A theoretical framework of the scaled Gaussian stochastic process in prediction and calibration

Mengyang Gu^{*}, Fangzheng Xie^{**} and Long Wang^{**}

*Department of Statistics and Applied Probability, UC Santa Barbara **Department of Applied Mathematical and Statistics, Johns Hopkins University

Abstract

Model calibration or data inversion is one of fundamental tasks in uncertainty quantification. In this work, we study the theoretical properties of the scaled Gaussian stochastic process (S-GaSP), to model the discrepancy between reality and imperfect mathematical models. We establish the explicit connection between Gaussian stochastic process (GaSP) and S-GaSP through the orthogonal series representation. The predictive mean estimator in the S-GaSP calibration model converges to the reality at the same rate as the GaSP with a suitable choice of the regularization and scaling parameters. We also show the calibrated mathematical model in the S-GaSP calibration converges to the one that minimizes the L_2 loss between the reality and mathematical model, whereas the GaSP model with other widely used covariance functions does not have this property. Numerical examples confirm the excellent finite sample performance of our approaches compared to a few recent approaches.

KEYWORDS: model misspecification; Bayesian prior; scaled Gaussian stochastic process prior; convergence; interpretability; orthogonal series representation.

1 Introduction

In scientific and engineering studies, mathematical models are developed by scientists and engineers based on their expert knowledge to reproduce the physical reality. With the rapid development of the computational technique in recent years, many mathematical models are implemented in computer code, often referred as computer models or simulators.

Some parameters of the mathematical model are often unknown or unobservable in experiments. For example, the Kīlauea volcano recently has one of the biggest eruptions in 2018. The location and volume of the magma chamber, as well as the magma supply and storage rate of this volcano, however, is unobservable. Some field data, such as the satellite interferograms and GPS measurement of the ground deformation were used to estimate these parameters for the Kīlauea volcano (Anderson and Poland, 2017; Anderson et al., 2019). Using the field observations to estimate the parameters in the mathematical model, and to identify the possible discrepancy between the mathematical model and the reality is widely known as the model calibration or data inversion.

For any observable input $\mathbf{x} \in \mathcal{X}$, denote $y^F(\mathbf{x})$ as the field observation and $f^M(\mathbf{x}, \boldsymbol{\theta})$ as a mathematical model with unobservable calibration parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Furthermore, let $y^R(\mathbf{x}) = \mathbb{E}[y^F(\mathbf{x})]$ represent the reality. A routinely used framework to calibrate imperfect mathematical model is

$$y^{F}(\mathbf{x}) = f^{M}(\mathbf{x}, \boldsymbol{\theta}) + \delta(\mathbf{x}) + \epsilon, \qquad (1)$$

where ϵ is the noise and $\delta(\cdot)$ is a discrepancy function between the reality and mathematical model. Since the mathematical model is often developed by experts, we assume the mean and trend of the observations are properly modeled in the mathematical model. The discrepancy function was modeled as a Gaussian stochastic process (GaSP) in Kennedy and O'Hagan (2001) and the framework has been widely studied in recent years (Goldstein and Rougier, 2004; Bayarri et al., 2007; Higdon et al., 2008). As both the mathematical model and discrepancy function are jointly estimated in the calibration, the predictions of the field data were found to be more accurate compared to the ones based on the mathematical model or a nonparametric regression alone. It was found in following-up studies that, however, the variability of the observations can be explained mostly by the discrepancy function in this approach, leaving the calibrated mathematical model far away from the reality, which results in an identifiability problem between the calibration parameters and discrepancy function (Tuo and Wu, 2016; Plumlee, 2017).

A few recent studies measure the goodness of calibration in terms of the L_2 loss between the calibrated mathematical model and reality (Tuo and Wu, 2015; Wong et al., 2017). These studies seek to find an estimator of $\boldsymbol{\theta}$ that converges to $\boldsymbol{\theta}_{L_2}$, which minimizes the L_2 distance between the reality and mathematical model, i.e.,

$$\boldsymbol{\theta}_{L_2} := \operatorname{argmin}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \int_{\mathbf{x} \in \mathcal{X}} [y^R(\mathbf{x}) - f^M(\mathbf{x}, \boldsymbol{\theta})]^2 d\mathbf{x}.$$
(2)

In Tuo and Wu (2015), for instance, the reality is first estimated through a nonparametric regression model without the assistance of the mathematical model. The calibration parameters are then estimated by minimizing the L_2 loss between the calibrated mathematical model and the estimator of the reality. Consequently, the calibrated mathematical model by the two-step approach typically fits the observation in terms of L_2 distance. For some complex applications, however, it is crucial to jointly estimate the reality and calibration parameters, as the mathematical model is often developed based on expert knowledge, and thus helpful for predicting the reality.

In this work, we study the theoretical properties of the scaled Gaussian stochastic process (S-GaSP), a new approach for modeling the discrepancy function proposed in Gu and Wang (2018). We establish the connection between GaSP and S-GaSP through the orthogonal representation of the process. We show that the predictive mean from S-GaSP converges to the reality at the same rate as the one from GaSP with a suitable choice of the regularization

and scaling parameters. Furthermore, with the same regularization and scaling parameters, the calibration parameters in the S-GaSP can also converge to θ_{L_2} , whereas GaSP calibration with other widely used kernels does not enjoy this property. Although these two convergence properties can be achieved using the aforementioned two-step approaches (Tuo and Wu, 2015; Wong et al., 2017), finite sample studies suggest that the predictive accuracy of the reality improves in the S-GaSP calibration, as the calibration parameters and discrepancy are jointly estimated. Besides, since the sampling model is fully specified, the model and parameter uncertainty in the S-GaSP calibration can be naturally assessed through the posterior distributions in a Bayesian approach. A close comparison of S-GaSP and other approaches is detailed in Section 5.

This paper is organized as follows. In Section 2, we introduce the S-GaSP along with the orthogonal series representation and joint estimation in calibration. Two convergence properties are discussed in Section 3. In Section 4, we introduce the discretized S-GaSP along with the parameter estimation under the Frequentist framework and Bayesian framework. A comparison between the S-GaSP calibration with other alternatives are discussed in Section 5. Section 6 provides some numerical studies comparing the S-GaSP calibration approach and other approaches. We conclude this work in Section 7. The proof of the theoretical results and other supporting evidence of our approaches are given in the supplementary materials. The GaSP calibration, S-GaSP calibration and calibration without a discrepancy function are implemented in the **RobustCalibration R** package available on CRAN.

2 The scaled Gaussian stochastic process

Denote $\delta(\cdot) \sim \text{GaSP}(0, \sigma^2 K(\cdot, \cdot))$ with variance σ^2 and correlation function $K(\cdot, \cdot)$ such that, for any inputs $\{\mathbf{x}_i\}_{i=1}^n$, the marginal distribution $(\delta(\mathbf{x}_1), ..., \delta(\mathbf{x}_n))^T$ follows a multivariate normal distribution with covariance $\text{Cov}(\mathbf{x}_i, \mathbf{x}_j) = \sigma^2 K(\mathbf{x}_i, \mathbf{x}_j)$. In order to have the mathematical model explain more variability, we introduce a new prior distribution of the

discrepancy function, which places more probability mass on the smaller random L_2 distance between the mathematical model and reality, as this measure quantifies how well a mathematical model fits the reality. The scaled Gaussian stochastic process calibration model is defined as the following hierarchical model:

$$y^{F}(\mathbf{x}) = f^{M}(\mathbf{x}, \boldsymbol{\theta}) + \delta_{z}(\mathbf{x}) + \epsilon,$$

$$\delta_{z}(\mathbf{x}) = \left\{ \delta(\mathbf{x}) \mid \int_{\boldsymbol{\xi} \in \mathcal{X}} \delta^{2}(\boldsymbol{\xi}) d\boldsymbol{\xi} = Z \right\},$$

$$\delta(\cdot) \sim \text{GaSP}(0, \sigma^{2}K(\cdot, \cdot)),$$

$$Z \sim p_{Z}(\cdot), \quad \epsilon \sim \text{N}(0, \sigma_{0}^{2}),$$

(3)

where conditional on all parameters, the default choice of $p_Z(\cdot)$ is defined as

$$p_Z(z) = \frac{g_Z(z)p_\delta(Z=z)}{\int_0^\infty g_Z(t)p_\delta(Z=t)dt},\tag{4}$$

with $g_Z(z)$ being a non-increasing scaling function and $p_{\delta}(Z=z)$ being the density of Z at z induced by a GaSP with mean 0 and covariance $\sigma^2 K(\cdot, \cdot)$.

We call $\delta_z(\cdot)$ in (3) the scaled Gaussian stochastic process (S-GaSP). Given Z = z, the S-GaSP becomes a GaSP constrained at the space $\int_{\mathbf{x}\in\mathcal{X}} \delta^2(\mathbf{x}) d\mathbf{x} = z$. Note that Z is the L_2 distance between the reality and mathematical model. By construction, the measure for Z induced by S-GaSP has more prior probability mass near 0 than the one by GaSP as $g(\cdot)$ is a non-decreasing function, reflecting one's belief that the mathematical model should be calibrated to fit the reality.

It is easy to see that when $g_Z(\cdot)$ is a constant function, S-GaSP reduces to GaSP without any constraint. Conditioning on all parameters, we assume

$$g_Z(z) = \frac{\lambda_z}{2\sigma^2} \exp\left(-\frac{\lambda_z z}{2\sigma^2}\right),\tag{5}$$

with a scaling parameter λ_z . We select $p_Z(\cdot)$ in (4) and $g_Z(\cdot)$ in (5) for the computational

results in this work, as any marginal distribution of δ_z still follows a multivariate normal distribution (Gu and Wang, 2018, Lemma 2.3). Other scaling functions may also be used, but we do not pursue this direction in this study.

2.1 Orthogonal series representation and marginal distribution

Based on Karhunen-Loève theorem, GaSP with a stationary kernel admits the following representation for any $\mathbf{x} \in \mathcal{X}$

$$\delta(\mathbf{x}) = \sigma \sum_{k=1}^{\infty} \sqrt{\rho_k} Z_k \phi_k(\mathbf{x}), \tag{6}$$

where $Z_k \overset{i.i.d.}{\sim} N(0,1)$, ρ_k and $\phi_k(\cdot)$ are the *k*th eigenvalue and eigenfunction of the kernel $K(\cdot, \cdot)$, respectively. The S-GaSP can also be represented as an orthogonal series given below.

Lemma 1 (Karhunen-Loève expansion for the S-GaSP). Assume $p_Z(\cdot)$ and $g_Z(\cdot)$ are defined in (4) and (5), respectively. For any $\mathbf{x} \in \mathcal{X}$, the S-GaSP defined in (3) has the following representation

$$\delta_z(\mathbf{x}) = \sigma \sum_{k=1}^{\infty} \sqrt{\frac{\rho_k}{1 + \lambda_z \rho_k}} Z_k \phi_k(\mathbf{x}),$$

where $Z_k \stackrel{i.i.d.}{\sim} N(0,1)$, ρ_k and $\phi_k(\cdot)$ are the kth eigenvalue and eigenfunction of the kernel $K(\cdot, \cdot)$, respectively.

The covariance function of the S-GaSP can also be decomposed as an infinite orthogonal series, which is an immediate consequence of the fact that the S-GaSP is indeed a GaSP with a transformed kernel (see Lemma 2.3 in Gu and Wang (2018) and Lemma 1).

Corollary 1. Assume $p_Z(\cdot)$ and $g_Z(\cdot)$ are defined in (4) and (5), respectively. The marginal distribution of the S-GaSP defined in (3) follows a multivariate normal distribution

$$[\delta_z(\mathbf{x}_1), ..., \delta_z(\mathbf{x}_n) \mid \sigma^2 \mathbf{R}_z] \sim MN(\mathbf{0}, \sigma^2 \mathbf{R}_z),$$

where the (i, j) entry of \mathbf{R}_z is

$$K_z(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k=1}^{\infty} \frac{\rho_k}{1 + \lambda_z \rho_k} \phi(\mathbf{x}_i) \phi(\mathbf{x}_j).$$
(7)

Corollary 1 implies that the *i*th eigenvalue of the kernel function $K_z(\cdot, \cdot)$ in the S-GaSP is $\rho_{z,k} := \rho_k/(1 + \lambda_z \rho_k)$ and the *k*th eigenfunction $\phi_k(\cdot)$ is the same as the one in the GaSP. The form (7) does not give an explicit expression for the kernel in the S-GaSP. Instead of truncating the series, one may discretize the integral $\int_{\mathbf{x} \in \mathcal{X}} \delta^2(\mathbf{x}) d\mathbf{x}$, which leads to an explicit expression of the covariance matrix, discussed in Section 4.

The following Corollary 2 provides a decomposition of Z in the S-GaSP, which follows from Lemma 2.1 in Gu and Wang (2018) and Corollary 1.

Corollary 2. Assume the same conditions in Lemma 1 hold. The distribution of $Z = \int_{\mathbf{x} \in \mathcal{X}} \delta^2(\mathbf{x}) d\mathbf{x}$ induced by the S-GaSP follows

$$Z \sim \sigma^2 \sum_{k=1}^{\infty} \frac{\rho_k}{1 + \lambda_z \rho_k} \chi_k^2(1),$$

where $\{\chi_k^2(1)\}_{k=1}^{\infty}$ are independent chi-squared random variables with one degree of freedom.

Denote \mathcal{H} and \mathcal{H}_z as the reproducing kernel Hilbert space attached to GaSP with kernel $K(\cdot, \cdot)$ and S-GaSP with kernel $K_z(\cdot, \cdot)$, respectively. Let the native norm associated with $K(\cdot, \cdot)$ and $K_z(\cdot, \cdot)$ be $\langle \cdot, \cdot \rangle_H$ and $\langle \cdot, \cdot \rangle_{H_z}$, respectively. We conclude this subsection by the explicit connection between the inner product of GaSP and that of S-GaSP.

Lemma 2. Assume $p_Z(\cdot)$ and $g_Z(\cdot)$ are defined in (4) and (5), respectively. Let $h(\cdot) = \sum_{i=1}^{\infty} h_i \phi_i(\cdot)$ and $g(\cdot) = \sum_{i=1}^{\infty} g_i \phi_i(\cdot)$ be the elements in \mathcal{H} . It holds that

$$\langle h, g \rangle_{\mathcal{H}_z} = \langle h, g \rangle_{\mathcal{H}} + \lambda_z \langle h, g \rangle_{L_2(\mathcal{X})}.$$

2.2 Joint estimation in the S-GaSP calibration

With the specification of $p_Z(\cdot)$ in (4) and $g_Z(\cdot)$ in (5), after marginalizing out $\boldsymbol{\delta}_z = [\delta_z(\mathbf{x}_1), ..., \delta_z(\mathbf{x}_n)]$, the marginal distribution of the field observations in (3) follows a multivariate normal distribution

$$[\mathbf{y}^F \mid \boldsymbol{\theta}, \, \sigma_0^2, \, \lambda, \, \lambda_z] \sim \mathrm{MN}(\mathbf{f}_{\boldsymbol{\theta}}^M, \, \sigma_0^2((n\lambda)^{-1}\mathbf{R}_z + \mathbf{I}_n)) \tag{8}$$

with the regularization parameter $\lambda := \sigma_0^2/(n\sigma^2)$ and the (i, j) entry of \mathbf{R}_z defined in (7).

Denote $\mathcal{L}_z(\boldsymbol{\theta})$ as the likelihood for $\boldsymbol{\theta}$ in (8). We show below that the following joint estimator of $(\boldsymbol{\theta}, \delta_z(\cdot))$ can be written as a penalized kernel ridge regression estimator (KRR), where both the RKHS norm and L_2 norm of the discrepancy function are penalized simultaneously.

Lemma 3. The maximum likelihood estimator $\hat{\theta}_{\lambda,\lambda_z,n} := \operatorname{argmax}_{\theta \in \Theta} \mathcal{L}_z(\theta)$ and predictive mean $\hat{\delta}_{\lambda,\lambda_z,n}(\cdot) := \mathbb{E}[\delta_z(\cdot) \mid \mathbf{y}^F, \hat{\theta}_{\lambda,\lambda_z,n}, \lambda, \lambda_z]$ are the same as the estimators of the penalized KRR,

$$(\hat{\boldsymbol{\theta}}_{\lambda,\lambda_{z},n},\hat{\delta}_{\lambda,\lambda_{z},n}(\cdot)) = \operatorname*{argmin}_{\delta(\cdot)\in\mathcal{H},\,\boldsymbol{\theta}\in\boldsymbol{\Theta}} \frac{1}{n} \sum_{i=1}^{n} (y^{F}(\mathbf{x}_{i}) - f^{M}(\mathbf{x}_{i},\boldsymbol{\theta}) - \delta(\mathbf{x}_{i}))^{2} + \lambda \|\delta\|_{\mathcal{H}_{z}}^{2}$$
(9)

with $\|\delta\|_{\mathcal{H}_z}^2 = \|\delta\|_{\mathcal{H}}^2 + \lambda_z \|\delta\|_{L_2(\mathcal{X})}^2$.

In Lemma 3, both the L_2 norm and native norm of the discrepancy function are penalized in S-GaSP calibration. When the discrepancy function is modeled as a GaSP, however, the L_2 norm of the discrepancy function is not penalized (see supplementary materials). This property of the S-GaSP calibration is the key to guarantee that, under some regularity conditions, the estimated calibration parameters in joint estimator converges to θ_{L_2} defined in (2). A more detailed discussion is provided in Section 3.

3 Convergence properties of the S-GaSP calibration

We discuss two convergence properties of the S-GaSP calibration in this section. First, the predictive mean estimator of the reality converges to the truth at the optimal rate with a suitable choice of the regularization and scaling parameters. Second, the estimated calibration parameters by S-GaSP calibration converge to θ_{L_2} when sample size increases. These two properties are obtained by jointly estimating the discrepancy function and calibration parameters in (9).

3.1 Convergence to the reality

Let us first consider the following nonparametric regression model,

$$y(\mathbf{x}_i) = f(\mathbf{x}_i) + \epsilon_i, \quad \epsilon_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma_0^2), \quad i = 1, \dots, n,$$
(10)

where f is assumed to follow the zero-mean S-GaSP prior with the default choice of $p_Z(\cdot)$ and $g_Z(\cdot)$ in (4) and (5), respectively. This is a special case where the mathematical model is zero and we will soon extend it to the general case when the mathematical model is not zero. For illustration purposes, we follow Tuo and Wu (2015) to assume that $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are independently sampled from Unif([0, 1]^p).

Assume the underlying truth $f_0(\cdot) := \mathbb{E}_y[y(\cdot)]$ resides in the *p*-dimensional Sobolev space

$$\mathcal{W}_2^m(\mathcal{X}) = \left\{ f(\cdot) = \sum_{k=1}^\infty f_k \phi_k(\cdot) \in L_2(\mathcal{X}) : \sum_{k=1}^\infty k^{2m/p} f_k^2 < \infty \right\}$$
(11)

with smoothness m > p/2 and $\{\phi_k(\cdot)\}_{k=1}^{\infty}$ being a sequence of the orthonormal basis of $L_2(\mathcal{X})$. For any integer vector $\mathbf{k} = (k_1, ..., k_p)^T$ and a function $f(x_1, ..., x_p) : \mathcal{X} \to \mathbb{R}$, denote $D^{\mathbf{k}}$ the mixed partial derivative operator $D^{\mathbf{k}}f(\cdot) := \partial^{|\mathbf{k}|}f(\cdot)/\partial^{k_1}x_1...\partial^{k_p}x_p$ with $|\mathbf{k}| = \sum_{i=1}^p k_i$. For any function in $\mathcal{W}_2^m(\mathcal{X})$, we have $\|D^{\mathbf{k}}f(\cdot)\|_{L_2(\mathcal{X})} < \infty$ for any $|\mathbf{k}| < m$.

Recall that $\lambda = \sigma_0^2/(n\sigma^2)$ in (8). By Lemma 3, the posterior mean estimator of $f(\cdot)$ with

a S-GaSP prior is equivalent to the KRR estimator below

$$\hat{f}_{\lambda,\lambda_z,n} = \operatorname*{argmin}_{f \in \mathcal{H}} \left[\frac{1}{n} \sum_{i=1}^n (y(\mathbf{x}_i) - f(\mathbf{x}_i))^2 + \lambda \|f\|_{\mathcal{H}}^2 + \lambda \lambda_z \|f\|_{L_2(\mathcal{X})}^2 \right].$$
(12)

Recall $\{\rho_k\}_{k=1}^{\infty}$ and $\{\phi_k\}_{k=1}^{\infty}$ are the sequence of the eigenvalues and eigenfunctions of the reproducing kernel $K(\cdot, \cdot)$ associated with \mathcal{H} , respectively. For all k, we assume the eigenvalues satisfy

$$c_{\rho}k^{-2m/p} \le \rho_k \le C_{\rho}k^{-2m/p},\tag{13}$$

for some constants c_{ρ} and $C_{\rho} > 0$. For all $k \in \mathbb{N}^+$ and $\mathbf{x} \in \mathcal{X}$, we assume the eigenfunctions are bounded uniformly,

$$\sup_{\mathbf{x}\in\mathcal{X}} |\phi_k(\mathbf{x})| \le C_{\phi},\tag{14}$$

where $C_{\phi} > 0$ is a constant depending on the kernel $K(\cdot, \cdot)$.

We are now ready to state the convergence rate of the S-GaSP for the nonparametric regression model in (10).

Theorem 1. Assume the eigenvalues and eigenfunctions of $K(\cdot, \cdot)$ satisfy (13) and (14), respectively. Further assume $f_0 \in W_2^m(\mathcal{X})$ and denote $\beta := (2m - p)^2 / \{2m(2m + p)\}$. Consider the nonparametric regression model (10). For sufficiently large n, any $\alpha > 2$ and $C_\beta \in (0, 1)$, with probability at least $1 - \exp\{-(\alpha - 2)/3\} - \exp(-n^{C_\beta\beta})$,

$$\|\hat{f}_{\lambda,\lambda_{z},n} - f_{0}\|_{L_{2}(\mathcal{X})} \leq 2\left[\sqrt{2}\left(\|f_{0}\|_{L_{2}(\mathcal{X})} + \|f_{0}\|_{\mathcal{H}}\right) + C_{K}\sigma_{0}\sqrt{\alpha}\right]n^{-\frac{m}{2m+\mu}}$$

and

$$\|\hat{f}_{\lambda,\lambda_z,n} - f_0\|_{\mathcal{H}} \le 2\left[\sqrt{2}\left(\|f_0\|_{L_2(\mathcal{X})} + \|f_0\|_{\mathcal{H}}\right) + C_K \sigma_0 \sqrt{\alpha}\right]$$

by choosing $\lambda = n^{-2m/(2m+p)}$ and $\lambda_z = \lambda^{-1/2}$, where C_K is a constant only depending on the kernel $K(\cdot, \cdot)$.

The proof of Theorem 1 is more challenging compared to the proof of convergence of Gaussian stochastic process regression in Yang et al. (2017a). First λ_z can go to infinity in Theorem 1, and consequently $||\delta||_{\mathcal{H}_z}$ is unbounded, wheres $||\delta||_{\mathcal{H}}$ was typically assumed bounded in proving the convergence of a nonparametric regression approach. Second we generalize the proof to multivariate inputs. Thus, we substantially modify the tools to prove Theorem 1, given in the supplementary materials.

