ON MESHFREE NUMERICAL DIFFERENTIATION

LEEVAN LING* AND QI YE†

Abstract. We combine techniques in meshfree methods and Gaussian process regressions to construct kernel-based estimators for numerical derivatives from noisy data. Specially, we construct meshfree estimators from normal random variables, which are defined by kernel-based probability measures induced from symmetric positive definite kernels, to reconstruct the unknown partial derivatives from scattered noisy data. Our developed theories give rise to Tikhonov regularization methods with a priori parameter, but the shape parameters of the kernels remain tunable. For that, we propose an error measure that is computable without the exact values of the derivative. This allows users to obtain a quasi-optimal kernel-based estimator by comparing the approximation quality of kernel-based estimators. Numerical examples in two-dimensions and three-dimensions are included to demonstrate the convergence behaviour and effectiveness of the proposed numerical differentiation scheme.

1. Introduction. Numerical differentiation aims to reconstruct partial derivatives of a function from its discrete values. Its applications can be found in many branches of science and engineering. Image processing [15, 22], mechanical systems [13], solving integral equations [4, 10] are a few examples besides of many other applications in scientific computing. Differentiation is a typical ill-posed process in the sense that small errors in the data will be greatly amplified. However, the presence of noise in the data is unavoidable in many applications. Numerical methods for stable numerical differentiation, which could be derived from the finite difference method [14, 18, 23], wavelet regularization [1, 9] and etc., can be found in literature.

In this paper, we are interested in the meshfree kernel-based approach that allows an easy treatment to scattered noisy data in higher dimensions. Existing algorithms are mostly due to the work of Wei and collaborators that are based on cubic spline [24], radial spline [27], thin plate spline [26], and a class of radial basis functions [25] including the Gaussian, multiquadrics, radial splines, thin-plate splines, and Matérn functions. In the following, we adopt a new approach based on the close connections between meshfree methods and Gaussian process regressions, that is recently reported in [20] and [29–31]. Materials discussed here are different from the classical meshfree techniques (i.e., the ones used by Wei *et al.*). By beginning with a simple example in Section 2, we help readers understand the fundamental ideas which are needed for this work: *meshfree interpolation and Gaussian process regression are identical.* That is the construction of kernel-based estimators and the related error analysis can be done in the context of normal random variables and probability measures. In Section 3, we generalize this equivalence via some kernel-based probability measures $\mathbb{P}_K$ induced from any sufficiently smooth symmetric positive definite (SPD) kernels $K : \Omega \times \Omega \to \mathbb{R}$ to deal with the scientific computing of numerical differentiation:

> *Let $\Omega \subset \mathbb{R}^d$ be some bounded and regular domain and $X \subset \Omega$ be a discrete set of data centers. For any smooth function $f \in C^m(\Omega)$, we compute the α partial derivative of f from noisy data $\left(X, \mathbf{f}^\delta\right) \to D^\alpha f(\mathbf{x})$, for any $\mathbf{x} \in \Omega$, with $\alpha \in \mathbb{N}_0^d$, $|\alpha| \leq m$, $\mathbf{f} := f(X)$, and $\|\mathbf{f}^\delta - \mathbf{f}\|_\infty \leq \delta$.*

By making a detour to stochastic theories, we derive the Tikhonov-regularized linear systems for identifying estimators for noisy data in Section 3. *A priori* regu-

*Department of Mathematics, Hong Kong Baptist University, Hong Kong (lling@hkbu.edu.hk).
†School of Mathematical Sciences, Laboratory for Machine Learning and Computational Optimization, South China Normal University, Guangzhou, Guangdong, China (yeqi@m.scnu.edu.cn).

larization parameter within can be determined by the noise level of the data and the degree of regularization can still be adjusted by the shape parameter of the employed kernel. In other words, there is only one (shape) parameter to tune in the proposed method, instead of two (shape and regularization) that is typically seen in regularized meshfree approaches for inverse problems. In addition, the variance provide us to obtain (function independent and pointwise) error estimates, which mimics the power function in meshfree theories. In Section 4, we propose an error measure that analogizes the pointwise error estimates in meshfree interpolation. Numerical examples are included to demonstrate the effectiveness of the proposed strategy for finding quasi-optimal kernels.

2. Deterministic and Stochastic Interpolation. The theoretical connections between meshfree methods and Gaussian process regressions for the classical interpolations were recently recognized [7, 20] and explored [29–31]. By seeing from two different viewpoints, kernel-based approximation methods can be analyzed by theories in approximation theories or probability theories. Here, we use such unified theoretical structure to study both the local geometric features of kernel-based estimators and derive simple formulation for numerical differentiation in the latter sections. We begin with a simple example to help readers get deeper insights into the topics.

Consider the standard interpolation problem. Suppose that we have an ordered set of observed data values $f_1, \dots, f_n \in \mathbb{R}$ given at an order set of distinct data points $\mathbf{x}_1, \dots, \mathbf{x}_n \in \Omega \subseteq \mathbb{R}^d$. Let $\mathbf{f} := (f_1, \cdots, f_n)^T$ and $X := \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$. We aim to estimate the "unknown value" at some other locations $\mathbf{x} \in \Omega \backslash X$ based on the observed data at X; see Figure 2.1 (a)for a schematic demonstration in $\Omega = [0, 1]$ with $n = 7$ data.

From the deterministic point of view, the observed data is viewed as the evaluation of an unknown function $f : \Omega \to \mathbb{R}$, which generates the observed data by

$$f(\mathbf{x}_1) = f_1, \dots, f(\mathbf{x}_n) = f_n \quad \text{or} \quad f(X) = \mathbf{f},$$

if we use a more compact vector notation. To obtain a meshfree interpolant s_{mf} to the data $(X, \mathbf{f})$, we can employ a symmetric positive definite (SPD) kernel $K : \Omega \times \Omega \to \mathbb{R}$. Then, the meshfree interpolant s_{mf} is given as a linear combination of the basis functions $K(\cdot, \mathbf{x}_1), \dots, K(\cdot, \mathbf{x}_n)$ in the form of

$$s_{mf}(\mathbf{x}) := c_1 K(\mathbf{x}, \mathbf{x}_1) + \cdots + c_n K(\mathbf{x}, \mathbf{x}_n), \tag{2.1}$$

with the coefficients $\mathbf{c} := (c_1, \cdots, c_n)^T$ uniquely determined by the linear system

$$\begin{pmatrix} K(\mathbf{x}_1, \mathbf{x}_1) & \cdots & K(\mathbf{x}_1, \mathbf{x}_n) \\ & \ddots & \\ K(\mathbf{x}_n, \mathbf{x}_1) & \cdots & K(\mathbf{x}_n, \mathbf{x}_n) \end{pmatrix} \begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix} = \begin{pmatrix} f_1 \\ \vdots \\ f_n \end{pmatrix} \quad \text{or} \quad K(X, X)\, \mathbf{c} = \mathbf{f}.$$

See [2, 6, 28] for more detailed discussions. Since the interpolation matrix $K(X, X)$ is SPD, the interpolant s_{mf} satisfies all interpolation conditions $s_{mf}(\mathbf{x}_1) = f_1, \dots, s_{mf}(\mathbf{x}_n) = f_n$. In approximation theory, we use $s_{mf}(\mathbf{x})$ to approximate the unknown value $f(\mathbf{x})$ for some $\mathbf{x} \in \Omega \backslash X$ and study its convergence behaviour. As a demonstration, Figure 2.1 (b) shows a meshfree interpolant s_{mf} using the Gaussian kernel $K^\theta(\mathbf{x}, \mathbf{y}) := \exp(-\theta^2 \|\mathbf{x} - \mathbf{y}\|_2^2)$.

In the context of Gaussian process regression, we compute the estimator $s_{kg}(\mathbf{x})$ of the realization of the normal random variable $S_{\mathbf{x}}$ conditioned on the observations

$$S_{\mathbf{x}_1} = f_1, \ldots, S_{\mathbf{x}_n} = f_n \quad \text{or} \quad \mathbf{S}_X = \mathbf{f}.$$

We aim to identify the unknown value of the realization of S at some other location $\mathbf{x} \in \Omega \backslash X$. In the Gaussian process regression [21], a.k.a the simple kriging method, S is a Gaussian process with a mean 0 and covariance kernel K. We usually assume that such covariance kernel K is SPD. The simple kriging method provides the best linear unbiased estimator $\hat{S}_{\mathbf{x}}$ of $S_{\mathbf{x}}$ in the form of

$$\hat{S}_{\mathbf{x}} := w_1(\mathbf{x}) S_{\mathbf{x}_1} + \cdots + w_n(\mathbf{x}) S_{\mathbf{x}_n},$$

where the weighting $\mathbf{w}(\mathbf{x}) := (w_1(\mathbf{x}), \ldots, w_n(\mathbf{x}))^T$ is uniquely determined by the linear system

$$\begin{pmatrix} K(\mathbf{x}_1, \mathbf{x}_1) & \cdots & K(\mathbf{x}_1, \mathbf{x}_n) \\ & \ddots & \\ K(\mathbf{x}_n, \mathbf{x}_1) & \cdots & K(\mathbf{x}_n, \mathbf{x}_n) \end{pmatrix} \begin{pmatrix} w_1(\mathbf{x}) \\ \vdots \\ w_n(\mathbf{x}) \end{pmatrix} = \begin{pmatrix} K(\mathbf{x}, \mathbf{x}_1) \\ \vdots \\ K(\mathbf{x}, \mathbf{x}_n) \end{pmatrix},$$

or in matrix form $K(X, X)\ \mathbf{w}(\mathbf{x}) = K(\mathbf{x}, X)^T$. Thus, the estimator $s_{kg}(\mathbf{x})$ can be written as

$$s_{kg}(\mathbf{x}) := w_1(\mathbf{x}) f_1 + \cdots + w_n(\mathbf{x}) f_n. \tag{2.2}$$

Obviously, we also have all interpolation conditions $s_{kg}(\mathbf{x}_1) = f_1, \ldots, s_{kg}(\mathbf{x}_n) = f_n$. The unknown value is then approximated by $s_{kg}(\mathbf{x})$. Figure 2.1 (c) shows the unbiased prediction $s_{kg}(\mathbf{x})$ and the associated 99% confident intervals obtained by the Gaussian process S with the covariance kernel K^θ same as in Figure 2.1 (b).

