Journal of Multivariate Analysis 150 (2016) 245–266

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

On the consistency of inversion-free parameter estimation for Gaussian random fields☆

Hossein Keshavarz [a,*], Clayton Scott [b,a], XuanLong Nguyen [a,b]

[a] Department of Statistics, University of Michigan, United States

[b] Department of Electrical Engineering and Computer Science, University of Michigan, United States

ARTICLE INFO

Article history:
Received 15 January 2016
Available online 4 July 2016

AMS subject classification:
primary 62M30, 62M40
secondary 60G15

Keywords:
Inversion-free estimation
Covariance function
Stationary Gaussian process
Asymptotic analysis

ABSTRACT

Gaussian random fields are a powerful tool for modeling environmental processes. For high dimensional samples, classical approaches for estimating the covariance parameters require highly challenging and massive computations, such as the evaluation of the Cholesky factorization or solving linear systems. Recently, Anitescu et al. (2014) proposed a fast and scalable algorithm which does not need such burdensome computations. The main focus of this article is to study the asymptotic behavior of the algorithm of Anitescu et al. (ACS) for regular and irregular grids in the increasing domain setting. Consistency, minimax optimality and asymptotic normality of this algorithm are proved under mild differentiability conditions on the covariance function. Despite the fact that ACS's method entails a non-concave maximization, our results hold for any stationary point of the objective function. A numerical study is presented to evaluate the efficiency of this algorithm for large data sets.

1. Introduction

Gaussian processes have plethora of applications, ranging from the modeling of environmental processes in *geostatistics* (e.g., [9,11]) to supervised regression and classification in *machine learning* [5,8,24], and the simulation of complex computer models [18]. The versatility of the correlation structure of Gaussian processes provides a tractable and powerful tool for the modeling of large and highly dependent environmental variables. As a common approach in the field of spatial statistics, the covariance functions of Gaussian processes are assumed to belong to a parametric family. High precision estimates of the covariance parameters are pivotal for interpolating Gaussian processes which is the ultimate goal in many geostatistical problems [9,19].

In the last two decades, there has been extensive research regarding the statistical and computational facets of the estimation of the Gaussian processes' covariance parameters. Maximum likelihood estimation (MLE) was the earliest favored algorithm in the geostatistics community, e.g., Mardia et al. [14] and Ying [25]. However, solving systems of linear equations is inevitable to evaluate the Gaussian likelihood. Notwithstanding the recent advances toward scalable and efficient solution of the system of linear equation (e.g., iterative *Krylov* subspace method or block preconditioned conjugate gradient algorithm [16]) which moderately reduces the computational and memory costs of the direct evaluation of the

☆ This research is partially supported by NSF grant ACI-1047871. Additionally, CS is partially supported by NSF grants 1422157, 1217880, and 0953135, and LN by NSF CAREER award DMS-1351362, NSF CNS-1409303, and NSF CCF-1115769.

* Corresponding author.
E-mail addresses: hksh@umich.edu (H. Keshavarz), clayscot@umich.edu (C. Scott), xuanlong@umich.edu (X. Nguyen).

http://dx.doi.org/10.1016/j.jmva.2016.06.003
0047-259X/

precision matrix, obtaining the MLE of unknown covariance parameters using such linear systems solvers is still a strenuous task, especially for a generic Gaussian spatial process observed at numerous and possibly irregularly spaced locations. Approximating the likelihood function by tapering the covariance matrix is another class of algorithms aiming to reduce the numerical burden of MLE (see Kaufman et al. [12]). Due to the sparsity of the tapered covariance matrix, its inverse can be computed in a faster and more stable way. Recent studies [10,12,22] demonstrate the consistency and asymptotic normality of this algorithm under some mild conditions on the taper function.

Because of the obstacles of solving system of linear equations for massive data, which is necessary for tapered and exact MLE, it is of great interest to develop estimation techniques without requiring such extensive computations. Such class of algorithms, which will be referred to as *inversion-free*, are based upon optimizing a loss function whose form (and its derivatives) is independent of the precision matrix of data. The first attempt toward such a goal has been done by Anitescu, Chen and Stein [2] in 2014 (referred here to as ACS). Their proposed procedure is faster and more stable than likelihood based algorithms. In [2], the covariance parameters are estimated by computing the global maximizer of a non-concave program. Simulation studies verify the efficiency of ACS's approach in the case that the covariance matrix has a bounded condition number. The main purpose of this paper is to appraise the asymptotic properties of the ACS's algorithm such as consistency, minimax optimality and asymptotic normality. The developed theory in this paper shows that ACS's algorithm has the same asymptotic rate of convergence as the MLE. In practice, the solution of ACS's optimization problem may also serve as the starting point (initial guess) of a likelihood maximization procedure.

In geostatistics, there are two common asymptotic regimes: *increasing-domain* and *fixed-domain*, the latter sometimes referred to as infill asymptotics (see [19], Section 3.3 or [9], p. 480). In the former setting, the minimum distance among the sampling points is bounded away from zero and more samples are collected by increasing the diameter of the spatial domain. In the latter regime, the data are sampled in a fixed and bounded domain, and the observations get denser as the sample size n increases.

Zhang [26] showed that not all the covariance parameters are consistently estimable in the fixed domain regime. Strictly speaking, there is no asymptotically consistent algorithm for estimating the *non-micro ergodic* covariance parameters, which do not asymptotically affect the interpolation mean square error (see [19] for a precise definition). On the other hand, it is known in the literature that subject to some mild regularity conditions, maximizing the likelihood provides a strongly consistent and asymptotically normal estimate for all the covariance parameters in the increasing domain setting [4,14].

Increasing domain asymptotic analysis of covariance estimation has two significant benefits. First, unlike the infill asymptotic setting, the geometry of the spatial sampling has a crucial impact on the asymptotic distribution of the parameter estimate. Thus, this regime is a natural asymptotic framework for assessing the role of irregularity of spatial sampling on the covariance parameter estimation [4]. This claim can be verified by a deeper look at the asymptotic distribution of the microergodic parameter estimates in the fixed domain (see e.g., [19,25] for MLE and [10,12,22] for tapered MLE). Another significant characteristic of increasing domain regime is that the covariance matrix has a universally bounded condition number as n grows under some mild regularity conditions. This feature of the covariance matrix plays a major role in our asymptotic analysis. Although in many geostatistical applications in a fixed bounded domain the condition number of the covariance matrix increases at least linearly with respect to n, preconditioning filters is commonly used to uniformly control the condition number independent of n [7,20]. Therefore, we believe that our developed increasing domain asymptotics can be useful for the fixed domain analysis of preconditioned inversion-free algorithms.

Outline of main results. This paper studies the increasing domain asymptotic behavior of ACS's estimation algorithm introduced in [2]. Specifically, suppose that $\mathfrak{G}$ is a zero mean stationary Gaussian process in $\mathbb{R}^d$ with covariance function $\operatorname{cov}\left(\mathfrak{G}(s), \mathfrak{G}\left(s'\right)\right) = R\left(s - s', \eta\right)$ in which $\eta \in \Omega$ denotes the vector of unknown covariance parameters. One realization of $\mathfrak{G}$ has been observed on a *d-dimensional perturbed regular lattice* of $n = N^d$ points, which will be formally defined in Section 2. The specific contributions of this work are given as follows:

(a) Assuming the polynomial decay of $R(s, \eta)$ and its gradient (with respect to η) in terms of the Euclidean norm of s, and under some mild identifiability condition on R, we prove that the global maximizer of ACS's method consistently estimates η. Furthermore, the estimation error is of order $\sqrt{n^{-1} \ln n}$ which is shown to be minimax optimal up to some $\sqrt{\ln n}$ term.
(b) As the proposed loss function in Anitescu et al. [2] is not jointly concave in η, finding its global maximizer is challenging. For a large enough sample size and under an additional condition regarding the polynomial decay of the second derivative of $R(s, \eta)$ with respect to η, we show that any *stationary point* of this non-concave program is concentrated around the true η with radius of order $\sqrt{n^{-1} \ln n}$.
(c) The asymptotic normality of the stationary points of the aforementioned algorithm will be substantiated under some mild restriction on the third derivative of R with respect to η.

Furthermore, the Appendix contains several easy-to-reference nonasymptotic results on the global and local behavior of the quadratic forms of Gaussian processes which may come in hand for the analysis of certain problems in statistics and machine learning.

Plan of the paper. In Section 2, we formulate ACS's inversion-free estimation method and precisely introduce the geometry of the sampling points. Section 3 expresses the necessary assumptions and studies the asymptotic properties of the estimation algorithm. Section 3.1 presents the convergence rate of the global and local maximizers of the optimization problem

introduced in Section 2. We investigate the minimax optimality and the asymptotic normality of the local maximizers in Section 3.2. The objective of Section 4 is to assess the performance of ACS's algorithm and verify the developed theory using simulation studies on synthetic data. Section 5 serves as the conclusion and discusses the future directions. Section 6 presents the proof of the main results. Finally, the Appendix contains some auxiliary technicalities on the nonasymptotic behavior of the quadratic forms of Gaussian processes and of large covariance matrices with polynomially decaying off-diagonal entries, which are essential in Section 6.

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

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

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

2. Problem set up and ACS's estimation algorithm

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

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

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

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

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

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

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

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

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

The condition $\delta \in [0, 1/2)$ guarantees the existence of a strictly positive minimum distance $(1 - 2\delta)$ between the distinct points in $\mathcal{D}_n$. The scalar δ quantifies the amount of irregularity in $\mathcal{D}_n$. In the case of $\delta = 0$, $\mathcal{D}_n$ forms a regular lattice and the irregularity can be more apparent as δ increases. Although the absence of randomness in Assumption 2.1 may appear problematic at first sight, our theoretical contributions are not restricted to any further set of strong conditions on p_i's. For

instance, the presented results in the next section hold almost surely if p_i's are independent (or even dependent) draws of a distribution supported on $[-1, 1]^d$ which is absolute continuous with respect to the Lebesgue measure.