The conditions in Theorem 1 can be relaxed in various ways. From the proof of Theorem 1, it is easy to see that if $\lambda = O(n^{-2m/(2m+p)})$ and $\lambda_z \leq O(\lambda^{-1/2})$, the estimator still converges to the truth in L_2 distance with the same rate $O(n^{-m/(2m+p)})$. Second the design can be generalized to other space filling design. Besides, although the stationarity of the process is often assumed for the computational purpose, it is not required in Theorem 1. Note the regularity parameter and kernel parameters are held fixed in Theorem 1. We discuss the estimation of λ and the parameters in the kernel function in Section 4.1.

We are ready to discuss the convergence of estimating the reality in the calibration. The estimator for the reality in the S-GaSP calibration model is defined as follows

$$\hat{y}^R_{\lambda,\lambda_z,n}(\mathbf{x}) := f^M(\mathbf{x}, \hat{\boldsymbol{\theta}}_{\lambda,\lambda_z,n}) + \hat{\delta}_{\lambda,\lambda_z,n}(\mathbf{x})$$

for any $\mathbf{x} \in \mathcal{X}$, where $(\hat{\boldsymbol{\theta}}_{\lambda,\lambda_z,n}, \hat{\delta}_{\lambda,\lambda_z,n})$ is the estimator of the penalized KRR obtained by minimizing the loss in (9). The following Corollary 3 gives the convergence rate of the S-GaSP calibration model in predicting the reality. Similar to the extensions for Theorem 1, the conditions in Corollary 3 can be relaxed by letting $\lambda = O(n^{-2m/(2m+p)})$ and $\lambda_z \leq O(\lambda^{-1/2})$ to obtain the same convergence rate.

Corollary 3. Assume $y^{R}(\cdot) - f^{M}(\cdot, \theta) \in \mathcal{W}_{2}^{m}(\mathcal{X})$ for any $\theta \in \Theta$ and $\sup_{\theta \in \Theta} ||y^{R}(\cdot) - f^{M}(\cdot, \theta)||_{\mathcal{H}} < \infty$. Let the eigenvalues and eigenfunctions of $K(\cdot, \cdot)$ satisfy (13) and (14), respectively. For sufficiently large n, any $\alpha > 2$ and $C_{\beta} \in (0, 1)$, with probability at least

 $1-\exp\{-(\alpha-2)/3\}-\exp(-n^{C_\beta\beta}),$

$$\begin{aligned} \|\hat{y}_{\lambda,\lambda_{z},n}^{R}(\cdot) - y^{R}(\cdot)\|_{L_{2}(\mathcal{X})} &\leq 2 \bigg[\sqrt{2} \bigg(\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|y^{R}(\cdot) - f^{M}(\cdot,\boldsymbol{\theta})\|_{L_{2}(\mathcal{X})} \\ &+ \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|y^{R}(\cdot) - f^{M}(\cdot,\boldsymbol{\theta})\|_{\mathcal{H}}\bigg) + C_{K}\sigma_{0}\sqrt{\alpha}\bigg] n^{-\frac{m}{2m+p}}. \end{aligned}$$

by choosing $\lambda = n^{-2m/(2m+p)}$ and $\lambda_z = \lambda^{-1/2}$, where C_K is a constant only depending on the kernel $K(\cdot, \cdot)$ and $\beta = (2m - p)^2/(2m(2m + p))$.

We illustrate the convergence using the following function studied in Yang et al. (2017a), where $y^{R}(\cdot)$ lies in the Sobolev space $\mathcal{W}_{2}^{m}(\mathcal{X})$ with m = 3 and $\mathcal{X} = [0, 1]$.

Example 1. Let the reality be $y^{R}(x) = 2 \sum_{j=1}^{\infty} j^{-6} \cos(5\pi(j-0.5)x) \sin(5j)$, and consider $y^{F}(x) = y^{R}(x) + \epsilon$, where $\epsilon \sim N(0, 0.05^{2})$ independently. Let $f^{M}(x, \theta) = \theta$. The goal is to predict $y^{R}(x)$ at $x \in [0, 1]$ and estimate θ .

As a motivating example, we let $K(\cdot, \cdot)$ follow the Matérn kernel in (19) with the range parameter $\gamma = 1$, as the reproducing kernel Hilbert space attached to the GaSP with this kernel is equal to Sobolev space $\mathcal{W}_2^3(\mathcal{X})$. We test 50 configurations with the number of observations $n \in [\exp(5), \exp(10)]$, and the design points $\{x_i\}_{i=1}^n$ are equally spaced in [0, 1]. In each configuration, N = 100 simulation replicates are implemented. We first compute the average root of the mean squared error below:

AvgRMSE<sub>*f*^{*M*}+
$$\delta$$</sub> = $\frac{1}{N} \sum_{i=1}^{N} \sqrt{\frac{1}{n^*} \sum_{j=1}^{n^*} (\hat{y}_i^R(\mathbf{x}_j^*) - y_i^R(\mathbf{x}_j^*))^2},$ (15)

where $\hat{y}_i^R(\mathbf{x}_j^*)$ is an estimator of the reality at \mathbf{x}_j^* for $j = 1, ..., n^*$. The subscript $f^M + \delta$ indicates both the calibrated mathematical model and discrepancy function can be used for prediction.

For both GaSP and S-GaSP calibration, the joint estimator, i.e. the predictive mean of the reality and maximum marginal likelihood estimator of the calibration parameters



Figure 1: Prediction and calibration by the GaSP and discretized S-GaSP calibration models for Example 1. In the left panel, the AvgRMSE_{$f^{M}+\delta$} of the GaSP calibration and that of the discretized S-GaSP calibration are graphed as the red triangles and blue dots, respectively; the black curve is $n^{-m/(2m+p)}/5$, representing the upper bound by Corollary 3 (up to a constant). In the right panel, the natural logarithm of the RMSE_{θ} of the GaSP calibration and that of the discretized S-GaSP calibration are graphed as the red triangles and blue circles, respectively; the black line is $\log(n^{-m/(2m+p)}/40)$, the upper bound from Theorem 2 (up to a constant). $\lambda = n^{-2m/(2m+p)} \times 10^{-4}$ and $\lambda_z = \lambda^{-1/2}$ are assumed.

discussed in Lemma 3, is implemented for each experiment at each configuration. In the left panel of Figure 1, the predictive mean estimator of the reality in both GaSP and S-GaSP estimator converges to the reality at the same rate, which matches the theoretical upper bound from Corollary 3. Here for computational purpose, we graph the results of discretized S-GaSP calibration, which replaces the integral $\int_{\mathbf{x}\in\mathcal{X}} \delta^2(\mathbf{x})d\mathbf{x}$ in the S-GaSP model in (3) by $(1/n)\sum_{i=1}^n \delta^2(\mathbf{x}_i)$ in Figure 1. The discretized S-GaSP calibration is discussed in Section 4.

To evaluate whether the calibrated mathematical model (here only a mean parameter) fits the data, we use the root of the mean squared error between the estimator of the calibration parameters and the L_2 minimizer $\boldsymbol{\theta}_{L_2}$ as follows $\text{RMSE}_{\boldsymbol{\theta}} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_{L_2})^2}$, where $\hat{\boldsymbol{\theta}}_i$ is an estimator of $\boldsymbol{\theta}$ in the *i*th experiment.

Although the GaSP and the S-GaSP perform equally well in prediction for Example 1, the estimator of the calibration parameter in the discretized S-GaSP calibration converges to the L_2 minimizer, but that in the GaSP calibration does not converge to θ_{L_2} , shown in the right panel of Figure 1. This problem is caused by the difference between the RKHS norm and the L_2 norm. As illustrated in Lemma 3, both the RKHS norm and L_2 norm of the discrepancy function are penalized in the S-GaSP calibration model, whereas the GaSP calibration model does not penalize the L_2 norm of the discrepancy function. In Section 3.2, we further show that under some regularity conditions, the calibrated parameters in the S-GaSP calibration do converge to the L_2 minimizer with the same choice of the regularization parameter and scaling parameter in Corollary 3.

3.2 Convergence of calibration parameters

We first list some regularity conditions for the convergence of calibration parameters.

A1 $\boldsymbol{\theta}_{L_2}$ is the unique solution of (2) and it is an interior point of $\boldsymbol{\Theta}$.

- A2 The Hessian matrix $\int \frac{\partial^2 (y^R(\mathbf{x}) f^M(\mathbf{x}, \boldsymbol{\theta}))^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} d\mathbf{x}$ is invertible in a neighborhood of $\boldsymbol{\theta}_{L_2}$.
- A3 For all j = 1, ..., q, it holds that $\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left\| \frac{\partial f^{M}(\cdot, \boldsymbol{\theta})}{\partial \theta_{j}} \right\|_{\mathcal{H}} < \infty$.
- A4 The function class $\{y^R(\cdot) f^M(\cdot, \theta) : \theta \in \Theta\}$ is Donsker.
- A5 $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|y^{R}(\cdot) f^{M}(\cdot,\boldsymbol{\theta})\|_{\mathcal{H}} < \infty.$
- A6 The eigenvalues and eigenfunctions of $K(\cdot, \cdot)$ satisfy (13) and (14), respectively.

Assumptions A1 to A3 are regularity conditions of θ_{L_2} and the mathematical model $y^M(\cdot, \theta)$ around θ_{L_2} . Assumptions A4 to A6 guarantee the KRR estimator $\hat{\delta}_{z,\theta}$ converges to $y^R(\cdot) - f^M(\cdot, \theta)$ uniformly for each $\theta \in \Theta$ in terms of the L_2 loss. We have the following result in Theorem 2 that guarantees the convergence of calibration parameters. As the calibration parameters and discrepancy function are estimated jointly in our approach, we extend the tools of proving the convergence of the two-step calibration approach (Tuo and Wu, 2015) to prove Theorem 2, given in the supplementary materials.

Theorem 2. Under assumptions A1 to A6, the estimated parameters in (9) follow

$$\hat{\boldsymbol{\theta}}_{\lambda,\lambda_z,n} = \boldsymbol{\theta}_{L_2} + O_p(n^{-\frac{m}{2m+p}}),$$

by choosing $\lambda = O(n^{-2m/(2m+p)})$ and $\lambda_z = O(\lambda^{-1/2})$.

Note that $\lambda = O(n^{-2m/(2m+p)})$ and $\lambda_z = O(\lambda^{-1/2})$ also guarantee the predictive mean estimator in the S-GaSP calibration converges to the reality at the rate $O(n^{-m/(2m+p)})$ in terms of the L_2 loss. The convergence rate of the calibration parameter is slightly slower than $O(n^{-1/2})$ obtained in the two-step approach (Tuo and Wu, 2015). Though the $O(n^{-1/2})$ rate may be obtained by choosing $\lambda = O(n^{-2m/(2m+p)})$ and $\lambda_z = O(\lambda^{-1/2})$, we should be aware that, however, the L_2 minimizer is not the true calibration parameter, but the one that minimizes the L_2 distance between the calibrated mathematical model and reality. At the finite sample scenarios, the residuals between reality and calibrated mathematical model with the L_2 minimizer may behave like noises, rather than a smooth function, which may be hard to be accurately estimated by a nonparametric model of the discrepancy function. In comparison, the joint estimate of the discrepancy function and calibration parameters was found have a smaller predictive error in Example 2, 3 and 4.

On the contrary, the calibrated parameters of the GaSP calibration typically do not converge to the L_2 minimizer. Let $\frac{\partial f^M(\cdot,\hat{\theta})}{\partial \theta_j} := \frac{\partial f^M(\cdot,\theta)}{\partial \theta_j}|_{\theta=\hat{\theta}}$. A key difference between the GaSP and the S-GaSP calibration is stated in the following Corollary 4, which is an immediate consequence from the proof of Theorem 2.

Corollary 4. Under assumptions A1 to A6, the estimator for the calibration parameters in the S-GaSP calibration in (9) satisfies

$$\frac{1}{\lambda_z} \left\langle \hat{\delta}_{\lambda,\lambda_z,n}(\cdot), \frac{\partial f^M(\cdot, \hat{\boldsymbol{\theta}}_{\lambda,\lambda_z,n})}{\partial \theta_j} \right\rangle_{\mathcal{H}} + \left\langle \hat{\delta}_{\lambda,\lambda_z,n}(\cdot), \frac{\partial f^M(\cdot, \hat{\boldsymbol{\theta}}_{\lambda,\lambda_z,n})}{\partial \theta_j} \right\rangle_{L_2(\mathcal{X})} = 0;$$

Further assuming the mathematical model is differentiable at $\hat{\theta}_{\lambda,n}$ in (S7), the estimator of the calibration parameters in the GaSP calibration satisfies

$$\left\langle \hat{\delta}_{\lambda,n}(\cdot), \frac{\partial f^M(\cdot, \hat{\boldsymbol{\theta}}_{\lambda,n})}{\partial \theta_j} \right\rangle_{\mathcal{H}} = 0,$$

for any θ_j , j = 1, ..., q.

To ensure the convergence of an estimator $\hat{\boldsymbol{\theta}}$ to the L_2 minimizer, one typical requirement is that $\langle \hat{\delta}_{L_2}(\cdot), \frac{\partial f^M(\cdot, \hat{\boldsymbol{\theta}})}{\partial \theta_j} \rangle_{L_2(\mathcal{X})} = o_p(1)$. It is easy to see that the S-GaSP satisfies this condition with $1/\lambda_z = o(1)$. However, because of the difference between the RKHS norm and L_2 norm, the estimated parameters $\hat{\boldsymbol{\theta}}_{\lambda,n}$ in the GaSP calibration model can be far away from the L_2 minimizer. As a result, the calibrated mathematical model may not fit the data in the GaSP calibration model, as found in previous studies (Tuo and Wu, 2015; Wong et al., 2017). For Example 1, the estimated parameters in the discretized S-GaSP calibration converges to the L_2 minimizer when the sample size increases, whereas the parameters in GaSP calibration with an unscaled kernel function do not converge, as shown in the right panel of Figure 1.

4 Discretized scaled Gaussian stochastic process

We address the computational issue in the S-GaSP calibration in this section. Instead of truncating the kernel function in (7) by the first several terms, we select N_C distinct points to discretize the input space $[0,1]^p$ and replace $\int_{\boldsymbol{\xi}\in\mathcal{X}}\delta(\boldsymbol{\xi})^2d\boldsymbol{\xi}$ by $(1/N_C)\sum_{i=1}^{N_C}\delta(\mathbf{x}_i^C)^2$ in the S-GaSP model in (3).

Here we let the discretization points be the observed inputs, i.e. $\mathbf{x}_i^C = \mathbf{x}_i$ for $i = 1, ..., N_C$ and $N_C = n$. The discretized S-GaSP is to replace δ_z in Equation (3) by

$$\delta_{z_d}(\mathbf{x}) = \left\{ \delta(\mathbf{x}) \mid \frac{1}{n} \sum_{i=1}^n \delta(\mathbf{x}_i)^2 = Z_d \right\}$$
(16)

with density $p_{Z_d}(\cdot)$ defined in (4). After marginalizing out Z_d , it follows from Lemma 2.4 in Gu and Wang (2018) that $\delta_{z_d}(\cdot)$ is a zero-mean GaSP with the covariance function

$$\sigma^2 K_{z_d}(\mathbf{x}_a, \mathbf{x}_b) = \sigma^2 (K(\mathbf{x}_a, \mathbf{x}_b) - \mathbf{r}^T(\mathbf{x}_a) \tilde{\mathbf{R}}^{-1} \mathbf{r}(\mathbf{x}_b))$$
(17)

for any $\mathbf{x}_a, \mathbf{x}_b \in \mathcal{X}$, where $\tilde{\mathbf{R}} := \mathbf{R} + n\mathbf{I}_n/\lambda_z$.

Recall $\lambda = \sigma_0^2/(n\sigma^2)$. We have the following predictive distribution of the discretized S-GaSP calibration model.

Theorem 3. Assume $\delta_{z_d}(\cdot)$ in (16) with $p_{Z_d}(\cdot)$ and $g_{Z_d}(\cdot)$ defined in (4) and (5), respectively. The predictive distribution of the field data at any $\mathbf{x} \in \mathcal{X}$ by the discretized S-GaSP calibration model in (16) is a multivariate normal distribution

$$y^F(\mathbf{x}) \mid \mathbf{y}^F, \boldsymbol{\theta}, \sigma_0^2, \lambda, \lambda_z \sim \mathit{MN}(\hat{\mu}_{z_d}(\mathbf{x}), \sigma_0^2((n\lambda)^{-1}K^*_{z_d}(\mathbf{x}, \mathbf{x}) + 1)),$$

where $\hat{\mu}_{z_d}(\mathbf{x}) = f^M(\mathbf{x}, \boldsymbol{\theta}) + \frac{\mathbf{r}^T(\mathbf{x})}{1+\lambda\lambda_z} \left(\mathbf{R} + \frac{n\lambda}{1+\lambda\lambda_z}\mathbf{I}_n\right)^{-1} \left(\mathbf{y}^F - \mathbf{f}^M_{\boldsymbol{\theta}}\right), \text{ and } K^*_{z_d}(\mathbf{x}, \mathbf{x}) = K(\mathbf{x}, \mathbf{x}) - \mathbf{r}^T(\mathbf{x}) \left[\mathbf{I}_n + \left(\mathbf{R} + \frac{n\lambda}{1+\lambda\lambda_z}\mathbf{I}_n\right)^{-1} \frac{n}{(1+\lambda\lambda_z)\lambda_z}\right] \tilde{\mathbf{R}}^{-1}\mathbf{r}(\mathbf{x}), \text{ with } \mathbf{r}(\mathbf{x}) = (K(\mathbf{x}, \mathbf{x}_1), ..., K(\mathbf{x}, \mathbf{x}_n))^T \text{ and} \\ \tilde{\mathbf{R}} = \mathbf{R} + \frac{n}{\lambda_z}\mathbf{I}_n \text{ with the } (i, j) \text{ entry of } \mathbf{R} \text{ being } K(\mathbf{x}_i, \mathbf{x}_j).$

Theorem 3 indicates that the predictive mean of the discretized S-GaSP calibration model shrinks the predictive mean towards the mean function. When $\lambda_z = 0$, the shrinkage is zero and the discretized S-GaSP becomes the GaSP.

Interestingly, when the observations contain no noise, the predictive mean and variance of the field data from the GaSP calibration model and the discretized S-GaSP calibration model are *exactly* the same, stated in the following Lemma 4.

Lemma 4. Assume the conditions in Theorem 3 hold. If $\sigma_0^2 = 0$, the predictive distribution of the field data at any $\mathbf{x} \in \mathcal{X}$ by the discretized S-GaSP model in (16) is a multivariate normal distribution with the predictive mean and variance as follows

$$\mathbb{E}[y^{F}(\mathbf{x}) \mid \mathbf{y}^{F}, \boldsymbol{\theta}, \lambda, \lambda_{z}] = f^{M}(\mathbf{x}, \boldsymbol{\theta}) + \mathbf{r}^{T}(\mathbf{x})\mathbf{R}^{-1}(\mathbf{y}^{F} - \mathbf{f}_{\boldsymbol{\theta}}^{M}),$$
$$\mathbb{V}[y^{F}(\mathbf{x}) \mid \mathbf{y}^{F}, \boldsymbol{\theta}, \lambda, \lambda_{z}] = \sigma^{2}\left(K(\mathbf{x}, \mathbf{x}) - \mathbf{r}^{T}(\mathbf{x})\mathbf{R}^{-1}\mathbf{r}(\mathbf{x})\right),$$

where $\mathbf{r}(\mathbf{x}) = (K(\mathbf{x}, \mathbf{x}_1), ..., K(\mathbf{x}, \mathbf{x}_n))^T$ and the (i, j) entry of \mathbf{R} is $K(\mathbf{x}_i, \mathbf{x}_j)$.

4.1 Emulating slow computer models and parameter estimation

We discuss the computational issue and the parameter estimation in this section. All the approaches are implemented in the new RobustCalibration R package available on CRAN Gu (2018b). First, some mathematical models are the numerical solutions of partial differential equations implemented as computer code, which is computationally expensive to run. In these cases, one often uses a statistical emulator to approximate the computer model based on a set of model runs (Sacks et al., 1989; Santner et al., 2003). The GaSP emulator from the RobustGaSP R package is used to emulate the computer model when it is expensive to run.

We next discuss estimating the regularization parameters and kernel parameters, which were held fixed in some studies. In practice, estimating these parameters can improve predictive performance. For any $\mathbf{x}_a = (x_{a1}, ..., x_{ap})^T$ and $\mathbf{x}_b = (x_{b1}, ..., x_{bp})^T$, the kernel is often assumed to have a product form in model calibration (Kennedy and O'Hagan, 2001):

$$K(\mathbf{x}_a, \mathbf{x}_b) = \prod_{i=1}^p K_i(d_i), \tag{18}$$

where $d_i = |x_{ai} - x_{bi}|$ for i = 1, ..., p, and $K_i(\cdot)$ is a one dimensional kernel function. One widely used kernel function is the Matérn kernel (Handcock and Stein, 1993). When the roughness parameter is a half-integer, the Matérn kernel has an explicit expression. For example, the Matérn kernel with roughness parameter being 5/2 has the following expression

$$K_i(d_i) = \left(1 + \frac{\sqrt{5}d_i}{\gamma_i} + \frac{5d_i^2}{3\gamma_i^2}\right) \exp\left(-\frac{\sqrt{5}d_i}{\gamma_i}\right), \qquad (19)$$

for i = 1, ..., p. A good feature of the Matérn kernel is the sample path of a GaSP is $\lfloor \nu_i - 1 \rfloor$ times differentiable, where ν_i is the roughness parameters.

Denote $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_p)^T$ the unknown range parameters in the covariance. The parameters in the discretized S-GaSP calibration model are the calibration parameters $\boldsymbol{\theta}$, the

variance parameter of the noise σ_0^2 , regularization parameter $\lambda = \sigma_0^2/(n\sigma^2)$, scaling parameter λ_z and range parameters γ . Simple algebra shows $\hat{\sigma}_{0,MLE}^2 = \lambda S_{z_d}^2$, where $S_{z_d}^2 = (\mathbf{y}^F - \mathbf{f}_{\theta}^M)^T \mathbf{\tilde{R}}_{z_d}^{-1} (\mathbf{y}^F - \mathbf{f}_{\theta}^M)$ with $\mathbf{\tilde{R}}_{z_d} = (\mathbf{R}^{-1} + \lambda_z \mathbf{I}_n/n)^{-1} + \lambda n \mathbf{I}_n$ and $\mathbf{\tilde{R}}_{z_d}^{-1} = \lambda_z/(ng) + (\mathbf{R} + n\lambda \mathbf{I}_n/g)^{-1}/g^2$ with $g = \lambda \lambda_z + 1$. Marginalizing out $\delta_z(\cdot)$ and plugging $\hat{\sigma}_{0,MLE}^2$ into the likelihood of the discretized S-GaSP calibration model in (16), one has the profile likelihood

$$\ell_{z_d}(\boldsymbol{\theta}, \boldsymbol{\gamma}, \lambda, \lambda_z) \propto -\frac{1}{2} \log |\tilde{\mathbf{R}}_{z_d}| - \frac{n}{2} \log(S_{z_d}^2).$$
(20)

One may numerically maximize the profile likelihood in (20) to estimate the parameters. Note λ_z reflects one's tolerance of how good a mathematical model should predict the reality without the discrepancy function and thus this parameter may be chosen based on the expert knowledge. Because of the conditions discussed in Theorem 2, λ_z may be fixed to be proportional to $\lambda^{-1/2}$ or be related to the sample size. For all numerical examples, we fix $\lambda_z =$ $(\lambda || \tilde{\gamma} ||)^{-1/2}$, where $\lambda = \sigma_0^2 / (\sigma^2 n)$ and $\tilde{\gamma} = (\tilde{\gamma}_1, ..., \tilde{\gamma}_p)^T$, with $\tilde{\gamma}_i$ being the normalized range parameter (normalized by the length of each coordinate of the input variable). The inclusion of the range parameters is due to the confounding issue between the range parameter and the variance parameter of the process, whereas the ratio of these parameters can typically be estimated consistently from the data (Zhang, 2004).

