The Annals of Statistics
2011, Vol. 39, No. 6, 2974–3002
DOI: 10.1214/11-AOS929

ACCURATE EMULATORS FOR LARGE-SCALE COMPUTER EXPERIMENTS

BY BEN HAALAND[1] AND PETER Z. G. QIAN[2]

Duke-NUS Graduate Medical School and National University of Singapore, and University of Wisconsin, Madison

Large-scale computer experiments are becoming increasingly important in science. A multi-step procedure is introduced to statisticians for modeling such experiments, which builds an accurate interpolator in multiple steps. In practice, the procedure shows substantial improvements in overall accuracy, but its theoretical properties are not well established. We introduce the terms nominal and numeric error and decompose the overall error of an interpolator into nominal and numeric portions. Bounds on the numeric and nominal error are developed to show theoretically that substantial gains in overall accuracy can be attained with the multi-step approach.

1. Introduction. Computer experiments use complex mathematical models implemented in large computer codes to study real systems. In many situations, a physical experiment is not feasible because it is unethical, impossible, inconvenient or too expensive. A mathematical model of the system can often be developed and input/output pairs can be produced with the help of computers. Typically, the input/output pairs are expensive in the sense that they require a great deal of time and computing to obtain and they are nearly deterministic in the sense that a particular input will produce almost the same output if given to the computer experiment on another occasion. Computer experiments are widely used in systems biology, engineering design, computational biochemistry, climatology and epidemiology and their pervasiveness in science, engineering and medicine is only growing. When using a computer experiment to study a real system, a thorough exploration of the surface is typically wanted. However, obtaining input/output pairs is often too expensive for a complete exploration. A solution is to evaluate the computer experiment at several well-distributed data sites given by a *space-filling design* [9, 21, 32, 34, 37, 38, 55]. Then build an interpolator which can be used as

Received May 2011; revised September 2011.

[1]Supported by Singapore National Medical Research Council Grant IRG10nov006 and an NUS Initiative to Improve Health in Asia, Global Asia Institute grant.

[2]Supported by NSF Grant CMMI 0969616, NSF Career Award DMS-10-55214 and an IBM Faculty Award.

MSC2010 subject classifications. Primary 41A17; secondary 65M12, 65G50.

Key words and phrases. Computer experiment, emulation, interpolation, Gaussian process, large-scale problem, multi-step procedure, numerical technique, radial basis function, reproducing kernel Hilbert space.

a stand-in, or emulator, for the actual computer experiment. The thorough exploration of the complex surface can then be carried out on the interpolator. Excellent overviews on data collection and modeling for computer experiments can be found in [6, 8, 23, 48–50].

To emulate the output from a computer experiment, Gaussian process (GP) models or reproducing kernel Hilbert space (RKHS) interpolators are often used. These interpolators have a simple form and control the smoothness of the emulator. In particular, let f denote the output of a run of the computer experiment, so that the functional link between input x and output y is $y = f(x)$. Take $\Phi : \Omega \times \Omega \to \mathbb{R}$ to be symmetric in its two arguments and positive definite. The kernel Φ is *positive definite* on a domain of interest Ω if

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \Phi(x_i, x_j) > 0$$

for every nonzero $\alpha \in \mathbb{R}^n$ and distinct $\{x_1, \ldots, x_n\} \subseteq \Omega$. Then, given distinct input sites $X = \{x_1, \ldots, x_n\}$, the GP or RKHS interpolator has the simple form

$$\mathcal{P}(x) = \sum_{i=1}^{n} \alpha_i \Phi(x, x_i),$$

where α has $A_X \alpha = f|_X$, $A_X = \{\Phi(x_i, x_j)\}$ and $f|_X = (f(x_1) \cdots f(x_n))'$. Associated with each symmetric, positive definite kernel is exactly one Hilbert space of functions whose norm, in the case that the kernel is smooth, measures both size and smoothness. For a particular kernel Φ, this associated function space will be called its *native space* and will be denoted $\mathcal{N}_\Phi(\Omega)$. Native spaces will be discussed further in Section 5. The smoothness of the emulator is controlled in the sense that the RKHS interpolator has the smallest possible native space norm of any function interpolating $f|_X$ [10, 59, 60]. It is worth noting that the GP models often used in practice to build emulators for computer experiments are essentially a special case of RKHS emulators. In the GP context, the kernel Φ is a, possibly scaled, correlation function. In the case that a nonzero mean function $\hat{\mu}$ is estimated in the GP model, the interpolator is actually the sum of this estimated mean function and an RKHS interpolator of the residual $(f - \hat{\mu})|_X$. Here, we consider *translation invariant*, or stationary, kernels so that Φ is a function of only the difference between its arguments. Hereafter, $\Phi(x, y)$ will be written as $\Phi(x - y)$. Note that the connection between Gaussian processes and RKHS was also discussed in [59].

Many of the systems which scientists, engineers and medical researchers use computer experiments to study exhibit extremely complex behavior in portions of the input space. To discover and understand these regions requires a large-scale computer experiment with many input sites which are potentially very near one another. Unfortunately, most methods for building emulators, including RKHS and GP interpolators, suffer from increasingly poor predictive accuracy due to numerical problems as the number of observations of the computer experiment becomes

larger. Throughout, we refer to large-scale computer experiments as those with a large number of runs. Such experiments appear frequently in various fields such as aerospace engineering [4], information technology [20], biology, high-energy physics, nanotechnology and security. The essential difficulty in emulation of a large-scale computer experiment is that as input sites become nearer to one another the problem of finding an interpolator becomes ill-conditioned and so less amenable to accurate calculation. Several techniques exist for numerically stabilizing kernel-based interpolators, including adding a nugget effect [27, 50], using compactly supported kernels [10, 13], covariance tapering [22], decomposing the correlation matrix [5] and approximating likelihoods [53]. The multi-step procedure [12] described below also addresses the vital issue of numerical stability and can be used alone or in concert with additional numeric measures such as those mentioned above.

The multi-step procedure is not new to the field of applied mathematics, yet the exposure of statisticians to this method is relatively limited. Further, while the procedure often improves overall predictive accuracy substantially in practice, minimal work has been done on its theoretical properties [10]. Notable exceptions include [33], [11] and [16]. The existing theoretical work in the literature examines numerical accuracy in a relatively qualitative manner. Here, we introduce the concepts of *nominal* and *numeric* accuracy. Nominal accuracy refers to the accuracy which would be attained if computations could be performed without floating point rounding. Numeric accuracy refers to how close computed quantities are to their corresponding nominal counterparts. Then, we introduce a decomposition of the error of an interpolator into nominal and numeric portions. This gives a complete description of the computed interpolator's error while separating the contributing sources of error to allow for more straight-forward analysis. Bounds on the numeric and nominal error of the multi-step interpolator are developed. The numeric bound is the only complete, rigorous bound on the numeric error of the multi-step interpolator. The result is very general and makes very few assumptions about the kernels used in different steps. The nominal bound is similar to the error bound developed in [33], but more general in that it allows the kernels at different stages to be re-scaled in a flexible manner. In practice, the kernel re-scalings can have a large impact on accuracy.

2. Multi-step interpolator. The multi-step procedure explored here is a generalization of the procedure introduced in [12]. Their idea was to form well-spread nested subsets of the data. Then interpolate the first subset using a wide kernel and form residuals of this interpolator on the next subset. The residuals are then interpolated using a narrower kernel and the current stage, and previous stage interpolators are added together, giving an interpolator on the larger subset. This procedure is repeated an appropriate number of times, at each stage updating the interpolator, until an interpolator of the complete data is obtained. We introduce a separation of the error into nominal and numeric portions and derive bounds on

each type of error. We adopt a slightly different notation than [12]. Let f denote the unknown function to be interpolated and $\Omega \subseteq \mathbb{R}^d$ denote the domain of interest. Throughout, the following assumption is made about the kernel Φ.

ASSUMPTION 1. The kernel Φ is continuous, positive definite and translation invariant.

Note that with minor modifications, the development and results in Sections 1–4.3 only require that Φ is positive definite.

In the below description of the multi-step interpolation procedure, J denotes the number of stages, and Φ_j denotes the kernel used for interpolation in stage j. Now, take

$$X_1 \subset \cdots \subset X_J = X \tag{1}$$

and initialize $\mathcal{P}^0 \equiv 0$. Then, for $j = 1, \ldots, J$, let

$$\begin{aligned} \mathcal{P}^j(x) &= \sum_{u=1}^{n_j} \alpha_u^j \Phi_j(x - x_u), \\ \alpha^j &= A_{X_j,\Phi_j}^{-1} \left(f - \sum_{k=0}^{j-1} \mathcal{P}^k \right) \Bigg|_{X_j}, \\ A_{X_j,\Phi_j} &= \{\Phi_j(x_u - x_v)\}, \qquad u, v = 1, \ldots, n_j, \\ n_j &= \operatorname{card} X_j. \end{aligned} \tag{2}$$

Then the multi-step interpolator,

$$\mathcal{P}(x) = \sum_{j=1}^{J} \mathcal{P}^j(x) \tag{3}$$

satisfies the interpolation conditions $\mathcal{P}(x_u) = f(x_u)$, $u = 1, \ldots, n$, where $n = \operatorname{card} X$. Here, X is the complete set of input sites. The results in this article indicate that the best performance will be achieved if each of the nested designs, $X_1, \ldots, X_J$, are chosen to have well-separated data sites, uniform low-dimensional projections and small data-free regions. Note that α^j should not be calculated using the formula $A_{X_j,\Phi_j}^{-1}(f - \sum_{k=0}^{j-1} \mathcal{P}^k)|_{X_j}$, but instead as the solution to the linear system $A_{X_j,\Phi_j}\alpha^j = (f - \sum_{k=0}^{j-1} \mathcal{P}^k)|_{X_j}$. In general, the solution to the linear system is subject to smaller numeric error. Also, in the situation where n is large and A_{X_j,Φ_j} is sparse due to memory constraints, A_{X_j,Φ_j}^{-1} will often be too dense to be stored.

It is commonly the situation that each kernel Φ_j depends on parameters Θ_j. For example, in Section 5 it is assumed that Φ_j is a known kernel Ψ_j whose inputs $x - y$ are re-scaled by a matrix Θ_j, so that $\Phi_j(x - y) = \Psi_j(\Theta_j(x - y))$.

The form of the underlying kernels Φ_j is often fixed in advance to achieve an interpolator with prespecified smoothness and numerical properties. In particular, the results in Sections 4 and 5 indicate that smoother underlying kernels have better nominal properties and worse numeric properties, as defined in (8), and vice versa. The accuracy of the interpolator can depend significantly on the choice of parameter values. A few possible criteria for choosing the parameters Θ_j are cross-validation, maximum likelihood and sparsity of the interpolation matrices. Most procedures for choosing the Θ_j are simplified by considering each stage sequentially. In particular, Θ_j can be chosen to minimize the cross-validation error, maximize the likelihood or restrain the number of nonzero entries in the interpolation matrix A_{X_j,Φ_j} at stage j. For smaller problems, where a dense $A^{-1}_{X_j,\Phi_j}$ can be stored, the short-cut formula in (34) can be used to make leave-one-out cross-validation computationally efficient. For larger problems, an option such as 10-fold cross-validation is more appropriate. If the residuals from the previous stage $(f - \sum_{k=0}^{j-1} \mathcal{P}^k)|_{X_j}$ are modeled as a GP, then maximum likelihood can be used to choose the parameters Θ_j. Maximizing the likelihood at each stage is equivalent to minimizing

$$n_j \log\left[\frac{1}{n_j}\left(f - \sum_{k=0}^{j-1} \mathcal{P}^k\right)' \Big|_{X_j} \alpha^j\right] + \log\det(A_{X_j,\Phi_j}). \tag{4}$$

Restricted maximum likelihood estimates can be obtained by replacing the n_j in the objective function (4) by $n_j - n_{j-1}$, with $n_0 = 0$. For large problems, a storage and computation efficient algorithm such as [2] should be used in calculating $\log\det(A_{X_j,\Phi_j})$. For very large problems, memory constraints demand that the sparsity of A_{X_j,Φ_j} be considered. One possibility for compactly supported kernels is to choose *fixed* Θ_j to ensure that the number of nonzero entries in A_{X_j,Φ_j} is manageable as in (35). Another possibility is to incorporate a penalty for nonsparsity into the objective function such as (4).