If the meshfree kernel and the covariance kernel coincide, then (2.1) and (2.2) suggest that

$$s_{mf}(\mathbf{x}) = K(\mathbf{x}, X)^T K(X, X)^{-1} \mathbf{f} = (K(X, X)^{-1} K(\mathbf{x}, X))^T \mathbf{f} = s_{kg}(\mathbf{x}).$$

It means that the meshfree estimator and the unbiased estimator shown respectively in Figure 2.1 (b) and (c) are indeed identical. This equivalence was first observed in [20]. So, we can recall s_{mf} and s_{kg} as s_X. Moreover, by [28, Theorem 13.2], the meshfree interpolant s_{mf} is the unique minimizer in the reproducing kernel Hilbert space $\mathcal{H}_K(\Omega)$ of K, a.k.a. the native space of K, with respect to the associated native space norm so that

$$s_{mf} = \underset{f \in \mathcal{H}_K(\Omega)}{\operatorname{argmin}} \|f\|_{\mathcal{H}_K(\Omega)} \quad \text{subject to} \quad f(X) = \mathbf{f}.$$

By the method of Bayesian estimation, we have that

$$\hat{S}_{\mathbf{x}} = \underset{U \in \operatorname{span}\{\mathbf{S}_X\}}{\operatorname{argmin}} \mathrm{E} \left| S_{\mathbf{x}} - U \right|^2, \tag{2.3}$$

and the minimization of mean squared errors can be written as

$$\sigma_X(\mathbf{x})^2 := \mathrm{E}|S_{\mathbf{x}} - \hat{S}_{\mathbf{x}}|^2 = K(\mathbf{x}, \mathbf{x}) - K(\mathbf{x}, X)^T K(X, X)^{-1} K(\mathbf{x}, X). \tag{2.4}$$

Here, the linear span $\operatorname{span}\{\mathbf{S}_X\} := \operatorname{span}\{S_{\mathbf{x}_1}, \ldots, S_{\mathbf{x}_n}\}$.

In [29–31], we continue to explore the connections between meshfree methods and Gaussian process regressions. Suppose that Ω is a regular compact domain and K is any $2m$ continuously differentiable SPD kernel. By [29, Theorem 6.1] and [30, Theorem 1], there exists a kernel-based probability measure $\mathbb{P}_K$ defined on the Sobolev space $\mathcal{H}^m(\Omega)$ of the order $m > d/2$ such that a Gaussian process S with the mean 0 and the covariance kernel K is constructed by

$$S_{\mathbf{x}}(\omega) = \omega(\mathbf{x}) \quad \text{for all} \quad \omega \in \mathcal{H}^m(\Omega), \tag{2.5}$$

see [31, Definition 2.2] and [30, Definition 2] there for details. In other words, for any sufficiently smooth kernel K, we have a probability space consisting the Sobolev space $\mathcal{H}^m(\Omega)$, its Borel σ-algebra as its sample space and set of events. The likelihood of happening is given by the kernel-based probability measure $\mathbb{P}_K$ as in probability theories. Based on (2.5), any function $\omega \in \mathcal{H}^m(\Omega)$ can be seen as a sample path, a.k.a. a trajectory, of the Gaussian process S. This shows that $f(\mathbf{x})$ can be viewed as the realization of $S_{\mathbf{x}}$ conditioned on $\mathbf{S}_X = \mathbf{f}$. Since S is Gaussian, we also have that $\mathrm{E}(S_{\mathbf{x}}|\mathbf{S}_X = \mathbf{f}) = \mathrm{E}(\hat{S}_{\mathbf{x}}|\mathbf{S}_X = \mathbf{f})$. This means that $s_{kg}(\mathbf{x})$ is the $\mathbb{P}_K$–average of $S_{\mathbf{x}}$ given $\mathbf{S}_X = \mathbf{f}$ and s_{kg} can be viewed as the "center" with respect to the kernel-based probability measure $\mathbb{P}_K$ of the collection of all sample paths in $\mathcal{H}^m(\Omega)$ satisfying the interpolation conditions.

At this point, we have a compatible connection between the interpolation conditions and the realizations of the Gaussian process. Let the map $\Gamma_K(V) := \int_{\mathcal{B}} \omega V(\omega)\mathbb{P}_K(\omega)$ for $V \in \mathcal{H}_S$, where $\mathcal{H}_S$ is the completion of the linear span of the Gaussian process S by the finite second moments (see [31, Section 3.3]). This shows that Γ_K is a map from $\mathcal{H}_S$ into $\mathcal{H}^m(\Omega)$. By [31, Theorem 3.15], the error of $|f(\mathbf{x}) - s_{kg}(\mathbf{x})|$ can be bounded by $\sigma_X(\mathbf{x})$ if $f \in \mathrm{range}(\Gamma_K)$. Here is another connection, such as $\sigma_X(\mathbf{x})$ is known as the power function in meshfree method [28, Section 11.1], which gives an upper bound for the error $|f(\mathbf{x}) - s_{mf}(\mathbf{x})|$ when $f \in \mathcal{H}_K(\Omega)$ and $f(X) = \mathbf{f}$. Moreover, [31, Theorem 3.10] assures that $s_{kg}(\mathbf{x})$ is also convergent to $f(\mathbf{x})$ even if $f \in \mathcal{H}^m(\Omega)\backslash\mathrm{range}(\Gamma_K)$. In general, the unknown function $f \in \mathcal{H}^m(\Omega)$ does not guarantee $f \in \mathcal{H}_K(\Omega)$ unless K reproduces some Sobolev spaces. In such cases, convergence analysis can be carried out by the structure of scattered zeros [17].

3. Numerical Differentiation by Kernel-based Approximation Methods. We generalize the ideas in the previous section to the numerical differentiation for exact data in Section 3.1 and extend the theories to noisy data in Section 3.2. To begin, let us overview the notations used and assumptions needed for the ease of reading. We suppose that $\Omega \subset \mathbb{R}^d$ is regular and compact, such as a Lipschitz domain and $K \in \mathrm{C}^{2m,1}(\Omega \times \Omega)$ with $m > d/2$ is a SPD kernel. We denote the exact and noisy observed data at $X := \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subset \Omega$ respectively by

$$\mathbf{f} := (f_1, \cdots, f_n)^T \in \mathbb{R}^n \quad \text{or} \quad \mathbf{f}^\delta := \left(f_1^\delta, \ldots, f_n^\delta\right)^T \in \mathbb{R}^n.$$

The exact data $(X, \mathbf{f})$ consists evaluations of some unknown function $f \in \mathcal{H}^m(\Omega)$ at X, i.e., $\mathbf{f} = f(X)$. We reserve the subscript n to denote the number of observed data throughout the paper and omit it for simplicity unless confusion may arise. For any noise level $\delta > 0$, we assume that the noise is additive and

$$\|\mathbf{f} - \mathbf{f}^\delta\|_\infty = \max_{k=1,\ldots,n} |f_k - f_k^\delta| \le \delta. \tag{3.1}$$

It is straightforward to modify the theorems to come to work on multiplicative noise by replacing δ by $\delta\|\mathbf{f}\|_\infty$ provided $\|\mathbf{f}\|_\infty$ is bounded away from zero.

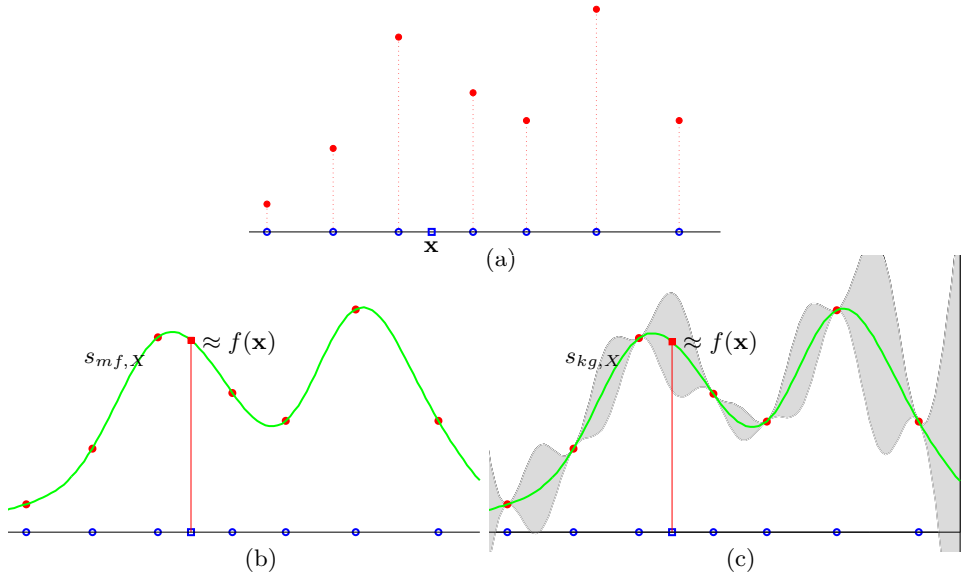

FIG. 2.1. *Example 1: Schematic demonstration for deterministic interpolation and Gaussian process regression. (a) Observed data values* $\mathbf{f}$ *(red) at the given data points* X *(blue). (b) Meshfree interpolant, (c) Gaussian process regression and the normally distributed* 99% *confidence intervals obtained by the Gaussian kernel* K^θ *with the shape parameter* $\theta = 6$.

Denote $\mathcal{H}^m(\Omega)$ to be the standard L_2–based Sobolev space of the order m. For any multi-index $\alpha = (\alpha_1, \cdots, \alpha_d) \in \mathbb{N}_0^d$ of nonnegative integers with $|\alpha| < m - d/2$, we denote the partial differentiate operator of the order α by

$$D^\alpha := \frac{\partial^{|\alpha|}}{\partial^{\alpha_1} \cdots \partial^{\alpha_d}}.$$

We use the notations D_1^α and D_2^α to indicate that D^α acts on the first argument $\mathbf{x}$ and second argument $\mathbf{y}$ of a kernel $K(\mathbf{x}, \mathbf{y})$, respectively.

By the Sobolev imbedding theorem, the linear functional $\delta_\mathbf{x} \circ D^\alpha$ is continuous on $\mathcal{H}^m(\Omega)$ whenever $|\alpha| < m - d/2$. Using the construction technique in [29–31], we can obtain the kernel-based probability measure $\mathbb{P}_K$ on the Sobolev space $\mathcal{H}^m(\Omega)$ and get a framework similar to the one shown in Section 2. To be precise, there exists a kernel-based probability measure $\mathbb{P}_K$ defined on the Sobolev space $\mathcal{H}^m(\Omega)$ such that the Gaussian process S^α equipped with the mean 0 and the covariance kernel $D_1^\alpha D_2^\alpha K$ has the representation

$$S_\mathbf{x}^\alpha(\omega) := D^\alpha \omega(\mathbf{x}), \quad \text{for } \omega \in \mathcal{H}^m(\Omega). \tag{3.2}$$

More details of the proofs of the above constrictions can be found in [29, Theorem 6.1] and [30, Theorem 1].

