SIAM/ASA J. UNCERTAINTY QUANTIFICATION
Vol. 8, No. 4, pp. 1522–1547

On the Improved Rates of Convergence for Matérn-Type Kernel Ridge Regression with Application to Calibration of Computer Models[*]

Rui Tuo[†], Yan Wang[‡], and C. F. Jeff Wu[§]

Abstract. Kernel ridge regression is an important nonparametric method for estimating smooth functions. We introduce a new set of conditions under which the actual rates of convergence of the kernel ridge regression estimator under both the L_2 norm and the norm of the reproducing kernel Hilbert space exceed the standard minimax rates. An application of this theory leads to a new understanding of the Kennedy–O'Hagan approach [*J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 63 (2001), pp. 425–464] for calibrating model parameters of computer simulation. We prove that, under certain conditions, the Kennedy–O'Hagan calibration estimator with a known covariance function converges to the minimizer of the norm of the residual function in the reproducing kernel Hilbert space.

Key words. nonparametric regression, reproducing kernel Hilbert space, kriging, calibration of parameters

AMS subject classifications. 62G08, 62M30, 62M40

DOI. 10.1137/19M1304222

1. Introduction.

A major challenge in computer simulation of complex systems is to choose suitable model parameters. These parameters usually represent specific intrinsic attributes of the system. The input values of the model parameters can significantly affect the accuracy and usefulness of the computer output. When physical observations are available, one can adjust the computer model parameters so that the computer outputs match the physical data. We call this activity the calibration of computer models.

The celebrated Bayesian calibration method by Kennedy and O'Hagan [13] is one of the major and widely used approaches for the calibration of computer models. A remarkable contribution of [13] is to incorporate a "discrepancy function" to model the difference between the computer outputs and the physical process. This discrepancy does exist in most computer simulation problems because we have to resort to simplifications and unrealistic assumptions when building the computer models.

Without an informative prior, the Kennedy–O'Hagan (K-O) model is nonidentifiable because one cannot determine the model parameters and the discrepancy function simultane-

[*] Received by the editors December 4, 2019; accepted for publication (in revised form) September 2, 2020; published electronically December 8, 2020.
https://doi.org/10.1137/19M1304222
Funding: The work of the first author was supported by National Science Foundation grants DMS-1914636, CCF-1934904, and DMS-1564438, the National Center for Mathematics and Interdisciplinary Sciences in CAS, and the National Natural Science Foundation of China grants 11501551, 11271355 and 11671386. The work of the third author was supported by National Science Foundation grants DMS-1564438 and DMS-1914632.

[†] Department of Industrial and Systems Engineering, Texas A&M University, College Station, TX 77843 USA (ruituo@tamu.edu).
[‡] College of Applied Sciences, Beijing University of Technology, Beijing 100124, China (yanwang@bjut.edu.cn).
[§] School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (jeff.wu@isye.gatech.edu).

ously. Kennedy and O'Hagan [13] used a Gaussian process as a prior for the discrepancy function. Tuo and Wu [26] conducted a theoretical study on a simplified version of the K-O method when the physical data are noiseless. Under this condition, the radial basis functions approximation can be regarded as a frequentist version of Gaussian process regression. With the help of related mathematical tools, Tuo and Wu [26] identified the limit value of the K-O method as well as the rate of convergence.

A primary goal of this work is to establish an asymptotic theory for the K-O method with noisy physical data. The frequentist version of the Gaussian process regression, in this situation, is the kernel ridge regression [19]. With an improved rate of convergence for kernel ridge regression, we prove that, under certain conditions, the K-O estimator tends to the parameter value which minimizes the norm of the residual function in the reproducing kernel Hilbert space. We also present the rate of convergence of the K-O estimator. As a consequence, we relax a key and rather restrictive assumption in [26]. Tuo and Wu had to assume that the physical experiments have no random errors, which is not realistic.

There is a vast literature on the theoretical properties of ridge kernel regression. It is known that the rate of convergence of this method can be improved by imposing extra smoothness conditions on the underlying function; see, e.g., [8]. We refer to [2, 5, 11, 15] and the references therein for the recent advances in this area. In this work, we present some results on the improved rates similar to the above works. Compared with the existing ones, our model settings are closer to the practical applications in engineering and computer experiments. First, the existing methods focus on kernels constructed by a set of eigenvalues and orthonormal basis functions. This construction has, albeit mathematically general, not been widely used in practice because the computational cost is high, and an orthonormal basis may be difficult to obtain for a general input domain. In this work, we consider the widely used Matérn kernels. Second, the existing results focus on random designs, which are not usually adopted in engineering. The present work considers fixed designs satisfying some space-filling properties. Third, the existing results for the improved rates are not sufficient to develop an asymptotic theory for the K-O method. We obtain a strengthened version of the improved rates, which lead to the desired asymptotic theory for calibration. It is worth noting that the mathematical treatments in this paper differ from those in the works mentioned above, and our work provides some new insight on kernel ridge regression.

This article is organized as follows. In section 2, we introduce some background of this work and present the improved rates of convergence for kernel ridge regression. In section 3, we establish an asymptotic theory for the K-O calibration estimators. In section 4, we validate our theoretical assertions with two numerical studies. Concluding remarks are made in section 5. Appendix A contains the long proofs in this article.

2. Improved rates for kernel ridge regression.

In this section, we discuss the mathematical tool and our results on the improved rates of convergence for kernel ridge regression.

2.1. Overview.

Consider a nonparametric regression model

$$y_i = f(x_i) + e_i, \tag{2.1}$$

where f is a smooth function whose domain of definition Ω is a convex and compact subset of

$\mathbb{R}^d$ and e_i's are independent and identically distributed random sequence with mean zero and finite variance. The problem of interest is to recover f from the data $(x_i, y_i), i = 1, \ldots, n$. Kernel ridge regression is one of the important methods to deal with this problem. This method has been widely used in statistics and machine learning [19]. It also has close relationships with classic kernel-based regression methods like smoothing splines or thin-plate splines [31].

Suppose f lies in the Sobolev space $H^m(\Omega)$ with $m > d/2$. By choosing a kernel function with m degree of smoothness, the kernel ridge regression, as defined in (2.18), can reach the standard rates of convergence

$$\|\hat{f}_n - f\|_{L_2(\Omega)} = O_p(n^{-\frac{m}{2m+d}}), \tag{2.2}$$

$$\|\hat{f}_n - f\|_{H^m(\Omega)} = O_p(1), \tag{2.3}$$

where $\|\cdot\|_{L_2(\Omega)}$ and $\|\cdot\|_{H^m(\Omega)}$ denote the corresponding L_2 and Sobolev norm, respectively. See, for example, [8, 29] for details. These rates are known to be the minimax rates in the current context [23]. That is, these rates are in general not improvable.

From (2.2), we can see that the convergence rate depends on the smoothness of the underlying function. If we assume a higher smoothness condition for f, we can achieve a better rate by applying the kernel ridge regression with a kernel function as smooth as f. However, the smoothness of most practical underlying functions is unknown. Therefore, usually we cannot identify the optimal kernel functions. In practice, kernel functions with relatively low smoothness are frequently used. For instance, in spatial statistics and computer experiments, Matérn kernels (see section 2.2 for the definition) with smoothness parameter 3/2 or 5/2 are widely used [22, 18]. In this article, we show that if the underlying function f is smoother than the kernel function, the rate of convergence of the kernel ridge regression may be improved. Specifically, we identify a dense subset $S \subset H^m(\Omega)$ in such a way that if $f \in S$, we can reach the improved rates of convergence

$$\|\hat{f}_n - f\|_{L_2(\Omega)} = O_p(n^{-\frac{2m}{4m+d}}), \tag{2.4}$$

$$\|\hat{f}_n - f\|_{H^m(\Omega)} = O_p(n^{-\frac{m}{4m+d}}). \tag{2.5}$$

Clearly, there is a substantial improvement from (2.3) to (2.5) because (2.3) does not entail convergence. We also prove an improved rate of convergence under the norm of the reproducing kernel Hilbert space generated by the kernel function, denoted by $\mathcal{N}$, as

$$\|\hat{f}_n - f\|_{\mathcal{N}} = O_p(n^{-\frac{m}{4m+d}}). \tag{2.6}$$

2.2. Reproducing kernel Hilbert spaces.

Our study will employ the reproducing kernel Hilbert spaces (also called the native spaces) as the mathematical tool.

We consider functions defined on $\Omega \subset \mathbb{R}^d$ and adopt Assumption 2.1 throughout this article.

Assumption 2.1. The set $\Omega \subset \mathbb{R}^d$ satisfies the following conditions:

1. Ω is compact.
2. The interior of Ω, denoted as Ω°, is nonempty and connected. Besides, Ω is the closure of Ω°.

3. Ω° is convex.[1]

Let Ω be a subset of $\mathbb{R}^d$.

Assume that $K : \Omega \times \Omega \to \mathbb{R}$ is a symmetric positive definite kernel. Define the linear space

$$F_K(\Omega) = \left\{ \sum_{i=1}^{n} \beta_i K(\cdot, x_i) : \beta_i \in \mathbb{R}, x_i \in \Omega, n \in \mathbb{N} \right\}, \tag{2.7}$$

and equip this space with the bilinear form

$$\left\langle \sum_{i=1}^{n} \beta_i K(\cdot, x_i), \sum_{j=1}^{m} \gamma_j K(\cdot, x_j') \right\rangle_K := \sum_{i=1}^{n} \sum_{j=1}^{m} \beta_i \gamma_j K(x_i, x_j'). \tag{2.8}$$

Then the *reproducing kernel Hilbert space* $\mathcal{N}_K(\Omega)$ generated by the kernel function K is defined as the closure of $F_K(\Omega)$ under the inner product $\langle \cdot, \cdot \rangle_K$, and the norm of $\mathcal{N}_K(\Omega)$ is $\|f\|_{\mathcal{N}_K(\Omega)} = \sqrt{\langle f, f \rangle_{\mathcal{N}_K(\Omega)}}$, where $\langle \cdot, \cdot \rangle_{\mathcal{N}_K(\Omega)}$ is induced by $\langle \cdot, \cdot \rangle_K$. More detail about reproducing kernel Hilbert space can be found in [31, 32].

In this work, we suppose the kernel function K is stationary; i.e., $K(x, y)$ depends only on $x - y$. We denote $K(x, y) =: \Phi(x - y)$ and also denote the reproducing kernel Hilbert space $\mathcal{N}_K(\Omega)$ by $\mathcal{N}_\Phi(\Omega)$. Specifically, we focus on the Matérn kernel function [18, 22] defined by

$$\Phi(x; \nu, \phi) = \frac{1}{\Gamma(\nu) 2^{\nu-1}} (2\sqrt{\nu}\phi|x|)^\nu K_\nu(2\sqrt{\nu}\phi|x|), \tag{2.9}$$

where K_ν is the modified Bessel function of the second kind, ν and ϕ are *fixed* parameters, and $|\cdot|$ is Euclidean distance meaning that (2.9) is an isotropic kernel. In (2.9), ϕ is a scale parameter, and ν is often called the smoothness parameter because it is related to the smoothness of the Gaussian processes associated with this kernel (covariance) function.

The smoothness of the kernel Φ is somehow inherited by the reproducing kernel Hilbert space $\mathcal{N}_\Phi(\Omega)$ [32, Theorem 10.45]. Specifically, if Φ is a Matérn kernel in (2.9), $\mathcal{N}_\Phi(\Omega)$ is equal to the (fractional) Sobolev space $H^{\nu+d/2}(\Omega)$,[2] with equivalent norms. See also Corollary 1 of [26]. Here we see that the smoothness parameter ν is also related to the smoothness of the Sobolev space.

2.3. An improved rate in scattered data approximation.

The current work is partially inspired by a result (section 11.5 of [32]) in scattered data approximation, which gives an improved rate of convergence for radial basis function interpolation. In this section, we briefly review this result.

[1] Technically, this convexity assumption can be relaxed to that Ω° has a Lipschitz boundary and satisfies a uniform cone condition. We refer to [1] for the detailed description of the aforementioned concepts.

