Electronic Journal of Statistics
Vol. 13 (2019) 4224–4279
ISSN: 1935-7524
https://doi.org/10.1214/19-EJS1592

# Local inversion-free estimation of spatial Gaussian processes*,†

**Hossein Keshavarz and XuanLong Nguyen**

*Institute for Mathematics and its Applications, University of Minnesota, Minneapolis*
*Department of Statistics, University of Michigan, Ann Arbor*
*e-mail:* hkeshava@umn.edu; xuanlong@umich.edu

**Clayton Scott**

*Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor*
*e-mail:* clayscot@umich.edu

**Abstract:** Maximizing the likelihood has been widely used for estimating the unknown covariance parameters of spatial Gaussian processes. However, evaluating and optimizing the likelihood function can be computationally intractable, particularly for large number of (possibly) irregularly spaced observations, due to the need to handle the inverse of ill-conditioned and large covariance matrices. Extending the "inversion-free" method of Anitescu, Chen and Stein [1], we investigate a broad class of covariance parameter estimation based on inversion-free surrogate losses and block diagonal approximation schemes of the covariance structure. This class of estimators yields a spectrum for negotiating the trade-off between statistical accuracy and computational cost. We present fixed-domain asymptotic properties of our proposed method, establishing $\sqrt{n}$-consistency and asymptotic normality results for isotropic Matern Gaussian processes observed on a multi-dimensional and irregular lattice. Simulation studies are also presented for assessing the scalability and statistical efficiency of the proposed algorithm for large data sets.

**MSC 2010 subject classifications:** Primary 62M30, 62M40; secondary 60G15.
**Keywords and phrases:** Local inversion-free covariance estimation, Gaussian process, computationally scalable, fixed-domain asymptotic analysis, irregularly spaced observations.

Received November 2018.

## 1. Introduction

Gaussian processes (GPs) are one of the most common modelling tools for the analysis of spatiotemporal data (see e.g., [6, 8]). A crucial aspect of GP-based inference is the estimation of its covariance function. The covariance function is

*This research is partially supported by NSF grant ACI-1047871. Additionally, CS is partially supported by NSF Grants 1838179 and 1422157, and LN by NSF CAREER award DMS-1351362, NSF CNS-1409303, and NSF CCF-1115769.
†We are grateful to Mihai Anitescu and Michael Stein for valuable discussions and suggestions.

typically specified up to a finite number of parameters, the estimation of which is pivotal for performing interpolation and prediction tasks.

While there are a number of likelihood-based techniques for covariance estimation, they do not scale well. Indeed, exact evaluation of the Gaussian likelihood requires computing the inverse of the covariance matrix, which generally requires $\mathcal{O}\left(n^3\right)$ operations and $\mathcal{O}\left(n^2\right)$ space storage. A number of authors have proposed ways of getting around this challenge, by working instead with an approximate version of the likelihood function. Vecchia [20] considered an approximation by ignoring the conditional correlation of distant sites given their nearest neighbours. This idea was further extended by Stein et al. [19] who studied more flexible choices of conditioning sets. The key to evaluating the exact log-likelihood function and its partial derivatives boils down to solving large and dense systems of linear equations. To accelerate such linear solvers, e.g., using the *Krylov subspace iteration method*, Furrer et al. [7] and Kaufman et al. [10] exploit the tapering technique to sparsify the dense covariance matrix. More recently, several authors investigated a stochastic optimization technique for implementing the MLE [3, 18]. Their proposed algorithms are statistically comparable to MLE, if the condition number of the covariance matrix has a uniform upper bound (independent of the sample size).

An attractive alternative to likelihood based techniques is to abandon the likelihood function altogether, and consider instead surrogate loss functions which may be evaluated and optimized more efficiently. Anitescu, Chen and Stein [2] proposed one such surrogate loss based method for covariance estimation, and showed that it is considerably computationally more efficient than the standard MLE, especially for irregularly spaced observations. Indeed, their loss function, which we call *inversion-free (IF)* in this manuscript, does not require computing the precision matrix (covariance inverse), and so it can be evaluated in $\mathcal{O}\left(n^2\right)$ time. It was established by the authors that when the covariance matrix has a bounded condition number, the resulting estimate possesses consistency and asymptotic normality [11]. It is noted that the boundedness of condition number holds in the increasing domain setting, where the minimum distance among the sampling points is bounded away from zero. This is in contrast to the scenarios in which the GP is observed in a fixed and bounded domain, where the observations get denser as the sample size $n$ increases. In this new regime, which is referred to as *fixed-domain* (or *infill*) setting, because of strong spatial correlation the condition number often grows without bound with $n$. This points to an unresolved question regarding the statistical efficiency of the inversion-free algorithm in the fixed-domain setting, including the situation of irregularly spaced observations.

In this article, we adopt and extend the basic surrogate loss based approach of [2], while looking to address the theoretical questions described above. A natural adaptation of the IF loss function is to apply it to a transformed version of the data using a transformation technique that helps to reduce the strong correlation among the (original) observations. A fast and root-$n$ consistent estimator studied by Anderes [1] can be viewed this way, as it is based on squared increments of the observed Gaussian process. In his work samples are transformed

using directional increments of the Gaussian process. However this method is applicable only to regularly spaced observations. A general scheme for dependence reduction, which we refer to as *preconditioning*, was introduced in [5, 17] and chapter 3 of [14]. The preconditioning technique is one of the building blocks of our proposed estimation algorithm. It will be shown that this preconditioner provides a suitable transformation in the case of irregularly placed observations.

The second ingredient of our approach is to apply a divide-and-conquer technique to design of the surrogate loss function, which will be referred to as the *local inversion-free (LIF)* loss. Specifically, the (preconditioned) samples are divided into $b_n$ possibly overlapping clusters (bins). The LIF loss is composed by taking a weighted average of the IF loss functions over these bins. The covariance estimates are obtained by optimizing with respect to the LIF loss function. The aforementioned preconditioning technique is crucial for the statistical efficiency of the LIF algorithm as it helps reduce the correlation between distant clusters.

The resulting LIF procedure comprises a rich and flexible class of estimation algorithms, depending on the number of bins $b_n$, and specific binning scheme determined by the size and shape of each bin. When $b_n = 1$, our algorithm reduces to the inversion-free method of [2], but applied with the preconditioning scheme that we will describe. Furthermore, the quadratic variation-based approach of [1] is a special instance in the LIF class, specifically corresponding to the other extreme scenario of $b_n = n$. Thus, the LIF class can be viewed as a spectrum of algorithms bridging between two distinct approaches in the literature. A noted advantage of our procedure in exploiting the divide and conquer strategy is to significantly expedite the estimation procedure, while preserving favorable statistical properties. Indeed, the LIF loss can be evaluated in order $n^2/b_n \ll n^2$ operations.

A considerable portion of this article is devoted to the investigation of the asymptotic behavior of the proposed LIF based estimation method in the fixed-domain regime. Theoretical analysis for several specific instances of LIF based estimation have been carried out before, by [1] on his quadratic variation based method on regularly spaced observations in the fixed-domain framework, and in the increasing domain regime by the authors [11]. The asymptotic theory for the fixed-domain regime is considerably more involved than the increasing domain regime, especially for irregularly spaced observations.

It is established by [23] that for the isotropic Matern GP, the variance $\phi$ and the range parameter $\rho$ are not identifiable when dimension $d \leq 3$. Thus we only concentrate on estimating the so-called *microergodic* parameter (see page 163 of [16] for the exact definition), namely $\phi\rho^{-2\nu}$ where $\nu$ quantifies the smoothness of GP. The microergodic parameter is of great interest as it determines the asymptotic mean square estimation error in the fixed-domain setting (e.g., pages 174–175 of [16]). We show that under some regularity conditions 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 establish the asymptotic normality of this estimate. Hence, the LIF loss does not sacrifice asymptotic rate for increasing the computational speed and memory efficiency, even for irregularly spaced observations.

The treatment of observations on irregular lattices distinguishes our theoretical contribution from the previous works of [1, 21, 22].

Following the theoretical study, a comprehensive set of synthetic numerical experiments are conducted for assessing the role of preconditioning, the irregularity of sampling locations, and the binning scheme in the performance of the LIF estimate. Despite the robustness of the asymptotic rate to changes of $b_n$ and the shape of the bins, such factors can still affect the bias and variance of the LIF estimator, particularly for moderate sample sizes. Our simulation studies serve to corroborate the asymptotic theory, but also reveal the stability of the LIF estimate with respect to the size and shape of the bins. We evaluate the efficiency of our method for data sets up to $2.5 \times 10^5$ data points.

**Plan of the paper** Section 2 describes the geometry of sampling sites, preconditioning, and the IF method. In Section 3, we propose the family of the LIF loss functions and introduce an efficient parallel technique for evaluating such functions. Section 4 establishes the infill asymptotic properties of the LIF algorithm such as $\sqrt{n}$-consistency and asymptotic normality, given samples in a $d$-dimensional space with $d \leq 3$. In Section 5 we present a series of simulation studies to assess the performance of the LIF estimator. Section 6 serves as the conclusion and discusses future directions. We substantiate the main results of the paper in Section 7. Finally, Appendices A and B not only contain some auxiliary technicalities which are crucial in Section 7, but also present a comprehensive sensitivity analysis of the correlation matrix of the preconditioned data with respect to the range parameter, which may be 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 the all zeros column vector of length $m$. Furthermore, for any $p \in \{1, \ldots, m\}$, $\boldsymbol{e}_p$ denotes the unit vector along the $p^{\text{th}}$ coordinate. If $\boldsymbol{u}$ and $\boldsymbol{v}$ are vectors of length $m$, then $\boldsymbol{u}^{\boldsymbol{v}}$ denotes $\Pi_{i=1}^m u_i^{v_i}$ (we define $0^0$ to be 1). For square matrices $A$ and $B$ of the same size, by writing $A \succeq B$, we mean that $A - B$ is symmetric positive semi-definite. Furthermore, $\langle A, B \rangle \coloneqq \operatorname{tr}\left(A^\top B\right)$ refers to their trace inner product. We use various types of matrix norms on $A \in \mathbb{R}^{n \times n}$ in this paper. For any $p \in [1, \infty)$, $\|A\|_{\ell_p} \coloneqq \left(\sum_{i,j} |A_{ij}|^p\right)^{1/p}$ 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. We also write $\operatorname{diam}(\Omega) = \sup_{\omega_1, \omega_2 \in \Omega} \|\omega_2 - \omega_1\|_{\ell_2}$ 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 $A = U \Lambda U^\top$, $\sqrt{A} \coloneqq U \Lambda^{1/2} U^\top$ 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 $\nu$ and the Gamma function.

## 2. Preconditioning and inversion-free surrogate loss

### 2.1. Gaussian processes observed on irregular lattices

Consider a zero mean, real valued, and stationary Gaussian process $G$ on domain $\mathcal{D}$, where $\mathcal{D}$ is a bounded subset of $\mathbb{R}^d$ such as $[0,1]^d$. The dependence structure of $G$ is typically parametrized by a variance parameter $\phi_0 > 0$ and a (correlation) range parameter $\rho_0$. Specifically, 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

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

The objective is to estimate the microergodic parameters of the covariance function, given $n$ measurements from one realization of $G$ at locations $\mathcal{D}_n = \{\boldsymbol{s}_1,\ldots,\boldsymbol{s}_n\} \subset \mathcal{D}$. Throughout the paper, we assume that $\rho_0$ belongs to a compact, connected space $\Theta_0$ (with respect to the Euclidean distance). We also restrict $d$ to be less than or equal 3.

As the first step we precisely formulate $\mathcal{D}_n$. $\mathcal{D}_n$ is called a $d$-dimensional regular (rectangular) lattice with $n = N^d$ point, 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 into irregular ones, which can be formalized as follows (see [14]):

**Assumption 2.1.** Let $\mathcal{D}_n \subset \mathcal{D}$ be a set of size $n$. For any $\boldsymbol{s} \in \mathcal{D}_n$, let $r_{\boldsymbol{s},i}$ denote the distance from $\boldsymbol{s}$ to its $i^{\text{th}}$ closest neighbor in $\mathcal{D}_n \setminus \{\boldsymbol{s}\}$. There are positive scalars $C_{\min}$ and $C_{\max}$ such that

$$C_{\min}\left(\frac{i}{n}\right)^{\frac{1}{d}} \le r_{\boldsymbol{s},i} \le C_{\max}\left(\frac{i}{n}\right)^{\frac{1}{d}}, \quad \forall\, \boldsymbol{s}\in\mathcal{D}_n, \text{ and } i = 1,\cdots,(n-1). \quad (2.1)$$

The properties required by the assumption enlarge the notion of regular lattice in three aspects. First, in contrast to the number of points in a regular lattice, there is no restriction on $n$. Moreover, $\mathcal{D}$ is not restricted to be $[0,1]^d$. For instance, $\mathcal{D}$ might be the union of a finite number of connected components, as long as each of them satisfy condition (2.1) and encompasses a non-vanishing fraction of samples, as $n$ tends to infinity. Finally, $\mathcal{D}_n$ needs not form a $d$-dimensional regular lattice.

### 2.2. Preconditioning

Controlling the strong spatial dependence between the observed samples $\{G(\boldsymbol{s}_1),\ldots,G(\boldsymbol{s}_n)\}$ via preconditioning is essential for reducing the condition

number of the covariance matrix. It plays a crucial role in the estimation procedure we will propose. Various types of preconditioners have been studied for GPs observed on regular and irregular lattices in the literature (see e.g., [5, 14, 17]).

We shall adopt a preconditioning scheme proposed by Lee [14] for irregularly spaced observations. Before proceeding further, it is convenient to define $N \coloneqq \lfloor n^{1/d} \rfloor$. Furthermore for any $\boldsymbol{s} \in \mathcal{D}_n$, $\mathcal{N}_m(\boldsymbol{s})$ represents a set points (in $\mathcal{D}_n$) in a small neighbourhood of radius $\mathcal{O}\left(N^{-1}\right)$ around $\boldsymbol{s}$ whose size depends on $m$. Namely, $\|\boldsymbol{t}-\boldsymbol{s}\|_{\ell_2} \lesssim 1/N$ for any $\boldsymbol{t} \in \mathcal{N}_m(\boldsymbol{s})$.

**Definition 2.1.** Let $m \in \mathbb{N}$ (which does not grow with $n$). Suppose that there are sets of real coefficients $\{a_{m,\boldsymbol{s}}(\boldsymbol{t}) : \boldsymbol{t} \in \mathcal{N}_m(\boldsymbol{s})\}$, $\boldsymbol{s} \in \mathcal{D}_n$, satisfying the following conditions:

1. For any $\boldsymbol{r} \in \mathbb{Z}_+^d$ (the entries of $\boldsymbol{r}$ are non-negative) and $\|\boldsymbol{r}\|_{\ell_1} < m$, $\sum_{\boldsymbol{t} \in \mathcal{N}_m(\boldsymbol{s})} a_{m,\boldsymbol{s}}(\boldsymbol{t})(\boldsymbol{t}-\boldsymbol{s})^{\boldsymbol{r}} = 0$.
2. There is $\boldsymbol{r} \in \mathbb{Z}_+^d$ with $\|\boldsymbol{r}\|_{\ell_1} \geq m$ such that $\sum_{\boldsymbol{t} \in \mathcal{N}_m(\boldsymbol{s})} a_{m,\boldsymbol{s}}(\boldsymbol{t})(\boldsymbol{t}-\boldsymbol{s})^{\boldsymbol{r}} \neq 0$.
3. $\sum_{\boldsymbol{t} \in \mathcal{N}_m(\boldsymbol{s})} a_{m,\boldsymbol{s}}^2(\boldsymbol{t}) = 1$ and $a_{m,\boldsymbol{s}}(\boldsymbol{t}) \neq 0$ for all $\boldsymbol{t} \in \mathcal{N}_m(\boldsymbol{s})$.

We say $G_m$ is a *preconditioned process of order* $m$, if

$$G_m(\boldsymbol{s}) \coloneqq N^{\nu} \sum_{\boldsymbol{t} \in \mathcal{N}_m(\boldsymbol{s})} a_{m,\boldsymbol{s}}(\boldsymbol{t}) G(\boldsymbol{t}), \quad \forall\, \boldsymbol{s} \in \mathcal{D}_n. \tag{2.2}$$

**Remark 2.1.** Since $\mathcal{N}_m(\boldsymbol{s})$ is constructed by the nearest neighbors of $\boldsymbol{s}$, 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 2.1. 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 a simple combinatorial argument that the first condition in Definition 2.1 is translated as $\binom{d+m-1}{d}$ linear constraints on the 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 $\mathcal{N}_m(\boldsymbol{s})$.

**Remark 2.2.** A preconditioning method for the $d$-dimensional regular lattices $\mathcal{D}_n = \{1/N, \ldots, 1\}^d$ has been studied in Stein et al. [17]. 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 via

$$\begin{aligned} &G_0(\boldsymbol{s}) = N^{\nu} G(\boldsymbol{s}), \quad \forall\, \boldsymbol{s} \in \mathcal{D}_n, \\ &G_{2k}(\boldsymbol{s}) = \sum_{r=1}^{d} \left[ G_{2k-2}\left(\boldsymbol{s} + \frac{\boldsymbol{e}_r}{N}\right) - 2G_{2k-2}(\boldsymbol{s}) + G_{2k-2}\left(\boldsymbol{s} - \frac{\boldsymbol{e}_r}{N}\right) \right], \\ &\quad \boldsymbol{s} \in \mathcal{D}_n^{\circ},\ k = 1, \ldots, m. \end{aligned} \tag{2.3}$$

To avoid unnecessary algebraic complexity in Eq. (2.3), 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 2.1 with order $2m$. Namely, (2.3) gives a recursive way of constructing the preconditioned process of even orders for regular lattices. It is also worth mentioning that although $G_{2m}$ defined by (2.3) is a stationary process, preconditioning does not necessarily preserve stationarity for irregular lattices.

