariance functions. We first introduce a slightly stronger identifiability condition than $(A2)$, which will be referred to as $(A4)$. Note that a slightly modified version of $(A4)$ has been first introduced in [Bac14] for studying the increasing domain asymptotics of the maximum likelihood and cross validation algorithms. That is to say, these identifiability conditions are not exclusive to IF method and pop up

in the asymptotic analysis of other algorithms. The proof of the subsequent results in this section will be omitted. We refer the reader to [KSN16] for a detailed proof.

Proposition 2.1. ($A2$) is satisfied, whenever

(A4) (a) There are positive scalars $r_2 > 1$ and M_2 such that for any $\eta \in \Omega$ and $\lambda \in \mathcal{S}^m$,

$$\min_{s\in\mathcal{D}_n} \max_{s'\in\mathcal{D}_n(s,r_2)} \left| \sum_{j=1}^{m+1} \lambda_j \frac{\partial}{\partial \eta_j} R(s-s',\eta) \right| \geq M_2.$$

(b) The following inequality holds for any distinct pair of points $\eta_1, \eta_2 \in \Omega$

$$\min_{s\in\mathcal{D}_n} \max_{s'\in\mathcal{D}_n(s,r_2)} \left| R(s-s',\eta_2) - R(s-s',\eta_1) \right| > 0.$$

Clearly, ($A4.b$) is necessary for any algorithm consistently estimating η and it can be verified for all typical classes of geometrical anisotropic covariance function. However, understanding the role and restrictiveness of ($A4.a$) is more subtle than that of ($A4.b$). Note that unlike ($A3$), all the introduced identifiability conditions not only depend on the choice of the covariance function but also to the observed locations $\mathcal{D}_n$. It may be excessive to seek the class of covariances satisfying ($A4.a$) for any perturbed lattice $\mathcal{D}_n$. So, a more pertinent question is: which class of covariance functions do almost surely satisfy ($A4.a$) for a randomly generated perturbed lattice? The following result responds to question by rigorously characterizing a broad subclass of the geometrically anisotropic covariances (as defined in Definition 2.1) fulfilling ($A4.a$).

Proposition 2.2. Let P be a distribution in $[-1,1]^d$ which is absolutely continuous with respect to the Lebesgue measure. Suppose that $R(\cdot,\eta) : \mathbb{R}^d \mapsto [0,\infty)$ is a geometrically anisotropic covariance function with a known ν_0 (if exists). Then, ($A4$) almost surely holds if

(a) $K : [0,\infty) \mapsto [0,\infty)$ is a nonzero, differentiable and strictly decreasing function (K may only have right derivative at zero).

(b) $\mathcal{D}_n$ is a randomly generated $\delta-$perturbed lattice associated to P. That is, p_i are independent draws of P in Assumption 2.1.

Corollary 2.1. Let $\mathcal{D}_n$ be $d-$dimensional regular lattice (associated to $\delta = 0$) and assume that $R(\cdot,\eta) : \mathbb{R}^d \mapsto [0,\infty)$ is a geometrically anisotropic covariance function with known ν_0. Then, ($A4$) holds if $K : [0,\infty) \mapsto [0,\infty)$ admits the condition (a) in Proposition 2.2.

Although the conditions of Proposition 2.2 trivially hold for the non-compactly supported covariance function introduced in Remark 2.2, deploying analogous proof techniques can lead to a similar result for compactly supported covariance function.

Proposition 2.3. Suppose that P, $R(\cdot,\eta)$ and $\mathcal{D}_n$ satisfy the same conditions as Proposition 2.2. Then, $(A4)$ almost surely holds if there exists a large enough positive scalar r_0 for which

- $K : [0,\infty) \mapsto [0,\infty)$ is a nonzero, differentiable and strictly decreasing function in the interval $[0, r_0]$ and $K(r) = 0$ for any $r > r_0$.

Although the required conditions on the covariance function's formulation, in Propositions 2.2 and 2.3, are very minimal, we assume that the fractal index ν_0 (if exists) is fully known. However in the following result, ν_0 is one of the unknown parameters to be estimated. Here the central emphasis is on the powered exponential and rational quadratic classes, as their partial derivative with respect to the fractal index have a somewhat simple closed form that can be handled without great difficulty.

Proposition 2.4. Let P be a distribution in $[-1,1]^d$ which is absolutely continuous with respect to the Lebesgue measure. Suppose that $R(\cdot,\eta) : \mathbb{R}^d \mapsto [0,\infty)$ is a geometrically anisotropic covariance function. Then, $(A4)$ almost surely holds if

(a) $K : [0,\infty) \mapsto [0,\infty)$ is either a powered exponential or rational quadratic covariance functions in Remark 2.2 with unknown $\nu_0 > 0$.

(b) $\mathcal{D}_n$ is a randomly generated $\delta-$perturbed lattice associated to P. That is, p_i are independent draws of P in Assumption 2.1.

Remark 2.5. A prudent look at the proof of Proposition 2.4 reveals that the following property (which is satisfied by the powered exponential and rational quadratic families) has the crucial role.

$$\frac{\partial K}{\partial \nu}\Big|_{\{r=\mathbf{0}_d,\,\theta\}} = 0, \qquad \forall\, \theta \in \Theta. \tag{2.13}$$

For the Matern class, not only $\frac{\partial K}{\partial \nu}$ not satisfy (2.13), it does not have a tractable algebraic form. We believe that $(A4)$ holds true for the geometric anisotropic Matern family with unknown ν, even though it is beyond the reach of our current proof technique.

2.3.2 Minimax Optimality and Asymptotic Normality

Now, we further investigate the asymptotic statistical properties of IF algorithm. Near minimax optimality and asymptotic normality are respectively presented in Theorems 2.3 and 2.4.

Theorem 2.3. Suppose that Assumptions 2.1 and 2.2 hold for $\mathcal{D}_n$, Ω and K. Then there exist $n_0 \in \mathbb{N}$ and a bounded scalar $C > 0$ such that

$$\sup_{\eta_0 \in \Omega} \Pr\left(\|\hat{\eta}_n - \eta_0\|_{\ell_2} \geq \frac{C}{\sqrt{n}} \right) \geq \frac{1}{4},$$

for any estimator $\hat{\eta}_n$ and any $n \geq n_0$.

Theorem 2.3 reveals that the established bounds in Theorems 2.1 and 2.2 are sharp up to order $\sqrt{\ln n}$. This means that for the perturbed regular lattice sampling scheme, no algorithm can achieve a significantly better rate than IF method.

Theorem 2.4. Suppose that $\mathcal{D}_n$ is a perturbed lattice introduced in Assumption 2.1. Furthermore, $(A1)$, $(A3)$ with $q = 3$ and $(A4)$ are fulfilled by Ω and R. There is a positive definite matrix $\Sigma \in \mathbb{R}^{(m+1)\times(m+1)}$ with bounded operator norm such that

$$\sqrt{n}\,(\hat{\eta}_n - \eta_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}_{m+1}, \Sigma). \tag{2.14}$$

The exact formulation of Σ has been omitted in this section due to its complicated algebraic form. We refer the reader to the proof of Theorem 2.4 in Section 2.6 for further details. It is worthwhile to mention that the entries of Σ heavily depend on the configuration of points in $\mathcal{D}_n$, which is a major disparity between fixed and increasing domain asymptotics. Comparing Theorems 2.2 and 2.4, here we impose a slightly stronger differentiability condition (polynomially decaying of the third derivative) for establishing asymptotic normality. This condition has been formerly introduced in [Bac14] and holds for the geometrically anisotropic covariances in Remark 2.2.

2.4 Simulation Studies

The relatively large-scaled numerical studies in this section give a fairly comprehensive appraisal of the statistical and computational performance of the optimization problem (2.3). Despite the popularity of *R* language, running the iterative programs such as loops in R is much slower than that of C++ (around 250 times slower according to some studies [AFV14]). Taking advantage of the *Rcpp* package and hybrid programming techniques in R can considerably expedite the execution time (up to 50 times in our simulation studies). In order to get the maximum speed, the *open MP* application programming interface has been used to exploit the multi-threaded programing technology. All the numerical experiments in this section have been executed on 12 processors, except for the second simulation study ($n = 10^6$) which has been implemented on 60 cores.

Generating high dimensional samples from a GP on an irregularly spaced grid is the foremost challenge that we confronted in our synthetic data simulations. Applying the traditional method based upon the Cholesky decomposition of the covariance matrix is almost infeasible in the case that $n \approx 10^5$ or larger. Hence, we use the considerably faster spectral method (pp. 203 – 205, [Cre15]) for generating stationary Gaussian processes. For completeness, this algorithm will be concisely presented here. Strictly speaking, the objective is to simulate a real valued zero mean stationary GP G in $\mathbb{R}^d$ with the covariance function $\phi_0 K(\cdot, \theta_0)$ over a $\delta-$perturbed lattice $\mathcal{D}_n = \{s_1, \ldots, s_n\}$. For the purpose of generating a realization of G on a perturbed grid, without loss of generality we can assume that the samples are all of unit variance, i.e., $\phi_0 = 1$. We also assume that G is geometric anisotropic. Recalling from Definition 2.6, there is a symmetric positive definite matrix $B_0 \in \mathbb{R}^{d\times d}$ which represents the symmetric square root of A_0, such that $K(r, \theta_0) = K\left(\|B_0 r\|_{\ell_2}\right)$. Throughout this section $d = 2$ and K is either the Matern or rational quadratic covariance function which have been previously introduced in Remark 2.2.

Let $p \in \mathbb{N}$ be a large enough number and $\{\xi_k\}_{k=1}^p$ be i.i.d. uniform random variables on $[-\pi, \pi]$. Let $\hat{K} : \mathbb{R}^d \mapsto \mathbb{R}$ denotes the spectral density of G defined by

$$\hat{K}(\omega) \coloneqq (2\pi)^{-d} \int_{\mathbb{R}^d} K(r, \theta_0) \cos(\langle \omega, r\rangle)\, dr = (2\pi)^{-d} \int_{\mathbb{R}^d} K\left(\|B_0 r\|_{\ell_2}\right) \cos(\langle \omega, r\rangle)\, dr.$$

The non-negative mapping $\hat{K}(\cdot)$ is a density function in $\mathbb{R}^d$ (since it integrates to $K(0, \theta_0) = 1$). Furthermore, let $\{\omega_k\}_{k=1}^N$ be independent draws from the density $\hat{K}(\cdot)$. Now, define

$$G(s) = \sqrt{\frac{2}{p}} \sum_{k=1}^{p} \cos(\langle \omega_k, s\rangle + \xi_k), \quad \forall\ s \in \mathbb{R}^d. \tag{2.15}$$

It is known that G is an anisotropic process with $\operatorname{cov}\{G(s), G(s')\} = K\left(\|B_0(s - s')\|_{\ell_2}\right)$ for any pair $s, s' \in \mathbb{R}^d$ (p. 204, [Cre15]), converging in distribution to a GP as p tends to infinity. Next, we explain how to generate the random variables $\{\omega_k\}_{k=1}^p$. The following fact which can be proved using the integration by substitution plays a principal role in our algorithm.

Remark 2.6. Let $\omega' \in \mathbb{R}^d$ be a draw from the following density function

$$\hat{K}_I(u) = (2\pi)^{-d} \int_{\mathbb{R}^d} K\left(\|r\|_{\ell_2}\right) \cos(\langle u, r\rangle)\, dr.$$

Then ω and $B_0\omega'$ have the same distribution, i.e., $\omega \overset{d}{=} B_0\omega'$. Note that $\hat{K}_I$ is an isotropic function. Namely, there is a function $\Phi : \mathbb{R} \mapsto [0, \infty)$ for which $\hat{K}_I(u) = \Phi\left(\|u\|_{\ell_2}\right)$. Moreover $\omega' \overset{d}{=} \mathfrak{r}\psi_d / \|\psi_d\|_{\ell_2}$ in which ψ_d is a standard $d-$ dimensional Gaussian vector and $\mathfrak{r}$ is a non-

negative random variable with the density function

$$f_{\mathfrak{r}}(r) = \frac{d\pi^{d/2}}{\Gamma(d/2+1)} r^{d-1}\Phi(r) = \frac{2\pi^{d/2}}{\Gamma(d/2)} r^{d-1}\Phi(r).$$

For instance in the case of $d = 2$, we have $f_{\mathfrak{r}}(r) = 2\pi r\Phi(r)$. Hence, $\omega \overset{d}{=} \mathfrak{r}B_0\psi_d / \|\psi_d\|_{\ell_2}$.

For Matern covariance function in two dimensional plane ($d = 2$), generating independent samples of the random variable $\mathfrak{r}$ is a straightforward task. In this case

$$\begin{aligned} f_{\mathfrak{r}}(r) &= \frac{2\pi^{d/2}}{\Gamma(d/2)} r^{d-1}\Phi(r) = \frac{2\pi^{d/2}}{\Gamma(d/2)} r^{d-1} \frac{\pi^{-d/2}\Gamma(d/2+\nu)}{\Gamma(\nu)} \left(1+r^2\right)^{-(\nu+d/2)} \\ &= \frac{2r\nu}{(1+r^2)^{1+\nu}}. \end{aligned}$$

Thus the cumulative distribution is of the form $\Pr(\mathfrak{r} \le r) := F_{\mathfrak{r}}(r) = 1 - \left(1+r^2\right)^{-\nu}$. So

$$\mathfrak{r} \overset{d}{=} F_{\mathfrak{r}}^{-1}(u) = \sqrt{1-(1-u)^{-1/\nu}},$$

in which u is a uniform random variable in $[0,1]$. One can find a closed from expression for $\mathfrak{r}$ in terms of u for the rational quadratic covariance function, in the case that $(\tau + 1/2) \in \mathbb{N}$ (Recall τ from Remark 2.2). In this case, $\Phi(\cdot)$ has a form of the Matern covariance function (2.8) (with different constants) due to the duality principle of the Fourier transform.

Throughout this section, G is assumed to be a zero mean GP in $\mathbb{R}^2$, whose covariance function is a member of either Matern or rational quadratic families with a known fractal index. In the first experiment, G is an isotropic spatial process. In other words, we set $A_0 = \theta_0^{-2} I_2$ in the Definition 2.6. θ_0 is a strictly positive scalar known as the *range parameter*. Furthermore, $\mathcal{D}_n$ is a randomly generated $\delta-$perturbed lattice of size 320^2, i.e., $n = 102400 \approx 10^5$. The approximated realizations of G are generated using (2.15) with $p = 1.5 \times 10^5$. To investigate the role of spatial irregularity in the computational and statistical performance of IF algorithm, we vary δ in the set $\{0.1, 0.3\}$. The range parameter and the standard deviation, which is represented by $\sigma_0 = \sqrt{\phi_0}$, are respectively estimated solving the optimization problem (2.11) and closed form formula (2.10). The range parameter space is chosen as $\Theta = [0.1, 15]$. The single variable constrained optimization problem (2.11) is solved using the *optimize* function in *R* ,which exploits a combination of golden section search and successive parabolic interpolation. We stop the iteration of the solver when the relative change in the objective is below 10^{-3}. Table 2.1 displays the summary of our first simulation study.

		$\delta = 0.1$		$\delta = 0.3$	
Matern	$\nu_0 = 0.5$	$\eta_0 = (1,4)$ $\hat{\eta} = (0.993, 4.420)$	$\eta_0 = (1,7)$ $\hat{\eta} = (0.978, 7.565)$	$\eta_0 = (1,4)$ $\hat{\eta} = (1.005, 3.941)$	$\eta_0 = (1,7)$ $\hat{\eta} = (1.028, 6.553)$
	$\nu_0 = 1.5$	$\eta_0 = (1,4)$ $\hat{\eta} = (0.973, 3.618)$	$\eta_0 = (1,7)$ $\hat{\eta} = (1.023, 6.568)$	$\eta_0 = (1,4)$ $\hat{\eta} = (1.026, 4.343)$	$\eta_0 = (1,7)$ $\hat{\eta} = (1.005, 7.102)$
	$\nu_0 = 2.5$	$\eta_0 = (1,4)$ $\hat{\eta} = (1.014, 4.186)$	$\eta_0 = (1,7)$ $\hat{\eta} = (1.010, 7.428)$	$\eta_0 = (1,4)$ $\hat{\eta} = (1.044, 4.037)$	$\eta_0 = (1,7)$ $\hat{\eta} = (0.993, 6.505)$
Rational quadratic	$\nu_0 = 0.5$	$\eta_0 = (1,4)$ $\hat{\eta} = (1.017, 4.100)$	$\eta_0 = (1,7)$ $\hat{\eta} = (0.976, 7.415)$	$\eta_0 = (1,4)$ $\hat{\eta} = (1.013, 3.916)$	$\eta_0 = (1,7)$ $\hat{\eta} = (1.001, 6.639)$
	$\nu_0 = 1.5$	$\eta_0 = (1,4)$ $\hat{\eta} = (1.000, 4.001)$	$\eta_0 = (1,7)$ $\hat{\eta} = (1.014, 7.210)$	$\eta_0 = (1,4)$ $\hat{\eta} = (1.017, 3.920)$	$\eta_0 = (1,7)$ $\hat{\eta} = (0.994, 7.063)$

Table 2.1: Estimation of $\eta_0 = (\sigma_0, \theta_0)$ for the isotropic Matern and rational quadratic covariance functions, where $\mathcal{D}_n$ is a perturbed lattice of size 320^2 with $\delta \in \{0.1, 0.3\}$.

The required CPU times for the numerical experiments in Table 2.1 are approximately 30 and 60 minutes for the rational quadratic and Matern kernels, respectively. However evaluating the full MLE for a such large sample size is intractable. As is apparent from Table 2.1, the estimated parameters, $\hat{\eta}$, are in a close neighborhood of η_0. Moreover, estimating σ_0 has a significantly higher precision than that of the range parameters, since as the distant samples in $\mathcal{D}_n$ carry negligible information about θ_0. Lastly, the condition number of the covariance matrix increases with the value of range parameter, leading to a higher estimation error $\|\eta_0 - \hat{\eta}\|_{\ell_2}$ for larger θ_0. In the second simulation study which has the same set up as the first experiment, $\mathcal{D}_n$ is a irregular grid of size $1000^2 = 10^6$. We also set $p = 5 \times 10^5$ in (2.15). Table 2.2 encapsulates the results of this experiment. The evaluation of $\hat{\eta}$ for this very high dimensional numerical study takes 8 hours on 60 cores with $4GB$ RAM.

	$\delta = 0.1,\ \nu_0 = 0.5$	$\delta = 0.3,\ \nu_0 = 1.5$
Rational quadratic	$\eta_0 = (1,4)$ $\hat{\eta} = (1.004, 4.053)$	$\eta_0 = (1,7)$ $\hat{\eta} = (0.999, 6.948)$

Table 2.2: Estimation of $\eta_0 = (\sigma_0, \theta_0)$ for the isotropic rational quadratic covariance functions, where $\mathcal{D}_n$ is a perturbed lattice of size 1000^2 with $\delta \in \{0.1, 0.3\}$.

In the next set of experiments featuring GPs with isotropic covariance functions, we set $\mathcal{D}_n$ to be a two dimensional perturbed lattice of size 100^2. For such a scenario, $\hat{\eta}$ can be estimated in a few minutes. Thus, we simulated $T = 100$ independent realizations and η_0 is estimated using the same procedure as previous studies, for each realization. The mean and Root Mean-Squared Error (RMSE) have been computed across T experiments. Table 2.3 displays the average and RMSE for the standard deviation and range parameters for different values of η, δ and covariance kernels. The instances for which $\hat{\eta}$ hits the boundary

points of Θ have been excluded in the procedure of calculating the mean and RMSE of the estimates.

Looking at the left and right panels of the Table 2.3 reveals that the RMSE of $\hat{\sigma}$ and $\hat{\theta}$ are slightly larger for the higher values of δ. Moreover, as we have discussed before, the RMSE and the average norm of $(\hat{\eta}-\eta_0)$ directly depends on the range parameter. It is immediately clear that there is a considerable reduction in RMSE for rational quadratic kernel comparing to the Matern class. This observation may look surprising for the reader, as the condition number of the covariance matrix associated to Matern kernel is smaller than that of rational quadratic due to its faster decay. Thus at the first glance, it may not corroborate our developed theory regarding the consistency of IF estimation algorithm for covariance matrices with bounded condition number. However, obtaining a highly accurate estimate of the dependence parameters are more difficult for a rapidly decaying covariance function, as more samples are almost independent. In the extreme case θ_0 is unidentifiable if $K(\cdot,\theta_0)$ is a compactly supported covariance function whose support size is strictly less than $(1-2\delta)$ (In this case all the samples are independent).

	$\delta = 0.1$		$\delta = 0.3$	
	$(\sigma_0,\theta_0) = (1,4)$	$(\sigma_0,\theta_0) = (1,7)$	$(\sigma_0,\theta_0) = (1,4)$	$(\sigma_0,\theta_0) = (1,7)$
Matern covariance	$\hat{\theta} \pm \text{RSME} = 4.107 \pm 1.224$	$\hat{\theta} \pm \text{RSME} = 7.259 \pm 2.462$	$\hat{\theta} \pm \text{RSME} = 3.982 \pm 0.980$	$\hat{\theta} \pm \text{RSME} = 6.814 \pm 2.233$
($\nu_0 = 0.5$)	$\hat{\sigma} \pm \text{RSME} = 0.999 \pm 0.067$	$\hat{\sigma} \pm \text{RSME} = 0.991 \pm 0.089$	$\hat{\sigma} \pm \text{RSME} = 0.995 \pm 0.062$	$\hat{\sigma} \pm \text{RSME} = 1.003 \pm 0.096$
	$(\sigma_0,\theta_0) = (1,4)$	$(\sigma_0,\theta_0) = (1,7)$	$(\sigma_0,\theta_0) = (1,4)$	$(\sigma_0,\theta_0) = (1,7)$
Matern covariance	$\hat{\theta} \pm \text{RSME} = 3.936 \pm 1.127$	$\hat{\theta} \pm \text{RSME} = 6.588 \pm 2.060$	$\hat{\theta} \pm \text{RSME} = 4.180 \pm 1.181$	$\hat{\theta} \pm \text{RSME} = 6.519 \pm 2.127$
($\nu_0 = 1.5$)	$\hat{\sigma} \pm \text{RSME} = 1.002 \pm 0.070$	$\hat{\sigma} \pm \text{RSME} = 0.992 \pm 0.096$	$\hat{\sigma} \pm \text{RSME} = 0.995 \pm 0.072$	$\hat{\sigma} \pm \text{RSME} = 1.018 \pm 0.107$
	$(\sigma_0,\theta_0) = (1,4)$	$(\sigma_0,\theta_0) = (1,7)$	$(\sigma_0,\theta_0) = (1,4)$	$(\sigma_0,\theta_0) = (1,7)$
Rational quadratic	$\hat{\theta} \pm \text{RSME} = 3.889 \pm 0.599$	$\hat{\theta} \pm \text{RSME} = 6.855 \pm 1.507$	$\hat{\theta} \pm \text{RSME} = 4.032 \pm 0.647$	$\hat{\theta} \pm \text{RSME} = 6.793 \pm 1.373$
covariance ($\nu_0 = 0.5$)	$\hat{\sigma} \pm \text{RSME} = 1.002 \pm 0.062$	$\hat{\sigma} \pm \text{RSME} = 0.986 \pm 0.082$	$\hat{\sigma} \pm \text{RSME} = 0.992 \pm 0.046$	$\hat{\sigma} \pm \text{RSME} = 0.990 \pm 0.069$
	$(\sigma_0,\theta_0) = (1,4)$	$(\sigma_0,\theta_0) = (1,7)$	$(\sigma_0,\theta_0) = (1,4)$	$(\sigma_0,\theta_0) = (1,7)$
Rational quadratic	$\hat{\theta} \pm \text{RSME} = 3.984 \pm 0.342$	$\hat{\theta} \pm \text{RSME} = 7.160 \pm 1.010$	$\hat{\theta} \pm \text{RSME} = 4.016 \pm 0.348$	$\hat{\theta} \pm \text{RSME} = 7.127 \pm 1.116$
covariance ($\nu_0 = 1.5$)	$\hat{\sigma} \pm \text{RSME} = 0.999 \pm 0.028$	$\hat{\sigma} \pm \text{RSME} = 0.994 \pm 0.074$	$\hat{\sigma} \pm \text{RSME} = 0.994 \pm 0.026$	$\hat{\sigma} \pm \text{RSME} = 0.995 \pm 0.049$

Table 2.3: Mean and RMSE of $\hat{\eta}$ over 100 independent experiments for the isotropic Matern and rational quadratic covariance functions, where $\mathcal{D}_n$ is a perturbed lattice of size 100^2 with $\delta \in \{0.1, 0.3\}$.

Now we turn to investigate the precision and RMSE of estimation algorithm (2.3) for the geometric anisotropic covariance structure. Same as before, G is a zero mean stationary GP in $\mathbb{R}^2$ observed on a perturbed lattice of size 100^2. G has a geometric anisotropic covariance kernel (Matern or rational quadratic) with

$$B_0 = \begin{pmatrix} \theta_0^{-1} & 0 \\ 0 & \rho_0^{-1} \end{pmatrix}, \quad \theta_0 = 4, \text{ and } \rho_0 = 6.$$

The parameter space $\Theta = \{(\theta_0,\rho_0) \in \Theta\}$ is a two dimensional box chosen as $[0.1, 15]^2$. The

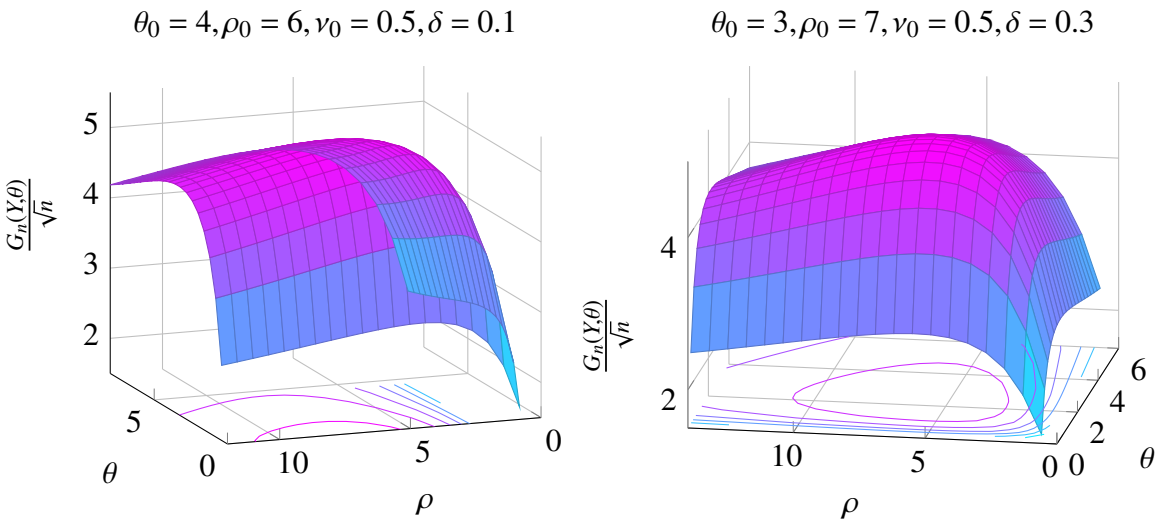

Figure 2.2: The above figures exhibit $n^{-1/2}G_n(Y,\theta)$ for geometric anisotropic Matern covariance function with $\nu_0 = 0.5$. The spatial samples form a two dimensional randomly $\delta-$perturbed regular lattice of size $N = 100$. In the left panel, $(\theta_0,\rho_0) = (4,6)$ and $\delta = 0.1$. In the right panel, $(\theta_0,\rho_0) = (3,7)$ and $\delta = 0.3$.

range parameters θ_0 and ρ_0 are estimated by solving the optimization problem (2.11). The Figure 2.2 exhibits the objective function in (2.11) and its contours for a GP with geometric anisotropic Matern covariance with $\nu_0 = 0.5$ which has been sampled on a $\delta-$perturbed regular grid. In the left and right panels of Figure 2.2, the other parameters are respectively given by $(\theta_0,\rho_0,\delta) = (4,6,0.1)$ and $(\theta_0,\rho_0,\delta) = (3,7,0.3)$. It can be seen from Figure 2.2 that G_n is a unimodal function with only one stationary point. We perform the maximization using the *optim* function in *R* and with *L-BFGS-B* algorithm [BLNZ95] (box constrained BFGS). The maximum iteration and the initial guess of the *L-BFGS-B* method are respectively 100 and $(2,2)$. The components of the gradient function are computed using the finite difference approximation with the step size of 10^{-3}. We cease the iteration when the relative change in the objective function is below 10^{-5}. The computation procedure of the average and RMSE of $\hat{\eta} = (\hat{\theta},\hat{\rho},\hat{\sigma})$ is exactly the same as the former simulation study. Table 2.4 presents a summary of the final results of this simulation study. It is clear from Table 2.4 that the RMSE for Matern covariance is significantly larger in comparison to rational quadratic class. Furthermore, increasing ν_0 for each covariance kernel leads to a slightly larger RMSE .

	$\delta = 0.1$	$\delta = 0.3$
Matern covariance ($\nu_0 = 0.5$)	$(\sigma_0, \rho_0, \theta_0) = (1, 6, 4)$ $\hat{\sigma} \pm \text{RSME} = 0.988 \pm 0.096$ $\hat{\rho} \pm \text{RSME} = 6.042 \pm 1.885$ $\hat{\theta} \pm \text{RSME} = 4.091 \pm 1.110$	$(\sigma_0, \rho_0, \theta_0) = (1, 6, 4)$ $\hat{\sigma} \pm \text{RSME} = 0.993 \pm 0.097$ $\hat{\rho} \pm \text{RSME} = 6.478 \pm 1.908$ $\hat{\theta} \pm \text{RSME} = 4.038 \pm 1.272$
Matern covariance ($\nu_0 = 1.5$)	$(\sigma_0, \rho_0, \theta_0) = (1, 6, 4)$ $\hat{\sigma} \pm \text{RSME} = 0.993 \pm 0.108$ $\hat{\rho} \pm \text{RSME} = 05.965 \pm 1.981$ $\hat{\theta} \pm \text{RSME} = 3.740 \pm 1.146$	$(\sigma_0, \rho_0, \theta_0) = (1, 6, 4)$ $\hat{\sigma} \pm \text{RSME} = 0.984 \pm 0.104$ $\hat{\rho} \pm \text{RSME} = 6.160 \pm 1.890$ $\hat{\theta} \pm \text{RSME} = 3.970 \pm 1.243$
Rational quadratic covariance ($\nu_0 = 0.5$)	$(\sigma_0, \rho_0, \theta_0) = (1, 6, 4)$ $\hat{\sigma} \pm \text{RSME} = 0.992 \pm 0.071$ $\hat{\rho} \pm \text{RSME} = 5.978 \pm 1.241$ $\hat{\theta} \pm \text{RSME} = 4.092 \pm 0.843$	$(\sigma_0, \rho_0, \theta_0) = (1, 6'4)$ $\hat{\sigma} \pm \text{RSME} = 0.989 \pm 0.076$ $\hat{\rho} \pm \text{RSME} = 5.921 \pm 1.208$ $\hat{\theta} \pm \text{RSME} = 4.037 \pm 1.064$
Rational quadratic covariance ($\nu_0 = 1.5$)	$(\sigma_0, \rho_0, \theta_0) = (1, 6, 4)$ $\hat{\sigma} \pm \text{RSME} = 0.996 \pm 0.036$ $\hat{\rho} \pm \text{RSME} = 6.116 \pm 0.821$ $\hat{\theta} \pm \text{RSME} = 4.045 \pm 0.543$	$(\sigma_0, \rho_0, \theta_0) = (1, 6, 4)$ $\hat{\sigma} \pm \text{RSME} = 0.998 \pm 0.036$ $\hat{\rho} \pm \text{RSME} = 6.158 \pm 0.766$ $\hat{\theta} \pm \text{RSME} = 4.150 \pm 0.524$

Table 2.4: Mean and RMSE of $\hat{\eta}$ over 100 independent experiments for the geometric anisotropic Matern and rational quadratic covariance functions, where $\mathcal{D}_n$ is a perturbed lattice of size 100^2 with $\delta \in \{0.1, 0.3\}$.

2.5 Discussion

Investigation of the asymptotic properties of the non-likelihood based optimization algorithms for estimating covariance parameters has remained relatively intact. Notwithstanding the thorough study of the consistency, minimax optimality and asymptotic normality of the stationary points of the IF loss function, there is much future work to be done to determine the computational and statistical strengths and weaknesses of this algorithm in either of the two frequently used asymptotic regimes. Here we mention a few among the many future directions which were beyond the scope of this paper.

(a) As indicated in Remark 2.1, the IF loss function can be viewed as an approximate minorizer for the likelihood loss (in the expected value sense) in the increasing domain setting. However, more work needs to be done to know how to precisely characterize a rich class of minorizers for the likelihood loss. We believe that responding to this question will provide a flexible class of fast and consistent estimators of covariance parameters.

(b) Spatial statisticians usually cast doubt upon the benefits of increasing domain asymptotics as spatial processes are unlikely to be stationary over a large domain. However

the developed theory in this chapter has persuaded us that the IF algorithm can consistently estimate microergodic covariance parameters, when applied on the preconditioned samples of a Gaussian random field in a fixed domain.

2.6 Proofs of the Main Results

We first introduce a few notation to simplify the algebra in the forthcoming sections. For any strictly positive scalar r and any $\theta_0 \in \Theta$, the ball of radius r (with respect to the Euclidean norm) centered at θ_0 and its complement are defined by

$$\Theta_{\theta_0}(r) := \Big\{\theta \in \Theta : \ \|\theta - \theta_0\|_{\ell_2} \leq r\Big\}, \quad \Theta^c_{\theta_0}(r) := \Theta \setminus \Theta_{\theta_0}(r).$$

Furthermore, for any $\theta_1, \theta_2 \in \Theta$ define

$$M_{\theta_1,\theta_2} := \frac{K_n^{1/2}(\theta_1)\, K_n(\theta_2)\, K_n^{1/2}(\theta_1)}{\|K_n(\theta_2)\|^2_{\ell_2}}, \quad H_{\theta_1,\theta_2} := \|K_n(\theta_2)\|_{\ell_2} M_{\theta_1,\theta_2}. \tag{2.16}$$

Proof of Theorem 2.1. Our proof has two major parts. In the first part, the consistency of $\hat{\theta}_n$ (correlation function's parameters) will be substantiated. In the second part, we establish the consistency of $\hat{\phi}_n$, which has a closed form solution in terms of Y, correlation function and $\hat{\theta}_n$, by conditioning on the consistency of $\hat{\theta}_n$. To this end, various types of concentration inequalities regarding the quadratic forms (and their supremum over a bounded space) of GPs are of the indispensable importance. Such results will be presented in the Section 2.7. Let Z be a standard Gaussian vector in $\mathbb{R}^n$. As Y and $\sqrt{\phi_0} K_n^{1/2}(\theta_0) Z$ have the same distribution, (2.3) can be equivalently written by

$$\Big(\hat{\phi}_n, \hat{\theta}_n\Big) = \underset{(\phi,\theta)\in\mathcal{I}\times\Theta}{\arg\max} \left\{ \phi\phi_0 Z^\top K_n^{1/2}(\theta_0) K_n(\theta)\, K_n^{1/2}(\theta_0) Z - \frac{\phi^2}{2} \|K_n(\theta)\|^2_{\ell_2} \right\}. \tag{2.17}$$

The objective function in (2.17) is quadratic in terms of ϕ and its maximizer $\hat{\phi}_n$ has a simple closed form. Replacing $\hat{\phi}_n$ to (2.17) gives a surrogate form for $\hat{\theta}_n$. Omitting the cumbersome algebra, the final results are given by

$$\frac{\hat{\phi}_n}{\phi_0} = Z^\top M_{\theta_0,\hat{\theta}_n} Z, \qquad \hat{\theta}_n = \underset{\theta\in\Theta}{\arg\max}\, Z^\top H_{\theta_0,\theta} Z. \tag{2.18}$$

We first show (as Claim 1) the consistency of $\hat{\theta}_n$, which is the supremum of a generalized chi-square random variable. The purpose of Claim 2 is to find the estimation rate of ϕ_0.

Claim 1. Choose $\xi > (m-1)$ and let $r_n \coloneqq C_{\min}\sqrt{\frac{\ln n}{n}}$ for some bounded positive scalar $C_{\min}$ (See Lemma 2.3 for its exact form). Then,

$$\Pr\left\{\hat{\theta}_n \in \Theta^c_{\theta_0}(r_n)\right\} \to 0, \quad \text{as } n \to \infty.$$

Proof of Claim 1. Consider the sequence $r'_n = n^{-1}\sqrt{\ln n}$, $\forall n$. The boundedness of Θ guarantees the existence of some $R_0 > 0$ such that a ball of radius R_0 contains Θ. So, the classical volume argument implies that

$$\left|\mathcal{N}_{r'_n}\left(\Theta^c_{\theta_0}(r_n)\right)\right| \leq \left|\mathcal{N}_{r'_n}(\Theta)\right| \lesssim \left(\frac{R_0}{r'_n}\right)^m = o\left(n^m\right). \tag{2.19}$$

So, there is $n_1 \in \mathbb{N}$ such that $\mathcal{N}_{r'_n}\left(\Theta^c_{\theta_0}(r_n)\right) \leq n^m$ for any $n \geq n_1$. It follows from (2.18) that

$$\Pr\left\{\hat{\theta}_n \in \Theta^c_{\theta_0}(r_n)\right\} \leq \text{RHS} \coloneqq \Pr(A_n) \coloneqq \Pr\left(Z^\top H_{\theta_0,\theta_0} Z \leq \sup_{\theta \in \Theta^c_{\theta_0}(r_n)} Z^\top H_{\theta_0,\theta} Z\right).$$

Thus, it suffices to control RHS from above. For a properly chosen positive scalar C_2, define the event π_n by

$$\pi_n \coloneqq \left\{\sup_{\theta \in \Theta^c_{\theta_0}(r_n)} Z^\top H_{\theta_0,\theta} Z \leq \sup_{\theta \in \mathcal{N}_{r'_n}\left(\Theta^c_{\theta_0}(r_n)\right)} Z^\top H_{\theta_0,\theta} Z + C_2\sqrt{\frac{\ln n}{n}}\right\}.$$

Notice that $r_n\sqrt{n} = \mathcal{O}\left(\sqrt{n^{-1}\ln n}\right)$. According to Lemma 2.1, there is a bounded $C_2 > 0$ for which $\tau_n \coloneqq \Pr(\pi_n^c) \to 0$. We refer the reader to Lemma 2.1 for the closed form expression of C_2. An upper bound on RHS is obtained by conditioning A_n on π_n.

$$\begin{aligned}
\text{RHS} &= \Pr(A_n \cap \pi_n) + \tau_n \Pr(A_n \mid \pi_n^c) \leq \tau_n + \Pr(A_n \cap \pi_n) \\
&\overset{(A)}{\leq} \tau_n + \Pr\left(Z^\top H_{\theta_0,\theta_0} Z \leq \sup_{\theta \in \mathcal{N}_{r'_n}\left(\Theta^c_{\theta_0}(r_n)\right)} Z^\top H_{\theta_0,\theta} Z + C_2\sqrt{\frac{\ln n}{n}}\right) \\
&\overset{(B)}{\leq} \tau_n + n^m \sup_{\theta \in \mathcal{N}_{r'_n}\left(\Theta^c_{\theta_0}(r_n)\right)} \Pr\left(Z^\top H_{\theta_0,\theta_0} Z \leq Z^\top H_{\theta_0,\theta} Z + C_2\sqrt{\frac{\ln n}{n}}\right)
\end{aligned} \tag{2.20}$$

The way that π_n and A_n have been defined trivially justifies inequality (A). Furthermore (B) is inferred from the combination of (2.19) and the union bound. Applying Lemma 2.3

guarantees the following result for any $\theta \in \mathcal{N}_{r_n'}\left(\Theta^c_{\theta_0}(r_n)\right)$.

$$\Pr\left(Z^\top H_{\theta_0,\theta_0} Z \leq Z^\top H_{\theta_0,\theta} Z + C_2\sqrt{\frac{\ln n}{n}}\right) \leq n^{-(1+\xi)}. \tag{2.21}$$

Finally, substituting (2.21) into (2.20) yields

$$\text{RHS} \leq \left(\tau_n + n^{m-(1+\xi)}\right) \to 0, \quad \text{as } n \to \infty.$$

□

Claim 2. There exists a bounded scalar $C > 0$, depending on $\mathcal{D}_n$, K and Ω, such that

$$\pi_n' := \Pr\left(\left|\frac{\hat{\phi}_n}{\phi_0} - 1\right| \geq r_n'' := C\sqrt{\frac{\ln n}{n}}\right) \to 0, \quad \text{as } n \to \infty.$$

(Here $C = 2\left(\mathfrak{D}_{\max} C_{\min} + C'\Lambda_{\max}\sqrt{m}\right)$ in which C' is a large enough positive universal constant and $C_{\min}$ has been defined in Claim 1. $\Lambda_{\max}$ and $\mathfrak{D}_{\max}$ are given in the Proposition 2.5.)

Proof of Claim 2. Recall r_n and r_n' form Claim 1. Obviously,

$$\pi_n' \leq T_1 + T_2 := \Pr\left\{\hat{\theta}_n \in \Theta^c_{\theta_0}(r_n)\right\} + \Pr\left\{\left(\left|\frac{\hat{\phi}_n}{\phi_0} - 1\right| \geq C\sqrt{\frac{\ln n}{n}}\right) \bigcap \left(\hat{\theta}_n \in \Theta_{\theta_0}(r_n)\right)\right\}.$$

Since T_1 tends to zero (via Claim 1), it suffices to show that T_2 is a diminishing sequence as $n \to \infty$. Let $\beta_{\hat{\theta}_n}$ be the closest point in $\mathcal{N}_{r_n'}\left(\Theta_{\theta_0}(r_n)\right)$ (which is a deterministic set) to $\hat{\theta}_n$. Based upon identity (2.18), we have

$$\left|\frac{\hat{\phi}_n}{\phi_0} - 1\right| = \left|Z^\top M_{\theta_0,\hat{\theta}_n} Z - 1\right| \tag{2.22}$$

Given that $\hat{\theta}_n$ belongs to $\Theta_{\theta_0}(r_n)$, applying the triangle inequality on the right hand side of the identity (2.22) yields that, almost surely

$$
\begin{aligned}
\left|\frac{\hat{\phi}_n}{\phi_0}-1\right| &\leq \left|Z^\top M_{\theta_0,\hat{\theta}_n} Z - Z^\top M_{\theta_0,\beta_{\hat{\theta}_n}} Z\right| + \left|Z^\top M_{\theta_0,\beta_{\hat{\theta}_n}} Z - tr\left(M_{\theta_0,\beta_{\hat{\theta}_n}}\right)\right| \\
&+ \left|tr\left(M_{\theta_0,\beta_{\hat{\theta}_n}}\right)-1\right| \\
&\overset{(D)}{\leq} \sup_{\theta\in\Theta_{\theta_0}(r_n)} \left\{ \begin{array}{c} \left|tr\left(M_{\theta_0,\beta_\theta}-1\right)\right| + \left|Z^\top M_{\theta_0,\beta_\theta} Z - tr\left(M_{\theta_0,\beta_\theta}\right)\right| \\ + \left|Z^\top M_{\theta_0,\theta} Z - Z^\top M_{\theta_0,\beta_\theta} Z\right| \end{array} \right\} \\
&\coloneqq \sup_{\theta\in\Theta_{\theta_0}(r_n)} \left\{T_{21}(\theta)+T_{22}(\theta)+T_{23}(\theta)\right\}. \qquad (2.23)
\end{aligned}
$$

Replacing the random quantities $\hat{\theta}_n$ and $\beta_{\hat{\theta}_n}$ with the nonrandom parameters θ and β_θ is the key advantage of (D). Now we control the terms T_{21}, T_{22} and T_{23} from above, uniformly over $\Theta_{\theta_0}(r_n)$. Lemma 2.2 guarantees the existence of a scalar C_0, for which

$$
\lim_{n\to\infty} \Pr\left(\sup_{\theta\in\Theta_{\theta_0}(r_n)} T_{23}(\theta) \leq C_0 r_n' \right) \to 1. \qquad (2.24)
$$

C_0 depends on $\Lambda_{\max}$ and $\mathfrak{D}_{\max}$. See Lemma 2.2 for its exact formulation. For large enough n, we have

$$
C_0 r_n' = \mathcal{O}\left(n^{-1}\sqrt{\ln n}\right) < \frac{1}{2}\mathfrak{D}_{\max} C_{\min}\sqrt{\frac{\ln n}{n}}. \qquad (2.25)
$$

Now we control $T_{21}(\theta)$ uniformly from above using Proposition 2.5. The goal is to show that

$$
\sup_{\theta\in\Theta_{\theta_0}(r_n)} T_{21}(\theta) < \frac{3}{2}\mathfrak{D}_{\max} C_{\min}\sqrt{\frac{\ln n}{n}}. \qquad (2.26)
$$

Applying the Cauchy-Schwartz inequality shows that (recalling M_{θ_1,θ_2} from (2.16))

$$
\begin{aligned}
\sup_{\theta\in\Theta_{\theta_0}(r_n)} T_{21}(\theta) &= \sup_{\theta\in\Theta_{\theta_0}(r_n)} \left| \frac{\langle K_n(\beta_\theta), K_n(\theta_0)\rangle}{\|K_n(\beta_\theta)\|_{\ell_2}^2} - 1 \right| \\
&= \sup_{\theta\in\Theta_{\theta_0}(r_n)} \left| \frac{\langle K_n(\beta_\theta), K_n(\beta_\theta) - K_n(\theta_0)\rangle}{\|K_n(\beta_\theta)\|_{\ell_2}^2} \right| \\
&\leq \sup_{\theta\in\Theta_{\theta_0}(r_n)} \frac{\|K_n(\beta_\theta) - K_n(\theta_0)\|_{\ell_2}}{\|K_n(\beta_\theta)\|_{\ell_2}} \\
&\leq \sup_{\theta\in\Theta_{\theta_0}(r_n)} \frac{\|K_n(\beta_\theta) - K_n(\theta_0)\|_{\ell_2}}{\sqrt{n}}. \qquad (2.27)
\end{aligned}
$$

Furthermore, using the part (b) of the Proposition 2.5, we get

$$\begin{aligned}\sup_{\theta\in\Theta_{\theta_0}(r_n)}\frac{\|K_n(\beta_\theta)-K_n(\theta_0)\|_{\ell_2}}{\sqrt{n}} &\leq \mathfrak{D}_{\max}\sup_{\theta\in\Theta_{\theta_0}(r_n)}\|\theta_0-\beta_\theta\|_{\ell_2}\\ &\leq \mathfrak{D}_{\max}\sup_{\theta\in\Theta_{\theta_0}(r_n)}\left(\|\theta-\beta_\theta\|_{\ell_2}+\|\theta_0-\theta\|_{\ell_2}\right)\\ &\leq \mathfrak{D}_{\max}(r_n+r_n') < \frac{3}{2}\mathfrak{D}_{\max}C_{\min}\sqrt{\frac{\ln n}{n}}. \qquad (2.28)\end{aligned}$$

Note that the last inequality holds for large enough n. So (2.25) follows from replacing (2.28) into (2.27). In the sequel we achieve a uniform upper bound on T_{22}. For brevity define $u_n \coloneqq \Lambda_{\max}\sqrt{mn^{-1}\ln n}$ and select a large enough universal constant C'. Recall that β_θ, by its definition, is an element of the finite set $\mathcal{N}_{r_n'}\left(\Theta_{\theta_0}(r_n)\right)$. Thus,

$$\Pr\left(\sup_{\theta\in\Theta_{\theta_0}(r_n)}T_{22}(\theta)\geq C'u_n\right)\leq\left|\mathcal{N}_{r_n'}\left(\Theta_{\theta_0}(r_n)\right)\right|\sup_{\theta\in\Theta_{\theta_0}(r_n)}\Pr\left(T_{22}(\theta)\geq C'u_n\right).$$

The same trick as (2.19) leads to $\left|\mathcal{N}_{r_n'}\left(\Theta_{\theta_0}(r_n)\right)\right| = o(n^m)$. So, it is adequate to show that

$$\Pr\left(T_{22}(\theta)\geq C'u_n\right)\leq n^{-m},\quad \forall\,\theta\in\Theta_{\theta_0}(r_n). \qquad (2.29)$$

We employ Hanson-Wright inequality (Theorem 1.1, [RV$^+$13]) for obtaining a probabilistic upper bound on $T_{22}(\theta)$ (for a fixed θ).

$$\Pr\left\{T_{22}(\theta)\geq C'\sqrt{m\ln n}\left(\left\|M_{\theta_0,\beta_\theta}\right\|_{\ell_2}\vee\left\|M_{\theta_0,\beta_\theta}\right\|_{2\to2}\sqrt{m\ln n}\right)\right\}\leq n^{-m},\quad\forall\,\theta\in\Theta_{\theta_0}(r_n).$$

For simplifying the upper bound on $T_{22}(\theta)$, we control the operator and Frobenius norms of $M_{\theta_0,\beta_\theta}$ from above. The following inequalities can be easily justified by Proposition 2.5.

$$\left\|M_{\theta_0,\beta_\theta}\right\|_{\ell_2}\leq\Lambda_{\max}n^{-1/2},\quad\left\|M_{\theta_0,\beta_\theta}\right\|_{2\to2}\leq\Lambda_{\max}^2n^{-1}. \qquad (2.30)$$

Replacing (2.30) into Hanson-Wright inequality justifies (2.29). Hence

$$\Pr\left(\sup_{\theta\in\Theta_{\theta_0}(r_n)}T_{22}(\theta)\geq C'u_n\right)\leq\left|\mathcal{N}_{r_n'}\left(\Theta_{\theta_0}(r_n)\right)\right|n^{-m}\to0,\quad\text{as } n\to\infty. \qquad (2.31)$$

Substituting inequalities (2.24), (2.25), (2.26) and (2.31) into (2.23) concludes the proof by confirming that T_2 goes to zero as $n\to\infty$. □

Combining Claims 1 and 2 ends the proof. □

Proof of Theorem 2.2. Let $r_n = C\sqrt{n^{-1}\ln n}$ for some strictly positive C whose exact form will be given shortly. Let $(\hat{\phi}_n, \hat{\theta}_n)$ be any stationary point of the optimization problem (2.3). Due to the space constraint, we just show that $\hat{\theta}_n \in \Theta_{\theta_0}(r_n)$. The same technique as Claim **??** in the proof of Theorem 2.1 attains the convergence rate of $\hat{\phi}_n$. Notice that $\hat{\theta}_n$ is a stationary point of the optimization problem (2.18). We show that with a high probability there is no $\theta \in \Theta^c_{\theta_0}(r_n)$ for which the gradient of the objective function in (2.18) be exactly zero. In order to substantiate our claim, we prove that the absolute value of the inner product of the gradient and a fixed non-zero vector is uniformly greater than zero on $\Theta^c_{\theta_0}(r_n)$.
Let $Z \in \mathbb{R}^n$ is a standard Gaussian vector and θ_l, $l = 1, \ldots, m$ is the l^{th} component of θ. We first give a closed form for the gradient function in (2.18), which will be denoted by $[G_l(\theta)]_{l=1}^m$.

$$\begin{aligned} G_l(\theta) &:= Z^\top P^l_{\theta_0,\theta} Z := \frac{\partial}{\partial \theta_l} Z^\top H_{\theta_0,\theta} Z \\ &= Z^\top K_n^{1/2}(\theta_0) \left\{ \frac{\partial}{\partial \theta_l} K_n(\theta) - \langle \frac{\partial}{\partial \theta_l} K_n(\theta), K_n(\theta) \rangle \frac{K_n(\theta)}{\|K_n(\theta)\|^2_{\ell_2}} \right\} K_n^{1/2}(\theta_0) Z, \end{aligned}$$

Clearly, $\left[G_l(\hat{\theta}_n)\right]_{l=1}^m = \mathbf{0}_m$. Choose any $\lambda \in \mathcal{S}^{m-1}$ and let $Y := K_n^{1/2}(\theta_0) Z$. Observe that

$$W(\theta) := \sum_{j=1}^m \lambda_j G_j(\theta) = Y^\top \left\{ \sum_{j=1}^m \lambda_j \frac{\partial}{\partial \theta_j} K_n(\theta) - \langle \sum_{j=1}^m \lambda_j \frac{\partial}{\partial \theta_j} K_n(\theta), K_n(\theta) \rangle \frac{K_n(\theta)}{\|K_n(\theta)\|^2_{\ell_2}} \right\} Y.$$

We can conclude that $\hat{\theta}_n \in \Theta_{\theta_0}(r_n)$ in probability, if we can prove that

$$\Pr\left(\inf_{\theta \in \Theta^c_{\theta_0}(r_n)} |W(\theta)| > 0 \right) \to 1, \quad \text{as } n \to \infty. \tag{2.32}$$

(2.32) follows from the succeeding claims, whose proofs have been thoroughly presented in [KSN17].

Claim 1. There exists a positive finite constant C_0 (depending on K, Θ and $\mathcal{D}_n$) such that

$$\lim_{n\to\infty} \Pr\left(\sup_{\theta \in \Theta} \left| W(\theta) - \sum_{j=1}^m \lambda_j tr\left(P^j_{\theta_0,\theta}\right) \right| \geq C_0 \sqrt{n \ln n} \right) = 0. \tag{2.33}$$

Claim 2. The succeeding inequality holds for large enough n (C_0 is from the previous

claim).

$$\inf_{\theta\in\Theta^c_{\theta_0}(r_n)}\left|\sum_{j=1}^{m}\lambda_j tr\left(P^j_{\theta_0,\theta}\right)\right| > C_0\sqrt{n\ln n}$$

The Claim 1 provides a uniform concentration inequality regarding the random function $W(\theta)$. In the Claim 2, we obtain a uniform lower bound on the expected value of $W(\theta)$ over $\Theta^c_{\theta_0}(r_n)$. □

Proof of Theorem 2.3. We follow the standard techniques presented in the Chapter 2 of [Tsy09] for bounding the minimax risk from below. For any $\theta\in\Theta$, $\mathbb{P}_\theta$ stands for the associated distribution to a zero mean Gaussian vector with the covariance function $K_n(\theta)$. Finding far enough (with respect to the Euclidean distance) pair of the correlation parameters, $\theta_i\in\Theta$, $i=1,2$, for which $D(\mathbb{P}_{\theta_1}\,\|\,\mathbb{P}_{\theta_2})<\alpha=1/2$ lies at the heart of our proof. The two bounded positive scalars $\mathfrak{D}_{\max}$ and $\Lambda_{\min}$ appearing here are defined in Proposition 2.5. To ease notation let $r_n\coloneqq\frac{\Lambda_{\min}}{8\mathfrak{D}_{\max}\sqrt{n}}$ (Choose n large enough so that $4r_n\le\operatorname{diam}(\Theta)$). Choose $\theta_1,\theta_2\in\Theta$ with $2r_n\le\|\theta_2-\theta_1\|_{\ell_2}\le 4r_n$. The connectedness of Θ guarantees the existence of such pair of points. We first use the Proposition 2.8 to show that $D(\mathbb{P}_{\theta_1}\,\|\,\mathbb{P}_{\theta_2})\le\alpha$.

$$D(\mathbb{P}_{\theta_1}\,\|\,\mathbb{P}_{\theta_2})\le 2n\left(\frac{\mathfrak{D}_{\max}}{\Lambda_{\min}}\|\theta_2-\theta_1\|_{\ell_2}\right)^2\le 32n\left(\frac{\mathfrak{D}_{\max}}{\Lambda_{\min}}r_n\right)^2=\alpha=\frac{1}{2}.$$

As $\alpha\ge D(\mathbb{P}_{\theta_1}\,\|\,\mathbb{P}_{\theta_2})$, Theorem 2.2 of [Tsy09] yields

$$\inf_{\hat\theta_n}\sup_{\theta_0\in\Theta}\Pr\left(\left\|\hat\theta_n-\theta_0\right\|_{\ell_2}\ge r_n\right)\ge\left(\frac{1}{4}e^{-\alpha}\right)\vee\left(\frac{1-\sqrt{\frac{\alpha}{2}}}{2}\right)=\frac{1}{4}. \tag{2.34}$$

The desired statement follows from the fact that $\|\hat\eta_n-\eta_0\|_{\ell_2}\ge\left\|\hat\theta_n-\theta_0\right\|_{\ell_2}$. □

Proof of Theorem 2.4. Let $g:\Omega\mapsto\mathbb{R}^{m+1}$ represents the gradient of the objective function in (2.2) with respect to η. Here g_j, $j=11,\ldots,(m+1)$ stands for the j^{th} entry of g. Analyzing the exact second order Taylor expansion of $\sqrt{n}g(\eta)$ around η_0 at $\eta=\hat\eta_n$ is the integral part of the proof. We argue that the second order term of the expansion, which involves the third order derivatives of the covariance function, converges to zero in probability as n grows to infinity. We also show that the first term (zeroth order term) in the expansion converges weakly to a Gaussian random variable. These two ingredients lead to the desirable result by showing the asymptotic normality of the first order term in the expansion, which directly depends on $\sqrt{n}(\hat\eta_n-\eta_0)$.

For simplicity, define $R_n^J(\eta) = \frac{\partial R_n(\eta)}{\partial \eta_{j_q} \partial \eta_{j_1}}$ for any $q \in \{1,2\}$ and $J \in \{1,\dots,m+1\}^q$. In fact

$$n g_j(\eta) = Y^\top R_n^j(\eta)\, Y - \langle R_n^j(\eta), R_n(\eta) \rangle. \tag{2.35}$$

Let $\hat{\eta}_n$ be an arbitrary stationary point of optimization problem (2.2). Clearly, $g(\hat{\eta}_n) = \mathbf{0}_{m+1}$. The second order approximation of g_j around $\hat{\eta}_n$ yields

$$\sqrt{n} g_j(\hat{\eta}_n) = \sqrt{n} g_j(\eta_0) + \langle \sqrt{n}(\hat{\eta}_n - \eta_0), \nabla_\eta g_j(\eta)\Big|_{\eta=\eta_0} \rangle + \sqrt{n}\Delta_j(\eta_0, \hat{\eta}_n),$$

for some residual function $\Delta_j(\cdot,\cdot)$. Note that $\Delta_j(\eta_0, \hat{\eta}_n)$ is given by

$$\Delta_j(\eta_0, \hat{\eta}_n) = (\hat{\eta}_n - \eta_0)^\top \left[\frac{\partial g_j(\eta)}{\partial \eta_{l_1} \partial \eta_{l_2}}\Big|_{\eta=z_j} \right]_{l_1,l_2=1}^{m+1} (\hat{\eta}_n - \eta_0)$$

in which z_j lies on the line segment between η_0 and $\hat{\eta}_n$. Proposition D.10 of [Bac14] guarantees the statement (2.14) for $\Sigma = \Sigma_2^{-1}\Sigma_1\Sigma_2^{-1}$ and hence concludes the proof, if

(a) The matrix Σ_2 defined as the following, is well defined and positive definite.

$$V^n := \left[-\frac{\partial}{\partial \eta_l} g_k(\eta)\Big|_{\eta=\eta_0} \right]_{l,k=1}^{m+1} \xrightarrow{\Pr} \Sigma_2, \text{ as } n \to \infty.$$

(b) $\sqrt{n} g(\eta_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}_{m+1}, \Sigma_1)$ for a positive semidefinite matrix $\Sigma_1 \in \mathbb{R}^{(m+1)\times(m+1)}$.

(c) $\Pr\left(\lim_{n\to\infty} \sqrt{n}\Delta_j(\eta_0, \hat{\eta}_n) = 0 \right) = 1$, for any $j \in \{1,\dots,m+1\}$.

The reminder of the proof hinges on the following technicalities which verify conditions $(a)-(c)$.

