Journal of Computational Physics 228 (2009) 7159–7180

Contents lists available at ScienceDirect

Journal of Computational Physics

journal homepage: www.elsevier.com/locate/jcp

Padé–Legendre approximants for uncertainty analysis with discontinuous response surfaces

T. Chantrasmi, A. Doostan, G. Iaccarino *

Mechanical Engineering Department, Stanford University, CA 94305, USA

ARTICLE INFO

Article history:
Received 3 August 2008
Received in revised form 9 June 2009
Accepted 9 June 2009
Available online 26 June 2009

Keywords:
Uncertainty quantification
Padé–Legendre approximation
Gibbs phenomenon
Shock capturing
Dual throat nozzle
RAE2822

ABSTRACT

A novel uncertainty propagation method for problems characterized by highly non-linear or discontinuous system responses is presented. The approach is based on a Padé–Legendre (PL) formalism which does not require modifications to existing computational tools (non-intrusive approach) and it is a global method. The paper presents a novel PL method for problems in multiple dimensions, which is non-trivial in the Padé literature. In addition, a filtering procedure is developed in order to minimize the errors introduced in the approximation close to the discontinuities. The numerical examples include fluid dynamic problems characterized by shock waves: a simple dual throat nozzle problem with uncertain initial state, and the turbulent transonic flow over a transonic airfoil where the flight conditions are assumed to be uncertain. Results are presented in terms of statistics of both shock position and strength and are compared to Monte Carlo simulations.

1. Introduction

The aerodynamic performance of vehicles flying at high speed are profoundly affected by the presence of shock waves. Among the most dominant effects, the added resistance associated to the energy losses (wave drag), and the interaction of shocks and the viscous boundary layers have received considerable attention. In particular, the latter can potentially lead to massive flow separation and thus requires careful analysis even in the early design phase. Shock/boundary layer interaction is affected by a variety of factors including the conditions of the incoming flow, the surface geometry and the fluid properties within the boundary layer. It is known that changes in few parameters, namely the flight Mach and Reynolds numbers can lead to sharp non-monotonic variation of the length and characteristics of the interaction zone [13] with obvious impact on the overall flow behavior.

Numerical simulations are routinely employed to perform aerodynamic design and analysis of transonic and supersonic configurations. Reynolds averaged Navier–Stokes computations have been used in a variety of studies (see for example, [14]) identifying the effect of the incoming flow conditions but also the potential inaccuracy related to the turbulence modeling [15]. More recently, direct numerical simulation investigations have studied the details of the shock/boundary layer interaction at a more fundamental level [16].

Most numerical investigations of shock dominated flows focus on the predictive capabilities of various approaches; the operating conditions, for example, are defined according to relevant experimental studies with the objective of performing *controlled* validation. On the other hand various factors can introduce uncertainty in the measurement setup thus creating difficulties in identifying the appropriate conditions to be used in the corresponding simulations. In transonic flows, one of the most commonly used test-cases is the flow around the RAE2822 airfoil [12]. In this specific problem, the presence of the

* Corresponding author. Tel.: +1 650 723 9599; fax: +1 650 723 9617.
E-mail address: jops@stanford.edu (G. Iaccarino).

0021-9991/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jcp.2009.06.024

wind-tunnel walls is known to have an effect on the overall flow characteristics and also the inflow conditions are known in the test chamber with a limited precision – 4% uncertainty in Mach number. Given the sensitivity of the shock–boundary layer interaction it is indeed important to assess the influence of uncertainty in the *numerical* flow conditions on the predictions.

It must be noted that, when the problem is strongly non-linear such as in transonic flows, it is generally difficult to establish the sensitivity of the output of interest by performing perturbation analysis on the input parameters. In the present work, we develop a novel numerical approach to overcome this difficulty. More specifically, the objective of the proposed method is to accurately propagate the uncertainty in the problem definition on the statistics of the shock position and strength. The approach is based on a probabilistic representation of uncertainties where the *conventional* deterministic inputs are replaced by random variables. This process inherently creates additional independent variables in the problem, and therefore we typically refer to a *deterministic* domain characterized by the physical dimensions and a stochastic space identified by the random variables. In the proposed method, we construct the response surface on the combination of these physical and stochastic spaces. Clearly, the total number of dimensions is at least two and a multi-dimensional method is required.

The proposed method is not limited to applications related to high speed flows and is a general approach for the solution of stochastic partial differential equations that exhibit discontinuities or solutions with high gradients in the physical domain and/or in the stochastic dimensions.

In the problems presented herein, the location of the shocks as well as their strength are not known *a priori* and are functions of uncertain parameters which depend on the operating conditions. In the following, the proposed method is explained in details. Firstly, we present the approach in one-dimensional settings for conceptual understandings. Then we formulate its extension to multi-dimensional problems. Next, applications to discontinuous solution of the Burgers equations are presented. The final application is the analysis of the turbulent flow around the RAE2822 airfoil with uncertainty in the specification of the incoming Mach number.

2. Proposed method

Probabilistic assessment of uncertainty in computational models consists of three major phases: (i) data assimilation in which the input parameters are characterized (in terms of probability distributions) from observations and physical evidence; (ii) uncertainty propagation in which the input variabilities are propagated through the mathematical model; and (iii) certification in which the output of the numerical predictions are characterized in terms of their statistical properties and confidence bounds are derived [17].

In the present work we focus on the propagation phase and assume distributions for the input parameters. The simplest approach to obtain the output statistics in response to input distributions is the Monte Carlo method, in which a large number of independent calculations are computed; in many practical cases the number of realizations required is too large and results in prohibitively high computational cost, especially for complex fluid dynamics computations. In recent years, two alternative approaches have found relatively widespread use: stochastic Galerkin [18,34,29,22] and stochastic collocation [31,19,20,30]. Stochastic Galerkin approaches are generally based on an expansion of the random quantities in terms of suitable global (or local) basis. These schemes are *intrusive*, in the sense that the deterministic solvers are modified to incorporate the stochastic expansions. On the other hand, in the stochastic collocation method only few computations are carried out corresponding to precise specification of the input variables, typically corresponding to *classical* quadrature points; this approach is therefore *non-intrusive*.

In their *original* form, both stochastic Galerkin and stochastic collocation schemes become inaccurate in the presence of discontinuities in probability space due to the Gibbs phenomenon. A number of extensions to these two methods have been proposed to remedy the problem. Among those the basis enrichment of polynomial chaos expansions [21] in which prior knowledge about the behavior of the solution is incorporated in the selection of enriching basis functions. With an educated selection of the additional bases, the method is shown to improve the accuracy and convergence rate of the stochastic approximations. Multi-element generalized polynomial chaos (ME-gPC) [23] and multi-element probabilistic collocation method (ME-PCM) are extensions of generalized polynomial chaos and stochastic collocation technique, respectively, that can improve the accuracy of those techniques provided that an efficient *a posteriori* error estimates are available.

In ME-gPC, the stochastic domain is decomposed into small elements. The generalized polynomial chaos expansion is then employed in each of these subdomains [23] thus possibly leading to *hp*-type convergence in the probability space. The partitioning of the probability space can be performed in an iterative manner by considering relative error in the response variance [22]. In ME-PCM, similar to ME-gPC, the stochastic domain is divided into subdomains. In each subdomain, the probabilistic collocation method is then employed [24,25]. Both ME-PCM and ME-gPC are *local* methods.

In the present work, we have developed an alternative *global* approach to deal with discontinuities in probability space. One potential benefit is the decrease in the computational cost compared to a local method. The proposed approach is coupled with a stochastic collocation scheme although it is possible, in principle, to formulate it as an *intrusive* Galerkin-type approach. Our method is based on Padé approximation of discontinuous functions. A Padé approximation is a *rational* function – a ratio of two polynomials – that can be thought as a generalization of a Taylor series expansion. In deterministic context, Padé approximants have been used in a variety of fields such as network theory, optimal control and quantum

mechanics [2]. In its original form, one calculates the Padé approximant of a function from its given power series. In recent extensions on Padé approximation, one writes the numerator and denominator polynomials as finite sums of orthogonal basis polynomials whose coefficients can be calculated from the function values at a predefined set of points. Examples of these recent extensions include Padé–Jacobi [32], Padé–Chebyshev [33] and Padé–Legendre approximants [1] which are the basis of the present study.

The method proposed here is simple, efficient and non-intrusive. The first step is to perform a number of deterministic calculations with input parameters chosen as quadrature points in the stochastic dimensions based on which a PL approximant is constructed as a response surface. In the applications presented here, this surface is function of both the uncertain parameters and the physical coordinates. Finally we sample from the response surface and extract the quantities of interest and their statistics.

In the following, Legendre polynomial expansions and the required quadrature rules are briefly introduced; in-depth discussions of these topics are available in the literature, see for example [7,1]. The one-dimensional Padé–Legendre method is then presented as a conceptual starting point. Finally, a novel extension of the PL method to multiple dimensions is formulated.

2.1. Legendre expansions

Let $L^2(-1,1)$ be the Hilbert space of square integrable functions u, defined on $(-1,1)$ and equipped with the scalar product

$$\langle u, v\rangle := \int_{-1}^{1} u(x)\,v(x)\,dx. \tag{1}$$

A complete basis of $L^2(-1,1)$ is formed by the Legendre polynomial $P_n(x)$ defined by

$$\langle P_n, P_m\rangle = \frac{1}{n+1/2}\,\delta_{nm}, \quad n,\ m \in \mathbb{N}\cup\{0\}. \tag{2}$$

where δ is the Kronecker delta and $\mathbb{N}$ is the set of positive integers. Eqs. (1) and (2) uniquely define the Legendre polynomial series: $1, x, \frac{1}{2}(3x^2-1), \frac{1}{2}(5x^3-3x), \ldots$ with the first term being P_0. One can also construct Legendre polynomials from the three-term recursive relation:

$$(n+1)P_{n+1}(x) = (2n+1)xP_n(x) - nP_{n-1}(x). \tag{3}$$

It follows that every function in this Hilbert space has an L^2-convergent expansion in Legendre basis:

$$u = \sum_{n=0}^{\infty} \hat{u}_n P_n, \tag{4}$$

where $\hat{u}_n$ is the nth Legendre coefficient of u defined by $\hat{u}_n := \langle u, P_n\rangle / \langle P_n, P_n\rangle$. For $N \in \mathbb{N}$, one can approximate u with the truncated series of order N as $u \approx \sum_{n=0}^{N} \hat{u}_n P_n$. The rate of convergence of the error is related solely to the smoothness of u with the error estimate:

$$\left\| u - \sum_{n=0}^{N} \hat{u}_n P_n \right\| \leqslant cN^{-s}\|u\|_{H^s} \quad \forall N \in \mathbb{N}, \tag{5}$$

when u belongs to the Sobolev space $H^s(-1,1)$. Here, c is a constant independent of N [1]. It is well understood that, for a discontinuous function u, the expansion (4) converges slowly, $\mathcal{O}(N^{-1})$ for the mean, due to the Gibbs phenomenon.