We denote a multiple normal random vector composed of $S_{\mathbf{x}_1}, \ldots, S_{\mathbf{x}_n}$ as

$$\mathbf{S}_X := (S_{\mathbf{x}_1}, \cdots, S_{\mathbf{x}_n})^T.$$

Clearly $S^\mathbf{0} = S$ and equation (3.2) is a generalization of (2.5). For the ease of computing the means, we want to have a matrix formula similar to the one in (2.4)

and need a well-defined vector function

$$\mathbf{k}_X^\alpha(\mathbf{x}) := D^\alpha K(\mathbf{x}, X)^T = (D_1^\alpha K(\mathbf{x}, \mathbf{x}_1), \cdots, D_1^\alpha K(\mathbf{x}, \mathbf{x}_n))^T \in \mathbb{R}^n, \tag{3.3}$$

and the covariance matrix of the normal random variables $(S_{\mathbf{x}}^\alpha, \mathbf{S}_X)$ given as

$$\mathsf{A}_X^\alpha(\mathbf{x}) := \begin{pmatrix} D_1^\alpha D_2^\alpha K(\mathbf{x},\mathbf{x}) & D_2^\alpha K(\mathbf{x}_1,\mathbf{x}) & \cdots & D_2^\alpha K(\mathbf{x}_n,\mathbf{x}) \\ D_1^\alpha K(\mathbf{x},\mathbf{x}_1) & K(\mathbf{x}_1,\mathbf{x}_1) & \cdots & K(\mathbf{x}_1,\mathbf{x}_n) \\ \vdots & & \ddots & \\ D_1^\alpha K(\mathbf{x},\mathbf{x}_n) & K(\mathbf{x}_n,\mathbf{x}_1) & \cdots & K(\mathbf{x}_n,\mathbf{x}_n) \end{pmatrix} \in \mathbb{R}^{(n+1)\times(n+1)}. \tag{3.4}$$

For convenience, we extend definitions (3.3) and (3.4) to $\alpha = \mathbf{0}$ by $\mathbf{k}_X(\mathbf{x}) := K(\mathbf{x}, X)^T$ and $\mathsf{A}_X := K(X, X) \in \mathbb{R}^{n\times n}$ is the covariance matrix of $\mathbf{S}_X$. Clearly $\mathrm{E}(S_{\mathbf{x}}^\alpha \mathbf{S}_X) = \mathbf{k}_X^\alpha(\mathbf{x}) = D^\alpha \mathbf{k}_X(\mathbf{x})$ and

$$\mathsf{A}_X^\alpha(\mathbf{x}) = \begin{pmatrix} D_1^\alpha D_2^\alpha K(\mathbf{x},\mathbf{x}) & \mathbf{k}_X^\alpha(\mathbf{x})^T \\ \mathbf{k}_X^\alpha(\mathbf{x}) & \mathsf{A}_X \end{pmatrix} \quad \text{for all} \quad 0 < |\alpha| \le m - \frac{d}{2}.$$

3.1. Numerical Derivatives of Exact Data. We now extend the technique of the kernel-based probability measure in Section 2 to estimate the partial derivatives of f from exact data $(X, \mathbf{f})$. The noise-free results in the section serve as points of reference to ensure that our proposed method for noisy data has the proper asymptotic behaviour as the noise level δ goes to zero.

Using the arguments in Section 2, we can relate the value of its partial derivative $D^\alpha f(\mathbf{x})$ at some $\mathbf{x} \in \Omega$ to the realization of a random variable $S_{\mathbf{x}}^\alpha$ conditioned on $\mathbf{S}_X = \mathbf{f}$. Since the meshfree and simple kriging estimators are identical, we can safely omit the subscripts mf and kg to simplify the notation from here onwards. Same as (2.3), we obtain the estimator $\hat{S}_{\mathbf{x}}^\alpha$ of $S_{\mathbf{x}}^\alpha$ by the mean squared errors so that

$$\hat{S}_{\mathbf{x}}^\alpha = \underset{U \in \mathrm{span}\{\mathbf{S}_X\}}{\operatorname{argmin}} \mathrm{E}\left|S_{\mathbf{x}}^\alpha - U\right|^2. \tag{3.5}$$

Thus, the estimator $s_X^\alpha(\mathbf{x})$ is constructed by the realization of $\hat{S}_{\mathbf{x}}^\alpha$ conditioned on $\mathbf{S}_X = \mathbf{f}$. It is expected that $s_X^\alpha(\mathbf{x})$ is related to some kernel-based functions; the following theorem asserts that this condition mean is simply the α derivative of the interpolant s_X in Section 2.

THEOREM 3.1. *The estimator $s_X^\alpha(\mathbf{x})$ is the D^α partial derivative of the kernel-based interpolant $s_X(\mathbf{x}) = \mathbf{k}_X(\mathbf{x})^T\mathbf{c}$, in which the coefficients $\mathbf{c} \in \mathbb{R}^n$ is uniquely determined by the linear system $\mathsf{A}_X\mathbf{c} = \mathbf{f}$.*

Proof. We know that the normal random variables $(S_{\mathbf{x}}^\alpha, \mathbf{S}_X)$ have the mean $\mathbf{0}$ and the covariance matrix $\mathsf{A}_X^\alpha(\mathbf{x})$.

For any $U \in \mathrm{span}\{S_{\mathbf{x}_k}\}_{k=1}^n$, there exists a $\mathbf{c} \in \mathbb{R}^n$ such that $U = \mathbf{c}^T\mathbf{S}_X$. Since $\mathrm{E}(S_{\mathbf{x}}^\alpha S_{\mathbf{x}}^\alpha) = D_1^\alpha D_2^\alpha K(\mathbf{x},\mathbf{x})$, $\mathrm{E}(S_{\mathbf{x}}^\alpha \mathbf{S}_X) = \mathbf{k}_X^\alpha(\mathbf{x})$, and $\mathrm{E}(\mathbf{S}_X\mathbf{S}_X^T) = \mathsf{A}_X$, we have that $\mathrm{E}\left|S_{\mathbf{x}}^\alpha - U\right|^2 = D_1^\alpha D_2^\alpha K(\mathbf{x},\mathbf{x}) - 2\mathbf{c}^T\mathbf{k}_X^\alpha(\mathbf{x}) + \mathbf{c}^T\mathsf{A}_X\mathbf{c}$. By using Lagrange multipliers and seeking the extremum of the Lagrangian, the minimizer $\hat{S}_{\mathbf{x}}^\alpha$ can be written as

$$\hat{S}_{\mathbf{x}}^\alpha = \mathbf{w}_X^\alpha(\mathbf{x})^T\mathbf{S}_X,$$

where

$$\mathbf{w}_X^\alpha(\mathbf{x}) = \mathsf{A}_X^{-1}\mathbf{k}_X^\alpha(\mathbf{x}).$$

Therefore, we have that

$$s_X^\alpha(\mathbf{x}) = \mathbf{w}_X^\alpha(\mathbf{x})^T\mathbf{f} = D^\alpha\left(\mathbf{k}_X(\mathbf{x})^T\mathsf{A}_X^{-1}\mathbf{f}\right) = D^\alpha s_X(\mathbf{x}).$$

The proof is completed. □

Within native spaces, approximation theory guarantees that $s_X^\alpha(\mathbf{x}) \to D^\alpha f(\mathbf{x})$ as the fill distance of X shrinks. The following theorem ensures the convergence of the kernel-based estimators to the function $f \in \mathcal{H}^m(\Omega)$.

THEOREM 3.2. *Suppose that the sequence of data points $X_1 \subseteq \cdots \subseteq X_n \subseteq \cdots$ is getting dense in Ω in the sense that their fill distance*[1] *$h_n \to 0$ as $n \to \infty$. Suppose further that the observed data values $\mathbf{f}_n := f(X_n)$ are evaluated by some function $f \in \mathcal{H}^m(\Omega)$ at X_n for all $n \in \mathbb{N}$, then the estimator $s_{X_n}^\alpha(\mathbf{x})$ converges pointwise to $D^\alpha f(\mathbf{x})$ as $n \to \infty$.*

Proof. The Sobolev space $\mathcal{H}^m(\Omega)$ can be imbedded into $\mathrm{C}(\Omega)$. Since the separable points $X_\infty := \bigcup_{n=1}^\infty X_n$ are dense in Ω, the compactness of Ω guarantees that for any $\omega \in \mathrm{C}(\Omega)$, we have that $\omega(X_\infty) = \mathbf{f}_\infty$ if and only if $\omega = f$. Let $\mathcal{A}_X(\mathbf{f}) := \{\omega \in \mathcal{H}^m(\Omega) : \omega(X) = \mathbf{f}\}$. Then the sequence of collections satisfies

$$\mathcal{A}_{X_1}(\mathbf{f}_1) \supseteq \cdots \supseteq \mathcal{A}_{X_n}(\mathbf{f}_n) \supseteq \cdots \supseteq \bigcap_{n=1}^\infty \mathcal{A}_{X_n}(\mathbf{f}_n) = \mathcal{A}_{X_\infty}(\mathbf{f}_\infty) = \{f\}.$$

Therefore, [31, Theorem 3.10] assures that the sequence of $s_{X_n}^\alpha(\mathbf{x})$, that can be seen as a sequence with respect to data $(X_n, \mathbf{f}_n)$, satisfies that

$$\lim_{n\to\infty} s_{X_n}^\alpha(\mathbf{x}) = \lim_{n\to\infty} \mathrm{E}\big(S_\mathbf{x}^\alpha \,\big|\, \mathbf{S}_{X_n} = \mathbf{f}_n\big) = D^\alpha f(\mathbf{x}).$$

□

If $f \in \mathrm{range}(\Gamma_K)$, then [31, Corollary 3.15] shows that we can analyze the square error $|D^\alpha f(\mathbf{x}) - s_X^\alpha(\mathbf{x})|^2$ by using the averages of $|S_\mathbf{x}^\alpha - s_X^\alpha(\mathbf{x})|^2$ conditioned on $\mathbf{S}_X = \mathbf{f}$ computed by the kernel-based probability measured $\mathbb{P}_K$, such as the variance

$$\sigma_X^\alpha(\mathbf{x})^2 := \mathrm{E}\left(|S_\mathbf{x}^\alpha - s_X^\alpha(\mathbf{x})|^2 \,\middle|\, \mathbf{S}_X = \mathbf{f}\right).$$

THEOREM 3.3. *The variance $\sigma_X^\alpha(\mathbf{x})^2$ has the form*

$$\sigma_X^\alpha(\mathbf{x})^2 = D_1^\alpha D_2^\alpha K(\mathbf{x},\mathbf{x}) - \mathbf{k}_X^\alpha(\mathbf{x})^T\mathsf{A}_X^{-1}\mathbf{k}_X^\alpha(\mathbf{x}). \tag{3.6}$$

Proof. Since $S_\mathbf{x}^\alpha$ and $\mathbf{S}_X$ have the multivariate normal distributions, we have that

$$\mathrm{E}\left(|S_\mathbf{x}^\alpha - s_X^\alpha(\mathbf{x})|^2 \,\middle|\, \mathbf{S}_X = \mathbf{f}\right) = \mathrm{E}\left(\left|S_\mathbf{x}^\alpha - \hat{S}_\mathbf{x}^\alpha\right|^2 \,\middle|\, \mathbf{S}_X = \mathbf{f}\right) = \mathrm{E}\left|S_\mathbf{x}^\alpha - \hat{S}_\mathbf{x}^\alpha\right|^2.$$

Moreover, the constructions of $S_\mathbf{x}^\alpha$ and $\hat{S}_\mathbf{x}^\alpha$ show that

$$\mathrm{E}\left|S_\mathbf{x}^\alpha - \hat{S}_\mathbf{x}^\alpha\right|^2 = D_1^\alpha D_2^\alpha K(\mathbf{x},\mathbf{x}) - \mathbf{k}_X^\alpha(\mathbf{x})^T\mathsf{A}_X^{-1}\mathbf{k}_X^\alpha(\mathbf{x}).$$

[1] Denote that h is the fill distance of the data points X for the domain Ω, i.e., $h := \sup_{\mathbf{x}\in\Omega}\min_{k=1,\ldots,n}\|\mathbf{x}-\mathbf{x}_k\|_2$ (see [28, Definition 1.4]). In another word, the fill distance h is the radius of the largest ball which is completely contained in Ω and which does not contain a data site. Obviously the data points X become a collection of densely separable points of the domain Ω when $h \to 0$.