Validating condition (a)**.** The entries of V^n has the following explicit form.

$$V_{lk}^n = \frac{1}{n}\langle R_n^l(\eta_0), R_n^k(\eta_0) \rangle + \frac{1}{n}\left\{ Y^\top R_n^{lk}(\eta_0)\, Y - \langle R_n^{lk}(\eta_0), R_n(\eta_0) \rangle \right\}.$$

Now define

$$\Sigma_2 := \left[\lim_{n\to\infty} \frac{\langle R_n^l(\eta_0), R_n^k(\eta_0) \rangle}{n} \right]_{l,k=1}^{m+1}.$$

Notice that the entries of Σ_2 are well defined and bounded due to part (a) of Proposition 2.6. The proof will be presented in two steps: First we show that Σ_2 is a positive definite matrix. Second, we prove that $\Phi_{lk}^n := \left\{ Y^\top R_n^{lk}(\eta_0)\, Y - \langle R_n^{lk}(\eta_0), R_n(\eta_0) \rangle \right\}/n$ converges to zero

in probability. To substantiate the first claim, consider an arbitrary $\lambda \in \mathcal{S}^m$. It is required to show that $\lambda^\top \Sigma_2 \lambda > c$, for some constant $c > 0$. The condition $(A4.a)$ guarantees the existence of positive scalars M_2, r_2 such that for any $s \in \mathcal{D}_n$,

$$\max_{s' \in \mathcal{D}_n(s,r_2)} \left| \sum_{l=1}^{m+1} \lambda_l \frac{\partial}{\partial \eta_l} R(s-s',\eta)\Big|_{\eta=\eta_0} \right| \geq M_2. \tag{2.36}$$

Thus,

$$\begin{aligned} \lambda^\top \Sigma_2 \lambda &= \lim_{n\to\infty} \sum_{l,k=1}^{m+1} \lambda_l \lambda_k \frac{\langle R_n^l(\eta_0), R_n^k(\eta_0)\rangle}{n} = \lim_{n\to\infty} \frac{1}{n} \left\| \sum_{l=1}^{m+1} \lambda_l R_n^l(\eta_0) \right\|_{\ell_2}^2 \\ &= \lim_{n\to\infty} \frac{1}{n} \left\| \left[\sum_{l=1}^{m+1} \lambda_l \frac{\partial}{\partial \eta_l} R(s'-s,\eta)\Big|_{\eta=\eta_0} \right]_{s,s'\in\mathcal{D}_n} \right\|_{\ell_2}^2 \overset{(A)}{\geq} M_2^2. \end{aligned} \tag{2.37}$$

Here, inequality (A) is an easy consequence of (2.36). The rest of the proof is devoted to prove the second claim. Choose an arbitrary strictly positive ϵ. As Φ_{lk}^n is a zero mean random variable, using Chebyshev's inequality we get

$$\begin{aligned} \Pr\left(\left|\Phi_{lk}^n\right| \geq \epsilon\right) &\leq \frac{\operatorname{var}\left(\Phi_{lk}^n\right)}{\epsilon^2} = \frac{2\left\|R_n^{1/2}(\eta_0) R_n^{lk}(\eta_0) R_n^{1/2}(\eta_0)\right\|_{\ell_2}^2}{n^2\epsilon^2} \overset{(B)}{\lesssim} \left(\frac{\phi_0}{n\epsilon}\left\|R_n^{lk}(\eta_0)\right\|_{\ell_2}\right)^2 \\ &\overset{(C)}{=} \mathcal{O}\left(n^{-1}\right) \to 0, \end{aligned}$$

in which (B) and (C) are implied by Propositions 2.5 and 2.6, respectively (See Section 2.7 for more details about the constants).

Validating condition (b)**.** Define $Q_n^j := n^{-1/2} R_n^{1/2}(\eta_0) R_n^j(\eta_0) R_n^{1/2}(\eta_0)$ for $1 \leq j \leq m+1$, and write $\Psi_{n,\lambda} := \lambda_1 Q_n^1 + \ldots + \lambda_{m+1} Q_n^{m+1}$ for any $\lambda = (\lambda_1, \ldots, \lambda_{m+1}) \in \mathcal{S}^m$. Rewriting (2.35) yields

$$\sqrt{n} g_j(\eta_0) \overset{d}{=} Z^\top Q_n^j Z - tr\left(Q_n^j\right).$$

The asymptotic normality of $\sqrt{n} g(\eta_0)$ is justified if there is a positive semi-definite Σ_1 such that $\langle \lambda, \sqrt{n} g(\eta_0)\rangle \xrightarrow{d} \mathcal{N}(0, \lambda^\top \Sigma_1 \lambda)$ for any $\lambda \in \mathcal{S}^m$. This statement trivially holds for zero $\Psi_{n,\lambda}$. So, without loss of generality assume that $\Psi_{n,\lambda}$ is non-zero. Observe that

$$\langle \lambda, \sqrt{n} g(\eta_0)\rangle = \frac{\{Z^\top \Psi_{n,\lambda} Z - tr(\Psi_{n,\lambda})\}}{\left\|\Psi_{n,\lambda}\right\|_{\ell_2}} \left\|\Psi_{n,\lambda}\right\|_{\ell_2}.$$

We claim that $\lim_{n\to\infty} 2\left\|\Psi_{n,\lambda}\right\|_{\ell_2}^2 = \lambda^\top \Sigma_1 \lambda$ for a covariance matrix Σ_1. The construction of $\Psi_{n,\lambda}$

yields

$$2\left\|\Psi_{n,\lambda}\right\|_{\ell_2}^2 = 2\lambda^\top \left[\langle Q_n^k, Q_n^l\rangle\right]_{l,k=1}^{m+1} \lambda. \tag{2.38}$$

Thus, it is enough to show that the matrix defined by $\Sigma_1 \coloneqq \lim_{n\to\infty} 2\left[\langle Q_n^k, Q_n^l\rangle\right]_{l,k=1}^{m+1}$ is well defined (with bounded entries) and positive semi-definite. Well definiteness of Σ_1 can be proved using the same techniques as the proof of Claim 1 and by employing Propositions 2.5 and 2.6. The positive semi-definite property of Σ_1 is an immediate consequence of (2.38). We conclude the proof by showing that

$$\frac{\{Z^\top \Psi_{n,\lambda} Z - tr(\Psi_{n,\lambda})\}}{\sqrt{2}\left\|\Psi_{n,\lambda}\right\|_{\ell_2}} \xrightarrow{d} \mathcal{N}(0,1).$$

According to Lemma 2.4, this statement is valid if the following claim holds.

Claim 1. $\left\|\Psi_{n,\lambda}\right\|_{\ell_2}^{-1} \left\|\Psi_{n,\lambda}\right\|_{2\to 2} \le C/\sqrt{n}$ for some positive scalar C.

Proof of Claim 1. We show that C depends on m, $\Lambda_{\max}$, $\Lambda_{\min}$ and $\Lambda'_{\max}$ (Except m, all the constant are introduced in Propositions 2.5 and 2.6). Obviously $\Psi_{n,\lambda}$ can be rewritten as,

$$\Psi_{n,\lambda} = \frac{1}{\sqrt{n}} R_n^{1/2}(\eta_0) \left\{ \sum_{j=1}^{m+1} \lambda_j R_n^j(\eta_0) \right\} R_n^{1/2}(\eta_0)$$

Applying Propositions 2.5, we get

$$\begin{aligned} \frac{\left\|\Psi_{n,\lambda}\right\|_{2\to 2}}{\left\|\Psi_{n,\lambda}\right\|_{\ell_2}} &\le \frac{\|R_n(\eta_0)\|_{2\to 2} \left\|\sum_{j=1}^{m+1} \lambda_j R_n^j(\eta_0)\right\|_{2\to 2}}{\left\|\sum_{j=1}^{m+1} \lambda_j R_n^j(\eta_0)\right\|_{\ell_2} \lambda_{\min}\{R_n(\eta_0)\}} \\ &\le \frac{\Lambda_{\max}}{\Lambda_{\min}} \left\|\sum_{j=1}^{m+1} \lambda_j R_n^j(\eta_0)\right\|_{2\to 2} \left\|\sum_{j=1}^{m+1} \lambda_j R_n^j(\eta_0)\right\|_{\ell_2}^{-1}. \end{aligned} \tag{2.39}$$

Furthermore, using Proposition 2.6 leads to

$$\left\|\sum_{j=1}^{m+1} \lambda_j R_n^j(\eta_0)\right\|_{2\to 2} \le \sum_{j=1}^{m+1} |\lambda_j| \left\|R_n^j(\eta_0)\right\|_{2\to 2} \le \Lambda'_{\max} \|\lambda\|_{\ell_1} \le \Lambda'_{\max} \sqrt{m+1}.$$

From (2.37) we know that there is a scalar $C_0 \in (0,\infty)$ for which $\left\|\sum_{j=1}^{m+1} \lambda_j R_n^j(\eta_0)\right\|_{\ell_2} \ge C_0\sqrt{n}$. Replacing the last two inequalities into (2.39) ends the proof. □

In conclusion we state that the condition (c) can be proved using akin techniques as the

proof of Proposition 3.2 of [Bac14]. We omit the technical details due to the space constraints. □

2.7 Technical Results

In this section, we only prove the auxiliary results of independent interest and the proof of other technical lemmas and propositions will be omitted. The detailed proof of all results appearing in this section can be found in [KSN16].
The first result examines the perturbation of some norms of $K_n(\theta)$ with respect to θ. It appears to be of great importance for proving Theorems 2.1–2.4 in the Section 2.6.

Proposition 2.5. Suppose that $\mathcal{D}_n$ admits Assumption 2.1. Moreover, Assumption 2.2 holds for Θ and K. Construct $n \times n$ correlation matrix $K_n(\theta) := [K(s-s',\theta)]_{s,s' \in \mathcal{D}_n}$ for any $\theta \in \Theta$.

(a) There are bounded positive scalars $\Lambda_{\min}$ and $\Lambda_{\max}$ (depending on K, Θ, d and δ) such that

$$\Lambda_{\min} \leq \min_{n \in \mathbb{N}} \min_{\theta \in \Theta} \frac{1}{\left\| K_n^{-1}(\theta) \right\|_{2 \to 2}}, \quad \max_{n \in \mathbb{N}} \max_{\theta \in \Theta} \| K_n(\theta) \|_{2 \to 2} \leq \Lambda_{\max}.$$

(b) There exist scalars $\mathfrak{D}_{\min}, \mathfrak{D}_{\max} \in (0, \infty)$ (depending on K, Θ, d and δ) such that

$$\| K_n(\theta_2) - K_n(\theta_1) \|_{2 \to 2} \leq \mathfrak{D}_{\max} \| \theta_2 - \theta_1 \|_{\ell_2}, \tag{2.40}$$

$$\frac{1}{\sqrt{n}} \| K_n(\theta_2) - K_n(\theta_1) \|_{\ell_2} \leq \mathfrak{D}_{\max} \| \theta_2 - \theta_1 \|_{\ell_2}, \tag{2.41}$$

and

$$\frac{1}{\sqrt{n}} \| K_n(\theta_2) - K_n(\theta_1) \|_{\ell_2} \geq \mathfrak{D}_{\min} \| \theta_2 - \theta_1 \|_{\ell_2}, \tag{2.42}$$

for any $\theta_1, \theta_2 \in \Theta$.

For ease of reference, we present the following result as a standalone Proposition. Its proof is akin to that of Proposition 2.5 and will be skipped to avoid redundancy.

Proposition 2.6. Suppose that $\mathcal{D}_n$ admits Assumption 2.1. Moreover, Θ and K satisfy Assumption 2.3. Construct the matrix $\partial K_n(\theta)/\partial \theta_j := \left[\partial K(s-s',\theta)/\partial \theta_j \right]_{s,s' \in \mathcal{D}_n}$,for $\theta \in \Theta$ and $j = 1, \ldots, m$.

(a) There is a bounded strictly positive scalar $\Lambda'_{\max}$ (depending on K, Θ, d and δ) such that

$$\max_{n \in \mathbb{N}} \max_{\theta \in \Theta} \left\| \frac{\partial}{\partial \theta_j} K_n(\theta) \right\|_{2 \to 2} \leq \Lambda'_{\max},$$

(b) There is $\mathfrak{D}'_{\max} > 0$ such that for any $\theta_1, \theta_2 \in \Theta$

$$\left\| \frac{\partial}{\partial \theta_j} K_n(\theta_2) - \frac{\partial}{\partial \theta_j} K_n(\theta_1) \right\|_{2\to 2} \leq \mathfrak{D}'_{\max} \|\theta_2 - \theta_1\|_{\ell_2}. \tag{2.43}$$

The bounded positive scalars $\mathfrak{D}_{\max}, \mathfrak{D}_{\min}$ and $\Lambda_{\max}$, which have been introduced in the Proposition 2.5, become frequently apparent in the subsequent results in this section. It is also proper to remind the reader that $\mathcal{N}_\epsilon(\mathcal{A})$ stands for the $\epsilon-$net of $\mathcal{A}$ with respect to the Euclidean distance. Furthermore, the matrices H_{θ_1,θ_2} and M_{θ_1,θ_2} have been formerly defined in (2.16) for any pair of the correlation function parameters θ_1, θ_2. The succeeding two Lemmas (2.1 and 2.2), which come in handy in the proof of Theorem 2.1, establish a probabilistic upper bound on the maximum of a quadratic Gaussian expression over a uncountable set Θ in terms of its largest value over one of its finite subset.

Lemma 2.1. Let $Z \in \mathbb{R}^n$ be a standard Gaussian vector and suppose that $\mathcal{D}_n$ satisfies Assumption 2.1. Furthermore, assume that Θ and K admit Assumption 2.2. For any vanishing positive sequence $\{r_n\}_{n\in\mathbb{N}}$, any non-empty $\bar{\Theta} \subset \Theta$ and each $\theta_0 \in \Theta$,

$$\lim_{n\to\infty} \Pr\left\{ \left(\sup_{\theta\in\bar{\Theta}} Z^\top H_{\theta_0,\theta} Z - \sup_{\theta\in\mathcal{N}_{r_n}(\bar{\Theta})} Z^\top H_{\theta_0,\theta} Z \right) \geq C r_n \sqrt{n} \right\} = 0, \tag{2.44}$$

where $C = 2\Lambda_{\max}(1 + \mathfrak{D}_{\max})$.

Lemma 2.2. Let $Z \in \mathbb{R}^n$ be a standard Gaussian vector. Suppose that Assumption 2.1 and Assumption 2.2 hold for $\mathcal{D}_n$, Θ and K. For any strictly positive vanishing sequence $\{r_n\}_{n=1}^\infty$, any non-empty $\bar{\Theta} \subset \Theta$ and arbitrary $\theta_0 \in \Theta$,

$$\Pr\left\{ \sup_{\theta\in\bar{\Theta}} \left| Z^\top \left(M_{\theta_0,\theta} - M_{\theta_0,\beta_\theta} \right) Z \right| \geq C r_n \right\} \to 0, \quad \text{as } n \to \infty.$$

Here β_θ represents the nearest element of $\mathcal{N}_{r_n}\left(\bar{\Theta}\right)$ to θ and $C = 2\mathfrak{D}_{\max}\left(1 + 2\Lambda^2_{\max}\right)$.

Now we state a lemma which plays a crucial role in the proof of Theorem 2.1.

Lemma 2.3. Let $Z \in \mathbb{R}^n$ be a standard Gaussian vector and let $C, \xi > 0$. Suppose that Θ and K satisfy Assumption 2.2. Select $\theta_1, \theta_2 \in \Theta$ such that

$$\|\theta_2 - \theta_1\|_{\ell_2} \geq C_{\min} \sqrt{\frac{\ln n}{n}}, \tag{2.45}$$

in which $C_{\min} \coloneqq 4\mathfrak{D}^{-1}_{\min} \Lambda^2_{\max} \sqrt{C'(1+\xi)}$ (Recall $\mathfrak{D}_{\min}$ and $\Lambda_{\max}$, from the Proposition 2.5),

for some appropriately chosen universal constant $C' > 0$. There exists $n_0 = \mathcal{O}(1)$ (depending on C, ξ, K, $\mathcal{D}_n$ and Θ) such that for any $n \geq n_0$

$$p \coloneqq \Pr\left\{ Z^\top \left(H_{\theta_2,\theta_2} - H_{\theta_2,\theta_1} \right) Z \leq C \sqrt{\frac{\ln n}{n}} \right\} \leq n^{-(1+\xi)}. \tag{2.46}$$

Refer to the identity (2.16) for the definition of H_{θ_2,θ_1}.

The next proposition rigorously expresses the uniform concentration of the Euclidean squared norm of Gaussian vectors with the covariance matrix $K_n(\theta)$, $\theta \in \Theta$ around their mean. We employ such inequality for proving Theorem 2.2.

Proposition 2.7. Let $\Theta \subset \mathbb{R}^m$ be a bounded set. Consider the class of n by n matrices $\{\Pi_n(\theta)\}_{\theta\in\Theta}$ parametrized by $\theta \in \Theta$. Suppose that the following conditions hold

(a) The operator norm of $\Pi_n(\theta)$ is uniformly bounded in Θ. Namely,

$$M \coloneqq \sup_n \sup_{\theta\in\Theta} \|\Pi_n(\theta)\|_{2\to 2} < \infty.$$

(b) The mapping $\left(\theta, \|\cdot\|_{\ell_2}\right) \mapsto \left(\Pi_n(\theta), \|\cdot\|_{2\to 2}\right)$ is Lipschitz. Namely, there is $C > 0$ for which

$$\|\Pi_n(\theta_2) - \Pi_n(\theta_1)\|_{2\to 2} \leq C \|\theta_2 - \theta_1\|_{\ell_2}, \quad \forall\, \theta_1, \theta_2 \in \Theta. \tag{2.47}$$

(c)

$$\frac{\|\Pi_n(\theta)\|_{2\to 2}}{\|\Pi_n(\theta)\|_{\ell_2}} = o\left(\frac{1}{\sqrt{\ln n}}\right), \quad \forall\, \theta \in \Theta.$$

Then, there is a constant $C' > 0$ such that

$$\lim_{n\to\infty} \Pr\left(\sup_{\theta\in\Theta} \left| Z^\top \Pi_n(\theta) Z - tr\{\Pi_n(\theta)\} \right| \geq C' \sqrt{n \ln n} \right) = 0.$$

Proof. Let $r_n = C^{-1}\sqrt{\ln n / n}$ in which C has been defined in (2.47) and let $\mathcal{N}_{r_n}(\Theta)$ denote the r_n-covering set of Θ. As before, for any θ let β_θ represents the closest element of $\mathcal{N}_{r_n}(\Theta)$ to θ. Observe that,

$$\begin{aligned}
\text{RHS} &\coloneqq \left| Z^\top \Pi_n(\theta) Z - tr\{\Pi_n(\theta)\} - Z^\top \Pi_n(\beta_\theta) Z + tr\{\Pi_n(\beta_\theta)\} \right| \\
&= \left| \langle \Pi_n(\theta) - \Pi_n(\beta_\theta), ZZ^\top + I_n \rangle \right| \leq \|\Pi_n(\theta) - \Pi_n(\beta_\theta)\|_{2\to 2} \left\| ZZ^\top + I_n \right\|_{\mathcal{S}_1} \\
&\overset{(A)}{\leq} C \|\theta - \beta_\theta\|_{\ell_2} \left\| ZZ^\top + I_n \right\|_{\mathcal{S}_1} \leq C r_n \left\| ZZ^\top + I_n \right\|_{\mathcal{S}_1} = \sqrt{\frac{\ln n}{n}} \left(n + \|Z\|_{\ell_2} \right),
\end{aligned}$$

in which $\|\cdot\|_{\mathcal{S}_1}$ stands for the nuclear norm (absolute sum of eigenvalues). Note that the obtained upper bound does not depend on θ (uniform upper bound). Moreover, based upon Hanson-Wright inequality there is $c > 0$ for which $\left(n + \|Z\|_{\ell_2}\right) \leq 3n$ with probability at least $1 - \exp(-cn)$. Thus, RHS $\geq 3\sqrt{n \ln n}$ with probability at most $\exp(-cn)$. Hence, as $n \to \infty$ we get

$$\Pr\left(\begin{array}{l} \sup_{\theta\in\Theta} \left|Z^\top \Pi_n(\theta) Z - tr\{\Pi_n(\theta)\}\right| \geq \\ \sup_{\theta\in\mathcal{N}_{r_n}(\Theta)} \left|Z^\top \Pi_n(\theta) Z - tr\{\Pi_n(\theta)\}\right| + 3\sqrt{n \ln n} \end{array} \right) \to 0. \tag{2.48}$$

In the sequel we find a tight upper bound on $\sup_{\theta\in\mathcal{N}_{r_n}(\Theta)} \left|Z^\top \Pi_n(\theta) Z - tr\{\Pi_n(\theta)\}\right|$. Applying condition (c) on Hanson-wright inequality and using union bound leads to

$$\Pr\left(\sup_{\theta\in\mathcal{N}_{r_n}(\Theta)} \left|Z^\top \Pi_n(\theta) Z - tr\{\Pi_n(\theta)\}\right| \geq C_0 \sup_{\theta\in\Theta} \|\Pi_n(\theta)\|_{\ell_2} \sqrt{\ln n} \right) \leq \left|\mathcal{N}_{r_n}(\Theta)\right| n^{-m},$$

for some constant $C_0 > 0$ depending on m. Notice that $\sup_{\theta\in\Theta} \|\Pi_n(\theta)\|_{\ell_2} \leq M\sqrt{n}$ according to condition (a). Moreover, as we argued in (2.29), $\left|\mathcal{N}_{r_n}(\Theta)\right| = o(n^m)$. Thus,

$$\lim_{n\to\infty} \Pr\left(\sup_{\theta\in\mathcal{N}_{r_n}(\Theta)} \left|Z^\top \Pi_n(\theta) Z - tr\{\Pi_n(\theta)\}\right| \geq C_0 M \sqrt{n \ln n} \right) = 0.$$

Replacing the last inequality into (2.48) concludes the proof. □

The following result gives an upper bound on the *Kullback-Leibler* divergence of two zero mean multivariate Gaussian distributions respectively associated with the two covariance matrices $K_n(\theta_i)$, $i = 1, 2$. Such upper bound is extremely useful for establishing Theorem 2.3.

Proposition 2.8. Choose $\theta_1, \theta_2 \in \Theta$ in such a way that $\|\theta_2 - \theta_1\|_{\ell_2} \leq \Lambda_{\min}/(2\mathfrak{D}_{\max})$. Let $P_i, i = 1, 2$, denotes the associated probability distribution to a zero mean Gaussian vector with the covariance matrix $K_n(\theta_i) \in \mathbb{R}^{n\times n}$, $i = 1, 2$. Then,

$$D(P_1 \parallel P_2) \leq 2n \left(\frac{\mathfrak{D}_{\max}}{\Lambda_{\min}} \|\theta_2 - \theta_1\|_{\ell_2} \right)^2.$$

Proof. For any symmetric matrix $A \in \mathbb{R}^{n\times n}$, let $\lambda_i(A), i = 1, \ldots, n$, denotes its i^{th} eigenvalue

in decreasing order. The *von Neumann's trace inequality* [Mir75] yields

$$\begin{aligned} D(P_1 \parallel P_2) &= \langle K_n^{-1}(\theta_2), K_n(\theta_1)\rangle - n + \ln\left(\frac{\det K_n(\theta_2)}{\det K_n(\theta_1)}\right) \\ &\leq Q \coloneqq \sum_{j=1}^{n}\left\{\frac{\lambda_j(K_n(\theta_1))}{\lambda_j(K_n(\theta_2))} - 1 - \ln\frac{\lambda_j(K_n(\theta_1))}{\lambda_j(K_n(\theta_1))}\right\}. \end{aligned}$$

We finish the proof by acquiring a proper upper bound on Q. Define $f : (0,\infty) \mapsto \mathbb{R}$ by $f(x) = |x - 1 - \ln x|$. Applying the second order Taylor's expansion around $x = 1$ shows that $f(x) \leq 2(x-1)^2$ for $|x-1| \leq 1/2$.

Claim 1. The succeeding inequality holds for any $j = 1, \ldots, n$.

$$\left|\frac{\lambda_j(K_n(\theta_1))}{\lambda_j(K_n(\theta_2))} - 1\right| \leq \frac{\mathfrak{D}_{\max}}{\Lambda_{\min}}\|\theta_2 - \theta_1\|_{\ell_2} \leq \frac{1}{2}.$$

Claim 1 provides the key tool to control Q from above.

$$\begin{aligned} Q &= \sum_{j=1}^{n} f\left[\frac{\lambda_j\{K_n(\theta_1)\}}{\lambda_j\{K_n(\theta_2)\}}\right] \leq 2\sum_{j=1}^{n}\left[\frac{\lambda_j\{K_n(\theta_1)\}}{\lambda_j\{K_n(\theta_2)\}} - 1\right]^2 \leq 2\sum_{j=1}^{n}\left(\frac{\mathfrak{D}_{\max}}{\Lambda_{\min}}\|\theta_2 - \theta_1\|_{\ell_2}\right)^2 \\ &= 2n\left(\frac{\mathfrak{D}_{\max}}{\Lambda_{\min}}\|\theta_2 - \theta_1\|_{\ell_2}\right)^2. \end{aligned}$$

In conclusion, we substantiate Claim 1. The first inequality can be established using Proposition 2.5 and the second one is obvious.

$$\begin{aligned} \left|\frac{\lambda_j\{K_n(\theta_1)\}}{\lambda_j\{K_n(\theta_2)\}} - 1\right| &= \left|\frac{\lambda_j\{K_n(\theta_1)\} - \lambda_j\{K_n(\theta_2)\}}{\lambda_j\{K_n(\theta_2)\}}\right| \leq \frac{\|K_n(\theta_2) - K_n(\theta_1)\|_{2\to 2}}{\lambda_n\{K_n(\theta_2)\}} \\ &\leq \frac{\mathfrak{D}_{\max}\|\theta_2 - \theta_1\|_{\ell_2}}{\Lambda_{\min}}. \end{aligned}$$

□

Now we demonstrate the asymptotic normality of the normalized quadratic Gaussian forms. We exploit this fact in the proof of Theorem 2.4.

Lemma 2.4. For $n \in \mathbb{N}$, let $Z_n \in \mathbb{R}^n$ be a standard Gaussian vector and let $A_n \in \mathbb{R}^{n\times n}$. Then,

$$\Psi_n \coloneqq \left\{\frac{Z_n^\top A_n Z_n - tr(A_n)}{\|A_n\|_{\ell_2}}\right\} \xrightarrow{d} \mathcal{N}(0,2),$$

provided that $\lim_{n\to\infty}\|A_n\|_{\ell_2}^{-1}\|A_n\|_{2\to 2} = 0$.

Proof. Let Ψ_∞ be a zero mean Gaussian random variable with variance 2. So, $\ln \mathrm{E} \exp(t\Psi_\infty) = t^2$ for any $t \in \mathbb{R}$. The basic properties of the quadratic forms of Gaussian vectors yields

$$\begin{aligned}
\ln \mathrm{E} \exp(t\Psi_n) &= -\frac{1}{2} \ln \det\left(I_n - 2t \frac{A_n}{\|A_n\|_{\ell_2}} \right) - \frac{t\, tr(A_n)}{\|A_n\|_{\ell_2}} \\
&= -\frac{1}{2} \sum_{j=1}^{n} \left\{ \ln\left(1 - \frac{2t\lambda_j(A_n)}{\|A_n\|_{\ell_2}} \right) + \frac{2t\lambda_j(A_n)}{\|A_n\|_{\ell_2}} \right\} \\
&\overset{(A)}{=} \sum_{j=1}^{n} \left[\left(\frac{t\lambda_j(A_n)}{\|A_n\|_{\ell_2}} \right)^2 + o\left\{ \left(\frac{t\lambda_j(A_n)}{\|A_n\|_{\ell_2}} \right)^2 \right\} \right] \to t^2, \quad \text{as } n \to \infty.
\end{aligned}$$

Here (A) follows from expanding $\ln(1-x)$ around 1 for infinitesimal x (since $\lambda_j(A_n)/\|A_n\|_{\ell_2}$ vanishes as $n \to \infty$). Consequently, Ψ_n converges in distribution to Ψ_∞ by the continuity theorem of moment generating functions. □

The last result of this section studies the shrinkage behaviour of the partial derivatives of Matern covariance function with respect to its fractal index. In turns out to be useful for corroborating the part (a) of Remark 2.2.

Lemma 2.5. Let $K_\nu : \mathbb{R}^d \mapsto \mathbb{R}$ be a geometric anisotropic (Recall from Definition 2.1) Matern correlation function given by

$$K_\nu(r) = \frac{2^{1-\nu}}{\Gamma(\nu)} \mathcal{K}_\nu\left(\sqrt{r^\top A r} \right),$$

in which $\mathcal{K}_\nu$ stands for the modified Bessel function of the second kind and A satisfies the condition 2.7. Then for any $\beta \in \mathbb{N}$ and $m \in \mathbb{N}$, there is a bounded constant $C_{\beta,A}$ such that

$$\left| \frac{\partial^m}{\partial \nu^m} K_\nu(r) \right| \leq \frac{C_{\beta,A}}{1 + \|r\|_{\ell_2}^{2\beta}}, \quad \forall\, r = (r_1, \ldots, r_d) \in \mathbb{R}^d. \tag{2.49}$$

CHAPTER 3

Local Inversion-Free (LIF) Covariance Estimation

3.1 Introduction

In Chapter 2, we have studied the computationally efficient IF covariance estimation algorithm proposed in Anitescu et al. [ACS16]. The IF loss function is independent of the precision matrix and can be computed in $\mathcal{O}\left(n^2\right)$ flops. When the covariance matrix has a bounded condition number, studies in [ACS16, KSN16] have established the consistency and asymptotic normality of IF estimate. However in the presence of spatial correlation, the condition number often grows without bound with the sample size, specifically if large number of samples are collected in a fixed and bounded domain. In other words, the inversion-free algorithm is ineffective, especially for large data sets, without a proper preprocessing step for reducing the strong correlation between the samples. Some dependence reduction schemes, which we refer to as *preconditioning*, are introduced in [Che13,SCA12] and chapter 3 of [Lee12] to decrease the condition number of the covariance matrix.

In this chapter we present a versatile and computationally efficient class of inversion-free optimization algorithms that can open new horizons to broader classes of the covariance estimators in the geostatistics. Our proposed loss function, which will be referred to as the *Local-Inversion Free (LIF)*, is closely connected to the local moment matching procedure applied to the preconditioned data. For constructing the LIF loss we split the observed samples into b_n (possibly overlapping) clusters and take the weighted average of the IF loss functions for the different bins. The preconditioning is crucial for statistical efficiency of the LIF algorithm as it significantly reduces the correlation between the distant clusters. Note that the LIF procedure comprises a rich and flexible class of estimation algorithms, depending on b_n, size, and the shape of each cluster. For instance the inversion-free loss in [ACS16] is indeed an element of the LIF class in which there is only a single cluster, i.e., $b_n = 1$. Furthermore, the quadratic variation-based approach proposed by Anderes [And10]

is a special instance in the LIF class (in the other extreme scenario of $b_n = n$). So the LIF class can be viewed as a spectrum of computationally scalable and statistically consistent algorithms building a bridge between the well-known approaches in the literature. Finally, exploiting the divide and conquer strategy can significantly expedite the estimation procedure, while preserving the key statistical properties. Strictly speaking, the LIF loss can be evaluated in order n^2/b_n operations. We present fairly comprehensive numerical studies on large data to examine the computational advantage of our method.

With an appropriate choice of b_n, the LIF function can be much easier to compute than the log-likelihood. However multiple evaluation of such function, which is necessary for any gradient-based optimization algorithm, is still formidable for large data sets, particularly on a single computing core. Another advantage of the our proposed estimator is a fairly simple parallel implementation of evaluating the objective function on a shared or distributed memory system, that is imperative for high resolution spatial processes.

We also analyze the large sample properties of the LIF algorithm such as $\sqrt{n}$-consistency and asymptotic normality given one realization of the GP on a d-dimensional irregular lattice. The covariance function of the GP is assumed to be isotropic Matern with known smoothness parameter. This class of covariance functions has numerous application in geostatistics and has attracted considerable attention on the theoretical side of the field (see e.g., [And10, KSN08, Ste12, Zha04]). Fixed domain asymptotics are generally more realistic as Gaussian random fields are rarely stationary over a large region. It is known that not all the covariance parameters are consistently estimable under the fixed domain regime, e.g., [Zha04]. It is also worthwhile to mention that studying the infill asymptotics for covariance estimation is much more difficult comparing to the increasing domain, and there are only few papers in the literature, particularly for GPs observed on multidimensional irregular grids (see e.g., [KSN08, WL$^+$11, Zha04]).

Our contributions. The aim of this chapter is two-fold: presenting the local inversion-free spectrum of covariance estimation algorithms, and investigating its fixed-domain asymptotic properties for d-dimensional isotropic Matern GP, observed on an irregular grid $\mathcal{D}_n$ with n points. We now summarize the main contribution of this chapter with further details.

1. We combine the divide and conquer technique and the inversion-free algorithm [ACS16] to propose the flexible and computationally efficient class of LIF covariance estimation algorithm. The proposed loss function can be effectively optimized as

- Breaking the preconditioned sample into b_n bins, the LIF loss function (and its gradient with respect to the covariance parameters) can be computed in order n^2/b_n operations.
- Evaluating the LIF loss can be accelerated on both shared and distributed memory machines. Specifically on a machine with p cores, implementing each iteration of the gradient descent algorithm can be p times faster.

2. We know from [Zha04] that for the isotropic Matern GP, the variance and the range parameter are not separately consistently estimable when $d \leq 3$. Thus we only concentrate on estimating the microergodic parameter which is of great interest in the literature. We show that under some regularity conditions on $\mathcal{D}_n$ and for any binning scheme, all the stationary points of the LIF objective function are concentrated around the true parameter on a ball of radius $\mathcal{O}(\sqrt{n^{-1}\log n})$, with high probability. We also substantiate the asymptotic normality of this estimate. In other words, the LIF loss does not sacrifice the asymptotic rate for increasing the speed and memory efficiency.

3. A fairly comprehensive set of synthetic numerical experiments are conducted for assessing the role of preconditioning, the irregularity of sampling locations, and the clustering scheme in the performance of the LIF estimate. Our simulation studies corroborate the developed asymptotic theory also reveals the stability of the LIF estimate with respect to the size and shape of the clusters. We also demonstrate the efficiency of our proposed algorithm for data sets of 2.5×10^5 data points.

Plan of the chapter. Section 3.2 describes the geometry of sampling sites, the preconditioning, and the IF method. In Section 3.3, we propose the family of the LIF loss functions and introduce an efficient parallel technique for evaluating such functions. Section 3.4 establishes the fixed domain asymptotic properties of the LIF algorithm such as $\sqrt{n}$-consistency and the asymptotic normality, given samples in a d-dimensional space with $d \leq 3$. In Section 3.5 we present a series of simulation studies to assess the performance of the LIF estimator. Section 3.6 serves as the conclusion and discusses future directions. We substantiate the main results of this chapter in Section 3.7. Finally, Sections 3.8.1 and 3.8 not only contain some auxiliary technicalities which are crucial in Section 3.7, also present a comprehensive sensitivity analysis of the correlation matrix of the preconditioned data with respect to the range parameter, which may becomes useful for the asymptotic analysis of other estimation algorithms in geostatistics.

Notation. For the convenience of the reader, we collect standard pieces of notation here. $j = \sqrt{-1}$ denotes the imaginary unit. Boldface symbols denote vectors. $\wedge$ and $\vee$ stand for the minimum and maximum operators. For any $m \in \mathbb{N}$, $\mathbf{0}_m$ denotes all zeros column vector of length m. Furthermore, for any $p \in \{1, \ldots, m\}$, $\boldsymbol{e}_p$ denotes the unit vector along the p^{th} coordinate. If $\boldsymbol{u}$ and $\boldsymbol{v}$ are vectors of length m, then $\boldsymbol{u}^{\boldsymbol{v}}$ is a compact way of referring to $\Pi_{i=1}^{m} u_i^{v_i}$ (we define 0^0 to be 1). For matrices A and B of the same size, by writing $A \succeq B$, we mean that $A - B$ is a symmetric positive semi-definite matrix. Furthermore, $\langle A, B \rangle := tr(A^\top B)$ refers to their trace inner product. We use various types of matrix norms on $A \in \mathbb{R}^{n \times n}$ in this chapter. For any $p \in [1, \infty)$, $\|A\|_{\ell_p} := \left(\sum_{i,j} \left|A_{ij}\right|^p\right)^{1/p}$ stands for the element-wise $p-$norm of A. We also write $\|A\|_{2\to 2}$ to denote the usual operator norm (largest singular value) of A. Moreover $\|A\|_{\mathcal{S}_1}$ represents the sum of the singular values of A, which is called the nuclear norm. For $\Omega_1, \Omega_2 \subset \mathbb{R}^m$, $\text{dist}(\Omega_1, \Omega_2) := \inf_{\omega_i \in \Omega_i,\, i=1,2} \|\omega_1 - \omega_2\|_{\ell_2}$ refers to their mutual distance with respect to the Euclidean norm. Moreover, for $\mathcal{A} \subset \mathbb{R}^m$ and $r > 0$, $\mathcal{N}_r(\mathcal{A})$ denotes a subset of $\mathcal{A}$ (of minimal size) such that for each $a \in \mathcal{A}$, $\text{dist}(\{a\}, \mathcal{N}_r(\mathcal{A})) \leq r$. The cardinality of such set is called the *covering number* of $\mathcal{A}$. We also write $\text{diam}(\Omega) = \sup_{\omega_1, \omega_2 \in \Omega} \|\omega_2 - \omega_1\|_{\ell_2}$ to denote the diameter of a bounded set $\Omega \subset \mathbb{R}^m$. For a symmetric, positive semi-definite $A \in \mathbb{R}^{n \times n}$ with spectral decomposition $A = U \Lambda U^\top$, $\sqrt{A} := U \Lambda^{1/2} U^\top$ represents its symmetric square root. For two non-negative sequences $\{a_m\}_{m=1}^{\infty}$ and $\{b_m\}_{m=1}^{\infty}$, we write $a_m \asymp b_m$ if there are strictly positive and bounded scalars $C_{\min}, C_{\max}$ such that $C_{\min} \leq \lim_{m \to \infty} a_m / b_m \leq C_{\max}$. Moreover, $a_m \lesssim b_m$ refers to the case that $a_m / b_m \leq C_{\max} < \infty$ as $m \to \infty$. Lastly, $\mathcal{K}_\nu(\cdot)$ and $\Gamma(\cdot)$ respectively represent the modified Bessel function of the second kind of order ν and the Gamma function.

3.2 Problem Formulation and Background

Let $\mathcal{D}$ be a bounded subset of $\mathbb{R}^d$ such as $[0,1]^d$. Consider a zero mean, real valued, and stationary GP G on $\mathcal{D}$. The strictly positive quantity ϕ_0 refers to the variance of G and ρ_0 denotes the unknown correlation parameters (which will be referred to as the range parameters). For instance if G is a geometric anisotropic process on $\mathcal{D}$, then there are a fully known covariance function K and a matrix $\rho_0 \in \mathbb{R}^{d \times d}$ such that

$$\mathbb{E} G(\boldsymbol{s}) G(\boldsymbol{t}) = \phi_0 K\left(\left\|\rho_0^{-1}(\boldsymbol{t} - \boldsymbol{s})\right\|_{\ell_2}\right), \quad \forall\, \boldsymbol{s}, \boldsymbol{t} \in \mathcal{D}$$

Throughout this chapter, we assume that ρ_0 belong to a compact, connected space Θ_0 (with respect to the Euclidean distance). We also restrict d to be less than or equal 3. The objective is to estimate the covariance parameters, given n samples (from one realization)

of G at the locations $\mathcal{D}_n = \{s_1, \ldots, s_n\} \subset \mathcal{D}$. As the first step we precisely formulate $\mathcal{D}_n$. $\mathcal{D}_n$ is called a d-dimensional regular (rectangular) lattice with $n = N^d$ points, if $\mathcal{D}_n = \{1/N, \ldots, 1\}^d$. In such a lattice the smallest distance between neighboring locations decreases with the rate of N^{-1}. This fact provides a clue for extending the notion of the regular lattice.

Assumption 3.1. Let $\mathcal{D}_n \subset \mathcal{D}$ be a set of size n. For any $\boldsymbol{s} \in \mathcal{D}_n$, let $r_{\boldsymbol{s},i}$ denotes the distance from $\boldsymbol{s}$ to its i^{th} closest neighbor in $\mathcal{D}_n \setminus \{\boldsymbol{s}\}$. There are positive scalars $C_{\min}$ and $C_{\max}$ (depending on d) such that

$$C_{\min}\left(\frac{i}{n}\right)^{\frac{1}{d}} \leq r_{\boldsymbol{s},i} \leq C_{\max}\left(\frac{i}{n}\right)^{\frac{1}{d}}, \quad \forall\, \boldsymbol{s} \in \mathcal{D}_n, \text{ and } i = 1, \cdots, (n-1). \tag{3.1}$$

Assumption 3.1, which is obviously satisfied by a d-dimensional regular lattice, has been introduced in Section 3 of [Lee12]. It generalizes the concept of regular lattice in two aspects. First, on the contrary to the number of points in a regular lattice, there is no restriction on n. Moreover, $\mathcal{D}$ is not restricted to be $[0,1]^d$. Indeed $\mathcal{D}$ can even be the union of a finite number of connected components, as long as each of them satisfy condition (3.1) and encompasses a non-vanishing fraction of samples, as n tends to infinity.

3.2.1 Preconditioning

As we argued in Section 3.1, controlling the strong spatial dependence between the observed samples $\{G(\boldsymbol{s}_1), \ldots, G(\boldsymbol{s}_n)\}$ is essential for reducing the condition number of the covariance matrix. More precisely, preconditioning is a surjective linear mapping from $\mathbb{R}^n$ to $\mathbb{R}^{n'}$ for some $n' \leq n$, designed to reduce the correlation between the transformed samples. Note that as $n' \leq n$, we may lose a meager fraction of information in the preconditioning procedure, which can be viewed as the price we pay for reducing the correlation in the transformed data. Various types of preconditioners has been studied for GPs observed on regular and irregular lattices in the literature (see e.g., [Che13, Lee12, SCA12]). Now, we precisely formulate the preconditioner proposed by [Lee12] for irregularly spaced observations. Before proceeding further, it would be more convenient to define $N := \lfloor n^{1/d} \rfloor$.

Definition 3.1. Let $m \in \mathbb{N}$ (which does not grow with the sample size). For any $\boldsymbol{s} \in \mathcal{D}_n$, consider $\mathcal{N}_m(\boldsymbol{s}) \subset \mathcal{D}_n$ and a set of real coefficients $\{a_{m,\boldsymbol{s}}(\boldsymbol{t}) : \boldsymbol{t} \in \mathcal{N}_m(\boldsymbol{s})\}$ satisfying the following conditions:

1. The maximum distance between $\boldsymbol{s} \in \mathcal{D}_n$ and the other points in $\mathcal{N}_m(\boldsymbol{s})$ is of order N^{-1}. Namely, $\|\boldsymbol{t} - \boldsymbol{s}\|_{\ell_2} \lesssim 1/N$ for any $\boldsymbol{t} \in \mathcal{N}_m(\boldsymbol{s})$.

2. For any $\boldsymbol{r} \in \mathbb{Z}^d$ with non-negative entries and $\|\boldsymbol{r}\|_{\ell_1} < m$, $\sum_{\boldsymbol{t} \in \mathcal{N}_m(\boldsymbol{s})} a_{m,\boldsymbol{s}}(\boldsymbol{t})(\boldsymbol{t} - \boldsymbol{s})^{\boldsymbol{r}} = 0$.

3. There is a vector $\boldsymbol{r} \in \{0, 1, \ldots\}^d$ with $\|\boldsymbol{r}\|_{\ell_1} \geq m$ such that $\sum_{\boldsymbol{t} \in \mathcal{N}_m(\boldsymbol{s})} a_{m,\boldsymbol{s}}(\boldsymbol{t})(\boldsymbol{t}-\boldsymbol{s})^{\boldsymbol{r}} \neq 0$.

4. $\sum_{\boldsymbol{t} \in \mathcal{N}_m(\boldsymbol{s})} a_{m,\boldsymbol{s}}^2(\boldsymbol{t}) = 1$ and $a_{m,\boldsymbol{s}}(\boldsymbol{t}) \neq 0$ for any $\boldsymbol{t} \in \mathcal{N}_m(\boldsymbol{s})$.

We define the *preconditioned process of order m*, which is represented by G_m, as

$$G_m(\boldsymbol{s}) \coloneqq N^{\nu} \sum_{\boldsymbol{t} \in \mathcal{N}_m(\boldsymbol{s})} a_{m,\boldsymbol{s}}(\boldsymbol{t}) G(\boldsymbol{t}), \quad \forall\, \boldsymbol{s} \in \mathcal{D}. \tag{3.2}$$

Apart from the first condition in Definition 3.1, No other restriction is imposed on the choice of $\mathcal{N}_m(\boldsymbol{s})$, $\boldsymbol{s} \in \mathcal{D}_n$. However, constructing $\mathcal{N}_m(\boldsymbol{s})$ by the nearest neighbors of $\boldsymbol{s}$ is the most common setting in this chapter. Because of the first condition, the preconditioned process is approximately proportional to the m-th derivative of G at $\boldsymbol{s}$, for large N. We also normalize the coefficients $\{a_{m,\boldsymbol{s}}(\boldsymbol{t}) : \boldsymbol{t} \in \mathcal{N}_m(\boldsymbol{s})\}$ by their Euclidean norm to uniformly control the magnitude of G_m over $\mathcal{D}_n$. Moreover, for reducing ambiguity in the definition of G_m, $\mathcal{N}_m(\boldsymbol{s})$ is chosen to be a minimal set, with respect to the inclusion ordering, satisfying the conditions in Definition 3.1. The cardinality of $\mathcal{N}_m(\boldsymbol{s})$ depends on d, m and the geometric structure of neighboring observations around $\boldsymbol{s}$ in $\mathcal{D}_n$ and may vary across $\mathcal{D}_n$. The reader can deduce from simple combinatorial tricks that the second condition in Definition 3.1 is translated as $\binom{d+m-1}{d}$ linear constraints on set of coefficients $\{a_{m,\boldsymbol{s}}(\boldsymbol{t}) : \boldsymbol{t} \in \mathcal{N}_m(\boldsymbol{s})\}$. This fact gives a rough estimate of the size of $\mathcal{N}_m(\boldsymbol{s})$. It is also noteworthy to mention that G_m is a non-stationary process, particularly for irregular lattices.

Remark 3.1. The preconditioning method for the d-dimensional regular lattices $\mathcal{D}_n = \{1/N, \ldots, 1\}^d$ has been studied in Stein et al. [SCA12]. Discarding the boundary points of $\mathcal{D}_n$, the preconditioned process is constructed on $\mathcal{D}_n^{\circ} = \{(m+1)/N, \ldots, 1 - m/N\}^d$ by $m-$times application of the discrete Laplace operator. More specifically, the preconditioner is recursively defined as the following.

$$\begin{aligned} &G_0(\boldsymbol{s}) = N^{\nu} G(\boldsymbol{s}), \quad \forall\, \boldsymbol{s} \in \mathcal{D}_n, \\ &G_{2k}(\boldsymbol{s}) = \sum_{r=1}^{d} \left[G_{2k-2}\left(\boldsymbol{s} + \frac{\boldsymbol{e}_r}{N}\right) - 2G_{2k-2}(\boldsymbol{s}) + G_{2k-2}\left(\boldsymbol{s} - \frac{\boldsymbol{e}_r}{N}\right) \right], \quad \boldsymbol{s} \in \mathcal{D}_n^{\circ},\ k = 1, \ldots, m. \end{aligned} \tag{3.3}$$

For avoiding unnecessary algebraic complexity in Eq. (3.3), the preconditioning coefficients have not been normalized to be of norm one. It can be shown that after proper normalization, G_{2m} admits the conditions of Definition 3.1 with order $2m$. Simply put, (3.3) gives a recursive way of constructing the preconditioned process of even orders for regular lattices.

3.2.2 The IF Algorithm

With a precise formulation of the preconditioned GP in hand, we now present the IF algorithm. Let Y_m represent the column vector of the preconditioned samples, i.e., $Y_m = [G_m(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{D}_n]^\top$. We use K_m to denote the normalized covariance function of G_m by factor ϕ_0. K_m can be easily expressed in terms of the covariance function of G and the preconditioning coefficients.

$$K_m(\boldsymbol{s},\boldsymbol{t};\rho_0) = \frac{\mathbb{E}G_m(\boldsymbol{s})G_m(\boldsymbol{t})}{\phi_0} = N^{2\nu} \sum_{\boldsymbol{s}' \in \mathcal{N}_m(\boldsymbol{s})} \sum_{\boldsymbol{t}' \in \mathcal{N}_m(\boldsymbol{t})} a_{m,\boldsymbol{s}}(\boldsymbol{s}') a_{m,\boldsymbol{t}}(\boldsymbol{t}') K(\boldsymbol{t}' - \boldsymbol{s}').$$

We also use $\phi_0 K_{n,m}(\rho_0)$ to denote the covariance matrix of Y_m. That is

$$\mathbb{E}Y_m Y_m^\top = \phi_0 K_{n,m}(\rho_0) \coloneqq \phi_0 \left[K_m(\boldsymbol{s},\boldsymbol{t};\rho_0)\right]_{\boldsymbol{s},\boldsymbol{t} \in \mathcal{D}_n}. \tag{3.4}$$

Recall that ρ_0 lies in a compact and connected space Θ_0. The IF estimator [ACS16] of the covariance parameters (ϕ_0, ρ_0) is given by

$$\left(\hat{\phi}_n, \hat{\rho}_n\right) = \underset{\phi>0, \rho \in \Theta_0}{\arg\max} \left\{ \phi Y_m^\top K_{n,m}(\rho)\, Y_m - \frac{\phi^2}{2} \left\| K_{n,m}(\rho) \right\|_{\ell_2}^2 \right\}. \tag{3.5}$$

The loss function in (3.5) does not depend on the Cholesky factorization of $K_{n,m}$ and can be evaluated in order n^2 flops even for the irregularly spaced observations. Furthermore, storing the whole matrix $K_{n,m}$ is not necessary for computing the objective function and its directional derivatives. More specifically, storing Y_m and $\mathcal{D}_n$, which needs $\mathcal{O}(n)$ storage, suffices for estimating the covariance parameters. Finally (3.5) can be reformulated as a moment matching minimization problem.

$$\left(\hat{\phi}_n, \hat{\rho}_n\right) = \underset{\phi>0, \rho \in \Theta_0}{\arg\min} \left\| Y_m Y_m^\top - \phi K_{n,m}(\rho) \right\|_{\ell_2}.$$

3.3 The LIF Algorithm

In this section we build a versatile spectrum of scalable covariance estimation algorithms upon the IF approach introduced in Section 3.2.2 and the block diagonal approximation of the covariance matrix of the preconditioned data $K_{n,m}(\rho)$. The block diagonal sparsification of $K_{n,m}(\rho)$ can speed up the method proposed in [ACS16] without sacrificing the asymptotic rate.

We previously used $Y_m = [G_m(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{D}_n]^\top$ to denote the column vector of the precondi-

tioned samples of order m. Let $\mathcal{B} = \{B_t : t = 1\ldots, b_n\}$ be a partition of $\mathcal{D}_n$ into b_n bins, i.e., $B_i \cap B_j = \emptyset$ for distinct $i, j \in \{1, \ldots, b_n\}$ and $\cup_{t=1}^{b_n} B_t = \mathcal{D}_n$. We write $Y_{B_t,m} = [G_m(\boldsymbol{s}) : \boldsymbol{s} \in B_t]^\top$ to represent the column vector of the preconditioned data in B_t, $t = 1\ldots, b_n$. Furthermore let $\phi_0 K_{B_t,m}(\rho_0)$ denote the covariance matrix of $Y_{B_t,m}$. Namely,

$$\mathbb{E} Y_{B_t,m} Y_{B_t,m}^\top = \phi_0 K_{B_t,m}(\rho_0) \coloneqq \phi_0 \left[K_m(\boldsymbol{s}, \boldsymbol{t}; \rho_0)\right]_{\boldsymbol{s},\boldsymbol{t} \in B_t}, \quad \forall\, t = 1\ldots, b_n, \tag{3.6}$$

in which $\phi_0 K_m(\cdot, \cdot, \rho_0)$ stands for the covariance function of G_m with the parameters (ϕ_0, ρ_0). The LIF objective function associated to a partitioning scheme $\mathcal{B}$ is constructed by summing the IF loss functions corresponding to B_t's over $\mathcal{B}$. The unknown covariance parameters are estimated by maximizing the LIF function. Strictly speaking

$$\left(\hat{\phi}_{n,\mathcal{B}}, \hat{\rho}_{n,\mathcal{B}}\right) = \underset{\phi>0, \rho \in \Theta_0}{\arg\max} \left\{ \sum_{t=1}^{b_n} \left(\phi Y_{B_t,m}^\top K_{B_t,m}(\rho)\, Y_{B_t,m} - \frac{\phi^2}{2} \left\| K_{B_t,m}(\rho) \right\|_{\ell_2}^2 \right) \right\}, \tag{3.7}$$

in which $\hat{\phi}_{n,\mathcal{B}}$ and $\hat{\rho}_{n,\mathcal{B}}$ respectively denote the estimated variance and the range parameters. For the trivial partition $\mathcal{B} = \{\mathcal{D}_n\}$, the optimization problem (3.7) is exactly same as the IF algorithm. Note that the objective function in Eq. (3.7) can be evaluated in $\sum_{t=1}^{b_n} |B_t|^2$ floating point operations. For instance if all $|B_t|$'s have the same order (as n grows), then $\sum_{t=1}^{b_n} |B_t|^2 \asymp n^2/b_n$. Thus in such a case, the LIF objective function can be computed almost b_n times faster than the one in (3.5). In Section 3.5, we numerically assess the connection between the partitioning scheme of $\mathcal{D}_n$ and the estimation performance of (3.7).

Remark 3.2. The LIF objective function is much easier to compute than the log-likelihood with a proper choice of b_n and the bins. However, implementing one iteration any gradient-based optimizer for (3.7), such as the BFGS method, can still be very challenging on a single computing core, particularly for large data sets ($n \approx 10^6$ or more), as it may require multiple evaluation of the LIF loss. Thus developing effective parallel schemes for computing the LIF function is a necessity for high resolution spatial GPs. For simplicity assume that all the bins have roughly the same size and we have access to p identical processor with q cores. For any $t = 1, \ldots, b_n$, let $f_t(Y_{B_t,m}; \phi, \rho)$ stands for the IF function, with the parameters (ϕ, ρ), associated to B_t. In the following we introduce a distributed memory parallel scheme for evaluating the LIF function.

1. The master processor assigns a label in $\{1, \ldots, p\}$ to each bin (each processor roughly receives b_n/p bins). More specifically if B_t is labelled as i, then the local memory of processor i stores $G_m(\boldsymbol{s})$, $\mathcal{N}_m(\boldsymbol{s})$, and the preconditioning coefficients $\{a_{m,\boldsymbol{s}}(\boldsymbol{t}) : \boldsymbol{t} \in \mathcal{N}_m(\boldsymbol{s})\}$ for any $\boldsymbol{s} \in B_t$.

2. Inside each processor, the terms $f_t(Y_{B_t,m};\phi,\rho)$ can be evaluated by employing the basic shared memory parallel schemes for computing $\left\|K_{B_t,m}(\rho)\right\|_{\ell_2}$ and $K_{B_t,m}(\rho)\,Y_{B_t,m}$. Finally the master processor aggregatesa the received quantities $\{f_t(Y_{B_t,m};\phi,\rho):\ t=1,\ldots,b_n\}$ from the slave processors to compute the LIF objective function.

Remark 3.3. The LIF class of estimators can be enriched in the two possible ways. First we can drop the assumption that $\{B_t\}_{t=1}^{b_n}$ forms a partition for $\mathcal{D}_n$. Namely, the distinct clusters may not be mutually exclusive. The LIF loss can also be extended by considering a weighted average of the IF functions. Strictly speaking, given a b_n-dimensional vector of strictly positive entries $w\in\mathbb{R}^{b_n}$,

$$\left(\hat{\phi}_{n,\mathcal{B},w},\hat{\rho}_{n,\mathcal{B},w}\right)=\underset{\phi>0,\rho\in\Theta_0}{\arg\max}\left\{\sum_{t=1}^{b_n}w_t\left(\phi Y_{B_t,m}^{\top}K_{B_t,m}(\rho)\,Y_{B_t,m}-\frac{\phi^2}{2}\left\|K_{B_t,m}(\rho)\right\|_{\ell_2}^2\right)\right\}.$$

However throughout this chapter and for simplifying the theoretical analysis, we only consider the case of non-overlapping bins. It will also be assumed that $w_i=1$ for any $i\in\{1,\ldots,b_n\}$.

Remark 3.4. We now introduce an alternative viewpoint to the LIF objective function in (3.7). The block diagonal approximation of $K_{n,m}(\rho)$ corresponding to partitioning scheme $\mathcal{B}$, which is denoted by $K_{n,m}^{\mathcal{B}}(\rho)$, has the canonical role in the new formulation. Choose any $\boldsymbol{s},\boldsymbol{s'}\in\mathcal{D}_n$, and let t,t' denote the index of the elements in $\mathcal{B}$ containing $\boldsymbol{s}$ and $\boldsymbol{s'}$, i.e., $\boldsymbol{s}\in B_t$ and $\boldsymbol{s'}\in B_{t'}$. The entries of $K_{n,m}^{\mathcal{B}}(\rho)$ can be equivalently represented by

$$\left(K_{n,m}^{\mathcal{B}}(\rho)\right)_{\boldsymbol{s},\boldsymbol{s'}}=\left[K_{n,m}(\rho)\right]_{\boldsymbol{s},\boldsymbol{s'}}\mathbb{1}_{\{t=t'\}}. \tag{3.8}$$

Observe that

$$\sum_{t=1}^{b_n}\left\|K_{B_t,m}(\rho)\right\|_{\ell_2}^2=\left\|K_{n,m}^{\mathcal{B}}(\rho)\right\|_{\ell_2}^2,\quad\text{and}\quad\sum_{t=1}^{b_n}Y_{B_t,m}^{\top}K_{B_t,m}(\rho)\,Y_{B_t,m}=Y_m^{\top}K_{n,m}^{\mathcal{B}}(\rho)\,Y_m.$$

These identities provide an alternative form for Eq. (3.7) in terms of $K_{n,m}^{\mathcal{B}}(\rho)$.

$$\left(\hat{\phi}_{n,\mathcal{B}},\hat{\rho}_{n,\mathcal{B}}\right)=\underset{\phi>0,\rho\in\Theta_0}{\arg\max}\left(\phi Y_m^{\top}K_{n,m}^{\mathcal{B}}(\rho)\,Y_m-\frac{\phi^2}{2}\left\|K_{n,m}^{\mathcal{B}}(\rho)\right\|_{\ell_2}^2\right), \tag{3.9}$$

Simply put any member of the LIF class is equivalent to applying the IF procedure on an appropriate block diagonal approximation of the covariance matrix.

We finally present a new formulation for the optimization problem in (3.9) which is more convenient for our theoretical analysis. Due to the quadratic dependence of the LIF loss to ϕ, $\hat{\phi}_{n,\mathcal{B}}$ can be explicitly expressed in terms of $\hat{\rho}_{n,\mathcal{B}}$ as the following:

$$\hat{\phi}_{n,\mathcal{B}} = \frac{Y_m^\top K_{n,m}^{\mathcal{B}}(\hat{\rho}_{n,\mathcal{B}})\, Y_m}{\left\| K_{n,m}^{\mathcal{B}}(\hat{\rho}_{n,\mathcal{B}}) \right\|_{\ell_2}^2}, \quad \text{where} \quad \hat{\rho}_{n,\mathcal{B}} = \underset{\rho\in\Theta_0}{\arg\max} \frac{Y_m^\top K_{n,m}^{\mathcal{B}}(\rho)\, Y_m}{\left\| K_{n,m}^{\mathcal{B}}(\rho) \right\|_{\ell_2}}. \tag{3.10}$$

The profile LIF loss in (3.10) is indeed proportional to the angle between $K_{n,m}^{\mathcal{B}}(\rho)$ and $Y_m Y_m^\top$.

3.4 Fixed Domain Asymptotic Analysis

The core emphasis of this section is to investigate the fixed domain asymptotic properties of the LIF optimization problems (3.10). Throughout this section we assume that G is a real valued GP with *isotropic Matern* covariance function observed on a bounded domain $\mathcal{D} \subset \mathbb{R}^d$ with $d \leq 3$. Strictly speaking, for any $\boldsymbol{s}, \boldsymbol{s'} \in \mathcal{D}$

$$\operatorname{cov}\Big(G(\boldsymbol{s}), G(\boldsymbol{t})\Big) = \frac{\phi_0}{2^{\nu-1}\Gamma(\nu)} \left(\frac{\|\boldsymbol{s}-\boldsymbol{t}\|_{\ell_2}}{\rho_0} \right)^{\nu_0} \mathcal{K}_\nu \left(\frac{\|\boldsymbol{s}-\boldsymbol{t}\|_{\ell_2}}{\rho_0} \right).$$

Here $\nu > 0$ is a known bounded constant controlling the mean squared smoothness of G; larger ν corresponds to smoother GP. The strictly positive scalars ϕ_0 and ρ_0 respectively stand for the variance and the range parameter of G. Despite the complicated form of covariance function, the Matern spectral density has a fairly simple form given by

$$\hat{K}(\boldsymbol{\omega}; \phi_0, \rho_0) = \frac{\phi_0 \rho_0^{-2\nu}}{\pi^{d/2}} \left(\frac{1}{\rho_0^2} + \|\boldsymbol{\omega}\|_{\ell_2}^2 \right)^{-(\nu+d/2)}. \tag{3.11}$$

It has been discussed in [Zha04] that for any bounded region $\mathcal{D} \subset \mathbb{R}^d$ with $d \leq 3$, the Matern covariance models with parameters (ϕ_1, ρ_1) and (ϕ_2, ρ_2) yield absolutely continuous measures whenever $\phi_1 \rho_1^{-2\nu} = \phi_2 \rho_2^{-2\nu}$. In this case, (ϕ_1, ρ_1) and (ϕ_2, ρ_2) are almost surely not distinguishable when observing a single realization of G. In other words, given a single realization of G in $\mathcal{D}$, we are only able to estimate $\phi_0 \rho_0^{-2\nu}$ in (3.11) (which is usually referred to as the *microergodic* parameter).