**Remark 2.3.** The preconditioned coefficients in Definition 2.1 are carefully chosen so that $G_m(\cdot)$ carries no information about the directional derivatives of $G$ of order less than $m$. Strictly speaking, the Taylor expansion of $G$ around $\boldsymbol{s}$ ensures the existence of an stochastic process $\Delta_m$ such that for any $\boldsymbol{t} \in \mathcal{N}_m(\boldsymbol{s})$,

$$G(\boldsymbol{t}) = \sum_{b=0}^{m-1} \sum_{\boldsymbol{r}\in\mathbb{Z}_+^d,\ \|r\|_{\ell_1}=b} \frac{1}{b!}\langle (\boldsymbol{t}-\boldsymbol{s})^{\boldsymbol{r}}, D^{\boldsymbol{r}}G(\boldsymbol{s})\rangle + \Delta_m(\boldsymbol{t}).$$

Here $D^{\boldsymbol{r}}G(\cdot)$ denotes the $\boldsymbol{r}^{\text{th}}$ directional derivative of $G$. Replacing this representation of $G$ into $G_m$ yields

$$\begin{aligned} G_m(\boldsymbol{s}) &= N^{\nu} \sum_{b=0}^{m-1} \sum_{\boldsymbol{r}\in\mathbb{Z}_+^d,\ \|r\|_{\ell_1}=b} \frac{1}{b!}\Big\langle \sum_{\boldsymbol{t}\in\mathcal{N}_m(\boldsymbol{s})} a_{m,\boldsymbol{s}}(\boldsymbol{t})(\boldsymbol{t}-\boldsymbol{s})^{\boldsymbol{r}}, D^{\boldsymbol{r}}G(\boldsymbol{s})\Big\rangle \\ &+ N^{\nu} \sum_{\boldsymbol{t}\in\mathcal{N}_m(\boldsymbol{s})} a_{m,\boldsymbol{s}}(\boldsymbol{t})\Delta_m(\boldsymbol{t}). \end{aligned}$$

The first condition in Definition 2.1 implies that

$$G_m(\boldsymbol{s}) = N^{\nu} \sum_{\boldsymbol{t}\in\mathcal{N}_m(\boldsymbol{s})} a_{m,\boldsymbol{s}}(\boldsymbol{t})\Delta_m(\boldsymbol{t}).$$

We finally present a concrete example satisfying the conditions in Definition 2.1. Note that Remark 2.2 constructs the preconditioning coefficients for regularly observed GPs. It is also easy to show that Definition is almost surely well-defined for randomly perturbed lattices (if the perturbation vector is absolutely continuous with respect to the Lebesgue measure). We refer the reader to Chapter 3 of [14] for further discussion.

### *2.3. The IF algorithm*

Anitescu, Stein and Chen [2] introduced a parameter estimation method based on an "inversion-free" surrogate loss for the Gaussian process that is both easy to compute and optimize. 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 covariance function of $G_m$ normalized by factor $\phi_0$. $K_m$ can be easily expressed in terms of the correlation function of $G$, $K(\cdot, \rho_0)$, and the preconditioning coefficients.

$$K_m(\boldsymbol{s},\boldsymbol{t};\rho_0) = \frac{\mathbb{E}G_m(\boldsymbol{s})G_m(\boldsymbol{t})}{\phi_0}$$

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

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

$$\mathbb{E} Y_m Y_m^\top = \phi_0 K_{n,m}(\rho_0) \coloneqq \phi_0 \left[K_m(\boldsymbol{s}, \boldsymbol{t}; \rho_0)\right]_{\boldsymbol{s},\boldsymbol{t} \in \mathcal{D}_n}. \tag{2.4}$$

Recall that $\rho_0$ lies in a compact and connected space $\Theta_0$. The IF estimator [2] of the covariance parameters $(\phi_0, \rho_0)$ is given by

$$\left(\hat{\phi}_n, \hat{\rho}_n\right) = \underset{\phi>0, \rho \in \Theta_0}{\arg\max} \left\{ \phi Y_m^\top K_{n,m}(\rho) Y_m - \frac{\phi^2}{2} \left\| K_{n,m}(\rho) \right\|_{\ell_2}^2 \right\}. \tag{2.5}$$

Note that (2.5) can be alternatively formulated as a moment matching minimization problem,

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

**Remark 2.4.** From a computational perspective, the loss function in (2.5) does not depend on the Cholesky factorization of $K_{n,m}$ and can be evaluated in order $n^2$ flops even for irregularly spaced observations. Moreover, storing the whole matrix $K_{n,m}$ is not necessary for computing the objective function and its directional derivatives. In particular, storing $Y_m$ and $\mathcal{D}_n$, which need $\mathcal{O}(n)$ storage, suffices for estimating the covariance parameters.

## 3. The local inversion-free (LIF) algorithm

We are ready to present in this section a broad class of scalable covariance estimation algorithms, building on the IF surrogate loss approach and the preconditioning technique described in the previous section. The asymptotic theory for our estimator will be presented in the following section.

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

$$\mathbb{E} Y_{B_t,m} Y_{B_t,m}^\top = \phi_0 K_{B_t,m}(\rho_0) \coloneqq \phi_0 \left[K_m(\boldsymbol{s}, \boldsymbol{t}; \rho_0)\right]_{\boldsymbol{s},\boldsymbol{t} \in B_t}, \quad \forall\, t = 1 \ldots, b_n, \tag{3.1}$$

in which $\phi_0 K_m(\cdot, \cdot, \rho_0)$ stands for the covariance function of $G_m$ with the parameters $(\phi_0, \rho_0)$.

The LIF objective function associated to a binning scheme $\mathcal{B}$ is constructed by summing the IF loss functions corresponding to the $B_t$'s over $\mathcal{B}$. The unknown covariance parameters are estimated by maximizing the LIF function, with

$$\left(\hat{\phi}_{n,\mathcal{B}}, \hat{\rho}_{n,\mathcal{B}}\right) = \underset{\phi>0, \rho \in \Theta_0}{\arg\max} \left\{ \sum_{t=1}^{b_n} \left( \phi Y_{B_t,m}^\top K_{B_t,m}(\rho) Y_{B_t,m} - \frac{\phi^2}{2} \left\| K_{B_t,m}(\rho) \right\|_{\ell_2}^2 \right) \right\}, \tag{3.2}$$

where $\hat{\phi}_{n,\mathcal{B}}$ and $\hat{\rho}_{n,\mathcal{B}}$ respectively denote the estimated variance and range parameters.

Several remarks are in order.

**Remark 3.1.** The LIF class of estimators can be enriched in 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. Given a $b_n$-dimensional vector of strictly positive entries $w \in \mathbb{R}^{b_n}$, we may define

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

However throughout the paper 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 all $i \in \{1, \ldots, b_n\}$.

**Remark 3.2.** It is informative to take an alternative viewpoint of the LIF objective function in (3.2) as corresponding to a block diagonal approximation of the covariance matrix. Interestingly, as a consequence of the asymptotic theory developed in the next section, this approximation does not affect the asymptotic estimation rate, but it can substantially help to speed up the computation.

The block diagonal approximation of $K_{n,m}(\rho)$ corresponding to partitioning scheme $\mathcal{B}$, to be denoted by $K_{n,m}^{\mathcal{B}}(\rho)$, can be described as follows. 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

$$\left(K_{n,m}^{\mathcal{B}}(\rho)\right)_{\boldsymbol{s},\boldsymbol{s'}} = \left[K_{n,m}(\rho)\right]_{\boldsymbol{s},\boldsymbol{s'}} \mathbb{1}_{\{t=t'\}}. \tag{3.3}$$

Observe that

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

These identities provide an alternative form for Eq. (3.2) in terms of $K_{n,m}^{\mathcal{B}}(\rho)$, namely

$$\left(\hat{\phi}_{n,\mathcal{B}}, \hat{\rho}_{n,\mathcal{B}}\right) = \underset{\phi>0,\rho\in\Theta_0}{\arg\max} \left( \phi Y_m^\top K_{n,m}^{\mathcal{B}}(\rho) Y_m - \frac{\phi^2}{2} \left\|K_{n,m}^{\mathcal{B}}(\rho)\right\|_{\ell_2}^2 \right). \tag{3.4}$$

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.

**Remark 3.3.** The following equivalent formulation for the optimization problem in (3.4) is more convenient for our subsequent theoretical analysis. Due to the quadratic dependence of the LIF loss on $\phi$, $\hat{\phi}_{n,\mathcal{B}}$ can be explicitly expressed in terms of $\hat{\rho}_{n,\mathcal{B}}$ as

$$\hat{\phi}_{n,\mathcal{B}} = \frac{Y_m^\top K_{n,m}^{\mathcal{B}}\left(\hat{\rho}_{n,\mathcal{B}}\right) Y_m}{\left\|K_{n,m}^{\mathcal{B}}\left(\hat{\rho}_{n,\mathcal{B}}\right)\right\|_{\ell_2}^2}, \quad \text{where} \quad \hat{\rho}_{n,\mathcal{B}} = \arg\max_{\rho\in\Theta_0} \frac{Y_m^\top K_{n,m}^{\mathcal{B}}\left(\rho\right) Y_m}{\left\|K_{n,m}^{\mathcal{B}}\left(\rho\right)\right\|_{\ell_2}}. \quad (3.5)$$

The term *profile LIF loss* refers to the objective function in Eq. (3.5), whose maximizer is $\hat{\rho}_{n,\mathcal{B}}$. The profile LIF loss is indeed proportional to the angle between $K_{n,m}^{\mathcal{B}}(\rho)$ and $Y_m Y_m^\top$.

Finally, the following remarks focus on computational and numerical properties of the LIF method.

**Remark 3.4.** For the trivial partition $\mathcal{B} = \{\mathcal{D}_n\}$, the optimization problem (3.2) is exactly the same as the IF algorithm. Note that the objective function in Eq. (3.2) 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 the LIF objective function can be computed almost $b_n$ times faster than the one in (2.5). In Section 5, we numerically assess the connection between the partitioning scheme of $\mathcal{D}_n$ and the estimation performance of (3.2).

**Remark 3.5.** 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 of any gradient-based optimizer for (3.2), such as the *Broyden–Fletcher–Goldfarb–Shanno (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 evaluations 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 $f_t\left(Y_{B_t,m}; \phi, \rho\right)$ stand for the IF function, with the parameters $(\phi, \rho)$, 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 $f_t\left(Y_{B_t,m}; \phi, \rho\right)$ can be evaluated by employing basic shared memory parallel schemes for computing $\|K_{B_t,m}(\rho)\|_{\ell_2}$ and $K_{B_t,m}(\rho) Y_{B_t,m}$. Finally the master processor aggregates the received quantities $\{f_t\left(Y_{B_t,m}; \phi, \rho\right) : t = 1, \ldots, b_n\}$ from the slave processors to compute the LIF objective function.

## 4. Fixed-domain asymptotic theory

The goal of this section is to investigate the fixed-domain asymptotic properties of the LIF estimator (3.5). 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 $d \leq 3$. In particular, for any $\boldsymbol{s}, \boldsymbol{s'} \in \mathcal{D}$

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

Recall that $\nu > 0$ is a known bounded constant controlling the mean squared smoothness of $G$; larger $\nu$ corresponds to smoother GP. The strictly positive scalars $\phi_0$ and $\rho_0$ respectively stand for the variance and the range parameters of $G$.

Recall that the Matern covariance function admits a relatively simple form for its spectral density:

$$\hat{K}\left(\boldsymbol{\omega}; \phi_0, \rho_0\right) = \frac{\phi_0 \rho_0^{-2\nu}}{\pi^{d/2}}\left(\frac{1}{\rho_0^2} + \left\|\boldsymbol{\omega}\right\|_{\ell_2}^2\right)^{-(\nu+d/2)}. \tag{4.1}$$

It is known that (see e.g., [23, 12]) for any bounded region $\mathcal{D} \subset \mathbb{R}^d$ with $d \leq 3$, the Matern covariance models with parameters $(\phi_1, \rho_1)$ and $(\phi_2, \rho_2)$ yield absolutely continuous measures (with respect to each other) whenever $\phi_1 \rho_1^{-2\nu} = \phi_2 \rho_2^{-2\nu}$. In this case, $(\phi_1, \rho_1)$ and $(\phi_2, \rho_2)$ 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 (4.1). The quantity $\phi_0 \rho_0^{-2\nu}$, which is usually referred to as the *microergodic* parameter, is sufficient for interpolation purposes [23]. Thus, it suffices to focus on the estimation rate for $\phi_0 \rho_0^{-2\nu}$ in our asymptotic analysis.

Recall from Remark 3.2 that $K_{n,m}^{\mathcal{B}}\left(\cdot\right)$ stands for the block diagonal approximation $K_{n,m}\left(\cdot\right)$. Define a real valued (stochastic) mapping over $\Theta_0$ by

$$\hat{\phi}_{n,\mathcal{B}}\left(\rho\right) \coloneqq \frac{Y_m^{\top} K_{n,m}^{\mathcal{B}}\left(\rho\right) Y_m}{\left\|K_{n,m}^{\mathcal{B}}\left(\rho\right)\right\|_{\ell_2}^2}, \quad \forall\, \rho \in \Theta_0. \tag{4.2}$$

For ease of presentation, we omit the dependence of $\hat{\phi}_{n,\mathcal{B}}\left(\cdot\right)$ on $m$ in our notation. It is also apparent from (3.5) that $\hat{\phi}_{n,\mathcal{B}} = \hat{\phi}_{n,\mathcal{B}}\left(\hat{\rho}_{n,\mathcal{B}}\right)$.

Before presenting the main results let us consider an interesting special instance in the LIF class of estimators that reveals a key reason behind the $\sqrt{n}$-consistency of any LIF estimation method.

**Remark 4.1.** Suppose that $\mathcal{B}$ comprises only singleton sets, i.e. $|B_t| = 1$ for any $B_t \in \mathcal{B}$. In this case $\phi K_{B_t,m}\left(\rho\right)$ (the covariance matrix of $\left[G_m\left(\boldsymbol{s}\right) : \boldsymbol{s} \in B_t\right]^{\top}$ associated to $\phi$ and $\rho$) is a scalar which is approximately proportional to $\phi\rho^{-2\nu}$. More specifically, using a similar approach as in the proof of Proposition A.1 shows that for $B_t = \{\boldsymbol{s}\}$

$$\phi K_{B_t,m}\left(\rho\right) = C_{\boldsymbol{s}} \phi \rho^{-2\nu} + \varepsilon_n\left(\boldsymbol{s}, \rho, \phi\right), \tag{4.3}$$

in which $C_{\boldsymbol{s}}$ is a known scalar, independent of $\phi$ and $\rho$, and $\varepsilon_n(\boldsymbol{s}, \rho, \phi)$ is a vanishing sequence in $n$ (which also depends on $m, d, \nu$ as well). Substituting Eq. (4.3) into Eq. (4.2) leads to

$$\hat{\phi}_{n,\mathcal{B}}(\rho)\,\rho^{-2\nu} = \left(\frac{\sum_{\boldsymbol{s}\in\mathcal{D}_n} C_{\boldsymbol{s}} G_m^2(\boldsymbol{s})}{\sum_{\boldsymbol{s}\in\mathcal{D}_n} C_{\boldsymbol{s}}^2}\right) + o(1), \quad \forall\, \rho \in \Theta_0. \tag{4.4}$$

$\hat{\phi}_{n,\mathcal{B}}(\rho)\,\rho^{-2\nu}$ 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 2.2 and denotes the interior of $\mathcal{D}_n$). Furthermore, the profile LIF loss has (roughly) no dependence on $\rho$, since

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

Simply put, there is no need to estimate $\rho$ using the profile LIF loss, for this particular scenario. For an arbitrarily chosen $\rho$, $\phi_0\rho_0^{-2\nu}$ can indeed be estimated by $\hat{\phi}_{n,\mathcal{B}}(\rho)\,\rho^{-2\nu}$. The estimator in Eq. (4.4) is in fact identical to the one proposed by Anderes [1]. He also investigated its fixed-domain asymptotic properties for regular lattices employing some techniques for studying the quadratic variation of stationary spatial Gaussian processes

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

**Theorem 4.1.** Let $G$ be observed on a lattice $\mathcal{D}_n$ satisfying Assumption 2.1. Suppose that the preconditioning order $m$ satisfies $m \geq (\nu + d/2)$. For a given binning scheme $\mathcal{B}$ of $\mathcal{D}_n$, there are bounded positive scalars $C_{\mathcal{B}}$ and $n_0$, depending on $m, d, \nu, \Theta_0, \mathcal{B}$ and the geometric structure of $\mathcal{D}_n$, such that

$$\mathbb{P}\left(\sup_{\rho\in\Theta_0}\left|\frac{\hat{\phi}_{n,\mathcal{B}}(\rho)\,\rho^{-2\nu}}{\phi_0\rho_0^{-2\nu}} - 1\right| \geq C_{\mathcal{B}}\sqrt{\frac{\log n}{n}}\right) \leq \frac{1}{n}, \quad \forall\, n \geq n_0. \tag{4.5}$$

Theorem 4.1 establishes 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 an immediate consequence of Theorem 4.1. It is known that an analogous bound as in Eq. (4.5) holds for the MLE, regardless of how $m$ is chosen. Namely, the MLE is $\sqrt{n}$-consistent even for raw data, $m = 0$. Thus Theorem 4.1 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 efficiency.

In the case that $\nu$ is either known or can be rather precisely estimated, Theorem 4.1 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 2.2 that for the regular lattices, if $m'$ represents the number

of times the Laplace operator is applied to the data, then the transformed process is a preconditioned GP of order $2m'$. Thus for Gaussian processes observed on $d$-dimensional regular lattices, $m = 2m'$ and so $m'$ should not be smaller than $\nu/2 + d/4$.

**Remark 4.2.** For pedagogical reasons, we outline a brief sketch of the proof of Theorem 4.1; full details are postponed to Section 7. The bias-variance decomposition plays a canonical role in our analysis. In particular,

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

We show that $P_1 = o(1/\sqrt{n})$ by employing a novel approach to investigate the large sample properties of the eigenvalues of $K^{\mathcal{B}}_{n,m}(\rho)$. On the other hand, $P_2$ is in fact the supremum of a chi-squared process over $\Theta_0$. Employing the classical chaining argument it can be shown that $P_2$ is of order $\sqrt{n^{-1}\log n}$, with high probability. We refer the reader to Appendix A for further details.