2.2. Gauss quadratures

In most problems, one knows the function values only at the finite number of points in the domain. Thus, a *discrete* scalar product is required. In analogy to the continuous scalar product equation (1), the discrete scalar product of functions ϕ and ψ is defined by

$$\langle \phi, \psi\rangle_N = \sum_{j=0}^{N} \phi(x_j)\psi(x_j)w_j, \tag{6}$$

where the quadrature points x_j and the associated weights w_j are predefined.

The formulation shown above is for one-dimensional problem, however, Gauss quadratures can be easily extended to higher dimensions by using tensor products of one-dimensional Legendre polynomials as the basis. The Legendre expansions and discrete scalar products then follow in a straightforward manner.

2.3. One-dimensional Padé–Legendre approximation

Based on the discrete scalar product, the definition of a Padé–Legendre (PL) approximant for one-dimensional problems follows the one used in [1]. For any non-negative integer k, let $\mathbb{P}_k$ be the set of all polynomials whose degrees are less than or

equal to k. Let u be a function to be represented on $[-1,1]$. Given the integers M and L, the pair of polynomials $P \in \mathbb{P}_M$ and $Q \in \mathbb{P}_L$ are said to be a solution of the (N,M,L) Padé–Legendre interpolation problem of u if:

$$\langle P - Qu, \phi \rangle_N = 0 \quad \forall \phi \in \mathbb{P}_N \tag{7}$$

and

$$\forall x \in [-1,1], \quad Q(x) > 0. \tag{8}$$

The rational function $R(u) := P/Q$ is then defined as an approximation of u. If (7) and (8) admit a solution, one can show that the solution is unique provided that $M + L \leqslant N$ [1]. In addition, it can be shown that the rational function $R(u)$ is an interpolation of u, i.e.,

$$R(u)(x_j) = u(x_j), \tag{9}$$

at all quadrature points x_j [1]. In this work, we use Gauss quadrature to evaluate the discrete scalar products. Our choice is dictated by the need to achieve high-order accuracy in the computation presented below.

Remark. It is worth noting that although the PL formulation formally yields the same order of accuracy as typical polynomial expansion, it can better represent cases with singularities and discontinuities in which the ordinary polynomial expansions generally fail [2]. Intuitively, Eq. (7) shows that given a discontinuous function u, the polynomial Q represents a *pre-conditioner* that produces a smooth function Qu which then can be effectively approximated by the polynomial P.

The rational approximation $R(u)$ is constructed by first defining the functions P and Q as linear combinations of Legendre polynomials of order M and L, respectively,

$$P = \sum_{j=0}^{M} \hat{p}_j P_j, \tag{10}$$

$$Q = \sum_{j=0}^{L} \hat{q}_j P_j. \tag{11}$$

In the following, we present a procedure to compute the coefficients in the Legendre expansions of P and Q such that Eq. (7) is satisfied.

Assume $N = M + L$ from now on. First, we compute the denominator Q. From (7) and the fact that $P \in \mathbb{P}_M$, choosing ϕ to be Legendre polynomials P_n where $n > M$ gives

$$\langle Qu, P_n \rangle_N = \langle P, P_n \rangle_N = 0 \quad \forall n = M+1, \ldots, M+L. \tag{12}$$

Plugging (11) and (6) into (8), we arrive at a linear system whose solution returns $\hat{q}_j$.

$$\begin{bmatrix} \langle uP_0, P_{M+1} \rangle_N & \cdots & \langle uP_L, P_{M+1} \rangle_N \\ \vdots & \ddots & \vdots \\ \langle uP_0, P_{M+L} \rangle_N & \cdots & \langle uP_L, P_{M+L} \rangle_N \end{bmatrix} \begin{bmatrix} \hat{q}_0 \\ \vdots \\ \hat{q}_L \end{bmatrix} = \underline{0}. \tag{13}$$

We define the above matrix A and the solution vector $\mathbf{q}$. Note that A is $L \times (L+1)$ matrix. We seek a non-zero solution to $A\mathbf{q} = 0$ with $\|\mathbf{q}\| = 1$. Moreover, for computation efficiency, we decompose the matrix A into the product of three matrices $A = BCD$ with $B \in \mathbb{R}^{L,N+1}$, the diagonal matrix $C \in \mathbb{R}^{N+1,N+1}$, and $D \in \mathbb{R}^{N+1,L+1}$ defined by

$$\begin{aligned} B_{i,j} &= P_{M+i}(x_j) w_j, \quad 1 \leqslant i \leqslant L, \ 0 \leqslant j \leqslant N, \\ C_{i,j} &= u_i(x_j) \delta_{ij}, \quad 0 \leqslant i \leqslant N, \ 0 \leqslant j \leqslant N \\ D_{i,j} &= P_j(x_i), \quad 0 \leqslant i \leqslant N, \ 0 \leqslant j \leqslant L. \end{aligned} \tag{14}$$

Note that the matrices B and D only depend on the parameters N, M and L and not on the data u.

Once the denominator Q is known, the computation of the coefficients of the numerator P is straightforward, i.e.,

$$\hat{p}_n = \frac{\langle P, P_n \rangle_N}{\langle P_n, P_n \rangle_N} = \frac{\langle Qu, P_n \rangle_N}{\langle P_n, P_n \rangle_N}, \quad n = 0, 1, \ldots, M, \tag{15}$$

where the discrete inner products $\langle \cdot, \cdot \rangle_N$ is given by (6).

Remark. The above procedure does not guarantee the condition (8). To the authors' knowledge it is not possible to formally ensure this condition. However, as will be described in Section 2.6, a simple Q-dependent filter is proposed to alleviate this problem.

2.4. Multi-dimensional Padé–Legendre approximation

All the cases we consider have two or more dimensions since we construct the Padé response surface on the combination of physical and stochastic spaces. Therefore, in the examples, we always have a minimum of two dimensions.

It is not trivial to generalize the Padé formulation to multiple dimensions. In the literature, many generalizations exist but with certain restrictions on their applications. These include homogeneous Padé [9], Padé–Padé [6], nested Padé [10], etc. One of the main difficulties is the fact that in higher dimensions, there is no obvious correspondence between the number of polynomial coefficients (coefficients of the expansions of P and Q) and the number of equations one can use [8], in other words there is no equivalent relationship to the equation $N = M + L$ in the one-dimensional case. Interested readers are referred to [6,8,9] on this subject. In the present work, we restrict ourselves to problems in few dimensions (up to 4).

Out of many possible multivariate formalisms, the least-squares Padé approximation [11] is presented in this work. For the sake of simplicity, the approximation is formulated in a two-dimensional setting; however, the generalization to higher dimensions is similar. In addition, we will consider only the isotropic cases, i.e., we consider the same number of data points in each direction on a tensor grid. As mentioned earlier, these correspond to two-dimensional Gauss quadrature points. Let $N+1$ be the number of data points in each direction (we have $(N+1)^2$ total data points).

Denote the set of two-dimensional polynomial $p(x,y)$ whose total degree is less than or equal to $S \in \mathbb{N} \cup \{0\}$ as $\mathbb{P}^2_S$. Let us define two-dimensional Legendre polynomials as the product of one-dimensional polynomials. Let us also assume an ordering identified with a sub-index: $P_1, P_2, P_3, \ldots$ where the total degrees of polynomials in the sequence are in non-decreasing order. This ordering is not unique and the indices do not correspond directly to the total degrees of the polynomials. In the two-dimensional case, there are $s+1$ polynomials of degree s and there are $(s+1)(s+2)/2$ polynomials of degree less than or equal to s. For notation compactness, we define $c(s) = (s+1)(s+2)/2$.

Let $\Phi^{(a,b)}$ be the set of all two-dimensional Legendre polynomials whose total degrees are higher than a but do not exceed b, i.e.,

$$\Phi^{(a,b)} = \left\{\phi \in \mathbb{P}^2_b \setminus \mathbb{P}^2_a | \phi \text{ is a two-dimensional Legendre polynomial}\right\}. \tag{16}$$

Let us also define $\mathbf{v}^{(a,b)}$ as a vector of the same size as $\Phi^{(a,b)}$ and whose elements are

$$v_i = \langle P - Qu, \phi_i \rangle_N, \quad i = 1, 2, 3, \ldots, |\Phi^{(a,b)}|, \tag{17}$$

where ϕ_i is the ith member of $\Phi^{(a,b)}$ (the order is not important). We are now ready to state the two-dimensional Padé–Legendre problem.

Given integers $M,\ L,\ K$ and N such that $M + K \leqslant N$, the pair of polynomials $P \in \mathbb{P}^2_M$ and $Q \in \mathbb{P}^2_L$ is said to be a solution of the (N, M, L, K) two-dimensional least-squares PL approximation problem of u if

$$\langle P - Qu, \phi \rangle_N = 0 \quad \forall \phi \in \mathbb{P}^2_M, \tag{18}$$

$$\|\mathbf{v}^{(M,M+K)}\| \text{ is minimal}, \tag{19}$$

and

$$\forall (x,y) \in [-1,1] \times [-1,1], \quad Q(x,y) > 0. \tag{20}$$

Fig. 1 shows the schematic relation among the parameters N, M, and K. Each stencil point in the diagram represents the multi-index of the two-dimensional Legendre polynomial basis, ϕ_i, used in the computation. The notation $c(s)$ represents the total number of the polynomials in the triangular area up to degree s. The problem definition stated above then requires that $v_i = \langle P - Qu, \phi_i \rangle_N$ be exactly zero in the lower triangle in the diagram (Eq. (18)) and minimized in the least-squares sense in the strip of width K away from this triangle (condition (19)).

Remark. It is worth noting that we can no longer require that $\langle P - Qu, \phi \rangle_N = 0$ for all polynomials ϕ up to (total) degree N as in the one-dimensional case since there would be more constraints (equations) than unknown coefficients. Thus, the formulation is based on finding a solution that is optimal in a sense that it minimizes $\|\mathbf{v}\|$. This has a noticeable impact on the accuracy of the approximation near discontinuities and is one of the main reasons we recommend the use of filters as explained in Section 2.6.