In the RobustCalibration package, we implement the Bayesian method of estimating the parameters and making predictions. The prior is assumed to follow $\pi(\theta, \gamma, \eta, \sigma^2) \propto \pi(\theta)\pi(\gamma, \eta)/\sigma^2$ with $\eta = \sigma_0^2/\sigma^2$ being the nugget parameter. Here $\pi(\gamma, \eta)$ is chosen as the joint robust prior for the kernel parameters (Gu, 2018a; Gu et al., 2018), and $\pi(\theta)$ may be specified by the user to reflect experts' knowledge. We assume a uniform distribution of $\pi(\theta)$ in the numerical examples for demonstration purposes. Compared to the MLE approach, the uncertainty of the parameters can be assessed naturally through the posterior samples.

5 Comparison between different calibration approaches

We compare a few calibration approaches in this section. First, one of most popular framework is the GaSP calibration approach (Kennedy and O'Hagan, 2001), which models the discrepancy in (1) as a GaSP. The mathematical model and discrepancy function are jointly estimated under the Bayesian framework. The S-GaSP approach is an extended version of GaSP calibration by placing more prior probability mass of the L_2 norm of the discrepancy near zero. Consequently, the calibrated mathematical models in the S-GaSP calibration fit the data better than the ones in the GaSP calibration.

The S-GaSP calibration approach was inspired by a few pioneering approaches seeking to find the L_2 minimizer of the calibration parameters (Tuo and Wu, 2015; Wong et al., 2017; Plumlee, 2017). The orthogonal Gaussian process proposed in Plumlee (2017) constrains the derivatives of the random L_2 norm of the discrepancy to be zeros, equivalently giving more prior probability mass of calibration parameters at the stationary points of the calibration parameters in terms of the L_2 loss. The S-GaSP model explores another transformation that avoids putting more prior probability mass at local maxima of L_2 loss of the discrepancy, and that has a closed-form likelihood function. The L_2 calibration was proposed in (Tuo and Wu, 2015), where the reality is first estimated by nonparametric regression, and the calibration parameters are estimated by minimizing the L_2 loss between the reality and mathematical model. The LS calibration is proposed in (Wong et al., 2017), where the calibration parameters are estimated by first minimizing the squared loss between the mathematical model and observations, and a nonparametric approach is applied to estimate the difference between the reality and mathematical model.

We use the following example to illustrate that jointly estimating the discrepancy function and calibration parameters can be helpful for predicting the reality.

Example 2. Let $y^F(x) = y^R(x) + \epsilon$, where ϵ independently follows $N(0, 0.05^2)$ and $y^R(x) = g_1 + g_2$, with $g_1 = \sum_{j=1}^{\infty} j^{-1} \cos(5\pi(j-0.5)x) \sin(5j)$ and $g_2 = \sum_{j=1}^{\infty} j^{-6} \cos(5\pi(j-0.5)x) \sin(5j)$.



Figure 2: Comparison of different approaches in Example 2. In the left panel, the logarithm of the AvgRMSE_{$f^M+\delta$} of four calibration approaches are graphed at the logarithm of different sample sizes. The histogram of the estimated calibration parameter of each experiment of different approaches are given in the middle panel and the right panel.

Let mathematical model be $f^M(x) = g_1 \theta$. The goal is to predict $y^R(\cdot)$ and estimate θ .

We compare the GaSP, S-GaSP, L_2 and LS calibration approaches using Example 2. The GaSP model is used as the nonparametric regression approach in the first step of L_2 calibration and the second step of the LS calibration. We assume kernel function $K(\cdot, \cdot)$ follows Matérn kernel function in (19) for all methods. The calibration parameter, variance and kernel parameters are estimated by the maximum likelihood estimator.

In the left panel of Figure 2, we found the predictive error by the GaSP and S-GaSP calibration is considerably smaller than the LS calibration and L_2 calibration approach. This is because the reality contains $g_1 = \sum_{j=1}^{\infty} j^{-1} \cos(5\pi(j-0.5)x) \sin(5j)$, which is hard to be predicted by a nonparametric regression approach alone. The mathematical model specified herein, however, can explain this term with calibration parameter close to 1. The estimated calibrated parameter in both GaSP and S-GaSP is indeed close 1 in all experiments, which leads to better predictive performance.

The L_2 minimizer of the calibration parameters in this example is around 1.775. Note that the estimated calibration parameter S-GaSP calibration may not converge to the L_2 minimizer when sample size increases. This is because $||y^R(\cdot) - f^M(\cdot, \theta)||_{\mathcal{H}}$ is unbounded when $\theta \neq 1$, which violates the assumption A5. The calibrated computer computer with calibration parameter being around 1 improves the predictive accuracy and interpretation, as the residual discrepancy term is a smooth term that is easy to be predicted.

The GaSP and S-GaSP calibration approaches are always more accurate in predicting the reality than the two-step approaches. Indeed, for Example 5 discussed in the numerical comparison, the predictions in the L_2 calibration approach is the best among all methods, as the observations can be easily predicted through a nonparametric regression without the mathematical model. When the reality is complicated, the joint estimation by GaSP and S-GaSP calibration approaches seems to have a smaller predictive error, which will be illustrated by a few more numerical examples.

As the sampling model of the observations is well-defined in both GaSP calibration and S-GaSP calibration, a Bayesian approach can be implemented and the uncertainty of the parameters can be assessed naturally through their posterior distributions. For the L_2 calibration, the asymptotic distribution of the estimator of the calibration parameter may be used to approximately quantify the uncertainty in parameter estimation (Tuo and Wu, 2015). A bootstrap approach is developed to assess the uncertainty of parameters for the LS calibration approach (Wong et al., 2017).

For a fast mathematical model, The computational complexity of all methods are typically dominated by nonparametric estimation of the discrepancy function, which is $O(n^3)$ in general with n being the number of observations. Though the computation complexity is similar, the two-step approaches such as L_2 calibration and LS calibration are typically faster than the joint estimation in GaSP and S-GaSP calibration approach, as the kernel parameters and calibration parameters are estimated in two different steps.

6 Numerical study

In this section, we numerically compare the performance of several methods for calibration and prediction. We consider two loss functions as follows.

i. The L_2 loss between the reality and its estimator: $L_2(\hat{y}^R(\cdot)) = \|y^R(\cdot) - \hat{y}^R(\cdot)\|_{L_2}^2$.

ii. The L_2 loss between the reality and calibrated mathematical model: $L_2(\hat{\theta}) = \|y^R(\cdot) - \hat{f}^M(\cdot, \hat{\theta})\|_{L_2}^2$, where $\hat{\theta}$ is the estimator of the calibration parameter.

The first criterion is our primary consideration, because the out of sample prediction for the reality examines how well we can reproduce the reality. The second criterion describes how well the calibrated mathematical model fits reality in terms of the L_2 loss (Tuo and Wu, 2015). The parameters in a mathematical model often have scientific interpretation, whereas a non-linear discrepancy function might not be interpretable. Thus the discrepancy function is not used for prediction in the second criterion.

To numerically evaluate the performance under two prediction criteria, we first compute $\operatorname{AvgRMSE}_{f^M+\delta}$ in Equation (15) of different methods. We compute $\operatorname{AvgRMSE}_{f^M}$ through replacing $\hat{y}_i^R(\mathbf{x}_j^*)$ in Equation (15) by the calibrated computer model output. For assessing the uncertainty in predictions, we also compute the average length of predictive interval and proportion of the observations covered by the 95% predictive interval defined below:

$$L_{CI}(95\%) = \frac{1}{Nn^*} \sum_{j=1}^{N} \sum_{i=1}^{n^*} \operatorname{length}\{CI_{ij}(95\%)\},\$$
$$P_{CI}(95\%) = \frac{1}{Nn^*} \sum_{j=1}^{N} \sum_{i=1}^{n^*} 1\{y_j^R(\mathbf{x}_i^*) \in CI_{ij}(95\%)\},\$$

where $CI_{ij}(95\%)$ is the 95% predictive interval; N and n^* are the total number of experiments and the number of held-out test data in each experiment, respectively. An efficient method should have small AvgRMSE_{$f^M+\delta$} and AvgRMSE_{f^M}, short predictive interval and $P_{CI}(95\%)$ close to the nominal 95%. For the real examples, we replace the test reality output by the held-out observations for out-of-sample predictions.

We numerically compare 5 different methods. The first and second methods are the GaSP calibration and S-GaSP calibration, respectively, both implemented in the full Bayesian framework. The scaling parameter of the S-GaSP is fixed to be $\lambda_z = (\lambda || \tilde{\gamma} ||)^{-1/2}$, where $\lambda = \sigma_0^2/(\sigma^2 n)$ and $\tilde{\gamma} = (\tilde{\gamma}_1, ..., \tilde{\gamma}_p)^T$, with $\tilde{\gamma}_i$ being the normalized range parameter (normalized

by the length of each coordinate of the input variable). The rest of model parameters and calibration parameters are sampled from posterior distributions. The third and fourth approaches are the L_2 calibration and LS calibration, respectively. The GaSP regression using the **RobustGaSP** R package is used to estimate the reality in the first step of the L_2 calibration, and the residual (between the calibrated mathematical model and reality) in the second step of the LS calibration. The kernel function $K(\cdot, \cdot)$ is assumed to follow (19) in all methods for demonstration purposes. We also include the method with no discrepancy function computed under the Bayesian framework. The GaSP, S-GaSP and no-discrepancy calibration approaches are implemented in the **RobustCalibration** R package.

6.1 Simulated example

Example 3. Let $y^F(\mathbf{x}) = y^R(\mathbf{x}) + \epsilon$, where $\mathbf{x} = (x_1, x_2) \in [0, 1]^2$, $y^R(\mathbf{x}) = \sin(0.2\pi x_1)x_2 + \sin(2\pi x_1)x_2 + 1$ and $\epsilon \stackrel{i.i.d.}{\sim} N(0, 0.1^2)$. The mathematical model is $f^M(\mathbf{x}, \boldsymbol{\theta}) = \sin(\theta_1 x_1)x_2 + \theta_2$ with $\theta_1 \in [0, 10]$ and $\theta_2 \in \mathbb{R}$.

We first consider a simulated study in Example 3 and test two configurations, where with sample sizes of the observations are taken to be n = 25 and n = 50, respectively. For each configuration, we test N = 100 experiments with $n^* = 2500$ reality output equally spaced in each interval held-out for testing. The input variable in each experiment is generated from the maximin Latin hypercube design (Santner et al. (2003)).

The predictive errors by different approaches are given in Table 1. First the AvgRMSE_{$f^{M}+\delta$} of all methods are better than AvgRMSE_{f^{M}} for almost all methods, indicating that a nonparametric model can improve the predictive performance. Second, the GaSP and S-GaSP calibration approaches performs better than the two-step LS and L_2 calibration methods in terms of AvgRMSE_{$f^{M}+\delta$}. Though around 95% of the held out test data are covered by the 95% predictive interval in all methods, the average lengths of the predictive interval by the GaSP and S-GaSP calibration are shorter than the two-step approaches.

For the L_2 calibration approach, the mathematical model is not used in the first step, as

n=25	GaSP	S-GaSP	L_2	LS	No-discrepancy
AvgRMSE $_{f^M+\delta}$	0.0556	0.0558	0.103	0.0702	/
$AvgRMSE_{f^M}$	0.143	0.131	0.131	0.131	0.130
$L_{CI}(95\%)$	0.206	0.209	0.434	0.389	0.654
$P_{CI}(95\%)$	0.931	0.932	0.921	0.958	0.987
n=50	GaSP	S-GaSP	L_2	LS	No-discrepancy
AvgRMSE $_{f^M+\delta}$	0.0405	0.0403	0.0655	0.0437	/
$AvgRMSE_{f^M}$	0.144	0.130	0.128	0.128	0.128
$L_{CI}(95\%)$	0.153	0.152	0.439	0.402	0.632
$P_{CI}(95\%)$	0.938	0.935	0.992	0.997	0.991

Table 1: Predictive performance by different methods for Example 3.



Figure 3: The reality in an experiment of Example 3 is graphed in the left panel. The residuals between the reality and prediction by a GaSP regression (without the mathematical model) are graphed in the middle panel, whereas those by the S-GaSP calibration are graphed in right panel. The middle and right panels have the same scale of color.

the parameters in the mathematical model are estimated in the second step to minimize the L_2 loss. The high frequency term $(\sin(2\pi x_1)x_2)$ makes the predictions by the nonparametric regression less accurate than jointly estimate of the calibration parameters and discrepancy function, as the high frequency term can be explained by the mathematical model. The comparison between predictions by using the nonparameteric regression alone and S-GaSP calibration in the first experiment with n = 60, for example, is shown in Figure 3. Indeed the S-GaSP calibration seems to have a smaller predictive error.

The estimated calibration parameters of Example 3 are graphed in Figure 4. In the left panel, the estimation of θ_1 using the LS, L_2 and no-discrepancy methods is close to the L_2



Figure 4: Boxplots of the estimated calibration parameters from the GaSP, the S-GaSP, L_2 and LS calibration approaches for each experiment in Example 3. The solid lines are the L_2 minimizer, which is around 6.48 and 1.15 for θ_1 and θ_2 , respectively.

minimizer (graphed as the solid line), whereas the estimation of θ_1 using the GaSP and the S-GaSP is, in fact, closer to 2π . This is because the model complexity is naturally built into the calibration: an estimated θ_1 that is close to 2π makes the prediction better, since the high frequency term can be approximately explained by the mathematical model.

The GaSP calibration approach produces the largest $\operatorname{AvgRMSE}_{f^M}$ compared to other calibration methods $\operatorname{AvgRMSE}_{f^M}$, as illustrated in Table 1. This is probably caused by the estimation of θ_2 (a mean parameter) shown in the right panel in Figure 4. The estimates of θ_2 by the S-GaSP calibration, LS, L_2 and no-discrepancy methods are all close to the L_2 minimizer, whereas the one in the GaSP calibration seems to be smaller the other ones.

Example 3 indicates that the calibrated mathematical model that minimizes the L_2 loss between the reality and mathematical model might not always be the optimal choice for predictions. In this example, when θ_1 is estimated to be close to 2π , the predictive accuracy can be improved significantly, and the calibrated computer model only produces slightly larger error than the one by the L_2 minimizer. The other parameter, θ_2 , is a mean parameter. The two types of errors are both small when θ_2 is estimated to be close to the L_2 minimizer. The S-GaSP calibration model seems to do well in both sides. It predicts the reality as accurately as the GaSP calibration model with the assistance of the calibrated mathematical model. The calibrated mathematical model using the S-GaSP is also closer to the reality than the one using the GaSP calibration.

6.2 Chemical system interaction

Example 4. Consider the system interaction between two chemical substances y_1 and y_2 :

$$\dot{y}_1(t) = 10^{\theta_1 - 3} y_1(t)$$
$$\dot{y}_2(t) = 10^{\theta_1 - 3} y_1(t) - 10^{\theta_2 - 3} y_2(t)$$

where 2 repeated observations of the second chemical substance are measured at each of the 6 time points t = 10, 20, 40, 80, 160, 320 in Box and Coutie (1956). The goal is to estimate θ_1 and θ_2 , and to predict the values of the chemical substance across time.

We consider $(\theta_1, \theta_2) \in [0.5, 1.5]^2$. As the number of observations are limited, we first perform a leave-one-out comparison by holding out two repeated observations for prediction at each time point. We replace the reality in each criterion by the held-out observations to test each approach. The predictive performance of the leave-one-out comparison for Example 4 is given in Table 2.

	GaSP	S-GaSP	L_2	LS	No-discrepancy
$AvgRMSE_{f^M+\delta}$	4.98	5.05	10.2	7.80	/
$AvgRMSE_{f^M}$	11.8	10.4	9.34	10.9	9.63
$L_{CI}(95\%)$	32.8	44.6	52.5	42.6	47.0
$P_{CI}(95\%)$	100%	100%	91.7%	100%	100%

Table 2: Predictive performance by different methods for Example 4.

First the predictive error $\operatorname{AvgRMSE}_{f^M+\delta}$ is smaller than $\operatorname{AvgRMSE}_{f^M}$ in both GaSP and S-GaSP calibration, indicating that jointly estimating the mathematical model and discrepancy function improves the predictive accuracy than the two-step approaches. On



Figure 5: Posteriors samples of calibration parameters and predictions from the GaSP, S-GaSP and no-discrepancy calibration approaches. The shaded area in the right panel is the predictive interval of the reality from the S-GaSP calibration approach.

the other hand, the GaSP calibration has the largest predictive error using the calibrated mathematical model alone. The L_2 calibration and no-discrepancy calibration have a smaller error AvgRMSE_f compared to other methods. The S-GaSP calibration has relatively small predictive error under both predictive criteria.

The left panel in Figure 5 gives posterior samples of calibration parameters by different methods in Example 4. When there is no discrepancy function, the posterior samples of the calibration parameters are close to (1.1, 0.8), which is inline with the result in Box and Coutie (1956). The estimated calibration parameters of the no-discrepancy method is close to the L_2 and LS methods, whereas the posterior samples of the calibration parameters from the GaSP calibration and S-GaSP calibration reflects more uncertainty.

The predictions in the GaSP, S-GaSP and no-discrepancy calibration are graphed in right panel of Figure 5. The calibrated mathematical model in the S-GaSP calibration seems to fit the reality better than the one by the GaSP calibration. The calibrated mathematical model in the no-discrepancy calibration fits the observations the best, as the residuals between the mathematical model and observations are assumed to be random noises. Although we do not know the reality of this example, the predictive performance of GaSP and S-GaSP calibration is better than the other approaches as shown in the first row in Table 2, suggesting that there might be a systematic discrepancy between the observations and the mathematical model.

6.3 Ion Channel experiments

In the last example, we consider calibrating the mathematical model for sodium ion channels using real observations from whole cell voltage clamp experiments (Plumlee et al., 2016).

Example 5. The data sets consists of 19 observations of normalized current needed to maintain the membrane potential at -35mV over time (Plumlee, 2017). Denote the input variable x the natural logarithm of time. The mathematical model has the following expression:

$$f^M(x, \theta) = \mathbf{e}_1^T \exp[\exp(x)\mathbf{A}(\theta)]\mathbf{e}_4,$$

where \mathbf{e}_i is a column vector with 1 at the *i*th element and 0 for the rest of components, the first exp is the matrix exponential, $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)^T$, and

$$\mathbf{A}(\boldsymbol{\theta}) = \begin{pmatrix} -\theta_2 - \theta_3 & \theta_1 & 0 & 0 \\ \theta_2 & -\theta_1 - \theta_2 & \theta_1 & 0 \\ 0 & \theta_2 & -\theta_1 - \theta_2 & \theta_1 \\ 0 & 0 & \theta_2 & -\theta_1 \end{pmatrix}$$

The range of calibration parameter considered herein is $\theta_i \in [0, 10]$ for i = 1, 2, 3.

We first consider a leave-one-out comparison for the ion channel experiment. The predictive errors by different approaches are given in Table 3. First the average root of mean squared errors under two criteria are all very small. Here since the input variable is only one dimensional and the real observations (graphed in the bottom left panel in Figure 6) seems to be very smooth, using the GaSP regression by the RobustGaSP R package gives very high predictive accuracy already. Including the mathematical model does not improves the prediction accuracy in this example. Therefore, the predictive error by the L_2 calibration is the smallest, as the GaSP regression without the mathematical model is used for predicting the reality in the first step.

	GaSP	S-GaSP	L_2	LS	No-discrepancy
$AvgRMSE_{f^M+\delta}$	1.77×10^{-3}	1.52×10^{-3}	8.26×10^{-4}	2.17×10^{-3}	/
$AvgRMSE_{f^M}$	1.13×10^{-2}	5.62×10^{-3}	4.83×10^{-3}	6.49×10^{-3}	6.45×10^{-3}
$L_{CI}(95\%)$	5.64×10^{-3}	9.58×10^{-3}	9.45×10^{-3}	5.80×10^{-3}	2.58×10^{-2}
$P_{CI}(95\%)$	94.7%	100%	84.2%	78.9%	84.2%

Table 3: Predictive performance by different methods for Example 5.

In the second place, the GaSP calibration approach has a substantial larger predictive error by using the calibrated computer model alone, shown in the second row in Table 3. Besides, the lengths of predictive intervals are all very small in all approaches. The predictive interval of the GaSP calibration seems to be most faithful for uncertainty assessment, as the proportion of the observations covered in the interval are closest to the nominal level.

The posterior samples and estimation of the reality by different calibration approaches are graphed in Figure 6. For better visualization, we thin the posterior samples by 10 and only graph one tenth of the total posterior samples. In this example, the posterior samples of the S-GaSP are closer to the ones with the no-discrepancy calibration. Due to this reason, the calibrated computer model by the S-GaSP calibration (graphed as the dashed blue curve in the lower right panel) fits the observations well, whereas the calibrated computer model by the GaSP calibration (graphed as the dashed red curve in the lower right panel) underestimates the values of the observations. This identifiability issue of the GaSP calibration in this example was also reported in (Plumlee, 2017), whereas the calibrated mathematical model fits the observations reasonably well in the S-GaSP calibration.

7 Concluding remarks

We have introduced the scaled Gaussian stochastic process (S-GaSP) for the calibration and prediction. We showed that under certain some routinely used assumptions, the predictive mean of the S-GaSP calibration model converges to the reality as fast as the GaSP calibration with some suitable choice of the regularization parameter and scaling parameter. The



Figure 6: The posterior samples and predictions from no-discrepancy calibration, S-GaSP and GaSP calibration. The shaded area in the bottom right panel is the 95% predictive credible interval of the mean of the observations.

estimated calibration parameters in the S-GaSP calibration converges to the L_2 minimizer with the same choice of the regularization parameter and scaling parameter, whereas those in the GaSP calibration typically do not converge to the L_2 minimizer. The results rely on the the orthogonal series representation of the processes studied in this work. The computational complexity of the discretized S-GaSP calibration is the same as the GaSP calibration. Both GaSP, S-GaSP calibration and calibration without a discrepancy function were implemented in the RobustCalibration R package under the Bayesian framework.

The numerical studies indicate that jointly estimating the calibration parameters and discrepancy function in GaSP and S-GaSP calibration can improve the predictive accuracy, if the reality are too complicated to be predicted precisely by nonparametric regression alone. The mathematical model, which typically contains some information of the trend and shape of the reality function, can be helpful in predictions. We also empirically found that the calibrated mathematical model that minimizes the L_2 distance between the reality and mathematical model may not always be the best for reducing the predictive error, as the residuals may contain information that could be hard to be modeled by the nonparametric regression. The S-GaSP calibration give as least as accurate predictions as the GaSP calibration, and the calibrated mathematical model by the S-GaSP calibration fits the real observations much closer than the one by the S-GaSP calibration in almost all examples.