**Corollary 4.1.** Under the same notation and conditions as in Theorem 4.1, the following inequality holds for any stationary point $(\hat{\phi}_{n,\mathcal{B}}, \hat{\rho}_{n,\mathcal{B}})$ of the LIF loss (3.2).

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

It has been argued in [9] that estimating $\rho_0$ can improve the statistical performance, especially for small $n$. The first advantage of Corollary 4.1 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 crucial in practice, is another advantage of Corollary 4.1.

Remark 3.2 may induce a false impression that the convergence rate of $\hat{\phi}_{n,\mathcal{B}}\hat{\rho}_{n,\mathcal{B}}^{-2\nu}$ is determined by how well the covariance matrix of the preconditioned samples $K_{n,m}(\rho)$ can be approximated by $K^{\mathcal{B}}_{n,m}(\rho)$. Yet, Corollary 4.1 discloses the 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 the partitioning scheme may become more pronounced in practical situations with moderate sample sizes.

**Remark 4.3.** It has been discussed in [2] that the global solution of the IF optimization problem, in Eq. (2.5), has the same convergence rate as the MLE, when

the covariance matrix of the preconditioned samples has a uniformly bounded condition number over $\Theta_0$. Such a 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 [16]). However Corollary 4.1 requires much weaker restrictions on the covariance matrix. Two sufficient conditions on $K_{n,m}^{\mathcal{B}}(\cdot)$ can be spotted by going through our proof of Theorem 4.1.

1. The largest eigenvalue of $K_{n,m}^{\mathcal{B}}(\cdot)$ should be uniformly bounded over $\Theta_0$. Namely,
$$\max_{\rho\in\Theta_0} \left\| K_{n,m}^{\mathcal{B}}(\cdot) \right\|_{2\to 2} \asymp 1.$$
2. $K_{n,m}^{\mathcal{B}}(\rho)$ must have $\mathcal{O}(n)$ non-negligible positive eigenvalues, for any $\rho \in \Theta_0$. That is,
$$\inf_{\rho\in\Theta_0} \left\| K_{n,m}^{\mathcal{B}}(\rho) \right\|_{\ell_2} \asymp \sqrt{n}.$$

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 expands the applicability of inversion-free techniques.

Now we establish the asymptotic distribution of all the stationary points of the LIF loss function.

**Theorem 4.2.** Under the same notation and conditions as in Theorem 4.1, there is 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
$$\frac{\sqrt{n}}{\sigma_{n,\mathcal{B}}} \left( \frac{\hat{\phi}_{n,\mathcal{B}} \hat{\rho}_{n,\mathcal{B}}^{-2\nu}}{\phi_0 \rho_0^{-2\nu}} - 1 \right) \xrightarrow{d} \mathcal{N}(0,1).$$

Theorem 4.2 formulates the asymptotic distribution of the LIF estimator for joint estimation of $\phi_0$ and $\rho_0$. To our knowledge, for the MLE, such a result has only appeared in [9]. Note that unlike the full or tapered MLE, in which $\sigma_{n,\mathcal{B}} = \sqrt{2}$ (see Theorem 2 of [21]), here $m, d, \nu$, 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 $\sigma_{n,\mathcal{B}}$. A complicated expression is stated in the proof of Theorem 4.2.

**Remark 4.4.** We conclude this section with a succinct discussion of the role of $\Theta_0$ in the optimization problem presented in Eq. (3.4). 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., $\Theta_p \subseteq \Theta_q$ $\forall\, p \le q$, whose diameter grows polynomially in $n$. Namely, $\operatorname{diam}(\Theta_n) \lesssim n^{\zeta}$ for an arbitrary bounded scalar $\zeta \ge 0$. As sample size grows, such a formulation of the LIF algorithm demands less restrictive assumptions on the range parameter and bears more resemblance to an unconstrained maximization problem.

## 5. Simulation studies

This section is devoted to appraising the computational and statistical properties of the LIF algorithm on synthetic stationary Gaussian process data[1]. The purpose of our study is two-fold: investigating the scalability and efficiency of the proposed method in large datasets, as well as corroborating the fixed-domain asymptotic theory presented in Section 4. We consider two different scenarios regarding the sample size $n$. In moderate-size settings 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 approximate evaluation of the likelihood function are 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 times of the simulations (up to 100 times), the LIF algorithm has been implemented in *C++* and *R* using the *RcppParallel*[2] package.

Throughout this section $G$ is a real-valued stationary Matern GP observed on an 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 $\mathcal{D}_n$ [11]. We now concisely describe the geometry of $\mathcal{D}_n$. Let $\mathcal{D} = [0, T]^2$ be a square of side-length $T$. $\mathcal{D}_n$ is a two dimensional randomly perturbed lattice of size $n = N^2$ if there exists a non-negative $\delta$, 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 $\delta$ controls the amount of irregularity in the set of sampling locations.

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 $[1/6, 1/6, 1/3, 1/3]^\top$. Thus on average half of the bins are twice a big as 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.

[1] See Section 3.5 of [13] for more complete numerical studies.
[2] https://cran.r-project.org/web/packages/RcppParallel/index.html.

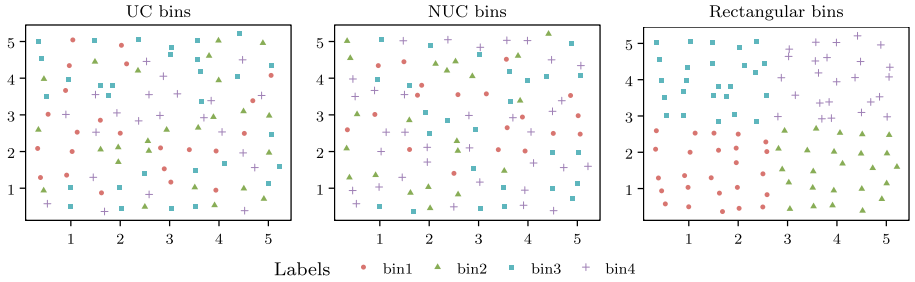


FIG 1. *Three partitioning schemes of* $10^2$ *points of a perturbed lattices on* $\mathcal{D} = [0,5]^2$ *with* $\delta = 0.5$

Figure 1 illustrates the three methods of constructing subgroups for a randomly perturbed lattice of size 100 and $\delta = 0.5$. For illustration, $b_n$ is chosen to be 4 for each scenario in Figure 1.

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 developed asymptotic insight in Section 4 is rather limited, as it is restricted to isotropic GPs. Therefore we present two sets of numerical studies for evaluating the performance of our proposed method for the geometric anisotropic processes (multiple range parameters). Note that the claim in Remark 4.1 is not valid for geometric anistotropic GPs. In other words the profile LIF loss directly depends on range parameters and therefore needs to be numerically maximized. The L-BFGS-B (limited-memory BFGS with bound constraints [4]) algorithm is utilized for maximizing the profile LIF loss. The finite difference 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.

### *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 4.1 and 4.2. 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 $\boldsymbol{s}$ ($|\mathcal{N}_m(\boldsymbol{s})| = 7$). For any $\boldsymbol{s} \in \mathcal{D}_n$, we adopt the following procedure for choosing the preconditioning coefficients $\{a_{m,\boldsymbol{s}}(\boldsymbol{t}) : \boldsymbol{t} \in \mathcal{N}_m(\boldsymbol{s})\}$.

1. Let $a_{m,\boldsymbol{s}}(\boldsymbol{s}) = 1$ and solve the system of linear equations introduced in the second condition of Definition 2.1 to compute $\{a_{m,\boldsymbol{s}}(\boldsymbol{t}) : \boldsymbol{t} \in \mathcal{N}_m(\boldsymbol{s}) \setminus \boldsymbol{s}\}$.
2. Each coefficient is normalized by dividing by the quantity $\sqrt{\sum_{\boldsymbol{t} \in \mathcal{N}_m(\boldsymbol{s})} a^2_{m,\boldsymbol{s}}(\boldsymbol{t})}$.

The goal is to estimate $\phi_0\rho_0^{-2\nu}$, which has the central role in the asymptotic analysis in Section 4. According to Theorems 4.1 and 4.2, estimating $\rho_0$ is not necessary for the isotropic Matern covariance functions. In other words, $\rho$ can be fixed in the optimization problem in Eq. (3.2). Therefore we select $\rho = 10$ and maximize the LIF function with respect to $\phi$, 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

$$\hat{\xi}_{n,\mathcal{B}} = \frac{\hat{\phi}_{n,\mathcal{B}}\hat{\rho}_{n,\mathcal{B}}^{-2\nu}}{\phi_0\rho_0^{-2\nu}}. \tag{5.1}$$

Theorem 4.2 suggests that $\hat{\xi}_{n,\mathcal{B}}$ is normally distributed centered at 1. Figures 2 and 3 respectively exhibit the histogram of $\hat{\xi}_{n,\mathcal{B}}$ for the 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 1 presents the mean and standard deviation of each histogram in Figures 2 and 3. According to Table 1, for different values of $\delta, b_n$ 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.

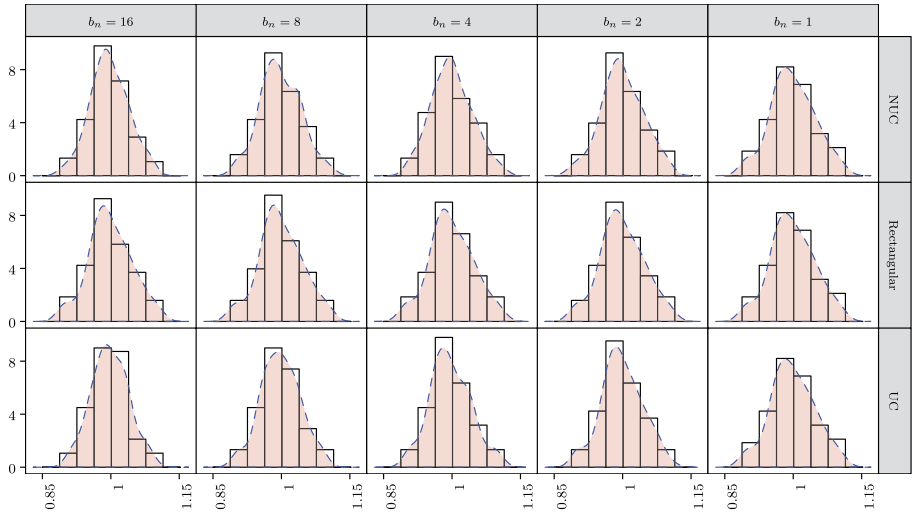


FIG 2. *The histogram of* $\hat{\xi}_{n,\mathcal{B}}$ *with* $m = 2$, $b_n = 1, 2, 4, 8, 16$ *and* 3 *binning schemes for isotropic Matern GP with* $(\phi_0, \rho_0, \nu) = (1, 5, 0.5)$ *observed on a perturbed lattice with* $\delta = 1$ *and* $n = 10^4$.

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

TABLE 1. *The mean and standard deviation of $\hat{\xi}_{n,\mathcal{B}}$ exhibited in histograms in Figures 2 and 3.*

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

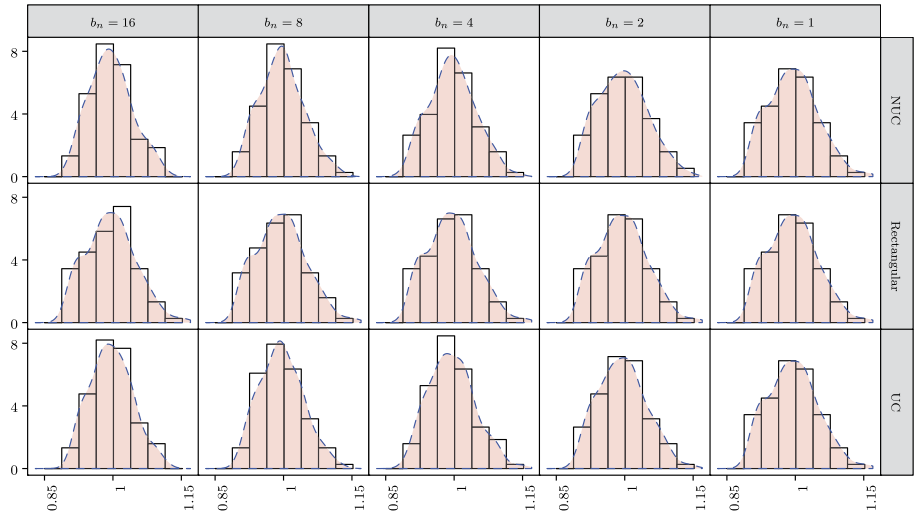


FIG 3. *The histogram of* $\hat{\xi}_{n,\mathcal{B}}$ *with* $m = 2$, $b_n = 1, 2, 4, 8, 16$ *and* 3 *binning schemes for isotropic Matern GP with* $(\phi_0, \rho_0, \nu) = (1, 5, 0.5)$ *observed on a perturbed lattice with* $\delta = 3$ *and* $n = 10^4$.

of dealing with larger conditioning sets ($|\mathcal{N}_3(\boldsymbol{s})| = 11$ for any $\boldsymbol{s} \in \mathcal{D}_n$). Table 2 summarizes the mean and standard deviation of $\hat{\xi}_{n,\mathcal{B}}$ for the different choices of $m, b_n, \delta$, and partitioning schemes.

**Remark 5.1.** The above experiments explicate some aspects of the LIF method which were 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 1 and 2, the bias of $\hat{\xi}_{n,\mathcal{B}}$ is considerably smaller than its standard deviation. We have shown that (see the proof of Theorem 4.1 for further details) for isotropic Matern GPs observed in a $d$-dimensional space

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

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.

(c) Comparing the results in Tables 1 and 2 shows that $\hat{\xi}_{n,\mathcal{B}}$ has larger bias and standard deviation for $\nu = 1$. Namely estimating $\phi_0 \rho_0^{-2\nu}$ is more difficult when $\nu = 1$. We give a qualitative justification for this phenomenon. It has been argued in Remark 4.3 that the LIF algorithm is consistent when

TABLE 2. *The mean and standard deviation of $\hat{\xi}_{n,\mathcal{B}}$ in experiments with $m = 2, 3$, $b_n = 1, 2, 4, 8, 16$ and 3 binning schemes for isotropic Matern GP with $(\phi_0, \rho_0, \nu) = (1, 2.5, 1)$ observed on a perturbed lattice with $\delta = 1, 3$.*

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

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 deficient 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 $\rho_0 = 1.25$. Note that $\rho_0$, the diameter of $\mathcal{D}$ and $\delta$ have been chosen in such a way that the lattice of size $50^2$ imitates the local neighbouring properties of $\mathcal{D}_n$ in Tables 1 and 2. Figure 4 displays $\Psi^{\mathcal{B}}_{n,m}$ in four different scenarios of $(\nu, \delta)$. It is 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.

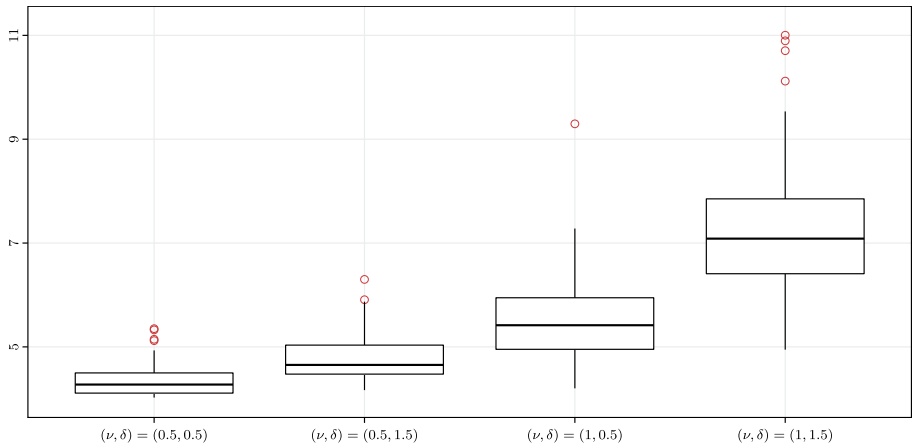


FIG 4. *The box-plot of $\Psi_{n,m}$ for different values of $\delta$ and $\nu$. Here $\mathcal{D}_n$ is a perturbed lattice of size* 2500 *and $G$ is an isotropic Matern GP with $\phi_0 = 1$ and $\rho_0 = 1.25$.*

Now we 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, with respect to the case of $b_n = 1$, is optimized with respect to $\phi$ and for a fixed $\rho = 10$. We refer the reader to Table 3 for the sample average and standard deviation of

TABLE 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.*

| | | $N = 20$ | $N = 30$ | $N = 50$ | $N = 70$ | $N = 100$ | $N = 150$ |
|---|---|---|---|---|---|---|---|
| $\delta = 1$ | 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 |

$\hat{\xi}_{n,\mathcal{B}}$ for different values of $n$. The results in Table 3 shows that the LIF estimate becomes more accurate as $n$ increases (in a fixed domain), when $b_n$ does not grow with $n$.

