International Journal for Uncertainty Quantification 1(1), xxx–xxx (2010)
Preprint ANL/MCS-P1806-1110

ORTHOGONAL BASES FOR POLYNOMIAL REGRESSION WITH DERIVATIVE INFORMATION IN UNCERTAINTY QUANTIFICATION

Yiou Li[1], Mihai Anitescu[2,*], Oleg Roderick[2] & Fred Hickernell[1]

[1]*Department of Applied Mathematics, Illinois Institute of Technology, 10 W. 32nd, Chicago, IL, 60616, USA*

[2]*Mathematics and Computer Science Division, Argonne National Laboratory, 9600 S. Cass Ave., Argonne, IL, 60439, USA*

Original Manuscript Submitted: 11/08/2010; Final Draft Received: 11/08/2010

We discuss the choice of polynomial basis for approximation of uncertainty propagation through complex simulation models with capability to output derivative information. Our work is part of a larger research effort in uncertainty quantification using sampling methods augmented with derivative information. The approach has new challenges compared with standard polynomial regression. In particular, we show that a tensor product multivariate orthogonal polynomial basis of an arbitrary degree may no longer be constructed.
We provide sufficient conditions for an orthonormal set of this type to exist, a basis for the space it spans. We demonstrate the benefits of the basis in the propagation of material uncertainties through a simplified model of heat transport in a nuclear reactor core. Compared with the tensor product Hermite polynomial basis, the orthogonal basis results in a better numerical conditioning of the regression procedure, a modest improvement in approximation error when basis polynomials are chosen a priori, and a significant improvement when basis polynomials are chosen adaptively, using a stepwise fitting procedure.

KEY WORDS: *Uncertainty quantification; Representation of uncertainty; Stochastic collocation; Heat transfer; energy and the environment*

1. INTRODUCTION

We discuss the choice of polynomial basis in polynomial regression with derivative (PRD) information. In PRD a model of the system response is computed by regressing both the output information and its derivative with respect to the physical parameters computed at a small number of sample points in the parameter space. In turn, this model can be used to efficiently estimate the system response under parametric uncertainty.

For several nuclear reactor system simulations, we found that approximation of the uncertainty effect by PRD is more precise than linear approximation by an order of magnitude or more [1]. Moreover, we have shown that the PRD model can be used as a control variate to reduce the variance of certain statistical estimators. In turn, this results in far fewer system samples being used to obtain a reasonable confidence interval for those estimators.

Our approach hinges on the observation that adjoint techniques can be used to efficiently compute gradient information. In particular, the required derivatives can be computed by algorithmic, or automatic, differentiation: a procedure that reads source code of the model, augments algebraic operations with their partial derivatives, and then assembles the gradients using the chain rule. The adjoint (reverse) mode of automatic differentiation computes the gradient of the system response in a time that is at most 5 times the cost of one function evaluation (system simulation for a given choice of parameters), *irrespective of the dimension of the parameter space* [2]. Hence, in principle, we

*Correspond to: Mihai Anitescu, E-mail: anitescu@mcs.anl.gov, URL: /http://www.mcs.anl.gov/ anitescu/

obtain no less than $d/5$ more information for the same computational cost when compared with samples of the function values alone. As a result, the use of derivative information allows one to build approximations based on smaller training sets (or, equivalently, by using fewer computationally expensive model runs).

As we have demonstrated in prior work, the use of derivative information in PRD relaxes the limitations of the "curse of dimensionality" and allows uncertainty quantification of models with 10 – 100 uncertainty quantifiers. At the high end of this range, a reasonable approximation precision requires a very large polynomial basis, and the regression procedure becomes numerically ill-conditioned for the Hermite polynomials basis, one of the most commonly used in uncertainty quantification. This raises the following important challenges: *How do we choose a basis that reduces or eliminates the ill-conditioning in the polynomial regression with the derivative information procedure? How do we take advantage of this basis?* Answering these questions becomes the central objective of this work.

To demonstrate our findings on an example that exhibits some of the complexity encountered in advanced engineering codes, we use a three-dimensional model of heat transport in a sodium-cooled reactor core, described in Section 4.1. The uncertainty in the model originates from the experimental error in measurement of dependency of material properties on temperature In the computational experiments described in this work the uncertainty space has dimension 12; a 66-dimensional version is also available. We compare the performance of the new basis with such standard choices as Hermite polynomials, and we show that the resulting information matrix is much better conditioned. In our numerical experiments, the use of the new basis results in a small improvement in precision when the basis polynomials are chosen a priori, and a significant improvement (of several orders of magnitude) when the basis polynomials are chosen adaptively, using a stepwise fitting procedure.

The rest of the paper is organized as follows. In Section 2, we explain the general task of uncertainty quantification for simulation models and the polynomial regression with derivatives (PRD) approach in particular, as well as the place of PRD in the context of techniques for uncertainty propagation. In Section 3, we analyze the features of tensor-product orthonormal multivariate bases for use in PRD and describe procedures for building them. In Section 4, we describe the nuclear reactor model used in our numerical experiments and apply the PRD technique both in standard form and as part of stepwise regression. In Section 5, we discuss the significance of the performed work and future steps needed to extend the technique.

1.1 Introduction

The quantitative characterization and reduction of uncertainties in large models are an important area of research, related to statistical analysis of random phenomena and to physical study of complex systems. One of the most general tasks is to describe the influence of random inputs on an arbitrary but given output. Any improvement to the existing tools of uncertainty quantification will have both mathematical and industrial benefits.

This is particularly true in the area of nuclear reactor design and control, where greater engineering precision results in significant preservation of resources. The usual difficulties in modeling the work of the nuclear reactor include the large size of the associated systems of equations, the nonlinearity, and the implicit dependence of equations on parameters. As a result, one can afford to run the computational model of a nuclear reactor only for a small number of scenarios involving the values of its physical parameters. In addition, although the information on the behavior of parameters is available in formats convenient for experimental physics and engineering purposes, such formats are not necessarily appropriate for uncertainty analysis.

As a test case for our method, we consider a mathematical model of heat transport in the nuclear reactor core. We create a computationally efficient method to describe the dependence of a merit function, the maximum centerline temperature, on the model uncertainties.

1.2 Problem Definition

We first present our mathematical problem in the most general form. Consider an arbitrary system of (discretized, algebraic-differential) equations with main variables $\mathbf{T} = (T_1, T_2, .., T_n)$ and *intermediate parameters* $\mathbf{R} = (R_1, R_2, ..., R_N)$:

$$F(\mathbf{T}, \mathbf{R}) = 0. \tag{1}$$

Here and in the sequel, we denote by R a generic intermediate parameter, one of $R_1, R_2, \ldots, R_N$, and by $\mathbf{R}$ the vector of all such parameters.

The parameter set $\mathbf{R}$ is not independent. It is related to the variables by a set of expressions

$$R := R(T) + \Delta\mathbf{R}(\mathbf{T}; \mathbf{x}) \tag{2}$$

with the experimental error $\Delta\mathbf{R}(\mathbf{T}; \mathbf{x})$, which is also dependent on temperature and on a set of parameters $\mathbf{x}$ that quantifies the uncertainty. The parameters $\mathbf{x}$ become the *primary* uncertainty parameters. Then the structural equation of the nonlinear system (1) becomes

$$F(\mathbf{T}, \mathbf{R}(\mathbf{T}; \mathbf{x})) = 0. \tag{3}$$

Strictly speaking, equation (3) now results in the primary variable T being a function of $\mathbf{x}$ and not of $\mathbf{R}$ (which is itself a function of temperature). To abide by the physical meaning of the respective parameters R, we may still write $\mathbf{T} = \mathbf{T}(\mathbf{R})$.

For a given *merit function*

$$J = J(\mathbf{T}) \quad : R^n \to R \tag{4}$$

we need to find the influence of uncertainties in the parameters on the uncertainty of the output. To find the effects of the uncertainty on the merit function J,

$$\Delta J = J(\mathbf{T}(\mathbf{R})) - \mathbf{J}(\mathbf{T}(\mathbf{R} + \Delta\mathbf{R})), \tag{5}$$

we express the output as a function of uncertainties of the inputs, represented by the parameters $\mathbf{x}$:

$$J \approx \hat{J} = \hat{J}(\mathbf{T}(\mathbf{R}(\mathbf{T}) + \Delta\mathbf{R}(\mathbf{T}; \mathbf{x}))) = \hat{J}(\mathbf{x}). \tag{6}$$

This representation, or *surrogate model*, is created by using *a polynomial regression with derivative information* (PRD), a technique closely related to stochastic finite-element approximation [3–5]; also see Section 2.2.

That is, we create a set of polynomials in $\mathbf{x}$, $\{\Psi_q(mathbfx)\}_{q\in Q}$, and we define $\hat{J}(\mathbf{x}) = \sum_{q\in Q} x_q \Psi_q(\mathbf{x})$. The coefficients β_q are obtained by requiring that the function and derivative values of the surrogate model $\hat{J}(\mathbf{x})$ match the ones of the real model $J(\mathbf{x})$, in a least-squares sense.

The main motivation for our approach is the observation that, for most computational multiphysics applications, there exist methods for computing the full gradient of a merit function J that require an effort comparable to or even smaller than the effort of computing J itself. We have in mind primarily the generic reverse automatic differentiation approach [6] or the adjoint sensitivity method [7]. Both approaches are applicable to arbitrary nonlinear systems with continuous solutions and differentiable merit functions. Therefore, as the number of parameters increases, using the output $\mathbf{J}$ and its derivative information is appealing because it provides more information than does the evaluation of J by a factor equal to the number of parameters, while having a similar overall comparable cost. We thus expect to have to compute the solution of (3) far fewer times.

2. UNCERTAINTY QUANTIFICATION BY POLYNOMIAL REGRESSION WITH DERIVATIVE INFORMATION

2.1 Problem Definition

We view a generic model with uncertainty as a discretized system of algebraic-differential equations:

$$\begin{aligned} F(\mathbf{T},\mathbf{R}) &= 0 && (7)\\ \mathbf{R} &= \mathbf{R}(\mathbf{T},\mathbf{x}) = \mathbf{R}_0(\mathbf{T})\cdot(1+\Delta\mathbf{R}(\mathbf{T},\mathbf{x})) && (8)\\ J &= J(\mathbf{T}), && (9)\end{aligned}$$

where the variables $\mathbf{T} = (T_1, T_2, ..., T_n)$ characterize the model state; the dependence of physical parameters of the model $\mathbf{R} = (R_1, R_2, ..., R_N)$ includes errors $\Delta\mathbf{R} = (\Delta R_1, \Delta R_2, ..., \Delta R_N)$; an output of interest is expressed by the merit function $J(\mathbf{T})$; and uncertainty in the physical description of the model is described by a set of stochastic variables $\mathbf{x} = (x_1, x_2, ..., x_d)$.

The algebraic structure under which uncertainty is introduced into the model can be as simple as $\Delta\mathbf{R}(\mathbf{T},\mathbf{x}) = \mathbf{x}$ or more complex depending on the modelling principles. Our models of interest typically use a complicated form of definition of the uncertainty, including dependence on the state variable $\mathbf{T}$. One such example is presented in Section 4.1.