Therefore, we complete the proofs. □

Clearly, the variance is consistent with the power function given in [28, Definition 11.2]. Interested readers can find the native space analogy of Theorem 3.3 in [28, Chapter 11]. Moreover, the recent papers [3,16] generalized the universal kriging for uncertainty propagation by the gradient of the Gaussian process, which can be viewed as a special case of Theorem 3.1. The difference of meshfree method and Gaussian process regression is the analysis of continuous and discrete features, respectively. The meshfree method is to solve the optimal function by the global norm, such as the minimization of the native norm. The Gaussian process regression is to estimate the local unknown value by the observation of the random variable, such as the minimization of the conditional variance. By the concept of kernel-based probability introduced here, we show that the classical results of meshfree methods for differentiation and the generalized algorithms of Gaussian process regression for gradients are strongly connected in a natural form of approximation theory. By the same idea, we can investigate the open problem of meshfree method using the technique of statistics. For examples, we can measure the quality of the meshfree estimator by the kernel-based probability measure.

3.2. Numerical Derivatives of Noisy Data. By equation (3.1), we know that $\mathbf{f}^\delta = \mathbf{f} + \xi^\delta$ where $\|\xi^\delta\|_\infty \leq \delta$. Since the noise ξ^δ is usually unknown, we naturally suppose that ξ^δ has the uniform distribution, such as $\xi^\delta := (\xi_1, \ldots, \xi_n)^T$ composed of the independent and uniform random variables $\xi_1, \ldots, \xi_n \sim \text{i.i.d.Unif}\,[-\delta, \delta]$. Here, we mainly look at the uniform noises. Actually, many general noises can be handled by the same method of this section. Let $\mathbf{V}_X^\delta := \mathbf{S}_X + \xi^\delta$. By the construction of $\mathbf{S}_X$, the noisy data $\mathbf{f}^\delta$ can be viewed as the observation of $\mathbf{V}_X^\delta$. Same as (3.5), we compute the estimator $S_{\mathbf{x}}^{\delta,\alpha}$ of $S_{\mathbf{x}}^\alpha$ to minimize the mean squared errors based on $\mathbf{V}_X^\delta$, such as

$$\hat{S}_{\mathbf{x}}^{\delta,\alpha} = \underset{U \in \text{span}\{\mathbf{V}_X^\delta\}}{\text{argmin}} \, \mathrm{E}\,|S_{\mathbf{x}}^\alpha - U|^2 . \tag{3.7}$$

The realization of $\hat{S}_{\mathbf{x}}^{\delta,\alpha}$ conditioned on $\mathbf{V}_X^\delta = \mathbf{f}^\delta$ will give rise to the estimator $s_X^{\delta,\alpha}(\mathbf{x})$ for the noisy data $\mathbf{f}^\delta$ and we will show that $s_X^{\delta,\alpha}(\mathbf{x})$ approximates $D^\alpha f(\mathbf{x})$. First, we provide a formulation to compute $s_X^{\delta,\alpha}(\mathbf{x})$ by the D^α partial derivative of the approximate function $s_X^\delta(\mathbf{x})$.

THEOREM 3.4. *The estimator $s_X^{\delta,\alpha}(\mathbf{x})$ is the D^α partial derivative of the kernel-based approximate function $s_X^\delta(\mathbf{x}) = \mathbf{k}_X(\mathbf{x})^T \mathbf{c}^\delta$, in which the coefficients $\mathbf{c}^\delta$ is uniquely determined by the linear system*

$$\left(\mathsf{A}_X + \frac{\delta^2}{3} \mathsf{I}_n \right) \mathbf{c}^\delta = \mathbf{f}^\delta,$$

where I_n is the identity matrix.

Proof. Since we have a normal distribution $(S_{\mathbf{x}}^\alpha, \mathbf{S}_X) \sim \mathcal{N}(\mathbf{0}, \mathsf{A}_X^\alpha(\mathbf{x}))$ and a uniform distribution $\xi^\delta \sim \text{Unif}\,[-\delta, \delta]^n$, the random variables $S_{\mathbf{x}}^\alpha$ and $\mathbf{V}_X^\delta$ have the mean 0 and the covariance matrix

$$\mathsf{A}_X^{\delta,\alpha}(\mathbf{x}) := \begin{pmatrix} D_1^\alpha D_2^\alpha K(\mathbf{x},\mathbf{x}) & \mathbf{k}_X^\alpha(\mathbf{x})^T \\ \mathbf{k}_X^\alpha(\mathbf{x}) & \mathsf{A}_X + \frac{\delta^2}{3}\mathsf{I}_n \end{pmatrix}.$$

This assures that the minimum-mean-square-error estimation of $S_{\mathbf{x}}^\alpha$ on the space spanned by $\{V_{\mathbf{x}_k}^\delta\}$ has the form

$$\hat{S}_{\mathbf{x}}^{\delta,\alpha} = \mathbf{w}_X^{\delta,\alpha}(\mathbf{x})^T \mathbf{V}_X^\delta,$$

where

$$\mathbf{w}_X^{\delta,\alpha}(\mathbf{x}) = \left(\mathsf{A}_X + \frac{\delta^2}{3}\mathsf{I}_n\right)^{-1} \mathbf{k}_X^{\alpha}(\mathbf{x}).$$

Therefore, we have that

$$s_X^{\delta,\alpha}(\mathbf{x}) = \mathbf{w}_X^{\delta,\alpha}(\mathbf{x})^T \mathbf{f}^{\delta}.$$

Moreover, since

$$D^{\alpha} s_X^{\delta}(\mathbf{x}) = D^{\alpha}\mathbf{k}_X(\mathbf{x}) \left(\mathsf{A}_X + \frac{\delta^2}{3}\mathsf{I}_n\right)^{-1} \mathbf{f}^{\delta},$$

we conclude that

$$s_X^{\delta,\alpha}(\mathbf{x}) = D^{\alpha} s_X^{\delta}(\mathbf{x}).$$

□

Now, we show that the proposed estimator for noisy data is δ–stable in the sense that the noise-free estimator is the limit when the noise vanishes.

THEOREM 3.5. *For any data site $X \subset \Omega$, the kernel-based estimator $s_X^{\delta,\alpha}(\mathbf{x})$ converges pointwise to the estimator $s_X^{\alpha}(\mathbf{x})$ as $\delta \to 0$.*

Proof. Using the formulas of $\mathbf{c}$ and $\mathbf{c}^{\delta}$ given in Theorems 3.1 and 3.4, we know that

$$\left(\mathsf{A}_X + \frac{\delta^2}{3}\mathsf{I}_n\right)(\mathbf{c} - \mathbf{c}^{\delta}) = \mathbf{f} - \mathbf{f}^{\delta} + \frac{\delta^2}{3}\mathbf{c}.$$

Thus, based on our assumption of the noisy data in (3.1), we have that

$$\left\|\mathbf{c} - \mathbf{c}^{\delta}\right\|_{\infty} \leq \left\|\left(\mathsf{A}_X + \frac{\delta^2}{3}\mathsf{I}_n\right)^{-1}\right\|_{\infty} \left(\left\|\mathbf{f} - \mathbf{f}^{\delta}\right\|_{\infty} + \frac{\delta^2}{3}\left\|\mathbf{c}\right\|_{\infty}\right) \leq \frac{3\delta + \delta^2\left\|\mathbf{c}\right\|_{\infty}}{3\lambda_{\min}\left(\mathsf{A}_X\right) + \delta^2},$$

where $\lambda_{\min}\left(\mathsf{A}_X\right)$ denotes the minimum eigenvalue of the SPD matrix A_X.

Theorems 3.1 and 3.4 provide that the both estimators $s_X^{\alpha}(\mathbf{x})$ and $s_X^{\delta,\alpha}(\mathbf{x})$ are the linear combinations of functions in $\mathbf{k}_X^{\alpha}(\mathbf{x})$ but with coefficients $\mathbf{c}$ and $\mathbf{c}^{\delta}$ respectively. Thus, the desired pointwise convergence is proven. □

THEOREM 3.6. *Suppose all the assumptions in Theorem 3.2 hold. Further suppose that the sequence of noise levels $\delta_m := \left\|\mathbf{f}_n^{\delta_m} - \mathbf{f}_n\right\|_{\infty}$ decreases monotonically to 0 as $m \to \infty$ for all n. Then, the estimator $s_{X_n}^{\delta_m,\alpha}(\mathbf{x})$ converges pointwise to $D^{\alpha} f(\mathbf{x})$ as $m, n \to \infty$.*

Proof. Combining Theorems 3.2 and 3.5, we immediately complete the proofs. □

We now extend the results for noise-free data in Theorem 3.3 to $s_X^{\delta,\alpha}(\mathbf{x})$. This allows us to evaluate the "approximate quality" *a posteriori* error without the knowledge of the unknown function $f \in \text{range}(\Gamma_K)$.