### *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 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 $\boldsymbol{t} = (t_1, t_2)$,

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

Here $f_\nu$ stands for the Matern standard correlation function with the smoothness parameter $\nu$. 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. It is known that $\phi_0$ and $\rho_0$ are not fully discernible in the infill setting (see [16], p. 120). Therefore the focus of our simulation studies is to estimate the quantities $\phi_0 \rho_{0,1}^{-2\nu}$ and $\phi_0 \rho_{0,2}^{-2\nu}$ (or equivalently $\phi_0 (\rho_{0,1}\rho_{0,2})^{-\nu}$ and $\rho_{0,1}/\rho_{0,2}$). We refer the reader to [1] for a comprehensive discussion regarding the identifiability of covariance parameters in multi-dimensional geometric anisotropic Matern GPs. For brevity we reformulate $\hat{\xi}_{n,\mathcal{B}}$ as the following:

$$\hat{\xi}_{n,\mathcal{B}} = \left(\frac{\hat{\phi}_{n,\mathcal{B}} \hat{\rho}_{1,n,\mathcal{B}}^{-2\nu}}{\phi_0 \rho_{0,1}^{-2\nu}}, \frac{\hat{\phi}_{n,\mathcal{B}} \hat{\rho}_{2,n,\mathcal{B}}^{-2\nu}}{\phi_0 \rho_{0,2}^{-2\nu}}\right) \in [0, \infty)^2. \tag{5.2}$$

Again, we let $\mathcal{D}_n$ be a perturbed lattice of size $n = 10^4$ and with $\delta \in \{1, 3\}$ on $\mathcal{D} = [0, 5]^2$. We simulate 100 independent realizations of a Matern GP with $\phi_0 = 1$, $\rho_0 = (1.5, 4)$ and $\nu = 0.5$ on 100 realizations of $\mathcal{D}_n$. The L-BFGS-B method with the initial guess $\rho = (10, 10)$ is used for maximizing the profile LIF loss function in a constrained box $[0.1, 50]^2$. In our experiments the boundary points were not touched during optimization, so the final results do not change even when the box constraints are not enforced. The scatter plots of $\hat{\xi}_{n,\mathcal{B}}$ is depicted in Figure 5 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 4 also accumulates the mean and standard deviation of $\hat{\xi}_{n,\mathcal{B}}$ displayed in Figure 5.

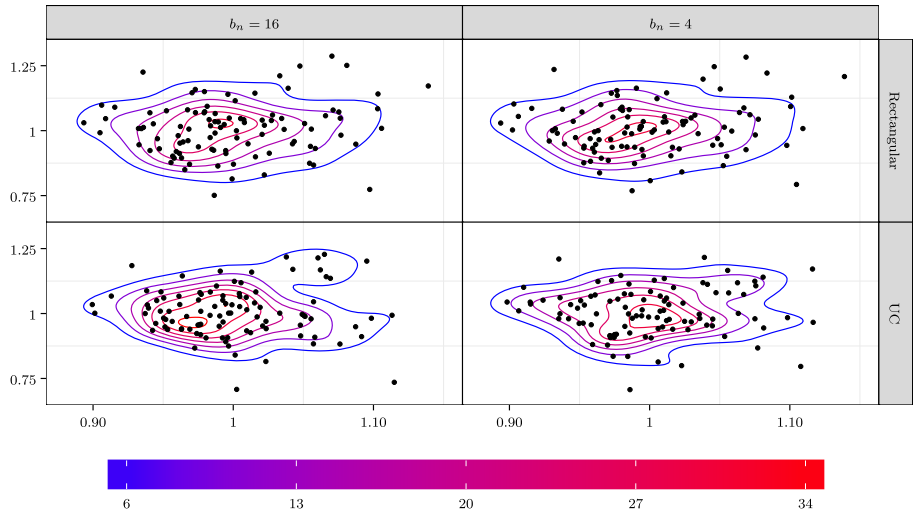


FIG 5. *The scatter plot and two dimensional KDE of $\hat{\xi}_{n,\mathcal{B}}$ for an anisotropic Matern GP with $\phi_0 = 1, \rho_0 = (1.5, 4)$, and $\nu_0 = 0.5$ observed on a perturbed lattice with $\delta = 1$ and $n = 10^4$.*

TABLE 4
*The mean and standard deviation of $\hat{\xi}_{n,\mathcal{B}}$ exhibited in scatter plots in Figures 5.*

| | $b_n = 16$ | $b_n = 4$ |
|---|---|---|
| UC | $\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = (0.9996, 1.0063)$<br>$\operatorname{std}\hat{\xi}_{n,\mathcal{B}} = (0.0467, 0.0966)$ | $\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = (1.0002, 1.0049)$<br>$\operatorname{std}\hat{\xi}_{n,\mathcal{B}} = (0.0482, 0.0932)$ |
| Rectangular | $\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = (0.9993, 1.0081)$<br>$\operatorname{std}\hat{\xi}_{n,\mathcal{B}} = (0.0507, 0.1026)$ | $\mathbb{E}\hat{\xi}_{n,\mathcal{B}} = (0.9994, 1.0104)$<br>$\operatorname{std}\hat{\xi}_{n,\mathcal{B}} = (0.0515, 0.0998)$ |

### *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 follows.

1. We fix $\mathcal{D} = [0, 25]^2$ and choose a perturbed lattice $\mathcal{D}_n$ of size $2.5 \times 10^5$, 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 infill behaviour, as the diameter of $\mathcal{D}$ is considerably smaller than $N$. We report the LIF estimates 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 ($N = 500$). In these simulations, the estimates of all unknown parameters will be reported.

Recall $\hat{\xi}_{n,\mathcal{B}}$ from Eq. (5.2). Tables 5 encapsulates $\hat{\xi}_{n,\mathcal{B}}$ and the running time of maximizing the profile 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 $b_n$.

TABLE 5
*The summary of the large-sample simulations for the first category.*

| | | $b_n = 200$ | $b_n = 50$ | $b_n = 10$ |
|---|---|---|---|---|
| $\nu = 0.5$ | $\hat{\xi}_{n,\mathcal{B}}$ | $(0.9978, 1.0434)$ | $(0.9988, 1.04085)$ | $(1.0011, 1.0280)$ |
| | 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 6 presents the summary of results 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 as slow as the former case.

TABLE 6
*The summary of the large-sample simulations for the second category.*

| | | $b_n = 200$ | $b_n = 50$ | $b_n = 10$ |
|---|---|---|---|---|
| $\nu = 1$ | $\hat{\phi}_{n,\mathcal{B}}$ | 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 |

Comparing the different columns in Table 5 and 6 reveals insensitivity of the LIF estimate to $b_n$. We believe that for large $n$, increasing the number of bins does not improve the statistical accuracy, as long as each bin can separately encode the local dependence structure. For instance when $n = 500^2$ and $b_n = 200$, there are more than 1000 samples in each bin, which is roughly enough for learning the local dependence structure in a geometric anisotropic GP with two range parameters. We observe that there is a large range of $b_n$ in which decreasing the bin size (which is equivalent to increasing $b_n$) barely degrades the statistical performance of the LIF algorithm, but the computational saving is quite substantial.

Finally, for a systematic evaluation of the role of $b_n$ on the statistical accuracy of the LIF estimate we consider a Geometric anisotropic GP with $\phi_0 = 1$ and $\rho_0 = (5, 10)$ and $\nu \in \{0.5, 1\}$ on a regular lattice ($\delta = 0$) of size $n = 40^2$ on $\mathcal{D} = [0, 10]^2$. That is, $\mathcal{D}_n = \{i/5 : i = 1, \ldots, 40\}^2$. Similar to the results in Table 5, the L-BFGS-B algorithm with starting point $\rho = (4, 8)$ is used for estimating $\hat{\xi}_{n,\mathcal{B}}$. For each $b_n$, we run 100 independent experiments for evaluating the empirical mean and standard deviation of $\hat{\xi}_{n,\mathcal{B}}$. The summary results in Table 7 shows that the standard deviation of the LIF estimator increases for larger $b_n$.

TABLE 7
*The summary of simulations for assessing the role of $b_n$.*

| | | $b_n = 1$ | $b_n = 2$ | $b_n = 4$ | $b_n = 8$ |
|---|---|---|---|---|---|
| $\nu = 0.5$ | $\mathbb{E}\hat{\xi}_{n,\mathcal{B}}$ | (0.9931, 1.0214) | (0.9902, 1.0286) | (0.9914, 1.0324) | (0.9923, 1.0341) |
| | $\operatorname{std}\hat{\xi}_{n,\mathcal{B}}$ | (0.0201, 0.0372) | (0.0239, 0.0398) | (0.0272, 0.0448) | (0.0290, 0.0482) |
| $\nu = 1$ | $\mathbb{E}\hat{\xi}_{n,\mathcal{B}}$ | (0.9873, 1.0521) | (0.9821, 1.0593) | (0.9852, 1.0565) | (0.9813, 1.0591) |
| | $\operatorname{std}\hat{\xi}_{n,\mathcal{B}}$ | (0.0573, 0.1011) | (0.0611, 0.1098) | (0.0659, 0.1149) | (0.0682, 0.1178) |

## 6. Discussion

In this paper we have introduced a family of scalable covariance estimation algorithms, called the local inversion-free (LIF) algorithm, by amalgamating the ideas of the inversion-free estimation procedure in [2] and a block diagonal approximation of the covariance matrix of the preconditioned 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 only 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 [2]. 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 fraction of the eigenvalues are further away from zero. The removal of 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 of spatial Gaussian processes.

Despite the relatively low cost of computing the LIF estimate for GPs observed on irregularly spaced locations, it remains to investigate the applicability of LIF-based algorithms beyond parameter estimation, e.g., prediction. Furthermore, despite recent progresses in preconditioning of stationary GPs, an effective mechanism to reduce the condition number of the covariance matrix 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.

We end this discussion by briefly describing a potential way of adjusting the LIF loss function for non-stationary processes with smoothly varying variance and range parameters (with a known smoothness parameter). 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 a 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 of 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, namely,

$$\left(\hat{\boldsymbol{\phi}}_{n,\mathcal{B}}, \hat{\boldsymbol{\rho}}_{n,\mathcal{B}}\right) = \underset{\boldsymbol{\phi},\boldsymbol{\rho}}{\arg\min}\left\{\sum_{t=1}^{b_n}\left\|Y_{B_t,m}Y_{B_t,m}^{\top} - \phi_t K_{B_t,m}\left(\rho_t\right)\right\|_{\ell_2}^2 + J_{\boldsymbol{\phi}}\left(\phi_1,\ldots,\phi_{b_n}\right) + J_{\boldsymbol{\rho}}\left(\rho_1,\ldots,\rho_{b_n}\right)\right\},$$

in which $J_{\boldsymbol{\phi}}$ and $J_{\boldsymbol{\rho}}$ are non-negative functions penalizing rapidly varying variance and range parameters. Such a penalized loss function may be optimized using the coordinate descent method.

## 7. Proofs

All the constants appearing in this section (including those implicitly defined in $\lesssim$, and $\asymp$), are bounded and depend on $m, \nu, d, \Theta_0$, and the geometric structure of the sampling locations.

*Proof of Theorem 4.1.* Applying the triangle inequality, we get

$$\sup_{\rho\in\Theta_0}\left|\frac{\hat{\phi}_{n,\mathcal{B}}\left(\rho\right)\rho^{-2\nu}}{\phi_0\rho_0^{-2\nu}} - 1\right| \leq \sup_{\rho\in\Theta_0}\left|\frac{\mathbb{E}\hat{\phi}_{n,\mathcal{B}}\left(\rho\right)\rho^{-2\nu}}{\phi_0\rho_0^{-2\nu}} - 1\right| + \sup_{\rho\in\Theta_0}\frac{\left|\hat{\phi}_{n,\mathcal{B}}\left(\rho\right)\rho^{-2\nu} - \mathbb{E}\hat{\phi}_{n,\mathcal{B}}\left(\rho\right)\rho^{-2\nu}\right|}{\phi_0\rho_0^{-2\nu}}. \quad (7.1)$$

Let $P_1$ and $P_2$ respectively stand for the two terms in the right hand side of (7.1). 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 [11] are crucial for bounding $P_2$ from above.

Choose an arbitrary $(\phi,\rho)\in\mathcal{I}\times\Theta_0$. Recall $K_{n,m}^{\mathcal{B}}\left(\rho\right)\in\mathbb{R}^{n\times n}$ from (3.3) and $\hat{\phi}_{n,\mathcal{B}}\left(\rho\right)$ from Eq. (4.2). Define $L_{n,m}^{\mathcal{B}}\left(\rho\right) \coloneqq \rho^{2\nu}K_{n,m}^{\mathcal{B}}\left(\rho\right)$. Observe that

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

Thus,

$$P_1 = \sup_{\rho\in\Theta_0}\left|\frac{\left\langle L_{n,m}^{\mathcal{B}}\left(\rho\right), L_{n,m}^{\mathcal{B}}\left(\rho_0\right)\right\rangle}{\left\|L_{n,m}^{\mathcal{B}}\left(\rho\right)\right\|_{\ell_2}^2} - 1\right|$$

$$
\begin{aligned}
&= \sup_{\rho\in\Theta_0}\left|\frac{\langle L^{\mathcal{B}}_{n,m}(\rho)-L^{\mathcal{B}}_{n,m}(\rho_0),L^{\mathcal{B}}_{n,m}(\rho)\rangle}{\left\|L^{\mathcal{B}}_{n,m}(\rho)\right\|^2_{\ell_2}}\right| \\
&\overset{(a)}{\leq} \sup_{\rho\in\Theta_0}\left[\frac{\left\|L^{\mathcal{B}}_{n,m}(\rho)-L^{\mathcal{B}}_{n,m}(\rho_0)\right\|_{\mathcal{S}_1}\left\|L^{\mathcal{B}}_{n,m}(\rho)\right\|_{2\to2}}{\left\|L^{\mathcal{B}}_{n,m}(\rho)\right\|^2_{\ell_2}}\right]. \qquad (7.2)
\end{aligned}
$$

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 (7.2) in Appendix A. Lemma A.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 A.3 implies that

$$
\begin{aligned}
\sup_{\rho\in\Theta_0}\left\|\Delta^{\mathcal{B}}(\rho,\rho_0)\right\|_{\mathcal{S}_1} &\lesssim \left(\mathbb{1}_{\{d=1\}}+\mathbb{1}_{\{d=2\}}\log n+\mathbb{1}_{\{d\geq3\}}n^{1-2/d}\right)\operatorname{diam}(\Theta_0) \\
&\asymp \left(\mathbb{1}_{\{d=1\}}+\mathbb{1}_{\{d=2\}}\log n+\mathbb{1}_{\{d\geq3\}}n^{1-2/d}\right). \qquad (7.3)
\end{aligned}
$$

Thus the upper bound on $P_1$ in (7.2) can be rewritten as

$$
P_1\lesssim\left(\frac{\mathbb{1}_{\{d=1\}}}{n}+\mathbb{1}_{\{d=2\}}\frac{\log n}{n}+\mathbb{1}_{\{d\geq3\}}n^{-2/d}\right)\sup_{\rho\in\Theta_0}\left\|L^{\mathcal{B}}_{n,m}(\rho)\right\|_{2\to2}. \qquad (7.4)
$$

So we need to find a uniform upper bound on the largest eigenvalue of $L^{\mathcal{B}}_{n,m}(\rho)$ on $\Theta_0$. Notice that $L^{\mathcal{B}}_{n,m}(\rho)$ is a block diagonalized version of $L_{n,m}(\rho)$. Hence

$$
\left\|L^{\mathcal{B}}_{n,m}(\rho)\right\|_{2\to2}\leq\left\|L_{n,m}(\rho)\right\|_{2\to2},\quad\forall\,\rho\in\Theta_0
$$

Thus, 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 [17] demonstrates that all the eigenvalues of $L_{n,m}(\rho)$ are universally bounded. Namely,

$$
\sup_{\rho\in\Theta_0}\lambda_j\left(L_{n,m}(\rho)\right)\leq\alpha_{\max},\quad\forall\,j=1,\ldots,|\mathcal{D}_n| \qquad (7.5)
$$

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

$$
P_1\lesssim\left(\frac{\mathbb{1}_{\{d=1\}}}{n}+\mathbb{1}_{\{d=2\}}\frac{\log n}{n}+\mathbb{1}_{\{d\geq3\}}n^{-2/d}\right). \qquad (7.6)
$$

However the operator norm of $L_{n,m}(\rho)$ is not necessarily uniformly bound on $\Theta_0$, for a general irregular lattice satisfying Assumption 2.1. For such case, we show in Proposition A.1 that

$$
\left|\left(L_{n,m}(\rho)\right)_{\boldsymbol{s},\boldsymbol{t}}\right|\lesssim\left(1+\lfloor n^{1/d}\rfloor\left\|\boldsymbol{t}-\boldsymbol{s}\right\|_{\ell_2}\right)^{-2(m-\nu)},\quad\boldsymbol{s},\boldsymbol{t}\in\mathcal{D}_n. \qquad (7.7)
$$

Lemma B.2 also introduces an upper bound on the operator norm of the matrices satisfying (7.7). Applying Lemma B.2 yields

$$
\sup_{\rho\in\Theta_0}\left\|L^{\mathcal{B}}_{n,m}(\rho)\right\|_{2\to2}\leq\sup_{\rho\in\Theta_0}\left\|L_{n,m}(\rho)\right\|_{2\to2}\lesssim\left(1+\mathbb{1}_{\{m=\nu+d/2\}}\log n\right). \qquad (7.8)
$$

The desired bound on $P_1$ is obtained by combining (7.4) and (7.8). 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

$$M^{\mathcal{B}}_{n,m}(\rho) = \sqrt{L_{n,m}(\rho_0)}\left[\frac{nL^{\mathcal{B}}_{n,m}(\rho)}{\left\|L^{\mathcal{B}}_{n,m}(\rho)\right\|^2_{\ell_2}}\right]\sqrt{L_{n,m}(\rho_0)}, \quad \forall\, \rho \in \Theta_0. \tag{7.9}$$

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,

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

So $P_2$ can be rewritten as the supremum of a centered $\chi^2$ process over $\Theta_0$, i.e.,

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

So if $M^{\mathcal{B}}_{n,m}(\rho)$ admits the three conditions in Proposition B.1, then there are bounded scalars $C$ and $n_0 \in \mathbb{N}$ such that for any $n \geq n_0$, we have

$$\begin{aligned}
&\mathbb{P}\left(P_2 \geq C\sqrt{\frac{\log n}{n}}\right)\\
&\quad= \mathbb{P}\left(\sup_{\rho\in\Theta_0}\left|Z^\top M^{\mathcal{B}}_{n,m}(\rho)\,Z - \operatorname{tr}\left\{M^{\mathcal{B}}_{n,m}(\rho)\right\}\right| \geq C\sqrt{n\log n}\right)\\
&\quad\leq \frac{1}{n}.
\end{aligned} \tag{7.10}$$

Thus we require to verify the conditions $(a)-(c)$ in Proposition B.1.