Our problem is to efficiently characterize the uncertainty in the merit function $J(\mathbf{T})$. We are given

- a probability structure on the physical uncertainty space (though some further modeling may be necessary to properly characterize it [8]) of the variables $\mathbf{x}$, and

- a numerical implementation of the physical phenomenon that computes $\mathbf{T}$ given $\mathbf{R}(\mathbf{T},\mathbf{x})$ and, subsequently, J.

For this work, we assume that $\nabla_x J$ can be obtained in this setup either by the adjoint derivative implementation provided by the domain scientist or by the backward automated differentiation mode [9]. We discuss the suitability of the assumption at the end of this section.

The challenge in our endeavor is that for a model of more than trivial complexity, the dependence of the output J on the uncertainty $\mathbf{x}$ cannot be described explicitly. A straightforward approach to understanding this dependence would be to evaluate the model over a large, representative subset of the uncertainty space. However, one can afford to run the model for only a limited number of scenarios. Therefore, a practical approach to uncertainty quantification is to approximate J based on small-scale sampling using the code, followed by large-scale exploration of the approximate model.

To that end, we redefine the output as a function of uncertainty quantifiers, $J(\mathbf{T}) := J(\mathbf{x})$, and choose a set $\mathbf{\Psi}$ of polynomials on the uncertainty quantifiers $\mathbf{x} = \{x_i\}$, $i = 1, ..., d$. A subset $\{\Psi_l\}$ is used to approximate the merit function:

$$J \approx \widetilde{J} = \sum_l \beta_l \Psi_l(\mathbf{x}). \tag{10}$$

The coefficients β_l are obtained by requiring that the function and the derivative values of the surrogate model $\widetilde{J}(x)$ match the ones of the real model $J(x)$, in a least-squares sense. Approximation of the uncertain effects by a flexible basis of functions on uncertainty quantifiers is closely related to the stochastic finite element method (SFEM); such a basis is sometimes called *polynomial chaos*, as we discuss in Section §2.2.

We extend the idea by using derivative information ∇Jat every training point, in addition to the function values

J. The polynomial fitting equations are as follows

$$\begin{pmatrix} \Psi_1(S_1) & \Psi_2(S_1) & \cdots & \Psi_k(S_1) \\ \vdots & \vdots & \vdots & \vdots \\ \Psi_1(S_m) & \Psi_2(S_m) & \cdots & \Psi_k(S_m) \\ \frac{\partial \Psi_1(S_1)}{\partial \mathbf{x}_1} & \frac{\partial \Psi_2(S_1)}{\partial \mathbf{x}_1} & \cdots & \frac{\partial \Psi_k(S_1)}{\partial \mathbf{x}_1} \\ \frac{\partial \Psi_1(S_1)}{\partial \mathbf{x}_2} & \frac{\partial \Psi_2(S_1)}{\partial \mathbf{x}_2} & \cdots & \frac{\partial \Psi_k(S_1)}{\partial \mathbf{x}_2} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial \Psi_1(S_1)}{\partial \mathbf{x}_n} & \frac{\partial \Psi_2(S_1)}{\partial \mathbf{x}_n} & \cdots & \frac{\partial \Psi_k(S_1)}{\partial \mathbf{x}_n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial \Psi_1(S_2)}{\partial \mathbf{x}_1} & \frac{\partial \Psi_2(S_2)}{\partial \mathbf{x}_1} & \cdots & \frac{\partial \Psi_k(S_2)}{\partial \mathbf{x}_1} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial \Psi_1(S_m)}{\partial \mathbf{x}_n} & \frac{\partial \Psi_2(S_m)}{\partial \mathbf{x}_n} & \cdots & \frac{\partial \Psi_k(S_m)}{\partial \mathbf{x}_n} \end{pmatrix} \cdot \beta = \begin{pmatrix} J(S_1) \\ \vdots \\ J(S_m) \\ \frac{\partial J(S_1)}{\partial \mathbf{x}_1} \\ \frac{\partial J(S_1)}{\partial \mathbf{x}_2} \\ \vdots \\ \frac{\partial J(S_1)}{\partial \mathbf{x}_n} \\ \frac{\partial J(S_2)}{\partial \mathbf{x}_1} \\ \vdots \\ \frac{\partial J(S_m)}{\partial \mathbf{x}_n} \end{pmatrix}, \tag{11}$$

where $S_1, S_2, ...S_m$ are sample training points in the uncertainty space: $S_i = (x_1^{(i)}, x_2^{(i)}, ..., x_m^{(i)})$. An evaluation of J and its first derivatives at $\mathbf{S}_i$ generates a subcolumn of entries, that is, information for several rows at once. In our recent work [8], the system of equations is slightly overdetermined, and is resolved in a least-squares sense.

We justify our use of derivative information (as opposed to adding more sample points by using only function values) by the fact that it is possible to obtain complete gradient information of the model with a limited computational overhead, independent of the model complexity. A computability theory result puts this overhead at 500% at most [9], making it advantageous to use PRD for models with uncertainty dimension higher than 5. In practice, the overhead is less. In our experiments the gradient was typically obtained in less time than one model evaluation. This situation is not unusual in cases where a nonlinear iteration is present to compute the system state and, subsequently, the response function $J(\mathbf{x})$. The sensitivity equations involve only one such system, whose cost may be comparable to one of the iterations.

A downside of the approach is that one has to make the derivative information available. In our numerical experiments, the adjoint differentiation was hand-coded. In a related effort [10], we have investigated the application of our approach when the gradient information is computed by automatic (or algorithmic) differentiation (AD), which uses the same chain-rule differentiation approach but with minimal human involvement. Our early investigations indicate that, while a nontrivial endeavor, gradient information can be obtained by AD, even when legacy code is involved, with a small to moderate amount of development time.

2.2 Connection between PRD and Collocation Stochastic Finite Element Approaches

Our work has originated in investigations of SFEM approaches for uncertainty quantification [11, 12], and, particularly, their application to nuclear engineering models. In the case where a SFEM approach with a polynomial basis is used, one constructs an approximation from (7):

$$\widetilde{\mathbf{T}}(\mathbf{x}) = \sum_l \beta_l^T \Psi_l(\mathbf{x}). \tag{12}$$

One such technique is the Galerkin approach [13]: the coefficients β^T are determined by requiring that the projection of the residual of (7) on the space V spanned by the polynomials Ψ_l be zero. We have demonstrated that the approach can be extended to constrained optimization problems as well, while maintaining the optimization structure as opposed

to converting the problem to the nonlinear equation (7) [12].

More relevant for our discussion, however, is the collocation approach. In this approach, the coefficients β^T are determined by enforcing that the stochastic finite-element approximation $\widetilde{\mathbf{T}}(\mathbf{x}) = \sum_l \beta_l^T \Psi_l(\mathbf{x})$ have a zero residual at a set of collocation points, x_i, $i = 1, 2, \ldots, M$. That is,

$$F(\widetilde{\mathbf{T}}(x_i), R(\widetilde{\mathbf{T}}(x_i), x_i)) = 0, \quad i = 1, 2, \ldots, m. \tag{13}$$

Assuming that the system $F(\mathbf{T}(\mathbf{x}), \mathbf{R}(\mathbf{T}(\mathbf{x}), \mathbf{x}))$ has a unique solution for a given $\mathbf{x}$, it follows that for each sample point S_i, there is a unique T_i such that $F(T_i, R(T_i, S_i)) = 0$. In turn the collocation problem (13) becomes equivalent to the interpolation problem

$$\sum_l \beta_l^T \Psi_l(x_i) = \widetilde{\mathbf{T}}(x_i) = T_i, \quad i = 1, 2, \ldots, M. \tag{14}$$

We can interpret (14) as an interpolation problem in each of the components of the vector set $\{\beta_l^T\}_l$. Effectively, based on (13), we can state that solving (14), and thus (13), is equivalent to building a surface response in each of the components of $\widetilde{\mathbf{T}}(\mathbf{x})$.

In this work, we are interested in one particular response function, $J(\mathbf{T})$. Assume, for the purpose of our argument, that the response function is *linear* in $\mathbf{T}$ and that we are seeking a response to the uncertainty variable $\mathbf{x}$ in the same polynomial space as we did for $\widetilde{\mathbf{x}}$, that is,

$$J(\mathbf{x}) = \sum_l \beta_l^J \Psi_l(\mathbf{x}). \tag{15}$$

Assume now that we carry out collocation on the state space (13), to which we apply the response function $J = J(\mathbf{T}(\mathbf{x}))$. It then immediately follows that $J(\mathbf{T}(\mathbf{x}))$ *also* satisfies the interpolation conditions

$$J(\widetilde{\mathbf{T}}(x_i)) = J(\mathbf{T}(x_i)), i = 1, 2, \ldots, M. \tag{16}$$

If, in addition, the function J is linear in the state variables $\mathbf{T}$, it immediately follows that the response function satisfies

$$\sum_l \beta_l^J \Psi_l(x_i) = \sum_l J(\beta_l^T) d\Psi_l(x_i) = J(x_i) = J(\widetilde{\mathbf{T}}(x_i)). \tag{17}$$

Therefore, if the interpolation problem (14) is well posed and thus has a unique solution, it follows that using collocation for the state and applying the response function J are *equivalent to* determining the coefficients β_l^J from imposing the collocation-interpolation conditions (17) directly on J. Moreover, the solution to (17) can be obtained by the least-squares regression approach,

$$\min_{\beta_i^J} \sum_{i=1}^{M} \left(J(x_i) - \sum_l \beta_l^J \Psi_l(x_i) \right)^2, \tag{18}$$

since the latter problem has a unique solution if the interpolation problems (17) and (14) have a unique solution. We also point out that obtaining an approximation of the response function (17) that satisfies (16) directly also carries the name (at least for some of its variants) of the response surface approach. Therefore, the approach described above can be seen simultaneously as an SFEM collocation approach, an interpolation approach, a surface response approach, and a regression approach.

When the response $J(\mathbf{T}(\mathbf{x}))$ is nonlinear, the equivalence among the approaches ceases to hold; but if the function J is smooth, one can demonstrate by polynomial approximation arguments that (18) will produce an approximation of the similar quality to using collocation and then using $J(\widetilde{\mathbf{x}})$ as the approximation.

An additional advantage of using (18) over (13) consists of far lower memory overhead, since multiple values of

the potentially large state vector $\mathbf{x}$ do not need to be stored.

In addition, in the case where gradient information is sought and J is real valued (or vector valued of low dimension), adjoint methods can be used to efficiently compute derivative information. We note that either advantage disappears if J is vector valued of large dimension.

In this work, we focus on the widely encountered case where J is real valued (though the approach is immediately extensible to vector response J, but the effort versus precision analysis will not be carried out in that case). We choose the regression ansatz (18), which is more flexible about the type of information included in creating an approximate model of J. In particular, we are interested in the case where derivative information for J is available and the formulation (18) naturally extends to

$$\min_{\beta_l^J} \sum_{i=1}^{m} \left[\left(J(x_i) - \sum_l \beta_l^J \Psi_l(x_i) \right)^2 + \sum_{j=1}^{d} \left(\frac{\partial J(x_i)}{\partial x_j} - \sum_l \beta_l^J \frac{\partial \Psi_l(x_i)}{\partial x_j} \right)^2 \right]. \tag{19}$$

We note that the optimality conditions of (19) are the same as the least-squares version of (11). It is easy to derive other forms of the regression approach (19) that include incomplete derivative information or weighting; but, for this paper, we will include only the standard approach (19).