[2] As pointed out by a reviewer, Sobolev spaces are conventionally defined only on open subsets of $\mathbb{R}^d$. Here we regard the functions in $H^{\nu+d/2}(\Omega)$ as the natural extensions of the functions in $H^{\nu+d/2}(\Omega^\circ)$, as these functions are continuous according to the Sobolev embedding theorem [1].

Let f be an underlying deterministic function. Suppose we have observed the function values of f over some scattered points $X = \{x_1, \ldots, x_n\}$. Then an interpolant of f is constructed by solving the optimization problem

$$\begin{aligned} &\min \|g\|_{\mathcal{N}_\Phi(\Omega)} \\ &\text{s.t. } g(x_i) = f(x_i) \text{ for } i = 1, \ldots, n. \end{aligned} \tag{2.10}$$

We denote this interpolant by $s_{f,X}$, which is commonly known as the radial basis function interpolant. Formula (2.10) is the limit case of the kernel ridge regression estimator introduced later in (2.18), with $\lambda_n \downarrow 0$.

The error estimate for radial basis function interpolant is well established in the literature. See [32]. Suppose $\mathcal{N}_\Phi(\Omega)$ is continuously embedded into a (fractional) Sobolev space $H^m(\Omega)$ and the design X is quasi-uniform (see section 2.4 for the formal definition). Then a standard error bound is

$$\|f - s_{f,X}\|_{L_2(\Omega)} \leq Cn^{-m/d}\|f - s_{f,X}\|_{\mathcal{N}_\Phi(\Omega)} \tag{2.11}$$

for some constant C independent of f, n and the choice of a quasi-uniform design. It is worth noting that the finiteness of the right-hand side of (2.11) requires $f \in \mathcal{N}_\Phi(\Omega)$. Radial basis function interpolation satisfies the orthogonality condition

$$\langle f - s_{f,X}, s_{f,X}\rangle_{\mathcal{N}_\Phi(\Omega)} = 0. \tag{2.12}$$

Thus, we have the Pythagorean identity

$$\|f - s_{f,X}\|^2_{\mathcal{N}_\Phi(\Omega)} + \|s_{f,X}\|^2_{\mathcal{N}_\Phi(\Omega)} = \|f\|^2_{\mathcal{N}_\Phi(\Omega)} = \text{constant},$$

which implies $\|f - s_{f,X}\|_{\mathcal{N}_\Phi(\Omega)} = O(1)$ as n tends to infinity. Therefore, $\|f - s_{f,X}\|_{L_2(\Omega)}$ decays at least with the order $O(n^{-m/d})$ according to (2.11).

To pursue an improved rate of convergence, one may ask whether $\|f - s_{f,X}\|_{\mathcal{N}_\Phi(\Omega)} = o(1)$. Although this does not hold generally [7, 17], we do have an improved rate if there exists $v \in L_2(\Omega)$, so that

$$f(x) = \int_\Omega \Phi(x - t)v(t)dt. \tag{2.13}$$

Proposition 10.28 of [32] shows that functions with the form (2.13) are a dense subset of $\mathcal{N}_\Phi(\Omega)$. It shows that in this case, for any $g \in \mathcal{N}_\Phi(\Omega)$,

$$\langle f, g\rangle_{\mathcal{N}_\Phi(\Omega)} = \langle v, g\rangle_{L_2(\Omega)}. \tag{2.14}$$

Combining (2.11), (2.12), and (2.14) and applying the Cauchy–Schwarz inequality yields

$$\begin{aligned} \|f - s_{f,X}\|^2_{L_2(\Omega)} &\leq C^2n^{-2m/d}\|f - s_{f,X}\|^2_{\mathcal{N}_\Phi(\Omega)} \qquad (2.15)\\ &= C^2n^{-2m/d}\langle f - s_{f,X}, f\rangle_{\mathcal{N}_\Phi(\Omega)} \\ &= C^2n^{-2m/d}\langle f - s_{f,X}, v\rangle_{L_2(\Omega)} \\ &\leq C^2n^{-2m/d}\|f - s_{f,X}\|_{L_2(\Omega)}\|v\|_{L_2(\Omega)}. \qquad (2.16) \end{aligned}$$

Canceling $\|f - s_{f,X}\|_{L_2(\Omega)}$ from both sides of (2.16) and comparing (2.15) and (2.16) yields the improved error bounds

$$\begin{aligned}\|f - s_{f,X}\|_{L_2(\Omega)} &\leq C^2 n^{-2m/d}\|v\|_{L_2(\Omega)},\\ \|f - s_{f,X}\|_{\mathcal{N}_\Phi(\Omega)} &\leq C n^{-m/d}\|v\|_{L_2(\Omega)}.\end{aligned}$$

In section 2.4, we will use the same assumption (2.13) to derive improved rates of convergence for kernel ridge regression.

If a Matérn kernel (2.9) is used, (2.13) is equivalent to imposing a certain higher-order smoothness condition. Before introducing the condition, we discuss the extension theorem of reproducing kernel Hilbert spaces.

Proposition 2.2. *Each $h \in \mathcal{N}_\Phi(\Omega)$ has an extension $h_e \in \mathcal{N}_\Phi(\mathbb{R}^d)$, which defines an isometric map from $\mathcal{N}_\Phi(\Omega)$ to $\mathcal{N}_\Phi(\mathbb{R}^d)$. In other words, $h_e|_\Omega = h$, and $\langle h_e, h'_e\rangle_{\mathcal{N}_\Phi(\mathbb{R}^d)} = \langle h, h'\rangle_{\mathcal{N}_\Phi(\Omega)}$ for all $h, h' \in \mathcal{N}_\Phi(\Omega)$, where $h_e|_\Omega$ denotes the restriction of h_e on the region Ω.*

The main steps in proving Proposition 2.2 are as follows. First, we consider the map from $F_\Phi(\Omega)$ defined in (2.7) to $F_\Phi(\mathbb{R}^d)$ given by

$$\sum_{i=1}^{n}\beta_i\Phi(x - x_i), x \in \Omega \mapsto \sum_{i=1}^{n}\beta_i\Phi(x - x_i), x \in \mathbb{R}^d,$$

which defines an extension for each function in $F_\Phi(\Omega)$. Clearly, this map preserves the inner product (2.8). Next, by using some functional analysis machinery such as taking Cauchy sequences, we can extend the domain of definition of this map from $F_\Phi(\Omega)$ to its closure, the Hilbert space $\mathcal{N}_\Phi(\Omega)$, and the extended map is also isometric. We refer the reader to Theorem 10.46 of [32] for details of the proof.

Theorem 2.3 gives an equivalent statement of the condition (2.13).

Theorem 2.3. *Suppose Φ is a Matérn kernel (2.9) with smoothness parameter $\nu = m - d/2$ and $f \in \mathcal{N}_\Phi(\Omega)$. Then the integral equation*

$$f(x) = \int_\Omega \Phi(x - t)v(t)dt \tag{2.17}$$

has a solution $v \in L_2(\Omega)$ if and only if the extended function $f_e \in H^{2m}(\mathbb{R}^d)$.

To maintain flow of the paper, all the long proofs are given in Appendix A.

Remark 2.4. Obviously, $f_e \in H^{2m}(\mathbb{R}^d)$ implies $f \in H^{2m}(\Omega)$. However, the converse is not necessarily true. The stronger condition $f_e \in H^{2m}(\mathbb{R}^d)$ essentially requires the smoothness of the function across the boundary of Ω. To illustrate this point, we consider a simple example. Suppose $\Omega = [-1, 1], \nu = 1/2, \phi = 1$. Then the Matérn kernel becomes $\Phi(x) = e^{-|x|}$. Let $f(x) = e^{1-x}$, $x \in [-1, 1]$. Then $f \in C^\infty[-1, 1]$. However, since $f(x) = \Phi(x-1)$ for $x \in [-1, 1]$, according to the discussion after Proposition 2.2, we have $f_e = \Phi(x - 1) = e^{-|x-1|}$, which is not in $H^2(\mathbb{R})$.

2.4. Rates of convergence for kernel ridge regression. In this section, we return to model (2.1). The goal is to estimate the underlying function f from the data $\{(x_i, y_i)\}_{i=1}^n$. As in section 2.2, we choose a positive definite kernel function Φ. The *kernel ridge regression* estimator of f is defined as

$$\hat{f}_n = \operatorname*{argmin}_{g\in\mathcal{N}_\Phi(\Omega)} \frac{1}{n}\sum_{i=1}^n (y_i - g(x_i))^2 + \lambda_n \|g\|^2_{\mathcal{N}_\Phi(\Omega)}, \tag{2.18}$$

where $\lambda_n > 0$ is a tuning parameter to balance the bias and the variance.

The optimization problem (2.18) can be solved analytically. With the help of the representer theorem [31, 21], we find that $\hat{f}$ has the form

$$\hat{f}_n(x) = \sum_{i=1}^n c_i \Phi(x - x_i), \tag{2.19}$$

where c_i's are undetermined coefficients. Substituting (2.19) into (2.18) and invoking (2.8), the estimation becomes a ridge regression problem weighted by the kernel matrix, and this is where the name “kernel ridge regression” comes from. After some calculations, we can find that the vector $c = (c_1, \dots, c_n)^T$ is given by

$$c = (\mathbf{\Phi} + n\lambda_n I_n)^{-1} Y, \tag{2.20}$$

where $\mathbf{\Phi} = (\Phi(x_i, x_j))_{ij}$, $Y = (y_1, \dots, y_n)^T$, and I_n is the identity matrix.

2.4.1. Standard rates of convergence. In this paper, we are interested in the conditions that ensure a consistent estimation for f using the kernel ridge regression and the rate of convergence. First, we review the existing results and the standard proof.

Throughout the paper, we assume that the reproducing kernel Hilbert space $\mathcal{N}_\Phi(\Omega)$ is equal to some (fractional) Sobolev space $H^m(\Omega)$ with equivalent norms for some $m > d/2$. Recall that if Φ is a Matérn kernel in (2.9), $\mathcal{N}_\Phi(\Omega)$ is $H^{\nu+d/2}(\Omega)$. We also assume that the random error e_i's are sub-Gaussian in the sense that there exists universal constants $K, \sigma_0 > 0$ such that

$$\mathbb{P}(|e_i| > t) \le K e^{-t^2/\sigma_0^2} \tag{2.21}$$

holds for all $t > 0$. This condition can be relaxed, but the technical details will become more involved, and we do not pursue such a treatment here.

Define the *empirical seminorm* by

$$\|f\|_n^2 = \frac{1}{n}\sum_{i=1}^n f^2(x_i),$$

and write $a \vee b = \max\{a, b\}$. The standard convergence results are stated in Proposition 2.5.

Proposition 2.5. *Suppose $f \in H^m(\Omega)$ and $\lambda_n^{-1} = O(n^{\frac{2m}{2m+d}})$. Then the estimator $\hat{f}_n$ given by* (2.18) *satisfies*

$$\begin{aligned} \|\hat{f}_n - f\|_n &= O_p(\lambda_n^{1/2} \vee n^{-\frac{1}{2}} \lambda_n^{-\frac{d}{4m}}), \\ \|\hat{f}_n\|_{\mathcal{N}_\Phi(\Omega)} &= O_p(1 \vee n^{-\frac{1}{2}} \lambda_n^{-\frac{2m+d}{4m}}). \end{aligned} \tag{2.22}$$

Because the main idea of proving Proposition 2.5 is also useful in establishing the improved rate of convergence, we give a sketch of proof for Proposition 2.5. A detailed version can be found in Theorem 10.2 of [29].