We outline a few extensions of the S-GaSP model below. From the theoretical perspective, we did not study the convergence of the discretized S-GaSP, whereas the numerical studies indicate the convergence rate from the discretized S-GaSP is the same as the S-GaSP. Secondly we illustrated that the S-GaSP calibration can be implemented in both Frequentist and Bayesian ways. The contraction rate of the S-GaSP under the Bayesian framework, however, was not studied. The studies of the contraction rate of the GaSP regression may be extended to achieve this goal (Van der Vaart and Van Zanten, 2009; Bhattacharya et al., 2014). Thirdly, we fix the scaling parameter in the S-GaSP calibration as a function of the sample size, whereas historical information may be used to develop a reasonable prior for this parameter (Salter et al., 2019). Furthermore, the S-GaSP calibration framework may be extended to include a model of correlated noises, as the field observations may contain time series and images (Anderson and Poland, 2017).

References

- Anderson, K. R., Johanson, I. A., Patrick, M. R., Gu, M., Segall, P., Poland, M. P., Montgomery-Brown, E. K., and Miklius, A. (2019). Magma reservoir failure and the onset of caldera collapse at kīlauea volcano in 2018. *Science*, 366(6470).
- Anderson, K. R. and Poland, M. P. (2017). Abundant carbon in the mantle beneath Hawai'i. Nature Geoscience, 10(9):704–708.

Bayarri, M. J., Berger, J. O., Paulo, R., Sacks, J., Cafeo, J. A., Cavendish, J., Lin, C.-

H., and Tu, J. (2007). A framework for validation of computer models. *Technometrics*, 49(2):138–154.

- Bhattacharya, A., Pati, D., and Dunson, D. (2014). Anisotropic function estimation using multi-bandwidth gaussian processes. Annals of statistics, 42(1):352.
- Box, G. and Coutie, G. (1956). Application of digital computers in the exploration of functional relationships. *Proceedings of the IEE-Part B: Radio and Electronic Engineering*, 103(1S):100–107.
- Edmunds, D. E. and Triebel, H. (2008). Function spaces, entropy numbers, differential operators, volume 120. Cambridge University Press.
- Ghosal, S. and Van der Vaart, A. (2017). Fundamentals of nonparametric Bayesian inference, volume 44. Cambridge University Press.
- Goldstein, M. and Rougier, J. (2004). Probabilistic formulations for transferring inferences from mathematical models to physical systems. *SIAM journal on scientific computing*, 26(2):467–487.
- Gu, M. (2018a). Jointly robust prior for Gaussian stochastic process in emulation, calibration and variable selection. *Bayesian Analysis*, 14(1).
- Gu, M. (2018b). RobustCalibration: Robust Calibration of Imperfect Mathematical Models.R package version 0.5.2.
- Gu, M. and Wang, L. (2018). Scaled Gaussian stochastic process for computer model calibration and prediction. SIAM/ASA Journal on Uncertainty Quantification, 6(4):1555–1583.
- Gu, M., Wang, X., and Berger, J. O. (2018). Robust Gaussian stochastic process emulation. Annals of Statistics, 46(6A):3038–3066.
- Handcock, M. S. and Stein, M. L. (1993). A Bayesian analysis of Kriging. *Technometrics*, 35(4):403–410.

- Higdon, D., Gattiker, J., Williams, B., and Rightley, M. (2008). Computer model calibration using high-dimensional output. Journal of the American Statistical Association, 103(482):570–583.
- Kennedy, M. C. and O'Hagan, A. (2001). Bayesian calibration of computer models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63(3):425–464.
- Kosorok, M. R. (2008). Introduction to empirical processes and semiparametric inference. Springer.
- Mammen, E. and Van de Geer, S. (1997). Penalized quasi-likelihood estimation in partial linear models. The Annals of Statistics, pages 1014–1035.
- Pinelis, I. (1994). Optimum bounds for the distributions of martingales in banach spaces. The Annals of Probability, pages 1679–1706.
- Plumlee, M. (2017). Bayesian calibration of inexact computer models. Journal of the American Statistical Association, 112(519):1274–1285.
- Plumlee, M., Joseph, V. R., and Yang, H. (2016). Calibrating functional parameters in the ion channel models of cardiac cells. *Journal of the American Statistical Association*, 111(514):500–509.
- Rasmussen, C. E. (2006). Gaussian processes for machine learning. MIT Press.
- Rudelson, M., Vershynin, R., et al. (2013). Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18.
- Sacks, J., Welch, W. J., Mitchell, T. J., Wynn, H. P., et al. (1989). Design and analysis of computer experiments. *Statistical science*, 4(4):409–423.
- Salter, J. M., Williamson, D. B., Scinocca, J., and Kharin, V. (2019). Uncertainty quantification for computer models with spatial output using calibration-optimal bases. *Journal* of the American Statistical Association, pages 1–24.

- Santner, T. J., Williams, B. J., and Notz, W. I. (2003). The design and analysis of computer experiments. Springer Science & Business Media.
- Tuo, R. and Wu, C. J. (2015). Efficient calibration for imperfect computer models. The Annals of Statistics, 43(6):2331–2352.
- Tuo, R. and Wu, C. J. (2016). A theoretical framework for calibration in computer models: parametrization, estimation and convergence properties. SIAM/ASA Journal on Uncertainty Quantification, 4(1):767–795.
- Van de Geer, S. A. (2000). Empirical Processes in M-estimation, volume 6. Cambridge university press.
- Van der Vaart, A. W. (2000). Asymptotic statistics, volume 3. Cambridge university press.
- Van der Vaart, A. W. and Van Zanten, J. H. (2009). Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *The Annals of Statistics*, pages 2655–2675.
- Wahba, G. (1990). Spline models for observational data, volume 59. SIAM.
- Wong, R. K., Storlie, C. B., and Lee, T. (2017). A frequentist approach to computer model calibration. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 79:635–648.
- Yang, Y., Bhattacharya, A., and Pati, D. (2017a). Frequentist coverage and sup-norm convergence rate in Gaussian process regression. arXiv preprint arXiv:1708.04753.
- Yang, Y., Shang, Z., and Cheng, G. (2017b). Non-asymptotic theory for nonparametric testing. arXiv preprint arXiv:1702.01330.
- Zhang, H. (2004). Inconsistent estimation and asymptotically equal interpolations in modelbased geostatistics. Journal of the American Statistical Association, 99(465):250–261.

Supplementary materials for a theoretical framework of the scaled Gaussian stochastic process in prediction and calibration

All the formulas in this supplementary materials are cross-referenced in the main body of the article. We first give a brief introduction of the Gaussian stochastic process model and reproducing kernel Hilbert space in Section S1. The proof for Section 2 is given in Section S2. The proof for Theorem 1 and two auxiliary lemmas are provided in Section S3. Section S4 encloses the proof for Theorem 2 and provides additional results regarding the convergence of S-GaSP calibration when kernel parameters are estimated.

S1 Background: Gaussian stochastic process

Assume the mean and trend of the reality are properly modeled in the mathematical model. Consider to model the unknown discrepancy function in the calibration model (1) via a real-valued zero-mean Gaussian stochastic process $\delta(\cdot)$ on a *p*-dimensional input domain \mathcal{X} ,

$$\delta(\cdot) \sim \text{GaSP}(0, \sigma^2 K(\cdot, \cdot)), \tag{S1}$$

where σ^2 is a variance parameter and $K(\mathbf{x}_a, \mathbf{x}_b)$ is the correlation for any $\mathbf{x}_a, \mathbf{x}_b \in \mathcal{X}$, parameterized by a kernel function. For simplicity, we assume $\mathcal{X} = [0, 1]^p$ in this work.

For any $\{\mathbf{x}_1, ..., \mathbf{x}_n\}$, the outputs $(\delta(\mathbf{x}_1), ..., \delta(\mathbf{x}_n))^T$ follow a multivariate normal distribution

$$[\delta(\mathbf{x}_1), ..., \delta(\mathbf{x}_n) \mid \sigma^2, \mathbf{R}] \sim \mathrm{MN}(\mathbf{0}, \sigma^2 \mathbf{R}),$$
(S2)

where the (i, j) entry of **R** is $K(\mathbf{x}_i, \mathbf{x}_j)$. Some frequently used kernel functions include the power exponential kernel and the Matérn kernel. We defer the issue of estimating the parameters in the kernel function in Section 4 and assume $K(\cdot, \cdot)$ is known for now.

The reproducing kernel Hilbert space (RKHS), denoted as \mathcal{H} , attached to the Gaussian

stochastic process $GaSP(0, \sigma^2 K(\cdot, \cdot))$, is the completion of the space of all functions

$$\mathbf{x} \to \sum_{i=1}^{k} w_i K(\mathbf{x}_i, \mathbf{x}), \quad w_1, ..., w_k \in \mathbb{R}, \, \mathbf{x}_1, ..., \mathbf{x}_k, \mathbf{x} \in \mathcal{X}, \, k \in \mathbb{N},$$

with the inner product

$$\left\langle \sum_{i=1}^{k} w_i K(\mathbf{x}_i, \cdot), \sum_{j=1}^{m} w_j K(\mathbf{x}_j, \cdot) \right\rangle_{\mathcal{H}} = \sum_{i=1}^{k} \sum_{j=1}^{m} w_i w_j K(\mathbf{x}_i, \mathbf{x}_j).$$

For any function $f(\cdot) \in \mathcal{H}$, denote $||f||_{\mathcal{H}} = \sqrt{\langle f, f \rangle_{\mathcal{H}}}$ the RKHS norm or the native norm. Because the evaluation maps in RKHS are bounded linear, it follows from the Riesz representation theorem that for each $\mathbf{x} \in \mathcal{X}$ and $f(\cdot) \in \mathcal{H}$, one has $f(\mathbf{x}) = \langle f(\cdot), K(\cdot, \mathbf{x}) \rangle_{\mathcal{H}}$.

Denote $L_2(\mathcal{X})$ the space of square-integrable functions $f : \mathcal{X} \to \mathbb{R}$ with $\int_{\mathbf{x} \in \mathcal{X}} f^2(\mathbf{x}) d\mathbf{x} < \infty$. We denote $\langle f, g \rangle_{L_2(\mathcal{X})} := \int_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}) g(\mathbf{x}) d\mathbf{x}$ the usual inner product in $L_2(\mathcal{X})$. By Mercer's theorem, there exists an orthonormal sequence of continuous eigenfunctions $\{\phi_k\}_{k=1}^{\infty}$ with a sequence of non-increasing and non-negative eigenvalues $\{\rho_k\}_{k=1}^{\infty}$ such that

$$K(\mathbf{x}_a, \mathbf{x}_b) = \sum_{k=1}^{\infty} \rho_k \phi_k(\mathbf{x}_a) \phi_k(\mathbf{x}_b),$$
(S3)

for any $\mathbf{x}_a, \mathbf{x}_b \in \mathcal{X}$.

The RKHS \mathcal{H} contains all functions $f(\cdot) = \sum_{k=1}^{\infty} f_k \phi_k(\cdot) \in L_2(\mathcal{X})$ with $f_k = \langle f, \phi_k \rangle_{L_2(\mathcal{X})}$ and $\sum_{k=1}^{\infty} f_k^2 / \rho_k < \infty$. For any $g(\cdot) = \sum_{k=1}^{\infty} g_k \phi_k(\cdot) \in \mathcal{H}$ and $f(\cdot)$, the inner product can be represented as $\langle f, g \rangle_{\mathcal{H}} = \sum_{k=1}^{\infty} f_k g_k / \rho_k$. For more properties of the RKHS, we refer to Chapter 1 of Wahba (1990) and Chapter 11 of Ghosal and Van der Vaart (2017).

S1.1 The equivalence between the maximum likelihood estimator and the kernel ridge regression estimator in calibration

Assume one has a set of observations $\mathbf{y}^F := (y^F(\mathbf{x}_1), ..., y^F(\mathbf{x}_n))^T$ and mathematical model outputs $\mathbf{f}^M_{\boldsymbol{\theta}} := (f^M(\mathbf{x}_1, \boldsymbol{\theta}), ..., f^M(\mathbf{x}_n, \boldsymbol{\theta}))^T$, where $\boldsymbol{\theta} = (\theta_1, ..., \theta_q)^T \in \boldsymbol{\Theta} \subset \mathbb{R}^q$ is a *q*dimensional vector of the calibration parameters.

Denote the regularization parameter $\lambda := \sigma_0^2/(n\sigma^2)$. For the calibration model (1) with δ modeled as a GaSP in (S1), the marginal distribution of \mathbf{y}^F follows a multivariate normal after marginalizing out δ ,

$$[\mathbf{y}^F \mid \boldsymbol{\theta}, \, \sigma_0^2, \, \lambda] \sim \mathrm{MN}(\mathbf{f}_{\boldsymbol{\theta}}^M, \, \sigma_0^2((n\lambda)^{-1}\mathbf{R} + \mathbf{I}_n)).$$
(S4)

Let $\mathcal{L}(\boldsymbol{\theta})$ be the likelihood for $\boldsymbol{\theta}$ in (S4) given the other parameters in the model. For any given λ , the maximum likelihood estimator (MLE) of $\boldsymbol{\theta}$ is denoted as

$$\hat{\boldsymbol{\theta}}_{\lambda,n} := \operatorname*{argmax}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathcal{L}(\boldsymbol{\theta}).$$
(S5)

Conditioning on the observations, $\hat{\theta}_{\lambda,n}$ and λ , the predictive mean of the discrepancy function at any $\mathbf{x} \in \mathcal{X}$ has the following expression

$$\hat{\delta}_{\lambda,n}(\mathbf{x}) := \mathbb{E}[\delta(\mathbf{x}) \mid \mathbf{y}^F, \hat{\boldsymbol{\theta}}_{\lambda,n}, \lambda] = \mathbf{r}^T(\mathbf{x})(\mathbf{R} + n\lambda \mathbf{I}_n)^{-1} \left(\mathbf{y}^F - \mathbf{f}_{\hat{\boldsymbol{\theta}}_{\lambda,n}}^M\right)$$
(S6)

with $\mathbf{r}(\mathbf{x}) = (K(\mathbf{x}_1, \mathbf{x}), ..., K(\mathbf{x}_n, \mathbf{x}))^T$ and \mathbf{I}_n being the *n*-dimensional identity matrix.

It is well-known that the predictive mean in (S6) can be written as the estimator for the kernel ridge regression (KRR). In the following lemma, we show that $(\hat{\theta}_{\lambda,n}, \hat{\delta}_{\lambda,n}(\cdot))$ is equivalent to the KRR estimator.

Lemma S1. The maximum likelihood estimator $\hat{\theta}_{\lambda,n}$ defined in (S5) and predictive mean estimator $\hat{\delta}_{\lambda,n}(\cdot)$ defined in (S6) can be expressed as the estimator of the kernel ridge regression

as follows

$$(\hat{\boldsymbol{\theta}}_{\lambda,n},\hat{\delta}_{\lambda,n}(\cdot)) = \operatorname*{argmin}_{\boldsymbol{\theta}\in\boldsymbol{\Theta},\delta(\cdot)\in\mathcal{H}} \ell_{\lambda,n}(\boldsymbol{\theta},\delta),$$

where

$$\ell_{\lambda,n}(\boldsymbol{\theta},\delta) = \frac{1}{n} \sum_{i=1}^{n} (y^F(\mathbf{x}_i) - f^M(\mathbf{x}_i,\boldsymbol{\theta}) - \delta(\mathbf{x}_i))^2 + \lambda \|\delta\|_{\mathcal{H}}^2.$$
(S7)

Proof of Lemma S1. By the representer lemma (Rasmussen, 2006; Wahba, 1990), for any $\theta \in \Theta$ and $\mathbf{x} \in \mathcal{X}$, one has

$$\hat{\delta}_{\lambda,n,\boldsymbol{\theta}}(\mathbf{x}) = \sum_{i=1}^{n} w_i(\boldsymbol{\theta}) K(\mathbf{x}_i, \mathbf{x}).$$
(S8)

Denote $\mathbf{w}_{\boldsymbol{\theta}} = (w_1(\boldsymbol{\theta}), ..., w_n(\boldsymbol{\theta}))^T$. Since $\langle K(\mathbf{x}_i, \cdot), K(\mathbf{x}_j, \cdot) \rangle_{\mathcal{H}} = K(\mathbf{x}_i, \mathbf{x}_j)$, (S7) becomes to find $\boldsymbol{\theta}$ and $\mathbf{w}_{\boldsymbol{\theta}}$ that minimize

$$\frac{1}{n}(\mathbf{y}^F - \mathbf{f}_{\theta}^M - \mathbf{R}\mathbf{w}_{\theta})^T(\mathbf{y}^F - \mathbf{f}_{\theta}^M - \mathbf{R}\mathbf{w}_{\theta}) + \lambda \mathbf{w}_{\theta}^T \mathbf{R}\mathbf{w}_{\theta}.$$
 (S9)

For any $\boldsymbol{\theta}$, solving the minimization for (S9) with regard to $\mathbf{w}_{\boldsymbol{\theta}}$ gives

$$\hat{\mathbf{w}}_{\boldsymbol{\theta}} = (\mathbf{R} + n\lambda \mathbf{I}_n)^{-1} (\mathbf{y}^F - \mathbf{f}_{\boldsymbol{\theta}}^M).$$
(S10)

Then plugging $\hat{\mathbf{w}}_{\theta}$ into (S9), based on the Woodbury matrix identity, one has

$$\frac{1}{n} (\mathbf{y}^{F} - \mathbf{f}_{\boldsymbol{\theta}}^{M} - \mathbf{R}\hat{\mathbf{w}}_{\boldsymbol{\theta}})^{T} (\mathbf{y}^{F} - \mathbf{f}_{\boldsymbol{\theta}}^{M} - \mathbf{R}\hat{\mathbf{w}}_{\boldsymbol{\theta}}) + \lambda \hat{\mathbf{w}}_{\boldsymbol{\theta}}^{T} \mathbf{R}\hat{\mathbf{w}}_{\boldsymbol{\theta}}$$

$$= \frac{1}{n} (\mathbf{y}^{F} - \mathbf{f}_{\boldsymbol{\theta}}^{M})^{T} [(\mathbf{I}_{n} - \mathbf{R}(\mathbf{R} + n\lambda \mathbf{I}_{n})^{-1})^{T} (\mathbf{I}_{n} - \mathbf{R}(\mathbf{R} + n\lambda \mathbf{I}_{n})^{-1})]$$

$$+ \lambda (\mathbf{y}^{F} - \mathbf{f}_{\boldsymbol{\theta}}^{M})^{T} (\mathbf{R} + n\lambda \mathbf{I}_{n})^{-1} \mathbf{R} (\mathbf{R} + n\lambda \mathbf{I}_{n})^{-1} (\mathbf{y}^{F} - \mathbf{f}_{\boldsymbol{\theta}}^{M})$$

$$= \lambda (\mathbf{y}^{F} - \mathbf{f}_{\boldsymbol{\theta}}^{M})^{T} (\mathbf{R} + n\lambda \mathbf{I}_{n})^{-1} (\mathbf{y}^{F} - \mathbf{f}_{\boldsymbol{\theta}}^{M}), \qquad (S11)$$

which shows that the minimizer of $\boldsymbol{\theta}$ on right-hand side of (S11) is the same as the MLE of $\boldsymbol{\theta}$ in (S5). Finally, plugging the estimator $\hat{\boldsymbol{\theta}}_{\lambda,n}$ into (S9), the result follows from the KRR estimator of $\delta(\cdot)$ in (S8) with the weights in (S10).

Although modeling the discrepancy function by the GaSP typically improves the prediction accuracy of the reality, the penalty term of (S7) only contains $\|\delta\|_{\mathcal{H}}$ to control the complexity of the discrepancy. As the RKHS norm is not equivalent to the L_2 norm, the calibrated computer model could deviate a lot from the best performed mathematical model in terms of the L_2 loss (Tuo and Wu, 2015). In Section 2, we introduce the scaled Gaussian stochastic process that predicts the reality as accurately as the GaSP with the aid of the mathematical model, but has more prior mass on the small L_2 distance between the reality and mathematical model. As a consequence, the KRR estimator of the new model penalizes both $\|\delta\|_{\mathcal{H}}$ and $\|\delta\|_{L_2(\mathcal{X})}$ simultaneously.

S2 Proof for Section 2

Proof of Lemma 1. By Karhunen-Loève expansion, we have

$$\delta(\mathbf{x}) = \sigma \sum_{i=1}^{\infty} \sqrt{\rho_i} Z_i \phi_i(\mathbf{x})$$

with $Z_i \overset{i.i.d}{\sim} N(0,1)$. Denote $W_k = \sum_{i=k+1}^{\infty} \rho_i Z_i^2$ for any $k \in \mathbb{N}^+$. From the definition of $Z = \int_{\mathbf{x} \in \mathcal{X}} \delta^2(\mathbf{x}) d\mathbf{x}$ and $\int_{\mathbf{x} \in \mathcal{X}} \phi_i^2(\mathbf{x}) d\mathbf{x} = 1$ for any $i \in \mathbb{N}^+$, it is straightforward to see that

$$Z = \sigma^2 (\rho_1 Z_1^2 + \dots + \rho_k Z_k^2 + W_k).$$
 (S12)

In the following expressions, we are conditioning on all parameters and they are dropped

for simplicity. From the construction of

$$\delta_z(\mathbf{x}) = \left\{ \delta(\mathbf{x}) \mid \int_{\mathbf{x} \in \mathcal{X}} \delta^2(\mathbf{x}) d\mathbf{x} = Z \right\}$$

and $Z \sim p_Z(\cdot)$, the joint density of $(Z_1, ..., Z_k, W_k)$ in the S-GaSP can be expressed as

$$\begin{split} p_{\delta_z}(Z_1 &= z_1, ..., Z_k = z_k, W_k = w_k) \\ &= \int_0^\infty p_\delta \left(Z_1 = z_1, ..., Z_k = z_k, W_k = w_k \mid Z = z \right) p_Z(Z = z) dz \\ &\propto \int_0^\infty \frac{p_\delta (Z = z, Z_1 = z_1, ..., Z_k = z_k, W_k = w_k)}{p_\delta (Z = z)} g_Z(Z = z) p_\delta (Z = z) dz \\ &\propto p_\delta (Z_1 = z_1, ..., Z_k = z_k) p_\delta (W_k = w_k) \times \\ &\int_0^\infty p_\delta (Z = z \mid Z_1 = z_1, ..., Z_k = z_k, W_k = w_k) \exp\left(-\frac{\lambda_z z}{2\sigma^2}\right) dz \\ &\propto p_\delta (Z_1 = z_1, ..., Z_k = z_k) p_\delta (W_k = w_k) \times \\ &\int_0^\infty \mathbbm{1} \left\{ z = \sigma^2 (\rho_1 z_1^2 + \dots + \rho_k z_k^2 + w_k) \right\} \exp\left(-\frac{\lambda_z z}{2\sigma^2}\right) dz \\ &\propto \exp\left(-\frac{1}{2}\sum_{i=1}^k z_i^2\right) p_\delta (W_k = w_k) \exp\left[-\frac{\lambda_z}{2}\left(\sum_{i=1}^k \rho_i z_i^2 + w_k\right)\right] \\ &= \left\{\prod_{i=1}^k \exp\left[-\frac{1}{2}(1 + \lambda_z \rho_i) z_i^2\right]\right\} p_\delta (W_k = w_k) \exp(-\lambda_z w_k/2) \,, \end{split}$$

where 1 in the fourth step is a Dirac delta function.

