Applied Numerical Mathematics 90 (2015) 91–110

Contents lists available at ScienceDirect

Applied Numerical Mathematics

www.elsevier.com/locate/apnum

# Adaptive multi-element polynomial chaos with discrete measure: Algorithms and application to SPDEs

Mengdi Zheng [a], Xiaoliang Wan [b], George Em Karniadakis [a,*]

[a] *Division of Applied Mathematics, Brown University, United States*
[b] *Department of Mathematics, Louisiana State University, United States*

ARTICLE INFO

*Article history:*
Received 4 January 2013
Received in revised form 14 July 2014
Accepted 28 November 2014
Available online 15 December 2014

*Keywords:*
Stochastic collocation
Data-driven model
KdV equation
Discrete–continuous inputs
Charlier polynomials

ABSTRACT

We develop a multi-element probabilistic collocation method (ME-PCM) for arbitrary discrete probability measures with finite moments and apply it to solve partial differential equations with random parameters. The method is based on numerical construction of orthogonal polynomial bases in terms of a discrete probability measure. To this end, we compare the accuracy and efficiency of five different constructions. We develop an adaptive procedure for decomposition of the parametric space using the local variance criterion. We then couple the ME-PCM with sparse grids to study the Korteweg–de Vries (KdV) equation subject to random excitation, where the random parameters are associated with either a discrete or a continuous probability measure. Numerical experiments demonstrate that the proposed algorithms lead to high accuracy and efficiency for hybrid (discrete–continuous) random inputs.

## 1. Introduction

Stochastic partial differential equations (SPDEs) are widely used for stochastic modeling in diverse applications from physics, to engineering, biology and many other fields. In this paper, we concentrate on parametric uncertainty associated with *arbitrary discrete* probability measures with finite moments, where the source of uncertainty includes random coefficients, and stochastic forcing. In many cases, the random parameters are only observed at discrete values, which implies that a discrete probability measure is more appropriate from the modeling point of view. More generally, random processes with jumps are of fundamental importance in stochastic modeling, e.g., stochastic-volatility jump-diffusion models in finance [29], stochastic simulation algorithms for modeling diffusion, reaction and taxis in biology [7], fluid models with jumps [25], quantum-jump models in physics [6], etc.

In this work we extend the multi-element probabilistic collocation method (ME-PCM) [12] to deal with discrete probability measures. The ME-PCM decomposes the parametric space [26] and implements generalized polynomial chaos (gPC) element-wise using the collocation approach [28], such that both $h$- and $p$-convergence can be obtained. In particular, we follow the gPC correspondence principle, where orthogonal polynomials with respect to the discrete measure are generated to improve efficiency [28]. We first evaluate the accuracy and efficiency of different methods for numerical construction of gPC bases by performing systematic tests on functions from the GENZ suite [15]. We then couple ME-PCM with sparse grids and develop an adaptive procedure using the local variance criterion [12]. The developed algorithms are applied to KdV equation [19] subject to stochastic excitation.

[*] Corresponding author.
*E-mail address:* George_Karniadakis@brown.edu (G.E. Karniadakis).

http://dx.doi.org/10.1016/j.apnum.2014.11.006
0168-9274/© 2014 IMACS. Published by Elsevier B.V. All rights reserved.

The paper is organized as follows. The methods of generating orthogonal polynomial bases with respect to discrete measures are presented in Section 2 followed by a discussion about the error of numerical integration in Section 3. Numerical solutions of the stochastic reaction equation and KdV equation, including adaptive procedures, are explained in Section 4, and in Section 5 we summarize the work. In the appendices, we provide more details about the deterministic KdV equation solver, and the adaptive procedure.

## 2. Generation of orthogonal polynomials for discrete measures

Let $\mu$ be a positive measure with infinite support $S(\mu) \subset \mathbb{R}$ and finite moments at all orders, i.e.,

$$\int_S \xi^n \mu(d\xi) < \infty, \quad \forall n \in \mathbb{N}_0, \tag{1}$$

where $\mathbb{N}_0 = \{0, 1, 2, ...\}$, and it is defined as a Riemann–Stieltjes integral. There exists one unique [14] set of orthogonal monic polynomials $\{P_i\}_{i=0}^{\infty}$ with respect to the measure $\mu$ such that

$$\int_S P_i(\xi)P_j(\xi)\mu(d\xi) = \delta_{ij}\gamma_i^{-2}, \quad i = 0, 1, 2, \dots, \tag{2}$$

where $\gamma_i \neq 0$ are constants. In particular, the orthogonal polynomials satisfy a three-term recurrence relation [5,9]

$$P_{i+1}(\xi) = (\xi - \alpha_i)P_i(\xi) - \beta_i P_{i-1}(\xi), \quad i = 0, 1, 2, \dots \tag{3}$$

The uniqueness of the set of orthogonal polynomials with respect to $\mu$ can be also derived by constructing such set of polynomials starting from $P_0(\xi) = 1$. We typically choose $P_{-1}(\xi) = 0$ and $\beta_0$ to be a constant. Then the full set of orthogonal polynomials is completely determined by the coefficients $\alpha_i$ and $\beta_i$.

If the support $S(\mu)$ is a finite set with data points $\{\tau_1, ..., \tau_N\}$, i.e., $\mu$ is a discrete measure defined as

$$\mu = \sum_{i=1}^{N} \lambda_i \delta_{\tau_i}, \quad \lambda_i > 0, \tag{4}$$

the corresponding orthogonality condition is finite, up to order $N - 1$ [11,14], i.e.,

$$\int_S P_i^2(\xi)\mu(d\xi) = 0, \quad i \geq N, \tag{5}$$

where $\delta_{\tau_i}$ indicates the empirical measure at $\tau_i$, although by the recurrence relation (3) we can generate polynomials at any order greater than $N - 1$. Furthermore, one way to test whether the coefficients $\alpha_i$ are well approximated is to check the following relation [10,11]

$$\sum_{i=0}^{N-1} \alpha_i = \sum_{i=1}^{N} \tau_i. \tag{6}$$

One can prove that the coefficient of $\xi^{N-1}$ in $P_N(\xi)$ is $-\sum_{i=0}^{N-1} \alpha_i$, and $P_N(\xi) = (\xi - \tau_1)...(\xi - \tau_N)$, therefore Eq. (6) holds [11].

We subsequently examine five different approaches of generating orthogonal polynomials for a discrete measure and point out the pros and cons of each method. In Nowak method, the coefficients of the polynomials are directly derived from solving a linear system; in the other four methods, we generate coefficients $\alpha_i$ and $\beta_i$ by four different numerical methods, and the coefficients of polynomials are derived from the recurrence relation in Eq. (3).

### 2.1. Nowak method [22]

Define the $k$-th order moment as

$$m_k = \int_S \xi^k \mu(d\xi), \quad k = 0, 1, ..., 2d - 1. \tag{7}$$

The coefficients of the $d$-th order polynomial $P_d(\xi) = \sum_{i=0}^{d} a_i \xi^i$ are determined by the following linear system

$$\begin{pmatrix} m_0 & m_1 & \dots & m_d \\ m_1 & m_2 & \dots & m_{d+1} \\ \dots & \dots & \dots & \dots \\ m_{d-1} & m_d & \dots & m_{2d-1} \\ 0 & 0 & \dots & 1 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \dots \\ a_{d-1} \\ a_d \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \dots \\ 0 \\ 1 \end{pmatrix}, \tag{8}$$

where the $(d+1)$ by $(d+1)$ Vandermonde matrix needs to be inverted.

Although this method is straightforward to implement, it is well known that the matrix may be ill conditioned when $d$ is very large.

The total computational complexity for solving the linear system in Eq. (8) is $\mathcal{O}(d^2)$ to generate $P_d(\xi)$.[1]

### 2.2. *Stieltjes method [14]*

Stieltjes method is based on the following formulas of the coefficients $\alpha_i$ and $\beta_i$

$$\alpha_i = \frac{\int_S \xi P_i^2(\xi)\mu(d\xi)}{\int_S P_i^2(\xi)\mu(d\xi)}, \qquad \beta_i = \frac{\int_S \xi P_i^2(\xi)\mu(d\xi)}{\int_S P_{i-1}^2(\xi)\mu(d\xi)}, \quad i = 0, 1, .., d-1. \tag{9}$$

For a discrete measure, the Stieltjes method is quite stable [14,13]. When the discrete measure has a finite number of elements in its support ($N$), the above formulas are exact. However, if we use Stieltjes method on a discrete measure with infinite support, i.e. Poisson distribution, we approximate the measure by a discrete measure with finite number of points; therefore, each time when we iterate for $\alpha_i$ and $\beta_i$, the error accumulates by neglecting the points with less weights. In that case, $\alpha_i$ and $\beta_i$ may suffer from inaccuracy when $i$ is close to $N$ [14].

