SIAM/ASA J. UNCERTAINTY QUANTIFICATION
Vol. 5, pp. 138–175

Maximum Likelihood Estimation for a Smooth Gaussian Random Field Model[*]

Wanting Xu[†] and Michael L. Stein[†]

Abstract. Gaussian processes are commonly used for modeling the output of deterministic computer models. We consider the behavior of maximum likelihood estimators (MLEs) of parameters of the commonly used squared exponential covariance function when the computer model has some simple deterministic form. We prove that for regularly spaced observations on the line, the MLE of the scale parameter converges to zero if the computer model is a constant function and diverges to infinity for linear functions. When observing successive derivatives of a pth order monomial at zero, we find the asymptotic orders of the MLE of the scale parameter for all $p \geq 0$. For some commonly used test functions, we compare the MLE with cross validation in a prediction problem and explore the joint estimation of range and scale parameters. The correlation matrix is nearly numerically singular even when the sample size is moderate. To overcome numerical difficulties, we perform exact computation by making use of exact results for the correlation matrix and restricting ourselves to parameter values and test functions that yield rational correlations and function values at the observation locations. We also consider the common approach of including a nugget effect to deal with the numerical difficulties, and explore its consequences on model fitting and prediction.

Key words. computer experiment, cross validation, best linear prediction, squared exponential covariance function

AMS subject classifications. 62M30, 60G15, 97N50

DOI. 10.1137/15M105358X

1. Introduction.

Computer experiments have been used extensively in investigating complex scientific phenomena. The responses of many computer experiments are deterministic, in the sense that rerunning the same code with the same inputs will give identical outputs. Often, each run of the code is computationally expensive, so a common alternative to running the code at all input values of interest is to run the code at some inputs and make cheaper predictions at others. [25] and [26] propose to model the deterministic computer experiment outputs as a realization of a Gaussian random field with covariance

$$\operatorname{Cov}\left(f(x), f(y)\right) = \theta_0 \prod_{u=1}^{d} e^{-\frac{|x_u - y_u|^\gamma}{\theta_u}}, \tag{1.1}$$

where x_u, $y_u \in [0,1]$, $u = 1, \ldots, d$, $\theta_0 > 0$ is the scale parameter, and the $\theta_u > 0$ are range parameters. The use of stochastic models provides a statistical basis for experimental design, parameter estimation, interpolation, and uncertainty calibration.

[*] Received by the editors December 21, 2015; accepted for publication (in revised form) November 1, 2016; published electronically January 26, 2017.

http://www.siam.org/journals/juq/5/M105358.html

Funding: This material is based upon work supported by the U.S. Department of Energy Office of Science, Office of Advanced Scientific Computing Research, Applied Mathematics program under Award DE-SC0011087.

[†] Department of Statistics, University of Chicago, Chicago, IL 60637 (wxu@galton.uchicago.edu, stein@galton.uchicago.edu).

When $\gamma = 2$, the Gaussian process with covariance function (1.1) is infinitely mean square differentiable and thus is an attractive choice when the output surface is known to be smooth [11, 21, 22, 27, 29]. This covariance function is sometimes called "Gaussian" because of its functional form, but we prefer the name "squared exponential" to avoid confusion with a Gaussian process. In fact, smooth test functions composed of elementary functions (e.g., polynomials, trigonometric functions, and exponential functions) are often used as test cases for studying the effectiveness of Gaussian processes in modeling computer experiments. However, little is known about properties of maximum likelihood estimators (MLEs) when observations are generated by these test functions. In this article, we are interested in the asymptotic; properties of the MLE when more and more observations are taken on a fixed domain (fixed domain asymptotics; see [28]) for the Gaussian process when the computer model is some simple deterministic function. We aim to understand the implications of modeling smooth deterministic functions using the squared exponential covariance function.

We consider a mean zero Gaussian random field with covariance function (1.1) and $\gamma = 2$. For $d = 1$, we prove some asymptotic properties of the MLE for the scale parameter θ_0 when the range parameter θ_1 is fixed and the computer experiment response is a pth order monomial $f(x) = x^p$. We consider two situations for the observations. In the first case, observations $\mathbf{z}$ are taken on a regular grid on $[0, 1]$ so that $\mathbf{z} = (f(\frac{1}{n}), f(\frac{2}{n}), \dots, f(1))^T$. In the second case, the observations are successive derivatives of the response function at zero, namely, $\mathbf{z} = (f(0), f^{(1)}(0), \dots, f^{(n-1)}(0))^T$. Automatic differentiation (AD) techniques [14] can be used to obtain derivatives of computer model output and there are certain problems for which higher order derivatives are needed [7, 13, 30]. Therefore considering what happens when one observes successive derivatives at a single location may be of some practical interest.

The rest of the article is organized as follows. Section 2 deals with regularly spaced observations on $[0, 1]$. The key finding is that the asymptotic order of the MLE $\hat{\theta}_0$ is $n^{-1/2}$ when $p = 0$ and at least $n^{1/2}$ when $p = 1$. In particular, $\hat{\theta}_0 \to 0$ when $p = 0$ and $\hat{\theta}_0 \to \infty$ when $p = 1$. Section 3 deals with the case where observations are derivatives at zero. An exact expression for the inverse Cholesky factor for the correlation matrix is obtained. For all $p \geq 0$, we prove that $\lim_{n\to\infty} n^{1/2-p}\hat{\theta}_0$ exists and is positive so that $\hat{\theta}_0 \to 0$ for $p = 0$ and $\hat{\theta}_0 \to \infty$ for all $p \geq 1$. Section 4 demonstrates the theoretical findings in sections 2 and 3, and explores numerically three commonly used two-dimensional test functions in the computer experiments literature. For estimating the scale parameter θ_0 and the two range parameters θ_1 and θ_2, we compare the maximum likelihood method with leave-one-out cross validation in a prediction problem. We find that the likelihood method and cross validation perform differently for different test functions in terms of magnitude and calibration of prediction errors. We also explore the MLE of the range parameter for pth order monomials when treating both scale and range parameters as unknown, and investigate its implications for practical test functions. In the numerical experiments, to deal with numerical singularity of the correlation matrix, we choose parameters such that the correlation matrix and observations are both rational, and do symbolic computation with *Mathematica* [31] to obtain exact results. Since a common approach to overcome the near singularity is to include a nugget effect, we investigate the effect of adding a nugget on the likelihood and prediction. We found that the likelihood generally decreases substantially with even a very small nugget, but prediction error can sometimes decrease a bit at first as the nugget size increases. All proofs of the theoretical results are presented in the appendix.

2. Regularly spaced observations. In this section, we consider the observations as outputs of the model function $f(x) = x^p$ regularly spaced on $[0,1]$. Fixing the range parameter θ_1, we prove that $\hat{\theta}_0 \to 0$ when the model function is constant and $\hat{\theta}_0 \to \infty$ when it is linear. Though some intermediate steps apply to all $p \geq 0$, we are only able to derive the asymptotic order and a lower bound on the asymptotic order for $p = 0$ and $p = 1$, respectively.

The observations are outputs of the model function $f(x) = x^p$ taken on a regular grid on $[0,1]$ so that $\boldsymbol{z} = ((\frac{1}{n})^p, (\frac{2}{n})^p, \ldots, 1)^T$. The covariance matrix can be written as

$$\Sigma(\theta_0, \theta_1, n) = \theta_0 R(\theta_1, n), \tag{2.1}$$

where the (i,j)th element of $R(\theta_1, n)$ is

$$R(\theta_1, n)_{ij} = w^{(i-j)^2}, \qquad w = e^{-1/(\theta_1 n^2)}. \tag{2.2}$$

[19] gives the exact form of the inverse of the Cholesky factor for $R(\theta_1, n)$. Letting $R(\theta_1, n) = LL^T$, where L is the lower triangular Cholesky factor with positive diagonal elements, then

$$(L^{-1})_{ij} = \begin{cases} \frac{(-w)^{i-j} \left[{i-1 \atop j-1} \right]_{w^2}}{\prod_{k=1}^{i-1} (1-w^{2k})^{1/2}}, & i \geq j, \\ 0, & i < j, \end{cases} \tag{2.3}$$

where $\left[{k \atop m} \right]_q$ is the q-binomial coefficient defined by

$$\begin{bmatrix} k \\ m \end{bmatrix}_q = \frac{(1-q^{k-m+1})(1-q^{k-m+2})\ldots(1-q^k)}{(1-q)(1-q^2)\ldots(1-q^m)}$$

if $0 \leq m \leq k$ and 0 otherwise.

The log-likelihood function of θ_0 is

$$2l(\theta_0) = -n\log 2\pi - n\log\theta_0 - \log|R(\theta_1, n)| - \frac{1}{\theta_0}\boldsymbol{z}^T R(\theta_1, n)^{-1}\boldsymbol{z}$$

and the MLE of θ_0 is

$$\hat{\theta}_0 = \frac{1}{n}\boldsymbol{z}^T R(\theta_1, n)^{-1}\boldsymbol{z}.$$

With the form of L^{-1} in (2.3), the exact form of $\hat{\theta}_0$ can be written as

$$\hat{\theta}_0 = \frac{1}{n}\sum_{i=1}^{n} \frac{\left(\sum_{j=1}^{i} (-w)^{i-j} \left[{i-1 \atop j-1} \right]_{w^2} j^p\right)^2}{n^{2p}\prod_{k=1}^{i-1}(1-w^{2k})}.$$

For convenience we make the following notation for the rest of this article:

$$a_{ip}(w) := \frac{\left(\sum_{j=1}^{i} (-w)^{i-j} \left[{i-1 \atop j-1} \right]_{w^2} j^p\right)^2}{n^{2p}\prod_{k=1}^{i-1}(1-w^{2k})}, \tag{2.4}$$

where $p \geq 0$ and $i \geq 1$. Note that $\hat{\theta}_0 = \frac{1}{n}\sum_{i=1}^{n} a_{ip}(w)$ for $w = e^{-1/(\theta_1 n^2)}$.

By considering the limit of the summand of $\hat{\theta}_0$, we obtain the following proposition.

Proposition 2.1. *Denote*

$$l_{ip} := \lim_{n\to\infty} a_{ip}(w) = \lim_{n\to\infty} \frac{\left(\sum_{j=1}^{i}(-w)^{i-j}\begin{bmatrix} i-1 \\ j-1 \end{bmatrix}_{w^2} j^p\right)^2}{n^{2p}\prod_{k=1}^{i-1}(1-w^{2k})}$$

then

$$l_{ip} = \begin{cases} \frac{(i-1)!\theta_1^p}{2^{i-1}\left(\frac{i-p-1}{2}!\right)^2}, & i-p \text{ odd}, \\ 0, & i-p \text{ even}, \end{cases} \tag{2.5}$$

where $w = e^{-1/(\theta_1 n^2)}$, $p \geq 0$, *and* $i > p$.

Proof. See section A.1. ■

Lemma 2.2. $\frac{1}{n}\sum_{i=p+1}^{n} l_{ip} \sim \frac{n^{p-\frac{1}{2}}\theta_1^p}{\sqrt{2\pi}2^p(p+\frac{1}{2})}$ *as* $n \to \infty$.

Proof. See section A.2. ■

The previous results deal with the general case where $p \geq 0$. The following results concentrate on $p = 0$ and $p = 1$. We now derive the asymptotic order for $\hat{\theta}_0$ when $p = 0$ and a lower bound on the asymptotic order for $\hat{\theta}_0$ when $p = 1$.

Theorem 2.3. *If* $p = 0$, $\hat{\theta}_0 \sim \sqrt{\frac{2}{\pi}}\frac{1}{\sqrt{n}}$ *as* $n \to \infty$.

Proof. See section A.3. ■

Theorem 2.4. *If* $p = 1$,

$$\liminf_{n\to\infty} \frac{\hat{\theta}_0}{\sqrt{n}} \geq \frac{\theta_1}{3\sqrt{2\pi}}.$$

Proof. See section A.4. ■

The difficulty of the proof lies in the fact that the dimension and elements of the correlation matrix $R(\theta_1, n)$ change with n. In particular, we cannot simply apply an elementwise limit theorem to $\hat{\theta}_0 = \boldsymbol{z}^T R(\theta_1, n)^{-1}\boldsymbol{z}/n$. Proposition 2.1 proves the limits of $a_{ip}(w)$ for fixed i as $n \to \infty$ and Lemma 2.2 proves the asymptotic order of the average of those limits. However, to derive the asymptotic order of $\hat{\theta}_0 = \frac{1}{n}\sum_{i=1}^{n} a_{ip}(w)$ we would need to have some results on the uniform convergence of $a_{ip}(w)$ for $1 \leq i \leq n$, which we have been unable to obtain.

We now state a conjecture about the asymptotic order of $\hat{\theta}_0$ for general $p \geq 0$. The case for $p = 0$ is proved in Theorem 2.3 with $C(0) = \sqrt{2/\pi}$, and Theorem 2.4 is a weaker version of the conjecture when $p = 1$.

Conjecture 2.5. *For all* $p \geq 0$, $\lim_{n\to\infty} n^{1/2-p}\hat{\theta}_0 = C(p)$, *where* $C(p) = \theta_1^p/\sqrt{2\pi}2^p(p+1/2)$.

It might also be of interest to consider functions that are continuous but have some form of singularities. These functions are not smooth and hence not the main focus of our work. However, we present one example here to illustrate what can happen.

Proposition 2.6. *If*

$$f(x) = \begin{cases} 0, & x \leq 1/2, \\ g(x - 1/2), & x > 1/2, \end{cases}$$

for some continuous function $g(x)$ satisfying $g(0) = 0$ and $c := \lim_{x\to 0} \frac{g(x)}{x^p} > 0$ for some $p \geq 1$, then

$$\liminf_{n\to\infty} \frac{\hat{\theta}_0}{n} > 0$$

as $n \to \infty$. In particular, $\hat{\theta}_0 \to \infty$ as $n \to \infty$.

Proof. See section A.5. ∎

We recognize that fixing the range parameter as we have done here is rather artificial, but the mathematical difficulties of analyzing even this problem are formidable and we believe that the resulting asymptotic theory is interesting and informative despite its limitations. As shown in [28, pp. 120-121], two nonidentical squared exponential covariance functions for a Gaussian process on a finite interval correspond to orthogonal measures, suggesting that, unlike the case for Matérn covariance functions [32], it might be possible to estimate both the scale and range parameters consistently based on fixed domain asymptotics if in fact the Gaussian process model is correct (see, for example, [1]). In the present setting when the process is just a simple deterministic function, it is not at all clear what should happen, so we investigate the properties of joint estimates of scale and range parameters through numerical experiments in section 4.

We do not have an intuitive explanation for the quantitative aspects of our asymptotic results, even for $p = 0$. Comparing this to two settings for which asymptotic calculations can be easily done provides us with a clue to the qualitative behavior of $\hat{\theta}_0$ as p increases. If $\Sigma = \theta_0 I_n$, where I_n is the $n \times n$ identity matrix, and f is a continuous function on $[0, 1]$, then $\hat{\theta}_0 \sim \int_0^1 f(x)^2\, dx$ as $n \to \infty$, so $\hat{\theta}_0$ tends to a nonzero constant for any nontrivial f. For the exponential covariance function ($\gamma = 1$ and $d = 1$ in (1.1)), if f has a bounded second derivative on $[0, 1]$ then

$$\hat{\theta}_0 \sim \frac{1}{n} f(0)^2 + \frac{1}{2\theta_1 n} \int_0^1 \left\{ f(x) + \theta_1 f'(x) \right\}^2 dx \tag{2.6}$$

as $n \to \infty$ (see section A.6), so that $n\hat{\theta}_0$ tends to a positive finite constant as $n \to \infty$ when $f(x) = x^p$ for any nonnegative integer p. These results are in stark contrast to what we have proven and conjectured here for the squared exponential covariance function, that $n^{1/2-p}\hat{\theta}_0$ tends to a positive, finite constant. Thus, there must be something about the squared exponential model that makes us think θ_0, the variance of the process, is large when p is large. A possible intuitive explanation for this result is that if we think the underlying function is very smooth (which is the case when we use the squared exponential model) and we observe that the function just happens to equal x^p at n densely spaced points, then we will conclude that this function must at least very nearly equal x^p over some broad interval, so that the larger p is, the more we think the function varies over this broad interval and the larger we think θ_0 is.

3. Derivatives at zero.

We obtain the asymptotic order of $\hat{\theta}_0$ when the observations are the first $n - 1$ derivatives at 0 for the response function $f(x) = x^p$, $p \geq 0$. Denote $\boldsymbol{z} = (f(0), f'(0), \ldots, f^{(n-1)}(0))^T$, and the covariance matrix of the observations $\boldsymbol{z}$ as $\Sigma_1(\theta_0, \theta_1, n)$.

Now we introduce a notation that is frequently used in the rest of this section.

Definition 3.1. *For a positive integer m, define the double factorial of m as*

$$m!! = \begin{cases} \prod_{k=1}^{m/2}(2k) = m(m-2)\dots 2, & m \text{ even}, \\ \prod_{k=1}^{(m+1)/2}(2k-1) = m(m-2)\dots 1, & m \text{ odd}, \end{cases}$$

and set $0!! = 1$. This double factorial notation is commonly used in combinatorics [12]. Note that $m!!$ is the product of all positive integers no larger than and having the same parity as m, which is different from the successive factorial $(m!)!$.