After integrating out W_k , it is clear that Z_i 's are independently distributed as N(0, 1/(1 + $\lambda_z \rho_i)$) under the measure induced by the S-GaSP. Since k is arbitrary, we have

$$\delta_z(\mathbf{x}) = \sigma \sum_{i=1}^{\infty} \sqrt{\frac{\rho_i}{1 + \lambda_z \rho_i}} Z_i \phi_i(\mathbf{x})$$

with $Z_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0,1)$, from which the proof is complete.

Proof of Lemma 2. First note that for any $\mathbf{x}_a, \mathbf{x}_b \in \mathcal{X}$, we have $K(\mathbf{x}_a, \mathbf{x}_b) = \sum_{i=1}^{\infty} \rho_i \phi_i(\mathbf{x}_a) \phi_i(\mathbf{x}_b)$ and $K_z(\mathbf{x}_a, \mathbf{x}_b) = \sum_{i=1}^{\infty} \rho_{z,i} \phi_i(\mathbf{x}_a) \phi_i(\mathbf{x}_b)$ with $\rho_{z,i} = \rho_i / (1 + \lambda_z \rho_i)$. For $h(\cdot) = \sum_{i=1}^{\infty} h_i \phi_i(\cdot) \in \mathcal{H}$

and $g(\cdot) = \sum_{i=1}^{\infty} g_i \phi_i(\cdot) \in \mathcal{H}$, one has

$$\langle h,g \rangle_{\mathcal{H}_z} = \sum_{i=1}^{\infty} \frac{1}{\rho_{z,i}} h_i g_i = \sum_{i=1}^{\infty} \frac{1}{\rho_i} h_i g_i + \lambda_z \sum_{i=1}^{\infty} h_i g_i = \langle h,g \rangle_{\mathcal{H}} + \lambda_z \langle h,g \rangle_{L_2}.$$

Proof of Lemma 3. We show below that for any $\theta \in \Theta$ and any $\mathbf{x} \in \mathcal{X}$, one has

$$\hat{\delta}_{\lambda,\lambda_z,n}(\mathbf{x}) = \sum_{i=1}^{n} w_{z,i}(\boldsymbol{\theta}) K_z(\mathbf{x}_i, \mathbf{x}).$$
(S13)

For any $\delta(\cdot) \in \mathcal{H}$, decomposing it into the linear combination of the basis $\{\lambda_z K_z(\mathbf{x}_i, \cdot)\}_{i=1}^n$ and the orthogonal complement $v(\cdot)$ gives

$$\delta(\cdot) = \sum_{i=1}^{n} \tilde{w}_{z,i}(\boldsymbol{\theta}) \lambda_z K_z(\mathbf{x}_i, \cdot) + v(\cdot),$$

where $\langle v(\cdot), \lambda_z K_z(\cdot, \mathbf{x}_i) \rangle_{\mathcal{H}_z} = 0$ for i = 1, ..., n.

To evaluate $\delta(\cdot)$ at \mathbf{x}_j for any j = 1, ..., n, we have

$$\delta(\mathbf{x}_j) = \left\langle \sum_{i=1}^n \tilde{w}_{z,i}(\boldsymbol{\theta}) \lambda_z K_z(\mathbf{x}_i, \cdot) + v(\cdot), K_z(\mathbf{x}_j, \cdot) \right\rangle_{\mathcal{H}_z}$$
$$= \sum_{i=1}^n \tilde{w}_{z,i}(\boldsymbol{\theta}) \lambda_z K_z(\mathbf{x}_i, \mathbf{x}_j),$$

which is independent from $v(\cdot)$. Hence the first term on right-hand side of (9) is also independent from $v(\cdot)$. For the second term on right-hand side of (9), since $v(\cdot)$ is orthogonal to $\{K_z(\mathbf{x}_i, \cdot)\}_{i=1}^n$, plugging in the decomposition of $\delta(\cdot)$, we have

$$\lambda \|\delta\|_{\mathcal{H}_{z}}^{2} = \lambda \left(\|\sum_{i=1}^{n} \tilde{w}_{z,i}(\boldsymbol{\theta})\lambda_{z}K_{z}(\mathbf{x}_{i},\cdot)\|_{\mathcal{H}_{z}}^{2} + \|v\|_{\mathcal{H}_{z}}^{2} \right)$$
$$\geq \lambda \|\sum_{i=1}^{n} \tilde{w}_{z,i}(\boldsymbol{\theta})\lambda_{z}K_{z}(\mathbf{x}_{i},\cdot)\|_{\mathcal{H}_{z}}^{2}.$$

Thus choosing $v(\cdot) = 0$ does not change the first term on right-hand side of (9), but also minimizes the second term on right-hand side of (9). Letting $w_{z,i}(\boldsymbol{\theta}) = \tilde{w}_{z,i}(\boldsymbol{\theta})\lambda_z$, we have proved (S13). The rest of the proof can be derived similarly as the proof for Lemma S1, so it is omitted here.

S3 Proof for Section 3.1

We prove Theorem 1 in this Section. Two auxiliary lemmas used for the proof of Theorem 1 are given after the proof.

Proof for Theorem 1. Define a new inner product on \mathcal{H} as

$$\langle f, g \rangle_{\lambda} = (1 + \sqrt{\lambda}) \langle f, g \rangle_{L_2(\mathcal{X})} + \lambda \langle f, g \rangle_{\mathcal{H}}$$
 (S14)

Let $f = \sum_{k=1}^{\infty} f_k \phi_k$ and $g = \sum_{k=1}^{\infty} g_k \phi_k$ be elements in \mathcal{H} . Then

$$\langle f,g\rangle_{\lambda} = (1+\sqrt{\lambda})\sum_{k=1}^{\infty} f_k g_k + \lambda \sum_{k=1}^{\infty} \frac{f_k g_k}{\rho_k} = \sum_{k=1}^{\infty} \left(1+\sqrt{\lambda}+\frac{\lambda}{\rho_k}\right) f_k g_k.$$

By letting μ_k by $\mu_k^{-1} = 1 + \sqrt{\lambda} + \lambda/\rho_k$, we can define a new reproducing kernel

$$K_{\lambda}(\mathbf{x}, \mathbf{x}') = \sum_{k=1}^{\infty} \mu_k \phi_k(\mathbf{x}) \phi_k(\mathbf{x}')$$
(S15)

Since $c_{\rho}^{-1}k^{-2m/p} \leq \rho_k \leq C_{\rho}^{-1}k^{-2m/p}$ and $|\phi_i(\cdot)| < C_{\phi}$ for some positive constants c_{ρ} , C_{ρ} and C_{ϕ} , bounding the sums by integrals, we have

$$\sup_{\mathbf{x},\mathbf{x}'} K_{\lambda}(\mathbf{x},\mathbf{x}') \leq C_{\phi}^{2} \sum_{k=1}^{\infty} \frac{1}{1 + \lambda c_{\rho} k^{2m/p}} \leq C_{\phi}^{2} \sum_{k=1}^{\infty} \int_{k-1}^{k} \frac{1}{1 + \lambda c_{\rho} x^{2m/p}} dx$$
$$= C_{\phi}^{2} c_{\rho}^{-p/2m} \lambda^{-p/2m} \int_{0}^{\infty} \frac{(\lambda c_{\rho})^{p/2m}}{1 + \{(\lambda c_{\rho})^{p/2m} x\}^{2m/p}} dx$$
$$= C_{\phi}^{2} c_{\rho}^{-p/2m} \lambda^{-p/2m} \int_{0}^{\infty} \frac{1}{1 + x^{2m/p}} dx.$$

Thus

$$\sup_{\mathbf{x},\mathbf{x}'} K_{\lambda}(\mathbf{x},\mathbf{x}') \le C_K^2 \lambda^{-p/(2m)},\tag{S16}$$

for some constant C_K depending on K. Define the following linear operators $F_{\lambda} : \mathcal{H} \to \mathcal{H}$ and $P_{\lambda} : \mathcal{H} \to \mathcal{H}$ via

$$(F_{\lambda}g)(\mathbf{x}) = \int_{\mathcal{X}} g(\mathbf{x}') K_{\lambda}(\mathbf{x}, \mathbf{x}') d\mathbf{x}', \text{ and } (P_{\lambda}g)(\mathbf{x}) = g(\mathbf{x}) - (F_{\lambda}g)(\mathbf{x}),$$

Clearly, we have

$$\langle f, F_{\lambda}g \rangle_{\lambda} = \sum_{k=1}^{\infty} \langle f, \phi_k \rangle_{L_2(\mathcal{X})} \langle g, \phi_k \rangle_{L_2(\mathcal{X})} = \langle f, g \rangle_{L_2(\mathcal{X})},$$

$$\langle f, P_{\lambda}g \rangle_{\lambda} = \langle f, g \rangle_{\lambda} - \langle f, F_{\lambda}g \rangle_{\lambda} = \sqrt{\lambda} \langle f, g \rangle_{L_2(\mathcal{X})} + \lambda \langle f, g \rangle_{\mathcal{H}}.$$
 (S17)

Denote the loss function

$$\ell_{n\lambda}(f) = \frac{1}{n} \sum_{i=1}^{n} (y_i - f(\mathbf{x}_i))^2 + \sqrt{\lambda} \|f\|_{L_2(\mathcal{X})}^2 + \lambda \|f\|_{\mathcal{H}}^2,$$

and the estimator $\hat{f}_{n\lambda} := \operatorname{argmin}_{f \in \mathcal{H}} \ell_{n\lambda}(f)$. Let $D\ell_{n\lambda}(f) : \mathcal{H} \to \mathcal{H}$ be the Frechét derivative of $\ell_{n\lambda}$ evaluated at f. Clearly, for any $g \in \mathcal{H}$,

$$D\ell_{n\lambda}(f)g = \frac{2}{n} \sum_{i=1}^{n} (f(\mathbf{x}_{i}) - y_{i}) \langle K_{\lambda}(\mathbf{x}_{i}, \cdot), g(\cdot) \rangle_{\lambda} + 2 \langle P_{\lambda}f, g \rangle_{\lambda}$$
$$= \left\langle \frac{2}{n} \sum_{i=1}^{n} (f(\mathbf{x}_{i}) - y_{i}) K_{\lambda}(\mathbf{x}_{i}, \cdot) + 2(P_{\lambda}f)(\cdot), g(\cdot) \right\rangle_{\lambda}.$$
(S18)

It follows that $D\ell_{n\lambda}(\hat{f}_{n\lambda})g = 0$ for all $g \in \mathcal{H}$, and hence, $S_{n\lambda}(\hat{f}_{n\lambda})(\cdot) = 0$, where

$$S_{n\lambda}(f)(\cdot) = \frac{1}{n} \sum_{i=1}^{n} (y_i - f(\mathbf{x}_i)) K_{\lambda}(\mathbf{x}_i, \cdot) - (P_{\lambda}f)(\cdot).$$

Define $S_{\lambda}(f)(\cdot) = \mathbb{E}_{y,\mathbf{x}}(S_{n\lambda}(f)(\cdot))$. Then

$$S_{\lambda}(f)(\cdot) = \int_{\mathbf{x}\in\mathcal{X}} (f_0(\mathbf{x}) - f(\mathbf{x})) K_{\lambda}(\mathbf{x}, \cdot) d\mathbf{x} - (P_{\lambda}f)(\cdot)$$
$$= (F_{\lambda}(f_0 - f))(\cdot) - (P_{\lambda}f)(\cdot) = (F_{\lambda}f_0)(\cdot) - f(\cdot),$$

and therefore, $S_{\lambda}(F_{\lambda}f_0)(\cdot) = 0$. Let $\Delta f = \hat{f}_{n\lambda} - F_{\lambda}f_0$. By the definitions of $S_{n\lambda}$ and S_{λ} , we have

$$\begin{cases} S_{n\lambda}(\hat{f}_{n\lambda}) - S_{\lambda}(\hat{f}_{n\lambda}) \\ &= \left\{ S_{n\lambda}(\hat{f}_{n\lambda}) - S_{n\lambda}(F_{\lambda}f_{0}) \right\} (\cdot) - \left\{ S_{\lambda}(\hat{f}_{n\lambda}) - S_{\lambda}(F_{\lambda}f_{0}) \right\} (\cdot) \\ &= \frac{1}{n} \sum_{i=1}^{n} \left\{ F_{\lambda}f_{0}(\mathbf{x}_{i}) - \hat{f}_{n\lambda}(\mathbf{x}_{i}) \right\} K_{\lambda}(\mathbf{x}_{i}, \cdot) + P_{\lambda} \left\{ (F_{\lambda}f_{0})(\cdot) - \hat{f}_{n\lambda}(\cdot) \right\} \\ &+ \hat{f}_{n\lambda}(\cdot) - (F_{\lambda}f_{0})(\cdot) \\ &= -\frac{1}{n} \sum_{i=1}^{n} \Delta f(\mathbf{x}_{i}) K_{\lambda}(\mathbf{x}_{i}, \cdot) - (P_{\lambda}\Delta f)(\cdot) + (\Delta f)(\cdot) \\ &= -\frac{1}{n} \sum_{i=1}^{n} \Delta f(\mathbf{x}_{i}) K_{\lambda}(\mathbf{x}_{i}, \cdot) + \mathbb{E}_{\mathbf{x}} \left\{ \Delta f(\mathbf{x}) K_{\lambda}(\mathbf{x}, \cdot) \right\}. \end{cases}$$

On the other hand, $S_{n\lambda}(\hat{f}_{n\lambda})(\cdot) = S_{\lambda}(F_{\lambda}f_0)(\cdot) = 0$ and $S_{\lambda}(\hat{f}_{n\lambda})(\cdot) = (F_{\lambda}f_0)(\cdot) - \hat{f}_{n\lambda}(\cdot) = -\Delta f(\cdot)$. Therefore,

$$\left\{S_{n\lambda}(\hat{f}_{n\lambda}) - S_{\lambda}(\hat{f}_{n\lambda})\right\}(\cdot) - \left\{S_{n\lambda}(F_{\lambda}f_{0}) - S_{\lambda}(F_{\lambda}f_{0})\right\}(\cdot) = \Delta f(\cdot) - S_{n\lambda}(F_{\lambda}f_{0})(\cdot).$$

Define the event

$$A_n(t) = \left\{ \left\| \frac{1}{n} \sum_{i=1}^n g(\mathbf{x}_i) K_\lambda(\mathbf{x}_i, \cdot) - \mathbb{E}_{\mathbf{x}} \{ g(\mathbf{x}) K_\lambda(\mathbf{x}, \cdot) \} \right\|_\lambda < t \|g\|_\lambda \text{ for all } g \in \mathcal{H} \right\}.$$

Applying Lemma S3 on $g(\cdot)/\|g\|_{\lambda},$

$$\mathbb{P}_{\mathbf{x}}\{A_n(t)\} \ge 1 - 2\exp\left\{-\lambda^{p(6m-p)/(4m^2)}\left(\frac{nt^2}{\kappa_K}\right)\right\}$$

for some constant $\kappa_K > 0$. The deviation threshold t will be specified later, and from now we consider data points $(\mathbf{x}_i, y_i)_{i=1}^n$ over the event $A_n(t)$.

Over the event $A_n(t)$, we have

$$\left\{S_{n\lambda}(\hat{f}_{n\lambda}) - S_{\lambda}(\hat{f}_{n\lambda})\right\}(\cdot) - \left\{S_{n\lambda}(F_{\lambda}f_{0}) - S_{\lambda}(F_{\lambda}f_{0})\right\}(\cdot)$$
$$= -\frac{1}{n}\sum_{i=1}^{n}\Delta f(\mathbf{x}_{i})K_{\lambda}(\mathbf{x}_{i}, \cdot) + \mathbb{E}_{\mathbf{x}}\left\{\Delta f(\mathbf{x})K_{\lambda}(\mathbf{x}, \cdot)\right\}$$
$$= \Delta f(\cdot) - S_{n\lambda}(F_{\lambda}f_{0})(\cdot),$$

implying that

$$\|\Delta f - S_{n\lambda}(F_{\lambda}f_0)\|_{\lambda} = \left\|\frac{1}{n}\sum_{i=1}^{n}\Delta f(\mathbf{x}_i)K_{\lambda}(\mathbf{x}_i,\cdot) - \mathbb{E}_{\mathbf{x}}\left\{\Delta f(\mathbf{x})K_{\lambda}(\mathbf{x},\cdot)\right\}\right\|_{\lambda}$$
$$\leq t\|\Delta f\|_{\lambda}.$$

Now we proceed to bound $||S_{n\lambda}(F_{\lambda}f_0)||_{\lambda}$. Write

$$\begin{split} \|S_{n\lambda}(F_{\lambda}f_{0})\|_{\lambda} &= \left\|\frac{1}{n}\sum_{i=1}^{n} \{y_{i} - F_{\lambda}f_{0}(\mathbf{x}_{i})\}K_{\lambda}(\mathbf{x}_{i}, \cdot) - (P_{\lambda}F_{\lambda}f_{0})(\cdot)\right\|_{\lambda} \\ &\leq \left\|\frac{1}{n}\sum_{i=1}^{n} \{f_{0}(\mathbf{x}_{i}) - F_{\lambda}f_{0}(\mathbf{x}_{i})\}K_{\lambda}(\mathbf{x}_{i}, \cdot) - \{F_{\lambda}(f_{0} - F_{\lambda}f_{0})\}(\cdot)\right\|_{\lambda} \\ &+ \left\|\frac{1}{n}\sum_{i=1}^{n} \epsilon_{i}K_{\lambda}(\mathbf{x}_{i}, \cdot)\right\|_{\lambda} \\ &= \left\|\frac{1}{n}\sum_{i=1}^{n} \{f_{0}(\mathbf{x}_{i}) - F_{\lambda}f_{0}(\mathbf{x}_{i})\}K_{\lambda}(\mathbf{x}_{i}, \cdot) - \mathbb{E}_{\mathbf{x}}[\{f_{0}(\mathbf{x}) - F_{\lambda}f_{0}(\mathbf{x})\}K_{\lambda}(\mathbf{x}, \cdot)]\right\|_{\lambda} \\ &+ \left\|\frac{1}{n}\sum_{i=1}^{n} \epsilon_{i}K_{\lambda}(\mathbf{x}_{i}, \cdot)\right\|_{\lambda} \\ &\leq t\|f_{0} - F_{\lambda}f_{0}\|_{\lambda} + \left\|\frac{1}{n}\sum_{i=1}^{n} \epsilon_{i}K_{\lambda}(\mathbf{x}_{i}, \cdot)\right\|_{\lambda}, \end{split}$$

where the last inequality is due to the construction of the event $A_n(t)$. To bound the second term of the preceding display, we let $\boldsymbol{\sigma}_{\lambda} = [K_{\lambda}(\mathbf{x}_i, \mathbf{x}_j)]_{n \times n}$ and $\boldsymbol{\epsilon} = [\epsilon_1, \ldots, \epsilon_n]^T$. By the Hanson-Wright inequality (Rudelson et al., 2013), for all x > 0, we have

$$\mathbb{P}_{\mathbf{x}}\left[\boldsymbol{\epsilon}^{T}\boldsymbol{\sigma}_{\lambda}\boldsymbol{\epsilon} \geq \sigma_{0}^{2}\left\{\operatorname{tr}(\boldsymbol{\sigma}_{\lambda}) + 2\sqrt{\operatorname{tr}(\boldsymbol{\sigma}_{\lambda}^{2})}x + 2\|\boldsymbol{\sigma}_{\lambda}\|_{\mathrm{F}}x^{2}\right\}\right] \leq \mathrm{e}^{-x^{2}}.$$

Since by the Cauchy-Schwarz inequality,

$$\operatorname{tr}(\boldsymbol{\Sigma}_{\lambda}) = \sum_{i=1}^{n} \|K_{\lambda}(\mathbf{x}_{i}, \cdot)\|_{\lambda}^{2} = \sum_{i=1}^{n} K_{\lambda}(\mathbf{x}_{i}, \mathbf{x}_{i}) \leq C_{K}^{2} n \lambda^{-p/(2m)},$$

$$\operatorname{tr}(\boldsymbol{\Sigma}_{\lambda}^{2}) \leq \sum_{i=1}^{n} \sum_{j=1}^{n} \|K_{\lambda}(\mathbf{x}_{i}, \cdot)\|_{L_{2}(\mathcal{X})} \|K_{\lambda}(\mathbf{x}_{j}, \cdot)\|_{L_{2}(\mathcal{X})} \leq C_{K}^{4} n^{2} \lambda^{-p/m},$$

$$\|\boldsymbol{\sigma}_{\lambda}\|_{\mathrm{F}} = \sqrt{\operatorname{tr}(\boldsymbol{\Sigma}_{\lambda}^{2})} \leq C_{K}^{2} n \lambda^{-p/(2m)},$$

it follows that

$$\operatorname{tr}(\boldsymbol{\Sigma}_{\lambda}) + 2\sqrt{\operatorname{tr}(\boldsymbol{\Sigma}_{\lambda}^{2})}x + 2\|\boldsymbol{\Sigma}_{\lambda}\|_{\mathrm{F}}x^{2} \leq C_{K}^{2}n\lambda^{-p/(2m)}(1 + 2x + 2x^{2}).$$
(S19)