The computational complexity for integral evaluation in Eq. (9) is of the order $\mathcal{O}(N)$.

### 2.3. *Fischer method [10,11]*

Fischer proposed a procedure for generating the coefficients $\alpha_i$ and $\beta_i$ by adding data points one-by-one [10,11]. Assume that the coefficients $\alpha_i$ and $\beta_i$ are known for the discrete measure $\mu = \sum_{i=1}^N \lambda_i \delta_{\tau_i}$. Then, if we add another data point $\tau$ to the discrete measure $\mu$ and define a new discrete measure $\nu = \mu + \lambda\delta_\tau$, $\lambda$ being the weight of the newly added data point $\tau$, the following relations hold:

$$\alpha_i^\nu = \alpha_i + \lambda \frac{\gamma_i^2 P_i(\tau) P_{i+1}(\tau)}{1 + \lambda \sum_{j=0}^{i} \gamma_j^2 P_j^2(\tau)} - \lambda \frac{\gamma_{i-1}^2 P_i(\tau) P_{i-1}(\tau)}{1 + \lambda \sum_{j=0}^{i-1} \gamma_j^2 P_j^2(\tau)}, \tag{10}$$

$$\beta_i^\nu = \beta_i \frac{[1 + \lambda \sum_{j=0}^{i-2} \gamma_j^2 P_j^2(\tau)][1 + \lambda \sum_{j=0}^{i} \gamma_j^2 P_j^2(\tau)]}{[1 + \lambda \sum_{j=0}^{i-1} \gamma_j^2 P_j^2(\tau)]^2} \tag{11}$$

for $i < N$, and

$$\alpha_N^\nu = \tau - \lambda \frac{\gamma_{N-1}^2 P_N(\tau) P_{N-1}(\tau)}{1 + \lambda \sum_{j=0}^{N-1} \gamma_j^2 P_j^2(\tau)}, \tag{12}$$

$$\beta_N^\nu = \frac{\lambda \gamma_{N-1}^2 P_N^2(\tau)[1 + \lambda \sum_{j=0}^{N-2} \gamma_j^2 P_j^2(\tau)]}{[1 + \lambda \sum_{j=0}^{N-1} \gamma_j^2 P_j^2(\tau)]^2}, \tag{13}$$

where $\alpha_i^\nu$ and $\beta_i^\nu$ indicate the coefficients in the three-term recurrence formula (3) for the measure $\nu$. The numerical stability of this algorithm depends on the stability of the recurrence relations above, and on the sequence of data points added [11]. For example, the data points can be in either ascending or descending order. Fischer's method basically modifies the available coefficients $\alpha_i$ and $\beta_i$ using the information induced by the new data point. Thus, this approach is very practical when an empirical distribution for stochastic inputs is altered by an additional possible value. For example, let us consider that we have already generated $d$ probability collocation points with respect to the given discrete measure with $N$ data points, and we want to add another data point into the discrete measure to generate $d$ new probability collocation points with respect to the new measure. Using the Nowak method, we will need to reconstruct the moment matrix and invert the matrix again with $N+1$ data points; however by Fischer's method, we will only need to update $2d$ values of $\alpha_i$ and $\beta_i$ by adding this new data point, which is more convenient.

We generate a new sequence of $\{\alpha_i, \beta_i\}$ by adding a new data point into the measure, therefore the computational complexity for calculating the coefficients $\{\gamma_i^2, i = 0, .., d\}$ for $N$ times is $\mathcal{O}(N^2)$.

[1] Here we notice that the Vandermonde matrix is in a Toeplitz matrix form. Therefore the computational complexity of solving this linear system is $\mathcal{O}(d^2)$ [16,24].

### 2.4. Modified Chebyshev method [14]

Compared to the Chebyshev method [14], the modified Chebyshev method computes moments in a different way. Define the quantities:

$$\mu_{i,j} = \int_S P_i(\xi)\xi^j \mu(d\xi), \quad i, j = 0, 1, 2, \dots \tag{14}$$

Then, the coefficients $\alpha_i$ and $\beta_i$ satisfy:

$$\alpha_0 = \frac{\mu_{0,1}}{\mu_{0,0}}, \qquad \beta_0 = \mu_{0,0}, \qquad \alpha_i = \frac{\mu_{i,i+1}}{\mu_{i,i}} - \frac{\mu_{i-1,i}}{\mu_{i-1,i-1}}, \qquad \beta_i = \frac{\mu_{i,i}}{\mu_{i-1,i-1}}. \tag{15}$$

Note that due to the orthogonality, $\mu_{i,j} = 0$ when $i > j$. Starting from the moments $\mu_j$, $\mu_{i,j}$ can be computed recursively as

$$\mu_{i,j} = \mu_{i-1,j+1} - \alpha_{i-1}\mu_{i-1,j} - \beta_{i-1}\mu_{i-2,j}, \tag{16}$$

with

$$\mu_{-1,0} = 0, \qquad \mu_{0,j} = \mu_j, \tag{17}$$

where $j = i, i+1, \dots, 2d - i - 1$.

However, this method suffers from the same effects of ill-conditioning as the Nowak method [22] does, because they both rely on calculating moments. To stabilize the algorithm we introduce another way of defining moments by polynomials:

$$\hat{\mu}_{i,j} = \int_S P_i(\xi) p_j(\xi) \mu(d\xi), \tag{18}$$

where $\{p_i(\xi)\}$ is chosen to be a set of orthogonal polynomials, e.g., Legendre polynomials. Define

$$\nu_i = \int_S p_i(\xi)\mu(d\xi). \tag{19}$$

Since $\{p_i(\xi)\}_{i=0}^{\infty}$ is not a set of orthogonal polynomials with respect to the measure $\mu(d\xi)$, $\nu_i$ is, in general, not equal to zero. For all the following numerical experiments we used the Legendre polynomials for $\{p_i(\xi)\}_{i=0}^{\infty}$.[2] Let $\hat{\alpha}_i$ and $\hat{\beta}_i$ be the coefficients in the three-term recurrence formula associated with the set $\{p_i\}$ of orthogonal polynomials.

Then, we initialize the process of building up the coefficients as

$$\begin{aligned}
&\hat{\mu}_{-1,j} = 0, \quad j = 1, 2, \dots, 2d - 2, \\
&\hat{\mu}_{0,j} = \nu_j, \quad j = 0, 2, \dots, 2d - 1, \\
&\alpha_0 = \hat{\alpha}_0 + \frac{\nu_1}{\nu_0}, \qquad \beta_0 = \nu_0,
\end{aligned}$$

and compute the following coefficients:

$$\hat{\mu}_{i,j} = \hat{\mu}_{i-1,j+1} - (\alpha_{i-1} - \hat{\alpha}_j)\hat{\mu}_{i-1,j} - \beta_{i-1}\hat{\mu}_{i-2,j} + \hat{\beta}_j \hat{\mu}_{i-1,j-1}, \tag{20}$$

where $j = i, i+1, \dots, 2d - i - 1$. The coefficients $\alpha_i$ and $\beta_i$ can be obtained as

$$\alpha_i = \hat{\alpha}_i + \frac{\hat{\mu}_{i,i+1}}{\hat{\mu}_{i,i}} - \frac{\hat{\mu}_{i-1,i}}{\hat{\mu}_{i-1,i-1}}, \qquad \beta_i = \frac{\hat{\mu}_{i,i}}{\hat{\mu}_{i-1,i-1}}. \tag{21}$$

Based on the modified moments, the ill-conditioning issue related to moments can be improved, although such an issue can still be severe especially when we consider orthogonality on infinite intervals.

The computational complexity for generating $\mu_{i,j}$ and $\nu_i$ is $\mathcal{O}(N)$.

[2] Legendre polynomials $\{p_i(\xi)\}_{i=0}^{\infty}$ are defined on $[-1, 1]$, therefore in implementation of the Modified Chebyshev method, we scale the measure onto $[-1, 1]$ first.