Now we present ACS's estimation algorithm introduced in [2]. Define,

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

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

$$\left(\hat{\phi}_n, \hat{\theta}_n\right) = \underset{(\phi,\theta)\in\mathcal{A}\times\Theta}{\operatorname{argmax}} F_n(Y,\phi,\theta), \quad \text{where } F_n(Y,\phi,\theta) := \frac{1}{n}\left\{\phi Y^\top K_n(\theta) Y - \frac{\phi^2}{2}\|K_n(\theta)\|_{\ell_2}^2\right\}. \tag{2.3}$$

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

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

1. As stated in [2], the true parameter η_0 is a stationary point of the expected value of $F_n(Y,\eta)$. That is,

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

   Roughly speaking, η_0 is located in a small neighborhood of a stationary point of (2.2) if the gradient of $F_n(Y,\eta)$ is smooth enough and concentrated around its expected value. The next fact, which has not been stated in [2], reveals a profound connection of (2.2) to MLE.
2. For brevity, define $H_n(Y,\eta) := F_n(Y,\eta) - F_n(Y,\eta_0)$ and $L_n(\eta,\eta_0) := R_n^{1/2}(\eta_0) R_n^{-1}(\eta) R_n^{1/2}(\eta_0)$. Also, let $\tilde{\eta}_n$ denotes the MLE of η. Obvious calculations lead to

$$\hat{\eta}_n = \underset{\eta\in\Omega}{\operatorname{argmax}} H_n(Y,\eta),$$

$$\tilde{\eta}_n = \underset{\eta\in\Omega}{\operatorname{argmax}} H_n'(Y,\eta), \quad \text{where } H_n'(Y,\eta) := \frac{1}{n}\left\{-\log\det L_n(\eta,\eta_0) - n + Y^\top R_n^{-1}(\eta) Y\right\}.$$

   Notice that $\mathrm{E}H_n(Y,\eta_0) = \mathrm{E}H_n'(Y,\eta_0) = 0$. Under Assumption 2.1 and using similar techniques as Appendix, one can guarantee the existence of a universal scalar $C \in (0,\infty)$ such that

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

   Namely, in the increasing domain regime, the objective function proposed in [2] can be viewed as an approximate *minorizing surrogate* of the likelihood function in the expected value sense (it forms a perfect minorizer whenever $C = 1$ in (2.4)).

3. Main results

We establish the asymptotic characteristics of the estimation algorithm in (2.2). Section 3.1 examines the consistency of the global maximizer and the stationary points of (2.2) under some sufficient conditions on Ω and the correlation function $K(\cdot,\theta)$. The near minimax optimality and the asymptotic normality of the stationary points will be covered in Section 3.2

3.1. Consistency and the convergence rate

The following assumptions are assumed on the parameter space $\Omega = \mathcal{A}\times\Theta$ and the correlation function $K(\cdot,\theta)$ for studying the asymptotic behavior of the global maximizer of (2.2). Similar but slightly stronger conditions have been used in [4] for the increasing domain asymptotic analysis of MLE.

Assumption 3.1. The following conditions are satisfied by Ω and K.

(A1) Θ and $\mathcal{A}$ are *compact connected* subsets of $\mathbb{R}^m$ and $(0,\infty)$, respectively.
(A2) There are bounded scalars $M > 0$ and $r_1 > 1$ such that for any $s\in\mathcal{D}_n$,

$$\max_{s'\in\mathcal{D}_n(s,r_1)} \left|K(s'-s,\theta_2) - K(s'-s,\theta_1)\right| \ge M\|\theta_2-\theta_1\|_{\ell_2}, \quad \forall\,\theta_1,\theta_2\in\Theta. \tag{3.1}$$

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

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

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

Condition (A2), assuring the *identifiability* of θ from the K, holds for the standard class of correlation functions such as Matern, powered exponential and rational quadratic. A detailed look at (A2) is postponed to the end of this section. Before commenting on (A3), let us define the family of geometric anisotropic covariance functions.

Definition 3.1. Let $\mathfrak{G} : \mathbb{R}^d \mapsto \mathbb{R}$ be a zero mean stationary Gaussian process in $\mathbb{R}^d$. Then $\mathfrak{G}$ is called *geometric anisotropic* if

$$R\left(s-s', \eta_0\right) := \mathrm{E}\mathfrak{G}(s)\,\mathfrak{G}\left(s'\right) = \phi_0 K\left(\sqrt{(s-s')^\top A_0 (s-s')}\right), \quad \forall s, s' \in \mathbb{R}^d, \tag{3.2}$$

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

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

$$\Lambda_{\min,\Theta} \leq \min_{A_0\in\Theta}\frac{1}{\left\|A_0^{-1}\right\|_{2\to 2}} \leq \max_{A_0\in\Theta}\|A_0\|_{2\to 2} \leq \Lambda_{\max,\Theta}, \tag{3.3}$$

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

Remark 3.1. Any compactly supported correlation function, such as *spherical or Wendland family* [23] on $\mathbb{R}^d$ trivially admits (A3). Assumption (A3) with $q = 1$ also holds for some classical families of geometric anisotropic covariances such as:

(a) *Matern*: The Gaussian process $\mathfrak{G}$ has Matern covariance function if it fulfills (3.2) with

$$K(r) = \frac{2^{1-\nu_0}}{\Gamma(\nu_0)} r^{\nu_0} \mathfrak{K}_{\nu_0}(r), \tag{3.4}$$

in which ν_0 is an unknown, strictly positive scalar lies in a compact space. Moreover, $\Gamma(\cdot)$ and $\mathfrak{K}_{\nu_0}(\cdot)$ represent the Gamma function and the modified Bessel function of the second kind, respectively. The parametric Matern family satisfies (A3) provided condition (3.3).

(b) *Powered exponential*: A covariance function in this class satisfies (3.2) with $K(r) = e^{-r^{\nu_0}}$ and $\nu_0 \in (0, 2)$. Like Matern class, assuming (3.3), any member of powered exponential family fulfills (A3) with $q = 1$.

(c) *Rational quadratic*: The elements of this class are of the form (3.2) with $K(r) = \left(1+r^2\right)^{-\left(\frac{d}{2}+\nu_0\right)}$ and $\nu_0 > 0$. For the case of known fractal index, (A3) with $q = 1$ is valid, if (3.3) holds. Note that for unknown ν_0 the exact same statement is satisfied under a slightly stronger condition of $\nu_0 > 1/2$.

Parts (b) and (c) of Remark 3.1 are verifiable by straightforward algebra and differentiation rules. In order to demonstrate part (a), see [1] for the derivative properties of the Bessel function of the second kind (with respect to the entries of A_0) and see Lemma A.5 for the asymptotic behavior of the partial derivatives of the Matern covariance with respect to ν_0. Now, we state the first significant result of this section regarding the consistency of the global maximizer of (2.3) under Assumption 3.1 and perturbed regular lattice sampling.

Theorem 3.1. *Suppose that Assumptions 2.1 and 3.1 with $q = 1$ hold for $\mathcal{D}_n$, Ω and K. Then the maximizer of (2.3) satisfies*

$$\Pr\left(\left\|\hat{\theta}_n - \theta_0\right\|_{\ell_2} \vee \left|\frac{\hat{\phi}_n}{\phi_0} - 1\right| \geq C\sqrt{\frac{\ln n}{n}}\right) \to 0, \quad \text{as } n \to \infty, \tag{3.5}$$

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

Remark 3.2. Let $\phi_{\min}$ and $\phi_{\max}$ denote the smallest and largest elements in l. Obviously, $\phi_{\min}$ and $\phi_{\max}$ are well defined and finite due to (A1). Moreover,

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

Thus (3.5) is a stronger statement than $\left\| \hat{\eta}_n - \eta_0 \right\|_{\ell_2} = \mathcal{O}_{\mathbb{P}} \left(\sqrt{n^{-1} \ln n} \right)$ and they are equivalent when $\phi_{\min} > 0$ (which is true under A1).

An analogous consistency result has been proved recently by Bachoc [4] for the MLE and cross validation estimator. Based upon Theorem 3.1, the asymptotic rate of ACS's algorithm has not been sacrificed for increasing the speed and memory efficiency compared to the MLE.

Finally, we concisely address the role of the identifiability condition (A2) in Theorem 3.1. Actually, (A2) plays a decisive role in translating consistent estimation of the correlation matrix (in the relative sense) to η_0. Strictly speaking, (A2) is required to deduce (3.5) from the probabilistic statement

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

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

$$\hat{\phi}_n = \frac{Y^\top K_n(\hat{\theta}_n) Y}{\left\| K_n(\hat{\theta}_n) \right\|_{\ell_2}^2}. \tag{3.6}$$

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

$$\hat{\theta}_n = \underset{\theta \in \Theta}{\operatorname{argmax}} \; G_n(Y, \theta), \quad \text{where } G_n(Y, \theta) = \frac{Y^\top K_n(\theta) Y}{\|K_n(\theta)\|_{\ell_2}}. \tag{3.7}$$

Note that for large n, computing the global maximizer of (3.7) can be less intensive than (2.2) due to searching over a smaller space Θ. We first visually assess the key properties of F_n in some simple scenarios. In Fig. 1, $G_n(Y, \theta)$ (which is a univariate function of scalar θ) has been plotted versus θ for the two dimensional ($d = 2$) isotropic Matern covariance function in two different scenarios. In the left panel the isotropic Gaussian process $\mathfrak{G}$ has been generated with the parameters $(\phi_0, \theta_0, \nu_0) = (1, 4, 0.5)$ and has been sampled in a randomly perturbed regular lattice with $\delta = 0.2$ and of size $N = 100$. In the right panel, the covariance parameters are given by $(\phi_0, \theta_0, \nu_0) = (1, 6, 1.5)$ and the GP is sampled at a randomly generated perturbed regular lattice with $N = 100$ and $\delta = 0.2$. As is apparent from Fig. 1, for these two parsimonious scenarios $G_n(Y, \theta)$ is not a concave function of θ and has a single *inflection point*. However, $G_n(Y, \theta)$ has only one stationary point which coincides with its global maximizer. In the following, we rigorously study the large sample behavior of the stationary points of $G_n(Y, \theta)$ (as well as F_n) in a generic case. We initiate our analysis by stating the sufficient conditions on K and Ω.

Assumption 3.2. (A1) holds for Ω, and K fulfills (A2) and (A3) with $q = 2$ in Assumption 3.1.

Remark 3.3. The analysis of the stationary points of (2.3) requires a slightly stronger conditions than that of the global maximizer in Assumption 3.1. The main distinction is the polynomial decay of the second order derivative of K with respect to θ. Note that the new condition on the second derivative of K is not too restrictive. For instance the same analysis as Remark 3.1 validates this condition for all covariance families introduced in Remark 3.1 (with a larger constant $C_{K,\Theta}$).