Remark 3.5. Given one realization of isotropic Matern G at $\mathcal{D}_n$, the distinct pairs of parameters (ϕ_1, ρ_1) and (ϕ_2, ρ_2) are not discernible if $\phi_1 \rho_1^{-2\nu} = \phi_2 \rho_2^{-2\nu}$. Let us concisely explain the reason behind non-separability of (ϕ_1, ρ_1) and (ϕ_2, ρ_2). Let $\mathbb{P}_j(x_1, \ldots, x_n),\ j = 1, 2$

represent the distribution of $[G(\boldsymbol{s}_1,\ldots,\boldsymbol{s}_n)]$ under covariance parameters $\left(\phi_j,\rho_j\right),\ j=1,2$, respectively. Furthermore, we use $p_j(x_1,\ldots,x_n),\ j=1,2$ to denote the density function of $P_j(x_1,\ldots,x_n),\ j=1,2$ and $\rho_n \coloneqq p_2(x_1,\ldots,x_n)/p_1(x_1,\ldots,x_n)$ for referring to the likelihood ratio function. It is known that if $\phi_1\rho_1^{-2\nu}=\phi_2\rho_2^{-2\nu}$ then ρ_n almost surely converges (with respect to probability measure $\mathbb{P}_1$) to a random quantity ρ (see Section 3.2.1 of [IR12] and Section 4.2 of [Ste12]). In addition,

$$\mathbb{P}_1(0<\rho<\infty)=1,\quad \left|\mathbb{E}_{\mathbb{P}_1}\log\rho\right|<\infty,\quad \text{and}\quad \lim_{n\to\infty}\mathbb{E}_{\mathbb{P}_1}\log\rho_n=\mathbb{E}_{\mathbb{P}_1}\log\rho. \tag{3.12}$$

In other words, the likelihood ratio is almost surely bounded as n tends to infinity.

The discussion in Remark 3.5 indirectly implies that (ϕ_1,ρ_1) and (ϕ_2,ρ_2) with $\phi_1\rho_1^{-2\nu}=\phi_2\rho_2^{-2\nu}$ are still non-distinguishable, even from multiple number of i.i.d realizations of G at $\mathcal{D}_n$. In particular given k i.i.d realizations of G at $\mathcal{D}_n$, the likelihood ratio almost surely converges to ρ^k, which satisfies the inequalities in Eq. (3.12). So we solely focus on estimating $\phi_0\rho_0^{-2\nu}$ from one realization of G in our asymptotic analysis.
Recall from Remark 3.4 that $K^{\mathcal{B}}_{n,m}$ stands for the block diagonal approximation of the pre-conditioned data. Define a real valued (stochastic) mapping over Θ_0 by

$$\hat{\phi}_{n,\mathcal{B}}(\rho)\coloneqq\frac{Y_m^\top K^{\mathcal{B}}_{n,m}(\rho)\,Y_m}{\left\|K^{\mathcal{B}}_{n,m}(\rho)\right\|^2_{\ell_2}},\quad \forall\,\rho\in\Theta_0. \tag{3.13}$$

For ease of presentation, we omit the dependence of $\hat{\phi}_{n,\mathcal{B}}(\cdot)$ on m in our notation. It is also obviously apparent from (3.10) that $\hat{\phi}_{n,\mathcal{B}}=\hat{\phi}_{n,\mathcal{B}}(\hat{\rho}_{n,\mathcal{B}})$. Before presenting the main results let us consider a simple extreme example in the LIF class which can reveal a key reason behind the $\sqrt{n}$-consistency of any LIF estimator.

Remark 3.6. Suppose that $\mathcal{B}$ only comprises the singelton sets, i.e. $|B_t|=1$ for any $B_t\in\mathcal{B}$. In this case $\phi K_{B_t,m}(\rho)$ (the covariance matrix of $[G_m(\boldsymbol{s}):\ \boldsymbol{s}\in B_t]^\top$ associated to ϕ and ρ) is a scalar which is approximately proportional to $\phi\rho^{-2\nu}$. More specifically it can be shown that for $B_t=\{\boldsymbol{s}\}$

$$\phi K_{B_t,m}(\rho)=C_{\boldsymbol{s}}\phi\rho^{-2\nu}+\varepsilon_n(\boldsymbol{s},\rho,\phi), \tag{3.14}$$

in which $C_{\boldsymbol{s}}$ is a known scalar, independent of ϕ and ρ, and $\varepsilon_n(\boldsymbol{s},\rho,\phi)$ is a vanishing sequence of n (which also depends on m,d,ν as well). Replacing Eq. (3.14) into Eq. (3.13) leads to

$$\hat{\phi}_{n,\mathcal{B}}(\rho)\rho^{-2\nu}=\left(\frac{\sum_{\boldsymbol{s}\in\mathcal{D}_n}C_{\boldsymbol{s}}G^2_m(\boldsymbol{s})}{\sum_{\boldsymbol{s}\in\mathcal{D}_n}C^2_{\boldsymbol{s}}}\right)+o(1),\quad \forall\,\rho\in\Theta_0. \tag{3.15}$$

$\hat{\phi}_{n,\mathcal{B}}(\rho)\rho^{-2\nu}$ has a simpler representation for regular lattices as $C_{\boldsymbol{s}}$ is constant over $\mathcal{D}_n^\circ$ ($\mathcal{D}_n^\circ$

has been defined in Remark 3.1 and denotes the interior of $\mathcal{D}_n$). Furthermore, the profile LIF loss has (roughly) no dependence to ρ, since

$$\frac{\sum_{t=1}^{b_n} Y_{B_t,m}^\top K_{B_t,m}(\rho)\, Y_{B_t,m}}{\sqrt{\sum_{t=1}^{b_n} \left\| K_{B_t,m}(\rho) \right\|_{\ell_2}^2}} = \frac{\sum_{\boldsymbol{s}\in\mathcal{D}_n} C_{\boldsymbol{s}} G_m^2(\boldsymbol{s})}{\sqrt{\sum \boldsymbol{s} \in \mathcal{D}_n C_{\boldsymbol{s}}^2}} + o(1).$$

Simply put, there is no need to estimate ρ using the profile LIF loss. For any ρ, $\phi_0\rho_0^{-2\nu}$ can indeed be estimated by $\hat{\phi}_{n,\mathcal{B}}(\rho)\rho^{-2\nu}$. This estimator is identical to the one proposed by Anderes [And10]. He also examined the fixed-domain asymptotic properties of (3.15) for regular lattices employing some techniques for studying the quadratic variation of stationary Gaussian spatial processes

The first main result of this section states that for appropriately chosen preconditioning order m, regardless of the choice of $\mathcal{B}$ and ρ, $\hat{\phi}_{n,\mathcal{B}}(\rho)\rho^{-2\nu}$ is a $\sqrt{n}$-consistent estimate of $\phi_0\rho_0^{-2\nu}$.

Theorem 3.1. Let G be observed on a lattice $\mathcal{D}_n$ satisfying Assumption 3.1. Suppose that the preconditioning order m satisfies $m \geq (\nu + d/2)$. For any partition $\mathcal{B}$ of $\mathcal{D}_n$, there exists a bounded positive scalar $C_{\mathcal{B}}$, depending on $m, d, \nu, \Theta_0, \mathcal{B}$ and geometric structure of $\mathcal{D}_n$, such that

$$\mathbb{P}\left(\sup_{\rho\in\Theta_0} \left| \frac{\hat{\phi}_{n,\mathcal{B}}(\rho)\rho^{-2\nu}}{\phi_0\rho_0^{-2\nu}} - 1 \right| \geq C_{\mathcal{B}} \sqrt{\frac{\log n}{n}} \right) \leq \frac{1}{n}, \quad \text{as } n \to \infty. \tag{3.16}$$

Theorem 3.1 establishes (high probability) uniform concentration of $\hat{\phi}_{n,\mathcal{B}}(\rho)\rho^{-2\nu}$ around $\phi_0\rho_0^{-2\nu}$ in a small ball of radius $\mathcal{O}(\sqrt{n^{-1}\log n})$. The $\sqrt{n}$-consistency of the global (or local) maximizers of the LIF objective function is a trivial consequence of Theorem 3.1. It is known that an analogous bound as in Eq. (3.16) holds fo the MLE, regardless of how m is chosen. Namely, the MLE is $\sqrt{n}$-consistent even for raw data, $m = 0$. Thus Theorem 3.1 implicitly says that, for sufficiently decorrelated samples, there are surrogates losses that can be optimized considerably faster than the log-likelihood on a wide range of irregular grids, and without sacrificing the asymptotic rate.

In the case that ν is either known or can be rather precisely estimated, Theorem 3.1 gives a straightforward way of choosing m. For instance the choice of $m = \lceil \nu + 1 \rceil$ is sufficient when G is observed within a two dimensional region. Recall from Remark 3.1 that for the regular lattices, if m' represents the number of times that Laplace operator is applied to data, then the transformed process is a preconditioned GP of order $2m'$. Thus for GPs observed on d-dimensional regular lattices, $m = 2m'$ and so m' should not be smaller than $\nu/2 + d/4$.

Remark 3.7. For the interested reader we present a very concise sketch of the proof of

Theorem 3.1 and the detailed explanation will be postponed to Section 3.7. The bias-variance decomposition has the canonical role in our analysis. Strictly speaking,

$$\begin{aligned}\sup_{\rho\in\Theta_0}\left|\frac{\hat{\phi}_{n,\mathcal{B}}(\rho)\rho^{-2\nu}}{\phi_0\rho_0^{-2\nu}}-1\right| &\leq P_1+P_2\\ &:= \sup_{\rho\in\Theta_0}\left|\frac{\mathbb{E}\hat{\phi}_{n,\mathcal{B}}(\rho)\rho^{-2\nu}}{\phi_0\rho_0^{-2\nu}}-1\right|+\sup_{\rho\in\Theta_0}\left|\frac{\hat{\phi}_{n,\mathcal{B}}(\rho)\rho^{-2\nu}-\mathbb{E}\hat{\phi}_{n,\mathcal{B}}(\rho)\rho^{-2\nu}}{\phi_0\rho_0^{-2\nu}}\right|.\end{aligned}$$

We show that $P_1 = o\left(1/\sqrt{n}\right)$ by employing a novel approach to investigate the large sample properties of the eigenvalues of $K^{\mathcal{B}}_{n,m}(\rho)$. Moreover, P_2 is in fact the supremum of a chi-squared process over Θ_0. The classical chaining argument demonstrates that P_2 is of order $\sqrt{n^{-1}\log n}$, with high probability. We refer the reader to Section 3.8.1 for further details.

Corollary 3.1. Under the same notation and conditions as in Theorem 3.1, the following inequality holds for any stationary point $(\hat{\phi}_{n,\mathcal{B}},\hat{\rho}_{n,\mathcal{B}})$ of the LIF loss (3.7).

$$\mathbb{P}\left(\left|\frac{\hat{\phi}_{n,\mathcal{B}}\hat{\rho}^{-2\nu}_{n,\mathcal{B}}}{\phi_0\rho_0^{-2\nu}}-1\right|\geq C_{\mathcal{B}}\sqrt{\frac{\log n}{n}}\right)\leq\frac{1}{n},\quad \text{as } n\to\infty.$$

It has been argued in [KS+13] that estimating ρ_0 can improve the statistical performance, especially for small n. The first advantage of Corollary 3.1 is that it establishes the consistency of an arbitrary stationary point of the LIF objective function. Allowing the range parameter to be estimated in a large bounded space, which is very crucial in practice, is another advantage of Corollary 3.1.

Remark 3.4 may induce a false impression that the convergence rate of $\hat{\phi}_{n,\mathcal{B}}\hat{\rho}^{-2\nu}_{n,\mathcal{B}}$ is determined by how well the covariance matrix of the preconditioned samples $K_{n,m}(\rho)$ can be approximated by $K^{\mathcal{B}}_{n,m}(\rho)$. However, Corollary 3.1 discloses a somewhat surprising fact that the LIF algorithm is $\sqrt{n}$-consistent, regardless of the choice of $\mathcal{B}$. The fast enough decay rate of the off-diagonal entries of $K_{n,m}(\rho)$ is a heuristic explanation for the $\sqrt{n}$-consistency of the LIF estimator. In other words since $K_{n,m}(\rho)$ can be suitably approximated by any block diagonal matrix induced by a partitioning scheme, splitting the preconditioned data into different bins does not affect the convergence rate of the LIF estimate. However the influence of partitioning scheme may become more apparent in the practical situations with the moderate sample size.

Remark 3.8. It has been discussed in [ACS16] that the global solution of the IF optimization problem, in Eq. (3.5), has the same convergence rate as the MLE, when the covariance matrix of the preconditioned samples has a uniformly bounded condition number over Θ_0.

Such restriction on the covariance matrix rarely holds in practice, unless under some strong conditions on the spectral density and the geometric structure of $\mathcal{D}_n$ (see [SCA12]). However Corollary 3.1 requires much weaker restrictions on the covariance matrix. Two necessary conditions on $K^{\mathcal{B}}_{n,m}(\cdot)$ can be spotted by a meticulous reader in our proof of Theorem 3.1.

1. The largest eigenvalue of $K^{\mathcal{B}}_{n,m}(\cdot)$ should be uniformly bounded over Θ_0. Namely,

$$\max_{\rho\in\Theta_0} \left\| K^{\mathcal{B}}_{n,m}(\cdot) \right\|_{2\to 2} \asymp 1.$$

2. $K^{\mathcal{B}}_{n,m}(\rho)$ must have $\mathcal{O}(n)$ non-negligible positive eigenvalues, for any $\rho \in \Theta_0$. That is,

$$\inf_{\rho\in\Theta_0} \left\| K^{\mathcal{B}}_{n,m}(\rho) \right\|_{\ell_2} \asymp \sqrt{n}.$$

Note that the above conditions do not rule out the existence of near zero eigenvalues and so the conditions number is still allowed to diverge as n tends to infinity. In this regard, our asymptotic understanding can expand the applicability of the LIF algorithm.

Now we introduce the asymptotic distribution of all the stationary point of the LIF loss function.

Theorem 3.2. Under the same notation and conditions as in Theorem 3.1, there exists a bounded sequence $\sigma_{n,\mathcal{B}}$ such that for any stationary point $(\hat{\phi}_{n,\mathcal{B}}, \hat{\rho}_{n,\mathcal{B}})$ of the LIF loss

$$\frac{\sqrt{n}}{\sigma_{n,\mathcal{B}}} \left(\frac{\hat{\phi}_{n,\mathcal{B}} \hat{\rho}^{-2\nu}_{n,\mathcal{B}}}{\phi_0 \rho_0^{-2\nu}} - 1 \right) \xrightarrow{d} \mathcal{N}(0,1).$$

Theorem 3.2 formulates the asymptotic distribution of the LIF algorithm for joint estimation of ϕ_0 and ρ_0. To our knowledge for the MLE, such result has only been appeared in [KS$^+$13]. Note that unlike the full or tapered MLE, in which $\sigma_{n,\mathcal{B}} = \sqrt{2}$ (see Theorem 2 of [WL$^+$11]), here m, d, ν, the geometric structure and the portioning scheme of $\mathcal{D}_n$ also affect the asymptotic standard deviation. We could not obtain a simple closed form expression for $\sigma_{n,\mathcal{B}}$. However its complicated formulation is stated in the proof of Theorem 3.2.

Remark 3.9. We conclude this section with a succinct discussion on the role of Θ_0 in the optimization problem presented in Eq. (3.9). The main results in this section can be

generalized to the following constrained optimization problem

$$\left(\hat{\phi}_{n,\mathcal{B}}, \hat{\rho}_{n,\mathcal{B}}\right) = \underset{\phi>0, \rho\in\Theta_n}{\arg\max} \left(\phi Y_m^\top K_{n,m}^{\mathcal{B}}(\rho)\, Y_m - \frac{\phi^2}{2} \left\| K_{n,m}^{\mathcal{B}}(\rho) \right\|_{\ell_2}^2 \right).$$

Here $\{\Theta_n\}_{n=1}^{\infty}$ represents a class of nested subsets of $(0,\infty)$, i.e., $\Theta_p \subseteq \Theta_q \; \forall \; p \leq q$, whose diameter grows polynomially in n. Namely, $\mathrm{diam}(\Theta_n) \lesssim n^{\zeta}$ for some bounded positive scalar ζ. As sample size grows, such formulation of the LIF algorithm demands less restrictive assumptions on the range parameter and bears more resemblance to an unconstrained maximization problem.

3.5 Simulation Studies

This section is devoted to appraise the computational and statistical features of the LIF algorithm on synthetic stationary GP data. The purpose of our numerical analysis is two fold: investigating the scalability and efficiency of the proposed method in large datasets, as well as corroborating the infill asymptotic theory presented in Section 3.4. We primarily focus on two different scenarios regarding the sample size n. In moderate-size studies which are designed for constructing confidence intervals of unknown parameters through independent experiments, $n = 10^4$. Moreover, large-scale simulations with $n = 2.5 \times 10^5$ are conducted to study the numerical capabilities of the LIF algorithm, particularly when the exact and approximated evaluation of the likelihood function is extremely challenging. The computations have been performed on a UM Flux Ivy bridge compute node with 20 cores (Intel Xeon processor) and 3 GB memory per core. For expediting execution time of the simulations (up to 100 times), the LIF algorithm has been implemented in *C++* and *R* using *RcppParallel*[1] package.

Throughout this section G is a real valued stationary Matern GP observed on irregularly spaced lattice $\mathcal{D}_n$. We consider two cases of isotropy and geometric anisotropy for the covariance function. For circumventing the obstacles of computing the Cholesky factorization of the covariance matrix, spectral methods are used for constructing G on $\mathcal{D}_n$ [KSN16]. We now concisely describe the geometry of $\mathcal{D}_n$. Let $\mathcal{D} = [0,T]^2$ be a square of size T. $\mathcal{D}_n$ is a two dimensional randomly perturbed lattice of size $n = N^2$ if there exists a non-negative δ, representing the perturbation parameter, such that for any point $\boldsymbol{t} \in \mathcal{D}_n$, there are a corresponding point in the regular lattice $\boldsymbol{s} \in \{T/N, 2T/N, \ldots, T\}^2$ and a randomly chosen $\boldsymbol{p} \in [-T/N, T/N]^2$ (with uniform distribution) for which $\boldsymbol{t} = \boldsymbol{s} + \delta \boldsymbol{p}$. The scalar quantity

[1] https://cran.r-project.org/web/packages/RcppParallel/index.html

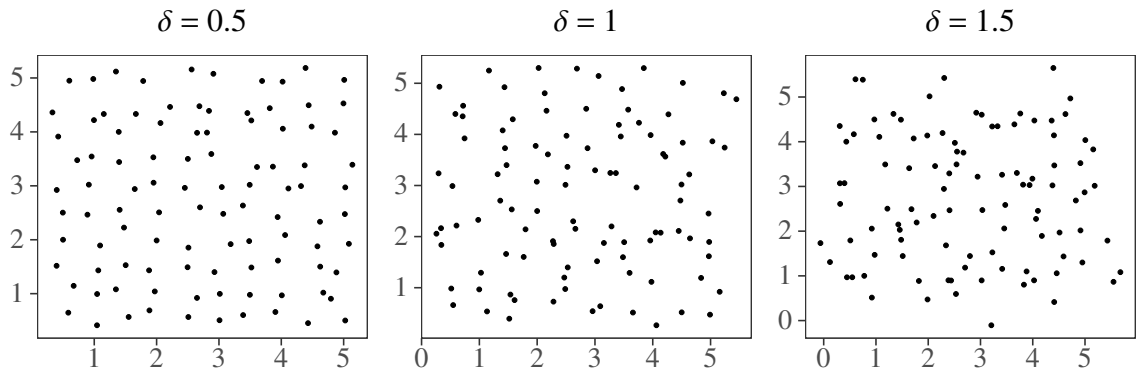

Figure 3.1: Each figure displays a perturbed lattices on $\mathcal{D} = [0,5]^2$ with $\delta = 0.5, 1$, and 2 from left to right. Each figure contains 10^2 points.

δ controls the amount of irregularity in the set of sampling locations. For instance Figure 3.1 exhibits three randomly perturbed lattices with $T = 5$, $N = 10$, and $\delta \in \{0.5, 1, 1.5\}$, from left to right.

Partitioning $\mathcal{D}_n$ into b_n bins is necessary for implementing the LIF algorithm. For brevity the bins are labelled 1 to b_n. In the following, we elucidate three schemes for constructing the bins.

1. *Uniformly Chosen (UC) bins*: Any $\boldsymbol{s} \in \mathcal{D}_n$ is randomly assigned to a bin in $\{1, \ldots, b_n\}$ with a uniform distribution. So the average size of all bins are the same.

2. *Non Uniformly Chosen (NUC) bins*: The points in $\mathcal{D}_n$ are independently assigned to bins labelled with $\{1, \ldots, b_n\}$, according to a non-uniform distribution Q. Throughout this section, we assume that Q is proportional to $[1, \ldots, 1, 2, \ldots, 2]^\top$. For instance in the case that $b_n = 4$, an arbitrary point $\boldsymbol{s} \in \mathcal{D}_n$ belongs to each bin with probabilities $[1/3, 1/3, 1/6, 1/6]$. Thus on average half of the bins are twice bigger than the other half.

3. *Rectangular bins*: $\mathcal{D}_n$ is segregated into b_n rectangular subregions and all the points in each subregion belong to the same bin.

Figure 3.2 illustrates the three methods of constructing subgroups for a randomly perturbed lattice of size 100 and $\delta = 0.5$. For a simple comparison, b_n is chosen to be 4 for each scenario in Figure 3.2.

We present three sets of simulation studies to assess the performance of the LIF algorithm. In all the experiments, G is a Matern GP observed on a randomly perturbed lattice. The *L-BFGS-B* algorithm is utilized for maximizing the LIF loss function. The finite difference

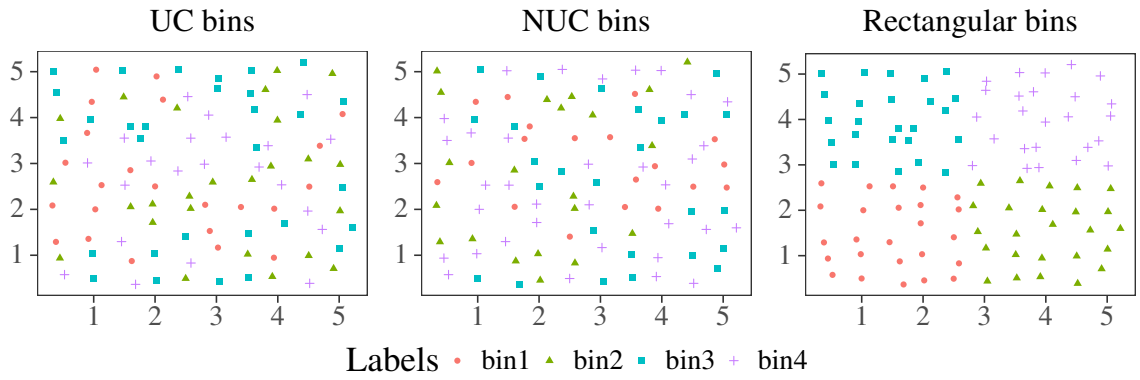

Figure 3.2: Three binning schemes of 10^2 points on a perturbed grid on $\mathcal{D} = [0,5]^2$ with $\delta = 0.5$

approximation with step size 10^{-3} is used for computing the gradient. We stop the optimization procedure if either the relative change in the objective function is below 10^{-5} or it reaches 50 iterations.

3.5.1 Moderate-Scale Simulations for Isotropic GPs

In all the experiments of this section, $\mathcal{D} = [0,5]^2$ and $\mathcal{D}_n$ is a perturbed lattice with $\delta \in \{1,3\}$ and 100^2 points, i.e. $n = 10^4$. We generate 100 realizations of an isotropic Matern GP G with parameters $\phi_0 = 1, \rho_0 = 5$, and $\nu = 0.5$ on 100 independent realizations of $\mathcal{D}_n$. The preconditioning order $m = 2$ is chosen for satisfying the condition $m \geq \nu + d/2$ in the statement of Theorems 3.1 and 3.2. Furthermore for any $\boldsymbol{s} \in \mathcal{D}_n$, $\mathcal{N}_m(\boldsymbol{s})$ consists of the seven closest points in $\mathcal{D}_n$ to $\boldsymbol{s}$ ($|\mathcal{N}_m(\boldsymbol{s})| = 7$). The goal is to estimate $\phi_0 \rho_0^{-2\nu}$, which has the central role in the asymptotic analysis in Section 3.4. According to Theorems 3.1 and 3.2, estimating ρ_0 is not necessary for the isotropic Matern covariance functions. In other words, ρ can be fixed in the optimization problem in Eq. (3.7). Therefore we select $\rho = 10$ and maximize the LIF function with respect to ϕ, i.e. $\hat{\rho}_{n,\mathcal{B}} = 10$. For each realization of G, $\hat{\phi}_{n,\mathcal{B}}$ is evaluated for $b_n \in \{1,2,4,8,16\}$ and three partitioning approaches UC, NUC, and rectangular. For brevity define

$$\hat{\xi}_{n,\mathcal{B}} = \frac{\hat{\phi}_{n,\mathcal{B}} \hat{\rho}_{n,\mathcal{B}}^{-2\nu}}{\phi_0 \rho_0^{-2\nu}}. \tag{3.17}$$

Theorem 3.2 suggests that $\hat{\xi}_{n,\mathcal{B}}$ is normally distributed centered at 1. Figures 3.3 and 3.4 respectively exhibit the histogram of $\hat{\xi}_{n,\mathcal{B}}$ for two cases of $\delta = 1$ and 3, different choices of b_n and partitioning schemes. Each plot also shows a kernel density estimate (KDE) of the histogram for a simpler comparison with the normal distribution. Table 3.1 presents the

mean and standard deviation of each histogram in Figures 3.3 and 3.4. According to Table 3.1, for different values of δ, b_n and bin shapes, $\hat{\xi}_{n,\mathcal{B}}$ is concentrated around 1 with the bias of order 10^{-3} and the standard deviation near 0.04, with a bell shaped density.

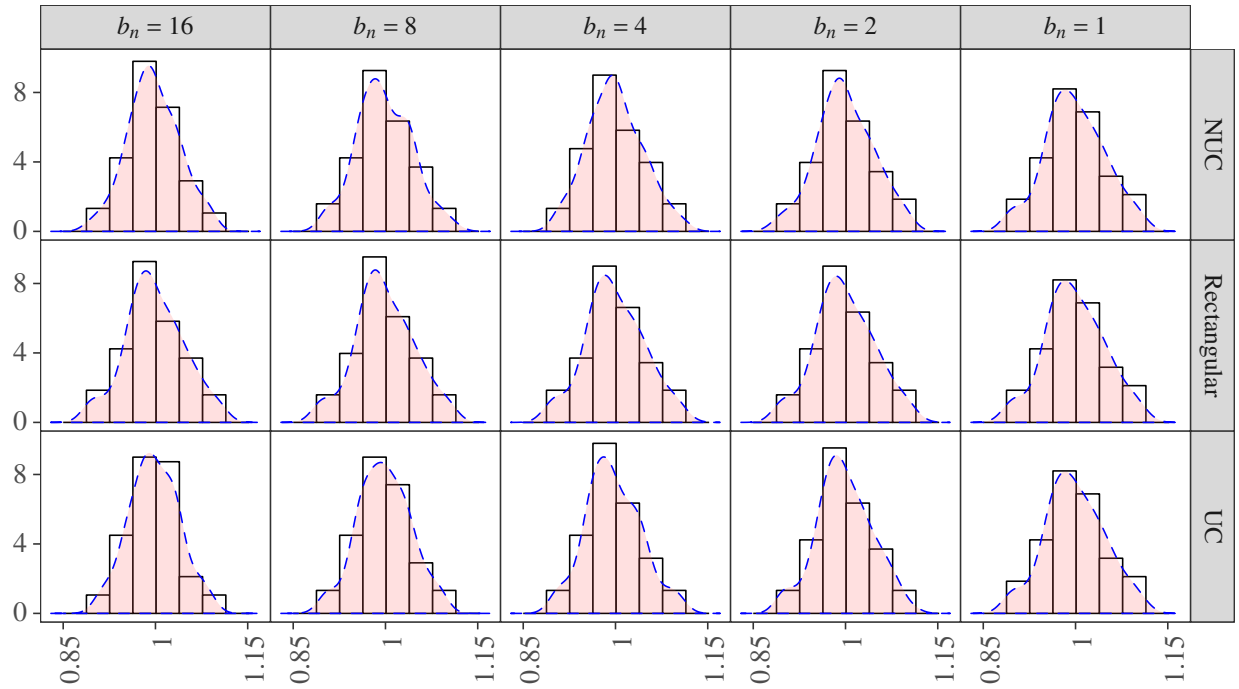

Figure 3.3: The histogram of $\hat{\xi}_{n,\mathcal{B}}$ with $m = 2$, $b_n = 1, 2, 4, 8, 16$ and 3 binning schemes for isotropic Matern GP with $(\phi_0, \rho_0, \nu) = (1, 5, 0.5)$ observed on a perturbed grid with $\delta = 1$ and $n = 10^4$.

		$b_n = 16$	$b_n = 8$	$b_n = 4$	$b_n = 2$	$b_n = 1$
$\delta = 1$	NUC	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 0.9968$ $\text{std}\,\hat{\xi}_{n,\mathcal{B}} = 0.0417$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 0.9979$ $\text{std}\,\hat{\xi}_{n,\mathcal{B}} = 0.0442$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 0.9993$ $\text{std}\,\hat{\xi}_{n,\mathcal{B}} = 0.0448$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 0.9993$ $\text{std}\,\hat{\xi}_{n,\mathcal{B}} = 0.0459$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 0.9990$ $\text{std}\,\hat{\xi}_{n,\mathcal{B}} = 0.0481$
	Rectangular	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 0.9989$ $\text{std}\,\hat{\xi}_{n,\mathcal{B}} = 0.0475$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 0.9990$ $\text{std}\,\hat{\xi}_{n,\mathcal{B}} = 0.0476$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 0.9991$ $\text{std}\,\hat{\xi}_{n,\mathcal{B}} = 0.0477$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 0.9992$ $\text{std}\,\hat{\xi}_{n,\mathcal{B}} = 0.0478$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 0.9990$ $\text{std}\,\hat{\xi}_{n,\mathcal{B}} = 0.0481$
	UC	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 0.9980$ $\text{std}\,\hat{\xi}_{n,\mathcal{B}} = 0.0403$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 0.9980$ $\text{std}\,\hat{\xi}_{n,\mathcal{B}} = 0.0424$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 0.9965$ $\text{std}\,\hat{\xi}_{n,\mathcal{B}} = 0.0443$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 0.9984$ $\text{std}\,\hat{\xi}_{n,\mathcal{B}} = 0.0450$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 0.9990$ $\text{std}\,\hat{\xi}_{n,\mathcal{B}} = 0.0481$
$\delta = 3$	NUC	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 0.9953$ $\text{std}\,\hat{\xi}_{n,\mathcal{B}} = 0.0463$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 0.9962$ $\text{std}\,\hat{\xi}_{n,\mathcal{B}} = 0.0472$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 0.9962$ $\text{std}\,\hat{\xi}_{n,\mathcal{B}} = 0.0500$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 0.9965$ $\text{std}\,\hat{\xi}_{n,\mathcal{B}} = 0.0524$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 0.9955$ $\text{std}\,\hat{\xi}_{n,\mathcal{B}} = 0.0534$
	Rectangular	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 0.9955$ $\text{std}_{n,\mathcal{B}} = 0.0536$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 0.9953$ $\text{std}\,\hat{\xi}_{n,\mathcal{B}} = 0.0536$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 0.9954$ $\text{std}\,\hat{\xi}_{n,\mathcal{B}} = 0.0534$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 0.9954$ $\text{std}\,\hat{\xi}_{n,\mathcal{B}} = 0.0535$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 0.9955$ $\text{std}\,\hat{\xi}_{n,\mathcal{B}} = 0.0534$
	UC	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 0.9966$ $\text{std}\,\hat{\xi}_{n,\mathcal{B}} = 0.0456$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 0.9954$ $\text{std}\,\hat{\xi}_{n,\mathcal{B}} = 0.0465$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 0.9954$ $\text{std}\,\hat{\xi}_{n,\mathcal{B}} = 0.0496$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 0.9952$ $\text{std}\,\hat{\xi}_{n,\mathcal{B}} = 0.0513$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 0.9955$ $\text{std}\,\hat{\xi}_{n,\mathcal{B}} = 0.0534$

Table 3.1: The mean and standard deviation of $\hat{\xi}_{n,\mathcal{B}}$ exhibited in histograms in Figures 3.3 and 3.4.

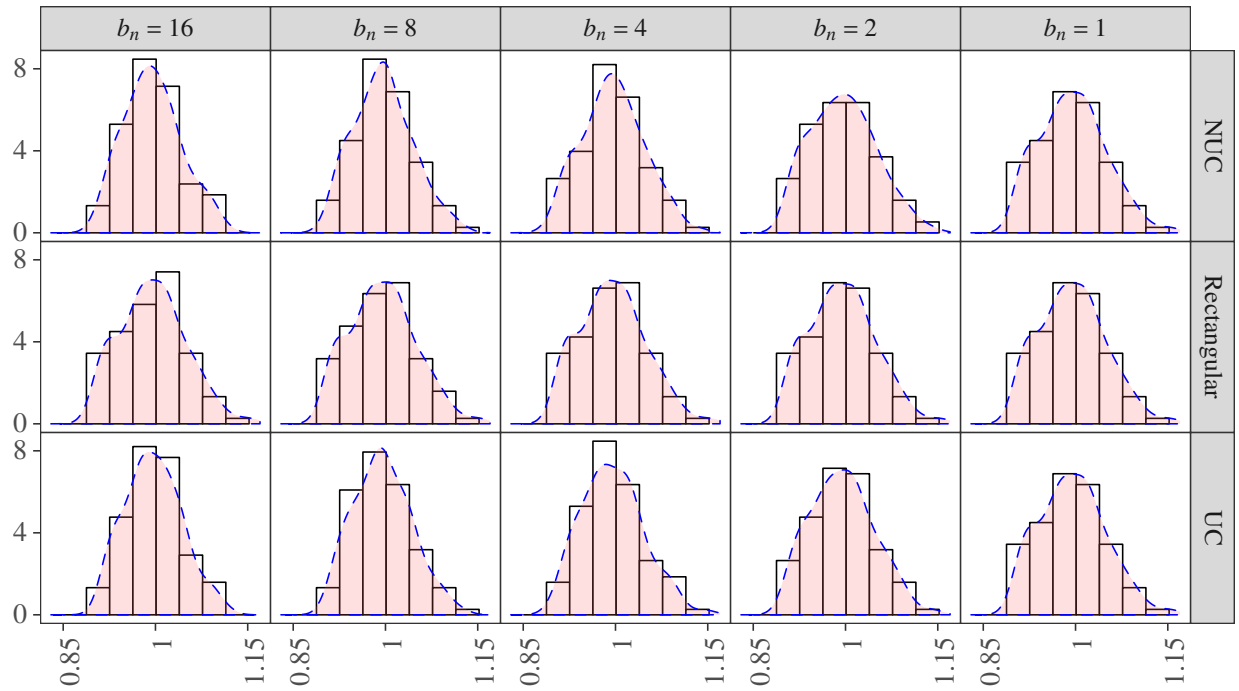

Figure 3.4: The histogram of $\hat{\xi}_{n,\mathcal{B}}$ with $m = 2$, $b_n = 1, 2, 4, 8, 16$ and 3 binning schemes for isotropic Matern GP with $(\phi_0, \rho_0, \nu) = (1, 5, 0.5)$ observed on a perturbed lattice with $\delta = 3$ and $n = 10^4$.

Next we conduct the same experiment on a smoother isotropic Matern GP with $\phi_0 = 1, \rho_0 = 2.5$, and $\nu = 1$. We seek to gauge the sensitivity of our estimation algorithm to the preconditioning order m by considering two cases of $m = 2$ and 3. Notice that the condition $m \geq \nu + d/2$ holds for both choices of m. However evaluating the LIF loss is a more difficult task for $m = 3$ because of dealing with larger conditioning sets ($|\mathcal{N}_3(\boldsymbol{s})| = 11$ for any $\boldsymbol{s} \in \mathcal{D}_n$). The histograms (and KDE) of $\hat{\xi}_{n,\mathcal{B}}$ associated witha $m = 2$ and 3 are successively displayed in Figures 3.5-3.6 and 3.7-3.8. Almost all the histograms in 3.5-3.8 are concentrated around 1 with the bias ranged in from 0.02 to 0.04 and the standard deviation between 0.3 and 0.45.Table 3.2 summarizes the mean and standard deviation of $\hat{\xi}_{n,\mathcal{B}}$ for the different choices of m, b_n, δ, and partitioning schemes.

			$b_n = 16$	$b_n = 8$	$b_n = 4$	$b_n = 2$	$b_n = 1$
$m = 2$	$\delta = 1$	NUC	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0465$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.3188$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0459$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.3222$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0478$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.3315$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0481$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.3439$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0489$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.3555$
		Rectangular	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0491$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.3548$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0489$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.3550$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0487$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.3554$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0491$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.3556$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.04889$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.3555$
		UC	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0458$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.3173$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0464$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.3215$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0470$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.3289$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0488$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.3418$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0489$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.3555$
	$\delta = 3$	NUC	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0302$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.3790$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0315$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.3847$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0329$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.4926$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0366$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.4075$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0393$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.4105$
		Rectangular	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0396$ $\text{std}_{n,\mathcal{B}} = 0.4196$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0392$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.4196$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0393$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.4201$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0394$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.4204$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0393$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.4105$
		UC	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0304$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.3789$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0323$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.3846$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0337$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.3927$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0363$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.4048$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0393$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.4105$
$m = 3$	$\delta = 1$	NUC	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0237$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.4104$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0237$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.4177$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0262$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.4285$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0279$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.4464$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0315$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.4635$
		Rectangular	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0311$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.4616$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0312$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.4620$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0313$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.4626$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0316$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.4633$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0315$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.4635$
		UC	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0232$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.4096$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0239$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.4156$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0267$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.41275$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0296$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.4463$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0315$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.4635$
	$\delta = 3$	NUC	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0206$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.3771$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0228$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.3835$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0223$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.3934$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0255$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.4069$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0271$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.4216$
		Rectangular	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0271$ $\text{std}_{n,\mathcal{B}} = 0.4202$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0276$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.4215$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0274$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.4219$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0273$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.4218$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0271$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.4216$
		UC	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0214$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.3764$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0204$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.3798$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.02037$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.3921$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0249$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.4045$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0271$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.4216$

Table 3.2: The mean and standard deviation of $\hat{\xi}_{n,\mathcal{B}}$ displayed in histograms in Figures 3.5-3.8.

Remark 3.10. The above experiments explicate some aspects of the LIF method which did not thoroughly explained by the asymptotic theory. In the following we list some critical observations of the simulation studies in this section.

(a) In most of the entries in Tables 3.1 and 3.2, the bias of $\hat{\xi}_{n,\mathcal{B}}$ is considerably smaller than its standard deviation. This phenomenon has been predicted in the proof of Theorem 3.1. The asymptotic analysis shows that for isotropic GPs observed in a d-dimensional space

$$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} - 1 = \mathcal{O}\left(n^{-2/d}\right), \text{ and } \quad \text{std}\,\hat{\xi}_{n,\mathcal{B}} = \mathcal{O}\left(n^{-1/2}\right).$$

So for $d = 2$, the bias to standard deviation ratio is order $n^{-1/2}$, converging to zero as $n \to \infty$.

(b) As long as m is chosen to satisfy $m \geq \nu + d/2$, increasing the preconditioning order does not improve the estimation performance. On the other hand larger m requires more challenging computation for evaluating the LIF loss function. So choosing $m = \lceil \nu + d/2 \rceil$ can optimally balance between statistical efficiency and computational tractability.

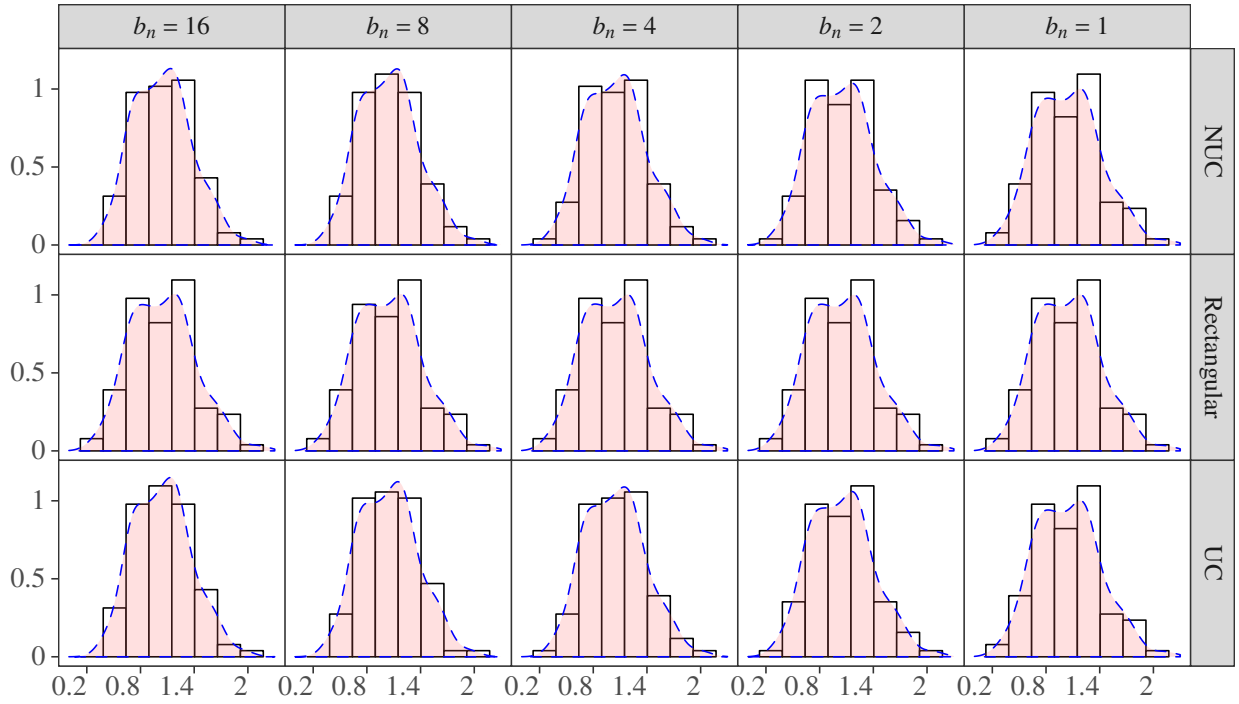

Figure 3.5: The histogram of $\hat{\xi}_{n,\mathcal{B}}$ with $m = 2$, $b_n = 1, 2, 4, 8, 16$ and 3 binning schemes for isotropic Matern GP with $(\phi_0, \rho_0, \nu) = (1, 2.5, 1)$ observed on a perturbed lattice with $\delta = 1$ and $n = 10^4$.

(c) Comparing the results in Tables 3.1 and 3.2 shows that $\hat{\xi}_{n,\mathcal{B}}$ has larger bias and standard deviation for $\nu = 1$. Namely estimating $\phi_0 \rho^{-2\nu}$ is more difficult when $\nu = 1$. We give a qualitative justification for this phenomenon. It has been argued in Remark 3.8 that the LIF algorithm is consistent when the largest eigenvalue of $K^{\mathcal{B}}_{n,m}(\cdot)$ is uniformly bounded (independent of n) and its Frobenius norm is of order $\sqrt{n}$. Simply put the effective rank of $K^{\mathcal{B}}_{n,m}(\cdot)$ should be of order n. Define the quantity $\Psi^{\mathcal{B}}_{n,m}$ as

$$\Psi^{\mathcal{B}}_{n,m} := \frac{\left\|K^{\mathcal{B}}_{n,m}\right\|_{2\to 2} \sqrt{n}}{\left\|K^{\mathcal{B}}_{n,m}\right\|_{\ell_2}},$$

Observe that $\Psi^{\mathcal{B}}_{n,m}$ is no smaller than 1 and attains its minimum for the identity matrix. If $K^{\mathcal{B}}_{n,m}(\cdot)$ can be well approximated by a rank deficit matrix of rank $r_n = o(n)$, then $\Psi^{\mathcal{B}}_{n,m}$ grows with the same rate as $\sqrt{n/r_n}$. So roughly speaking the LIF algorithm works better for smaller $\Psi^{\mathcal{B}}_{n,m}$. Here we compare $\Psi^{\mathcal{B}}_{n,m}$ for the two cases of $\nu = 0.5$ and 1. For avoiding the computational challenges of evaluating the operator norm of large matrices, we focus on smaller size perturbed grids on $\mathcal{D} = [0, 2.5]^2$ of size 2500 ($N = 50$) and with $\delta \in (0.5, 1.5)$. The range parameter of G is assumed to be $\rho_0 = 1.25$.

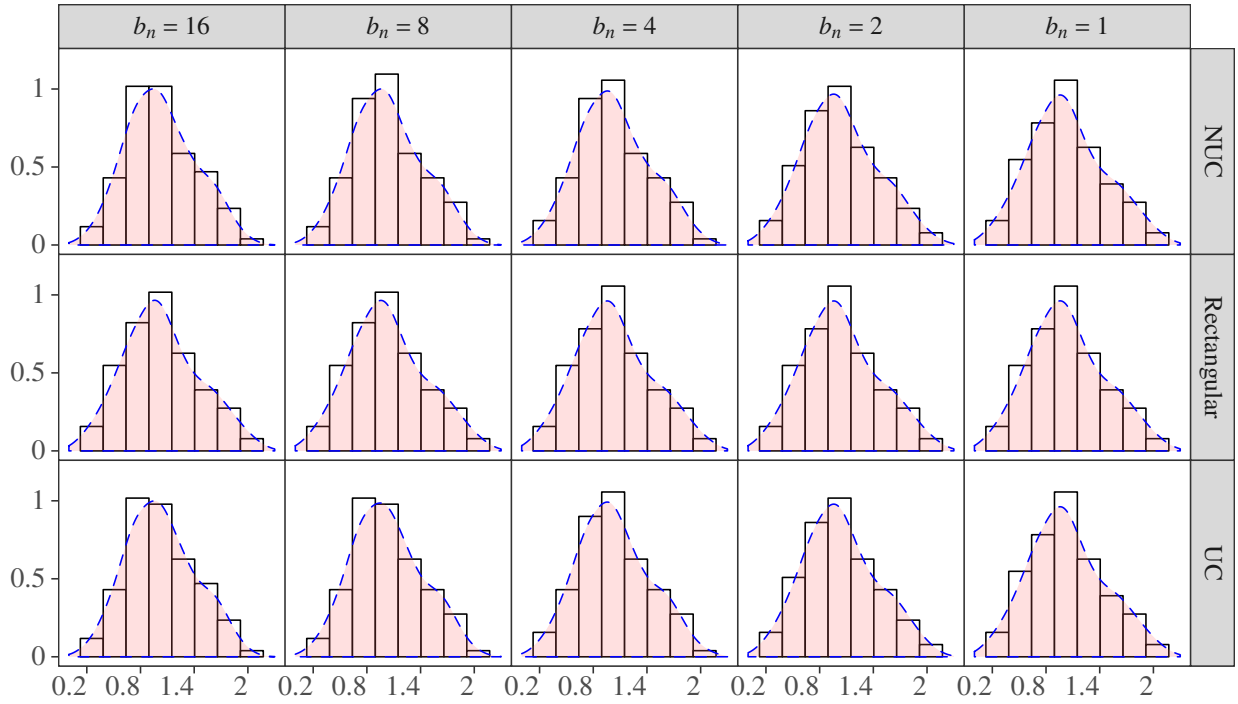

Figure 3.6: The histogram of $\hat{\xi}_{n,\mathcal{B}}$ with $m = 2$, $b_n = 1, 2, 4, 8, 16$ and 3 binning schemes for isotropic Matern GP with $(\phi_0, \rho_0, \nu) = (1, 2.5, 1)$ observed on a perturbed lattice with $\delta = 3$ and $n = 10^4$.

Note that in the new experiment ρ_0, the diameter of $\mathcal{D}$ and δ have been chosen in such a way that the lattice of size 50^2 imitates the local neighbouring properties of $\mathcal{D}_n$ in Figures 3.3-3.8. Figure 3.9 displays $\Psi^{\mathcal{B}}_{n,m}$ in four different scenarios of (ν, δ). It is obviously apparent that $\Psi^{\mathcal{B}}_{n,m}$ is always larger for $\nu = 1$, which can explain the higher bias and variance of the LIF estimate.

(d) A prudent look at Figures 3.3-3.8 discloses a notable similarity between all the histograms in each row. This observation recommends that varying b_n from 16 to 1 narrowly affects the variance and bias of $\hat{\xi}_{n,\mathcal{B}}$. In fact the covariance matrix of the preconditioned process can be well approximated by a banded sparse matrix in our simulations. For a more vivid explanation, define the quantity $r_{n,m}$ by

$$r_{n,m} := \frac{\sum_{\boldsymbol{s}\in\mathcal{D}_n} \sum_{\boldsymbol{t}\in\mathcal{N}_m(\boldsymbol{s})} \left[\operatorname{cov}\left(G(\boldsymbol{s}), G(\boldsymbol{t})\right)\right]^2}{\sum_{\boldsymbol{s},\boldsymbol{t}\in\mathcal{D}_n} \left[\operatorname{cov}\left(G(\boldsymbol{s}), G(\boldsymbol{t})\right)\right]^2}.$$

Notice that in this section any preconditioning set $\mathcal{N}_m(\boldsymbol{s})$ comprises only the closest points to $\boldsymbol{s}$. Thus if $r_{n,m}$ lies in the vicinity of 1, then the covariance structure of the pre-

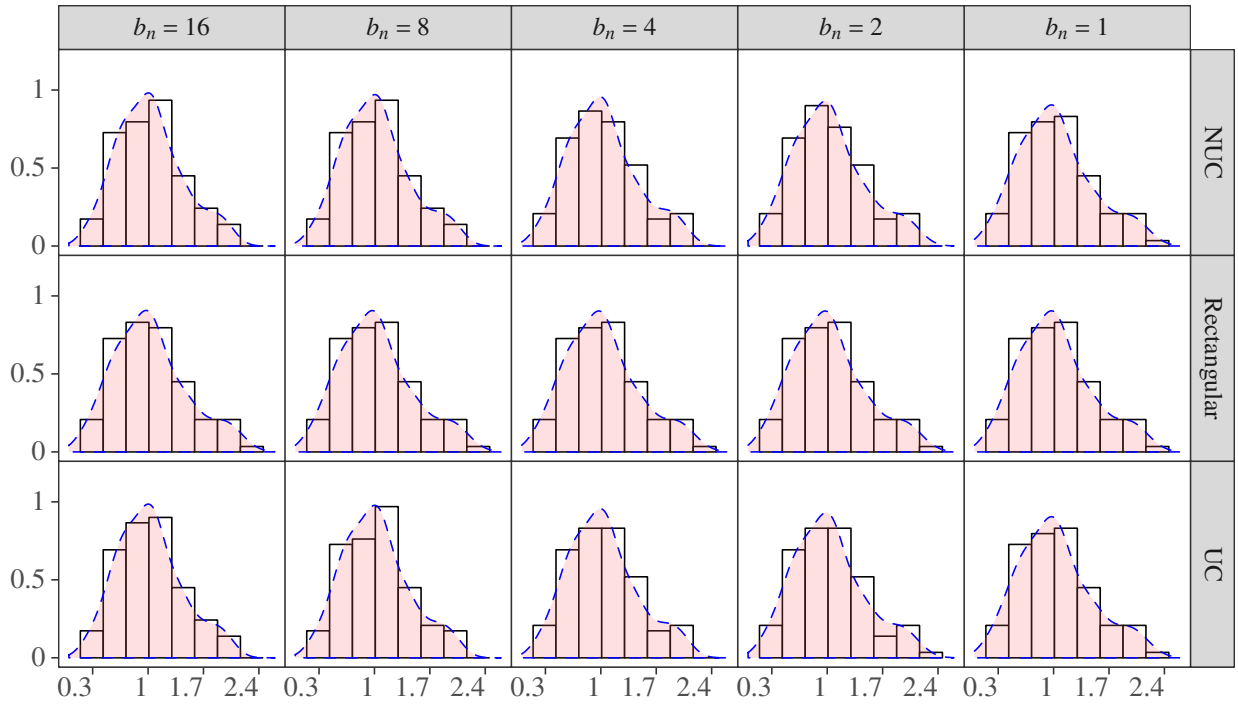

Figure 3.7: The histogram of $\hat{\xi}_{n,\mathcal{B}}$ with $m = 3$, $b_n = 1, 2, 4, 8, 16$ and 3 binning schemes for isotropic Matern GP with $(\phi_0, \rho_0, \nu) = (1, 2.5, 1)$ observed on a perturbed lattice with $\delta = 1$ and $n = 10^4$.

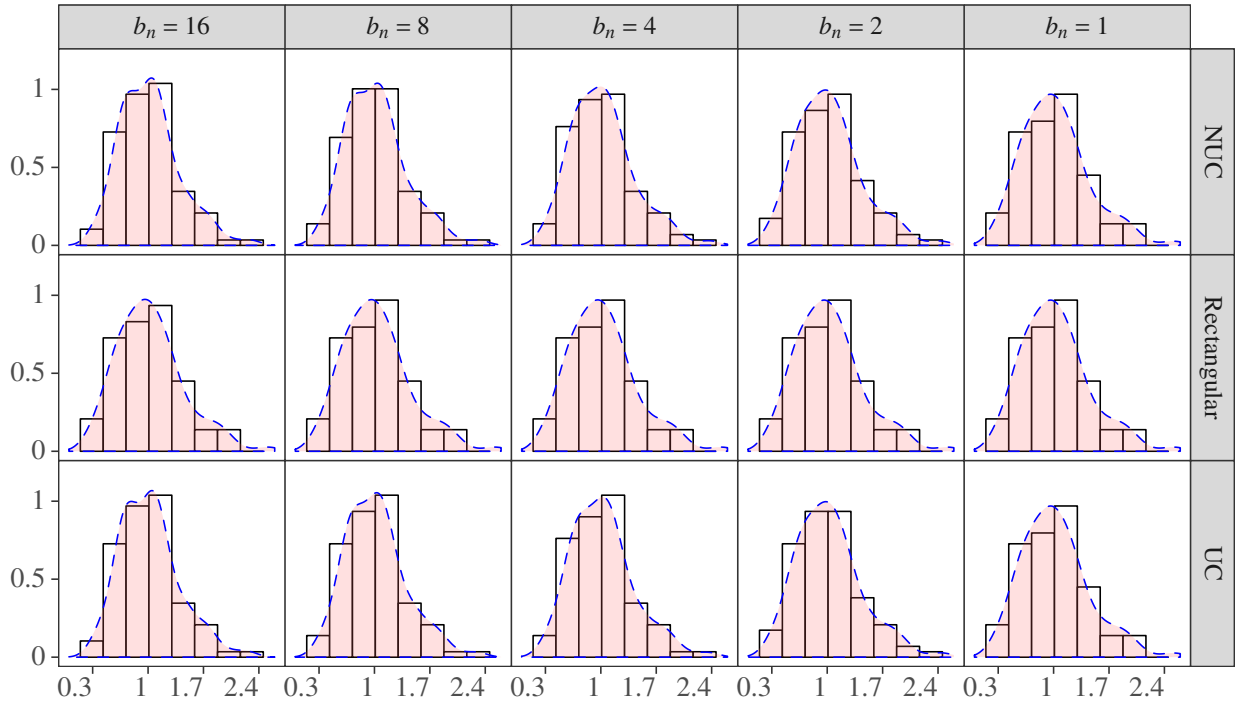

Figure 3.8: The histogram of $\hat{\xi}_{n,\mathcal{B}}$ with $m = 3$, $b_n = 1, 2, 4, 8, 16$ and 3 binning schemes for isotropic Matern GP with $(\phi_0, \rho_0, \nu) = (1, 2.5, 1)$ observed on a perturbed lattice with $\delta = 3$ and $n = 10^4$.

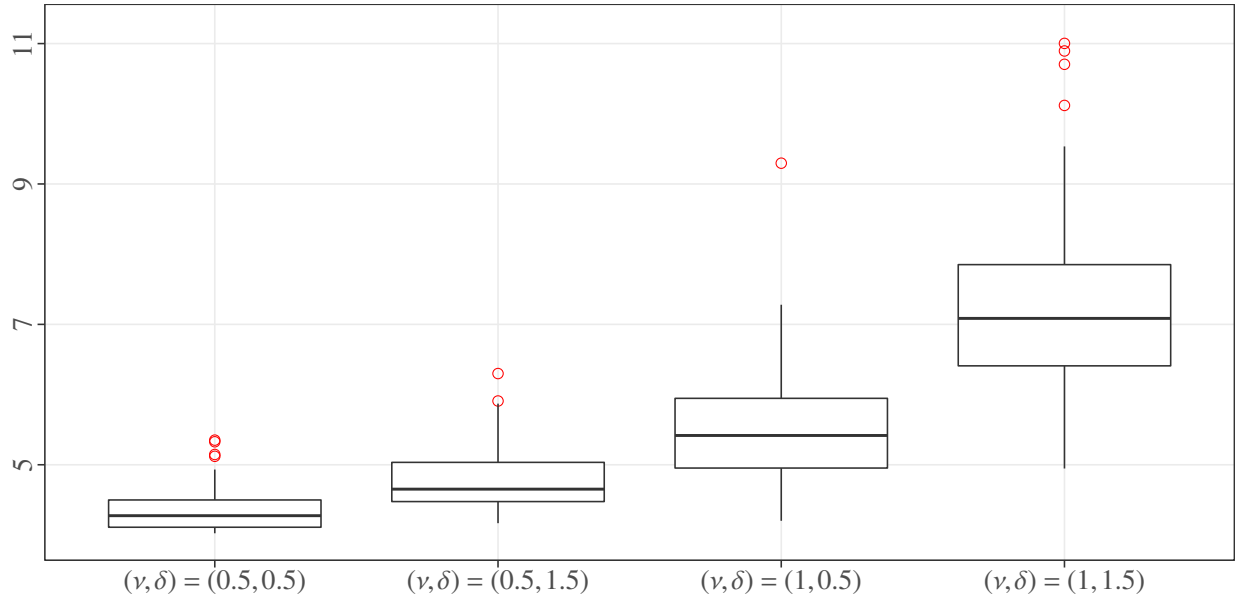

Figure 3.9: The box-plot of $\Psi_{n,m}$ for different values of δ and ν. Here $\mathcal{D}_n$ is a perturbed lattice of size 2500 and G is an isotropic Matern GP with $\phi_0 = 1$ and $\rho_0 = 1.25$.

conditioned GP can be suitably explained by the nearest neighbours. For the isotropic Matern GPs in the previous simulation studies, we have computed the average $r_{n,m}$ for $\nu = 0.5$ and 1. $r_{n,m} = 0.9569$ and 0.9702 for $\nu = 0.5$ and 1, respectively. Simply put, the covariance matrix of G_m is near banded, meaning that the LIF estimate is robust to the changes of b_n, particularly for large bins.

We now numerically gauge the asymptotic behaviour of the LIF estimate. For doing so we generate 100 independent realizations of an isotropic Matern GP with $(\phi_0, \rho_0, \nu) = (1, 5, 0.5)$ on 100 independently generated perturbed lattices of size $n = N^2$ and with $\delta \in \{1, 3\}$ on $\mathcal{D} = [0, 5]^2$. The LIF loss function is optimized with respect to ϕ and for a fixed $\rho = 10$. We refer the reader to Table 3.3 for the sample average and standard deviation of $\hat{\xi}_{n,\mathcal{B}}$ for different values of n. The results in Table 3.3 shows that the LIF estimate becomes more accurate as n increases (in a fixed domain).

		$N = 20$	$N = 30$	$N = 50$	$N = 70$	$N = 100$	$N = 150$
$\delta = 1$	bias of $\hat{\xi}_{n,\mathcal{B}}$	0.8643	0.5891	0.2955	0.1593	0.0299	0.0198
	std of $\hat{\xi}_{n,\mathcal{B}}$	0.3716	0.2305	0.1093	0.0700	0.0480	0.0233
$\delta = 3$	bias of $\hat{\xi}_{n,\mathcal{B}}$	3.2033	1.0161	0.5133	0.2157	0.0634	0.0187
	std of $\hat{\xi}_{n,\mathcal{B}}$	1.4174	0.4070	0.1218	0.0984	0.0519	0.0355

Table 3.3: The mean and standard deviation of $\hat{\xi}_{n,\mathcal{B}}$ over 100 independent experiments for isotropic Matern GP with $(\phi_0, \rho_0, \nu) = (1, 5, 0.5)$ and for different size of lattice.

From Definition 3.1, recall the notion of the preconditioning sets $\mathcal{N}_m(\cdot)$. The first condition in Definition 3.1, which we state here for convenience, says that $\|\boldsymbol{t} - \boldsymbol{s}\|_{\ell_2} \lesssim 1/N$ for any $\boldsymbol{t} \in \mathcal{N}_m(\boldsymbol{s})$ and any $\boldsymbol{s} \in \mathcal{D}_n$. Roughly speaking, the points in $\mathcal{N}_m(\boldsymbol{s})$ should be in the vicinity of $\boldsymbol{s}$. Note that this condition is the only restriction on the choice of $\boldsymbol{t}$. We intend to validate the necessity of such condition through a simulation study. For this experiment G is assumed to be isotropic Matern with parameters $(\phi_0, \rho_0, \nu) = (1, 5, 0.5)$, sampled on a perturbed lattice of size 10^4 with $\delta \in \{1, 3\}$. For any $\boldsymbol{s} \in \mathcal{D}_n$, $\mathcal{N}_m(\boldsymbol{s})$ includes $\boldsymbol{s}$ and the six most distant points to $\boldsymbol{s}$ in $\mathcal{D}_n$. Thus each preconditioning sets is of size 7. Notice that such setting trivially violates the assumption in Definition 3.1. Furthermore, similar to the previous experiments in this section ρ is fixed at 10. Table 3.4 reports the bias and standard deviation of $\hat{\xi}_{n,\mathcal{B}}$ (Eq. (3.17)) for different values of b_n and the formerly described partitioning schemes. Comparing the results in Table 3.4 and 3.1 shows that accuracy of the LIF estimator significantly declined in the new framework, which corroborates our understanding of the preconditioning sets.