*Validating condition* $(a)$. We should substantiate the uniform boundedness of $n^{-1/2}\left\|M^{\mathcal{B}}_{n,m}(\rho)\right\|_{\ell_2}$ over $\Theta_0$. 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 A.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 A.7 ensures the boundedness of $U'$ (and consequently $U$).

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

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

We argued in (7.8) that $\|L_{n,m}(\rho_0)\|_{2\to 2} \lesssim \left(1 + \mathbb{1}_{\{m=\nu+d/2\}} \log n\right) \leq \log n$. So,

$$\frac{\left\|M^{\mathcal{B}}_{n,m}(\rho_2) - M^{\mathcal{B}}_{n,m}(\rho_1)\right\|_{2\to 2}}{n\,|\rho_2 - \rho_1| \log n} \lesssim \frac{\left\| \frac{L^{\mathcal{B}}_{n,m}(\rho_2)}{\left\|L^{\mathcal{B}}_{n,m}(\rho_2)\right\|^2_{\ell_2}} - \frac{L^{\mathcal{B}}_{n,m}(\rho_1)}{\left\|L^{\mathcal{B}}_{n,m}(\rho_1)\right\|^2_{\ell_2}} \right\|_{2\to 2}}{|\rho_2 - \rho_1|}. \tag{7.11}$$

Furthermore, we know from the triangle inequality that

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

Let $\Psi^1_n(\rho_1, \rho_2)$ and $\Psi^2_n(\rho_1, \rho_2)$ denote the first and second terms in the right hand side of the above identity. The fact that $\min_{\rho\in\Theta_0} n^{-1} \left\|L^{\mathcal{B}}_{n,m}(\rho)\right\|^2_{\ell_2} > 0$ (see Lemma A.6) comes in handy for finding a simpler upper bound on $\Psi^1_n(\rho_1, \rho_2)$ and $\Psi^2_n(\rho_1, \rho_2)$.

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

Furthermore, Lemma A.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

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

$$\times \left\| L^{\mathcal{B}}_{n,m}(\rho_2) - L^{\mathcal{B}}_{n,m}(\rho_1) \right\|_{\ell_2}.$$

It is known from (7.8) that $\left\| L^{\mathcal{B}}_{n,m}(\rho_2) \right\|_{2\to 2} \lesssim \log n$. Moreover, it is easy to verify that

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

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 A.5. In summary, (7.11) can be rewritten as

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

showing that the condition $(b)$ of Proposition B.1 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.

$$V_n \coloneqq \left\| M^{\mathcal{B}}_{n,m}(\rho) \right\|_{2\to 2} \sqrt{\frac{\log n}{n}}. \tag{7.12}$$

$V_n$ 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 A.6, which says the Frobenius norm of $L_{n,m}(\rho)$ is of order $\sqrt{n}$ (uniformly on $\Theta_0$) 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. (7.8) for an upper bound on the operator norm of $L_{n,m}$ and $L^{\mathcal{B}}_{n,m}$ matrices over $\Theta_0$. So, $V_n$ can be bounded above by

$$V_n \lesssim \left(1+\mathbb{1}_{\{m=\nu+d/2\}}\log n\right)^2\sqrt{\frac{\log n}{n}} \to 0, \quad \text{as } n\to\infty. \tag{7.13}$$

□

*Proof of Theorem 4.2.* Let $\rho_{\max}$ and $\rho_{\min}$ respectively denote the largest and smallest element of $\Theta_0$. 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

$$T_n(\rho,Y) \coloneqq \sqrt{n}\left(\frac{\hat{\phi}_{n,\mathcal{B}}(\rho)\,\rho^{-2\nu}}{\phi_0\rho_0^{-2\nu}}-1\right) = \sqrt{n}\left(\frac{Y^\top L^{\mathcal{B}}_{n,m}(\rho)\,Y}{\phi_0\rho_0^{-2\nu}\left\|L^{\mathcal{B}}_{n,m}(\rho)\right\|^2_{\ell_2}}-1\right). \tag{7.14}$$

For notational convenience, the dependence to $\phi_0,\rho_0$ 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 $\sigma_n$. 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 $\rho_0,m,d,\nu$, and $\Theta_0$) such that for any $n\geq n_0$

$$T_n(\rho_{\min},Y)(1-p_n) \leq T_n(\hat{\rho}_n,Y) \leq T_n(\rho_{\max},Y)(1+q_n). \tag{7.15}$$

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 (7.15) weakly converge to a random variable distributed as $N(0,1)$, the squeeze theorem for the weak convergence (see Lemma B.4 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

$$T'_n(\rho_{\min})(1-p'_n) \leq T'_n(\hat{\rho}_n) \leq T'_n(\rho_{\max})(1+q'_n), \tag{7.16}$$

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 $\rho$. Then,

$$T'_n(\rho) = \frac{U^\top L^{\mathcal{B}}_{n,m}(\rho)\, U}{\left\|L^{\mathcal{B}}_{n,m}(\rho)\right\|^2_{\ell_2}}, \tag{7.17}$$

We only prove the right hand side inequality in Eq. (7.16) and the other side can be shown similarly. We separately analyze the numerator and denominator in (7.17). We know that $L^{\mathcal{B}}_{n,m}(\rho) \preceq L^{\mathcal{B}}_{n,m}(\rho_{\max})$ for any $\rho \in \Theta_0$ (see (A.17) for the details). Thus, $U^\top L^{\mathcal{B}}_{n,m}(\rho)\, U \leq U^\top L^{\mathcal{B}}_{n,m}(\rho_{\max})\, U$ almost surely. Namely,

$$\begin{aligned} T'_n(\rho) \leq \frac{U^\top L^{\mathcal{B}}_{n,m}(\rho_{\max})\, U}{\left\|L^{\mathcal{B}}_{n,m}(\rho)\right\|^2_{\ell_2}} \quad &\Leftrightarrow \\ \left\{\frac{T'_n(\rho)}{T'_n(\rho_{\max})} - 1\right\} \leq \frac{\left\|L^{\mathcal{B}}_{n,m}(\rho_{\max})\right\|^2_{\ell_2} - \left\|L^{\mathcal{B}}_{n,m}(\rho)\right\|^2_{\ell_2}}{\left\|L^{\mathcal{B}}_{n,m}(\rho)\right\|^2_{\ell_2}}&. \end{aligned} \tag{7.18}$$

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

$$q'_n \coloneqq \frac{\left\|L^{\mathcal{B}}_{n,m}(\rho_{\max})\right\|^2_{\ell_2} - \left\|L^{\mathcal{B}}_{n,m}(\rho)\right\|^2_{\ell_2}}{\left\|L^{\mathcal{B}}_{n,m}(\rho)\right\|^2_{\ell_2}} = o\left(\frac{1}{\sqrt{n}}\right), \quad \text{as } n \to \infty. \tag{7.19}$$

As we know from Lemma A.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

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

Eq. (7.8) provides an upper bounds on $\left\|L^{\mathcal{B}}_{n,m}(\rho_{\max})\right\|_{2\to 2}$. So

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

We now employ analogous techniques as Eq. (7.3) (see also Lemma A.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. (7.3) can be rewritten as the following.

$$\exists\, 0 < \gamma < \frac{1}{2}, \;\text{ s.t. }\; \left\|\Delta^{\mathcal{B}}(\rho_{\max}, \rho)\right\|_{\mathcal{S}_1} \lesssim n^{\gamma}. \tag{7.20}$$

Thus $\psi_n$ 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 $\rho$ and $Y$ on $\xi_n$. Let us decompose $T_n(\rho, Y)$ into two parts as

$$\begin{aligned} T_n(\rho, Y) &= \left( \frac{T_n(\rho, Y) - \mathbb{E}T_n(\rho, Y)}{\sqrt{\operatorname{var} T_n(\rho, Y)}} \right) \sqrt{\operatorname{var} T_n(\rho, Y)} + \mathbb{E}T_n(\rho, Y) \\ &= \left( \frac{\xi_n - \mathbb{E}\xi_n}{\sqrt{\operatorname{var} \xi_n}} \right) \sqrt{\operatorname{var} \xi_n} + \mathbb{E}T_n(\rho, Y). \end{aligned} \tag{7.21}$$

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 4.1. A prudent look at Eqs. (7.4) and (7.6) reveals that $P_1 \lesssim n^{\gamma - 1} \log n$ for some $\gamma < 1/2$ ($\gamma$ is the same as in (7.20)). 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 (7.21). Now we express $\xi_n$ as a quadratic term of a Gaussian random vector. Using identity (7.14), one can easily show that

$$\xi_n \overset{d}{=} Z^{\top} \frac{M^{\mathcal{B}}_{n,m}(\rho)}{\sqrt{n}} Z, \tag{7.22}$$

in which $Z$ is a standard Gaussian vector of proper size and $M^{\mathcal{B}}_{n,m}(\rho)$ has been defined in (7.9). The explicit expressions for the expected value and standard deviation of $\xi_n$ are given by

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

We showed in the proof of Theorem 4.1 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 (7.12) and (7.13)). Thus applying Lemma $A.4$ of [11], on asymptotic normality of the normalized generalized $\chi^2$ 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

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

We claim that

$$\lim_{n\to\infty}\frac{\sigma_{n,m}\left(\rho,\rho_0\right)}{\sigma_{n,m}\left(\rho_1,\rho_2\right)}=1,\quad\forall\,\rho_1,\rho_2\in\Theta_0. \tag{7.23}$$

Thus, $\sigma_{n,m}$ has no dependence to $\rho$, $\rho_0$, and $\Theta_0$. In other words, $\sigma_{n,m}$ only depends on $m,d,\nu$, 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 $\sigma_{n,m}\left(\rho,\rho_0\right)\asymp 1$ for some $\rho_1',\rho_2'\in\Theta_0$. Applying Lemma A.6 on the denominator of $\sigma_{n,m}\left(\rho_1',\rho_2'\right)$, we get,

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

So, $\sigma_{n,m}\left(\rho_1',\rho_2'\right)\asymp 1$ as a result of Lemma A.7. We now turn to substantiate (7.23). It is sufficient to verify the following identities for any $\rho_1,\rho_2\in\Theta_0$.

$$\lim_{n\to\infty}\frac{\sigma_{n,m}\left(\rho,\rho_0\right)}{\sigma_{n,m}\left(\rho_1,\rho_0\right)}=1,\quad\lim_{n\to\infty}\frac{\sigma_{n,m}\left(\rho_1,\rho_0\right)}{\sigma_{n,m}\left(\rho_1,\rho_2\right)}=1. \tag{7.24}$$

To avoid repetition, we only demonstrate the left hand side identity in (7.24) and the other one can be substantiated using analogous techniques. Observe that

$$\begin{aligned}\frac{\sigma_{n,m}\left(\rho,\rho_0\right)}{\sigma_{n,m}\left(\rho_1,\rho_0\right)} &= \left[\frac{\left\|L_{n,m}^{\mathcal{B}}\left(\rho_1\right)\right\|_{\ell_2}}{\left\|L_{n,m}^{\mathcal{B}}\left(\rho\right)\right\|_{\ell_2}}\right]^2\frac{\left\|\sqrt{L_{n,m}\left(\rho_0\right)}L_{n,m}^{\mathcal{B}}\left(\rho\right)\sqrt{L_{n,m}\left(\rho_0\right)}\right\|_{\ell_2}}{\left\|\sqrt{L_{n,m}\left(\rho_0\right)}L_{n,m}^{\mathcal{B}}\left(\rho_1\right)\sqrt{L_{n,m}\left(\rho_0\right)}\right\|_{\ell_2}}\\ &\coloneqq a_nb_n.\end{aligned}$$

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 (7.19). 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.

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

$$\overset{(a)}{\lesssim} \frac{\log n \left\|\Delta^{\mathcal{B}}(\rho,\rho_0)\right\|_{\mathcal{S}_1}}{\left\|\sqrt{L_{n,m}(\rho_0)}L^{\mathcal{B}}_{n,m}(\rho_1)\sqrt{L_{n,m}(\rho_0)}\right\|_{\ell_2}} \overset{(b)}{\lesssim} \frac{\left\|\Delta^{\mathcal{B}}(\rho,\rho_0)\right\|_{\mathcal{S}_1}\log n}{\sqrt{n}}.$$

Here $(a)$ and $(b)$ are successively implied from Eq. (7.8) and Lemma A.7. Using similar techniques as Eq. (7.20) 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. □

□

## Appendix A: Large sample behavior 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 $m, d, \nu, \Theta_0$ and the topological structure of $\mathcal{D}_n$.

- $\mathcal{D}_n$ is a $d$-dimensional grid satisfying Assumption 2.1. 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 2.1.

Before jumping into stating the theoretical results in the subsequent sections, we recall some key assumptions and notations that we have used in the body of the paper. $G$ represents a centered, isotropic Matern GP whose one time realization has been observed at $\mathcal{D}_n$. The range parameters $\rho$ 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 2.1). $m$ 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 $K^{\mathcal{B}}_{n,m}(\rho)$ in Eq. (3.3), 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 $L^{\mathcal{B}}_{n,m}(\rho) := \rho^{2\nu} K^{\mathcal{B}}_{n,m}(\rho)$ for notational convenience.

### *A.1. How do the off-diagonal entries of $K^{\mathcal{B}}_{n,m}(\rho)$ decay?*

The main objective of this section is to study the decay rate of the off-diagonal entries of $K^{\mathcal{B}}_{n,m}(\rho)$, which comes in handy for analyzing the asymptotic behavior of different norms of $K^{\mathcal{B}}_{n,m}(\rho)$ in Section 7.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

$$f_{\boldsymbol{s}}^{N}(\boldsymbol{\omega}):=\|\boldsymbol{\omega}\|_{\ell_{2}}^{-(\nu+d/2)}\sum_{\boldsymbol{s'}\in\mathcal{N}_{m}(\boldsymbol{s})}a_{m,\boldsymbol{s}}(\boldsymbol{s'})\exp\left(j\langle N\boldsymbol{\omega},\boldsymbol{s'}-\boldsymbol{s}\rangle\right),\quad\forall\,\boldsymbol{\omega}\neq\mathbf{0}_{d}, \tag{A.1}$$

and the strictly increasing function $h_{N}:(0,\infty)\mapsto(0,1)$ with

$$h_{N}(x):=\left[1+(Nx)^{-2}\right]^{-(\nu+d/2)}. \tag{A.2}$$

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.

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

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

$$\begin{aligned}\left(K_{n,m}(\rho)\right)_{\boldsymbol{s},\boldsymbol{t}} &= \frac{1}{\rho^{2\nu}}\sum_{\boldsymbol{s'}\in\mathcal{N}_{m}(\boldsymbol{s})}\sum_{\boldsymbol{t'}\in\mathcal{N}_{m}(\boldsymbol{t})}a_{m,\boldsymbol{s}}(\boldsymbol{s'})\,a_{m,\boldsymbol{t}}(\boldsymbol{t'})\\ &\quad\times\int_{\mathbb{R}^{d}}\frac{\exp\left(j\langle N\boldsymbol{\omega},\boldsymbol{t'}-\boldsymbol{s'}\rangle\right)}{\|\boldsymbol{\omega}\|_{\ell_{2}}^{2\nu+d}}h_{N}\left(\rho\,\|\boldsymbol{\omega}\|_{\ell_{2}}\right)d\boldsymbol{\omega}\\ &= \rho^{-2\nu}\int_{\mathbb{R}^{d}}\exp\left(j\langle\boldsymbol{t}-\boldsymbol{s},\boldsymbol{\omega}\rangle\right)f_{\boldsymbol{s}}^{N}(\boldsymbol{\omega})\,\overline{f_{\boldsymbol{t}}^{N}(\boldsymbol{\omega})}h_{N}\left(\rho\,\|\boldsymbol{\omega}\|_{\ell_{2}}\right)d\boldsymbol{\omega}.\end{aligned} \tag{A.3}$$

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 A.1.** There exists $\beta\in(1,\infty)$ (depending on $m,\nu,d$ and $\mathcal{D}_{n}$) such that

$$\max_{\boldsymbol{s}\in\mathcal{D}_{n}}\left|f_{\boldsymbol{s}}^{N}(\boldsymbol{\omega})\right|^{2}\leq\frac{\beta}{1+\|\boldsymbol{\omega}\|_{\ell_{2}}^{2\nu+d}},\quad\forall\,\boldsymbol{\omega}\neq\mathbf{0}_{d}. \tag{A.4}$$

*Proof.* Define the bounded integer $g_{m}$ by $g_{m}:=\max_{\boldsymbol{s}\in\mathcal{D}_{n}}|\mathcal{N}_{m}(\boldsymbol{s})|$. 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. (A.4), 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 is a bounded scalar $\pi_m$ 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 2.1, we normalize the coefficients $a_{m,\boldsymbol{s}}(\boldsymbol{s'})$'s to have unit Euclidean norm. Thus

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

For proving the other claim, we study the Taylor expansion of $f_{\boldsymbol{s}}^N$ near the origin. Definition 2.1 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

$$\begin{aligned}&\limsup_{\boldsymbol{\omega}\to\mathbf{0}_d}\left|f_{\boldsymbol{s}}^N(\boldsymbol{\omega})\right|^2 \\ =\ &\lim_{\boldsymbol{\omega}\to\mathbf{0}_d}\frac{1}{\|\boldsymbol{\omega}\|_{\ell_2}^{2\nu+d}}\left|\sum_{r=0}^{\infty}\frac{(jN)^r}{r!}\sum_{\boldsymbol{s'}\in\mathcal{N}_m(\boldsymbol{s})}a_{m,\boldsymbol{s}}(\boldsymbol{s'})\left(\langle\boldsymbol{\omega},\boldsymbol{s'}-\boldsymbol{s}\rangle\right)^r\right|^2 \\ =\ &\limsup_{\boldsymbol{\omega}\to\mathbf{0}_d}\frac{1}{\|\boldsymbol{\omega}\|_{\ell_2}^{2\nu+d}}\left|\sum_{r=m}^{\infty}\frac{(jN)^r}{r!}\sum_{\boldsymbol{s'}\in\mathcal{N}_m(\boldsymbol{s})}a_{m,\boldsymbol{s}}(\boldsymbol{s'})\left(\langle\boldsymbol{\omega},\boldsymbol{s'}-\boldsymbol{s}\rangle\right)^r\right|^2 \\ =\ &\limsup_{\boldsymbol{\omega}\to\mathbf{0}_d}\frac{N^{2m}}{m!\,\|\boldsymbol{\omega}\|_{\ell_2}^{2\nu+d}}\left|\sum_{\boldsymbol{s'}\in\mathcal{N}_m(\boldsymbol{s})}a_{m,\boldsymbol{s}}(\boldsymbol{s'})\left(\langle\boldsymbol{\omega},\boldsymbol{s'}-\boldsymbol{s}\rangle\right)^m\right|^2. \qquad \text{(A.5)}\end{aligned}$$

The Cauchy-Schwartz inequality simplifies the complex expressions in Eq. (A.5).