Given the connection we have pointed out between our approach and collocation approaches, we will still refer to the optimality conditions of (19), implied when solving (11), as collocation equations since, as pointed out in the preceding paragraphs, for the linear response case and unique solution of (11) they are equivalent to the SFEM collocation approach.

2.3 Comparison with Previous Approaches

As described in Section 2.2, the PRD method is related to polynomial approximations of complex systems with uncertain parameters and stochastic finite-element methods [3–5, 12]. An important class of SFEM is SFEM-Galerkin methods [3]. Such methods are robust, but they also are demanding in terms of computational effort and require substantial storage. SFEM collocation methods [14–16] are closely related to our approach. They are nonintrusive and do not need specialized solvers, but they still use a state variable approximation and, in most circumstances do not explore the use of gradient information. We also point out that using a state variable approximation makes the use of adjoint calculation much less efficient since the number of dependent variables is now very large [9].

To a great extent, our method can be thought as a hybrid between a Monte Carlo method [17, 18] and a sensitivity surface response method [19, 20]. Such approaches have recently been proposed in the context of Gaussian process methods [21]. Closer to our approach, other authors have also proposed SFEM-based hybrid methods [20, 22]. In particular, both references point out the potential information efficiency that can be obtained from the gradient and demonstrate the reduction in number of samples for the same quality of the uncertainty assessment, as we do in [8]. Reference [20] uses a singular value decomposition approach to determine the coefficients of the model, which would, in principle, result in a model equivalent to the regression approach. Nevertheless, our recent work [8] enhanced the approach in several ways. Specifically, we presented new ways to prune the polynomial basis to mitigate the effects of the curse of dimensionality and described the use of the approach as a control variate to reduce the bias. The regression–least squares interpretation that we posited is essential to determine the advances in polynomial basis that we develop in the rest of this work. Moreover, we have been — to our knowledge — the first group to investigate the issues of applying the method in the nuclear engineering field [8, 10, 23].

Our work shares some characteristics with surface response approximation [20, 22, 24, 25]. Such approaches have been successfully used in nuclear engineering applications, including in the USNRC licensing process [26], [27], [28], [29], [30]. Nevertheless, our method is different in its use of gradient information as an enhancement to Monte Carlo

sampling.

3. ORTHOGONAL BASIS FOR POLYNOMIAL REGRESSION WITH DERIVATIVE INFORMATION

In this section, we discuss the theoretical considerations that lead to the construction of a polynomial basis for doing regression with derivative information. In this work, for simplicity, we call a basis an orthonormal system of polynomials (which is a basis for the linear space it spans).

3.1 Modeling Framework

Choose a set $\mathbf{\Theta}$ of multivariable orthonormal polynomials of the variables $\mathbf{x} = (x_1, \cdots, x_d)^T$. A subset $\{\Psi_l\} \subset \mathbf{\Theta}$ is used to approximate the merit function:

$$J \approx \widetilde{J} = \sum_l \beta_l \Psi_l. \tag{20}$$

Define an operator $\mathbf{L_x}$ that, when applied to a d-variate scalar function f, returns its value and gradient information:

$$\mathbf{L_x} f = \left(f(\mathbf{x}), \frac{\partial f}{\partial x_1}(\mathbf{x}), \cdots, \frac{\partial f}{\partial x_d}(\mathbf{x}) \right)^T. \tag{21}$$

For a vector function $\mathbf{f} = (f_1, \cdots, f_k)^T$, we extend the definition of the operator as follows:

$$\mathbf{L_x f}^T = \begin{pmatrix} f_1(\mathbf{x}) & f_2(\mathbf{x}) & \cdots & f_k(\mathbf{x}) \\ \frac{\partial f_1}{\partial x_1}(\mathbf{x}) & \frac{\partial f_2}{\partial x_1}(\mathbf{x}) & \cdots & \frac{\partial f_k}{\partial x_1}(\mathbf{x}) \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial f_1}{\partial x_d}(\mathbf{x}) & \frac{\partial f_2}{\partial x_d}(\mathbf{x}) & \cdots & \frac{\partial f_k}{\partial x_d}(\mathbf{x}) \end{pmatrix}. \tag{22}$$

We now use this notation to define the collocation matrix in this framework. For the considered choice of polynomials $\mathbf{\Psi} = (\Psi_1, \cdots, \Psi_k)^T$, we define the define collocation matrix $\mathbf{F}$ as follows:

$$\mathbf{F} = \begin{pmatrix} \mathbf{L_{x_1}} \mathbf{\Psi}^T \\ \mathbf{L_{x_2}} \mathbf{\Psi}^T \\ \vdots \\ \mathbf{L_{x_m}} \mathbf{\Psi}^T \end{pmatrix}. \tag{23}$$

Here, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$ are the m points at which the system output function J is sampled. Then our regression model becomes

$$\mathbf{L_x} J = \mathbf{L_x} \mathbf{\Psi}^T \boldsymbol{\beta} + \boldsymbol{\varepsilon}(\mathbf{x}), \tag{24}$$

where $\boldsymbol{\varepsilon}(\mathbf{x}) \in \mathbb{R}^{m+1}$ is the error term, which we assume here to be a random variable such that $\boldsymbol{\varepsilon}(\mathbf{x}_1)$ is independent from $\boldsymbol{\varepsilon}(\mathbf{x}_2)$ if $\mathbf{x}_1 \neq \mathbf{x}_2$. Moreover, we also will assume that the components of $\boldsymbol{\varepsilon}(\mathbf{x})$ are independently distributed with mean zero and the same variance σ. We will discuss the suitability of this assumption shortly.

To determine the parameters $\boldsymbol{\beta}$ of the model, we compute the values of the output function J and its derivatives at the m sample points. Then a single sample point $\mathbf{x}_i$ will generate a subvector of entries with components $J(\mathbf{x}_i)$ and $(\frac{\partial J(\mathbf{x}_i)}{\partial x_j})$, $j = 1, 2, \ldots, d$, providing right-side information for several collocation equations at once. By matching the values of J and its derivatives with the corresponding polynomial representation, we build an extended system of collocation equations

$$\mathbf{F}\boldsymbol{\beta} = \mathbf{y}, \tag{25}$$

where $\mathbf{y} = \left(\mathbf{L}_{\mathbf{x}_1}^T J, \cdots, \mathbf{L}_{\mathbf{x}_m}^T J\right)^T$. This is equivalent to (11), but now using matrix-vector notation.

The system equation in (25) is overdetermined. The least squares solution, that is, the one satisfying (19), is given by

$$\widehat{\boldsymbol{\beta}} = (\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T\mathbf{y}, \tag{26}$$

provided that the matrix $\mathbf{F}$ has full column rank.

We now discuss the implications and suitability of several assumptions we have made. We observe that the estimator (26) is unbiased for the model (24) for any mean zero noise, irrespective of the other properties of the noise [31]. That is, from (26) $E[\widehat{\boldsymbol{\beta}}] = E[(\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T\mathbf{y}] = (\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T E[\mathbf{y}] = \boldsymbol{\beta}$. Therefore, our assumption that $\boldsymbol{\varepsilon}(\mathbf{x})$ has independent identically distributed entries has no bearing over the biasedness, even if incorrect for a particular model. Moreover, consistency (that is, convergence of $\widehat{\boldsymbol{\beta}}$ to $\boldsymbol{\beta}$ in probability for increasingly large data sets) would also follow under fairly weak conditions even if the distribution of $\boldsymbol{\varepsilon}(\mathbf{x})$ is misspecified.

Naturally, any confidence test will be affected if the covariance assumption on $\boldsymbol{\varepsilon}(\mathbf{x})$ is incorrect. On the other hand, absent other information about the problem, assuming that $\boldsymbol{\varepsilon}(\mathbf{x})$ has independent, identically distributed components is a reasonable starting assumption. In addition, it seems the correct assumption if the error is due to rounding. While assumptions about the proper noise form are clearly not without consequences, the latter observation, the robustness of several of the properties of classical regression with respect with several of its assumptions [31], and the fact that bias is not affected by the particular form of the noise prompt us to continue the analysis of the consequences of the independently, identically distributed component noise model at this time.

3.2 Design Consequences

Toward the end of making (26) a robust estimate, a crucial assumption is the one that $\left(\mathbf{F}^T\mathbf{F}\right)^{-1}$ not be singular. Moreover, assuming that the regression model (24) and the assumptions on $\boldsymbol{\varepsilon}$ are correct, then the estimator described in (26) satisfies $\mathrm{cov}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{F}^T\mathbf{F})^{-1}$. Therefore, obtaining a good regression estimate means obtaining a small $\mathrm{cov}(\widehat{\boldsymbol{\beta}})$, subject to a normalization constraint (such as prescribed trace). Such problems are the subject of experimental design [32]; and one design strategy, the D-optimal approach, attempts to maximize the determinant of the covariance matrix. Unfortunately, the D-optimal and other alphabetic optimal designs are highly dependent on the choice of the basis, $\boldsymbol{\Psi}$. In our situation, the final choice of $\boldsymbol{\Psi}$ is made after the data is observed. Such a data dependent basis selection procedure is common in linear regression. The effect of removing a Ψ_l from the surrogate model is confounded by the presence of other $\Psi_{l'}$ in the model. The magnitude of this confounding is proportional to the l, l' element of $\mathrm{cov}(\widehat{\boldsymbol{\beta}})$. Therefore, we aim to choose $\boldsymbol{\Psi}$ and a design $\mathbf{x}_1, ..., \mathbf{x}_m$ that makes $\mathrm{cov}(\widehat{\boldsymbol{\beta}})$ close to a multiple of the identity matrix.

Suppose that the design $\mathbf{x}_1, ..., \mathbf{x}_m$ is chosen to approximate a probability distribution with density ρ, in other words, the empirical distribution of the design is close to the distribution of a continuous random variable with density ρ. Therefore, the information matrix may be approximated by an integral involving the basis but independent of the details of the design:

$$\frac{1}{m}\mathbf{F}^T\mathbf{F} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{L}_{\mathbf{x}_i}\boldsymbol{\Psi}\left(\mathbf{L}_{\mathbf{x}_i}\boldsymbol{\Psi}^T\right) \approx$$

$$\approx \int_{\boldsymbol{\Omega}}\mathbf{L}_{\mathbf{x}}\boldsymbol{\Psi}\left(\mathbf{L}_{\mathbf{x}}\boldsymbol{\Psi}^T\right)\rho(\mathbf{x})d\mathbf{x} = \left(\int_{\boldsymbol{\Omega}}\left(\Psi_j(\mathbf{x})\Psi_h(\mathbf{x}) + \sum_{i=1}^{d}\frac{\partial\Psi_j}{\partial x_i}(\mathbf{x})\frac{\partial\Psi_h}{\partial x_i}(\mathbf{x})\right)\rho(\mathbf{x})d\mathbf{x}\right)_{j,h=1}^{k}. \tag{27}$$

This relationship suggests the definition of the inner product that depends on both the function and its derivative:

$$\langle f, g\rangle = \int_{\mathbf{\Omega}} \mathbf{L}_{\mathbf{x}} f \left(\mathbf{L}_{\mathbf{x}} g\right)^T \rho(\mathbf{x}) d\mathbf{x} = \int_{\mathbf{\Omega}} \left(f(\mathbf{x}) g(\mathbf{x}) + \sum_{i=1}^{d} \frac{\partial f}{\partial x_i}(\mathbf{x}) \frac{\partial g}{\partial x_i}(\mathbf{x}) \right) \rho(\mathbf{x}) d\mathbf{x}. \tag{28}$$

This approximation for the information matrix implies that it is approximately a multiple of the identity matrix if the Ψ_l are chosen to be orthonormal with respect to the inner product defined in (28): $\langle \Psi_j, \Psi_h \rangle = \delta_{jh}$.