### 2.5. Lanczos method [4]

The idea of Lanczos method is to tridiagonalize a matrix to obtain the coefficients of the recurrence relation $\alpha_j$ and $\beta_j$. Suppose the discrete measure is $\mu = \sum_{i=1}^{N} \lambda_i \delta_{\tau_i}, \lambda_i > 0$. With weights $\lambda_i$ and $\tau_i$ in the expression of the measure $\mu$, the first step of this method is to construct a matrix:

$$\begin{pmatrix} 1 & \sqrt{\lambda_1} & \sqrt{\lambda_2} & \dots & \sqrt{\lambda_N} \\ \sqrt{\lambda_1} & \tau_1 & 0 & \dots & 0 \\ \sqrt{\lambda_2} & 0 & \tau_2 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ \sqrt{\lambda_N} & 0 & 0 & \dots & \tau_N \end{pmatrix}. \tag{22}$$

After we triagonalize it by the Lanczos algorithm, which is a process that reduces a symmetric matrix into a tridiagonal form with unitary transformations [16], we can obtain:

$$\begin{pmatrix} 1 & \sqrt{\beta_0} & 0 & \dots & 0 \\ \sqrt{\beta_0} & \alpha_0 & \sqrt{\beta_1} & \dots & 0 \\ 0 & \sqrt{\beta_1} & \alpha_1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \alpha_{N-1} \end{pmatrix}, \tag{23}$$

where the non-zero entries correspond to the coefficients $\alpha_i$ and $\beta_i$. Lanczos method is motivated by the interest in the inverse Sturm–Liouville problem: given some information on the eigenvalues of the matrix with a highly structured form, or some of its principal sub-matrices, this method is able to generate a symmetric matrix, either Jacobi or banded, in a finite number of steps. It is easy to program but can be considerably slow [4].

The computational complexity for the unitary transformation is $\mathcal{O}(N^2)$.

### 2.6. Gaussian quadrature rule associated with $\mu$

Here we describe how to utilize the above five methods to perform integration over a discrete measure numerically, using the Gaussian quadrature rule [17] associated with $\mu$.

We consider integrals of the form

$$\int_S f(\xi)\mu(d\xi) < \infty. \tag{24}$$

With respect to $\mu$, we generate the $\mu$-orthogonal polynomials up to order $d$ ($d \leq N-1$), denoted as $\{P_i(\xi)\}_{i=0}^{d}$, by one of the five methods in Sections 2.1–2.5. We calculated the zeros $\{\xi_i\}_{i=1}^{d}$ from $P_d(\xi) = a_d \xi^d + a_{d-1}\xi^{d-1} + \ldots + a_0$, as Gaussian quadrature points, and Gaussian quadrature weights $\{w_i\}_{i=1}^{d}$ by

$$w_i = \frac{a_d}{a_{d-1}} \frac{\int_S \mu(d\xi) P_{d-1}(\xi)^2}{P'_d(\xi_i) P_{d-1}(\xi_i)}. \tag{25}$$

Therefore, numerically the integral is approximated by

$$\int_S f(\xi)\mu(d\xi) \approx \sum_{i=1}^{d} f(\xi_i) w_i. \tag{26}$$

In the case when zeros for polynomial $P_d(\xi)$ do not have explicit formulas, Newton–Raphson is used [3,30], with a specified tolerance as $10^{-16}$ (in double precision). In order to ensure that at each search we find a new root, the polynomial deflation method [18] is applied, where the searched roots are factored out of the initial polynomial once they have been determined. All the calculations are done with double precision in this paper.

### 2.7. Orthogonality tests

To investigate the stability of the five methods, we perform an orthogonality test, where the orthogonality is defined as:

$$\mathrm{orth}(i) = \frac{1}{i} \sum_{j=0}^{i-1} \frac{\left| \int_S P_i(\xi) P_j(\xi) \mu(d\xi) \right|}{\sqrt{\int_S P_j^2(\xi)\mu(d\xi) \int_S P_i^2(x)\mu(d\xi)}}, \quad i \leq N-1, \tag{27}$$

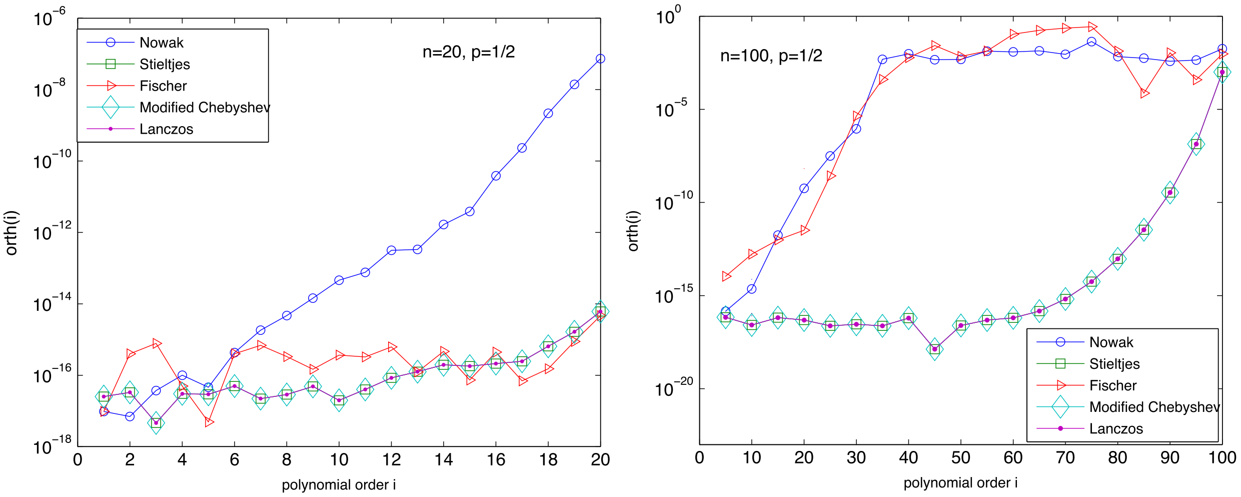

**Fig. 1.** Orthogonality defined in (27) with respect to the polynomial order $i$ up to 20 with distribution defined in (28) ($n = 20$, $p = 1/2$) (left) and $i$ up to 100 with ($n = 100$, $p = 1/2$) (right).

for the set $\{P_j(\xi)\}_{j=0}^{i}$ of orthogonal polynomials constructed numerically. Note that $\int_S P_i(\xi)P_j(\xi)\mu(d\xi) \neq 0$, $0 \leq j < i$, for orthogonal polynomials constructed numerically due to round-off errors, although they should be orthogonal theoretically.

We compare the numerical orthogonality given by the aforementioned five methods in Fig. 1 for the following distribution[3]:

$$f(k; n, p) = \mathbb{P}\left(\xi = \frac{2k}{n} - 1\right) = \frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}, \quad k = 0, 1, 2, ..., n. \tag{28}$$

We see that Stieltjes, Modified Chebyshev, and Lanczos methods preserve the best numerical orthogonality when the polynomial order $i$ is close to $N$. We notice that when $N$ is large, the numerical orthogonality is preserved up to the order of 70, indicating the robustness of these three methods. The Nowak method exhibits the worst numerical orthogonality among the five methods, due to the ill-conditioning nature of the matrix in Eq. (8). The Fischer method exhibits better numerical orthogonality when the number of data points $N$ in the discrete measure is small. The numerical orthogonality is lost when $N$ is large, which serves as a motivation to use ME-PCM instead of PCM for numerical integration over discrete measures. Our results suggest that we shall use Stieltjes, Modified Chebyshev, and Lanczos methods for more accuracy.

We also compare the cost by tracking the CPU time to evaluate (27) in Fig. 2: for a fixed polynomial order $i$, we track the CPU time with respect to $N$, the number of points in the discrete measure defined in (28); for a fixed $N$, we track the CPU time with respect to $i$. We observe that the Stieltjes method has the least computational cost while the Fischer method has the largest computational cost. Asymptotically, we observe that the computational complexity to evaluate (27) is $\mathcal{O}(i^2)$ for Nowak method, $\mathcal{O}(N)$ for the Stieltjes method, $\mathcal{O}(N^2)$ for the Fischer method, $\mathcal{O}(N)$ for the Modified Chebyshev method, and $\mathcal{O}(N^2)$ for the Lanczos method.

To conclude we recommend Stieltjes method as the most accurate and efficient in generating orthogonal polynomials with respect to discrete measures, especially when higher orders are required. However, for generating polynomials at lower orders (for ME-PCM), the five methods are equally effective.

We noticed from Figs. 1 and 2 that the Stieltjes method exhibits the most accuracy and efficiency in generating orthogonal polynomials with respect to a discrete measure $\mu$. Therefore, here we investigate the minimum polynomial order $i$ ($i \leq N - 1$) that the orthogonality orth($i$) defined in Eq. (27) of the Stieltjes method is larger than a threshold $\epsilon$. In Fig. 3, we perform this test on the distribution given by (28) with different parameters for $n$ ($n \geq i$). The highest polynomial order $i$ for polynomial chaos shall be less than the minimum $i$ that orth($i$) exceeds a certain desired $\epsilon$, for practical computations. The cost for numerical orthogonality is, in general, negligible compared to the cost for solving a stochastic problem by either Galerkin or collocation approaches. Hence, we can pay more attention on the accuracy, rather than the cost, of these five methods.

[3] We rescale the support for Binomial distribution with parameters $(n, p)$, $\{0, .., n\}$, onto $[-1, 1]$.