THEOREM 3.7. *If $f \in$ range(Γ_K), then the estimator $s_X^{\delta,\alpha}(\mathbf{x})$ has the error bound*

$$\left|D^{\alpha} f(\mathbf{x}) - s_X^{\delta,\alpha}(\mathbf{x})\right|^2 \leq C\sigma_X^{\delta,\alpha}(\mathbf{x})^2,$$

where C is the positive constant independent of X, δ, α and the noisy power function

$$\sigma_X^{\delta,\alpha}(\mathbf{x})^2 := D_1^{\alpha} D_2^{\alpha} K(\mathbf{x},\mathbf{x}) - \mathbf{k}_X^{\alpha}(\mathbf{x})^T \left(\mathsf{A}_X + \frac{\delta^2}{3}\mathsf{I}_n\right)^{-1} \mathbf{k}_X^{\alpha}(\mathbf{x}). \tag{3.8}$$

Proof. Since $f \in \text{range}(\Gamma_K)$, there exists $U \in \mathcal{H}_S$ such that $f = \Gamma_K(U)$. By the structure theorem of Gaussian measures, we have that $D^\alpha f(\mathbf{x}) = \mathrm{E}(S_{\mathbf{x}}^\alpha U)$ and $\mathbf{f} = f(X) = \mathrm{E}(\mathbf{S}_X U)$. Since the noise ξ^δ is independent of U and S, there exists a random variable Λ independent of U and S such that $\mathrm{E}(\Lambda^2) < \infty$ and $\mathbf{f}^\delta - \mathbf{f} = \mathrm{E}(\xi^\delta \Lambda)$. Therefore, we have that $D^\alpha f(\mathbf{x}) = \mathrm{E}(S_{\mathbf{x}}^\alpha U) + \mathrm{E}(S_{\mathbf{x}}^\alpha \Lambda) = \mathrm{E}(S_{\mathbf{x}}^\alpha (U+\Lambda))$ and $\mathbf{f}^\delta = \mathrm{E}(\mathbf{S}_X U) + \mathrm{E}(\xi^\delta \Lambda) = \mathrm{E}((\mathbf{S}_X + \xi^\delta)(U+\Lambda)) = \mathrm{E}(\mathbf{V}_X^\delta (U+\Lambda))$. This shows that

$$\begin{aligned}
&\left| D^\alpha f(\mathbf{x}) - s_X^{\delta,\alpha}(\mathbf{x}) \right|^2 = \left| D^\alpha f(\mathbf{x}) - \mathbf{w}_X^{\delta,\alpha}(\mathbf{x})^T \mathbf{f}^\delta \right|^2 \\
&= \left| \mathrm{E}(S_{\mathbf{x}}^\alpha (U+\Lambda)) - \mathbf{w}_X^{\delta,\alpha}(\mathbf{x})^T \mathrm{E}(\mathbf{V}_X^\delta (U+\Lambda)) \right|^2 \\
&\leq \left| \mathrm{E}\left(S_{\mathbf{x}}^\alpha (U+\Lambda) - \mathbf{w}_X^{\delta,\alpha}(\mathbf{x})^T \mathbf{V}_X^\delta (U+\Lambda) \right) \right|^2 \\
&\leq \left(\mathrm{E}|U+\Lambda|^2 \right) \left(\mathrm{E} \left| S_{\mathbf{x}}^\alpha - \mathbf{w}_X^{\delta,\alpha}(\mathbf{x})^T \mathbf{V}_X^\delta \right|^2 \right).
\end{aligned}$$

Moreover, we denote the positive constant $C := \mathrm{E}|U+\Lambda|^2$. Thus, C is independent of X, δ, α. Finally, we compute the noisy power function $\sigma_X^{\delta,\alpha}(\mathbf{x})^2$, such as

$$\begin{aligned}
&\mathrm{E} \left| S_{\mathbf{x}}^\alpha - \mathbf{w}_X^{\delta,\alpha}(\mathbf{x})^T \mathbf{V}_X^\delta \right|^2 \\
&= \mathrm{E}(S_{\mathbf{x}}^\alpha S_{\mathbf{x}}^\alpha) - 2\mathbf{w}_X^{\delta,\alpha}(\mathbf{x})^T \mathrm{E}(S_{\mathbf{x}}^\alpha \mathbf{V}_X^\delta) + \mathbf{w}_X^{\delta,\alpha}(\mathbf{x})^T \mathrm{E}(\mathbf{V}_X^\delta \mathbf{V}_X^{\delta T}) \mathbf{w}_X^{\delta,\alpha}(\mathbf{x}) \\
&= D_1^\alpha D_2^\alpha K(\mathbf{x},\mathbf{x}) - 2\mathbf{w}_X^{\delta,\alpha}(\mathbf{x})^T \mathbf{k}_X^\alpha(\mathbf{x}) + \mathbf{w}_X^{\delta,\alpha}(\mathbf{x})^T \left(\mathsf{A}_X + \frac{\delta^2}{3} \mathsf{I}_n \right) \mathbf{w}_X^{\delta,\alpha}(\mathbf{x}) \\
&= D_1^\alpha D_2^\alpha K(\mathbf{x},\mathbf{x}) - \mathbf{k}_X^\alpha(\mathbf{x})^T \left(\mathsf{A}_X + \frac{\delta^2}{3} \mathsf{I}_n \right)^{-1} \mathbf{k}_X^\alpha(\mathbf{x}) = \sigma_X^{\delta,\alpha}(\mathbf{x})^2.
\end{aligned}$$

The proof is completed. ∎

COROLLARY 3.8. *The noisy power function $\sigma_X^{\delta,\alpha}(\mathbf{x})^2$ is the sum of the noise-free power function $\sigma_X^\alpha(\mathbf{x})^2$ and some noise-induced residual term as*

$$\sigma_X^{\delta,\alpha}(\mathbf{x})^2 = \sigma_X^\alpha(\mathbf{x})^2 + \varrho_X^{\delta,\alpha}(\mathbf{x})^2, \tag{3.9}$$

where

$$\varrho_X^{\delta,\alpha}(\mathbf{x})^2 := \delta^2 \mathbf{k}_X^\alpha(\mathbf{x})^T \left(3\mathsf{A}_X^2 + \delta^2 \mathsf{A}_X \right)^{-1} \mathbf{k}_X^\alpha(\mathbf{x}) \leq \frac{\delta^2 D_1^\alpha D_2^\alpha K(\mathbf{x},\mathbf{x})}{3\lambda_{\min}(\mathsf{A}_X) + \delta^2}, \tag{3.10}$$

and $\lambda_{\min}(\mathsf{A}_X)$ denotes the minimum eigenvalue of the SPD matrix A_X. Thus, the noisy variance vanishes as $\mathcal{O}(\delta^2)$.

Proof. Note that the power functions (3.6) and (3.8) share an identical first term. It is straightforward to verify (3.9). By the eigen-decomposition of the SPD matrix A_X, we can rewrite and bound the exact residual term in (3.10) as

$$\begin{aligned}
&\mathbf{k}_X^\alpha(\mathbf{x})^T \mathsf{A}_X^{-1} \mathbf{k}_X^\alpha(\mathbf{x}) - \mathbf{k}_X^\alpha(\mathbf{x})^T \left(\mathsf{A}_X + \frac{\delta^2}{3} \mathsf{I}_n \right)^{-1} \mathbf{k}_X^\alpha(\mathbf{x}) \\
&= \delta^2 \mathbf{k}_X^\alpha(\mathbf{x})^T \left(3\mathsf{A}_X^2 + \delta^2 \mathsf{A}_X \right)^{-1} \mathbf{k}_X^\alpha(\mathbf{x}) = \varrho_X^{\delta,\alpha}(\mathbf{x})^2.
\end{aligned}$$

This indicates another upper bound of the residual terms

$$\varrho_X^{\delta,\alpha}(\mathbf{x})^2 \leq \left(\frac{\delta^2}{3\lambda_{\min}(\mathsf{A}_X) + \delta^2} \right) \mathbf{k}_X^\alpha(\mathbf{x})^T \mathsf{A}_X^{-1} \mathbf{k}_X^\alpha(\mathbf{x}) \leq \frac{\delta^2 D_1^\alpha D_2^\alpha K(\mathbf{x},\mathbf{x})}{3\lambda_{\min}(\mathsf{A}_X) + \delta^2}. \tag{3.11}$$

It is now obvious that $\varrho_X^{\delta,\alpha}(\mathbf{x})^2 \to 0$ as $\delta \to 0$. □

The noisy power function $\sigma_X^{\delta,\alpha}(\mathbf{x})^2$ can be controlled by making $\sigma_X^{\alpha}(\mathbf{x})^2 + \varrho_X^{\delta,\alpha}(\mathbf{x})^2$ small. An appropriate kernel selection strategy (i.e., our regularization strategy) should have $\lambda_{\min}(A_X)$ bounded away from zero in order to avoid the residual term $\varrho_X^{\delta,\alpha}(\mathbf{x})$ from blowing up. This idea will be further elaborated in the coming section

By the classical meshfree method, the kernel-based approximate function s_X^{δ} is the optimizer of the regularization problem defined in the reproducing kernel Hilbert space $\mathcal{H}_K(\Omega)$, i.e.,

$$s_X^{\delta} := \operatorname*{argmin}_{f \in \mathcal{H}_K(\Omega)} \|f(X) - \mathbf{f}^{\delta}\|_2^2 + \frac{\delta^2}{3}\|f\|_{\mathcal{H}_K(\Omega)}^2.$$

If there exists a function $f^{\delta} \in \mathcal{H}_K(\Omega)$ such that $f^{\delta}(X) = \mathbf{f}^{\delta}$, then, by the reproducing property, we have that

$$\begin{aligned} |D^{\alpha} f(\mathbf{x}) - s_X^{\delta,\alpha}(\mathbf{x})| &\le |D^{\alpha} f(\mathbf{x}) - s_X^{\alpha}(\mathbf{x})| + |s_X^{\alpha}(\mathbf{x}) - s_X^{\delta,\alpha}(\mathbf{x})| \\ &\le \sigma_X^{\alpha}(\mathbf{x})\|f\|_{\mathcal{H}_K(\Omega)} + \varrho_X^{\delta,\alpha}(\mathbf{x})\ell_X(f) + D_1^{\alpha} D_2^{\alpha} K(\mathbf{x},\mathbf{x})^{\frac{1}{2}} \ell_X(f - f^{\delta}), \end{aligned}$$

where $\ell_X(g) := \left(\mathbf{g}^T \mathsf{A}_X^{-1} \mathbf{g}\right)^{1/2}$ for $\mathbf{g} := g(X)$ can be viewed as the approximate norm of the native norm $\|g\|_{\mathcal{H}_K(\Omega)}$ for $g \in \mathcal{H}_K(\Omega)$; more discussions can be found in the coming section. According to Corollary 3.8, the noisy power function can be seen as the extended error bound of the classical error bound of the meshfree estimator. This shows that the stochastic approach can supplement the knowledge of the unknown area of meshfree method.

4. Identifying quasi-optimal kernels. In practice, we need to fix a kernel before constructing any kernel-based estimator. The aim of the section is to "identify" a quasi-optimal kernel from a family of some continuously differentiable SPD kernels

$$\Xi := \left\{K^{\theta}(\mathbf{x},\mathbf{y}) : \theta > 0\right\}, \tag{4.1}$$

based on the noisy data $(X, \mathbf{f}^{\delta})$. To do so, a computable strategy is needed. The noise-free and noisy power functions in the previous section could be such a tool. However, as discussed in Section 2, the power function is only dependent on the data points X but independent of the observations $\mathbf{f}^{\delta}$. For a reliable strategy, we proposed a data dependency quasi-optimization strategy.