If the error at stage j, $f - \sum_{k=0}^{j-1} \mathcal{P}^k$, is modeled as a GP, then confidence intervals on the function's values $f(x)$ can be obtained in much the same manner as a single stage interpolator [58]. In particular, model the output as

$$f(x) = \sum_{j=1}^{J} Z_j(x),$$

where the Z_j are mean zero Gaussian processes with $\mathrm{Cov}(Z_j(x_1), Z_j(x_2)) = \sigma_j^2 \Phi_j(x_1 - x_2)$. Note that the Z_j are *not* independent. For point sets X and Y, denote the $\mathrm{card}\, X \times \mathrm{card}\, Y$ matrix of pairwise kernel evaluations of points in X and Y as

$$\Phi(X - Y) = \{\Phi(x_u - y_v)\}, \tag{5}$$

where $x_u \in X$, $y_v \in Y$. Take $Z_0 \equiv 0$ to simplify the development below. Conditional on $f|_{X_J}, Z_1, \ldots, Z_{J-1}$,

$$f(x) - \sum_{j=0}^{J-1} Z_j(x) \sim \mathcal{N}\Bigg(\Phi_J(X_J - x)' A_{X_J,\Phi_J}^{-1} \Bigg(f - \sum_{j=0}^{J-1} Z_j\Bigg)\Bigg|_{X_J}, \sigma_J^2\big(\Phi_J(0) - \Phi_J(X_J - x)' A_{X_J,\Phi_J}^{-1} \Phi(X_J - x)\big)\Bigg)$$

(6)

$$\Longrightarrow \quad f(x) \sim \mathcal{N}\Bigg(\Phi_J(X_J - x)' A_{X_J,\Phi_J}^{-1} \Bigg(f - \sum_{j=0}^{J-1} Z_j\Bigg)\Bigg|_{X_J} + \sum_{j=0}^{J-1} Z_j(x), \sigma_J^2\big(\Phi_J(0) - \Phi_J(X_J - x)' A_{X_J,\Phi_J}^{-1} \Phi(X_J - x)\big)\Bigg).$$

Let $\tilde{X}_J = \{X_J, x\}$. Then, conditional on $f|_{X_J}, Z_1, \ldots, Z_{j-1}$

$$Z_j|_{\tilde{X}_J} \sim \mathcal{N}\Bigg(\Phi_j(X_j - \tilde{X}_J)' A_{X_j,\Phi_j}^{-1} \Bigg(f - \sum_{k=0}^{j-1} Z_k\Bigg)\Bigg|_{X_j}, \sigma_j^2\big(\Phi_j(\tilde{X}_J - \tilde{X}_J) - \Phi_j(X_j - \tilde{X}_J)' A_{X_j,\Phi_j}^{-1} \Phi_j(X_j - \tilde{X}_J)\big)\Bigg).$$

(7)

Note that the distribution in (7) is singular and $\Phi_j(\tilde{X}_J - \tilde{X}_J) = A_{\tilde{X}_J,\Phi_j}$ in the notation of (2). The first n_j components of these conditional distributions are trivial and given by

$$Z_j|_{X_j} = \Bigg(f - \sum_{k=0}^{j-1} Z_k\Bigg)\Bigg|_{X_j}, \qquad j = 1, \ldots, J.$$

The remaining $n_J - n_j + 1$ components have the nontrivial distribution, conditional on $f|_{X_J}, Z_1, \ldots, Z_{j-1}$, given by

$$Z_j|_{\tilde{X}_J \setminus X_j} \sim \mathcal{N}\Bigg(\Phi_j(X_j - \tilde{X}_J \setminus X_j)' A_{X_j,\Phi_j}^{-1} \Bigg(f - \sum_{k=0}^{j-1} Z_k\Bigg)\Bigg|_{X_j}, \sigma_j^2\big(\Phi_j(\tilde{X}_J \setminus X_j - \tilde{X}_J \setminus X_j) - \Phi_j(X_j - \tilde{X}_J \setminus X_j)' A_{X_j,\Phi_j}^{-1} \Phi_j(X_j - \tilde{X}_J \setminus X_j)\big)\Bigg).$$

After estimates of the σ_j^2 and any parameters in the Φ_j have been plugged in, the results in (6) and (7) can be combined to obtain a Gaussian estimated predictive distribution for $f(x)$ conditional on $f|_{X_J}$ with mean given by (3). For generating confidence intervals, the variance of the estimated predictive distribution,

conditional on $f|_{X_J}$, can be calculated in a backwards recursive manner using (6) and (7). Once again, note that $A^{-1}b$ should be taken as shorthand for the solution to the linear system $Ax = b$.

3. Nominal and numeric error. Now, we develop some intuition for why the multi-step procedure can improve accuracy in many situations in practice. First, computed quantities, which are subject to floating point error, are distinguished from the idealized quantities that could be obtained if a computer performed calculations with full accuracy. Hereafter, computed quantities will be distinguished with a tilde, such as $\tilde{y}$. We introduce the following separation of error into *nominal* and *numeric* portions:

$$\begin{aligned} |f(x) - \tilde{\mathcal{P}}(x)| &= |f(x) - \mathcal{P}(x) + \mathcal{P}(x) - \tilde{\mathcal{P}}(x)| \\ &\leq |f(x) - \mathcal{P}(x)| + |\mathcal{P}(x) - \tilde{\mathcal{P}}(x)|. \end{aligned} \tag{8}$$

Note that the absolute values in inequality (8) can be replaced with the norm of one's choosing. It is necessary to account for both nominal and numeric error since the trade-off between the two is very important. In most situations, reducing one will increase the other. The following proposition shows that the native space norm of the nominal error is always reduced by the addition of new data sites. Throughout, let $\mathcal{N}_\Phi(\Omega)$ denote the reproducing kernel Hilbert space corresponding to the positive definite kernel Φ, and let $\|\cdot\|_{\mathcal{N}_\Phi(\Omega)}$ denote the norm on that space [1].

PROPOSITION 3.1. *If $f \in \mathcal{N}_\Phi(\Omega)$ and $X_1 \subseteq X_2$, then*

$$\|f - \mathcal{P}_2\|_{\mathcal{N}_\Phi(\Omega)} \leq \|f - \mathcal{P}_1\|_{\mathcal{N}_\Phi(\Omega)},$$

where $\mathcal{P}_1$ and $\mathcal{P}_2$ denote the single-stage interpolators on the sets X_1 and X_2, respectively.

PROOF. It can be shown that the interpolator is orthogonal to its error with respect to the native space inner product. This implies that the result holds if and only if

$$\begin{aligned} &\|f\|^2_{\mathcal{N}_\Phi(\Omega)} - \|\mathcal{P}_2\|^2_{\mathcal{N}_\Phi(\Omega)} \leq \|f\|^2_{\mathcal{N}_\Phi(\Omega)} - \|\mathcal{P}_1\|^2_{\mathcal{N}_\Phi(\Omega)} \\ &\iff \quad \|\mathcal{P}_2\|^2_{\mathcal{N}_\Phi(\Omega)} \geq \|\mathcal{P}_1\|^2_{\mathcal{N}_\Phi(\Omega)} \\ &\iff \quad f|'_{X_2} A^{-1}_{X_2,\Phi} f|_{X_2} \geq f|'_{X_1} A^{-1}_{X_1,\Phi} f|_{X_1}, \end{aligned}$$

where the last equivalent condition follows from the definition of the native space norm and the fact that $\alpha^j = A^{-1}_{X_j,\Phi} f|_{X_j}$ for $A_{X_j,\Phi} = \{\Phi(x_u - x_v)\}$, $x_u, x_v \in X_j$, $j = 1, 2$. Then, write the interpolation matrix $A_{X_2,\Phi}$ as

$$A_{X_2,\Phi} = \begin{pmatrix} A_{X_1,\Phi} & A_{12} \\ A_{21} & A_{22} \end{pmatrix},$$

where $A_{12} = \Phi(X_1 - X_2 \setminus X_1)$, $A_{21} = \Phi(X_2 \setminus X_1 - X_1)$, and $A_{22} = \Phi(X_2 \setminus X_1 - X_2 \setminus X_1)$, using the notation in (5). Using partitioned matrix inverse and binomial inverse results [18], it can be shown that

$$\begin{aligned} &f|'_{X_2} A^{-1}_{X_2,\Phi} f|_{X_2} \\ &\quad = f|'_{X_1} A^{-1}_{X_1,\Phi} f|_{X_1} \\ &\qquad + (f|_{X_2\setminus X_1} - A_{21} A^{-1}_{X_1,\Phi} f|_{X_1})' A^{-1}_{22\cdot 1} (f|_{X_2\setminus X_1} - A_{21} A^{-1}_{X_1,\Phi} f|_{X_1}), \end{aligned}$$

where $A_{22\cdot 1} = A_{22} - A_{21} A^{-1}_{X_1,\Phi} A_{12}$. Since $A^{-1}_{22\cdot 1}$ is a block on the diagonal of $A^{-1}_{X_2,\Phi}$, it must be positive definite and the result follows. □

On the other hand, the numeric error can become arbitrarily large by the addition of new data sites. Throughout, let $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ denote the maximum and minimum eigenvalues, respectively, of a positive definite matrix A. Note that $\lambda_{\min}(A_{X,\Phi}) \to 0$ as $\min_{x_u \neq x_v} \|x_u - x_v\|_2 \to 0$. Therefore, $\lambda_{\max}(A^{-1}_{X,\Phi}) \to \infty$ as $\min_{x_u \neq x_v} \|x_u - x_v\|_2 \to 0$. An unboundedly large maximum eigenvalue of $A^{-1}_{X,\Phi}$ can enormously amplify small errors in the function and kernel evaluations. Consider the numeric error of the interpolator at a new point x,

$$\begin{aligned} \mathcal{P}(x) - \tilde{\mathcal{P}}(x) &= \sum_{i=1}^{n} [\alpha_i \Phi(x - x_i) - \tilde{\alpha}_i \tilde{\Phi}(x - x_i)] \\ &= \sum_{i=1}^{n} \big[(\alpha_i - \tilde{\alpha}_i)\Phi(x - x_i) - \tilde{\alpha}_i \big(\tilde{\Phi}(x - x_i) - \Phi(x - x_i)\big)\big]. \end{aligned}$$

Let $\varepsilon^{\alpha} = \alpha - \tilde{\alpha}$ and $\varepsilon^{\Phi} = \tilde{\Phi}(X - x) - \Phi(X - x)$ using the notation in (5). Then

$$\begin{aligned} \mathcal{P}(x) - \tilde{\mathcal{P}}(x) &= \sum_{i=1}^{n} [\varepsilon_i^{\alpha} \Phi(x - x_i) - (\alpha_i - \varepsilon_i^{\alpha}) \varepsilon_i^{\Phi}] \\ &= \sum_{i=1}^{n} [\varepsilon_i^{\alpha} \Phi(x - x_i) - \alpha_i \varepsilon_i^{\Phi} + \varepsilon_i^{\alpha} \varepsilon_i^{\Phi}]. \end{aligned}$$

So,

$$|\mathcal{P}(x) - \tilde{\mathcal{P}}(x)| \geq |f|'_X A^{-1}_{X,\Phi} \varepsilon^{\Phi}| - \|\varepsilon^{\alpha}\|_2 \|\Phi(x - X)\|_2 - \|\varepsilon^{\alpha}\|_2 \|\varepsilon^{\Phi}\|_2 \tag{9}$$

since $A_{X,\Phi}\alpha = f|_X$. If, for example, ε^{Φ} is proportional to the eigenvector corresponding to $\lambda_{\min}(A_{X,\Phi})$, and $f|_X$ is not orthogonal to ε^{Φ}, then the right-hand side of (9) can be made unboundedly large by taking $\lambda_{\min}(A_{X,\Phi}) \to 0$.