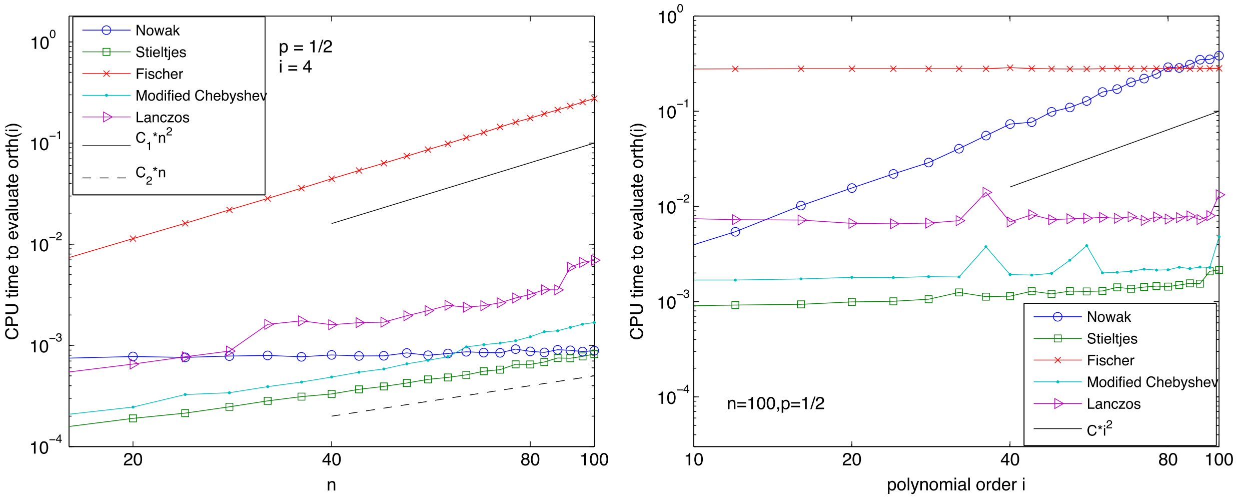

**Fig. 2.** CPU time (in seconds) on Intel (R) Core(TM) i5-3470 CPU @ 3.20 GHz in Matlab to evaluate orthogonality in (27) at the order $i = 4$ for distribution defined in (28) with parameter $n$ and $p = 1/2$ (left). CPU time to evaluate orthogonality in (27) at the order $i$ for distribution defined in (28) with parameter $n = 100$ and $p = 1/2$ (right).

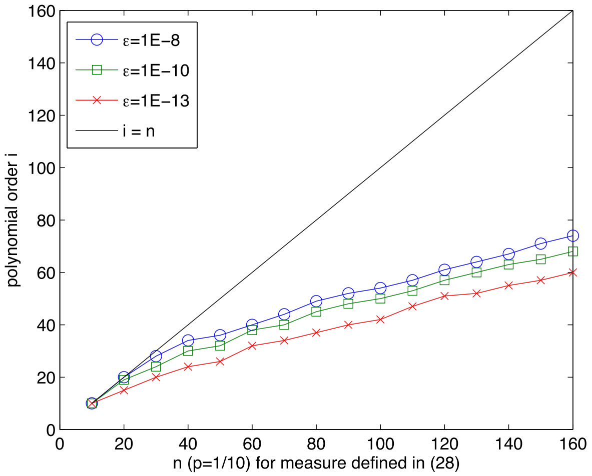

**Fig. 3.** Minimum polynomial order $i$ (vertical axis) such that orth($i$) defined in (27) is greater than a threshold value $\varepsilon$ (here $\varepsilon = 1E{-}8, 1E{-}10, 1E{-}13$), for distribution defined in (28) with $p = 1/10$. Orthogonal polynomials are generated by the Stieltjes method.

## 3. Discussion about the error of numerical integration

### 3.1. Theorem of numerical integration on discrete measure

In [12], the $h$-convergence rate of ME-PCM [18] for numerical integration in terms of continuous measures was established with respect to the degree of exactness given by the quadrature rule.

Let us first define the Sobolev space $W^{m+1,p}(\Gamma)$ to be the set of all functions $f \in L^p(\Gamma)$ such that for every multi-index $\gamma$ with $|\gamma| \le m+1$, the weak partial derivative $D^\gamma f$ belongs to $L^p(\Gamma)$ [1,8], i.e.

$$W^{m+1,p}(\Gamma) = \left\{ f \in L^p(\Gamma) : D^\gamma f \in L^p(\Gamma), \forall |\gamma| \le m+1 \right\}. \tag{29}$$

Here $\Gamma$ is an open set in $\mathbb{R}^n$ and $1 \le p \le +\infty$. The natural number $m+1$ is called the order of the Sobolev space $W^{m+1,p}(\Gamma)$. Here the Sobolev space $W^{m+1,\infty}(A)$ in the following theorem is defined for functions $f : A \to \mathbb{R}$ subject to the norm:

$$\|f\|_{m+1,\infty,A} = \max_{|\boldsymbol{\gamma}| \le m+1} \operatorname*{ess\,sup}_{\xi \in A} \left| D^\gamma f(\xi) \right|,$$

and the seminorm is defined as:

$$|f|_{m+1,\infty,A} = \max_{|\boldsymbol{\gamma}|=m+1} \operatorname*{ess\,sup}_{\xi\in A} \left|D^{\gamma} f(\xi)\right|,$$

where $A \subset \mathbb{R}^n$, $\gamma \in \mathbb{N}_0^n$, $|\boldsymbol{\gamma}| = \gamma_1 + \ldots + \gamma_n$ and $m+1 \in \mathbb{N}_0$.

We first consider a one-dimensional discrete measure $\mu = \sum_{i=1}^{N} \lambda_i \delta_{\tau_i}$, where $N$ is a finite number. For simplicity and without loss of generality, we assume that $\{\tau_i\}_{i=1}^{N} \subset (0,1)$. Otherwise, we can use a linear mapping to map $(\min\{\tau_i\}_{i=1}^{N} - c, \max\{\tau_i\}_{i=1}^{N} + c)$ to $(0,1)$ with $c$ being a arbitrarily small positive number. We then construct the approximation of the Dirac measure as

$$\mu_\varepsilon = \sum_{i=1}^{N} \lambda_i \eta_{\tau_i}^{\varepsilon}, \tag{30}$$

where $\varepsilon$ is a small positive number and $\eta_{\tau_i}^{\varepsilon}$ is defined as

$$\eta_{\tau_i}^{\varepsilon} = \begin{cases} \frac{1}{\varepsilon} & \text{if } |\xi - \tau_i| < \varepsilon/2, \\ 0 & \text{otherwise.} \end{cases} \tag{31}$$

First of all, $\eta_{\tau_i}^{\varepsilon}$ defines a continuous measure in $(0,1)$ with a finite number of discontinuous points, where a uniform distribution is taken on the interval $(\tau_i - \varepsilon/2, \tau_i + \varepsilon/2)$. Second, $\eta_{\tau_i}^{\varepsilon}$ converges to $\delta_{\tau_i}$ in the weak sense, i.e.,

$$\lim_{\varepsilon\to 0^+} \int_0^1 g(\xi)\eta_{\tau_i}^{\varepsilon}(d\xi) = \int_0^1 g(\xi)\delta_{\tau_i}(d\xi), \tag{32}$$

for all bounded continuous functions $g(\xi)$. We write that

$$\lim_{\varepsilon\to 0^+} \eta_{\tau_i}^{\varepsilon} = \delta_{\tau_i}. \tag{33}$$

It is seen that when $\varepsilon$ is small enough, the intervals $(\tau_i - \varepsilon/2, \tau_i + \varepsilon/2)$ can be mutually disjoint for $i = 1, \ldots, N$. Due to the linearity, we have

$$\lim_{\varepsilon\to 0^+} \mu_\varepsilon = \mu, \tag{34}$$

and the convergence is defined in the weak sense as before. Then, $\mu_\varepsilon$ is also a continuous measure with a finite number of discontinuous points.

**Remark 1.** The choice for $\eta_{\tau_i}^{\varepsilon}$ is not unique. Another choice is

$$\eta_{\tau_i}^{\varepsilon} = \frac{1}{\varepsilon}\eta\left(\frac{\xi - \tau_i}{\varepsilon}\right), \qquad \eta(\xi) = \begin{cases} e^{-\frac{1}{1-|\xi|^2}} & \text{if } |\xi| < 1, \\ 0 & \text{otherwise.} \end{cases} \tag{35}$$

Such a choice is smooth. When $\varepsilon$ is small enough, the domains defined by $|\frac{\xi-\tau_i}{\varepsilon}| < 1$ are also mutually disjoint.

We then have the following proposition.