As before, f is modeled as a stationary Gaussian random field with covariance function $K(u) = \theta_0 e^{-u^2/\theta_1}$. Then the (i,j)th element of $\Sigma_1(\theta_0, \theta_1, n)$ can be computed as

$$\begin{aligned} \Sigma_1(\theta_0, \theta_1, n)_{ij} &= \frac{\partial^{i+j-2}}{\partial x^{i-1}\partial y^{j-1}} K(x-y)\Big|_{x=y=0} \\ &= (-1)^{j-1}\theta_0 \frac{d^{i+j-2}}{du^{i+j-2}} e^{-u^2/\theta_1}\Big|_{u=0} \\ &= \theta_0 \theta_1^{-\frac{i+j-2}{2}} (-1)^{i-1} H_{i+j-2}, \end{aligned}$$

where

$$H_m = \begin{cases} 0, & m \text{ odd}, \\ (-2)^{\frac{m}{2}}(m-1)!!, & m \text{ even}, \end{cases}$$

is mth order Hermite polynomial at 0. The mth order Hermite polynomial is defined as

$$H_m(x) = (-1)^m e^{x^2} \frac{d^m}{dx^m} e^{-x^2}.$$

Defining $R_1(\theta_1, n)$ so that $\Sigma_1(\theta_0, \theta_1, n) = \theta_0 R_1(\theta_1, n)$, the MLE of θ_0 is $\hat{\theta}_0 = \frac{1}{n}\boldsymbol{z}^T R_1(\theta_1, n)^{-1}\boldsymbol{z}$.

The following proposition gives an exact form of the reverse Cholesky factorization [17] of $R_1(\theta_1, n)^{-1}$.

Proposition 3.2. *Let $D(\theta_1, n)$ be the lower triangular matrix with positive diagonal elements such that $R_1(\theta_1, n)^{-1} = D(\theta_1, n)^T D(\theta_1, n)$. Then for all $1 \le i, j \le n$, the (i,j)th element $D(\theta_1, n)_{ij} = d_{ij}$, where*

$$d_{ij} = \begin{cases} \dfrac{\theta_1^{\frac{j-1}{2}}\sqrt{(i-1)!}}{2^{\frac{j-1}{2}}(j-1)!(i-j)} & \text{if} \quad i \ge j, \quad \text{and} \quad i+j \text{ is even}, \\ 0 & \text{otherwise}. \end{cases} \tag{3.1}$$

Proof. See section A.7. ■

Note that d_{ij} depends only on i and j but not n. So the matrices $D(\theta_1, n)$ are nested as n increases. In fact, in this case $R_1(\theta_1, n)$ is nested and, in general, the reverse Cholesky factors of inverses of a sequence of nested matrices are nested (see section A.8). This feature simplifies the proof of the asymptotic order of $\hat{\theta}_0$ for all $p \ge 0$ and is not shared by the setting considered in section 2.

Theorem 3.3. *Suppose* $f(x) = x^p$, $p \geq 0$, *then*

$$\hat{\theta}_0 \sim \frac{n^{p-\frac{1}{2}}\theta_1^p}{\sqrt{2\pi}2^p(p+\frac{1}{2})}$$

as $n \to \infty$.

Proof. See section A.9. ■

AD can be used to obtain derivatives of functions coded as computer programs. AD exploits the fact that function evaluation can be broken down into elementary operations (e.g., addition, multiplication, $\exp(\cdot)$) and applies the chain rule. A comprehensive reference for AD can be found in [14]. Although, in practice, only lower order derivatives are usually found, there are efforts to obtain higher order Taylor coefficients in one direction for certain problems [7, 13, 30].

For smooth functions, estimates based on multiple derivatives at zero should be closely related to estimates based on observing more frequently on a fixed domain, since more observations enable the calculation of higher order finite differences that approximate derivatives as the spacing gets small. More specifically, consider observing at k inputs $\boldsymbol{z}_k = ((\frac{1}{n})^p, \ldots, (\frac{k}{n})^p)^T$ for some $k \leq n$. For this case, MLE $\hat{\theta}_0(k) = \frac{1}{k}\sum_{i=1}^k a_{ip}(w)$, where $a_{ip}(w)$ is defined in (2.4) and $w = e^{-1/(\theta_1 n^2)}$, is the same as the MLE $\hat{\theta}_0'(k)$ when observing the first k finite differences $\boldsymbol{z}_k'$ at zero, since these first k finite differences at zero are a linear transformation of the first k observations, namely, $\boldsymbol{z}_k' = C\boldsymbol{z}_k$ for some $C \in \mathbb{R}^{k\times k}$. It follows that $\hat{\theta}_0'(k) = \frac{1}{n}\boldsymbol{z}_k'^T\left(CRC^T\right)^{-1}\boldsymbol{z}_k' = \frac{1}{n}\boldsymbol{z}_k R^{-1}\boldsymbol{z}_k = \hat{\theta}_0(k)$. Fixing k and letting $n \to \infty$ gives that finite differences converge to derivatives and $\hat{\theta}_0(k) \to \frac{1}{k}\sum_{i=1}^k l_{ip}$. The asymptotic order of the limit $\frac{1}{k}\sum_{i=1}^k l_{ip}$ as $k \to \infty$ is given by Lemma 2.2 and is exactly the same as what is obtained in Theorem 3.3, which is the asymptotic order of the MLE when observations are derivatives at zero. However, this heuristic argument does not directly imply $\hat{\theta}_0$ has the same asymptotics for the two situations of observations. In particular, taking $k \to \infty$ at the same time as $n \to \infty$ is a different and harder problem.

4. Numerical results.

In this section, first we illustrate our theoretical findings in sections 2 and 3 numerically. Then we compare MLEs with leave-one-out cross validation (CV) in a prediction problem for two commonly used test functions. We also show that for the Branin function, the MLE for the range parameter along one coordinate does not appear to exist. The parameters in the numerical experiments are chosen to make both the correlation matrix and the observations rational so that the matrix calculations can be done exactly with symbolic computations. A common approach to overcome the near singularity of the correlation matrix is to include a small nugget effect in the hope of improving conditioning at the same time introducing minimal modification to the matrix. We also investigate the effect of this approach on the likelihood and prediction.

4.1. Asymptotic behavior of the MLE for the scale parameter.

As in sections 2 and 3, we consider the test function as a pth order monomial $f(x) = x^p$ and two situations for the observations. In the first case, the observations $\boldsymbol{z} = (f(\frac{1}{n}), f(\frac{2}{n}), \ldots, f(1))^T$ are taken on a regular grid on $[0,1]$ and, in the second case, the observations $\boldsymbol{z}' = (f(0), f^{(1)}(0), \ldots, f^{(n-1)}(0))^T$

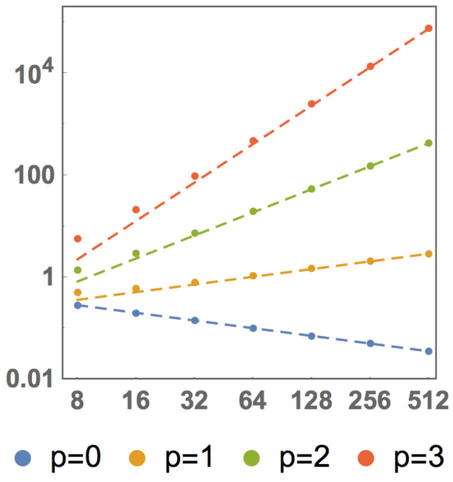

Figure 1. *$\hat{\theta}_0$ when $n = 2^k$, $k = 3, \dots, 9$, for $\boldsymbol{z} = (f(\frac{1}{n}), f(\frac{2}{n}), \dots, f(1))^T$. The slopes of the reference lines are the asymptotic orders $p - 1/2$ when $p = 0, 1, 2, 3$. Both axes are in log scale.*

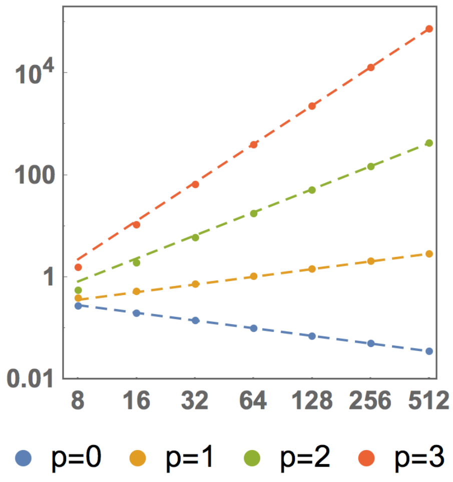

Figure 2. *$\hat{\theta}_0'$ when $n = 2^k$, $k = 3, \dots, 9$, for $\boldsymbol{z}' = (f(0), f^{(1)}(0), \dots, f^{(n-1)}(0))^T$. The slopes of the reference lines are the asymptotic orders $p - 1/2$ when $p = 0, 1, 2, 3$. Both axes are in log scale.*

are the first $n-1$ derivatives of the test function at zero. Denote the MLE of the scale parameter for the two situations of observations as $\hat{\theta}_0$ and $\hat{\theta}_0'$, respectively. We consider $n = 2^k$, $k = 3, \dots, 9$, and the range parameter $\theta_1 \approx 0.95$ is chosen to make the correlation matrix rational for all choices of n so that exact computations can be done. Note that exact computation is needed here to prevent numerical overflow even if the exact form of the Cholesky factor (2.3) is used.

Theorems 2.3 and 2.4 state that

$$\lim_{n\to\infty} \hat{\theta}_0 n^{1/2-p} = \sqrt{2/\pi},\ p = 0, \quad \liminf_{n\to\infty} \hat{\theta}_0 n^{1/2-p} \geq \theta_1/3\sqrt{2\pi},\ p = 1.$$

Conjecture 2.5 in section 2 states that $\hat{\theta}_0 n^{1/2-p}$ converges to some limit $C(p)$ as $n \to \infty$ for all $p \geq 0$. Figure 1 shows $\hat{\theta}_0$ for increasing n when $p = 0, 1, 2, 3$ in log scale. Theorem 2.3 and Conjecture 2.5 imply that $(\log n, \log \hat{\theta}_0)$ will be close to the reference line $y = (p - 1/2)x + \log C(p)$ for n large. The numerical results show clear agreement with the theoretical results in Theorem 2.3 and Conjecture 2.5. When observations are the first $n-1$ derivatives at zero, Theorem 3.3 states that

$$\lim_{n\to\infty} \hat{\theta}_0' n^{1/2-p} = \theta_1^p/\sqrt{2\pi}2^p(p+1/2) = C(p), \qquad p \geq 0.$$

Figure 2 shows $\hat{\theta}_0'$ for increasing n when $p = 0, 1, 2, 3$ in log scale with the same reference lines as those in Figure 1. For all four cases shown here, the agreement between the numerical and asymptotic results is good, even for $n = 8$.

4.2. Comparing the MLE and CV in a prediction problem. We consider the first two functions on a 23×23 regular grid on $[0, 1] \times [0, 1]$ and let $\delta = 1/23$ be the spacing between neighboring points. The observations are taken on a 12×12 regular subgrid, and the remaining 385 points are predictands. An illustration of the setup is shown in Figure 3. Observations are taken at every other location along each dimension to facilitate the use of the inverse Cholesky factor (2.3), so that exact computations can be done with rational correlations. The first test function we experiment with is a mixture of Gaussians [10, 15],

$$f(x_1, x_2) = c_1 e^{-s_1\left((x_1/\delta-\mu_1)^2+(x_2/\delta-\mu_2)^2\right)} + c_2 e^{-s_2\left((x_1/\delta-\widetilde{\mu}_1)^2+(x_2/\delta-\widetilde{\mu}_2)^2\right)}. \tag{4.1}$$

We choose $e^{-s_1} = 399/400$ and $e^{-s_2} = 99/100$ to be rationals and $\mu_1 = \mu_2 = 8$, $\widetilde{\mu}_1 = \widetilde{\mu}_2 = 17$ to be integers so that all the observations are rational. We set $c_1 = 1$ and $c_2 = -1/2$ so that the function consists of a peak and a small dip. The second function we consider is a product of trigonometric and exponential functions [8],

$$f(x_1, x_2) = \cos(c_1 x_1 + c_2 x_2) e^{c_0 x_1 x_2}. \tag{4.2}$$

We choose $c_1 = c_2$ such that $\cos(c_1\delta) = \cos(c_2\delta) = 24/25$. With this choice, $\sin(c_1\delta) = \sin(c_2\delta) = 7/25$ and $\cos(c_1 x_1 + c_2 x_2)$ is rational for all grid points (x_1, x_2) by trigonometric identities. c_0 is chosen to satisfy $e^{c_0\delta^2} = 500/499$ and with this choice $c_0 \approx 1.06$, which approximates $c_0 = 1$ used in [8]. Both the mixture of Gaussians (4.1) and trig-exponential function (4.2) are symmetric about the diagonal. The third test function we consider is the Branin function [6, 23] on a 27×27 regular grid on $[-5, 10] \times [0, 15]$,

$$f(x_1, x_2) = e(x_2 - f x_1^2 + g x_1 - r)^2 + s(1-t)\cos(c_0 x_1) + s. \tag{4.3}$$

The spacing between neighboring points is $\delta_b = 5/9$. We choose parameters as $e = 1$, $f = 5/36$, $g = 5/3$, $r = 6$, $s = 10$, $t = 1/24$, and $\cos(c_0\delta_b) = 4/5$. The observations are taken on the 14×14 regular subgrid. The different setting of the grid for the Branin function is to make the observations rational at all grid points. The three test functions with the aforementioned parameters are shown in Figure 4. The parameters for the three test functions are rounded from the commonly used values to ensure rationality. For example, the recommended parameter values for the Branin function that are different from our choices are $f = 5.1/4\pi^2$, $g = 5/\pi$, $t = 1/8\pi$ [9, 23].

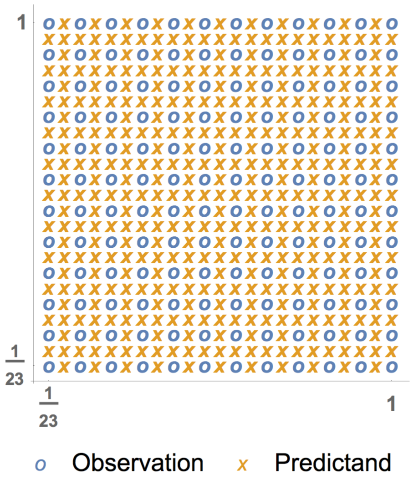

Figure 3. *Locations of observations and predictands.*

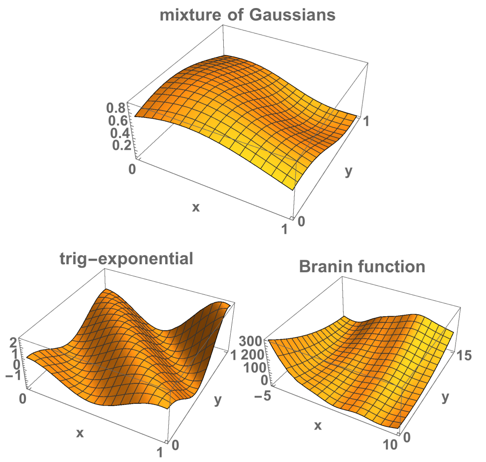

Figure 4. *Surface for test functions* (4.1), (4.2), *and* (4.3).

4.2.1. Exact computation results. We compare the MLE and CV for a mixture of Gaussians (4.1) and for the trig-exponential function (4.2) in predicting at the 385 points not used to fit the model. Denote the observations as $\mathbf{z}$ and the log-likelihood function as $\mathcal{L}(\theta_0, \theta_1, \theta_2)$. The profile log-likelihood function of (θ_1, θ_2) is defined as

$$l_n(\theta_1, \theta_2) = \mathcal{L}\left(\hat{\theta}_0(\theta_1, \theta_2), \theta_1, \theta_2\right),$$

where $\hat{\theta}_0(\theta_1, \theta_2) = \operatorname{argmax}_{\theta_0} \mathcal{L}(\theta_0, \theta_1, \theta_2)$. The profile log-likelihood function satisfies

$$\begin{aligned} 2l_n(\theta_1, \theta_2) = &-n \log 2\pi - n \log \hat{\theta}_0 - \log |R(\theta_1, m) \otimes R(\theta_2, m)| \\ &- \frac{1}{\hat{\theta}_0} \boldsymbol{z}^T \left(R(\theta_1, m) \otimes R(\theta_2, m)\right)^{-1} \boldsymbol{z}, \end{aligned} \tag{4.4}$$

where $n = m^2$, $m = 12$, R is as defined in (2.2), and the MLE for the scale parameter is

$$\hat{\theta}_0(\theta_1, \theta_2) = \frac{1}{n} \boldsymbol{z}^T \left(R(\theta_1, m) \otimes R(\theta_2, m)\right)^{-1} \boldsymbol{z}. \tag{4.5}$$

Leave-one-out CV error is

$$p_n(\theta_1, \theta_2) = \sum_{i=1}^{n} \left(z_i - \hat{z}_{-i}(\theta_1, \theta_2)\right)^2, \tag{4.6}$$

where $\hat{z}_{-i}(\theta_1, \theta_2)$ is the best linear predictor (BLP) of z_i given z_j, $1 \le j \le m$ and $j \ne i$ under the Gaussian process model. The functions l_n and p_n are, respectively, maximized and minimized to obtain estimates of (θ_1, θ_2). Though both l_n and p_n are continuous functions, only certain values of (θ_1, θ_2) correspond to rational correlations. We search over grids consisting of values that allow exact computation to optimize the corresponding functions.