Sketch of proof for Proposition 2.5. The optimization condition (2.18) implies the basic inequality

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^n (y_i - \hat{f}_n(x_i))^2 + \lambda_n \|\hat{f}_n\|^2_{\mathcal{N}_\Phi(\Omega)} \\ \le &\frac{1}{n}\sum_{i=1}^n (y_i - f(x_i))^2 + \lambda_n \|f\|^2_{\mathcal{N}_\Phi(\Omega)}. \end{aligned} \tag{2.23}$$

After some rearrangement, we can see that (2.23) is equivalent to

$$\|\hat{f}_n - f\|_n^2 + \lambda_n \|\hat{f}_n\|^2_{\mathcal{N}_\Phi(\Omega)} \le 2\langle e, \hat{f}_n - f\rangle_n + \lambda_n \|f\|^2_{\mathcal{N}_\Phi(\Omega)}, \tag{2.24}$$

where

$$\langle e, g\rangle_n := \frac{1}{n}\sum_{i=1}^n e_i g_i(x_i). \tag{2.25}$$

It follows from a standard result in empirical process theory that

$$\frac{\langle e, \hat{f}_n - f\rangle_n}{\|\hat{f}_n - f\|_n^{1-\frac{d}{2m}} \|\hat{f}_n - f\|^{\frac{d}{2m}}_{\mathcal{N}_\Phi(\Omega)}} = O_p(n^{-1/2}); \tag{2.26}$$

see Lemma A.1 in Appendix A for details. With some elementary algebraic calculations, also seeing Lemma A.2 and the proof of Theorem 3.2 in Appendix A, it is not hard to find that (2.24) and (2.26) yield the desired results. ■

The smoothing parameter λ_n makes a trade-off between the bias and the variance of the estimator. If λ_n decays no faster than $O_p(n^{-\frac{2m}{2m+d}})$, the bias term dominates the variance term, and the rate of convergence under the empirical seminorm is $O_p(\lambda_n^{1/2})$. On the other hand, if λ_n decays faster than $O_p(n^{-\frac{2m}{2m+d}})$, the variance term dominates the bias term, and the rate of convergence under the L_2 norm is $O_p(n^{-\frac{1}{2}}\lambda_n^{-\frac{d}{4m}})$. In this case, $\|\hat{f}_n - f\|_{\mathcal{N}_\Phi(\Omega)}$ may go to infinity. Therefore, to reach the best rates of convergence, one needs to balance the bias and the variance. By choosing $\lambda_n \sim n^{-\frac{2m}{2m+d}}$, one can obtain the best rates

$$\begin{aligned} \|\hat{f}_n - f\|_n &= O_p(n^{-\frac{m}{2m+d}}), \\ \|\hat{f}_n - f\|_{\mathcal{N}_\Phi(\Omega)} &= O_p(1). \end{aligned}$$

An important question is whether the convergence results in (2.22) imply a convergence under a more commonly used norm, like the L_2 norm. Such a result relies on whether the design points $\{x_1, \ldots, x_n\}$ are allocated in a space-filling manner. To address this point, we introduce the concept of quasi-uniformity [3, 28].

Definition 2.6. *For a set of design points $\{x_1,\dots,x_n\}\subset\Omega$, define its fill distance as*

$$h_n=\max_{x\in\Omega}\min_i\|x-x_i\|$$

and its separation distance as

$$q_n=\min_{i\neq j}\|x_j-x_i\|,$$

where $\|\cdot\|$ denotes the Euclidean distance. Call a design sequence $x_1,\dots,x_n,\dots$ quasi-uniform if there exists a universal constant $B>0$ such that

$$h_n/q_n\le B \tag{2.27}$$

holds for all $n>1$.

Remark 2.7. Most commonly used space-filling designs are quasi-uniform, like regular lattices and Sobol sequences [18].

For any $\{x_1,\dots,x_n\}\subset\Omega$, the balls centered at x_i's with radius $q_n/2$ are disjoint. By comparing the volume of these balls and that of Ω, we find that the inequality

$$nV_d(q_n/2)^d\le 2Vol(\Omega) \tag{2.28}$$

holds if $q_n=O(n^{-1/d})$, where V_d denotes the volume of d-dimensional unit ball and $Vol(\Omega)$ denotes the volume of Ω. If $\{x_1,\dots,x_n\}$ also satisfies (2.27), (2.28) yields

$$h_n\le 2B\left(\frac{2Vol(\Omega)}{V_d}\right)^{1/d}n^{-1/d}=:B'n^{-1/d}. \tag{2.29}$$

Under certain conditions, the empirical seminorm and the L_2 norm are equivalent. The following Proposition comes from Theorems 3.3 and 3.4 of [28].

Proposition 2.8. *Suppose the design sequence is quasi-uniform. Then there exist constants C_1 and C_2 (depending only on m, d, Ω, and B) and h_0 such that, for any $g\in H^m(\Omega)$ and $h_n\le h_0$, we have*

$$\begin{aligned}\|g\|^2_{L_2(\Omega)}&\le C_1\left\{\|g\|_n^2+h_n^{2m}\|g\|^2_{H^m(\Omega)}\right\},\\ \|g\|_n^2&\le C_2\left\{\|g\|^2_{L_2(\Omega)}+h_n^{2m}\|g\|^2_{H^m(\Omega)}\right\}.\end{aligned} \tag{2.30}$$

Corollary 2.9 gives the standard results for the rates of convergence of ridge kernel regression, which is a direct consequence of Proposition 2.5, (2.29), and Proposition 2.8.

Corollary 2.9. *Under the condition of Proposition 2.5, suppose the design sequence is quasi-uniform. Then the estimator $\hat f_n$ given by (2.18) satisfies*

$$\begin{aligned}\|\hat f_n-f\|_{L_2(\Omega)}&=O_p(\lambda_n^{1/2}\vee n^{-\frac12}\lambda_n^{-\frac{d}{4m}}),\\ \|\hat f_n\|_{\mathcal{N}_\Phi(\Omega)}&=O_p(1\vee n^{-\frac12}\lambda_n^{-\frac{2m+d}{4m}}).\end{aligned} \tag{2.31}$$

2.4.2. Improved rates of convergence. We can regard the rates of convergence (2.31) as a stochastic version of the error bound (2.11). They are both standard convergence results under their respective settings. In view of the improved rate of convergence in interpolation discussed in section 2.3, we also expect an improved rate of convergence for the regression problem (2.1) by imposing the same assumption that there exists $v \in L_2(\Omega)$ so that (2.13) holds.

Now we give more details about the intuition of why improved rates of convergence can be obtained. Note the identity

$$\|f\|^2_{\mathcal{N}_\Phi(\Omega)} - \|\hat{f}_n\|^2_{\mathcal{N}_\Phi(\Omega)} = 2\langle f, f - \hat{f}_n\rangle_{\mathcal{N}_\Phi(\Omega)} - \|f - \hat{f}_n\|^2_{\mathcal{N}_\Phi(\Omega)}, \tag{2.32}$$

which, together with the basic inequality (2.24), yields

$$\begin{aligned}&\|\hat{f}_n - f\|^2_n + \lambda_n\|\hat{f}_n - f\|^2_{\mathcal{N}_\Phi(\Omega)}\\ &\le 2\langle e, \hat{f}_n - f\rangle_n + 2\lambda_n\langle f, f - \hat{f}_n\rangle_{\mathcal{N}_\Phi(\Omega)}.\end{aligned} \tag{2.33}$$

Invoking identity (2.14) and the Cauchy–Schwarz inequality, we obtain

$$\langle f, f - \hat{f}_n\rangle_{\mathcal{N}_\Phi(\Omega)} = \langle v, f - \hat{f}_n\rangle_{L_2(\Omega)} \le \|v\|_{L_2(\Omega)}\|\hat{f}_n - f\|_{L_2(\Omega)},$$

which, together with (2.33), implies

$$\begin{aligned}&\|\hat{f}_n - f\|^2_n + \lambda_n\|\hat{f}_n - f\|^2_{\mathcal{N}_\Phi(\Omega)}\\ &\le 2\langle e, \hat{f}_n - f\rangle_n + 2\lambda_n\|v\|_{L_2(\Omega)}\|\hat{f}_n - f\|_{L_2(\Omega)}.\end{aligned} \tag{2.34}$$

We call (2.34) the *improved basic inequality* because it gives a refined version of the basic inequality (2.24). Compared to (2.24), the right-hand side of (2.34) is significantly deflated because $\|\hat{f}_n\|^2_{\mathcal{N}_\Phi(\Omega)}$ in (2.24) has the order $O_p(1)$ according to Proposition 2.5, while in (2.34), $\|\hat{f}_n - f\|_{L_2(\Omega)} = o_p(1)$ if $\lambda_n = o_p(1)$. This explains why we can expect improved rates of convergence for the two terms on the left-hand side of (2.34). These rates can be obtained by employing additional algebraic calculations. We summarize our findings in Proposition 2.10.

Proposition 2.10. *Suppose there exists $v \in L_2(\Omega)$ such that*

$$f(x) = \int_\Omega \Phi(x - t)v(t)dt. \tag{2.35}$$

Moreover, suppose the sequence of design points is quasi-uniform and the random error e_i's are sub-Gaussian satisfying (2.21). Then

$$\begin{aligned}\|\hat{f}_n - f\|_n &= O_p\left(\lambda_n \vee n^{-\frac{1}{2}}\lambda_n^{-\frac{d}{4m}}\right),\\ \|\hat{f}_n - f\|_{\mathcal{N}_\Phi(\Omega)} &= O_p\left(\lambda_n^{1/2} \vee n^{-\frac{1}{2}}\lambda_n^{-\frac{2m+d}{4m}}\right).\end{aligned} \tag{2.36}$$

Proof. This result is a special case of Corollary 3.3 in section 3. ■

Remark 2.11. The improved rates in Proposition 2.10 are known; see [2, 8, 5, 11, 15] and the references therein. Despite these known rates, the conditions in Proposition 2.10 differ from these works. These works focus on kernels represented by eigenvalues and eigenfunctions and random designs. We consider Matérn kernels and quasi-uniform designs, which are widely used in engineering and computer experiment applications. Also, the mathematical tools used here are different from those in the above works, and our analysis yields a stronger result, given in Theorem 3.2, which leads to an asymptotic theory for the K-O calibration estimator.

Corollary 2.12. *Under the conditions of Proposition* 2.10*, we can apply Proposition* 2.8 *to derive*

$$\|\hat{f}_n - f\|_{L_2(\Omega)} = O_p\left(\lambda_n \vee n^{-\frac{1}{2}}\lambda_n^{-\frac{d}{4m}}\right).$$

3. Calibration of computer models.

In this section, we use the improved convergence theory established in section 2.4 to study the asymptotic theory for the K-O method for the calibration of computer models.

In computer experiments, calibration is the activity of identifying the computer model parameters by matching the computer and physical outputs. Consider a physical experiment, with a vector of input variable denoted as x. To reduce the cost of the physical experiment, researchers often conduct a computer simulation to mimic the physical system as well. Usually, the computer code input consists of the physical input x and model parameters θ. The model parameters are not observed in the physical experiment; they commonly represent certain intrinsic attributes of the system. Here we consider only deterministic computer experiments; i.e., the computer output is a deterministic function of the inputs, denoted by $y^s(x,\theta)$.

In the K-O approach, the physical experimental data are modeled as

$$y_i = \xi(x_i) + e_i, i = 1, \dots, n, \tag{3.1}$$

where ξ is an underlying function called the true process, x_i's are fixed input points, and e_i's are independent and identically distributed random error with mean zero.

Because the computer models are built under inevitable simplification and approximation, their outputs cannot coincide with the true process; [13] used the following model to link these functions:

$$\xi(x) = y^s(x,\theta_0) + \delta(x), \tag{3.2}$$

where θ_0 is the "optimal choice" of the model parameter and δ denotes the discrepancy function. The model (3.2) is clearly nonidentifiable because both θ_0 and δ are unknown. We refer the reader to [16, 24, 25, 26, 27] for related theoretical discussions regarding the identifiability. Kennedy and O'Hagan [13] proposed to impose a Gaussian process prior on δ to facilitate the estimation of θ_0.

Given the widespread use of the K-O method in computer experiments and related scientific and engineering problems, understanding the asymptotic properties of this method is of interest. In this work, we *do not* assume that δ (or (ξ, y^s)) is random; that is, we regard the Gaussian process modeling technique in the K-O's approach only as a computational method. This nonrandom model setting can be justified as follows. Because the computer code is deterministic, y^s should be nonrandom. Also, the true process ξ is usually presumed as nonrandom in industrial statistics, for example, in the response surface methodology [33]. The main objective of this section is to study the asymptotic behavior of the K-O calibration estimator under the above deterministic setting. Our findings in the section *should not* be interpreted under the usual framework of Gaussian process regression, where the underlying function is truly random.