Set the event B_n to be

$$B_n = \left\{ \left\| \frac{1}{n} \sum_{i=1}^n e_i K_\lambda(\mathbf{x}_i, \cdot) \right\|_\lambda < \sigma_0 C_K n^{-1/2} \lambda^{-p/(4m)} \alpha^{1/2} \right\},\$$

where $\alpha = 2 + 3x^2$. Since $1 + 2x + 2x^2 \le 2 + 3x^2 = \alpha$, by taking $x = \sqrt{(\alpha - 2)/3}$, we have $\mathbb{P}(B_n) \ge 1 - \exp(-(\alpha - 2)/3)$ for any $\alpha > 2$. Putting all pieces obtained above together, we have

$$\begin{split} \|\Delta f\|_{\lambda} &\leq \|\Delta f - S_{n\lambda}(F_{\lambda}f_{0})\|_{\lambda} + \|S_{n\lambda}(F_{\lambda}f_{0})\|_{\lambda} \\ &\leq t\|\Delta f\|_{\lambda} + t\|f_{0} - F_{\lambda}f_{0}\|_{\lambda} + \sigma_{0}C_{K}n^{-1/2}\lambda^{-p/(4m)}\alpha^{1/2} \\ &= t\|\Delta f\|_{\lambda} + t\|P_{\lambda}f_{0}\|_{\lambda} + \sigma_{0}C_{K}n^{-1/2}\lambda^{-p/(4m)}\alpha^{1/2}, \end{split}$$
(S20)

over the event $A_n(t) \cap B_n$. Now take $\lambda = n^{-2m/(2m+p)}$. Choose any $C_\beta \in (0,1)$ and let $t = \sqrt{\kappa_K/(n^{(1-C_\beta)\beta}\log(2))}$. Then, for sufficiently large n,

$$\mathbb{P}_{\mathbf{x}}\{A_n(t)\} \ge 1 - 2\exp\left\{-\frac{n^{\beta}t^2}{\kappa_K}\right\} \ge 1 - \exp\left\{-n^{C_{\beta}\beta}\right\},\$$

where $\beta = (2m - p)^2/(2m(2m + p))$, and therefore,

$$\|\Delta f\|_{\lambda} \le \|P_{\lambda}f_0\|_{\lambda} + 2\sigma_0 C_K n^{-m/(2m+p)} \alpha^{1/2},$$

with probability at least

$$\mathbb{P}\{A_n(t) \cap B_n\} = 1 - \mathbb{P}\{A_n^c(t) \cup B_n^c\}$$

$$\geq 1 - \mathbb{P}\{A_n^c(t)\} - \mathbb{P}(B_n^c) = 1 - \exp\{-(\alpha - 2)/3\} - \exp\{-n^{C_\beta\beta}\}$$

for sufficiently large n. Observe that

$$\begin{split} \|P_{\lambda}f_{0}\|_{\lambda}^{2} &= \left\|\sum_{k=1}^{\infty} (1-\mu_{k})\langle f_{0},\phi_{k}\rangle_{L_{2}(\mathcal{X})}\phi_{k}(\cdot)\right\|_{\lambda}^{2} = \sum_{k=1}^{\infty} \frac{(1-\mu_{k})^{2}}{\mu_{k}}\langle f_{0},\phi_{k}\rangle_{L_{2}(\mathcal{X})}^{2} \\ &= \sum_{k=1}^{\infty} \frac{(\sqrt{\lambda}+\lambda/\rho_{k})^{2}}{1+\sqrt{\lambda}+\lambda/\rho_{k}}\langle f_{0},\phi_{k}\rangle_{L_{2}(\mathcal{X})}^{2} \leq \sum_{k=1}^{\infty} \frac{2\lambda+2(\lambda/\rho_{k})^{2}}{1+\lambda/\rho_{k}}\langle f_{0},\phi_{k}\rangle_{L_{2}(\mathcal{X})}^{2} \\ &\leq 2\lambda\sum_{k=1}^{\infty}\langle f_{0},\phi_{k}\rangle_{L_{2}(\mathcal{X})}^{2} + 2\lambda\sum_{k=1}^{\infty} \frac{1}{\rho_{k}}\langle f_{0},\phi_{k}\rangle_{L_{2}(\mathcal{X})}^{2} \\ &= 2\lambda\|f_{0}\|_{L_{2}(\mathcal{X})}^{2} + 2\lambda\|f_{0}\|_{\mathcal{H}}^{2} \\ &\leq 2n^{-2m/(2m+p)}\left(\|f_{0}\|_{L_{2}(\mathcal{X})} + \|f_{0}\|_{\mathcal{H}}\right)^{2}. \end{split}$$

Hence, we proceed to compute

$$\begin{aligned} \|\hat{f}_{n\lambda} - f_0\|_{L_2(\mathcal{X})} &\leq \|\hat{f}_{n\lambda} - f_0\|_{\lambda} \\ &\leq \|\hat{f}_{n\lambda} - F_{\lambda}f_0\|_{\lambda} + \|F_{\lambda}f_0 - f_0\|_{\lambda} \\ &= \|\Delta f\|_{\lambda} + \|P_{\lambda}f_0\|_{\lambda} \\ &\leq \left(2\sqrt{2}(\|f_0\|_{L_2(\mathcal{X})} + \|f_0\|_{\mathcal{H}}) + 2\sigma_0 C_K \alpha^{1/2}\right) n^{-m/(2m+p)} \end{aligned}$$

with probability at least $1 - \exp\{-(\alpha - 2)/3\} - \exp(-n^{C_{\beta}\beta})$ for sufficiently large *n*. The bound for $\|\hat{f}_{\lambda,\lambda_z,n} - f_0\|_{\mathcal{H}}$ follows immediately by the definition of $\|\cdot\|_{\lambda}$, completing the proof.

The following the Lemma S2 is Theorem 3.6 in Pinelis (1994), which is needed for the proof of Lemma S3.

Lemma S2. Let $(X_j)_{j=0}^{\infty}$ be a sequence of random elements in a Hilbert space \mathcal{H} with norm $\|\cdot\|_{\mathcal{H}}$. Suppose that $(X_j)_{j=0}^{\infty}$ forms a martingale in the sense that $\mathbb{E}(X_j \mid X_0, \ldots, X_{j-1}) = X_j$ a.s., and that the difference sequence $(D_j)_{j=1}^{\infty} = (X_j - X_{j-1})_{j=1}^{\infty}$ satisfies $\|D_j\|_{\mathcal{H}}^2 \leq b_j^2$ a.s. and $\sum_{j=1}^{\infty} b_j^2 \leq b_*^2$. Then for any $t \geq 0$,

$$\mathbb{P}\left(\sup_{j\geq 1} \|X_j\|_{\mathcal{H}} \geq t\right) \leq 2\exp\left(-\frac{t^2}{2b_*^2}\right).$$

The following maximum inequality for functional empirical processes in the Sobolev space $\mathcal{W}_2^m(\mathcal{X}, 1)$, which generalizes Lemma 5.1 in Yang et al. (2017a) to multivariate functions, is of fundamental importance to the proof of Theorem 1.

Lemma S3. Denote $\mathcal{W}_2^m(\mathcal{X}, 1) = \{f \in \mathcal{W}_2^m(\mathcal{X}) : ||f||_{\lambda} \leq 1\}$. Suppose $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are independently and uniformly drawn from \mathcal{X} . Then there exists some constant κ_K depending on the kernel K, such that for any t > 0,

$$\mathbb{P}_{\mathbf{x}}\left(\sup_{g\in\mathcal{W}_{2}^{m}(\mathcal{X},1)}\left\|\frac{1}{n}\sum_{i=1}^{n}\left[g(\mathbf{x}_{i})K_{\lambda}(\mathbf{x}_{i},\cdot)-\mathbb{E}_{\mathbf{x}}\left\{g(\mathbf{x})K_{\lambda}(\mathbf{x},\cdot)\right\}\right]\right\|_{\lambda}\geq t\right)$$
$$\leq 2\exp\left\{-\frac{\lambda^{d(6m-d)/(4m^{2})}nt^{2}}{\kappa_{K}}\right\}.$$

Proof of Lemma S3. We follow the argument used in the proof of Lemma 6.1 in Yang et al. (2017b). Denote

$$\{Z_{n\lambda}(g)\}(\cdot) = \frac{1}{n} \sum_{i=1}^{n} [g(\mathbf{x}_i) K_{\lambda}(\mathbf{x}_i, \cdot) - \mathbb{E}_{\mathbf{x}} \{g(\mathbf{x}) K_{\lambda}(\mathbf{x}, \cdot)\}].$$

Fix $g, h \in \mathcal{H}, n$, and λ , consider the following sequence of martingale $(X_j)_{j=0}^{\infty}$ in \mathcal{H} :

$$X_{j} = \begin{cases} 0, & \text{if } j = 0, \\ j\{Z_{j\lambda}(g) - Z_{j\lambda}(h)\}, & \text{if } j = 1, \dots, n \\ X_{n}, & \text{if } j \ge n+1. \end{cases}$$

Clearly, for $j = 1, \ldots, n$,

$$(X_j - X_{j-1})(\cdot) = \{g(\mathbf{x}_j) - h(\mathbf{x}_j)\}K_{\lambda}(\mathbf{x}_j, \cdot) - \mathbb{E}_{\mathbf{x}}[\{g(\mathbf{x}_j) - h(\mathbf{x}_j)\}K_{\lambda}(\mathbf{x}_j, \cdot)]$$

and $X_j - X_{j-1} = 0$ for $j \ge n + 1$. Observe that

$$\|K_{\lambda}(\mathbf{x}_{j},\cdot)\|_{\lambda} = \sqrt{\langle K_{\lambda}(\mathbf{x}_{j},\cdot), K_{\lambda}(\mathbf{x}_{j},\cdot)\rangle_{\lambda}} = \sqrt{K_{\lambda}(\mathbf{x}_{j},\mathbf{x}_{j})} \leq C_{K}\lambda^{-d/(4m)}$$

with probability one. Therefore, with probability one, we have

$$||X_j - X_{j-1}||_{\lambda}^2 \le 4C_K^2 \lambda^{-d/(2m)} ||g - h||_{L_{\infty}}^2$$

for j = 1, ..., n, and hence, we invoke the bounded difference inequality for martingales in Banach space (Lemma S2) to derive

$$\mathbb{P}\left(\|Z_{n\lambda}(g) - Z_{n\lambda}(h)\|_{\lambda} \ge t\right) = \mathbb{P}\left(\|n\{Z_{n\lambda}(g) - Z_{n\lambda}(h)\}\|_{\lambda} \ge nt\right)$$
$$\leq \mathbb{P}\left(\sup_{j\ge 1} \|j\{Z_{j\lambda}(g) - Z_{j\lambda}(h)\}\|_{\lambda} \ge nt\right)$$
$$\leq 2\exp\left\{-\frac{nt^2}{8C_K^2\lambda^{-d/(2m)}\|g - h\|_{L_{\infty}}^2}\right\}.$$

Applying Lemma 8.1 in Kosorok (2008), we obtain the following bound

$$\|\|Z_{n\lambda}(g) - Z_{n\lambda}(h)\|_{\lambda}\|_{\psi_2} \le \frac{\sqrt{24}C_K^2}{\sqrt{n}}\lambda^{-d/(4m)}\|g - h\|_{L_{\infty}},$$
(S21)

where $\|\cdot\|_{\psi_2}$ is the Orlicz norm associated with $\psi_2(s) = \exp(s^2) - 1$.

Now let $\tau = \{\log(3/2)\}^{1/2}$ and set $\phi(x) = \psi_2(\tau x)$. Clearly, $\phi(1) = 1/2$, and $\phi(x)\phi(y) \le \phi(xy)$ for any $x, y \ge 1$. Applying Lemma 8.2 in Kosorok (2008), the Orlicz norm of the maximum of finitely many random variables can be bounded by the maximum of these

Orlicz norms as follows:

$$\left\| \max_{1 \le i \le k} (\tau \xi_i) \right\|_{\psi_2} = \left\| \max_{1 \le i \le k} \xi_i \right\|_{\phi} \le 2\phi^{-1}(k) \max_{1 \le i \le k} \|\xi_i\|_{\phi} = \frac{2}{\tau} \psi_2^{-1}(k) \max_{1 \le i \le k} \|\tau \xi_i\|_{\psi_2},$$

namely,

$$\left\| \max_{1 \le i \le k} \xi_i \right\|_{\psi_2} \le \frac{2}{\tau} \psi_2^{-1}(k) \max_{1 \le i \le k} \|\xi_i\|_{\psi_2},$$
(S22)

where $\{\xi_i\}_{i=1}^k$ are finitely many random variables.

Next we apply the "chaining" argument. Let $\varepsilon > 0$ be some constant to be determined later. Construct a sequence of function classes $(\mathcal{G}_j)_{j=0}^{\infty}$ in $\mathcal{H}_{\lambda}(1)$ satisfying the following conditions:

- (i) For any \mathcal{G}_j and any $h_j, g_j \in \mathcal{G}_j$, $||h_j g_j||_{L_{\infty}} \ge \varepsilon/2^j$, and \mathcal{G}_j is maximal in the sense that for any $g_j \notin \mathcal{G}_j$, there exists some $h_j \in \mathcal{G}_j$ such that $||h_j g_j|| < \varepsilon/2^j$.
- (ii) For any \mathcal{G}_{j+1} , and any $g_{j+1} \in \mathcal{G}_{j+1}$, select a unique element $g_j \in \mathcal{G}_j$ such that $||g_{j+1} g_j||_{L_{\infty}} \leq \varepsilon/2^j$. Thus, there exists a finite sequence $(g_0, g_1, \ldots, g_{j+1})$ such that $||g_i g_{i+1}||_{L_{\infty}} \leq \varepsilon/2^i$ for $i = 0, \ldots, j$, and $g_i \in \mathcal{G}_i$.

Therefore, for any $g_{j+1}, h_{j+1} \in \mathcal{G}_{j+1}$ with $\|g_{j+1} - h_{j+1}\|_{L_{\infty}} \leq \varepsilon$, there exists two sequences $(g_i)_{i=0}^{j+1}, (h_i)_{i=0}^{j+1}$, such that $g_i, h_i \in \mathcal{G}_i$, $\max\{\|g_i - g_{i+1}\|_{L_{\infty}}, \|h_i - h_{i+1}\|_{L_{\infty}}\} \leq \varepsilon/2^i$, and that

$$||g_0 - h_0||_{L_{\infty}} \le \sum_{i=0}^{j} (||g_i - g_{i+1}||_{L_{\infty}} + ||h_i - h_{i+1}||_{L_{\infty}}) + ||h_{j+1} - g_{j+1}||_{L_{\infty}} \le 2\sum_{i=0}^{j} \frac{\varepsilon}{2^i} + \varepsilon \le 5\varepsilon,$$

and hence, by (S21) one has

$$||||Z_{n\lambda}(g_0) - Z_{n\lambda}(h_0)||_{\lambda}||_{\psi_2} \le \frac{5\sqrt{24}C_K^2}{\sqrt{n}}\lambda^{-p/(4m)}\varepsilon.$$
 (S23)

We also notice that $\mathcal{G}_j \subset \mathcal{H}_{\lambda}(1) \subset \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq \lambda^{-1/2}\}$, and therefore, the cardinality of \mathcal{G}_j can be bounded by the metric entropy of $\{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq \lambda^{-1/2}\}$, which is known in the literature (Edmunds and Triebel, 2008):

$$\log |\mathcal{G}_j| \le \log \mathcal{N}_{[\cdot]} \left(\varepsilon/2^j, \{ f \in \mathcal{H} : \|f\|_{\mathcal{H}} \le \lambda^{-1/2} \}, \|\cdot\|_{L_{\infty}} \right)$$
$$\le c_0 \lambda^{-p/(2m)} \left(\frac{\varepsilon}{2^j} \right)^{-p/m},$$

where c_0 is some absolute constant.

Now suppose g, h are arbitrary functions in $\mathcal{H}_{\lambda}(1)$ such that $||g - h||_{L_{\infty}} \leq \varepsilon/2$. For any $j \geq 2$, there exists $g_j, h_j \in \mathcal{G}_j$ such that

$$\max\{\|g_j - g\|_{L_{\infty}}, \|h_j - h\|_{L_{\infty}}\} \le \varepsilon/2^j,$$

and hence, $||g_j - h_j||_{L_{\infty}} \leq \varepsilon$. Therefore, for any $j \geq 2$,

$$\begin{aligned} \left\| \sup_{g,h\in\mathcal{W}_{2}^{m}(\mathcal{X},1),\|g-h\|_{L_{\infty}}\leq\varepsilon} \|Z_{n\lambda}(g) - Z_{n\lambda}(h)\|_{\lambda} \right\|_{\psi_{2}} \\ &\leq \left\| \sup_{g,h\in\mathcal{W}_{2}^{m}(\mathcal{X},1),\|g-h\|_{L_{\infty}}\leq\varepsilon} \left(\|Z_{n\lambda}(g) - Z_{n\lambda}(g_{j})\|_{\lambda} + \|Z_{n\lambda}(g_{j}) - Z_{n\lambda}(h_{j})\|_{\lambda} \right) \right\|_{\psi_{2}} \\ &+ \|Z_{n\lambda}(h_{j}) - Z_{n\lambda}(h)\|_{\lambda} \right) \right\|_{\psi_{2}} \\ &\leq \frac{2\sqrt{24}C_{K}^{2}}{\sqrt{n}}\lambda^{-d/(4m)} \max\left\{ \|g - g_{j}\|_{L_{\infty}}, \|h - h_{j}\|_{L_{\infty}} \right\} \\ &+ \left\| \sup_{g,h\in\mathcal{W}_{2}^{m}(\mathcal{X},1),\|g-h\|_{L_{\infty}}\leq\varepsilon} \|Z_{n\lambda}(g_{j}) - Z_{n\lambda}(h_{j})\|_{\lambda} \right\|_{\psi_{2}} \\ &\leq \frac{2\sqrt{24}C_{K}^{2}}{\sqrt{n}}\frac{\lambda^{-d/(4m)}\varepsilon}{2^{j}} + \left\| \sup_{g_{j},h_{j}\in\mathcal{G}_{j},\|g_{j}-h_{j}\|_{L_{\infty}}\leq\varepsilon} \|Z_{n\lambda}(g_{j}) - Z_{n\lambda}(h_{j})\|_{\lambda} \right\|_{\psi_{2}}. \end{aligned}$$

We focus on the second term of the preceding display. Fix $j \ge 2$, for any $g_j, h_j \in \mathcal{G}_j$, consider the finite sequences (g_0, g_1, \ldots, g_j) and (h_0, h_1, \ldots, h_j) such that $g_i, h_i \in \mathcal{G}_i$ and $||g_i - g_{i+1}||_{L_{\infty}} \leq \varepsilon/2^i, i = 1, \dots, j-1$. Invoking the inequality (S22), we have

$$\begin{aligned} \left\| \max_{g_{j},h_{j}\in\mathcal{G}_{j},\|g_{j}-h_{j}\|_{L_{\infty}}\leq\varepsilon} \|Z_{n\lambda}(g_{j})-Z_{n\lambda}(h_{j})\|_{\lambda} \right\|_{\psi_{2}} \\ &\leq \left\| \max_{g_{j},h_{j}\in\mathcal{G}_{j},\|g_{j}-h_{j}\|_{L_{\infty}}\leq\varepsilon} \|\{Z_{n\lambda}(g_{j})-Z_{n\lambda}(h_{j})\}-\{Z_{n\lambda}(g_{0})-Z_{n\lambda}(h_{0})\}\|_{\lambda} \right\|_{\psi_{2}} \\ &+ \left\| \max_{g_{0},h_{0}\in\mathcal{G}_{0},\|g_{j}-h_{j}\|_{L_{\infty}}\leq\varepsilon} \|Z_{n\lambda}(g_{0})-Z_{n\lambda}(h_{0})\|_{\lambda} \right\|_{\psi_{2}} \\ &\leq \left\| \max_{g_{j},h_{j}\in\mathcal{G}_{j},\|g_{j}-h_{j}\|_{L_{\infty}}\leq\varepsilon} \|\{Z_{n\lambda}(g_{j})-Z_{n\lambda}(h_{j})\}-\{Z_{n\lambda}(g_{0})-Z_{n\lambda}(h_{0})\}\|_{\lambda} \right\|_{\psi_{2}} \\ &+ \frac{2}{\tau}\sqrt{\log\left(1+|\mathcal{G}_{0}\times\mathcal{G}_{0}|\right)} \max_{(g_{0},h_{0})\in\mathcal{G}_{0}\times\mathcal{G}_{0},\|g_{j}-h_{j}\|_{L_{\infty}}\leq\varepsilon} \|\|Z_{n\lambda}(g_{0})-Z_{n\lambda}(h_{0})\|_{\lambda}\|_{\psi_{2}}. \end{aligned}$$

Clearly, the second term can be bounded by inequality (S23):

$$\frac{2}{\tau}\sqrt{\log\left(1+|\mathcal{G}_{0}\times\mathcal{G}_{0}|\right)}\max_{\substack{(g_{0},h_{0})\in\mathcal{G}_{0}\times\mathcal{G}_{0},\|g_{j}-h_{j}\|_{L_{\infty}}\leq\varepsilon}}\|\|Z_{n\lambda}(g_{0})-Z_{n\lambda}(h_{0})\|_{\lambda}\|_{\psi_{2}}}$$
$$\leq \left(\frac{10\sqrt{24}C_{K}^{2}}{\tau}\right)\frac{\lambda^{-p/(4m)}\varepsilon}{\sqrt{n}}\sqrt{\log\left\{1+\exp\left(2c_{0}\lambda^{-p/(2m)}\varepsilon^{-p/m}\right)\right\}},$$

since |

$$|\mathcal{G}_0 \times \mathcal{G}_0| = |\mathcal{G}_0|^2 \le \exp(2\log \mathcal{N}_{[\cdot]}(\varepsilon, \{\|f\|_\lambda \le 1\}, \|\cdot\|_{L_\infty}))$$

$$\le \exp(2c_0\lambda^{-p/(2m)}\varepsilon^{-p/m}), \tag{S24}$$

it suffices to bound the first term. Write

$$\begin{aligned} & \left\| \max_{g_{j},h_{j}\in\mathcal{G}_{j},\|g_{j}-h_{j}\|_{L_{\infty}}\leq\varepsilon} \left\| \{Z_{n\lambda}(g_{j})-Z_{n\lambda}(h_{j})\} - \{Z_{n\lambda}(g_{0})-Z_{n\lambda}(h_{0})\} \right\|_{\lambda} \right\|_{\psi_{2}} \\ & \leq 2\sum_{i=0}^{j-1} \left\| \max_{(g_{i},g_{i+1})\in\mathcal{G}_{i}\times\mathcal{G}_{i+1},\|g_{i}-g_{i+1}\|_{L_{\infty}}\leq\varepsilon/2^{i}} \left\| Z_{n\lambda}(g_{i+1})-Z_{n\lambda}(g_{i}) \right\|_{\lambda} \right\|_{\psi_{2}} \\ & \leq 2\sum_{i=0}^{j-1} \frac{2}{\tau} \sqrt{\log(1+|\mathcal{G}_{i}|\times|\mathcal{G}_{i+1}|)} \\ & \times \max_{(g_{i},g_{i+1})\in\mathcal{G}_{i}\times\mathcal{G}_{i+1},\|g_{i}-g_{i+1}\|_{L_{\infty}}\leq\varepsilon/2^{i}} \left\| \left\| Z_{n\lambda}(g_{i+1})-Z_{n\lambda}(g_{i}) \right\|_{\lambda} \right\|_{\psi_{2}} \\ & \leq \frac{4\sqrt{24}C_{K}^{2}\lambda^{-p/(4m)}}{\tau\sqrt{n}} \sum_{i=0}^{j-1} \sqrt{\log\left[1+\exp\left\{2c_{0}\lambda^{-p/(2m)}\left(\varepsilon/2^{i}\right)^{-p/m}\right\}\right]}\varepsilon/2^{i}, \end{aligned}$$

where inequalities (S21) and (S22) are applied. Bounding the sum by integral, we have

$$\sum_{i=0}^{j-1} \sqrt{\log\left[1 + \exp\left\{2c_0\lambda^{-p/(2m)}\left(\varepsilon/2^i\right)^{-p/m}\right\}\right]} \varepsilon/2^i$$
$$\leq \sum_{i=0}^{j-1} \int_{\varepsilon/2^{i+1}}^{\varepsilon/2^i} \sqrt{\log\{1 + \exp(2c_0\lambda^{-p/(2m)}x^{-p/m})\}} dx$$
$$\leq \int_0^\varepsilon \sqrt{\log\{1 + \exp(2c_0\lambda^{-p/(2m)}x^{-p/m})\}} dx.$$