We perform symbolic computations because the correlation matrix is very nearly singular. For example, if we take the set of observations as in Figure 3 with $\theta_1 = \theta_2$ chosen so that the correlation between neighboring points is 0.99, when doing double precision computations, the resulting correlation matrix is found to be not positive definite, nor is its inverse even when using the exact formula for the inverse [19].

Since the correlations between neighboring grid points $w_1 = e^{-\delta^2/\theta_1}$ and $w_2 = e^{-\delta^2/\theta_2}$ are uniquely identifiable with θ_1 and θ_2, we carry out the optimization in terms of (w_1, w_2). Denote by $C(w_1, w_2)$ the function to be optimized, either $\exp(l_n)$ or p_n. Throughout the optimization algorithm, we only consider rational w_1 and w_2 to allow exact computations. Successive grids with shrinking sizes are defined on $[0,1] \times [0,1]$ over which $C(w_1, w_2)$ is optimized. Once an optimizer (w_1^*, w_2^*) is found in the interior of a grid, we compare the log ratio of function values of the current iterate and the previous iterate with a convergence tolerance. Moreover, we define the 3×3 subgrid with (w_1^*, w_2^*) at the center as $S(w_1^*, w_2^*)$, and compare the log ratio of maximal and minimal function values over $S(w_1^*, w_2^*)$ with the convergence tolerance. This comparison is done to help ensure the grid points are taken densely enough in a neighborhood of (w_1^*, w_2^*) so that a local optimum is obtained. We iterate until convergence. Details are provided in Algorithm 1.

Algorithm 1. Grid search.

```
Require: Convergence tolerance $\varepsilon$.
 1: Initialize grid $l_1 = l_2 = 0.01$, $r_1 = r_2 = 0.99$.
 2: Define $m_1 \times m_2$ regular grid $G = \{w_1^1, \dots, w_{m_1}^1\} \times \{w_1^2, \dots, w_{m_2}^2\} \subset [l_1, r_1] \times [l_2, r_2]$.
 3: Obtain optimizer $(w_1^*, w_2^*) \in G$ and corresponding function value $C^{(i)}(w_1^*, w_2^*)$ of $i$th iteration.
 4: ($\textit{Test for convergence}$)
 5: if $\left|\log\left(\frac{C^{(i)}}{C^{(i-1)}}\right)\right| \le \varepsilon$, $\left|\log\left(\frac{\max\{c^{(i)}(w_1,w_2),(w_1,w_2)\in S(w_1^*,w_2^*)\}}{\min\{c^{(i)}(w_1,w_2),(w_1,w_2)\in S(w_1^*,w_2^*)\}}\right)\right| \le \varepsilon$, and $(w_1^*, w_2^*) \in \mathrm{int}(G)$ then
 6:     Return with $(w_1^*, w_2^*)$
 7: else
 8:     ($\textit{Update searching grid}$)
 9:     for $k = 1, 2$ do
10:         if $w_k^* = w_{m_k}^k$ then        ($\textit{optimized at right boundary}$)
11:             $dr_k \leftarrow 1 - w_{m_k}^k$
12:             $dl_k \leftarrow \frac{w_{m_k}^k - w_1^k}{m_k - 1}$
13:         else if $w_k^* = w_1^k$ then        ($\textit{optimized at left boundary}$)
14:             $dr_k \leftarrow \frac{w_{m_k}^k - w_1^k}{m_k - 1}$
15:             $dl_k \leftarrow w_1^k$
16:         else        ($\textit{optimized in interior}$)
17:             $dr_k \leftarrow \frac{w_{m_k}^k - w_1^k}{m_k - 1}$
18:             $dl_k \leftarrow \frac{w_{m_k}^k - w_1^k}{m_k - 1}$
19:         end if
20:         $l_k \leftarrow w_k^* - dl_k$
21:         $r_k \leftarrow w_k^* + dr_k$
22:     end for
23:     $i \leftarrow i + 1$
24:     Repeat steps 2 - 5.
25: end if
```

We take $\varepsilon = 10^{-7}$ in all experiments. The initial grid search in step 1 of Algorithm 1 led to $(0.99, 0.99)$ as the optimizer in step 3 for both functions considered here. To check for multiple optima, we also started the algorithm with different initial grids. In addition, we search over smaller and denser grids inside $(0.01, 0.99) \times (0.01, 0.99)$ and see if an optimum could be obtained in the interior. For neither method did we find evidence for multiple local optima up to symmetry, in the sense that (w_1^*, w_2^*) generates the same CV error as (w_2^*, w_1^*) because of the symmetry in the observations and functions. When choosing the minimizer of the CV error for a grid in step 3 of Algorithm 1, we select (w_1^*, w_2^*) with the convention that $w_1^* \le w_2^*$.

Table 1

Estimates of parameters (first row), standard deviations (sd) of standardized prediction errors (second row), and root mean squared prediction errors (last row) for mixture of Gaussians (4.1)*.*

	MLE	CV
$(\hat{\theta}_0, \hat{\theta}_1, \hat{\theta}_2)$	$(0.024, 0.39, 0.39)$	$(3.29, 0.19, 0.66)$
$\mathrm{sd}\left(\frac{\hat{p}_i - p_i}{\sqrt{\mathrm{EMSE}(\hat{p}_i)}}\right)$	1.56	9.42
$\sqrt{\frac{1}{n_1}\sum_{i=1}^{n_1}(\hat{p}_i - p_i)^2}$	3.90×10^{-8}	3.99×10^{-7}

Table 2

Estimates of parameters (first row), sd of standardized prediction errors (second row), and root mean squared prediction errors (last row) for trig-exponential function (4.2)*.*

	MLE	CV
$(\hat{\theta}_0, \hat{\theta}_1, \hat{\theta}_2)$	$(1611.13, 0.42, 0.42)$	$(1.09 \times 10^{13}, 1.87, 1.87)$
$\mathrm{sd}\left(\frac{\hat{p}_i - p_i}{\sqrt{\mathrm{EMSE}(\hat{p}_i)}}\right)$	0.17	7.12×10^{-3}
$\sqrt{\frac{1}{n_1}\sum_{i=1}^{n_1}(\hat{p}_i - p_i)^2}$	7.94×10^{-7}	4.75×10^{-7}

Denote the true predictand value as $\mathbf{p} \in \mathbb{R}^{n_1}$ with $n_1 = 385$, and covariance matrices as $\Sigma_{zz} = \mathrm{Cov}(\boldsymbol{z}, \boldsymbol{z}^T)$, $\Sigma_{pp} = \mathrm{Cov}(\boldsymbol{p}, \boldsymbol{p}^T)$ and $\Sigma_{zp} = \mathrm{Cov}(\boldsymbol{z}, \boldsymbol{p}^T)$. The predictions $\hat{\mathbf{p}}$ are obtained using the empirical BLP (EBLP) and calibrated with empirical mean squared error (EMSE). EBLP is BLP with θ replaced by its estimate $\hat{\theta}$ and is given by

$$\hat{\boldsymbol{p}} = \Sigma_{zp}^T(\hat{\theta})\Sigma_{zz}^{-1}(\hat{\theta})\boldsymbol{z}, \tag{4.7}$$

and EMSE is the mean squared error (MSE) with θ replaced by its estimate $\hat{\theta}$ and is given by

$$\mathrm{EMSE}(\hat{\boldsymbol{p}}) = \Sigma_{pp}(\hat{\theta}) - \Sigma_{zp}^T(\hat{\theta})\Sigma_{zz}^{-1}(\hat{\theta})\Sigma_{zp}(\hat{\theta}).$$

For CV, we estimate the scale parameter $\widetilde{\theta}_0$ by $\hat{\theta}_0(\widetilde{\theta}_1, \widetilde{\theta}_2)$ using (4.5) as suggested by [20], where $(\widetilde{\theta}_1, \widetilde{\theta}_2)$ are CV estimates for the range parameters. For the two functions, Tables 1 and 2 show the estimates, the standard deviations of standardized prediction errors, and the root MSEs/(RMSEs).

Note that the CV estimates of the two range parameters for the mixture of Gaussians are not equal. Switching the two range parameter estimates (namely, $(0.66, 0.19)$) generates the same CV error, and if we were to employ the convention that $w_1^* \geq w_2^*$ in step 3 of Algorithm 1, we would end up with the CV estimates for this case being $(0.66, 0.19)$. We find the unequal estimated range parameters somewhat surprising, so we did a further careful search along the diagonal $w_1 = w_2$ and could not find any points on this diagonal with smaller CV error than that produced by $(0.19, 0.66)$. We also experimented with estimating the mean of the Gaussian process. For the mixture of Gaussians, the MLEs are $(\hat{\theta}_0, \hat{\theta}_1, \hat{\theta}_2, \hat{\mu}) = (0.018, 0.38, 0.38, 0.20)$ when treating the mean as unknown and is estimated. The standard deviation of the standardized prediction errors and the root mean squared error of predictions are 0.94 and 2.15×10^{-8}, respectively, which are roughly similar to the results in Table 1 when fixing the mean at 0.

We now consider how the MLE and CV estimates of the parameters perform when used for prediction. One of the important features of Gaussian processes is that they provide uncertainty estimates for the predictions, so we will look at both the quality of the point predictions and whether the standardized prediction errors (i.e., the errors divided by their estimated standard deviations) have standard deviation close to 1. If the Gaussian process model under consideration were correct, we should expect the MLE to do better than CV, but since the deterministic functions we consider are obviously not realizations of a Gaussian process, it is unclear which method will perform better. In terms of root mean squared error, MLE is much better (by an order of magnitude; see Table 1) than CV for the mixture of Gaussians and is moderately worse (67% larger; see Table 2) for trig-exponential. For the mixture of Gaussians, the standardized prediction errors under MLE are reasonably well calibrated with a standard deviation of 1.56, whereas for CV, their standard deviation is 9.42, so that CV badly underestimates the variability of the prediction errors. For trig-exponential, both the MLE and CV seriously overestimate the variability of the prediction errors, but CV much more so (Table 2).

Figure 5 shows histograms of the standardized prediction errors for both functions and both estimates. Ideally, we might hope these histograms will approximate a standard normal distribution, but even if the truth were a Gaussian process, we should not be surprised to see something that does not look approximately normal because of possible strong dependencies between prediction errors at different locations. We see that in all four cases, the standardized prediction errors follow a vaguely symmetric distribution about 0. The most noteworthy feature in these plots occurs for the mixture of Gaussians based on CV, which was the case where the estimated range parameters were not equal. In this plot, we see that the standard deviation of the standardized prediction errors is much larger than 1 for predictands in odd columns (sd=16.10) and much smaller than 1 for predictands in even columns (sd=0.15). The tensor product form of the squared exponential covariance function implies that the EBLP of a predictand in an odd column only depends on observations in that column (see section A.10), so that the form of the EBLP is entirely determined by the range parameter along columns. Similarly, the EBLP in an odd row is only a function of observations in that row, whereas EBLPs in an even row and even column depend on all of the observations. Note that the estimated range parameter is large along the columns, so the fitted model thinks observations are much more strongly correlated in this direction. Since the EBLPs within odd columns only depend on within column correlations, it is then perhaps not surprising that the model is overoptimistic about the quality of interpolations in the direction with strong estimated correlations.

Figure 6 shows the (unstandardized) prediction RMSE averaged over each row and column of the large grid. For each function, we focus on the method yielding smaller prediction error. Since the estimates of the two range parameters are equal for both of these cases and the observations and function are symmetric about the diagonal, the prediction errors are also symmetric. Hence, averaging over rows and columns gives identical results, so we only present the root mean squared error averaged over the rows. Figures 7 and 8 show the prediction errors $\hat{p}_i - p_i$ at the corresponding locations for the two functions. Note that the magnitudes of the errors are only comparable for each function; they are not normalized across functions.

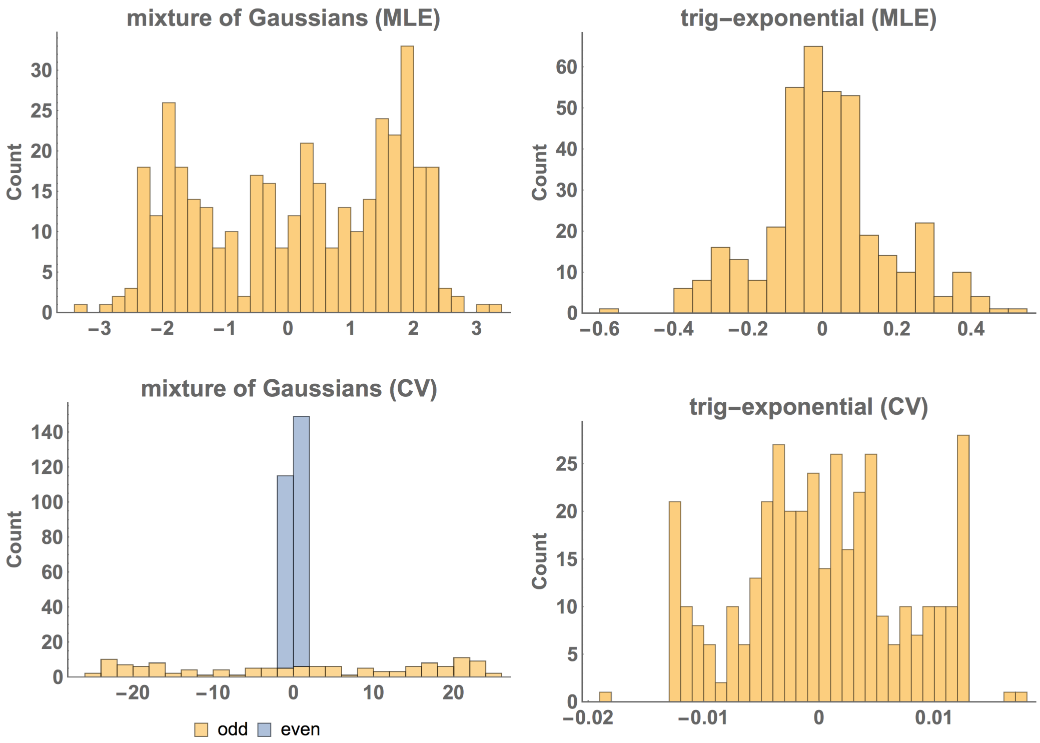

Figure 5. *From left to right: histograms of standardized prediction errors generated by the MLE for the mixture of Gaussians, CV for the mixture of Gaussians, MLE for trig-exponential function, and CV for trig-exponential function. The standardized prediction errors generated by CV for the mixture of Gaussians are grouped into those in odd and even numbered columns. The histograms for the two groups are stacked.*

Figures 6–8 show that, for both test functions, the prediction errors are largest for the predictions on the second and second to last rows and columns (i.e., rows and columns 2 and 22 out of 23). We should generally expect prediction errors to be larger near a border of an observation domain than in the interior, but it is interesting to note that, at least for these functions, the errors tend to be larger, for example, in row 2 than row 1, even when comparing a predictand in row 2 and and an odd column to one in row 1 and an even column, so that the distance from the predictand to the nearest observation equals $\frac{1}{23}$ in both cases.

Next, we show numerically that the MLE for the Branin function (4.3) does not appear to exist. First of all, let us consider the estimation of the range parameter when $f(x) = x^p$ with $p \geq 1$. The profile log-likelihood $L_n(\theta_1)$ satisfying

$$2L_n(\theta_1) = -n\log(\hat{\theta}_0(\theta_1)) - \log|R(\theta_1, n)| - n\log 2\pi - n$$

is maximized to obtain MLE $\hat{\theta}_1$.