In classical meshfree method, the pointwise error [28, Theorem 11.4] of the kernel-based estimator is bounded by

$$|D^{\alpha} f(\mathbf{x}) - s_X^{\alpha}(\mathbf{x})|^2 \le \sigma_X^{\alpha}(\mathbf{x})^2 \|f\|_{\mathcal{H}_K(\Omega)}^2 \quad \text{for all} \quad f \in \mathcal{H}_K(\Omega).$$

This motivates us to include the native norm of f from the observed data. Then, we use some objective functions like $\sigma_X^{\delta,\alpha}(\mathbf{x})^2 \|f\|_{\mathcal{H}_K(\Omega)}^2$ or in some other equivalent forms (denoted by $\sim$) to search for a quasi-optimal kernel $K_{\theta^*} \in \Xi$. Since the exact value $f(X)$ is unavailable, we must use a statistical analogy to approximate $\|f\|_{\mathcal{H}_K(\Omega)}$.

By the optimal recovery property of the meshfree interpolant of $f \in \mathcal{H}_K(\Omega)$, we have $\|s_X\|_{\mathcal{H}_K(\Omega)} \le \|f\|_{\mathcal{H}_K(\Omega)}$ for all $f \in \mathcal{H}_K(\Omega)$. Instead of $\|f\|_{\mathcal{H}_K(\Omega)}^2$, the first approximation we make is to minimize $\|s_X\|_{\mathcal{H}_K(\Omega)}$, which makes sense for all $f \in \mathcal{H}^m(\Omega)$. Based on the probability density function p_X of $\mathbf{S}_X$, i.e.,

$$p_X(\mathbf{z}) = \frac{1}{\sqrt{\det(2\pi \mathsf{A}_X)}} \exp\left(-\frac{1}{2}\mathbf{z}^T \mathsf{A}_X^{-1} \mathbf{z}\right), \quad \text{for } \mathbf{z} \in \mathbb{R}^n,$$

we can rewrite

$$\|s_X\|^2_{\mathcal{H}_K(\Omega)} = \mathbf{f}^T A_X^{-1}\mathbf{f} = -\log\det(2\pi A_X) - 2\log p_X(\mathbf{f}).$$

Note that the magnitude of $\log\det(2\pi A_X)$ depends heavily on $\lambda_{\min}(A_X)$ that is already controlled by $\varrho_X^{\delta,\alpha}(\mathbf{x})^2$ in (3.11). For the exact data, $f(X) = \mathbf{f}$ and $\mathbb{P}_K(\mathbf{S}_X = \mathbf{f}) \sim p_X(\mathbf{f})$. For the noisy data, $f(X) \in \mathbf{f}^\delta + [-\delta,\delta]^n$. To obtain a computational formula, we integrate the probability density function over the set $\mathbf{f}^\delta + [-\delta,\delta]^n$, i.e.,

$$\mathbb{P}_K\left(\mathbf{S}_X \in \mathbf{f}^\delta + [-\delta,\delta]^n\right) = \int_{\mathbf{f}^\delta+[-\delta,\delta]^n} p_X(\mathbf{z})\mathrm{d}\mathbf{z},$$

which allows us to work with the kernel-based probability measure. Then, we approximate the probability by the Taylor expansion as in

$$\begin{aligned}\int_{\mathbf{f}^\delta+[-\delta,\delta]^n} p_X(\mathbf{z})\mathrm{d}\mathbf{z} &= \int_{\mathsf{A}_X^{-1/2}(\mathbf{f}^\delta+[-\delta,\delta]^n)} \frac{1}{(2\pi)^{n/2}}\exp\left(-\frac{1}{2}\mathbf{z}^T\mathbf{z}\right)\mathrm{d}\mathbf{z}\\ &= \frac{2^n\delta^n}{(2\pi)^{n/2}}\exp\left(-\frac{1}{2}\mathbf{f}^{\delta T}\mathsf{A}_X^{-1}\mathbf{f}^\delta\right) + \mathcal{O}\left(\delta^{n+2}\right),\end{aligned}$$

where the covariance matrix $\mathsf{A}_X = K(X,X)$ is related to the kernel $K \in \Xi$ of choice. Altogether, this yields the sequence of approximations

$$\sigma_X^\alpha(\mathbf{x})^2\,\|f\|^2_{\mathcal{H}_K(\Omega)} \sim \sigma_X^{\delta,\alpha}(\mathbf{x})^2\,(-\log p_X(\mathbf{f})) \sim \sigma_X^{\delta,\alpha}(\mathbf{x})^2\mathbf{f}^{\delta T}\mathsf{A}_X^{-1}\mathbf{f}^\delta,$$

and we define the following minimization problem for finding a pointwise quasi-optimal kernel for any $\mathbf{x} \in \Omega$, such as

$$K^*_{\mathbf{x}} := \underset{K\in\Xi}{\operatorname{argmin}} \left(D_1^\alpha D_2^\alpha K(\mathbf{x},\mathbf{x}) - \mathbf{k}_X^\alpha(\mathbf{x})^T\left(\mathsf{A}_X + \frac{\delta^2}{3}\mathsf{I}_n\right)^{-1}\mathbf{k}_X^\alpha(\mathbf{x})\right)\mathbf{f}^{\delta T}\mathsf{A}_X^{-1}\mathbf{f}^\delta. \quad (4.2)$$

Here, the native norm of f is approximated by the observed data, i.e., $\|f\|^2_{\mathcal{H}_K(\Omega)} \approx \mathbf{f}^{\delta T}\mathsf{A}_X^{-1}\mathbf{f}^\delta$. Since $\sigma_X^{\delta,\alpha}(\mathbf{x})^2 \to \sigma_X^\alpha(\mathbf{x})^2$ and $\mathbf{f}^{\delta T}\mathsf{A}_X^{-1}\mathbf{f}^\delta \to \mathbf{f}^T\mathsf{A}_X^{-1}\mathbf{f}$ when $\delta \to 0$, the minimization problem (4.2) is consistent with the minimization of the classical upper bound of meshfree method.

To identify a quasi-optimal kernel for Ω, the corresponding minimization problem can be defined by integrating (4.2) over Ω, such as

$$K^*_\Omega := \underset{K\in\Xi}{\operatorname{argmin}} \int_\Omega \left(D_1^\alpha D_2^\alpha K - \mathbf{k}_X^{\alpha,T}\left(\mathsf{A}_X + \frac{\delta^2}{3}\mathsf{I}_n\right)^{-1}\mathbf{k}_X^\alpha\right)\mathbf{f}^{\delta T}\mathsf{A}_X^{-1}\mathbf{f}^\delta\mathrm{d}\mathbf{x}, \quad (4.3)$$

or by minimization in the L^∞–sense, such as

$$K^*_\Omega := \underset{K\in\Xi}{\operatorname{argmin}} \sup_{\mathbf{x}\in\Omega} \left(D_1^\alpha D_2^\alpha K - \mathbf{k}_X^{\alpha,T}\left(\mathsf{A}_X + \frac{\delta^2}{3}\mathsf{I}_n\right)^{-1}\mathbf{k}_X^\alpha\right)\mathbf{f}^{\delta T}\mathsf{A}_X^{-1}\mathbf{f}^\delta, \quad (4.4)$$

Both (4.3) and (4.4) can be computed without any knowledge of the exact f.

In the area of machine learning, one can find many algorithms to learn the optimal parameters dependent of stochastic variables. In this section, we mainly use the combination of minimum variance estimation and maximum likelihood estimation to solve the quasi-optimal kernel.

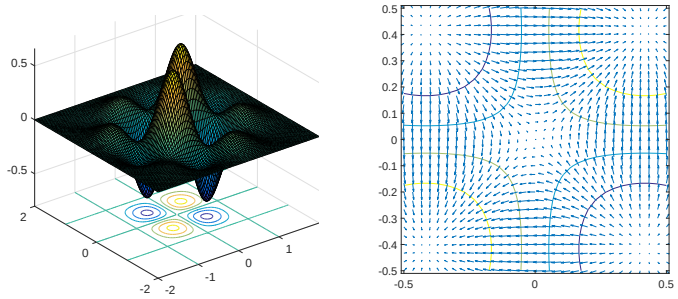

FIG. 4.1. *Example 1: Unknown function* $f : [-2,2]^2 \to \mathbb{R}$ *and its gradient field in* $[-0.5,0.5] \times [-0.5,0.5]$.

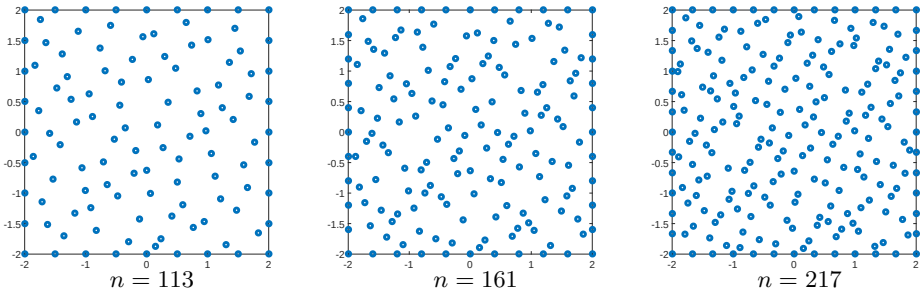

FIG. 4.2. *Example 1: Three sets of scattered data point* X *used for testings.*

4.1. Numerical examples. Now we investigate the numerical performance and the accuracy of the proposed quasi-optimal estimator. We focus on the family of Gaussian kernels

$$\Xi = \left\{K^\theta(\mathbf{x},\mathbf{y}) : \theta > 0\right\} \quad \text{with} \quad K^\theta(\mathbf{x},\mathbf{y}) = \exp(-\theta^2 \left\|\mathbf{x}-\mathbf{y}\right\|_2^2),$$

with different shape parameter θ. The n exact data in $\mathbf{f}$ are evaluated by test functions (used in [24, 25] for 1- and 2-D, see Figure 4.1) in the form of

$$f(\mathbf{x}) = e^{-\|\mathbf{x}\|_2^2} \prod_{j=1}^{d} \sin(\pi x_j) \quad \text{for} \quad \mathbf{x} = (x_1,\ldots,x_d) \in \Omega = [-2,2]^d,$$

at some sets of scattered data points X generated by Halton sequences; see Figure 4.2 for examples.

The $L^\infty(\Omega)$ norms of f are 0.66 and 0.54 for $d = 2$ and $d = 3$ respectively. The n noisy data in $\mathbf{f}^\delta$ are generated (for each tested n) by the uniform random noise ξ^δ composed of uniform random variables $\xi_1,\ldots,\xi_n \sim$ i.i.d.Unif $[-\delta,\delta]$, that is, $\mathbf{f}^\delta := \mathbf{f} + \xi^\delta$. Reported $L^\infty(\Omega)$ errors are approximated using a set of (denser than all tested X) regularly distributed evaluation points. To identify the optimal kernel, we search the set Ξ for θ from 0.5 by 0.05 to 8.0. The true "optimal" (for each tested case) is the one that yields the smallest $L^\infty(\Omega)$ error. Numerical experiments show that the numerical performance of quasi-optimal kernels selected by (4.3) and (4.4) are similar. In this section, all quasi-optimal kernels are selected based on the L^∞–minimization given in (4.4) whose functional is evaluated at the set of points used for error evaluation.