Theorem 3.2. *Suppose that $\mathcal{D}_n$ admits Assumption 2.1 and Assumption 3.2 holds for Ω and K. Then any stationary point of the optimization problem (2.3) satisfies*

$$\lim_{n \to \infty} \Pr \left(\left\| \hat{\theta}_n - \theta_0 \right\|_{\ell_2} \vee \left| \frac{\hat{\phi}_n}{\phi_0} - 1 \right| \geq C \sqrt{\frac{\ln n}{n}} \right) = 0, \tag{3.8}$$

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

Theorem 3.2 shows that any stationary point of F_n is concentrated in a small neighborhood of (ϕ_0, θ_0), with high probability. In other words, $F_n(Y, \phi, \theta)$ shows a similar behavior as Fig. 1 in the general case. In addition, the comparison between (3.5) and (3.8) reveals that stationary points converge to (ϕ_0, θ_0) with the same rate as the global maximizer.

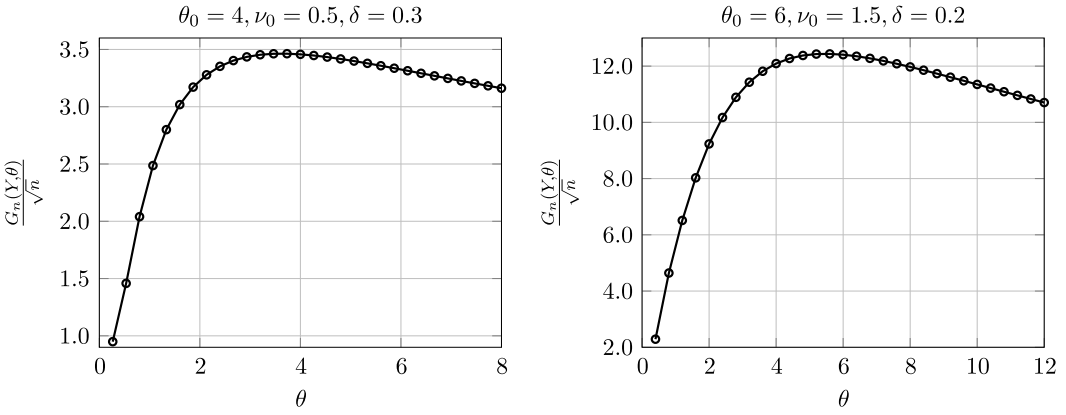

Fig. 1. The above figures exhibit $n^{-1/2}G_n(Y,\theta)$ for the isotropic Matern covariance function (with known ν_0). In the left panel, $\theta_0 = 4$, $\nu_0 = 0.5$ and the spatial samples form a two dimensional randomly perturbed regular lattice of size $N = 100$ with $\delta = 0.3$. In the right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

We conclude this section by illustrating how restrictive the identifiability assumption (A2) can be for the frequently used classes of the covariance functions. We first introduce a slightly stronger identifiability condition than (A2), which will be referred to as (A4). Note that a slightly modified version of (A4) has been first introduced in [4] for studying the increasing domain asymptotics of the maximum likelihood and cross validation algorithms. That is to say, these identifiability conditions are not exclusive to ACS method and pop up in the asymptotic analysis of other algorithms. The proof of the subsequent results in this section will be omitted. We refer the reader to [13] for a detailed proof.

Proposition 3.1. (A2) *is satisfied, whenever*

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

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

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

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

Clearly, (A4.b) is necessary for any algorithm consistently estimating η and it can be verified for all typical classes of geometrical anisotropic covariance function. However, understanding the role and restrictiveness of (A4.a) is more subtle than that of (A4.b). Note that unlike (A3), all the introduced identifiability conditions not only depend on the choice of the covariance function but also to the observed locations $\mathcal{D}_n$. It may be excessive to seek the class of covariances satisfying (A4.a) for any perturbed lattice $\mathcal{D}_n$. So, a more pertinent question is: which class of covariance functions do almost surely satisfy (A4.a) for a randomly generated perturbed lattice? The following result responds to question by rigorously characterizing a broad subclass of the geometrically anisotropic covariances (as defined in Definition 3.1) fulfilling (A4.a).

Proposition 3.2. *Let* $\mathfrak{P}$ *be a distribution in* $[-1, 1]^d$ *which is absolutely continuous with respect to the Lebesgue measure. Suppose that* $R(\cdot, \eta) : \mathbb{R}^d \mapsto [0, \infty)$ *is a geometrically anisotropic covariance function with a known* ν_0 *(if exists). Then,* (A4) *almost surely holds if*

(a) $K : [0, \infty) \mapsto [0, \infty)$ *is a nonzero, differentiable and strictly decreasing function (K may only have right derivative at zero).*
(b) $\mathcal{D}_n$ *is a randomly generated* δ*-perturbed lattice associated to* $\mathfrak{P}$*. That is,* p_i *are independent draws of* $\mathfrak{P}$ *in Assumption 2.1.*

Corollary 3.1. *Let* $\mathcal{D}_n$ *be d-dimensional regular lattice (associated to* $\delta = 0$*) and assume that* $R(\cdot, \eta) : \mathbb{R}^d \mapsto [0, \infty)$ *is a geometrically anisotropic covariance function with known* ν_0*. Then,* (A4) *holds if* $K : [0, \infty) \mapsto [0, \infty)$ *admits the condition* (a) *in Proposition 3.2.*

Although the conditions of Proposition 3.2 trivially hold for the non-compactly supported covariance function introduced in Remark 3.1, deploying analogous proof techniques can lead to a similar result for compactly supported covariance function.

Proposition 3.3. *Suppose that* $\mathfrak{P}$*,* $R(\cdot, \eta)$ *and* $\mathcal{D}_n$ *satisfy the same conditions as Proposition 3.2. Then,* (A4) *almost surely holds if there exists a large enough positive scalar* r_0 *for which*

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

Although the required conditions on the covariance function's formulation, in Propositions 3.2 and 3.3, are very minimal, we assume that the fractal index ν_0 (if exists) is fully known. However in the following result, ν_0 is one of the unknown parameters to be estimated. Here the central emphasis is on the powered exponential and rational quadratic classes, as their partial derivative with respect to the fractal index have a somewhat simple closed form that can be handled without great difficulty.

Proposition 3.4. *Let* $\mathfrak{P}$ *be a distribution in* $[-1, 1]^d$ *which is absolutely continuous with respect to the Lebesgue measure. Suppose that* $R(\cdot, \eta) : \mathbb{R}^d \mapsto [0, \infty)$ *is a geometrically anisotropic covariance function. Then,* (A4) *almost surely holds if*

(a) $K : [0, \infty) \mapsto [0, \infty)$ *is either a powered exponential or rational quadratic covariance functions in Remark 3.1 with unknown* $\nu_0 > 0$.
(b) $\mathcal{D}_n$ *is a randomly generated* δ*-perturbed lattice associated to* $\mathfrak{P}$. *That is,* p_i *are independent draws of* $\mathfrak{P}$ *in Assumption 2.1.*

Remark 3.4. A prudent look at the proof of Proposition 3.4 reveals that the following property (which is satisfied by the powered exponential and rational quadratic families) has the crucial role.

$$\frac{\partial K}{\partial \nu}\Big|_{\{r=\mathbf{0}_d,\, \theta\}} = 0, \quad \forall\, \theta \in \Theta. \tag{3.9}$$

For the Matern class, not only $\frac{\partial K}{\partial \nu}$ not satisfy (3.9), it does not have a tractable algebraic form. We believe that (A4) holds true for the geometric anisotropic Matern family with unknown ν, even though it is beyond the reach of our current proof technique.

3.2. Minimax optimality and asymptotic normality

Now, we further investigate the asymptotic statistical properties of ACS's algorithm. Near minimax optimality and asymptotic normality are respectively presented in Theorems 3.3 and 3.4.

Theorem 3.3. *Suppose that Assumptions 2.1 and 3.1 hold for* $\mathcal{D}_n$, Ω *and* K. *Then there exist* $n_0 \in \mathbb{N}$ *and a bounded scalar* $C > 0$ *such that*

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

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

Theorem 3.3 reveals that the established bounds in Theorems 3.1 and 3.2 are sharp up to order $\sqrt{\ln n}$. This means that for the perturbed regular lattice sampling scheme, no algorithm can achieve a significantly better rate than the one considered in this paper.

Theorem 3.4. *Suppose that* $\mathcal{D}_n$ *is a perturbed lattice introduced in Assumption 2.1. Furthermore,* (A1), (A3) *with* $q = 3$ *and* (A4) *are fulfilled by* Ω *and* R. *There is a positive definite matrix* $\Sigma \in \mathbb{R}^{(m+1)\times(m+1)}$ *with bounded operator norm such that*

$$\sqrt{n}\left(\hat{\eta}_n - \eta_0\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}_{m+1}, \Sigma\right). \tag{3.10}$$

The exact formulation of Σ has been omitted in this section due to its complicated algebraic form. We refer the reader to the proof of Theorem 3.4 in Section 6 for further details. It is worthwhile to mention that the entries of Σ heavily depend on the configuration of points in $\mathcal{D}_n$, which is a major disparity between fixed and increasing domain asymptotics. Comparing Theorems 3.2 and 3.4, here we impose a slightly stronger differentiability condition (polynomially decaying of the third derivative) for establishing asymptotic normality. This condition has been formerly introduced in [4] and holds for the geometrically anisotropic covariances in Remark 3.1.

4. Simulation studies

The relatively large-scaled numerical studies in this section give a fairly comprehensive appraisal of the statistical and computational performance of the optimization problem (2.3). Despite the popularity of *R* language, running the iterative programs such as loops in R is much slower than that of C++ (around 250 times slower according to some studies [3]). Taking advantage of the *Rcpp* package and hybrid programming techniques in R can considerably expedite the execution time (up to 50 times in our simulation studies). In order to get the maximum speed, the *open MP* application programming interface has been used to exploit the multi-threaded programming technology. All the numerical experiments in this section have

been executed on 12 processors, except for the second simulation study ($n = 10^6$) which has been implemented on 60 cores.