Putting all pieces above together, we obtain the following bound:

$$\left\| \sup_{g,h\in\mathcal{W}_{2}^{m}(\mathcal{X},1), \|g-h\|_{L_{\infty}}\leq\varepsilon} \|Z_{n\lambda}(g) - Z_{n\lambda}(h)\|_{\lambda} \right\|_{\psi_{2}}$$

$$\lesssim \frac{\lambda^{-p/(4m)}}{\sqrt{n}} \left[\frac{\varepsilon}{2^{j}} + \int_{0}^{\varepsilon} \sqrt{\log\{1 + \exp(2c_{0}\lambda^{-p/(2m)}x^{-p/m})\}} dx + \varepsilon \sqrt{\log\{1 + \exp(2c_{0}\lambda^{-p/(2m)}\varepsilon^{-p/m})\}} \right].$$

By taking $j \to \infty$, we can let the first term in the squared bracket tend to 0, and hence,

$$\begin{aligned} \left\| \sup_{g,h \in \mathcal{W}_{2}^{m}(\mathcal{X},1), \|g-h\|_{L_{\infty}} \leq \varepsilon} \|Z_{n\lambda}(g) - Z_{n\lambda}(h)\|_{\lambda} \right\|_{\psi_{2}} \\ &\lesssim \frac{\lambda^{-p/(4m)}}{\sqrt{n}} \left[\int_{0}^{\varepsilon} \sqrt{\log\{1 + \exp(2c_{0}\lambda^{-p/(2m)}x^{-p/m})\}} dx + \varepsilon \sqrt{\log\{1 + \exp(2c_{0}\lambda^{-p/(2m)}\varepsilon^{-p/m})\}} \right] \\ &\lesssim \frac{\lambda^{-p/(4m)}}{\sqrt{n}} \int_{0}^{\varepsilon} \sqrt{\log\{1 + \exp(2c_{0}\lambda^{-p/(2m)}x^{-p/m})\}} dx. \end{aligned}$$

Now we take h = 0, which implies $Z_{n\lambda}(h) = 0$ by the construction of $Z_{n\lambda}$. Furthermore, by the property of reproducing kernel K_{λ} and the Cauchy-Schwarz inequality,

$$\|g - h\|_{L_{\infty}} \leq \sup_{\mathbf{x} \in \mathcal{X}} |g(\mathbf{x})| = \sup_{\mathbf{x} \in \mathcal{X}} |\langle g(\cdot), K_{\lambda}(\mathbf{x}, \cdot) \rangle_{\lambda}|$$
$$\leq \sup_{\mathbf{x} \in \mathcal{X}} \|g\|_{\lambda} \sqrt{\langle K_{\lambda}(\mathbf{x}, \cdot), K_{\lambda}(\mathbf{x}, \cdot) \rangle_{\lambda}} \leq C_{K} \lambda^{-p/(4m)}.$$

Taking $\varepsilon = C_K \lambda^{-p/(4m)}$, we obtain

$$\begin{aligned} \left\| \sup_{g \in \mathcal{W}_{2}^{m}(\mathcal{X},1)} \|Z_{n\lambda}(g)\|_{\lambda} \right\|_{\psi_{2}} \\ &\leq \left\| \sup_{g \in \mathcal{W}_{2}^{m}(\mathcal{X},1), \|g\|_{L_{\infty}} \leq \varepsilon} \|Z_{n\lambda}(g)\|_{\lambda} \right\|_{\psi_{2}} \\ &\leq \left\| \sup_{g,h \in \mathcal{W}_{2}^{m}(\mathcal{X},1), \|g-h\|_{L_{\infty}} \leq \varepsilon} \|Z_{n\lambda}(g) - Z_{n\lambda}(h)\|_{\lambda} \right\|_{\psi_{2}} \\ &\lesssim n^{-1/2} \lambda^{-p/(4m)} \int_{0}^{\epsilon} \sqrt{\log\{1 + \exp(2c_{0}\lambda^{-p/(2m)}x^{-p/m})} \mathrm{d}x \\ &\lesssim n^{-1/2} \lambda^{-p(6m-p)/(8m^{2})}. \end{aligned}$$

Hence, invoking Lemma 8.1 in Kosorok (2008), we finally obtain

$$\mathbb{P}\left(\sup_{g\in\mathcal{W}_2^m(\mathcal{X},1)}\|Z_{n\lambda}(g)\|_{\lambda}>t\right)\leq 2\exp\left\{-\frac{nt^2}{\kappa_K^2\lambda^{-p(6m-p)/(4m^2)}}\right\},\,$$

for some absolute constant κ_K depending on K only, completing the proof.

S4 Proof for Section 3.2

Denote $\hat{\boldsymbol{\theta}}_{z} := \hat{\boldsymbol{\theta}}_{\lambda,\lambda_{z},n}, \, \hat{\delta}_{z}(\cdot) := \hat{\delta}_{\lambda,\lambda_{z},n}(\cdot) \text{ and } \ell_{z}(\boldsymbol{\theta},\delta) := \ell_{\lambda,\lambda_{z},n}(\boldsymbol{\theta},\delta) \text{ in } (9).$

We need the following Corollary 5 and Lemma S4 to prove theorem 2. Corollary 5 is a direct consequence of Theorem 1. We repeatedly use the fact that for any $f(\cdot) \in L_2(\mathcal{X})$, there exists a constant C_{ρ} such that $\|f\|_{L_2(\mathcal{X})} \leq C_{\rho} \|f\|_{\mathcal{H}}$ in the following proof.

Corollary 5. Denote $\hat{\delta}_{z,\theta} = \operatorname{argmin}_{\delta \in \mathcal{H}} \ell_z(\theta, \delta)$ for each $\theta \in \Theta$. Under the Assumptions A1 to A6, for sufficiently large n and any $\alpha > 2$ and $C_\beta \in (0,1)$, with probability at least $1 - \exp\{-(\alpha - 2)/3\} - \exp\{-n^{C_\beta\beta}\}$, one has

$$\begin{split} \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\hat{\delta}_{z,\boldsymbol{\theta}}(\cdot) - (y^{R}(\cdot) - f^{M}(\cdot,\boldsymbol{\theta}))\|_{L_{2}(\mathcal{X})} \\ &\leq 2 \bigg[\sqrt{2} \bigg(\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|y^{R}(\cdot) - f^{M}(\cdot,\boldsymbol{\theta})\|_{L_{2}(\mathcal{X})} \\ &+ \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|y^{R}(\cdot) - f^{M}(\cdot,\boldsymbol{\theta})\|_{\mathcal{H}} \bigg) + C_{K} \sigma_{0} \alpha^{1/2} \bigg] n^{-\frac{m}{2m+p}}, \end{split}$$

and

$$\begin{split} \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\hat{\delta}_{z,\boldsymbol{\theta}}(\cdot)\|_{\mathcal{H}} &\leq (2\sqrt{2}+1) \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|(y^{R}(\cdot)-f^{M}(\cdot,\boldsymbol{\theta}))\|_{\mathcal{H}} \\ &+ 2\sqrt{2} \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|y^{R}(\cdot)-f^{M}(\cdot,\boldsymbol{\theta})\|_{L_{2}(\mathcal{X})} + 2\sqrt{2}C_{K}\sigma_{0}\alpha^{1/2} \end{split}$$

by choosing $\lambda = n^{-2m/(2m+p)}$ and $\lambda_z = \lambda^{-1/2}$, where C_K is a constant depending on the kernel $K(\cdot, \cdot)$.

Lemma S4. Under assumptions A1 to A6,

(i) it holds that

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left| \frac{1}{n} \sum_{i=1}^{n} (y^{R}(\mathbf{x}_{i}) - f^{M}(\mathbf{x}_{i}, \boldsymbol{\theta}) - \hat{\delta}_{z, \boldsymbol{\theta}}(\mathbf{x}_{i}))^{2} - \int_{\mathbf{x}\in\mathcal{X}} (y^{R}(\mathbf{x}) - f^{M}(\mathbf{x}, \boldsymbol{\theta}) - \hat{\delta}_{z, \boldsymbol{\theta}}(\mathbf{x}))^{2} d\mathbf{x} \right| = o_{p}(n^{-1/2}),$$

and

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left| \frac{1}{n} \sum_{i=1}^{n} (y^{R}(\mathbf{x}_{i}) - f^{M}(\mathbf{x}_{i}, \boldsymbol{\theta}) - \hat{\delta}_{z, \boldsymbol{\theta}}(\mathbf{x}_{i})) \epsilon_{i} \right| = o_{p}(n^{-1/2});$$

(ii) for any j = 1, ..., q, one has

$$\frac{1}{n} \sum_{i=1}^{n} (y^{R}(\mathbf{x}_{i}) - f^{M}(\mathbf{x}_{i}, \hat{\boldsymbol{\theta}}_{z}) - \hat{\delta}_{z}(\mathbf{x}_{i})) \frac{\partial f^{M}(\mathbf{x}_{i}, \hat{\boldsymbol{\theta}}_{z})}{\partial \theta_{j}} \\ = \int_{\mathbf{x} \in \mathcal{X}} (y^{R}(\mathbf{x}) - f^{M}(\mathbf{x}, \hat{\boldsymbol{\theta}}_{z}) - \hat{\delta}_{z}(\mathbf{x})) \frac{\partial f^{M}(\mathbf{x}, \hat{\boldsymbol{\theta}}_{z})}{\partial \theta_{j}} d\mathbf{x} + o_{p}(n^{-1/2}).$$

Proof. Denote

$$\mathcal{W}_2^m(\mathcal{H}, B) := \left\{ f(\cdot) = \sum_{j=1}^\infty f_j \phi(\cdot) \in L_2(\mathcal{X}) : \sum_{j=1}^\infty j^{2m/p} f_j^2 \le B^2 \right\},$$

and

$$s_i^2(\boldsymbol{\theta}, \delta) := (y^R(\mathbf{x}_i) - f^M(\mathbf{x}_i, \boldsymbol{\theta}) - \delta(\mathbf{x}_i))^2,$$
$$u_i(\boldsymbol{\theta}, \delta) := (y^R(\mathbf{x}_i) - f^M(\mathbf{x}_i, \boldsymbol{\theta}) - \delta(\mathbf{x}_i))\epsilon_i,$$
$$r_i^2(\boldsymbol{\theta}, \delta) := (y^R(\mathbf{x}_i) - f^M(\mathbf{x}_i, \boldsymbol{\theta}) - \delta(\mathbf{x}_i))\frac{\partial f^M(\mathbf{x}, \boldsymbol{\theta})}{\partial \theta_j}$$

for $(\boldsymbol{\theta}, \delta) \in \boldsymbol{\Theta} \times \mathcal{W}_2^m(\mathcal{X}, B)$ and some B > 0 that will be specified later. Define the empirical

processes

$$\bar{s}^{2}(\boldsymbol{\theta}, \delta) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{s_{i}^{2}(\delta, \boldsymbol{\theta}) - \mathbb{E}_{\mathbf{x}_{i}}[s_{i}^{2}(\delta, \boldsymbol{\theta})]\},\\ \bar{u}(\boldsymbol{\theta}, \delta) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{u_{i}(\delta, \boldsymbol{\theta}) - \mathbb{E}_{\mathbf{x}_{i}, \epsilon_{i}}[u_{i}(\delta, \boldsymbol{\theta})]\},\\ \bar{r}(\boldsymbol{\theta}, \delta) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{r_{i}(\delta, \boldsymbol{\theta}) - \mathbb{E}_{\mathbf{x}_{i}}[r_{i}(\delta, \boldsymbol{\theta})]\},$$

where

$$\mathbb{E}_{\mathbf{x}_{i}}[s_{i}^{2}(\boldsymbol{\theta}, \delta)] = \int_{\mathbf{x}\in\mathcal{X}} (y^{R}(\mathbf{x}) - f^{M}(\mathbf{x}, \boldsymbol{\theta}) - \delta(\mathbf{x}))^{2} d\mathbf{x},$$
$$\mathbb{E}_{\mathbf{x}_{i},\epsilon_{i}}[u_{i}(\boldsymbol{\theta}, \delta)] = \int_{\mathbf{x}\in\mathcal{X}} (y^{R}(\mathbf{x}) - f^{M}(\mathbf{x}, \boldsymbol{\theta}) - \delta(\mathbf{x}))\epsilon_{i}d\mathbf{x} = 0,$$
$$\mathbb{E}_{\mathbf{x}_{i}}[r_{i}(\boldsymbol{\theta}, \delta)] = \int_{\mathbf{x}\in\mathcal{X}} (y^{R}(\mathbf{x}) - f^{M}(\mathbf{x}, \boldsymbol{\theta}) - \delta(\mathbf{x}))\frac{\partial f^{M}(\mathbf{x}, \boldsymbol{\theta})}{\partial \theta_{j}}d\mathbf{x}.$$

By Assumptions A3 and A4, the function classes $\{\partial f^M(\cdot, \boldsymbol{\theta})/\partial \theta_j : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ and $\mathcal{F} = \{y^R(\cdot) - f^M(\cdot, \boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ are Donsker. Note that, by definition, $\mathcal{W}_2^m(\mathcal{X}, B)$ is also Donsker. Since both $\mathcal{W}_2^m(\mathcal{X}, B)$ and \mathcal{F} are uniformly bounded, the function classes

$$\left\{ (y^{R}(\cdot) - f^{M}(\cdot, \boldsymbol{\theta}) - \delta(\cdot))^{2} : \boldsymbol{\theta} \in \boldsymbol{\Theta}, \delta \in \mathcal{W}_{2}^{m}(\mathcal{X}, B) \right\}, \text{ and} \\ \left\{ (y^{R}(\cdot) - f^{M}(\cdot, \boldsymbol{\theta}) - \delta(\cdot)) \frac{\partial f^{M}(\cdot, \boldsymbol{\theta})}{\partial \theta_{j}} : \boldsymbol{\theta} \in \boldsymbol{\Theta}, \delta \in \mathcal{W}_{2}^{m}(\mathcal{X}, B) \right\}$$

are also Donsker classes. Furthermore, letting $f_{\theta,\delta}(\epsilon, \mathbf{x}) = (y^R(\mathbf{x}) - f^M(\mathbf{x}, \theta) - \delta(\mathbf{x}))\epsilon$, observe that for any (θ_1, δ_1) and (θ_2, δ_2) , the distance

$$\begin{split} \left\{ \mathbb{E}_0 \left[(f_{\theta_1,\delta_1} - f_{\theta_2,\delta_2})^2 \right] \right\}^{1/2} \\ &= \sigma_0 \| f^M(\cdot, \theta_1) - \delta_1(\cdot) - f^M(\cdot, \theta_2) + \delta_2(\cdot) \|_{L_2(\mathcal{X})} \\ &\leq \sigma_0 \left[\| f^M(\cdot, \theta_1) - f^M(\cdot, \theta_2) \|_{L_2(\mathcal{X})} + \| \delta_1(\cdot) - \delta_2(\cdot) \|_{L_2(\mathcal{X})} \right] \end{split}$$

can be bounded by the $L_2(\mathcal{X})$ -distance of functions in $\{f^M(\cdot, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ and $\delta(\cdot) \in \mathcal{W}_2^m(\mathcal{X}, B)$. In addition, by Assumption A4 $\{f^M(\cdot, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ and $\mathcal{W}_2^m(\mathcal{X}, B)$ are Donsker classes, it follows that the function class

$$\{f_{\theta,\delta} \in C(\mathbb{R} \times \mathcal{X}) : \theta \in \Theta, \delta \in \mathcal{W}_2^m(\mathcal{X}, B)\}$$

is also Donsker, since its metric entropy can be upper bounded by those of $\{f^M(\cdot, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ and $\mathcal{W}_2^m(\mathcal{X}, B)$. By Theorem 2.4 in Mammen and Van de Geer (1997), for any $t_1 > 0$ and any B > 0, there exists $t_2, t'_2, t''_2 > 0$ such that

$$\limsup_{n \to \infty} P\left(\sup_{\|\delta\|_{\mathcal{H}} \le B, \boldsymbol{\theta} \in \boldsymbol{\Theta}, \|y^{R}(\cdot) - f^{M}(\cdot, \boldsymbol{\theta}) - \delta(\cdot)\|_{L_{2}(\mathcal{X})} \le t_{2}} |\bar{s}^{2}(\boldsymbol{\theta}, \delta)| > t_{1}\right) < t_{1}, \quad (S25)$$

$$\limsup_{n \to \infty} P\left(\sup_{\|\delta\|_{\mathcal{H}} \le B, \,\boldsymbol{\theta} \in \boldsymbol{\Theta}, \, \|y^{R}(\cdot) - f^{M}(\cdot,\boldsymbol{\theta}) - \delta(\cdot)\|_{L_{2}(\mathcal{X})} \le t_{2}'} |\bar{r}(\boldsymbol{\theta}, \delta)| > t_{1}\right) < t_{1}, \quad (S26)$$

$$\limsup_{n \to \infty} P\left(\sup_{\|\delta\|_{\mathcal{H}} \le B, \,\boldsymbol{\theta} \in \boldsymbol{\Theta}, \, \|y^{R}(\cdot) - f^{M}(\cdot,\boldsymbol{\theta}) - \delta(\cdot)\|_{L_{2}(\mathcal{X})} \le t_{2}''} |\bar{u}(\boldsymbol{\theta}, \delta)| > t_{1}\right) < t_{1}.$$
(S27)

Note that by Corollary 5, $\sup_{\theta \in \Theta} \|\hat{\delta}_{z,\theta}\|_{\mathcal{H}}$ is asymptotically tight, and therefore for any $\varepsilon > 0$, there exists $B_0 > 0$ and some integer $N \in \mathbb{N}_+$, both depending on ϵ , such that $P(\sup_{\theta \in \Theta} \|\hat{\delta}_{z,\theta}\|_{\mathcal{H}} > B_0) \le \varepsilon/3$ for all n > N. Now take $B = B_0$, $t_1 = \epsilon/3$. Then we can choose t_2 to be a value that satisfies (S25), t'_2 satisfying (S26), and t''_2 satisfying (S27). By Corollary 5 and Assumption A5, $\sup_{\theta} \|\hat{\delta}_{z,\theta}(\cdot) - (y^R(\cdot) - f^M(\cdot, \theta))\|_{L_2(\mathcal{X})} = O_P(n^{-2m/(2m+p)})$, and hence there exists $t_3 > 0$, depending on ϵ and n, such that for all n > N, it holds that

$$P\left(\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\|y^{R}(\cdot)-f^{M}(\cdot,\boldsymbol{\theta})-\hat{\delta}_{z,\boldsymbol{\theta}}(\cdot)\|_{L_{2}(\mathcal{X})}\geq t_{3}\right)<\varepsilon/3.$$

Without loss of generality, we may require $t_3 \leq \min\{t_2, t'_2, t''_2\}$ by taking sufficiently large n.

Then for sufficiently large n, we obtain

$$P\left(\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left|\bar{s}^{2}(\boldsymbol{\theta},\hat{\delta}_{z,\boldsymbol{\theta}})\right|>\varepsilon\right)$$

$$\leq P\left(\sup_{\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left\|\hat{\delta}_{z,\boldsymbol{\theta}}\right\|_{\mathcal{H}}\leq B_{0},\boldsymbol{\theta}\in\boldsymbol{\Theta},\|y^{R}(\cdot)-f^{M}(\cdot,\boldsymbol{\theta})-\hat{\delta}_{z,\boldsymbol{\theta}}(\cdot)\|_{L_{2}(\mathcal{X})}\leq t_{2}}\left|\bar{s}^{2}(\boldsymbol{\theta},\hat{\delta}_{z,\boldsymbol{\theta}})\right|>t_{1}\right)$$

$$+P\left(\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\|y^{R}(\cdot)-f^{M}(\cdot,\boldsymbol{\theta})-\hat{\delta}_{z,\boldsymbol{\theta}}(\cdot)\|_{L_{2}(\mathcal{X})}>t_{2}\right)$$

$$+P\left(\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\|\hat{\delta}_{z,\boldsymbol{\theta}}\|_{\mathcal{H}}>B_{0}\right)$$

$$<\varepsilon/3+\varepsilon/3+\varepsilon/3=\varepsilon,$$

and similarly,

$$P\left(\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}|\bar{r}(\boldsymbol{\theta},\hat{\delta}_{z,\boldsymbol{\theta}})|>\varepsilon\right)<\epsilon\quad\text{and}\quad P\left(\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}|\bar{u}(\boldsymbol{\theta},\hat{\delta}_{z,\boldsymbol{\theta}})|>\varepsilon\right)<\epsilon.$$

Therefore,

$$\begin{aligned} \frac{1}{\sqrt{n}} \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left| \bar{s}(\boldsymbol{\theta}, \hat{\delta}_{z, \boldsymbol{\theta}}) \right| \\ &= \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left| \frac{1}{n} \sum_{i=1}^{n} (y^{R}(\mathbf{x}_{i}) - f^{M}(\mathbf{x}_{i}, \boldsymbol{\theta}) - \hat{\delta}_{z, \boldsymbol{\theta}}(\mathbf{x}_{i}))^{2} - \int_{\mathbf{x} \in \mathcal{X}} (y^{R}(\mathbf{x}) - f^{M}(\mathbf{x}, \boldsymbol{\theta}) - \hat{\delta}_{z, \boldsymbol{\theta}}(\mathbf{x}))^{2} d\mathbf{x} \right| = o_{p}(n^{-1/2}), \end{aligned}$$

and

$$\frac{1}{\sqrt{n}}\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left|\bar{u}(\boldsymbol{\theta},\hat{\delta}_{z,\boldsymbol{\theta}})\right| = \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left|\frac{1}{n}\sum_{i=1}^{n}(y^{R}(\mathbf{x}_{i}) - f^{M}(\mathbf{x}_{i},\boldsymbol{\theta}) - \hat{\delta}_{z,\boldsymbol{\theta}}(\mathbf{x}_{i}))\epsilon_{i}\right| = o_{p}(n^{-1/2}),$$

completing the proof of (i). The proof of (ii) can be completed by observing that

$$\begin{aligned} \left| \frac{1}{n} \sum_{i=1}^{n} (y^{R}(\mathbf{x}_{i}) - f^{M}(\mathbf{x}_{i}, \hat{\boldsymbol{\theta}}_{z}) - \hat{\delta}_{z}(\mathbf{x}_{i})) \frac{\partial f^{M}(\mathbf{x}_{i}, \hat{\boldsymbol{\theta}}_{z})}{\partial \theta_{j}} \right. \\ &\left. - \int_{\mathbf{x} \in \mathcal{X}} (y^{R}(\mathbf{x}) - f^{M}(\mathbf{x}, \hat{\boldsymbol{\theta}}_{z}) - \hat{\delta}_{z}(\mathbf{x})) \frac{\partial f^{M}(\mathbf{x}, \hat{\boldsymbol{\theta}}_{z})}{\partial \theta_{j}} d\mathbf{x} \right| \\ &\left. = \frac{1}{\sqrt{n}} |\bar{r}(\hat{\boldsymbol{\theta}}_{z}, \hat{\delta}_{z, \hat{\boldsymbol{\theta}}_{z}})| \leq \frac{1}{\sqrt{n}} \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} |\bar{r}(\boldsymbol{\theta}, \hat{\delta}_{z, \boldsymbol{\theta}})| = o_{p}(n^{-1/2}). \end{aligned}$$