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

Since $N\|\boldsymbol{s'}-\boldsymbol{s}\|_{\ell_2}\asymp 1$ for any $\boldsymbol{s'}\in\mathcal{N}_m(\boldsymbol{s})$, then

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

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 easy to obtain a closed form expression for $\beta$ in terms of $g_m$ and $\pi_m$. □

**Proposition A.1.** For any pair $\boldsymbol{s},\boldsymbol{t}\in\mathcal{D}_n$ and any partition $\mathcal{B}$ of $\mathcal{D}_n$,

$$\left|\left(K_{n,m}^{\mathcal{B}}(\rho)\right)_{\boldsymbol{s},\boldsymbol{t}}\right| \lesssim \rho^{-2\nu}\left(1+N\left\|\boldsymbol{t}-\boldsymbol{s}\right\|_{\ell_2}\right)^{-2(m-\nu)}. \tag{A.6}$$

*Proof.* Without loss of generality we can 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. (A.6) for the entries of $K_{n,m}(\rho)$. For simplicity, let $f_{\nu,\rho}$ denotes the Matern correlation function with parameters $(\rho,\nu)$. Notice that $f_{\nu,\rho}(x)=f_{\nu,1}(x/\rho)$. We first prove the inequality (A.6) for the case of $\left\|\boldsymbol{t}-\boldsymbol{s}\right\|_{\ell_2}=O\left(N^{-1}\right)$. 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 (A.3) for $\boldsymbol{s}=\boldsymbol{t}$. Trivially,

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

We finish the proof of this part by using Lemma A.1.

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

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

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

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

$$R_r(\boldsymbol{t}-\boldsymbol{s}) = 2m\int_0^1 (1-x)^{2m-1} D^r f_{\nu,\rho}\Big( (\boldsymbol{t}-\boldsymbol{s}) + x\left[(\boldsymbol{t}'-\boldsymbol{t}) - (\boldsymbol{s}'-\boldsymbol{s})\right]\Big) dx. \tag{A.7}$$

Thus,

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

The first term in the right hand side vanishes, as for any $|r| < 2m$

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

which is implied from the constraint on $\{a_{m,\boldsymbol{s}}(\boldsymbol{t}) : \boldsymbol{s}\in\mathcal{D}_n, \boldsymbol{t}\in\mathcal{N}_m(\boldsymbol{s})\}$ in Definition 2.1. We now control the second term from above. Observe that

$$\begin{aligned}|\psi| \quad \leq \quad & \sum_{|r|=2m} \left| \sum_{\boldsymbol{s}'\in\mathcal{N}_m(\boldsymbol{s})} \sum_{\boldsymbol{t}'\in\mathcal{N}_m(\boldsymbol{t})} a_{m,\boldsymbol{s}}(\boldsymbol{s}') a_{m,\boldsymbol{t}}(\boldsymbol{t}') \left[(\boldsymbol{t}'-\boldsymbol{t}) - (\boldsymbol{s}'-\boldsymbol{s})\right]^r \right| \\ & \times \max_{|r|=2m} \left| \frac{R_r(\boldsymbol{t}-\boldsymbol{s})}{r!} \right|. \end{aligned} \tag{A.8}$$

The next step is to introduce a uniform upper bound on the residual functions using Eq. (A.7) and the chain rule of derivative.

$$\begin{aligned}\max_{|r|=2m} |R_r(\boldsymbol{t}-\boldsymbol{s})| \quad &\leq \quad \max_{\substack{|r|=2m \\ x\in[0,1]}} \left| D^r f_{\nu,\rho}\Big\{ (\boldsymbol{t}-\boldsymbol{s}) + x\left[(\boldsymbol{t}'-\boldsymbol{t}) - (\boldsymbol{s}'-\boldsymbol{s})\right] \Big\} \right| \\ &\leq \quad \max_{\substack{|r|=2m \\ x\in[0,1]}} \frac{\left| D^r f_{\nu,1}\Big\{ \frac{(\boldsymbol{t}-\boldsymbol{s}) + x\left[(\boldsymbol{t}'-\boldsymbol{t}) - (\boldsymbol{s}'-\boldsymbol{s})\right]}{\rho} \Big\} \right|}{\rho^{2m}}. \end{aligned} \tag{A.9}$$

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

$$\min_{x\in[0,1]} \|(\boldsymbol{t}-\boldsymbol{s}) + x\left[(\boldsymbol{t}'-\boldsymbol{t})\right]\|_{\ell_2} \geq \frac{\|\boldsymbol{t}-\boldsymbol{s}\|_{\ell_2}}{2}. \tag{A.10}$$

Now we apply Lemma 4 of [1] 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

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

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

Combining this inequality and Eq. (A.10) shows that for any pair $(\boldsymbol{s},\boldsymbol{t})$ with $\|\boldsymbol{t}-\boldsymbol{s}\|_{\ell_2} \geq h/N$

$$\max_{|r|=2m} |R_r(\boldsymbol{t}-\boldsymbol{s})| \lesssim \rho^{-2m}\left(\frac{\|\boldsymbol{t}-\boldsymbol{s}\|_{\ell_2}}{\rho}\right)^{2(\nu-m)} \lesssim \rho^{-2\nu}\left(\frac{1}{N}+\|\boldsymbol{t}-\boldsymbol{s}\|_{\ell_2}\right)^{2(\nu-m)}. \tag{A.11}$$

Substituting (A.11) into (A.8) yields (in which $\hat{C}_{m,d}^{\rho,\nu}$ is another bounded scalar)

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

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

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

Here, $(a)$ is an obvious implication of the Holder inequality. The identity $(b)$ is exactly same as the second condition in Definition 2.1 and $(c)$ holds for the class of non-regular lattices satisfying Assumption 2.1. Hence

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

□

### *A.2. Sensitivity of $L_{n,m}^{\mathcal{B}}(\rho)$ with respect to $\rho$*

Recall that we defined $L_{n,m}^{\mathcal{B}}(\rho)$ 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 $\rho$, 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 7. The quantity $Q_N$, which will be defined in the next lemma, appears numerous times in this section.

**Lemma A.2.** Let $\rho_1, \rho_2$ be distinct points in $\Theta_0$ such that $\rho_2 > \rho_1$. Define

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

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 A.1 provides an upper bound on the term $f^N_{\boldsymbol{s}}(\boldsymbol{\omega})\, f^N_{\boldsymbol{t}}(\boldsymbol{\omega})$.

$$\left|f^N_{\boldsymbol{s}}(\boldsymbol{\omega})\, f^N_{\boldsymbol{t}}(\boldsymbol{\omega})\right| \lesssim \left(1 + \|\boldsymbol{\omega}\|^{2\nu+d}_{\ell_2}\right)^{-1}. \tag{A.12}$$

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

$$(1+x)^{-\alpha} - (1+y)^{-\alpha} < [\alpha(y-x)] \wedge \left(x^{-\alpha} - y^{-\alpha}\right), \quad \forall\, 0 < x < y < \infty, \quad \alpha > 0. \tag{A.13}$$

Using (A.13) (with $\alpha = \nu + \frac{d}{2}$) yields

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

The fact that $\Theta_0$ is compact and does not contain zero simplify the last inequality as the following.

$$\left|\frac{h_N\left(\rho_2 \|\boldsymbol{\omega}\|_{\ell_2}\right) - h_N\left(\rho_1 \|\boldsymbol{\omega}\|_{\ell_2}\right)}{\rho_2 - \rho_1}\right| \lesssim \left[\left(N\|\boldsymbol{\omega}\|_{\ell_2}\right)^{2\nu+d} \wedge \left(N\|\boldsymbol{\omega}\|_{\ell_2}\right)^{-2}\right]. \tag{A.14}$$

Combining (A.12) and (A.14) leads to

$$\frac{Q_N}{(\rho_2 - \rho_1)} \lesssim \int_{\mathbb{R}^d} \left[\left(N\|\boldsymbol{\omega}\|_{\ell_2}\right)^{2\nu+d} \wedge \left(N\|\boldsymbol{\omega}\|_{\ell_2}\right)^{-2}\right] \frac{d\boldsymbol{\omega}}{1 + \|\boldsymbol{\omega}\|^{2\nu+d}_{\ell_2}}$$

$$\overset{(b)}{\asymp} \int_0^\infty \left[(Nu)^{2\nu+d} \wedge (Nu)^{-2}\right] \frac{u^{d-1}du}{1+u^{2\nu+d}}$$
$$= N^{2\nu+d} \int_0^{\frac{1}{N}} \frac{u^{2\nu+2d-1}}{1+u^{2\nu+d}}du + \frac{1}{N^2}\int_{\frac{1}{N}}^\infty \frac{u^{d-3}}{1+u^{2\nu+d}}du \quad \text{(A.15)}$$

The change of variable $u = \|\boldsymbol{\omega}\|_{\ell_2}$ in the integral validates $\overset{(b)}{\asymp}$. For brevity, let $\psi_1$ and $\psi_2$ stand for the two expressions in the last line of (A.15), respectively from left to right. We ultimately introduce tight upper bounds on $\psi_1$ and $\psi_2$. 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,

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

Replacing the upper bounds on $\psi_1$ and $\psi_2$ into (A.15) 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. (A.13). 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 (A.13) 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

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

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), $\Delta$ and $L$ respectively refer to $\Delta^{\mathcal{B}}$ and $L^{\mathcal{B}}$.

**Lemma A.3.** Choose $\rho_1,\rho_2 \in \Theta_0$ such that $\rho_2 \neq \rho_1$. Then

$$\frac{\left\|\Delta^{\mathcal{B}}(\rho_1,\rho_2)\right\|_{\mathcal{S}_1}}{|\rho_2-\rho_1|} \lesssim \left(\mathbb{1}_{\{d=1\}} + \mathbb{1}_{\{d=2\}}\log N + \mathbb{1}_{\{d\geq 3\}} N^{d-2}\right). \tag{A.16}$$

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. (A.1) and (A.2), respectively. Now choose an arbitrary unit norm vector $v \in \mathbb{R}^n$ ($n = |D_n|$). Observe that

$$\begin{aligned}
&v^\top \Delta^{\mathcal{B}}(\rho_1,\rho_2)\, v \\
&\quad= \sum_{\boldsymbol{s},\boldsymbol{t}\in\mathcal{D}_n} v_{\boldsymbol{s}} v_{\boldsymbol{t}} \left(\Delta^{\mathcal{B}}(\rho_1,\rho_2)\right)_{\boldsymbol{s},\boldsymbol{t}} \\
&\quad= \sum_{\boldsymbol{s},\boldsymbol{t}\in\mathcal{D}_n} v_{\boldsymbol{s}} v_{\boldsymbol{t}} \left[\rho_2^{2\nu}\left(K^{\mathcal{B}}_{n,m}(\rho_2)\right)_{\boldsymbol{s},\boldsymbol{t}} - \rho_1^{2\nu}\left(K^{\mathcal{B}}_{n,m}(\rho_1)\right)_{\boldsymbol{s},\boldsymbol{t}}\right] \\
&\quad\overset{(a)}{=} \sum_{t=1}^{b_n}\sum_{\boldsymbol{s}\in B_t} v_{\boldsymbol{s}} v_{\boldsymbol{t}} \left[\rho_2^{2\nu}\left(K_{n,m}(\rho_2)\right)_{\boldsymbol{s},\boldsymbol{t}} - \rho_1^{2\nu}\left(K_{n,m}(\rho_1)\right)_{\boldsymbol{s},\boldsymbol{t}}\right] \\
&\quad\overset{(b)}{=} \sum_{t=1}^{b_n}\int_{\mathbb{R}^d}\sum_{\boldsymbol{s},\boldsymbol{t}\in B_t} v_{\boldsymbol{s}} v_{\boldsymbol{t}} e^{j\langle \boldsymbol{t}-\boldsymbol{s}, N\boldsymbol{\omega}\rangle} f^N_{\boldsymbol{s}}(\boldsymbol{\omega})\,\overline{f^N_{\boldsymbol{t}}(\boldsymbol{\omega})} \\
&\qquad\times \left[h_N\left(\rho_2\|\boldsymbol{\omega}\|_{\ell_2}\right) - h_N\left(\rho_1\|\boldsymbol{\omega}\|_{\ell_2}\right)\right] d\boldsymbol{\omega} \\
&\quad= \sum_{t=1}^{b_n}\int_{\mathbb{R}^d}\left|\sum_{\boldsymbol{s}\in B_t} v_{\boldsymbol{s}} e^{j\langle \boldsymbol{s}, N\boldsymbol{\omega}\rangle} f^N_{\boldsymbol{s}}(\boldsymbol{\omega})\right|^2 \left[h_N\left(\rho_2\|\boldsymbol{\omega}\|_{\ell_2}\right) - h_N\left(\rho_1\|\boldsymbol{\omega}\|_{\ell_2}\right)\right] d\boldsymbol{\omega} \\
&\quad\overset{(c)}{>} 0.
\end{aligned} \tag{A.17}$$

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. (A.3). Furthermore, inequality $(c)$ follows from the monotonicity of $h_N$. Now obviously we have

$$\begin{aligned}
|\mathcal{D}_n| \min_{\boldsymbol{s}\in\mathcal{D}_n}\left|\left(\Delta^{\mathcal{B}}(\rho_1,\rho_2)\right)_{\boldsymbol{s},\boldsymbol{s}}\right| &\leq \left\|\Delta^{\mathcal{B}}(\rho_1,\rho_2)\right\|_{\mathcal{S}_1} = \operatorname{tr}\left(\Delta^{\mathcal{B}}(\rho_1,\rho_2)\right) \\
&\leq |\mathcal{D}_n| \max_{\boldsymbol{s}\in\mathcal{D}_n}\left|\left(\Delta^{\mathcal{B}}(\rho_1,\rho_2)\right)_{\boldsymbol{s},\boldsymbol{s}}\right|.
\end{aligned}$$

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

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

Applying similar techniques as (A.17) as well as Lemma A.2 yields

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

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

$$\left(\Delta^{\mathcal{B}}(\rho_1,\rho_2)\right)_{\boldsymbol{s},\boldsymbol{s}}\geq\int_{\|\boldsymbol{\omega}\|_{\ell_2}\geq 1}\left|f_{\boldsymbol{s}}^N(\boldsymbol{\omega})\right|^2\left[h_N\left(\rho_2\|\boldsymbol{\omega}\|_{\ell_2}\right)-h_N\left(\rho_1\|\boldsymbol{\omega}\|_{\ell_2}\right)\right]d\boldsymbol{\omega}. \tag{A.18}$$

Let us control $h_N\left(\rho_2\|\boldsymbol{\omega}\|_{\ell_2}\right)-h_N\left(\rho_1\|\boldsymbol{\omega}\|_{\ell_2}\right)$ from below. Due to the fact that (its proof is similar to (A.13) 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

$$\frac{h_N\left(\rho_2\|\boldsymbol{\omega}\|_{\ell_2}\right)-h_N\left(\rho_1\|\boldsymbol{\omega}\|_{\ell_2}\right)}{\rho_2-\rho_1}\geq\frac{\left(\nu+\frac{d}{2}\right)}{2N^2\|\boldsymbol{\omega}\|_{\ell_2}^2}\frac{\rho_1+\rho_2}{\rho_1^2\rho_2^2}\gtrsim\left(N\|\boldsymbol{\omega}\|_{\ell_2}\right)^{-2}. \tag{A.19}$$

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 $\|\boldsymbol{\omega}\|_{\ell_2}\geq 1$, and decay with the polynomial rate given in Lemma A.1. So

$$\int_{\|\boldsymbol{\omega}\|_{\ell_2}\geq 1}\left|\frac{f_{\boldsymbol{s}}^N(\boldsymbol{\omega})}{\|\boldsymbol{\omega}\|_{\ell_2}}\right|^2 d\boldsymbol{\omega}\asymp 1, \quad \forall\, \boldsymbol{s}\in\mathcal{D}_n. \tag{A.20}$$

Replacing (A.20) and (A.19) into Eq. (A.18) gives the desirable lower bound. □

**Lemma A.4.** Let $\rho_1, \rho_2 \in \Theta_0$. Then

$$\left\|\Delta^{\mathcal{B}}(\rho_1,\rho_2)\right\|_{2\to 2} \lesssim \Big(1 \wedge |\rho_2-\rho_1|\Big)\left(1+\mathbb{1}_{\{m=\nu+d/2\}}\log N\right). \tag{A.21}$$

Moreover, if $\mathcal{D}_n$ be a $d$-dimensional regular lattice, then

$$\left\|\Delta^{\mathcal{B}}(\rho_1,\rho_2)\right\|_{2\to 2} \lesssim \Big(1 \wedge |\rho_2-\rho_1|\Big). \tag{A.22}$$

*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

$$\left\|\Delta^{\mathcal{B}}(\rho_1,\rho_2)\right\|_{2\to 2} \le \left\|\Delta(\rho_1,\rho_2)\right\|_{2\to 2}.$$