		$b_n = 16$	$b_n = 8$	$b_n = 4$	$b_n = 2$	$b_n = 1$
$\delta = 1$	NUC	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0674$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.8383$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0663$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.8377$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0675$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.8396$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0665$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.8377$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0669$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.8384$
	Rectangular	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0637$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.8062$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0651$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.8203$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0668$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.8373$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0666$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.8372$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0669$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.8384$
	UC	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0665$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.8369$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0667$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.8377$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0668$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.8384$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0668$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.8383$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0669$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.8384$
$\delta = 3$	NUC	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0315$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.7707$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0321$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.7707$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0320$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.7679$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0315$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.7712$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0321$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.7711$
	Rectangular	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0234$ $\text{std}_{n,\mathcal{B}} = 0.7708$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0300$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.7797$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0336$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.7743$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0324$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.7717$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0321$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.7711$
	UC	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0317$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.7692$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0319$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.7703$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0320$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.7708$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0321$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.7713$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = 1.0321$ $\text{std}\hat{\xi}_{n,\mathcal{B}} = 0.7711$

Table 3.4: The mean and standard deviation of $\hat{\xi}_{n,\mathcal{B}}$ for 100 independent experiments of isotropic Matern GP with $(\phi_0, \rho_0, \nu) = (1, 5, 0.5)$, sampled on a perturbed lattice of size 10^4 with $\delta \in \{1, 3\}$. For any $\boldsymbol{s}$ in the perturbed lattice, $\mathcal{N}_m(\cdot)$ includes $\boldsymbol{s}$ and six most distant points to $\boldsymbol{s}$.

3.5.2 Moderate-Scale Simulations for Geometric Anisotropic GPs

This subsection is devoted to assess the performance of the LIF method for geometric anisotropic Matern GPs lie in two dimensional fixed domains. Particularly, there is $\rho_0 = (\rho_{0,1}, \rho_{0,2})$ such that for any $\boldsymbol{s} = (s_1, s_2)$ and $\boldsymbol{t} = (t_1, t_2)$,

$$\text{cov}\Big(G(\boldsymbol{s}), G(\boldsymbol{t})\Big) = \phi_0 f_\nu(r), \text{ in which } r^2 = \left(\frac{t_1 - s_1}{\rho_{0,1}}\right)^2 + \left(\frac{t_2 - s_2}{\rho_{0,2}}\right)^2.$$

Here f_ν stands for the Matern standard correlation function with the smoothness parameter ν. The quantities $\hat{\phi}_{n,\mathcal{B}} \in \mathbb{R}$ and $\hat{\rho}_{n,\mathcal{B}} \in \mathbb{R}^2$ are obtained by maximizing the LIF loss. As ϕ_0 and ρ_0 are not discernible in the infill setting, the core emphasis of our simulation studies in to estimate $\phi_0 \rho_{0,1}^{-2\nu}$ and $\phi_0 \rho_{0,2}^{-2\nu}$. For brevity we reformulate $\hat{\xi}_{n,\mathcal{B}}$ as the following:

$$\hat{\xi}_{n,\mathcal{B}} = \left(\frac{\hat{\phi}_{n,\mathcal{B}} \hat{\rho}_{1,n,\mathcal{B}}^{-2\nu}}{\phi_0 \rho_{0,1}^{-2\nu}}, \frac{\hat{\phi}_{n,\mathcal{B}} \hat{\rho}_{2,n,\mathcal{B}}^{-2\nu}}{\phi_0 \rho_{0,2}^{-2\nu}}\right) \in [0, \infty)^2. \tag{3.18}$$

Again, we let $\mathcal{D}_n$ to be a perturbed lattice of size $n = 10^4$ and with $\delta \in \{1, 3\}$. We simulate 100 independent realizations of a Matern GP with $\phi_0 = 1$, $\rho_0 = (1.5, 4)$ and $\nu \in (0.5, 1)$ on 100 realizations of $\mathcal{D}_n$. The L-BFGS-B method with the initial guess $\rho = (10, 10)$ maximizes the LIF loss function in a constrained box $[0.1, 50]^2$. In our experiments the boundary points have not been touched during optimization, so the final results do not change even when the box constraints are not enforced. The scatter plots of $\hat{\xi}_{n,\mathcal{B}}$ are depicted in Figures 3.10-3.11

for $b_n \in \{4, 16\}$ and two partitioning approaches. It appears that $\hat{\xi}_{n,\mathcal{B}}$ is concentrated around $(1, 1)$ for all the scenarios. Table 3.5 also accumulates the mean and standard deviation of $\hat{\xi}_{n,\mathcal{B}}$ displayed in Figures 3.10-3.11. It turns out from the scatter plots that the LIF estimates have higher variance when $\nu = 1$. As we have discussed in Remark 3.10, this observation is because the covariance matrix of the preconditioned data has larger effective rank, in the case that $\nu = 1$.

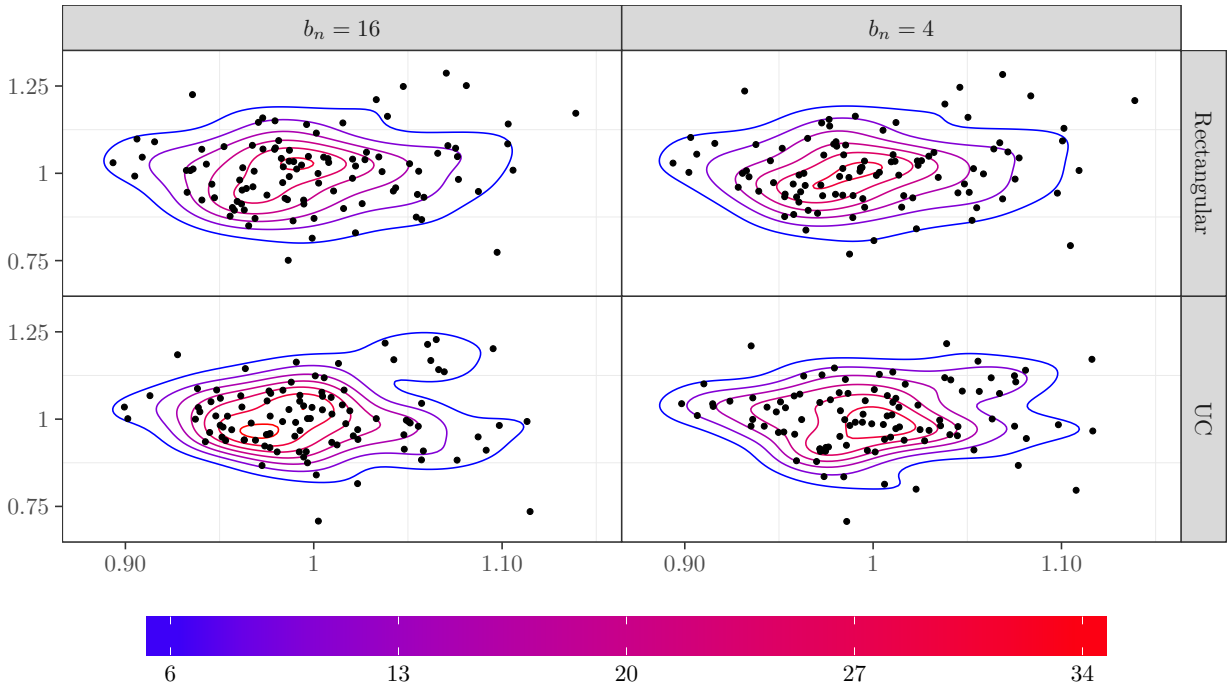

Figure 3.10: The scatter plot and two dimensional KDE of $\hat{\xi}_{n,\mathcal{B}}$ for an anisotropic Matern GP with $\phi_0 = 1, \rho_0 = (1.5, 4)$, and $\nu_0 = 0.5$ observed on a perturbed lattice with $\delta = 1$ and $n = 10^4$.

		$b_n = 16$	$b_n = 4$
$\nu = 0.5$	UC	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = (0.9996, 1.0063)$ $\text{std}\,\hat{\xi}_{n,\mathcal{B}} = (0.0467, 0.0966)$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = (1.0002, 1.0049)$ $\text{std}\,\hat{\xi}_{n,\mathcal{B}} = (0.0482, 0.0932)$
	Rectangular	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = (0.9993, 1.0081)$ $\text{std}\,\hat{\xi}_{n,\mathcal{B}} = (0.0507, 0.1026)$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = (0.9994, 1.0104)$ $\text{std}\,\hat{\xi}_{n,\mathcal{B}} = (0.0515, 0.0998)$
$\nu = 1$	UC	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = (1.004, 1.0800)$ $\text{std}\,\hat{\xi}_{n,\mathcal{B}} = (0.2776, 0.5656)$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = (1.0016, 1.0891)$ $\text{std}\,\hat{\xi}_{n,\mathcal{B}} = (0.2864, 0.5954)$
	Rectangular	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = (0.9992, 1.1083)$ $\text{std}\,\hat{\xi}_{n,\mathcal{B}} = 0.3059, 0.6432$	$\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = (0.9991, 1.1086)$ $\text{std}\,\hat{\xi}_{n,\mathcal{B}} = (0.3061, 0.6487)$

Table 3.5: The mean and standard deviation of $\hat{\xi}_{n,\mathcal{B}}$ exhibited in scatter plots in Figures 3.10 and 3.11.

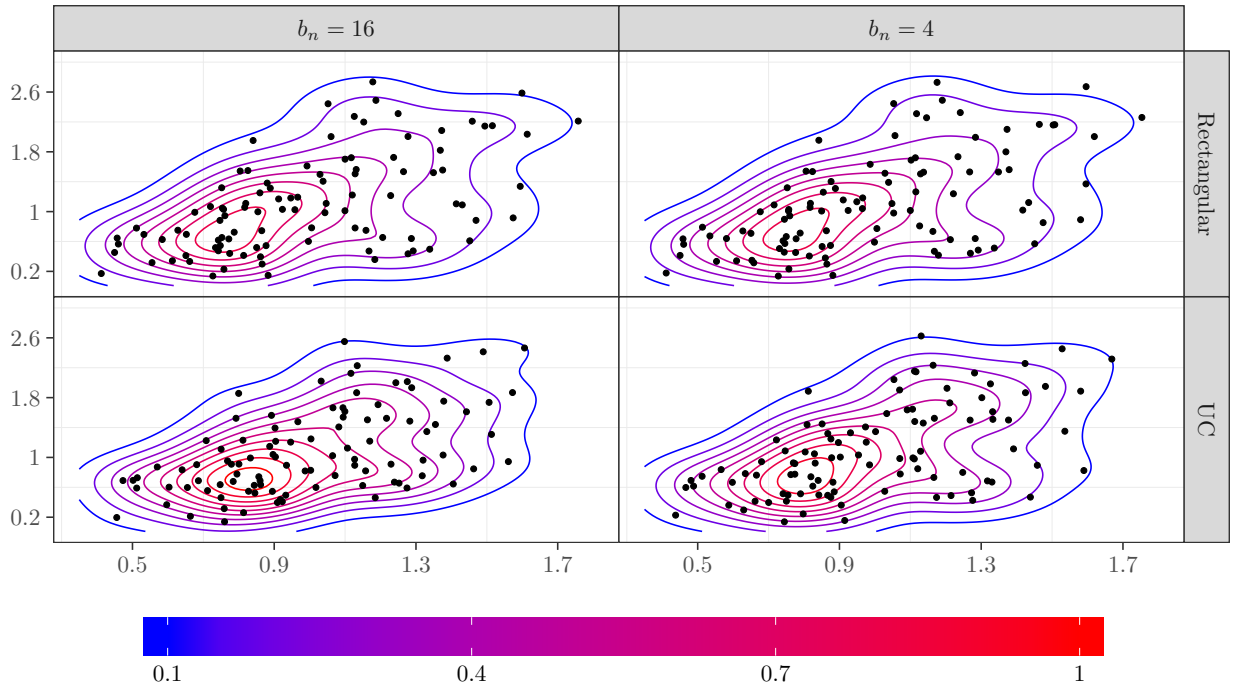

Figure 3.11: The scatter plot and two dimensional KDE of $\hat{\xi}_{n,\mathcal{B}}$ for an anisotropic Matern GP with $\phi_0 = 1, \rho_0 = (1.5, 4)$, and $\nu_0 = 1$ observed on a perturbed lattice with $\delta = 3$ and $n = 10^4$.

3.5.3 Large-Scale Simulations for Geometric Anisotropic GPs

To obtain further insights into the estimation accuracy of the LIF algorithm on large data sets, we carry out a few simulation studies on Matern GPs observed on perturbed lattices. The simulations are separated into two categories described as the following:

1. We fix $\mathcal{D} = [0,25]^2$ and choose a perturbed lattice $\mathcal{D}_n$ of size 2.5×10^5, i.e. $N = 500$, with $\delta = 5$ on $\mathcal{D}$. G is a geometric anisotropic Matern GP with $\rho_0 = (\rho_{0,1}, \rho_{0,2}) = (2,5)$ and $\phi_0 = 1$ observed on $\mathcal{D}_n$. Such simulation imitates the large-sample fixed domain behaviour, as the diameter of $\mathcal{D}$ is considerably smaller than N. So we report the LIF estimate of $\phi_0 \rho_{0,1}^{-2\nu}$ and $\phi_0 \rho_{0,2}^{-2\nu}$.

2. In the second class which emulates the increasing domain setting, we select $\mathcal{D} = [0,500]^2$. Furthermore, the variance and range parameter of G are given by $\phi_0 = 1$ and $\rho_0 = (10,20)$ and $\nu = 1$. $\mathcal{D}_n$ is also treated the same as the first category. In these simulations, the estimate of all the unknown parameters will be reported.

Recall $\hat{\xi}_{n,\mathcal{B}}$ from Eq. (3.18). Tables 3.6 encapsulates $\hat{\xi}_{n,\mathcal{B}}$ and the running time of maximizing the LIF loss in the box-constrained region $[0.1,50]$ by L-BFGS-B algorithm and with the initial guess $\rho = (4,8)$. Comparing to the case of $\nu = 0.5$, the optimization algorithm is three times slower for $\nu = 1$, which is due to the more complicated form of the covariance function. Furthermore the running time of the LIF loss optimizer is inversely proportional to b_n.

		$b_n = 200$	$b_n = 50$	$b_n = 10$
$\nu = 0.5$	$\hat{\xi}_{n,\mathcal{B}}$	(0.9978, 1.0434)	(0.9988, 1.04085)	(1.0011, 1.0280)
	Running time (hour)	0.5016	2.1747	4.8055
$\nu = 1$	$\hat{\xi}_{n,\mathcal{B}}$	(0.9910, 1.1060)	(0.9951, 1.0858)	(0.9928, 1.0899)
	Running time (hour)	1.4128	5.4449	13.2018

Table 3.6: The summary of the large-sample simulations for the first category.

In the sequel we present the summary of the results in Table 3.7, for the case that $\mathcal{D} = [0,500]^2$. The L-BFGS-B optimizer starts at $\rho = (25,40)$. We only consider the case that $\nu = 1$, because of the more challenging computation. Note that obtaining the estimated parameters in this setting is around twice slower than the former case.

		$b_n = 200$	$b_n = 50$	$b_n = 10$
$\nu = 1$	$\hat{\phi}_{n,\mathcal{B}}$	1.0179	1.0072	1.0125
	$\hat{\rho}_{n,\mathcal{B}}$	(10.4457, 19.8137)	(10.3789, 19.8433)	(10.4203, 19.8278)
	Running time (hour)	2.7441	10.5585	25.6577

Table 3.7: The summary of the large-sample simulations for the second category.

3.6 Discussion

This chapter introduced a family of scalable covariance estimation algorithm, which is called the LIF algorithm, by amalgamating the idea of inversion-free estimation procedure in [ACS16] and the block diagonal approximation of the covariance matrix of the pre-conditioned data. We have established $\sqrt{n}$-consistency and asymptotic normality of our method for the isotropic Matern covariance function on a d-dimensional irregular lattice (with $d \leq 3$). Prior to this work, it had been asserted that the inversion-free estimator is statistically comparable to the MLE, when there exists a linear transformation to uniformly control the condition number of the covariance matrix below some constant, independent of the sample size. However, our analysis demonstrates that the LIF algorithm has the same convergence rate as the MLE, as long as the largest eigenvalue remains uniformly bounded and a non-negligible percent of the eigenvalues are further away from zero. Refuting the necessity of uniformly controlling the condition number of the covariance matrix in our asymptotic theory can expand the applicability of surrogate loss maximization methods for estimating the covariance estimation of Gaussian spatial processes.

Despite the relatively low cost of computing the LIF estimate for GPs observed on irregularly spaced locations, it still has several drawbacks, particularly if the goal is beyond parameter estimation. For instance, as opposed to the likelihood based methods, still little is known about using proxy losses such as LIF for performing the model selection. Furthermore, despite recent progresses in reducing the condition number of the covariance matrix for stationary GPs, the effective preconditioning mechanism for non-stationary random fields is still obscure. However, we have only scratched the surface of scalable non-likelihood based estimation algorithms and still much needs to be done for developing an efficient class of algorithms for a broad family of spatial processes.

3.7 Proof of the Main Results

All the constants appearing in this section (including those implicitly defined in $\lesssim$, and $\asymp$), are bounded and depend on m, ν, d, Θ_0, and the geometric structure of the sampling locations.

Proof of Theorem 3.1. Applying the triangle inequality, we get

$$\sup_{\rho\in\Theta_0}\left|\frac{\hat{\phi}_{n,\mathcal{B}}(\rho)\rho^{-2\nu}}{\phi_0\rho_0^{-2\nu}}-1\right|\leq\sup_{\rho\in\Theta_0}\left|\frac{\mathbb{E}\hat{\phi}_{n,\mathcal{B}}(\rho)\rho^{-2\nu}}{\phi_0\rho_0^{-2\nu}}-1\right|+\sup_{\rho\in\Theta_0}\frac{\left|\hat{\phi}_{n,\mathcal{B}}(\rho)\rho^{-2\nu}-\mathbb{E}\hat{\phi}_{n,\mathcal{B}}(\rho)\rho^{-2\nu}\right|}{\phi_0\rho_0^{-2\nu}}. \tag{3.19}$$

Let P_1 and P_2 respectively stand for the two terms in the right hand side of (3.19). For clarity, we break the proof into two parts. The first part is devoted to uniformly control P_1. Strictly speaking, we prove that

$$P_1\lesssim\left(\mathbb{1}_{\{d=1\}}\frac{1}{n}+\mathbb{1}_{\{d=2\}}\frac{\log n}{n}+\mathbb{1}_{\{d\geq 3\}}n^{-2/d}\right)\left(1+\mathbb{1}_{\{m=\nu+d/2\}}\log n\right).$$

We then show that the stochastic quadratic quantity P_2 is of order $\sqrt{n^{-1}\log n}$, with high probability. The concentration inequalities involving the quadratic forms (and their supremum over a bounded space) of GPs presented in [KSN16] are crucial for bounding P_2 from above.

Choose an arbitrary $(\phi,\rho)\in\mathcal{I}\times\Theta_0$. Recall $K^{\mathcal{B}}_{n,m}(\rho)\in\mathbb{R}^{n\times n}$ from (3.8) and $\hat{\phi}_{n,\mathcal{B}}(\rho)$ from Eq. (3.13). For brevity, define $L^{\mathcal{B}}_{n,m}(\rho):=\rho^{2\nu}K^{\mathcal{B}}_{n,m}(\rho)$. Observe that

$$\begin{aligned}\frac{\mathbb{E}\hat{\phi}_{n,\mathcal{B}}(\rho)\rho^{-2\nu}}{\phi_0\rho_0^{-2\nu}} &= \frac{\rho^{-2\nu}}{\phi_0\rho_0^{-2\nu}}\frac{\mathbb{E}Y^\top K^{\mathcal{B}}_{n,m}(\rho)Y}{\left\|K^{\mathcal{B}}_{n,m}(\rho)\right\|^2_{\ell_2}}=\left(\frac{\rho_0}{\rho}\right)^{2\nu}\frac{\langle K^{\mathcal{B}}_{n,m}(\rho),K^{\mathcal{B}}_{n,m}(\rho_0)\rangle}{\left\|K^{\mathcal{B}}_{n,m}(\rho)\right\|^2_{\ell_2}}\\ &= \frac{\langle L^{\mathcal{B}}_{n,m}(\rho),L^{\mathcal{B}}_{n,m}(\rho_0)\rangle}{\left\|L^{\mathcal{B}}_{n,m}(\rho)\right\|^2_{\ell_2}}.\end{aligned}$$

Thus,

$$\begin{aligned}P_1 &= \sup_{\rho\in\Theta_0}\left|\frac{\langle L^{\mathcal{B}}_{n,m}(\rho),L^{\mathcal{B}}_{n,m}(\rho_0)\rangle}{\left\|L^{\mathcal{B}}_{n,m}(\rho)\right\|^2_{\ell_2}}-1\right|=\sup_{\rho\in\Theta_0}\left|\frac{\langle L^{\mathcal{B}}_{n,m}(\rho)-L^{\mathcal{B}}_{n,m}(\rho_0),L^{\mathcal{B}}_{n,m}(\rho)\rangle}{\left\|L^{\mathcal{B}}_{n,m}(\rho)\right\|^2_{\ell_2}}\right|\\ &\overset{(a)}{\leq} \sup_{\rho\in\Theta_0}\left[\frac{\left\|L^{\mathcal{B}}_{n,m}(\rho)-L^{\mathcal{B}}_{n,m}(\rho_0)\right\|_{\mathcal{S}_1}\left\|L^{\mathcal{B}}_{n,m}(\rho)\right\|_{2\to 2}}{\left\|L^{\mathcal{B}}_{n,m}(\rho)\right\|^2_{\ell_2}}\right].\end{aligned} \tag{3.20}$$

Here (a) is implied by the generalized Cauchy-Schwartz inequality. We assess the large sample behaviour of the terms appearing in the second line of (3.20) in Section 3.8.1. Lemma 3.6 states that $\min_{\rho\in\Theta_0}\left\|L^{\mathcal{B}}_{n,m}(\rho)\right\|_{\ell_2}\gtrsim\sqrt{n}$. For brevity define $\Delta^{\mathcal{B}}(\rho,\rho_0):=L^{\mathcal{B}}_{n,m}(\rho)-$

$L^{\mathcal{B}}_{n,m}(\rho_0)$. Furthermore, Lemma 3.3 implies that

$$\begin{aligned}\sup_{\rho\in\Theta_0}\left\|\Delta^{\mathcal{B}}(\rho,\rho_0)\right\|_{\mathcal{S}_1} &\lesssim \left(\mathbb{1}_{\{d=1\}}+\mathbb{1}_{\{d=2\}}\log n+\mathbb{1}_{\{d\geq 3\}}n^{1-2/d}\right)\operatorname{diam}(\Theta_0)\\ &\asymp \left(\mathbb{1}_{\{d=1\}}+\mathbb{1}_{\{d=2\}}\log n+\mathbb{1}_{\{d\geq 3\}}n^{1-2/d}\right).\end{aligned} \tag{3.21}$$

Thus the upper bound on P_1 in (3.20) can be rewritten as

$$P_1 \lesssim \left(\frac{\mathbb{1}_{\{d=1\}}}{n}+\mathbb{1}_{\{d=2\}}\frac{\log n}{n}+\mathbb{1}_{\{d\geq 3\}}n^{-2/d}\right)\sup_{\rho\in\Theta_0}\left\|L^{\mathcal{B}}_{n,m}(\rho)\right\|_{2\to 2}. \tag{3.22}$$

So it is only needed to find a uniform upper bound on the largest eigenvalue of $L^{\mathcal{B}}_{n,m}(\rho)$ on Θ_0. Notice that $L^{\mathcal{B}}_{n,m}(\rho)$ is a block diagonalized version of $L_{n,m}(\rho)$. Hence

$$\left\|L^{\mathcal{B}}_{n,m}(\rho)\right\|_{2\to 2}\leq\left\|L_{n,m}(\rho)\right\|_{2\to 2},\quad \forall\,\rho\in\Theta_0$$

In other words, we only need to focus on the case of no partitioning. For d-dimensional regular lattices, the exact procedure as Theorems 2.1 and 2.3 of [SCA12] demonstrates that all the eigenvalues of $L_{n,m}(\rho)$ are universally bounded. Namely,

$$\sup_{\rho\in\Theta_0}\lambda_j(L_{n,m}(\rho))\leq\alpha_{\max},\quad \forall\ j=1,\ldots,|\mathcal{D}_n| \tag{3.23}$$

for some bounded $\alpha^{\max}>0$. Thus P_1 admits the following inequality for regular lattices.

$$P_1\lesssim\left(\frac{\mathbb{1}_{\{d=1\}}}{n}+\mathbb{1}_{\{d=2\}}\frac{\log n}{n}+\mathbb{1}_{\{d\geq 3\}}n^{-2/d}\right). \tag{3.24}$$

However the operator norm of $L_{n,m}(\rho)$ is not necessarily uniformly bound on Θ_0, for a general irregular lattice satisfying Assumption 3.1. For such case, we show in Proposition 3.1 that

$$\left|(L_{n,m}(\rho))_{\boldsymbol{s},\boldsymbol{t}}\right|\lesssim\left(1+\lfloor n^{1/d}\rfloor\|\boldsymbol{t}-\boldsymbol{s}\|_{\ell_2}\right)^{-2(m-\nu)},\quad \boldsymbol{s},\boldsymbol{t}\in\mathcal{D}_n. \tag{3.25}$$

Lemma 3.9 also introduces an upper bound on the operator norm of the matrices satisfying (3.25). Applying Lemma 3.9 yields

$$\sup_{\rho\in\Theta_0}\left\|L^{\mathcal{B}}_{n,m}(\rho)\right\|_{2\to 2}\leq\sup_{\rho\in\Theta_0}\left\|L_{n,m}(\rho)\right\|_{2\to 2}\lesssim\left(1+\mathbb{1}_{\{m=\nu+d/2\}}\log n\right). \tag{3.26}$$

The desired bound on P_1 is obtained by combining (3.22) and (3.26). The next goal is control P_2 from above. Let $Z\in\mathbb{R}^n$ be a standard Gaussian vector and define the symmetric

matrix $M^{\mathcal{B}}_{n,m}(\rho)$ by

$$M^{\mathcal{B}}_{n,m}(\rho) = \sqrt{L_{n,m}(\rho_0)}\left[\frac{nL^{\mathcal{B}}_{n,m}(\rho)}{\left\|L^{\mathcal{B}}_{n,m}(\rho)\right\|^2_{\ell_2}}\right]\sqrt{L_{n,m}(\rho_0)}, \quad \forall\, \rho \in \Theta_0. \tag{3.27}$$

We first introduce an equivalent representation for $\hat{\phi}_{n,\mathcal{B}}(\rho)\rho^{-2\nu}$ in terms of Z and $M^{\mathcal{B}}_{n,m}(\rho)$. Obviously, the Gaussian vectors Y and $\sqrt{\phi_0 K_{n,m}(\rho_0)}Z = \phi_0^{1/2}\rho_0^{-\nu}\sqrt{L_{n,m}(\rho_0)}Z$ have the same distribution. Thus,

$$\hat{\phi}_{n,\mathcal{B}}(\rho)\rho^{-2\nu} = \rho^{-2\nu}\frac{Y^\top K^{\mathcal{B}}_{n,m}(\rho)\,Y}{\left\|K^{\mathcal{B}}_{n,m}(\rho)\right\|^2_{\ell_2}} = \frac{Y^\top L^{\mathcal{B}}_{n,m}(\rho)\,Y}{\left\|L^{\mathcal{B}}_{n,m}(\rho)\right\|^2_{\ell_2}} \overset{d}{=} \frac{Z^\top M^{\mathcal{B}}_{n,m}(\rho)Z}{n}\phi_0\rho_0^{-2\nu},$$

and so P_2 can be rewritten as the supremum of a centered χ^2 process over Θ_0, i.e.,

$$P_2 = \frac{1}{n}\sup_{\rho\in\Theta_0}\left|Z^\top M^{\mathcal{B}}_{n,m}(\rho)Z - tr\left\{M^{\mathcal{B}}_{n,m}(\rho)\right\}\right|.$$

So if $M^{\mathcal{B}}_{n,m}(\rho)$ admits the three conditions in Proposition 3.2, then there is a bounded scalar C such that as $n \to \infty$, we have

$$\mathbb{P}\left(P_2 \geq C\sqrt{\frac{\log n}{n}}\right) = \mathbb{P}\left(\sup_{\rho\in\Theta_0}\left|Z^\top M^{\mathcal{B}}_{n,m}(\rho)Z - tr\left\{M^{\mathcal{B}}_{n,m}(\rho)\right\}\right| \geq C\sqrt{n\log n}\right) \to 0. \tag{3.28}$$

Thus we require to verify the conditions $(a)-(c)$ in Proposition 3.2.

Validating condition (a). We should substantiate the uniform boundedness of $n^{-1/2}\left\|M^{\mathcal{B}}_{n,m}(\rho)\right\|_{\ell_2}$ over Θ_0. Namely, we must prove that U defined as the following is bounded.

$$U \coloneqq \sup_{\rho\in\Theta_0}\frac{\left\|M^{\mathcal{B}}_{n,m}(\rho)\right\|_{\ell_2}}{\sqrt{n}} = \sup_{\rho\in\Theta_0}\frac{\sqrt{n}\left\|\sqrt{L_{n,m}(\rho_0)}L^{\mathcal{B}}_{n,m}(\rho)\sqrt{L_{n,m}(\rho_0)}\right\|_{\ell_2}}{\left\|L^{\mathcal{B}}_{n,m}(\rho)\right\|^2_{\ell_2}}.$$

We prove in Lemma 3.6 that $\min_{\rho\in\Theta_0} n^{-1}\left\|L^{\mathcal{B}}_{n,m}(\rho)\right\|^2_{\ell_2} > 0$ for large enough n. Thus, U can be bounded above by some U' given by

$$U \lesssim U' \coloneqq \sup_{\rho\in\Theta_0}\frac{\left\|\sqrt{L_{n,m}(\rho_0)}L^{\mathcal{B}}_{n,m}(\rho)\sqrt{L_{n,m}(\rho_0)}\right\|_{\ell_2}}{\sqrt{n}}.$$

Finally, Lemma 3.7 ensures the boundedness of U' (and consequently U).

Validating condition (b). Pick two arbitrary distinct $\rho_1, \rho_2 \in \Theta_0$ with $|\rho_2 - \rho_1| \leq 1$. Our

objective is to demonstrate the Lipschitz property of $\left\|M^{\mathcal{B}}_{n,m}(\rho_2)-M^{\mathcal{B}}_{n,m}(\rho_1)\right\|_{2\to 2}$ (with a constant of order $\log^2 n$). Obviously

$$\frac{\left\|M^{\mathcal{B}}_{n,m}(\rho_2)-M^{\mathcal{B}}_{n,m}(\rho_1)\right\|_{2\to 2}}{n|\rho_2-\rho_1|}\leq\frac{\left\|L_{n,m}(\rho_0)\right\|_{2\to 2}}{|\rho_2-\rho_1|}\left\|\frac{L^{\mathcal{B}}_{n,m}(\rho_2)}{\left\|L^{\mathcal{B}}_{n,m}(\rho_2)\right\|^2_{\ell_2}}-\frac{L^{\mathcal{B}}_{n,m}(\rho_1)}{\left\|L^{\mathcal{B}}_{n,m}(\rho_1)\right\|^2_{\ell_2}}\right\|_{2\to 2}.$$

We have argued in (3.26) that $\left\|L_{n,m}(\rho_0)\right\|_{2\to 2}\lesssim\left(1+\mathbb{1}_{\{m=\nu+d/2\}}\log n\right)\leq\log n$. Hence,

$$\frac{\left\|M^{\mathcal{B}}_{n,m}(\rho_2)-M^{\mathcal{B}}_{n,m}(\rho_1)\right\|_{2\to 2}}{n|\rho_2-\rho_1|\log n}\lesssim\frac{1}{|\rho_2-\rho_1|}\left\|\frac{L^{\mathcal{B}}_{n,m}(\rho_2)}{\left\|L^{\mathcal{B}}_{n,m}(\rho_2)\right\|^2_{\ell_2}}-\frac{L^{\mathcal{B}}_{n,m}(\rho_1)}{\left\|L^{\mathcal{B}}_{n,m}(\rho_1)\right\|^2_{\ell_2}}\right\|_{2\to 2}. \tag{3.29}$$

Furthermore, we know from the triangle inequality that

$$\begin{aligned}\left\|\frac{L^{\mathcal{B}}_{n,m}(\rho_2)}{\left\|L^{\mathcal{B}}_{n,m}(\rho_2)\right\|^2_{\ell_2}}-\frac{L^{\mathcal{B}}_{n,m}(\rho_1)}{\left\|L^{\mathcal{B}}_{n,m}(\rho_1)\right\|^2_{\ell_2}}\right\|_{2\to 2}&\leq\frac{\left\|L^{\mathcal{B}}_{n,m}(\rho_2)-L^{\mathcal{B}}_{n,m}(\rho_1)\right\|_{2\to 2}}{\left\|L^{\mathcal{B}}_{n,m}(\rho_2)\right\|^2_{\ell_2}}\\&+\left\|\frac{L^{\mathcal{B}}_{n,m}(\rho_1)}{\left\|L^{\mathcal{B}}_{n,m}(\rho_2)\right\|^2_{\ell_2}}-\frac{L^{\mathcal{B}}_{n,m}(\rho_1)}{\left\|L^{\mathcal{B}}_{n,m}(\rho_1)\right\|^2_{\ell_2}}\right\|_{2\to 2}.\end{aligned} \tag{3.30}$$

Let $\Psi^1_n(\rho_1,\rho_2)$ and $\Psi^2_n(\rho_1,\rho_2)$ stand for the first and second terms in the right hand side of (3.30), which we aim to control from above. The fact that $\min_{\rho\in\Theta_0} n^{-1}\left\|L^{\mathcal{B}}_{n,m}(\rho)\right\|^2_{\ell_2}>0$ (see Lemma 3.6) comes in handy for finding a simpler upper bound on $\Psi^1_n(\rho_1,\rho_2)$ and $\Psi^2_n(\rho_1,\rho_2)$.

$$\Psi^1_n(\rho_1,\rho_2):=\frac{\left\|L^{\mathcal{B}}_{n,m}(\rho_2)-L^{\mathcal{B}}_{n,m}(\rho_1)\right\|_{2\to 2}}{\left\|L^{\mathcal{B}}_{n,m}(\rho_2)\right\|^2_{\ell_2}}\lesssim\frac{\left\|L^{\mathcal{B}}_{n,m}(\rho_2)-L^{\mathcal{B}}_{n,m}(\rho_1)\right\|_{2\to 2}}{n}.$$

Furthermore, Lemma 3.4 indicates that

$$\left\|L^{\mathcal{B}}_{n,m}(\rho_2)-L^{\mathcal{B}}_{n,m}(\rho_1)\right\|_{2\to 2}\lesssim\left(1+\mathbb{1}_{\{m=\nu+d/2\}}\log n\right)|\rho_2-\rho_1|\leq|\rho_2-\rho_1|\log n.$$

So $\Psi^1_n(\rho_1,\rho_2)\lesssim n^{-1}\log n|\rho_2-\rho_1|$. Now we consider $\Psi^2_n(\rho_1,\rho_2)$. Observe that

$$\begin{aligned}\Psi^2_n(\rho_1,\rho_2)&:=\left\|L^{\mathcal{B}}_{n,m}(\rho_1)\right\|_{2\to 2}\left(\frac{\left\|L^{\mathcal{B}}_{n,m}(\rho_2)\right\|^2_{\ell_2}-\left\|L^{\mathcal{B}}_{n,m}(\rho_1)\right\|^2_{\ell_2}}{\left\|L^{\mathcal{B}}_{n,m}(\rho_1)\right\|^2_{\ell_2}\left\|L^{\mathcal{B}}_{n,m}(\rho_2)\right\|^2_{\ell_2}}\right)\\&\leq\left\|L^{\mathcal{B}}_{n,m}(\rho_2)\right\|_{2\to 2}\frac{\left\|L^{\mathcal{B}}_{n,m}(\rho_2)\right\|_{\ell_2}+\left\|L^{\mathcal{B}}_{n,m}(\rho_1)\right\|_{\ell_2}}{\left\|L^{\mathcal{B}}_{n,m}(\rho_1)\right\|^2_{\ell_2}\left\|L^{\mathcal{B}}_{n,m}(\rho_2)\right\|^2_{\ell_2}}\left\|L^{\mathcal{B}}_{n,m}(\rho_2)-L^{\mathcal{B}}_{n,m}(\rho_1)\right\|_{\ell_2}.\end{aligned}$$

It is known from (3.26) that $\left\|L^{\mathcal{B}}_{n,m}(\rho_2)\right\|_{2\to 2} \lesssim \log n$. Moreover, it is easy to verify that

$$\begin{aligned}\frac{\left\|L^{\mathcal{B}}_{n,m}(\rho_2)\right\|_{\ell_2} + \left\|L^{\mathcal{B}}_{n,m}(\rho_1)\right\|_{\ell_2}}{\left\|L^{\mathcal{B}}_{n,m}(\rho_1)\right\|^2_{\ell_2}\left\|L^{\mathcal{B}}_{n,m}(\rho_2)\right\|^2_{\ell_2}} &= \frac{1/\left\|L^{\mathcal{B}}_{n,m}(\rho_1)\right\|_{\ell_2} + 1/\left\|L^{\mathcal{B}}_{n,m}(\rho_2)\right\|_{\ell_2}}{\left\|L^{\mathcal{B}}_{n,m}(\rho_1)\right\|_{\ell_2}\left\|L^{\mathcal{B}}_{n,m}(\rho_2)\right\|_{\ell_2}} \lesssim \frac{1/\sqrt{n}+1\sqrt{n}}{\sqrt{n}\sqrt{n}} \\ &\asymp n^{-3/2}.\end{aligned}$$

Thus, the upper bound on $\Psi^2_n(\rho_1,\rho_2)$ can be simplified as

$$\begin{aligned}\frac{\Psi^2_n(\rho_1,\rho_2)}{|\rho_2-\rho_1|} &\leq \frac{\log n}{n^{3/2}}\left(\frac{\left\|L^{\mathcal{B}}_{n,m}(\rho_2)-L^{\mathcal{B}}_{n,m}(\rho_1)\right\|_{\ell_2}}{|\rho_2-\rho_1|}\right) \\ &\overset{(c)}{\lesssim} \frac{\log n}{n^{3/2}}\left(\mathbb{1}_{\{d=1\}} + \mathbb{1}_{\{d=2\}}\log n + \mathbb{1}_{\{d=3\}}n^{1/3} + \mathbb{1}_{\{d\geq 4\}}n^{1/2}\right) \\ &= \frac{\log n}{n}\left(\mathbb{1}_{\{d=1\}}\frac{1}{\sqrt{n}} + \mathbb{1}_{\{d=2\}}\frac{\log n}{\sqrt{n}} + \mathbb{1}_{\{d=3\}}n^{-1/6} + \mathbb{1}_{\{d>3\}}\right) \lesssim \frac{\log n}{n},\end{aligned}$$

where the inequality (c) follows from Lemma 3.5. In summary, (3.29) can be rewritten as

$$\frac{\left\|M^{\mathcal{B}}_{n,m}(\rho_2)-M^{\mathcal{B}}_{n,m}(\rho_1)\right\|_{2\to 2}}{|\rho_2-\rho_1|} \leq n\log n\left(\frac{\Psi^1_n(\rho_1,\rho_2)+\Psi^2_n(\rho_1,\rho_2)}{|\rho_2-\rho_1|}\right) \lesssim n\log n\frac{\log n}{n} = \log^2 n,$$

showing that the condition (b) of Proposition 3.2 holds.

Validating condition (c). Choose an arbitrary $\rho \in \Theta_0$. We should prove that V_n, which is defined as the following, converging to zero as n goes to infinity.

$$V_n \coloneqq \left\|M^{\mathcal{B}}_{n,m}(\rho)\right\|_{2\to 2}\sqrt{\frac{\log n}{n}}. \tag{3.31}$$

V_n can be equivalently written as

$$V_n = \frac{\left\|\sqrt{L_{n,m}(\rho_0)}L^{\mathcal{B}}_{n,m}(\rho)\sqrt{L_{n,m}(\rho_0)}\right\|_{2\to 2}\sqrt{n\log n}}{\left\|L^{\mathcal{B}}_{n,m}(\rho)\right\|^2_{\ell_2}}.$$

Lemma 3.6, which says the Frobenius norm of $L_{n,m}(\rho)$ is of order $\sqrt{n}$ (uniformly on Θ_0) provides a simpler asymptotic expression for V_n.

$$V_n \asymp \left\|\sqrt{L_{n,m}(\rho_0)}L^{\mathcal{B}}_{n,m}(\rho)\sqrt{L_{n,m}(\rho_0)}\right\|_{2\to 2}\sqrt{\frac{\log n}{n}} \leq \left\|L^{\mathcal{B}}_{n,m}(\rho)\right\|_{2\to 2}\left\|L_{n,m}(\rho_0)\right\|_{2\to 2}\sqrt{\frac{\log n}{n}}.$$

We refer the reader to Eq. (3.26) for an upper bound on the operator norm of $L_{n,m}$ and $L^{\mathcal{B}}_{n,m}$

matrices over Θ_0. So, V_n can be bounded above by

$$V_n \lesssim \left(1 + \mathbb{1}_{\{m=\nu+d/2\}} \log n\right)^2 \sqrt{\frac{\log n}{n}} \to 0, \quad \text{as } n \to \infty. \tag{3.32}$$

□

Proof of Theorem 3.2. Let $\rho_{\max}$ and $\rho_{\min}$ respectively denote the largest and smallest element of Θ_0. Recall the positive semi-definite class of matrices $L^{\mathcal{B}}_{n,m}(\rho) \coloneqq \rho^{2\nu} K^{\mathcal{B}}_{n,m}(\rho)$, $\rho \in \Theta_0$. Moreover, define

$$T_n(\rho, Y) \coloneqq \sqrt{n}\left(\frac{\hat{\phi}_{n,\mathcal{B}}(\rho)\rho^{-2\nu}}{\phi_0 \rho_0^{-2\nu}} - 1\right) = \sqrt{n}\left(\frac{Y^\top L^{\mathcal{B}}_{n,m}(\rho)\, Y}{\phi_0 \rho_0^{-2\nu} \left\| L^{\mathcal{B}}_{n,m}(\rho) \right\|^2_{\ell_2}} - 1\right). \tag{3.33}$$

For notational convenience, the dependence to ϕ_0, ρ_0 and m has been dropped in T_n. We aim to show that $\sigma_n^{-1} T_n(\hat{\rho}_n, Y) \xrightarrow{d} N(0,1)$ for some scalar bounded sequence σ_n. The proof is broken into two parts for easier digestion. We first find probabilistic upper and lower bounds on $T_n(\hat{\rho}_n, Y)$ in terms of $T_n(\rho_{\max}, Y)$ and $T_n(\rho_{\min}, Y)$. The precise statement of this claim is as following.

Claim 1. There are non-negative sequences of random variables $\{p_n\}_{n=1}^\infty$ and $\{q_n\}_{n=1}^\infty$ converging to zero in probability and scalar $n_0 \in \mathbb{N}$ (depending on ρ_0, m, d, ν, and Θ_0) such that for any $n \geq n_0$

$$T_n(\rho_{\min}, Y)(1 - p_n) \leq T_n(\hat{\rho}_n, Y) \leq T_n(\rho_{\max}, Y)(1 + q_n). \tag{3.34}$$

Next, we substantiate the asymptotic normality of $T_n(\rho, Y)$ for an arbitrary $\rho \in \Theta_0$.

Claim 2. There is a bounded sequence $\sigma_{n,m}$ such that $\frac{1}{\sigma_{n,m}} T_n(\rho, Y) \xrightarrow{d} N(0,1)$, for any fixed $\rho \in \Theta_0$.

As both upper and lower bounds on $\sigma_{n,m}^{-1} T_n(\hat{\rho}_n, Y)$ in (3.34) weakly converge to a random variable distributed as $N(0,1)$, the squeeze theorem for the weak convergence (see Lemma 3.11 for its rigorous statement) concludes the proof. The rest of the proof serves to establish Claims 1 and 2.

Proof of Claim 1. Define $T'_n(\rho) \coloneqq 1 + T_n(\rho, Y)/\sqrt{n}$. Claim 2 obviously holds if we can show that

$$T'_n(\rho_{\min})(1 - p'_n) \leq T'_n(\hat{\rho}_n) \leq T'_n(\rho_{\max})(1 + q'_n), \tag{3.35}$$

for any realization of Y and for sequences $\{p'_n\}_{n=1}^{\infty}, \{q'_n\}_{n=1}^{\infty}$ converging to zero faster than $n^{-1/2}$. Let Z be a standard Gaussian column vector with the same length as Y. Define $U \coloneqq \sqrt{L_{n,m}(\rho_0)}Z$, which obviously has no dependence on ρ. Then,

$$T'_n(\rho) = \frac{U^\top L^{\mathcal{B}}_{n,m}(\rho)\, U}{\left\|L^{\mathcal{B}}_{n,m}(\rho)\right\|^2_{\ell_2}}, \tag{3.36}$$

We only prove the right hand side inequality in Eq. (3.35) and the other side can be shown similarly. We separately analyze the numerator and denominator in (3.36). We know that $L^{\mathcal{B}}_{n,m}(\rho) \le L^{\mathcal{B}}_{n,m}(\rho_{\max})$ for any $\rho \in \Theta_0$ (see (3.61) for the details). Thus, $U^\top L^{\mathcal{B}}_{n,m}(\rho)\, U \le U^\top L^{\mathcal{B}}_{n,m}(\rho_{\max})\, U$ almost surely. Namely,

$$T'_n(\rho) \le \frac{U^\top L^{\mathcal{B}}_{n,m}(\rho_{\max})\, U}{\left\|L^{\mathcal{B}}_{n,m}(\rho)\right\|^2_{\ell_2}} \quad \Leftrightarrow \quad \left\{\frac{T'_n(\rho)}{T'_n(\rho_{\max})} - 1\right\} \le \frac{\left\|L^{\mathcal{B}}_{n,m}(\rho_{\max})\right\|^2_{\ell_2} - \left\|L^{\mathcal{B}}_{n,m}(\rho)\right\|^2_{\ell_2}}{\left\|L^{\mathcal{B}}_{n,m}(\rho)\right\|^2_{\ell_2}}. \tag{3.37}$$

Recall that we have defined $\Delta^{\mathcal{B}}(\rho_2,\rho_1) \coloneqq L^{\mathcal{B}}_{n,m}(\rho_2) - L^{\mathcal{B}}_{n,m}(\rho_1)$, for any $\rho_1,\rho_2 \in \Theta_0$. It is sufficient to show that

$$q'_n \coloneqq \frac{\left\|L^{\mathcal{B}}_{n,m}(\rho_{\max})\right\|^2_{\ell_2} - \left\|L^{\mathcal{B}}_{n,m}(\rho)\right\|^2_{\ell_2}}{\left\|L^{\mathcal{B}}_{n,m}(\rho)\right\|^2_{\ell_2}} = o\left(\frac{1}{\sqrt{n}}\right), \quad \text{as } n \to \infty. \tag{3.38}$$

As we know from Lemma 3.6 that $\left\|L^{\mathcal{B}}_{n,m}(\rho)\right\|_{\ell_2} \gtrsim \sqrt{n}$, we just need to show that

$$\psi_n \coloneqq \left\|L^{\mathcal{B}}_{n,m}(\rho_{\max})\right\|^2_{\ell_2} - \left\|L^{\mathcal{B}}_{n,m}(\rho)\right\|^2_{\ell_2} = o\left(\sqrt{n}\right), \quad \text{as } n \to \infty.$$

On the other hand we have

$$\begin{aligned}
\left\|L^{\mathcal{B}}_{n,m}(\rho_{\max})\right\|^2_{\ell_2} - \left\|L^{\mathcal{B}}_{n,m}(\rho)\right\|^2_{\ell_2} &= \left\|L^{\mathcal{B}}_{n,m}(\rho_{\max})\right\|^2_{\ell_2} - \left\|L^{\mathcal{B}}_{n,m}(\rho_{\max}) - \Delta^{\mathcal{B}}(\rho_{\max},\rho)\right\|^2_{\ell_2} \\
&\le 2\langle L^{\mathcal{B}}_{n,m}(\rho_{\max}), \Delta^{\mathcal{B}}(\rho_{\max},\rho)\rangle \\
&\le 2\left\|L^{\mathcal{B}}_{n,m}(\rho_{\max})\right\|_{2\to 2}\left\|\Delta^{\mathcal{B}}(\rho_{\max},\rho)\right\|_{\mathcal{S}_1}.
\end{aligned}$$

Eq. (3.26) provides an upper bounds on $\left\|L^{\mathcal{B}}_{n,m}(\rho_{\max})\right\|_{2\to 2}$. So

$$\begin{aligned}
\psi_n &\le 2\left\|L^{\mathcal{B}}_{n,m}(\rho_{\max})\right\|_{2\to 2}\left\|\Delta^{\mathcal{B}}(\rho_{\max},\rho)\right\|_{\mathcal{S}_1} \lesssim \left(1 + \mathbb{1}_{\{m=\nu+d/2\}}\log n\right)\left\|\Delta^{\mathcal{B}}(\rho_{\max},\rho)\right\|_{\mathcal{S}_1} \\
&\le \left\|\Delta^{\mathcal{B}}(\rho_{\max},\rho)\right\|_{\mathcal{S}_1}\log n.
\end{aligned}$$

We now employ analogous techniques as Eq. (3.21) (see also Lemma 3.3) to control

$\left\|\Delta^{\mathcal{B}}(\rho_{\max},\rho)\right\|_{\mathcal{S}_1}$ from above. Since we only consider the case of $d \leq 3$, the bound in Eq. (3.21) can be rewritten as the following.

$$\exists\, 0 < \gamma < \frac{1}{2},\ \text{s.t.}\ \left\|\Delta^{\mathcal{B}}(\rho_{\max},\rho)\right\|_{\mathcal{S}_1} \lesssim n^{\gamma}. \tag{3.39}$$

Thus ψ_n can be upper bounded by $\psi_n \lesssim n^{\gamma}\log n = o\left(\sqrt{n}\right)$, which concludes the proof. □

Proof of Claim 2. For brevity let $\xi_n \coloneqq T_n(\rho,Y) + \sqrt{n}$. We suppress the dependence of ρ and Y on ξ_n. Let us decompose $T_n(\rho,Y)$ into two parts as

$$\begin{aligned} T_n(\rho,Y) &= \left(\frac{T_n(\rho,Y) - \mathbb{E}T_n(\rho,Y)}{\sqrt{\operatorname{var} T_n(\rho,Y)}}\right)\sqrt{\operatorname{var} T_n(\rho,Y)} + \mathbb{E}T_n(\rho,Y) \\ &= \left(\frac{\xi_n - \mathbb{E}\xi_n}{\sqrt{\operatorname{var}\xi_n}}\right)\sqrt{\operatorname{var}\xi_n} + \mathbb{E}T_n(\rho,Y). \end{aligned} \tag{3.40}$$

Recall that we defined $P_1 \coloneqq \sup_{\rho\in\Theta_0} n^{-1/2}\mathbb{E}T_n(\rho,Y)$ in the proof of Theorem 3.1. A prudent look at Eqs. (3.22) and (3.24) reveals that $P_1 \lesssim n^{\gamma-1}\log n$ for some $\gamma < 1/2$ (γ is the same as in (3.39)). Hence,

$$\mathbb{E}T_n(\rho,Y) \leq \sqrt{n}P_1 \lesssim n^{-1/2+\gamma}\log n \to 0, \quad \text{as } n \to \infty.$$

Namely, $\mathbb{E}T_n(\rho,Y)$ tends to zero as n grows to infinity. Thus, it is sufficient to obtain the asymptotic distribution of the first term in the right hand side of (3.40). Now we express ξ_n as a quadratic term of a Gaussian random vector. Using identity (3.33), one can easily show that

$$\xi_n \overset{d}{=} Z^{\top}\frac{M^{\mathcal{B}}_{n,m}(\rho)}{\sqrt{n}}Z, \tag{3.41}$$

in which Z is a standard Gaussian vector of proper size and $M^{\mathcal{B}}_{n,m}(\rho)$ has been defined in (3.27). The explicit expressions for the expected value and standard deviation of ξ_n are given by

$$\mathbb{E}\xi_n = \sqrt{\frac{1}{n}}\,tr\left\{M^{\mathcal{B}}_{n,m}(\rho)\right\}, \quad \sqrt{\operatorname{var}\xi_n} = \sqrt{\frac{2}{n}}\left\|M^{\mathcal{B}}_{n,m}(\rho)\right\|_{\ell_2}.$$

We showed in the proof of Theorem 3.1 that $\left\|M^{\mathcal{B}}_{n,m}(\rho)\right\|_{2\to 2} / \left\|M^{\mathcal{B}}_{n,m}(\rho)\right\|_{\ell_2} \to 0$ when $n \to \infty$ (see (3.31) and (3.32)). Thus applying Lemma $A.4$ of [KSN16], on asymptotic normality of the normalized generalized χ^2 random variables, leads to

$$\left(\frac{\xi_n - \mathbb{E}\xi_n}{\sqrt{\operatorname{var}\xi_n}}\right) \overset{d}{\to} N(0,1).$$

Finally we study the limiting behaviour of $\sqrt{\operatorname{var}\xi_n}$, which is denoted by $\sigma_{n,m}(\rho,\rho_0)$. Notice that

$$\sigma_{n,m}(\rho,\rho_0) \coloneqq \sqrt{\frac{2}{n}}\left\|M^{\mathcal{B}}_{n,m}(\rho)\right\|_{\ell_2} = \frac{\sqrt{2n}}{\left\|L^{\mathcal{B}}_{n,m}(\rho)\right\|^2_{\ell_2}}\left\|\sqrt{L_{n,m}(\rho_0)}L^{\mathcal{B}}_{n,m}(\rho)\sqrt{L_{n,m}(\rho_0)}\right\|_{\ell_2}.$$

We claim that

$$\lim_{n\to\infty}\frac{\sigma_{n,m}(\rho,\rho_0)}{\sigma_{n,m}(\rho_1,\rho_2)} = 1, \quad \forall\,\rho_1,\rho_2\in\Theta_0. \tag{3.42}$$

Thus, $\sigma_{n,m}$ has no dependence to ρ, ρ_0, and Θ_0. In other words, $\sigma_{n,m}$ only depends on m,d,ν, and the topology of $\mathcal{D}_n$. Assuming that the claim holds, for proving the boundedness of $\sigma_{n,m}$, we just need to check that $\sigma_{n,m}(\rho,\rho_0)\asymp 1$ for some $\rho_1',\rho_2'\in\Theta_0$. Applying Lemma 3.6 on the denominator of $\sigma_{n,m}\left(\rho_1',\rho_2'\right)$, we get,

$$f_{n,m}\left(\rho_1',\rho_2'\right)\lesssim\frac{\left\|\sqrt{L_{n,m}\left(\rho_2'\right)}L^{\mathcal{B}}_{n,m}\left(\rho_1'\right)\sqrt{L_{n,m}\left(\rho_2'\right)}\right\|_{\ell_2}}{\sqrt{n}}.$$

So, $\sigma_{n,m}\left(\rho_1',\rho_2'\right)\asymp 1$ as a result of Lemma 3.7. We now turn to substantiate (3.42). It is sufficient to verify the following identities for any $\rho_1,\rho_2\in\Theta_0$.

$$\lim_{n\to\infty}\frac{\sigma_{n,m}(\rho,\rho_0)}{\sigma_{n,m}(\rho_1,\rho_0)} = 1, \quad \lim_{n\to\infty}\frac{\sigma_{n,m}(\rho_1,\rho_0)}{\sigma_{n,m}(\rho_1,\rho_2)} = 1. \tag{3.43}$$

To avoid repetition, we only demonstrate the left hand side identity in (3.43) and the other one can be substantiated using analogous techniques. Observe that

$$\frac{\sigma_{n,m}(\rho,\rho_0)}{\sigma_{n,m}(\rho_1,\rho_0)} = \left[\frac{\left\|L^{\mathcal{B}}_{n,m}(\rho_1)\right\|_{\ell_2}}{\left\|L^{\mathcal{B}}_{n,m}(\rho)\right\|_{\ell_2}}\right]^2\frac{\left\|\sqrt{L_{n,m}(\rho_0)}L^{\mathcal{B}}_{n,m}(\rho)\sqrt{L_{n,m}(\rho_0)}\right\|_{\ell_2}}{\left\|\sqrt{L_{n,m}(\rho_0)}L^{\mathcal{B}}_{n,m}(\rho_1)\sqrt{L_{n,m}(\rho_0)}\right\|_{\ell_2}} \coloneqq a_n b_n.$$

We prove that both a_n and b_n converge to one as n tends to infinity. Notice that $|a_n-1|$ has the same limiting behaviour as q_n' defined at (3.38). So for avoiding the redundancy we just state that $|a_n-1|\lesssim n^{\gamma-1}\log n = o\left(n^{-1/2}\right)$ and refer the reader to the proof of Claim 1. The last step of the proof is devoted to control $|b_n-1|$ from above.

$$
\begin{aligned}
|b_n - 1| &= \left| \frac{\left\| \sqrt{L_{n,m}(\rho_0)} L^{\mathcal{B}}_{n,m}(\rho) \sqrt{L_{n,m}(\rho_0)} \right\|_{\ell_2}}{\left\| \sqrt{L_{n,m}(\rho_0)} L^{\mathcal{B}}_{n,m}(\rho_1) \sqrt{L_{n,m}(\rho_0)} \right\|_{\ell_2}} - 1 \right| \\
&\leq \frac{\left\| L_{n,m}(\rho_0) \right\|_{2\to 2} \left\| L^{\mathcal{B}}_{n,m}(\rho) - L^{\mathcal{B}}_{n,m}(\rho_1) \right\|_{\ell_2}}{\left\| \sqrt{L_{n,m}(\rho_0)} L^{\mathcal{B}}_{n,m}(\rho_1) \sqrt{L_{n,m}(\rho_0)} \right\|_{\ell_2}} \\
&= \frac{\left\| L_{n,m}(\rho_0) \right\|_{2\to 2} \left\| \Delta^{\mathcal{B}}(\rho, \rho_0) \right\|_{\ell_2}}{\left\| \sqrt{L_{n,m}(\rho_0)} L^{\mathcal{B}}_{n,m}(\rho_1) \sqrt{L_{n,m}(\rho_0)} \right\|_{\ell_2}} \leq \frac{\left\| L_{n,m}(\rho_0) \right\|_{2\to 2} \left\| \Delta^{\mathcal{B}}(\rho, \rho_0) \right\|_{\mathcal{S}_1}}{\left\| \sqrt{L_{n,m}(\rho_0)} L^{\mathcal{B}}_{n,m}(\rho_1) \sqrt{L_{n,m}(\rho_0)} \right\|_{\ell_2}} \\
&\overset{(a)}{\lesssim} \frac{\log n \left\| \Delta^{\mathcal{B}}(\rho, \rho_0) \right\|_{\mathcal{S}_1}}{\left\| \sqrt{L_{n,m}(\rho_0)} L^{\mathcal{B}}_{n,m}(\rho_1) \sqrt{L_{n,m}(\rho_0)} \right\|_{\ell_2}} \overset{(b)}{\lesssim} \frac{\left\| \Delta^{\mathcal{B}}(\rho, \rho_0) \right\|_{\mathcal{S}_1} \log n}{\sqrt{n}}.
\end{aligned}
$$

Here (a) and (b) are successively implied from Eq. (3.26) and Lemma 3.7. Using similar techniques as Eq. (3.39) implies that

$$
|b_n - 1| \lesssim \frac{\|\Delta(\rho, \rho_0)\|_{\mathcal{S}_1} \log n}{\sqrt{n}} \lesssim \frac{n^{\gamma} \log n}{\sqrt{n}} \to 0, \quad \text{as } n \to \infty.
$$

Namely $\lim_{n\to\infty} b_n = 1$, which concludes the proof. □

□

3.8 Technical Results

3.8.1 Large Sample Behaviour of Covariance Matrices of GPs Observed on Irregular Grids

Throughout this section, we put the following restrictions on the irregular lattice $\mathcal{D}_n$ with n points. To avoid repetition, we omit these common assumptions in the statement of all the results in this section. Moreover, the scalars implicitly expressed in $\asymp$ and $\lesssim$ relations are bounded and generally depend on m, d, ν, Θ_0 and the topological structure of $\mathcal{D}_n$.

- $\mathcal{D}_n$ is a d-dimensional grid satisfying Assumption 3.1. It is expedient to define $N := \lfloor n^{1/d} \rfloor$.
- The set of coefficients $\{a_{m,\boldsymbol{s}}(\boldsymbol{t}) : \boldsymbol{s} \in \mathcal{D}_n, \boldsymbol{t} \in \mathcal{N}_m(\boldsymbol{s})\}$, admit the conditions in Definition 3.1.