Our empirical evidence suggests that for each p and $n^*(p) = 2p+1$, when $n \geq n^*(p)$, $L_n(\theta_1)$ monotonically increases for $\theta_1 \in (0, \infty)$ so that the MLE for θ_1 does not exist. Moreover, $L_n(\theta_1)$ is bounded when $n = n^*(p)$ and increases to ∞ when $n > n^*(p)$. As noted in [18], this

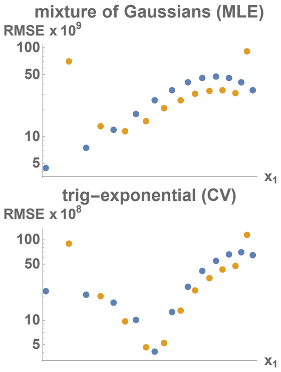

Figure 6. *Root mean squared errors for each column and row of the 23 × 23 grid. Blue: odd; yellow: even.*

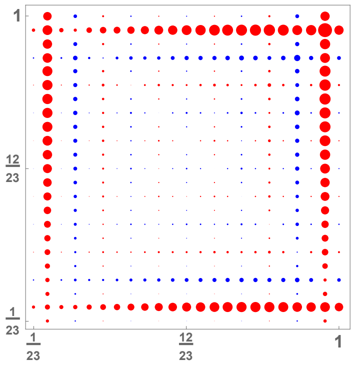

Figure 7. *Prediction errors of MLE for mixture of Gaussians. Red: $\hat{p}_i > p_i$; Blue: $\hat{p}_i \leq p_i$. Area of disk is proportional to $|\hat{p}_i - p_i|$.*

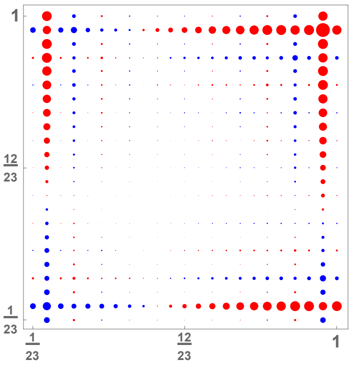

Figure 8. *Prediction errors of CV for trig-exponential function. Red: $\hat{p}_i > p_i$; Blue: $\hat{p}_i \leq p_i$. Area of disk is proportional to $|\hat{p}_i - p_i|$.*

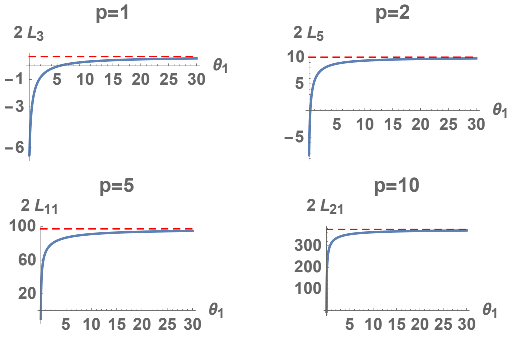

Figure 9. *Profile log-likelihood $2L_{n^*(p)}(\theta_1)$ for $f(x) = x^p$, $n^*(p) = 2p + 1$, and $p = 1, 2, 5, 10$.*

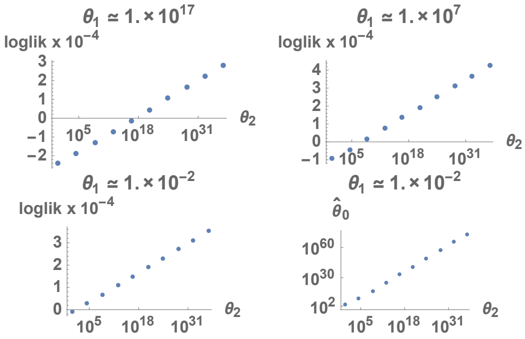

Figure 10. *Profile log-likelihood l_{14} (4.4) and estimated scale parameter $\hat{\theta}_0$ for Branin function (4.3).*

finding also appears to be documented in a thesis [16]. Figure 9 shows that at the critical value $n^*(p)$, $2L_{n^*(p)}(\theta_1)$ monotonically increases with a finite asymptote for some choices of p. Since the Branin function is a quadratic polynomial along its second dimension, we expect that the MLE for the second range parameter $\hat{\theta}_2$ does not exist. In fact, Figure 10 shows that for different choices of θ_1, the profile log-likelihood (4.4) increases for increasing θ_2. Also, the estimated scale parameter appears to be unbounded above as θ_2 increases.

4.2.2. Experiments with nugget effect.

A common approach to overcome the numerical difficulties in computing with the covariance function (1.1) is to include a small nugget effect to stabilize the computation of the covariance matrix inversion [3, 24]. In the following, we add a small nugget effect δ_0 so that the covariance matrix of the observations has the form

$$\theta_0\{R(\theta_1, m) \otimes R(\theta_2, m) + \delta_0 I_{m^2}\}, \tag{4.8}$$

where $m = 12$ is the number of observations along each dimension. In the above formulation, we treat the nugget size δ_0 fixed when fitting the model. For the two test functions (4.1) and (4.2), we investigate the effect of including a nugget on model fitting and prediction.

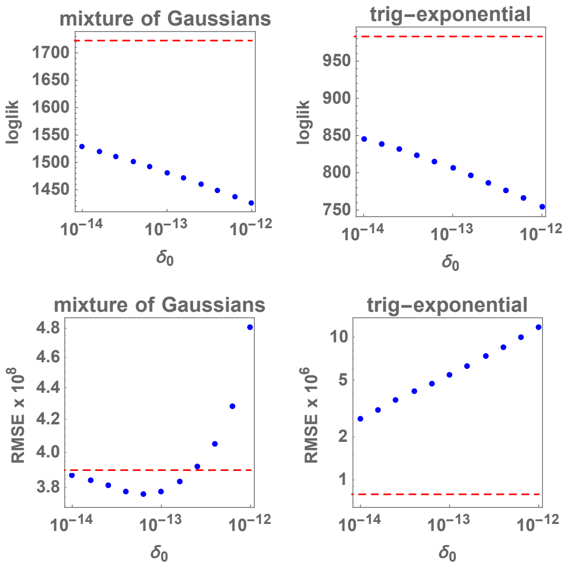

Figure 11. *Log-likelihoods and root mean squared prediction errors of $(\hat\theta_1, \hat\theta_2)$ obtained with exact computations for model (4.8). Reference line indicates nugget-free case.*

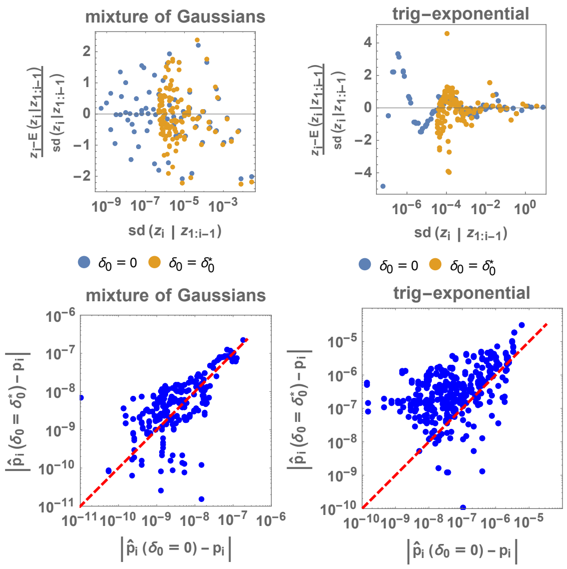

Figure 12. *Successive conditional standardized errors for ordered observations (top); absolute prediction error for $\delta_0 = 0$ and $\delta_0 = \delta_0^*$ yielding the smallest RMSE for each function (bottom).*

With the alternative model (4.8), we evaluate the log-likelihood and prediction errors of the MLE $(\hat\theta_1, \hat\theta_2)$ obtained with exact computations in section 4.2.1 for 11 values of δ_0 equally spaced on the log scale between 10^{-14} and 10^{-12}. The value 14 is the largest integer k for which a nugget of 10^{-k} generally yields a covariance matrix that is found to be numerically nonsingular by the Cholesky Decomposition routine of *Mathematica*. For each δ_0, the scale parameter θ_0 is refitted with the model (4.8) but the range parameters are not changed. Figure 11 shows the log-likelihood and RMSEs of $(\hat\theta_1, \hat\theta_2)$ for the model (4.8).

For both test functions, the likelihood is substantially reduced when including even a nugget of 10^{-14} and decreases for increasing nugget size. To help see why the log-likelihood changes so much, note that log-likelihood can also be obtained with the conditional distributions of successive ordered observations so that

$$l(\theta|\mathbf{z}) = -\frac{n}{2}\log(2\pi) - \sum_{i=1}^{n} \log\left(\mathrm{sd}(z_i|z_1,\ldots,z_{i-1})\right) - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{z_i - E(z_i|z_1,\ldots,z_{i-1})}{\mathrm{sd}(z_i|z_1,\ldots,z_{i-1})}\right)^2,$$

where $\mathrm{sd}(z_i|z_1,\ldots,z_{i-1})$ denotes conditional standard deviation and $E(z_i|z_1,\ldots,z_{i-1})$ denotes conditional mean. The log-likelihood can hence be expressed by conditional standardized errors and conditional standard deviations. We order the observations lexicographically so that $f(\frac{i_1}{m}, \frac{i_2}{m})$ precedes $f(\frac{j_1}{m}, \frac{j_2}{m})$ if and only if $i_1 < j_1$ or whenever $i_1 = j_1$, $i_2 < j_2$. The top panels of Figure 12 show, for each observation, the standardized errors and standard deviations conditional on previous observations. For each test function, we compare the case with $\delta_0 = 0$ and the case with the nugget size $\delta_0^* > 0$ yielding the smallest prediction error among the 11 positive values for δ_0 we considered, which for the mixture of Gaussians, is

$\delta_0^* = 10^{-66/5} \approx 6.3 \times 10^{-14}$ and for the trig-exponential function, is $\delta_0^* = 10^{-14}$. For both test functions, the conditional standardized errors are similarly calibrated for $\delta_0 = 0$ and $\delta_0 = \delta_0^*$. However, some of the conditional standard deviations are much smaller when there is no nugget. Successive predictions of the ordered observations are more accurate at some locations when the model does not have a nugget.

For the trig-exponential function, Figure 11 shows root mean squared prediction error increases for increasing nugget size. However, better successive predictions at the test observations does not necessarily imply better predictions at other locations: we do see that for the mixture of Gaussians, the prediction error is slightly smaller for $\delta_0 = 10^{-14}$ compared with the nugget-free case and further decreases as δ_0 increases before eventually increasing. The bottom panels of Figure 12 show the absolute prediction errors for each predictand when $\delta_0 = 0$ and $\delta_0 = \delta_0^*$. For the mixture of Gaussians, there is no obvious dominance for either $\delta_0 = 0$ and $\delta_0 = \delta_0^*$. In contrast, for the trig-exponential function, it is evident that the nugget-free model predicts better at most locations, perhaps especially for locations with the largest error, which tend to be near the boundaries of the observation region.

5. Conclusions.

Since the approach was proposed by [25, 26], it has become quite common practice to model the deterministic output of a computer experiment as a realization of a Gaussian process. The Gaussian process with squared exponential covariance function is infinitely differentiable and thus is attractive if the computer model output is known to be smooth. In this article, we investigated the asymptotics for the MLE of the scale parameter for this covariance when the computer response is a pth order monomial. Using exact computation, we investigated and compared the MLE and CV estimates in a prediction problem.

Using the exact expression for the Cholesky factor and its inverse of the correlation matrix derived in [19], we proved that for regularly spaced observations, when the test function is a pth order monomial and the range parameter is fixed, the MLE of the scale parameter $\hat{\theta}_0 \to 0$ when $p = 0$ and $\hat{\theta}_0 \to \infty$ when $p = 1$ as the number of observations $n \to \infty$. When the observations are derivatives of the model function at zero, we derived the exact expression of the inverse Cholesky factor of the correlation matrix and proved asymptotic orders of $\hat{\theta}_0$ for all $p \geq 0$. We are unable to prove an asymptotic order for general $p > 1$ for regularly spaced observations. However, we conjecture that the asymptotic order is the same as that of the derivative case with a possibly different constant.

Though both MLE and CV are used in the computer experiment literature, it is not clear under what circumstances each method will yield smaller prediction errors with more calibrated standardized errors. When a model is misspecified, CV can sometimes be a good way of choosing parameters for prediction, since the CV criterion is based on prediction. For deterministic computer experiments, we know that, in fact, the outputs are not realizations of some Gaussian process model. Nevertheless, our experiments show that CV does not always improve upon the likelihood method. For example, CV estimates produce much larger prediction errors and poorer calibration for the mixture of Gaussians and modestly better predictions but far worse calibration for the trig-exponential function. This finding is consistent with the findings in [4] which shows CV appears to be less robust than the MLE to model misspecification under regular grid design.

These numerical experiments used exact arithmetic, which will not generally be possible. Adding a small nugget is a common approach to alleviate the numerical instabilities in decomposing nearly singular covariance matrices. Our experiments suggest model fitting (as measured by the likelihood) deteriorates substantially by adding even a nugget that barely makes the covariance matrix numerically positive definite, whereas prediction can sometimes be slightly improved by adding a small nugget. Another interesting finding of our work is that, for the Branin test function, the MLEs do not appear to exist. This result can be viewed as an implication of the numerical finding that when the model function is a pth order monomial, the MLE of the range parameter does not exist when the number of observations exceeds a critical value (see also [16, 18]). For test functions that are a polynomial in one of its dimensions, the use of the likelihood method is not expected to produce meaningful estimates and inference. Though our results regarding the estimation of range parameters is empirical and based on limited numerical experiments, we believe that the examples shown give an indication of possible issues and consequences when using simple smooth test functions to study how well Gaussian process models work for deterministic computer models. Whether similar results might hold for, say, the numerical solution of a complex system of differential equations deserves further study.

Appendix A. Proofs of statements in sections 2–4. In order to prove Proposition 2.1, Lemma 2.2, Theorems 2.3 and 2.4, we need the following lemmas.

Lemma A.1. *For* $0 \le r \le m$ *and* $i \ge 1$,

(a) $\begin{bmatrix} m \\ r \end{bmatrix}_q = \begin{bmatrix} m \\ m-r \end{bmatrix}_q$,

(b) $\begin{bmatrix} m \\ r \end{bmatrix}_q = q^r \begin{bmatrix} m-1 \\ r \end{bmatrix}_q + \begin{bmatrix} m-1 \\ r-1 \end{bmatrix}_q = \begin{bmatrix} m-1 \\ r \end{bmatrix}_q + q^{(m-r)} \begin{bmatrix} m-1 \\ r-1 \end{bmatrix}_q$,

(c) $\lim_{q\to 1} \begin{bmatrix} m \\ r \end{bmatrix}_q = \binom{m}{r}$,

(d) $\sum_{j=1}^{i} (-w)^{i-j} \begin{bmatrix} i-1 \\ j-1 \end{bmatrix}_{w^2} = \prod_{k=1}^{i-1} \left(1 + (-w)^k\right)$.

We refer the readers to [2] for the proof of Lemma A.1.

Lemma A.2. *If* $a_n \sim b_n$, $a_n > 0$, *and* $\sum_{n=1}^{\infty} a_n = \infty$, *then* $\sum_{n=1}^{N} a_n \sim \sum_{n=1}^{N} b_n$ *as* $N \to \infty$.

Proof. We carry out a standard $\varepsilon - N$ proof.

Since $a_n \sim b_n$, for any $\varepsilon > 0$, there exists $N_0(\varepsilon) > 0$ such that for any $n \ge N_0(\varepsilon)$,

$$|a_n - b_n| < \varepsilon a_n / 2;$$

then for any $N \ge N_0(\varepsilon)$,

$$\sum_{n=N_0(\varepsilon)}^{N} |a_n - b_n| < \frac{\varepsilon}{2} \sum_{n=N_0(\varepsilon)}^{N} a_n. \tag{A.1}$$

Since $\sum_{n=1}^{\infty} a_n = \infty$, there exists $N_1(\varepsilon) \ge N_0(\varepsilon)$ such that for any $N \ge N_1(\varepsilon)$,

$$\sum_{n=1}^{N_0(\varepsilon)} |a_n - b_n| < \frac{\varepsilon}{2} \sum_{n=N_0(\varepsilon)}^{N} a_n. \tag{A.2}$$

As a result, for any $N \geq N_1(\varepsilon) \geq N_0(\varepsilon)$,

$$\frac{\left|\sum_{n=1}^N (a_n - b_n)\right|}{\sum_{n=1}^N a_n} \leq \frac{\sum_{n=1}^{N_0(\varepsilon)-1} |a_n - b_n| + \sum_{n=N_0(\varepsilon)}^N |a_n - b_n|}{\sum_{n=N_0(\varepsilon)}^N a_n} \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon,$$

where the second inequality follows from (A.2) and (A.1). ■

Lemma A.3. *$\sum_{j=1}^i (-w)^{i-j} \begin{bmatrix} i-1 \\ j-1 \end{bmatrix}_{w^2} j^p - \sum_{j=1}^{i-1} (-w)^{i-j-1} \begin{bmatrix} i-2 \\ j-1 \end{bmatrix}_{w^2} j^p = A_{ip}(w) + B_{ip}(w)$, where*

$$A_{ip}(w) = (-w)^{i-1} \sum_{j=1}^{i-1} (-w)^{i-1-j} \begin{bmatrix} i-2 \\ j-1 \end{bmatrix}_{w^2} (i-1-j)^p,$$

$$B_{ip}(w) = \sum_{j=1}^{i} (-w)^{i-j} \begin{bmatrix} i-1 \\ j-1 \end{bmatrix}_{w^2} (j^p - (j-1)^p)$$

for $p \geq 0$ and $i - 1 > p$.