Given any initial polynomial basis, one can use the Gram-Schmidt method to construct an orthonormal basis with respect to the inner product (28). However, as shown, such a basis might not be of tensor product form. A tensor product basis has the important advantage of facilitating the inclusion or exclusion of terms involving the variable x_i without adversely affecting the terms involving other variables. For example, the basis $1, x_1, x_2, x_1x_2$ is of tensor product form, and removing the variable x_2 means only removing the last two basis elements. The resulting basis still allows for general linear polynomials in x_1. The basis $1, x_1 + x_2, x_2, x_1x_2$ spans the same space as the first one, but now removing all terms involving x_2 leaves only the constant term.

Thus, one would like to have a tensor product basis that is orthonormal with respect to inner product (28). If the derivative terms are not included in the definition of the inner product, then one naturally obtains a tensor product of common orthogonal polynomials such as the Legendre polynomials in the case of the uniform distribution, or the Hermite polynomials in the Gaussian distribution [33]. Indeed, tensor product bases are the most routinely considered bases in uncertainty quantification. However, since the derivative terms must be included in the inner product, reflecting the derivative values in the the information matrix, it may not be possible to retain orthogonality and a tensor product basis for arbitrary orders of polynomials. This is an important issue to address, given the observation in previous work [1] that some of the original variables may exhibit higher degrees of nonlinearity than others. The next subsection explores this problem.

3.3 Characterizing a Tensor Product Basis

To gain a taste of the difficulties involved, consider the case for $d = 2$, with both variables uniformly distributed on $[-1, 1]$. The univariate polynomials with respect to the inner product in (28) are 1, x_1, $x_1^2 - \frac{1}{3}$, $x_1^3 - \frac{9}{10}x_1$, and 1, x_2, $x_2^2 - \frac{1}{3}$, $x_2^3 - \frac{9}{10}x_2$. Unfortunately, it can be shown that the multilinear polynomial $x_1 \cdot x_2$ is not orthogonal to the fourth degree polynomial $x_1 \cdot (x_2^3 - \frac{9}{10}x_2)$ under this inner product. Therefore, *a tensor-product orthogonal polynomial basis of an arbitrary degree may not exist when the inner product contains gradient information*, as is the case for our choice of inner product (28).

We thus proceed to investigate the circumstances under which tensor product bases can be defined, which necessarily must include constraints on the polynomial degrees that are considered. The following Theorems 2 and 3 and Corollary 2 provide sufficient conditions under the assumption that the variables are symmetrically distributed on their domain. We first characterize the one-variable polynomials orthogonal under the inner product (28).

Theorem 1. [ ]
Let $w_j(x)$ be univariate ($d = 1$) orthonormal polynomials with respect to the inner product (28) such that the degree of $w_j(x)$ is j. Then, $w_j(x)$ has the form $a_{j,0}x^j + a_{j,2}x^{j-2} + \cdots + a_{j,2\lfloor \frac{j}{2} \rfloor} x^{j-2\lfloor \frac{j}{2} \rfloor}$, for $\forall j \in N$ (where $\lfloor\ \rfloor$ is the floor function, that is, it rounds down to the nearest integer).

Proof. The $w_j(x)$ are computed recursively by using the Gram-Schmidt orthogonalization and the inner product in (28):

$$w_0(x) = 1, \qquad w_j(x) = \frac{g_j(x) - \sum_{i=0}^{j-1} \langle g_j, w_i \rangle w_i(x)}{\| g_j(x) - \sum_{i=0}^{j-1} \langle g_j, w_i \rangle w_i(x) \|}, \tag{29}$$

where $g_j(x) = x^j$. Note that for any non-negative integers j and h,

$$\langle g_j, g_h \rangle = \langle x^j, x^h \rangle = \int_\Omega [x^{j+h} + jhx^{j+h-2}]\rho(x)dx = 0 \tag{30}$$

if j and h have opposite parity, since ρ is an even function. The proof of this theorem proceeds by induction.

For the cases $j = 1, 2$, property (30) implies that

$$w_1(x) = \frac{x - \langle x, 1\rangle}{\|x - \langle x, 1\rangle\|} = \frac{x}{\|x\|} = a_{1,0}x, \tag{31}$$

$$w_2(x) = \frac{x^2 - \big(\langle x^2, 1\rangle + \langle x^2, a_{1,0}x\rangle a_{1,0}x\big)}{\|x^2 - (\langle x^2, 1\rangle + \langle x^2, a_{1,0}x\rangle a_{1,0}x)\,\|} = \frac{x^2 - \langle x^2, 1\rangle}{\|x^2 - \langle x^2, 1\rangle\|} = a_{2,0}x^2 + a_{2,2}, \tag{32}$$

both of which satisfy the conclusion of the theorem.

Now assume that $w_j(x) = a_{j,0}x^j + a_{j,2}x^{j-2} + \cdots + a_{j,2\lfloor \frac{j}{2} \rfloor}x^{j-2\lfloor \frac{j}{2} \rfloor}$ for $j < n$. Since $\langle g_j, g_h\rangle = 0$ for j and h of opposite parity, it also follows that $\langle g_j, w_h\rangle = \langle x^j, w_h\rangle = 0$ for $h < n$ and j and h of opposite parity. Then, for $j = n + 1$, by definition of the Gram-Schmidt orthogonalization it follows that

$$\begin{aligned} w_n(x) &= \frac{x^n - \sum_{j=1}^{n-1}\langle x^n, w_j\rangle w_j(x)}{\|x^n - \sum_{j=1}^{n-1}\langle x^n, w_j\rangle w_j(x)\|} \\ &\propto x^n - \sum_{\substack{0\le j<n \\ j \ \&\ n \text{ same parity}}} \langle x^n, w_j\rangle w_j(x) \\ &= a_{n,0}x^n + a_{n,2}x^{n-2} + \cdots + a_{n,2\lfloor \frac{n}{2} \rfloor}x^{n-2\lfloor \frac{n}{2} \rfloor}. \end{aligned}$$

So, the statement is proved by induction. □

Corollary 1. []
For the inner product defined in (28) and orthonormal basis $w_0, w_1, w_2, \ldots$ defined above, the two components of the inner product, $\int_\Omega w_i(x)w_j(x)\rho(x)dx$ and $\int_\Omega w_i'(x)w_j'(x)\rho(x)dx$, both vanish if i and j are of different parity.

Proof. By Theorem 1 w_j is a sum of terms of the form $a_{j,j-h}x^h$, where j and h have the same parity. Thus, $\int_\Omega w_i(x)w_j(x)\rho(x)dx$ and $\int_\Omega w_i'(x)w_j'(x)\rho(x)dx$ may both be written as integrals of monomials x^h, where h has the parity of $i + j$. If $i + j$ is odd, then the integrals vanish because ρ is symmetric, as was noted in the derivation of (30). □

We now tackle the issue of the restrictions on the polynomial degree that allow for the definition of a tensor-product orthogonal polynomial basis for the inner product (28). Sufficient conditions for such a basis to exist are provided by the following Theorem 2.

Theorem 2. []
Consider the set of multivariate polynomials $\{w_{\mathbf{p}} : w_{\mathbf{p}}(\mathbf{x}) = c_{\mathbf{p}} \prod_{j=1}^d w_{j,p_j}(x_j),\ \mathbf{p} \in \Gamma\}$. Here $\{w_{j,p}\}_{p=0}^\infty$ is the set of orthogonal univariate polynomials constructed according to Theorem 1 using the symmetric probability density ρ_j defined on the domain Ω_j. Also, $\Gamma \subset \mathbb{N}_0^d$ is the set of possible indices (degrees) of the multivariate polynomials, where $\mathbf{p} = (p_1, p_2, \cdots, p_d)^T$ is one such index set. Moreover, $c_{\mathbf{p}}$ is the normalizing factor to make $\|w_{\mathbf{p}}\| = 1$; and the inner product, (28), for these multivariate polynomials is defined by the product density function $\rho(x) = \prod_{j=1}^d \rho_j(x_j)$ defined on the Cartesian product sample space $\mathbf{\Omega} = \Omega_1 \times \Omega_2 \times \cdots \times \Omega_d$. Under these assumptions, if $\mathbf{\Psi} = (\Psi_1, \cdots, \Psi_k)^T$ is the vector basis whose elements are taken from the above set of multivariate

polynomials, then these Ψ_l are orthonormal, that is, $\langle \Psi_l, \Psi_l' \rangle = \delta_{l,l'}$, provided that the index set Γ satisfies the following condition.

For all *distinct* pairs of indices, $\mathbf{p}, \mathbf{q} \in \Gamma$ there exists some $i \in \{1, 2, \cdots, d\}$ such that one of the following criteria is satisfied:

i. The polynomials $w_{\mathbf{p}}$ and $w_{\mathbf{q}}$ are univariate polynomials of x_i, i.e., $p_i \neq q_i$, $p_j = q_j = 0$, for $\forall j \neq i$.

ii. One of the two polynomials $w_{\mathbf{p}}$ or $w_{\mathbf{q}}$ does not depend on x_i, while the other does, i.e., $p_i = 0 \neq q_i$ or $p_i \neq 0 = q_i$.

iii. The two polynomials $w_{\mathbf{p}}$ or $w_{\mathbf{q}}$ have different parity in the variable x_i, i.e., p_i and q_i have opposite parity.

Proof. The proof proceeds by showing that $\langle w_{\mathbf{p}}, w_{\mathbf{q}} \rangle = 0$ for any $\mathbf{p}$ and $\mathbf{q}$ matching the criteria above.

Case i: Since $w_{\mathbf{p}}(\mathbf{x}) = w_{i,p_i}(x_i)$ and $w_{\mathbf{q}}(\mathbf{x}) = w_{i,q_i}(x_i)$, it follows that $\langle w_{\mathbf{p}}, w_{\mathbf{q}} \rangle = \langle w_{i,p_i}, w_{i,q_i} \rangle = 0$.

Case ii: Without loss of generality, we can take $p_i = 0 \neq q_i$. The inner product $\langle w_{\mathbf{p}}, w_{\mathbf{q}} \rangle$ is shown to vanish by showing that each term in its definition in (28) vanishes. In particular, the orthogonality of the univariate polynomials $w_{i,0}$ and w_{i,q_i} implies that

$$\int_{\boldsymbol{\Omega}} w_{\mathbf{p}}(\mathbf{x}) w_{\mathbf{q}}(\mathbf{x}) \rho(\mathbf{x}) d\mathbf{x} \propto \int_{\Omega_i} w_{i,0}(x) w_{i,q_i}(x) \rho(x) dx = \langle w_{i,0}, w_{i,q_i} \rangle = 0,$$

$$\int_{\boldsymbol{\Omega}} \frac{\partial w_{\mathbf{p}}}{\partial x_s}(\mathbf{x}) \frac{\partial w_{\mathbf{q}}}{\partial x_s}(\mathbf{x}) \rho(\mathbf{x}) d\mathbf{x} \propto \begin{cases} \int_{\Omega_i} w_{i,0}'(x_i) w_{i,q_i}'(x_i) \rho(x_i) dx_i = 0, & i = s, \quad \text{since } w_{i,0}'(x_i) = 0, \\ \int_{\Omega_i} w_{i,0}(x_i) w_{i,q_i}(x_i) \rho(x_i) dx_i = 0, & i \neq s. \end{cases}$$

It then follows that every term in the definition of $\langle w_{\mathbf{p}}, w_{\mathbf{q}} \rangle$ vanishes, so $\langle w_{\mathbf{p}}, w_{\mathbf{q}} \rangle = 0$.