Before jumping into stating the theoretical results in the subsequent sections, we recall some key assumptions and notations that we have used in Chapter 3. G represents a cen-

tered, isotropic Matern GP whose one time realization has been observed at $\mathcal{D}_n$. The range parameters ρ belongs to a compact $\Theta_0 \subset (0,\infty)$. We also write $\{G_m(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{D}_n\}$ to denote the preconditioned process of order-m (see Definition 3.1). m is chosen in such a way that $m \geq (\nu + d/2)$. Let $\mathcal{B} = \{B_t\}_{t=1}^{b_n}$ be an arbitrary partition of $\mathcal{D}_n$. We have defined $K_{n,m}^{\mathcal{B}}(\rho)$ in Eq. (3.8), a matrix which is proportional to the block diagonal approximation of to the covariance of $[G_m(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{D}_n]$, associated to the partitioning scheme $\mathcal{B}$. We also define $L_{n,m}^{\mathcal{B}}(\rho) \coloneqq \rho^{2\nu} K_{n,m}^{\mathcal{B}}(\rho)$ for notational convenience.

3.8.1.1 The Decay of Off-diagonal Entries of $K_{n,m}^{\mathcal{B}}(\rho)$

The main objective of this section is to study the decay rate of the off-diagonal entries of $K_{n,m}^{\mathcal{B}}(\rho)$, which comes in handy for analyzing the asymptotic behavior of different norms of $K_{n,m}^{\mathcal{B}}(\rho)$ in Section 2.6.For achieving this goal, we need a spectral representation for the entries of $K_{n,m}^{\mathcal{B}}(\rho)$. For brevity define the complex valued function $f_{\boldsymbol{s}}^N : \mathbb{R}^d \setminus \{\mathbf{0}_d\} \mapsto \mathbb{C}$, for any $\boldsymbol{s} \in \mathcal{D}_n$, by

$$f_{\boldsymbol{s}}^N(\boldsymbol{\omega}) \coloneqq \|\boldsymbol{\omega}\|_{\ell_2}^{-(\nu+d/2)} \sum_{\boldsymbol{s}' \in \mathcal{N}_m(\boldsymbol{s})} a_{m,\boldsymbol{s}}(\boldsymbol{s}') \exp\big(j\langle N\boldsymbol{\omega}, \boldsymbol{s}' - \boldsymbol{s}\rangle\big), \quad \forall\, \boldsymbol{\omega} \neq \mathbf{0}_d, \tag{3.44}$$

and the strictly increasing function $h_N : (0,\infty) \mapsto (0,1)$ with

$$h_N(x) \coloneqq \Big[1 + (Nx)^{-2}\Big]^{-(\nu+d/2)}. \tag{3.45}$$

Choose $\boldsymbol{s}, \boldsymbol{t} \in \mathcal{D}_n$ arbitrarily. The entries of $K_{n,m}$ (corresponding to the single bin scenario) can be expressed in terms of the Matern spectral density.

$$\begin{aligned}
(K_{n,m}(\rho))_{\boldsymbol{s},\boldsymbol{t}} &= \frac{N^{2\nu}}{\rho^{2\nu}} \sum_{\boldsymbol{s}' \in \mathcal{N}_m(\boldsymbol{s})} \sum_{\boldsymbol{t}' \in \mathcal{N}_m(\boldsymbol{t})} a_{m,\boldsymbol{s}}(\boldsymbol{s}')\, a_{m,\boldsymbol{t}}(\boldsymbol{t}') \int_{\mathbb{R}^d} e^{j\langle \boldsymbol{\omega}, \boldsymbol{t}' - \boldsymbol{s}'\rangle} \left(\|\boldsymbol{\omega}\|_{\ell_2}^2 + \frac{1}{\rho^2}\right)^{-(\nu+d/2)} d\boldsymbol{\omega} \\
&= \frac{N^{2\nu}}{\rho^{2\nu}} \sum_{\boldsymbol{s}' \in \mathcal{N}_m(\boldsymbol{s})} \sum_{\boldsymbol{t}' \in \mathcal{N}_m(\boldsymbol{t})} a_{m,\boldsymbol{s}}(\boldsymbol{s}')\, a_{m,\boldsymbol{t}}(\boldsymbol{t}') \int_{\mathbb{R}^d} \frac{\exp\big(j\langle \boldsymbol{\omega}, \boldsymbol{t}' - \boldsymbol{s}'\rangle\big)}{\|\boldsymbol{\omega}\|_{\ell_2}^{2\nu+d}} h_N\left(\frac{\rho\|\boldsymbol{\omega}\|_{\ell_2}}{N}\right) d\boldsymbol{\omega}.
\end{aligned}$$

Change of variable method introduces an equivalent form of the above identity (replace $N\boldsymbol{\omega}$ instead of $\boldsymbol{\omega}$).

$$\begin{aligned}
(K_{n,m}(\rho))_{\boldsymbol{s},\boldsymbol{t}} &= \frac{1}{\rho^{2\nu}} \sum_{\boldsymbol{s}' \in \mathcal{N}_m(\boldsymbol{s})} \sum_{\boldsymbol{t}' \in \mathcal{N}_m(\boldsymbol{t})} a_{m,\boldsymbol{s}}(\boldsymbol{s}')\, a_{m,\boldsymbol{t}}(\boldsymbol{t}') \int_{\mathbb{R}^d} \frac{\exp\big(j\langle N\boldsymbol{\omega}, \boldsymbol{t}' - \boldsymbol{s}'\rangle\big)}{\|\boldsymbol{\omega}\|_{\ell_2}^{2\nu+d}} h_N\Big(\rho\|\boldsymbol{\omega}\|_{\ell_2}\Big) d\boldsymbol{\omega} \\
&= \rho^{-2\nu} \int_{\mathbb{R}^d} \exp\big(j\langle \boldsymbol{t} - \boldsymbol{s}, \boldsymbol{\omega}\rangle\big)\, f_{\boldsymbol{s}}^N(\boldsymbol{\omega})\, \overline{f_{\boldsymbol{t}}^N(\boldsymbol{\omega})} h_N\Big(\rho\|\boldsymbol{\omega}\|_{\ell_2}\Big) d\boldsymbol{\omega}.
\end{aligned} \tag{3.46}$$

Next we examine the behavior of $f_{\boldsymbol{s}}^N(\cdot)$ for large $\boldsymbol{\omega}$. Such analysis is decisive for controlling the entries of $K_{n,m}^{\mathcal{B}}(\rho)$ from above.

Lemma 3.1. There exists $\beta \in (1,\infty)$ (depending on m, ν, d and $\mathcal{D}_n$) such that

$$\max_{\boldsymbol{s}\in\mathcal{D}_n} \left|f_{\boldsymbol{s}}^N(\boldsymbol{\omega})\right|^2 \leq \frac{\beta}{1+\|\boldsymbol{\omega}\|_{\ell_2}^{2\nu+d}}, \quad \forall\, \boldsymbol{\omega} \neq \mathbf{0}_d. \tag{3.47}$$

Proof. Define the bounded integer g_m by $g_m \coloneqq \max_{\boldsymbol{s}\in\mathcal{D}_n} |\mathcal{N}_m(\boldsymbol{s})|$. Choose an arbitrary $\boldsymbol{s} \in \mathcal{D}_n$. $f_{\boldsymbol{s}}^N$ is trivially continuous and well defined at any $\boldsymbol{\omega} \neq \mathbf{0}_d$, so is the function $\max_{\boldsymbol{s}\in\mathcal{D}_n} \left|f_{\boldsymbol{s}}^N\right|^2$ (due to the continuity of the max operator). Thus for validating Eq. (3.47), we only require to show that

1. $\max_{\boldsymbol{s}\in\mathcal{D}_n} \left|f_{\boldsymbol{s}}^N(\boldsymbol{\omega})\right|^2 \lesssim \left(1+\|\boldsymbol{\omega}\|_{\ell_2}^{2\nu+d}\right)^{-1}$, for any $\boldsymbol{\omega}$ with $\|\boldsymbol{\omega}\|_{\ell_2}^{2\nu+d} \geq g_m$.
2. There exists a bounded constant π_m such that $\max_{\boldsymbol{s}\in\mathcal{D}_n} \limsup_{\boldsymbol{\omega}\to\mathbf{0}_d} \left|f_{\boldsymbol{s}}^N(\boldsymbol{\omega})\right|^2 \leq \pi_m$.

The first claim is an implication of the Cauchy-Schwartz inequality. In Definition 3.1, we normalized the coefficients $a_{m,\boldsymbol{s}}(\boldsymbol{s}')$'s to have unit Euclidean norm. Thus

$$\begin{aligned}
\left|f_{\boldsymbol{s}}^N(\boldsymbol{\omega})\right|^2 &\leq \|\boldsymbol{\omega}\|_{\ell_2}^{-(2\nu+d)} |\mathcal{N}_m(\boldsymbol{s})| \sum_{\boldsymbol{s}'\in\mathcal{N}_m(\boldsymbol{s})} a_{m,\boldsymbol{s}}^2(\boldsymbol{s}') = \|\boldsymbol{\omega}\|_{\ell_2}^{-(2\nu+d)} |\mathcal{N}_m(\boldsymbol{s})| \\
&\leq g_m \|\boldsymbol{\omega}\|_{\ell_2}^{-(2\nu+d)} \leq \frac{1+g_m}{1+\|\boldsymbol{\omega}\|_{\ell_2}^{2\nu+d}}.
\end{aligned}$$

For proving the other claim we need to study the Taylor expansion of $f_{\boldsymbol{s}}^N$ near the origin. The second condition of Definition 3.1 implies that for any natural number $r < m$,

$$\sum_{\boldsymbol{s}'\in\mathcal{N}_m(\boldsymbol{s})} a_{m,\boldsymbol{s}}(\boldsymbol{s}') \left(\langle \boldsymbol{\omega}, \boldsymbol{s}'-\boldsymbol{s}\rangle\right)^r = 0, \quad \forall\, \boldsymbol{\omega} \in \mathbb{R}^d,\ \forall\, \boldsymbol{s} \in \mathcal{D}_n.$$

So

$$\begin{aligned}
\limsup_{\boldsymbol{\omega}\to\mathbf{0}_d} \left|f_{\boldsymbol{s}}^N(\boldsymbol{\omega})\right|^2 &= \lim_{\boldsymbol{\omega}\to\mathbf{0}_d} \frac{1}{\|\boldsymbol{\omega}\|_{\ell_2}^{2\nu+d}} \left| \sum_{r=0}^{\infty} \frac{(jN)^r}{r!} \sum_{\boldsymbol{s}'\in\mathcal{N}_m(\boldsymbol{s})} a_{m,\boldsymbol{s}}(\boldsymbol{s}') \left(\langle \boldsymbol{\omega}, \boldsymbol{s}'-\boldsymbol{s}\rangle\right)^r \right|^2 \\
&= \limsup_{\boldsymbol{\omega}\to\mathbf{0}_d} \frac{1}{\|\boldsymbol{\omega}\|_{\ell_2}^{2\nu+d}} \left| \sum_{r=m}^{\infty} \frac{(jN)^r}{r!} \sum_{\boldsymbol{s}'\in\mathcal{N}_m(\boldsymbol{s})} a_{m,\boldsymbol{s}}(\boldsymbol{s}') \left(\langle \boldsymbol{\omega}, \boldsymbol{s}'-\boldsymbol{s}\rangle\right)^r \right|^2 \\
&= \frac{N^{2m}}{m!} \limsup_{\boldsymbol{\omega}\to\mathbf{0}_d} \frac{1}{\|\boldsymbol{\omega}\|_{\ell_2}^{2\nu+d}} \left| \sum_{\boldsymbol{s}'\in\mathcal{N}_m(\boldsymbol{s})} a_{m,\boldsymbol{s}}(\boldsymbol{s}') \left(\langle \boldsymbol{\omega}, \boldsymbol{s}'-\boldsymbol{s}\rangle\right)^m \right|^2. \qquad (3.48)
\end{aligned}$$

Cauchy-Schwartz inequality helps to further simplify the complex expressions in Eq. (3.48).

$$\begin{aligned}\limsup_{\boldsymbol{\omega}\to\mathbf{0}_d}\left|f_{\boldsymbol{s}}^N(\boldsymbol{\omega})\right|^2 &\leq \limsup_{\boldsymbol{\omega}\to\mathbf{0}_d}\frac{N^{2m}\|\boldsymbol{\omega}\|_{\ell_2}^{2m-2\nu-d}}{m!}\sum_{\boldsymbol{s}'\in\mathcal{N}_m(\boldsymbol{s})}a_{m,\boldsymbol{s}}^2(\boldsymbol{s}')\sum_{\boldsymbol{s}'\in\mathcal{N}_m(\boldsymbol{s})}\left\|\boldsymbol{s}'-\boldsymbol{s}\right\|_{\ell_2}^{2m}\\ &= \frac{\sum_{\boldsymbol{s}'\in\mathcal{N}_m(\boldsymbol{s})}\|N(\boldsymbol{s}'-\boldsymbol{s})\|_{\ell_2}^{2m}}{m!}\mathbb{1}_{\{2m=2\nu+d\}}.\end{aligned}$$

We can argue based on the first condition in Definition 3.1 that

$$\exists\,\pi_m\in(0,\infty)\ \text{ s.t. }\ \max_{s\in\mathcal{D}_n}\left(\frac{\sum_{\boldsymbol{s}'\in\mathcal{N}_m(\boldsymbol{s})}\|N(\boldsymbol{s}'-\boldsymbol{s})\|_{\ell_2}^{2m}}{m!}\right)\leq\pi_m.$$

Hence,

$$\limsup_{\boldsymbol{\omega}\to\mathbf{0}_d}\left|f_{\boldsymbol{s}}^N(\boldsymbol{\omega})\right|^2\leq Q_m\mathbb{1}_{\{2m=2\nu+d\}}\leq Q_m.$$

It is straightforward to find a closed form expression for β in terms of g_m and π_m. □

Proposition 3.1. For any pair $\boldsymbol{s},\boldsymbol{t}\in\mathcal{D}_n$ and any partition $\mathcal{B}$ of $\mathcal{D}_n$,

$$\left|\left(K_{n,m}^{\mathcal{B}}(\rho)\right)_{\boldsymbol{s},\boldsymbol{t}}\right|\lesssim\rho^{-2\nu}\left(1+N\|\boldsymbol{t}-\boldsymbol{s}\|_{\ell_2}\right)^{-2(m-\nu)}. \tag{3.49}$$

Proof. Without loss of generality we cane assume that $\mathcal{B}$ has only a single bin, i.e. $\mathcal{B}=\{\mathcal{D}_n\}$. In other words, we just need to validate Eq. (3.49) for the entries of $K_{n,m}(\rho)$. For simplicity, let $f_{\nu,\rho}$ denotes the Matern correlation function with parameters (ρ,ν). Notice that $f_{\nu,\rho}(x)=f_{\nu,1}(x/\rho)$. We first prove the inequality (3.49) for the case of $\|\boldsymbol{t}-\boldsymbol{s}\|_{\ell_2}=O\left(N^{-1}\right)$. It suffices to show that the largest diagonal entry of $K_{n,m}(\rho)$ is of order $\rho^{-2\nu}$. That is,

$$\rho^{2\nu}\max_{\boldsymbol{s}\in\mathcal{D}_n}\left|\left(K_{n,m}(\rho)\right)_{\boldsymbol{s},\boldsymbol{s}}\right|\lesssim 1.$$

The proof of this result hinges on the inequality (3.46) for $\boldsymbol{s}=\boldsymbol{t}$. Trivially,

$$\rho^{2\nu}\max_{\boldsymbol{s}\in\mathcal{D}_n}\left|\left(K_{n,m}(\rho)\right)_{\boldsymbol{s},\boldsymbol{s}}\right|=\max_{\boldsymbol{s}\in\mathcal{D}_n}\int_{\mathbb{R}^d}\left|f_{\boldsymbol{s}}^N(\boldsymbol{\omega})\right|^2h_N\left(\rho\|\boldsymbol{\omega}\|_{\ell_2}\right)d\boldsymbol{\omega}\leq\max_{\boldsymbol{s}\in\mathcal{D}_n}\int_{\mathbb{R}^d}\left|f_{\boldsymbol{s}}^N(\boldsymbol{\omega})\right|^2d\boldsymbol{\omega}.$$

We finish the proof of this part by using Lemma 3.1.

$$\rho^{2\nu}\max_{\boldsymbol{s}\in\mathcal{D}_n}\left|\left(K_{n,m}(\rho)\right)_{\boldsymbol{s},\boldsymbol{s}}\right|\leq\max_{\boldsymbol{s}\in\mathcal{D}_n}\int_{\mathbb{R}^d}\left|f_{\boldsymbol{s}}^N(\boldsymbol{\omega})\right|^2d\boldsymbol{\omega}\lesssim\int_{\mathbb{R}^d}\frac{d\boldsymbol{\omega}}{1+\|\boldsymbol{\omega}\|_{\ell_2}^{2\nu+d}}\asymp\int_0^\infty\frac{x^{d-1}}{1+x^{2\nu+d}}dx\asymp 1.$$

So without loss of generality we can assume that $\|\boldsymbol{t}-\boldsymbol{s}\|_{\ell_2}>h/N$, for some large enough h.

Trivially,

$$\psi := \frac{(K_{n,m}(\rho))_{\boldsymbol{s},\boldsymbol{t}}}{N^{2\nu}} = \sum_{\boldsymbol{s}' \in \mathcal{N}_m(\boldsymbol{s})} \sum_{\boldsymbol{t}' \in \mathcal{N}_m(\boldsymbol{t})} a_{m,\boldsymbol{s}}(\boldsymbol{s}')\, a_{m,\boldsymbol{t}}(\boldsymbol{t}')\, f_{\nu,\rho}(\boldsymbol{t}' - \boldsymbol{s}').$$

The key step of the proof is to replace $f_{\nu,\rho}(\cdot)$ with its exact Taylor expansion of order $2m$. Strictly speaking, we have

$$\begin{aligned} f_{\nu,\rho}(\boldsymbol{t}' - \boldsymbol{s}') &= \sum_{|r|<2m} \frac{D^r f_{\nu,\rho}(\boldsymbol{t} - \boldsymbol{s})}{r!} \left[(\boldsymbol{t}' - \boldsymbol{t}) - (\boldsymbol{s}' - \boldsymbol{s})\right]^r \\ &+ \sum_{|r|=2m} R_r(\boldsymbol{t} - \boldsymbol{s}) \frac{\left[(\boldsymbol{t}' - \boldsymbol{t}) - (\boldsymbol{s}' - \boldsymbol{s})\right]^r}{r!}, \end{aligned}$$

in which R_r denotes the residual function given by

$$R_r(\boldsymbol{t} - \boldsymbol{s}) = 2m \int_0^1 (1-x)^{2m-1} D^r f_{\nu,\rho}\Big((\boldsymbol{t} - \boldsymbol{s}) + x\left[(\boldsymbol{t}' - \boldsymbol{t}) - (\boldsymbol{s}' - \boldsymbol{s})\right] \Big) dx. \tag{3.50}$$

Thus,

$$\begin{aligned} \psi &= \sum_{|r|<2m} \frac{D^r f_{\nu,\rho}(\boldsymbol{t} - \boldsymbol{s})}{r!} \sum_{\boldsymbol{s}' \in \mathcal{N}_m(\boldsymbol{s})} \sum_{\boldsymbol{t}' \in \mathcal{N}_m(\boldsymbol{t})} a_{m,\boldsymbol{s}}(\boldsymbol{s}')\, a_{m,\boldsymbol{t}}(\boldsymbol{t}') \left[(\boldsymbol{t}' - \boldsymbol{t}) - (\boldsymbol{s}' - \boldsymbol{s})\right]^r \\ &+ \sum_{|r|=2m} \frac{R_r(\boldsymbol{t} - \boldsymbol{s})}{r!} \sum_{\boldsymbol{s}' \in \mathcal{N}_m(\boldsymbol{s})} \sum_{\boldsymbol{t}' \in \mathcal{N}_m(\boldsymbol{t})} a_{m,\boldsymbol{s}}(\boldsymbol{s}')\, a_{m,\boldsymbol{t}}(\boldsymbol{t}') \left[(\boldsymbol{t}' - \boldsymbol{t}) - (\boldsymbol{s}' - \boldsymbol{s})\right]^r. \end{aligned} \tag{3.51}$$

The first constraint on $\{a_{m,\boldsymbol{s}}(\boldsymbol{t}) : \boldsymbol{s} \in \mathcal{D}_n, \boldsymbol{t} \in \mathcal{N}_m(\boldsymbol{s})\}$ in Definition 3.1 easily implies that

$$\sum_{\boldsymbol{s}' \in \mathcal{N}_m(\boldsymbol{s})} \sum_{\boldsymbol{t}' \in \mathcal{N}_m(\boldsymbol{t})} a_{m,\boldsymbol{s}}(\boldsymbol{s}')\, a_{m,\boldsymbol{t}}(\boldsymbol{t}') \left[(\boldsymbol{t}' - \boldsymbol{t}) - (\boldsymbol{s}' - \boldsymbol{s})\right]^r = 0.$$

for any $|r| < 2m$. So the first term in the right hand side of (3.51) vanishes. Henceforth, we only need control the second term from above. Observe that

$$|\psi| \le \sum_{|r|=2m} \left| \sum_{\boldsymbol{s}' \in \mathcal{N}_m(\boldsymbol{s})} \sum_{\boldsymbol{t}' \in \mathcal{N}_m(\boldsymbol{t})} a_{m,\boldsymbol{s}}(\boldsymbol{s}')\, a_{m,\boldsymbol{t}}(\boldsymbol{t}') \left[(\boldsymbol{t}' - \boldsymbol{t}) - (\boldsymbol{s}' - \boldsymbol{s})\right]^r \right| \max_{|r|=2m} \left| \frac{R_r(\boldsymbol{t} - \boldsymbol{s})}{r!} \right|. \tag{3.52}$$

The next step is to introduce a uniform upper bound on the residual functions using Eq.

(3.50) and the chain rule of derivative.

$$\begin{aligned}\max_{|r|=2m}|R_r(\boldsymbol{t}-\boldsymbol{s})| &\leq \max_{|r|=2m}\max_{x\in[0,1]}\left|D^r f_{\nu,\rho}\Big\{(\boldsymbol{t}-\boldsymbol{s})+x\big[(\boldsymbol{t}'-\boldsymbol{t})-(\boldsymbol{s}'-\boldsymbol{s})\big]\Big\}\right| \\ &\leq \rho^{-2m}\max_{|r|=2m}\max_{x\in[0,1]}\left|D^r f_{\nu,1}\Big\{\frac{(\boldsymbol{t}-\boldsymbol{s})+x\big[(\boldsymbol{t}'-\boldsymbol{t})-(\boldsymbol{s}'-\boldsymbol{s})\big]}{\rho}\Big\}\right|. \quad (3.53)\end{aligned}$$

As the maximum distance between $\boldsymbol{s}$ and the points $\boldsymbol{s}' \in \mathcal{N}_m(\boldsymbol{s})$ is of order $1/N$, so we can choose h large enough such that

$$\min_{x\in[0,1]}\Big\|(\boldsymbol{t}-\boldsymbol{s})+x\big[(\boldsymbol{t}'-\boldsymbol{t})\big]\Big\|_{\ell_2} \geq \frac{\|\boldsymbol{t}-\boldsymbol{s}\|_{\ell_2}}{2}. \quad (3.54)$$

Now we apply Lemma 4 of [And10] to get an upper bound on $D^r f_{\nu,1}(\cdot)$ in terms of the Euclidean norm of its argument. So for any $x \in [0,1]$, we have

$$\left|D^r f_{\nu,1}\Big\{\frac{(\boldsymbol{t}-\boldsymbol{s})+x\big[(\boldsymbol{t}'-\boldsymbol{t})-(\boldsymbol{s}'-\boldsymbol{s})\big]}{\rho}\Big\}\right| \lesssim \left\|\frac{(\boldsymbol{t}-\boldsymbol{s})+x\big[(\boldsymbol{t}'-\boldsymbol{t})-(\boldsymbol{s}'-\boldsymbol{s})\big]}{\rho}\right\|_{\ell_2}^{2(\nu-m)}.$$

Combining this inequality and Eq. (3.54) shows that for any pair $(\boldsymbol{s},\boldsymbol{t})$ with $\|\boldsymbol{t}-\boldsymbol{s}\|_{\ell_2} \geq h/N$

$$\max_{|r|=2m}|R_r(\boldsymbol{t}-\boldsymbol{s})| \lesssim \rho^{-2m}\left(\frac{\|\boldsymbol{t}-\boldsymbol{s}\|_{\ell_2}}{\rho}\right)^{2(\nu-m)} \lesssim \rho^{-2\nu}\left(\frac{1}{N}+\|\boldsymbol{t}-\boldsymbol{s}\|_{\ell_2}\right)^{2(\nu-m)}. \quad (3.55)$$

Substituting (3.55) into (3.52) yields (in which $\hat{C}^{\rho,\nu}_{m,d}$ is another bounded scalar)

$$|\psi| \lesssim \rho^{-2\nu}\left(\frac{1}{N}+\|\boldsymbol{t}-\boldsymbol{s}\|_{\ell_2}\right)^{2(\nu-m)}\sum_{|r|=2m}\underbrace{\left|\sum_{\boldsymbol{s}'\in\mathcal{N}_m(\boldsymbol{s})}\sum_{\boldsymbol{t}'\in\mathcal{N}_m(\boldsymbol{t})}a_{m,\boldsymbol{s}}(\boldsymbol{s}')\,a_{m,\boldsymbol{t}}(\boldsymbol{t}')\big[(\boldsymbol{t}'-\boldsymbol{t})-(\boldsymbol{s}'-\boldsymbol{s})\big]^r\right|}_{\varpi_r}.$$

In the sequel, we prove that $\varpi_r = \mathcal{O}\left(N^{-2m}\right)$ for any $|r| = 2m$ using the following series of inequalities.

$$\begin{aligned}\varpi_r &\overset{(a)}{\leq} \left(\sum_{\boldsymbol{s}'\in\mathcal{N}_m(\boldsymbol{s})}a^2_{m,\boldsymbol{s}}(\boldsymbol{s}')\right)^{1/2}\left(\sum_{\boldsymbol{t}'\in\mathcal{N}_m(\boldsymbol{t})}a^2_{m,\boldsymbol{t}}(\boldsymbol{t}')\right)^{1/2}\max\left\{\Big|(\boldsymbol{t}'-\boldsymbol{t})-(\boldsymbol{s}'-\boldsymbol{s})\Big|^r:\ \begin{matrix}\boldsymbol{s}'\in\mathcal{N}_m(\boldsymbol{s})\\ \boldsymbol{t}'\in\mathcal{N}_m(\boldsymbol{t})\end{matrix}\right\} \\ &\overset{(b)}{=} \max\left\{\Big|(\boldsymbol{t}'-\boldsymbol{t})-(\boldsymbol{s}'-\boldsymbol{s})\Big|^r:\ \begin{matrix}\boldsymbol{s}'\in\mathcal{N}_m(\boldsymbol{s})\\ \boldsymbol{t}'\in\mathcal{N}_m(\boldsymbol{t})\end{matrix}\right\} \overset{(c)}{=} \mathcal{O}\left(N^{-2m}\right).\end{aligned}$$

Here, (a) is an obvious implication of the Holder inequality. The identity (b) is exactly same as the third condition in Definition 3.1 and (c) holds for the class of non-regular

lattices satisfying Assumption 3.1. Hence

$$\begin{aligned}\left|\left(K_{n,m}(\rho)\right)_{\boldsymbol{s},\boldsymbol{t}}\right| &= N^{2\nu}|\psi| \lesssim \left(\frac{N}{\rho}\right)^{2\nu}\left(\frac{1}{N}+\|\boldsymbol{t}-\boldsymbol{s}\|_{\ell_2}\right)^{2(\nu-m)}\sum_{|r|=2m} N^{-2m} \\ &\lesssim \rho^{-2\nu}\left(1+N\|\boldsymbol{t}-\boldsymbol{s}\|_{\ell_2}\right)^{-2(m-\nu)}\end{aligned}$$

□

3.8.2 Sensitivity of $L^{\mathcal{B}}_{n,m}(\rho)$ with Respect to ρ

Recall that we defined $L^{\mathcal{B}}_{n,m}(\rho)$ as the block diagonal approximation of $L_{n,m}(\rho) = \rho^{2\nu}K_{n,m}(\rho)$, corresponding to the partitioning scheme $\mathcal{B} = \{B_t\}_{t=1}^{b_n}$ of $\mathcal{D}_n$. This section is dedicated to study the sensitivity of $L^{\mathcal{B}}_{n,m}(\rho)$ with respect to ρ, for large n. In other words, we are interested to study the quantity

$$\frac{\left\|L^{\mathcal{B}}_{n,m}(\rho_2)-L^{\mathcal{B}}_{n,m}(\rho_1)\right\|}{|\rho_2-\rho_1|}, \quad \rho_1,\rho_2\in\Theta_0,$$

as n tends to infinity. Here $\|\cdot\|$ represents either nuclear, Frobenius or operator norm. The presented results are decisive in Section 2.6. The quantity Q_N, which will be defined in the next lemma, appears numerous times in this section.

Lemma 3.2. Let ρ_1,ρ_2 be distinct points in Θ_0 such that $\rho_2>\rho_1$. Define

$$Q_N \coloneqq \int_{\mathbb{R}^d}\left|f^N_{\boldsymbol{s}}(\boldsymbol{\omega})\,f^N_{\boldsymbol{t}}(\boldsymbol{\omega})\right|\left|h_N\left(\rho_2\|\boldsymbol{\omega}\|_{\ell_2}\right)-h_N\left(\rho_1\|\boldsymbol{\omega}\|_{\ell_2}\right)\right|d\boldsymbol{\omega}$$

Choose an arbitrary pairs of $\boldsymbol{s},\boldsymbol{t}\in\mathcal{D}_n$.

$$\frac{Q_N}{\rho_2-\rho_1}\lesssim\frac{\left(\mathbb{1}_{\{d\geq 3\}}+\mathbb{1}_{\{d=2\}}\log N+\mathbb{1}_{\{d=1\}}N\right)}{N^2}.$$

Proof. Lemma 3.1 provides an upper bound on the term $f^N_{\boldsymbol{s}}(\boldsymbol{\omega})\,f^N_{\boldsymbol{t}}(\boldsymbol{\omega})$.

$$\left|f^N_{\boldsymbol{s}}(\boldsymbol{\omega})\,f^N_{\boldsymbol{t}}(\boldsymbol{\omega})\right|\lesssim\left(1+\|\boldsymbol{\omega}\|^{2\nu+d}_{\ell_2}\right)^{-1}. \tag{3.56}$$

For controlling the other term of the integrand from above, we employ the following inequality, which will be justified later.

$$(1+x)^{-\alpha}-(1+y)^{-\alpha}<[\alpha(y-x)]\wedge\left(x^{-\alpha}-y^{-\alpha}\right),\quad\forall\ 0<x<y<\infty,\ \alpha>0. \tag{3.57}$$

Using (3.57) (with $\alpha = \nu + \frac{d}{2}$) yields

$$\left|\frac{h_N\left(\rho_2\|\omega\|_{\ell_2}\right)-h_N\left(\rho_1\|\omega\|_{\ell_2}\right)}{\rho_2-\rho_1}\right| \leq \left[\left(N\|\omega\|_{\ell_2}\right)^{2\nu+d}\left(\frac{\rho_2^{2\nu+d}-\rho_1^{2\nu+d}}{\rho_2-\rho_1}\right)\right] \wedge \left[\frac{(\nu+d/2)\left(1/\rho_1^2-1/\rho_2^2\right)}{\left(N\|\omega\|_{\ell_2}\right)^2(\rho_2-\rho_1)}\right].$$

The fact that Θ_0 is compact and does not contain zero simplify the last inequality as the following.

$$\left|\frac{h_N\left(\rho_2\|\omega\|_{\ell_2}\right)-h_N\left(\rho_1\|\omega\|_{\ell_2}\right)}{\rho_2-\rho_1}\right| \lesssim \left[\left(N\|\omega\|_{\ell_2}\right)^{2\nu+d} \wedge \left(N\|\omega\|_{\ell_2}\right)^{-2}\right]. \tag{3.58}$$

Combining (3.56) and (3.58) leads to

$$\begin{aligned}
\frac{Q_N}{(\rho_2-\rho_1)} &\lesssim \int_{\mathbb{R}^d}\left[\left(N\|\omega\|_{\ell_2}\right)^{2\nu+d} \wedge \left(N\|\omega\|_{\ell_2}\right)^{-2}\right]\frac{d\omega}{1+\|\omega\|_{\ell_2}^{2\nu+d}} \\
&\overset{(b)}{\asymp} \int_0^\infty \left[(Nu)^{2\nu+d} \wedge (Nu)^{-2}\right]\frac{u^{d-1}du}{1+u^{2\nu+d}} \\
&= N^{2\nu+d}\int_0^{\frac{1}{N}}\frac{u^{2\nu+2d-1}}{1+u^{2\nu+d}}du + \frac{1}{N^2}\int_{\frac{1}{N}}^\infty \frac{u^{d-3}}{1+u^{2\nu+d}}du
\end{aligned} \tag{3.59}$$

The change of variable $u = \|\omega\|_{\ell_2}$ in the integral validates $\overset{(b)}{\asymp}$. For brevity, let ψ_1 and ψ_2 stand for the two expressions in the last line of (3.59), respectively from left to right. We ultimately introduce tight upper bounds on ψ_1 and ψ_2. Observe that

$$\psi_1 = N^{2\nu+d}\int_0^{\frac{1}{N}}\frac{u^{2\nu+2d-1}}{1+u^{2\nu+d}}du \leq N^{2\nu+d}\int_0^{\frac{1}{N}} u^{2\nu+2d-1}du \asymp N^{2\nu+d}N^{-2(\nu+d)} = N^{-d}.$$

Furthermore,

$$\begin{aligned}
\psi_2 &= \frac{1}{N^2}\int_{\frac{1}{N}}^\infty \frac{u^{d-3}}{1+u^{2\nu+d}}du = \frac{1}{N^2}\left[\int_{\frac{1}{N}}^1 \frac{u^{d-3}}{1+u^{2\nu+d}}du + \int_1^\infty \frac{u^{d-3}}{1+u^{2\nu+d}}du\right] \\
&\leq \frac{1}{N^2}\left[\int_{\frac{1}{N}}^1 u^{d-3}du + \int_1^\infty u^{-(2\nu+3)}du\right] \lesssim \frac{1}{N^2}\left(\int_{\frac{1}{N}}^1 u^{d-3}du + 1\right) \\
&\asymp \frac{\left(1+\mathbb{1}_{\{d=2\}}\log N + \mathbb{1}_{\{d=1\}}N\right)}{N^2}.
\end{aligned}$$

Replacing the upper bounds on ψ_1 and ψ_2 into (3.59) yields

$$\frac{Q_N}{(\rho_2-\rho_1)} \lesssim \frac{\left(1+\mathbb{1}_{\{d=2\}}\log N+\mathbb{1}_{\{d=1\}}N\right)}{N^2}+N^{-d} \asymp \frac{\left(1+\mathbb{1}_{\{d=2\}}\log N+\mathbb{1}_{\{d=1\}}N\right)}{N^2}$$

In the sequel, we prove Eq. (3.57). Choose an arbitrary $\alpha>0$ and define $g_1, g_2 : (0,\infty) \mapsto \mathbb{R}$ by

$$g_1(u) = \alpha u - (1+u)^{-\alpha}, \quad g_2(u) = u^{-\alpha} - (1+u)^{-\alpha}.$$

Notice that (3.57) is equivalent to the two inequalities $g_1(x) < g_1(y)$ and $g_2(y) < g_2(x)$. Namely, we need to show that both g_1 and $-g_2$ are strictly increasing function. For any $u \in (0,\infty)$, we have

$$g_1'(u) = \alpha\left(1-(1+u)^{-(\alpha+1)}\right) > 0, \quad g_2'(u) = -\alpha\left(u^{-(\alpha+1)} - (1+u)^{-(\alpha+1)}\right) < 0,$$

which concludes the proof. □

For notational convenience and from now on define, $\Delta^{\mathcal{B}}(\rho_1,\rho_2) \coloneqq L^{\mathcal{B}}_{n,m}(\rho_2) - L^{\mathcal{B}}_{n,m}(\rho_1)$, for any $\rho_1,\rho_2 \in \Theta_0$. When we deal with a single bin (no partitioning), Δ and L respectively refer to $\Delta^{\mathcal{B}}$ and $L^{\mathcal{B}}$.

Lemma 3.3. Choose $\rho_1,\rho_2 \in \Theta_0$ such that $\rho_2 \neq \rho_1$. Then

$$\frac{\left\|\Delta^{\mathcal{B}}(\rho_1,\rho_2)\right\|_{\mathcal{S}_1}}{|\rho_2-\rho_1|} \lesssim \left(\mathbb{1}_{\{d=1\}} + \mathbb{1}_{\{d=2\}}\log N + \mathbb{1}_{\{d\geq 3\}}N^{d-2}\right). \tag{3.60}$$

Furthermore for any $d \geq 3$, $\left\|\Delta^{\mathcal{B}}(\rho_1,\rho_2)\right\|_{\mathcal{S}_1} \asymp N^{d-2}|\rho_2-\rho_1|$.

Proof. Without loss of generality assume that $\rho_2 > \rho_1$. We claim that $\Delta^{\mathcal{B}}(\rho_1,\rho_2)$ is a positive semi-definite matrix. If such property holds then $\mathcal{S}_1$ norm and trace are the same. Namely the absolute sum of eigenvalues can be expressed only in terms of the diagonal entries. To see this is so begin by obtaining the spectral representation for the entries of $\Delta^{\mathcal{B}}$. Recall $f^N_{\boldsymbol{s}}(\cdot)$ and $h_N(\cdot)$ from Eq. (3.44) and (3.45), respectively. Now choose an arbitrary

unit norm vector $v \in \mathbb{R}^n$ ($n = |D_n|$). Observe that

$$\begin{aligned}
v^\top \Delta^{\mathcal{B}}(\rho_1,\rho_2) v &= \sum_{\boldsymbol{s},\boldsymbol{t}\in\mathcal{D}_n} v_{\boldsymbol{s}} v_{\boldsymbol{t}} \Big(\Delta^{\mathcal{B}}(\rho_1,\rho_2)\Big)_{\boldsymbol{s},\boldsymbol{t}} \\
&= \sum_{\boldsymbol{s},\boldsymbol{t}\in\mathcal{D}_n} v_{\boldsymbol{s}} v_{\boldsymbol{t}} \Big[\rho_2^{2\nu}\Big(K^{\mathcal{B}}_{n,m}(\rho_2)\Big)_{\boldsymbol{s},\boldsymbol{t}} - \rho_1^{2\nu}\Big(K^{\mathcal{B}}_{n,m}(\rho_1)\Big)_{\boldsymbol{s},\boldsymbol{t}}\Big] \\
&\overset{(a)}{=} \sum_{t=1}^{b_n} \sum_{\boldsymbol{s}\in B_t} v_{\boldsymbol{s}} v_{\boldsymbol{t}} \Big[\rho_2^{2\nu}\big(K_{n,m}(\rho_2)\big)_{\boldsymbol{s},\boldsymbol{t}} - \rho_1^{2\nu}\big(K_{n,m}(\rho_1)\big)_{\boldsymbol{s},\boldsymbol{t}}\Big] \\
&\overset{(b)}{=} \sum_{t=1}^{b_n} \int_{\mathbb{R}^d} \sum_{\boldsymbol{s},\boldsymbol{t}\in B_t} v_{\boldsymbol{s}} v_{\boldsymbol{t}} e^{j\langle \boldsymbol{t}-\boldsymbol{s}, N\boldsymbol{\omega}\rangle} f_{\boldsymbol{s}}^N(\boldsymbol{\omega}) \overline{f_{\boldsymbol{t}}^N(\boldsymbol{\omega})} \Big[h_N\Big(\rho_2 \|\boldsymbol{\omega}\|_{\ell_2}\Big) - h_N\Big(\rho_1 \|\boldsymbol{\omega}\|_{\ell_2}\Big)\Big] d\boldsymbol{\omega} \\
&= \sum_{t=1}^{b_n} \int_{\mathbb{R}^d} \left| \sum_{\boldsymbol{s}\in B_t} v_{\boldsymbol{s}} e^{j\langle \boldsymbol{s}, N\boldsymbol{\omega}\rangle} f_{\boldsymbol{s}}^N(\boldsymbol{\omega}) \right|^2 \Big[h_N\Big(\rho_2 \|\boldsymbol{\omega}\|_{\ell_2}\Big) - h_N\Big(\rho_1 \|\boldsymbol{\omega}\|_{\ell_2}\Big)\Big] d\boldsymbol{\omega} \\
&\overset{(c)}{>} 0. \qquad (3.61)
\end{aligned}$$

in which (a) follows from the fact that $(K^{\mathcal{B}}_{n,m}(\rho_2))_{\boldsymbol{s},\boldsymbol{t}} = 0$ when $\boldsymbol{s}$ and $\boldsymbol{t}$ belong to distinct bins. The identity (b) is a simple application of Eq. (3.46). Furthermore, inequality (c) follows from the monotonicity of h_N. Now obviously we have

$$|\mathcal{D}_n| \min_{\boldsymbol{s}\in\mathcal{D}_n} \Big|\Big(\Delta^{\mathcal{B}}(\rho_1,\rho_2)\Big)_{\boldsymbol{s},\boldsymbol{s}}\Big| \le \Big\|\Delta^{\mathcal{B}}(\rho_1,\rho_2)\Big\|_{\mathcal{S}_1} = tr\Big(\Delta^{\mathcal{B}}(\rho_1,\rho_2)\Big) \le |\mathcal{D}_n| \max_{\boldsymbol{s}\in\mathcal{D}_n} \Big|\Big(\Delta^{\mathcal{B}}(\rho_1,\rho_2)\Big)_{\boldsymbol{s},\boldsymbol{s}}\Big|.$$

The rest of the proof is devoted to study the behavior of the diagonal entries of $\Delta^{\mathcal{B}}(\rho_1,\rho_2)$. We need to show that

$$\begin{aligned}
&\left|\frac{\Big(\Delta^{\mathcal{B}}(\rho_1,\rho_2)\Big)_{\boldsymbol{s},\boldsymbol{s}}}{\rho_2-\rho_1}\right| \lesssim N^{-2}\Big(\mathbb{1}_{\{d\ge 3\}} + \mathbb{1}_{\{d=2\}}\log N + \mathbb{1}_{\{d=1\}} N\Big), \quad \forall\, \boldsymbol{s}\in\mathcal{D}_n, \\
&\left|\frac{\Big(\Delta^{\mathcal{B}}(\rho_1,\rho_2)\Big)_{\boldsymbol{s},\boldsymbol{s}}}{\rho_2-\rho_1}\right| \gtrsim N^{-2}, \quad \forall\, \boldsymbol{s}\in\mathcal{D}_n, \text{and } \forall\, d\ge 3.
\end{aligned}$$

Applying similar techniques as (3.61) as well as Lemma 3.2 yields

$$\begin{aligned}
\max_{\boldsymbol{s}\in\mathcal{D}_n} \left|\frac{\Big(\Delta^{\mathcal{B}}(\rho_1,\rho_2)\Big)_{\boldsymbol{s},\boldsymbol{s}}}{\rho_2-\rho_1}\right| &= \max_{\boldsymbol{s}\in\mathcal{D}_n} \left| \int_{\mathbb{R}^d} \Big|f_{\boldsymbol{s}}^N(\boldsymbol{\omega})\Big|^2 \Big[\frac{h_N\Big(\rho_2\|\boldsymbol{\omega}\|_{\ell_2}\Big) - h_N\Big(\rho_1\|\boldsymbol{\omega}\|_{\ell_2}\Big)}{\rho_2-\rho_1}\Big] d\boldsymbol{\omega} \right| \\
&\lesssim N^{-2}\Big(\mathbb{1}_{\{d\ge 3\}} + \mathbb{1}_{\{d=2\}}\log N + \mathbb{1}_{\{d=1\}} N\Big).
\end{aligned}$$

We now proceed to establish the desired lower bound on $tr(\Delta^{\mathcal{B}}(\rho_1,\rho_2))$. Choose any $s\in\mathcal{D}_n$.

Then,

$$\begin{aligned}\left(\Delta^{\mathcal{B}}(\rho_1,\rho_2)\right)_{\boldsymbol{s},\boldsymbol{s}} &= \int_{\mathbb{R}^d}\left|f_{\boldsymbol{s}}^N(\boldsymbol{\omega})\right|^2\left[h_N\left(\rho_2\left\|\boldsymbol{\omega}\right\|_{\ell_2}\right)-h_N\left(\rho_1\left\|\boldsymbol{\omega}\right\|_{\ell_2}\right)\right]d\boldsymbol{\omega}\\ &\geq \int_{\left\|\boldsymbol{\omega}\right\|_{\ell_2}\geq 1}\left|f_{\boldsymbol{s}}^N(\boldsymbol{\omega})\right|^2\left[h_N\left(\rho_2\left\|\boldsymbol{\omega}\right\|_{\ell_2}\right)-h_N\left(\rho_1\left\|\boldsymbol{\omega}\right\|_{\ell_2}\right)\right]d\boldsymbol{\omega} \qquad (3.62)\end{aligned}$$

Let us control $h_N\left(\rho_2\left\|\boldsymbol{\omega}\right\|_{\ell_2}\right)-h_N\left(\rho_1\left\|\boldsymbol{\omega}\right\|_{\ell_2}\right)$ from below. Due to the fact that (its proof is similar to (3.57) and we left it to the reader)

$$(1+x)^{-\alpha}-(1+y)^{-\alpha}\geq\frac{\alpha(y-x)}{2},\quad\forall\,0<x\leq y<2^{1/(\alpha+1)}-1,$$

it is possible to write

$$\frac{h_N\left(\rho_2\left\|\boldsymbol{\omega}\right\|_{\ell_2}\right)-h_N\left(\rho_1\left\|\boldsymbol{\omega}\right\|_{\ell_2}\right)}{\rho_2-\rho_1}\geq\frac{\left(\nu+\frac{d}{2}\right)}{2N^2\left\|\boldsymbol{\omega}\right\|_{\ell_2}^2}\frac{\rho_1+\rho_2}{\rho_1^2\rho_2^2}\gtrsim\left(N\left\|\boldsymbol{\omega}\right\|_{\ell_2}\right)^{-2}. \qquad (3.63)$$

for large enough N. Moreover, the class of functions $\left\{f_{\boldsymbol{s}}^N(\boldsymbol{\omega})\right\}_{\boldsymbol{s}\in\mathcal{D}_n}$ are nonzero (in a large enough neighborhood of the origin), continuously differentiable, with a uniformly bounded derivative when $\left\|\boldsymbol{\omega}\right\|_{\ell_2}\geq 1$, and decay with the polynomial rate given in Lemma 3.1. So

$$\int_{\left\|\boldsymbol{\omega}\right\|_{\ell_2}\geq 1}\left|\frac{f_{\boldsymbol{s}}^N(\boldsymbol{\omega})}{\left\|\boldsymbol{\omega}\right\|_{\ell_2}}\right|^2 d\boldsymbol{\omega}\asymp 1,\quad\forall\,\boldsymbol{s}\in\mathcal{D}_n. \qquad (3.64)$$

Replacing (3.64) and (3.63) into Eq. (3.62) gives the desirable lower bound. □

Lemma 3.4. Let $\rho_1,\rho_2\in\Theta_0$. Then

$$\left\|\Delta^{\mathcal{B}}(\rho_1,\rho_2)\right\|_{2\to 2}\lesssim\left(1\wedge|\rho_2-\rho_1|\right)\left(1+\mathbb{1}_{\{m=\nu+d/2\}}\log N\right). \qquad (3.65)$$

Moreover, if $\mathcal{D}_n$ be a d-dimensional regular lattice, then

$$\left\|\Delta^{\mathcal{B}}(\rho_1,\rho_2)\right\|_{2\to 2}\lesssim\left(1\wedge|\rho_2-\rho_1|\right). \qquad (3.66)$$

Proof. Consider any arbitrary partitioning $\mathcal{B}$. We know that $\Delta^{\mathcal{B}}(\rho_1,\rho_2)$ is a block diagonal approximation of $\Delta(\rho_1,\rho_2)$. The basic properties of operator norm implies that

$$\left\|\Delta^{\mathcal{B}}(\rho_1,\rho_2)\right\|_{2\to 2}\leq\left\|\Delta(\rho_1,\rho_2)\right\|_{2\to 2}.$$

Hence, we just need to find an upper bound on $\left\|\Delta(\rho_1,\rho_2)\right\|_{2\to 2}$. Without loss of generality,

suppose that $\rho_2 > \rho_1$. If $\rho_2 - \rho_1 > 1$ then the positive definiteness of $\Delta(\rho_1,\rho_2)$ (see (3.61)) implies that

$$\|\Delta(\rho_1,\rho_2)\|_{2\to 2} \le \left\|L_{n,m}(\rho_2)\right\|_{2\to 2}. \tag{3.67}$$

Now assume that $(\rho_2 - \rho_1)$ is strictly less than 1. We also showed that for any unit norm column vector v (of the proper size)

$$\frac{v^\top \Delta(\rho_1,\rho_2)\, v}{\rho_2-\rho_1} = \int_{\mathbb{R}^d} \left| \sum_{s\in\mathcal{D}_n} v_s e^{j\langle s, N\omega\rangle} f_s^N(\omega) \right|^2 \left\{ \frac{h_N\left(\rho_2 \|\omega\|_{\ell_2}\right) - h_N\left(\rho_1 \|\omega\|_{\ell_2}\right)}{\rho_2-\rho_1} \right\} d\omega.$$

The mean value theorem gives an alternative form for $h_N\left(\rho_2 \|\omega\|_{\ell_2}\right) - h_N\left(\rho_1 \|\omega\|_{\ell_2}\right)$.

$$\begin{aligned} \exists\, \rho \in (\rho_1,\rho_2) \ \text{ s.t. } \ \frac{h_N\left(\rho_2 \|\omega\|_{\ell_2}\right) - h_N\left(\rho_1 \|\omega\|_{\ell_2}\right)}{\rho_2-\rho_1} &= \dot{h}_N\left(\rho \|\omega\|_{\ell_2}\right) \\ &= \frac{2\nu+d}{\rho} \frac{h_N\left(\rho \|\omega\|_{\ell_2}\right)}{1+\left(N\rho \|\omega\|_{\ell_2}\right)^2}. \end{aligned}$$

In following identity we show that $\sup_{\rho\in[\rho_1,\rho_2]} \dot{h}_N\left(\rho \|\omega\|_{\ell_2}\right) \lesssim h_N\left(\rho_2 \|\omega\|_{\ell_2}\right)$.

$$\begin{aligned} \frac{h_N\left(\rho_2 \|\omega\|_{\ell_2}\right) - h_N\left(\rho_1 \|\omega\|_{\ell_2}\right)}{\rho_2-\rho_1} &\le \frac{2\nu+d}{\rho_1} \frac{h_N\left(\rho \|\omega\|_{\ell_2}\right)}{1+\left(N\rho \|\omega\|_{\ell_2}\right)^2} \le \frac{2\nu+d}{\rho_1} h_N\left(\rho \|\omega\|_{\ell_2}\right) \\ &\lesssim h_N\left(\rho_2 \|\omega\|_{\ell_2}\right). \end{aligned} \tag{3.68}$$

The last inequality in (3.68) is an easy consequence of the fact that $\inf(\Theta_0) > 0$. Thus,

$$0 \le \frac{v^\top \Delta(\rho_1,\rho_2)\, v}{\rho_2-\rho_1} \lesssim \int_{\mathbb{R}^d} \left| \sum_{s\in\mathcal{D}_n} v_s e^{j\langle s, N\omega\rangle} f_s^N(\omega) \right|^2 h_N\left(\rho_2 \|\omega\|_{\ell_2}\right) d\omega = v^\top L_{n,m}(\rho_2)\, v.$$

In other words, there is a bounded constant $c > 1$ for which

$$\frac{\Delta(\rho_1,\rho_2)}{\rho_2-\rho_1} \le c L_{n,m}(\rho_2) \quad \Rightarrow \quad \frac{\|\Delta(\rho_1,\rho_2)\|_{2\to 2}}{\rho_2-\rho_1} \lesssim \left\|L_{n,m}(\rho_2)\right\|_{2\to 2}. \tag{3.69}$$

Combining (3.67) and (3.69) leads to

$$\|\Delta(\rho_1,\rho_2)\|_{2\to 2} \lesssim \left(1 \wedge |\rho_2-\rho_1|\right) \left\|L_{n,m}(\rho_2)\right\|_{2\to 2}.$$

In the case that $\mathcal{D}_n$ is a regular lattice, $\left\|L_{n,m}(\rho_2)\right\|_{2\to 2}$ is known to be less than some bounded

scalar C (see [SCA12], Theorem 3.1), which justifies (3.66). For arbitrary irregular lattices satisfying Assumption 3.1, Proposition 3.1 characterizes the decay rate of the off diagonal entries of $L_{n,m}(\rho_2)$. Thus, applying Lemma 3.9 immediately substantiates (3.66) and ends the proof. □

Lemma 3.5. Let $N \coloneqq \lfloor n^{1/d} \rfloor$ and select two distinct ρ_1 and ρ_2 in Θ_0. Then,

$$\frac{\left\|\Delta^{\mathcal{B}}(\rho_1,\rho_2)\right\|_{\ell_2}}{|\rho_2-\rho_1|} \lesssim \left(\mathbb{1}_{\{d=1\}} + \mathbb{1}_{\{d=2\}}\log n + \mathbb{1}_{\{d=3\}} n^{1/3} + \mathbb{1}_{\{d\geq 4\}} n^{1/2}\right).$$

Proof. The same logic as in the proof of Lemma 3.4 leads to

$$\left\|\Delta^{\mathcal{B}}(\rho_1,\rho_2)\right\|_{\ell_2} \leq \|\Delta(\rho_1,\rho_2)\|_{\ell_2}.$$

So it suffices to control $\|\Delta(\rho_1,\rho_2)\|_{\ell_2}$ from above. When $d \leq 4$, it is trivial that

$$\|\Delta(\rho_1,\rho_2)\|_{\ell_2} \leq \|\Delta(\rho_1,\rho_2)\|_{\mathcal{S}_1}.$$

Substituting the bound on $\|\Delta(\rho_1,\rho_2)\|_{\mathcal{S}_1}$ from Lemma 3.3 in the above inequality leads to the desired result. Now suppose that $d \geq 5$. In this case, $1-2/d > 1/2$ and so we inevitably need new proof techniques. Without loss of generality assume that $\rho_2 \geq \rho_1$. In (3.69), we showed that

$$\|\Delta(\rho_1,\rho_2)\|_{\ell_2} \leq \left\|L_{n,m}(\rho_2)\right\|_{\ell_2}(\rho_2-\rho_1).$$

We also know from Proposition 3.1 that

$$\left|\left(L_{n,m}(\rho_2)\right)_{\boldsymbol{s},\boldsymbol{t}}\right| \lesssim \left(1+N\|\boldsymbol{t}-\boldsymbol{s}\|_{\ell_2}\right)^{-2(m-\nu)}, \tag{3.70}$$

which means that $\left\|L_{n,m}(\rho_2)\right\|_{\ell_2} \lesssim \sqrt{n}$ (see the second part of Lemma 3.9). In summary for $d \geq 5$,

$$\|\Delta(\rho_1,\rho_2)\|_{\ell_2} \leq \left\|L_{n,m}(\rho_2)\right\|_{\ell_2}|\rho_2-\rho_1| \lesssim n^{1/2}|\rho_2-\rho_1|.$$

□

Lemma 3.6. There exists a large enough N_0 such that for any $N \geq N_0$,

$$\min_{\rho\in\Theta_0} \frac{\left\|L^{\mathcal{B}}_{n,m}(\rho)\right\|_{\ell_2}}{\sqrt{n}} > 0.$$

Proof. Let $\rho_{\min}$ represents the smallest member of Θ_0. We have shown in the proof of

Lemma 3.3 (inequality (3.61)) that

$$L^{\mathcal{B}}_{n,m}(\rho) \succcurlyeq L^{\mathcal{B}}_{n,m}(\rho_{\min}), \quad \forall\, \rho \in \Theta_0$$

Henceforth, all the eigenvalues of $L^{\mathcal{B}}_{n,m}(\rho)$ are greater than or equal to the corresponding eigenvalues of $L^{\mathcal{B}}_{n,m}(\rho_{\min})$. So $n^{-1/2}\left\|L^{\mathcal{B}}_{n,m}(\rho)\right\|_{\ell_2}$ attains its minimum at $\rho = \rho_{\min}$, due to the positive definiteness of $L_{n,m}(\rho)$ and $L^{\mathcal{B}}_{n,m}(\rho_{\min})$. As $L^{\mathcal{B}}_{n,m}(\rho_{\min})$ is a square matrix of size n, it suffices to show that all of its diagonal entries are bounded away from zero.

$$\left\|L^{\mathcal{B}}_{n,m}(\rho_{\min})\right\|^2_{\ell_2} \geq \sum_{\boldsymbol{s}\in\mathcal{D}_n} \left|\left(L^{\mathcal{B}}_{n,m}(\rho_{\min})\right)_{\boldsymbol{s},\boldsymbol{s}}\right|^2 = \sum_{\boldsymbol{s}\in\mathcal{D}_n} \left|\left(L_{n,m}(\rho_{\min})\right)_{\boldsymbol{s},\boldsymbol{s}}\right|^2.$$

Recall the two functions $f^N_{\boldsymbol{s}}$ and h_N from Eq. (3.44) and (3.45), respectively. Now choose an arbitrary $\boldsymbol{s} \in \mathcal{D}_n$ and a large enough $R \in (0,\infty)$. From the identity (3.46), we have a closed form expression for the diagonal entries of $L_{n,m}(\rho_{\min})$.

$$\left(L_{n,m}(\rho_{\min})\right)_{\boldsymbol{s},\boldsymbol{s}} = \int_{\mathbb{R}^d} \left|f^N_{\boldsymbol{s}}(\boldsymbol{\omega})\right|^2 h_N\left(\rho_{\min}\|\boldsymbol{\omega}\|_{\ell_2}\right) d\boldsymbol{\omega} > \int_{\|\boldsymbol{\omega}\|_{\ell_2}\leq R} \left|f^N_{\boldsymbol{s}}(\boldsymbol{\omega})\right|^2 h_N\left(\rho_{\min}\|\boldsymbol{\omega}\|_{\ell_2}\right) d\boldsymbol{\omega}.$$

We trivially can choose N_0 (depending on Θ_0 and R) such that $\inf_{\|\boldsymbol{\omega}\|_{\ell_2}\leq R} h_N\left(\rho_{\min}\|\boldsymbol{\omega}\|_{\ell_2}\right) \geq \frac{1}{2}$ for any $N \geq N_0$. Thus,

$$\left|\left(L_{n,m}(\rho_{\min})\right)_{\boldsymbol{s},\boldsymbol{s}}\right| > \frac{1}{2}\int_{\|\boldsymbol{\omega}\|_{\ell_2}\leq R} \left|f^N_{\boldsymbol{s}}(\boldsymbol{\omega})\right|^2 d\boldsymbol{\omega}.$$

□

Lemma 3.7. There exist a strictly positive scalars C_1 and C_2 such that

$$C_1\sqrt{n} \geq \left\|\sqrt{L_{n,m}(\rho_1)}L^{\mathcal{B}}_{n,m}(\rho_2)\sqrt{L_{n,m}(\rho_1)}\right\|_{\ell_2} \geq C_2\sqrt{n}, \quad \forall\, \rho_1,\rho_2 \in \Theta_0. \tag{3.71}$$

Proof. For brevity we use Q to refer the Frobenius norm in Eq. (3.71). The cyclic permutation property of trace operator implies that

$$\left\|\sqrt{L_{n,m}(\rho_1)}L^{\mathcal{B}}_{n,m}(\rho_2)\sqrt{L_{n,m}(\rho_1)}\right\|_{\ell_2} = \left\|\sqrt{L^{\mathcal{B}}_{n,m}(\rho_2)}L_{n,m}(\rho_1)\sqrt{L^{\mathcal{B}}_{n,m}(\rho_2)}\right\|_{\ell_2}.$$

The inequality (3.61) indicates that $L_{n,m}(\rho_1\vee\rho_2) \succcurlyeq L_{n,m}(\rho_1)$ and $L^{\mathcal{B}}_{n,m}(\rho_1\vee\rho_2) \succcurlyeq L^{\mathcal{B}}_{n,m}(\rho_2)$.