With the above-mentioned problem definition, we are now ready to formulate the algorithm to solve for the coefficients of P and Q. The numerator P and denominator Q can then be written as

$$P(x,y) = \sum_{j=1}^{c(M)} \hat{p}_j P_j(x,y) \tag{21}$$

and

$$Q(x,y) = \sum_{j=1}^{c(L)} \hat{q}_j P_j(x,y). \tag{22}$$

The orthogonality condition holds with respect to the sub-index

$$\langle P_n, P_m \rangle = \frac{1}{(n_x + 1/2)(n_y + 1/2)} \delta_{nm}, \quad n,\ m \in \mathbb{N} \cup \{0\}, \tag{23}$$

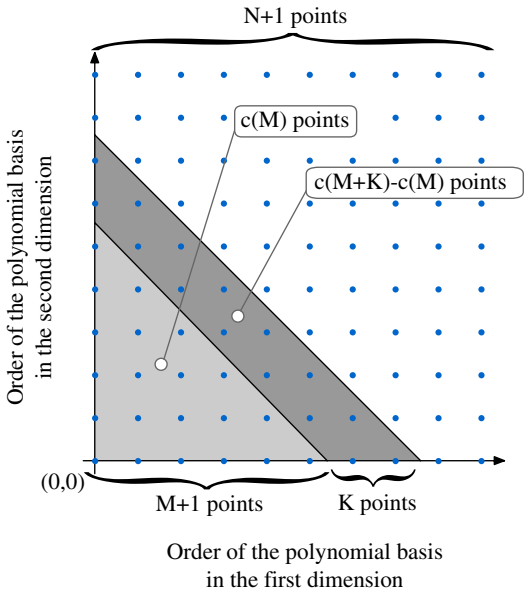


Fig. 1. The schematic representation of various parameters in the Padé–Legendre surface construction and their relationship. The parameter L is not presented in the diagram but has to satisfy the inequality $c(M+K)-c(M) > c(L)$.

where n_x and n_y are the polynomial degrees in x and y of P_n, respectively. Following a construction similar to the one-dimensional case, we obtain a linear system to solve for coefficients $\hat{q}_j$,

$$\begin{bmatrix} \langle uP_1, P_{c(M)+1}\rangle_N & \cdots & \langle uP_{c(L)}, P_{c(M)+1}\rangle_N \\ \vdots & \ddots & \vdots \\ \langle uP_1, P_{c(M+K)}\rangle_N & \cdots & \langle uP_{c(L)}, P_{c(M+K)}\rangle_N \end{bmatrix} \begin{bmatrix} \hat{q}_1 \\ \vdots \\ \hat{q}_{c(L)} \end{bmatrix} = \underline{0}. \tag{24}$$

As before, we can also decompose the matrix A into product of three matrices $A = BCD$ with

$$\begin{aligned} B_{i,j} &= P_i(x_j)w_j, \quad c(M)+1 \leqslant i \leqslant c(M+K),\ 0 \leqslant j \leqslant N, \\ C_{i,j} &= u_i(x_j)\delta_{ij}, \quad 0 \leqslant i \leqslant N,\ 0 \leqslant j \leqslant N, \\ D_{i,j} &= P_j(x_i), \quad 0 \leqslant i \leqslant N,\ 0 \leqslant j \leqslant c(L). \end{aligned} \tag{25}$$

Note that the matrix A in this case has the dimension of $(c(M+K)-c(M)) \times c(L)$. We require that the system is over-constrained:

$$c(M+K)-c(M) > c(L). \tag{26}$$

This condition is easily satisfied since L is small and M is reasonably large in the cases we consider. In fact, $K = 1$ generally satisfies (26); however, it is beneficial to keep $K \geqslant 2$, as will be explained in Section 2.5.

With condition (26), we have an over-constrained linear system of equations from which we can obtain the optimal solution $\mathbf{q}$ in the least-squares sense. Thus, the condition (19) can be restated as choosing $\mathbf{q}$ to be the solution of the minimization problem:

$$\min_{\|\mathbf{q}'\|=1} \|A\mathbf{q}'\|, \tag{27}$$

where $\|\cdot\|$ is the L_2-norm. This is easily done using the singular value decomposition of the matrix A [26]. (See Appendix A for details.) Once the denominator Q is known, the computation of the numerator coefficients is similar to the one-dimensional case:

$$\hat{p}_n = \frac{\langle P, P_n\rangle_N}{\langle P_n, P_n\rangle_N} = \frac{\langle Qu, P_n\rangle_N}{\langle P_n, P_n\rangle_N}, \quad n = 1, 2, \ldots, c(M), \tag{28}$$

thus leading to the approximation

$$u(x,y) \approx \frac{\sum_{j=1}^{c(M)} \hat{p}_j P_j(x,y)}{\sum_{j=1}^{c(L)} \hat{q}_j P_j(x,y)}, \tag{29}$$

where $\hat{q}_j$ and $\hat{p}_j$ are obtained from (27) and (28), respectively. The approximation on the right hand side of (29) is referred to as the PL response surface and is used to extract a large number of samples for computing statistics of u such as mean and variance.

2.5. A perspective on error

From the definition (7) in the one-dimensional case and (18) and (19) in higher dimensions, we see that the present method seeks to minimize the error of the linear Padé problem, $e = P - Qu$, projected on Legendre basis. In one dimension, the projected error on Legendre basis up to order N is exactly zero. In the multi-dimensional settings, the projected error on Legendre basis is zero only up to the total order M and the projected error on the Legendre basis of order from $M+1$ to $M+K$ is minimized in least-squares sense.

However, when applying the rational approximation, we are interested in the error of the non-linear Padé problem $E = P/Q - u = e/Q$. Clearly, this is the error of the linear Padé construction amplified by a factor of $1/Q$. This amplification could become large where Q is close to zero; in fact if $Q = 0$ at some point on its support, there is a singularity in the approximation, thus violating the conditions (8) and (20).

It is useful to analyze the algorithm step by step; we consider one-dimensional problem for compactness in writing. First, we construct Q such that $\langle Qu, P_n\rangle_N = 0$, for $n = M+1, \ldots, M+L$. In other words, the nth coefficient of the Legendre expansion of Qu vanishes for $n = M+1, \ldots, M+L$. This step can be viewed as a way to construct a *pre-conditioner* Q such that Qu may have better properties (in this case, continuity of the function) than u. As the second step, we compute P from $\langle P - Qu, P_n\rangle_N = 0$ for $n = 0, 1, \ldots, M$, which is a standard polynomial approximation, $P \approx Qu$. As the last step, we divide P by Q to obtain our approximation of u.

There is an implicit relation between the zeroes of the polynomial Q and the discontinuity region(s) of u. The construction of Q requires that the highest coefficients of Legendre expansion of Qu become small (or exactly zero in the one-dimensional case). This suggests that Q is small near the discontinuity regions of u, thus implying the proximity of the zeros of the polynomial Q and the discontinuity regions. It is possible that the algorithm yields a discontinuous Qu that has a PL expansion with certain high-order coefficients having small values. However, it is less likely that Qu will be discontinuous if a larger number of high-order coefficients in consecutive orders are required to be small. Thus, we prefer to choose $K > 1$ even though with $K = 1$, we might already have enough equations to solve for $\mathbf{q}$.

It is useful to outline that $1/Q$ must be small; otherwise it might lead to error amplification in the rationale approximation. On the other hand, Q is required to be very small at the discontinuities so that those are *suppressed* in the product Qu.

Without extra consideration, the method will therefore be effective if the error in the approximation of $R = P/Q$ is small while Qu remains smooth. In one-dimensional problem it is possible to construct polynomials whose zeroes coincide with the discontinuity locations (provided that L is large enough). In higher dimensions, this is not possible and there are few additional difficulties including:

- there might be no polynomial Q whose zeroes are close to the discontinuity locations, and conversely Q might have zeros where there are no discontinuities;
- the condition $Q > 0$ might be violated at some points on the support;
- the approximation $P \approx Qu$ is generally not as accurate as in one-dimensional cases. In particular, the values of P on the quadrature points can be different than the values of Qu, i.e., the approximant is not an interpolant.

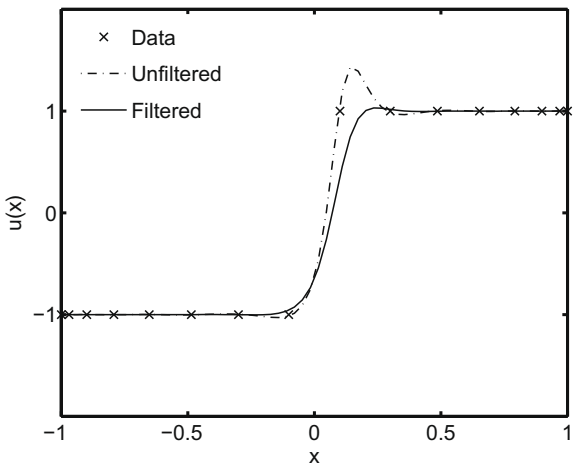


Fig. 2. The result of using Q-dependent filter for the Padé approximation of one-dimensional step function. The denominator Q is intentionally modified – by shifting the zero of Q by 0.1 – to yield a "bad" approximation in order to mimic possible situations in high-dimensional cases.

Most of these problems are essentially due to the limitation of polynomials to represent arbitrary functions. To address above issues we introduce a spatially varying filter whose weights depend on Q.

2.6. Q-dependent filter

As it was mentioned earlier, the error $E = P/Q - u$ in the approximation is amplified by a factor $1/Q$ from the error of the linear Padé problem $e = P - Qu$. The problem arises when Q becomes small or changes sign, away from the discontinuity locations. The error will then be amplified by a large factor. This situation is likely in multi-dimensional cases, since the calculation of the denominator Q is a minimization to an over-constraint problem (thus no solution exists for all modes

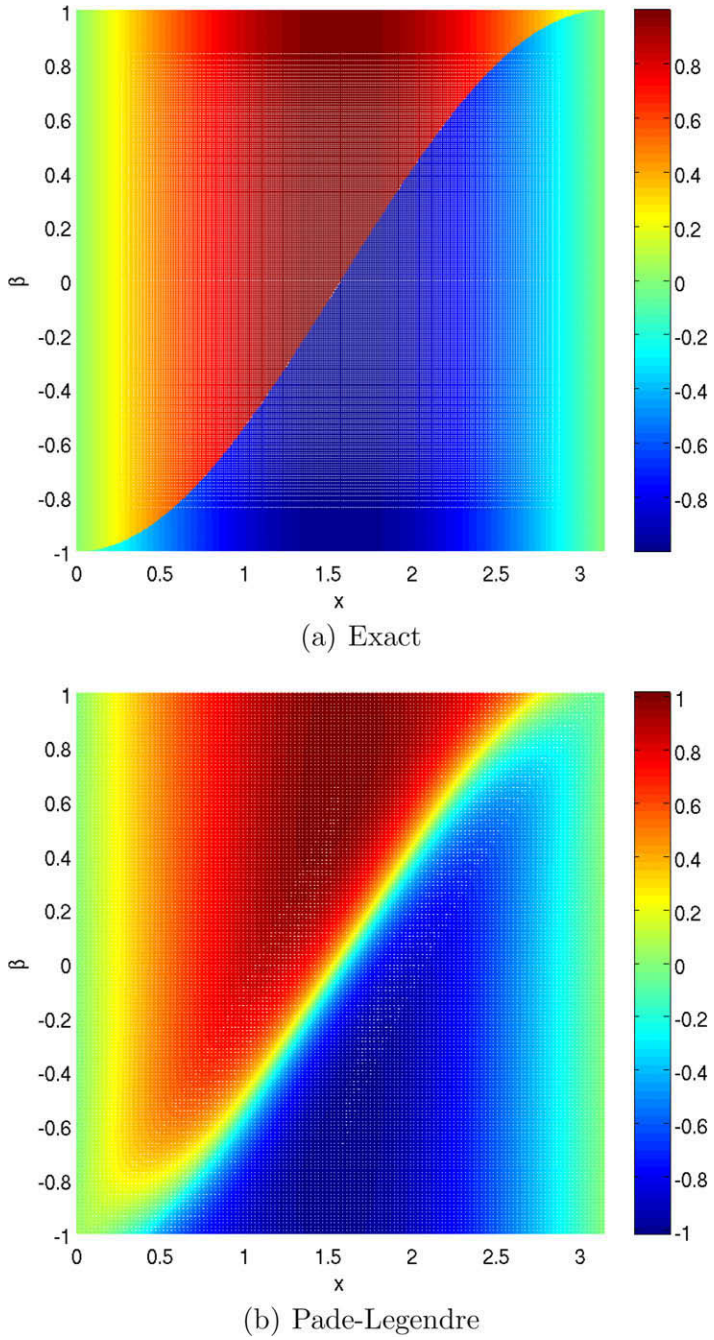


Fig. 3. Surfaces from the dual throat nozzle problem with one uncertain parameter using $(N, M, K, L) = (40, 24, 3, 2)$. No Q-dependent filter is used.