**Proposition 1.** *For the continuous measure $\mu_\varepsilon$, we let $\alpha_{i,\varepsilon}$ and $\beta_{i,\varepsilon}$ indicate the coefficients in the three-term recurrence formula (3), which is valid for both continuous and discrete measures. For the discrete measure $\mu$, we let $\alpha_i$ and $\beta_i$ indicate the coefficients in the three-term recurrence formula. We then have*

$$\lim_{\varepsilon\to 0^+} \alpha_{i,\varepsilon} = \alpha_i, \qquad \lim_{\varepsilon\to 0^+} \beta_{i,\varepsilon} = \beta_i. \tag{36}$$

*In other words, the monic orthogonal polynomials defined by $\mu_\varepsilon$ will converge to those defined by $\mu$, i.e*

$$\lim_{\varepsilon\to 0^+} P_{i,\varepsilon}(\xi) = P_i(\xi), \tag{37}$$

*where $P_{i,\varepsilon}$ and $P_i$ are monic polynomials of order i corresponding to $\mu_\varepsilon$ and $\mu$, respectively.*

**Proof.** The coefficients $\alpha_{i,\varepsilon}$ and $\beta_{i,\varepsilon}$ are given by the formula, see Eq. (9),

$$\alpha_{i,\varepsilon} = \frac{(\xi P_{i,\varepsilon}, P_{i,\varepsilon})_{\mu_\varepsilon}}{(P_{i,\varepsilon}, P_{i,\varepsilon})_{\mu_\varepsilon}}, \quad i = 0, 1, 2, \ldots, \tag{38}$$

$$\beta_{i,\varepsilon} = \frac{(P_{i,\varepsilon}, P_{i,\varepsilon})_{\mu_\varepsilon}}{(P_{i-1,\varepsilon}, P_{i-1,\varepsilon})_{\mu_\varepsilon}}, \quad i = 1, 2, \ldots, \tag{39}$$

where $(\cdot,\cdot)_{\mu_\varepsilon}$ indicates the inner product with respect to $\mu_\varepsilon$. Correspondingly, we have

$$\alpha_i = \frac{(\xi P_i, P_i)_\mu}{(P_i, P_i)_\mu}, \quad i = 0, 1, 2, \dots, \tag{40}$$

$$\beta_i = \frac{(P_i, P_i)_\mu}{(P_{i-1,i-1})_\mu}, \quad i = 1, 2, \dots. \tag{41}$$

By definition,

$$\beta_{0,\varepsilon} = (1,1)_{\mu_\varepsilon} = 1, \qquad \beta_0 = (1,1)_\mu = 1.$$

The argument is based on induction. We assume that Eq. (37) is true for $k = i$ and $k = i - 1$. When $i = 0$, this is trivial. To show that Eq. (37) holds for $k = i + 1$, we only need to prove Eq. (36) for $k = i$ based on the observation that $P_{i+1,\varepsilon} = (\xi - \alpha_{i,\varepsilon})P_{i,\varepsilon} - \beta_{i,\varepsilon}P_{i-1,\varepsilon}$. We now show that all inner products in Eqs. (38) and (39) converges to the corresponding inner products in Eqs. (40) and (41) as $\varepsilon \to 0^+$. We here only consider $(P_{i,\varepsilon}, P_{i,\varepsilon})_{\mu_\varepsilon}$ and other inner products can be dealt with in a similar way. We have

$$(P_{i,\varepsilon}, P_{i,\varepsilon})_{\mu_\varepsilon} = (P_i, P_i)_{\mu_\varepsilon} + 2(P_i, P_{i,\varepsilon} - P_i)_{\mu_\varepsilon} + (P_{i,\varepsilon} - P_i, P_{i,\varepsilon} - P_i)_{\mu_\varepsilon}$$

We then have $(P_i, P_i)_{\mu_\varepsilon} \to (P_i, P_i)_\mu$ due to the definition of $\mu_\varepsilon$. The second term on the right-hand side can be bounded as

$$\left|(P_i, P_{i,\varepsilon} - P_i)_{\mu_\varepsilon}\right| \le \operatorname*{ess\,sup}_\xi P_i \operatorname*{ess\,sup}_\xi (P_{i,\varepsilon} - P_i)(1,1)_{\mu_\varepsilon}.$$

According to the assumption that $P_{i,\varepsilon} \to P_i$, the right-hand side of the above inequality goes to zero. Similarly, $(P_{i,\varepsilon} - P_i, P_{i,\varepsilon} - P_i)_{\mu_\varepsilon}$ goes to zero. We then have $(P_{i,\varepsilon}, P_{i,\varepsilon})_{\mu_\varepsilon} \to (P_i, P_i)_\mu$. The conclusion is then achieved by induction. □

**Remark 2.** Since as $\varepsilon \to 0^+$, the orthogonal polynomials defined by $\mu_\varepsilon$ will converge to those defined by $\mu$. The (Gauss) quadrature points and weights defined by $\mu_\varepsilon$ should also converge to those defined by $\mu$.

We then recall the following theorem for continuous measures.

**Theorem 1.** *(See [12].) Suppose $f \in W^{m+1,\infty}(\Gamma)$ with $\Gamma = (0,1)^n$, and $\{B^i\}_{i=1}^{N_e}$ is a non-overlapping mesh of $\Gamma$. Let h indicate the maximum side length of each element and $\mathcal{Q}_m^\Gamma$ a quadrature rule with degree of exactness m in domain $\Gamma$. (In other words $\mathcal{Q}_m$ exactly integrates polynomials up to order m.) Let $\mathcal{Q}_m^A$ be the quadrature rule in subset $A \subset \Gamma$, corresponding to $\mathcal{Q}_m^\Gamma$ through an affine linear mapping. We define a linear functional on $W^{m+1,\infty}(A)$:*

$$E_A(g) \equiv \int_A g(\xi)\mu(d\xi) - \mathcal{Q}_m^A(g), \tag{42}$$

*whose norm in the dual space of $W^{m+1,\infty}(A)$ is defined as*

$$\|E_A\|_{m+1,\infty,A} = \sup_{\|g\|_{m+1,\infty,A} \le 1} \left|E_A(g)\right|. \tag{43}$$

*Then, the following error estimate holds:*

$$\left|\int_\Gamma f(\xi)\mu(d\xi) - \sum_{i=1}^{N_e} \mathcal{Q}_m^{B^i} f\right| \le Ch^{m+1}\|E_\Gamma\|_{m+1,\infty,\Gamma}|f|_{m+1,\infty,\Gamma} \tag{44}$$

*where C is a constant and $\|E_\Gamma\|_{m+1,\infty,\Gamma}$ refers to the norm in the dual space of $W^{m+1,\infty}(\Gamma)$, which is defined in Eq. (43).*

For discrete measures, we have the following theorem.

**Theorem 2.** *Suppose the function f satisfies all assumptions required by Theorem 1. We add the following three assumptions for discrete measures: 1) The measure $\mu$ can be expressed as a product of n one-dimensional discrete measures, i.e., we consider n independent discrete random variables; 2) The quadrature rule $\mathcal{Q}_m^A$ can be generated from the quadrature rules given by the n one-dimensional discrete measures by the tensor product; 3) The number of all the possible values for the discrete measure $\mu$ is finite and they are located within $\Gamma$. We then have*

$$\left|\int_\Gamma f(\xi)\mu(d\xi) - \sum_{i=1}^{N_e} \mathcal{Q}_m^{B^i} f\right| \le CN_{es}^{-m-1}\|E_\Gamma\|_{m+1,\infty,\Gamma}|f|_{m+1,\infty,\Gamma}, \tag{45}$$

*where $N_{es}$ indicates the number of integration elements for each random variable.*

**Proof.** The argument is based on Theorem 1 and the approximation $\mu_\varepsilon$ of $\mu$. Since we assume that $\mu$ is given by $n$ independent discrete random variables, we can define a continuous approximation (see Eq. (30)) for each one-dimensional discrete measure and $\mu_\varepsilon$ can be naturally chosen as the product of these $n$ continuous one-dimensional measures.