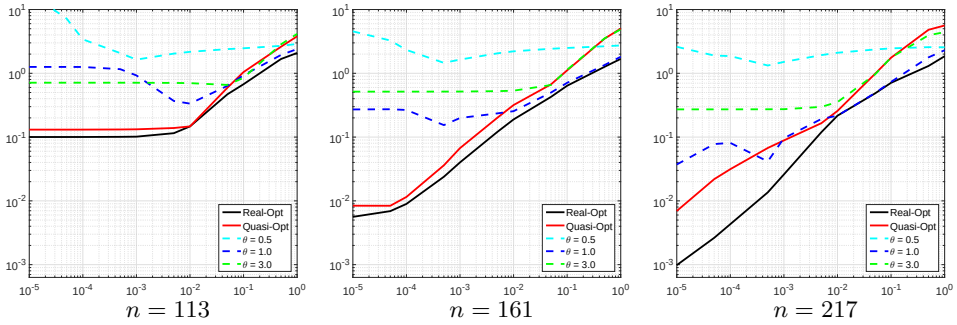

FIG. 4.3. *Example 1: The absolute L^∞ error profiles of various kernels verse different noise levels δ.*

Example 1: Two dimensional gradient. We compare our proposed method and the numerical methods proposed in [24,25] by solving the same test problem, that is, to find ∇f from noisy data. Figure 4.3 show the δ–convergence of five different kernel-based estimators based on the three sets of scattered data X_n in Figure 4.2. The `Real-Opt` is the true optimal in Ξ, which is available only if we know f. The `Quasi-Opt` is obtained by (4.4). Note that both estimators use adaptive shape parameters $\theta(\delta)$ that change with δ. For comparison, we also include the results of three estimators with fixed $\theta = 0.5$, 1.0 and 3.0 respectively. Here are two worth noting observations:

- adaptive shape parameters are of a higher necessity for small numbers of data n, and
- using more data could be harmful to both efficiency and accuracy for large noise level δ.

To see the first point, we can focus on the curve for $\theta = 1.0$. The larger the n the longer it can stay close to the optimal. For the second, we can carefully compare the accuracies of `Real-Opt` and `Quasi-Opt` for different n in the range of δ between 10^{-2} to 10^{0} (approx. 1.5% to 15% noise). The `Quasi-Opt` result of using $n = 113$ data is very close to the optimal. By comparing the errors of `Real-Opt` estimations for different n, we can actually see that $n = 113$ yields more accurate approximation. The accuracy of our method is competitive with those obtained by other meshfree numerical differentiation methods, see [24, Tab.2] and [25, Exmp.2]. Note that the presented numerical results in these papers are not resulted from uniform random noises; the added noises are generated either by the sin function or random numbers following the normal distribution $\mathcal{N}(0, 0.01)$. Moreover, the number of data points there, n, were about 600 and 900 and the reported errors in both are root mean square (RMS) norm, which is always smaller than our reported $L^\infty(\Omega)$ errors. Most importantly, they use thin plate and cubic splines while we are using infinitely smooth Gaussian kernels. The RMS errors of our quasi-optimal estimators obtained from noisy data with $\delta = 10^{-3}$ are 0.138, 0.0658, and 0.0457 respectively, for $n = 113$, 161, and 217.

For $n = 217$, the performance of our quasi-optimal estimators deteriorates and the accuracy could be an order larger than the true optimal. Figure 4.4 shows some details in the quasi-optimal search. In each snapshot, we show the exact error (in blue) and a posteriori estimator (in red), i.e., the functional in (4.4). Starting from the left, the case of $n = 113$ and $\delta = 10^{-3}$ shows a typical behavior for "small" noise, in which we can see both curves are convex (up to some oscillations due to

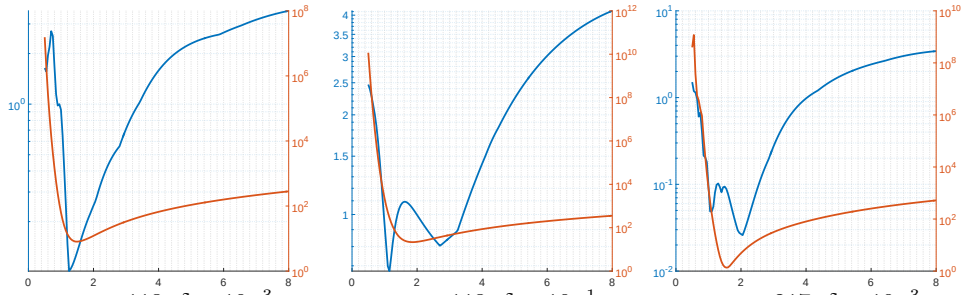

FIG. 4.4. *Example 1: The absolute L^∞ error profiles (y-left in blue) and the proposed posteriori error measure (y-right in red) of various X_n and δ verse different shape parameter θ.*

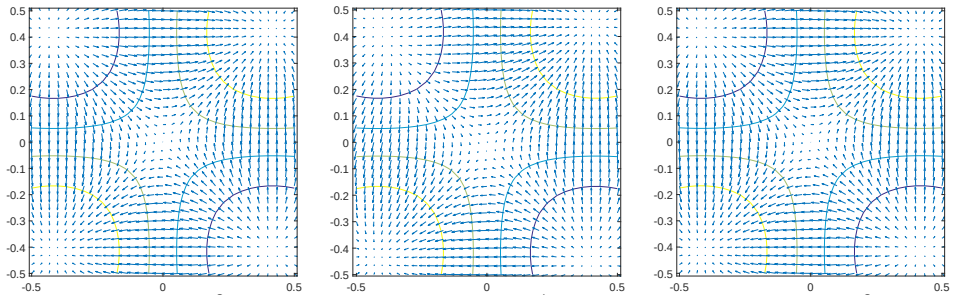

FIG. 4.5. *Example 1: Gradient fields in $[-0.5, 0.5]\times[-0.5, 0.5]$ obtained by quasi-optimal kernels associated with Figure 4.4.*

the problem of ill-conditioning). When the noise level increases, say to $\delta = 10^{-1}$ in the middle plot, the exact error is no longer convex in such a way that as if a region near the minimum is being flipped upside down. Yet, our proposed estimator remains convex and hence, select the local maximum. The noise level for such phenomenon is relative; when $n = 217$, $\delta = 10^{-3}$ is not "small" enough. This explains the gap between the red and blue curves in Figure 4.3($n = 217$) for $10^{-2} \le \delta \le 10^{0}$. We omit the graphic presentation for the case $\delta < 10^{-2}$; in such case, both the red and blue curves are similar in shape. The selected quasi-optimal θ is at most 0.5 away from the true optimal shape-parameter; such a difference in shape parameter results in one order of accuracy loss.

The goal of this test problem is to obtain an approximation to the gradient of f. We end this example by showing the gradient fields obtained by our quasi-optimal kernels in Figure 4.5 associated with the three settings in Figure 4.4. We zoom-in to a smaller region near the origin where the function f varies most to enhance resolution. Readers can compare these results with the exact one shown in Figure 4.2. Even with $\delta = 10^{-1}$ (approx. 15% noise), the resulting vector field shows great resemblance to the exact one with a noticeable error in the region near $(-0.5, 0)$.

Example 2: Three dimensional Laplacian. Meshfree methods are more competitive in higher dimensions. This example aims to approximate the Laplacian of $f : \mathbb{R}^3 \to \mathbb{R}$ from noisy data. We construct noisy data on a set of 1115 scattered data in $[-2, 2]^3$ with noise level 0.01 and 0.05 (approx. 2% and 10% noise, respectively). All other settings are similar to those in Example 1. To visualize the volumetric data, all slice-plots in the example show the orthogonal planes that slice at the x_1-, x_2-,

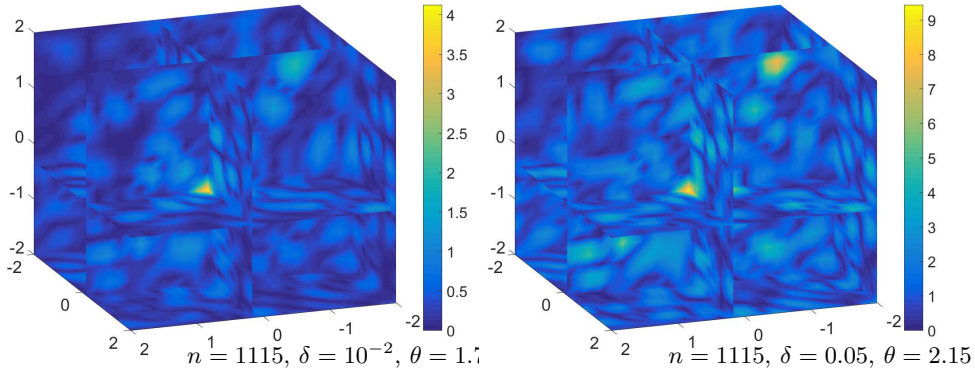

FIG. 4.6. *Example 2: The absolute L^∞ error profiles for the quasi-optimal estimators associated with two different noise level δ.*

x_3-coordinates at which the maximum error of the estimator is attainted.

Figure 4.6 shows the maximum error profiles of the quasi-optimal estimators for the two test cases. The maximum errors are 4.12 and 9.43 respectively, where as $\|\Delta f\|_{L^\infty(\Omega)} \approx 18$. The corresponding RMS errors are much more impressive: 0.0631 and 0.120 respectively. In Figure 4.6, we can see that the error function is very localized. This reflects the fact that the estimators cannot capture the exact "amplitude" of the Laplacian with the presence of noise.

To see how well the proposed quasi-optimal estimator can approximate Δf, Figures 4.7–4.8 show the exact and approximated values of the Laplacian. Readers should focus at the intersection of the interior slices where the maximum error occurs. The approximation associated with $\delta = 0.01$ in Figure 4.7 matches closely with the exact one. A noticeable numerical error can be seen near the bright yellow spot on the right side where ∇f is large. The quasi-optimal estimator can capture the shape but not the amplitude. The approximation associated with $\delta = 0.05$ in Figure 4.8 is not accurate up to fine details, but it certainly can capture all the basic features.