So

$$\begin{aligned}\left\|\sqrt{L^{\mathcal{B}}_{n,m}(\rho_2)}L_{n,m}(\rho_1)\sqrt{L^{\mathcal{B}}_{n,m}(\rho_2)}\right\|^2_{\ell_2} &\leq \left\|\sqrt{L^{\mathcal{B}}_{n,m}(\rho_2)}L_{n,m}(\rho_1\vee\rho_2)\sqrt{L^{\mathcal{B}}_{n,m}(\rho_2)}\right\|^2_{\ell_2}\\ &= \left\|\sqrt{L_{n,m}(\rho_1\vee\rho_2)}L^{\mathcal{B}}_{n,m}(\rho_2)\sqrt{L_{n,m}(\rho_1\vee\rho_2)}\right\|^2_{\ell_2}\\ &\leq \left\|\sqrt{L_{n,m}(\rho_1\vee\rho_2)}L^{\mathcal{B}}_{n,m}(\rho_1\vee\rho_2)\sqrt{L_{n,m}(\rho_1\vee\rho_2)}\right\|^2_{\ell_2}.\end{aligned}$$

Thus we may suppose that $\rho_2 \geq \rho_1$ without losing the generality. Namely $\rho_1 \vee \rho_2 = \rho_2$. In summary, so far we have

$$Q \leq \left\|\sqrt{L_{n,m}(\rho_2)}L^{\mathcal{B}}_{n,m}(\rho_2)\sqrt{L_{n,m}(\rho_2)}\right\|_{\ell_2}.$$

On the other hand,

$$\left\|\sqrt{L_{n,m}(\rho_2)}L^{\mathcal{B}}_{n,m}(\rho_2)\sqrt{L_{n,m}(\rho_2)}\right\|^2_{\ell_2} = \text{RHS} := tr\left\{L_{n,m}(\rho_2)\,L^{\mathcal{B}}_{n,m}(\rho_2)\,L_{n,m}(\rho_2)\,L^{\mathcal{B}}_{n,m}(\rho_2)\right\}.$$

For any matrix A, define its absolute value by $|A| = \left[\left|A_{s,t}\right|\right]$. The triangle inequality says that for matrices $A_1,\ldots,A_b$, for some $b \in \mathbb{N}$, we have

$$tr(A_1\ldots A_b) \leq tr(|A_1|\ldots|A_b|).$$

This fact help us to find an upper bound on RHS.

$$\text{RHS} \leq tr\left\{\left|L_{n,m}(\rho_2)\right|\left|L^{\mathcal{B}}_{n,m}(\rho_2)\right|\left|L_{n,m}(\rho_2)\right|\left|L^{\mathcal{B}}_{n,m}(\rho_2)\right|\right\}.$$

Finally, since $\left|L^{\mathcal{B}}_{n,m}(\rho_2)\right|$ is the block diagonalized version of $\left|L_{n,m}(\rho_2)\right|$ and both of these matrices have non-negative entries, we get

$$\begin{aligned}tr\left\{\left|L_{n,m}(\rho_2)\right|\left|L^{\mathcal{B}}_{n,m}(\rho_2)\right|\left|L_{n,m}(\rho_2)\right|\left|L^{\mathcal{B}}_{n,m}(\rho_2)\right|\right\} &\leq tr\left\{\left|L_{n,m}(\rho_2)\right|\left|L_{n,m}(\rho_2)\right|\left|L_{n,m}(\rho_2)\right|\left|L_{n,m}(\rho_2)\right|\right\}\\ &= \left\|\left|L_{n,m}(\rho_2)\right|^2\right\|^2_{\ell_2}.\end{aligned}$$

Combining the above inequalities yields

$$Q \leq \left\|\left|L_{n,m}(\rho_2)\right|^2\right\|_{\ell_2}.$$

Notice that the off-diagonal entries of $L_{n,m}(\rho_2)$ and $\left|L_{n,m}(\rho_2)\right|$ decay with the same rate. Thus applying Lemma 3.8 can determine an bound on the entries of $\left|L_{n,m}(\rho_2)\right|^2$ as the

following.

$$\left|\left(\left|L_{n,m}(\rho_2)\right|^2\right)_{\boldsymbol{s},\boldsymbol{t}}\right| \lesssim \left(1+N\|\boldsymbol{t}-\boldsymbol{s}\|_{\ell_2}\right)^{-2(m-\nu)}\left\{1+\mathbb{1}_{\{m=\nu+d/2\}}\log\left(1+N\|\boldsymbol{t}-\boldsymbol{s}\|_{\ell_2}\right)\right\}.$$

Finally, Lemma 3.10 guarantees the existence of a bounded scalar c for which $\left\|L^2_{n,m}(\rho_2)\right\|_{\ell_2} \leq c\sqrt{n}$, finishing the proof of the first part. We now turn to the proof of the other side. Using the same trick as before implies that

$$Q \geq \left\|\sqrt{L_{n,m}(\rho_1)}L^{\mathcal{B}}_{n,m}(\rho_1)\sqrt{L_{n,m}(\rho_1)}\right\|_{\ell_2}.$$

□

3.8.3 The Basic Properties of Matrices with Polynomial Decaying Off-diagonals

We showed in Section 3.8.1.1 that the off-diagonal entries of $K^{\mathcal{B}}_{n,m}(\rho)$ decay polynomially in terms of the distance to the main diagonal. In this section, we show that such class of matrices are close to multiplication. We also investigate the large sample properties of their norms.

Lemma 3.8. Let $N = \lfloor n^{1/d} \rfloor$ and suppose that $A_n \in \mathbb{R}^{n\times n}$ whose entries satisfy

$$\left|A_{\boldsymbol{s},\boldsymbol{t}}\right| \leq C\left(1+N\|\boldsymbol{t}-\boldsymbol{s}\|_{\ell_2}\right)^{-(d+\zeta)}, \quad \forall\, \boldsymbol{s},\boldsymbol{t} \in \mathcal{D}_n. \tag{3.72}$$

for some bounded $C > 0$ and $\zeta \geq 0$. Then, the entries of $B = A^2$ are bounded above by

$$\left|B_{\boldsymbol{s},\boldsymbol{t}}\right| \lesssim \left(1+N\|\boldsymbol{t}-\boldsymbol{s}\|_{\ell_2}\right)^{-(d+\zeta)}\left\{1+\mathbb{1}_{\{\zeta=0\}}\log\left(1+N\|\boldsymbol{t}-\boldsymbol{s}\|_{\ell_2}\right)\right\}, \quad \forall \boldsymbol{s},\boldsymbol{t} \in \mathcal{D}_n. \tag{3.73}$$

Proof. For simplicity let $\Delta = N(\boldsymbol{t}-\boldsymbol{s})$. Without loss of generality assume that $C = 1$. We first justify Eq. (3.73) for the special case of $\Delta = \mathbf{0}_d$ (associated to the diagonal entries of B). Indeed we need to show that all the diagonal entries of B are smaller than some bounded scalar C', which depends on d, C, and $\mathcal{D}_n$, i.e., $\left|B_{\boldsymbol{s},\boldsymbol{s}}\right| \leq C'$ for any $\boldsymbol{s} \in \mathcal{D}_n$. Notice that the pairwise distances among two points in $\mathcal{D}_n$ have a similar behaviour to that of a

d-dimensional regular lattice. Thus,

$$\begin{aligned}\left|B_{\boldsymbol{s},\boldsymbol{s}}\right| &= \left|\sum_{\boldsymbol{r}\in\mathcal{D}_n} A_{\boldsymbol{s},\boldsymbol{r}}^2\right| \leq \sum_{\boldsymbol{r}\in\mathcal{D}_n}\left(1+N\|\boldsymbol{r}-\boldsymbol{s}\|_{\ell_2}\right)^{-2(d+\zeta)} \lesssim \int_0^\infty x^{d-1}\,(1+x)^{-2(d+\zeta)}\,dx \\ &\lesssim \int_1^\infty x^{-(d+1+2\zeta)}dx \asymp 1.\end{aligned}$$

Now suppose that Δ is a non-zero vector. Clearly $1 \lesssim \|\Delta\|_{\ell_2} \lesssim N$ and so $1+\|\Delta\|_{\ell_2}^{d+\zeta} \asymp \|\Delta\|_{\ell_2}^{d+\zeta}$. We replace Eq. (3.72) with the following more algebraically convenient alternative form.

$$\left|A_{\boldsymbol{s},\boldsymbol{t}}\right| \lesssim \left[1+\|\Delta\|_{\ell_2}^{d+\zeta}\right]^{-1}, \quad \forall \boldsymbol{s},\boldsymbol{t}\in\mathcal{D}_n, \quad \left(\boldsymbol{t}=\boldsymbol{s}+\frac{\Delta}{N}\right).$$

Next we obtain an upper bound on $\left|B_{\boldsymbol{s},\boldsymbol{t}}\right|$ as the sum of two terms.

$$\begin{aligned}\left|B_{\boldsymbol{s},\boldsymbol{t}}\right| &\lesssim \sum_{\boldsymbol{r}\in\mathcal{D}_n} \frac{1}{\left(1+\|N(\boldsymbol{s}-\boldsymbol{r})\|_{\ell_2}^{d+\zeta}\right)\left(1+\|N(\boldsymbol{t}-\boldsymbol{r})\|_{\ell_2}^{d+\zeta}\right)} \\ &= \sum_{\boldsymbol{r}\in\mathcal{D}_n} \frac{\left(1+\|N(\boldsymbol{s}-\boldsymbol{r})\|_{\ell_2}^{d+\zeta}\right)^{-1}}{2+\|N(\boldsymbol{s}-\boldsymbol{r})\|_{\ell_2}^{d+\zeta}+\|N(\boldsymbol{t}-\boldsymbol{r})\|_{\ell_2}^{d+\zeta}} \\ &+ \sum_{\boldsymbol{r}\in\mathcal{D}_n} \frac{\left(1+\|N(\boldsymbol{t}-\boldsymbol{r})\|_{\ell_2}^{d+\zeta}\right)^{-1}}{2+\|N(\boldsymbol{s}-\boldsymbol{r})\|_{\ell_2}^{d+\zeta}+\|N(\boldsymbol{t}-\boldsymbol{r})\|_{\ell_2}^{d+\zeta}}.\end{aligned}$$

We write ξ_1 and ξ_2 to denote the first and second terms in the last line of the above expression. The next step serves as controlling ξ_1 from above. A similar upper bound can be found on ξ_2. For doing so, we introduce a lower bound on the expression in the denominator of ξ_1. Define $c = 2^{d+\zeta-1} \geq 1$. Applying Jensen's inequality on the convex univariate function $f(x) = x^{d+\zeta}$ implies that

$$\begin{aligned}\|N(\boldsymbol{s}-\boldsymbol{r})\|_{\ell_2}^{d+\zeta}+\|N(\boldsymbol{t}-\boldsymbol{r})\|_{\ell_2}^{d+\zeta} &\geq \frac{\|N(\boldsymbol{s}-\boldsymbol{r})\|_{\ell_2}^{d+\zeta}}{c+1}+\frac{c}{c+1}\left(\|N(\boldsymbol{s}-\boldsymbol{r})\|_{\ell_2}^{d+\zeta}+\|N(\boldsymbol{t}-\boldsymbol{r})\|_{\ell_2}^{d+\zeta}\right) \\ &\geq \frac{\|N(\boldsymbol{s}-\boldsymbol{r})\|_{\ell_2}^{d+\zeta}}{c+1}+\frac{\left(\|N(\boldsymbol{s}-\boldsymbol{r})\|_{\ell_2}+\|N(\boldsymbol{t}-\boldsymbol{r})\|_{\ell_2}\right)^{d+\zeta}}{c+1} \\ &\geq \frac{\|N(\boldsymbol{s}-\boldsymbol{r})\|_{\ell_2}^{d+\zeta}+\|\Delta\|_{\ell_2}^{d+\zeta}}{c+1}.\end{aligned}$$

Thus

$$\xi_1 \lesssim \sum_{\boldsymbol{r}\in\mathcal{D}_n} \frac{1}{\left(1+\|N(\boldsymbol{s}-\boldsymbol{r})\|_{\ell_2}^{d+\zeta}\right)\left(1+\|N(\boldsymbol{s}-\boldsymbol{r})\|_{\ell_2}^{d+\zeta}+\|\Delta\|_{\ell_2}^{d+\zeta}\right)}. \tag{3.74}$$

Notice that the points in $\{N(\boldsymbol{s}-\boldsymbol{r}),\ \boldsymbol{r}\in\mathcal{D}_n\}$ belong to a scaled (with the factor N) and translated version of $\mathcal{D}_n$. Assumption 3.1 states that the pairwise distances in $\mathcal{D}_n$ and a regular lattice look alike. Hence, the summation in the right hand side of Eq. (3.74), which only depends on the norm of the elements in $\mathcal{D}_n-\boldsymbol{s}$, can be upper bounded by an integral. Strictly speaking (in the following x represents $\|N(\boldsymbol{s}-\boldsymbol{r})\|_{\ell_2}$)

$$\begin{aligned}
\xi_1 &\lesssim \int_0^N \frac{x^{d-1}dx}{(1+x^{d+\zeta})\left(1+x^{d+\zeta}+\|\Delta\|_{\ell_2}^{d+\zeta}\right)} \\
&= \frac{1}{\|\Delta\|_{\ell_2}^{d+\zeta}}\int_0^N\left(\frac{x^{d-1}}{1+x^{d+\zeta}}-\frac{x^{d-1}}{1+x^{d+\zeta}+\|\Delta\|_{\ell_2}^{d+\zeta}}\right)dx \\
&\lesssim \|\Delta\|_{\ell_2}^{-(d+\zeta)}\left[1+\mathbb{1}_{\{\zeta=0\}}\log\left(\frac{N^d\|\Delta\|_{\ell_2}^d}{N^d+\|\Delta\|_{\ell_2}^d}\right)\right] \asymp \|\Delta\|_{\ell_2}^{-(d+\zeta)}\left(1+\mathbb{1}_{\{\zeta=0\}}\log\|\Delta\|_{\ell_2}^d\right).
\end{aligned}$$

An analogous bound holds for ξ_2. Replacing these upper bounds in $\left|B_{\boldsymbol{s},\boldsymbol{t}}\right|\lesssim\xi_1+\xi_2$ ends the proof. □

Lemma 3.9. Let $\mathcal{D}_n$ be a irregular lattice of size n satisfying Assumption 3.1. Define $N\coloneqq\lfloor n^{1/d}\rfloor$ and let $\Psi^n\in\mathbb{R}^{n\times n}$ be a symmetric matrix associated to $\mathcal{D}_n$ whose entries satisfy

$$\left|\Psi^n_{\boldsymbol{s},\boldsymbol{t}}\right|\le C\left(1+N\|\boldsymbol{s}-\boldsymbol{t}\|_{\ell_2}\right)^{-(d+\zeta)},\quad\forall\ \boldsymbol{s},\boldsymbol{t}\in\mathcal{D}_n$$

for some non-negative ζ and $C\in(0,\infty)$. Then there exist bounded scalar $A,A'>0$ (depending on C,d and ζ) for which

1. $\|\Psi^n\|_{2\to2}\le A\left(1+\mathbb{1}_{\{\zeta=0\}}\log n\right)$.
2. $\|\Psi^n\|_{\ell_2}\le A'\sqrt{n}$.

Proof. We first focus on the operator norm of Ψ^n. The symmetry of Ψ^n implies that

$$\begin{aligned}
\left\|\Psi^n\right\|_{2\to2} &\le \sqrt{\|\Psi^n\|_{1\to1}\|\Psi^n\|_{\infty\to\infty}}=\left\|\Psi^n\right\|_{1\to1}=\max_{\boldsymbol{s}\in\mathcal{D}_n}\sum_{\boldsymbol{t}\in\mathcal{D}_n}\left|\Psi^n_{\boldsymbol{s},\boldsymbol{t}}\right| \\
&\le C\max_{\boldsymbol{s}\in\mathcal{D}_n}\sum_{\boldsymbol{t}\in\mathcal{D}_n}\left(1+N\|\boldsymbol{s}-\boldsymbol{t}\|_{\ell_2}\right)^{-(d+\zeta)}. \qquad (3.75)
\end{aligned}$$

Choose $\boldsymbol{s}\in\mathcal{D}_n$. Reorder the points in $\mathcal{D}_n$ based on their distance from $\boldsymbol{s}$. Define the non-overlapping sets $\Pi_{\boldsymbol{s},l}$ by

$$\Pi_{\boldsymbol{s},l}=\left\{\boldsymbol{t}\in\mathcal{D}_n:\frac{l}{N}\le\|\boldsymbol{s}-\boldsymbol{t}\|_{\ell_2}<\frac{l+1}{N}\right\},\quad\forall\ l\in\mathbb{N}\cup\{0\}.$$

The following facts are trivial implications of Assumption 3.1.

- There exists a bounded constant $a > 0$ such that $\Pi_{\boldsymbol{s},l} = \emptyset$ for any $l > aN$.
- $\left|\Pi_{\boldsymbol{s},l}\right| \lesssim (l+1)^d - l^d \lesssim (l+1)^{d-1}$ for any $l \leq aN$.

Thus,

$$\sum_{\boldsymbol{t}\in\mathcal{D}_n} \left(1 + N\|\boldsymbol{s}-\boldsymbol{t}\|_{\ell_2}\right)^{-(d+\zeta)} \leq \sum_{l=0}^{\infty} \left|\Pi_{\boldsymbol{s},l}\right| (l+1)^{-(d+\zeta)} \lesssim \sum_{l=0}^{aN} (l+1)^{-(1+\zeta)}. \tag{3.76}$$

We conclude the proof by substituting Eq. (3.76) into Eq. (3.75). Now we turn into finding an upper bound on $n^{-1}\|\Psi^n\|_{\ell_2}^2$. Using similar techniques as (3.76) yields

$$\begin{aligned} n^{-1}\left\|\Psi^n\right\|_{\ell_2}^2 &\leq n^{-1}\sum_{\boldsymbol{s}\in\mathcal{D}_n}\sum_{l=0}^{\infty}\left|\Pi_{\boldsymbol{s},l}\right| \sup_{\boldsymbol{t}\in\Pi_{\boldsymbol{s},l}} \left|\Psi^n_{\boldsymbol{s},\boldsymbol{t}}\right|^2 \leq \sum_{l=0}^{\infty}\left|\Pi_{\boldsymbol{s},l}\right| \sup_{\boldsymbol{t}\in\Pi_{\boldsymbol{s},l}} \left|\Psi^n_{\boldsymbol{s},\boldsymbol{t}}\right|^2 \\ &\leq C^2 \sum_{l=0}^{aN}\left|\Pi_{\boldsymbol{s},l}\right|(l+1)^{-2(d+\zeta)} \lesssim \sum_{l=0}^{\infty}(l+1)^{-(d+1+2\zeta)} \asymp 1. \end{aligned} \tag{3.77}$$

□

The next result has a similar flavor as the second part of Lemma 3.9. We omit its proof for avoiding the repetition.

Lemma 3.10. Let $\mathcal{D}_n$ be a irregular lattice of size n satisfying Assumption 3.1. Define $N \coloneqq \lfloor n^{1/d} \rfloor$ and let $\Psi^n \in \mathbb{R}^{n\times n}$ be a symmetric matrix associated to $\mathcal{D}_n$ whose entries satisfy

$$\left|\Psi^n_{\boldsymbol{s},\boldsymbol{t}}\right| \leq C\left(1 + N\|\boldsymbol{s}-\boldsymbol{t}\|_{\ell_2}\right)^{-(d+\zeta)} \left\{1 + \mathbb{1}_{\{\zeta=0\}} \log\left(1 + N\|\boldsymbol{s}-\boldsymbol{t}\|_{\ell_2}\right)\right\}, \quad \forall\, \boldsymbol{s},\boldsymbol{t} \in \mathcal{D}_n$$

for some non-negative ζ and $C \in (0,\infty)$. Then there exists a bounded scalar $A > 0$ (depending on C, d and ζ) for which

$$\left\|\Psi^n\right\|_{\ell_2} \leq A\sqrt{n}.$$

3.8.4 Probabilistic Inequalities

We first extend Proposition $A.3$ of [KSN16] regarding the uniform concentration of generalized χ^2 random processes around its mean. It provides a powerful tool in the proof of Theorems 3.1 and 3.2.

Proposition 3.2. Let $\Theta_0 \subset \mathbb{R}^b$, $\forall\, n \in \mathbb{N}$ be a compact space with respect to the Euclidean metric. Consider the class of $n \times n$ matrices $\{\Pi_n(\theta)\}_{\theta\in\Theta_0}$ parametrized by $\theta \in \Theta_0$. Suppose that the following conditions hold

(a) The normalized Frobenius norm of $\Pi_n(\theta)$ is uniformly bounded on Θ_0, i.e.,

$$J_{\max} := \sup_n \sup_{\theta\in\Theta_0} n^{-1/2} \|\Pi_n(\theta)\|_{\ell_2} < \infty.$$

(b) The mapping $\left(\theta, \|\cdot\|_{\ell_2}\right) \mapsto (\Pi_n(\theta), \|\cdot\|_{2\to 2})$ is Lipschitz with constant of order $\log^2 n$. Namely, there is $C > 0$ for which

$$\|\Pi_n(\theta_2) - \Pi_n(\theta_1)\|_{2\to 2} \le C \log^2 n \|\theta_2 - \theta_1\|_{\ell_2}, \quad \forall\, \theta_1, \theta_2 \in \Theta_0 \ \text{ s.t.} |\theta_2 - \theta_1| \le 1. \quad (3.78)$$

(c)

$$\lim_{n\to\infty} \|\Pi_n(\theta)\|_{2\to 2} \sqrt{\frac{\log n}{n}} = 0, \quad \forall\, \theta \in \Theta_0.$$

Then, there is a finite positive constant C', depending on C, $J_{\max}$ and b, such that

$$\mathbb{P}\left(\sup_{\theta\in\Theta_0} \left| Z^\top \Pi_n(\theta) Z - tr\{\Pi_n(\theta)\} \right| \ge C' \sqrt{n \log n} \right) \le \frac{1}{n}, \quad \text{as } n \to \infty. \quad (3.79)$$

Proof. Let $r_n = 1/(C\sqrt{n \log^3 n})$ for C defined in Eq. (3.78). For large enough n, we have $r_n \le 1$. Let $\mathcal{N}_{r_n}(\Theta_0)$ represents the r_n-covering number of Θ_0. The simple volume argument implies that

$$\left|\mathcal{N}_{r_n}(\Theta_0)\right| \lesssim \left(\frac{\operatorname{diam}(\Theta_0)}{r_n} \right)^b = \mathcal{O}\left\{ \left(n \log^3 n \right)^{b/2} \right\}. \quad (3.80)$$

The key idea is to reduce the supremum over Θ_0 in (3.79) to the discrete finite space $\mathcal{N}_{r_n}(\Theta_0)$. Applying union bounded provides an upper bound on a probabilistic statement over $\mathcal{N}_{r_n}(\Theta_0)$. Using the Hanson-Wright concentration inequality concludes the proof.
For any $\theta \in \Theta_0$, let γ_θ stands for the closest element of $\mathcal{N}_{r_n}(\Theta_0)$ to θ. Thus, $\|\theta - \gamma_\theta\|_{\ell_2} \le r_n$. Observe that

$$\begin{aligned}
\text{RHS} &:= \left| Z^\top \Pi_n(\theta) Z - tr\{\Pi_n(\theta)\} - Z^\top \Pi_n(\gamma_\theta) Z + tr\{\Pi_n(\gamma_\theta)\} \right| \\
&= \left| \langle \Pi_n(\theta) - \Pi_n(\gamma_\theta), ZZ^\top + I_n \rangle \right| \le \|\Pi_n(\theta) - \Pi_n(\beta_\theta)\|_{2\to 2} \left\| ZZ^\top + I_n \right\|_{\mathcal{S}_1} \\
&\overset{(a)}{\le} C \log^2 n \|\theta - \beta_\theta\|_{\ell_2} \left\| ZZ^\top + I_n \right\|_{\mathcal{S}_1} \le C r_n \log^2 n \left\| ZZ^\top + I_n \right\|_{\mathcal{S}_1} \\
&= \sqrt{\frac{\log n}{n}} \left(n + \|Z\|_{\ell_2}^2 \right).
\end{aligned}$$

Here (a) is implied from Eq. (3.78). The Bernestein's inequality for the sub-exponential

random variables states that

$$\mathbb{P}\Big(\|Z\|_{\ell_2}^2 \geq n + nt\Big) \leq e^{-\frac{nt^2}{8}}, \quad \forall\, t > 0. \tag{3.81}$$

Choosing $t = 1$ in (3.81) shows that RHS $\geq 3\sqrt{n\log n}$ with probability at most $\exp(-n/8)$. Hence,

$$\mathbb{P}\left(\sup_{\theta\in\Theta_0}\Big|Z^\top\Pi_n(\theta)Z - tr(\Pi_n(\theta))\Big| \geq \sup_{\theta\in\mathcal{N}_{r_n}(\Theta_0)}\Big|Z^\top\Pi_n(\theta)Z - tr(\Pi_n(\theta))\Big| + 3\sqrt{n\log n}\right) \leq e^{-n/8}.$$

Recall $J_{\max}$ from the condition (a). Choose an arbitrary bounded ξ such that $\xi > 1 + b/2$. Eq. (3.80) can be rewritten as $\big|\mathcal{N}_{r_n}(\Theta_0)\big| n^{-\xi} = o\Big(n^{-1}\Big)$, when n tends to infinity. The proof will be terminated if we show that (for some bounded scalar C_0)

$$\mathbb{P}\left(\sup_{\theta\in\mathcal{N}_{r_n}(\Theta_0)}\Big|Z^\top\Pi_n(\theta)Z - tr\{\Pi_n(\theta)\}\Big| \geq C_0 J_{\max}\sqrt{n\log n}\right) \leq \big|\mathcal{N}_{r_n}(\Theta_0)\big| n^{-\xi} = o\left(\frac{1}{n}\right),$$

as n goes to infinity. For proving this claim, it suffices to obtain an appropriate probabilistic upper bound on $\Big|Z^\top\Pi_n(\theta)Z - tr\{\Pi_n(\theta)\}\Big|$ for any $\theta \in \mathcal{N}_{r_n}(\Theta_0)$ and then exploiting the union bound trick. Hanson-Wright inequality says that for some $C_0 < \infty$ (depending on ξ), we have

$$\mathbb{P}\Big[\Big|Z^\top\Pi_n(\theta)Z - tr\{\Pi_n(\theta)\}\Big| \geq C_0\Big(\|\Pi_n(\theta)\|_{\ell_2}\sqrt{\log n} \vee \|\Pi_n(\theta)\|_{2\to 2}\log n\Big)\Big] \leq n^{-\xi}. \tag{3.82}$$

The condition (c) means that, $\|\Pi_n(\theta)\|_{2\to 2}\log n = o\Big(\sqrt{n\log n}\Big)$ as n tends to infinity. So

$$\begin{aligned}\Big(\|\Pi_n(\theta)\|_{\ell_2}\sqrt{\log n} \vee \|\Pi_n(\theta)\|_{2\to 2}\log n\Big) &= \Big(\|\Pi_n(\theta)\|_{\ell_2}\sqrt{\log n} \vee o\Big(\sqrt{n\log n}\Big)\Big)\\ &\leq J_{\max}\sqrt{n\log n}, \quad \text{as } n\to\infty,\end{aligned}$$

due to the condition (a). Thus Eq. (3.82) can be rewritten as

$$\mathbb{P}\Big(\Big|Z^\top\Pi_n(\theta)Z - tr\{\Pi_n(\theta)\}\Big| \geq C_0 J_{\max}\sqrt{n\log n}\Big) \leq n^{-\xi}, \quad \forall\, \theta \in \mathcal{N}_{r_n}(\Theta_0),$$

ending the proof of the claim. □

Next we rigorously state the squeeze theorem for weak convergence. It is beneficial in the proof of Theorem 3.2.

Lemma 3.11. Let $\{X_n\}_{n=1}^\infty, \{Y_n\}_{n=1}^\infty$ be two real valued sequences converging to U in distri-

bution. Suppose that $\{Z_n\}_{n=1}^{\infty}$ satisfies the following inequality

$$X'_n := X_n(1-p_n) \leq Z_n \leq Y'_n := Y_n(1+q_n), \quad \forall\, n \in \mathbb{N}, \tag{3.83}$$

in which $p_n, q_n \xrightarrow{\mathbb{P}} 0$. Then $Z_n \xrightarrow{d} U$.

Proof. Let $t \in \mathbb{R}$ be a continuity point of U. It suffices to show that $\mathbb{P}(Z_n \geq t) \to \mathbb{P}(U \geq t)$ as n tends to infinity. Eq. (3.83) obviously means that

$$\mathbb{P}(X'_n \geq t) \leq \mathbb{P}(Z_n \geq t) \leq \mathbb{P}(Y'_n \geq t), \quad \forall\, n \in \mathbb{N}.$$

Both X'_n and Y'_n weakly converge to U by *Slutsky's theorem*. Hence, $\mathbb{P}(Y'_n \geq t) \to \mathbb{P}(U \geq t)$ and $\mathbb{P}(X'_n \geq t) \to \mathbb{P}(U \geq t)$ as $n \to \infty$. Namely, both upper and lower bounds on $\mathbb{P}(Z_n \geq t)$ converge to the same limit. Thus, $\lim_{n\to\infty} \mathbb{P}(Z_n \geq t) \to \mathbb{P}(U \geq t)$ as a result of the usual *squeeze theorem*. □

CHAPTER 4

Optimal Change-Point Detection

4.1 Introduction

Change-point detection is the problem of detecting an abrupt change or changes arising in a sequence of observed samples. A common problem of this type involves detecting shifts in the mean of a temporal or spatial process. This problem has found a variety of applications in many fields, including audio analysis [GER07], EEG segmentation [Lav05], structural health monitoring [NRK12, HQI07] and environment sciences [LS08, VHNC10]. Despite advances in the development of algorithms [KS09, Lav05, LYY10, Rig10] and asymptotic theory [BFG11, TRBK06, SZ10, LlH08] for a number of contexts, such studies are mainly confined to the setting of (conditionally) independently distributed data. Existing works on optimal detection of shifts in the mean in temporal data with statistically dependent observations are far less common.

Incorporating dependence structures into the modelling of random processes is a natural approach. In fact, this has been considered in detecting changes of remotely collected data [CV11, AGB05]. For instance, Chandola et al. [CV11] proposed a GP based algorithm to identify changes in Normalized Difference Vegetation Index (NDVI) time series for a particular location in California. It is therefore of interest to study how the dependence structures of the underlying process can be accounted for, e.g., its covariance function and spectral density, in designing statistically efficient detection procedures.

In this chapter, we shall focus on the detection of a single change in the mean of a GP data sequence. Consider a simplified setting in which we let G be a GP on a domain $\mathcal{D} \subseteq \mathbb{R}$ and $\mathcal{D}_n \coloneqq \{t_k\}_{k=1}^n \subset \mathcal{D}$ represent a finite index set of sampling points. Denote the observed samples by $\boldsymbol{X} = \{X_k\}_{k=1}^n$ in which $X_k = G(t_k)$ for $k = 1, \ldots, n$. Moreover, let $t \in \mathcal{C}_{n,\alpha} \subseteq \{1, \ldots, n\}$ (the parameter α is a positive scalar which will be introduced in Section 4.2.1) and $b > 0$ represents the point of sudden change and the jump/shift value, respectively.

Namely, there is $\mu \in \mathbb{R}$ (which will be assumed to be 0 for now) such that

$$\mathbb{E}X_k = \left(\mu - \frac{b}{2}\right)\mathbb{1}(k < t) + \left(\mu + \frac{b}{2}\right)\mathbb{1}(k \geq t), \quad k \in \{1, \ldots, n\}. \tag{4.1}$$

To design a detection procedure and analyze its performance as sample size n grows to infinity, one is confronted with two fundamentally different frameworks, the framework of *increasing domain asymptotics* and that of *fixed domain (infill) asymptotics*, cf., e.g., [RRR12]. The former arises naturally in time series analysis, which is distinguished by the constraint that the distance between consecutive sampling *time* points are bounded away from zero. The simplest instance of the sampling scenario in this regime arises when the diameter of $\mathcal{D}_n$ is of order n and $\min|t_{i+1} - t_i| > \epsilon$ for some strictly positive, fixed scalar ϵ. In our notation the index set for the GP represents the sampling time points. Typically we set $\mathcal{D} = \mathbb{R}$ and $\bigcup_{n=1}^{\infty} \mathcal{D}_n = \mathbb{N}$ or $\mathbb{Z}$. There is a large literature on change-point detection via the increasing domain asymptotics [AHP97, Hor97, HH12, KL98, REN09, YD86] — which we shall return to in a moment. Fixed domain (or infill) asymptotics, one the other hand, is a more suitable setting when the index set of sampling points $\mathcal{D}$ is bounded, so that the observations get denser in $\mathcal{D}$ as n increases. Particularly for $\mathcal{D} \subset \mathbb{R}$, we have that $\min|t_{i+k} - t_i| = \mathcal{O}(k/n)$ for positive integers i, k with $i, (i+k) \in \{1, \ldots, n\}$, and it can be extended to multidimensional domains in a straightforward way. The development of detection algorithms and theory for fixed domain asymptotics are relatively rare.

To gain a quick intuition on how the different asymptotic setting can affect the detection of a change in the observed sequence $\boldsymbol{X} = \{X_k\}_{k=1}^n$, one can look into the correlation among nearby samples in the sequence. In the increasing domain setting, even for long range dependent processes the correlation among samples X_i and X_j is small when $|j - i|$ is large. By contrast, in the fixed domain regime, regardless of how large the sample size is, if $|j - i|$ is of order n^β for some $\beta \in (0, 1)$, the correlation among X_i and X_j is still close to one. This entails that the effective sample size is much smaller than n. As a consequence, standard techniques that work well in the increasing domain setting do not work as well in the fixed domain setting. In the latter, we shall need more effective techniques to account for the strong dependence in the observed samples.

Previous works. An early attempt to study shift in mean detection was that of Chernoff et al. [CZ64]. More general settings of this problem have been studied in subsequent works, e.g., [Mac74, DP86, YD86]. For instance it is assumed in [YD86] that the sequence of X_k's are independent Gaussian variables. They proposed a detection method based on the Generalized Likelihood Ratio Test (GLRT), also known as the Cumulative Sum (CUSUM)

test, and given by

$$T_{CUSUM} = \mathbb{1}\left\{\max_{t\in\mathcal{C}_{n,\alpha}}\left\{\sqrt{\frac{t(n-t)}{n}}\left|\frac{1}{n-t}\sum_{k=t+1}^{n}X_k - \frac{1}{t}\sum_{k=1}^{t}X_k\right|\right\} \geq R_n\right\}. \tag{4.2}$$

CUSUM compares the maximum of a test statistic over $C_{n,\alpha}$ with a critical value R_n. Non-asymptotic upper bounds on the error probabilities of this simple test were obtained by the authors under the Gaussian and i.i.d. assumptions. Due to its simplicity, the CUSUM test is very popular, and has been applied to a variety of settings.

For example, subsequent works studied the behaviour of the CUSUM test under weaker assumptions in the increasing domain regime [AHP97, Hor97, HH12, REN09]. We wish to mention Rencova ([REN09], chapter 4), who studied the same test as [YD86], but working with the assumption that $\boldsymbol{X}$ is a strong mixing time series. Kokoszka [KL98] also analyzed the CUSUM test, but working with a different dependent observation model with sub-squared growth of the variance of partial sums, i.e., there is $\delta \in (0,2)$ such that for any $k < m$, $\text{var}\sum_{j=k}^{m} X_j \lesssim (m-k+1)^{\delta}$. Horváth et al. [Hor97, HH12] and Antoch [AHP97] studied the performance of the CUSUM test for the detection of a sudden change in the mean in linear processes, i.e. $X_t = \sum_{j=0}^{\infty} w_j \epsilon_{t-j}$, in which $\{\epsilon_t\}_{t=-\infty}^{\infty}$ are i.i.d. and zero mean random variables and the weights $\left\{w_j\right\}_{j=0}^{\infty}$ satisfy some properties such as absolute or square summability.

The CUSUM test may also be applied to one dimensional processes with correlated samples, after a proper standardization. For instance, Horváth et al. [HH12] used a different normalizing factor for applying CUSUM to one dimensional Gaussian time series with long range dependence. However apart from the standardizing factor, they do not directly incorporate the correlation structures of the data in the formulation of the test statistics. Furthermore, different forms of the CUSUM test were proposed to detect abrupt changes in the sequential detection literature, see e.g., Lai [Lai98]. At first sight, it may seem puzzling how the CUSUM test attains nearly optimal detection performance in the increasing domain even as its test statistic apparently ignore the dependence among data samples (see e.g. [AHP97, Hor97, HH12, REN09]). As noted earlier, the covariance $\text{cov}(X_s, X_t) \to 0$ as $|t-s|$ grows to infinity. As a result, the percentage of pairs $(X_s, X_t)_{s,t=1}^{n}$ whose covariance is non-negligible tends to zero as $n \to \infty$. Thus, there is not much gain in accounting for the dependence structures underlying the sequence, and so the CUSUM statistic provides a good approximation of the likelihood ratio test for large n, leading to the asymptotic optimality of T_{CUSUM} in the increasing domain setting.

One may consider applying the CUSUM test to detect a change-point in the fixed domain setting, but we will see in this chapter that the CUSUM test has suboptimal performance.

We shall consider instead a generalized likelihood ratio test and show that this achieves the asymptotically optimal rate. In comparison to the increasing domain regime, the theoretical analysis for the fixed domain setting is considerably more involved from a technical standpoint, as one needs to take into account the statistical dependence in the data sequence in a more fundamental way.

Overview of main results. Our focus is on the change-point detection problem given samples collected from a fixed and bounded domain. In particular, the data sequence is assumed to be drawn from a GP that experiences a change in the mean. We consider two scenarios for the covariance functions: fully known or with some unknown parameters. We seek to achieve the following:

(a) Given an n-sample drawn from a one dimensional GP with a known covariance structure, we propose a generalized likelihood ratio test for detecting a sudden shift in the mean. This method requires the knowledge of the dependence structure (via the covariance matrix), and will be shown to achieve asymptotically near optimal detection performance in the fixed domain setting. Our theory holds for a variety of covariance structures, such as the Matern class, powered exponential class, and several others specified in terms of the covariance kernel's spectral density. The smoothness parameter for the GP (which determines how fast the corresponding spectral density decays) plays a central role in characterizing the minimax optimal detection performance.

(b) We establish an upper bound guarantee for the CUSUM detection method. This result suggests that the CUSUM is suboptimal in the fixed domain setting. The suboptimality is confirmed in our simulation study, which exhibits a wide gap between the CUSUM and GLRT. This result makes sense, in light of the minimax result described earlier.

(c) Next, the GP covariance structure is assumed unknown. To address this scenario, we propose a Plug-in Generalized Likelihood Ratio Test (PGLRT) method, and investigate its performance. Quite remarkably, we show that as long as a consistent covariance estimate is employed (the notion of consistency will be defined in Section 4.4), regardless of its estimation rate, the PGLRT achieves asymptotically near optimal detection performance.

(d) For completeness we have also derived the performance of CUSUM and GLRT based algorithms in the increasing domain regime which confirms near minimax optimality this regime. Due to space constraints, such results are included in the Section 4.10.

The interested reader is referred to the technical report [KSN17] for a more comprehensive treatment of the increasing domain setting along with technical proofs.

In addition to studying the change-point detection problem for dependent data in the fixed domain regime and distinguishing this setting from the increasing domain regime, the work carried out in this chapter may serve as a starting point in the study of optimal detection of discontinuities in Gaussian spatial processes on domains of higher dimensions, as initiated by [AGB05, SHC02]. At a more technical level, our asymptotic analysis contain several useful proof techniques worth mentioning: they include properties of mutually orthogonal Gaussian measures, the decorrelation of samples drawn from GPs in a fixed and bounded domain, and the classical theory of minimax detection.

Structure of the chapter. Section 4.2 presents the problem of detection of a shift in the mean in an one-dimensional GP, and then introduces detection procedures for the cases that the covariance structure is known and unknown. When the covariance structure is unknown, i.e., the spectral density is given with an unknown parameter, a PGLRT will be introduced. Section 4.3 studies sufficient conditions on shift value b and spectral density to detect the existence of shift in mean with high probability. The analysis of the PGLRT and the CUSUM test in the fixed domain setting is given in Section 4.4 and Section 4.5, respectively. The minimax optimality of the proposed algorithms is established in Section 4.6. The empirical evaluation of these tests is carried out by a simulation study in Section 4.7. Section 4.8 contains the proofs of the main results and Section 4.9 presents and proves some auxiliary results used in Section 4.8. Results on the asymptotic behaviour of both CUSUM and GLRT in the increasing domain setting are stated in Section 4.10.

Notation. $\wedge$ and $\vee$ stand for minimum and maximum operators and the indicator function is represented by $\mathbb{1}(\cdot)$. For any $m \in \mathbb{N}$, I_m, $\mathbf{0}_m$ and $\mathbb{1}_m$ respectively denote the m by m identity matrix, all zeros column vector of length m, and all ones column vector of length m. For two matrices of the same size M_1 and M_2, $\langle M_1, M_2 \rangle := \sum_{i,j} (M_1)_{ij} (M_2)_{ij}$ denotes their usual inner product. For any symmetric matrix M, $\lambda_{\min}(M)$ represents the smallest eigenvalue of M. We will use the following matrix norms on $M \in \mathbb{R}^{m \times n}$. For any $1 \leq p < \infty$, $\|M\|_{\ell_p} := \left(\sum_{i,j} \left|M_{ij}\right|^p\right)^{1/p}$ stands for the element-wise ℓ_p norm of M, while $\|M\|_{\ell_\infty} := \max_{i,j} \left|M_{ij}\right|$ represents the sup norm of M. For a function $f : \mathcal{D} \mapsto \mathbb{R}$ and $p > 0$, $\|f\|_p^p := \int_{\mathcal{D}} |f(u)|^p \, du$. The special case of $p = \infty$ is defined by $\|f\|_\infty := \sup_{u \in \mathcal{D}} |f(u)|$. For any

$f \in \mathbb{L}^1(\mathbb{R})$, $\hat{f}$ represents its Fourier transform defined by

$$\hat{f}(\omega) = \int_{-\infty}^{\infty} f(t) e^{-j\omega t} dt, \quad \forall\, \omega \in \mathbb{R},$$

where $j^2 = -1$ denotes the imaginary unit. For two functions f and g on $\mathbb{R}$, we write $f(t) \asymp g(t)$ as $t \to t_0$, if $C_1 \leq \lim_{t \to t_0} \left| \frac{f(t)}{g(t)} \right| \leq C_2$ for some strictly positive bounded scalars $C_1 \leq C_2$. In particular, we write $f(t) \sim g(t)$ as $t \to t_0$ to indicate the case that $C_1 = C_2 = 1$. Furthermore, for sequences a_n and b_n, we write $b_n = \Omega(a_n)$ when b_n is bounded below by a_n asymptotically, i.e. $\lim_{n \to \infty} |b_n / a_n| \geq C$ for some positive C. In case a_n and b_n are random, $b_n = o_{\mathbb{P}}(a_n)$ means that b_n / a_n converges in probability to zero as $n \to \infty$. For two sets $\Omega_1, \Omega_2 \subset \mathbb{R}^m$, $\operatorname{dist}(\Omega_1, \Omega_2) \coloneqq \inf_{\omega_i \in \Omega_i,\, i=1,2} \|\omega_1 - \omega_2\|_{\ell_2}$ stands for their mutual distance with respect to Euclidean distance. Lastly, $\Gamma(\cdot)$ denotes the gamma function.

4.2 Change-Point Detection Procedures

In this section we describe the formulation of the shift-in-mean detection problem associated with GP data, and then present detection procedures that account for the underlying process modelling assumptions. From here on, we assume that G is a one dimensional GP in $\mathcal{D} = [0,1]$ with n regularly spaced samples, i.e. $\mathcal{D}_n = \{k/n\}_{k=1}^n$. Let the symmetric real functions $K : \mathbb{R} \mapsto \mathbb{R}$ and $\hat{K} : \mathbb{R} \mapsto \mathbb{R}$ respectively denote the covariance function and spectral density of G. Accordingly, the covariance matrix $\Sigma_n \coloneqq \operatorname{cov}\left(\{X_k\}_{k=1}^n\right)$ is a symmetric Toeplitz matrix given by

$$\Sigma_n = \{\operatorname{cov}(X_r, X_s)\}_{r,s=1}^n = \left[K\left(\frac{r-s}{n}\right) \right]_{r,s=1}^n . \tag{4.3}$$

Some regularity conditions on K will be introduced the sequel.

4.2.1 Detection Procedure Based on GLRT

The problem of detecting a sudden shift in the mean of a one dimensional GP can be formulated as a composite hypothesis testing problem. Under the null hypothesis $\mathbb{H}_0$, all the random variables have zero mean, i.e. $\mathbb{E}\boldsymbol{X} = \mathbf{0}_n$. To specify the alternative hypothesis $\mathbb{H}_1$ we first introduce a few additional notations. Let $t \in \mathcal{C}_{n,\alpha}$ denote the occurrence times of the single change-point. The set $\mathcal{C}_{n,\alpha} \subseteq \{1, \ldots, n\}$ contains plausible occurrence time of the change. Assume there is $\alpha \in (0, 1/2)$ such that $\mathcal{C}_{n,\alpha} = \{t :\ t \wedge (n-t) > \alpha n\}$. The amount of shift in the mean before and after the change-point is denoted by real parameter b. Thus,

for a fixed change-point $t \in \mathcal{C}_n$, the associated alternative hypothesis associated with t can be stated as,

$$H_{1,t} : \exists\, b \neq 0,\, \mathbb{E}\boldsymbol{X} = \frac{b}{2}\zeta_t, \tag{4.4}$$

where $\zeta_t \in \mathbb{R}^n$ is given by $\zeta_t(k) \coloneqq \operatorname{sign}(k-t)$ for any $t \in \mathcal{C}_{n,\alpha}$. We adopt the convention $\operatorname{sign}(0) = 1$. Since t is not known a priori, the alternative hypothesis is specified by taking the union of $\mathbb{H}_{1,t}$. We arrive at the following composite hypothesis testing problem:

$$\mathbb{H}_0 : \mathbb{E}\boldsymbol{X} = \mathbf{0}_n, \quad \text{v.s.} \quad \mathbb{H}_1 = \bigcup_{t \in \mathcal{C}_{n,\alpha}} \mathbb{H}_{1,t}, \text{ i.e., } \exists\, t \in \mathcal{C}_{n,\alpha},\, b \neq 0, \text{ s.t. } \mathbb{E}\boldsymbol{X} = \frac{b}{2}\zeta_t. \tag{4.5}$$

Next, we propose a test statistic which is constructed by the generalized likelihood ratio (GLR). Note that the GLR is an explicit function of the joint density of samples and so the GP assumption is essential to its calculation.

Proposition 4.1. Assuming that Σ_n is known, there exists $R_{n,\delta} > 0$ for which the GLRT is given by

$$T_{GLRT} = \mathbb{1}\left(\max_{t \in \mathcal{C}_{n,\alpha}} \left| \frac{\zeta_t^\top (\Sigma_n)^{-1} \boldsymbol{X}}{\sqrt{\zeta_t^\top (\Sigma_n)^{-1} \zeta_t}} \right|^2 \geq R_{n,\delta} \right). \tag{4.6}$$

The threshold value $R_{n,\delta}$ depends only on n and some parameter δ determining the false alarm rate. The precise form of $R_{n,\delta}$ will be presented in subsequent sections. Setting $\mu = 0$ in (4.1) results in a substantially simplified expression of the GLR, which eases the exposition of our analysis of the computational and theoretical properties of the proposed test. The general form of the GLRT, when μ is unknown, is presented as Proposition 4.2 in Section 4.8.

Unlike the CUSUM test, cf. Eq. (4.2), the covariance function of G is explicitly taken into account in the GLRT. As a result, it will be shown in the sequel that the proposed detection method is optimal, while the same cannot be said for the CUSUM test, specifically in the setting of fixed domain asymptotics.

Unknown covariance In practice, however, the covariance is not known and needs to be estimated. To address such scenarios, we propose to approximate the likelihood ratio by plugging into Eq. (4.6) a positive definite estimate of the covariance matrix, which will be denoted by $\tilde{\Sigma}_n$. This results in a PGLRT procedure.

Definition 4.1. Let $\tilde{\Sigma}_n$ be a positive definite estimate of Σ_n. The PGLRT is given by

$$\tilde{T}_{GLRT} = \mathbb{1}\left(\max_{t \in \mathcal{C}_{n,\alpha}} \left| \frac{\zeta_t^\top \left(\tilde{\Sigma}_n\right)^{-1} \boldsymbol{X}}{\sqrt{\zeta_t^\top \left(\tilde{\Sigma}_n\right)^{-1} \zeta_t}} \right|^2 \geq \tilde{R}_{n,\delta} \right), \tag{4.7}$$

for some strictly positive threshold value $\tilde{R}_{n,\delta}$.

The specific choice of $\tilde{\Sigma}_n$ and the accompanying theory will be given later in Section 4.4.

4.3 Detection Rate of GLRT: Known Σ_n

In this section we shall establish the detection rate of the GLRT in a fixed domain regime, given that Σ_n is known. We adopt the following performance measure.

Definition 4.2. For any change detection algorithm $T \in \{0,1\}$, the conditional detection error probability (CDEP) of T, which is denoted by $\varphi_n(T)$, is defined as

$$\varphi_n(T) = \mathbb{P}(T = 1 \mid \mathbb{H}_0) + \max_{t \in \mathcal{C}_{n,\alpha}} \mathbb{P}(T = 0 \mid \mathbb{H}_{1,t}).$$

In words, φ_n is the sum of the false alarm error and the worst-case misdetection error (taken over the set of possible change-point locations $\mathcal{C}_{n,\alpha}$). Clearly, CDEP hinges on the choices of $\mathcal{C}_{n,\alpha}$ – the value of φ_n increases as $\mathcal{C}_{n,\alpha}$ becomes a larger proper subset of $\{1,\ldots,n\}$. CDEP as a risk measure has been adopted for detecting abnormal clusters in a network (see, e.g., [ACCD11, BI+13]). It also provides an upper bound on the Bayesian risk measure. We refer the reader to [ABBD+10] for a comparison of CDEP and the Bayesian risk measure. Given a fixed $\delta \in (0,1)$, we will present a sufficient condition expressed in terms of shift value b, sample size n, δ, and the spectral properties of the GP such that the proposed detection procedures can guarantee that CDEP is bounded from above by δ. This can be achieved by

- first, choose the critical value $R_{n,\delta}$ so that the false alarm error $\mathbb{P}(T = 1 \mid \mathbb{H}_0) < \delta/2$;
- second, the proposed sufficient condition (in terms of b, n, δ, and the parameters encoding the dependence structure of the data) guarantees that the worse-case misdetection error rate is also upper bounded by $\delta/2$.

Recall that G is a GP defined on $\mathcal{D} = [0,1]$ whose one realization has been observed at $\mathcal{D}_n = \{k/n\}_{k=1}^n$. The covariance function and the spectral density of G are respectively denoted by K and $\hat{K}$. We study two common classes of covariance functions, one of which

admits polynomially decaying spectral densities, and the other being the Gaussian covariance function.

Assumption 4.1. K is an integrable positive definite covariance function. Moreover, there exist $\nu \in (0,\infty)$ and C_K (depending on K) so the spectral density $\hat{K}$ satisfies the following condition:

$$C_K \coloneqq \sup_{\omega\in\mathbb{R}} \left|\hat{K}(\omega)\left(1+\omega^2\right)^{\nu+\frac{1}{2}}\right| < \infty. \tag{4.8}$$

We shall always choose the largest possible ν that satisfies (4.8). It is simple to see that Assumption 4.1 holds if and only if $\hat{K}$ is bounded at the origin and $\hat{K}(\omega) \asymp \omega^{-(2\nu+1)}$ as ω tends to infinity. It is well-known that the tail behavior of $\hat{K}$ is closely linked to the smoothness of K at the origin (see, e.g., Section 2.8 of [Ste12]). The following are a few examples of common covariance functions that will be studied in this chapter.

(a) *Matern*: This class is widely used in geostatistics, and has a fairly simple explicit form of spectral density.

$$\hat{K}(\omega) = \frac{\sqrt{4\pi}\Gamma(\nu+1/2)}{\Gamma(\nu)}\sigma^2\rho^{-2\nu}\left(\frac{1}{\rho^2}+\omega^2\right)^{-(\nu+1/2)}, \tag{4.9}$$

where $\rho,\nu,\sigma \in (0,\infty)$. Regardless of the choice of ρ and σ, condition (4.8) holds for Matern spectral density with parameter ν.

(b) *Powered exponential*: Another versatile class of covariance functions is

$$K(r) = \sigma^2 \exp\left(-\left|\frac{r}{\rho}\right|^{\beta}\right) \tag{4.10}$$

for some $\beta \in (0,2)$ and $\rho,\sigma \in (0,\infty)$. Although the spectral density does not have a closed form in terms of simple functions, Lemma 4.2 shows that $\hat{K}$ admits Assumption 4.1 with $\nu = \beta/2$.

(c) *Rational spectral densities*: Rational spectral densities form a general class admitting Assumption 4.8. For any $\hat{K}$ in this class, there are two polynomials, Q_n and Q_d, with real coefficients, unit leading coefficients and $p \coloneqq \deg(Q_d) - \deg(Q_n) \in \mathbb{N}$, such that

$$\hat{K}(\omega) = \lambda\frac{|Q_n(j\omega)|^2}{|Q_d(j\omega)|^2}. \tag{4.11}$$

Moreover, we assume that Q_d has no root on the imaginary axis and λ is a strictly positive scalar. Since $K(0) < \infty$ and $\hat{K}(\omega) \asymp \omega^{-2p}$ as $\omega \to \infty$, Assumption 4.1 holds with $\nu = p - 1/2$.

(d) *Triangular*: For $\rho, \sigma \in (0, \infty)$, the covariance function and spectral density are given by

$$K(r) = \sigma^2 \left(1 - \left|\frac{r}{\rho}\right|\right)_+, \quad \hat{K}(\omega) = \frac{\rho\sigma^2}{2} \left|\operatorname{sinc}\left(\frac{\rho\omega}{2}\right)\right|^2.$$

The triangular covariance is less favorable than the aforestated cases due to the oscillatory behaviour of $\hat{K}$ (p. 31, [Ste12]). One can easily show that this covariance fulfils Assumption 4.1 with $\nu = 1/2$.

The following theorem establishes a detection error guarantee for the GLRT, provided that the GP covariance function K is known.

Theorem 4.1. Let $\delta \in (0, 1)$. Suppose that G is a real-valued GP defined on domain $\mathcal{D} = [0, 1]$ whose associated spectral density $\hat{K}$ admits Assumption 4.1 for some ν and C_K. G is regularly sampled at i/n, $i = 1, \ldots, n$. There exist $R_{n,\delta} > 0$ (depending only on n and δ), $n_0 := n_0(K)$ and a positive universal constant C such that if $n \geq n_0$ and

$$|b| \geq C n^{-\nu} \sqrt{C_K \left(1 + \frac{1}{\nu}\right) \log\left(\frac{n(1-2\alpha)}{\delta}\right)}, \tag{4.12}$$

we have

$$\varphi_n(T_{GLRT}) \leq \delta.$$

In the theorem statement, *universal constant* is used to refer to a fixed, finite positive scalar independent of n, δ, and all the covariance parameters. See Section 4.8 for the proof of Theorem 4.1. The right hand side of Eq. (4.12) provides a bound on the smallest detectable shift using the GLRT algorithm associated to the CDEP risk measure. We make several comments regarding the roles of various quantities embedded in Theorem 4.1.

(a) $R_{n,\delta}$ in Theorem 4.1 can be chosen as

$$R_{n,\delta} = 1 + 2\left[\log\left(\frac{2n(1-2\alpha)}{\delta}\right) + \sqrt{\log\left(\frac{2n(1-2\alpha)}{\delta}\right)}\right]. \tag{4.13}$$

We guarantee that CDEP is less than or equal δ by controlling the false alarm and misdetection probabilities below $\delta/2$. To gain some insight into (4.13), notice that under the null hypothesis, the test statistic in Eq. (4.6) has the same distribution as the supremum of a χ^2_1 process over $\mathcal{C}_{n,\alpha}$, which is represented by $\{\Psi(t) : t \in \mathcal{C}_{n,\alpha}\}$. For controlling the false alarm probability below $\delta/2$, $R_{n,\delta}$ needs to be chosen such that

$$\mathbb{P}\left(\max_{t \in \mathcal{C}_{n,\alpha}} \Psi(t) \geq R_{n,\delta}\right) \leq \frac{\delta}{2}.$$

The standard χ^2_1 tail inequality in [Bir01] implies that if $R_{n,\delta}$ is chosen based upon Eq. (4.13), then $\Psi(t) \leq \delta/\{2n(1-2\alpha)\}$ for any $t \in \mathcal{C}_{n,\alpha}$. Thus, the union bound inequality yields

$$\mathbb{P}\left(\sup_{t\in\mathcal{C}_{n,\alpha}} \Psi(t) \geq R_{n,\delta}\right) \leq \left|\mathcal{C}_{n,\alpha}\right| \max_{t\in\mathcal{C}_{n,\alpha}} \mathbb{P}(\Psi(t) \geq R_{n,\delta}) \leq \frac{\delta\left|\mathcal{C}_{n,\alpha}\right|}{2n(1-2\alpha)} = \frac{\delta}{2}.$$

(b) The bound on the minimal detectable shift is proportional to $\sqrt{C_K}$, as defined in (4.8). Note that C_K is determined by both low frequency and tail behaviour of the spectral density via ν. C_K is obviously linearly proportional to $\sqrt{K(0)}$ (see Eq. (4.8)), meaning that C_K also captures the notion of the standard deviation of the observations. Thus, Theorem 4.1 implicitly expresses that change detection is more challenging for GPs with larger variance.

(c) Sample size n has two opposing effects on the detection rate of the GLRT algorithm. On the one hand, $n(1-2\alpha)$ appearing in the logarithmic function, is connected to the size of alternative hypothesis which is determined by $\left|\mathcal{C}_{n,\alpha}\right| = n(1-2\alpha)$. On the other hand, the term $n^{-\nu}$ indicates the possibility of small shift detection as more observations are available.

We note that the parameter δ, the variance of observations, and sample size have similar roles in the increasing domain setting. The main difference between the two asymptotic settings is the role of the decay rate of $\hat{K}$ in the fixed domain, which is encapsulated by ν. See Section 4.10 for further details in the increasing domain regime. Note that ν is closely related to the smoothness of G with larger values of ν corresponding to a smoother GP in the mean squared sense (cf. [Ste12], Chapter 2). For smooth GPs, $G(t_0)$ can be interpolated using the observations in the vicinity of t_0 with small estimation error. This leads to a simpler shift-in-mean detection for smoother processes. More precisely, as $n \to \infty$ the lower bound on detectable b, (4.12), vanishes more rapidly for larger ν.

Remark 4.1. As an easy consequence of the theorem, we can elaborate on the asymptotic behaviour (as $n \to \infty$) of the GLRT for several specific classes of spectral densities, all of which satisfy Assumption 4.1.

(a) *Matern*: The smallest detectable jump is $|b| \asymp n^{-\nu}\sqrt{\log(n(1-2\alpha)/\delta)}$.

(b) *Powered exponential*: $\hat{K}$ admits Assumption 4.1 with $\nu = \beta/2$. Namely the smallest detectable b has the same order as $\sqrt{n^{-\beta}\log(n(1-2\alpha)/\delta)}$.

(c) *Rational spectral densities*: It has been discussed previously that $\hat{K}(\omega) \asymp |\omega|^{-2p}$ as $\omega \to \infty$ and $\nu = p - 1/2$. Thus $|b| = \Omega\left(n^{-(p-1/2)}\sqrt{\log(n(1-2\alpha)/\delta)}\right)$ guarantees CDEP to remain below δ.

(d) *Triangular*: Assumption 4.1 with $\nu = 1/2$ holds for $\hat{K}$, so the smallest detectable jump is of order $\sqrt{n^{-1}\log(n(1-2\alpha)/\delta)}$.

Remark 4.2. Let us comment on the role of α in Theorem 4.1. The dependence on α in Eq. (4.12) is logarithmic, which encodes how the size of $\mathcal{C}_{n,\alpha}$ affects the detection rate. Strictly speaking, the asymptotic behavior of the smallest detectable jump remains unchanged, regardless of how small α has been chosen (even if α tends to zero). Note that we did *not* require that α is a fixed and strictly positive scalar in the theorem. For algebraic convenience, we assume that the mean of G fluctuates around $\mu = 0$ in Eq. (4.5). The fact that we assumed μ is known is the main reason behind the trifling effect of α in the detection rate of the GLRT, as we do not need to estimate μ from the data. That is why in this particular case α can even be chosen as small as $\mathcal{O}(1/n)$. The generic form of the GLRT test for unknown μ is presented in Proposition 4.2.

Gaussian covariance function. The Gaussian covariance function is given by

$$K(r) = \sigma^2 \exp\left[-\frac{1}{2}\left(\frac{r}{\rho}\right)^2\right], \quad \hat{K}(\omega) = \rho\sigma^2\sqrt{2\pi}\exp\left[-\frac{(\rho\omega)^2}{2}\right]. \tag{4.14}$$

This is also a popular modeling choice of smooth GPs [LL+00]. Regarding this covariance, we have the following result:

Theorem 4.2. Let G be a GP on $[0,1]$ which is observed at i/n, $i = 1,\ldots,n$, whose covariance function is given by Eq. (4.14). Let $\delta \in (0,1)$. There are $R_{n,\delta} > 0$, $n_0 := n_0(\rho)$, $C_0 := C_0(\rho) > 0$ and a universal constant $C > 0$, such that if $n \geq n_0$ and

$$|b| \geq C\sqrt{\exp\left[-n\log(C_0 n)\right]\log\left(\frac{n(1-2\alpha)}{\delta}\right)}, \tag{4.15}$$

then

$$\varphi_n(T_{GLRT}) \leq \delta.$$

The proof of this result is given in Section 4.8. Because of the super-exponential decay of the Gaussian spectral density, Assumption 4.1 is actually satisfied for any $\nu > 0$. This result shows that it is possible to detect exponentially small jump size b as n increases.

Theorem 4.2, along with Theorem 4.1 confirm the intuition that smoother the GP is (larger ν in Assumption 4.1), the easier it is to detect the presence of a shift in the mean.

Remark 4.3. We conclude this section noting the difference in the detection error rate guarantee in the fixed domain regime (Theorems 4.1 and 4.2) and the analogous results in the increasing domain setting. The GLRT can detect the jumps of magnitude $\mathcal{O}\left(\sqrt{n^{-1}\log\left(n\left(1-2\alpha\right)/\delta\right)}\right)$ for large n (that is $\varphi_n(T_{GLRT}) \leq \delta$) in the increasing domain regime, regardless of the covariance structure of G. Simply put the detection rate is not affected by the dependence structure of G. By contrast, we have seen in the earlier theorems how the detection error guarantee for the GLRT in the fixed domain setting is affected by dependence structure of G in a fundamental way.