Hence, we just need to find an upper bound on $\|\Delta(\rho_1,\rho_2)\|_{2\to 2}$. 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 (A.17)) implies that

$$\left\|\Delta(\rho_1,\rho_2)\right\|_{2\to 2} \le \left\|L_{n,m}(\rho_2)\right\|_{2\to 2}. \tag{A.23}$$

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)

$$\begin{aligned}&\frac{v^\top \Delta(\rho_1,\rho_2)\, v}{\rho_2-\rho_1}\\ &= \int_{\mathbb{R}^d}\left|\sum_{\boldsymbol{s}\in\mathcal{D}_n} v_{\boldsymbol{s}} e^{j\langle \boldsymbol{s}, N\boldsymbol{\omega}\rangle} f_{\boldsymbol{s}}^N(\boldsymbol{\omega})\right|^2 \left\{\frac{h_N(\rho_2\|\boldsymbol{\omega}\|_{\ell_2}) - h_N(\rho_1\|\boldsymbol{\omega}\|_{\ell_2})}{\rho_2-\rho_1}\right\} d\boldsymbol{\omega}.\end{aligned}$$

The mean value theorem gives an alternative form for $h_N(\rho_2\|\boldsymbol{\omega}\|_{\ell_2}) - h_N(\rho_1\|\boldsymbol{\omega}\|_{\ell_2})$.

$$\begin{aligned}\exists\, \rho\in(\rho_1,\rho_2) \quad \text{s.t.} \quad &\frac{h_N(\rho_2\|\boldsymbol{\omega}\|_{\ell_2}) - h_N(\rho_1\|\boldsymbol{\omega}\|_{\ell_2})}{\rho_2-\rho_1}\\ &= \dot{h}_N(\rho\|\boldsymbol{\omega}\|_{\ell_2}) = \frac{2\nu+d}{\rho}\frac{h_N(\rho\|\boldsymbol{\omega}\|_{\ell_2})}{1+\left(N\rho\|\boldsymbol{\omega}\|_{\ell_2}\right)^2}.\end{aligned}$$

In following identity we show that $\sup_{\rho\in[\rho_1,\rho_2]}\dot{h}_N(\rho\|\boldsymbol{\omega}\|_{\ell_2}) \lesssim h_N(\rho_2\|\boldsymbol{\omega}\|_{\ell_2})$.

$$\begin{aligned}\frac{h_N(\rho_2\|\boldsymbol{\omega}\|_{\ell_2}) - h_N(\rho_1\|\boldsymbol{\omega}\|_{\ell_2})}{\rho_2-\rho_1} &\le \frac{2\nu+d}{\rho_1}\frac{h_N(\rho\|\boldsymbol{\omega}\|_{\ell_2})}{1+\left(N\rho\|\boldsymbol{\omega}\|_{\ell_2}\right)^2}\\ &\le \frac{2\nu+d}{\rho_1}h_N(\rho\|\boldsymbol{\omega}\|_{\ell_2})\\ &\lesssim h_N(\rho_2\|\boldsymbol{\omega}\|_{\ell_2}).\end{aligned} \tag{A.24}$$

The last inequality in (A.24) is implied from the fact that $\inf(\Theta_0) > 0$. Thus,

$$\begin{aligned} 0 &\leq \frac{v^\top \Delta(\rho_1,\rho_2)\, v}{\rho_2-\rho_1} \lesssim \int_{\mathbb{R}^d} \left| \sum_{\boldsymbol{s}\in\mathcal{D}_n} v_{\boldsymbol{s}} e^{j\langle \boldsymbol{s}, N\boldsymbol{\omega}\rangle} f_{\boldsymbol{s}}^N(\boldsymbol{\omega}) \right|^2 h_N\left(\rho_2 \|\boldsymbol{\omega}\|_{\ell_2}\right) d\boldsymbol{\omega} \\ &= v^\top L_{n,m}(\rho_2)\, v. \end{aligned}$$

In other words, there is a bounded constant $c > 1$ for which

$$\frac{\Delta(\rho_1,\rho_2)}{\rho_2-\rho_1} \preceq c L_{n,m}(\rho_2) \quad \Rightarrow \quad \frac{\|\Delta(\rho_1,\rho_2)\|_{2\to 2}}{\rho_2-\rho_1} \lesssim \|L_{n,m}(\rho_2)\|_{2\to 2}. \tag{A.25}$$

Combining (A.23) and (A.25) leads to

$$\|\Delta(\rho_1,\rho_2)\|_{2\to 2} \lesssim \left(1 \wedge |\rho_2-\rho_1|\right) \|L_{n,m}(\rho_2)\|_{2\to 2}.$$

In the case that $\mathcal{D}_n$ is a regular lattice, $\|L_{n,m}(\rho_2)\|_{2\to 2}$ is known to be less than some bounded scalar $C$ (see [16], Theorem 3.1), which justifies (A.22). For arbitrary irregular lattices satisfying Assumption 2.1, Proposition A.1 characterizes the decay rate of the off diagonal entries of $L_{n,m}(\rho_2)$. Thus, applying Lemma B.2 immediately substantiates (A.22) and ends the proof. □

**Lemma A.5.** Let $N \coloneqq \lfloor n^{1/d} \rfloor$ and select two distinct $\rho_1$ and $\rho_2$ in $\Theta_0$. 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 A.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 A.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 (A.25), we showed that

$$\|\Delta(\rho_1,\rho_2)\|_{\ell_2} \leq \|L_{n,m}(\rho_2)\|_{\ell_2} (\rho_2-\rho_1).$$

We also know from Proposition A.1 that

$$\left|\left(L_{n,m}(\rho_2)\right)_{\boldsymbol{s},\boldsymbol{t}}\right| \lesssim \left(1 + N\|\boldsymbol{t}-\boldsymbol{s}\|_{\ell_2}\right)^{-2(m-\nu)}, \tag{A.26}$$

which means that $\|L_{n,m}(\rho_2)\|_{\ell_2} \lesssim \sqrt{n}$ (see the second part of Lemma B.2). In summary for $d \geq 5$,

$$\|\Delta(\rho_1,\rho_2)\|_{\ell_2} \leq \|L_{n,m}(\rho_2)\|_{\ell_2} |\rho_2-\rho_1| \lesssim n^{1/2} |\rho_2-\rho_1|.$$ □

**Lemma A.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 $\Theta_0$. We have shown in the proof of Lemma A.3 (inequality (A.17)) 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. (A.1) and (A.2), respectively. Now choose an arbitrary $\boldsymbol{s} \in \mathcal{D}_n$ and a large enough $R \in (0, \infty)$. From the identity (A.3), we have a closed form expression for the diagonal entries of $L_{n,m}(\rho_{\min})$.

$$\begin{aligned} \left( L_{n,m}(\rho_{\min}) \right)_{\boldsymbol{s},\boldsymbol{s}} &= \int_{\mathbb{R}^d} \left| f^N_{\boldsymbol{s}}(\boldsymbol{\omega}) \right|^2 h_N\left( \rho_{\min} \left\| \boldsymbol{\omega} \right\|_{\ell_2} \right) d\boldsymbol{\omega} \\ &> \int_{\left\| \boldsymbol{\omega} \right\|_{\ell_2} \leq R} \left| f^N_{\boldsymbol{s}}(\boldsymbol{\omega}) \right|^2 h_N\left( \rho_{\min} \left\| \boldsymbol{\omega} \right\|_{\ell_2} \right) d\boldsymbol{\omega}. \end{aligned}$$

We can opt $N_0$ (depending on $\Theta_0$ and $R$) so that $\inf_{\left\| \boldsymbol{\omega} \right\|_{\ell_2} \leq R} h_N\left( \rho_{\min} \left\| \boldsymbol{\omega} \right\|_{\ell_2} \right) \geq \frac{1}{2}$ for any $N \geq N_0$. Thus,

$$\left| \left( L_{n,m}(\rho_{\min}) \right)_{\boldsymbol{s},\boldsymbol{s}} \right| > \frac{1}{2} \int_{\left\| \boldsymbol{\omega} \right\|_{\ell_2} \leq R} \left| f^N_{\boldsymbol{s}}(\boldsymbol{\omega}) \right|^2 d\boldsymbol{\omega}. \qquad \square$$

**Lemma A.7.** There exist a strictly positive scalars $C_1$ and $C_2$ such that

$$C_1 \sqrt{n} \geq \left\| \sqrt{L_{n,m}(\rho_1)} L^{\mathcal{B}}_{n,m}(\rho_2) \sqrt{L_{n,m}(\rho_1)} \right\|_{\ell_2} \geq C_2 \sqrt{n}, \quad \forall\, \rho_1, \rho_2 \in \Theta_0. \tag{A.27}$$

*Proof.* For brevity we use $Q$ to refer the Frobenius norm in Eq. (A.27). The cyclic permutation property of trace operator implies that

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

Based on Eq. (A.17), $L_{n,m}(\rho_1 \vee \rho_2) \succcurlyeq L_{n,m}(\rho_1)$ and $L^{\mathcal{B}}_{n,m}(\rho_1 \vee \rho_2) \succcurlyeq L^{\mathcal{B}}_{n,m}(\rho_2)$. So

$$\left\| \sqrt{L^{\mathcal{B}}_{n,m}(\rho_2)} L_{n,m}(\rho_1) \sqrt{L^{\mathcal{B}}_{n,m}(\rho_2)} \right\|^2_{\ell_2}$$

$$\leq \left\| \sqrt{L^{\mathcal{B}}_{n,m}(\rho_2)} L_{n,m}(\rho_1 \vee \rho_2) \sqrt{L^{\mathcal{B}}_{n,m}(\rho_2)} \right\|^2_{\ell_2}$$
$$= \left\| \sqrt{L_{n,m}(\rho_1 \vee \rho_2)} L^{\mathcal{B}}_{n,m}(\rho_2) \sqrt{L_{n,m}(\rho_1 \vee \rho_2)} \right\|^2_{\ell_2}$$
$$\leq \left\| \sqrt{L_{n,m}(\rho_1 \vee \rho_2)} L^{\mathcal{B}}_{n,m}(\rho_1 \vee \rho_2) \sqrt{L_{n,m}(\rho_1 \vee \rho_2)} \right\|^2_{\ell_2}.$$

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,

$$\left\| \sqrt{L_{n,m}(\rho_2)} L^{\mathcal{B}}_{n,m}(\rho_2) \sqrt{L_{n,m}(\rho_2)} \right\|^2_{\ell_2}$$
$$= \text{RHS} \coloneqq \operatorname{tr}\left\{ L_{n,m}(\rho_2) L^{\mathcal{B}}_{n,m}(\rho_2) L_{n,m}(\rho_2) L^{\mathcal{B}}_{n,m}(\rho_2) \right\}.$$

For any matrix $A$, define its absolute value by $|A| = [|A_{\boldsymbol{s},\boldsymbol{t}}|]$. The triangle inequality says that for matrices $A_1, \ldots, A_b$, for some $b \in \mathbb{N}$, we have

$$\operatorname{tr}(A_1 \ldots A_b) \leq \operatorname{tr}(|A_1| \ldots |A_b|).$$

This fact help us to find an upper bound on RHS.

$$\text{RHS} \leq \operatorname{tr}\left\{ |L_{n,m}(\rho_2)| \left|L^{\mathcal{B}}_{n,m}(\rho_2)\right| |L_{n,m}(\rho_2)| \left|L^{\mathcal{B}}_{n,m}(\rho_2)\right| \right\}.$$

Finally, since $\left|L^{\mathcal{B}}_{n,m}(\rho_2)\right|$ is the block diagonalized version of $|L_{n,m}(\rho_2)|$ and both of these matrices have non-negative entries, we get

$$\operatorname{tr}\left\{ |L_{n,m}(\rho_2)| \left|L^{\mathcal{B}}_{n,m}(\rho_2)\right| |L_{n,m}(\rho_2)| \left|L^{\mathcal{B}}_{n,m}(\rho_2)\right| \right\}$$
$$\leq \operatorname{tr}\left\{ |L_{n,m}(\rho_2)| |L_{n,m}(\rho_2)| |L_{n,m}(\rho_2)| |L_{n,m}(\rho_2)| \right\}$$
$$= \left\| |L_{n,m}(\rho_2)|^2 \right\|^2_{\ell_2}.$$

Combining the above inequalities yields

$$Q \leq \left\| |L_{n,m}(\rho_2)|^2 \right\|_{\ell_2}.$$

Notice that the off-diagonal entries of $L_{n,m}(\rho_2)$ and $|L_{n,m}(\rho_2)|$ decay with the same rate. Thus applying Lemma B.1 can determine an bound on the entries of $|L_{n,m}(\rho_2)|^2$ as the following.

$$\left|\left(\left|L_{n,m}\left(\rho_2\right)\right|^2\right)_{\boldsymbol{s},\boldsymbol{t}}\right| \lesssim \left(1+N\left\|\boldsymbol{t}-\boldsymbol{s}\right\|_{\ell_2}\right)^{-2(m-\nu)}\left\{1+\mathbb{1}_{\{m=\nu+d/2\}}\log\left(1+N\left\|\boldsymbol{t}-\boldsymbol{s}\right\|_{\ell_2}\right)\right\}.$$

Finally, Lemma B.3 guarantees the existence of a bounded scalar $c$ for which $\left\|L^2_{n,m}\left(\rho_2\right)\right\|_{\ell_2} \leq c\sqrt{n}$, 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

$$Q \geq \left\|\sqrt{L_{n,m}\left(\rho_1\right)}L^{\mathcal{B}}_{n,m}\left(\rho_1\right)\sqrt{L_{n,m}\left(\rho_1\right)}\right\|_{\ell_2}. \qquad \square$$

## Appendix B: Auxiliary results

In this section we collect the auxiliary propositions and lemmas which come in handy to substantiate the results in Section 7 and Appendix A.

### B.1. The basic properties of matrices with polynomial decaying off-diagonals

We showed in Appendix A.1 that the off-diagonal entries of $K^{\mathcal{B}}_{n,m}\left(\rho\right)$ 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 B.1.** Let $N = \lfloor n^{1/d} \rfloor$ and suppose that $A_n \in \mathbb{R}^{n\times n}$ whose entries satisfy

$$\left|A_{\boldsymbol{s},\boldsymbol{t}}\right| \leq C\left(1+N\left\|\boldsymbol{t}-\boldsymbol{s}\right\|_{\ell_2}\right)^{-(d+\zeta)}, \quad \forall\, \boldsymbol{s},\boldsymbol{t}\in\mathcal{D}_n. \tag{B.1}$$

for some bounded $C > 0$ and $\zeta \geq 0$. Then, the entries of $B = A^2$ are bounded above by

$$\left|B_{\boldsymbol{s},\boldsymbol{t}}\right| \lesssim \left(1+N\left\|\boldsymbol{t}-\boldsymbol{s}\right\|_{\ell_2}\right)^{-(d+\zeta)}\left\{1+\mathbb{1}_{\{\zeta=0\}}\log\left(1+N\left\|\boldsymbol{t}-\boldsymbol{s}\right\|_{\ell_2}\right)\right\}, \quad \forall \boldsymbol{s},\boldsymbol{t}\in\mathcal{D}_n. \tag{B.2}$$

*Proof.* For simplicity let $\Delta = N\left(\boldsymbol{t}-\boldsymbol{s}\right)$. Without loss of generality assume that $C = 1$. We first justify Eq. (B.2) 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 $\boldsymbol{s}\in\mathcal{D}_n$. The pairwise distances among two points in $\mathcal{D}_n$ have a similar behaviour to that of a $d$-dimensional regular lattice. Thus,

$$\begin{aligned}\left|B_{\boldsymbol{s},\boldsymbol{s}}\right| &= \left|\sum_{\boldsymbol{r}\in\mathcal{D}_n} A^2_{\boldsymbol{s},\boldsymbol{r}}\right| \leq \sum_{\boldsymbol{r}\in\mathcal{D}_n}\left(1+N\left\|\boldsymbol{r}-\boldsymbol{s}\right\|_{\ell_2}\right)^{-2(d+\zeta)} \\ &\lesssim \int_0^\infty x^{d-1}\left(1+x\right)^{-2(d+\zeta)}dx\end{aligned}$$

$$\lesssim \int_1^{\infty} x^{-(d+1+2\zeta)} dx \asymp 1.$$

Now suppose that $\Delta$ 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. (B.1) with the following more algebraically convenient alternative form.

$$|A_{\boldsymbol{s},\boldsymbol{t}}| \lesssim \left[1 + \|\Delta\|_{\ell_2}^{d+\zeta}\right]^{-1}, \quad \forall \boldsymbol{s}, \boldsymbol{t} \in \mathcal{D}_n, \quad \left(\boldsymbol{t} = \boldsymbol{s} + \frac{\Delta}{N}\right).$$

Next we obtain an upper bound on $|B_{\boldsymbol{s},\boldsymbol{t}}|$ as the sum of two terms.

$$\begin{aligned}
|B_{\boldsymbol{s},\boldsymbol{t}}| &\lesssim \sum_{\boldsymbol{r}\in\mathcal{D}_n} \frac{1}{\left(1 + \|N(\boldsymbol{s}-\boldsymbol{r})\|_{\ell_2}^{d+\zeta}\right)\left(1 + \|N(\boldsymbol{t}-\boldsymbol{r})\|_{\ell_2}^{d+\zeta}\right)} \\
&= \sum_{\boldsymbol{r}\in\mathcal{D}_n} \frac{\left(1 + \|N(\boldsymbol{s}-\boldsymbol{r})\|_{\ell_2}^{d+\zeta}\right)^{-1}}{2 + \|N(\boldsymbol{s}-\boldsymbol{r})\|_{\ell_2}^{d+\zeta} + \|N(\boldsymbol{t}-\boldsymbol{r})\|_{\ell_2}^{d+\zeta}} \\
&\quad + \sum_{\boldsymbol{r}\in\mathcal{D}_n} \frac{\left(1 + \|N(\boldsymbol{t}-\boldsymbol{r})\|_{\ell_2}^{d+\zeta}\right)^{-1}}{2 + \|N(\boldsymbol{s}-\boldsymbol{r})\|_{\ell_2}^{d+\zeta} + \|N(\boldsymbol{t}-\boldsymbol{r})\|_{\ell_2}^{d+\zeta}}.
\end{aligned}$$