5. Conclusions. We derive and analyze kernel-based estimators that approximate partial derivatives of an unknown function from scattered noisy data. We obtain some statistical error estimates for a class of estimators that can be obtained by solving a simple regularized meshfree linear system with a fixed *a prior* regularized parameter depending on noise level. Instead of tuning the regularization parameters, the problem of parameter selection is transferred to selecting a parameter of the employed kernel. By interacting with various meshfree and probability theories, we propose an *a posteriori* strategy for finding quasi-optimal kernels that depend on all the inputs of numerical derivatives problems (i.e., data location, noise level, noisy data, etc.) but not the exact solution. Numerical examples show that the proposed quasi-optimal estimator is particularly powerful when the number of data is small.

When more data is available, we can modify the proposed method and adaptively select a subset for numerical differentiation. Such adaptive technique is already available for direct problem [19]. Using the generalized power function we developed, it is hopeful that a scientific computing problem analogy can be developed. Although the presented numerical results focus on the Gaussian kernel and we use brute-force approaches for finding the quasi-optimal, more efficient algorithms can be developed for some kernels. For the Gaussian and Sobolev kernels, one could begin with the

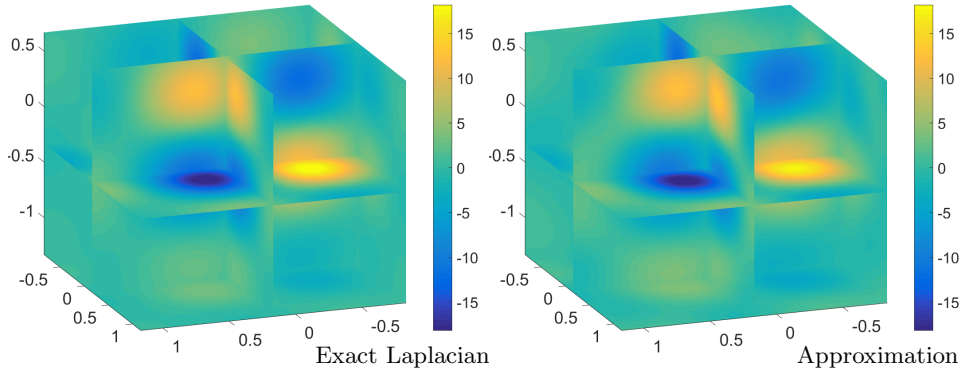

FIG. 4.7. *Example 2: The exact and approximated Laplacian using the quasi-optimal estimators associated with* $\delta = 10^{-2}$ *(approx. 2% noise) around the point where the maximum error occurs.*

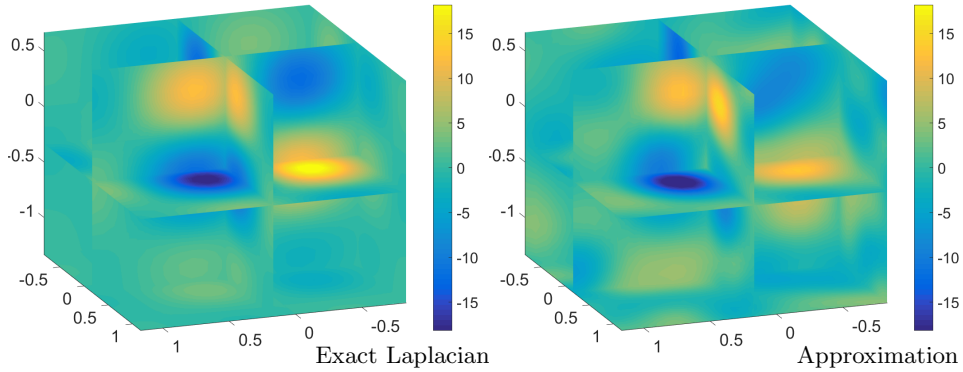

FIG. 4.8. *Example 2: The exact and approximated Laplacian using the quasi-optimal estimators associated with* $\delta = 0.05$ *(approx. 10% noise) around the point where the maximum error occurs.*

decomposition for stable evaluation [8] and develop a fast algorithm to solve linear systems with different shape parameters. For compactly supported functions, powers and thin-plate splines (note: the latter two are conditionally SPD), one can make use of the polynomial-type upper and lower bounds for their power functions and minimum eigenvalues of interpolation matrices to get fast algorithm. We leave both ideas open for now as possible future research.

This work is the first attempt to develop meshfree algorithms and theories for numerical differentiation from a stochastic point of view. We focus on the problem of numerical differentiation with integer order as it is closely related the meshfree interpolation. It is straightforward to extended the presented framework to fractional orders [15]. We believe the mixed meshfree and stochastic techniques used in this paper can be generalized to other scientific computing problems, such as inverse problems. The quasi-optimal method, which is difficult to introduce in the deterministic models, is well-defined by the kernel-based probability measures. This shows that the kernel-based probability measure gives a new direction of meshfree method to solve the open problems of classical approximation theory. The technique has high potential values in new research areas for identifying the quasi-optimal kernels for scientific computing problems by different learning methods. By the same discussions in [5, 11, 12] of learning with the regularization schemes, the meshfree method is a numerical tool of a scaling parameter on machine learning. Since the complexities of

equations (4.3) and (4.4) are dependent of the global meshfree algorithm, it is difficult to compute the quasi-optimal parameter for the large size of data. It is a good research topic to investigate the large-scale data of numerical differentiations by the local meshfree technique.

Acknowledgments. This work was partially supported by a Hong Kong Research Grant Council GRF Grant, a Hong Kong Baptist University FRG Grant, the "Thousand Talents Program" of China, the National Natural Science Foundation of China (11601162), and the South China Normal University Grant.

REFERENCES

[1] M. Bozzini and M. Rossini, Numerical differentiation of 2D functions from noisy data, *Comput. Mat. Appl.* **45** (2003) 309-327.
[2] M. D. Buhmann, *Radial Basis Functions: Theory and Implementations* (Cambridge University Press, Cambridge, 2003).
[3] X. Chen, B. E. Ankenman, and B. L. Nelson, Enhancing stochastic kriging metamodels with gradient estimators, *Oper. Res.* **61** (2013) 512-528.
[4] J. Cheng, Y. C. Hon, and Y. B. Wang, A numerical method for the discontinuous solutions of Abel integral equations, in *Inverse Problems and Spectral Theory*, eds. H. Isozaki, (Amer. Math. Soc., Providence, RI, 2004), pp. 233-243.
[5] J. Fan, T. Hu, Q. Wu, and D. X. Zhou, Consistency analysis of an empirical minimum error entropy algorithm, *Appl. Comput. Harmonic Anal.* **41** (2016), 164-189.
[6] G. E. Fasshauer, *Meshfree Approximation Methods with Matlab* (World Scientific Publishing Co. Inc., Hackensack, NJ, 2007).
[7] G. E. Fasshauer and M. J. McCourt, *Kernel-based Approximation Methods using Matlab* (World Scientific Publishing Co. Inc., Hackensack, NJ, 2015).
[8] G. E. Fasshauer and M. J. McCourt, Stable evaluation of Gaussian radial basis function interpolants, *SIAM J. Sci. Comput.* **34** (2012) A737-A762.
[9] C.-L. Fu, X.-L. Feng, and Z. Qian, Wavelets and high order numerical differentiation, *Appl. Math. Modelling* **34** (2010) 3008-3021.
[10] R. Gorenflo and M. Yamamoto, Operator-theoretic treatment of linear Abel integral equations of first kind, *Japan J. Indust. Appl. Math.* **16** (1999) 137-161.
[11] Z. C. Guo, D. H. Xiang, X. Guo, and D. X. Zhou, Thresholded spectral algorithms for sparse approximations, *Anal. Appl.* **15** (2017), 433-455.
[12] T. Hu, J. Fan, Q. Wu, and D. X. Zhou, Regularization schemes for minimum error entropy principle, *Anal. Appl.* **13** (2015), 437-455.
[13] C. Itiki and J. J. Neto, Complete automation of the generalized inverse method for constrained mechanical systems of particles, *Appl. Math. Comput.* **152** (2004) 561-580.
[14] I. R. Khan and R. Ohba, New finite difference formulas for numerical differentiation, *J. Comp. Appl. Math.* **126** (2000) 269-276.
[15] M. Li, Y. Wang, and L. Ling, Numerical caputo differentiation by radial basis functions, *J. Sci. Compt.* **62** (2015) 300-315.
[16] B. A. Lockwood and M. Anitescu, Gradient-enhanced universal kriging for uncertainty propagation, *Nucl. Sci. Eng.* **170** (2012) 168-195.
[17] F. J. Narcowich, J. D. Ward, and H,. Wendland, Sobolev bounds on functions with scattered zeros, with applications to radial basis function surface fitting, *Math. Comp.* **74** (2005) 743-763.
[18] A. G. Ramm and A. B. Smirnova, On stable numerical differentaiation, *Math. Comp.* **70** (2001) 1131-1153.
[19] R. Schaback, A computational tool for comparining all inear PDE solvers, *Adv. comput. Math.* **41** (2014) 333-355.
[20] M. Scheuerer, R. Schaback, and M. Schlather, Interpolation of spatial data - a stochatic or a determinisitc problem? *European J. Appl. Math.* **24** (2013) 601-629.
[21] M. L. Stein, *Interpolation of Spartial Data: Some Theory for Kriging* (Springer-Verlg, New York, 1999).
[22] X. Q. Wan, Y. B. Wang, and M. Yamamoto, Detection of irregular points by regularization in numerical differentiation and application to edge detection, *Inverse Probl.* **22** (2006) 1089-1103.

[23] Y. B. Wang, X. Z. Jia, and J. Cheng, A numerical differentiation method and its application to reconstruction of discontinuity, *Inverse Probl.* **18** (2002) 1461-1476.
[24] T. Wei and Y. C. Hon, Numerical derivatives from one-dimensional scattered noisy data, *J. Phys. Conf. Ser.* **12** (2005) 171.
[25] T. Wei and Y. C. Hon, Numerical differentiation by radial bais functions approximation, *Adv. Comput. Math.* **27** (2006) 247-272.
[26] T. Wei, Y. C. Hon, and Y. B. Wang, Reconstruction of numerical derivatives from scattered noisy data, *Inverse Probl.* **21** (2005) 657-672.
[27] T. Wei and M. Li, High order numerical derivatives for one-dimensional scattered nosiy data, *Appl. Math. Comput.* **175** (2006) 1744-1759.
[28] H. Wendland, *Scattered Data Approximation* (Cambridge University Press, Cambridge, 2005).
[29] Q. Ye, Optimal designs of positive definite kernels for scattered data approximation, *Appl. Comput. Harmonic. Anal.* **41** (2016) 214-236.
[30] Q. Ye, Kernel-based approximation methods for partial differential equations: deterministic or stochastic problems? in *Approximation Theory XV: San Antonio 2016*, eds. G. E. Fasshauer and L. L. Schumaker, (Springer-Verlg, New York, 2017), pp. 145-166.
[31] Q. Ye, Kernel-based approximation methods for generalized interpolations: a deterministic or stochastic problem? submitted, pp. 1-30.