simultaneously). The problem also becomes more severe as the number of dimensions increases since the error in the least-square minimization increases.

To alleviate this problem, we introduce a spatial filter whose weights are modified by the local value of $|Q|$. The weight is increased if the value of $|Q|$ is relatively high at that point and decreased otherwise, that is, to say we *trust* the approximation more if the value of $|Q|$ is relatively high since the resulting error will be less amplified.

To construct the weights we consider a mean filter in the same dimension as the polynomials P and Q. Let h_x be the spatial filter kernel (weight) at the point y and $H(x)$ be the set of points involved in the kernel h_x. A conventional spatial filter would be applied to the approximation $R(u)$ as following:

$$\overline{R(u)(x)} = \sum_{y\in H(x)} R(u)(y)h_x(y), \tag{30}$$

where $\overline{R(u)}$ would be the filtered Padé approximation. To construct a Q-dependent filter, we modify the original spatial filter to take into account the value of Q by point-wise multiplying it with $|Q|$ value (and normalize the kernel). More specifically,

$$\overline{R(u)(x)} = \frac{\sum_{y\in H(x)} R(u)(y)|Q(y)|h_x(y)}{\sum_{y\in H(x)} |Q(y)|h_x(y)}. \tag{31}$$

Remark. One advantage of Q-dependent filtering over other types of filtering to remedy Gibbs phenomenon in spectral methods is that our filter does not require fine tuning of the parameters. Instead, it uses the information about the discontinuities readily available from the construction of the denominator Q to improve the original spatial filter. The only adjustable parameter of the filter is the kernel size. From our computational study, we found that a filter kernel width that is large enough to cover the neighboring *data* points produces a satisfactory result.

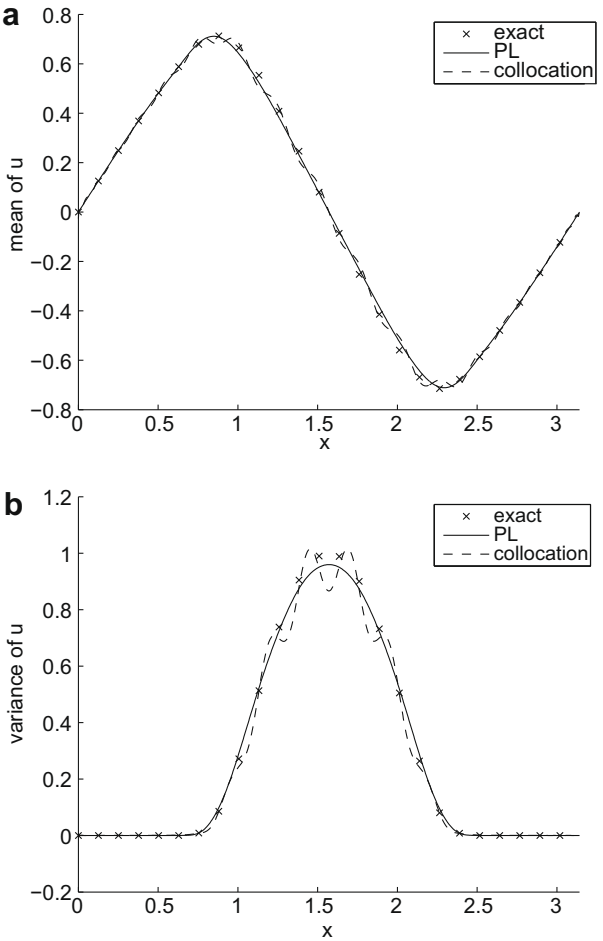


Fig. 4. Mean and variance of the solution u with $\sigma = 0.6$ computed from (a) analytical formula, (b) PL method with $(N, M, K, L) = (40, 24, 3, 2)$ and 10^6 samples from the PL response surface (29), and (c) stochastic collocation method with the same data and the number of samples. For the PL reconstruction, no Q-dependent filter is used.

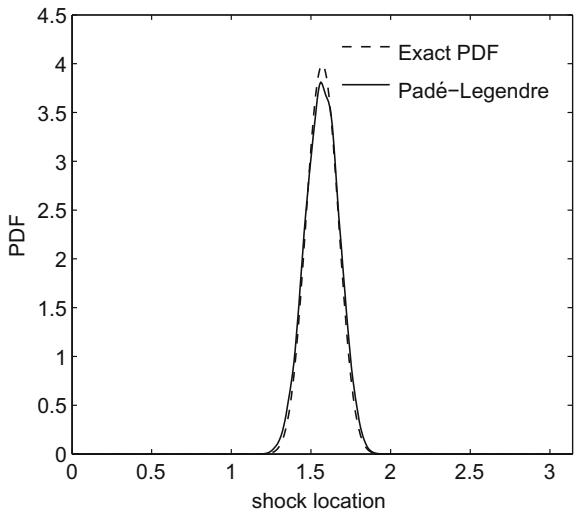
Fig. 2 illustrates the application of Q-dependent filter for the Padé approximation of one-dimensional step function:

$$u(x) = \begin{cases} 1 & x > 0, \\ -1 & x < 0. \end{cases} \tag{32}$$

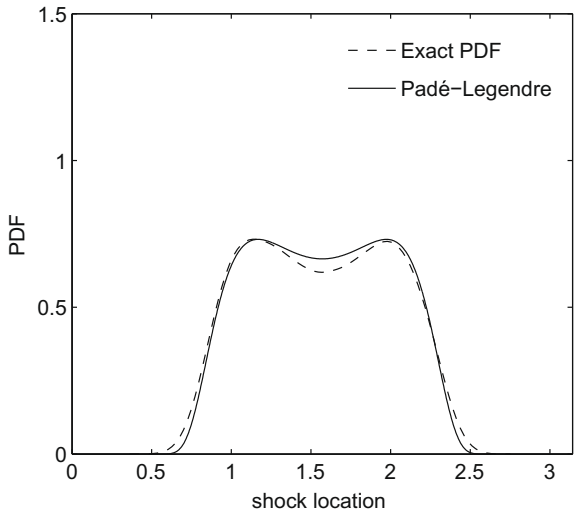
We start by computing Q from the algorithm provided in Section 2 with $L = 2$. This is a parabola with a single zero at $x = 0$. Next, we intentionally shift the zero of Q by 0.1 to mimic possible situations in high-dimensional cases (by replacing x with $x - 0.1$ in the expression of Q). Then we proceed to compute the numerator P and the rational approximant R from the shifted Q. The approximant exhibits a large overshoot near the discontinuity. This overshoot is roughly where the new zero of Q is; the error is large here because it is amplified by a large factor $1/Q$. When the Q-dependent filter is applied the overshoot is substantially decreased.

3. Numerical examples

The proposed PL formulation is applied to two problems dominated by discontinuities. In the first example, we consider a problem that resembles an isentropic flow in a dual throat nozzle [3]. It involves the solution of the Burgers equation with a source term. The steady-state solution depends on the initial condition in which we introduce one or more uncertain parameters. The second problem is related to the transonic flow around the RAE2822 airfoil and involves the solution of



(a) $\sigma = 0.1$



(b) $\sigma = 0.6$

Fig. 5. The PDF of shock locations drawn from 10^6 samples for two different variabilities, σ.

the compressible Reynolds averaged Navier–Stokes equations. In the latter case, the uncertainty is directly related to the flight conditions.

We note here that in all cases presented below a *multi-dimensional* PL formulation is required. In each problem, we consider one physical dimension and one or more stochastic dimension(s).

3.1. Dual throat nozzle problem with one uncertain parameter

According to isentropic theory of compressible flows in a dual throat nozzle with equal throat areas [3], the steady-state solution is completely subsonic or supersonic, or it contains a shock wave that determines a switch between the two flow regimes. If the shock exists, its location depends solely on the initial condition [4]. Here we use a simplified model based on the Burgers equations which was analyzed in [3]:

$$\frac{\partial u}{\partial t}+\frac{\partial}{\partial z}\left(\frac{u^2}{2}\right)=\frac{\partial}{\partial z}\left(\frac{\sin^2 z}{2}\right), \quad 0\leqslant z\leqslant \pi,\ t>0, \tag{33}$$

with the initial condition $u(z,0)=\beta\sin z$, and boundary conditions $u(0,t)=u(\pi,t)=0$. We are interested in the steady-state solution of (33) when β is a random variable $\beta\in[-1,1]$. Specifically,

$$\beta=\frac{-1+\sqrt{1+4\alpha^2}}{2\alpha}, \tag{34}$$

where α is a Gaussian random variable with zero mean and variance σ^2.

Analytical solutions for u as well as the probability distribution of the shock location corresponding to β, defined above, have been derived in [5].