Case iii: According to Corollary 1,

$$\int_{\Omega_i} w_{i,p_i}(x) w_{i,q_i}(x) \rho_i(x) dx = 0, \quad \text{and} \quad \int_{\Omega_i} w_{i,p_i}'(x) w_{i,q_i}'(x) \rho_i(x) dx = 0.$$

By the same argument used for case ii, it follows that $\langle w_{\mathbf{p}}, w_{\mathbf{q}} \rangle = 0$. □

Note that case iii. includes the case where $p_i = 1$ and $q_i = 2$ but not the case where $p_i = 1$ and $q_i = 3$. Some simple characterizations of sets of polynomial indices (degrees) that satisfy Theorem 2 are given by Corollary 2. The proof of this corollary follows by checking that Γ_3 satisfies the conditions of Theorem 2 and noting that Γ_3 is a superset of Γ_1 and Γ_2.

Corollary 2. []
The following choices of Γ all satisfy the satisfy the criteria of Theorem 2 that guarantees an orthonormal multivariate basis:

$$\Gamma_1 = \{\mathbf{p} \in \mathbb{N}_0^d : \|\mathbf{p}\|_1 \leq 3\}, \qquad \Gamma_2 = \{\mathbf{p} \in \mathbb{N}_0^d : \|\mathbf{p}\|_\infty \leq 2\},$$
$$\Gamma_3 = \Gamma_2 \cup \{\mathbf{p} = (0, \ldots, 0, p_i, 0, \ldots, 0)^T \in \mathbb{N}_0^d : i = 1, \ldots, d\}.$$

The sets in Corollary 2 are almost the best obtainable for practical purposes. Indeed, we note that the example provided at the beginning of Section 3.3 shows the the impossibility of constructing tensor product bases with all $\mathbf{p}$ satisfying either $\|\mathbf{p}\|_1 \leq 4$ or $\|\mathbf{p}\|_\infty \leq 3$.

We now discuss how the orthogonal basis is affected by rescaling. This issue is important because many times the parameters of interest have completely different physical units, yet they will be modeled on some reference domain, making rescaling necessary. It has been assumed in Section 3.1 that the function values and the first order partial

derivatives all have the same variance, σ^2. This assumption depends on the scaling of the variables x_i. Using a different scaling (different units) changes the first-order partial derivative by a constant and changes its variance accordingly. However, the main conclusions of Theorem 2 still hold under different scalings.

Theorem 3. **[ ]**
For $j = 1, \ldots, d$, let $\widetilde{\Omega}_j$, $\tilde{\rho}_j$, and $\{w_{j,p}\}_{j=0}^{\infty}$ be reference domains, probability density functions, and sequences of orthogonal polynomials satisfying the hypotheses of Theorem 2, respectively. For $j = 1, \ldots, d$, choose any $a_j > 0$ and any b_j to define new rescaled domains, probability density functions, and sets of multivariate polynomials:

$$\Omega_j = \left\{ x : \frac{x - b_j}{a_j} \in \widetilde{\Omega}_j \right\}, \quad \rho_j(x) = \frac{1}{a_j} \tilde{\rho}_j \left(\frac{x - b_j}{a_j} \right), \quad \left\{ v_{\mathbf{p}} : v_{\mathbf{p}}(\mathbf{x}) = c_{\mathbf{p}} \prod_{j=1}^{d} w_{j,p_j} \left(\frac{x_j - b_j}{a_j} \right), \ \mathbf{p} \in \Gamma \right\}.$$

Here $c_{\mathbf{p}}$ is the same constant as in Theorem 2, and the set Γ satisfies the same condition as in Theorem 2. Using the scaling constants a_j, redefine the operator L as:

$$\mathbf{L}_{\mathbf{x}} f = \left(f(\mathbf{x}), a_1 \frac{\partial f}{\partial x_1}(\mathbf{x}), \cdots, a_d \frac{\partial f}{\partial x_d}(\mathbf{x}) \right)^T, \tag{33}$$

$$\mathbf{L}_{\mathbf{x}} \mathbf{f}^T = \begin{pmatrix} f_1(\mathbf{x}) & f_2(\mathbf{x}) & \cdots & f_l(\mathbf{x}) \\ a_1 \frac{\partial f_1}{\partial x_1}(\mathbf{x}) & a_1 \frac{\partial f_2}{\partial x_1}(\mathbf{x}) & \cdots & a_1 \frac{\partial f_l}{\partial x_1}(\mathbf{x}) \\ \vdots & \vdots & \vdots & \vdots \\ a_d \frac{\partial f_1}{\partial x_d}(\mathbf{x}) & a_d \frac{\partial f_2}{\partial x_d}(\mathbf{x}) & \cdots & a_d \frac{\partial f_l}{\partial x_d}(\mathbf{x}) \end{pmatrix}. \tag{34}$$

Use this rescaled operator to redefine the inner product in (28) as

$$\langle f, g \rangle = \int_{\mathbf{\Omega}} \mathbf{L}_{\mathbf{x}} f \left(\mathbf{L}_{\mathbf{x}} g \right)^T \rho(\mathbf{x}) d\mathbf{x} = \int_{\mathbf{\Omega}} \left(f(\mathbf{x}) g(\mathbf{x}) + \sum_{i=1}^{d} a_i^2 \frac{\partial f}{\partial x_i}(\mathbf{x}) \frac{\partial g}{\partial x_i}(\mathbf{x}) \right) \rho(\mathbf{x}) d\mathbf{x}. \tag{35}$$

if the above set of multivariate polynomials is orthonormal with respect to this new inner product.

Proof. The proof proceeds by verifying that the inner product of two multivariate polynomials from this theorem, $v_{\mathbf{p}}$ and $v_{\mathbf{q}}$, equals the inner product of two multivariate polynomials from Theorem 2 by a change of variable. Since

$$\frac{\partial v_{\mathbf{p}}}{\partial x_k}(x) = c_{\mathbf{p}} \frac{1}{a_k} w'_{k,p_k} \left(\frac{x_k - b_k}{a_k} \right) \prod_{j=1,\, j \neq k}^{d} w_{j,p_j} \left(\frac{x_j - b_j}{a_j} \right),$$

it follows that

$$\begin{aligned}\langle v_{\mathbf{p}}, v_{\mathbf{q}}\rangle &= \int_{\mathbf{\Omega}} c_{\mathbf{p}} c_{\mathbf{q}} \left(\prod_{j=1}^{d} w_{j,p_j}\left(\frac{x_j - b_j}{a_j}\right) w_{j,q_j}\left(\frac{x_j - b_j}{a_j}\right) \right.\\ &\quad \left. + \sum_{k=1}^{d} w'_{k,p_k}\left(\frac{x_k - b_k}{a_k}\right) w'_{k,q_k}\left(\frac{x_k - b_k}{a_k}\right) \prod_{j=1,\, j\neq k}^{d} w_{j,p_j}\left(\frac{x_j - b_j}{a_j}\right) w_{j,q_j}\left(\frac{x_j - b_j}{a_j}\right) \right) \prod_{j=1}^{d} \rho_j(x_j) d\mathbf{x}\\ &= \int_{\tilde{\mathbf{\Omega}}} c_{\mathbf{p}} c_{\mathbf{q}} \left(\prod_{j=1}^{d} w_{j,p_j}(u_j) w_{j,q_j}(u_j) \right.\\ &\qquad \left. + \sum_{k=1}^{d} w'_{k,p_k}(u_k) w'_{k,q_k}(u_k) \prod_{j=1,\, j\neq k}^{d} w_{j,p_j}(u_j) w_{j,q_j}(u_j) \right) \prod_{j=1}^{d} \tilde{\rho}(u_j) d\mathbf{u}\\ &= \langle w_{\mathbf{p}}, w_{\mathbf{q}}\rangle = \delta_{\mathbf{p},\mathbf{q}},\end{aligned}$$

where the first inner product is the rescaled one in (35) and the second one is the original one in (28).

□

3.4 Construction of Orthogonal Bases

Provided that the required degrees of the multivariate polynomials satisfy the conditions in Theorems 2 and 3, one can always construct a basis of orthogonal multivariate polynomials as tensor products of orthogonal univariate polynomials. Given a family of distributions (e.g., uniform), and a reference domain (e.g., $[-1, 1]$), Theorem 3 may then be used to construct the multivariate tensor product orthogonal basis even if the domain is stretched, shrunk, or translated. At the same time, Theorem 3 also describes how to adjust the collocation matrix to account for the rescaling.

For example, suppose that $d = 2$, variable x_1 is uniformly distributed on $[-0.5, 0.5]$, and variable x_2 is uniformly distributed on $[-1, 1]$. The univariate orthogonal polynomials with respect to the uniform distribution on $[-1, 1]$ and inner product (28) are

$$1, \quad x, \quad x^2 - \frac{1}{3}, \quad x^3 - \frac{9}{10}x, \quad \ldots.$$

If the total degree of the multivariate polynomial is no larger than 3, then by Theorem 3 one may obtain the following vector of orthogonal basis functions with respect to the inner product (35):

$$\begin{aligned}\mathbf{\Psi}^T = (\Psi_1, \ldots, \Psi_{10}) = \Bigg(1,\ \frac{x_1}{0.5},\ x_2,\ \left(\frac{x_1}{0.5}\right)^2 - \frac{1}{3},\ \frac{x_1 x_2}{0.5},\ x_2^2 - \frac{1}{3},\\ \left(\frac{x_1}{0.5}\right)^3 - \frac{9}{10}\left(\frac{x_1}{0.5}\right),\ \frac{x_1}{0.5}\left(x_2^2 - \frac{1}{3}\right),\ \left(\left(\frac{x_1}{0.5}\right)^2 - \frac{1}{3}\right) x_2,\ x_2^3 - \frac{9}{10}x_2.\Bigg)\end{aligned}$$