This phenomenon can be illustrated by attempting to build an interpolator for the function

$$f(x) = \exp\{(x + 1/2)^2\} \sin\big(\exp\{(x + 1/2)^2\}\big)$$

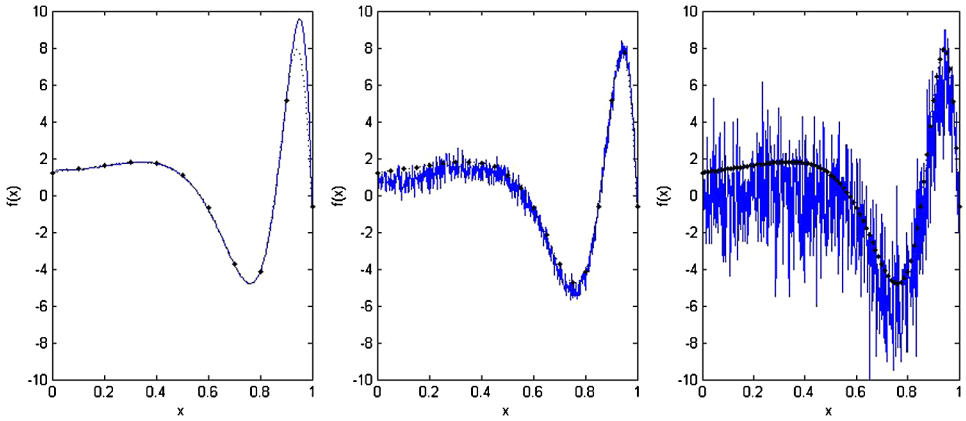

FIG. 1. *Panels* 1–3: *interpolator in solid blue and actual function in dotted black with collected data indicated by black dots.*

shown in Figure 1 using the Gaussian kernel

$$\Phi(x - y) = \exp\{-(x - y)^2\}.$$

Interpolators, shown in blue, are built on 11, 21 and 81 evenly spaced data points, shown in black dots, in the respective panels of Figure 1. As the density of points increases, so does the numeric error.

Suppose that one is in the situation where most of the data sites are well spread, but a few poorly separated data sites are causing small numeric errors to be amplified. Consider forming an interpolator in two stages. In the first stage, remove the data sites which are causing the ill-conditioning of the interpolation matrix and interpolate the remaining points with a relatively wide kernel. The nominal error will be only slightly larger than the error for the full data set, since the removed data sites were nearly equal to data sites which were included. However, the numeric error will be substantially less than that of an interpolator formed on the full data set. In the second stage, interpolate the residuals from the first-stage interpolator using a kernel which is narrow enough that numeric errors remain small. The second-stage interpolator will increase neither the nominal accuracy nor the numeric error substantially. When the two interpolators are added together to form the multi-step interpolator, the nominal accuracy may be slightly worse, but the numeric accuracy will be very much better.

For example, consider building an emulator for the Michalewicz function

$$f(x, y) = \sin(\pi x)\sin^{20}(\pi x^2) + \sin(\pi y)\sin^{20}(2\pi y^2)$$

using the third 925 point data set in Figure 2 with separation distance 5×10^{-11}. The separation distance of a point set X is half the distance between the closest two points,

$$q_X = \frac{1}{2} \min_{x_i, x_j \in X} \|x_i - x_j\|_2. \tag{10}$$

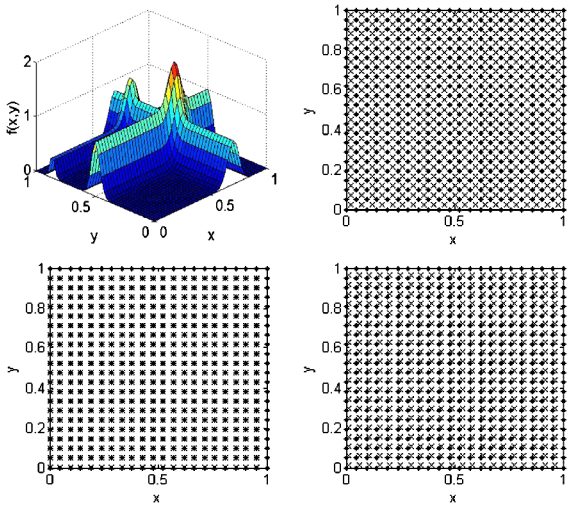

FIG. 2. *Panel* 1: *the Michalewicz function*. *Panels* 2–4 *in clockwise order*: 925 *point data sets with separation distances* 0.017, 0.009 *and* 5×10^{-11}, *respectively.*

Clearly, the $\times$'s do not contribute much information about the unknown surface. If an ordinary Gaussian kernel interpolator, corresponding to a single stage with

$$\Phi(x-y) = \exp\left\{-\sum_{j=1}^{2} \theta_j (x_j - y_j)^2\right\} \tag{11}$$

is built using all the data sites, the best possible mean squared prediction error over values of θ_1, θ_2 is ≈ 0.15, the square of the function's L_2 norm. This is because the kernel must be very narrow, or the interpolation matrix will be nearly singular. Throughout, the term *mean squared prediction error* is taken to be the average prediction error over the input domain. If, on the other hand, the $\cdot$'s are interpolated and then the residuals on the $\times$'s are interpolated, corresponding to two stages, the best possible mean squared prediction error over values of θ_1, θ_2 at each stage is $\approx 1.5 \times 10^{-5}$.

4. Numeric accuracy. The numeric accuracy of the multi-step interpolation procedure depends on the accuracy of floating point matrix manipulations. Floating point accuracy refers to the fact that computers do not perform calculations with real numbers, but instead with rounded versions thereof. For example, a typical computer has 15 digits of accuracy meaning that

$$\frac{\|\tilde{x} - x\|_2}{\|x\|_2} \leq 10^{-15},$$

where x denotes the actual value, and $\tilde{x}$ denotes the value that the computer stores.

4.1. *Numeric accuracy of matrix inversion.* The following lemma on the accuracy of floating point matrix inversion is a combination and generalization of results in [14].

DEFINITION 2. The matrix 2-norm $\|\cdot\|_2$ is defined as $\|A\|_2 = \sqrt{\lambda_{\max}(A'A)}$.

LEMMA 1. *Suppose* $Ax = b$ *and* $\tilde{A}\tilde{x} = \tilde{b}$ *with* $\|A - \tilde{A}\|_2 \le \delta_A\|A\|_2$, $\|b - \tilde{b}\|_2 \le \delta_b\|b\|_2$ *and* $\kappa(A) = r/\delta_A < 1/\delta_A$ *for some* $\delta_A, \delta_b > 0$. *Then* $\tilde{A}$ *is nonsingular,*

$$\begin{aligned} \frac{\|\tilde{x}\|_2}{\|x\|_2} &\le \frac{1 + r(\delta_b/\delta_A)}{1 - r}, \\ \frac{\|x - \tilde{x}\|_2}{\|x\|_2} &\le \frac{\delta_A + \delta_b}{1 - r}\kappa(A), \end{aligned} \tag{12}$$

where $\kappa(A) = \|A\|_2\|A^{-1}\|_2$.

PROOF. Suppose $\tilde{A}$ is singular. Then there is a $y \neq 0$ with $\tilde{A}y = 0$ so $(I - A^{-1}\tilde{A})y = y$. This implies $\|I - A^{-1}\tilde{A}\|_2 \ge 1$. On the other hand, the conditions $\|A - \tilde{A}\|_2 \le \delta_A\|A\|_2$ and $\kappa(A) < 1/\delta_A$ imply $\|I - A^{-1}\tilde{A}\|_2 < 1$ giving a contradiction.

Now, $\tilde{A}\tilde{x} = \tilde{b}$ implies $A^{-1}\tilde{A}\tilde{x} = A^{-1}(b - (b - \tilde{b})) = x + A^{-1}(\tilde{b} - b)$. The condition $\|I - A^{-1}\tilde{A}\|_2 \le r$ implies $\|A^{-1}\tilde{A}\|_2 \ge 1 - r$ and in turn

$$\begin{aligned} \|\tilde{x}\|_2 &\le \frac{1}{1 - r}(\|x\|_2 + \|A^{-1}\|_2\|\tilde{b} - b\|_2) \\ &\le \frac{1}{1 - r}(\|x\|_2 + \delta_b\|A^{-1}\|_2\|b\|_2) \\ &\le \frac{1}{1 - r}\left(\|x\|_2 + r\frac{\delta_b\|b\|_2}{\delta_A\|A\|_2}\right) \\ &\le \frac{1}{1 - r}\big(\|x\|_2 + r(\delta_b/\delta_A)\|x\|_2\big), \end{aligned}$$

where the first inequality follows from the stated condition, the triangle inequality, and the fact that $\|By\|_2 \le \|B\|_2\|y\|_2$, the second inequality follows from the condition $\|b - \tilde{b}\|_2 \le \delta_b\|b\|_2$, the third inequality follows from the condition $\kappa(A) = r/\delta_A$ and the final inequality follows from $\|b\|_2 \le \|A\|_2\|x\|_2$. Dividing by $\|x\|_2$ gives the first inequality in (12).

Note that $A(\tilde{x} - x) = \tilde{b} - b - (\tilde{A} - A)\tilde{x}$. So,

$$\begin{aligned} \|\tilde{x} - x\|_2 &\le \|A^{-1}\|_2\|\tilde{b} - b\|_2 + \|A^{-1}\|_2\|\tilde{A} - A\|_2\|\tilde{x}\|_2 \\ &\le \delta_b\|A^{-1}\|_2\|b\|_2 + \delta_A\|A^{-1}\|_2\|A\|_2\|\tilde{x}\|_2 \end{aligned}$$

$$\leq \delta_b \kappa(A) \frac{\|b\|_2}{\|A\|_2} + \delta_A \kappa(A) \|\tilde{x}\|_2$$
$$\leq \kappa(A) \|x\|_2 \left(\delta_b + \delta_A \frac{1 + r(\delta_b/\delta_A)}{1 - r} \right),$$

where the first inequality follows from the triangle inequality and the fact that $\|By\|_2 \leq \|B\|_2 \|y\|_2$, the second inequality follows from the conditions $\|b - \tilde{b}\|_2 \leq \delta_b \|b\|_2$ and $\|A - \tilde{A}\|_2 \leq \delta_A \|A\|_2$, the third inequality follows from the definition of $\kappa(A)$ and the final inequality follows from the fact that $\|b\|_2 \leq \|A\|_2 \|x\|_2$ and the first inequality in (12). Dividing by $\|x\|_2$ and simplifying gives the second part of (12). □

4.2. *Numeric accuracy of single-stage interpolator.* The above lemma can be used to bound the numeric error of an interpolator as follows.

THEOREM 4.1. *Suppose that* $\|A_{X,\Phi} - \tilde{A}_{X,\Phi}\|_2 \leq \delta_A \|A_{X,\Phi}\|_2$, $\|f|_X - \tilde{f}|_X\|_2 \leq \delta_f \|f|_X\|_2$, $\kappa(A_{X,\Phi}) = r/\delta_A < 1/\delta_A$ *and* $\sup_{x,y\in\Omega} |\Phi(x - y) - \tilde{\Phi}(x - y)| < D\delta_A$ *for some* $\delta_A, \delta_f, D > 0$, *then*

$$|\mathcal{P}(x) - \tilde{\mathcal{P}}(x)| \leq \|f|_X/\sqrt{n}\|_2 \frac{(\delta_A + \delta_f)}{1 - r} g(X, \Phi),$$
$$g(X, \Phi) = \frac{n}{\lambda_{\min}(A_{X,\Phi})} \big(\kappa(A_{X,\Phi})\Phi(0) + D\big),$$

where $\kappa(\cdot)$ *is defined in Lemma* 1.

Note that for large n and approximately uniform X, $\|f|_X/\sqrt{n}\|_2 \approx \|f\|_{L_2(\Omega)}$, where

$$\|f\|_{L_2(\Omega)} = \sqrt{\int_\Omega f(x)^2 \, dx}.$$

Further, the assumption $\sup_{x,y\in\Omega} |\Phi(x - y) - \tilde{\Phi}(x - y)| < D\delta_A$ requires that the kernel is computed in a relatively accurate manner.