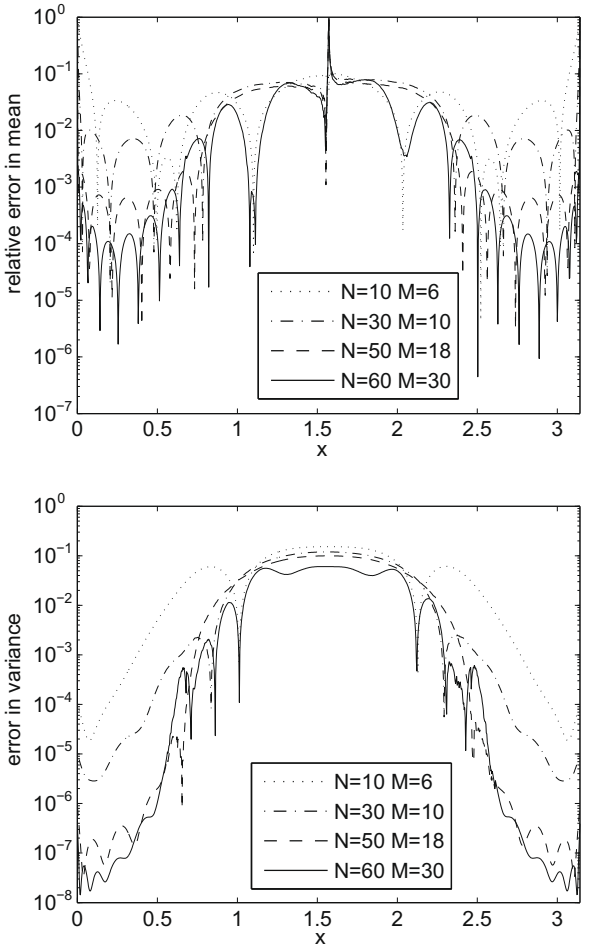


Fig. 6. Mean and variance errors as functions of the spatial coordinate x as the order of approximation increases for the high-variability case ($\sigma=0.6$). In all cases K and L are set to 3 and 2, respectively. The four cases shown here are $(N,M)=(10,6),\ (30,10),\ (50,18)$ and $(60,30)$.

For each β, the solution has a single shock. The shock location and strength vary with β as shown in Fig. 3(a).

To apply the Padé–Legendre method, we proceed as the following: (1) we perform deterministic simulations for different values of β in the spirit of stochastic collocation [19] and then extract the solutions only at a finite number of x location; (2) we construct a two-dimensional PL representation as a function of the scaled physical coordinate, x, where $x = z/\pi$, and stochastic coordinate, β; (3) we sample a large number of solutions from this representation according to the prescribed distribution of β; and (4) we extract from each of these solutions the shock location and build the desired statistics.

In this particular case, both the computed steady-state solutions and the PDF of the shock locations can be compared with the analytical values in [5]. The Padé–Legendre response surface is shown in Fig. 3(b) using $(N, M, K, L) = (40, 24, 3, 2)$.

Fig. 4 shows mean and variance of the solution u extracted from the PL surface with $\sigma = 0.6$. A large number of samples can be extracted from the surface (29) to ensure that the mean and variance converge with respect to the number of samples. In this case, 10^6 samples are used and found to be sufficient to achieve convergence. Mean and variance calculated from standard stochastic collocation method [19] on the same tensor-product grid are also presented for comparison. The collocation method employs the same data points and number of samples used in PL method. Clearly, the PL method performs much better than the standard collocation method in reducing the spurious oscillations due to the Gibbs phenomenon. Note that

Table 1
List of all (N, M, K, L) parameter sets shown in this section.

Case	N	M	K	L
1	40	18	3	2
2	60	18	3	2
3	40	26	3	2
4	60	30	3	2
5	40	18	5	2
6	40	18	3	4

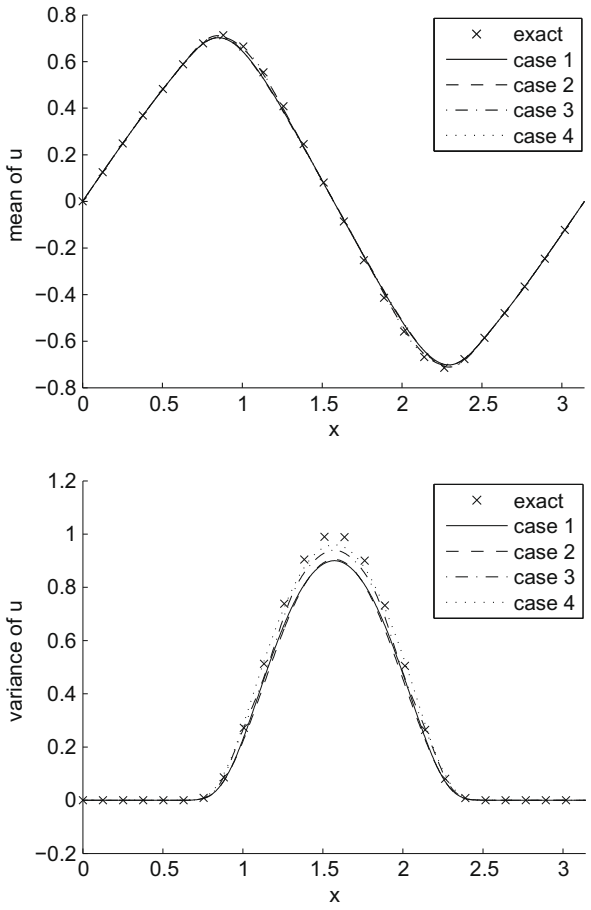


Fig. 7. The effects of increasing M and N on mean and variance.

the variance of the PL solution appears to be lower than the exact solution. This is due to the fact that the reconstructed surface is smoother than the actual discontinuous surface. This error can be reduced as we increase the parameters (N, M, K and L) as shown later in Section 3.1.1.

Fig. 6 shows errors of the approximated mean and variance of solutions as a function of x as the order of approximation increases. N is varied from 10 to 60 and M is increased accordingly, while K and L are set to 3 and 2, respectively. For each set of parameters, 10^6 samples are drawn from the *reconstructed* surface in order to obtain the statistics. The expected convergence in the mean and variance with respect to the order of approximation (N and M) is observed. The errors in both mean and variance are concentrated in the region where the shocks are likely to occur. For the mean, the relative error is used; this error measure is point-wise error normalized by the exact local mean. On the other hand, the relative error for the variance is a poor indicator since away from the shock, the variance is very close to zero; therefore, the absolute error in variance is reported here instead.

Fig. 5 shows the PDFs of the shock locations for the low-variability case $\sigma = 0.1$ and high-variability case $\sigma = 0.6$ compared with the corresponding exact PDFs. The results are satisfactory in both cases. We note here that a more sophisticated error representation for discontinuous functions is possible. In particular, one can first compare the shock location of the predicted solution to the exact solution, and then proceed to compare the entire solutions at the same position *relative* to its corresponding shock position as in [27]. In this paper, for the sake of simplicity, we choose to compare solutions at the same fixed positions.

In the sub-section below, the effects of the parameters N, M, K and L as well as the Q-dependent filter are discussed. In general, using higher-order polynomial reconstruction results in more accurate results; however, there are certain constraints and guidelines in the choices of these parameters.

3.1.1. Sensitivity of PL reconstruction

In this sub-section, we arbitrarily selected a base case of $(N, M, K, L) = (40, 18, 3, 2)$ for computing the mean and variance of the solution u with $\sigma = 0.6$. We increase each of the parameters and compare the results with the base case; initially no

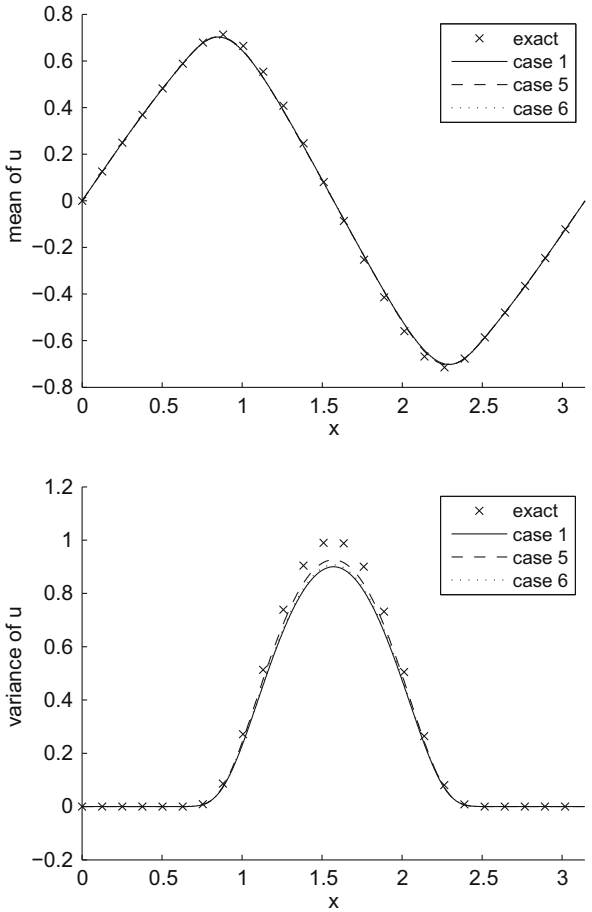


Fig. 8. The effects of increasing K and L on mean and variance.

Q-dependent filter is used. For each parameter, we have considered four different values. In general, increasing a parameter results in consistently more accurate results and, therefore, for the sake of compactness, we present below only one value in addition to the base case for each parameter. All the parameter sets presented here are listed in Table 1; the base case is case 1 in the table.

One constraint we have observed through this investigation is that the higher-order modes are severely contaminated and should not be used to compute the denominator. The rough guideline for this constraint is as following:

$$M + K < \frac{2N}{3}. \tag{35}$$

For all cases presented here the above relationship is satisfied. Fig. 7 shows the effects of increasing N and M (cases 2–4). In the first two cases, each parameter is increased while, in the last case, both are increased simultaneously. For both mean and variance, increasing N alone improves the solution only slightly. This suggests that the discrete inner product is calculated accurately, since the orders of the denominator and numerators as well as the equations to compute the denominator remain the same. Increasing M alone shows better improvement, especially in the variance. Increasing M affects the approximation in two ways: (1) the product Qu is approximated with more terms, and (2) the denominator is calculated with inner products with higher-order polynomials. While the former effect improves the accuracy of the approximation, the latter might cause an unexpected result due to change in the contamination in the higher-order modes. Increasing both N and M while still fulfilling the constraint (35) reduces the error substantially. In this case, the product Qu is more accurately approximately while the calculation of the denominator is kept away from such contamination.

Fig. 8 shows the effects of varying K and L (cases 5 and 6). The parameters K and L are varied independently from the base case. In both cases, small improvements are observed as expected. This suggests that the computation of the denominator is consistent and it does not change significantly with the parameters K and L.