Proof.

$$\begin{aligned}
&\sum_{j=1}^i (-w)^{i-j} \begin{bmatrix} i-1 \\ j-1 \end{bmatrix}_{w^2} j^p - \sum_{j=1}^{i-1} (-w)^{i-j-1} \begin{bmatrix} i-2 \\ j-1 \end{bmatrix}_{w^2} j^p \\
&= (-w)^{i-1} + (i^p - (i-1)^p) + \sum_{k=1}^{i-2} (-w)^{i-k-1} \left(\begin{bmatrix} i-1 \\ k \end{bmatrix}_{w^2} (k+1)^p - \begin{bmatrix} i-2 \\ k-1 \end{bmatrix} k^p \right) \\
&= (-w)^{i-1} + (i^p - (i-1)^p) + \sum_{k=1}^{i-2} (-w)^{i-k-1} \left(w^{2k} \begin{bmatrix} i-2 \\ k \end{bmatrix}_{w^2} k^p + \begin{bmatrix} i-1 \\ k \end{bmatrix}_{w^2} ((k+1)^p - k^p) \right) \\
&= (-w)^{i-1} \sum_{k=1}^{i-2} (-w)^k \begin{bmatrix} i-2 \\ k \end{bmatrix}_{w^2} k^p + \sum_{k=0}^{i-1} (-w)^{i-k-1} \begin{bmatrix} i-1 \\ k \end{bmatrix}_{w^2} ((k+1)^p - k^p) \\
&= (-w)^{i-1} \sum_{j=1}^{i-1} (-w)^{i-1-j} \begin{bmatrix} i-2 \\ j-1 \end{bmatrix}_{w^2} (i-1-j)^p + \sum_{j=1}^{i} (-w)^{i-j} \begin{bmatrix} i-1 \\ j-1 \end{bmatrix}_{w^2} (j^p - (j-1)^p) \\
&= A_{ip}(w) + B_{ip}(w),
\end{aligned}$$

where the second equality is obtained using Lemma A.1(b), and the fourth equality is obtained by a change of variable $j = i - 1 - k$ for $A_{ip}(w)$ and $j = k + 1$ for $B_{ip}(w)$. ■

The following lemma gives a factorization of $\sum_{j=1}^i (-w)^{i-j} \begin{bmatrix} i-1 \\ j-1 \end{bmatrix}_{w^2} j^p$ that enables simplification of $\hat{\theta}_0$.

Lemma A.4.

$$\sum_{j=1}^i (-w)^{i-j} \begin{bmatrix} i-1 \\ j-1 \end{bmatrix}_{w^2} j^p = \begin{cases} (1-w)^{\frac{i-p-1}{2}} f_{ip}(w), & i-p \text{ odd}, \\ (1-w)^{\frac{i-p}{2}} g_{ip}(w), & i-p \text{ even} \end{cases}$$

for any $w \in (0,1)$, where, for $p \geq 0$ and $i > p$, $f_{ip}(w)$ is a polynomial, $f_{ip}(1) = \frac{(i-1)!}{(\frac{i-p-1}{2})!}$, $g_{ip}(w)$ is a polynomial, and $g_{ip}(1) = \frac{i!(p+1)}{2(\frac{i-p}{2})!}$.

Proof. We denote for simplicity $m_{ip}(w) := \sum_{j=1}^{i}(-w)^{i-j}\begin{bmatrix} i-1 \\ j-1 \end{bmatrix}_{w^2} j^p$. Here we make an induction on (i,p) where $p \geq 0$ and $i > p$. Specifically, we prove the following three steps of which the first two serve as an induction basis:

(a) The statement holds for any $p = 0$ and $i > 0$.
With Lemma A.1(d), we have

$$m_{i0}(w) = \sum_{j=1}^{i}(-w)^{i-j}\begin{bmatrix} i-1 \\ j-1 \end{bmatrix}_{w^2} = \prod_{k=1}^{i-1}\left(1+(-w)^k\right) = \begin{cases} (1-w)^{\frac{i-1}{2}} f_{i0}(w), & i \text{ odd}, \\ (1-w)^{\frac{i}{2}} g_{i0}(w), & i \text{ even}, \end{cases}$$

where $f_{i0}(w)$ and $g_{i0}(w)$ are polynomials and

$$f_{i0}(1) = 2^{\frac{i-1}{2}}(i-2)!! = \frac{(i-1)!}{(\frac{i-1}{2})!},$$
$$g_{i0}(1) = 2^{\frac{i-2}{2}}(i-1)!! = \frac{i!}{2(\frac{i}{2})!}.$$

(b) The statement holds for any $i = p+1$ and $p \geq 1$.
When $i = p+1$, $i - p$ is odd and $m_{p+1,p}(w)$ is a polynomial by definition. Then by Lemma A.1(c),

$$\begin{aligned} m_{p+1,p}(1) &= \sum_{j=1}^{p+1}(-1)^{p+1-j}\binom{p}{j-1} j^p \\ &= (-1)^p \sum_{k=0}^{p}(-1)^k \binom{p}{k}(k+1)^p \\ &= (-1)^p(-1)^p p! \\ &= p!. \end{aligned}$$

(c) Suppose the statement holds for any (i', p') such that $(0 \leq p' < p$ and $i' > p')$ or $(p' = p$ and $p' < i' < i)$, then the statement also holds for (i,p).
Define the following polynomials in w:

$$\begin{aligned} h_{ip}(w) = (-w)^{i-1}\sum_{k=3}^{p}(-1)^{p-k}(i-1)^k\binom{p}{k}\Big((1-w)^{\frac{k-3}{2}} g_{i-1,p-k}(w)\mathbf{1}_{(k \text{ odd})} \\ + (1-w)^{\frac{k-4}{2}} f_{i-1,p-k}(w)\mathbf{1}_{(k \text{ even})}\Big), \end{aligned}$$

$$r_{ip}(w) = \sum_{k=0}^{p-2}(-1)^{p-k+1}\binom{p}{k}\left((1-w)^{\frac{p-k-3}{2}} f_{ik}(w)\mathbf{1}_{(i-k \text{ odd})} + (1-w)^{\frac{p-k-2}{2}} g_{ik}(w)\mathbf{1}_{(i-k \text{ even})}\right),$$

$$
\begin{aligned}
u_{ip}(w) = (-w)^{i-1}\sum_{k=2}^{p}(-1)^{p-k}(i-1)^k\binom{p}{k}\bigg(&(1-w)^{\frac{k-3}{2}}f_{i-1,p-k}(w)\mathbf{1}_{(k\text{ odd})}\\
&+(1-w)^{\frac{k-2}{2}}g_{i-1,p-k}\mathbf{1}_{(k\text{ even})}\bigg),\\
v_{ip}(w) = \sum_{k=0}^{p-1}(-1)^{p-k+1}\binom{p}{k}\Big(&(1-w)^{p-k-2}f_{ik}(w)\mathbf{1}_{(i-k\text{ odd})}+(1-w)^{p-k-1}g_{ik}(w)\mathbf{1}_{(i-k\text{ even})}\Big).
\end{aligned}
$$

When $i-p$ is even, letting $A_{ip}(w)$ and $B_{ip}(w)$ be defined as in Lemma A.3 and from binomial expansion, we have

$$
\begin{aligned}
A_{ip}(w) &= (-w)^{i-1}\sum_{j=1}^{i-1}(-w)^{i-1-j}\begin{bmatrix} i-2\\ j-1\end{bmatrix}_{w^2}(i-1-j)^p\\
&= (-w)^{i-1}\sum_{j=1}^{i-1}(-w)^{i-1-j}\begin{bmatrix} i-2\\ j-1\end{bmatrix}_{w^2}\\
&\quad\times\Bigg((-1)^p j^p+(-1)^{p-1}p(i-1)j^{p-1}+(-1)^{p-2}(i-1)^2\binom{p}{2}j^{p-2}\\
&\qquad+\sum_{k=3}^{p}(-1)^{p-k}(i-1)^k\binom{p}{k}j^{p-k}\Bigg)\\
&= (-w)^{i-1}\Bigg((-1)^p m_{i-1,p}(w)+(-1)^{p-1}p(i-1)m_{i-1,p-1}(w)\\
&\qquad\qquad+(-1)^{p-2}(i-1)^2\binom{p}{2}m_{i-1,p-2}(w)\Bigg)\\
&\quad+(-w)^{i-1}\sum_{k=3}^{p}(-1)^{p-k}(i-1)^k\binom{p}{k}m_{i-1,p-k}(w)\\
&= (-w)^{i-1}\left((-1)^p(1-w)^{\frac{i-p-2}{2}}f_{i-1,p}(w)+(-1)^{p-1}p(i-1)(1-w)^{\frac{i-p}{2}}g_{i-1,p-1}(w)\right)\\
&\quad+(-w)^{i-1}\left((-1)^{p-2}(i-1)^2\binom{p}{2}(1-w)^{\frac{i-p}{2}}f_{i-1,p-2}(w)\right)\\
&\quad+(-w)^{i-1}\sum_{k=3}^{p}(-1)^{p-k}(i-1)^k\binom{p}{k}\\
&\qquad\times\Big((1-w)^{\frac{i-1-p+k}{2}}g_{i-1,p-k}(w)\mathbf{1}_{(i-p+k-1\text{ even})}\\
&\qquad\quad+(1-w)^{\frac{i-p+k-2}{2}}f_{i-1,p-k}(w)\mathbf{1}_{(i-p+k-1\text{ odd})}\Big)\\
&= (-w)^{i-1}\left((-1)^p(1-w)^{\frac{i-p-2}{2}}f_{i-1,p}(w)+(-1)^{p-1}p(i-1)(1-w)^{\frac{i-p}{2}}g_{i-1,p-1}(w)\right)\\
&\quad+(-w)^{i-1}\left((-1)^{p-2}(i-1)^2\binom{p}{2}(1-w)^{\frac{i-p}{2}}f_{i-1,p-2}(w)\right)
\end{aligned}
$$

$$
\begin{aligned}
&+(-w)^{i-1}\sum_{k=3}^{p}(-1)^{p-k}(i-1)^k\binom{p}{k}(1-w)^{\frac{i-p}{2}+1}\\
&\quad\times\left((1-w)^{\frac{k-3}{2}}g_{i-1,p-k}(w)\mathbf{1}_{(k\text{ odd})}+(1-w)^{\frac{k-4}{2}}f_{i-1,p-k}(w)\mathbf{1}_{(k\text{ even})}\right)\\
&=(-w)^{i-1}\left((-1)^p(1-w)^{\frac{i-p-2}{2}}f_{i-1,p}(w)+(-1)^{p-1}p(i-1)(1-w)^{\frac{i-p}{2}}g_{i-1,p-1}(w)\right)\\
&+(-w)^{i-1}\left((-1)^{p-2}(i-1)^2\binom{p}{2}(1-w)^{\frac{i-p}{2}}f_{i-1,p-2}(w)\right)+(1-w)^{\frac{i-p}{2}+1}h_{ip}(w),
\end{aligned}
$$

and

$$
\begin{aligned}
B_{ip}(w)&=\sum_{j=1}^{i}(-w)^{i-j}\begin{bmatrix}i-1\\j-1\end{bmatrix}_{w^2}(j^p-(j-1)^p)\\
&=\sum_{j=1}^{i}(-w)^{i-j}\begin{bmatrix}i-1\\j-1\end{bmatrix}_{w^2}\left(\sum_{k=0}^{p-1}\binom{p}{k}j^k(-1)^{(p-k+1)}\right)\\
&=\sum_{j=1}^{i}(-w)^{i-j}\begin{bmatrix}i-1\\j-1\end{bmatrix}_{w^2}\left(pj^{p-1}+\sum_{k=0}^{p-2}\binom{p}{k}j^k(-1)^{(p-k+1)}\right)\\
&=pm_{i,p-1}(w)+\sum_{k=0}^{p-2}(-1)^{(p-k+1)}\binom{p}{k}m_{ik}(w)\\
&=p(1-w)^{\frac{i-p}{2}}f_{i,p-1}(w)\\
&\quad+\sum_{k=0}^{p-2}(-1)^{p-k+1}\binom{p}{k}\left((1-w)^{\frac{i-k-1}{2}}f_{ik}(w)\mathbf{1}_{(i-k\text{ odd})}+(1-w)^{\frac{i-k}{2}}g_{ik}(w)\mathbf{1}_{(i-k\text{ even})}\right)\\
&=p(1-w)^{\frac{i-p}{2}}f_{i,p-1}(w)\\
&\quad+\sum_{k=0}^{p-2}(-1)^{p-k+1}\binom{p}{k}(1-w)^{\frac{i-p}{2}+1}\Bigg((1-w)^{\frac{p-k-3}{2}}f_{ik}(w)\mathbf{1}_{(i-k\text{ odd})}\\
&\qquad\qquad+(1-w)^{\frac{p-k-2}{2}}g_{ik}(w)\mathbf{1}_{(i-k\text{ even})}\Bigg)\\
&=p(1-w)^{\frac{i-p}{2}}f_{i,p-1}(w)+(1-w)^{\frac{i-p}{2}+1}r_{ip}(w).
\end{aligned}
$$

Then by Lemma A.3, we have

$$
\begin{aligned}
m_{ip}(w)&=m_{i-1,p}(w)+A_{ip}(w)+B_{ip}(w)\\
&=(1-w)^{\frac{i-p-2}{2}}f_{i-1,p}(w)-w^{i-1}(1-w)^{\frac{i-p-2}{2}}f_{i-1,p}(w)\\
&\quad+w^{i-1}p(i-1)(1-w)^{\frac{i-p}{2}}g_{i-1,p-1}(w)\\
&\quad-w^{i-1}(i-1)^2\binom{p}{2}(1-w)^{\frac{i-p}{2}}f_{i-1,p-2}(w)+p(1-w)^{\frac{i-p}{2}}f_{i,p-1}(w)\\
&\quad+(1-w)^{\frac{i-p}{2}+1}h_{ip}(w)+(1-w)^{\frac{i-p}{2}+1}r_{ip}(w)
\end{aligned}
$$

$$
\begin{aligned}
&= (1-w)^{\frac{i-p}{2}}\Bigg(f_{i-1,p}(w)\left(\frac{1-w^{i-1}}{1-w}\right) + p(i-1)w^{i-1}g_{i-1,p-1}(w) \\
&\qquad - \binom{p}{2}(i-1)^2 w^{i-1} f_{i-1,p-2}(w) + pf_{i,p-1}(w) \\
&\qquad + (1-w)h_{ip}(w) + (1-w)r_{ip}(w)\Bigg).
\end{aligned}
$$

Let

$$
\begin{aligned}
g_{ip}(w) = f_{i-1,p}(w)&\left(\frac{1-w^{i-1}}{1-w}\right) + p(i-1)w^{i}g_{i-1,p-1}(w) \\
&- \binom{p}{2}(i-1)^2 w^{i-1} f_{i-1,p-2}(w) + pf_{i,p-1}(w) \\
&+ (1-w)h_{ip}(w) + (1-w)r_{ip}(w)
\end{aligned}
$$

which is a polynomial by the induction hypothesis, and

$$
\begin{aligned}
g_{ip}(1) &= (i-1)\frac{(i-2)!}{(\frac{i-p-2}{2})!} + p(i-1)\frac{(i-1)!p}{2(\frac{i-p}{2})!} - \binom{p}{2}(i-1)^2\frac{(i-2)!}{(\frac{i-p}{2})!} + p\frac{(i-1)!}{(\frac{i-p}{2})!} \\
&= \frac{(i-1)!}{(\frac{i-p}{2})!}\left(\frac{i-p}{2} + p + \frac{p^2(i-1)}{2} - \frac{p(p-1)(i-1)}{2}\right) \\
&= \frac{i!(p+1)}{2(\frac{i-p}{2})!}.
\end{aligned}
$$

When $i-p$ is odd, similarly by binomial expansion and the induction hypothesis, we have

$$
\begin{aligned}
A_{ip}(w) &= (-w)^{i-1}\left((-1)^p(1-w)^{\frac{i-p-1}{2}}g_{i-1,p}(w) + (-1)^{p-1}p(i-1)(1-w)^{\frac{i-p-1}{2}}f_{i-1,p-1}(w)\right) \\
&\quad + (-w)^{i-1}\sum_{k=2}^{p}(-1)^{p-k}(i-1)^k\binom{p}{k}m_{i-1,p-k}(w) \\
&= (-w)^{i-1}\left((-1)^p(1-w)^{\frac{i-p-1}{2}}g_{i-1,p}(w) + (-1)^{p-1}p(i-1)(1-w)^{\frac{i-p-1}{2}}f_{i-1,p-1}(w)\right) \\
&\quad + (-w)^{i-1}\sum_{k=2}^{p}(-1)^{p-k}(i-1)^k\binom{p}{k} \\
&\qquad \times (1-w)^{\frac{i-p+1}{2}}\left((1-w)^{\frac{k-3}{2}}f_{i-1,p-k}(w)\mathbf{1}_{(k\text{ odd})} + (1-w)^{\frac{k-2}{2}}g_{i-1,p-k}\mathbf{1}_{(k\text{ even})}\right) \\
&= (-w)^{i-1}\left((-1)^p(1-w)^{\frac{i-p-1}{2}}g_{i-1,p}(w) + (-1)^{p-1}p(i-1)(1-w)^{\frac{i-p-1}{2}}f_{i-1,p-1}(w)\right) \\
&\quad + (1-w)^{\frac{i-p+1}{2}}u_{ip}(w),
\end{aligned}
$$

and

$$
B_{ip}(w) = \sum_{k=0}^{p-1}(-1)^{p-k+1}\binom{p}{k}m_{ik}(w)
$$

$$
\begin{aligned}
&= \sum_{k=0}^{p-1} (-1)^{p-k+1} \binom{p}{k} \\
&\quad \times (1-w)^{\frac{i-p+1}{2}} \left((1-w)^{p-k-2} f_{ik}(w) \mathbf{1}_{(i-k \text{ odd})} + (1-w)^{p-k-1} g_{ik}(w) \mathbf{1}_{(i-k \text{ even})} \right) \\
&= (1-w)^{\frac{i-p+1}{2}} v_{ip}(w).
\end{aligned}
$$