Generating high dimensional samples from a Gaussian process on an irregularly spaced grid is the foremost challenge that we confronted in our synthetic data simulations. Applying the traditional method based upon the Cholesky decomposition of the covariance matrix is almost infeasible in the case that $n \approx 10^5$ or larger. Hence, we use the considerably faster spectral method (pp. 203–205, [9]) for generating stationary Gaussian processes. For completeness, this algorithm will be concisely presented here. Strictly speaking, the objective is to simulate a real valued zero mean stationary Gaussian process $\mathfrak{G}$ in $\mathbb{R}^d$ with the covariance function $\phi_0 K(\cdot, \theta_0)$ over a δ-perturbed lattice $\mathcal{D}_n = \{s_1, \ldots, s_n\}$. For the purpose of generating a realization of $\mathfrak{G}$ on a perturbed grid, without loss of generality we can assume that the samples are all of unit variance, i.e., $\phi_0 = 1$. We also assume that $\mathfrak{G}$ is geometric anisotropic. Recalling from Eq. (3.2), there is a symmetric positive definite matrix $B_0 \in \mathbb{R}^{d\times d}$ which represents the symmetric square root of A_0, such that $K(r, \theta_0) = K\left(\|B_0 r\|_{\ell_2}\right)$. Throughout this section $d = 2$ and K is either the Matern or rational quadratic covariance function which have been previously introduced in Remark 3.1.

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

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

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

$$\mathfrak{G}(s) = \sqrt{\frac{2}{p}} \sum_{k=1}^{p} \cos(\langle \omega_k, s\rangle + \xi_k), \quad \forall\, s \in \mathbb{R}^d. \tag{4.1}$$

It is known that $\mathfrak{G}$ is a geometric anisotropic process with $\operatorname{cov}\left\{\mathfrak{G}(s), \mathfrak{G}(s')\right\} = K\left(\left\|B_0(s - s')\right\|_{\ell_2}\right)$ for any pair $s, s' \in \mathbb{R}^d$ (p. 204, [9]), converging in distribution to a Gaussian process as p tends to infinity. Next, we explain how to generate the random variables $\{\omega_k\}_{k=1}^p$. The following fact which can be proved using the integration by substitution plays a principal role in our algorithm.

Remark 4.1. Let $\omega' \in \mathbb{R}^d$ be a draw from the following density function

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

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

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

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

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

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

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

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

in which u is a uniform random variable in [0, 1]. One can find a closed form expression for $\mathfrak{r}$ in terms of u for the rational quadratic covariance function, in the case that $(\tau + 1/2) \in \mathbb{N}$ (recall τ from Remark 3.1). In this case, $\Phi(\cdot)$ has a form of the Matern covariance function (3.4) (with different constants) due to the duality principle of the Fourier transform.

Throughout this section, $\mathfrak{G}$ is assumed to be a zero mean Gaussian process in $\mathbb{R}^2$, whose covariance function is a member of either Matern or rational quadratic families with a known fractal index. In the first experiment, $\mathfrak{G}$ is an isotropic spatial process. In other words, we set $A_0 = \theta_0^{-2} I_2$ in Eq. (3.2). θ_0 is a strictly positive scalar known as the *range parameter*. Furthermore, $\mathcal{D}_n$ is a randomly generated δ-perturbed lattice of size 320^2, i.e., $n = 102\,400 \approx 10^5$. The approximated realizations of $\mathfrak{G}$ are generated using (4.1) with $p = 1.5 \times 10^5$. To investigate the role of spatial irregularity in the

Table 1
Estimation of $\eta_0 = (\sigma_0, \theta_0)$ for the isotropic Matern and rational quadratic covariance functions, where $\mathcal{D}_n$ is a perturbed lattice of size 320^2 with associated $\delta \in \{0.1, 0.3\}$.

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

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

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

computational and statistical performance of ACS's algorithm, we vary δ in the set $\{0.1, 0.3\}$. The range parameter and the standard deviation, which is represented by $\sigma_0 = \sqrt{\phi_0}$, are respectively estimated solving the optimization problem (3.7) and closed form formula (3.6). The range parameter space is chosen as $\Theta = [0.1, 15]$. The single variable constrained optimization problem (3.7) is solved using the *optimize* function in *R*, which exploits a combination of golden section search and successive parabolic interpolation. We stop the iteration of the solver when the relative change in the objective is below 10^{-3}. Table 1 displays the summary of our first simulation study.

The required CPU times for the numerical experiments in Table 1 are approximately 30 and 60 min for the rational quadratic and Matern kernels, respectively. However evaluating the full MLE for such a large sample size is intractable. As is apparent from Table 1, the estimated parameters, $\hat{\eta}$, are in a close neighborhood of η_0. Moreover, estimating σ_0 has a significantly higher precision than that of the range parameters, since as the distant samples in $\mathcal{D}_n$ carry negligible information about θ_0. Lastly, the condition number of the covariance matrix increases with the value of range parameter, leading to a higher estimation error $\left\|\eta_0 - \hat{\eta}\right\|_{\ell_2}$ for larger θ_0. In the second simulation study which has the same setup as the first experiment, $\mathcal{D}_n$ is an irregular grid of size $1000^2 = 10^6$. We also set $p = 5 \times 10^5$ in (4.1). Table 2 encapsulates the results of this experiment. The evaluation of $\hat{\eta}$ for this very high dimensional numerical study takes 8 h on 60 cores with 4 GB RAM.

In the next set of experiments featuring Gaussian processes with isotropic covariance functions, we set $\mathcal{D}_n$ to be a two dimensional perturbed lattice of size 100^2. For such a scenario, $\hat{\eta}$ can be estimated in a few minutes. Thus, we simulated $T = 100$ independent realizations and η_0 is estimated using the same procedure as previous studies, for each realization. The mean and root mean squared error (RMSE) have been computed across T experiments. Table 3 displays the average and RMSE for the standard deviation and range parameters for different values of η, δ and covariance kernels. The instances for which $\hat{\eta}$ hits the boundary points of Θ have been excluded in the procedure of calculating the mean and RMSE of the estimates.

Looking at the left and right panels of Table 3 reveals that the RMSE of $\hat{\sigma}$ and $\hat{\theta}$ are slightly larger for the higher values of δ. Moreover, as we have discussed before, the RMSE and the average norm of $\left(\hat{\eta} - \eta_0\right)$ directly depend on the range parameter. It is immediately clear that there is a considerable reduction in RMSE for rational quadratic kernel comparing to the Matern class. This observation may look surprising for the reader, as the condition number of the covariance matrix associated to Matern kernel is smaller than that of rational quadratic due to its faster decay. Thus at the first glance, it may not corroborate our developed theory regarding the consistency of ACS's estimation algorithm for covariance matrices with bounded condition number. However, obtaining a highly accurate estimate of the dependence parameters is more difficult for a rapidly decaying covariance function, as more samples are almost independent. In the extreme case θ_0 is unidentifiable if $K(\cdot, \theta_0)$ is a compactly supported covariance function whose support size is strictly less than $(1 - 2\delta)$ (in this case all the samples are independent).

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

Table 3
Mean and RMSE of $\hat{\eta}$ over 100 independent experiments for the isotropic Matern and rational quadratic covariance functions, where $\mathcal{D}_n$ is a perturbed lattice of size 100^2 with associated $\delta \in \{0.1, 0.3\}$.

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

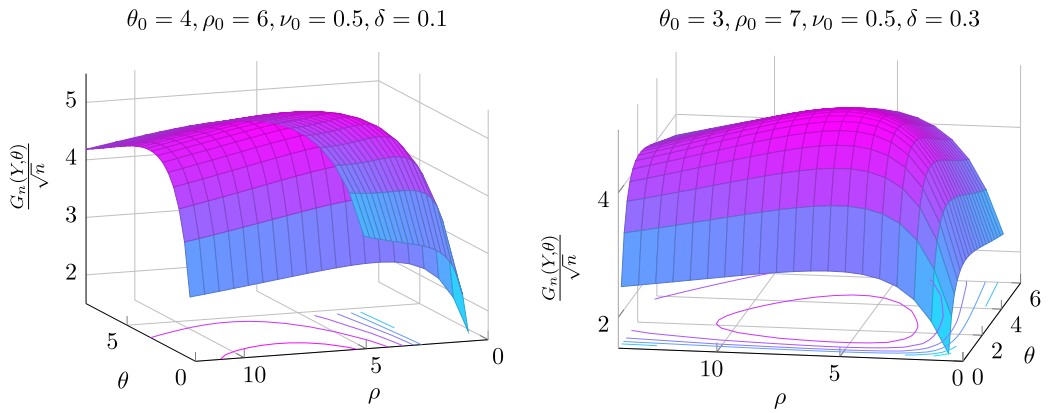

Fig. 2. The above figures exhibit $n^{-1/2}G_n(Y, \theta)$ for geometric anisotropic Matern covariance function with $\nu_0 = 0.5$. The spatial samples form a two dimensional randomly δ-perturbed regular lattice of size $N = 100$. In the left panel, $(\theta_0, \rho_0) = (4, 6)$ and $\delta = 0.1$. In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

$\mathfrak{G}$ has a geometric anisotropic covariance kernel (Matern or rational quadratic) with

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

The parameter space $\Theta = \{(\theta_0, \rho_0) \in \Theta\}$ is a two dimensional box chosen as $[0.1, 15]^2$. The range parameters θ_0 and ρ_0 are estimated by solving the optimization problem (3.7). Fig. 2 exhibits the objective function in (3.7) and its contours for a Gaussian process with geometric anisotropic Matern covariance with $\nu_0 = 0.5$ which has been sampled on a δ-perturbed regular grid. In the left and right panels of Fig. 2, the other parameters are respectively given by $(\theta_0, \rho_0, \delta) = (4, 6, 0.1)$ and $(\theta_0, \rho_0, \delta) = (3, 7, 0.3)$. It can be seen from Fig. 2 that G_n is a unimodal function with only one stationary point. We perform the maximization using the *optim* function in *R* and with *L-BFGS-B* algorithm [6] (box constrained BFGS). The maximum iteration and the initial guess of the *L-BFGS-B* method are respectively 100 and (2, 2). The components of the gradient function are computed using the finite difference approximation with the step size of 10^{-3}. We cease the iteration when the relative change in the objective function is below 10^{-5}. The computation procedure of the average and RMSE of $\hat{\eta} = (\hat{\theta}, \hat{\rho}, \hat{\sigma})$ is exactly the same as the former simulation study. Table 4 presents a summary of the final results of this simulation study. It is clear from Table 4 that the RMSE for Matern covariance is significantly larger in comparison to rational quadratic class. Furthermore, increasing ν_0 for each covariance kernel leads to a slightly larger RMSE.