Proof for Theorem 2. Without loss of generality, it suffices to prove the case when $\lambda_z = \lambda^{-1/2}$. For the general case when $\lambda_z = O(\lambda^{-1/2})$, the proof follows similarly. We first show $\hat{\theta}_z \to^p \theta_{L_2}$. By the definition of $\hat{\theta}_z$, θ_{L_2} , and the theory of M-estimators (see, Theorem 5.7 in Van der Vaart (2000)), it suffices to show that $\lambda^{-1/2}(\ell_z(\theta, \hat{\delta}_{z,\theta}) - \sigma_0^2) \to^p ||y^R(\cdot) - f^M(\cdot, \theta)||^2_{L_2(\mathcal{X})}$ uniformly for each $\theta \in \Theta$. Note that

$$\ell_{z}(\hat{\delta}_{z,\boldsymbol{\theta}}(\cdot),\boldsymbol{\theta})$$

$$=\frac{1}{n}\sum_{i=1}^{n}(y^{R}(\mathbf{x}_{i})-f^{M}(\mathbf{x}_{i},\boldsymbol{\theta})-\hat{\delta}_{z,\boldsymbol{\theta}}(\mathbf{x}_{i}))^{2}+\frac{1}{n}\sum_{i=1}^{n}\epsilon i lon_{i}^{2}$$

$$+\frac{2}{n}\sum_{i=1}^{n}(y^{R}(\mathbf{x}_{i})-f^{M}(\mathbf{x}_{i},\boldsymbol{\theta})-\hat{\delta}_{z,\boldsymbol{\theta}}(\mathbf{x}_{i}))\epsilon_{i}+\lambda\|\hat{\delta}_{z,\boldsymbol{\theta}}\|_{\mathcal{H}}^{2}+\sqrt{\lambda}\|\hat{\delta}_{z,\boldsymbol{\theta}}\|_{L_{2}(\mathcal{X})}^{2}$$

$$:=A_{n}+B_{n}+C_{n}+D_{n}+E_{n}.$$

For A_n , by Lemma S4 (i) and Corollary (5), one has

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left| \frac{1}{n} \sum_{i=1}^{n} (y^{R}(\mathbf{x}_{i}) - f^{M}(\mathbf{x}_{i}, \boldsymbol{\theta}) - \hat{\delta}_{z, \boldsymbol{\theta}}(\mathbf{x}_{i}))^{2} \right| = o_{p}(n^{-1/2})$$
(S28)

Since $\mathbb{E}[B_n] = \sigma_0^2$ and $\mathbb{V}[B_n] = O(n^{-1})$, Chebyshev's inequality implies $(1/n) \sum_{i=1}^n \epsilon_i lon_i^2 = 0$

 $\sigma_0^2 + O_p(n^{-1/2})$ for B_n . For C_n , Lemma S4 (i) guarantees that

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \frac{2}{n} \sum_{i=1}^{n} (y^{R}(\mathbf{x}_{i}) - f^{M}(\mathbf{x}_{i}, \boldsymbol{\theta}) - \hat{\delta}_{z, \boldsymbol{\theta}}(\mathbf{x}_{i})) \epsilon_{i} = o_{p}(n^{-1/2})$$

Since $\lambda = O(n^{-2m/(2m+p)})$, by the asymptotic tightness of $\sup_{\theta \in \Theta} \|\hat{\delta}_{z,\theta}\|_{\mathcal{H}}$ (Corollary 5), one has $\sup_{\theta \in \Theta} \lambda \|\hat{\delta}_{z,\theta}\|_{\mathcal{H}}^2 = o_p(n^{-1/2})$. By putting the above all pieces together, we obtain

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left| \lambda^{-1/2} (\ell_z(\hat{\delta}_{z,\boldsymbol{\theta}}(\cdot),\boldsymbol{\theta}) - \sigma_0^2) - \| \hat{\delta}_{z,\boldsymbol{\theta}} \|_{L_2(\mathcal{X})}^2 \right| = O_p((\lambda n)^{-1/2}).$$
(S29)

For any $\boldsymbol{\theta}$, by the Cauchy-Schwarz inequality, one has

$$\begin{aligned} \left\| \hat{\delta}_{z,\boldsymbol{\theta}} \right\|_{L_{2}(\mathcal{X})}^{2} &- \| y^{R}(\cdot) - f^{M}(\cdot,\boldsymbol{\theta}) \|_{L_{2}(\mathcal{X})}^{2} \\ &\leq \| (\hat{\delta}_{z,\boldsymbol{\theta}}(\cdot) - (y^{R}(\cdot) - f^{M}(\cdot,\boldsymbol{\theta})) \|_{L_{2}(\mathcal{X})} \| \hat{\delta}_{z,\boldsymbol{\theta}}(\cdot) + y^{R}(\cdot) - f^{M}(\cdot,\boldsymbol{\theta}) \|_{L_{2}(\mathcal{X})} \end{aligned}$$

Recall that

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|(\hat{\delta}_{z,\boldsymbol{\theta}}(\cdot) - (y^{R}(\cdot) - f^{M}(\cdot,\boldsymbol{\theta}))\|_{L_{2}(\mathcal{X})} = O_{p}(n^{-m/(2m+d)})$$

by Corollary 5 and Assumption A4. Using Assumptions A4 and the asymptotic tightness of $\sup_{\theta} \|\hat{\delta}_{z,\theta}\|_{\mathcal{H}}$ (Corollary 5), one has

$$\begin{aligned} \|\hat{\delta}_{z,\boldsymbol{\theta}}(\cdot) + y^{R}(\cdot) - f^{M}(\cdot,\boldsymbol{\theta})\|_{L_{2}(\mathcal{X})} \\ &\leq \|\hat{\delta}_{z,\boldsymbol{\theta}}(\cdot)\|_{L_{2}(\mathcal{X})} + \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|y^{R}(\cdot) - f^{M}(\cdot,\boldsymbol{\theta})\|_{L_{2}(\mathcal{X})} \\ &\leq C_{\rho} \|\hat{\delta}_{z,\boldsymbol{\theta}}(\cdot)\|_{\mathcal{H}} + \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|y^{R}(\cdot) - f^{M}(\cdot,\boldsymbol{\theta})\|_{\mathcal{H}} = O_{p}(1). \end{aligned}$$

Thus

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left| \| \hat{\delta}_{z,\boldsymbol{\theta}} \|_{L_2(\mathcal{X})}^2 - \| y^R(\cdot) - f^M(\cdot,\boldsymbol{\theta}) \|_{L_2(\mathcal{X})}^2 \right| = O_p(n^{-m/(2m+d)}),$$

and hence,

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left| \lambda^{-1/2} (\ell_z(\boldsymbol{\theta}, \hat{\delta}_{z, \boldsymbol{\theta}}) - \sigma_0^2) - \| y^R(\cdot) - f^M(\cdot, \boldsymbol{\theta}) \|_{L_2(\mathcal{X})}^2 \right| = o_p(1),$$

from which we conclude $\hat{\boldsymbol{\theta}}_z \rightarrow^p \boldsymbol{\theta}_{L_2}$.

Next we derive the convergence rate of $\hat{\theta}_z$. Apply the Fréchet derivative on ℓ_z with regard to $\delta(\cdot)$ and the partial derivative on ℓ_z with regard to θ_j , j = 1, ..., q. For any $g(\cdot) \in \mathcal{H}$, $\hat{\delta}_z$ and $\hat{\theta}_z$ satisfy

$$0 = -\frac{2}{n} \sum_{i=1}^{n} (y_i^F - f(\mathbf{x}_i, \hat{\boldsymbol{\theta}}_z) - \hat{\delta}_z(\mathbf{x}_i))g(\mathbf{x}_i) + 2\lambda \langle \hat{\delta}_z(\cdot), g(\cdot) \rangle_{\mathcal{H}} + 2\sqrt{\lambda} \langle \hat{\delta}_z(\cdot), g(\cdot) \rangle_{L_2(\mathcal{X})},$$
(S30)

$$0 = -\frac{2}{n} \sum_{i=1}^{n} (y_i^F - f(\mathbf{x}_i, \hat{\boldsymbol{\theta}}_z) - \hat{\delta}_z(\mathbf{x}_i)) \frac{\partial f^M(\mathbf{x}_i, \hat{\boldsymbol{\theta}}_z)}{\partial \theta_j}.$$
 (S31)

Choosing $g(\cdot) = \frac{\partial f^M(\cdot, \hat{\theta}_z)}{\partial \theta_j}$ and plugging (S31) into (S30), one has

$$\sqrt{\lambda} \left\langle \hat{\delta}_{z}(\cdot), \frac{\partial f^{M}(\cdot, \hat{\boldsymbol{\theta}}_{z})}{\partial \theta_{j}} \right\rangle_{\mathcal{H}} + \left\langle \hat{\delta}_{z}(\cdot), \frac{\partial f^{M}(\cdot, \hat{\boldsymbol{\theta}}_{z})}{\partial \theta_{j}} \right\rangle_{L_{2}(\mathcal{X})} = 0.$$
(S32)

Substituting (S31) into (S32) and by Lemma S4 (ii), we have

$$\begin{split} 0 &= -\frac{1}{n} \sum_{i=1}^{n} (y_{i}^{F} - f(\mathbf{x}_{i}, \hat{\boldsymbol{\theta}}_{z}) - \hat{\delta}_{z}(\mathbf{x}_{i})) \frac{\partial f^{M}(\mathbf{x}_{i}, \hat{\boldsymbol{\theta}}_{z})}{\partial \theta_{j}} \\ &= -\int (y^{R}(\mathbf{x}) - f^{M}(\mathbf{x}, \hat{\boldsymbol{\theta}}_{z})) \frac{\partial f^{M}(\mathbf{x}, \hat{\boldsymbol{\theta}}_{z})}{\theta_{j}} d\mathbf{x} + \left\langle \hat{\delta}_{z}(\cdot), \frac{\partial f^{M}(\cdot, \hat{\boldsymbol{\theta}}_{z})}{\partial \theta_{j}} \right\rangle_{L_{2}(\mathcal{X})} \\ &- \frac{1}{n} \sum_{i=1}^{n} \epsilon_{i} \frac{\partial f^{M}(\mathbf{x}_{i}, \hat{\boldsymbol{\theta}}_{z})}{\partial \theta_{j}} + o_{p}(n^{-1/2}) \\ &= \int \frac{\partial (y^{R}(\mathbf{x}) - f^{M}(\mathbf{x}, \hat{\boldsymbol{\theta}}_{z}))^{2}}{\partial \theta_{j}} d\mathbf{x} - \sqrt{\lambda} \left\langle \hat{\delta}_{z}(\cdot), \frac{\partial f^{M}(\cdot, \hat{\boldsymbol{\theta}}_{z})}{\partial \theta_{j}} \right\rangle_{\mathcal{H}} \\ &- \frac{1}{n} \sum_{i=1}^{n} \epsilon_{i} \frac{\partial f^{M}(\mathbf{x}_{i}, \hat{\boldsymbol{\theta}}_{z})}{\theta_{j}} + o_{p}(n^{-1/2}). \end{split}$$

Applying Taylor expansion to the first term on the right-hand side at θ_{L_2} , for any j = 1, ..., q, we obtain

$$\left\{ \int \frac{\partial^2 (y^R(\mathbf{x}) - f^M(\mathbf{x}, \tilde{\boldsymbol{\theta}}_z))^2}{\partial \theta_j \partial \boldsymbol{\theta}} d\mathbf{x} \right\}^T (\hat{\boldsymbol{\theta}}_z - \boldsymbol{\theta}_{L_2})
= \left\{ \int \frac{\partial^2 (y^R(\mathbf{x}) - f^M(\mathbf{x}, \boldsymbol{\theta}_{L_2}))^2}{\partial \theta_j \partial \boldsymbol{\theta}} d\mathbf{x} + o_p(1) \right\}^T (\hat{\boldsymbol{\theta}}_z - \boldsymbol{\theta}_{L_2})
= \sqrt{\lambda} \left\langle \hat{\delta}_z(\cdot), \frac{\partial f^M(\cdot, \hat{\boldsymbol{\theta}}_z)}{\partial \theta_j} \right\rangle_{\mathcal{H}} + \frac{1}{n} \sum_{i=1}^n \epsilon_i \frac{\partial f^M(\mathbf{x}_i, \hat{\boldsymbol{\theta}}_z)}{\partial \theta_j} + o_p(n^{-1/2}), \quad (S33)$$

where $\tilde{\theta}_z$ lies within the *q* dimensional rectangle between θ_{L_2} and $\hat{\theta}_z$. Observe that Corollary 5 and assumption A3 imply

$$\left| \left\langle \hat{\delta}_z, \frac{\partial f^M(\cdot, \hat{\boldsymbol{\theta}}_z)}{\partial \theta_j} \right\rangle_{\mathcal{H}} \right| \leq \| \hat{\delta}_z \|_{\mathcal{H}} \left\| \frac{\partial f^M(\cdot, \boldsymbol{\theta}_{L_2})}{\partial \theta_j} + o_p(1) \right\|_{\mathcal{H}} = O_p(1).$$

Now we consider the second term. Define the empirical process

$$G_n(\boldsymbol{\theta}) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left[\epsilon_i \frac{\partial f^M(\mathbf{x}_i, \boldsymbol{\theta})}{\partial \theta_j} - \epsilon_i \frac{\partial f^M(\mathbf{x}_i, \boldsymbol{\theta}_{L_2})}{\partial \theta_j} \right]$$

and denote

$$f_{\boldsymbol{\theta}}(\boldsymbol{\epsilon}, \mathbf{x}) = \boldsymbol{\epsilon} \frac{\partial f^{M}(\mathbf{x}, \boldsymbol{\theta})}{\partial \theta_{j}} - \boldsymbol{\epsilon} \frac{\partial f^{M}(\mathbf{x}, \boldsymbol{\theta}_{L_{2}})}{\partial \theta_{j}}$$

Since

$$\mathbb{E}_{\epsilon i lon, \mathbf{x}} \left\{ [f_{\boldsymbol{\theta}_1}(\epsilon, \mathbf{x}) - f_{\boldsymbol{\theta}_2}(\epsilon, \mathbf{x})]^2 \right\} = \mathbb{E}_{\epsilon i lon, \mathbf{x}} \left[\epsilon^2 \left(\frac{\partial f^M(\mathbf{x}, \boldsymbol{\theta}_1)}{\partial \theta_j} - \frac{\partial f^M(\mathbf{x}, \boldsymbol{\theta}_2)}{\partial \theta_j} \right)^2 \right] \\ = \sigma_0^2 \left\| \frac{\partial f^M(\mathbf{x}, \boldsymbol{\theta}_1)}{\partial \theta_j} - \frac{\partial f^M(\mathbf{x}, \boldsymbol{\theta}_2)}{\partial \theta_j} \right\|_{L_2(\mathcal{X})}^2,$$

therefore the function class $\{f_{\theta}(\epsilon, \mathbf{x}) \in C(\mathbb{R} \times \mathcal{X}) : \theta \in \Theta\}$ is Donsker by Assumption A3, and hence, $G_n(\theta)$ converges weakly to a tight Gaussian stochastic process, denoted by $G(\cdot)$. W.l.o.g., we may take $G(\cdot)$ a version that has uniformly continuous sample paths (see Chapter 6 in Van de Geer (2000)). Since $G_n(\boldsymbol{\theta}_{L_2}) = 0$ for all n, it follows that $G(\boldsymbol{\theta}_{L_2}) = 0$. By the consistency of $\hat{\boldsymbol{\theta}}_z$ and the continuous mapping theorem (Van der Vaart, 2000), $G_n(\hat{\boldsymbol{\theta}}_z) = G(\boldsymbol{\theta}_{L_2}) + o_p(1) = o_p(1)$. Therefore,

$$\frac{1}{n}\sum_{i=1}^{n}\epsilon_{i}\frac{\partial f^{M}(\mathbf{x}_{i},\hat{\boldsymbol{\theta}}_{z})}{\partial\theta_{j}} = \frac{1}{\sqrt{n}}G_{n}(\hat{\boldsymbol{\theta}}_{z}) + \frac{1}{n}\sum_{i=1}^{n}\epsilon_{i}\frac{\partial f^{M}(\mathbf{x}_{i},\boldsymbol{\theta}_{L_{2}})}{\partial\theta_{j}} = O_{p}(n^{-1/2})$$

To sum up,

$$\begin{split} &\left\{\int \frac{\partial^2 (y^R(\mathbf{x}) - f^M(\mathbf{x}, \boldsymbol{\theta}_{L_2}))^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} d\mathbf{x} + o_p(1)\right\} (\hat{\boldsymbol{\theta}}_z - \boldsymbol{\theta}_{L_2}) \\ &= O_p(n^{-m/(2m+p)}) + O_p(n^{-1/2}) + o_p(n^{-1/2}) = O_p(n^{-m/(2m+p)}), \end{split}$$

completing the proof.

S5 Proof and additional results for Section 4

The identities in the Lemma S5 are used repeatedly in the proof of the Theorem 3 and Lemma 4.

Lemma S5. Denote $\sigma^2 \mathbf{R}_{z_d}$ the covariance matrix of $(\delta_{z_d}(\mathbf{x}_1), ..., \delta_{z_d}(\mathbf{x}_n))^T$, where the (i, j)entry being $\sigma^2 K_{z_d}(\mathbf{x}_i, \mathbf{x}_j)$ defined in (17). Denote $r_{z_d}(\mathbf{x}) = (K_{z_d}(\mathbf{x}, \mathbf{x}_1), ..., K_{z_d}(\mathbf{x}, \mathbf{x}_n))^T$ for any $\mathbf{x} \in \mathcal{X}$. One has the following identities

$$\mathbf{R}_{z_d}^{-1} = \mathbf{R}^{-1} + \frac{\lambda_z}{n} \mathbf{I}_n,\tag{S34}$$

$$\mathbf{r}_{z_d}^T(\mathbf{x}) = \frac{n}{\lambda_z} \mathbf{r}^T(\mathbf{x}) \tilde{\mathbf{R}}^{-1} = \mathbf{r}^T(\mathbf{x}) \mathbf{R}^{-1} \mathbf{R}_{z_d},$$
(S35)

for any $\mathbf{x} \in \mathcal{X}$.

Proof. By the definitions of \mathbf{R}_{z_d} and the Woodbury Identity, one has

$$\mathbf{R}_{z_d} = \mathbf{R} - \mathbf{R}\tilde{\mathbf{R}}^{-1}\mathbf{R} = \mathbf{R}\left(\mathbf{I}_n - \left(\mathbf{I}_n + \frac{\lambda_z}{n}\mathbf{R}^{-1}\right)^{-1}\right)$$
$$= \mathbf{R}\left(\frac{\lambda_z}{n}\mathbf{R} + \mathbf{I}_n\right)^{-1} = \left(\mathbf{R}^{-1} + \frac{\lambda_z}{n}\mathbf{I}_n\right)^{-1},$$

from which (S34) follows.

Equation (S35) can be shown similarly by noting $\mathbf{r}_{z_d}^T(\mathbf{x}) = \mathbf{r}^T(\mathbf{x}) - \mathbf{r}^T(\mathbf{x})\mathbf{\tilde{R}}^{-1}\mathbf{R}$ and the Woodbury Identity.

Proof of Theorem 3. The predictive mean is as follows

$$\begin{split} \hat{\mu}_{z}(\mathbf{x}) &= \mathbb{E}[y^{F}(\mathbf{x}) \mid \mathbf{y}^{F}, \boldsymbol{\theta}, \sigma_{0}^{2}, \lambda, \lambda_{z}] \\ &= f^{M}(\mathbf{x}, \boldsymbol{\theta}) + \mathbf{r}_{z_{d}}(\mathbf{x})^{T} \left(\mathbf{R}_{z_{d}} + n\lambda \mathbf{I}_{n}\right)^{-1} \left(\mathbf{y}^{F} - \mathbf{f}_{\boldsymbol{\theta}}^{M}\right) \\ &= f^{M}(\mathbf{x}, \boldsymbol{\theta}) + \mathbf{r}(\mathbf{x})^{T} \mathbf{R}^{-1} \mathbf{R}_{z_{d}} \left(\mathbf{R}_{z_{d}} + n\lambda \mathbf{I}_{n}\right)^{-1} \left(\mathbf{y}^{F} - \mathbf{f}_{\boldsymbol{\theta}}^{M}\right) \\ &= f^{M}(\mathbf{x}, \boldsymbol{\theta}) + \mathbf{r}(\mathbf{x})^{T} \mathbf{R}^{-1} \left(\mathbf{I}_{n} + n\lambda \left(\mathbf{R}^{-1} + \frac{\lambda_{z}}{n} \mathbf{I}_{n}\right)^{-1}\right) \left(\mathbf{y}^{F} - \mathbf{f}_{\boldsymbol{\theta}}^{M}\right) \\ &= f^{M}(\mathbf{x}, \boldsymbol{\theta}) + \frac{\mathbf{r}(\mathbf{x})^{T}}{\left(1 + \lambda\lambda_{z}\right)} \left(\mathbf{R}^{-1} + \frac{n\lambda}{1 + \lambda\lambda_{z}} \mathbf{I}_{n}\right)^{-1} \left(\mathbf{y}^{F} - \mathbf{f}_{\boldsymbol{\theta}}^{M}\right), \end{split}$$

where the last two equalities follow from (S35) and (S34), respectively.

The predictive variance can be obtained using (S35) and (S34) as follows

$$\begin{aligned} K_z^*(\mathbf{x}, \mathbf{x}) &= K_{z_d}(\mathbf{x}, \mathbf{x}) - \mathbf{r}_{z_d}^T(\mathbf{x}) \left(\mathbf{R}_{z_d} + n\lambda \mathbf{I}_n\right)^{-1} \mathbf{r}_{z_d}(\mathbf{x}) \\ &= K(\mathbf{x}, \mathbf{x}) - \mathbf{r}^T(\mathbf{x}) \mathbf{\tilde{R}}^{-1} \mathbf{r}(\mathbf{x}) \\ &- (1 + \lambda\lambda_z)^{-1} \mathbf{r}(\mathbf{x})^T \left(\mathbf{R}^{-1} + \frac{n\lambda}{1 + \lambda\lambda_z} \mathbf{I}_n\right)^{-1} \frac{n}{\lambda_z} \mathbf{\tilde{R}}^{-1} \mathbf{r}(\mathbf{x}) \end{aligned}$$

from which the result follows.

L	_	_	_	

Proof of Lemma 4. When $\sigma_0^2 = 0$, the predictive mean is as follows

$$\begin{split} \mathbb{E}[y^{F}(\mathbf{x}) \mid \mathbf{y}^{F}, \boldsymbol{\theta}, \sigma_{0}^{2}, \lambda, \lambda_{z}] = & f^{M}(\mathbf{x}, \boldsymbol{\theta}) + \mathbf{r}_{z_{d}}(\mathbf{x})^{T} \mathbf{R}_{z_{d}}^{-1}(\mathbf{y}^{F} - \mathbf{f}_{\boldsymbol{\theta}}^{M}) \\ = & f^{M}(\mathbf{x}, \boldsymbol{\theta}) + \mathbf{r}(\mathbf{x})^{T} \mathbf{R}^{-1} \mathbf{R}_{z_{d}} \mathbf{R}_{z_{d}}^{-1}(\mathbf{y}^{F} - \mathbf{f}_{\boldsymbol{\theta}}^{M}) \\ = & f^{M}(\mathbf{x}, \boldsymbol{\theta}) + \mathbf{r}(\mathbf{x})^{T} \mathbf{R}^{-1}(\mathbf{y}^{F} - \mathbf{f}_{\boldsymbol{\theta}}^{M}). \end{split}$$

The predictive variance can be obtained similarly.