Then similarly by Lemma A.3, we have

$$
\begin{aligned}
m_{ip}(w) &= m_{i-1,p}(w) + A_{ip}(w) + B_{ip}(w) \\
&= (1-w)^{\frac{i-p-1}{2}} \left(g_{i-1,p}(w) + w^{i-1} g_{i-1,p}(w) - p(i-1) w^{i-1} f_{i-1,p-1}(w) \right) \\
&\quad + (1-w)^{\frac{i-p-1}{2}} \left((1-w) u_{ip}(w) + (1-w) v_{ip}(w) \right).
\end{aligned}
$$

Let

$$
\begin{aligned}
f_{ip}(w) = g_{i-1,p}(w) + w^{i-1} g_{i-1,p}(w) - p(i-1) w^{i-1} f_{i-1,p-1}(w) \\
+ (1-w) u_{ip}(w) + (1-w) v_{ip}(w)
\end{aligned}
$$

which is a polynomial, and

$$
\begin{aligned}
f_{ip}(1) &= 2 \frac{(i-1)!(p+1)}{2(\frac{i-p-1}{2})!} - p(i-1) \frac{(i-2)!}{(\frac{i-p-1}{2})!} \\
&= \frac{(i-1)!}{(\frac{i-p-1}{2})!}.
\end{aligned}
$$

■

Lemma A.5. *$a_{i0}(w)$ monotonically decreases for $w \in (0,1)$ and any $i \geq 1$.*

Proof. With Lemma A.1(d), we have

$$
\begin{aligned}
a_{i0}(w) &= \frac{\prod_{k=1}^{i-1} \left(1 + (-w)^k\right)^2}{\prod_{k=1}^{i-1} (1 - w^{2k})} \\
&= \prod_{k=1}^{i-1} \frac{1 + (-w)^k}{1 - (-w)^k}.
\end{aligned}
$$

For $k \geq 1$, let

$$
f_k(w) = \frac{(1-w^k)(1+w^{k+1})}{(1+w^k)(1-w^{k+1})}, \tag{A.3}
$$

then $f_k(w) = \left(\frac{2}{1+w^k} - 1\right)\left(\frac{2}{1-w^{k+1}} - 1\right)$ and

$$
\begin{aligned}
f_k'(w) &= \frac{-2kw^{k-1}}{(1+w^k)^2} \left(\frac{2}{1-w^{k+1}} - 1 \right) + \frac{2(k+1)w^k}{(1-w^{k+1})^2} \left(\frac{2}{1+w^k} - 1 \right) \\
&= \frac{1}{(1+w^k)^2 (1-w^{k+1})^2} 2w^{k-1} \left(kw^{2k+2} - (k+1) w^{2k+1} + (k+1) w - k \right).
\end{aligned}
$$

For $g_k(w) = kw^{2k+2} - (k+1)w^{2k+1} + (k+1)w - k$, we know $g_k(1) = 0$ and

$$g_k'(w) = (k+1)(1-w)(1 + w + \cdots + w^{2k-1} - 2kw^{2k}) > 0$$

for $w \in (0,1)$. As a result, $g_k(w) < 0$ for $w \in (0,1)$ and hence $f_k(w)$ monotonically decreases on $(0,1)$. Because

$$a_{i0}(w) = \begin{cases} \prod_{k=1}^{\frac{i-1}{2}} f_{2k-1}(w), & i \text{ odd}, \\ \prod_{k=1}^{\frac{i-2}{2}} f_{2k-1}(w) \frac{1-w^{i-1}}{1+w^{i-1}}, & i \text{ even}, \end{cases}$$

$a_{i0}(w)$ is monotonically decreasing over $(0,1)$. ■

Lemma A.6. *Denote $w_i^* := e^{-\frac{1}{2(i-2)}}$, then $a_{i1}(w)$ monotonically decreases for $w \in (w_i^*, 1)$ and any $i \geq 2$ and i even.*

Proof. We prove this by induction. Note that by the definition of w, we have $n^2 = \left(\theta_1 \log \frac{1}{w}\right)^{-1}$.

When $i = 2$,

$$a_{21}(w) = \frac{(2-w)^2 \theta_1 \log \frac{1}{w}}{(1-w^2)},$$
$$a_{21}'(w) = \frac{(2-w)\theta_1}{(1-w^2)^2} \left(-2(1-w^2)\log\frac{1}{w} + (2-w)\left(2w \log \frac{1}{w} + w - \frac{1}{w}\right)\right).$$

Consider $z(w) = 2w \log \frac{1}{w} + w - \frac{1}{w}$, then $z(1) = 0$ and $z'(w) = \frac{1}{w^2} + 2\log\frac{1}{w} - 1 > 0$ for $w \in (0,1)$. So $z(w) < 0$ for $w \in (0,1)$. As a result, $a_{21}'(w) < 0$ for $w \in (0,1)$. Note that $w_2^* = 0$. We denote for simplicity $m_{ip}(w) := \sum_{j=1}^{i} (-w)^{i-j} \begin{bmatrix} i-1 \\ j-1 \end{bmatrix}_{w^2} j^p$. Suppose the statement holds for some even i, then for $i+2$, by repeatedly applying Lemma A.3, we obtain

$$m_{i+2,1}(w) = (1 + w^{i+1})(1-w^i) m_{i1}(w) + \left(iw^i(1-w) + (1+w^i)(2 - w^{i+1})\right) m_{i0}(w).$$

We consider, by using Lemma A.1(d),

$$\begin{aligned}
\sqrt{a_{i+2,1}(w)} &= \frac{m_{i+2,1}(w)}{n\sqrt{\prod_{k=1}^{i+1}(1-w^{2k})}} \\
&= \frac{(1+w^{i+1})(1-w^i) m_{i1}(w)}{n\sqrt{(1-w^{2i})(1-w^{2(i+1)})\prod_{k=1}^{i-1}(1-w^{2k})}} \\
&\quad + \frac{m_{i0}(w)\left(iw^i(1-w) + (1+w^i)(2-w^{i+1})\right)}{n\sqrt{\prod_{k=1}^{i+1}(1-w^{2k})}} \\
&= \sqrt{a_{i1}(w)}\sqrt{\frac{(1-w^i)(1+w^{i+1})}{(1+w^i)(1-w^{i+1})}} \\
&\quad + \sqrt{a_{i0}(w)} \frac{iw^i(1-w) + (1+w^i)(2-w^{i+1})}{n\sqrt{(1-w^{2i})(1-w^{2(i+1)})}}
\end{aligned}$$

$$
\begin{aligned}
&= \sqrt{a_{i1}(w)}\sqrt{\frac{(1-w^i)(1+w^{i+1})}{(1+w^i)(1-w^{i+1})}} \\
&\quad + \sqrt{a_{i+2,0}(w)}\frac{iw^i(1-w)+(1+w^i)(1-w^{i+1})+(1+w^i)}{n(1+w^i)(1-w^{i+1})} \\
&= \sqrt{a_{i1}(w)}\sqrt{\frac{(1-w^i)(1+w^{i+1})}{(1+w^i)(1-w^{i+1})}} \\
&\quad + \frac{w^i}{n}\frac{i\sqrt{a_{i+2,0}}}{(1+w^i)(1+w+\cdots+w^i)} + \frac{\sqrt{a_{i+2,0}}}{n} + \frac{\sqrt{a_{i+2,0}}}{n(1-w^{i+1})} \\
&:= I_1(w)+I_2(w)+I_3(w)+I_4(w).
\end{aligned}
$$

Now we show the monotonicity for each $I_i(w)$, $i = 1, \ldots, 4$. First,

$$I_1(w) = \sqrt{a_{i1}(w)f_i(w)},$$

where $f_i(w)$ is defined in (A.3). $I_1(w)$ monotonically decreases for $w > w_i^*$ by the induction hypothesis and the fact that $f_i(w)$ decreases (proved in Lemma A.5). Second, $I_2(w)$ monotonically decreases for $w > w_{i+2}^*$ by Lemma A.5 and the fact that $\frac{w^{2i}}{n^2}$ monotonically decreases for $w > w_{i+2}^* = e^{-\frac{1}{2i}}$, which follows from

$$
\begin{aligned}
\frac{w^{2i}}{n^2} &= \theta_1 w^{2i}\log\frac{1}{w}, \\
\frac{d}{dw}\left(w^{2i}\log\frac{1}{w}\right) &= w^{2i-1}(2i\log\frac{1}{w}-1) < 0 \\
&\Leftrightarrow w > e^{-\frac{1}{2i}}.
\end{aligned}
$$

Third, $I_3(w)$ monotonically decreases for $w \in (0,1)$ by Lemma A.5. Finally,

$$I_4(w) = \frac{\sqrt{a_{i+1,0}(1-w^{i+1})}}{n(1-w^{i+1})\sqrt{1+w^{i+1}}} = \frac{\sqrt{a_{i+1,0}}}{n\sqrt{1-w^{2(i+1)}}},$$

where $a_{i+1,0}$ monotonically decreases for $w \in (0,1)$ by Lemma A.5. Then we show that $n^2(1-w^{2(i+1)})$ monotonically increases over $(0,1)$ as follows:

$$
\begin{aligned}
n^2(1-w^{2(i+1)}) &= \frac{1-w^{2(i+1)}}{\theta_1\log\frac{1}{w}}, \\
\frac{d}{dw}\left(\frac{1-w^{2(i+1)}}{\log\frac{1}{w}}\right) &= \frac{1}{w\log^2\frac{1}{w}}\left(1-w^{2(i+1)}-2(i+1)w^{2(i+1)}\log\frac{1}{w}\right),
\end{aligned}
$$

letting $t(w) = 1-w^{2(i+1)}-2(i+1)w^{2(i+1)}\log\frac{1}{w}$, then $t(1) = 0$ and $t'(w) = -(2i+2)^2w^{2i+1}\log\frac{1}{w} < 0$ for $w \in (0,1)$, so we have $t(w) > 0$ for $w \in (0,1)$. So $I_4(w)$ monotonically decreases for $w \in (0,1)$. As a result, we conclude that $a_{i+2,1}(w)$ monotonically decreases for $w > w_{i+2}^*$. ■

A.1. Proof of Proposition 2.1. With Lemma A.4, we can cancel some factors to obtain

$$\frac{\left(\sum_{j=1}^{i}(-w)^{i-j}\begin{bmatrix} i-1 \\ j-1 \end{bmatrix}_{w^2} j^p\right)^2}{n^{2p}\prod_{k=1}^{i-1}(1-w^{2k})} = \begin{cases} \frac{f_{ip}^2(w)}{(1-w)^p n^{2p}\prod_{k=1}^{i-1} z_k(w)}, & i-p \text{ odd}, \\ \frac{g_{ip}^2(w)(1-w)}{(1-w)^p n^{2p}\prod_{k=1}^{i-1} z_k(w)}, & i-p \text{ even}, \end{cases}$$

where $z_k(w) = (1+w^k)(1+w+\cdots+w^{k-1})$.

Note that

$$(1-w)n^2 \to 1/\theta_1$$

as $n \to \infty$. Then we have

$$l_{ip} = \begin{cases} \frac{f_{ip}^2(1)\theta_1^p}{\prod_{k=1}^{i-1} z_k(1)}, & i-p \text{ odd}, \\ 0, & i-p \text{ even}, \end{cases}$$

which equals the right-hand side of (2.5) by using Lemma A.4.

A.2. Proof of Lemma 2.2. When $i-p$ is odd, by Stirling's approximation, as $i \to \infty$, we have

$$\begin{aligned} l_{ip} &\sim \frac{\sqrt{2\pi(i-1)}(\frac{i-1}{e})^{i-1}\theta_1^p}{2^{i-1}\pi(i-p-1)(\frac{i-p-1}{2e})^{i-p-1}} \\ &\sim \sqrt{\frac{2}{\pi}}\frac{i^{p-\frac{1}{2}}\theta_1^p}{2^p}. \end{aligned}$$

Since $\sum_{i=1}^{\infty} i^{p-1/2} = \infty$ for all $p \geq 0$, by Lemma A.2, as $n \to \infty$, we have

$$\begin{aligned} \sum_{i=p+1}^{n} l_{ip} &= \sum_{k=1}^{\lceil \frac{n-p}{2} \rceil} l_{p+2k-1,p} \\ &\sim \sum_{k=1}^{\lceil \frac{n-p}{2} \rceil} \sqrt{\frac{2}{\pi}}\frac{(2k)^{p-\frac{1}{2}}\theta_1^p}{2^p} \\ &\sim \sqrt{\frac{2}{\pi}}\frac{2^{p-\frac{1}{2}}\theta_1^p}{2^p}\sum_{k=1}^{\lceil \frac{n-p}{2} \rceil} k^{p-\frac{1}{2}} \\ &\sim \sqrt{\frac{2}{\pi}}\frac{2^{p-\frac{1}{2}}\theta_1^p}{2^p}\frac{1}{(p+\frac{1}{2})}(\frac{n-p}{2})^{p+\frac{1}{2}} \\ &\sim \frac{n^{p+\frac{1}{2}}\theta_1^p}{\sqrt{2\pi}2^p(p+\frac{1}{2})} \end{aligned}$$

and Lemma 2.2 follows.

A.3. Proof of Theorem 2.3. Let $s(x) = \frac{1-w^x}{x}$ for some $w \in (0,1)$. Note that $s(x)$ monotonically decreases for $x > 0$. The series expansion is $s(x) = \sum_{l=1}^{\infty} \frac{(-x)^{l-1}}{l!} (\frac{1}{\theta_1 n^2})^l$ for $w = e^{-1/\theta_1 n^2}$. Then

$$\begin{aligned}
s(2k-1) - s(2k) &= \sum_{l=1}^{\infty} \frac{(-1)^l}{\theta_1^l l!} \frac{(2k)^{l-1} - (2k-1)^{l-1}}{n^{2l}} \\
&\leq \sum_{l=1}^{\infty} \frac{1}{\theta_1^{2l}(2l)!} \frac{(2k)^{2l-1} - (2k-1)^{2l-1}}{n^{4l}} \\
&\leq \sum_{l=1}^{\infty} \frac{1}{\theta_1^{2l}(2l)!} \frac{2l\binom{2l-1}{l}(2k)^{2l-2}}{n^{4l}} \\
&\leq \sum_{l=1}^{\infty} \frac{1}{\theta_1^{2l} l!} \frac{(2k)^{2l-2}}{n^{4l}} \\
&\leq e^{1/\theta_1^2} \sum_{l=1}^{\infty} \frac{(2k)^{2l-2}}{n^{4l}},
\end{aligned}$$

where the second inequality comes from the binomial expansion.

Referring back to Definition 3.1, since, when n is odd, $(n-1)!! = 2^{\frac{n-1}{2}} \left(\frac{n-1}{2}\right)!$ and $(n-1)! = (n-1)!!(n-2)!!$, we have

$$l_{n0} = \frac{(n-1)!}{2^{n-1}\left(\frac{n-1}{2}!\right)^2} = \frac{(n-2)!!}{(n-1)!!},$$

so

$$\frac{a_{n0}(w)}{l_{n0}} = \prod_{k=1}^{\frac{n-1}{2}} \frac{1+w^{2k}}{1+w^{2k-1}} \frac{s(2k-1)}{s(2k)} \leq \prod_{k=1}^{\frac{n-1}{2}} \frac{s(2k-1)}{s(2k)}$$

and

$$\begin{aligned}
\log \frac{a_{n0}(w)}{l_{n0}} &= \sum_{k=1}^{\frac{n-1}{2}} \log\left(1 + \frac{s(2k-1)-s(2k)}{s(2k)}\right) \\
&\leq \sum_{k=1}^{\frac{n-1}{2}} \frac{s(2k-1)-s(2k)}{s(n-1)} \\
&\leq \frac{e^{1/\theta_1^2}}{s(n-1)} \sum_{k=1}^{\frac{n-1}{2}} \sum_{l=1}^{\infty} \frac{(2k)^{2l-2}}{n^{4l}} \\
&= \frac{e^{1/\theta_1^2}}{s(n-1)} \sum_{l=1}^{\infty} \sum_{k=1}^{\frac{n-1}{2}} \frac{(2k)^{2l-2}}{n^{4l}}
\end{aligned} \tag{A.4}$$

$$\begin{aligned}&\le \frac{e^{1/\theta_1^2}}{s(n-1)}\sum_{l=1}^{\infty}\frac{n^{2l-1}}{n^{4l}}\\ &= \frac{e^{1/\theta_1^2}}{s(n-1)}\frac{n^2}{n^3(n^2-1)}\\ &= \frac{1}{(n+1)}\frac{e^{1/\theta_1^2}}{(1-w^{n-1})n}\to 0\end{aligned}$$

as $n\to\infty$, since $(1-w^{n-1})n\to 1/\theta_1$ as $n\to\infty$.

By Lemma A.5, $\frac{a_{n0}(w)}{l_{n0}}\ge 1$ for $n\ge 1$, and combined with (A.4), we obtain that for n odd,

$$\frac{a_{n0}(w)}{l_{n0}}\to 1$$

as $n\to\infty$. For n even,

$$a_{n0}(w)=a_{n-1,0}(w)\frac{1-w^{n-1}}{1+w^{n-1}}\sim\frac{1}{2\theta_1 n}l_{n-1,0}$$

as $n\to\infty$.