5. Discussion

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

Table 4
Mean and RMSE of $\hat{\eta}$ over 100 independent experiments for the geometric anisotropic Matern and rational quadratic covariance functions, where $\mathcal{D}_n$ is a perturbed lattice of size 100^2 with associated $\delta \in \{0.1, 0.3\}$.

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

(a) As indicated in Remark 2.1, the inversion-free loss function can be viewed as an approximate minorizer for the likelihood loss (in the expected value sense) in the increasing domain setting. However, more work needs to be done to know how to precisely characterize a rich class of minorizers for the likelihood loss. We believe that responding to this question will provide a flexible class of fast and consistent estimators of covariance parameters.
(b) Spatial statisticians usually cast doubt upon the benefits of increasing domain asymptotics as spatial processes are unlikely to be stationary over a large domain. However the developed theory in this manuscript has persuaded us that inversion free algorithm can consistently estimate microergodic covariance parameters, when applied on the preconditioned samples of a Gaussian random field in a fixed domain. In this regard, a modified version of the estimator studied in this paper can provide a powerful tool for interpolation of Gaussian processes.

6. Proofs of the main results

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

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

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

$$M_{\theta_1,\theta_2} := \frac{K_n^{1/2}(\theta_1) K_n(\theta_2) K_n^{1/2}(\theta_1)}{\|K_n(\theta_2)\|^2_{\ell_2}}, \qquad H_{\theta_1,\theta_2} := \|K_n(\theta_2)\|_{\ell_2} M_{\theta_1,\theta_2}. \tag{6.1}$$

Proof of Theorem 3.1. Our proof has two major parts. In the first part, the consistency of $\hat{\theta}_n$ (correlation function's parameters) will be substantiated. In the second part, we establish the consistency of $\hat{\phi}_n$, which has a closed form solution in terms of Y, correlation function and $\hat{\theta}_n$, by conditioning on the consistency of $\hat{\theta}_n$. To this end, various types of concentration inequalities regarding the quadratic forms (and their supremum over a bounded space) of Gaussian processes are of the indispensable importance. Such results will be presented in the Appendix.

Let Z be a standard Gaussian vector in $\mathbb{R}^n$. As Y and $\sqrt{\phi_0} K_n^{1/2}(\theta_0) Z$ have the same distribution, (2.3) can be equivalently written by

$$\left(\hat{\phi}_n, \hat{\theta}_n\right) = \underset{(\phi,\theta)\in \mathbb{1}\times\Theta}{\operatorname{argmax}} \left\{\phi\phi_0 Z^\top K_n^{1/2}(\theta_0) K_n(\theta) K_n^{1/2}(\theta_0) Z - \frac{\phi^2}{2} \|K_n(\theta)\|^2_{\ell_2}\right\}. \tag{6.2}$$

The objective function in (6.2) is quadratic in terms of ϕ and its maximizer $\hat{\phi}_n$ has a simple closed form. Replacing $\hat{\phi}_n$ to (6.2) gives a surrogate form for $\hat{\theta}_n$. Omitting the cumbersome algebra, the final results are given by

$$\frac{\hat{\phi}_n}{\phi_0} = Z^\top M_{\theta_0,\hat{\theta}_n} Z, \qquad \hat{\theta}_n = \underset{\theta\in\Theta}{\operatorname{argmax}} Z^\top H_{\theta_0,\theta} Z. \tag{6.3}$$

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

Claim 1. *Choose $\xi > (m-1)$ and let $r_n := C_{\min}\sqrt{\frac{\ln n}{n}}$ for some bounded positive scalar $C_{\min}$ (see Lemma A.3 for its exact form). Then,*

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

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

$$\left|\mathcal{N}_{r'_n}\left(\Theta^c_{\theta_0}(r_n)\right)\right| \le \left|\mathcal{N}_{r'_n}(\Theta)\right| \lesssim \left(\frac{R_0}{r'_n}\right)^m = o\left(n^m\right). \tag{6.4}$$

So, there is $n_1 \in \mathbb{N}$ such that $\mathcal{N}_{r'_n}\left(\Theta^c_{\theta_0}(r_n)\right) \le n^m$ for any $n \ge n_1$. It follows from (6.3) that

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

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

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

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

$$\begin{aligned}
\text{RHS} &= \Pr(A_n \cap \pi_n) + \tau_n \Pr\left(A_n \mid \pi_n^c\right) \le \tau_n + \Pr(A_n \cap \pi_n) \\
&\overset{(A)}{\le} \tau_n + \Pr\left(Z^\top H_{\theta_0,\theta_0} Z \le \sup_{\theta \in \mathcal{N}_{r'_n}\left(\Theta^c_{\theta_0}(r_n)\right)} Z^\top H_{\theta_0,\theta} Z + C_2\sqrt{\frac{\ln n}{n}}\right) \\
&\overset{(B)}{\le} \tau_n + n^m \sup_{\theta \in \mathcal{N}_{r'_n}\left(\Theta^c_{\theta_0}(r_n)\right)} \Pr\left(Z^\top H_{\theta_0,\theta_0} Z \le Z^\top H_{\theta_0,\theta} Z + C_2\sqrt{\frac{\ln n}{n}}\right).
\end{aligned} \tag{6.5}$$

The way that π_n and A_n have been defined trivially justifies inequality (A). Furthermore (B) is inferred from the combination of (6.4) and the union bound. Applying Lemma A.3 guarantees the following result for any $\theta \in \mathcal{N}_{r'_n}\left(\Theta^c_{\theta_0}(r_n)\right)$.

$$\Pr\left(Z^\top H_{\theta_0,\theta_0} Z \le Z^\top H_{\theta_0,\theta} Z + C_2\sqrt{\frac{\ln n}{n}}\right) \le n^{-(1+\xi)}. \tag{6.6}$$

Finally, substituting (6.6) into (6.5) yields

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

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

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

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

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

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

Since T_1 tends to zero (via Claim 1), it suffices to show that T_2 is a diminishing sequence as $n \to \infty$. Let $\beta_{\hat{\theta}_n}$ be the closest point in $\mathcal{N}_{r'_n}\left(\Theta_{\theta_0}(r_n)\right)$ (which is a deterministic set) to $\hat{\theta}_n$. Based upon identity (6.3), we have

$$\left|\frac{\hat{\phi}_n}{\phi_0} - 1\right| = \left|Z^\top M_{\theta_0,\hat{\theta}_n} Z - 1\right|. \tag{6.7}$$

Given that $\hat{\theta}_n$ belongs to $\Theta_{\theta_0}(r_n)$, applying the triangle inequality on the right hand side of the identity (6.7) yields that, almost surely

$$\begin{aligned}\left|\frac{\hat{\phi}_n}{\phi_0} - 1\right| &\le \left|Z^\top M_{\theta_0,\hat{\theta}_n} Z - Z^\top M_{\theta_0,\beta_{\hat{\theta}_n}} Z\right| + \left|Z^\top M_{\theta_0,\beta_{\hat{\theta}_n}} Z - \operatorname{tr}\left(M_{\theta_0,\beta_{\hat{\theta}_n}}\right)\right| + \left|\operatorname{tr}\left(M_{\theta_0,\beta_{\hat{\theta}_n}}\right) - 1\right| \\ &\overset{(D)}{\le} \sup_{\theta\in\Theta_{\theta_0}(r_n)} \left\{ \left|\operatorname{tr}\left(M_{\theta_0,\beta_\theta} - 1\right)\right| + \left|Z^\top M_{\theta_0,\beta_\theta} Z - \operatorname{tr}\left(M_{\theta_0,\beta_\theta}\right)\right| + \left|Z^\top M_{\theta_0,\theta} Z - Z^\top M_{\theta_0,\beta_\theta} Z\right| \right\} \\ &:= \sup_{\theta\in\Theta_{\theta_0}(r_n)} \left\{ T_{21}(\theta) + T_{22}(\theta) + T_{23}(\theta) \right\}. \end{aligned} \tag{6.8}$$

Replacing the random quantities $\hat{\theta}_n$ and $\beta_{\hat{\theta}_n}$ with the nonrandom parameters θ and β_θ is the key advantage of (D). Now we control the terms T_{21}, T_{22} and T_{23} from above, uniformly over $\Theta_{\theta_0}(r_n)$. Lemma A.2 guarantees the existence of a scalar C_0, for which

$$\lim_{n\to\infty} \Pr\left(\sup_{\theta\in\Theta_{\theta_0}(r_n)} T_{23}(\theta) \le C_0 r'_n \right) \to 1. \tag{6.9}$$

C_0 depends on $\Lambda_{\max}$ and $\mathfrak{D}_{\max}$. See Lemma A.2 for its exact formulation. For large enough n, we have

$$C_0 r'_n = \mathcal{O}\left(n^{-1}\sqrt{\ln n}\right) < \frac{1}{2}\mathfrak{D}_{\max} C_{\min}\sqrt{\frac{\ln n}{n}}. \tag{6.10}$$

Now we control $T_{21}(\theta)$ uniformly from above using Proposition A.1. The goal is to show that

$$\sup_{\theta\in\Theta_{\theta_0}(r_n)} T_{21}(\theta) < \frac{3}{2}\mathfrak{D}_{\max} C_{\min}\sqrt{\frac{\ln n}{n}}. \tag{6.11}$$

Applying the Cauchy–Schwarz inequality shows that (recalling M_{θ_1,θ_2} from (6.1))

$$\begin{aligned}\sup_{\theta\in\Theta_{\theta_0}(r_n)} T_{21}(\theta) &= \sup_{\theta\in\Theta_{\theta_0}(r_n)} \left| \frac{\langle K_n(\beta_\theta), K_n(\theta_0)\rangle}{\|K_n(\beta_\theta)\|_{\ell_2}^2} - 1 \right| = \sup_{\theta\in\Theta_{\theta_0}(r_n)} \left| \frac{\langle K_n(\beta_\theta), K_n(\beta_\theta) - K_n(\theta_0)\rangle}{\|K_n(\beta_\theta)\|_{\ell_2}^2} \right| \\ &\le \sup_{\theta\in\Theta_{\theta_0}(r_n)} \frac{\|K_n(\beta_\theta) - K_n(\theta_0)\|_{\ell_2}}{\|K_n(\beta_\theta)\|_{\ell_2}} \le \sup_{\theta\in\Theta_{\theta_0}(r_n)} \frac{\|K_n(\beta_\theta) - K_n(\theta_0)\|_{\ell_2}}{\sqrt{n}}. \end{aligned} \tag{6.12}$$