PROOF OF THEOREM 4.1. First,

$$\mathcal{P}(x) - \tilde{\mathcal{P}}(x) = \sum_{i=1}^n [\alpha_i \Phi(x - x_i) - \tilde{\alpha}_i \tilde{\Phi}(x - x_i)]$$
$$= \sum_{i=1}^n \big[(\alpha_i - \tilde{\alpha}_i)\Phi(x - x_i) - \tilde{\alpha}_i \big(\tilde{\Phi}(x - x_i) - \Phi(x - x_i)\big)\big].$$

So,

$$\begin{aligned}|\mathcal{P}(x)-\tilde{\mathcal{P}}(x)| &\\ \le \left|\sum_{i=1}^{n}(\alpha_i-\tilde{\alpha}_i)\Phi(x-x_i)\right| &\\ + \left|\sum_{i=1}^{n}\tilde{\alpha}_i\big(\tilde{\Phi}(x-x_i)-\Phi(x-x_i)\big)\right|.&\end{aligned}$$

Applying the Cauchy–Schwarz inequality to each term gives

$$\begin{aligned}|\mathcal{P}(x)-\tilde{\mathcal{P}}(x)| &\\ \le \|\alpha-\tilde{\alpha}\|_2\sqrt{\sum_{i=1}^{n}\Phi(x-x_i)^2} &\\ + \|\tilde{\alpha}\|_2\sqrt{\sum_{i=1}^{n}\big(\tilde{\Phi}(x-x_i)-\Phi(x-x_i)\big)^2}.&\end{aligned}$$

The terms under the radicals can be bounded to obtain

$$|\mathcal{P}(x)-\tilde{\mathcal{P}}(x)| \le \sqrt{n}\|\alpha-\tilde{\alpha}\|_2\Phi(0)+\sqrt{n}\|\tilde{\alpha}\|_2 D\delta_A.$$

Now, Lemma 1 can be applied to the coefficients, giving

$$\begin{aligned}|\mathcal{P}(x)-\tilde{\mathcal{P}}(x)| &\le \sqrt{n}\frac{\delta_A+\delta_f}{1-r}\kappa(A_{X,\Phi})\|\alpha\|_2\Phi(0)\\ &\quad+\sqrt{n}\frac{1+r(\delta_f/\delta_A)}{1-r}\|\alpha\|_2 D\delta_A.\end{aligned}$$

Noting that $\|\alpha\|_2 \le \|A_{X,\Phi}^{-1}\|_2\|f|_X\|_2$ and collecting terms shows that

$$\begin{aligned}|\mathcal{P}(x)-\tilde{\mathcal{P}}(x)| &\le \frac{\sqrt{n}\|A_{X,\Phi}^{-1}\|_2\|f|_X\|_2}{1-r}\\ &\quad\times\big((\delta_A+\delta_f)\kappa(A_{X,\Phi})\Phi(0)+D(\delta_A+r\delta_f)\big)\\ &\le \frac{\sqrt{n}\|A_{X,\Phi}^{-1}\|_2\|f|_X\|_2}{1-r}(\delta_A+\delta_f)\big(\kappa(A_{X,\Phi})\Phi(0)+D\big).\end{aligned}$$

Rearranging gives the result. □

4.3. *Numeric accuracy of multi-step interpolator.* The first numeric result for the multi-step interpolator follows from Theorem 4.1. Here, δ denotes the computer's floating point accuracy, typically $\delta \le 10^{-15}$.

THEOREM 4.2. *Suppose that for* $j=1,\dots,J$, $\|A_{X_j,\Phi_j}-\tilde{A}_{X_j,\Phi_j}\|_2 \le \delta_j\|A_{X_j,\Phi_j}\|_2$, $\|f|_{X_j}-\tilde{f}|_{X_j}\|_2 \le \delta\|f|_{X_j}\|_2$, $\kappa(A_{X_j,\Phi_j}) \le r/\delta_j < 1/\delta_j$ *and*

$\sup_{x,y\in\Omega}|\Phi_j(x-y)-\tilde{\Phi}_j(x-y)| < D\delta$ *for some* $\delta_j, \delta, D > 0$ *with* $\delta_j\|(f - \sum_{k=1}^{j-1}\mathcal{P}^k)|_{X_j}/\sqrt{n_j}\|_2 \le \delta\|f|_{X_j}/\sqrt{n_j}\|_2$, *then*

$$\begin{aligned}&\left|\sum_{j=1}^{J}\mathcal{P}^j(x)-\sum_{j=1}^{J}\tilde{\mathcal{P}}^j(x)\right|\\&\quad\le\delta\|f|_{X_J}/\sqrt{n_J}\|_2\left[\sum_{M=1}^{J}C^M\sum_{i\in\mathcal{S}_J(M)}\prod_{k=1}^{M}\rho(X_{i_k},X_{i_{k+1}})g(X_{i_k},\Phi_{i_k})\right],\end{aligned}\tag{13}$$

where $C = 2/(1-r)$, $\mathcal{S}_J(M) = \{i \in \mathbb{N}^{M+1} : 1 \le i_1 < \cdots < i_M \le i_{M+1} = J\}$ $\rho(X,Y) = \|f|_X/\sqrt{n_X}\|_2/\|f|_Y/\sqrt{n_Y}\|_2$, *and* g *is defined in Theorem* 4.1.

The assumption $\delta_j\|(f - \sum_{k=1}^{j-1}\mathcal{P}^k)|_{X_j}/\sqrt{n_j}\|_2 \le \delta\|f|_{X_j}/\sqrt{n_j}\|_2$ roughly requires that the nominal errors either shrink or are not much larger than the function values. In practice, combinations of functions and training data sets which do not meet this assumption are very rare.

PROOF OF THEOREM 4.2. The result can be shown using induction on the number of stages J. If $J = 1$, then the result follows immediately from Theorem 4.1. Take $J \ge 2$, and assume the result holds for $J-1$ stages. Then

$$\begin{aligned}&\left\|\left(f-\sum_{j=1}^{J-1}\mathcal{P}^j\right)\Big|_{X_J}-\left(\tilde{f}-\sum_{j=1}^{J-1}\tilde{\mathcal{P}}^j\right)\Big|_{X_J}\right\|_2\\&\quad\le\|f|_{X_J}-\tilde{f}|_{X_J}\|_2+\left\|\left(\sum_{j=1}^{J-1}\mathcal{P}^j-\sum_{j=1}^{J-1}\tilde{\mathcal{P}}^j\right)\Big|_{X_J}\right\|_2\\&\quad\le\delta\|f|_{X_J}\|_2+\sqrt{n_J}\left\|\sum_{j=1}^{J-1}\mathcal{P}^j-\sum_{j=1}^{J-1}\tilde{\mathcal{P}}^j\right\|_{L_\infty(\Omega)},\end{aligned}\tag{14}$$

where the first inequality follows from the triangle inequality, and the second inequality follows from the assumptions and by bounding the L_2 error with the maximum error. The induction hypothesis can be applied to the final term in (14) giving the bound

$$\begin{aligned}&\delta\|f|_{X_J}\|_2\\&\quad\times\left(1+\rho(X_{J-1},X_J)\sum_{M=1}^{J-1}C^M\sum_{i\in\mathcal{S}_{J-1}(M)}\prod_{k=1}^{M}\rho(X_{i_k},X_{i_{k+1}})g(X_{i_k},\Phi_{i_k})\right).\end{aligned}\tag{15}$$

In stage J, the error from the first $J-1$ stages are interpolated on X_J. After multiplying and dividing the above bound (15) by $\|(f - \sum_{j=1}^{J-1}\mathcal{P}^j)|_{X_J}\|_2$, Theorem 4.1

can be used to bound the error due to stage J. Note that the term δ_f in Theorem 4.1 is the above bound (15) divided by $\|(f - \sum_{j=1}^{J-1} \mathcal{P}^j)|_{X_J}\|_2$ and the term δ_A in Theorem 4.1 is δ_J. By assumption, δ_J is smaller than or equal to (15) divided by $\|(f - \sum_{j=1}^{J-1} \mathcal{P}^j)|_{X_J}\|_2$. Simplification and coarsening of the bound gives

$$
\begin{aligned}
&|\mathcal{P}^J(x) - \tilde{\mathcal{P}}^J(x)| \\
&\leq \delta \|f|_{X_J}/\sqrt{n_J}\|_2 \frac{2}{1-r} g(X_J, \Phi_J) \\
&\quad \times \left(1 + \rho(X_{J-1}, X_J) \sum_{M=1}^{J-1} C^M \sum_{i \in \mathcal{S}_{J-1}(M)} \prod_{k=1}^{M} \rho(X_{i_k}, X_{i_{k+1}}) g(X_{i_k}, \Phi_{i_k})\right).
\end{aligned} \tag{16}
$$

Now,

$$
\left|\sum_{j=1}^{J} \mathcal{P}^j(x) - \sum_{j=1}^{J} \tilde{\mathcal{P}}^j(x)\right| \leq \left|\sum_{j=1}^{J-1} \mathcal{P}^j(x) - \sum_{j=1}^{J-1} \tilde{\mathcal{P}}^j(x)\right| + |\mathcal{P}^J(x) - \tilde{\mathcal{P}}^J(x)|.
$$

So, the induction hypothesis can be applied again along with (16) giving

$$
\begin{aligned}
&\left|\sum_{j=1}^{J} \mathcal{P}^j(x) - \sum_{j=1}^{J} \tilde{\mathcal{P}}^j(x)\right| \\
&\quad \leq \delta \|f|_{X_J}/\sqrt{n_J}\|_2 \\
&\qquad \times \Bigg[\rho(X_{J-1}, X_J) \sum_{M=1}^{J-1} C^M \sum_{i \in \mathcal{S}_{J-1}(M)} \prod_{k=1}^{M} \rho(X_{i_k}, X_{i_{k+1}}) g(X_{i_k}, \Phi_{i_k}) \\
&\qquad\quad + C g(X_J, \Phi_J) \\
&\qquad\quad + C \rho(X_{J-1}, X_J) g(X_J, \Phi_J) \\
&\qquad\qquad \times \sum_{M=1}^{J-1} C^M \sum_{i \in \mathcal{S}_{J-1}(M)} \prod_{k=1}^{M} \rho(X_{i_k}, X_{i_{k+1}}) g(X_{i_k}, \Phi_{i_k})\Bigg].
\end{aligned} \tag{17}
$$

Note that the term in square brackets in (13) is the sum of the terms with $i_M < J$ and $i_M = J$ giving

$$
\begin{aligned}
&\sum_{M=1}^{J} C^M \sum_{i \in \mathcal{S}_J(M)} \prod_{k=1}^{M} \rho(X_{i_k}, X_{i_{k+1}}) g(X_{i_k}, \Phi_{i_k}) \\
&\quad = \sum_{M=1}^{J-1} C^M \sum_{i \in \mathcal{S}_J(M), i_M < J} \prod_{k=1}^{M} \rho(X_{i_k}, X_{i_{k+1}}) g(X_{i_k}, \Phi_{i_k}) \\
&\qquad + \sum_{M=1}^{J} C^M \sum_{i \in \mathcal{S}_J(M), i_M = J} \prod_{k=1}^{M} \rho(X_{i_k}, X_{i_{k+1}}) g(X_{i_k}, \Phi_{i_k})
\end{aligned}
$$

$$= \rho(X_{J-1}, X_J) \sum_{M=1}^{J-1} C^M \sum_{i \in \mathcal{S}_{J-1}(M)} \prod_{k=1}^{M} \rho(X_{i_k}, X_{i_{k+1}}) g(X_{i_k}, \Phi_{i_k})$$

$$+ Cg(X_j, \Phi_J) + \sum_{M=2}^{J} C^M \sum_{i \in \mathcal{S}_J(M), i_M = J} \prod_{k=1}^{M} \rho(X_{i_k}, X_{i_{k+1}}) g(X_{i_k}, \Phi_{i_k}),$$

which is exactly the term in square brackets in (17), proving the result. □

4.4. *Dependence on separation distance.* The terms

$$g(X_j, \Phi_j) = \frac{n_j}{\lambda_{\min}(A_{X_j, \Phi_j})} \big(\kappa(A_{X_j, \Phi_j}) \Phi(0) + D\big) \tag{18}$$

from Theorem 4.2 can be computed, at least approximately. However, by bounding (18) in terms of the separation distance, as defined in (10), the role of the data sites and the kernel's smoothness in the numeric accuracy are revealed. These results indicate that using poorly separated data or a wide kernel Φ with a rapidly decaying Fourier transform, implying more smoothness, has more potential to result in large numeric errors in interpolation. The Fourier transform can be defined as follows.