For each case above, two calculations have been carried out, with and without the Q-dependent filter. Suppose that $\mathbb{G}$ is the grid on which we want to obtain the filtered PL surface. We assume that this grid is uniform in all directions. Let m be the

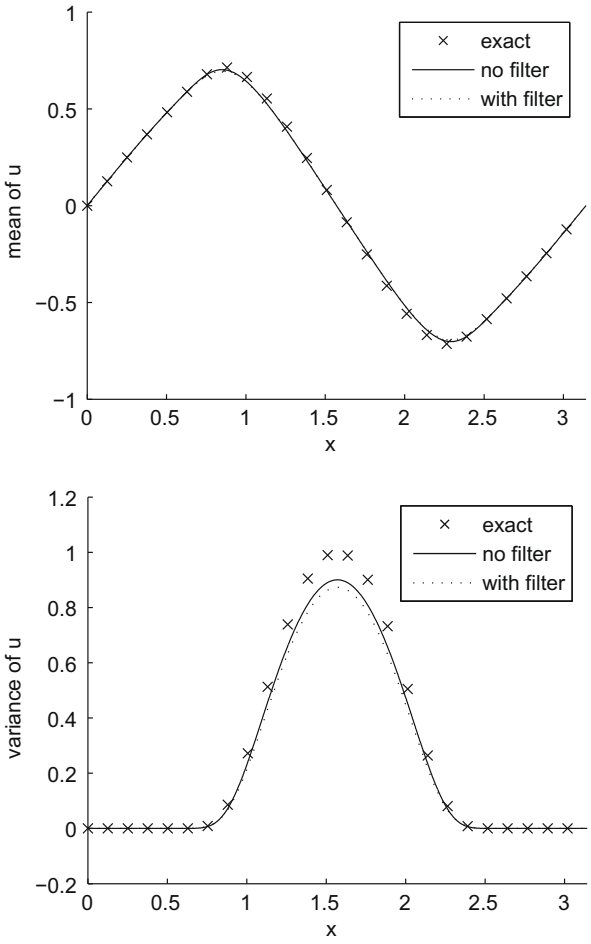


Fig. 9. Mean and variance of the solution u from the base case $(N, M, K, L) = (40, 18, 3, 2)$ with and without a Q-dependent filter.

number of points along each direction of $\mathbb{G}$. To construct the filter as described in Section 2.6, we first compute the kernel width from

$$n_H = 2\left\lceil \frac{m}{2N} \right\rceil + 1, \tag{36}$$

where $\lceil s \rceil$ is the smallest integer greater or equal to s. This makes the kernel large enough to cover the closest *data* points. The Q-dependent filter is based on the *mean* filter. This means that the kernel weight $h_x(y)$ is a constant. Since we know the size of the kernel from (36), we can compute this constant simply from:

$$h_x(y) = n_H^{-d}, \tag{37}$$

where d is the dimension of the problem (including both physical and stochastic dimensions). Plugging (36) into (31), we obtain the final Q-dependent filter:

$$\overline{R(u)(x)} = \frac{\sum_{y\in H(x)} R(u)(y)|Q(y)|n_H^{-d}}{\sum_{y\in H(x)} |Q(y)|n_H^{-d}}, \tag{38}$$

where $H(x) \subset \mathbb{G}$ includes all the points in the box, centered at x and covered $m/2N$ points away from x in each direction. Q and $R(u) = P/Q$ are the values obtained from the construction without a filter. We use this Q-dependent filter throughout the paper with fixed $m = 1000$.

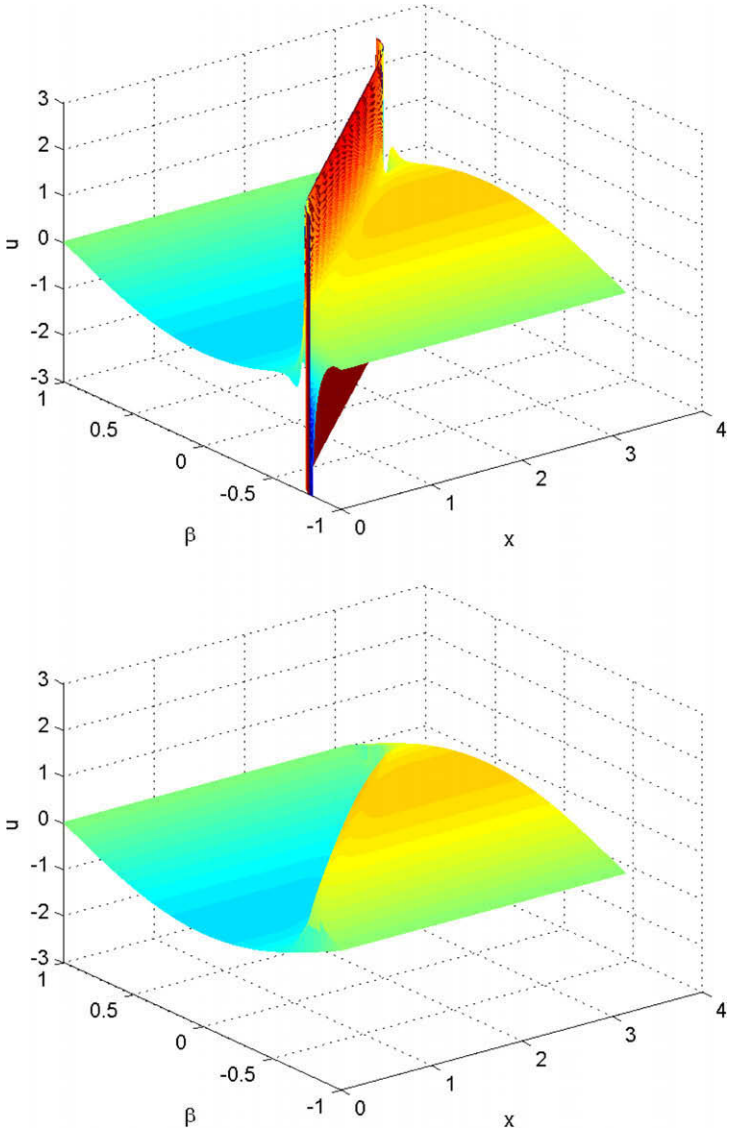


Fig. 10. Response surface of the case $(N, M, K, L) = (40, 20, 3, 3)$ with and without a Q-dependent filter.

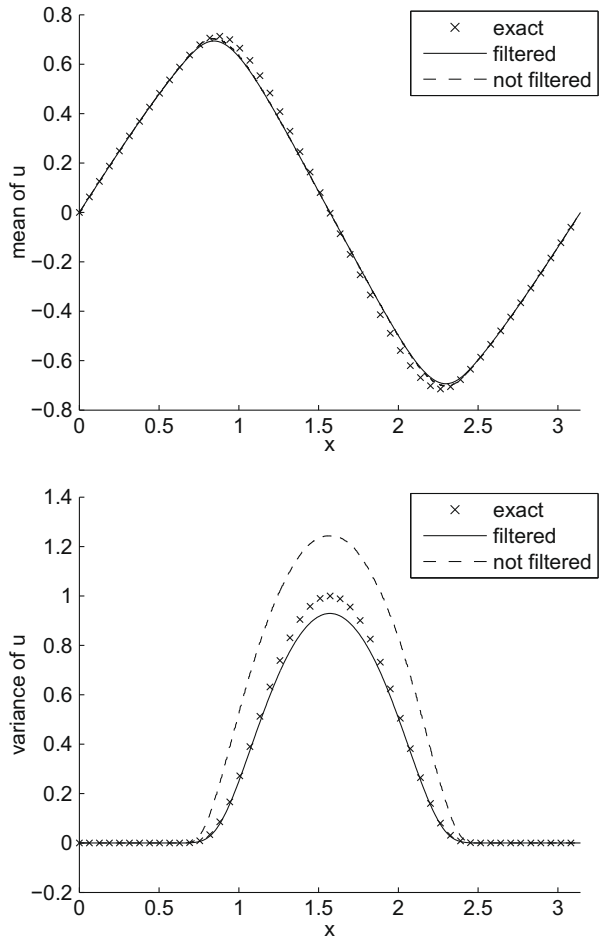
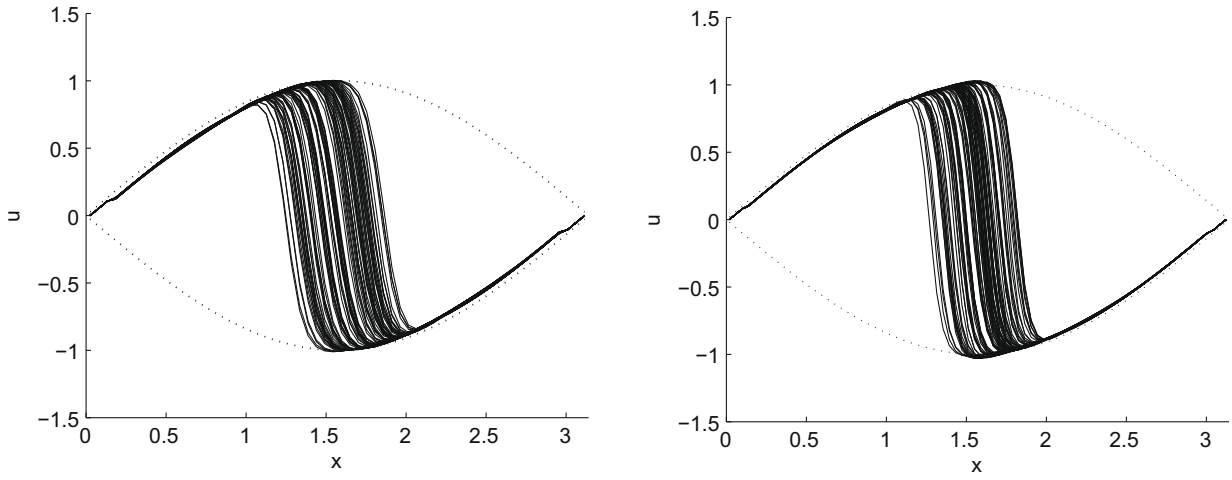


Fig. 11. Mean and variance of the solution u of the case $(N, M, K, L) = (40, 20, 3, 3)$ with and without a Q-dependent filter.



(a) $N = 15$ (b) $N = 20$

Fig. 12. Random samples of the solutions of the dual throat problem with two uncertain parameters. The solutions are extracted from Padé–Legendre reconstructed surfaces using $N = 15$ and $N = 20$ with an average Q-dependent filter.

In most cases, we observed the same trend when applying the filter: the mean of u does not visibly change while the variance is slightly reduced. This is expected as the filter tends to smooth out the surface thus reducing the variance of the samples from the surface. One example with and without filter is reported in Fig. 9.

In certain cases, however, using a Q-dependent filter improves the results substantially by removing the spurious errors caused by small or zero Q. Fig. 10 compares the response surfaces of one such case – $(N, M, K, L) = (40, 20, 3, 3)$ – with and without a Q-dependent filter. In this case, the denominator order L is odd and Q crosses zero along a certain curve on the support. As a result, the error is very high near that curve where Q is zero. Applying a Q-dependent filter effectively removes this large error. Fig. 11 shows the effects of the filter on mean and variance of the approximated solution u from this parameter set with $\sigma = 0.6$. It is hard to observe the improvement in the mean possibly because the support of the region with the large error is small. However, the application of the filter results in clear improvement in the variance.