Furthermore, using the part (b) of the Proposition A.1, we get

$$\begin{aligned}\sup_{\theta\in\Theta_{\theta_0}(r_n)} \frac{\|K_n(\beta_\theta) - K_n(\theta_0)\|_{\ell_2}}{\sqrt{n}} &\le \mathfrak{D}_{\max} \sup_{\theta\in\Theta_{\theta_0}(r_n)} \|\theta_0 - \beta_\theta\|_{\ell_2} \\ &\le \mathfrak{D}_{\max} \sup_{\theta\in\Theta_{\theta_0}(r_n)} \left(\|\theta - \beta_\theta\|_{\ell_2} + \|\theta_0 - \theta\|_{\ell_2}\right) \\ &\le \mathfrak{D}_{\max}\left(r_n + r'_n\right) < \frac{3}{2}\mathfrak{D}_{\max} C_{\min}\sqrt{\frac{\ln n}{n}}. \end{aligned} \tag{6.13}$$

Note that the last inequality holds for large enough n. So (6.10) follows from replacing (6.13) into (6.12). In the sequel we achieve a uniform upper bound on T_{22}. For brevity define $u_n := \Lambda_{\max}\sqrt{mn^{-1}\ln n}$ and select a large enough universal constant C'. Recall that β_θ, by its definition, is an element of the finite set $\mathcal{N}_{r'_n}\left(\Theta_{\theta_0}(r_n)\right)$. Thus,

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

The same trick as (6.4) leads to $\left|\mathcal{N}_{r_n'}\left(\Theta_{\theta_0}(r_n)\right)\right| = o(n^m)$. So, it is adequate to show that

$$\Pr\left(T_{22}(\theta) \geq C' u_n\right) \leq n^{-m}, \quad \forall\, \theta \in \Theta_{\theta_0}(r_n). \tag{6.14}$$

We employ Hanson–Wright inequality (Theorem 1.1, [17]) for obtaining a probabilistic upper bound on $T_{22}(\theta)$ (for a fixed θ).

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

For simplifying the upper bound on $T_{22}(\theta)$, we control the operator and Frobenius norms of $M_{\theta_0,\beta_\theta}$ from above. The following inequalities can be easily justified by Proposition A.1.

$$\left\|M_{\theta_0,\beta_\theta}\right\|_{\ell_2} \leq \Lambda_{\max} n^{-1/2}, \qquad \left\|M_{\theta_0,\beta_\theta}\right\|_{2\to 2} \leq \Lambda_{\max}^2 n^{-1}. \tag{6.15}$$

Replacing (6.15) into Hanson–Wright inequality justifies (6.14). Hence

$$\Pr\left(\sup_{\theta\in\Theta_{\theta_0}(r_n)} T_{22}(\theta) \geq C' u_n\right) \leq \left|\mathcal{N}_{r_n'}\left(\Theta_{\theta_0}(r_n)\right)\right| n^{-m} \to 0, \quad \text{as } n \to \infty. \tag{6.16}$$

Substituting inequalities (6.9), (6.10), (6.11) and (6.16) into (6.8) concludes the proof by confirming that T_2 goes to zero as $n \to \infty$. □

Combining Claims 1 and 2 ends the proof. □

Proof of Theorem 3.2. Let $r_n = C\sqrt{n^{-1}\ln n}$ for some strictly positive C whose exact form will be given shortly. Let $(\hat{\phi}_n, \hat{\theta}_n)$ be any stationary point of the optimization problem (2.3). Due to the space constraint, we just show that $\hat{\theta}_n \in \Theta_{\theta_0}(r_n)$. The same technique as Claim 2 in the proof of Theorem 3.1 attains the convergence rate of $\hat{\phi}_n$. Notice that $\hat{\theta}_n$ is a stationary point of the optimization problem (6.3). We show that with a high probability there is no $\theta \in \Theta^c_{\theta_0}(r_n)$ for which the gradient of the objective function in (6.3) be exactly zero. In order to substantiate our claim, we prove that the absolute value of the inner product of the gradient and a fixed non-zero vector is uniformly greater than zero on $\Theta^c_{\theta_0}(r_n)$.

Let $Z \in \mathbb{R}^n$ be a standard Gaussian vector and θ_l, $l = 1, \ldots, m$ be the lth component of θ. We first give a closed form for the gradient function in (6.3), which will be denoted by $[G_l(\theta)]_{l=1}^m$.

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

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

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

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

$$\Pr\left(\inf_{\theta\in\Theta^c_{\theta_0}(r_n)} |W(\theta)| > 0\right) \to 1, \quad \text{as } n \to \infty. \tag{6.17}$$

(6.17) follows from the succeeding claims, whose proofs have been thoroughly presented in [13].

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

$$\lim_{n\to\infty} \Pr\left(\sup_{\theta\in\Theta}\left|W(\theta) - \sum_{j=1}^m \lambda_j \mathrm{tr}\left(P^j_{\theta_0,\theta}\right)\right| \geq C_0\sqrt{n\ln n}\right) = 0. \tag{6.18}$$

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

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

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

Proof of Theorem 3.3. We follow the standard techniques presented in the Chapter 2 of [21] for bounding the minimax risk from below. For any $\theta \in \Theta$, $\mathbb{P}_\theta$ stands for the associated distribution to a zero mean Gaussian vector with the covariance function $K_n(\theta)$. Finding far enough (with respect to the Euclidean distance) pair of the correlation parameters, $\theta_i \in \Theta$, $i = 1, 2$, for which $D\left(\mathbb{P}_{\theta_1} \parallel \mathbb{P}_{\theta_2}\right) < \alpha = 1/2$ lies at the heart of our proof. The two bounded positive scalars $\mathfrak{D}_{\max}$ and $\Lambda_{\min}$ appearing here are defined in Proposition A.1. To ease notation let $r_n := \frac{\Lambda_{\min}}{8\mathfrak{D}_{\max}\sqrt{n}}$ (choose n large enough so that $4r_n \leq \operatorname{diam}(\Theta)$). Choose $\theta_1, \theta_2 \in \Theta$ with $2r_n \leq \|\theta_2 - \theta_1\|_{\ell_2} \leq 4r_n$. The connectedness of Θ guarantees the existence of such pair of points. We first use Proposition A.4 to show that $D\left(\mathbb{P}_{\theta_1} \parallel \mathbb{P}_{\theta_2}\right) \leq \alpha$.

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

As $\alpha \geq D\left(\mathbb{P}_{\theta_1} \parallel \mathbb{P}_{\theta_2}\right)$, Theorem 2.2 of [21] yields

$$\inf_{\hat{\theta}_n}\sup_{\theta_0\in\Theta}\Pr\left(\left\|\hat{\theta}_n - \theta_0\right\|_{\ell_2} \geq r_n\right) \geq \left(\frac{1}{4}e^{-\alpha}\right) \vee \left(\frac{1-\sqrt{\frac{\alpha}{2}}}{2}\right) = \frac{1}{4}. \tag{6.19}$$

The desired statement follows from the fact that $\left\|\hat{\eta}_n - \eta_0\right\|_{\ell_2} \geq \left\|\hat{\theta}_n - \theta_0\right\|_{\ell_2}$. □

Proof of Theorem 3.4. Let $g : \Omega \mapsto \mathbb{R}^{m+1}$ represent the gradient of the objective function in (2.2) with respect to η. Here g_j, $j = 11, \ldots, (m+1)$ stands for the jth entry of g. Analyzing the exact second order Taylor expansion of $\sqrt{n}g(\eta)$ around η_0 at $\eta = \hat{\eta}_n$ is the integral part of the proof. We argue that the second order term of the expansion, which involves the third order derivatives of the covariance function, converges to zero in probability as n grows to infinity. We also show that the first term (zeroth order term) in the expansion converges weakly to a Gaussian random variable. These two ingredients lead to the desirable result by showing the asymptotic normality of the first order term in the expansion, which directly depends on $\sqrt{n}\left(\hat{\eta}_n - \eta_0\right)$.

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

$$ng_j(\eta) = Y^\top R^j_n(\eta)\, Y - \left\langle R^j_n(\eta), R_n(\eta)\right\rangle. \tag{6.20}$$

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

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

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

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

in which z_j lies on the line segment between η_0 and $\hat{\eta}_n$. Proposition $D.10$ of [4] guarantees the statement (3.10) for $\Sigma = \Sigma_2^{-1}\Sigma_1\Sigma_2^{-1}$ and hence concludes the proof, if

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

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

(b) $\sqrt{n}g(\eta_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}_{m+1}, \Sigma_1)$ for some positive semidefinite matrix $\Sigma_1 \in \mathbb{R}^{(m+1)\times(m+1)}$.
(c) $\Pr\left(\lim_{n\to\infty}\sqrt{n}\Delta_j\left(\eta_0, \hat{\eta}_n\right) = 0\right) = 1$, for any $j \in \{1, \ldots, m+1\}$.

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

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

$$V^n_{lk} = \frac{1}{n}\left\langle R^l_n(\eta_0), R^k_n(\eta_0)\right\rangle + \frac{1}{n}\left\{Y^\top R^{lk}_n(\eta_0)\, Y - \left\langle R^{lk}_n(\eta_0), R_n(\eta_0)\right\rangle\right\}.$$

Now define

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

Notice that the entries of Σ_2 are well defined and bounded due to part (a) of Proposition A.2. The proof will be presented in two steps: First we show that Σ_2 is a positive definite matrix. Second, we prove that $\Phi_{lk}^n := \left\{Y^\top R_n^{lk}(\eta_0) Y - \left\langle R_n^{lk}(\eta_0), R_n(\eta_0)\right\rangle\right\}/n$ converges to zero in probability. To substantiate the first claim, consider an arbitrary $\lambda \in \mathcal{S}^m$. It is required to show that $\lambda^\top \Sigma_2 \lambda > c$, for some constant $c > 0$. The condition (A4.a) guarantees the existence of positive scalars M_2, r_2 such that for any $s \in \mathcal{D}_n$,

$$\max_{s'\in\mathcal{D}_n(s,r_2)} \left|\sum_{l=1}^{m+1} \lambda_l \frac{\partial}{\partial\eta_l} R\left(s-s',\eta\right)\Big|_{\eta=\eta_0}\right| \geq M_2. \tag{6.21}$$

Thus,

$$\begin{aligned}\lambda^\top\Sigma_2\lambda &= \lim_{n\to\infty}\sum_{l,k=1}^{m+1}\lambda_l\lambda_k\frac{\left\langle R_n^l(\eta_0), R_n^k(\eta_0)\right\rangle}{n} = \lim_{n\to\infty}\frac{1}{n}\left\|\sum_{l=1}^{m+1}\lambda_l R_n^l(\eta_0)\right\|_{\ell_2}^2\\ &= \lim_{n\to\infty}\frac{1}{n}\left\|\left[\sum_{l=1}^{m+1}\lambda_l\frac{\partial}{\partial\eta_l}R\left(s'-s,\eta\right)\Big|_{\eta=\eta_0}\right]_{s,s'\in\mathcal{D}_n}\right\|_{\ell_2}^2 \overset{(A)}{\geq} M_2^2.\end{aligned} \tag{6.22}$$

Here, inequality (A) is an easy consequence of (6.21). The rest of the proof is devoted to prove the second claim. Choose an arbitrary strictly positive ϵ. As Φ_{lk}^n is a zero mean random variable, using Chebyshev's inequality we get

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

in which (B) and (C) are implied by Propositions A.1 and A.2, respectively (see Appendix for more details about the constants).