As a result, denoting $w_i=e^{-1/\theta_1 i^2}$, we have, as $n\to\infty$,

$$\sum_{i=1}^{n}a_{i0}(w_i)\sim\sum_{i=1}^{n}l_{i0}+\sum_{i=1}^{n}\frac{1}{2\theta_1 i}l_{i0}\sim\sum_{i=1}^{n}l_{i0}. \tag{A.5}$$

Lemma A.5 implies

$$\sum_{i=1}^{n}l_{i0}\le\sum_{i=1}^{n}a_{i0}(w)\le\sum_{i=1}^{n}a_{i0}(w_i). \tag{A.6}$$

Combining (A.5) and (A.6), we have

$$\hat\theta_0=\frac{1}{n}\sum_{i=1}^{n}a_{i0}(w)\sim\frac{1}{n}\sum_{i=1}^{n}l_{i0}\sim\sqrt{\frac{2}{\pi}}\frac{1}{\sqrt{n}},$$

where the last asymptotic equivalence is obtained by taking $p=0$ in Lemma 2.2.

A.4. Proof of Theorem 2.4. Since there exists $N(\theta_1)$ such that for any $n>N(\theta_1)$,

$$e^{-\frac{1}{2(n-2)}}<e^{-\frac{1}{\theta_1 n^2}}=w,$$

with Lemma A.6, for any $n>N(\theta_1)$,

$$w>w_n^*\ge w_i^*$$

for any $2\le i\le n$ and i even.

As a result,

$$\sum_{i=2}^{n} a_{i1}(w) \geq \sum_{i=2}^{n} l_{i1},$$

then

$$\frac{\hat{\theta}_0}{\sqrt{n}} = \frac{\sum_{i=1}^{n} a_{i1}}{n^{3/2}} \geq \frac{\sum_{i=2}^{n} l_{i1}}{n^{3/2}}.$$

Theorem 2.4 follows by taking limit infimum on both sides of the above inequality and setting $p = 1$ in Lemma 2.2.

A.5. Proof of Proposition 2.6. For convenience, consider n to be even in the subsequent proof. The arguments only need to be slightly modified for n odd. Denote k as the index of the first nonzero element in the observations, then $k = n/2 + 1$ and $f(k/n) = g(1/n)$. The observations are $\mathbf{z} = (0, \ldots, 0, g(1/n), \ldots, g(1/2))^T$. Denote the (k, k)th element of the inverse Cholesky factor by $C^{-1}_{(k,k)}$. Referring back to (2.3) gives

$$\begin{aligned}\hat{\theta}_0 &= \frac{1}{n}\sum_{i=1}^{n} \|C^{-1}\mathbf{z}\|^2 \\ &\geq \frac{1}{n}\left(C^{-1}_{(k,k)} g(1/n)\right)^2 \\ &= \frac{g^2(1/n)}{n\prod_{l=1}^{k-1}(1-w^{2l})}.\end{aligned} \tag{A.7}$$

Let $L = p + 1$, then for all n sufficiently large,

$$\prod_{l=1}^{k-1}(1-w^{2l}) < \prod_{l=1}^{L}(1-w^{2l}). \tag{A.8}$$

Since $w = e^{1/(\theta_1 n^2)}$, $\prod_{l=1}^{L}(1-w^{2l}) \sim \frac{2^L L!}{\theta_1^L n^{2L}}$ as $n \to \infty$. Combining (A.7) and (A.8) gives

$$\begin{aligned}\liminf_{n\to\infty} \frac{\hat{\theta}_0}{n} &\geq \liminf_{n\to\infty} \frac{g^2(1/n)n^{2p}}{n^{2p+2}\prod_{l=1}^{L}(1-w^{2l})} \\ &= \frac{\theta_1^L c}{2^L L!} > 0.\end{aligned}$$

A.6. Proof of a statement in section 2. We state and prove the following proposition.

Proposition A.7. *With the exponential covariance function* $\mathrm{Cov}(f(x), f(y)) = \theta_0 e^{-|x-y|/\theta_1}$, *if the observations are* $\boldsymbol{z} = (f(\frac{1}{n}), f(\frac{2}{n}), \ldots, f(1))^T$ *for some f having a bounded second derivative on* $[0, 1]$, *then as* $n \to \infty$,

$$\hat{\theta}_0 \sim \frac{1}{n} f(0)^2 + \frac{1}{2\theta_1 n}\int_0^1 \left(f(x) + \theta_1 f'(x)\right)^2 \, dx.$$

Proof. Denote the correlation matrix as R and its Cholesky decomposition as $R = CC^T$ for some C lower triangular, then $R_{ij} = \rho^{|i-j|}$, where $\rho = e^{-1/n\theta_1}$. The Cholesky and inverse Cholesky factors are

$$C = \begin{bmatrix} 1 & & & & \\ \rho & \sqrt{1-\rho^2} & & \mathbf{0} & \\ \rho^2 & \rho\sqrt{1-\rho^2} & \sqrt{1-\rho^2} & & \\ \vdots & \vdots & \vdots & \ddots & \\ \rho^{n-1} & \rho^{n-2}\sqrt{1-\rho^2} & \rho^{n-3}\sqrt{1-\rho^2} & \dots & \sqrt{1-\rho^2} \end{bmatrix},$$

$$C^{-1} = \frac{1}{\sqrt{1-\rho^2}} \begin{bmatrix} \sqrt{1-\rho^2} & & & \mathbf{0} \\ -\rho & 1 & & \\ & \ddots & \ddots & \\ \mathbf{0} & & -\rho & 1 \end{bmatrix},$$

so

$$\begin{aligned} \hat{\theta}_0 &= \frac{1}{n}\boldsymbol{z}^T R^{-1}\boldsymbol{z} = \frac{1}{n}\|C^{-1}\boldsymbol{z}\|^2 \\ &= \frac{1}{n}f(0)^2 + \frac{1}{n}\sum_{j=2}^{n} \frac{\left(f(\frac{j}{n}) - \rho f(\frac{j-1}{n})\right)^2}{1-\rho^2}. \end{aligned}$$

A Taylor expansion gives that for some $x_j \in (\frac{j-1}{n}, \frac{j}{n})$, $2 \le j \le n$,

$$\begin{aligned} f\left(\frac{j}{n}\right) - \rho f\left(\frac{j-1}{n}\right) &= f\left(\frac{j-1}{n}\right) + \frac{1}{n}f'\left(\frac{j-1}{n}\right) + \frac{1}{n^2}f''(x_j) - f\left(\frac{j-1}{n}\right)\left(1 - \frac{1}{\theta_1 n} + \frac{\alpha_n}{n^2}\right) \\ &= \frac{1}{\theta_1 n} f\left(\frac{j-1}{n}\right) + \frac{1}{n}f'\left(\frac{j-1}{n}\right) + \frac{r_{jn}}{n^2}, \end{aligned} \tag{A.9}$$

where $|\alpha_n| \le \alpha = 1/(2\theta_1^2)$ and

$$r_{jn} = f''(x_j) - \alpha_n f\left(\frac{j-1}{n}\right).$$

Since $f''(x)$ is bounded on $[0,1]$, $f(x)$ and $f'(x)$ are continuous and bounded on $[0,1]$. Denote $A := \sup_{x\in[0,1]}\{|f(x)|, |f'(x)|, |f''(x)|\}$, then

$$|r_{jn}| \le A + A\alpha.$$

(A.9) gives that

$$\begin{aligned} \sum_{j=2}^{n}\left(f\left(\frac{j}{n}\right) - \rho f\left(\frac{j-1}{n}\right)\right)^2 &= \sum_{j=2}^{n}\left(\frac{1}{\theta_1 n}f\left(\frac{j-1}{n}\right) + \frac{1}{n}f'\left(\frac{j-1}{n}\right)\right)^2 + \frac{\sum_{j=2}^{n} r_{jn}^2}{n^4} \\ &\quad + \sum_{j=2}^{n}\frac{2r_{jn}}{n^2}\left(\frac{1}{\theta_1 n}f\left(\frac{j-1}{n}\right) + \frac{1}{n}f'\left(\frac{j-1}{n}\right)\right), \end{aligned}$$

where as $n \to \infty$,

$$\sum_{j=2}^{n}\left(\frac{1}{\theta_1 n}f\left(\frac{j-1}{n}\right)+\frac{1}{n}f'\left(\frac{j-1}{n}\right)\right)^2 = \frac{1}{n}\sum_{j=2}^{n}\frac{1}{n}\left(\frac{1}{\theta_1}f\left(\frac{j-1}{n}\right)+f'\left(\frac{j-1}{n}\right)\right)^2$$
$$\sim \frac{1}{n}\int_0^1\left(\frac{1}{\theta_1}f\left(\frac{j-1}{n}\right)+f'\left(\frac{j-1}{n}\right)\right)^2 dx,$$

since the integrand is continuous so the sum converges to the corresponding Riemann integral, and

$$\frac{\sum_{j=2}^{n} r_{jn}^2}{n^4} \leq \frac{A^2(1+\alpha)^2}{n^3},$$

$$\left|\sum_{j=2}^{n}\frac{2r_{jn}}{n^2}\left(\frac{1}{\theta_1 n}f\left(\frac{j-1}{n}\right)+\frac{1}{n}f'\left(\frac{j-1}{n}\right)\right)\right| \leq \sum_{j=2}^{n}\frac{2|r_{jn}|}{n^2}\left(\frac{A}{\theta_1 n}+\frac{A}{n}\right)$$
$$\leq \frac{2A^2(1+\alpha)}{n^2}\left(1+\frac{1}{\theta_1}\right).$$

As a result, as $n \to \infty$,

$$\sum_{j=2}^{n}\left(f\left(\frac{j}{n}\right)-\rho f\left(\frac{j-1}{n}\right)\right)^2 \sim \frac{1}{n}\int_0^1\left(\frac{1}{\theta_1}f\left(\frac{j-1}{n}\right)+f'\left(\frac{j-1}{n}\right)\right)^2 dx,$$

together with $\rho = e^{-1/\theta_1 n}$, $1-\rho^2 \sim \frac{2}{\theta_1 n}$ complete the proof. ■

A.7. Proof of Proposition 3.2. First of all, we prove the following lemma that is used frequently in the subsequent proofs.

Lemma A.8. *For $0 \leq p \leq m-1$ and $m \geq 1$,*
(a) $\sum_{l=0}^{m}(-1)^l\binom{m}{l}l^p = 0$,
(b) $\sum_{l=0}^{m}(-1)^l\binom{m}{l}l^m = (-1)^m m!$

Proof. The Stirling numbers of the second kind can be expressed as the sum [5]:

$$S(m,k) = \frac{1}{k!}\sum_{i=0}^{k}(-1)^i\binom{k}{i}(k-i)^m.$$

Lemma A.8 follows from $S(m,m)=1$ and $S(p,m)=0$ for $0 \leq p < m$. ■

For convenience we write $R_1(\theta_1, n) = R_n$ and $D(\theta_1, n) = D_n$. Since both R_n and D_n are nested, we use induction to prove the proposition. First of all, when $n=1$, $R_1 = D_1 = 1$. Suppose Proposition 3.2 is true for n, i.e., $D_n^T D_n R_n = I_n$, then for $n+1$, partition R_{n+1} and D_{n+1} as

$$R_{n+1} := \begin{bmatrix} R_n & r_{n+1} \\ r_{n+1}^T & R_{n+1,n+1} \end{bmatrix},$$

$$D_{n+1} := \begin{bmatrix} D_n & \mathbf{0} \\ d_{n+1}^T & D_{n+1,n+1} \end{bmatrix},$$

then

$$D_{n+1}^T D_{n+1} R_{n+1} = \begin{bmatrix} D_n^T D_n R_n + d_{n+1} A_n & C_n \\ D_{n+1,n+1} A_n & B_n \end{bmatrix},$$

where

$$\begin{aligned}
A_n &= d_{n+1}^T R_n + D_{n+1,n+1} r_{n+1}^T,\\
B_n &= D_{n+1,n+1}(d_{n+1}^T r_{n+1} + D_{n+1,n+1} R_{n+1,n+1}),\\
C_n &= D_n^T D_n r_{n+1} + d_{n+1} d_{n+1}^T r_{n+1} + d_{n+1} D_{n+1,n+1} R_{n+1,n+1}.
\end{aligned}$$

Here we claim that $A_n = \mathbf{0}^T$, $B_n = 1$, and $C_n = \mathbf{0}$ so that along with the induction hypothesis, we have

$$D_{n+1}^T D_{n+1} R_{n+1} = I_{n+1}.$$

(a) Proof of $A_n = \mathbf{0}^T$: note that if n and j are of the same parity, then $(d_{n+1}^T R_n)_j = (r_{n+1})_j = 0$, so we have that the jth element of A_n is 0. If n and j are of different parities,

$$\begin{aligned}
(d_{n+1}^T R_n)_j &= \sum_{i=1,(i+j)\text{even}}^{n} \frac{\sqrt{n!}\,2^{\frac{j-1}{2}}(i+j-3)!!(-1)^{i+\frac{i+j}{2}}}{\theta_1^{\frac{j-1}{2}}(i-1)!(n+1-i)!!},\\
-(D_{n+1,n+1} r_{n+1})_j &= -\frac{2^{\frac{j-1}{2}}(-1)^{n+1+\frac{n+1+j}{2}}(n+j-2)!!}{\theta_1^{\frac{j-1}{2}}\sqrt{n!}}.
\end{aligned}$$

(i) If n odd, j even, and $j < n$, applying Lemma A.8 and making the change of variable $l = \frac{i}{2} - 1$ gives

$$\begin{aligned}
(d_{n+1}^T R_n)_j &= (2/\theta_1)^{\frac{j-1}{2}}(-1)^{\frac{j}{2}}\sqrt{n!}\\
&\quad \times \sum_{l=0}^{\frac{n-1}{2}-1}(-1)^{l+1}\frac{(2l+1)!!}{(2l+1)!(n-1-2l)!!}\prod_{m=1}^{\frac{j-2}{2}}(2l+2m+1)\\
&= -\frac{(2/\theta_1)^{\frac{j-1}{2}}(-1)^{\frac{j}{2}}\sqrt{n!}}{2^{\frac{n-1}{2}}(\frac{n-1}{2})!}\sum_{l=0}^{\frac{n-1}{2}-1}(-1)^l\binom{\frac{n-1}{2}}{l}\prod_{m=1}^{\frac{j-2}{2}}(2l+2m+1)\\
&= -\frac{(2/\theta_1)^{\frac{j-1}{2}}(-1)^{\frac{j}{2}}\sqrt{n!}}{2^{\frac{n-1}{2}}(\frac{n-1}{2})!}\\
&\quad \times \left(\sum_{l=0}^{\frac{n-1}{2}}(-1)^l\binom{\frac{n-1}{2}}{l}\prod_{m=1}^{\frac{j-2}{2}}(2l+2m+1) - (-1)^{\frac{n-1}{2}}\prod_{m=1}^{\frac{j-2}{2}}(n+2m)\right)\\
&= -\frac{(2/\theta_1)^{\frac{j-1}{2}}(-1)^{\frac{j}{2}}\sqrt{n!}}{2^{\frac{n-1}{2}}(\frac{n-1}{2})!}\left(0 + (-1)^{\frac{n+1}{2}}\prod_{m=1}^{\frac{j-2}{2}}(n+2m)\right)
\end{aligned}$$

$$
\begin{aligned}
&= (2/\theta_1)^{\frac{j-1}{2}}(-1)^{\frac{n+j-1}{2}}\frac{\sqrt{n!}}{(n-1)!!}\prod_{m=1}^{\frac{j-2}{2}}(n+2m)\\
&= (2/\theta_1)^{\frac{j-1}{2}}(-1)^{\frac{n+j-1}{2}}\frac{(n+j-2)!!}{\sqrt{n!}}\\
&= -(D_{n+1,n+1}r_{n+1})_j.
\end{aligned}
$$

(ii) If n even, j odd, and $j < n$, similarly, applying the change of variable $l = \frac{i-1}{2}$ gives

$$
\begin{aligned}
(d_{n+1}^T R_n)_j &= -\sqrt{-1}\frac{(2/\theta_1)^{\frac{j-1}{2}}(-1)^{\frac{j}{2}}\sqrt{n!}}{2^{\frac{n}{2}}(\frac{n}{2})!}\sum_{l=0}^{\frac{n}{2}-1}(-1)^l\binom{\frac{n}{2}}{l}\prod_{m=1}^{\frac{j-1}{2}}(2l+2m-1)\\
&= \sqrt{-1}\frac{(2/\theta_1)^{\frac{j-1}{2}}(-1)^{\frac{j}{2}}\sqrt{n!}}{2^{\frac{n}{2}}(\frac{n}{2})!}(-1)^{\frac{n}{2}}\prod_{m=1}^{\frac{j-1}{2}}(n+2m-1)\\
&= \frac{(-1)^{\frac{n+j+1}{2}}(2/\theta_1)^{\frac{j-1}{2}}(n+j-2)!!}{\sqrt{n!}}\\
&= -(D_{n+1,n+1}r_{n+1})_j.
\end{aligned}
$$

(b) Proof of $B_n = 1$: (3.1) gives

$$
r_{n+1}^T d_{n+1} = \sum_{i=1,(n+i)\text{odd}}^{n}(2/\theta_1)^{\frac{n}{2}}\sqrt{n!}\frac{(-1)^{n+1+\frac{n+1+i}{2}}(n+i-2)!!}{(i-1)!(n+1-i)!!}.
$$