3.2. Dual throat nozzle problem with two uncertain parameters

The dual throat nozzle problem in the previous section can be extended (and made more challenging) by introducing additional uncertain parameters in the definition of the initial conditions. In this case, we consider the uncertain initial condition characterized by a random field

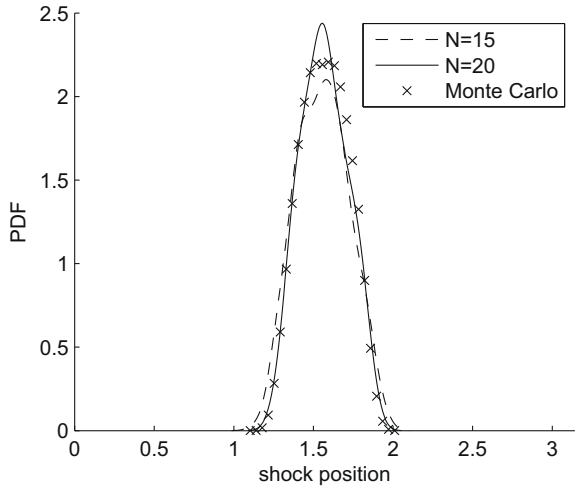


Fig. 13. The PDFs of the solutions of the dual throat problem with two uncertain parameters. The solutions are extracted from Padé–Legendre reconstructed surfaces using $N = 15$ and $N = 20$.

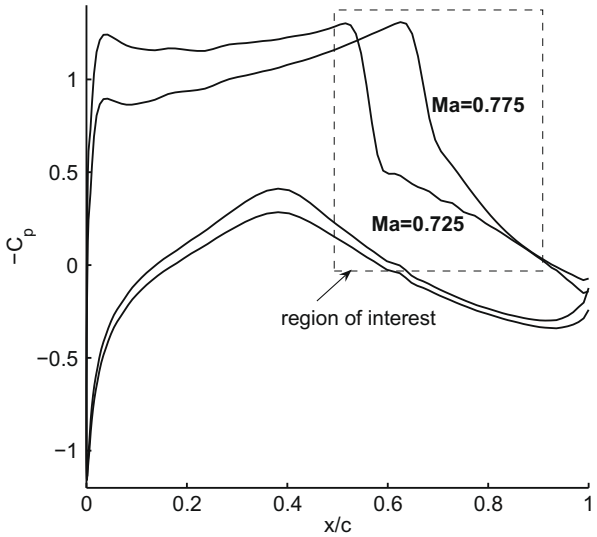


Fig. 14. Coefficient of pressures in the two extreme cases corresponding to $Ma = 0.725$ and $Ma = 0.775$. The rectangle shows the region of interest where the proposed method has been used.

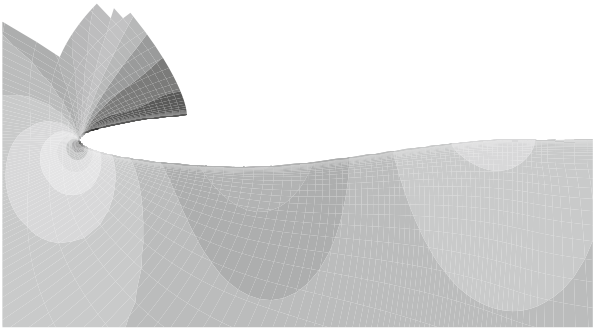
$$u(x,\beta_1,\beta_2,t=0)=\sigma\left(\sqrt{\lambda_1}f_1(x)\beta_1+\sqrt{\lambda_2}f_2(x)\beta_2\right), \tag{39}$$

where β_1 and β_2 are functions of two independent identically distributed Gaussian random variables α_1 and α_2, respectively. The relations between α_i and β_i are similar to the one-dimensional case (34). The definitions of λ_i and $f_i(x)$ are as following:

$$\lambda_n=\frac{2b}{1+b^2\omega_n^2},\quad n=1,2,3,\ldots \tag{40}$$

and

$$f_n(x)=\begin{cases}\frac{\cos(\omega_n(x-\pi/2))}{\sqrt{a+\frac{\sin(2\omega_n a)}{2\omega_n}}} & \text{if } n \text{ is odd},\\ \frac{\sin(\omega_n(x-\pi/2))}{\sqrt{a-\frac{\sin(2\omega_n a)}{2\omega_n}}} & \text{if } n \text{ is even}\end{cases} \tag{41}$$



(a) $Ma=0.725$

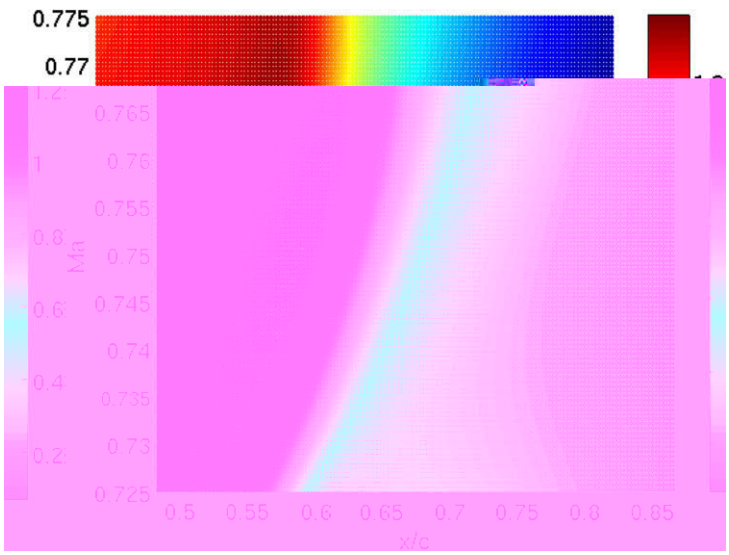


Fig. 16. The response surface of C_p of the region $x/c \in [0.5, 0.85]$ as Ma varies uniformly from 0.725 to 0.775.

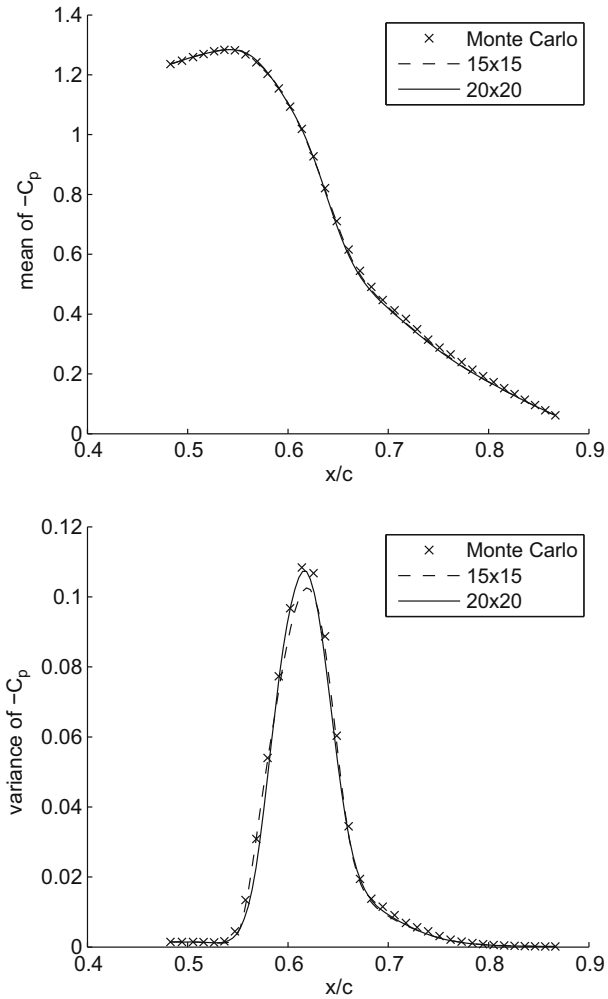


Fig. 17. Mean and variance of the C_p profiles near shock regions of the RAE2822 test case. The solutions are extracted from Padé–Legendre reconstructed surfaces using $N = 15$ and $N = 20$ compared to Monte Carlo simulations.

where $a = \pi/2,\ b = 10$, and the increasing sequence $\{\omega_n\}$ satisfies

$$\begin{cases} \frac{1}{b} - \omega_n \tan(\omega_n a) = 0 & \text{if } n \text{ is odd,} \\ \omega_n + \frac{1}{b} \tan(\omega_n a) = 0 & \text{if } n \text{ is even.} \end{cases} \tag{42}$$

Note that the individual steady-state solutions will be of the same form as in the previous case with one uncertain parameter, however, the probability distribution of the shock location is different. In this simulation, we use $\sigma = 0.4$ and $\alpha_1, \alpha_2 \sim N(0, 0.3)$ as in the high-variability case from [5]. Fig. 12 shows 100 samples from Padé–Legendre reconstructed surfaces using $N = 15$ $(L = 2,\ M = 10,\ K = 3)$ and $N = 20$ $(L = 2,\ M = 14,\ K = 3)$. A Q-dependent filter based on the mean spatial filter is used. Fig. 13 shows PDFs of the shock location obtained with the PL method and compared to Monte Carlo simulation with 20,000 samples.

3.3. Transonic airfoil

The second application of the proposed method involves the computation of the flow around a transonic airfoil, the RAE2822 airfoil. The flow conditions are specified in terms of Reynolds and Mach number as $Re = 6.2 \times 10^6$ and $Ma = 0.75$; the angle of attack is $\alpha = 2.8^\circ$. These conditions correspond to Case 10 in [12]. The flow is characterized by a strong interaction of the boundary layer developing on the airfoil and a shock wave on the upper surface; this can lead to separation and affect the overall aerodynamic forces. The objective of the present study is to identify the effect of the uncertainty in the flight conditions (specifically in the Mach number) on the flow characteristic and, more specifically, on the overall surface pressure distribution. The experimental uncertainty in the free-stream conditions $\approx 4\%$ [12]; in the present simulations we assumed Ma to be uniformly distributed in the interval [0.725,0.775].

The calculations presented are performed using SUmb [28], a parallel multiblock Reynolds averaged Navier–Stokes based on finite volume second-ordered discretization. The computational grid consists of 256×64 elements in a C-grid with strong clustering in the near-wall region to capture the boundary layer. Turbulence is modeled using k—ω Wilcox turbulence model [35]. In the present exercise we are not specifically focused on the accuracy of the turbulence model and its ability to represent the experimental observation, although ultimately this is of crucial importance. Our objective is to demonstrate that the present PL formulation allows us to efficiently construct a response surface for the wall pressure coefficient and to extract accurate statistics.

The Mach number range identified earlier leads to considerable variability in terms of the wall pressure coefficient C_p (Fig. 14) as well as of the flow field (Fig. 15). The calculations show that the shock on the upper surface moves considerably downstream while becoming stronger. On the other hand, the discontinuity in the wall pressure coefficient becomes weaker.