DEFINITION 3. For $f \in L_1(\mathbb{R}^d)$ define the Fourier transform [51]

$$\hat{f}(\omega) = (2\pi)^{-d/2} \int_{\mathbb{R}^d} f(x) e^{-i\omega' x} \, dx.$$

To generate the bound on (18), the following result from [60] can be used.

THEOREM 4.3. *Let* $\varphi_*(M, \Phi) = \inf_{\|\omega\|_2 \le 2M} \hat{\Phi}(\omega)$. *Then*

$$\lambda_{\min}(A_{X,\Phi}) \ge C_d \varphi_*(M_d/q, \Phi)/q^d,$$

$$M_d = 12\big(\pi \Gamma^2(d/2+1)/9\big)^{1/(d+1)},$$

$$C_d = (M_d/2^{3/2})^d / \big(2\Gamma(d/2+1)\big)$$

for any $q \le q_X$, *where* $A_{X,\Phi} = \{\Phi(x_i - x_j)\}$.

To bound $\lambda_{\max}(A_{X,\Phi})$ below, Gershgorin's theorem [57] can be used. Gershgorin's theorem states that the largest eigenvalue of $A_{X,\Phi}$ has

$$|\lambda_{\max}(A_{X,\Phi}) - \Phi(x_j - x_j)| \le \sum_{i=1, i \ne j}^{n} |\Phi(x_i - x_j)|.$$

Rearranging and coarsening the bound gives

$$\lambda_{\max}(A_{X,\Phi}) \le n\Phi(0). \tag{19}$$

Theorem 4.3 and inequality (19) can be combined to obtain the following theorem bounding (18).

THEOREM 4.4. *Under the assumptions in Theorem 4.2,*

$$g(X_j, \Phi_j) \le \kappa_{\text{upper}}(X_j, \Phi_j)\big(\kappa_{\text{upper}}(X_j, \Phi_j)\Phi(0) + D\big),$$

$$\kappa_{\text{upper}}(X_j, \Phi_j) = \frac{n_j q_{X_j}^d}{C_d \varphi_*(M_d/q_{X_j}, \Phi_j)}.$$

The nested sequence $X_1 \subset \cdots \subset X_J$ in (1) with large separation distance can be generated from *nested space-filling designs* [15, 41–44], which were originally developed for the purpose of conducting multi-fidelity computer experiments. Space-filling designs have shown particular merit in numerical integration [28–31, 35, 36, 39, 40, 52, 55]. Theorem 4.4 provides new insights into the use of such designs in interpolation.

5. Nominal accuracy. The results in this section indicate that the nominal error in interpolation converges to zero more quickly for wider, smoother kernels Φ, although the constant involved in this rate changes. This is in direct opposition to the numeric error, which tends to be smaller for narrower, less smooth kernels. In fact, it will be seen that convergence of the nominal error of an arbitrarily fast *rate* can be achieved with an infinitely smooth kernel, such as the Gaussian in (11).

A re-scaling is introduced in the following definition.

DEFINITION 4. For a nonsingular Θ, define $\Phi_\Theta(x) = \Phi(\Theta x)$.

5.1. *Point-wise bound.* Initially, consider a single stage with a fixed Φ which is re-scaled by a fixed Θ. For a set of input sites X of size n, define the cardinal basis functions

$$u_i(x) = \sum_{i=1}^{n} \beta_i \Phi_\Theta(x - x_j),$$

$$u_i(x_j) = \mathbb{1}_{\{i=j\}}$$

for $i, j = 1, \ldots, n$. Then

$$\mathcal{P}(x) = \sum_{i=1}^{n} f(x_i) u_i(x).$$

Since $f(x) = \langle f, \Phi_\Theta(\cdot - x)\rangle_{\mathcal{N}_{\Phi_\Theta}(\Omega)}$ if $f \in \mathcal{N}_{\Phi_\Theta}(\Omega)$,

$$f(x) - \mathcal{P}(x) = \langle f, \Phi_\Theta(\cdot - x)\rangle_{\mathcal{N}_{\Phi_\Theta}(\Omega)} - \sum_{i=1}^{n} u_i(x) \langle f, \Phi_\Theta(\cdot - x_i)\rangle_{\mathcal{N}_{\Phi_\Theta}(\Omega)}$$

$$= \left\langle f, \Phi_\Theta(\cdot - x) - \sum_{i=1}^{n} u_i(x) \Phi_\Theta(\cdot - x_i) \right\rangle_{\mathcal{N}_{\Phi_\Theta}(\Omega)}.$$

Now, the Cauchy–Schwarz inequality can be applied, giving the error bound

$$|f(x) - \mathcal{P}(x)| \le \|f\|_{\mathcal{N}_{\Phi_\Theta}(\Omega)} \left\| \Phi_\Theta(\cdot - x) - \sum_{i=1}^{n} u_i(x)\Phi_\Theta(\cdot - x_i) \right\|_{\mathcal{N}_{\Phi_\Theta}(\Omega)}. \tag{20}$$

The second term on the right-hand side of (20) is the so-called *power function*, $P_{\Phi_\Theta,X}$. It can be shown [60] that if the domain of interest Ω is bounded and convex, then

$$P^2_{\Phi_\Theta,X} \le C_1 \|\Phi_\Theta - p\|_{L_\infty(B(0,C_2 h_X))},$$

where $C_1, C_2 > 0$ are constants which may depend on Ω, p is any multivariate polynomial, $B(a,b) = \{x \in \mathbb{R}^d : \|x - a\|_2 < b\}$ and h_X denotes the *fill distance*

$$h_X = \sup_{x\in\Omega} \min_{x_u \in X} \|x - x_u\|_2.$$

Now, if Φ has k continuous derivatives, p can be taken to be the Taylor's polynomial of Φ_Θ of degree $k-1$. Then

$$\|\Phi_\Theta - p\|_{L_\infty(B(0,C_2 h_X))} \le C_3 \|\Theta\|_2^k h_X^k,$$

where C_3 is a constant which does not depend on Θ. Combining the above development gives the following.

THEOREM 5.1. *Suppose that Ω is bounded and convex, Φ satisfies Assumption 1 and has k continuous derivatives and Θ is nonsingular. Then*

$$|f(x) - \mathcal{P}(x)| \le C_\Phi \|\Theta\|_2^{k/2} h_X^{k/2} \|f\|_{\mathcal{N}_{\Phi_\Theta}(\Omega)}.$$

5.2. *Native space bound.* First, write $\Phi_\Theta * \Phi_\Theta$ as

$$\Phi_\Theta * \Phi_\Theta(x - y) = \int_\Omega \Phi_\Theta(x - t)\Phi_\Theta(y - t)\,\mathrm{d}t.$$

Then, for $f \in \mathcal{N}_{\Phi_\Theta * \Phi_\Theta}(\Omega)$ and $x \in \Omega$, express f in terms of the integral operator

$$f(x) = \int_\Omega u(y)\Phi_\Theta * \Phi_\Theta(x - y)\,\mathrm{d}y,$$

where $u \in L_2(\Omega)$. Combining these expressions gives

$$\begin{aligned} f(x) &= \int_\Omega u(y) \int_\Omega \Phi_\Theta(y - t)\Phi_\Theta(x - t)\,\mathrm{d}t\,\mathrm{d}y \\ &= \int_\Omega v(t)\Phi_\Theta(x - t)\,\mathrm{d}t, \end{aligned}$$

where $v \in L_2(\Omega)$ is given by

$$v(t) = \int_\Omega u(y)\Phi_\Theta(y - t)\,\mathrm{d}y$$

for $t \in \Omega$. Then

$$\begin{aligned} \|f - \mathcal{P}\|^2_{\mathcal{N}_{\Phi_\Theta}(\Omega)} &= \langle f - \mathcal{P}, f \rangle_{\mathcal{N}_{\Phi_\Theta}(\Omega)} \\ &= \langle f - \mathcal{P}, v \rangle_{L_2(\Omega)} \\ &\leq \|f - \mathcal{P}\|_{L_2(\Omega)} \|v\|_{L_2(\Omega)}, \end{aligned} \tag{21}$$

where the first equality follows from the orthogonality of the interpolator and its error with respect to the native space norm, the second equality follows from the properties of the integral operator and the inequality follows from the Cauchy–Schwarz inequality.

If Φ has k continuous derivatives, then the first term on the right-hand side of inequality (21) can be bounded using Theorem 5.1 as

$$\begin{aligned} \|f - \mathcal{P}\|_{L_2(\Omega)} &\leq \sqrt{\operatorname{vol} \Omega} \|f - \mathcal{P}\|_{L_\infty}(\Omega) \\ &\leq C_\Phi \|\Theta\|_2^{k/2} h_X^{k/2} \|f - \mathcal{P}\|_{\mathcal{N}_{\Phi_\Theta}(\Omega)}, \end{aligned} \tag{22}$$

where the first inequality follows by relating the $L_2(\Omega)$ and $L_\infty(\Omega)$ norms, and the second inequality follows by applying Theorem 5.1 to $f - \mathcal{P}$. Plugging inequality (22) into inequality (21) and canceling a single $\|f - \mathcal{P}\|_{\mathcal{N}_{\Phi_\Theta}(\Omega)}$ term gives

$$\|f - \mathcal{P}\|_{\mathcal{N}_{\Phi_\Theta}(\Omega)} \leq C_\Phi \|\Theta\|_2^{k/2} h_X^{k/2} \|v\|_{L_2(\Omega)}. \tag{23}$$

Using the properties of the integral operator, the square of the second term on the right-hand side of inequality (23) can be expressed as

$$\begin{aligned} \|v\|^2_{L_2(\Omega)} &= \int_{\Omega^3} u(x) u(y) \Phi_\Theta(y - t) \Phi_\Theta(x - t) \, \mathrm{d}x \, \mathrm{d}y \, \mathrm{d}t \\ &= \|f\|^2_{\mathcal{N}_{\Phi_\Theta * \Phi_\Theta}(\Omega)}. \end{aligned} \tag{24}$$

Combining inequality (23) and equality (24) gives the following theorem.

THEOREM 5.2. *Under the assumptions of Theorem 5.1,*

$$\|f - \mathcal{P}\|_{\mathcal{N}_{\Phi_\Theta}(\Omega)} \leq C_\Phi \|\Theta\|_2^{k/2} h_X^{k/2} \|f\|_{\mathcal{N}_{\Phi_\Theta * \Phi_\Theta}(\Omega)}.$$

To allow for individual re-scalings in different stages, we start with some notation. Define Ψ_k recursively as

$$\begin{aligned} \Psi^0 &= \Phi, \\ \Psi^k &= \Psi^{k-1} * \Psi^{k-1} \end{aligned} \tag{25}$$

for $k \in \mathbb{N}$. For the kernel on step j, take

$$\Phi_j = \Psi^{J-j}_{\Theta_j}. \tag{26}$$

We now develop a bound on $\|\cdot\|_{\mathcal{N}_{\Phi_j * \Phi_j}(\Omega)}$ in terms of $\|\cdot\|_{\mathcal{N}_{\Phi_{j-1}}(\Omega)}$. The basic assumptions on the re-scaling matrices Θ_j in this section are that they are nonsingular and *larger* than the Θ_{j-1} in the sense that $\lambda_{\max}(\Theta'_{j-1}\Theta_{j-1}\Xi'_j\Xi_j) \le 1$, where $\Xi'_j = \Theta_j^{-1}$.