We then consider

$$\begin{aligned}\left|\int_\Gamma f(\xi)\mu(d\xi) - \sum_{i=1}^{N_e} \mathcal{Q}_m^{B^i} f\right| \le & \left|\int_\Gamma f(\xi)\mu(d\xi) - \int_\Gamma f(\xi)\mu_\varepsilon(d\xi)\right| \\ & + \left|\int_\Gamma f(\xi)\mu_\varepsilon(d\xi) - \sum_{i=1}^{N_e} \mathcal{Q}_m^{\varepsilon,B^i} f\right| \\ & + \left|\sum_{i=1}^{N_e} \mathcal{Q}_m^{\varepsilon,B^i} f - \sum_{i=1}^{N_e} \mathcal{Q}_m^{B^i} f\right|,\end{aligned}$$

where $\mathcal{Q}_m^{\varepsilon,B^i}$ defines the corresponding quadrature rule for the continuous measure $\mu_\varepsilon$. Since we assume that the quadrature rules $\mathcal{Q}_m^{\varepsilon,B_i}$ and $\mathcal{Q}_m^{B_i}$ can be constructed by $n$ one-dimensional quadrature rules, $\mathcal{Q}_m^{\varepsilon,B_i}$ should converge to $Q_m^{B_i}$ as $\varepsilon$ goes to zero based on Proposition 1 and the fact that the construction procedure for $\mathcal{Q}_m^{\varepsilon,B_i}$ and $\mathcal{Q}_m^{B_i}$ to have a degree of exactness $m$ is measure independent. For the second term on the right-hand side, Theorem 1 can be applied with a well-defined $h$ because we assume that all possible values for $\mu$ are located within $\Gamma$, otherwise, this assumption can be achieved by a linear mapping. We then have

$$\left|\int_\Gamma f(\xi)\mu_\varepsilon(d\xi) - \sum_{i=1}^{N_e} \mathcal{Q}_m^{\varepsilon,B^i} f\right| \le Ch^{m+1}\|E_\Gamma^\varepsilon\|_{m+1,\infty,\Gamma}|f|_{m+1,\infty,\Gamma}, \tag{46}$$

where $E_\Gamma^\varepsilon$ is a linear functional defined with respect to $\mu_\varepsilon$. We then let $\varepsilon \to 0^+$. In the error bound given by Eq. (46), only $\|E_\Gamma^\varepsilon\|_{m+1,\infty,\Gamma}$ is associated with $\mu_\varepsilon$. According to its definition and noting that $\mathcal{Q}_m^{\varepsilon,A} \to \mathcal{Q}_m^A$,

$$\lim_{\varepsilon\to 0} E_A^\varepsilon(g) = \lim_{\varepsilon\to 0}\left(\int_A g(\xi)\mu_\varepsilon(d\xi) - \mathcal{Q}_m^{\varepsilon,A}(g)\right) = E_A(g),$$

which is a linear functional with respect to $\mu$. Since $\mu_\varepsilon \to \mu$ and $\mathcal{Q}_m^{\varepsilon,B_i} \to \mathcal{Q}_m^{B_i}$, the first and third term will go to zero. However, since we are working with discrete measures, it is not convenient to use the element size. Instead we use the number of elements since $h \propto N_{es}^{-1}$, where $N_{es}$ indicates the number of elements per side. Then the conclusion is reached. □

**Remark 3.** The $h$-convergence rate of ME-PCM for discrete measures takes the form $O(N_{es}^{-(m+1)})$. If we employ Gauss quadrature rule with $d$ points, the degree of exactness is $m = 2d - 1$, which corresponds to a $h$-convergence rate $N_{es}^{-2d}$.

**Remark 4.** The extra assumptions in Theorem 2 are actually quite practical. In applications, we often consider i.i.d. random variables and the commonly used quadrature rules for high-dimensional cases, such as tensor-product rule and sparse grids, are obtained from one-dimensional quadrature rules.

### *3.2. Testing numerical integration with one RV*

We now verify the $h$-convergence rate numerically. We employ the Lanczos method [4] to generate the Gauss quadrature points. We then approximate integrals of GENZ functions [15] with respect to the binomial distribution $Bino(n = 120, p = 1/2)$ using ME-PCM. We consider the following one-dimensional GENZ functions:

- GENZ1 function deals with oscillatory integrands:

$$f_1(\xi) = \cos(2\pi w + c\xi), \tag{47}$$

- GENZ4 function deals with Gaussian-like integrands:

$$f_4(\xi) = \exp\big(-c^2(\xi - w)^2\big), \tag{48}$$

where $c$ and $w$ are constants. Note that both GENZ1 and GENZ4 functions are smooth. In this section, we consider the absolute error defined as $|\int_S f(\xi)\mu(d\xi) - \sum_{i=1}^d f(\xi_i)w_i|$, where $\{\xi_i\}$ and $\{w_i\}$ $(i = 1, \ldots, d)$ are $d$ Gauss quadrature points and weights with respect to $\mu$, explained in Section 2.6.

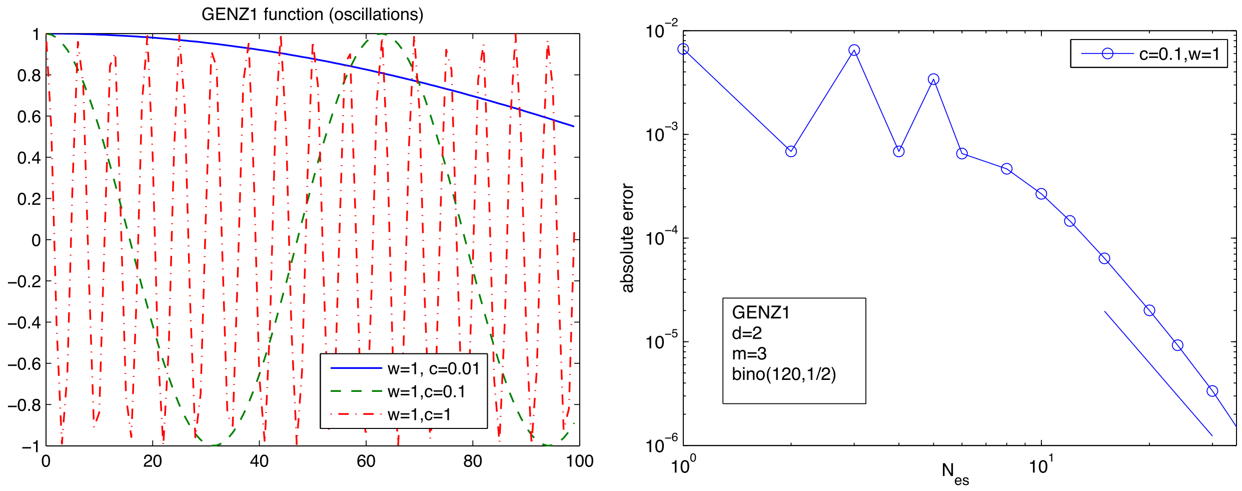

**Fig. 4.** Left: GENZ1 functions with different values of $c$ and $w$; Right: $h$-convergence of ME-PCM for function GENZ1. Two Gauss quadrature points, $d = 2$, are employed in each element corresponding to a degree $m = 3$ of exactness. $c = 0.1$, $w = 1$, $\xi \sim Bino(120, 1/2)$. Lanczos method is employed to compute the orthogonal polynomials.

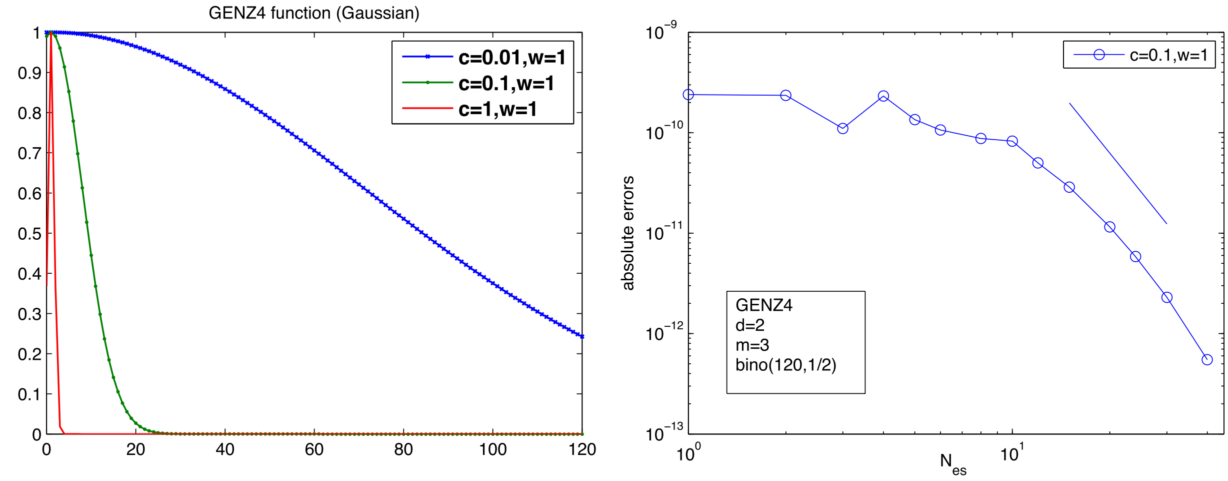

**Fig. 5.** Left: GENZ4 functions with different values of $c$ and $w$; Right: $h$-convergence of ME-PCM for function GENZ4. Two Gauss quadrature points, $d = 2$, are employed in each element corresponding to a degree $m = 3$ of exactness. $c = 0.1$, $w = 1$, $\xi \sim Bino(120, 1/2)$. Lanczos method is employed for numerical orthogonality.