Validating condition (b). Define $Q_n^j := n^{-1/2} R_n^{1/2}(\eta_0) R_n^j(\eta_0) R_n^{1/2}(\eta_0)$ for $1 \leq j \leq m+1$, and write $\Psi_{n,\lambda} := \lambda_1 Q_n^1 + \cdots + \lambda_{m+1} Q_n^{m+1}$ for any $\lambda = (\lambda_1, \ldots, \lambda_{m+1}) \in \mathcal{S}^m$. Rewriting (6.20) yields

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

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

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

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

$$2\left\|\Psi_{n,\lambda}\right\|_{\ell_2}^2 = 2\lambda^\top\left[\left\langle Q_n^k, Q_n^l\right\rangle\right]_{l,k=1}^{m+1}\lambda. \tag{6.23}$$

Thus, it is enough to show that the matrix defined by $\Sigma_1 := \lim_{n\to\infty} 2\left[\left\langle Q_n^k, Q_n^l\right\rangle\right]_{l,k=1}^{m+1}$ is well defined (with bounded entries) and positive semi-definite. Well definiteness of Σ_1 can be proved using the same techniques as the proof of Claim 1 and by employing Propositions A.1 and A.2. The positive semi-definite property of Σ_1 is an immediate consequence of (6.23). We conclude the proof by showing that

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

According to Lemma A.4, this statement is valid if the following claim holds.

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

Proof of Claim 1. We show that C depends on m, $\Lambda_{\max}$, $\Lambda_{\min}$ and $\Lambda'_{\max}$ (except m, all the constants are introduced in Propositions A.1 and A.2). Obviously $\Psi_{n,\lambda}$ can be rewritten as,

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

Applying Proposition A.1, we get

$$\frac{\left\|\Psi_{n,\lambda}\right\|_{2\to 2}}{\left\|\Psi_{n,\lambda}\right\|_{\ell_2}}\leq\frac{\left\|R_n\left(\eta_0\right)\right\|_{2\to 2}\left\|\sum_{j=1}^{m+1}\lambda_j R_n^j\left(\eta_0\right)\right\|_{2\to 2}}{\left\|\sum_{j=1}^{m+1}\lambda_j R_n^j\left(\eta_0\right)\right\|_{\ell_2}\lambda_{\min}\left\{R_n\left(\eta_0\right)\right\}}\leq\frac{\Lambda_{\max}}{\Lambda_{\min}}\left\|\sum_{j=1}^{m+1}\lambda_j R_n^j\left(\eta_0\right)\right\|_{2\to 2}\left\|\sum_{j=1}^{m+1}\lambda_j R_n^j\left(\eta_0\right)\right\|_{\ell_2}^{-1}. \tag{6.24}$$

Furthermore, using Proposition A.2 leads to

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

From (6.22) we know that there is a scalar $C_0\in(0,\infty)$ for which $\left\|\sum_{j=1}^{m+1}\lambda_j R_n^j\left(\eta_0\right)\right\|_{\ell_2}\geq C_0\sqrt{n}$. Replacing the last two inequalities into (6.24) ends the proof. □

In conclusion we state that the condition (c) can be proved using akin techniques as the proof of proposition 3.2 of [4]. We omit the technical details due to the space constraints. □

Appendix. Auxiliary results

In this section, we only prove the auxiliary results of independent interest and the proof of other technical lemmas and propositions will be omitted. The detailed proof of all results appearing in this section can be found in [13].

The first result examines the perturbation of some norms of $K_n\left(\theta\right)$ with respect to θ. It appears to be of great importance for proving Theorems 3.1–3.4 in Section 6.

Proposition A.1. *Suppose that* $\mathcal{D}_n$ *admits Assumption 2.1. Moreover, Assumption 3.1 holds for* Θ *and* K. *Construct* $n\times n$ *correlation matrix* $K_n\left(\theta\right):=\left[K\left(s-s',\theta\right)\right]_{s,s'\in\mathcal{D}_n}$ *for any* $\theta\in\Theta$.

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

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

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

$$\left\|K_n\left(\theta_2\right)-K_n\left(\theta_1\right)\right\|_{2\to 2}\leq\mathfrak{D}_{\max}\left\|\theta_2-\theta_1\right\|_{\ell_2}, \tag{A.1}$$

$$\frac{1}{\sqrt{n}}\left\|K_n\left(\theta_2\right)-K_n\left(\theta_1\right)\right\|_{\ell_2}\leq\mathfrak{D}_{\max}\left\|\theta_2-\theta_1\right\|_{\ell_2}, \tag{A.2}$$

and

$$\frac{1}{\sqrt{n}}\left\|K_n\left(\theta_2\right)-K_n\left(\theta_1\right)\right\|_{\ell_2}\geq\mathfrak{D}_{\min}\left\|\theta_2-\theta_1\right\|_{\ell_2}, \tag{A.3}$$

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

For ease of reference, we present the following result as a standalone proposition. Its proof is akin to that of Proposition A.1 and will be skipped to avoid redundancy.

Proposition A.2. *Suppose that* $\mathcal{D}_n$ *admits Assumption 2.1. Moreover,* Θ *and* K *satisfy Assumption 3.2. Construct the matrix* $\partial K_n\left(\theta\right)/\partial\theta_j:=\left[\partial K\left(s-s',\theta\right)/\partial\theta_j\right]_{s,s'\in\mathcal{D}_n}$ *for* $\theta\in\Theta$ *and* $j=1,\ldots,m$.

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

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

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

$$\left\| \frac{\partial}{\partial\theta_j} K_n(\theta_2) - \frac{\partial}{\partial\theta_j} K_n(\theta_1) \right\|_{2\to 2} \leq \mathfrak{D}'_{\max} \|\theta_2 - \theta_1\|_{\ell_2}. \tag{A.4}$$

The bounded positive scalars $\mathfrak{D}_{\max}$, $\mathfrak{D}_{\min}$ and $\Lambda_{\max}$, which have been introduced in Proposition A.1, become frequently apparent in the subsequent results in this section. It is also proper to remind the reader that $\mathcal{N}_\epsilon(\mathcal{A})$ stands for the ϵ-net of $\mathcal{A}$ with respect to the Euclidean distance. Furthermore, the matrices H_{θ_1,θ_2} and M_{θ_1,θ_2} have been formerly defined in (6.1) for any pair of the correlation function parameters θ_1, θ_2. The succeeding two Lemmas A.1 and A.2, which come in hand in the proof of Theorem 3.1, establish a probabilistic upper bound on the maximum of a quadratic Gaussian expression over an uncountable set Θ in terms of its largest value over one of its finite subset.

Lemma A.1. *Let* $Z \in \mathbb{R}^n$ *be a standard Gaussian vector and suppose that* $\mathcal{D}_n$ *satisfies Assumption 2.1. Furthermore, assume that* Θ *and* K *admit Assumption 3.1. For any vanishing positive sequence* $\{r_n\}_{n\in\mathbb{N}}$*, any non-empty* $\bar{\Theta} \subset \Theta$ *and each* $\theta_0 \in \Theta$*,*

$$\lim_{n\to\infty} \Pr\left\{ \left(\sup_{\theta\in\bar{\Theta}} Z^\top H_{\theta_0,\theta} Z - \sup_{\theta\in\mathcal{N}_{r_n}(\bar{\Theta})} Z^\top H_{\theta_0,\theta} Z \right) \geq C r_n \sqrt{n} \right\} = 0, \tag{A.5}$$

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

Lemma A.2. *Let* $Z \in \mathbb{R}^n$ *be a standard Gaussian vector. Suppose that Assumption 2.1 and Assumption 3.1 hold for* $\mathcal{D}_n$, Θ *and* K*. For any strictly positive vanishing sequence* $\{r_n\}_{n=1}^\infty$*, any non-empty* $\bar{\Theta} \subset \Theta$ *and arbitrary* $\theta_0 \in \Theta$*,*

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

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

Now we state a lemma which plays a crucial role in the proof of Theorem 3.1.

Lemma A.3. *Let* $Z \in \mathbb{R}^n$ *be a standard Gaussian vector and let* $C, \xi > 0$*. Suppose that* Θ *and* K *satisfy Assumption 3.1. Select* $\theta_1, \theta_2 \in \Theta$ *such that*

$$\|\theta_2 - \theta_1\|_{\ell_2} \geq C_{\min} \sqrt{\frac{\ln n}{n}}, \tag{A.6}$$

in which $C_{\min} := 4\mathfrak{D}_{\min}^{-1} \Lambda^2_{\max} \sqrt{C'(1+\xi)}$ *(recall* $\mathfrak{D}_{\min}$ *and* $\Lambda_{\max}$*, from Proposition A.1), for some appropriately chosen universal constant* $C' > 0$*. There exists* $n_0 = \mathcal{O}(1)$ *(depending on* C, ξ, K, $\mathcal{D}_n$ *and* Θ*) such that for any* $n \geq n_0$

$$p := \Pr\left\{ Z^\top \left(H_{\theta_2,\theta_2} - H_{\theta_2,\theta_1} \right) Z \leq C \sqrt{\frac{\ln n}{n}} \right\} \leq n^{-(1+\xi)}. \tag{A.7}$$

Refer to the identity (6.1) for the definition of H_{θ_2,θ_1}.