In the case $\Omega = \mathbb{R}^d$, the native space $\mathcal{N}_{\Phi_\Theta}(\mathbb{R}^d)$ has norm defined through the inner product

$$\langle f, g\rangle_{\mathcal{N}_{\Phi_\Theta}(\mathbb{R}^d)} = (2\pi)^{-d/2}\int_{\mathbb{R}^d} \frac{\hat{f}(\omega)\overline{\hat{g}(\omega)}}{\hat{\Phi}_\Theta(\omega)}\,\mathrm{d}\omega, \tag{27}$$

where $\hat{f}$ and $\overline{\hat{g}}$ denote the Fourier transform and complex conjugate of the Fourier transform, respectively, of $f, g \in \mathcal{N}_{\Phi_\Theta}(\mathbb{R}^d)$ [60]. This explicit representation of the native space inner product can be used to relate the native space norms for convolutions and re-scalings. Hereafter, take $\infty > c_2 \ge c_1 > 0$ and $\hat{\Upsilon}$ with

$$\omega'\omega \le \nu'\nu \implies \hat{\Upsilon}(\omega) \ge \hat{\Upsilon}(\nu), \qquad c_1\hat{\Upsilon}(\omega) \le \hat{\Phi}(\omega) \le c_2\hat{\Upsilon}(\omega). \tag{28}$$

Assumption 1 ensures that c_1, c_2 and $\hat{\Psi}$ satisfying (28) exist [60]. The bounds to follow are tightest for $c_2 - c_1$ as small as possible. Essentially, we want a *radially decreasing* envelop on the Fourier transform of the underlying kernel Φ to simplify development. Note that the Fourier transforms $\hat{\Phi}$ and $\hat{\Phi}_\Theta$ are related in the following manner:

$$\begin{aligned}
\hat{\Phi}_\Theta(\omega) &= (2\pi)^{-d/2}\int_{\mathbb{R}^d} \Phi_\Theta(x)e^{-i\omega' x}\,\mathrm{d}x \\
&= (2\pi)^{-d/2}\int_{\mathbb{R}^d} \Phi(\Theta x)e^{-i\omega'\Xi'\Theta x}\,\mathrm{d}x \\
&= (2\pi)^{-d/2}|\det(\Xi)|\int_{\mathbb{R}^d} \Phi(y)e^{-i\omega'\Xi' y}\,\mathrm{d}y \\
&= |\det(\Xi)|\hat{\Phi}(\Xi\omega),
\end{aligned} \tag{29}$$

where $\Xi' = \Theta^{-1}$ and the third equality follows by making the substitution $y = \Theta x$.

PROPOSITION 5.1. *If Assumption 1 is satisfied and Θ_{j-1}, Θ_j are nonsingular with respective inverses Ξ'_{j-1}, Ξ'_j, then*

$$\begin{aligned}
&\lambda_{\max}(\Theta'_{j-1}\Theta_{j-1}\Xi'_j\Xi_j) \le 1 \\
&\implies \|f\|^2_{\mathcal{N}_{\Phi_j * \Phi_j}(\mathbb{R}^d)} \le \left(\frac{c_2}{c_1}\right)^{2^{J-(j-1)}} \frac{|\det(\Xi_{j-1})|}{|\det(\Xi_j)|^2}\|f\|^2_{\mathcal{N}_{\Phi_{j-1}}(\mathbb{R}^d)}
\end{aligned} \tag{30}$$

for $1 \le j \le J$ where c_1 and c_2 satisfy (28), and Φ_{j-1} and Φ_j satisfy relations (25) and (26).

PROOF. If $f \notin \mathcal{N}_{\Phi_{j-1}}(\mathbb{R}^d)$, then $\|f\|^2_{\mathcal{N}_{\Phi_{j-1}}(\mathbb{R}^d)} = \infty$ and (30) is true. Now, assume $f \in \mathcal{N}_{\Phi_{j-1}}(\mathbb{R}^d)$, and note that

$$\frac{\omega' \Xi_j' \Xi_j \omega}{\omega' \Xi_{j-1}' \Xi_{j-1} \omega} \le \lambda_{\max}(\Theta_{j-1}' \Theta_{j-1} \Xi_j' \Xi_j).$$

If $\lambda_{\max}(\Theta_{j-1}' \Theta_{j-1} \Xi_j' \Xi_j) \le 1$, then

$$\begin{aligned}
&\omega' \Xi_j' \Xi_j \omega \le \omega' \Xi_{j-1}' \Xi_{j-1} \omega \\
&\quad \Longrightarrow \quad \frac{1}{c_1} \hat{\Phi}(\Xi_j \omega) \ge \hat{\Upsilon}(\Xi_j \omega) \ge \hat{\Upsilon}(\Xi_{j-1} \omega) \ge \frac{1}{c_2} \hat{\Phi}(\Xi_{j-1} \omega) \\
&\quad \Longrightarrow \quad \frac{1}{\hat{\Phi}(\Xi_j \omega)^{2^{J-j}}} \le \left(\frac{c_2}{c_1}\right)^{2^{J-j}} \frac{1}{\hat{\Phi}(\Xi_{j-1} \omega)^{2^{J-j}}} \\
&\quad \Longrightarrow \quad \frac{1}{\hat{\Psi}^{J-j}(\Xi_j \omega)} \le \left(\frac{c_2}{c_1}\right)^{2^{J-j}} \frac{1}{\hat{\Psi}^{J-j}(\Xi_{j-1} \omega)},
\end{aligned} \tag{31}$$

where the first implication follows from (28), the second implication follows since the right- and left-hand sides are positive and the final implication follows from the relations (25) and (26) and the properties of Fourier transforms of convolutions. So,

$$\begin{aligned}
&\|f\|^2_{\mathcal{N}_{\Phi_{j-1}}(\mathbb{R}^d)} \\
&\quad = \|f\|^2_{\mathcal{N}_{\Psi^{J-(j-1)}_{\Theta_{j-1}}(\mathbb{R}^d)}} \\
&\quad = (2\pi)^{-d/2} \int_{\mathbb{R}^d} \frac{|\hat{f}(\omega)|^2}{\hat{\Psi}^{J-(j-1)}_{\Theta_{j-1}}(\omega)} \, \mathrm{d}\omega \\
&\quad = \frac{(2\pi)^{-d/2}}{|\det(\Xi_{j-1})|} \int_{\mathbb{R}^d} \frac{|\hat{f}(\omega)|^2}{\hat{\Psi}^{J-(j-1)}(\Xi_{j-1} \omega)} \, \mathrm{d}\omega \\
&\quad = \frac{(2\pi)^{-d/2}}{|\det(\Xi_{j-1})|} \int_{\mathbb{R}^d} \frac{|\hat{f}(\omega)|^2}{\widehat{\Psi^{J-j} * \Psi^{J-j}}(\Xi_{j-1} \omega)} \, \mathrm{d}\omega \\
&\quad = \frac{(2\pi)^{-d}}{|\det(\Xi_{j-1})|} \int_{\mathbb{R}^d} \frac{|\hat{f}(\omega)|^2}{\hat{\Psi}^{J-j}(\Xi_{j-1} \omega)^2} \, \mathrm{d}\omega \\
&\quad = (2\pi)^{-d} \frac{|\det(\Xi_j)|^2}{|\det(\Xi_{j-1})|} \int_{\mathbb{R}^d} \frac{|\hat{f}(\omega)|^2}{|\det(\Xi_j)|^2 \hat{\Psi}^{J-j}(\Xi_{j-1} \omega)^2} \, \mathrm{d}\omega \\
&\quad \ge (2\pi)^{-d} \frac{|\det(\Xi_j)|^2}{|\det(\Xi_{j-1})|} \left(\frac{c_1}{c_2}\right)^{2^{J-j+1}} \int_{\mathbb{R}^d} \frac{|\hat{f}(\omega)|^2}{|\det(\Xi_j)|^2 \hat{\Psi}^{J-j}(\Xi_j \omega)^2} \, \mathrm{d}\omega
\end{aligned}$$

$$= (2\pi)^{-d/2} \frac{|\det(\Xi_j)|^2}{|\det(\Xi_{j-1})|} \left(\frac{c_1}{c_2}\right)^{2^{J-j+1}} \int_{\mathbb{R}^d} \frac{|\hat{f}(\omega)|^2}{\widehat{\Psi_{\Theta_j}^{J-j} * \Psi_{\Theta_j}^{J-j}}(\omega)} \, \mathrm{d}\omega$$

$$= \frac{|\det(\Xi_j)|^2}{|\det(\Xi_{j-1})|} \left(\frac{c_1}{c_2}\right)^{2^{J-j+1}} \|f\|^2_{\mathcal{N}_{\Phi_j * \Phi_j}(\mathbb{R}^d)},$$

where the first equality follows from relation (26), the second equality follows from the inner product representation (27), the third equality follows from the scaled Fourier transform relation (29), the fourth equality follows from the definition of $\Psi^{J-(j-1)}$ (25), the fifth equality follows from the properties of Fourier transforms of convolutions, the sixth equality follows by multiplying by $|\det(\Xi_j)|^2/|\det(\Xi_j)|^2$, the inequality follows from the development (31), the seventh equality follows from the scaled Fourier transform relation (29) and the properties of Fourier transforms of convolutions and the final equality follows from the inner product representation (27). □

In most applications, the domain of interest Ω is a strict subset of $\mathbb{R}^d$. If $f \in \mathcal{N}_{\Phi_1}(\Omega)$, then f can be extended to $Ef \in \mathcal{N}_{\Phi_1}(\mathbb{R}^d)$ [60] with

$$\begin{aligned} \|f\|_{\mathcal{N}_{\Phi_1}(\Omega)} &= \|Ef\|_{\mathcal{N}_{\Phi_1}(\mathbb{R}^d)}, \\ \|f\|_{\mathcal{N}_{\Phi_2}(\Omega)} &\le \|Ef\|_{\mathcal{N}_{\Phi_2}(\mathbb{R}^d)} \end{aligned} \tag{32}$$

for all Φ_2. Combining (32) with Proposition 5.1 gives the following corollary.

COROLLARY 1. *If the assumptions of Proposition 5.1 are satisfied, then*

$$\begin{aligned} &\lambda_{\max}(\Theta'_{j-1}\Theta_{j-1}\Xi'_j\Xi_j) \le 1 \\ &\implies \|f\|^2_{\mathcal{N}_{\Phi_j * \Phi_j}(\Omega)} \le \left(\frac{c_2}{c_1}\right)^{2^{J-(j-1)}} \frac{|\det(\Xi_{j-1})|}{|\det(\Xi_j)|^2} \|f\|^2_{\mathcal{N}_{\Phi_{j-1}}(\Omega)}. \end{aligned} \tag{33}$$

PROOF. If $f \notin \mathcal{N}_{\Phi_{j-1}}(\Omega)$, then $\|f\|^2_{\mathcal{N}_{\Phi_{j-1}}(\Omega)} = \infty$ and (33) is true. Now, assume $f \in \mathcal{N}_{\Phi_{j-1}}(\Omega)$ and extend f to $Ef \in \mathcal{N}_{\Phi_{j-1}}(\mathbb{R}^d)$ with $\|Ef\|^2_{\mathcal{N}_{\Phi_{j-1}}(\mathbb{R}^d)} = \|f\|^2_{\mathcal{N}_{\Phi_{j-1}}(\Omega)}$. Then

$$\begin{aligned} \|f\|^2_{\mathcal{N}_{\Phi_j * \Phi_j}(\Omega)} &\le \|Ef\|^2_{\mathcal{N}_{\Phi_j * \Phi_j}(\mathbb{R}^d)} \\ &\le \left(\frac{c_2}{c_1}\right)^{2^{J-(j-1)}} \frac{|\det(\Xi_{j-1})|}{|\det(\Xi_j)|^2} \|Ef\|^2_{\mathcal{N}_{\Phi_{j-1}}(\mathbb{R}^d)} \\ &= \left(\frac{c_2}{c_1}\right)^{2^{J-(j-1)}} \frac{|\det(\Xi_{j-1})|}{|\det(\Xi_j)|^2} \|f\|^2_{\mathcal{N}_{\Phi_{j-1}}(\Omega)}, \end{aligned}$$

where the first inequality follows from (32), the second inequality follows from Proposition 5.1 and the equality follows from the property of the chosen extension. □

5.3. *Error bound for multi-step interpolator.* Combining Theorem 5.2 with Corollary 1, we are able to obtain the following theorem bounding the native space norm of the multi-step interpolator's error.