Correspondingly, the collocation matrix and response vector should be adjusted as

$$\mathbf{F} = \begin{pmatrix} \Psi_1(S_1) & \Psi_2(S_1) & \cdots & \Psi_{10}(S_1) \\ 0.5\frac{\partial\Psi_1}{\partial x_1}(S_1) & 0.5\frac{\partial\Psi_2}{\partial x_1}(S_1) & \cdots & 0.5\frac{\partial\Psi_{10}}{\partial x_1}(S_1) \\ \frac{\partial\Psi_1}{\partial x_2}(S_1) & \frac{\partial\Psi_2}{\partial x_2}(S_1) & \cdots & \frac{\partial\Psi_{10}}{\partial x_2}(S_1) \\ \Psi_1(S_2) & \Psi_2(S_2) & \cdots & \Psi_{10}(S_2) \\ \vdots & \vdots & \vdots & \vdots \\ 0.5\frac{\partial\Psi_1}{\partial x_1}(S_m) & 0.5\frac{\partial\Psi_2}{\partial x_1}(S_m) & \cdots & 0.5\frac{\partial\Psi_{10}}{\partial x_1}(S_m) \\ \frac{\partial\Psi_1}{\partial x_2}(S_m) & \frac{\partial\Psi_2}{\partial x_2}(S_m) & \cdots & \frac{\partial\Psi_{10}}{\partial x_2}(S_m) \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} J(S_1) \\ 0.5\frac{\partial J(S_1)}{\partial x_1} \\ \frac{\partial J(S_1)}{\partial x_2} \\ J(S_2) \\ \vdots \\ 0.5\frac{\partial J(S_m)}{\partial x_1} \\ \frac{\partial J(S_m)}{\partial x_2} \end{pmatrix}. \tag{36}$$

Naturally, the conditions in Theorem 2 are somewhat restrictive, since they limit the degree of polynomials that can be used while still retaining orthogonality and the tensor product structure. However, to introduce polynomials of higher degree requires that we give up either orthogonality or the tensor product structure. Giving up the former may lead to the situation where the estimates of pairs of regression coefficients are highly correlated. Giving up the latter makes it awkward to remove one variable from the model without adversely affecting the dependence of the model on other variables. In practice, the restriction on the degree may not be too limiting since given a maximum total degree allowed of p, the number of possible polynomials increases as $O(d^p)$ as the dimension, d, tends to infinity. On the other hand, the number of polynomials used should not exceed the number of observations available, namely $m(d+1)$. In our numerical results in Sections 4.2 and 4.3 we will consider only the tensor product basis produced by PRD in order to assess its properties and potential.

4. NUMERICAL RESULTS

In this section, we investigate the results of using the tensor product basis developed in Section 3 with the PRD approach.

4.1 Applied Problem

As an applied example, we use a three-dimensional, steady-state reactor core model with uniform fuel elements, simple heat transport description (including convection and diffusion), uniform liquid coolant flow, and no control mechanisms. While our research extends to more complex systems, the idea was to work with a model that exhibits behavior typical for real-world nuclear reactors in sufficient measure to study uncertainty propagation and that avoids model-specific complexities of the nuclear reactor analysis. The operational parameters of the model were chosen to correspond to a sodium-cooled fast reactor with realistic temperature excursions. A cross section of the finite-volume geometric representation (with just 7 pins) is shown in Figure 1. In the following paragraphs we briefly describe the physical model. A more detailed description of this model is provided in [8].

To model uncertainty related to thermo-hydraulic description of the reactor core, we couple a three-dimensional heat conduction and convection equation

$$0 = -\nabla \cdot K\nabla T - \rho c_p \overrightarrow{u}\nabla T + q''' \tag{37}$$

represented by

$$0 = \int_{\partial\Omega} K\nabla\mathbf{T} \cdot \overrightarrow{n}\, dS + \int_{\partial\Omega} \rho c_p \mathbf{T}\vec{u} \cdot \vec{n} dS - \int_{\Omega} q''' dV \tag{38}$$

in every volume cell Ω with the dependencies of the material properties (heat conductivity in fuel and coolant K,

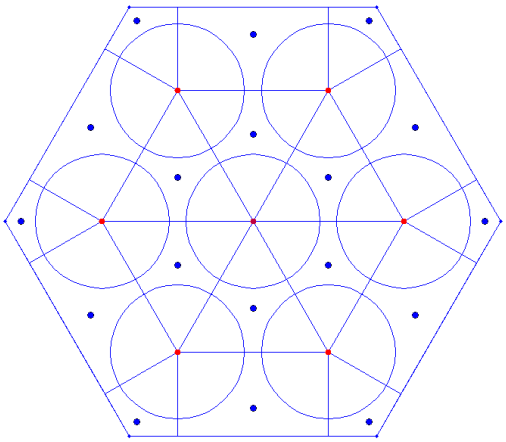

FIG. 1: Simplified 3-D model of the reactor core

specific coolant heat c_p , heat transfer coefficient h, and coolant density ρ) on temperature:

$$(K, c_p, h, \rho) = \mathbf{R} = \mathbf{R}(\mathbf{T}) \cdot (1 + \Delta\mathbf{R}(\mathbf{T}, \mathbf{x})) \tag{39}$$

with the error-free dependency functions $\mathbf{R}(\mathbf{T})$ taken from the available materials properties references [34, 35]. The coolant flow $\vec{u}$ and heat source term q''' were calibrated to represent a realistic situation. The heat transfer coefficient h appears in the discretization of $\nabla\mathbf{T}$ over the boundary between fuel and coolant.

We use a fairly complex uncertainty structure in which the uncertainty quantifiers are dimensionless coefficients in the representations of the dependency of material properties on temperature:

$$\Delta R(x, T) = x_0 \cdot C_0(T) + x_1 \cdot C_1(T) + x_2 \cdot C_2(T) \tag{40}$$

$$C_0(T) = 1; C_1(T) = T + 1; C_2(T) = 2T^2 - 1 \tag{41}$$

thus resulting in three uncertainty quantifiers per physical parameter; the dimension of the uncertainty space is 12.

Furthermore, x_0, x_1, x_2 are randomly distributed with a probability density function estimated from the published data [34, 35]. As a result, the experimental error ΔR in measurement of a material property R is, in this case, not randomly and uniformly distributed over the geometry of the reactor. Instead, it depends on temperature (and that dependence itself is uncertain as quantified by the parameters x_0, x_1, and x_2). Other expressions for uncertainty representation are, of course, possible. Our main approach admits any structure as long as the derivative $\frac{\partial J(x)}{\partial x}$ can be computed.

The solution of the coupled system is resolved. In the central fuel element we use a finer three-dimensional grid for evaluating the temperature distribution T_{pin}:

$$-\nabla \cdot K \nabla T_{pin} + q''' = 0. \tag{42}$$

We chose the maximal fuel centerline temperature as a merit function. We note that, over the continous spatial coordinates, this function is differentiable as long as the maximum is unique and regular in an optimization sense. Nevertheless, to protect against nondifferentiability that may be induced by the discretization of the partial differential equations, we use an approximation with another vector norm: $J(T) = max(T_{centerline}) \approx \|T_{centerline}\|_{100}$. The

approximation is differentiable since the argument of the norm never approaches zero.

For this model, the gradient information was obtained by direct coding. We are currently actively investigating the use of automatic differentiation techniques to obtain gradients for our method when applied to nuclear engineering applications [10].

4.2 Using the Tensor-Product Orthogonal Basis within Stepwise Regression

Once we produce an orthogonal basis as described in Section 3, an important issue is how to harness its potential promised by our analysis in Section 3.2. In particular, we are interested in identifying regression procedures using this basis, which has a small generalization error; that is, we seek procedures that do well on data on which they have not been trained.

Since small generalization error is connected with the ability to fit a model well on a small set of predictors [36], a natural question to ask is, what is the best subset of this basis that will predict the output? Stepwise regression [37] gives an approach to truncate the basis. It is a systematic method for adding and removing terms from a multilinear model based on their statistical significance in a regression, based on hypothesis testing, such as F and t tests [37].

For the model $\mathbf{L}_{\mathbf{x}}J = \mathbf{L}_{\mathbf{x}}\mathbf{\Psi}^T\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, stepwise regression is based on the F-test. The method begins with an initial model and then compares the explanatory power of incrementally larger and smaller models. At each step, we compute the F-statistic of each coefficient and then compute the p-value, which is the probability with respect to the F distribution to test models with and without a candidate term. If a term is not currently in the model, the null hypothesis is that the term would have a zero coefficient if added to the model. If there is sufficient evidence to reject the null hypothesis, the term is added to the model. Conversely, if a term is currently in the model, the null hypothesis is that the term has a zero coefficient. If there is insufficient evidence to reject the null hypothesis, the term is removed from the model. The method proceeds as follows.

Step 1 Fit the initial model.

Step 2 If any terms not in the model have p-values less than an entrance tolerance (that is, if it is unlikely that they would have zero coefficient if added to the model), add the one with the smallest p-value, and repeat this step; otherwise, go to step 3.

Step 3 If any terms in the model have p-values greater than an exit tolerance (that is, if it is unlikely that the hypothesis of a zero coefficient can be rejected), remove the one with the largest p-value, and go to Step 2; otherwise, end.

Depending on the terms included in the initial model and the order in which terms are moved in and out, the method may build different models from the same set of potential terms. The method terminates when no single step improves the model. There is no guarantee, however, that a different initial model or a different sequence of steps will not lead to a better fit. In this sense, stepwise models are locally optimal but may not be globally optimal.

Stepwise regression performs the modeling by analyzing a large number of terms, and selecting those that fit well. Thus, the F-values for the selected terms are likely to be significant, and hypothesis testing loses its inference power. If the objective of modelling is to test the validity of a relationship between certain terms or to test the significance of a particular term, stepwise regression is not recommended [37]. If the objective is to predict, however, stepwise regression is a convenient procedure for selecting terms, especially when a large number of terms are to be considered. We believe this to be one of the best variable selection procedures.

In our numerical experiments, we consider an applied model with uncertainty space dimension 12. The number of multivariate polynomials of maximal total degree 3 on 12 variables is 455. We use the `stepwisefit` function in Matlab (implementing an algorithm from [38]), and define the p-value for basis to enter and exit as 0.05. We tried both starting with nothing (no polynomials) in the model and everything (all 455 polynomials) in the model. We use 36 sample points for collocation and use 18 sample points for model testing.

For the orthogonal basis obtained from Corollary 2, starting with nothing in the model, we got 65 polynomials in the final model, while starting with all the 455 polynomials in the model, we got 371 polynomials in the final model. Figure 2 shows the function value errors, with and without basis truncation, with (O) standing for the "orthogonal basis case". The errors are ordered from smallest to largest. Starting with nothing in the model results in far fewer polynomials than starting with everything in the model. It also results in better estimation error for the testing data.

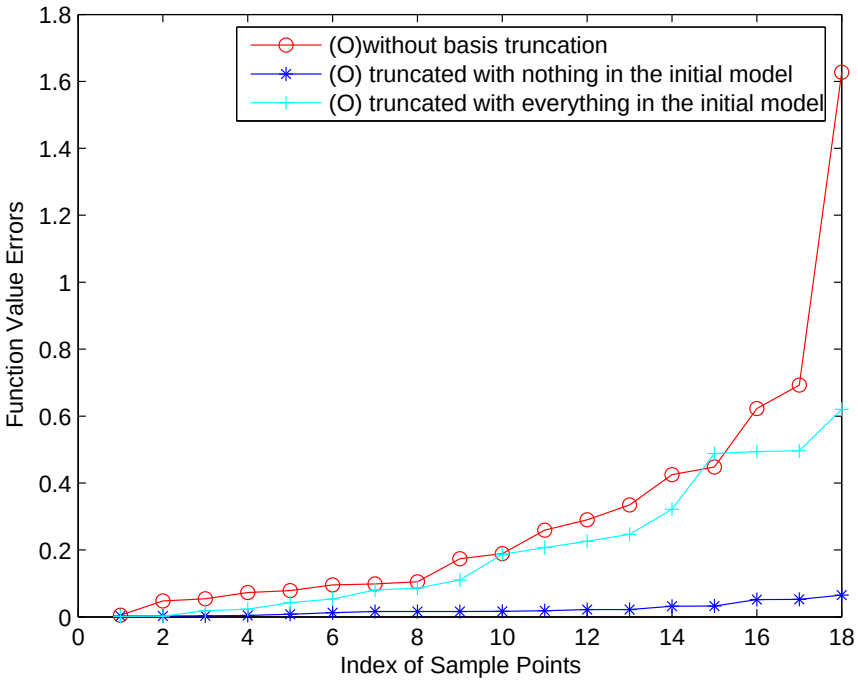

FIG. 2: Function value errors without basis truncation compared with those with basis truncations.