3.1. A frequentist version of the K-O approach.

We consider estimating θ by maximizing the following "likelihood function":

$$L(\theta,\sigma^2,\tau^2)=\det(\sigma^2\mathbf{\Phi}+\tau^2 I)^{-1/2} \times \exp\left\{-\frac{1}{2}(Y-y^s(X;\theta))^T(\sigma^2\mathbf{\Phi}+\tau^2 I)^{-1}(Y-y^s(X;\theta))\right\}, \tag{3.3}$$

where $\mathbf{\Phi}=(\Phi(x_i,x_j))_{ij}$, $Y=(y_1,\dots,y_n)^T$, $y^s(X;\theta)=(y^s(x_1;\theta),\dots,y^s(x_n;\theta))^T$, and I denotes the identity matrix.

Under some extra conditions, (3.3) is indeed the likelihood function induced by the K-O approach. First, we suppose that e_i's in (3.1) follow the normal distribution $N(0,\tau^2)$,[3] and we impose a Gaussian process prior on y^s. Second, suppose this Gaussian process has mean zero and covariance function $\sigma^2\Phi(\cdot,\cdot)$. Here we assume that Φ is given. Then it is easily shown that the likelihood function of (θ,σ^2,τ^2) is (3.3).

The maximum likelihood estimators (MLEs) of σ^2 and τ^2 in (3.3) do not have explicit expressions. To ease the mathematical treatments, we denote $\lambda=\tau^2/(n\sigma^2)$ in a non–data-driven manner. We will show that, a deterministic choice of λ (depending on n) can sufficiently lead to a desired asymptotic theory. Once λ is given, we have the following simplified expression of $\hat{\theta}$:

$$\hat{\theta}=\underset{\theta\in\Theta}{\operatorname{argmin}}(Y-y^s(X;\theta))^T(\mathbf{\Phi}+n\lambda I)^{-1}(Y-y^s(X;\theta)). \tag{3.4}$$

Our goal is to develop an asymptotic theory for $\hat{\theta}$ under the assumption that ξ and y^s are *deterministic* functions. We call $\hat{\theta}$ the *frequentist estimator of the K-O approach*. Of course, we adopt a totally different model setting compared with [13]. Computationally, the two methods are also different in the following aspects:

1. In [13], prior distributions are imposed on the parameters θ,σ^2,τ^2 and possibly the hyperparameters associated with Φ. In this work, we do not impose those distributions. Also, we do not introduce extra hyperparameters on the kernel Φ.
2. In [13], Bayesian analysis is conducted by calculating the posterior distribution. In this work, we focus on the MLE.
3. In [13], both σ^2 and τ^2 are estimated from the data. In this work, we choose $\lambda=\tau^2/(n\sigma^2)$ in a non–data-driven manner to facilitate our mathematical analysis.
4. In [13], the computer model can be expensive to run, so that a surrogate model is introduced to reconstruct y^s. In this work, we assume that y^s is a known function. This assumption is reasonable when the computer model is inexpensive.

[3] In our theoretical analysis in Theorems 3.2–3.4, we relax this assumption by incorporating sub-Gaussian noise.

3.2. Asymptotic theory. The MLE estimator $\hat{\theta}$ in (3.4) has a close relationship with the kernel ridge regression discussed in section 2.4. The following proposition is the same as Lemma 2.1 of [10].

Proposition 3.1. *The MLE estimator $\hat{\theta}$ and the estimator of discrepancy function $\hat{\delta}$ can be expressed as the estimator of the kernel ridge regression as follows:*

$$(\hat{\theta}, \hat{\delta}) = \operatorname*{argmin}_{\theta\in\Theta, \delta\in\mathcal{N}_\Phi(\Omega)} l(\theta, \delta),$$

where

$$l(\theta,\delta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - y^s(x_i;\theta) - \delta(x_i))^2 + \lambda\|\delta\|^2_{\mathcal{N}_\Phi(\Omega)}. \tag{3.5}$$

Define

$$\zeta^\theta(x) = \xi(x) - y^s(x;\theta), \zeta_i^\theta = y_i - y^s(x_i;\theta).$$

For each $\theta \in \Theta$, denote

$$\hat{\zeta}^\theta = \operatorname*{argmin}_{g\in\mathcal{N}_\Phi(\Omega)} \frac{1}{n}\sum_{i=1}^{n}(\zeta_i^\theta - g(x_i))^2 + \lambda\|g\|^2_{\mathcal{N}_\Phi(\Omega)},$$

which is the kernel ridge regression estimator for ζ^θ. Then $\hat{\theta}$ can be represented as

$$\begin{aligned}\hat{\theta} &= \operatorname*{argmin}_{\theta\in\Theta} l(\theta, \hat{\zeta}^\theta)\\ &= \operatorname*{argmin}_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n}(\zeta_i^\theta - \hat{\zeta}^\theta(x_i))^2 + \lambda\|\hat{\zeta}^\theta\|^2_{\mathcal{N}_\Phi(\Omega)}.\end{aligned} \tag{3.6}$$

To employ the theory developed in section 2.4, we assume that ζ^θ lies in $\mathcal{N}_\Phi(\Omega)$ or a subspace of it. This assumption does not hold when ζ^θ is sampled from a Gaussian process because the set $\mathcal{N}_\Phi(\Omega)$ has probability zero under the probability measure of the corresponding Gaussian process [6]. Our discussion, however, should not be affected because we are *not* adopting a Gaussian process model. Also, we believe that $\zeta^\theta \in \mathcal{N}_\Phi(\Omega)$ is a reasonable assumption in the context of computer experiments because the reproducing kernel Hilbert space is large enough, which covers all smooth functions.

For notational consistency with section 2.4, we write $\hat{\theta}$ as $\hat{\theta}_n$ to emphasize its dependency on n. Similarly, we write λ as λ_n. Then (3.6) becomes

$$\hat{\theta}_n = \operatorname*{argmin}_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n}(\zeta_i^\theta - \hat{\zeta}_n^\theta(x_i))^2 + \lambda_n\|\hat{\zeta}_n^\theta\|^2_{\mathcal{N}_\Phi(\Omega)},$$

with

$$\hat{\zeta}_n^\theta = \operatorname*{argmin}_{g} \frac{1}{n}\sum_{i=1}^{n}(\zeta_i^\theta - g(x_i))^2 + \lambda_n\|g\|^2_{\mathcal{N}_\Phi(\Omega)}.$$

Following the standard framework for establishing asymptotic theory for M-estimation, we should consider the limiting behavior of the objective function:

$$\frac{1}{n}\sum_{i=1}^{n}(\zeta_i^\theta - \hat{\zeta}_n^\theta(x_i))^2 + \lambda_n\|\hat{\zeta}_n^\theta\|^2_{\mathcal{N}_\Phi(\Omega)}. \tag{3.7}$$

Although this function is related to the kernel ridge regression, the standard rates of convergence for kernel ridge regression given by Corollary 2.9 are insufficient to provide an asymptotic result for $\hat{\theta}_n$. To see this, we note that according to Corollary 2.9, the second term in (3.7) is merely known to be $O_p(\lambda_n)$. This error bound is too crude to ensure a convergence result for $\hat{\theta}_n$.

In contrast, if the conditions of Proposition 2.10 are fulfilled, the improved rate of convergence gives the asymptotic representation

$$\|\hat{\zeta}_n^\theta\|^2_{\mathcal{N}_\Phi(\Omega)} = \|\zeta^\theta\|^2_{\mathcal{N}_\Phi(\Omega)} + O_p(\lambda_n),$$

which gives a much finer error bound. Thanks to the improved rates of convergence, we can establish an asymptotic theory for $\hat{\theta}_n$.

We first consider the prediction problem: how accurate $\hat{\zeta}_n^\theta$ can approximate ζ^θ in a uniform sense. The result, which is a generalization of Proposition 2.10, is given by Theorem 3.2. As in section 2.4, we assume that the reproducing kernel Hilbert space $\mathcal{N}_\Phi(\Omega)$ is equal to some (fractional) Sobolev space $H^m(\Omega)$ with equivalent norms for some $m > d/2$. Specifically, if Φ is a Matérn kernel in (2.9), then $m = \nu + d/2$.

In Theorem 3.2, we pursue nonasymptotic error bounds; that is, the sample size n is assumed to be fixed rather than tending to infinity. In the rest of this article, we use $c_1, c_2, c_3, \ldots$ to denote universal positive constants. They are independent of n. They may depend on m, d, Ω and the quasi-uniformity constant B in (2.27) but are independent of the specific collocation scheme of the design points. For simplicity, we may use the same c_i in different places to denote different constants.

Theorem 3.2. *Suppose the set of design points is quasi-uniform, i.e., (2.27) holds. Suppose for each $\theta \in \Theta$, we have $\zeta^\theta \in \mathcal{N}_\Phi(\Omega)$, and there exists $v_\theta \in L_2(\Omega)$ such that*

$$\begin{aligned} \zeta^\theta(x) &= \int_\Omega \Phi(x-t)v_\theta(t)dt, \\ \bar{v} &:= \sup_{\theta\in\Theta}\|v_\theta\|_{L_2(\Omega)} < +\infty. \end{aligned} \tag{3.8}$$

Then, for $n > c_1$, the inequalities

$$\sup_{\theta\in\Theta}\|\hat{\zeta}_n^\theta - \zeta^\theta\|_n \le c_2\bar{v}\lambda_n \vee c_3 t n^{-\frac{1}{2}}\lambda_n^{-\frac{d}{4m}},$$

$$\sup_{\theta\in\Theta}\|\hat{\zeta}_n^\theta - \zeta^\theta\|_{\mathcal{N}_\Phi(\Omega)} \le c_4\bar{v}\lambda_n^{\frac{1}{2}} \vee c_5 t n^{-\frac{1}{2}}\lambda_n^{-\frac{2m+d}{4m}}$$

hold simultaneously on the event

$$A_t := \left\{\sup_{g\in\mathcal{N}_\Phi(\Omega)}\frac{|\langle e, g\rangle_n|}{\|g\|_n^{1-\frac{d}{2m}}\|g\|^{\frac{d}{2m}}_{\mathcal{N}_\Phi(\Omega)}} \le t n^{-1/2}\right\}. \tag{3.9}$$

Condition (3.8) is a uniform version of the condition (2.13) because in (3.8) we require not only the existence of $v_\theta \in L_2(\Omega)$ but also the uniform boundedness of their L_2 norms. Suppose a Matérn kernel in (2.9) with $m = \nu + d/2$ is used. Theorem 2.3 shows that (2.13) is equivalent to $\|f_e\|_{H^{2m}(\mathbb{R}^d)} < \infty$. From the proof of Theorem 2.3, one can justify that (3.8) is equivalent to $\sup_{\theta\in\Theta} \|\zeta_e^\theta\|_{H^{2m}(\mathbb{R}^d)} < \infty$. From Theorem 3.2, we can establish the asymptotic rates of convergence as given in Corollary 3.3.

Corollary 3.3. *Suppose e_i's are sub-Gaussian. Then, under the conditions of Theorem 3.2, we have the rates of convergence*

$$\sup_{\theta\in\Theta} \|\hat{\zeta}_n^\theta - \zeta^\theta\|_n = O_p\left(\lambda_n \vee n^{-\frac{1}{2}} \lambda_n^{-\frac{d}{4m}}\right),$$

$$\sup_{\theta\in\Theta} \|\hat{\zeta}_n^\theta - \zeta^\theta\|_{\mathcal{N}_\Phi(\Omega)} = O_p\left(\lambda_n^{\frac{1}{2}} \vee n^{-\frac{1}{2}} \lambda_n^{-\frac{2m+d}{4m}}\right).$$

Proof. According to Lemma A.1, A_t has probability at least $1 - c_1 \exp\{-c_2 t^2\}$ for all $t > c_3$, which tends to one as $t \to +\infty$. The rates then follow from Theorem 3.2. ■

Next we state the convergence results for $\hat{\theta}_n$. We will show that under certain conditions, $\hat{\theta}_n$ will tend to

$$\theta' = \underset{\theta\in\Theta}{\operatorname{argmin}} \|\zeta^\theta\|_{\mathcal{N}_\Phi(\Omega)} \tag{3.10}$$

as $n \to \infty$. Here we only present the error bound of $\|\hat{\theta}_n - \theta'\|$ for the case $\lambda_n^{-1} = O_p(n^{\frac{2m}{4m+d}})$ because this case gives the best rate of convergence. By using similar but more cumbersome mathematical analysis, we can show that $\hat{\theta}_n$ converges to θ' if $\lambda_n^{-1} = o_p(n^{\frac{2m}{2m+d}})$. The general error bounds are more complicated, and we choose not to pursue them here.