4.4 Detection Rate of PGLRT

As we have shown in Proposition 4.1, full knowledge of Σ_n is central to computing the generalized likelihood ratio. In practice, the spectral density and covariance function of G are not known a priori, and so we take a plug-in approach, approximating the GLRT by plugging in the covariance estimate $\tilde{\Sigma}_n$ (see Definition 4.1). This section serves to investigate various ways of constructing PGLRTs and assessing their detection performance. In this section we will focus only on the fixed domain setting.

We first assume that G is a Matern GP on $\mathcal{D} = [0,1]$ with unknown parameters $\eta = (\sigma,\rho)$ in a compact space Ω, and is regularly observed on $\{k/n\}_{k=1}^{n}$ (see Eq. (4.9)). We use $\tilde{\eta}_m = (\tilde{\sigma}_m,\tilde{\rho}_m)$ to indicate the estimated parameters using m regularly spaced samples in $\mathcal{D}$. We also assume that $\mathcal{C}_{n,\alpha} = \{k : \alpha n \leq k \leq (1-\alpha)n\}$. Namely, the GP is under control for a certain number of observations. The controlled samples before the sudden change, $\boldsymbol{X}_{\mathcal{B}} \coloneqq \{X_k : k \leq \alpha n\}$, will be used to estimate η. The parameter estimation stage is typically called the *burn-in* period in the literature.

It is well-known that η is not consistently estimable in the fixed domain setting when the number of the observations in $\mathcal{D}$ grows to infinity (cf., e.g., [Yin91, Zha04]). In particular, Zhang [Zha04] showed that neither σ or ρ are consistently estimable but the quantity $\sigma\rho^{-\nu}$ can be consistently estimated using MLE. The reason behind the inconsistency is the existence of a class of mutually absolutely continuous models for G which are almost surely impossible to discern by observing one realization of G. The induced measures corresponding to two Matern GPes with parameters η and η' are absolutely continuous with respect to each other whenever $\sigma\rho^{-\nu} = \sigma'\rho'^{-\nu}$. Furthermore Zhang [Zha04] showed that if one fixes ρ at an arbitrary value, then the maximum likelihood estimator for $\sigma\rho^{-\nu}$ is consistent. We shall show that despite the inconsistency in estimating η, quite remarkably,

the PGLRT exhibits an analogous performance as the GLRT with fully known covariance function, provided that the estimate of η is consistent up to its equivalence class.
It has been noted in [KS+13] that fixing $\tilde{\rho}_m$ at large values has a trifling impact on predictive performance. Due to the complicated dependence of the Matern covariance function on ρ, estimating ρ is a computationally challenging task, particularly for large data sets. Fortunately we can accelerate the whole detection procedure without estimating ρ. In fact, our PGLRT change detector $\tilde{T}_{GLRT}$ is a two stage algorithm as follows:

- *Estimation step:*

  1. Fix $\tilde{\rho}_m$ at the largest possible element in Ω. Namely, $\tilde{\rho}_m$ is a deterministic quantity given by $\tilde{\rho}_m = \max\{\rho : (\sigma,\rho) \in \Omega\}$.
  2. Estimate $\sigma\rho^{-\nu}$ given the controlled samples $\boldsymbol{X}_{\mathcal{B}}$, using any consistent procedure such as MLE [Zha04], weighted local Whittle likelihood [WLX13], or averaging quadratic variation [And10]. By *consistent* we mean the condition that $\left|\sigma\rho^{-\nu} - \tilde{\sigma}_m\tilde{\rho}_m^{-\nu}\right| \xrightarrow{\mathbb{P}} 0$ as m tends to infinity.
  3. Construct the approximated covariance matrix of $\boldsymbol{X}$, as $\tilde{\Sigma}_n = \left[K\left(\frac{r-s}{n}, \tilde{\eta}_m\right)\right]_{r,s=1}^{n}$ (here $m := \lfloor \alpha n \rfloor$).

- *Detection step:*

  1. Apply the GLRTby plugging $\tilde{\Sigma}_n$ in place of Σ_n into (4.6), as described in Definition 4.1.

The following theorem establishes the detection performance for the PGLRT.

Theorem 4.3. Let $\delta \in (0,1)$. Let G be GP whose associated spectral density $\hat{K}$ has Matern form with unknown parameters $(\sigma,\rho) \in \Omega$, and sampled in $\{k/n\}$, $k = 1,\dots,n$. There are finite scalar $C, n_0 \in \mathbb{N}$, a non-negative sequence $\lim_{m\to\infty} \tau_m = 0$, and threshold level $\tilde{R}_{n,\delta} > 0$ such that for any $n \geq n_0$,

$$\varphi_n\left(\tilde{T}_{GLRT}\right) \leq \delta + 2\tau_m,$$

whenever

$$|b| \geq Cn^{-\nu}\sqrt{C_K\left(1+\frac{1}{\nu}\right)\log\left(\frac{n(1-2\alpha)}{\delta}\right)}. \tag{4.16}$$

See Section 4.8 for the proof of Theorem 4.3.

Remark 4.4. The threshold value for the PGLRT is chosen exactly the same as in Theorem 4.1;

$$\tilde{R}_{n,\delta} = \left[1 + 2\left(\log\left(\frac{2n(1-2\alpha)}{\delta}\right) + \sqrt{\log\left(\frac{2n(1-2\alpha)}{\delta}\right)}\right)\right]. \tag{4.17}$$

Since $\{\tau_k\}_{k\in\mathbb{N}}$ is a vanishing sequence and $m = \lfloor \alpha n \rfloor$ is an increasing function of n, $(\delta + 2\tau_m)$ lies in the vicinity of δ for large n. The most interesting aspect of Theorem 4.3 is that if some consistent estimate of $\sigma\rho^{-\nu}$ is available, the PGLRT has asymptotically the same rate as the GLRT with fully known covariance function, regardless of how efficient the point estimate for $\sigma\rho^{-\nu}$ is. The main asymptotic cost to pay for (potentially) mis-specifying ρ and σ is that that constant C appearing in Theorem 4.3 is larger than the constant C of Theorem 4.1.

We conclude this section by studying the performance of the PGLRT when both variance and range parameters are consistently estimable. Suppose that G has a powered exponential covariance function, introduced in Eq. (4.10). Anderes [And10] introduced a consistent estimate of covariance parameters using empirical average of the quadratic variation of G. According to Theorem 5 of [And10], unlike the Matern class, both σ_0 and ρ_0 are consistently estimable when $\beta \in (0, 1/2)$. Namely, $|\rho - \tilde{\rho}_m| \vee |\sigma - \tilde{\sigma}_m| \xrightarrow{\mathbb{P}} 0$, for the method introduced in [And10]. The following result, which has a similar flavor as Theorem 4.3 and can be substantiated in a similar way, determines the detection rate of the PGLRT for one-dimensional powered exponential GPs.

Theorem 4.4. Let $\delta \in (0,1)$. Let G be a GP with powered exponential covariance function with unknown parameters (σ,ρ) and known $\beta \in (0,1/2)$, sampled in $\{k/n\}$, $k = 1,\dots,n$. Given a consistent estimate of (σ,ρ) (e.g., the method in [And10]), there are finite scalars $n_0 \in \mathbb{N}$ and C (which depends on the covariance parameters β, σ and ρ), and a non-negative sequence $\lim_{m\to\infty} \tau_m = 0$, such that for any $n \geq n_0$,

$$\varphi_n\left(\tilde{T}_{GLRT}\right) \leq \delta + 2\tau_m,$$

whenever

$$|b| \geq C\sqrt{n^{-\beta}\log\left(\frac{n(1-2\alpha)}{\delta}\right)},$$

and

$$\tilde{R}_{n,\delta} = \left[1 + 2\left(\log\left(\frac{2n(1-2\alpha)}{\delta}\right) + \sqrt{\log\left(\frac{2n(1-2\alpha)}{\delta}\right)}\right)\right].$$

Theorem 4.4 states that given a consistent estimate of $\eta = (\sigma_0, \rho_0)$, the PGLRT procedure has the same asymptotic behavior as the GLRT with fully known parameters (see part (b) of Remark 4.1 for the detection rate of the GLRT with known σ_0 and ρ_0).

4.5 Detection Rate of CUSUM

In this section we revisit the classical CUSUM test and obtain its detection rate in the fixed domain setting. This result should be contrasted with our earlier theorems on the performance of the proposed exact and PGLRTs, and highlights the need for accounting for the dependence structures underlying the data. Theorem 4.5 introduces sufficient conditions on $|b|$ under which CUSUM can distinguish null and alternative hypotheses with high probability.

Theorem 4.5. Suppose that $\|K\|_1 < \infty$ and $\left\|\hat{K}'\right\|_\infty < \infty$. Moreover let $\delta \in (0,1)$, $\alpha \in (0, 1/2)$ and $\mathcal{C}_{n,\alpha} = [\alpha n, (1-\alpha)n] \cap \mathbb{N}$. There are $R_{n,\delta} > 0$, and $n_0 := n_0(\delta, \alpha)$ such that if $n \geq n_0$ and

$$|b| \geq 4\sqrt{\frac{\log\left(\frac{2n(1-2\alpha)}{\delta}\right)}{\alpha(1-\alpha)}}, \tag{4.18}$$

then,

$$\varphi_n(T_{CUSUM}) \leq \delta.$$

The proof of this theorem is deferred to Section 4.3. The risk of fixed domain-CUSUM has been controlled from above under mild conditions on K, which holds true for all the examples of covariance functions considered in this chapter. Due to the following inequality, K satisfies the assumptions in Theorem 4.5 if $a(r) := rK(r)$ is absolutely integrable:

$$\left\|\hat{K}'\right\|_\infty = \sup_{\omega \in \mathbb{R}} \left| \int_{-\infty}^{\infty} a(r) e^{-j\omega r} dr \right| \leq \int_{-\infty}^{\infty} |rK(r)| \, dr.$$

The main feature of the above theorem is the sufficient condition that the jump size increases (at the order of $\log n$ at least) in order to have an upper bound guarantee on the detection error. Although we do not have a definitive proof that this sufficient condition is also necessary, the theorem suggests that the CUSUM test is *inconsistent* in the fixed domain setting: the detection error may not vanish as data sample size increases, when the jump size is a constant. This statement is in fact verified by a careful simulation study. By contrast, we have shown earlier that using the GLRT, we can guarantee vanishing detection error as long as the jump size is either constant or (better yet) bounded from below by a

suitable vanishing term.

Remark 4.5. Let us give a qualitative argument for the inconsistency of the CUSUM test in the fixed domain setting. Suppose that b tends to zero as $n \to \infty$. Define

$$U_t := \sqrt{\frac{t(n-t)}{n}}\left[\frac{1}{n-t}\sum_{k=t+1}^{n} X_k - \frac{1}{t}\sum_{k=1}^{t} X_k\right].$$

The expected value of U_t is zero, under the null hypothesis and for any t. Regardless of the existence of a shift in the mean, the standard deviation of U_t remains the same. A careful look at the proof of Theorem 4.5 reveals that the smallest value of the standard deviation of U_t over $t \in \mathcal{C}_{n,\alpha}$ is order $\sqrt{n}$. Moreover, if there is a shift in the mean occuring at the change-point $\bar{t} \in \mathcal{C}_{n,\alpha}$, then the expected value of $U_{\bar{t}}$ is given by $b\sqrt{\bar{t}(n-\bar{t})/n} = \mathcal{O}\left(b\sqrt{n}\right)$ (Recall that $\alpha n \leq \bar{t} \leq (1-\alpha)n$). Generally speaking, as the mean of U_t under the null hypothesis, denoted by $\mathbb{E}(U_t \mid \mathbb{H}_0)$, is zero for any $t \in \mathcal{C}_{n,\alpha}$, the CUSUM test cannot distinguish between the null and the alternative (even for large sample size), since

$$\left|\frac{\mathbb{E}(U_{\bar{t}} \mid \mathbb{H}_1)}{\sqrt{\mathrm{var}(U_{\bar{t}})}}\right| = \mathcal{O}\left(\frac{b\sqrt{n}}{\sqrt{n}}\right) = \mathcal{O}(b) \to 0, \quad \text{as } n \nearrow \infty.$$

Here, $\mathbb{E}(U_{\bar{t}} \mid \mathbb{H}_1)$ represents the expected value of U_t under the alternative. This suggests that regardless of the sample size, the CUSUM test cannot detect the existence of a small shift in the mean in the fixed domain setting.

Remark 4.6. The threshold value of the CUSUM test in Theorem 4.5 is given by

$$R_{n,\delta} = \sqrt{n\left(1 + 2\log\left(\frac{2n(1-2\alpha)}{\delta}\right) + 2\sqrt{\log\left(\frac{2n(1-2\alpha)}{\delta}\right)}\right)}.$$

This threshold has a different form of dependence on n than that of the threshold of the GLRT in Eq. (4.13), since, unlike the GLRT, the CUSUM test does not reduce the correlation among the samples. In order to remove the gap between the threshold of GLRT and CUSUM in the fixed domain setting, we further normalize U_t by considering $U_n^\star = U_n/\sqrt{n}$. Equivalently, CUSUM test in this regime can be written as

$$T_{CUSUM} = \mathbb{1}\left(\max_{t \in \mathcal{C}_{n,\alpha}} \left|U_n^\star\right|^2 > R_{n,\delta}^\star := \frac{R_{n,\delta}^2}{n}\right).$$

Here, $R_{n,\delta}^\star$ is exactly the same as the critical value of the GLRT.

4.6 Minimax Lower Bound on Detection Rate

In this section, we establish minimax lower bounds on the detectable jump in the mean of G in the fixed domain regime. Theorem 4.6 shows that the obtained rate for the PGLRT (Theorem 4.3) is asymptotically near-optimal in a minimax sense. This result is applicable for rational spectral densities. First, let us formalize the notion of asymptotic (near)-optimality.

Definition 4.3. Given n samples, let $T \in \{0,1\}$ be a shift-in-mean detection algorithm whose CDEP is denoted by $\varphi_n(T)$. T is said to be asymptotically near-optimal in a minimax sense if, for any $\delta \in (0,2)$, there is sequences $\{h_n\}_{n=1}^{\infty}$ dependent on n, δ and the spectral density, such that

1. As $n \to \infty$, T can detect the existence of any abrupt change with the CDEP guarantee $\varphi_n(T) \leq \delta$, provided that the jump size b satisfies $h_n \log n = o(b)$.

2. There is a large enough n_0 (depending on the model parameters) such that if $n \geq n_0$ and $|b| \leq h_n$, then there is no algorithm whose CDEP is strictly less than δ.

Recall the fixed domain regime in which G is a GP defined on $[0,1]$ and is observed at $\{i/n\}_{i=1}^{n}$. We formally introduce a suitable class of spectral densities that we consider in this section. While somewhat more restrictive than Assumption 4.1, it still provides a sufficiently rich class of commonly used spectral densities.

Assumption 4.2. There are constants $p \in \mathbb{N}$ and $\beta \in (1/2, \infty)$ such that

1. $\lim_{\omega\to\infty} \hat{K}(\omega)|\omega|^{2p}$ exists and $C'_K := \lim_{\omega\to\infty} \hat{K}(\omega)|\omega|^{2p} \in (0,\infty)$.

2. $\limsup_{\omega\to\infty} \left| \left(\frac{\hat{K}(\omega)|\omega|^{2p}}{C'_K} - 1 \right) \omega^{\beta} \right| < \infty$.

Generally speaking, Assumption 4.2 contains the class of spectral densities $\hat{K}(\omega)$ for which there is some $p \in \mathbb{N}$ such that $\hat{K}(\omega) \asymp |\omega|^{-2p}$ as ω tends to infinity. Note that the second condition in Assumption 4.2 is of theoretical purposes and does not have a simple qualitative interpretation. It can be observed that Assumption 4.2 excludes any $\hat{K}(\omega)$ satisfying Assumption 4.1 with $(\nu + 1/2) \notin \mathbb{N}$. For instance, Assumption 4.2 does not hold for Matern covariance functions with $(\nu + 1/2) \notin \mathbb{N}$.

Remark 4.7. Here, we name a salient class of spectral densities satisfying Assumption 4.2.

- Simple calculations show that any rational spectral density $\hat{K}$ (See (4.11)) admits Assumption 4.2 with $C'_K = \lambda$, $\beta = 1$ and $p = \deg(Q_d) - \deg(Q_n) \in \mathbb{N}$. Moreover, $\hat{K}$

satisfies Assumption 4.1 with $\nu = p - 1/2$. Indeed the Matern covariance function with $p := (\nu + 1/2) \in \mathbb{N}$ has a rational spectral density. These particular instances of Matern covariance, which are commonly used in machine learning and geostatistics, are of the form $K(r) = Q(|r|)e^{-d|r|}$, where $Q(\cdot)$ is a polynomial of degree $p-1$.

Theorem 4.6. Let $\delta \in (0,2)$ and assume that Assumption 4.2 holds for K. Consider the change-point detection problem (4.5) in which $\mathrm{cov}(\boldsymbol{X}) = \left[K\left(\frac{r-s}{n}\right)\right]_{r,s=1}^{n}$. There are positive scalars $\bar{C}_K$ and $n_0 := n_0(K)$ such that if $n \geq n_0$ and

$$|b| \leq \bar{C}_K n^{-p+1/2} \sqrt{\log\left(\frac{1}{\delta(2-\delta)}\right)}, \tag{4.19}$$

then for any test T,

$$\varphi_n(T) \geq \delta.$$

See Section 4.8 for the proof of Theorem 4.6.

Remark 4.8. Comparing the detection rate of the GLRT (see Theorem 4.1) and PGLRT (see Theorem 4.3), with the rate established in Eq. (4.19) entails the asymptotically near optimality of the GLRT with known covariance structure and the PGLRT for the class of spectral densities considered in Remark 4.7. Strictly speaking, under the fixed domain setting, there is a gap of order $\sqrt{\log n}$ between (4.19) and the detection rate of the GLRT based algorithms. Although we do not have a proof to establish the asymptotic near optimality of the GLRT and PGLRT for the broader class of spectral densities admitting Assumption 4.1, our conjecture is that Theorem 4.6 can be extended to this broader class.

4.7 Simulation Study

To illustrate the performance of the proposed shift-in-mean detection algorithms, we conduct a set of controlled simulation studies for verifying the results in Sections 4.3, 4.4 and 4.5. We also present other simulation studies to assess the performance of CUSUM and GLRT in the increasing domain regime. Our goals are two-fold:

(a) comparing the performance of the GLRT based algorithms with the standard CUSUM test in the two asymptotic frameworks.

(b) assessing the sensitivity of algorithm (4.7) to the parameters of the covariance function.

In all the numerical studies in this section we fix $n = 500$ and $\alpha = 0.1$.
The area under the receiver operating characteristic (ROC) curve, which will be referred as Area Under Curve (AUC), is a standard way for assessing the performance of a test. The ROC curve plots the power against the false alarm probability. Since the ROC curve is confined in the unit square, the AUC ranges in [0,1]. The ROC curve of a test based on pure random guessing is the diagonal line between origin and $(1,1)$ and so the AUC of any realistic test is at least 0.5.
The subsequent figures in this section exhibit empirical AUC versus b. For a fixed value of b, covariance function K and a detection algorithm T, we apply the following method to compute the AUC of T:

1. Set $T_1 = 500$ and $T_2 = 50$.

2. For $k = 1$ to T_2 repeat independently

   (a) For $\ell = 1$ to T_1 repeat independently

      i. Choose $p \in \{0,1\}$ with equal probability which denotes null or alternative hypotheses. Thus, approximately $T_1/2 = 250$ experiments correspond to both null and alternative.

      ii. If $p = 0$, generate zero mean $\boldsymbol{X} \in \mathbb{R}^n$ according to covariance function K. That is, $\boldsymbol{X}$ are sampled from a GP with no abrupt shift in mean. Otherwise, choose $t \in [\alpha n, (1-\alpha)n] = \{50, 51, \cdots, 450\}$ uniformly at random (recall that t represents the location of the mean shift) and generate $\boldsymbol{X} \in \mathbb{R}^n$ according to $\mathbb{H}_{1,t}$.

      iii. Compute T score.

   (b) Numerically obtain the ROC curve of T based upon T_1 experiments in part *i*.

   (c) Given the ROC curve, compute AUC_k using trapezoidal integration method.

3. Compute the average AUC by $\overline{AUC} = \frac{1}{T_2} \sum_{k=1}^{T_2} AUC_k$.

The first simulation study aims to compare CUSUM and GLRT based algorithms in the fixed domain regime and assess the role of smoothness and other parameters of K in the performance of the GLRT. For this experiment G is a GP in $[0,1]$ which is observed at regularly spaced samples, $\mathcal{D}_n = \{k/n\}_{k=1}^n$, i.e., $X_k = G(k/n)$, $k = 1, \ldots, n$. The covariance

function of G is assumed to have Matern form with parameters (σ_0, ρ_0, ν). Strictly speaking,

$$\operatorname{cov}(X_i, X_l) = \sigma_0^2 K_\nu\left(\left|\frac{i-l}{n\rho_0}\right|\right), \quad i, l = 1, \ldots, n,$$
$$K_\nu(x) = \frac{\sqrt{4\pi}\Gamma(\nu + 1/2)}{\Gamma(\nu)} \int_{-\infty}^{\infty} e^{-j\omega x}\left(1+\omega^2\right)^{-(\nu+1/2)} d\omega, \quad \forall x \geq 0.$$

We consider $\nu = 0.5, 1$, and 1.5. We also set $\sigma_0 = 1$ and $\rho_0 = 1/2$. As customary in the literature, we assume that ν is known and so ν will not be estimated. For conducting the PGLRT procedure, both parameters (σ_0, ρ_0) are estimated using the MLE. Due to the low dimensionality of the unknown parameters, the most effective way to estimate (σ_0, ρ_0) is to apply a brute force grid search over a pre-specified set $\mathcal{P}$. Here, we choose $\mathcal{P} = \{0.2, 0.4, \cdots, 2\} \times \{1/4, 1/3.9, \ldots, 1/0.1\}$. The final results are exhibited in Figure 4.1. We observe the following:

- GLRT and PGLRT have a significantly better detection performance than CUSUM. This performance improvement is more pronounced for smoother covariance functions (larger ν). In particular, the CUSUM test is completely impractical for detection of a small change when $\nu = 1$ or 1.5.

- In each panel of Figure 4.1, the GLRT has a slightly larger AUC than that of the PGLRT suggesting a small gap between the smallest detectable jump of these two scenarios. Although the two GLRTs have the same rate, this gap is likely accounted for by the differing constants in Eqs. (4.12) and (4.16). In short, having full knowledge of the covariance parameters slightly improves the detection performance and so our proposed algorithm is robust to the estimation error of the unknown parameters of K.

- Comparing the range of b in each panel of Figure 4.1 discloses that more rapid decay of the spectral density can decrease the smallest detectable jump. This observation substantiates the role of ν in the theory established in Sections 4.3 and 4.4.

Next, we compare the performance of the GLRT with known parameters and the CUSUM in the increasing domain setting. We have concisely discussed that the two methods have analogous asymptotic rates (see Section 4.60 for further details). In the left panel of Figure 4.2, we choose an exponentially decaying covariance function

$$\operatorname{cov}(X_i, X_l) = \sigma_0^2 \exp\left(-\frac{|i-l|}{\rho_0}\right), \quad i, l = 1, \ldots, n,$$

in which $\sigma_0 = 1$ and $\rho_0 = 2$. That is Σ_n has exponentially decaying off-diagonal entries. However, in the right panel, the chosen covariance function has a polynomially decaying tail given by

$$\operatorname{cov}(X_i, X_l) = \sigma_0^2 \left(1 + \frac{|i-l|}{\rho_0}\right)^{-(1+\lambda)},$$

with $\sigma_0 = 1$, $\rho_0 = 2$ and $\lambda = 0.5$. In this case, Σ_n has heavier off-diagonal terms. It is evident from Figure 4.2 that the GLRT exhibits a slightly better performance than the CUSUM, and the gap between the two AUC curves is more visible in the case of polynomially decaying covariance function. Thus, we still recommend the use of GLRT in the presence of strong correlation among samples in applications described by the increasing domain regime.

4.8 Proof of the Main Results

Proof of Proposition 4.1. In the following $\mathfrak{L}$ stands for the generalized negative log-likelihood ratio.

$$2\mathfrak{L} = \boldsymbol{X}^\top (\Sigma_n)^{-1} \boldsymbol{X} - \min_{t \in \mathcal{C}_{n,\alpha}} \min_{b \neq 0} \left[\left(\boldsymbol{X} - \frac{b}{2}\zeta_t \right)^\top (\Sigma_n)^{-1} \left(\boldsymbol{X} - \frac{b}{2}\zeta_t \right) \right]. \tag{4.20}$$

Note that the objective function in (4.20) is quadratic in terms of b. The explicit form of $2\mathfrak{L}$ can be obtained with a bit of algebraic derivations. The algebra has been skipped to save space; we arrive at

$$2\mathfrak{L} = \max_{t \in \mathcal{C}_{n,\alpha}} \max_{b \neq 0} \left(-\frac{\zeta_t^\top (\Sigma_n)^{-1} \zeta_t}{4} b^2 + b \zeta_t^\top (\Sigma_n)^{-1} \boldsymbol{X} \right) = \max_{t \in \mathcal{C}_{n,\alpha}} \left| \frac{\zeta_t^\top (\Sigma_n)^{-1} \boldsymbol{X}}{\sqrt{\zeta_t^\top (\Sigma_n)^{-1} \zeta_t}} \right|^2.$$

So, there is a threshold value, $R_{n,\delta} > 0$, for which the MLE is given by (4.6). □

The following result expressing the form of MLE in the generic case of unknown μ can be proved in an analogous way as Proposition 4.1.

Proposition 4.2. There is $R_{n,\delta} > 0$ for which the MLE is given by

$$T_{GLRT} = \mathbb{1}\left(\max_{t \in \mathcal{C}_{n,\alpha}} \left| \frac{\langle \boldsymbol{Y}, \zeta_t - B_1(t)\, \mathbb{1}_n \rangle}{\sqrt{B_2(t)}} \right|^2 \geq R_{n,\delta} \right), \tag{4.21}$$

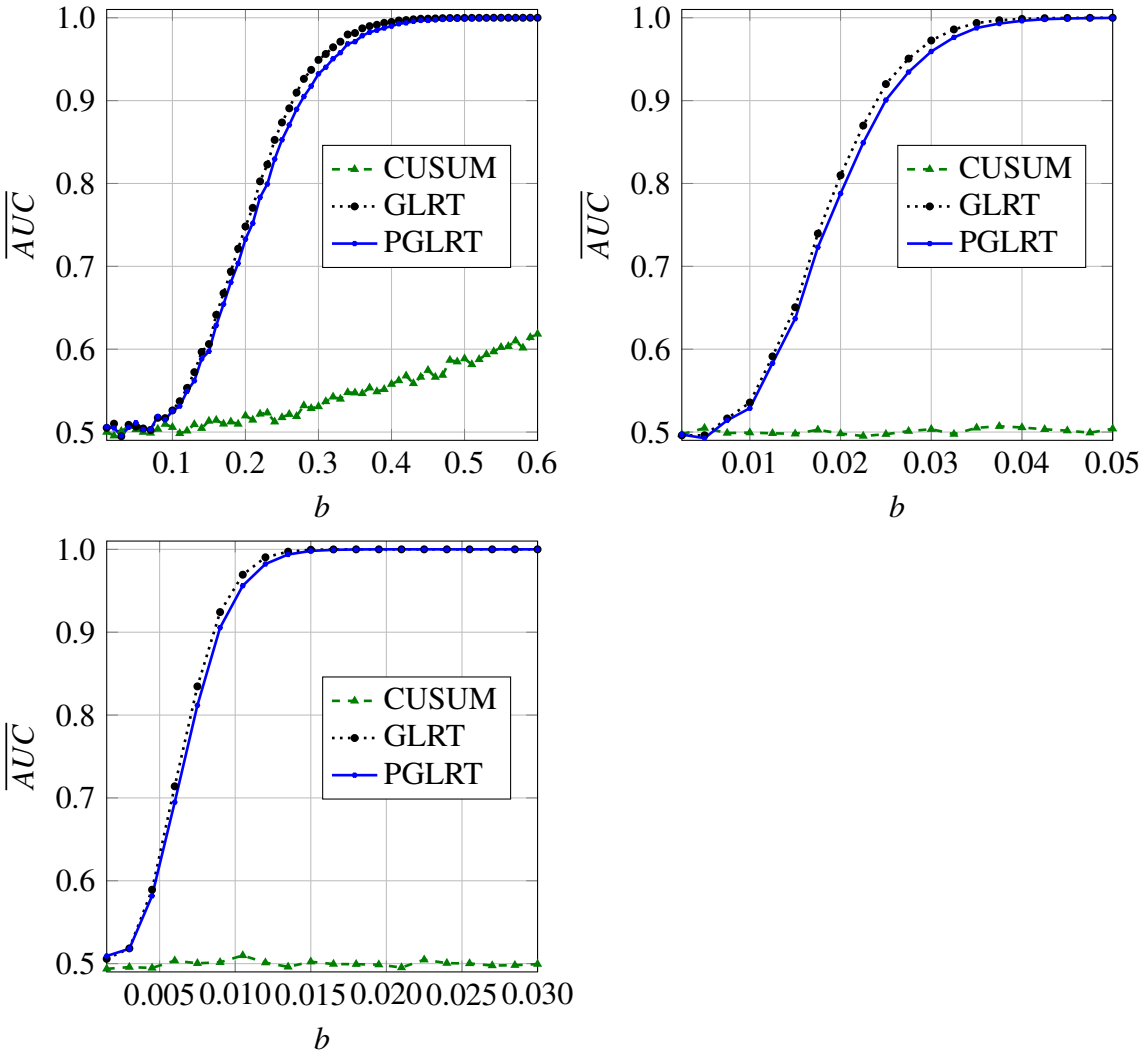

Figure 4.1: The above figures assess the performance of different detection algorithms when G is a one dimensional Matern GP, with parameters (ν, σ_0, ρ_0), and regularly sampled in $[0,1]$. From left to right then from top to bottom, $(\nu, \sigma_0, \rho_0) = (0.5,1,0.5), (1,1,0.5)$, and $(1.5,1,0.5)$. In each panel the horizontal axis displays b and the three curves (dashed black, solid blue and green) respectively exhibit the AUC of the GLRT with known covariance structure, PGLRT using full MLE and CUSUM.

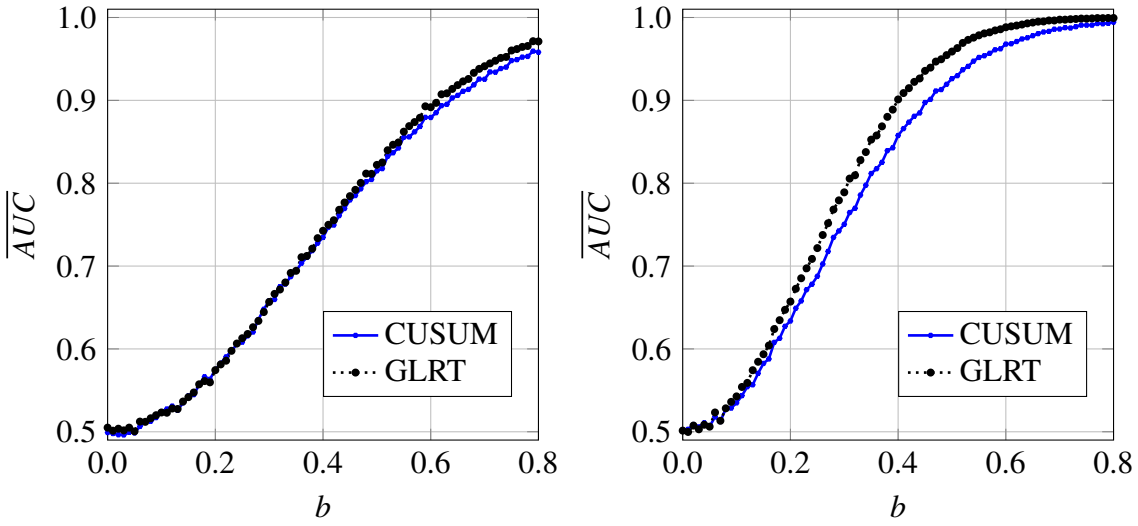

Figure 4.2: The above figure assesses the performance of increasing domain detection algorithms. In each panel the horizontal axis displays b and the two curves (dashed black and solid blue) respectively exhibit the AUC of the GLRT with known covariance structure and CUSUM. In the right panel, we choose $\mathrm{cov}(X_i, X_l) = \sigma_0^2(1+|i-l|/\rho_0)^{-(1+\lambda)}$ in which $(\sigma_0, \rho_0) = (1,2)$ and $\lambda = 0.5$. For the left panel, the covariance function is given by $\mathrm{cov}(X_i, X_l) = \sigma_0^2 \exp(-|i-l|/\rho_0)$ where $(\sigma_0, \rho_0) = (1,2)$.

where $\boldsymbol{Y} = (\Sigma_n)^{-1}\boldsymbol{X}$ and

$$B_1(t) = \frac{\zeta_t^\top (\Sigma_n)^{-1}\mathbb{1}_n}{\mathbb{1}_n^\top (\Sigma_n)^{-1}\mathbb{1}_n}, \quad B_2(t) = \zeta_t^\top (\Sigma_n)^{-1}\zeta_t - \frac{\left(\zeta_t^\top (\Sigma_n)^{-1}\mathbb{1}_n\right)^2}{\mathbb{1}_n^\top (\Sigma_n)^{-1}\mathbb{1}_n}.$$

Proof of Theorem 4.1. Let $p = \lceil \nu + 1/2 \rceil$, $P = \{1, \ldots, p\}$ and $\theta_n = \exp(-1/n)$. Construct a banded triangular matrix $A_n \in \mathbb{R}^{n\times n}$ by the following procedure.

$$A_n[k, k-j] = \binom{p}{j}(-\theta_n)^j, \; j \in \{0, \ldots, p\}, \; k \in \{p+1, \ldots, n\},$$
$$(A_n)_{P,P} = n^{-2\nu} I_p.$$

It is relatively simple to verify that A_n is invertible. In addition, for brevity let $Z_t = \frac{\zeta_t^\top (\Sigma_n)^{-1}\boldsymbol{X}}{\sqrt{\zeta_t^\top (\Sigma_n)^{-1}\zeta_t}}$ for any $t \in \mathcal{C}_{n,\alpha}$, in which ζ_t has been defined in (4.4). Lastly, define $U_{n,t} := A_n\zeta_t \in \mathbb{R}^n$, $W := A_n\boldsymbol{X}$ and $D_n := \mathrm{cov}(W)$.

Easy calculations show that under the null hypothesis $\{Z_t\}_{t\in\mathcal{C}_{n,\alpha}}$ is a set of standard Gaussian random variables and so, by Lemma 4.1, we have $\mathbb{P}\left(\max_{t\in\mathcal{C}_{n,\alpha}} Z_t^2 \geq R_{n,\delta}\right) \leq \delta/2$. That is, the false alarm probability is less than $\delta/2$. Moreover if the alternative hypothesis $\mathbb{H}_{1,\tilde{t}}$ (for some $\tilde{t} \in \mathcal{C}_{n,\alpha}$) holds then $\left\{Z_t^2\right\}_{t\in\mathcal{C}_{n,\alpha}}$ are non-central χ_1^2 random variables and the non-

centrality parameter of $Z_{\tilde{t}}^2$ is given by

$$\mathbb{E}\left(Z_{\tilde{t}} \mid \mathbb{H}_{1,\tilde{t}}\right) = \frac{|b|}{2}\sqrt{\zeta_{\tilde{t}}^\top (\Sigma_n)^{-1}\,\zeta_{\tilde{t}}}.$$

Applying Lemma 4.1 ($\sigma_0 = \sigma_k = 1$ for any k) demonstrates that $\varphi_n(T_2) \le \delta$, whenever

$$|b|\sqrt{\zeta_{\tilde{t}}^\top (\Sigma_n)^{-1}\,\zeta_{\tilde{t}}} \ge |b| \min_{t\in\mathcal{C}_{n,\alpha}} \sqrt{\zeta_t^\top (\Sigma_n)^{-1}\,\zeta_t} \ge 8\sqrt{\log\left(\frac{4n}{\delta}\right)}. \tag{4.22}$$

Thus, in order to get a sufficient condition on detectable b, it suffices to find a tight uniform lower bound on $\zeta_t^\top (\Sigma_n)^{-1}\,\zeta_t$ for $t \in \mathcal{C}_{n,\alpha}$.

The identity $\Sigma_n^{-1} = A_n^\top (D_n)^{-1} A_n$ can be shown using the linearity of covariance operator and non-singularity of A. Choose $t \in \mathcal{C}_{n,\alpha}$ in an arbitrary way. As a result of this alternative representation of Σ_n^{-1}, we have $\zeta_t^\top (\Sigma_n)^{-1}\,\zeta_t = U_{n,t}^\top (D_n)^{-1}\,U_{n,t}$. Applying *Kantorovich* inequality (cf. Section 4.9) and the triangle inequality yields

$$\zeta_t^\top (\Sigma_n)^{-1}\,\zeta_t = U_{n,t}^\top (D_n)^{-1}\,U_{n,t} \ge \frac{\left\|U_{n,t}\right\|_{\ell_2}^4}{U_{n,t}^\top D_n U_{n,t}} \ge \left[\frac{\left\|U_{n,t}\right\|_{\ell_2}^2}{\left\|U_{n,t}\right\|_{\ell_1}}\right]^2 \frac{1}{\|D_n\|_{\ell_\infty}}. \tag{4.23}$$

Now, we show that $\frac{\|U_{n,t}\|_{\ell_2}^2}{\|U_{n,t}\|_{\ell_1}} \ge \frac{1}{3}$, for large enough n. Indeed, after some algebra, we can get

$$\begin{aligned}
\left\|U_{n,t}\right\|_{\ell_2}^2 &\ge \sum_{k=t+1}^{t+p} U_{n,t}^2(k) = \sum_{k=1}^{p}\left[-(1-\theta_n)^p + 2\sum_{j=0}^{k-1}\binom{p}{j}(-\theta_n)^j\right]^2 \\
&\overset{(a)}{\ge} 2\sum_{k=1}^{p}\left[\sum_{j=0}^{k-1}\binom{p}{j}(-1)^j\right]^2 = 2\sum_{k=1}^{p}\left[\binom{p-1}{k-1}(-1)^{k-1}\right]^2 \\
&= 2\binom{2(p-1)}{p-1} \ge 2^p,
\end{aligned} \tag{4.24}$$

where inequality (a) follows from the fact that for large enough n, θ_n is arbitrarily close to

1. To get an upper bound on $\left\|U_{n,t}\right\|_{\ell_1}$,

$$\begin{aligned}
\left\|U_{n,t}\right\|_{\ell_1} &= \sum_{k=1}^{p}\left|U_{n,t}(k)\right| + \sum_{k=p+1}^{t}\left|U_{n,t}(k)\right| + \sum_{k=t+p+1}^{n}\left|U_{n,t}(k)\right| + \sum_{k=t+1}^{t+p}\left|U_{n,t}(k)\right| \\
&= \sum_{k=1}^{p} n^{-2\nu} + \sum_{k=p+1}^{t}(1-\theta_n)^p + \sum_{k=t+p+1}^{n}(1-\theta_n)^p \\
&+ \sum_{k=1}^{p}\left|-(1-\theta_n)^p + 2\sum_{j=0}^{k-1}\binom{p}{j}(-\theta_n)^j\right| \leq pn^{-2\nu} + n^{1-p} \\
&+ 2\sum_{k=1}^{p}\left|\sum_{j=0}^{k-1}\binom{p}{j}(-\theta_n)^j\right| \overset{(b)}{\leq} 2 + 2\sum_{k=1}^{p}\left|\sum_{j=0}^{k-1}\binom{p}{j}(-\theta_n)^j\right| \\
&\leq 2 + 4\sum_{k=1}^{p}\left|\sum_{j=0}^{k-1}\binom{p}{j}(-1)^j\right| = 2 + 4\sum_{k=1}^{p}\left|\binom{p-1}{k-1}(-1)^{k-1}\right| \\
&= 2 + 2^{p+1} \leq 3\,2^p. \qquad (4.25)
\end{aligned}$$

Note that inequality (b) is valid when $pn^{-2\nu} + n^{1-p} \leq 2$, which obviously holds for sufficiently large $n = \mathcal{O}(1)$. The remaining inequalities and identities in (4.25) can be easily verified via basic properties of the binomial coefficients. Combining (4.24) and (4.25) yields the desired goal. Now, inequality (4.23) can be rewritten as

$$\zeta_t^\top (\Sigma_n)^{-1}\zeta_t \geq \frac{1}{9\left\|D_n\right\|_{\ell_\infty}} = \left[9\max_{1\leq k\leq n}\operatorname{var}(W_k)\right]^{-1}. \qquad (4.26)$$

In the final phase of the proof, we achieve a tight upper bound on $\max_{1\leq k\leq n}\operatorname{var}(W_k)$. It is obvious from the formulation of A_n and the stationarity of $\boldsymbol{X} - \mathbb{E}\boldsymbol{X}$ that $\max_{1\leq k\leq n}\operatorname{var}(W_k) =$

$n^{-2\nu} \vee \mathrm{var}\Big(W_{p+1}\Big)$. So, the goal is reduced to give an upper bound on the variance of W_{p+1}.

$$
\begin{aligned}
\mathrm{var}\Big(W_{p+1}\Big) &= \mathrm{var}\left(\sum_{r=0}^{p}\binom{p}{r}(-\theta_n)^r X_{p+1-r}\right) \\
&\overset{(c)}{=} \frac{1}{2\pi}\int_{\mathbb{R}} \hat{K}(\omega)\left|\sum_{r=0}^{p}\binom{p}{r}(-\theta_n)^r \exp\left(\frac{-jr\omega}{n}\right)\right|^2 d\omega \\
&= \frac{1}{2\pi}\int_{\mathbb{R}} \hat{K}(\omega)\left|\sum_{r=0}^{p}\binom{p}{r}\left(-\exp\left(\frac{-(1+j\omega)}{n}\right)\right)^r\right|^2 d\omega \\
&= \frac{1}{2\pi}\int_{\mathbb{R}} \hat{K}(\omega)\left|1-e^{\frac{-(1+j\omega)}{n}}\right|^{2p} d\omega \\
&= \frac{1}{2\pi}\int_{\mathbb{R}} \hat{K}(\omega)\left[1+\theta_n^2-2\theta_n\cos(\omega/n)\right]^p d\omega \\
&\overset{(d)}{\leq} \frac{C_K}{2\pi}\int_{\mathbb{R}} \frac{\left[1+\theta_n^2-2\theta_n\cos\left(\frac{\omega}{n}\right)\right]^p}{(1+\omega^2)^{\nu+1/2}} d\omega, \qquad (4.27)
\end{aligned}
$$

where, identity (c) is implied by *Bochner theorem* (cf. [Ste12], Chapter 2) and (d) is immediate consequence of Assumption 4.1. Notice that

$$
\begin{aligned}
1+\theta_n^2-2\theta_n\cos\left(\frac{\omega}{n}\right) &\leq (1-\theta_n)^2+2\theta_n\left(1-\cos\left(\frac{\omega}{n}\right)\right) \leq \frac{1}{n^2}+2\left(1-\cos\left(\frac{\omega}{n}\right)\right) \\
&= \frac{1}{n^2}+\left[\frac{\omega}{n}\,\mathrm{sinc}\left(\frac{\omega}{2n}\right)\right]^2.
\end{aligned}
$$

Let $\xi = p - (\nu + 1/2) < 1$. Henceforth, for any $R > 0$,

$$\begin{aligned}
\frac{2\pi n^{2\nu}}{C_K}\operatorname{var}\left(W_{p+1}\right) &\leq n^{2\nu}\int\limits_{\mathbb{R}} \frac{\left\{\frac{1}{n^2}+\left[\frac{\omega}{n}\operatorname{sinc}\left(\frac{\omega}{2n}\right)\right]^2\right\}^p}{\left(1+\omega^2\right)^{\nu+1/2}}d\omega = \int\limits_{\mathbb{R}} \frac{\left\{1/n^2+\left[\omega\operatorname{sinc}\left(\frac{\omega}{2}\right)\right]^2\right\}^p}{\left(1/n^2+\omega^2\right)^{\nu+1/2}}d\omega \\
&= \int\limits_{-R}^{R} \frac{\left\{1/n^2+\left[\omega\operatorname{sinc}\left(\frac{\omega}{2}\right)\right]^2\right\}^p}{\left(1/n^2+\omega^2\right)^{\nu+1/2}}d\omega + \int\limits_{|\omega|\geq R} \frac{\left\{1/n^2+\left[\omega\operatorname{sinc}\left(\frac{\omega}{2}\right)\right]^2\right\}^p}{\left(1/n^2+\omega^2\right)^{\nu+1/2}}d\omega \\
&\overset{(e)}{\leq} \int\limits_{-R}^{R} \left(1/n^2+\omega^2\right)^{\xi} d\omega + \int\limits_{|\omega|\geq R} \frac{\left\{1/n^2+\left[\omega\operatorname{sinc}\left(\frac{\omega}{2}\right)\right]^2\right\}^p}{\left(1/n^2+\omega^2\right)^{\nu+1/2}}d\omega \\
&\overset{(f)}{\leq} \int\limits_{-R}^{R} \left(1/n^2+\omega^2\right)^{\xi} d\omega + 5^p \int\limits_{|\omega|\geq R} |\omega|^{-(2\nu+1)}\, d\omega \overset{(g)}{\leq} 3R^3 + 5^p\frac{R^{-2\nu}}{\nu}. \quad (4.28)
\end{aligned}$$

Inequality (e) follows form the fact that $\sup_{\omega\in\mathbb{R}}|\operatorname{sinc}(\omega/2)| \leq 1$. In order to justify (f), observe that $|\omega\operatorname{sinc}(\omega/2)| \leq 2$ for any $\omega \in \mathbb{R}$. Thus, for large enough n and $|\omega| \geq R$, we get

$$\frac{\left\{1/n^2+\left[\omega\operatorname{sinc}\left(\frac{\omega}{2}\right)\right]^2\right\}^p}{\left(1/n^2+\omega^2\right)^{\nu+1/2}} \leq |\omega|^{-(2\nu+1)}\left(1/n^2+4\right)^p \leq 5^p\,|\omega|^{-(2\nu+1)}.$$

Note that there is some $n_0 \coloneqq n_0(R,\nu)$ such that $\sup_{\omega\in\mathbb{R}}\left(1/n^2+\omega^2\right)^{\xi} \leq 3/2R^2$ for all $n > n_0$. This immediately entails inequality (g).

Finally, minimizing the obtained upper bound in (4.28) over $R > 0$, we get

$$\operatorname{var}\left(W_{p+1}\right) \leq CC_K n^{-2\nu}\left(1+\frac{1}{\nu}\right) \tag{4.29}$$

for some universal constant $C > 0$. Thus, there is another strictly positive universal constant, C', for which $\max_{1\leq k\leq n}\operatorname{var}(W_k) = n^{-2\nu} \vee \operatorname{var}\left(W_{p+1}\right) \leq C'C_K n^{-2\nu}\left(1+\frac{1}{\nu}\right)$. So, (4.26) implies that

$$\zeta_t^{\top}\left(\Sigma_n\right)^{-1}\zeta_t \gtrsim \frac{n^{2\nu}}{C_K\left(1+\frac{1}{\nu}\right)}. \tag{4.30}$$

The combination of (4.22) and (4.30) completes our proof. □

Proof of Theorem 4.2. The proof proceeds in a similar manner as that of the preceding theorem, in the sense that it is required to show that inequality (4.22) holds. Let $\theta_n = \exp\left(-\frac{\rho^2}{n^2}\right)$. A_n represents the inverse of the Cholesky factorization of Σ_n. For any $k \leq j$ and

$q \in [0,1]$, $G(k,j;q)$ denotes the following rational function.

$$G(k,j;q) = \prod_{\ell=j-k+1}^{j} \left(1-q^{\ell}\right) \left[\prod_{\ell=1}^{k} \left(1-q^{\ell}\right)\right]^{-1},$$

and $G(k,j;1) = \binom{j}{k}$. $G(k,j;q)$ is usually referred to Gaussian binomial coefficients in the combinatorics literature. Finally, let $U_{n,t} := A_n \zeta_t$. Similar to (4.22), the aim is to obtain a universal lower bound on $\zeta_t^{\top} (\Sigma_n)^{-1} \zeta_t$ for $t \in \mathcal{C}_{n,\alpha}$. Observe that, $\zeta_t^{\top} (\Sigma_n)^{-1} \zeta_t = \left\|U_{n,t}\right\|_{\ell_2}^2$. In order to achieve a tight lower bound on $\left\|U_{n,t}\right\|_{\ell_2}$, it is pivotal to study the non-asymptotic behaviour of the entries of A_n. According to Proposition 1 of [LL$^+$00], the entries of A_n are given by

$$(A_n)_{jk} = \left(-\sqrt{\theta_n}\right)^{(j-k)} \frac{G(k-1,j-1;\theta_n)}{\sqrt{\prod_{\ell=1}^{j-1}\left(1-\theta_n^{\ell}\right)}} \mathbb{1}_{\{j \geq k\}}.$$

Since $\frac{\ell\rho^2}{n^2}$ tends to 0 as n gets large for any $\ell \in \{0,\ldots,n\}$ and $\lim_{x \searrow 0} \frac{1-e^{-x}}{x} = 1$, we get

$$\left[\prod_{\ell=1}^{j-1}\left(1-\theta_n^{\ell}\right)\right]^{-1} = \left[\prod_{\ell=1}^{j-1}\left(1-\exp\left(-\frac{\ell\rho^2}{n^2}\right)\right)\right]^{-1} \asymp \frac{1}{(j-1)!}\left(\frac{n}{\rho}\right)^{2(j-1)}. \tag{4.31}$$

Direct calculations show that $G(k-1,j-1;\theta_n) \asymp \binom{j-1}{k-1}$ for any θ_n in a small neighborhood of 1 and $j,k \in \{1,\ldots,n\}$. Thus,

$$\left(-\sqrt{\theta_n}\right)^{(j-k)} G(k-1,j-1;\theta_n) \asymp (-1)^{(j-k)} \binom{j-1}{k-1}. \tag{4.32}$$

The asymptotic identities (4.31) and (4.32) come in handy to analyze $\|U_n\|_{\ell_2}^2$:

$$\begin{aligned}
\|U_n\|_{\ell_2}^2 &\geq \sum_{j=t+1}^{n} (U_n)_j^2 = \sum_{j=t+1}^{n} \left[\sum_{k=t+1}^{j} (A_n)_{jk} - \sum_{k=1}^{t} (A_n)_{jk} \right]^2 \\
&\asymp \sum_{j=t+1}^{n} \frac{1}{(j-1)!} \left(\frac{n}{\rho}\right)^{2(j-1)} \left[\sum_{k=t+1}^{j} (-1)^{(j-k)} \binom{j-1}{k-1} - \sum_{k=1}^{t} (-1)^{(j-k)} \binom{j-1}{k-1} \right]^2 \\
&= \sum_{j=t+1}^{n} \frac{1}{(j-1)!} \left(\frac{n}{\rho}\right)^{2(j-1)} \left[\sum_{k=1}^{j} (-1)^{(j-k)} \binom{j-1}{k-1} - 2\sum_{k=1}^{t} (-1)^{(j-k)} \binom{j-1}{k-1} \right]^2 \\
&= \sum_{j=t+1}^{n} \frac{1}{(j-1)!} \left(\frac{n}{\rho}\right)^{2(j-1)} \left[0 - 2(-1)^j \binom{j-1}{t} \right]^2 \asymp \sum_{j=t+1}^{n} \frac{\binom{j-1}{t}^2}{(j-1)!} \left(\frac{n}{\rho}\right)^{2(j-1)}.
\end{aligned}$$

Thus there are universal constants $C, C' > 0$ and C_0 depending on α and ρ such that

$$\begin{aligned}
\|U_n\|_{\ell_2}^2 &\geq C \sum_{j=t+1}^{n} \frac{\binom{j-1}{t}^2}{(j-1)!} \left(\frac{n}{\rho}\right)^{2(j-1)} \geq C \frac{\binom{n-1}{t}^2}{(n-1)!} \left(\frac{n^2}{\rho^2}\right)^{(n-1)} \\
&\overset{(a)}{\geq} C' \left(\frac{n}{t}\right)^t \frac{1}{\sqrt{n}} \left(\frac{en^2}{n\rho^2}\right)^{(n-1)} \overset{(b)}{\geq} (C_0 n)^n.
\end{aligned}$$

Note that inequality (a) can be shown using *Stirling's formula* and (b) is obvious implication of the fact that $t \leq (1-\alpha)n$ (Recall $\mathcal{C}_{n,\alpha}$ from Section 4.2.1). In summary, we have that

$$|b| \sqrt{\zeta_n^\top (\Sigma_n)^{-1} \zeta_n} \gtrsim |b| (C_0 n)^{n/2}.$$

We conclude the proof by appealing to Lemma 4.1. □

Proof of Theorem 4.3. For simplicity set $c := \sigma^2 \rho^{-2\nu}$ and use $\tilde{c}_m$ to represent its estimated quantity $\tilde{\sigma}_m^2 \tilde{\rho}_m^{-2\nu}$. Recall that $\tilde{\rho}_m$ is a fixed quantity which has been chosen as the largest possible range parameter in the space Ω. Since c is consistently estimable by the maximum likelihood algorithm, there are vanishing non-negative sequences $\{\tau_m\}_{m=1}^{\infty}$ and $\{\varepsilon_m\}_{m=1}^{\infty}$ and $n_0\mathbb{N}$ such that

$$\mathbb{P}(\mathcal{A}_m) := \mathbb{P}\left(\left| \frac{\tilde{c}_m}{c} - 1 \right| < \varepsilon_m \right) > 1 - \tau_m, \quad \forall\, n \geq n_0.$$

As the range of φ_n is $[0,2]$ (See Definition 4.2), we have

$$\varphi_n\left(\tilde{T}_{GLRT}\right) \leq 2\tau_m + \mathbb{E}\left(\varphi_n\left(\tilde{T}_{GLRT}\right) \mid \mathcal{A}_m\right). \tag{4.33}$$

Furthermore, for any $\eta' = (\sigma', \tilde{\rho}_m) \in \Omega$

$$Z_t(\eta') := \frac{\zeta_t^\top \Sigma_n^{-1}(\eta') \boldsymbol{X}}{\sqrt{\zeta_t^\top \Sigma_n^{-1}(\eta') \zeta_t}}.$$

Notice that the Matern covariance matrix associated to η' can be re-parametrized as

$$\begin{aligned} \Sigma_n(\eta') &= \left[K\left(\frac{r-s}{n}, \eta'\right)\right]_{r,s=1}^n \\ &= \sigma'^2 \rho'^{-2\nu} \left[\int_{\mathbb{R}} \left[\omega^2 + \rho'^{-2}\right]^{-(\nu+1/2)} \exp\left(-j\omega\left(\frac{r-s}{n}\right)\right) d\omega\right]_{r,s=1}^n . \end{aligned} \tag{4.34}$$

Notice that the matrix appearing in the second line of (4.34), which will be denoted by $\Gamma_n(\rho')$, only depends on ρ'. The following property of $\Gamma_n(\cdot)$ is essential in our proof.

$$\Gamma_n(\rho_1) \preceq \Gamma_n(\rho_2), \quad \forall\ (\rho_1, \rho_2) \in \Omega \text{ with } \rho_1 \leq \rho_2$$

We aim to obtain a sufficient condition on b to control the second term in the right hand side of (4.33) below δ. Similar to the proof of Theorem 4.1, it is necessary to study the two following quantities: 1. variance of $Z_t(\eta')$ and 2. expected value of $Z_t(\eta')$ under the alternative hypothesis, to control the false alarm and miss detection probabilities. Observe that

$$\begin{aligned} \operatorname{var} Z_t(\eta') &= \frac{\zeta_t^\top \Sigma_n^{-1}(\eta') \Sigma_n(\eta) \Sigma_n^{-1}(\eta') \zeta_t}{\zeta_t^\top \Sigma_n^{-1}(\eta') \zeta_t} = \frac{\sigma^2 \rho^{-2\nu}}{\sigma'^2 \tilde{\rho}_m^{-2\nu}} \frac{\zeta_t^\top \Gamma_n^{-1}(\tilde{\rho}_m) \Gamma_n(\rho) \Gamma_n^{-1}(\tilde{\rho}_m) \zeta_t}{\zeta_t^\top \Gamma_n^{-1}(\tilde{\rho}_m) \zeta_t} \\ &\overset{(a)}{\leq} \frac{1}{1-\varepsilon_m} \frac{\zeta_t^\top \Gamma_n^{-1}(\tilde{\rho}_m) \Gamma_n(\rho) \Gamma_n^{-1}(\tilde{\rho}_m) \zeta_t}{\zeta_t^\top \Gamma_n^{-1}(\tilde{\rho}_m) \zeta_t} \overset{(b)}{\leq} \frac{1}{1-\varepsilon_m}. \end{aligned} \tag{4.35}$$

Where (a) is an easy implication of the fact that $\eta' \in \mathcal{A}_m$. Furthermore, (b) follows from the fact that $\Gamma(\rho) \preceq \Gamma(\tilde{\rho}_m)$. In other words, $\operatorname{var} Z_t(\eta') < 1/(1-\varepsilon_m)$. Lemma 4.1 guarantees the existence of a vanishing sequence $\{\tau'_m\}_{m \in \mathbb{N}}$, which depends on ε_m, such that

$$\mathbb{P}\left(\max_{1 \leq t \leq n} Z_t^2(\tilde{\eta}_m) \geq \left[1 + 2\left(\log\left(\frac{2n}{\delta}\right) + \sqrt{\log\left(\frac{2n}{\delta}\right)}\right)\right] \mid \mathcal{A}_m\right) \leq \frac{\delta}{2} + \tau'_m. \tag{4.36}$$

So, we have controlled type one error from above in (4.36). Now we turn to control the type two error from above. Assume that there is a sudden change in the mean at $\bar{t} \in \mathcal{C}_{n,\alpha}$. According to Lemma 4.1, type II error is less than $\delta/2$ whenever for any $\eta' \in \mathcal{A}_m$

$$\frac{|b|}{2}\sqrt{\zeta_{\bar{t}}^{\top}\Sigma_n^{-1}(\eta')\zeta_{\bar{t}}} \geq \frac{4}{1-\varepsilon_m}\sqrt{\log\left(\frac{2n}{\delta}\right)}. \tag{4.37}$$

Having a lower bound on $\zeta_{\bar{t}}^{\top}\Sigma_n^{-1}(\eta')\zeta_{\bar{t}}$ is necessary to make sure that b satisfying (4.37). Notice that

$$\Sigma_n(\eta') = \sigma'^2\tilde{\rho}_m^{-2\nu}\Gamma_n(\tilde{\rho}_m) \leq (1+\varepsilon_m)\sigma^2\rho^{-2\nu}\Gamma_n(\tilde{\rho}_m) = (1+\varepsilon_m)\, c\Gamma_n(\tilde{\rho}_m).$$

Thus, (4.37) holds true whenever

$$\frac{|b|}{2}\sqrt{c\zeta_{\bar{t}}^{\top}\Gamma_n^{-1}(\tilde{\rho}_m)\zeta_{\bar{t}}} \geq \frac{4\sqrt{1+\varepsilon_m}}{1-\varepsilon_m}\sqrt{\log\left(\frac{2n}{\delta}\right)}. \tag{4.38}$$

Theorem 4 of [SY] conveys the equivalence of the associated Gaussian measures to matrices $\Gamma_n(\rho)$ and $\Gamma_n(\tilde{\rho}_m)$. Thus Lemma 4.3 ensures the existence of a bounded scalar $B > 1$ for which

$$\zeta_{\bar{t}}^{\top}\Gamma_n^{-1}(\tilde{\rho}_m)\zeta_{\bar{t}} \geq \frac{1}{B}\zeta_{\bar{t}}^{\top}\Gamma_n^{-1}(\rho)\zeta_{\bar{t}}. \tag{4.39}$$

Combining (4.38) and (4.39) yields a sufficient condition on b to control the type two error

$$\frac{|b|}{2}\sqrt{c\zeta_{\bar{t}}^{\top}\Gamma_n^{-1}(\rho)\zeta_{\bar{t}}} = \frac{|b|}{2}\sqrt{\zeta_{\bar{t}}^{\top}\Sigma_n^{-1}(\eta)\zeta_{\bar{t}}} \geq \frac{4\sqrt{B(1+\varepsilon_m)}}{1-\varepsilon_m}\sqrt{\log\left(\frac{2n}{\delta}\right)}.$$

In conclusion we employ the lower bound on $\zeta_{\bar{t}}^{\top}\Sigma_n^{-1}(\eta)\zeta_{\bar{t}}$ in (4.30) gives the proper rate of detectable b. □

Proof of Theorem 4.4. As the proof has much in common with the proof of Theorem 4.3, we skip the algebraic details to avoid repetition. $\tilde{\eta}_m = (\tilde{\sigma}_m, \tilde{\rho}_m)$ stands for an estimate of the unknown parameters $\eta = (\sigma, \rho)$. Note that in this case η is consistently estimable, i.e. there are two vanishing sequences ε_m and τ_m, $m \in \mathbb{N}$ and a large enough $n_0 \in \mathbb{N}$ such that

$$\mathbb{P}(\mathcal{A}_m) := \mathbb{P}(|\rho - \tilde{\rho}_m| \vee |\sigma - \tilde{\sigma}_m| < \varepsilon_m) > 1 - \tau_m, \quad \forall\, n \geq n_0$$