For Hermite polynomials, one of the choice polynomial sets used in uncertainty quantification [33], starting with nothing in the model, we got 65 polynomials in the final model, while starting with all the 455 polynomials in the model, we got 424 polynomials in the final model. Figure 3 displays function value errors, with and without basis truncation, with (H) standing for "Hermite polynomials". Again, the errors are ordered from smallest to largest. For Hermite polynomials starting with nothing in the model, the maximum error for the testing set is the smallest, but starting with everything in the model will gives a smaller error for much of the testing data.

Figure 4 compares the function value errors of truncated orthogonal basis starting with nothing in the model with those of the Hermite polynomials using the full basis and two methods of stepwise regression. We conclude from Figure 4 that stepwise regression works substantially better for the orthogonal basis compared with the Hermite basis, resulting in better estimates (by more than an order of magnitude) and fewer polynomials in the final model.

4.3 Quality of the Information Matrix

To determine the quality of the information matrix, we assume a uniform distribution for each of the 12 parameters of our model from Section 4.1. The range of values of each uniform marginal distribution is obtained by matching it with the mean and variance of each parameter from a full nuclear reactor simulation model. This uniform distribution

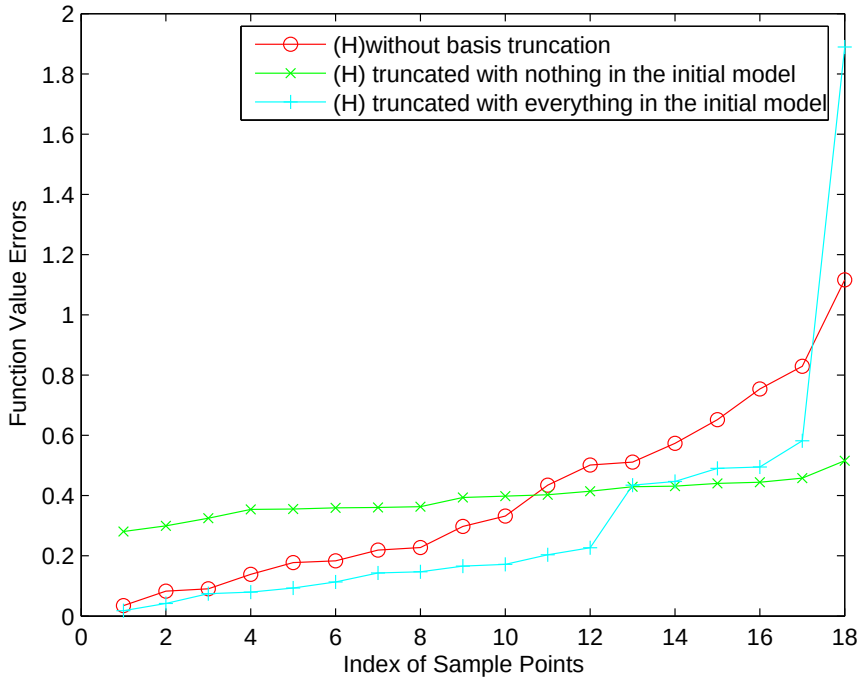

FIG. 3: Function value errors without basis truncation compared with those with basis truncations.

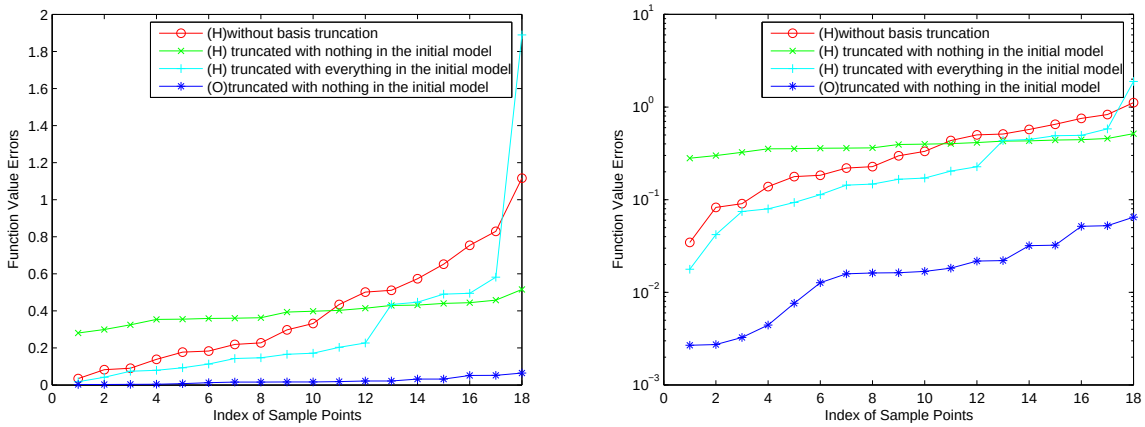

FIG. 4: Function value errors for the truncated orthogonal basis compared with those of Hermite polynomials.

does not necessarily match the full multivariate distribution of the full nuclear reactor simulation model, but it is more convenient to work with. Starting from a uniform distribution on $[-1,1]$, we use the following values of the scaling and shift parameters, as described in Theorem 3, where $\mathbf{A} = (a_1, \cdots, a_{12})^T$, and $\mathbf{B} = (b_1, \cdots, b_{12})^T$:

$$(\mathbf{A}, \mathbf{B}) = \begin{pmatrix} 0.0094 & 0 \\ 0.0094 & 0 \\ 0.0097 & 0 \\ 0.0807 & 0 \\ 0.0819 & 0 \\ 0.0865 & 0 \\ 0.0734 & 0 \\ 0.0768 & 0 \\ 0.0841 & 0 \\ 1.9868 \times 10^{-5} & 0 \\ 1.8514 \times 10^{-5} & 0 \\ 1.9047 \times 10^{-5} & 0 \end{pmatrix}. \tag{43}$$

Using this experimental setup, we analyze the properties of the information matrix $\mathbf{F}^T\mathbf{F}$. As described in Section 3.1 we expect the properties of this matrix to be a good indicator of the performance of the model. In particular, we would like this matrix to be at a substantial distance from singularity and as close to identity as possible (though, for random designs like the ones considered here, this can be achieved only in the limit of infinite number of sample points; whereas we will use the approach for a relatively small number of samples). Using the Hermite polynomial basis with 455 polynomials, with 36 sample points including function and derivative information, we observed that the numerical rank of the collocation matrix is 433, which means the information matrix is singular. The condition number of the collocation matrix is 1.9806×10^{17}.

We have run the same experiment for the orthogongal basis described in Corollary 2. We define the index set based on total degree of the polynomial basis, which we require to be less than or equal to 3, that is, $||\mathbf{p}||_1 \leq 3$. Using standard distribution $U[-1,1]$ and the Gram-Schmidt method, we can get the univariate basis up to degree 3 in one dimension as: $w_0(x) = 1$, $w_1(x) = x$, $w_2(x) = x^2 - \frac{1}{3}$, $w_3(x) = x^3 - \frac{9}{10}x$.

Then, based on Corollary 2, the tensor product of the univariate basis is an orthogonal basis for the multivariate. We use the same 36 sample points as for Hermite polynomials, which means the collocation matrix will be of size 468×455, and we get the full rank collocation matrix with condition number 1.4408×10^5, a far better result compared with the Hermite polynomial case.

Figure 5 plots the log of the singular values of the collocation matrix for both our tensor product orthogonal basis and the Hermite polynomial basis. We see that, for Hermite polynomials, the singular values drop more quickly, so the corresponding information matrix will be farther away than the one for our orthogonal design. Also, we can see that for our orthogonal basis, most of the singular values are large, which means the variance of the corresponding coefficient is small.

Figure 6 displays the function value errors when a model is used with the orthogonal basis compared with the one of Hermite Polynomials. There is a visible improvement, though perhaps not as large as suggested by Figure 5. Nevertheless, the stepwise regression results in Section 4.2 conclusively demonstrated that the superior properties of the basis can lead to far better prediction results.

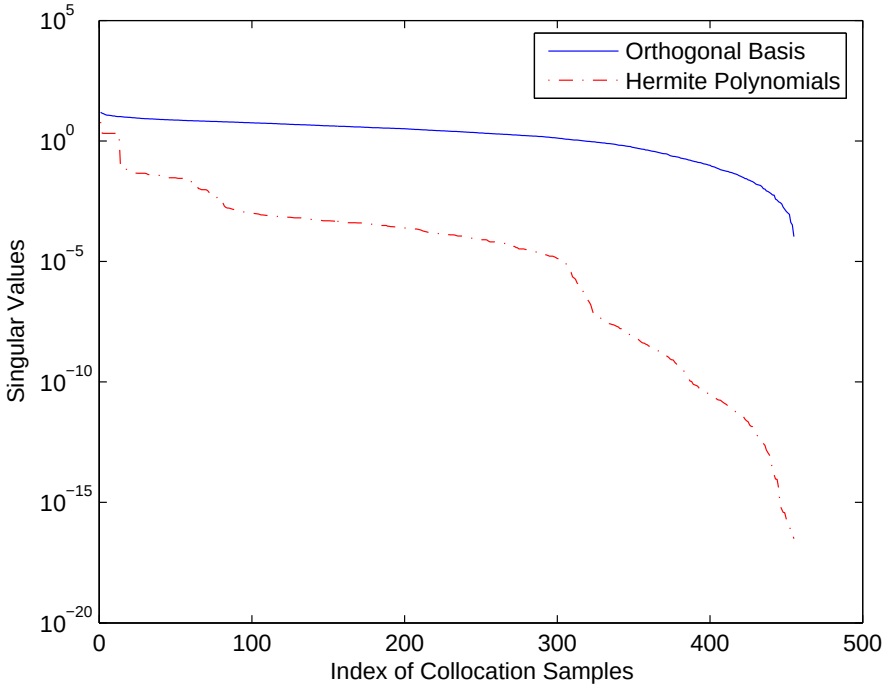

FIG. 5: Ordered singular values of the collocation matrix

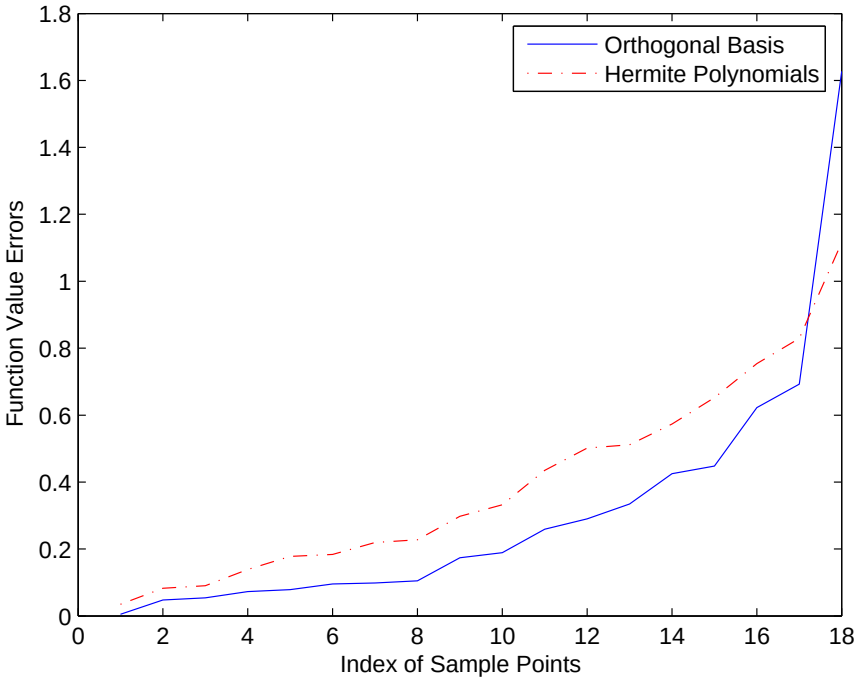

FIG. 6: Ordered function value errors at sample points reserved for model testing for orthogonal basis and Hermite polynomial; full model has 455 polynomials.