Theorem 3.4. *Suppose the conditions of Theorem 3.2 are fulfilled. In addition, we suppose that θ' is the unique solution to (3.10). Moreover, there exists constants $a_2, a_3, \gamma > 0$ such that*

$$\|\zeta^\theta\|^2_{\mathcal{N}_\Phi(\Omega)} - \|\zeta^{\theta'}\|^2_{\mathcal{N}_\Phi(\Omega)} \geq a_2 \min\{\|\theta - \theta'\|^\gamma, a_3\} \tag{3.11}$$

for all $\theta \in \Theta$, where $\|\cdot\|$ denotes the Euclidean distance. Let A_t be the event defined in (3.9) and

$$\lambda_n^{\frac{4m+d}{4m}} > a_1 \bar{v}^{-1} t n^{-1/2} \tag{3.12}$$

for some $a_1 > 0$. If $\bar{v}^2 \lambda_n < c_1$, then on the event A_t,

$$\|\hat{\theta}_n - \theta'\| \leq c_3 \bar{v}^{2/\gamma} \lambda_n^{1/\gamma}.$$

Remark 3.5. Condition (3.11) is a counterpart of the local strong convexity around the minimum point. Specifically, (3.11) requires $\|\zeta^\theta\|^2_{\mathcal{N}_\Phi(\Omega)} - \|\zeta^{\theta'}\|^2_{\mathcal{N}_\Phi(\Omega)} \geq a_2\|\theta - \theta'\|^\gamma$ for θ near θ', which is a Hölder condition. If $h(\theta) := \|\zeta^\theta\|^2_{\mathcal{N}_\Phi(\Omega)} - \|\zeta^{\theta'}\|^2_{\mathcal{N}_\Phi(\Omega)}$ is continuously twice differentiable around θ', then we can apply Taylor's theorem to conclude that (3.11) holds with $\gamma = 2$. When θ is far away from θ', (3.11) requires that $\|\zeta^{\theta}\|^2_{\mathcal{N}_\Phi(\Omega)} - \|\zeta^{\theta'}\|^2_{\mathcal{N}_\Phi(\Omega)}$ is bounded below from zero, which means that we do not require the whole function to be convex. If Θ is bounded, the only global assumption here is that the minimizer is unique. The condition $\bar{v}^2\lambda_n < c_1$ requires that λ_n should be small enough, which can be fulfilled asymptotically by choosing $\lambda_n \downarrow 0$ but not decreasing too fast so that (3.12) holds as well.

Corollary 3.6. *Under the conditions of Theorem 3.4 and $\lambda_n^{-1} = O(n^{\frac{2m}{4m+d}})$, we have the rate of convergence $\|\hat{\theta}_n - \theta'\| = O_p(\lambda_n^{1/\gamma})$. Specifically, if $h(\theta) := \|\zeta^{\theta}\|^2_{\mathcal{N}_\Phi(\Omega)} - \|\zeta^{\theta'}\|^2_{\mathcal{N}_\Phi(\Omega)}$ is continuously twice differentiable around θ', then $\|\hat{\theta}_n - \theta'\| = O_p(\lambda_n^{1/2})$.*

Proof. According to Lemma A.1, A_t has probability at least $1 - c_1 \exp\{-c_2 t^2\}$ for all $t > c_3$, which tends to one as $t \to +\infty$. The rate then follows from Theorem 3.4. ■

Remark 3.7. [26] observed that under certain conditions, the limit value of the K-O method is θ' defined in (3.10), i.e., $\theta_0 = \theta'$. In Theorem 4.2 of [26], they prove the limit result when the physical observations y_i have no random error, i.e., e_i's in (3.1) are zero. In Tuo–Wu's result, the condition (3.8) is also necessary in the mathematical treatments. In Theorem 3.4 of this paper, we generalize the Tuo–Wu theory by assuming that e_i's are independent and identically distributed sub-Gaussian random variables and obtain the rate of convergence. Given the fact that physical responses are always subject to random noise, Theorem 3.4 in this paper is much more useful than Theorem 4.2 of [26] for practical applications. Therefore, the result we obtain here can be viewed as a substantial improvement over the Tuo–Wu theory.

Remark 3.8. It is worth noting that the limit value of the K-O calibration estimator under the current framework differs from that of some other methods, including [9, 10, 24, 25]. These existing methods converge to the minimizer of $\|\zeta^{\theta}\|_{L_2(\Omega)}$, i.e.,

$$\theta^* = \operatorname*{argmin}_{\theta\in\Theta} \|\zeta^{\theta}\|_{L_2(\Omega)}. \tag{3.13}$$

Note that (3.13) differs from the definition of θ' in (3.10) because the reproducing kernel Hilbert space norm in (3.10) is replaced by the L_2 norm. Besides, these existing methods turn out to have faster rates of convergence. Especially, a $n^{-1/2}$ rate of convergence and a semiparametric efficiency can be achieved [24, 25]. We believe that such a high rate of convergence cannot be achieved in the current context, in which we require the limit value to be θ' instead of θ^*.

4. Numerical studies.

We conduct numerical studies in sections 4.1 and 4.2 to validate the theoretical results in sections 2.4 and 3.2, respectively.

4.1. Numerical results for kernel ridge regression.

Corollary 2.12 shows that under certain conditions and $\lambda_n \sim n^{-\frac{2m}{4m+d}}$, we have the rate of convergence of $\hat{f}_n$:

$$\|\hat{f}_n - f\|_{L_2(\Omega)} = O_p(n^{-\frac{2m}{4m+d}}). \tag{4.1}$$

In this section, we consider the following model to verify whether the rate of convergence $O_p(n^{-\frac{2m}{4m+d}})$ is sharp. We start by taking the logarithm on both sides of (4.1) to get

$$\log \|\hat{f}_n - f\|_{L_2(\Omega)} \lesssim -\frac{2m}{4m+d} \log n + c.$$

This inspires us to consider a set of sample sizes, denoted as $\{n_1, \ldots, n_k\}$, and for each n_j, we conduct an independent simulation and compute $L_j = \|\hat{f}_{n_j} - f\|_{L_2(\Omega)}$. Next we consider the regression problem given by

$$\log L_j = a + b \log n_j + e_j, \quad j = 1, \ldots, k. \tag{4.2}$$

We estimate the regression coefficients (a, b) by the least squares method and denote the estimator as $(\hat{a}, \hat{b})$. Then we can regard $O_p(n^{\hat{b}})$ as the estimated rate of convergence. We shall check whether $\hat{b}$ is close to $-\frac{2m}{4m+d}$.

In our simulation study, we need functions that satisfy the condition (2.35). Suppose $\Phi(x)$ is the exponential kernel function $\Phi(x) = \exp\{-|x|\}$, which is also the Matérn kernel function (2.9) with $\phi = 1$ and $\nu = 0.5$, and the experimental region $\Omega = [-1, 1]$. The corresponding Sobolev space is $H^1[-1, 1]$, that is, $m = 1$ and $d = 1$.

Suppose the true function f is

$$f(x) = \int_{-1}^{1} \Phi(x - t)\Phi(t)dt = e^{-|x|} + |x|e^{-|x|} - e^{x-2}/2 - e^{-(x+2)}/2. \tag{4.3}$$

Clearly, f satisfies the condition (2.35). Suppose we observe data

$$y_i = f(x_i) + e_i, i = 1, \ldots, n,$$

where e_i's are independent and identically distributed random errors following $N(0, \tau^2)$ and $\tau = 0.1$.

To estimate the regression coefficient in (4.2), we choose 30 different Sobol designs [18] with sample sizes $n_j = 20j, j = 1, \ldots, 30$. For each j, we use the Monte Carlo method to calculate $\|\hat{f}_{n_j} - f\|_{L_2(\Omega)}$, where $\hat{f}_{n_j}$ is computed by using (2.19). Following the theoretical guidance in Corollary 2.12, we choose $\lambda_j = \hat{\eta} n_j^{-2m/(4m+d)}$ and determine the constant $\hat{\eta}$ by the following cross-validation approach. We consider the largest sample size in the simulation $n_k = 600$, and the estimate $\hat{\eta}$ is obtained by the K-fold cross-validation method [12] in one simulation run with $K = 10$. We use the *caret* package [14] in R to find $\hat{\eta} = 0.0668$, and then we use this value in the rest of this simulation study.

We repeat the simulation 100 times and calculate the Monte Carlo sample mean to reduce the random error. The scattered plot of $\log L_j$ against $\log n_j$ is shown in Figure 1. The estimated regression coefficient is -0.399, which closely agrees with our theoretical assertion -0.4.

4.2. Numerical results for the K-O calibration.

In this section, we verify the rate of convergence given by Theorem 3.4. Theorem 3.4 asserts that under certain conditions and

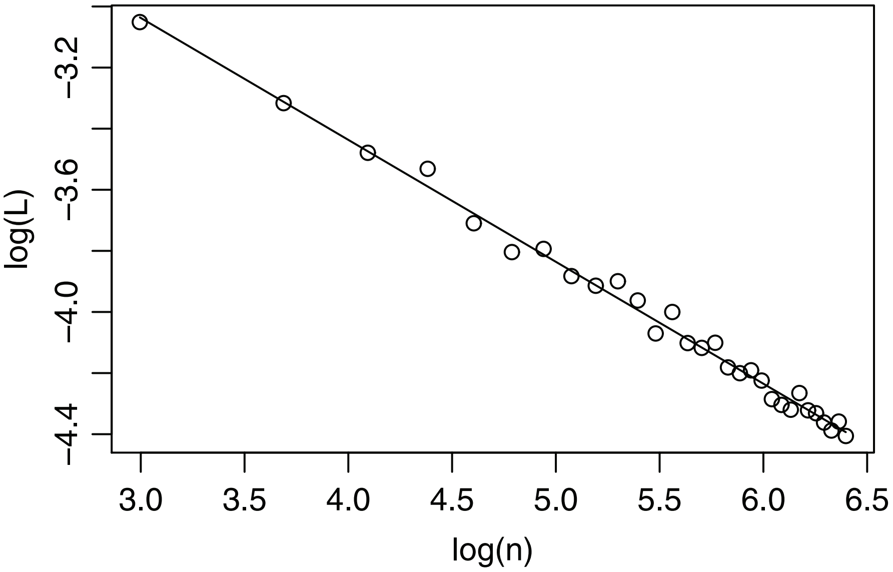

Figure 1. *The scattered plot and the regression line of the simulated data for kernel ridge regression.*

$\lambda_n \sim n^{-\frac{2m}{4m+d}}$, we have the rate of convergence $\|\hat\theta_n - \theta'\| = O_p(\lambda_n^{1/\gamma})$. Specifically, if $h(\theta) := \|\zeta^\theta\|^2_{\mathcal{N}_\Phi(\Omega)} - \|\zeta^{\theta'}\|^2_{\mathcal{N}_\Phi(\Omega)}$ is continuously twice differentiable around θ', then

$$\|\hat\theta_n - \theta'\| = O_p(\lambda_n^{1/2}) = O_p(n^{-\frac{m}{4m+d}}). \tag{4.4}$$

Denoting $E_j = \|\hat\theta_{n_j} - \theta'\|$, we consider the regression problem given by

$$\log E_j = a + b \log n_j + e_j, \quad j = 1, \dots, k. \tag{4.5}$$

We shall check whether $\hat b$ is close to $-\frac{m}{4m+d}$, the theoretical rate of convergence asserted by Theorem 3.4. Suppose the true process $\xi(x)$ is same as the function (4.3) in section 4.1 and the computer model is

$$y^s(x,\theta) = \xi(x) - \int_{-1}^{1} \Phi(x-y)(\theta y^2 + 0.8)dy,$$

where $\Phi(x) = \exp\{-|x|\}$ and θ is the model parameter to be calibrated.