Recall from (4.33) that $\varphi_n\left(\tilde{T}_{GLRT}\right) \leq 2\tau_m + \mathbb{E}\left(\varphi_n\left(\tilde{T}_{GLRT}\right) \mid \mathcal{A}_m\right)$. We use $\hat{K}_{\eta'}(\omega)$ to represent the spectral density of the powered exponential covariance function associated to $\eta' = (\sigma', \rho')$. That is

$$\hat{K}_{\eta'}(\omega) = \int_{\mathbb{R}} \sigma'^2 \exp\left(-\left|\frac{x}{\rho'}\right|^{\beta} - jx\omega\right) dx.$$

It is trivial that $\hat{K}_{\eta'}(\cdot)$ is a strictly positive, continuous function of both ω and η'. Further-

more, due to the compactness of the parameter space Ω, $\hat{K}_{\eta'}$ is uniformly continuous with respect to η'. Thus there is another vanishing sequence ε'_m (depending on ε_m) such that

$$\left\| \frac{\hat{K}_{\eta'}}{\hat{K}_{\eta}} - 1 \right\|_{\infty} \leq \varepsilon'_m, \quad \forall\, \eta' = (\sigma', \rho') \text{ with } \left|\rho - \rho'\right| \vee \left|\sigma - \sigma'\right| < \varepsilon_m.$$

So, the following inequality holds for any $\omega \in \mathbb{R}$, and any η' satisfying $|\rho - \rho'| \vee |\sigma - \sigma'| < \varepsilon_m$.

$$\hat{K}_{\eta}(\omega)(1 - \varepsilon'_m) \leq \hat{K}_{\eta'}(\omega) \leq \hat{K}_{\eta}(\omega)(1 + \varepsilon'_m). \tag{4.40}$$

Note that (4.40) can be easily translated in terms of the covariance matrices $\Sigma_n(\eta)$ and $\Sigma_n(\eta')$.

$$(1 - \varepsilon'_m)\Sigma_n(\eta) \leq \Sigma_n(\eta') \leq (1 + \varepsilon'_m)\Sigma_n(\eta). \tag{4.41}$$

Now using similar techniques as (4.35) and (4.36), we can bound the variance of $Z_t(\eta')$.

$$\begin{aligned} \operatorname{var} Z_t(\eta') &= \frac{\zeta_t^\top \Sigma_n^{-1}(\eta')\Sigma_n(\eta)\Sigma_n^{-1}(\eta')\zeta_t}{\zeta_t^\top \Sigma_n^{-1}(\eta')\zeta_t} \leq \frac{1}{1 - \varepsilon'_m} \frac{\zeta_t^\top \Sigma_n^{-1}(\eta')\Sigma_n(\eta')\Sigma_n^{-1}(\eta')\zeta_t}{\zeta_t^\top \Sigma_n^{-1}(\eta')\zeta_t} \\ &= \frac{1}{1 - \varepsilon'_m}, \end{aligned}$$

and control the type one error below $\delta/2 + \tau'_m$ for a vanishing sequence τ'_m (depending on ε'_m). Now we introduce a sufficient condition on b to keep the type two error below $\delta/2$. Using Lemma 4.1 b should satisfy

$$\frac{|b|}{2}\sqrt{\zeta_{\bar{t}}^\top \Sigma_n^{-1}(\eta')\zeta_{\bar{t}}} \geq \frac{4}{1 - \varepsilon'_m}\sqrt{\log\left(\frac{2n}{\delta}\right)}.$$

for any η' with $|\rho - \rho'| \vee |\sigma - \sigma'| < \varepsilon_m$. It follows from (4.41) that

$$\zeta_{\bar{t}}^\top \Sigma_n^{-1}(\eta')\zeta_{\bar{t}} \geq (1 - \varepsilon'_m)\zeta_{\bar{t}}^\top \Sigma_n^{-1}(\eta)\zeta_{\bar{t}}.$$

So, we can have a slightly stronger restriction on b by combining the last two inequalities.

$$\frac{|b|}{2}\sqrt{\zeta_{\bar{t}}^\top \Sigma_n^{-1}(\eta)\zeta_{\bar{t}}} \geq \frac{4}{(1 - \varepsilon'_m)^2}\sqrt{\log\left(\frac{2n}{\delta}\right)}. \tag{4.42}$$

Notice that as m increases we have $4/(1 - \varepsilon'_m)^2 < 5$. Thus, (4.42) holds when $n \geq n_0$ (for

some large enough n_0) and

$$\frac{|b|}{2}\sqrt{\zeta_{\bar{t}}^{\top}\Sigma_n^{-1}(\eta)\,\zeta_{\bar{t}}} \geq 5\sqrt{\log\left(\frac{2n}{\delta}\right)}. \tag{4.43}$$

Inequality (4.43) is same as the sufficient condition on b for the case of known covariance parameters, which leads to the same detection rate. □

Proof of Theorem 4.5. Choose $t \in C_{n,\alpha}$ and define

$$U_t^{\star} := \sqrt{\frac{t(n-t)}{n^2}}\left(\frac{1}{n-t}\sum_{k=t+1}^{n} X_k - \frac{1}{t}\sum_{k=1}^{n} X_k\right).$$

Moreover set

$$R_{n,\delta} = \sqrt{n\left(1 + 2\log\left(\frac{2n(1-2\alpha)}{\delta}\right) + 2\sqrt{\log\left(\frac{2n(1-2\alpha)}{\delta}\right)}\right)}.$$

Note that under the null hypothesis, $U_t^{\star}$ is a zero mean random variable and

$$\begin{aligned}
&\lim_{n\to\infty} \operatorname{var}\left(U_t^{\star}\right) \\
\overset{(a)}{=}\ &\lim_{n\to\infty} \frac{t(n-t)}{n^2}\int_{-\infty}^{\infty} \frac{\hat{K}(\omega)}{2\pi}\left|\frac{1}{n-t}\sum_{k=t+1}^{n}\exp(-jk\omega/n) - \frac{1}{t}\sum_{k=1}^{n}\exp(-jk\omega/n)\right|^2 d\omega \\
=\ &\lim_{n\to\infty}\int_{-\infty}^{\infty} \frac{\hat{K}(\omega)}{2\pi}\left|\sqrt{\frac{\beta}{1-\beta}}\sum_{k=t+1}^{n}\frac{\exp(-jk\omega/n)}{n} - \sqrt{\frac{1-\beta}{\beta}}\sum_{k=1}^{n}\frac{\exp(-jk\omega/n)}{n}\right|^2 d\omega \\
\overset{(a)}{=}\ &\int_{-\infty}^{\infty} \frac{\hat{K}(\omega)}{2\pi}\left|\sqrt{\frac{\beta}{1-\beta}}\int_{\beta}^{1} e^{-j\omega u}du - \sqrt{\frac{1-\beta}{\beta}}\int_{0}^{\beta} e^{-j\omega u}du\right|^2 d\omega \\
=\ &\int_{-\infty}^{\infty} \frac{\hat{K}(\omega)\,G_{\beta}(\omega)}{2\pi} d\omega,
\end{aligned}$$

where

$$\begin{aligned}
G_{\beta}(\omega) \ :=\ &\left[(1-\beta)\operatorname{sinc}\left(\frac{\beta\omega}{2}\right)\right]^2 + \left[\beta\operatorname{sinc}\left(\frac{(1-\beta)\omega}{2}\right)\right]^2 \\
+\ &4\beta(1-\beta)\operatorname{sinc}\left(\frac{\beta\omega}{2}\right)\operatorname{sinc}\left(\frac{(1-\beta)\omega}{2}\right)\sin^2\left(\frac{\omega}{2}\right).
\end{aligned} \tag{4.44}$$

The identity (a) is implied by Bochner Theorem and (b) follows from the dominated convergence theorem. It is easy to see that $\left\|G_\beta\right\|_\infty \le 1$ and so $\lim_{n\to\infty} \mathrm{var}(U_n^\star) \le 1$ by the triangle inequality. Moreover, Lemma 4.5 shows that the achieved upper bound on $\sigma_n^2 = \mathrm{var}(U_n^\star)$ is tight up to some constant whenever $\hat{K}$ has a uniformly bounded derivative. Namely, there is a universal constant $c \in (0,1)$ such that $c \le \lim_{n\to\infty} \mathrm{var}(U_n^\star) \le 1$ for any $\beta \in (0,1)$. Let $R_{n,\delta}^\star = R_{n,\delta}^2/n$. Thus

$$\mathbb{P}(T=1 \mid \mathbb{H}_0) = \mathbb{P}\left(\max_{t\in\mathcal{C}_{n,\alpha}} |U_t| \ge R_{n,\delta} \mid \mathbb{H}_0\right) = \mathbb{P}\left(\max_{t\in\mathcal{C}_{n,\alpha}} \left|U_t^\star\right|^2 \ge R_{n,\delta}^\star \mid \mathbb{H}_0\right). \tag{4.45}$$

For any $t \in \mathcal{C}_{n,\alpha}$, $\left|U_t^\star\right|^2$ is a (non-normalized) χ_1^2 random variable, as $\sigma_n^2 \le 1$. Moreover $\left|\mathcal{C}_{n,\alpha}\right| = n(1-2\alpha)$. So the part (a) of Lemma 4.1 says that

$$\mathbb{P}\left(\max_{t\in\mathcal{C}_{n,\alpha}} \left|U_t^\star\right|^2 \ge R_{n,\delta}^\star \mid \mathbb{H}_0\right) \le \frac{\delta}{2}.$$

Now we turn to control the miss detection probability. Without loss of generality assume that $b > 0$. Choose an arbitrary $t \in \mathcal{C}_{n,\alpha}$. A line of algebra shows that

$$\mathbb{E}\left(U_t^\star \mid \mathbb{H}_{1,t}\right) \ge b\sqrt{\alpha(1-\alpha)}. \tag{4.46}$$

Eq. (4.18) on b implies that $\mathbb{E}\left(U_t^\star \mid \mathbb{H}_{1,t}\right) \ge 4\sqrt{\log\left(2n(1-2\alpha)/\delta\right)}$. In other words, given a sudden jump at t, $\left|U_s^\star\right|^2$, $s \in \mathcal{C}_{n,\alpha}$ are non-central χ_1^2 random variables satisfying the conditions of the part (b) of Lemma 4.1. Hence

$$\mathbb{P}(T=0 \mid \mathbb{H}_{1,t}) = \mathbb{P}\left(\max_{s\in\mathcal{C}_{n,\alpha}} \left|U_s^\star\right|^2 \le R_{n,\delta}^\star \mid \mathbb{H}_{1,t}\right) \le \frac{\delta}{2}. \tag{4.47}$$

□

Proof of Theorem 4.6. We follow the standard method for bounding the Bayes risk from below. Observe that

$$\begin{aligned}
\inf_T \varphi_n(T) &= 1 - \sup_T \inf_{t\in\mathcal{C}_{n,\alpha}} \left[\mathbb{P}(T=0 \mid \mathbb{H}_0) - \mathbb{P}(T=0 \mid \mathbb{H}_{1,t})\right] \\
&\ge 1 - \inf_{t\in\mathcal{C}_{n,\alpha}} \sup_T \left|\mathbb{P}(T=0 \mid \mathbb{H}_0) - \mathbb{P}(T=0 \mid \mathbb{H}_{1,t})\right| \overset{(a)}{\ge} 1 - \inf_{t\in\mathcal{C}_{n,\alpha}} H(\mathbb{P}_0, \mathbb{P}_{1,t}),
\end{aligned}$$

where (a) follows from inequality 2.27 in [Tsy09]. So, it suffices to show that $\inf_{t\in\mathcal{C}_{n,\alpha}} H^2(\mathbb{P}_0, \mathbb{P}_{1,t}) \le (1-\delta)^2$. A few lines of straightforward algebra on the explicit form

of Hellinger distance of Gaussian measures indicates that $\inf_T \varphi_n(T) \geq \delta$, whenever

$$b^2 \inf_{t \in \mathcal{C}_{n,\alpha}} \zeta_t^\top (\Sigma_n)^{-1} \zeta_t \leq 32 \log\left(\frac{1}{\delta(2-\delta)}\right). \tag{4.48}$$

Henceforth, it is enough to obtain a tight upper bound on $\inf_{t \in \mathcal{C}_{n,\alpha}} \zeta_t^\top (\Sigma_n)^{-1} \zeta_t$.
Let $\sigma = 1$ and choose $\rho > 0$ by $\rho^{-2p+1} = \frac{C'_K \Gamma(p-1/2)}{\sqrt{4\pi}\Gamma(p)}$. Furthermore, let $\hat{F}_{\rho,p,\sigma} : \mathbb{R} \mapsto \mathbb{R}$ denote the Matern spectral density parametrized by p, ρ and σ as (4.9). Note that ρ is well defined due to the first condition in Assumption 4.2. Define $\xi_t \in \mathbb{R}^n$ by $\xi_t(k) = \mathbb{1}_{\{k>t\}}$ and let $\xi'_t = \xi_t - \zeta_t$ for any $t \in \mathcal{C}_{n,\alpha}$. Moreover, let $\theta_n = \exp(-1/n)$ and $S_t = \{t+1, \ldots, n\}$. Finally, define the covariance matrix $\Psi_n \in \mathbb{R}^{n \times n}$ by $\Psi_n = \left[F_{\rho,p,\sigma}((r-s)/n)\right]_{r,s=1}^n$. Observe that

$$\begin{aligned} \zeta_t^\top (\Sigma_n)^{-1} \zeta_t &= 2\left(\xi_t^\top (\Sigma_n)^{-1} \xi_t + \xi_t'^\top (\Sigma_n)^{-1} \xi'_t\right) - \mathbb{1}_n^\top (\Sigma_n)^{-1} \mathbb{1}_n \\ &\leq 4\left(\xi_t^\top (\Sigma_n)^{-1} \xi_t \vee \xi_t'^\top (\Sigma_n)^{-1} \xi'_t\right). \end{aligned} \tag{4.49}$$

We aim to prove that there is a constant $C := C(p) > 0$ for which $\xi_t^\top (\Sigma_n)^{-1} \xi_t \leq C n^{2p-1}$. The same upper bound can be obtained for $\xi_t'^\top (\Sigma_n)^{-1} \xi'_t$ in an analogous manner.
We first show that

$$\left(\frac{\hat{K}}{\hat{F}_{\rho,p,\sigma}} - 1\right) \in \mathbb{L}^2(\mathbb{R}). \tag{4.50}$$

Let M represent the finite lim sup in the second condition of Assumption 4.2. Without loss of generality, we can assume that $\beta < 2$ in Assumption 4.2. Using a few lines of algebra one can find a bounded positive scalar M for which the following inequality holds.

$$\begin{aligned} \limsup_{\omega \to \infty} \left|\omega^\beta \left(\frac{\hat{K}(\omega)}{\hat{F}_{\rho,p,\sigma}(\omega)} - 1\right)\right| &= \limsup_{\omega \to \infty} \left|\omega^\beta \left(\frac{\hat{K}(\omega)\omega^{2p}}{C'_K}\left(1 + \frac{1}{\rho^2\omega^2}\right)^p - 1\right)\right| \\ &\leq \limsup_{\omega \to \infty} \left|\omega^\beta \left(\frac{\hat{K}(\omega)\omega^{2p}}{C'_K} - 1\right)\right| \\ &+ \limsup_{\omega \to \infty} \left|\omega^\beta \left[\frac{\hat{K}(\omega)\omega^{2p}}{C'_K}\left(\left(1 + \frac{1}{\rho^2\omega^2}\right)^p - 1\right)\right]\right| \\ &\overset{(a)}{=} M + \frac{2p\rho^{-2}}{C'_K} \limsup_{\omega \to \infty} \hat{K}(\omega)|\omega|^{2p-2+\beta} \overset{(b)}{=} M. \end{aligned}$$

Notice that, identity (a) follows from Assumption 4.2 and first order Taylor expansion of $(1+x)^p$ for infinitesimal $x > 0$. Moreover, (b) follows from the combination of $\beta < 2$ and

the first condition in Assumption 4.2. Namely, there is $R > 0$ such that

$$\left| \frac{\hat{K}(\omega)}{\hat{F}_{\rho,p,\sigma}(\omega)} - 1 \right| \leq \frac{2M}{|\omega|^{\beta}}, \quad \forall\ |\omega| \geq R,$$

which substantiates (4.50) as $\beta > 1/2$.

It is known (4.31, Chapter *III*, [IR12]) that there is a function $\phi \in \mathbb{L}^2(\mathbb{R})$ with bounded support such that $\hat{F}_{\rho,p,\sigma}(\omega) \asymp \left|\hat{\phi}(\omega)\right|^2$ as $|\omega| \to \infty$. Theorem 4 of Skorokhod [SY] implies that the associated zero mean Gaussian measures to spectral densities $\hat{K}$ and $\hat{F}_{\rho,p,\sigma}$ are equivalent. Based upon Lemma 4.3, there exists a constant $h \in (0, \infty)$ such that

$$\frac{1}{h} \leq \left| \lim_{n\to\infty} \frac{\xi_t^{\top} (\Sigma_n)^{-1} \xi_t}{\xi_t^{\top} (\Psi_n)^{-1} \xi_t} \right| \leq h.$$

So, it suffices to show that $\xi_t^{\top} (\Psi_n)^{-1} \xi_t \leq C' n^{2p-1}$ for some appropriately chosen $C' > 0$ depending on h and C. Letting $\nu = p - 1/2$ and recalling A_n, W and D_n form the proof of Theorem 4.1, we have

$$\xi_t^{\top} (\Psi_n)^{-1} \xi_t = (A_n \xi_t)^{\top} D_n^{-1} (A_n \xi_t) \overset{(b)}{\leq} \frac{\|A_n \xi_t\|_{\ell_2}^2}{\lambda_{\min}(D_n(S_t, S_t))}. \tag{4.51}$$

Note that inequality (b) is inferred from $\mathrm{supp}(A_n \xi_t) = S_t$. Applying a similar technique as (4.24), we get

$$\begin{aligned} \|A_n \xi_t\|_{\ell_2}^2 &= (n-t-p)(1-\theta_n)^p + \sum_{k=1}^{p} \left(\sum_{j=0}^{k-1} \binom{p}{j} (-\theta_n)^j \right)^2 \\ &\leq n(1-\theta_n)^p + 2 \sum_{k=1}^{p} \left(\sum_{j=0}^{k-1} \binom{p}{j} (-1)^j \right)^2 \\ &\leq 2\binom{2p-2}{p-1} + n(1-\theta_n)^p \leq n^{-(p-1)} + 2(2e)^{p-1} \leq (2e)^p. \end{aligned} \tag{4.52}$$

So, $\xi_t^{\top} (\Psi_n)^{-1} \xi_t \leq (2e)^{2p} [\lambda_{\min}(D_n(S_t, S_t))]^{-1}$.

Next, we control the smallest eigenvalue of $D_n(S_t, S_t)$ from the below. We first control the diagonal entries from below. Note that all the diagonal entries of $D_n(S_t, S_t)$ are the same

and given by (cf. (4.27))

$$\begin{aligned} Q &= \int_{\mathbb{R}} \frac{\hat{F}_{\rho,p,\sigma}(\omega)}{2\pi} \left[1+\theta_n^2 - 2\theta_n \cos(\omega/n)\right]^p d\omega \\ &\overset{(c)}{\propto} \int_{\mathbb{R}} \rho^{-2\nu} \left(\frac{1}{\rho^2}+\omega^2\right)^{-p} \left[1+\theta_n^2 - 2\theta_n \cos(\omega/n)\right]^p d\omega \\ &= n^{-2\nu} \int_{\mathbb{R}} \left[\frac{(1-\theta_n)^2 + 4\theta_n \sin^2(\omega/2\rho)}{1/n^2+\omega^2}\right]^p d\omega \overset{(d)}{\geq} \frac{n^{-2\nu}\rho^{-2p}}{2} \int_{\mathbb{R}} \left[\operatorname{sinc}(\omega/2\rho)\right]^{2p} d\omega \\ &= C'_\rho n^{-2\nu}, \end{aligned} \tag{4.53}$$

where (c) is obtained from (4.9) and the inequality (d) follows from the fact that for any $\gamma \in (0,1)$ (here we put $\gamma = 2^{-\frac{1}{p}}$), there is $n_0(\gamma)$ such that for any $n \geq n_0(\gamma)$,

$$\frac{(1-\theta_n)^2 + 4\theta_n \sin^2(\omega/2\rho)}{1/n^2+\omega^2} \geq \frac{\gamma}{\rho^2} \left[\operatorname{sinc}(\omega/2\rho)\right]^2.$$

We skip the proof of this inequality due to the simplicity.
Now, let $\Xi \coloneqq D_n(S_t, S_t)/Q$. The combination of (4.51), (4.52) and (4.53) shows that

$$\xi_t^\top (\Psi_n)^{-1} \xi_t \leq \frac{C_0 n^{-2\nu}}{\lambda_{\min}(\Xi)} \;\Rightarrow\; \xi_t^\top (\Sigma_n)^{-1} \xi_t \leq \frac{C'_0 n^{2\nu}}{\lambda_{\min}(\Xi)} = \frac{C'_0 n^{2p-1}}{\lambda_{\min}(\Xi)},$$

for some constants, $C_0(p)$ and C'_0 depending on C_0, h and K. It can be shown using identity 1.2 of [BL11] that there is some integrable function $g : [-\pi,\pi] \mapsto \mathbb{R}$ with $m_g \coloneqq \operatorname{essinf}(g) > 0$ such that Ξ is a $p-$banded correlation matrix, i.e. $\Xi(r,s) = 0$ for $|r-s| \geq p$, and $\Xi = \mathcal{T}_n(f)$. In remains to note that Lemma 6 of [G$^+$06] implies that $\lambda_{\min}(\Xi) > m_g$ for any n, which concludes the proof. □

4.9 Auxiliary Results

We now present several technical results needed in Section 4.8.

Lemma 4.1. Let $\sigma_0 \geq 1$ and $n \geq 2$. Let $Z \in \mathbb{R}^n$ be a Gaussian random vector with $\mathbb{E}Z = \mu$ and $\operatorname{var} Z_k \leq \sigma_0^2$ for any $1 \leq k \leq n$. Moreover, let $R_n = 1 + 2\left(\log\left(\frac{2n}{\delta}\right) + \sqrt{\log\left(\frac{2n}{\delta}\right)}\right)$. For any $\delta \in (0,1)$ and any $n \in \mathbb{N}$, the following results hold.

1. If $\mu = 0$, then $\mathbb{P}\left[\max_{1\leq j\leq n} Z_j^2 \geq \sigma_0^2 R_n\right] \leq \frac{\delta}{2}$.

2. If $\max\limits_{1\le j\le n}\left|\mu_j\right| \ge 4\sigma_0\sqrt{\log\left(\frac{2n}{\delta}\right)}$, then $\mathbb{P}\left[\max\limits_{1\le j\le n} Z_j^2 \le \sigma_0^2 R_n\right] \le \frac{\delta}{2}$.

Proof. For brevity, let $\sigma_j = \mathrm{var} Z_j,\ j = 1,\dots,n$. Notice that $\left(\frac{Z_j}{\sigma_j}\right)^2$ are standard χ_1^2 random variables, for any $j = 1,\dots,n$. Lemma 8.1 in [Bir01] implies that $\mathbb{P}\left(Z_j^2 \ge \sigma_j^2 R_n\right) \le \frac{\delta}{2n}$. Thus, $\mathbb{P}\left(Z_j^2 \ge \sigma_0^2 R_n\right) \le \frac{\delta}{2n}$ due to $\sigma_j \le \sigma_0$. We conclude the proof of the first part by a union bound argument. Now, we turn to prove the second part. Define $k \coloneqq \arg\max\limits_{1\le j\le n}\left|\mu_j\right|$. It is easy to verify that $R_n \le 4\log\left(\frac{2n}{\delta}\right)$. Observe that

$$\mathbb{P}\left[\max_{1\le j\le n} Z_j^2 \le \sigma_0^2 R_n\right] \le \mathbb{P}\left[\frac{Z_k^2}{\sigma_k^2} \le \left(\frac{\sigma_0}{\sigma_k}\right)^2 R_n\right] \le \mathbb{P}\left[\frac{Z_k^2}{\sigma_k^2} \le 4\left(\frac{\sigma_0}{\sigma_k}\right)^2 \log\left(\frac{2n}{\delta}\right)\right].$$

Moreover, $\frac{Z_k^2}{\sigma_k^2}$ is a non-central χ_1^2 random variables with non-centrality parameter $B_k \coloneqq \left|\frac{\mu_k}{\sigma_k}\right|$. The lower bound condition on $|\mu_k|$ implies that $B_k \ge 4\frac{\sigma_0}{\sigma_k}\sqrt{\log\left(\frac{2n}{\delta}\right)}$. We finish the proof by the following inequality,

$$\mathbb{P}\left[\frac{Z_k^2}{\sigma_k^2} \le 4\left(\frac{\sigma_0}{\sigma_k}\right)^2 \log\left(\frac{2n}{\delta}\right)\right] \overset{(a)}{\le} \mathbb{P}\left[\frac{Z_k^2}{\sigma_k^2} \le 1 + B_k^2 - 2\sqrt{\left(1+2B_k^2\right)\log\left(\frac{2}{\delta}\right)}\right] \overset{(b)}{\le} \frac{\delta}{2}.$$

In order to demonstrate inequality (a), we need to show that $1 + B_k^2 - 2\sqrt{\left(1+2B_k^2\right)\log\left(\frac{2}{\delta}\right)} \ge 4\left(\frac{\sigma_0}{\sigma_k}\right)^2\log\left(\frac{2n}{\delta}\right)$ which can be shown by obvious inequality $\sigma_0/\sigma_k \ge 1$ and a few lines of algebra. Inequality (b) can be inferred from Lemma 8.1 of [Bir01]. □

Proposition 4.3 (*Kantorovich* inequality, (p. 452, [HJ12]))**.** Let $\Sigma \in \mathbb{R}^{n\times n}$ be a non-singular covariance matrix and let $V \in \mathbb{R}^n$ be a non-zero vector. Then, $V^\top\Sigma^{-1}V \ge \frac{\|V\|_{\ell_2}^4}{V^\top\Sigma V}$.

Lemma 4.2. Let $\delta \in (0,2)$, $d \in (0,\infty)$ and define $K : \mathbb{R} \mapsto \mathbb{R}$ by $K(r) = \sigma^2\exp\left(-\left|\frac{r}{\rho}\right|^\delta\right)$. Then,

$$\lim_{\omega\to\infty} \hat{K}(\omega)|\omega|^{1+\delta} = C_\delta(\rho,\sigma) \coloneqq \frac{\sigma^2\delta\Gamma(\delta)\sin\left(\frac{\pi\delta}{2}\right)}{\pi\rho^\delta}.$$

Proof. Obviously $C_\delta(\rho,\sigma) = \sigma^2 C_\delta(\rho,1)$, so without loss of generality assume that $\sigma = 1$. Moreover $K(r)$ is of index δ as $|r| \to 0$, i.e. $\lim\limits_{|r|\to 0}\frac{1-K(r\lambda)}{1-K(r)} = \lambda^\delta\ \forall\ \lambda > 0$. The Tauberian Theorem (p. 35, [SCA12]) says that

$$\lim_{\omega\to\infty}\left[1-K(1/\omega)\right]^{-1}\int_\omega^\infty \hat{K}(u)\,du = \frac{\Gamma(\delta)\sin\left(\frac{\pi\delta}{2}\right)}{\pi} = \frac{C_\delta(\rho,1)\rho^\delta}{\delta}. \tag{4.54}$$

Moreover, the first order Taylor expansion of e^{-x} is at 0, implies that $[1-K(1/\omega)]^{-1}(\rho\omega)^{-\delta} \to 1$ as $\omega \to 0$. Thus, (4.54) can be rewritten by last limiting identity and applying L'Hospital's rule.

$$C_\delta(\rho,1) = \lim_{\omega\to\infty} \delta\rho^{-\delta}(\rho\omega)^\delta \delta\omega^\delta \int_\omega^\infty \hat{K}(u)\,du = \lim_{\omega\to\infty} \delta\omega^\delta \int_\omega^\infty \hat{K}(u)\,du = \lim_{\omega\to\infty} \hat{K}(\omega)|\omega|^{1+\delta}.$$

□

The following Lemma is probably well-known in the literature of GPs (e.g. the identity 2 of [SCA12] (p.112) is analogous but not exactly same as the part (a) of Lemma 4.3). Because of the absence of direct references, we include and prove the following result in this section.

Lemma 4.3. Let G_i, $i = 1,2$ be two zero mean stationary GP in $[0,1]$ associated to covariance functions K_i, $i = 1,2$, respectively. For any $n \in \mathbb{N}$, define two positive definite covariance matrices by $\Sigma_n := \left[K_1\left(\frac{r-s}{n}\right)\right]$ and $\Psi_n := \left[K_2\left(\frac{r-s}{n}\right)\right]$. If G_1 and G_2 induce equivalent measures on the Hilbert space of $\mathbb{L}^2([0,1])$, then there exists an scalar $B \in [1,\infty)$ for which

1. $\frac{1}{B} \le \lim_{n\to\infty} \inf_{v\neq\mathbf{0}_n} \frac{v^\top\Sigma_n v}{v^\top\Psi_n v} \le \lim_{n\to\infty} \sup_{v\neq\mathbf{0}_n} \frac{v^\top\Sigma_n v}{v^\top\Psi_n v} \le B.$
2. $\frac{1}{B} \le \lim_{n\to\infty} \inf_{v\neq\mathbf{0}_n} \frac{v^\top\Sigma_n^{-1} v}{v^\top\Psi_n^{-1} v} \le \lim_{n\to\infty} \sup_{v\neq\mathbf{0}_n} \frac{v^\top\Sigma_n^{-1} v}{v^\top\Psi_n^{-1} v} \le B.$

Proof. We use $\mathbb{P}_i, i = 1,2$ to denote the probability measures with respect to G_i, $i = 1,2$, respectively. Abusing the notation, $\boldsymbol{X} \in \mathbb{R}^n$ represents the random vector generated by sampling GP at $\{k/n\}_{k=1}^n$ for any $n \in \mathbb{N}$. We prove the existence of a finite scalar B_1 for which $\lim_{n\to\infty} \sup_{v\neq\mathbf{0}_n} \frac{v^\top\Sigma_n v}{v^\top\Psi_n v} \le B_1$. Assume toward contradiction that $\lim_{n\to\infty} \sup_{v\neq\mathbf{0}_n} \frac{v^\top\Sigma_n v}{v^\top\Psi_n v}$ tends to infinity. So, there is a sequence of non-zero vectors $\{v_n \in \mathbb{R}^n\}_{n=1}^\infty$ such that

$$\limsup_{n\to\infty} \frac{v_n^\top\Sigma_n v_n}{v_n^\top\Psi_n v_n} = \infty. \tag{4.55}$$

Consider the measurable event $\mathrm{E}_n = \left[|\langle v_n, \boldsymbol{X}\rangle| \ge \sqrt{v_n^\top\Sigma_n v_n}\right]$. Simple calculations shows that

$$\mathbb{P}_1(\mathrm{E}_n) = Q(1), \quad \mathbb{P}_2(\mathrm{E}_n) = Q\left(\sqrt{\frac{v_n^\top\Sigma_n v_n}{v_n^\top\Psi_n v_n}}\right), \tag{4.56}$$

in which $Q(\cdot)$ stands for the Q-function, i.e. $Q(r) = \int \frac{1}{\sqrt{2\pi}} \exp\left(-x^2/2\right) \mathbb{1}(|x| \ge r)\,dx$. Combining (4.55) and (4.56) leads to $\limsup_{n\to\infty} \frac{\mathbb{P}_1(\mathrm{E}_n)}{\mathbb{P}_2(\mathrm{E}_n)} = \infty$ which contradicts the absolute con-

tinuity of $\mathbb{P}_1$ with respect to $\mathbb{P}_2$. One cam show using the same technique that there is $B_2 \in (1, \infty)$ such that

$$\frac{1}{B_2} \leq \lim_{n\to\infty} \inf_{v\neq \mathbf{0}_n} \frac{v^\top \Sigma_n v}{v^\top \Psi_n v}.$$

We conclude the proof by choosing $B = B_1 \vee B_2$. Now, we turn to substantiate the second claim. Pick a non-zero vector $v \in \mathbb{R}^n$. According to Lemma 4.4, there is an suitably chosen $n-$dimensional vector u (The inner product of u and v is necessarily 1) such that

$$\frac{v^\top \Sigma_n^{-1} v}{v^\top \Psi_n^{-1} v} = v^\top \Sigma_n^{-1} v u^\top \Psi_n u = \frac{u^\top \Psi_n u}{\max_{\langle \omega, v\rangle = 1} \omega^\top \Sigma_n \omega} \leq \frac{u^\top \Psi_n u}{u^\top \Sigma_n u} \overset{(a)}{\leq} B.$$

Note that the inequality (a) is obtained from the first part of this Lemma. Taking supremum over all non-zero $v \in \mathbb{R}^n$ and $n \in \mathbb{N}$ terminates the proof. □

Lemma 4.4. Let $\Sigma \in \mathbb{R}^{n\times n}$ be a non-singular covariance matrix and let $\omega \in \mathbb{R}^n$ be a non-zero vector. Then,

$$\left(\omega^\top \Sigma^{-1} \omega\right)^{-1} = \min_{\langle v, \omega\rangle = 1} v^\top \Sigma v. \tag{4.57}$$

Proof. Since the optimization problem in (4.57) is a convex program with continuously differentiable objective function and constraint, so its minimal value can be obtained solving the KKT equations. That is, there are $\hat{\lambda} \geq 0$ and $\hat{v}$ such that

$$2\Sigma\hat{v} - \hat{\lambda}\omega = 0, \quad \hat{\lambda}\left(\langle \hat{v}, \omega\rangle - 1\right) = 0.$$

Solving the above set of equations yields, $\hat{v} = \frac{\Sigma^{-1}\omega}{\omega^\top \Sigma^{-1}\omega}$. The desired result will be established by replacing $\hat{v}$ into the right hand side of (4.57). □

Lemma 4.5. Let K be a covariance function such that $\left\|\hat{K}'\right\|_\infty < \infty$ and define $G_\beta : \mathbb{R} \mapsto [0, 1]$ by (4.44). Then, there is a universal constant $c > 0$ such that

$$\inf_{\beta\in(0,1)} \int_{-\infty}^{\infty} \hat{K}(\omega) G_\beta(\omega)\, d\omega \geq c.$$

Proof. Observe that for any $\omega \in \mathbb{R}$, $G_\beta(\omega)$ is a quadratic function of β in the compact interval $[0, 1]$ and $\lim_{n\to\infty} \left\|G_{\beta_n} - G_\beta\right\|_\infty = 0$ for any convergent sequence $\beta_n \to \beta$. This property implies that

$$\inf_{\beta\in(0,1)} \int_{-\infty}^{\infty} \hat{K}(\omega) G_\beta(\omega)\, d\omega \geq \frac{1}{2}\left[\inf_{\beta\in(0,1),\, |\beta - 1/2|\geq r} \int_{-\infty}^{\infty} \hat{K}(\omega) G_\beta(\omega)\, d\omega \wedge \int_{-\infty}^{\infty} \hat{K}(\omega) G_{0.5}(\omega)\, d\omega\right]. \tag{4.58}$$

for some sufficiently small $r > 0$. Observe that, $G_\beta(0) = (1-2\beta)^2 > 0$ for $\beta \neq 1/2$. The differentiability of G_β and $\hat{K}(\omega)$ implies the existence of a non-degenerate open interval $\mathcal{I}_\beta$ centered at 0 such that,

$$\inf_{\omega \in \mathcal{I}_\beta} \hat{K}(\omega) G_\beta(\omega) \geq \frac{(1-2\beta)^2 \hat{K}(0)}{2} \quad \Rightarrow \quad \int_{-\infty}^{\infty} \frac{\hat{K}(\omega) G_\beta(\omega)}{2\pi} d\omega \geq \frac{(1-2\beta)^2 \hat{K}(0)}{4\pi} \left|\mathcal{I}_\beta\right|.$$

Notice that $\inf_{|\beta-1/2|\geq r} (1-2\beta)^2 \left|\mathcal{I}_\beta\right| > 0$. So, we just need to show that the corresponding term to $\beta = 1/2$ in the right hand side of (4.58) is strictly positive. For $\beta = 1/2$, $G_\beta(\omega) = [\mathrm{sinc}(\omega/4)\sin(\omega/2)]^2$ and so

$$\begin{aligned} \int_{-\infty}^{\infty} \frac{\hat{K}(\omega) G_\beta(\omega)}{2\pi} d\omega &\geq \int_{-2\pi}^{2\pi} \frac{\hat{K}(\omega) [\mathrm{sinc}(\omega/4)\sin(\omega/2)]^2}{2\pi} d\omega \\ &\overset{(b)}{\geq} \frac{2}{\pi^3} \int_{-2\pi}^{2\pi} \hat{K}(\omega) \sin^2(\omega/2)\, d\omega \overset{(c)}{>} 0. \end{aligned}$$

Note that (b) is a consequence of monotonicity of $\mathrm{sinc}(\cdot)$ in the interval $(0, \pi/2)$ and inequality (c) follows from the combination of $\left|\hat{K}'(0)\right| < \infty$ and $\hat{K}(0) > 0$. □

4.10 Change-Point Detection in the Increasing Domain Regime

In this section we briefly investigate the detection rate of both CUSUM and MLE algorithms in the increasing domain setting. Due to the space constraint, the proofs of all the results appearing in this section are omitted. We refer the reader to [KSN17] for the detailed proofs. Before proceeding further, we present the covariance structure of our GP model in the increasing domain framework.

In the increasing domain setting, $G - \mathbb{E}G$ is a mean-zero GP in $\mathcal{D} = [0, \infty)$ and $\mathcal{D}_n = \{1, 2, \ldots\}$. Define $\mathrm{cov}(X_1, X_k) = f_k$ for any k, in which $\{f_m\}_{m=0}^{\infty}$ is an absolutely summable sequence with $f_0 = 1$. Due to the stationarity assumption, $\Sigma_{\mathbb{N}} \coloneqq \mathrm{cov}\left(\{X_k\}_{k=1}^{\infty}\right)$ is an infinite symmetric Toeplitz matrix. We view $\{X_k\}_{k=1}^{n}$ as the observed part of an infinite stationary time series, $\{X_k\}_{k=1}^{\infty}$. Accordingly, the covariance matrix of $\{X_k\}_{k=1}^{n}$, denoted by Σ_n, is a symmetric (truncated) Toeplitz matrix.

It is a known fact (Chapter 4, [G$^+$06]) that there is a symmetric and almost surely (with respect to *Lebesgue* measure) positive function, $f : [-\pi, \pi] \mapsto \mathbb{R}$ such that $\Sigma_{\mathbb{N}} = \mathcal{T}_{\mathbb{N}}(f)$.

Thus $\Sigma_n = \mathcal{T}_n(f)$. For studying the asymptotic properties of the change detection algorithm, certain regularity conditions are required on f.

Assumption 4.3. $f : [-\pi, \pi] \mapsto \mathbb{R}$ is a real symmetric function such that

(a) There are two positive universal scalars, $0 < m_f \leq M_f < \infty$ such that

$$m_f \coloneqq \inf_{\omega \in [-\pi,\pi]} f(\omega) \leq M_f \coloneqq \sup_{\omega \in [-\pi,\pi]} f(\omega).$$

(b) There exist positive constants c and λ such that

$$|f_k| \leq c(1+k)^{-(1+\lambda)}. \tag{4.59}$$

The first condition regarding the infimum of f is necessary to have a positive definite infinite covariance matrix, i.e., $v^\top \Sigma_{\mathbb{N}} v > 0$ for any non-zero $v \in \mathbb{R}^{\mathbb{N}}$. Moreover, the polynomial decay of f_k's as stated in (4.59) is a sufficient condition to ensure that f can be equivalently expressed by its Fourier series. Such condition is common in the non-asymptotic analysis of Toeplitz matrices (see, e.g., [G+06]). We now present a result describing the detection rate of the MLE algorithm with known covariance matrix (see Eq. (4.6)).

Theorem 4.7. Let $\delta \in (0,1)$ and suppose that $\Sigma_n = \mathcal{T}_n(f)$ in which f admits Assumption 4.3 for some positive scalars c and λ. There exist $n_0 \in \mathbb{N}$, $C > 0$ (depending only on c and λ) and $R_{n,\delta} > 0$ such that for any $n \geq n_0$, if

$$|b| \geq C\sqrt{f(0)\, n^{-1} \log\left(\frac{n(1-2\alpha)}{\delta}\right)}, \tag{4.60}$$

then

$$\varphi_n(T_{GLRT}) \leq \delta.$$

Some comments are in order. First, the threshold $R_{n,\delta}$ in Theorem 4.7 is chosen in exactly the same way as in the fixed domain setting, as given by Eq. (4.13). Second, in contrast to the fixed domain setting, the dependence structure for G no longer plays the central role in the characterization of detection performance. In particular, $f(0)$ is the only factor in (4.60) that captures the correlation in the samples, but this scalar quantity evidently has an insignificant effect: the asymptotic behaviour of MLE remains the same (up to some constant factor) for different GPs satisfying Assumption 4.3. A related observation that arises by comparing between (4.12) and (4.60) is that the correlation structure of observations,

which is encapsulated into ν or $f(0)$, and the quantities encoding the marginal density information such as n have been completely decoupled in the rate of the MLE in the increasing domain. An examination of the proof reveals that the decoupling effect in the increasing domain setting arises due to the short-range correlation assumption ($\mathrm{cov}(X_r, X_s) \to 0$ polynomially in $|r-s|$). It follows that as n increases the correlation for most pairs of observed sample become negligible.
Now we aim to study the CUSUM test whose formulation is given in Eq. (4.2)

Theorem 4.8. Let $\delta \in (0,1)$, and $\mathcal{C}_{n,\alpha} = [\alpha n, (1-\alpha)n] \cap \mathbb{N}$. Assume that f satisfies Assumption 4.3 for some c and λ. There are $n_0 = n_0(f)$ and scalar $C(\lambda, c) > 0$, such that if $n \geq n_0$ and

$$|b| \geq C\sqrt{\frac{f(0)}{n\alpha(1-\alpha)}\log\left(\frac{n(1-2\alpha)}{\delta}\right)}, \tag{4.61}$$

then

$$\varphi_n(T_{CUSUM}) \leq \delta.$$

The presented detection rates in Theorems 4.7 and 4.8 reveals that the both CUSUM and MLE exhibit similar detection performance in the increasing domain setting. However, according to the numerical studies in Section 4.7, the MLE slightly outperforms the CUSUM test, especially in the presence of strong long range dependence.
In the sequel we give a condition on jump size $|b|$ according to which no algorithm in the increasing domain can properly detect the existence of a shift in the mean.

Theorem 4.9. Let $\delta \in (0,2)$, and $\mathcal{C}_{n,\alpha} = [\alpha n, (1-\alpha)n]$. Suppose that $\Sigma_n = \mathcal{T}_n(f)$ in which f satisfies Assumption 4.3. There exist $n_0 \coloneqq n_0(f)$ and $C > 0$ such that if $n \geq n_0$ and

$$|b| \leq C\sqrt{\frac{(1+\vartheta) f(0) \log\left(\frac{1}{\delta(2-\delta)}\right)}{\alpha n}},$$

then for any test T,

$$\varphi_n(T) \geq \delta.$$

The direct comparison between the detection rate of both CUSUM (in Theorem 4.8) and MLE (see Theorem 4.7) test with the above result indicates the minimax optimality (up to some order $\log n$ term) of both of these procedures in the increasing domain setting.

CHAPTER 5

Future Works

In this chapter we succinctly describe the potential ways of extending the problems studies in this thesis.

- *Adjusting the LIF loss function for non-stationary processes with smoothly varying variance and range parameters*: The main idea is to partition the set of sampling sites $\mathcal{D}_n$ into b_n small bins, so that the GP inside each bin can be well approximated by an stationary process. For any $\boldsymbol{s} \in \mathcal{D}_n$, construct the set $\mathcal{N}_m(\boldsymbol{s})$ using the nearest neighbours of $\boldsymbol{s}$ inside its associated bin. The vectors variance and range parameters, denoted by $\boldsymbol{\phi}_0 = [\phi_{0,1}, \ldots, \phi_{0,b_n}]^\top$ and $\boldsymbol{\rho}_0 = [\rho_{0,1}, \ldots, \rho_{0,b_n}]^\top$, can be simultaneously estimated by optimizing a penalized LIF objective function. Strictly speaking,

$$\left(\hat{\boldsymbol{\phi}}_{n,\mathcal{B}}, \hat{\boldsymbol{\rho}}_{n,\mathcal{B}}\right) = \underset{\boldsymbol{\phi},\boldsymbol{\rho}}{\arg\min} \left\{ \sum_{t=1}^{b_n} \left\| Y_{B_t,m} Y_{B_t,m}^\top - \phi_t K_{B_t,m}(\rho_t) \right\|_{\ell_2}^2 + J_{\boldsymbol{\phi}}(\phi_1, \ldots, \phi_{b_n}) + J_{\boldsymbol{\rho}}(\rho_1, \ldots, \rho_{b_n}) \right\},$$

  in which $J_{\boldsymbol{\phi}}$ and $J_{\boldsymbol{\rho}}$ are non-negative functions penalizing the rapidly varying variance and range parameters. Such penalized loss function may be optimized using the coordinate descent method.

- *LIF estimation for the multi-dimensional Gaussian processes*: Interpolation of vector-valued spatial Gaussian processes has been rarely studied in the statistics literature. For instance developing scalable estimation algorithms for such processes is a major computational and scientific challenges for the oceanographers. *Argo* project, which is an international collaborative partnership of more than 30 countries, has been operational since the early 2000s. More than 3000 free-drifting profiling Argo floats measure the temperature and salinity over 3° latitude by 3° longitude. Each float is launched as deep as 2000 meter below the ocean surface to monitor the temperature and salinity. After some suitable data preprocessing, the collected data can be modelled using a vector-valued Gaussian process of the form $\left[G_{\text{sal}}(\boldsymbol{s}), G_{\text{temp}}(\boldsymbol{s})\right]$ in a three dimensional space

(latitude by longitude by depth). Evaluating the full likelihood function for such spatial process is almost infeasible, as we have discussed in Chapters 2 and 3. An interesting future path is to design an inversion-free algorithm for simultaneous interpolation of the salinity and temperature fields. Note that due to the mutual correlation between G_{sal} and G_{temp} (which can not be neglected), generalizing the LIF estimator is not a trivial task.

- *The smoothness parameter adaptation in the LIF estimator*: Recall that the preconditioning order m in the LIF class of algorithms should satisfy the condition $m \geq \nu + d/2$. Increasing m directly affects the computational complexity of the proposed estimation algorithm, which is more pronounced for large data sets. Thus a prior knowledge of the smallest possible preconditioning order ($m = \lceil \nu + d/2 \rceil$) is imperative, particularly for practical scenarios. We assumed to fully know ν in our fixed domain asymptotic analysis in chapter 3, which is unlikely realistic. A thorough asymptotic and algorithmic analysis remains to be done on adjusting the LIF algorithm for the case of unknown ν. For stationary GPs, a consistent estimator of ν has been proposed in the literature. So a first possible choice of m can be $\lceil \hat{\nu} + d/2 \rceil$, in which $\hat{\nu}$ stands for the estimate of ν given the data. However adapting the LIF estimator to the case of unknown smoothness parameter for the smoothly varying non-stationary processes is yet to be discovered.

- *Change-zone detection in spatial GPs*: A systematic study of change-zone detection methods for multi-dimensional GPs remains unavailable. The fixed domain setting is clearly the natural way to study asymptotic behaviour of detection procedures for these spatial processes. The one dimensional abrupt change model in Chapter 4 can be easily extended to a multi-dimensional setting. Consider a spatial GP G in a bounded domain $\mathcal{D} \subset \mathbb{R}^d$ whose mean function can be formulated as the following:

$$\mathbb{E}G(\boldsymbol{s}) = \mu_0 \mathbb{1}_{s \in \Omega} + \mu_1 \mathbb{1}_{s \notin \Omega}, \ \ \forall\, \boldsymbol{s} \in \mathbb{R}^d. \tag{5.1}$$

  Here $\mu_0 \neq \mu_1$ are unknown scalars and $\Omega \subset \mathcal{D}$ denotes a zone with constant mean. Aside from the sample size and smoothness of the covariance function, the geometric properties of Ω also has a crucial role in the design and analysis of detection algorithms.

- *Detecting abrupt changes in more complex Gaussian models*: Taking the dependence structure of the Gaussian time series into account can significantly improve the performance of abrupt change detectors. Using the plug-in GLRT approach for detecting the abrupt changes in more complicated spatial-temporal processes, such as time varying Gaussian graphical models, can be very computationally challenging. Scalable surrogate statistics to likelihood ratio function for sequential and off-line change-point detection in

high-dimensional Gaussian graphical models is another viable future research direction. Employing pseudo-likelihood based approaches is a tractable approach for alleviating the computational burden of detection algorithms in complex dependent structures.

BIBLIOGRAPHY

[ABBD+10] Louigi Addario-Berry, Nicolas Broutin, Luc Devroye, Gábor Lugosi, et al. On combinatorial testing problems. *The Annals of Statistics*, 38(5):3063–3092, 2010.

[ACCD11] Ery Arias-Castro, Emmanuel J Candès, and Arnaud Durand. Detection of an anomalous cluster in a network. *The Annals of Statistics*, pages 278–304, 2011.

[ACS16] Mihai Anitescu, Jie Chen, and Michael L Stein. An inversion-free estimating equations approach for gaussian process models. *Journal of Computational and Graphical Statistics*, (just-accepted):1–42, 2016.

[ACW12] Mihai Anitescu, Jie Chen, and Lei Wang. A matrix-free approach for solving the parametric gaussian process maximum likelihood problem. *SIAM Journal on Scientific Computing*, 34(1):A240–A262, 2012.

[Ada98] Sudeshna Adak. Time-dependent spectral analysis of nonstationary time series. *Journal of the American Statistical Association*, 93(444):1488–1501, 1998.

[AFV14] S Borağan Aruoba and Jesús Fernández-Villaverde. A comparison of programming languages in economics. Technical report, National Bureau of Economic Research, 2014.

[AGB05] Denis Allard, Edith Gabriel, and Jean-Noël Bacro. Estimating and testing zones of abrupt change for spatial data. In *WNAR/IMS conference, Fairbanks*, 2005.

[AHP97] J Antoch, M Hušková, and Z Prášková. Effect of dependence on statistics for determination of change. *Journal of Statistical Planning and Inference*, 60(2):291–310, 1997.

[And10] Ethan Anderes. On the consistent separation of scale and variance for gaussian random fields. *The Annals of Statistics*, pages 870–893, 2010.

[AS+66] Milton Abramowitz, Irene A Stegun, et al. Handbook of mathematical functions. *Applied mathematics series*, 55(62):39, 1966.

[Bac14] François Bachoc. Asymptotic analysis of the role of spatial sampling for covariance parameter estimation of gaussian processes. *Journal of Multivariate Analysis*, 125:1–35, 2014.

[BFG11] Pierre Raphael Bertrand, Mehdi Fhima, and Arnaud Guillin. Off-line detection of multiple change points by the filtered derivative with p-value method. *Sequential Analysis*, 30(2):172–207, 2011.

[BGT87] NH Bringham, CM Goldie, and JL Teugels. Regular variation, cam-bridge univ. *Pres, Cambridge*, 1987.

[BI+13] Cristina Butucea, Yuri I Ingster, et al. Detection of a sparse submatrix of a high-dimensional noisy matrix. *Bernoulli*, 19(5B):2652–2688, 2013.

[Bir01] Lucien Birgé. An alternative point of view on lepski's method. *Lecture Notes-Monograph Series*, pages 113–133, 2001.

[BL11] David Bolin and Finn Lindgren. Spatial models generated by nested stochastic partial differential equations, with an application to global ozone mapping. *The Annals of Applied Statistics*, pages 523–550, 2011.

[BLNZ95] Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

[Che13] Jie Chen. On the use of discrete laplace operator for preconditioning kernel matrices. *SIAM Journal on Scientific Computing*, 35(2):A577–A602, 2013.

[Cre15] Noel Cressie. *Statistics for spatial data*. John Wiley & Sons, 2015.

[CV11] Varun Chandola and Ranga Raju Vatsavai. A gaussian process based online change detection algorithm for monitoring periodic time series. In *Proceedings of the 2011 SIAM International Conference on Data Mining*, pages 95–106. SIAM, 2011.

[CZ64] Herman Chernoff and Shelemyahu Zacks. Estimating the current mean of a normal distribution which is subjected to changes in time. *The Annals of Mathematical Statistics*, 35(3):999–1018, 1964.

[D+97] Rainer Dahlhaus et al. Fitting time series models to nonstationary processes. *The annals of Statistics*, 25(1):1–37, 1997.

[DP86] Jean Deshayes and Dominique Picard. Off-line statistical analysis of change-point models using non parametric and likelihood methods. *Detection of Abrupt Changes in Signals and Dynamical Systems*, pages 103–168, 1986.

[DZM+09] Juan Du, Hao Zhang, VS Mandrekar, et al. Fixed-domain asymptotic properties of tapered maximum likelihood estimators. *the Annals of Statistics*, 37(6A):3330–3361, 2009.

[Fel68] William Feller. *An introduction to probability theory and its applications: volume I*, volume 3. John Wiley & Sons New York, 1968.

[FGN06] Reinhard Furrer, Marc G Genton, and Douglas Nychka. Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics*, 15(3):502–523, 2006.

[G+06] Robert M Gray et al. Toeplitz and circulant matrices: A review. *Foundations and Trends® in Communications and Information Theory*, 2(3):155–239, 2006.

[GDFG10] Alan E Gelfand, Peter J Diggle, Montserrat Fuentes, and Peter Guttorp. *Handbook of spatial statistics*. CRC press, 2010.

[GER07] Olivier Gillet, Slim Essid, and Gal Richard. On the correlation of automatic audio and visual segmentations of music videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(3):347–355, 2007.

[HH12] Lajos Horváth and Marie Hušková. Change-point detection in panel data. *Journal of Time Series Analysis*, 33(4):631–648, 2012.

[HJ12] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

[Hor97] Lajos Horváth. Detection of changes in linear sequences. *Annals of the Institute of Statistical Mathematics*, 49(2):271–283, 1997.

[HQI07] Xiao Hu, Hai Qiu, and Naresh Iyer. Multivariate change detection for time series data in aircraft engine fault diagnostics. In *Systems, Man and Cybernetics, 2007. ISIC. IEEE International Conference on*, pages 2484–2489. IEEE, 2007.

[IR12] Ildar Abdulovič Ibragimov and Yurii Antol'evich Rozanov. *Gaussian random processes*, volume 9. Springer Science & Business Media, 2012.

[KL98] Piotr Kokoszka and Remigijus Leipus. Change-point in the mean of dependent observations. *Statistics & probability letters*, 40(4):385–393, 1998.

[KS09] Yoshinobu Kawahara and Masashi Sugiyama. Change-point detection in time-series data by direct density-ratio estimation. In *Proceedings of the 2009 SIAM International Conference on Data Mining*, pages 389–400. SIAM, 2009.

[KS+13] CG Kaufman, BA Shaby, et al. The role of the range parameter for estimation and prediction in geostatistics. *Biometrika*, 100(2):473–484, 2013.

[KSN08] Cari G Kaufman, Mark J Schervish, and Douglas W Nychka. Covariance tapering for likelihood-based estimation in large spatial data sets. *Journal of the American Statistical Association*, 103(484):1545–1555, 2008.

[KSN16] Hossein Keshavarz, Clayton Scott, and XuanLong Nguyen. On the consistency of inversion-free parameter estimation for gaussian random fields. *Journal of Multivariate Analysis*, 150:245–266, 2016.

[KSN17] Hossein Keshavarz, Clayton Scott, and XuanLong Nguyen. Optimal change point detection in gaussian processes. *arXiv preprint arXiv:1506.01338*, 2017.

[Lai98] Tze Leung Lai. Information bounds and quick detection of parameter changes in stochastic systems. *IEEE Transactions on Information Theory*, 44(7):2917–2929, 1998.

[Lav05] Marc Lavielle. Using penalized contrasts for the change-point problem. *Signal processing*, 85(8):1501–1510, 2005.

[Lee12] Myoungji Lee. *Local properties of irregularly observed Gaussian fields*, volume 74. 2012.

[LL+00] Wei-Liem Loh, Tao-Kai Lam, et al. Estimating structured correlation matrices in smooth gaussian random field models. *The Annals of Statistics*, 28(3):880–904, 2000.

[LlH08] Céline Levy-leduc and Zaïd Harchaoui. Catching change-points with lasso. In *Advances in Neural Information Processing Systems*, pages 617–624, 2008.

[LS08] Michael Last and Robert Shumway. Detecting abrupt changes in a piecewise locally stationary time series. *Journal of multivariate analysis*, 99(2):191–214, 2008.

[LYY10] Jun Liu, Lei Yuan, and Jieping Ye. An efficient algorithm for a class of fused lasso problems. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 323–332. ACM, 2010.

[Mac74] Ian B MacNeill. Tests for change of parameter at unknown times and distributions of some related functionals on brownian motion. *The Annals of Statistics*, pages 950–962, 1974.

[Mir75] Leon Mirsky. A trace inequality of john von neumann. *Monatshefte für mathematik*, 79(4):303–306, 1975.

[MM84] Kanti V Mardia and RJ Marshall. Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, pages 135–146, 1984.

[NRK12] Hae Young Noh, Ram Rajagopal, and Anne S Kiremidjian. Damage diagnosis algorithm using a sequential change point detection method with an unknown distribution for damage. In *SPIE Smart Structures and Materials+*

Nondestructive Evaluation and Health Monitoring, pages 834507–834507. International Society for Optics and Photonics, 2012.

[O'L80] Dianne P O'Leary. The block conjugate gradient algorithm and related methods. *Linear algebra and its applications*, 29:293–322, 1980.

[REN09] MONIKA RENCOVA. *CHANGE-POINT DETECTION IN TEMPERATURE SERIES*. PhD thesis, Doctoral dissertation, Czech Technical University, 2009.

[Rig10] Guillem Rigaill. Pruned dynamic programming for optimal multiple change-point detection, 2010.

[RRR12] Tata Subba Rao, Suhasini Subba Rao, and Calyampudi Radhakrishna Rao. *Handbook of statistics: time series analysis: methods and applications*, volume 30. Elsevier, 2012.

[RV+13] Mark Rudelson, Roman Vershynin, et al. Hanson-wright inequality and sub-gaussian concentration. *Electron. Commun. Probab*, 18(82):1–9, 2013.

[SCA12] Michael L Stein, Jie Chen, and Mihai Anitescu. Difference filter preconditioning for large covariance matrices. *SIAM Journal on Matrix Analysis and Applications*, 33(1):52–72, 2012.

[SCA+13] Michael L Stein, Jie Chen, Mihai Anitescu, et al. Stochastic approximation of score functions for gaussian processes. *The Annals of Applied Statistics*, 7(2):1162–1191, 2013.

[SCW04] Michael L Stein, Zhiyi Chi, and Leah J Welty. Approximating likelihoods for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(2):275–296, 2004.

[SHC02] Xiaotong Shen, Hsin-Cheng Huang, and Noel Cressie. Nonparametric hypothesis testing for a spatial signal. *Journal of the American Statistical Association*, 97(460):1122–1140, 2002.

[Ste12] Michael L Stein. *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media, 2012.

[SY] Anatolii Volodimirovich Skorokhod and Mikhaĭ Yadrenko. On absolute continuity of measures corresponding to homogeneous gaussian fields.

[SZ10] Xiaofeng Shao and Xianyang Zhang. Testing for change points in time series. *Journal of the American Statistical Association*, 105(491):1228–1240, 2010.

[TRBK06] Alexander G Tartakovsky, Boris L Rozovskii, Rudolf B Blazek, and Hongjoong Kim. A novel approach to detection of intrusions in computer networks via adaptive sequential and batch-sequential change-point detection methods. *IEEE Transactions on Signal Processing*, 54(9):3372–3382, 2006.

[Tsy09] Alexandre B Tsybakov. Introduction to nonparametric estimation. revised and extended from the 2004 french original. translated by vladimir zaiats, 2009.

[Vec88] Aldo V Vecchia. Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 297–312, 1988.

[VHNC10] Jan Verbesselt, Rob Hyndman, Glenn Newnham, and Darius Culvenor. Detecting trend and seasonal changes in satellite image time series. *Remote sensing of Environment*, 114(1):106–115, 2010.

[Wen95] Holger Wendland. Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. *Advances in computational Mathematics*, 4(1):389–396, 1995.

[WL+11] Daqing Wang, Wei-Liem Loh, et al. On fixed-domain asymptotics and covariance tapering in gaussian random field models. *Electronic Journal of Statistics*, 5:238–269, 2011.

[WLX13] Wei-Ying Wu, Chae Young Lim, and Yimin Xiao. Tail estimation of the spectral density for a stationary gaussian random field. *Journal of Multivariate Analysis*, 116:74–91, 2013.

[YD86] Yi-Ching Yao and Richard A Davis. The asymptotic behavior of the likelihood ratio statistic for testing a shift in mean in a sequence of independent normal variates. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 339–353, 1986.

[Yin91] Zhiliang Ying. Asymptotic properties of a maximum likelihood estimator with data from a gaussian process. *Journal of Multivariate Analysis*, 36(2):280–296, 1991.

[Zha04] Hao Zhang. Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association*, 99(465):250–261, 2004.