We focus our attention on the wall pressure coefficient C_p within the shock/boundary layer interaction region, roughly corresponding to x/c (x-position normalized by chord length) in the range [0.5,0.85] on the top surface. Fig. 16 shows the calculated surface response of C_p in this region of interest. The surface clearly indicates the shock location varying as a function of the uncertain parameter similar to previous problems, thus justifying the use of the PL formulation.

We compare our predictions using the PL approach with Monte Carlo results obtained using 5000 samples which were found to be sufficient for achieving convergence in the statistics.

Fig. 17 shows the computed mean and variance of C_p. Two sets of Padé–Legendre parameters are used to test the robustness of the method: (a) $(N, M, K, L) = (15, 10, 3, 2)$ and (b) $(N, M, K, L) = (20, 14, 3, 2)$; as usual the Q-dependent filter is used. The mean is captured accurately in both cases. There is some discrepancy in the peak variance for the former case. This is due to insufficient resolution in the construction of the surface response, as observed in the two-dimensional dual throat nozzle test case. In that case, the problem does not persist when we increase N to 20.

4. Conclusions

A Padé–Legendre approach for propagating input uncertainties in problems characterized by strong non-linearity and discontinuous response surfaces has been introduced. It is successfully applied to fluid-dynamic problems dominated by the presence of shocks. The method has higher efficiency and accuracy with respect to standard spectral stochastic Galerkin methods and does not require *a priori* knowledge of the discontinuity location. In the present approach, a rational polynomial is used to represent the response surface containing sharp interfaces. The convergence of the scheme has been verified with respect to Monte Carlo statistics and also by changing the order of the polynomial interpolants.

We have extended the method to multiple dimensions by using a non-uniform filter dependent on the denominator of the rational interpolation function. This has proved to be robust in the problems of interest.

Acknowledgments

This material is based upon work supported by the Department of Energy [National Nuclear Security Administration] under Award Number NA28614. We also would like to thank Prof. Jan S. Hesthaven for insightful discussions regarding the Padé–Legendre method and Dr. Edwin van der Weide for providing the SUmb code and advice on how to use it.

Appendix A. Singular value decomposition

Suppose A is a n-by-p matrix. The SVD theorem states that there exists a factorization of A of the form

$$A = USV^*, \tag{A.1}$$

where U is an n-by-n unitary matrix, V is an p-by-p unitary matrix and S is an n-by-p diagonal matrix with non-negative numbers on the diagonal. V^* denotes the conjugate transpose of V. This factorization is called a singular value decomposition of A.

Calculating the SVD consists of finding the eigenvales and eigenvectors of AA^* and A^*A. The eigenvectors of AA^* make up the columns of U and those of A^*A do those of V. The singular values in S are square roots of the eigenvalues from AA^*. Conventionally, the diagonal entries of S are arranged in descending order.

We would like to use SVD to solve the minimization problem (27), restated below

$$\min_{\|\mathbf{q}^*\|=1} \|A\mathbf{q}^*\|. \tag{A.2}$$

We start by decomposing A with SVD. Since U is unitary, it follows that

$$\|A\mathbf{q}'\|^2 = \mathbf{q}'^* A^* A\mathbf{q}' \tag{A.3}$$
$$= \mathbf{q}'^* VS^* U^* USV^* \mathbf{q}' \tag{A.4}$$
$$= \mathbf{q}'^* VS^* SV^* \mathbf{q}' \tag{A.5}$$
$$= \mathbf{x}^* \Lambda^2 \mathbf{x} \tag{A.6}$$
$$= \|\Lambda\mathbf{x}\|^2, \tag{A.7}$$

where $\mathbf{x} = V^*\mathbf{q}'$, $\Lambda^2 = S^*S$ and Λ is a diagonal square matrix with non-negative elements. Note further that, since V is unitary, it follows that

$$\|\mathbf{x}\|^2 = \mathbf{x}^*\mathbf{x} = \mathbf{q}'^* VV^* \mathbf{q}' = \mathbf{q}'^*\mathbf{q}' = \|\mathbf{q}'\|^2 = 1. \tag{A.8}$$

Thus, the minimization problem (A.2) reduces to

$$\min_{\|\mathbf{x}\|=1} \|\Lambda\mathbf{x}\|. \tag{A.9}$$

Clearly, the solution to this problem is $\mathbf{x} = e_1$, since the first diagonal element in S is the smallest (and thus that in Λ is the smallest). Therefore, $V^*\mathbf{q} = e_1$ and $\mathbf{q} = Ve_1$. In short, the solution to the minimization problem (A.2) is the last column of V.

References

[1] J.S. Hesthaven, S.M. Kaber, L. Lurati, Padé–Legendre interpolants for Gibbs reconstruction, J. Sci. Comput. 28 (2–3) (2006) 337–359.
[2] G.A. Baker, Padé Approximants, Cambridge University Press, 1996.
[3] M.D. Salas, S. Abarbanel, D. Gottlieb, Multiple steady states for characteristic initial value problems, Appl. Numer. Math. 2 (1986) 193–210.
[4] C.M. Dafermos, Trend to steady state in a conservation law with spatial inhomogeneity, Quart. Appl. Math. XLV (2) (1987) 313–319.
[5] Q.-Y. Chen, D. Gottlieb, J.S. Hesthaven, Uncertainty analysis for the steady-state flows in a dual throat nozzle, J. Comput. Phys. 204 (1) (2005) 378–398.
[6] Ph. Guillaume, A. Huard, Multivariate pad approximation, J. Comput. Appl. Math. 121 (2000) 197–219.
[7] A.C. Matos, Recursive computation of Padé–Legendre approximants and some acceleration properties, Numer. Math. 89 (2001) 535–560.
[8] J.S.R. Chisholm, Rational approximants defined from double power series, Math. Comput. 27 (1973) 841–848.
[9] A. Cuyt, Multivariate Padé approximants, J. Math. Anal. Appl. 96 (1983) 283–293.
[10] Ph. Guillaume, Nested multivariate Padé approximants, J. Comput. Appl. Math. 82 (1997) 149–158.
[11] Ph. Guillaume, A. Huard, V. Robin, Generalized multivariate Padé approximants, J. Approx. Theory 95 (2) (1998) 203–214.
[12] P.H. Cook, M.A. McDonald, M.C.P. Firmin, Aerofoil RAE2822 – Pressure Distributions and Boundary Layers and Wake Measurements, AGARD-AR-138, 1979.
[13] T.I. Malik, R.K. Tagirov, Calculation of the length of the shock–boundary layer interaction zone, Fluid Dynam. 22 (1987) 318–321.
[14] A.B. Oliver, R.P. Lillard, A.M. Schwing, G.A. Blaisdell, A.S. Lyrintzis, Assessment of Turbulent Shock–Boundary Layer Interaction Computations Using the OVERFLOW Code, AIAA Paper 2007-104.
[15] K. Sinha, K. Mahesh, G.V. Candler, Modeling the effect of shock unsteadiness in shock/turbulent boundary-layer interactions, AIAA J. 43 (2005) 587–594.
[16] S. Stolz, N.A. Adams, L. Kleiser, The approximate deconvolution model for large eddy simulation of compressible flows and its application to shock–turbulent-boundary-layer interaction, Phys. Fluid 13 (2001) 2985–3001.
[17] W.L. Oberkampf, M.F. Barone, Measures of agreement between computation and experiment: validation metrics, J. Comput. Phys. 217 (2006) 5–36.
[18] R. Ghanem, P. Spanos, Stochastic Finite Elements: A Spectral Approach, Springer-Verlag, New York, 1991.
[19] D. Xiu, J.S. Hesthaven, High order collocation methods for the differential equations with random inputs, SIAM J. Sci. Comput. 27 (2005) 1118–1139.
[20] F. Nobile, R. Tempone, C.G. Webster, A sparse grid stochastic collocation method for partial differential equations with random input data, SIAM J. Num. Analysis 46 (2008) 2309–2345.
[21] D. Ghosh, R. Ghanem, Stochastic convergence acceleration through basis enrichment of polynomial Chaos expansions, Int. J. Numer. Method Eng. 73 (2) (2007) 162–184.
[22] X. Wan, G.E. Karniadakis, An adaptive multi-element generalized polynomial chaos method for stochastic differential equations, J. Comput. Phys. 209 (2005) 617–642.
[23] X. Wan, G.E. Karniadakis, Multi-element generalized polynomial chaos for arbitrary probability measures, SIAM J. Sci. Comput. 28 (2006) 901–928.
[24] G. Lin, A.M. Tartakovsky, An efficient, high-order multi-element probabilistic collocation method on sparse grids for three-dimensional flow in random porous media, in: American Geophysical Union, Fall Meeting, Abstract Number H23B-1318, 2007.

[25] G. Lin, C.-H. Su, G.E. Karniadakis, Stochastic modeling of random roughness in shock scattering problems: theory and simulations, Comput. Method Appl. Mech. Eng. (2008).
[26] D. Bau III, L.N. Trefethen, Numerical Linear Algebra, SIAM, 1997.
[27] J. Glimm, J.W. Grove, Y. Kang, T. Lee, X. Li, D.H. Sharp, Y. Yu, K. Ye, M. Zhao, Error Analysis for Shock Interactions, Stony Brook AMS, Preprint Report Number SUNYSB-AMS-03-14, 2003.
[28] E. van der Weide, G. Kalitzin, J. Schluter, J.J. Alonso, Unsteady turbomachinery computations using massively parallel platforms, in: Proceedings of the 44th AIAA Aerospace Sciences Meeting, AIAA Paper 2006-421, Reno, January 2006.
[29] O.P. Le Maitre, H.N. Najm, R.G. Ghanem, O.M. Knio, Multi-resolution analysis of Wiener-type uncertainty propagation schemes, J. Comput. Phys. 197 (2004).
[30] I. Babuka, F. Nobile, R. Tempone, A stochastic collocation method for elliptic partial differential equations with random input data, SINUM 45 (3) (2007).
[31] L. Mathelin, M. Yousuff Hussaini, A Stochastic Collocation Algorithm for Uncertainty Analysis, NASA/CR-2003-212153, 2003.
[32] L. Emmel, S.M. Kaber, Y. Maday, Padé–Jacobi filtering for spectral approximations of discontinuous solutions, Numer. Algorithm 33 (2003) 251–264.
[33] S.M. Kaber, Y. Maday, Padé–Chebyshev approximants, SIAM J. Numer. Anal. 43 (2004) 437–454.
[34] D. Xiu, D. Lucor, C.-H. Su, G.E. Karniadakis, Stochastic modeling of flow–structure interactions using generalized polynomial chaos, J. Fluid Eng. 124 (2002) 51–59.
[35] D.C. Wilcox, Re-assessment of the scale-determining equation for advanced turbulence models, AIAA J. 26 (1988) 1414–1421.