Clearly, the discrepancy function $\zeta^\theta(x) = \int_{-1}^1 \Phi(x-y)(\theta y^2 + 0.8)dy$ satisfies all conditions of Theorem 3.4. The identity (2.14) implies

$$\|\zeta^\theta\|^2_{\mathcal{N}_\Phi(\Omega)} = \int_{-1}^1\int_{-1}^1 (\theta x^2 + 0.8)\Phi(x)\Phi(x-y)\Phi(y)(\theta y^2 + 0.8)dxdy.$$

By numerical search, we find that, as a function of θ, $\|\zeta^\theta\|^2_{\mathcal{N}_\Phi(\Omega)}$ is minimized at $\theta' = -2.348$.

Suppose we observe data according to

$$y_i = \xi(x_i) + e_i, i = 1, \dots, n,$$

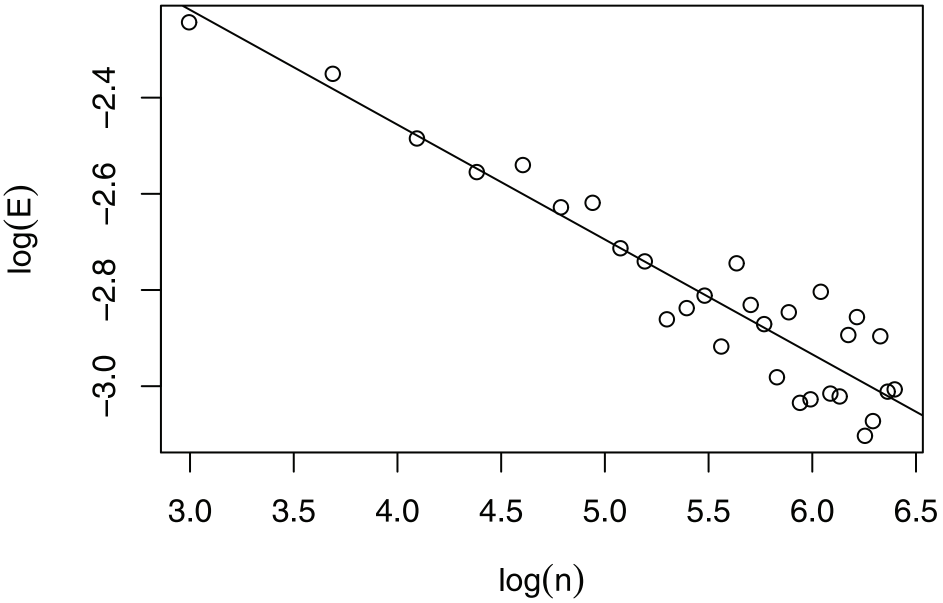

Figure 2. *The scattered plot and the regression line of the simulated data for K-O calibration.*

where x_i's are same as the ones in section 4.1 and e_i's are independent and identically distributed random errors following $N(0,\tau^2)$. Now we can compute $\hat{\theta}$ in (3.6). We choose $\lambda_j = \hat{\eta} n_j^{-2m/(4m+d)}$, and the constant $\hat{\eta}$ is determined by a cross-validation approach similar to that in section 4.1. The resulting $\hat{\eta} = 0.0615$. For each j, we repeat the simulation 100 times and calculate the Monte Carlo sample mean to reduce the random error.

The scattered plot of $\log E_j$ against $\log n_j$ is shown in Figure 2. The estimated regression coefficient is -0.227, which closely agrees with our theoretical assertion -0.2.

5. Discussion. In this work, we obtain some new results on the improved rates of convergence for kernel ridge regression. We apply this theory to study the asymptotic properties of the K-O calibration method for computer experiments. This new result generalizes the work of [26].

Several related problems can be studied in the future. In this article, we suppose the design set $\{x_1,\ldots,x_n\}$ is fixed and quasi-uniform. A further question is whether the improved rates still hold if the design points are random samples, for instance, if the design points are independent and follow the uniform distribution over Ω.

As is discussed in section 2.4, compared to the existing results, the bias of the kernel ridge regression estimator is reduced by imposing the condition (2.35), while the variance remains the same. Improved rates of convergence are achieved by rebalancing the bias and the variance. In other words, the choice of the smoothing parameters λ_n is crucial in achieving the optimal rate of convergence. Suppose condition (2.35) is fulfilled. Proposition 2.10 implies that the optimal tuning parameter is $\lambda_n \sim n^{-\frac{2m}{4m+d}}$. If condition (2.35) is not satisfied, we should return to the classic results given by Proposition 2.5. In this case, the optimal tuning parameter is $\lambda_n \sim n^{-\frac{2m}{2m+d}}$, and $\lambda_n \sim n^{-\frac{2m}{4m+d}}$ would render a suboptimal rate of convergence. In most practical scenarios, we do not know whether the condition (2.35) holds or not. Therefore, there is no a priori optimal choice of λ_n. One would ask whether the optimal order of magnitude for λ_n can be obtained by a data-driven approach. We conjecture that model selection criteria like generalized cross validation [31] can automatically adapt an optimal choice of λ_n.

Appendix A. Technical proofs.

Proof of Theorem 2.3. Without loss of generality, we can assume that the scale parameter ϕ in (2.9) is $1/(2\sqrt{\nu})$ because otherwise we can stretch the region Ω to make this happen. In this situation, the Matérn kernel becomes

$$\frac{1}{\Gamma(\nu)2^{\nu-1}}|x|^{\nu}K_{\nu}(|x|).$$

Suppose $f(x)=\int_\Omega \Phi(x-t)v(t)dt$ with $v\in L_2(\Omega)$. It can be justified that $f_e=\int_\Omega \Phi(x-t)v(t)dt$ for $x\in\mathbb{R}^d$. See Lemma 11.34 in [20] for details. Define

$$v_e(x)=\begin{cases} v(x), & x\in\Omega,\\ 0, & x\notin\Omega.\end{cases}$$

Clearly, $f_e(x)=\int_{\mathbb{R}^d}\Phi(x-t)v_e(t)dt$. For $h\in L_2(\mathbb{R}^d)$, denote its Fourier transform and inverse Fourier transform by $\mathcal{F}(h)$ and $\mathcal{F}^{-1}(h)$, respectively. Then by the convolution theorem, $\mathcal{F}(f_e)=(2\pi)^{d/2}\mathcal{F}(\Phi)\mathcal{F}(v_e)$. Direct calculations [22, 32] give

$$\begin{aligned}\mathcal{F}(\Phi)(\omega) &= (2\pi)^{d/2}\frac{\Gamma(\nu+d/2)}{\Gamma(\nu)}(1+\|\omega\|^2)^{-(\nu/2+d/4)}\\ &:= C_0(1+\|\omega\|^2)^{-m/2}.\end{aligned}\tag{A.1}$$

Note that $\mathcal{F}(f_e)/\mathcal{F}(\Phi)=(2\pi)^{d/2}\mathcal{F}(v_e)\in L_2(\mathbb{R}^d)$, which gives

$$\int_{\mathbb{R}^d}(1+\|\omega\|^2)^m|\mathcal{F}(f_e)(\omega)|^2d\omega<+\infty.\tag{A.2}$$

According to Paragraph 7.62 of [1], (A.2) is equivalent to $f_e\in H^{2m}(\mathbb{R})$.

Suppose $f_e\in H^{2m}(\mathbb{R})$. Then by (A.2) and (A.1), we have $\mathcal{F}(f_e)/\mathcal{F}(\Phi)\in L_2(\mathbb{R}^d)$. Theorem 4.3 of [20] proves that in this case, $h_f:=\mathcal{F}^{-1}(\mathcal{F}(f_e)/\mathcal{F}(\Phi))$ is zero almost everywhere outside Ω. Then according to the convolution theorem, $v:=h_f|_\Omega$ satisfies (2.17). ∎

Lemma A.1. *Suppose* $\{x_1,\dots,x_n\}\subset\Omega$, $e_1,\dots,e_n$ *are independent and identically distributed random variables which are sub-Gaussian. Then, for all* $t>c_1$, *we have*

$$\sup_{g\in\mathcal{N}_\Phi(\Omega)}\frac{|\langle e,g\rangle_n|}{\|g\|_n^{1-\frac{d}{2m}}\|g\|_{\mathcal{N}_\Phi(\Omega)}^{\frac{d}{2m}}}\le tn^{-1/2},\tag{A.3}$$

with probability at least $1-c_2\exp\{-c_3t^2\}$, *where* $\langle e,g\rangle_n$ *is defined in* (2.25).

Proof. For $g \in \mathcal{N}_\Phi(\Omega)$, let $h = g/\|g\|_{\mathcal{N}_\Phi(\Omega)}$. It is easily verified that

$$\frac{|\langle e, g\rangle_n|}{\|g\|_n^{1-\frac{d}{2m}}\|g\|_{\mathcal{N}_\Phi(\Omega)}^{\frac{d}{2m}}} = \frac{|\langle e, h\rangle_n|}{\|h\|_n^{1-\frac{d}{2m}}}.$$

Let $\mathcal{H} = \{h \in \mathcal{N}_\Phi(\Omega) : \|h\|_{\mathcal{N}_\Phi(\Omega)} = 1\}$. Noting that $\mathcal{N}_\Phi(\Omega)$ can be embedded into $H^m(\Omega)$, we can use the metric entropy of the Sobolev spaces [7, 25] to find an upper bound of the metric entropy of $\mathcal{H}$ as

$$H(\epsilon, \mathcal{H}, \|\cdot\|_n) \leq c_4\epsilon^{-d/m}.$$

We refer the reader to [30] for the definition and detailed discussions about the metric entropy of a function space. The remainder of the proof follows by invoking the concentration inequality given by Corollary 14.6 of [4]. ■

Proof of Theorem 3.1. Using (2.19) and (2.20), we find that $\hat{\zeta}^\theta$ can be expressed by

$$\hat{\zeta}^\theta = \sum_{i=1}^n c_i^\theta \Phi(x - x_i),$$

with $c^\theta = (c_1^\theta, \ldots, c_n^\theta)^T$ defined as

$$c^\theta = (\mathbf{\Phi} + n\lambda I_n)^{-1}Y_\theta,$$

where $\mathbf{\Phi} = (\Phi(x_i, x_j))_{ij}$ and $Y_\theta = (\zeta_1^\theta, \ldots, \zeta_n^\theta)^T$. The norm of $\hat{\zeta}^\theta$ in $\mathcal{N}_\Phi(\Omega)$ can be calculated using (2.8), given by

$$\|\hat{\zeta}^\theta\|_{\mathcal{N}_\Phi(\Omega)}^2 = Y_\theta^T(\mathbf{\Phi} + n\lambda I_n)^{-1}\mathbf{\Phi}(\mathbf{\Phi} + n\lambda I_n)^{-1}Y_\theta. \tag{A.4}$$

Also, $\hat{\zeta}^\theta(X) := (\hat{\zeta}^\theta(x_1), \ldots, \hat{\zeta}^\theta(x_n))^T$ can be expressed by

$$\hat{\zeta}^\theta(X) = \mathbf{\Phi}(\mathbf{\Phi} + n\lambda I_n)^{-1}Y_\theta,$$

which yields

$$\begin{aligned} Y_\theta - \hat{\zeta}^\theta(X) &= (I - \mathbf{\Phi}(\mathbf{\Phi} + n\lambda I_n)^{-1})Y_\theta \\ &= n\lambda(\mathbf{\Phi} + n\lambda I_n)^{-1}Y_\theta. \end{aligned}$$

Thus,

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^n(\zeta_i^\theta - \hat{\zeta}^\theta(x_i))^2 &= \frac{1}{n}(Y_\theta - \hat{\zeta}^\theta(X))^T(Y_\theta - \hat{\zeta}^\theta(X)) \\ &= n\lambda^2 Y_\theta^T(\mathbf{\Phi} + n\lambda I_n)^{-2}Y_\theta. \end{aligned} \tag{A.5}$$

From (A.4) and (A.5) we obtain

$$\frac{1}{n}\sum_{i=1}^n(\zeta_i^\theta - \hat{\zeta}^\theta(x_i))^2 + \lambda\|\hat{\zeta}^\theta\|_{\mathcal{N}_\Phi(\Omega)}^2 = \lambda Y_\theta^T(\mathbf{\Phi} + n\lambda I_n)^{-1}Y_\theta,$$

which implies the desired results. ■

Lemma A.2. *Let λ, P, Q, s, t be nonnegative numbers and n be a positive integer. If there exist constants $a_1, a_2, a_3, a_4 > 0$ such that*

$$P^2 + a_1\lambda Q^2 \leq a_2 s\lambda P + a_3 t n^{-\frac{1}{2}} P^{1-\frac{d}{2m}} Q^{\frac{d}{2m}}, \tag{A.6}$$

then we have

$$\begin{aligned} P &\leq b_1 s\lambda \vee b_3 t n^{-\frac{1}{2}} \lambda^{-\frac{d}{4m}}, \\ Q &\leq b_2 s\lambda^{\frac{1}{2}} \vee b_4 t n^{-\frac{1}{2}} \lambda^{-\frac{2m+d}{4m}}. \end{aligned}$$

Here b_1, b_2, b_3, b_4 are independent of $P, Q, \lambda, s,$ and t.