(i) If n odd, applying the change of variable $l = \frac{i}{2} - 1$ and using Lemma A.8,

$$
\begin{aligned}
&r_{n+1}^T d_{n+1}\\
&= (2/\theta_1)^{\frac{n}{2}}\sqrt{n!}\sum_{i=1,(i)\text{even}}^{n}\frac{(-1)^{\frac{n+1+i}{2}}(i-1)!!\prod_{m=1}^{\frac{n-1}{2}}(i+2m-1)}{(i-1)!(n+1-i)!!}\\
&= (2/\theta_1)^{\frac{n}{2}}\sqrt{n!}(-1)^{\frac{n-1}{2}}\sum_{l=0}^{\frac{n-1}{2}-1}\frac{(-1)^l(2l+1)!!}{(2l+1)!(n-1-2l)!!}\prod_{m=1}^{\frac{n-1}{2}}(2l+2m+1)\\
&= \frac{(2/\theta_1)^{\frac{n}{2}}\sqrt{n!}(-1)^{\frac{n-1}{2}}}{2^{\frac{n-1}{2}}(\frac{n-1}{2})!}\sum_{l=0}^{\frac{n-1}{2}-1}(-1)^l\binom{\frac{n-1}{2}}{l}\prod_{m=1}^{\frac{n-1}{2}}(2l+2m+1)\\
&= \frac{(2/\theta_1)^{\frac{n}{2}}\sqrt{n!}(-1)^{\frac{n-1}{2}}}{2^{\frac{n-1}{2}}(\frac{n-1}{2})!}\\
&\quad\times\left(\sum_{l=0}^{\frac{n-1}{2}}(-1)^l\binom{\frac{n-1}{2}}{l}\prod_{m=1}^{\frac{n-1}{2}}(2l+2m+1) - (-1)^{\frac{n-1}{2}}\prod_{m=1}^{\frac{n-1}{2}}(n+2m)\right)
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{(2/\theta_1)^{\frac{n}{2}}\sqrt{n!}(-1)^{\frac{n-1}{2}}}{2^{\frac{n-1}{2}}(\frac{n-1}{2})!}\left(\sum_{l=0}^{\frac{n-1}{2}}(-1)^l\binom{\frac{n-1}{2}}{l}(2l)^{\frac{n-1}{2}} - (-1)^{\frac{n-1}{2}}\prod_{m=1}^{\frac{n-1}{2}}(n+2m)\right)\\
&= \frac{(2/\theta_1)^{\frac{n}{2}}\sqrt{n!}(-1)^{\frac{n-1}{2}}}{2^{\frac{n-1}{2}}(\frac{n-1}{2})!}\left((-2)^{\frac{n-1}{2}}\left(\frac{n-1}{2}\right)! - (-1)^{\frac{n-1}{2}}\prod_{m=1}^{\frac{n-1}{2}}(n+2m)\right)\\
&= (2/\theta_1)^{\frac{n}{2}}\sqrt{n!} - \frac{(2\theta_2)^{\frac{n}{2}}}{\sqrt{n!}}(2n-1)!!\\
&= \frac{1}{D_{n+1,n+1}} - D_{n+1,n+1}R_{n+1,n+1}.
\end{aligned}
$$

(ii) If n even, similarly, applying the change of variable $l = \frac{i-1}{2}$ gives

$$
\begin{aligned}
r_{n+1}^T d_{n+1} &= \frac{(2/\theta_1)^{\frac{n}{2}}\sqrt{n!}(-1)^{\frac{n}{2}}}{2^{\frac{n}{2}}(\frac{n}{2})!}\sum_{l=0}^{\frac{n}{2}-1}(-1)^l\binom{\frac{n}{2}}{l}\prod_{m=1}^{\frac{n}{2}}(2l+2m-1)\\
&= \frac{(2/\theta_1)^{\frac{n}{2}}\sqrt{n!}(-1)^{\frac{n}{2}}}{2^{\frac{n}{2}}(\frac{n}{2})!}\left((-2)^{\frac{n}{2}}(\frac{n}{2})! - (-1)^{\frac{n}{2}}\prod_{m=1}^{\frac{n}{2}}(n+2m-1)\right)\\
&= (2/\theta_1)^{\frac{n}{2}}\sqrt{n!} - \frac{(2/\theta_1)^{\frac{n}{2}}}{\sqrt{n!}}(2n-1)!!\\
&= \frac{1}{D_{n+1,n+1}} - D_{n+1,n+1}R_{n+1,n+1}.
\end{aligned}
$$

(c) Proof of $C_n = \mathbf{0}$:

$$
\begin{aligned}
R_n^T C_n &= r_{n+1} + R_n^T d_{n+1}(d_{n+1}^T r_{n+1} + D_{n+1,n+1}R_{n+1,n+1})\\
&= r_{n+1} + R_n^T d_{n+1}\frac{B_n}{D_{n+1,n+1}}\\
&= \frac{1}{D_{n+1,n+1}}(D_{n+1,n+1}r_{n+1} + R_n^T d_{n+1})\\
&= \frac{1}{D_{n+1,n+1}}A_n^T\\
&= \mathbf{0}.
\end{aligned}
$$

Since R_n is nonsingular, $C_n = \mathbf{0}$.

A.8. Proof of a statement in section 3. We state and prove the following proposition.

Proposition A.9. *If $\{R_n\}$ is a sequence of nested positive definite matrices and we let $R_n^{-1} = D_n^T D_n$ be the reverse Cholesky decomposition of R_n^{-1}, then $\{D_n\}$ is nested.*

Proof. Letting $R_n = C_n C_n^T$ be the Cholesky decomposition of R_n, then $D_n = C_n^{-1}$ since $R_n^{-1} = D_n^T D_n$ and D_n is lower triangular as required. Since $\{R_n\}$ is nested, by construction of the Cholesky decompostion, C_n is a nested sequence of lower triangular matrices. By inspection of the relationship $D_n C_n = I_n$ when both D_n and C_n are lower triangular, it is apparent that $\{D_n\}$ is a nested sequence of matrices. ■

A.9. Proof of Theorem 3.3. Consider $k \geq i$ and $k+i$ is even (so that $k-i$ is also even and $d_{ki} \neq 0$). By Stirling's approximation, as $k \to \infty$,

$$\begin{aligned}
d_{ki}^2 &= \frac{(k-1)!}{(2/\theta_1)^{i-1}\left((i-1)!\right)^2\left((k-i)!!\right)^2} \\
&= \frac{(k-1)!}{(2/\theta_1)^{i-1}\left((i-1)!\right)^2 2^{k-i}\left(\frac{k-i}{2}!\right)^2} \\
&\sim \frac{\theta_1^{i-1}}{\left((i-1)!\right)^2 2^{k-1}} \frac{\sqrt{2\pi(k-1)}(\frac{k-1}{e})^{k-1}}{\pi(k-i)(\frac{k-i}{2e})^{k-i}} \\
&\sim \frac{\theta_1^{i-1}}{\left((i-1)!\right)^2 2^{k-1}} \sqrt{\frac{2}{\pi}}\frac{1}{\sqrt{k}} 2^{k-i} e^{1-i}(k-1)^{i-1}\left(1+\frac{i-1}{k-i}\right)^{k-i} \\
&\sim \sqrt{\frac{2}{\pi}} \frac{k^{i-\frac{3}{2}}\theta_1^{i-1}}{2^{i-1}\left((i-1)!\right)^2}.
\end{aligned}$$

Since $f(x) = x^p$, the nth order derivatives when $x=0$ are all 0 except for $n=p$, when it is $p!$. As $n \to \infty$,

$$\begin{aligned}
\hat{\theta}_0 &= \frac{1}{n}\|D(\theta_1, n)z\|^2 \\
&= \frac{(p!)^2}{n}\sum_{k=1}^{n} d_{k,p+1}^2 \\
&= \frac{(p!)^2}{n}\sum_{k=0}^{\lfloor\frac{n-p}{2}\rfloor} d_{p+1+2k,p+1}^2 \\
&\sim \frac{(p!)^2}{n}\sum_{k=0}^{\lfloor\frac{n-p}{2}\rfloor} \sqrt{\frac{2}{\pi}}\frac{(p+1+2k)^{p-\frac{1}{2}}\theta_1^p}{2^p(p!)^2} \\
&\sim \sqrt{\frac{2}{\pi}}\frac{2^{p-\frac{1}{2}}\theta_1^p}{2^p}\frac{(\lfloor\frac{n-p}{2}\rfloor)^{p+\frac{1}{2}}}{n(p+\frac{1}{2})} \\
&\sim \frac{n^{p-\frac{1}{2}}\theta_1^p}{\sqrt{2\pi}2^p(p+\frac{1}{2})},
\end{aligned}$$

where the first asymptotic equivalence follows from Lemma A.2.

A.10. Proof of a statement in section 4. We state and prove the following proposition.

Proposition A.10. *For some $m > 1$, consider a $(2m-1)\times(2m-1)$ regular grid on $[0,1]\times[0,1]$. Observations $\boldsymbol{z}$ are taken on the $m\times m$ regular subgrid. When $m = 12$, the setup is shown in Figure 3. Denote $p_{i,j}$ as the predictand at location (i,j) for some j odd, and $\hat{p}_{i,j}$ as the EBLP defined in (4.7), then*

$$\hat{p}_{i,j} = v_j^T(\hat{\theta}_2)\boldsymbol{z}_{(j-1)m+1:jm},$$

where $v_j^T(\hat{\theta}_2) = r_j(\hat{\theta}_2)R^{-1}(\hat{\theta}_2, m)$ for some $r_j(\hat{\theta}_2) \in \mathbb{R}^{1\times m}$ depending only on $\hat{\theta}_2$. That is, $\hat{p}_{i,j}$ only depends on observations on the jth column of the grid, and the range parameter estimate along columns.

Proof. Note that the covariance of $p_{i,j}$ and the observations $\boldsymbol{z}$ is

$$\text{Cov}\left(p_{i,j}, \boldsymbol{z}^T\right) = \hat{\theta}_0 R(\hat{\theta}_1, m)_{j,.} \otimes r_j(\hat{\theta}_2),$$

where $R(\hat{\theta}_1, m)_{j,.}$ is the jth row of $R(\hat{\theta}_1, m)$ and $r_j(\hat{\theta}_2)$ is the correlation of $p_{i,j}$ and observations on the jth column. Then we have

$$\begin{aligned} \hat{p}_{i,j} &= \text{Cov}\left(p_{i,j}, \boldsymbol{z}^T\right) \text{Cov}\left(\boldsymbol{z}, \boldsymbol{z}^T\right)^{-1} \boldsymbol{z} \\ &= \Big(R(\hat{\theta}_1, m)_{j,.} \otimes r_j(\hat{\theta}_2)\Big) \Big(R^{-1}(\hat{\theta}_1, m) \otimes R^{-1}(\hat{\theta}_2, m)\Big) \boldsymbol{z} \\ &= \Big(R(\hat{\theta}_1, m)_{j,.} R^{-1}(\hat{\theta}_1, m)\Big) \otimes \Big(r_j(\hat{\theta}_2) R^{-1}(\hat{\theta}_2, m)\Big) \boldsymbol{z} \\ &= \Big(\boldsymbol{e}_j^T \otimes r_j(\hat{\theta}_2) R^{-1}(\hat{\theta}_2, m)\Big) \boldsymbol{z} \\ &= v_j^T(\hat{\theta}_2) \boldsymbol{z}_{(j-1)m+1:jm}, \end{aligned}$$

where $\boldsymbol{e}_j$ is the jth standard base and $v_j^T(\hat{\theta}_2) = r_j(\hat{\theta}_2)R^{-1}(\hat{\theta}_2, m)$. ■

Acknowledgment. The authors would like to thank two anonymous referees and an associate editor for helpful suggestions that improve the quality of the paper.

REFERENCES

[1] M. Abt and W. J. Welch, *Fisher information and maximum-likelihood estimation of covariance parameters in Gaussian stochastic processes*, Canad. J. Statist., 26 (1998), pp. 127–137.

[2] G. E. Andrews, *The Theory of Partitions*, Cambridge University Press, Cambridge, UK, 1998.

[3] I. Andrianakis and P. G. Challenor, *The effect of the nugget on Gaussian process emulators of computer models*, Comput. Statist. Data Anal., 56 (2012), pp. 4215–4228.

[4] F. Bachoc, *Cross validation and maximum likelihood estimations of hyper-parameters of Gaussian processes with model misspecification*, Comput. Statist. Data Anal., 66 (2013), pp. 55–69.

[5] R. A. Beeler, *How to Count: An Introduction to Combinatorics and Its Applications*, Springer, Cham, Switzerland, 2015.

[6] F. Branin and S. Hoo, *A method for finding multiple extrema of a function of n variables*, Numer. Methods for Non-Linear Optimization/Academic, London, 1972, pp. 231–237.

[7] I. Charpentier, C. D. Cappello, and J. Utke, *Efficient higher-order derivatives of the hypergeometric function*, in Advances in Automatic Differentiation, Springer, Berlin, 2008, pp. 127–137.

[8] H. Cheng and A. Sandu, *Collocation least-squares polynomial chaos method*, in Proceedings of the 2010 Spring Simulation Multiconference, Society for Computer Simulation International, San Diego, CA, 2010, 80.

[9] A. Forrester, A. Sobester, and A. Keane, *Engineering Design via Surrogate Modeling: A Practical Guide*, Wiley, Chichester, England, 2008.

[10] R. Franke, *A Critical Comparison of Some Methods for Interpolation of Scattered Data*, Technical report NPS-53-79-003, DTIC, Fort Belvoir, VA, 1979.

[11] F. A. Gómez, C. E. Coleman-Smith, B. W. O'Shea, J. Tumlinson, and R. L. Wolpert, *Characterizing the formation history of milky way like stellar halos with model emulators*, Astrophys. J., 760 (2012), 112.

[12] H. Gould and J. Quaintance, *Double fun with double factorials*, Math. Mag., 85 (2012), pp. 177–192.

[13] A. Griewank, J. Utke, and A. Walther, *Evaluating higher derivative tensors by forward propagation of univariate Taylor series*, Math. Comput., 69 (2000), pp. 1117–1130.

[14] A. Griewank and A. Walther, *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation*, 2nd ed., Other Titles Appl. Math., SIAM, Philadelphia, 2008.

[15] B. Haaland, P. Z. Qian, *Accurate emulators for large-scale computer experiments*, Ann. Statist., 39 (2011), pp. 2974–3002.

[16] W. Huang and W. J. Welch, *Properties of Parameters in a Stochastic Process Model for Computer Experiments*. M. Math thesis, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada, 2000.

[17] C. L. Lawson and R. J. Hanson, *Solving Least Squares Problems*, classics 15, SIAM, Appl. Math. Philadelphia, 1995.

[18] Y. B. Lim, J. Sacks, W. Studden, and W. J. Welch, *Design and analysis of computer experiments when the output is highly correlated over the input space*, Canad. J. Statist., 30 (2002), pp. 109–126.

[19] W.-L. Loh and T.-K. Lam, *Estimating structured correlation matrices in smooth Gaussian random field models*, Ann. Statist., 28 (2000), pp. 880–904.

[20] J. D. Martin and T. W. Simpson, *On the use of kriging models to approximate deterministic computer models*, in ASME 2004 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, American Society of Mechanical Engineers, New York, 2004, pp. 481–492.

[21] J. Oakley, *Estimating percentiles of uncertain computer code outputs*, J. R. Stat. Soc. Ser. C. Appl. Stat., 53 (2004), pp. 83–93.

[22] O. Obrezanova, G. Csányi, J. M. Gola, and M. D. Segall, *Gaussian processes: A method for automatic QSAR modeling of ADME properties*, J. Chem. Inf. Model., 47 (2007), pp. 1847–1857.

[23] V. Picheny, T. Wagner, and D. Ginsbourger, *A benchmark of kriging-based infill criteria for noisy optimization*, Struct. Multidiscip. Optim., 48 (2013), pp. 607–626.

[24] P. Ranjan, R. Haynes, and R. Karsten, *A computationally stable approach to Gaussian process interpolation of deterministic computer simulation data*, Technometrics, 53 (2011), pp. 366–378.

[25] J. Sacks, S. B. Schiller, and W. J. Welch, *Designs for computer experiments*, Technometrics, 31 (1989), pp. 41–47.

[26] J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn, *Design and analysis of computer experiments*, Statist. Sci., 4 (1989), pp. 409–423.

[27] T. W. Simpson, T. M. Mauery, J. J. Korte, and F. Mistree, *Kriging models for global approximation in simulation-based multidisciplinary design optimization*, AIAA J., 39 (2001), pp. 2233–2241.

[28] M. L. Stein, *Interpolation of Spatial Data: Some Theory for Kriging*, Springer, New York, 1999.

[29] Q. Tang, Y. B. Lau, S. Hu, W. Yan, Y. Yang, and T. Chen, *Response surface methodology using Gaussian processes: Towards optimizing the trans-stilbene epoxidation over Co 2+–NaX catalysts*, Chem. Eng. J., 156 (2010), pp. 423–431.

[30] M. Wagner, A. Walther, and B.-J. Schaefer, *On the efficient computation of high-order derivatives for implicitly defined functions*, Comput. Phys. Commun., 181 (2010), pp. 756–764.

[31] S. Wolfram, *The MATHEMATICA Book, Version* 4, Cambridge University Press, Cambridge, UK, 1999.

[32] H. Zhang, *Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics*, J. Amer. Statist. Assoc., 99 (2004), pp. 250–261.