THEOREM 5.3. *Under the assumptions of Theorem 5.1 and Proposition 5.1,*

$$\left\| f - \sum_{j=1}^{J} \mathcal{P}^j \right\|_{\mathcal{N}_{\Phi_J}(\Omega)} \leq C_{\Phi,J} \|f\|_{\mathcal{N}_{\Phi_0}(\Omega)} \prod_{j=1}^{J} \left\{ \frac{\sqrt{|\det(\Xi_{j-1})|}}{|\det(\Xi_j)|} (\|\Theta_j\|_2^k h_{X_j}^k)^{2^{J-j-1}} \right\}.$$

PROOF. First applying Theorem 5.2 and then applying Proposition 5.1 gives

$$\begin{aligned} \left\| f - \sum_{j=1}^{J} \mathcal{P}^j \right\|_{\mathcal{N}_{\Phi_J}(\Omega)} &\leq C_{\Phi} \|\Theta_J\|_2^{k/2} h_{X_J}^{k/2} \left\| f - \sum_{j=1}^{J-1} \mathcal{P}^j \right\|_{\mathcal{N}_{\Phi_J * \Phi_J}(\Omega)} \\ &\leq C_{\Phi,J} \|\Theta_J\|_2^{k/2} h_{X_J}^{k/2} \frac{\sqrt{|\det(\Xi_{J-1})|}}{|\det(\Xi_J)|} \left\| f - \sum_{j=1}^{J-1} \mathcal{P}^j \right\|_{\mathcal{N}_{\Phi_{J-1}}(\Omega)}. \end{aligned}$$

For $J \geq 2$, repeat the above argument $J-1$ more times, and note that Φ_{J-j} has $k2^j$ continuous derivatives. □

By applying Theorem 5.1 to the error $f - \sum_{j=1}^{J} \mathcal{P}^j$, an additional multiple of $h_{X_J}^{k/2}$ is obtained in the following theorem.

THEOREM 5.4. *Under the assumptions of Theorem 5.1 and Proposition 5.1,*

$$\begin{aligned} &\left| f(x) - \sum_{j=1}^{J} \mathcal{P}^j(x) \right| \\ &\quad \leq C_{\Phi,J} \|f\|_{\mathcal{N}_{\Phi_0}(\Omega)} \|\Theta_J\|_2^{k/2} h_{X_J}^{k/2} \prod_{j=1}^{J} \left\{ \frac{\sqrt{|\det(\Xi_{j-1})|}}{|\det(\Xi_j)|} (\|\Theta_j\|_2^k h_{X_j}^k)^{2^{J-j-1}} \right\}. \end{aligned}$$

6. Examples. First, consider using the multi-step procedure to interpolate Franke's function

$$\begin{aligned} f(x,y) = {} & \tfrac{3}{4} \exp\{-((9x-2)^2 + (9y-2)^2)/4\} \\ & + \tfrac{3}{4} \exp\{-((9x+1)^2/49 - (9y+1)^2/10)\} \\ & + \tfrac{1}{2} \exp\{-((9x-7)^2 + (9y-3)^2)/4\} \\ & - \tfrac{1}{5} \exp\{-((9x-4)^2 + (9y-7)^2)\} \end{aligned}$$

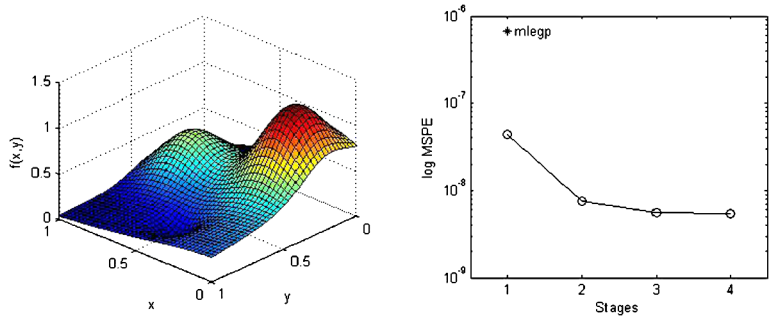

FIG. 3. *Left panel*: *Franke's function*. *Right panel*: *log mean squared prediction error versus number of stages* (*circles*) *and using* `mlegp` (*asterisk*).

shown in the left panel of Figure 3. Theorems 4.2 and 4.4 indicate that each of the nested data sets should have well-separated points in the full dimension as well as lower-dimensional projections to give small numeric error, and Theorem 5.4 indicates that each of the nested data sets should have small data-free regions in the full dimension as well as lower-dimensional projections to give small nominal error. Training data are collected from Franke's function using a randomized $(0, 4, 2)$-net in base 5 [38] with $5^4 = 625$ points, which has a convenient nested structure with both the full and each sub-design having small data-free regions and relatively well-spread points in both the full and projected space, making it ideal for the multi-step procedure. Theorem 4.4 indicates that a less smooth underlying kernel Φ will give more numerically accurate results, while Theorem 5.4 indicates that a more smooth kernel will give more nominally accurate results. To balance these opposing forces in this moderately sized example, the selected Φ is Wendland's compactly supported kernel with four continuous derivatives [60],

$$\Phi(x-y) = \phi\bigl(\sqrt{(x-y)'(x-y)}\bigr),$$

$$\phi(r) = (1-r)_+^{l+2}[(l^2+4l+3)r^2 + (3l+6)r + 3], \qquad l = \lfloor d/2 \rfloor + 3,$$

and the rescaling matrices $\Theta_1, \ldots, \Theta_J$ are restricted to be diagonal, so each input is re-scaled separately. The re-scalings for each stage are chosen by leave-one-out cross-validation, for which a simple short-cut formula holds making computation undemanding for this moderately sized problem, although $A^{-1}_{X_j,\Phi_j}$ needs to be calculated. In particular, the ith cross-validation error at stage j is [47]

$$e_{(i)} = \frac{\alpha_i^j}{B_{ii}^j}, \qquad B^j = A^{-1}_{X_j,\Phi_j}. \tag{34}$$

In this example, the single-stage sample size is $n_1 = 625$, the two-stage sample sizes are $n_1 = 250$ and $n_2 = 625$, the three-stage sample sizes are $n_1 = 250$, $n_2 = 375$ and $n_3 = 625$ and the four-stage sample sizes are $n_1 = 250$, $n_2 = 375$, $n_3 = 500$ and $n_4 = 625$. The nested data sets are $X_j = \{x_i \in X : i \le n_j\}$. The right panel

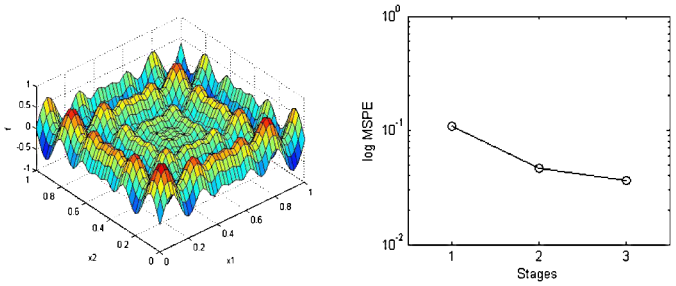

FIG. 4. *Left panel*: *two-dimensional projection of Schwefel's function*. *Right panel*: *log mean squared prediction error versus number of stages*.

of Figure 3 shows the logarithm of the mean squared prediction error on a test set of 1,000 randomly generated uniform points. Notice that the mean squared prediction error is improved from 4.4×10^{-8} to 5.4×10^{-9}. A Gaussian process fit using the `mlegp` package [7] in `R`, on the other hand, has mean squared prediction error 6.8×10^{-7}.

Next, consider using the multi-step procedure to interpolate Schwefel's function for $d = 5$

$$f(x) = -\sum_{j=1}^{d} (1{,}000x_j - 500) \sin\bigl(\sqrt{|1{,}000x_j - 500|}\bigr)/1{,}000,$$

a two-dimensional projection of which with the remaining variables fixed at $1/2$ is shown in the left panel of Figure 4. This function is relatively complex and a very large training set is needed to build an accurate emulator. To ensure easy nesting and good space-filling properties for sub-designs, data are collected from Schwefel's function using a randomized $(0, 8, 5)$-net in base 5 with $5^8 = 390{,}625$ points. In this example there is a great deal of potential for numeric problems so Wendland's continuous, compactly supported kernel,

$$\Phi(x - y) = \phi\bigl(\sqrt{(x - y)'(x - y)}\bigr),$$

$$\phi(r) = (1 - r)_+^{l+2}, \qquad l = \lfloor d/2 \rfloor + 1$$

with relatively little smoothness is selected. The re-scaling matrices $\Theta_1, \ldots, \Theta_J$ are chosen to be *fixed* scalar multiples of the identity, $\Theta_j = \theta_j I_d$, with

$$\theta_j = \left(\frac{n_j^2 \pi^{d/2}}{10^7 \Gamma(d/2 + 1)} \right)^{1/d}, \tag{35}$$

which ensures that each intepolation matrix A_{X_j, Φ_j} has less than 10^7 nonzero entries. Edge effects in the five-dimensional cube ensure that the number of nonzero entries is substantially less than 10^7. In this example, the single-stage sample size is $n_1 = 390{,}625$, the two-stage sample sizes are $n_1 = 5^7 = 78{,}125$ and $n_2 = 390{,}625$ and the three-stage sample sizes are $n_1 = 78{,}125$, $n_2 = 2 \times 5^7 = 156{,}250$

and $n_3 = 390{,}625$. The nested data sets are $X_j = \{x_i \in X : i \leq n_j\}$. The right panel of Figure 4 shows the logarithm of the mean squared prediction error on a test set of 10,000 randomly generated uniform points. Notice that the mean squared prediction error is improved from 0.11 to 0.036. On the other hand, the `mlegp` package runs out of memory trying to fit a GP.

7. Discussion. We have presented the intuitively appealing and practically useful multi-step interpolation procedure. This procedure is easy to use and offers substantial improvements in overall accuracy in the emulation of large-scale computer experiments. We introduced a decomposition of the error of *any* interpolator into nominal and numeric portions. This decomposition is important because it allows the two sources of error to be analyzed separately while emphasizing the interplay between the two types of errors. We proved a very general result bounding the numeric error of a multi-step interpolator, of which an ordinary interpolator is a special case. This result constitutes the only complete and rigorous bound on the numeric error of the multi-step interpolator. We proved that in the situation where the earlier stage kernels are convolutions of the later stage kernels, substantial nominal improvements can be realized. In the context of the multi-step interpolator, this result is the most general and explicit of its kind.

Further work on the multi-step interpolation method will be explored in the following directions. First, its implementation details, along with various examples, will be reported in a subsequent article, to illustrate the theoretical results derived here. The implementation of the method requires the generation of nested data sites, for which the typical choice in applied mathematics is nested grids. Nested space-filling designs [15, 41–43], originally constructed for running multiple computer experiments with different levels of accuracy, are a better choice because of their good uniformity properties. Such designs can be generated by exploiting *nesting* in orthogonal arrays [19], U designs [55, 56], orthogonal Latin hypercubes [3, 25, 26, 54, 62] or scrambled nets [38]. Second, emulation of computer models with qualitative and quantitative factors is currently getting increasing attention [17, 45, 46]. We plan to extend the multi-step procedure to accommodate these two types of factors. Third, beyond emulation of computer experiments, singularity issues arise in fitting many other large kernel models. We plan to introduce a general multi-step framework for fitting kernel based classification and regression methods with a large number of observations. As in the multi-step interpolation procedure, this framework obtains nested data sites and then fits a kernel model in multiple steps, where in each step interpolation is replaced by an appropriate procedure for the given problem. New theoretical bounds on the nominal and numeric accuracy, analogous to those in Sections 4 and 5, will be derived for this framework. The required well-spread nested data sites for the framework will be generated by using nested space-filling designs or the efficient *thinning* algorithm [12] for observational data. In the revision of this paper, we became aware of new theoretical developments of the multi-step method in applied mathematics, including [24] and [61].