Proof. Clearly, (A.6) implies either

$$P^2 + a_1\lambda Q^2 \leq 2a_2 s\lambda P$$

or

$$P^2 + a_1\lambda Q^2 \leq 2a_3 t n^{-1/2} P^{1-\frac{d}{2m}} Q^{\frac{d}{2m}}.$$

Next we consider these two cases separately.

Case I. Suppose $P^2 + a_1\lambda Q^2 \leq 2a_2 s\lambda P$. Then we have

$$\begin{aligned} P^2 &\leq 2a_2 s\lambda P, \\ a_1\lambda Q^2 &\leq 2a_2 s\lambda P, \end{aligned}$$

which yields

$$\begin{aligned} P &\leq 2a_2 s\lambda = b_1 s\lambda, \\ Q &\leq 2a_2 a_1^{-1/2} s\lambda^{1/2} = b_2 s\lambda^{1/2}. \end{aligned} \tag{A.7}$$

Case II. Suppose $P^2 + a_1\lambda Q^2 \leq 2a_3 n^{-1/2} P^{1-\frac{d}{2m}} Q^{\frac{d}{2m}}$. Then we have

$$\begin{aligned} P^2 &\leq 2a_3 t n^{-1/2} P^{1-\frac{d}{2m}} Q^{\frac{d}{2m}}, \\ a_1\lambda Q^2 &\leq 2a_3 t n^{-1/2} P^{1-\frac{d}{2m}} Q^{\frac{d}{2m}}. \end{aligned} \tag{A.8}$$

By elementary calculations, we find that (A.8) implies

$$\begin{aligned} P &\leq b_3 t n^{-\frac{1}{2}} \lambda^{-\frac{d}{4m}}, \\ Q &\leq b_4 t n^{-\frac{1}{2}} \lambda^{-\frac{2m+d}{4m}}. \end{aligned} \tag{A.9}$$

The desired results then follows by combining (A.7) and (A.9). ■

Proof of Theorem 3.2. Following similar arguments as those in (2.32)–(2.34), we can deduce the improved basic inequality

$$\begin{aligned} &\|\zeta^\theta - \hat{\zeta}_n^\theta\|_n^2 + \lambda_n \|\zeta^\theta - \hat{\zeta}_n^\theta\|^2_{\mathcal{N}_\Phi(\Omega)} \\ &\leq 2\langle e, \hat{\zeta}_n^\theta - \zeta^\theta\rangle_n + 2\lambda_n \|v_\theta\|_{L_2(\Omega)} \|\zeta^\theta - \hat{\zeta}_n^\theta\|_{L_2(\Omega)}, \end{aligned} \tag{A.10}$$

which holds for all $\theta \in \Theta$.

It follows from Proposition 2.8 and (2.29) that, for sufficiently large n,

$$\begin{aligned}\|\zeta^\theta-\hat{\zeta}_n^\theta\|_{L_2(\Omega)} &\leq C\sqrt{\|\zeta^\theta-\hat{\zeta}_n^\theta\|_n^2+n^{-\frac{2m}{d}}\|\zeta^\theta-\hat{\zeta}_n^\theta\|_{H^m(\Omega)}^2}\\ &\leq C\left\{\|\zeta^\theta-\hat{\zeta}_n^\theta\|_n+n^{-\frac{m}{d}}\|\zeta^\theta-\hat{\zeta}_n^\theta\|_{H^m(\Omega)}\right\}\\ &\leq C\|\zeta^\theta-\hat{\zeta}_n^\theta\|_n+C'n^{-\frac{m}{d}}\|\zeta^\theta-\hat{\zeta}_n^\theta\|_{\mathcal{N}_\Phi(\Omega)}, \end{aligned} \tag{A.11}$$

where the last inequality follows from the assumption that $\|\cdot\|_{H^m(\Omega)}$ and $\|\cdot\|_{\mathcal{N}_\Phi(\Omega)}$ are equivalent.

Combining (A.10), (A.11), and the condition $\bar{v}=\sup_{\theta\in\Theta}\|v_\theta\|_{L_2(\Omega)}<+\infty$ yields

$$\begin{aligned}&\|\zeta^\theta-\hat{\zeta}_n^\theta\|_n^2+\lambda_n\|\zeta^\theta-\hat{\zeta}_n^\theta\|_{\mathcal{N}_\Phi(\Omega)}^2\leq 2\langle e,\hat{\zeta}_n^\theta-\zeta^\theta\rangle_n\\ &+2C\lambda_n\bar{v}\|\zeta^\theta-\hat{\zeta}_n^\theta\|_{L_2(\Omega)}+2C'\lambda_n\bar{v}n^{-\frac{m}{d}}\|\zeta^\theta-\hat{\zeta}_n^\theta\|_{\mathcal{N}_\Phi(\Omega)}.\end{aligned} \tag{A.12}$$

Now we consider three different cases.

Case I. Suppose $n^{-\frac{m}{d}}\|\zeta^\theta-\hat{\zeta}_n^\theta\|_{\mathcal{N}_\Phi(\Omega)}\leq\|\zeta^\theta-\hat{\zeta}_n^\theta\|_{L_2(\Omega)}$. Then we obtain from (A.12) that

$$\begin{aligned}&\|\zeta^\theta-\hat{\zeta}_n^\theta\|_n^2+\lambda_n\|\zeta^\theta-\hat{\zeta}_n^\theta\|_{\mathcal{N}_\Phi(\Omega)}^2\leq 2\langle e,\hat{\zeta}_n^\theta-\zeta^\theta\rangle_n\\ &+2(C+C')\lambda_n\bar{v}\|\zeta^\theta-\hat{\zeta}_n^\theta\|_{L_2(\Omega)}.\end{aligned} \tag{A.13}$$

Case II. Suppose $n^{-\frac{m}{d}}\|\zeta^\theta-\hat{\zeta}_n^\theta\|_{\mathcal{N}_\Phi(\Omega)}>\|\zeta^\theta-\hat{\zeta}_n^\theta\|_{L_2(\Omega)}$ and $4C'\bar{v}n^{-\frac{m}{d}}\leq\|\zeta^\theta-\hat{\zeta}_n^\theta\|_{\mathcal{N}_\Phi(\Omega)}$. Then we can cancel the term $\lambda_n\|\zeta^\theta-\hat{\zeta}_n^\theta\|_{\mathcal{N}_\Phi(\Omega)}^2/2$ from both sides of (A.12) and get

$$\begin{aligned}&\|\zeta^\theta-\hat{\zeta}_n^\theta\|_n^2+\frac{1}{2}\lambda_n\|\zeta^\theta-\hat{\zeta}_n^\theta\|_{\mathcal{N}_\Phi(\Omega)}^2\leq 2\langle e,\hat{\zeta}_n^\theta-\zeta^\theta\rangle_n\\ &+2C\lambda_n\bar{v}\|\zeta^\theta-\hat{\zeta}_n^\theta\|_{L_2(\Omega)}.\end{aligned} \tag{A.14}$$

Case III. Suppose $n^{-\frac{m}{d}}\|\zeta^\theta-\hat{\zeta}_n^\theta\|_{\mathcal{N}_\Phi(\Omega)}>\|\zeta^\theta-\hat{\zeta}_n^\theta\|_{L_2(\Omega)}$ and $4C'\bar{v}n^{-\frac{m}{d}}>\|\zeta^\theta-\hat{\zeta}_n^\theta\|_{\mathcal{N}_\Phi(\Omega)}$. It follows directly from this assumption that

$$\begin{aligned}\|\zeta^\theta-\hat{\zeta}_n^\theta\|_{\mathcal{N}_\Phi(\Omega)}&<4C'\bar{v}n^{-\frac{m}{d}},\\ \|\zeta^\theta-\hat{\zeta}_n^\theta\|_{L_2(\Omega)}&<4C'\bar{v}n^{-\frac{2m}{d}},\end{aligned}$$

from which we have already arrived at the desired results.

Now we only need to consider the first two cases. Clearly, both (A.13) and (A.14) can be expressed as

$$\begin{aligned}&\|\zeta^\theta-\hat{\zeta}_n^\theta\|_n^2+B_1\lambda_n\|\zeta^\theta-\hat{\zeta}_n^\theta\|_{\mathcal{N}_\Phi(\Omega)}^2\\ &\quad\leq 2\langle e,\hat{\zeta}_n^\theta-\zeta^\theta\rangle_n+B_2\lambda_n\bar{v}\|\zeta^\theta-\hat{\zeta}_n^\theta\|_{L_2(\Omega)}.\end{aligned} \tag{A.15}$$

On the event A_t, we have the inequality

$$\begin{aligned}&|\langle e,\hat{\zeta}_n^\theta-\zeta^\theta\rangle_n|\\ &\leq\sup_{g\in\mathcal{N}_\Phi(\Omega)}\frac{|\langle e,g\rangle_n|}{\|g\|_n^{1-\frac{d}{2m}}\|g\|_{\mathcal{N}_\Phi(\Omega)}^{\frac{d}{2m}}}\cdot\|\hat{\zeta}_n^\theta-\zeta^\theta\|_n^{1-\frac{d}{2m}}\|\hat{\zeta}_n^\theta-\zeta^\theta\|_{\mathcal{N}_\Phi(\Omega)}^{\frac{d}{2m}}\\ &\leq tn^{-1/2}\|\hat{\zeta}_n^\theta-\zeta^\theta\|_n^{1-\frac{d}{2m}}\|\hat{\zeta}_n^\theta-\zeta^\theta\|_{\mathcal{N}_\Phi(\Omega)}^{\frac{d}{2m}}.\end{aligned} \tag{A.16}$$

Combining inequalities (A.15)–(A.16) yields

$$\begin{aligned}&\|\zeta^\theta-\hat{\zeta}^\theta_n\|_n^2+B_1\lambda_n\|\zeta^\theta-\hat{\zeta}^\theta_n\|^2_{\mathcal{N}_\Phi(\Omega)}\\&\leq B_2\lambda_n\bar{v}\|\zeta^\theta-\hat{\zeta}^\theta_n\|_n+2tn^{-1/2}\|\hat{\zeta}^\theta_n-\zeta^\theta\|_n^{1-\frac{d}{2m}}\|\hat{\zeta}^\theta_n-\zeta^\theta\|^{\frac{d}{2m}}_{\mathcal{N}_\Phi(\Omega)}.\end{aligned}$$

Then we obtain the desired results by applying Lemma A.2. ■

Proof of Theorem 3.4. Under the condition (3.12), it is not hard to verify that $\bar{v}\lambda_n$ and $\bar{v}\lambda_n^{1/2}$ are bounded by the product of $tn^{-1/2}\lambda_n^{-\frac{d}{4m}}$ and $tn^{-1/2}\lambda_n^{-\frac{2m+d}{4m}}$, respectively. Thus, Theorem 3.2 gives

$$\sup_{\theta\in\Theta}\|\hat{\zeta}^\theta_n-\zeta^\theta\|_n\leq c_2\bar{v}\lambda_n, \tag{A.17}$$

$$\sup_{\theta\in\Theta}\|\hat{\zeta}^\theta_n-\zeta^\theta\|_{\mathcal{N}_\Phi(\Omega)}\leq c_3\bar{v}\lambda_n^{1/2}. \tag{A.18}$$