In Figs. 4 and 5, we plot the $h$-convergence behavior of ME-PCM for GENZ1 and GENZ4 functions, respectively. In each element, two Gauss quadrature points are employed, corresponding to a degree 3 of exactness, which means that the $h$-convergence rate should be $N_{es}^{-4}$. In Figs. 4 and 5, we see that when $N_{es}$ is large enough, the $h$-convergence rate of ME-PCM approaches the theoretical prediction, demonstrated by the reference straight lines $CN_{es}^{-4}$.

### 3.3. Testing numerical integration with multiple RVs on sparse grids

An interesting question is if the sparse grid approach is as effective for discrete measures as it is for continuous measures [27], and how that compares to the tensor product grids. Let us denote the sparse grid level by $k$ and the dimension by $n$. Assume that each random dimension is independent. We apply the Smolyak algorithm [23,20,21] to construct sparse grids, i.e.,

$$A(k+n,n) = \sum_{k+1 \le |\mathbf{i}| \le k+n} (-1)^{k+n-|\mathbf{i}|} \binom{n-1}{k+n-|\mathbf{i}|} \left(U^{i_1} \otimes \ldots \otimes U^{i_n}\right), \tag{49}$$

where $A(k+n,n)$ defines a cubature formula with respect to the $n$-dimensional discrete measure and $U^{i_j}$ defines the quadrature rule of $i$-th level for the $j$-th dimension [27].

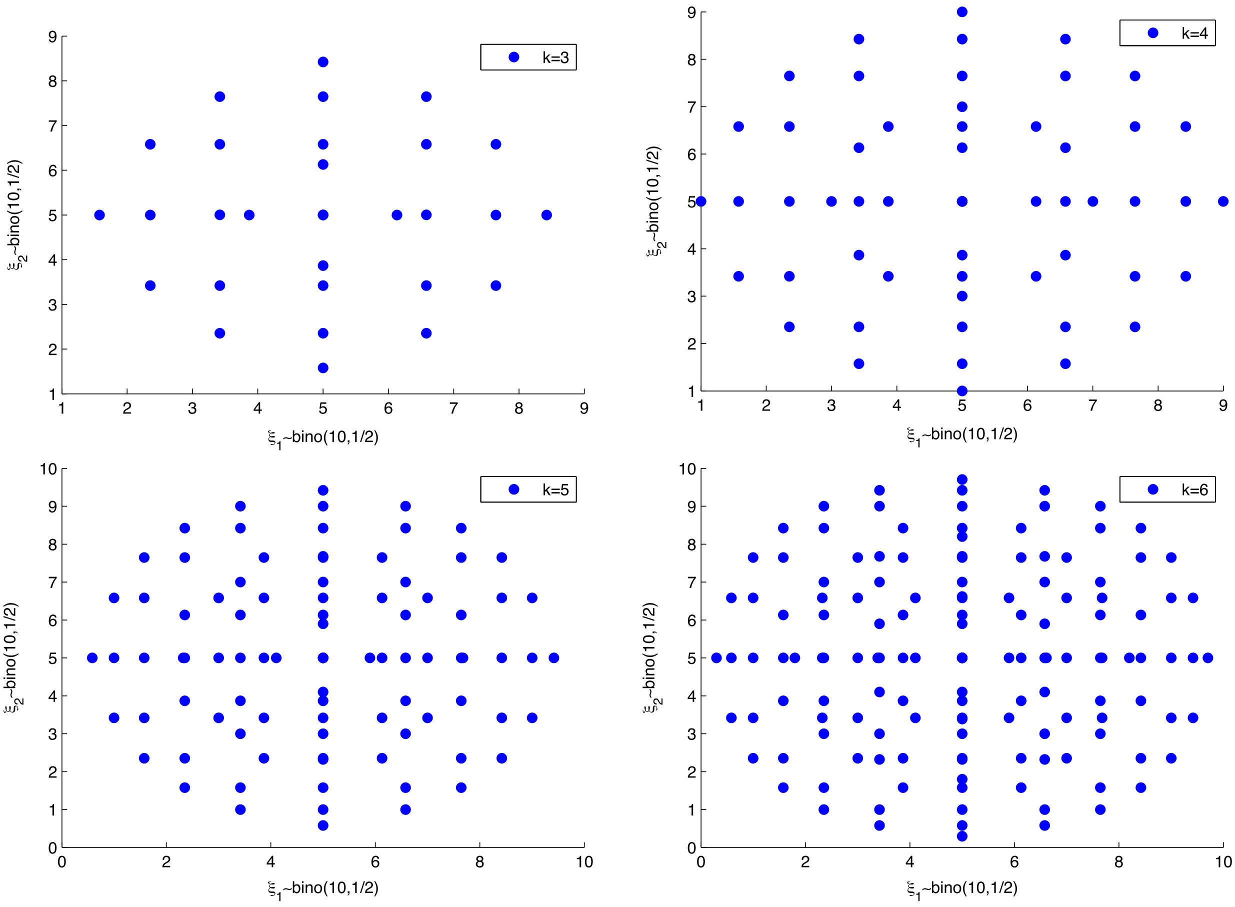

**Fig. 6.** Non-nested sparse grid points with respect to sparseness parameter $k = 3, 4, 5, 6$ for random variables $\xi_1, \xi_2 \sim Bino(10, 1/2)$, where the one-dimensional quadrature formula is based on Gauss quadrature rule.

We use Gauss quadrature rule to define $U^{i_j}$, which implies that the grids at different levels are not necessarily nested. Two-dimensional non-nested sparse grid points are plotted in Fig. 6, where each dimension has the same discrete measure as binomial distribution $Bino(10, 1/2)$. We then use sparse grids to approximate the integration of the following two GENZ functions with $M$ RVs [15]:

- GENZ1

$$f_1(\xi_1, \xi_2, ..., \xi_M) = \cos\left(2\pi w_1 + \sum_{i=1}^{M} c_i \xi_i\right) \tag{50}$$

- GENZ4

$$f_4(\xi_1, \xi_2, ..., \xi_M) = \exp\left[-\sum_{i=1}^{M} c_i^2 (\xi_i - w_i)^2\right] \tag{51}$$

where $c_i$ and $w_i$ are constants. We compute $\mathbb{E}[f_i(\xi_1, \xi_2, ..., \xi_M)]$ under the assumption that $\{\xi_i, i = 1, ..., M\}$ are $M$ independent identically distributed (i.i.d.) random variables. The absolute errors versus the total number of sparse grid points $r(k)$ with $k$ being the sparse grid level, are plotted in Figs. 7 and 8, for two RVs and eight RVs respectively. We see that the sparse grids for discrete measures work well for smooth GENZ1 and GENZ4 functions, and the convergence rate is much faster than the Monte Carlo simulations with a convergence rate $O(r(k)^{-1/2})$. In low dimensions, it is known that integration on sparse grids converges slower than on tensor product grids [27] for continuous measures based on numerical tests. We observe the same trend in Fig. 7 for discrete measures. The error line from the tensor product grid has a slight up bending at its tail because the error is near the machine error ($1E - 16$). In higher dimensions sparse grids are more efficient than tensor product grids as in Fig. 8 for discrete measures. In Section 4.2.3, we will obtain the numerical solution of the KdV equation with eight RVs, where sparse grids are also more accurate than tensor product grids.

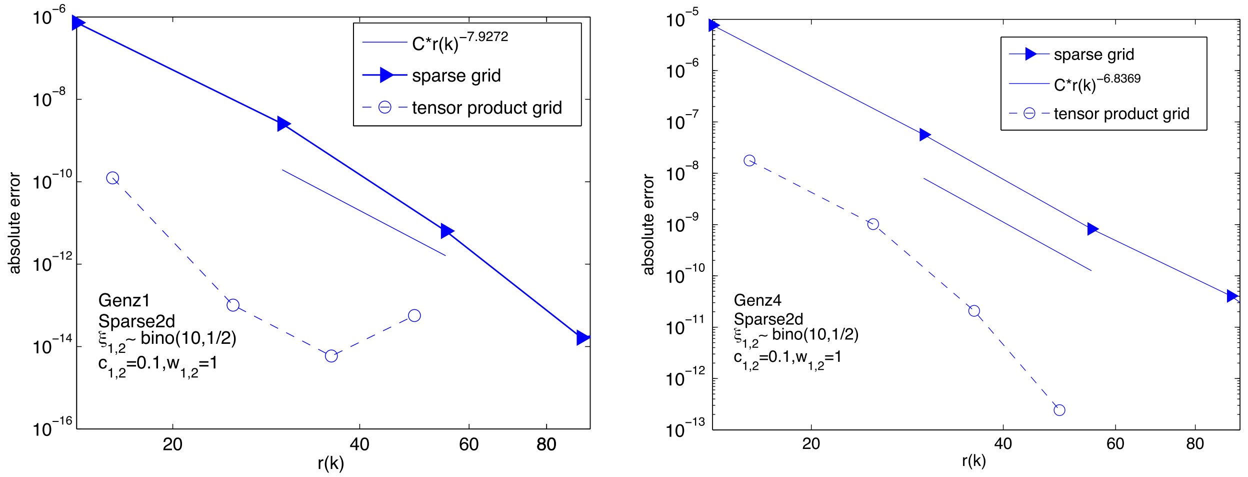

**Fig. 7.** Convergence of sparse grids and tensor product grids to approximate $\mathbb{E}[f_i(\xi_1,\xi_2)]$, where $\xi_1$ and $\xi_2$ are two i.i.d. random variables associated with a distribution $Bino(10,1/2)$. Left: $f_1$ is GENZ1; Right: $f_4$ is GENZ4. Orthogonal polynomials are generated by Lanczos method.