The next proposition rigorously expresses the uniform concentration of the Euclidean squared norm of Gaussian vectors with the covariance matrix $K_n(\theta)$, $\theta \in \Theta$ around their mean. We employ such inequality for proving Theorem 3.2.

Proposition A.3. *Let* $\Theta \subset \mathbb{R}^m$ *be a bounded set. Consider the class of* n *by* n *matrices* $\{\Pi_n(\theta)\}_{\theta\in\Theta}$ *parametrized by* $\theta \in \Theta$*. Suppose that the following conditions hold*

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

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

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

$$\|\Pi_n(\theta_2) - \Pi_n(\theta_1)\|_{2\to 2} \leq C \|\theta_2 - \theta_1\|_{\ell_2}, \quad \forall\, \theta_1, \theta_2 \in \Theta. \tag{A.8}$$

(c)

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

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

$$\lim_{n\to\infty} \Pr\left(\sup_{\theta\in\Theta} \left|Z^\top \Pi_n(\theta) Z - \operatorname{tr}\{\Pi_n(\theta)\}\right| \geq C'\sqrt{n \ln n}\right) = 0. \tag{A.9}$$

Proof. Let $r_n = C^{-1}\sqrt{\ln n/n}$ in which C has been defined in (A.8) and let $\mathcal{N}_{r_n}(\Theta)$ denote the r_n-covering set of Θ. As before, for any θ let β_θ represents the closest element of $\mathcal{N}_{r_n}(\Theta)$ to θ. Observe that,

$$\begin{aligned}
\text{RHS} &:= \left|Z^\top \Pi_n(\theta) Z - \operatorname{tr}\{\Pi_n(\theta)\} - Z^\top \Pi_n(\beta_\theta) Z + \operatorname{tr}\{\Pi_n(\beta_\theta)\}\right| \\
&= \left|\left\langle \Pi_n(\theta) - \Pi_n(\beta_\theta), ZZ^\top + I_n\right\rangle\right| \leq \|\Pi_n(\theta) - \Pi_n(\beta_\theta)\|_{2\to 2} \left\|ZZ^\top + I_n\right\|_{s_1} \\
&\overset{(A)}{\leq} C\,\|\theta - \beta_\theta\|_{\ell_2} \left\|ZZ^\top + I_n\right\|_{s_1} \leq C r_n \left\|ZZ^\top + I_n\right\|_{s_1} = \sqrt{\frac{\ln n}{n}}\left(n + \|Z\|_{\ell_2}\right),
\end{aligned}$$

in which $\|\cdot\|_{s_1}$ stands for the nuclear norm (absolute sum of eigenvalues). Note that the obtained upper bound does not depend on θ (uniform upper bound). Moreover, based upon Hanson–Wright inequality there is $c > 0$ for which $\left(n + \|Z\|_{\ell_2}\right) \leq 3n$ with probability at least $1 - \exp(-cn)$. Thus, RHS $\geq 3\sqrt{n \ln n}$ with probability at most $\exp(-cn)$. Hence, as $n \to \infty$ we get

$$\Pr\left(\sup_{\theta\in\Theta} \left|Z^\top \Pi_n(\theta) Z - \operatorname{tr}\{\Pi_n(\theta)\}\right| \geq \sup_{\theta\in\mathcal{N}_{r_n}(\Theta)} \left|Z^\top \Pi_n(\theta) Z - \operatorname{tr}\{\Pi_n(\theta)\}\right| + 3\sqrt{n \ln n}\right) \to 0. \tag{A.10}$$

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

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

for some constant $C_0 > 0$ depending on m. Notice that $\sup_{\theta\in\Theta} \|\Pi_n(\theta)\|_{\ell_2} \leq M\sqrt{n}$ according to condition (a). Moreover, as we argued in (6.14), $\left|\mathcal{N}_{r_n}(\Theta)\right| = o(n^m)$. Thus,

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

Replacing the last inequality into (A.10) concludes the proof. □

The following result gives an upper bound on the *Kullback–Leibler* divergence of two zero mean multivariate Gaussian distributions respectively associated with the two covariance matrices $K_n(\theta_i)$, $i = 1, 2$. Such upper bound is extremely useful for establishing Theorem 3.3.

Proposition A.4. *Choose $\theta_1, \theta_2 \in \Theta$ in such a way that $\|\theta_2 - \theta_1\|_{\ell_2} \leq \Lambda_{\min}/(2\mathfrak{D}_{\max})$. Let P_i, $i = 1, 2$, denote the associated probability distribution to a zero mean Gaussian vector with the covariance matrix $K_n(\theta_i) \in \mathbb{R}^{n\times n}$, $i = 1, 2$. Then,*

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

Proof. For any symmetric matrix $A \in \mathbb{R}^{n\times n}$, let $\lambda_i(A)$, $i = 1, \ldots, n$, denote its ith eigenvalue in decreasing order. The *von Neumann's trace inequality* [15] yields

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

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

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

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

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

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

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

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

Now we demonstrate the asymptotic normality of the normalized quadratic Gaussian forms. We exploit this fact in the proof of Theorem 3.4.

Lemma A.4. *For $n \in \mathbb{N}$, let $Z_n \in \mathbb{R}^n$ be a standard Gaussian vector and let $A_n \in \mathbb{R}^{n \times n}$. Then,*

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

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

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

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

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

The last result of this section studies the shrinkage behavior of the partial derivatives of Matern covariance function with respect to its fractal index. It turns out to be useful for corroborating the part (a) of Remark 3.1.

Lemma A.5. *Let $K_\nu : \mathbb{R}^d \mapsto \mathbb{R}$ be a geometric anisotropic (Recall from Definition 3.1) Matern correlation function given by*

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

in which $\mathfrak{K}_\nu$ stands for the modified Bessel function of the second kind and A satisfies the condition (3.3). Then for any $\beta \in \mathbb{N}$ and $m \in \mathbb{N}$, there is a bounded constant $C_{\beta,A}$ such that

$$\left| \frac{\partial^m}{\partial \nu^m} K_\nu \left(r \right) \right| \leq \frac{C_{\beta,A}}{1 + \left\| r \right\|_{\ell_2}^{2\beta}}, \quad \forall r = \left(r_1, \ldots, r_d \right) \in \mathbb{R}^d. \tag{A.11}$$

References

[1] M. Abramowitz, I.A. Stegun, Handbook of Mathematical Functions, in: Applied Mathematics Series, vol. 55, 1966, p. 62.
[2] M. Anitescu, J. Chen, M.L. Stein, An inversion-free estimating equation approach for Gaussian process models, (submitted for publication). 2014.
[3] S.B. Aruoba, J. Fernndez-Villaverde, A comparison of programming languages in economics, No. w20263, National Bureau of Economic Research, 2014.
[4] F. Bachoc, Asymptotic analysis of the role of spatial sampling for covariance parameter estimation of Gaussian processes, J. Multivariate Anal. 125 (2014) 1–35.
[5] J. Bernardo, J. Berger, A. Dawid, A. Smith, Regression and Classification using Gaussian Process Priors, Bayesian Stat. 6 (1998) 475.
[6] R.H. Byrd, P. Lu, J. Nocedal, C. Zhu, A limited memory algorithm for bound constrained optimization, SIAM J. Sci. Comput. 16 (5) (1995) 1190–1208.
[7] J. Chen, On the use of discrete Laplace operator for preconditioning kernel matrices, SIAM J. Sci. Comput. 35 (2) (2013) A577–A602.
[8] W. Chu, Z. Ghahramani, Gaussian processes for ordinal regression, J. Mach. Learn. Res. (2005) 1019–1041.
[9] N. Cressie, Statistics for Spatial Data, John Wiley & Sons, 2015.
[10] J. Du, H. Zhang, V.S. Mandrekar, Fixed-domain asymptotic properties of tapered maximum likelihood estimators, Ann. Statist. 37 (6A) (2009) 3330–3361.
[11] A.E. Gelfand, P. Diggle, P. Guttorp, M. Fuentes, Handbook of Spatial Statistics, CRC press, 2010.
[12] C.G. Kaufman, M.J. Schervish, D.W. Nychka, Covariance tapering for likelihood-based estimation in large spatial data sets, J. Amer. Statist. Assoc. 103 (484) (2008) 1545–1555.
[13] H. Keshavarz, C. Scott, X.L. Nguyen, On the consistency of inversion-free parameter estimation for Gaussian random fields, 2016. arXiv preprint arXiv:1601.03822.
[14] K.V. Mardia, R.J. Marshall, Maximum likelihood estimation of models for residual covariance in spatial regression, Biometrika 71 (1) (1984) 135–146.
[15] L. Mirsky, A trace inequality of John von Neumann, Monatsh. Math. 79 (4) (1975) 303–306.
[16] D.P. O'Leary, The block conjugate gradient algorithm and related methods, Linear Algebra Appl. 29 (1980) 293–322.
[17] M. Rudelson, R. Vershynin, Hanson-Wright inequality and sub-Gaussian concentration, 2013. arXiv preprint arXiv:1306.2872.
[18] J. Sacks, W.J. Welch, T.J. Mitchell, H.P. Wynn, Design and analysis of computer experiments, Statist. Sci. (1989) 409–423.
[19] M.L. Stein, Interpolation of Spatial Data: Some Theory for Kriging, Springer Science & Business Media, 2012.
[20] M.L. Stein, J. Chen, M. Anitescu, Difference filter preconditioning for large covariance matrices, SIAM J. Matrix Anal. Appl. 33 (1) (2012) 52–72.
[21] A.B. Tsybakov, Introduction to Nonparametric Estimation, in: Springer Series in Statistics, 2009.
[22] D. Wang, W.L. Loh, On fixed-domain asymptotics and covariance tapering in Gaussian random field models, Electron. J. Stat. 5 (2011) 238–269.
[23] H. Wendland, Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree, Adv. Comput. Math. 4 (1) (1995) 389–396.
[24] C.K. Williams, C.E. Rasmussen, Gaussian Processes for Machine Learning, The MIT Press, vol. 2 (3), p. 4, 2006.
[25] Z. Ying, Asymptotic properties of a maximum likelihood estimator with data from a Gaussian process, J. Multivariate Anal. 36 (2) (1991) 280–296.
[26] H. Zhang, Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics, J. Amer. Statist. Assoc. 99 (465) (2004) 250–261.