We write $\xi_1$ and $\xi_2$ to denote the first and second terms in the last line of the above expression. The next step serves as controlling $\xi_1$ from above. A similar upper bound can be found on $\xi_2$. For doing so, we introduce a lower bound on the expression in the denominator of $\xi_1$. 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

$$\xi_1 \lesssim \sum_{\boldsymbol{r}\in\mathcal{D}_n} \frac{1}{\left(1 + \|N(\boldsymbol{s}-\boldsymbol{r})\|_{\ell_2}^{d+\zeta}\right)\left(1 + \|N(\boldsymbol{s}-\boldsymbol{r})\|_{\ell_2}^{d+\zeta} + \|\Delta\|_{\ell_2}^{d+\zeta}\right)}. \tag{B.3}$$

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 2.1 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. (B.3), 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 $\|N(\boldsymbol{s}-\boldsymbol{r})\|_{\ell_2}$)

$$\begin{aligned}
\xi_1 &\lesssim \int_0^N \frac{x^{d-1}dx}{(1+x^{d+\zeta})\left(1+x^{d+\zeta}+\|\Delta\|_{\ell_2}^{d+\zeta}\right)} \\
&= \frac{1}{\|\Delta\|_{\ell_2}^{d+\zeta}} \int_0^N \left(\frac{x^{d-1}}{1+x^{d+\zeta}} - \frac{x^{d-1}}{1+x^{d+\zeta}+\|\Delta\|_{\ell_2}^{d+\zeta}}\right) dx \\
&\lesssim \|\Delta\|_{\ell_2}^{-(d+\zeta)} \left[1+\mathbb{1}_{\{\zeta=0\}} \log\left(\frac{N^d\|\Delta\|_{\ell_2}^d}{N^d+\|\Delta\|_{\ell_2}^d}\right)\right] \\
&\asymp \|\Delta\|_{\ell_2}^{-(d+\zeta)} \left(1+\mathbb{1}_{\{\zeta=0\}} \log\|\Delta\|_{\ell_2}^d\right).
\end{aligned}$$

A similar bound holds for $\xi_2$. Replacing these upper bounds in $|B_{\boldsymbol{s},\boldsymbol{t}}| \lesssim \xi_1+\xi_2$ ends the proof. □

**Lemma B.2.** Let $\mathcal{D}_n$ be a irregular lattice of size $n$ satisfying Assumption 2.1. 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 $\zeta$ and $C \in (0,\infty)$. Then there exist bounded scalar $A, A' > 0$ (depending on $C, d$ and $\zeta$) for which

1. $\|\Psi^n\|_{2\to 2} \le A\Big(1+\mathbb{1}_{\{\zeta=0\}}\log n\Big)$.
2. $\|\Psi^n\|_{\ell_2} \le A'\sqrt{n}$.

*Proof.* We first focus on $\|\Psi^n\|_{2\to 2}$. The symmetry of $\Psi^n$ implies that

$$\begin{aligned}
\|\Psi^n\|_{2\to 2} &\le \sqrt{\|\Psi^n\|_{1\to 1}\|\Psi^n\|_{\infty\to\infty}} = \|\Psi^n\|_{1\to 1} = \max_{\boldsymbol{s}\in\mathcal{D}_n} \sum_{\boldsymbol{t}\in\mathcal{D}_n} \left|\Psi^n_{\boldsymbol{s},\boldsymbol{t}}\right| \\
&\le C \max_{\boldsymbol{s}\in\mathcal{D}_n} \sum_{\boldsymbol{t}\in\mathcal{D}_n} \left(1+N\|\boldsymbol{s}-\boldsymbol{t}\|_{\ell_2}\right)^{-(d+\zeta)}. \qquad \text{(B.4)}
\end{aligned}$$

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 2.1.

- There exists a bounded constant $a > 0$ such that $\Pi_{\boldsymbol{s},l} = \emptyset$ for any $l > aN$.
- $|\Pi_{\boldsymbol{s},l}| \lesssim (l+1)^d - l^d \lesssim (l+1)^{d-1}$ for any $l \le aN$.

Thus,

$$\sum_{\boldsymbol{t}\in\mathcal{D}_n}\left(1+N\left\|\boldsymbol{s}-\boldsymbol{t}\right\|_{\ell_2}\right)^{-(d+\zeta)}\leq\sum_{l=0}^{\infty}\left|\Pi_{\boldsymbol{s},l}\right|(l+1)^{-(d+\zeta)}\lesssim\sum_{l=0}^{aN}(l+1)^{-(1+\zeta)}. \tag{B.5}$$

We conclude the proof by substituting Eq. (B.5) into Eq. (B.4). Now we turn into finding an upper bound on $n^{-1}\left\|\Psi^n\right\|_{\ell_2}^2$. Using similar techniques as (B.5) yields

$$\begin{aligned} n^{-1}\left\|\Psi^n\right\|_{\ell_2}^2 &\leq n^{-1}\sum_{\boldsymbol{s}\in\mathcal{D}_n}\sum_{l=0}^{\infty}\left|\Pi_{\boldsymbol{s},l}\right|\sup_{\boldsymbol{t}\in\Pi_{\boldsymbol{s},l}}\left|\Psi^n_{\boldsymbol{s},\boldsymbol{t}}\right|^2\leq\sum_{l=0}^{\infty}\left|\Pi_{\boldsymbol{s},l}\right|\sup_{\boldsymbol{t}\in\Pi_{\boldsymbol{s},l}}\left|\Psi^n_{\boldsymbol{s},\boldsymbol{t}}\right|^2 \\ &\leq C^2\sum_{l=0}^{aN}\left|\Pi_{\boldsymbol{s},l}\right|(l+1)^{-2(d+\zeta)} \\ &\lesssim \sum_{l=0}^{\infty}(l+1)^{-(d+1+2\zeta)}\asymp 1. \end{aligned} \tag{B.6}$$

□

The next result has a similar flavor as the second part of Lemma B.2. We omit its proof for avoiding the repetition.

**Lemma B.3.** Let $\mathcal{D}_n$ be a irregular lattice of size $n$ satisfying Assumption 2.1. 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

$$\left|\Psi^n_{\boldsymbol{s},\boldsymbol{t}}\right|\leq C\left(1+N\left\|\boldsymbol{s}-\boldsymbol{t}\right\|_{\ell_2}\right)^{-(d+\zeta)}\left\{1+\mathbb{1}_{\{\zeta=0\}}\log\left(1+N\left\|\boldsymbol{s}-\boldsymbol{t}\right\|_{\ell_2}\right)\right\},$$
$$\forall\ \boldsymbol{s},\boldsymbol{t}\in\mathcal{D}_n$$

for some non-negative $\zeta$ and $C\in(0,\infty)$. Then there exists a bounded scalar $A>0$ (depending on $C,d$ and $\zeta$) for which

$$\left\|\Psi^n\right\|_{\ell_2}\leq A\sqrt{n}.$$

### *B.2. Probabilistic inequalities*

We first extend Proposition $A.3$ of [11] regarding the uniform concentration of generalized $\chi^2$ random processes around its mean. It provides a powerful tool in the proof of Theorems 4.1 and 4.2.

**Proposition B.1.** 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 $\Theta_0$, i.e.,

$$J_{\max}\coloneqq\sup_n\sup_{\theta\in\Theta_0}n^{-1/2}\left\|\Pi_n(\theta)\right\|_{\ell_2}<\infty.$$

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

$$\begin{aligned} &\|\Pi_n(\theta_2) - \Pi_n(\theta_1)\|_{2\to 2} \le C \log^2 n \, \|\theta_2 - \theta_1\|_{\ell_2}, \\ &\forall\, \theta_1, \theta_2 \in \Theta_0 \quad \text{s.t.} \, |\theta_2 - \theta_1| \le 1. \end{aligned} \tag{B.7}$$

(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

$$\mathbb{P}\left( \sup_{\theta\in\Theta_0} \left| Z^\top \Pi_n(\theta) Z - \operatorname{tr}\{\Pi_n(\theta)\} \right| \ge C' \sqrt{n \log n} \right) \le \frac{1}{n}, \quad \text{as } n \to \infty. \tag{B.8}$$

*Proof.* Let $r_n = 1/(C\sqrt{n \log^3 n})$ for $C$ defined in Eq. (B.7). For large enough $n$, we have $r_n \le 1$. Let $\mathcal{N}_{r_n}(\Theta_0)$ represents the $r_n$-covering number of $\Theta_0$. The simple volume argument implies that

$$|\mathcal{N}_{r_n}(\Theta_0)| \lesssim \left( \frac{\operatorname{diam}(\Theta_0)}{r_n} \right)^b = \mathcal{O}\left\{ \left(n \log^3 n\right)^{b/2} \right\}. \tag{B.9}$$

The key idea is to reduce the supremum over $\Theta_0$ in (B.8) 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 [15] concludes the proof.

For any $\theta \in \Theta_0$, let $\gamma_\theta$ stands for the closest element of $\mathcal{N}_{r_n}(\Theta_0)$ to $\theta$. Thus, $\|\theta - \gamma_\theta\|_{\ell_2} \le r_n$. Observe that

$$\begin{aligned} \text{RHS} &:= \left| Z^\top \Pi_n(\theta) Z - \operatorname{tr}\{\Pi_n(\theta)\} - Z^\top \Pi_n(\gamma_\theta) Z + \operatorname{tr}\{\Pi_n(\gamma_\theta)\} \right| \\ &= \left| \langle \Pi_n(\theta) - \Pi_n(\gamma_\theta), ZZ^\top + I_n \rangle \right| \\ &\le \|\Pi_n(\theta) - \Pi_n(\beta_\theta)\|_{2\to 2} \left\| ZZ^\top + I_n \right\|_{\mathcal{S}_1} \\ &\overset{(a)}{\le} C \log^2 n \, \|\theta - \beta_\theta\|_{\ell_2} \left\| ZZ^\top + I_n \right\|_{\mathcal{S}_1} \le C r_n \log^2 n \left\| ZZ^\top + I_n \right\|_{\mathcal{S}_1} \\ &= \sqrt{\frac{\log n}{n}} \left( n + \|Z\|_{\ell_2}^2 \right). \end{aligned}$$

Here $(a)$ is implied from Eq. (B.7). The Bernestein's inequality for the sub-exponential random variables states that

$$\mathbb{P}\left( \|Z\|_{\ell_2}^2 \ge n + nt \right) \le e^{-\frac{nt^2}{8}}, \quad \forall\, t > 0. \tag{B.10}$$

Choosing $t = 1$ in (B.10) shows that RHS $\ge 3\sqrt{n \log n}$ with probability at most $\exp(-n/8)$. Hence,

$$\mathbb{P}\Bigg( \sup_{\theta\in\Theta_0} \left| Z^\top \Pi_n(\theta) Z - \operatorname{tr}(\Pi_n(\theta)) \right|$$

$$\geq \sup_{\theta\in\mathcal{N}_{r_n}(\Theta_0)} \left|Z^\top \Pi_n(\theta) Z - \operatorname{tr}(\Pi_n(\theta))\right| + 3\sqrt{n\log n}\Bigg) \leq e^{-n/8}.$$

Recall $J_{\max}$ from the condition $(a)$. Choose an arbitrary bounded $\xi$ such that $\xi > 1 + b/2$. Eq. (B.9) can be rewritten as $|\mathcal{N}_{r_n}(\Theta_0)|\, n^{-\xi} = o\left(n^{-1}\right)$, when $n$ tends to infinity. The proof will be terminated if we show that (for some bounded scalar $C_0$)

$$\mathbb{P}\left(\sup_{\theta\in\mathcal{N}_{r_n}(\Theta_0)} \left|Z^\top \Pi_n(\theta) Z - \operatorname{tr}\{\Pi_n(\theta)\}\right| \geq C_0 J_{\max}\sqrt{n\log n}\right)$$
$$\leq |\mathcal{N}_{r_n}(\Theta_0)|\, n^{-\xi} = o\left(\frac{1}{n}\right),$$

as $n$ goes to infinity. For proving this claim, it suffices to obtain an appropriate probabilistic upper bound on $\left|Z^\top \Pi_n(\theta) Z - \operatorname{tr}\{\Pi_n(\theta)\}\right|$ for any $\theta \in \mathcal{N}_{r_n}(\Theta_0)$ and then exploiting the union bound trick. Hanson-Wright inequality [15] says that for some $C_0 < \infty$ (depending on $\xi$), we have

$$\mathbb{P}\left[\left|Z^\top \Pi_n(\theta) Z - \operatorname{tr}\{\Pi_n(\theta)\}\right| \geq C_0\left(\|\Pi_n(\theta)\|_{\ell_2}\sqrt{\log n} \vee \|\Pi_n(\theta)\|_{2\to 2}\log n\right)\right]$$
$$\leq n^{-\xi}. \tag{B.11}$$

The condition $(c)$ means that, $\|\Pi_n(\theta)\|_{2\to 2}\log n = o\left(\sqrt{n\log n}\right)$ as $n$ tends to infinity. So

$$\left(\|\Pi_n(\theta)\|_{\ell_2}\sqrt{\log n} \vee \|\Pi_n(\theta)\|_{2\to 2}\log n\right)$$
$$= \left(\|\Pi_n(\theta)\|_{\ell_2}\sqrt{\log n} \vee o\left(\sqrt{n\log n}\right)\right)$$
$$\leq J_{\max}\sqrt{n\log n}, \quad \text{as } n\to\infty,$$

due to the condition $(a)$. Thus Eq. (B.11) can be rewritten as

$$\mathbb{P}\left(\left|Z^\top \Pi_n(\theta) Z - \operatorname{tr}\{\Pi_n(\theta)\}\right| \geq C_0 J_{\max}\sqrt{n\log n}\right) \leq n^{-\xi}, \quad \forall\, \theta \in \mathcal{N}_{r_n}(\Theta_0),$$

ending the proof of the claim. □

Next we rigorously state the squeeze theorem for weak convergence. It is beneficial in the proof of Theorem 4.2.

**Lemma B.4.** Let $\{X_n\}_{n=1}^\infty, \{Y_n\}_{n=1}^\infty$ be two real valued sequences converging to $U$ in distribution. Suppose that $\{Z_n\}_{n=1}^\infty$ satisfies the following inequality

$$X'_n := X_n(1-p_n) \leq Z_n \leq Y'_n := Y_n(1+q_n), \quad \forall\, n\in\mathbb{N}, \tag{B.12}$$

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. (B.12) 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*. □

## References


[1] E. Anderes. On the consistent separation of scale and variance for gaussian random fields. *The Annals of Statistics*, 38(2):870–893, 2010. MR2604700

[2] M. Anitescu, J. Chen, and M. L. Stein. An inversion-free estimating equations approach for gaussian process models. *Journal of Computational and Graphical Statistics*, 26(1):98–107, 2017. MR3610411

[3] M. Anitescu, J. Chen, and L. Wang. A matrix-free approach for solving the parametric gaussian process maximum likelihood problem. *SIAM Journal on Scientific Computing*, 34(1):A240–A262, 2012. MR2890265

[4] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995. MR1346301

[5] J. Chen. On the use of discrete laplace operator for preconditioning kernel matrices. *SIAM Journal on Scientific Computing*, 35(2):A577–A602, 2013. MR3033083

[6] N. Cressie. Statistics for spatial data. *Terra Nova*, 4(5):613–617, 1992. MR1127423

[7] R. Furrer, M. G. Genton, and D. Nychka. Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics*, 15(3):502–523, 2006. MR2291261

[8] A. E. Gelfand, P. Diggle, P. Guttorp, and M. Fuentes. *Handbook of spatial statistics*. CRC Press, 2010. MR2730964

[9] C. Kaufman and B. Shaby. The role of the range parameter for estimation and prediction in geostatistics. *Biometrika*, 100(2):473–484, 2013. MR3068447

[10] C. G. Kaufman, M. J. Schervish, and D. W. Nychka. Covariance tapering for likelihood-based estimation in large spatial data sets. *Journal of the American Statistical Association*, 103(484):1545–1555, 2008. MR2504203

[11] H. Keshavarz, C. Scott, and X. Nguyen. On the consistency of inversion-free parameter estimation for gaussian random fields. *Journal of Multivariate Analysis*, 150:245–266, 2016. MR3534913

[12] H. Keshavarz, C. Scott, and X. Nguyen. Optimal change point detection in gaussian processes. *Journal of Statistical Planning and Inference*, 193:151–178, 2018. MR3713470

[13] H. Keshavarz Shenastaghi. Detection and estimation in gaussian random fields: Minimax theory and efficient algorithms. 2017. MR3768547

[14] M. Lee. *Local properties of irregularly observed Gaussian fields*, volume 74. 2012. MR3078578

[15] M. Rudelson and R. Vershynin. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18, 2013. MR3125258

[16] M. L. Stein. *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media, 2012. MR1697409

[17] M. L. Stein, J. Chen, and M. Anitescu. Difference filter preconditioning for large covariance matrices. *SIAM Journal on Matrix Analysis and Applications*, 33(1):52–72, 2012. MR2902671

[18] M. L. Stein, J. Chen, M. Anitescu, et al. Stochastic approximation of score functions for gaussian processes. *The Annals of Applied Statistics*, 7(2):1162–1191, 2013. MR3113505

[19] M. L. Stein, Z. Chi, and L. J. Welty. Approximating likelihoods for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(2):275–296, 2004. MR2062376

[20] A. V. Vecchia. Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 297–312, 1988. MR0964183

[21] D. Wang and W.-L. Loh. On fixed-domain asymptotics and covariance tapering in gaussian random field models. *Electronic Journal of Statistics*, 5:238–269, 2011. MR2792553

[22] Z. Ying. Asymptotic properties of a maximum likelihood estimator with data from a gaussian process. *Journal of Multivariate Analysis*, 36(2):280–296, 1991. MR1096671

[23] H. Zhang. Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association*, 99(465):250–261, 2004. MR2054303