Acknowledgments. The authors thank the Editor, an Associate Editor and two referees for their comments that have led to improvements in the paper. They also thank Greg Fasshauer, Yizhi Zhang, Fred Hickernell and Grace Wahba for their comments.

REFERENCES

[1] ARONSZAJN, N. (1950). Theory of reproducing kernels. *Trans. Amer. Math. Soc.* **68** 337–404. MR0051437

[2] BARRY, R. P. and PACE, R. K. (1999). Monte Carlo estimates of the log determinant of large sparse matrices. *Linear Algebra Appl.* **289** 41–54. MR1670972

[3] BINGHAM, D., SITTER, R. R. and TANG, B. (2009). Orthogonal and nearly orthogonal designs for computer experiments. *Biometrika* **96** 51–65. MR2482134

[4] THE BOEING COMPANY (2010). Design space exploration software.

[5] BOOKER, A. J. (2000). Well-conditioned kriging models for optimization of computer simulations. Mathematics and Computing Technology Phantom Works, M&CT Technical Report 00-002, The Boeing Co., Bellvue, WA.

[6] CURRIN, C., MITCHELL, T., MORRIS, M. and YLVISAKER, D. (1991). Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments. *J. Amer. Statist. Assoc.* **86** 953–963. MR1146343

[7] DANCIK, G. M. (2011). mlegp: Maximum likelihood estimates of Gaussian processes. R package Version 3.1.2. Available at http://CRAN.R-project.org/package=mlegp.

[8] FANG, K.-T., LI, R. and SUDJIANTO, A. (2006). *Design and Modeling for Computer Experiments*. Chapman and Hall/CRC, Boca Raton, FL. MR2223960

[9] FANG, K.-T., LIN, D. K. J., WINKER, P. and ZHANG, Y. (2000). Uniform design: Theory and application. *Technometrics* **42** 237–248. MR1801031

[10] FASSHAUER, G. E. (2007). *Meshfree Approximation Methods with MATLAB. Interdisciplinary Mathematical Sciences* **6**. World Scientific, Hackensack, NJ. MR2357267

[11] FASSHAUER, G. E. and JEROME, J. W. (1999). Multistep approximation algorithms: Improved convergence rates through postconditioning with smoothing kernels. *Adv. Comput. Math.* **10** 1–27. MR1671863

[12] FLOATER, M. S. and ISKE, A. (1996). Multistep scattered data interpolation using compactly supported radial basis functions. *J. Comput. Appl. Math.* **73** 65–78. MR1424869

[13] GNEITING, T. (2002). Compactly supported correlation functions. *J. Multivariate Anal.* **83** 493–508. MR1945966

[14] GOLUB, G. H. and VAN LOAN, C. F. (1989). *Matrix Computations*, 2nd ed. *Johns Hopkins Series in the Mathematical Sciences* **3**. Johns Hopkins Univ. Press, Baltimore, MD. MR1002570

[15] HAALAND, B. and QIAN, P. Z. G. (2010). An approach to constructing nested space-filling designs for multi-fidelity computer experiments. *Statist. Sinica* **20** 1063–1075. MR2729853

[16] HALES, S. J. and LEVESLEY, J. (2002). Error estimates for multilevel approximation using polyharmonic splines. *Numer. Algorithms* **30** 1–10.

[17] HAN, G., SANTNER, T. J., NOTZ, W. I. and BARTEL, D. L. (2009). Prediction for computer experiments having quantitative and qualitative input variables. *Technometrics* **51** 278–288. MR2751072

[18] HARVILLE, D. A. (2008). *Matrix Algebra From a Statistician's Perspective*. Springer, New York.

[19] HEDAYAT, A. S., SLOANE, N. J. A. and STUFKEN, J. (1999). *Orthogonal Arrays*: *Theory and Applications*. Springer, New York. MR1693498

[20] IBM (2010). A smarter planet project. Available at http://www.ibm.com/smarterplanet/.

[21] JOSEPH, V. R. and HUNG, Y. (2008). Orthogonal-maximin Latin hypercube designs. *Statist. Sinica* **18** 171–186. MR2416907

[22] KAUFMAN, C. G., SCHERVISH, M. J. and NYCHKA, D. W. (2008). Covariance tapering for likelihood-based estimation in large spatial data sets. *J. Amer. Statist. Assoc.* **103** 1545–1555. MR2504203

[23] KOEHLER, J. R. and OWEN, A. B. (1996). Computer experiments. In *Design and Analysis of Experiments. Handbook of Statistics* **13** 261–308. North-Holland, Amsterdam. MR1492571

[24] LE GIA, Q. T., SLOAN, I. H. and WENDLAND, H. (2010). Multiscale analysis in Sobolev spaces on the sphere. *SIAM J. Numer. Anal.* **48** 2065–2090.

[25] LIN, C. D., BINGHAM, D., SITTER, R. R. and TANG, B. (2010). A new and flexible method for constructing designs for computer experiments. *Ann. Statist.* **38** 1460–1477. MR2662349

[26] LIN, C. D., MUKERJEE, R. and TANG, B. (2009). Construction of orthogonal and nearly orthogonal Latin hypercubes. *Biometrika* **96** 243–247. MR2482150

[27] LINKLETTER, C., BINGHAM, D., HENGARTNER, N., HIGDON, D. and YE, K. Q. (2006). Variable selection for Gaussian process models in computer experiments. *Technometrics* **48** 478–490. MR2328617

[28] LOH, W.-L. (1996). A combinatorial central limit theorem for randomized orthogonal array sampling designs. *Ann. Statist.* **24** 1209–1224. MR1401845

[29] LOH, W.-L. (1996). On Latin hypercube sampling. *Ann. Statist.* **24** 2058–2080. MR1421161

[30] LOH, W.-L. (2003). On the asymptotic distribution of scrambled net quadrature. *Ann. Statist.* **31** 1282–1324. MR2001651

[31] LOH, W.-L. (2008). A multivariate central limit theorem for randomized orthogonal array sampling designs in computer experiments. *Ann. Statist.* **36** 1983–2023. MR2435462

[32] MCKAY, M. D., CONOVER, W. J. and BECKMAN, R. J. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics* **21** 239–245. MR0533252

[33] NARCOWICH, F. J., SCHABACK, R. and WARD, J. D. (1999). Multilevel interpolation and approximation. *Appl. Comput. Harmon. Anal.* **7** 243–261. MR1721806

[34] NIEDERREITER, H. (1992). *Random Number Generation and Quasi-Monte Carlo Methods. CBMS-NSF Regional Conference Series in Applied Mathematics* **63**. SIAM, Philadelphia, PA. MR1172997

[35] OWEN, A. (1994). Lattice sampling revisited: Monte Carlo variance of means over randomized orthogonal arrays. *Ann. Statist.* **22** 930–945. MR1292549

[36] OWEN, A. B. (1992). A central limit theorem for Latin hypercube sampling. *J. Roy. Statist. Soc. Ser. B* **54** 541–551. MR1160481

[37] OWEN, A. B. (1992). Orthogonal arrays for computer experiments, integration and visualization. *Statist. Sinica* **2** 439–452. MR1187952

[38] OWEN, A. B. (1995). Randomly permuted (t, m, s)-nets and (t, s)-sequences. In *Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing* (*Las Vegas*, *NV*, 1994). *Lecture Notes in Statist.* **106** 299–317. Springer, New York. MR1445791

[39] OWEN, A. B. (1997). Monte Carlo variance of scrambled net quadrature. *SIAM J. Numer. Anal.* **34** 1884–1910. MR1472202

[40] OWEN, A. B. (1997). Scrambled net variance for integrals of smooth functions. *Ann. Statist.* **25** 1541–1562. MR1463564

[41] QIAN, P. Z. G. (2009). Nested Latin hypercube designs. *Biometrika* **96** 957–970. MR2767281

[42] QIAN, P. Z. G. and AI, M. (2010). Nested lattice sampling: A new sampling scheme derived by randomizing nested orthogonal arrays. *J. Amer. Statist. Assoc.* **105** 1147–1155. MR2752610

[43] QIAN, P. Z. G., AI, M. and WU, C. F. J. (2009). Construction of nested space-filling designs. *Ann. Statist.* **37** 3616–3643. MR2549572

[44] QIAN, P. Z. G., TANG, B. and WU, C. F. J. (2009). Nested space-filling designs for computer experiments with two levels of accuracy. *Statist. Sinica* **19** 287–300. MR2487890

[45] QIAN, P. Z. G. and WU, C. F. J. (2009). Sliced space-filling designs. *Biometrika* **96** 945–956. MR2767280

[46] QIAN, P. Z. G., WU, H. and WU, C. F. J. (2008). Gaussian process models for computer experiments with qualitative and quantitative factors. *Technometrics* **50** 383–396. MR2457574

[47] RIPPA, S. (1999). An algorithm for selecting a good value for the parameter c in radial basis function interpolation. *Adv. Comput. Math.* **11** 193–210. MR1731697

[48] SACKS, J., SCHILLER, S. B. and WELCH, W. J. (1989). Designs for computer experiments. *Technometrics* **31** 41–47. MR0997669

[49] SACKS, J., WELCH, W. J., MITCHELL, T. J. and WYNN, H. P. (1989). Design and analysis of computer experiments. *Statist. Sci.* **4** 409–423.

[50] SANTNER, T. J., WILLIAMS, B. J. and NOTZ, W. I. (2003). *The Design and Analysis of Computer Experiments*. Springer, New York. MR2160708

[51] STEIN, E. M. and WEISS, G. (1971). *Introduction to Fourier Analysis on Euclidean Spaces*. *Princeton Mathematical Series* **32**. Princeton Univ. Press, Princeton, NJ. MR0304972

[52] STEIN, M. (1987). Large sample properties of simulations using Latin hypercube sampling. *Technometrics* **29** 143–151. MR0887702

[53] STEIN, M. L., CHI, Z. and WELTY, L. J. (2004). Approximating likelihoods for large spatial data sets. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **66** 275–296. MR2062376

[54] STEINBERG, D. M. and LIN, D. K. J. (2006). A construction method for orthogonal Latin hypercube designs. *Biometrika* **93** 279–288. MR2278083

[55] TANG, B. (1993). Orthogonal array-based Latin hypercubes. *J. Amer. Statist. Assoc.* **88** 1392–1397. MR1245375

[56] TANG, B. (1994). A theorem for selecting OA-based Latin hypercubes using a distance criterion. *Comm. Statist. Theory Methods* **23** 2047–2058. MR1281902

[57] VARGA, R. S. (2004). *Geršgorin and His Circles*. *Springer Series in Computational Mathematics* **36**. Springer, Berlin. MR2093409

[58] WAHBA, G. (1983). Bayesian "confidence intervals" for the cross-validated smoothing spline. *J. Roy. Statist. Soc. Ser. B* **45** 133–150. MR0701084

[59] WAHBA, G. (1990). *Spline Models for Observational Data*. *CBMS-NSF Regional Conference Series in Applied Mathematics* **59**. SIAM, Philadelphia, PA. MR1045442

[60] WENDLAND, H. (2005). *Scattered Data Approximation*. *Cambridge Monographs on Applied and Computational Mathematics* **17**. Cambridge Univ. Press, New York. MR2131724

[61] WENDLAND, H. (2010). Multiscale analysis in Sobolev spaces on bounded domains. *Numerische Mathematik* **116** 493–517.

[62] YE, K. Q. (1998). Orthogonal column Latin hypercubes and their application in computer experiments. *J. Amer. Statist. Assoc.* **93** 1430–1439. MR1666638

CENTRE FOR QUANTITATIVE MEDICINE
OFFICE OF CLINICAL SCIENCES
DUKE-NUS GRADUATE MEDICAL SCHOOL
SINGAPORE 169857
AND
DEPARTMENT OF STATISTICS AND APPLIED PROBABILITY
NATIONAL UNIVERSITY OF SINGAPORE
SINGAPORE 117546
E-MAIL: benjamin.haaland@duke-nus.edu.sg

DEPARTMENT OF STATISTICS
UNIVERSITY OF WISCONSIN, MADISON
MADISON, WISCONSIN 53706
USA
E-MAIL: peterq@stat.wisc.edu