Using the definition of $\hat{\theta}_n$, we obtain the basic inequality

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n(\zeta_i^{\hat{\theta}_n}-\hat{\zeta}_n^{\hat{\theta}_n}(x_i))^2+\lambda_n\|\hat{\zeta}_n^{\hat{\theta}_n}\|^2_{\mathcal{N}_\Phi(\Omega)}\\&\leq\frac{1}{n}\sum_{i=1}^n(\zeta_i^{\theta'}-\hat{\zeta}_n^{\theta'}(x_i))^2+\lambda_n\|\hat{\zeta}_n^{\theta'}\|^2_{\mathcal{N}_\Phi(\Omega)},\end{aligned}$$

which is equivalent to

$$\begin{aligned}&\lambda_n\left\{\|\zeta^{\hat{\theta}_n}\|^2_{\mathcal{N}_\Phi(\Omega)}-\|\zeta^{\theta'}\|^2_{\mathcal{N}_\Phi(\Omega)}\right\}\\&\leq\left\{\|\hat{\zeta}_n^{\theta'}-\zeta^{\theta'}\|_n^2-\|\hat{\zeta}_n^{\hat{\theta}_n}-\zeta^{\hat{\theta}_n}\|_n^2\right\}\\&+2\left\{\langle e,\hat{\zeta}_n^{\hat{\theta}_n}-\zeta^{\hat{\theta}_n}\rangle_n-\langle e,\hat{\zeta}_n^{\theta'}-\zeta^{\theta'}\rangle_n\right\}\\&+\lambda_n\left\{\|\zeta^{\hat{\theta}_n}\|^2_{\mathcal{N}_\Phi(\Omega)}-\|\hat{\zeta}_n^{\hat{\theta}_n}\|^2_{\mathcal{N}_\Phi(\Omega)}-\|\zeta^{\theta'}\|^2_{\mathcal{N}_\Phi(\Omega)}+\|\hat{\zeta}_n^{\theta'}\|^2_{\mathcal{N}_\Phi(\Omega)}\right\}\\&=:D_1+2D_2+\lambda_nD_3.\end{aligned} \tag{A.19}$$

Now we bound D_1, D_2, and D_3 conditional on the event A_t. Using (A.17), we have

$$D_1\leq\sup_{\theta\in\Theta}\|\hat{\zeta}^\theta_n-\zeta^\theta\|_n^2\leq d_1\bar{v}^2\lambda_n^2. \tag{A.20}$$

By (A.17)–(A.18) and the definition of A_t, we obtain

$$\begin{aligned}D_2&\leq2\sup_{\theta\in\Theta}|\langle e,\hat{\zeta}^\theta_n-\zeta^\theta\rangle_n|\\&\leq2\sup_g\frac{|\langle e,g\rangle_n|}{\|g\|_n^{1-\frac{d}{2m}}\|g\|^{\frac{d}{2m}}_{\mathcal{N}_\Phi(\Omega)}}\sup_{\theta\in\Theta}\|\hat{\zeta}^\theta_n-\zeta^\theta\|_n^{1-\frac{d}{2m}}\sup_{\theta\in\Theta}\|\hat{\zeta}^\theta_n-\zeta^\theta\|^{\frac{d}{2m}}_{\mathcal{N}_\Phi(\Omega)}\\&\leq d_2tn^{-\frac{1}{2}}\bar{v}\lambda_n^{\frac{4m-d}{4m}}\\&\leq d_2'\bar{v}^2\lambda_n^2,\end{aligned} \tag{A.21}$$

where the last inequality follows from condition (3.12). For any $\theta \in \Theta$, we bound

$$\begin{aligned}
&\left| \|\hat{\zeta}_n^\theta\|^2_{\mathcal{N}_\Phi(\Omega)} - \|\zeta^\theta\|^2_{\mathcal{N}_\Phi(\Omega)} \right| \\
&= \left| \|\hat{\zeta}_n^\theta - \zeta^\theta\|^2_{\mathcal{N}_\Phi(\Omega)} + 2\langle \hat{\zeta}_n^\theta - \zeta^\theta, \zeta^\theta \rangle_{\mathcal{N}_\Phi(\Omega)} \right| \\
&\leq \|\hat{\zeta}_n^\theta - \zeta^\theta\|^2_{\mathcal{N}_\Phi(\Omega)} + 2\left| \langle \hat{\zeta}_n^\theta - \zeta^\theta, \zeta^\theta \rangle_{\mathcal{N}_\Phi(\Omega)} \right| \\
&= \|\hat{\zeta}_n^\theta - \zeta^\theta\|^2_{\mathcal{N}_\Phi(\Omega)} + 2\left| \langle \hat{\zeta}_n^\theta - \zeta^\theta, v_\theta \rangle_{L_2(\Omega)} \right| \\
&\leq \|\hat{\zeta}_n^\theta - \zeta^\theta\|^2_{\mathcal{N}_\Phi(\Omega)} + 2\|v_\theta\|_{L_2(\Omega)} \|\hat{\zeta}_n^\theta - \zeta^\theta\|_{L_2(\Omega)} \\
&\leq c_3^2 \bar{v}^2 \lambda_n + 2c_2 \bar{v}^2 \lambda_n \\
&= d_3 \bar{v}^2 \lambda_n,
\end{aligned}$$

where the second equality follows from (2.14), the second inequality follows from the Cauchy–Schwarz inequality, and the third inequality follows from (A.17) and (A.18). Therefore, we obtain the bound

$$D_3 \leq 2d_3 \bar{v}^2 \lambda_n. \tag{A.22}$$

Combining (A.19), (A.20), (A.21), and (A.22) and using the condition $t > 1$ yields

$$\|\zeta^{\hat{\theta}_n}\|^2_{\mathcal{N}_\Phi(\Omega)} - \|\zeta^{\theta'}\|^2_{\mathcal{N}_\Phi(\Omega)} \leq d_4 \bar{v}^2 \lambda_n. \tag{A.23}$$

The assumption (3.11) implies

$$a_2 \min\{\|\hat{\theta}_n - \theta'\|^\gamma, a_3\} \leq \|\zeta^{\hat{\theta}_n}\|^2_{\mathcal{N}_\Phi(\Omega)} - \|\zeta^{\theta'}\|^2_{\mathcal{N}_\Phi(\Omega)},$$

which, together with (A.23) and the condition $\bar{v}^2 \lambda_n < c_1 := a_3/(a_2 d_4)$, yields the desired results. ■

REFERENCES

[1] R. A. Adams and J. J. Fournier, *Sobolev Spaces*, Vol. 140, Academic Press, New York, 2003.

[2] G. Blanchard and N. Mücke, *Optimal rates for regularization of statistical inverse learning problems*, Found. Comput. Math., 18 (2018), pp. 971–1013.

[3] S. Brenner and R. Scott, *The Mathematical Theory of Finite Element Methods*, Vol. 15, Springer, New York, 2007.

[4] P. Bühlmann and S. Van De Geer, *Statistics for High-Dimensional Data: Methods, Theory and Applications*, Springer, New York, 2011.

[5] L. H. Dicker, D. P. Foster, D. Hsu, et al., *Kernel ridge vs. principal component regression: Minimax bounds and the qualification of regularization operators*, Electron. J. Stat., 11 (2017), pp. 1022–1047.

[6] M. F. Driscoll, *The reproducing kernel Hilbert space structure of the sample paths of a Gaussian process*, Probab. Theory Related Fields, 26 (1973), pp. 309–316.

[7] D. E. Edmunds and H. Triebel, *Function Spaces, Entropy Numbers, Differential Operators*, Vol. 120, Cambridge University Press, Cambridge, 1996.

[8] C. Gu, *Smoothing Spline ANOVA Models*, Vol. 297, Springer, New York, 2002.

[9] M. Gu and L. Wang, *Scaled Gaussian stochastic process for computer model calibration and prediction*, SIAM/ASA J. Uncertain. Quantif., 6 (2018), pp. 1555–1583.

[10] M. Gu, F. Xie, and L. Wang, *A Theoretical Framework of the Scaled Gaussian Stochastic Process in Prediction and Calibration*, preprint, https://arxiv.org/abs/1807.03829, 2018.

[11] Z.-C. Guo, S.-B. Lin, and D.-X. Zhou, *Learning theory of distributed spectral algorithms*, Inverse Problems, 33 (2017), 074009.

[12] G. James, D. Witten, T. Hastie, and R. Tibshirani, *Bias-variance trade-off for k-fold cross-validation*, in An Introduction to Statistical Learning: With Applications in R, Springer, New York, 2013, p. 183.

[13] M. C. Kennedy and A. O'Hagan, *Bayesian calibration of computer models*, J. Roy. Statist. Soc. Ser. B, 63 (2001), pp. 425–464.

[14] M. Kuhn, *A short introduction to the caret package*, R Found. Stat. Comput., 1 (2015), pp. 1–10.

[15] S.-B. Lin, X. Guo, and D.-X. Zhou, *Distributed learning with regularized least squares*, J. Mach. Learn. Res., 18 (2017), pp. 3202–3232.

[16] M. Plumlee, V. R. Joseph, and H. Yang, *Calibrating functional parameters in the ion channel models of cardiac cells*, J. Amer. Statist. Assoc., 111 (2015), pp. 500–509.

[17] G. Santin and R. Schaback, *Approximation of eigenfunctions in kernel-based spaces*, Adv. Comput. Math., 42 (2016), pp. 973–993.

[18] T. J. Santner, B. J. Williams, and W. I. Notz, *The Design and Analysis of Computer Experiments*, Springer, New York, 2003.

[19] C. Saunders, A. Gammerman, and V. Vovk, *Ridge regression learning algorithm in dual variables*, in International Conference on Machine Learning, Morgan Kaufmann, San Francisco, CA, 1998, pp. 515–521.

[20] R. Schaback, *Improved error bounds for scattered data interpolation by radial basis functions*, Math. Comp., (1999), pp. 201–216.

[21] B. Schölkopf, R. Herbrich, and A. J. Smola, *A generalized representer theorem*, in International Conference on Computational Learning Theory, Springer, New York, 2001, pp. 416–426.

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

[23] C. J. Stone, *Optimal global rates of convergence for nonparametric regression*, Ann. Statist., (1982), pp. 1040–1053.

[24] R. Tuo, *Adjustments to computer models via projected kernel calibration*, SIAM/ASA J. Uncertain. Quantif., 7 (2019), pp. 553–578.

[25] R. Tuo and C. F. J. Wu, *Efficient calibration for imperfect computer models*, Ann. Statist., 43 (2015), pp. 2331–2352.

[26] R. Tuo and C. F. J. Wu, *A theoretical framework for calibration in computer models: Parametrization, estimation and convergence properties*, SIAM/ASA J. Uncertain. Quantif., 4 (2016), pp. 767–795.

[27] R. Tuo and C. F. J. Wu, *Prediction based on the Kennedy-O'Hagan calibration model: Asymptotic consistency and other properties*, Statist. Sinica, 28 (2018), pp. 743–759.

[28] F. I. Utreras, *Convergence rates for multivariate smoothing spline functions*, J. Approx. Theory, 52 (1988), pp. 1–27.

[29] S. A. van de Geer, *Empirical Processes in M-estimation*, Vol. 6, Cambridge University Press, Cambridge, 2000.

[30] A. W. Van der Vaart and J. A. Wellner, *Weak Convergence and Empirical Processes with Applications to Statistics*, Springer, New York, 1996.

[31] G. Wahba, *Spline Models for Observational Data*, Vol. 59, SIAM, Philadelphia, 1990.

[32] H. Wendland, *Scattered Data Approximation*, Vol. 17, Cambridge University Press, Cambridge, 2004.

[33] C. F. J. Wu and M. S. Hamada, *Experiments: Planning, Analysis, and Optimization*, Vol. 552, John Wiley & Sons, New York, 2009.