5. CONCLUSIONS

We investigate polynomial approximations to the response of a system to a multidimensional uncertainty parameter. Specifically, we investigate a regression procedure for obtaining the best polynomial approximation in the presence of function and gradient information. Such an investigation is warranted by the increased availability of gradient information, for example, by use of automatic differentiation tools.

Nevertheless, the use of gradients to approximate the system response also poses new challenges that we address in this paper. We find that the use of the Hermite polynomial basis may result in an essentially singular information matrix for the regression procedure, especially when the number of function and derivative values only slightly exceeds the number of polynomials used. We remedy this situation by deriving an orthogonal basis with respect to an $\mathcal{L}_2$-type inner product that includes both the function and its derivative.

We are interested in particular in obtaining tensor product bases. These bases give us two advantages. First, they are easy to implement, regardless of the dimension. Second, when we want to do basis truncation according to the importance of a certain variable, we can directly remove an unimportant variable without inadvertently deleting the polynomials including important variables. We proved here that such bases can be obtained under some restriction of the maximum degree of the multivariate polynomials.

Numerical experiments demonstrate that the tensor product orthogonal bases constructed here result in substantially better-conditioned information matrices. In addition, stepwise regression performs much better using this new basis in terms of obtaining a smaller error in predicting function values and in a more parsimonious model. These findings are validated by using a nuclear reactor core simulation example.

The work presented here needs to be expanded in several directions in order to increase its generality. In this paper we have considered only random designs for sampling. A better-conditioned information matrix and more accurate function approximation might be obtained by choosing a more uniform design In the application discussed here, such designs must must constructed on somewhat nonrectangular domains. Another area for further study is model selection. The numerical experiments suggest that pruning of the basis leads to better model prediction. Besides the stepwise selection procedure used here, one might consider a shrinkage method, such as LASSO. This method chooses the regression coefficient to minimize an ℓ_1-penalized least squares:

$$\widehat{\boldsymbol{\beta}}^{\text{lasso}} = \underset{\mathbf{b}}{\operatorname{argmin}} \left\{ \|\mathbf{y} - \mathbf{F}\mathbf{b}\|^2 + \lambda \sum_{j=2}^{k} |b_j| \right\},$$

where $\lambda \geq 0$ is a complexity parameter that controls the amount of shrinkage and k is the number of polynomials in the basis. Note that the constant term, b_1, is not part of the penalty. Choosing λ sufficiently large will cause some of the $\widehat{\beta}_j^{\text{lasso}}$s to be exactly zero. Another important question is the generic issue of error models; we believe that in our case it may make sense to assume a correlation between the errors at different sampling points as well as between function and derivative information.

ACKNOWLEDGMENTS

This work was supported by the U.S. Department of Energy under Contract DE-AC02-06CH11357.

REFERENCES

1 O. Roderick, M. Anitescu, and P. Fischer, Polynomial regression approaches using derivative information for uncertainty quantification, *Nuclear Science and Engineering*, 2010, to appear.

2 A. Griewank. A mathematical view of automatic differentiation. In *Acta Numerica*, Vol. 12, pp. 321–398. Cambridge University Press, 2003.

3 R. Ghanem and P. Spanos, *The Stochastic Finite Element Method: A Spectral Approach*, Springer, New York, 1991.

4 R. Ghanem and P. Spanos, Polynomial chaos in stochastic finite elements, *Journal of Applied Mechanics*, 57:197–202, 1990.

5 P. Spanos and R. Ghanem, Stochastic finite element expansion for random media, *Journal of Engineering Mechanics*, 115(5):1035–1053, 1989.

6 A. Griewank and A. Walther, Algorithm 799: Revolve: An implementation of checkpointing for the reverse or adjoint mode of computational differentiation, *ACM Trans. on Math. Software*, 26(1):19–45, 2000.

7 D. Cacuci and M. Ionescu-Bujor, Adjoint sensitivity analysis of reliability models with application to ifmif, In *ANS/ENS 2003 Winter Meeting, New Orleans, LA*, Vol. 89, pp. 415–416, November 2003.

8 O. Roderick, M. Anitescu, and P. Fischer, Stochastic finite element approaches using derivative information for uncertainty quantification, *Nuclear Science and Engineering*, 164(2):122–139, 2010.

9 A. Griewank, *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation*, SIAM, 2000.

10 M. Alexe, O. Roderick, M. Anitescu, J. Utke, T. Fanning, and P. Hovland., Using automatic differentiation in sensitivity analysis of nuclear simulation models, *Transactions of American Nuclear Society*, 102:235–237, 2010.

11 M. Anitescu, G. Palmiotti, W. Yang, and M. Neda, Stochastic Finite-Element Approximation of the Parametric Dependence of Eigenvalue Problem Solution, In *Proc. of Mathematics, Computation and Supercomputing in Nuclear Applications Conference*. American Nuclear Society, 2007.

12 M. Anitescu, Spectral finite-element methods for parametric constrained optimization problems, *SIAM Journal on Numerical Analysis*, 47(3):1739–1759, 2009.

13 M. K. Deb, I. M. Babuska, and J. T. Oden, Solution of stochastic partial differential equations using Galerkin finite element techniques, *Computer Methods in Applied Mechanics and Engineering*, 190(48):6359–6372, 2001.

14 D. Xiu and J. Hesthaven, High-order collocation methods for differential equations with random inputs, *SIAM Journal on Scientific Computing*, 27:1118–1139, 2005.

15 I. Babuska, F. Nobile, and R. Tempone, Reliability of computational science, *Numerical Methods for Partial Differential Equations*, 23(4):753, 2007.

16 I. Babuska, R. Tempone, and G. Zouraris, Solving elliptic boundary value problems with uncertain coefficients by the finite element method: the stochastic formulation, *Computer Methods in Applied Mechanics and Engineering*, 194(12-16):1251–1294, 2005.

17 M. Anitescu, P. Hovland, G. Palmiotti, and W.-S. Yang, Randomized quasi monte carlo sampling techniques in nuclear reactor uncertainty assessment, *Transactions of American Nuclear Society*, 96:526–527, 2007.

18 K. P. Schmitt, M. Anitescu, and D. Negrut, Efficient sampling for spatial uncertainty quantification in multibody system dynamics applications, *International Journal for Numerical Methods in Engineering*, 80(5):537–564, 2009.

19 I.-B. M. N. I. Cacuci, D.G., *Sensitivity and Uncertainty Analysis: Applications to Large-Scale Systems, Volume II*, CRC Press, 2005.

20 S. Isukapalli, A. Roy, and P. Georgopoulos, Efficient sensitivity/uncertainty analysis using the combined stochastic response surface method and automated differentiation: Application to environmental and biological systems, *Risk Analysis*, 20(5):591–602, 2000.

21 E. Solak, R. Murray-Smith, W. Leithead, D. Leith, and C. Rasmussen, Derivative Observations in Gaussian Process Models of Dynamic Systems, *Advances in Neural Information Processing Systems*, pp. 1057–1064, 2003.

22 N. Kim, H. Wang, and N. Queipo, Adaptive reduction of random variables using global sensitivity in reliability-based optimisation, *International Journal of Reliability and Safety*, 1(1):102–119, 2006.

23 O. Roderick, M. Anitescu, P. Fischer, and W.-S. Yang, Stochastic finite-element approach in nuclear reactor uncertainty quantification, *Transactions of American Nuclear Society*, 100:317–318, 2009.

24 W. Roux, N. Stander, and R. Haftka, Response surface approximations for structural optimization, *International Journal for Numerical Methods in Engineering*, 42(3):517–534, 1998.

25 H. Cheng and A. Sandu, Numerical study of uncertainty quantification techniques for implicit stiff systems, In *ACM-SE 45: Proceedings of the 45th Annual Southeast Regional Conference*, pp. 367–372, New York, NY, USA, 2007. ACM.

26 C. Frepoli, *An Overview of Westinghouse Realistic Large Break LOCA EvaluationModel,*, Science and Technology of Nuclear Installations, Hindawi Publishing Corporation, 2008.

27 D. J. Downing, R. H. Gardner, and F. O. Hoffman, An examination of response-surface methodologies for uncertainty analysis in assessment models, *Technometrics*, 27:151–163, 1985.

28 B. Boyack, R. Duffey, P. Griffith, K. Katsma, G. Lellouche, S. Levy, U. Rohatgi, G. Wilson, W. Wulff, and N. Zuber, Quantifying reactor safety margins: Part 1: An overview of the code scaling, applicability, and uncertainty evaluation methodology, Tech. Rep., Los Alamos National Lab., NM (USA), 1988.

29 G. Wilson, B. Boyack, R. Duffey, P. Griffith, K. Katsma, G. Lellouche, S. Levy, U. Rohatgi, W. Wulff, and N. Zuber, Quantifying reactor safety margins: Part 2, characterization of important contributors to uncertainty, Tech. Rep., EG and G Idaho, Inc., Idaho Falls (USA), 1988.

30 R. P. Martin and L. D. O'Dell, Areva's realistic large break loca analysis methodology, *Nuclear Engineering and Design*, 235:1713–1725, 2005.

31 J. Rawlings, S. Pantula, and D. Dickey, *Applied regression analysis: a research tool*, springer New York, 1998.

32 F. Pukelsheim, *Optimal Design of Experiments*, Society for Industrial Mathematics, 2006.

33 R. Ghanem and P. Spanos, *The Stochastic Finite Element Method: A Spectral Approach*, Springer, New York, 1991.

34 J. Fink and L. Leibowitz, Thermodynamic and transport properties of sodium liquid and vapor, Tech. Rep., ANL/RE-95/2 (Argonne, IL: Argonne National Laboratory), 1995.

35 J. Fink, Thermophysical properties of uranium dioxide, *Journal of Nuclear Materials*, 279(1):1–18, 2000.

36 J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer, 2001.

37 S.-H. Draper, N.R., *Applied Regression Analysis*, Wiley-Interscience, 1998.

38 K. Atkinson, *An Introduction to Numerical Analysis*, John Wiley, New York, 1989.