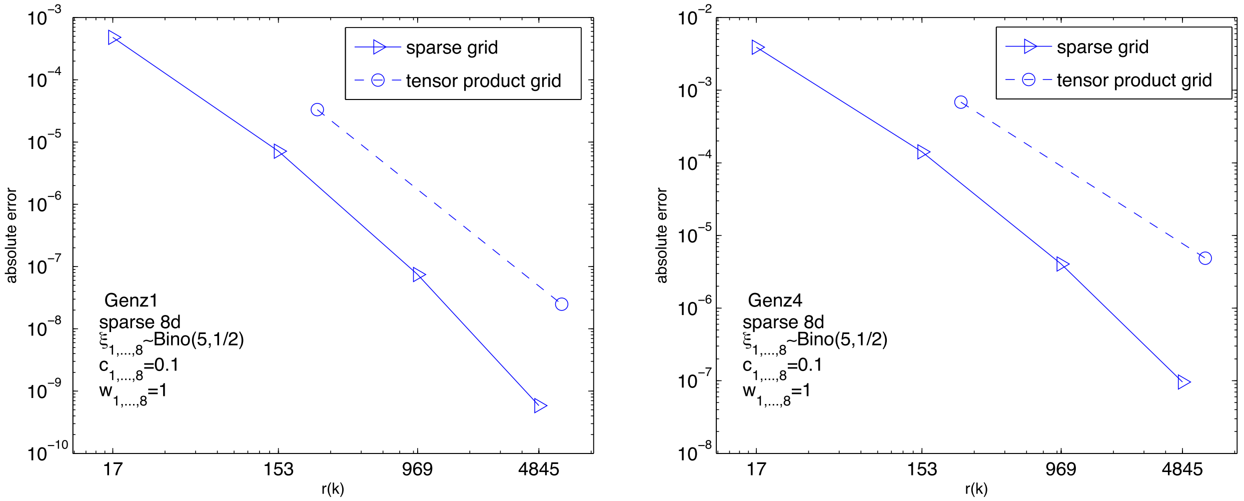

**Fig. 8.** Convergence of sparse grids and tensor product grids to approximate $\mathbb{E}[f_i(\xi_1,\xi_2,...,\xi_8)]$, where $\xi_1,...,\xi_8$ are eight i.i.d. random variables associated with a distribution $Bino(10,1/2)$. Left: $f_1$ is GENZ1; Right: $f_4$ is GENZ4. Orthogonal polynomials are generated by Lanczos method.

## 4. Application to stochastic reaction equation and KdV equation

For numerical experiments on SPDEs, we choose one method among Nowak, Stieltjes, Fischer, and Lanczos methods to generate orthogonal polynomials, in order to calculate the moment statistics by Gaussian quadrature rule associated with the discrete measure. Other methods will provide identical results.

### *4.1. Reaction equation with discrete random coefficients*

We first consider the reaction equation with a random coefficient:

$$\frac{dy(t;\xi)}{dt} = -\xi y(t;\xi), \tag{52}$$

with initial condition

$$y(0;\xi) = y_0, \tag{53}$$

where $\xi$ is a random coefficient. Let us define the error of mean and variance of the solution to be

$$\epsilon_{\mathrm{mean}}(t) = \left|\frac{\mathbb{E}_{\mathrm{PCM}}[y(t)] - \mathbb{E}_{\mathrm{exact}}[y(t)]}{\mathbb{E}_{\mathrm{exact}}[y(t)]}\right|, \tag{54}$$

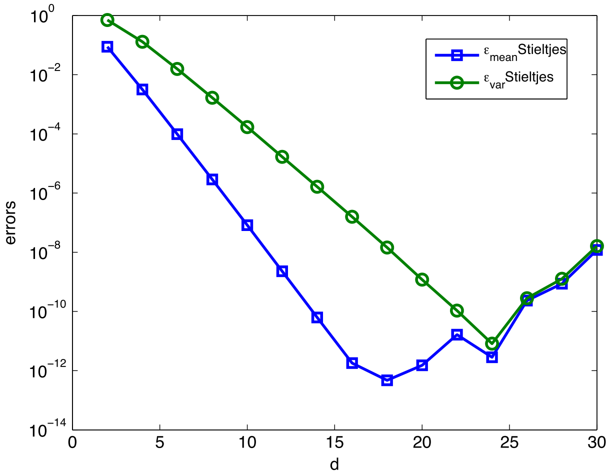

**Fig. 9.** $p$-Convergence of PCM with respect to errors defined in Eqs. (54) and (55) for the reaction equation with $t = 1$, $y_0 = 1$. $\xi$ is associated with negative binomial distribution with $c = \frac{1}{2}$ and $\beta = 1$. Orthogonal polynomials are generated by the Stieltjes method.

and

$$\epsilon_{\text{var}}(t) = \left| \frac{Var_{\text{PCM}}[y(t)] - Var_{\text{exact}}[y(t)]}{Var_{\text{exact}}[y(t)]} \right|. \tag{55}$$

The exact value of the $m$-th moment of the solution is:

$$\mathbb{E}\big[y^m(t;\xi)\big] = \mathbb{E}\big[\big(y_0 e^{-\xi t}\big)^m\big]. \tag{56}$$

The error defined in Eqs. (54) and (55) of solution for Eq. (52) has been considered in the literature by gPC [28] with Wiener–Askey polynomials [2] with respect to discrete measures. Here instead of using hypergeometric polynomials in the Wiener–Askey scheme, we solve Eq. (52) by PCM with collocation points generated by the Stieltjes method. The $p$-convergence is demonstrated in Fig. 9 for the negative binomial distribution with $\beta = 1$, $c = \frac{1}{2}$. We observe spectral convergence by polynomial chaos with orthogonal polynomials generated by the Stieltjes method, and the method is accurate up to order 15 here.

### 4.2. KdV equation with random forcing

We subsequently consider the KdV equation subject to stochastic forcing:

$$u_t + 6uu_x + u_{xxx} = \sigma\xi, \quad x \in \mathbb{R}, \tag{57}$$

with initial condition:

$$u(x,0) = \frac{a}{2}\operatorname{sech}^2\left(\frac{\sqrt{a}}{2}(x - x_0)\right), \tag{58}$$

where $a$ is associated with the speed of the soliton, $x_0$ is the initial position of the soliton, and $\sigma$ is a constant that scales the variance of the random variable (RV) $\xi$. The $m$-th moment of the solution is:

$$\mathbb{E}\big[u^m(x,t;\xi)\big] = \mathbb{E}\left[\left(\frac{a}{2}\operatorname{sech}^2\left(\frac{\sqrt{a}}{2}(x - 3\sigma\xi t^2 - x_0 - at)\right) + \sigma\xi t\right)^m\right]. \tag{59}$$

Derivation of the exact solution and verification of the accuracy of the deterministic solver are presented in Appendix A.

To examine the convergence of ME-PCM, we define the following normalized $L_2$ errors for the mean and the second-moment as:

$$l2u1 = \frac{\sqrt{\int dx(\mathbb{E}[u_{\text{num}}(x,t;\xi)] - \mathbb{E}[u_{\text{ex}}(x,t;\xi)])^2}}{\sqrt{\int dx(\mathbb{E}[u_{\text{ex}}(x,t;\xi)])^2}}, \tag{60}$$

$$l2u2 = \frac{\sqrt{\int dx(\mathbb{E}[u^2_{\text{num}}(x,t;\xi)] - \mathbb{E}[u^2_{\text{ex}}(x,t;\xi)])^2}}{\sqrt{\int dx(\mathbb{E}[u^2_{\text{ex}}(x,t;\xi)])^2}}, \tag{61}$$

where $u_{\text{num}}$ and $u_{\text{ex}}$ indicate the numerical and exact solutions, respectively.

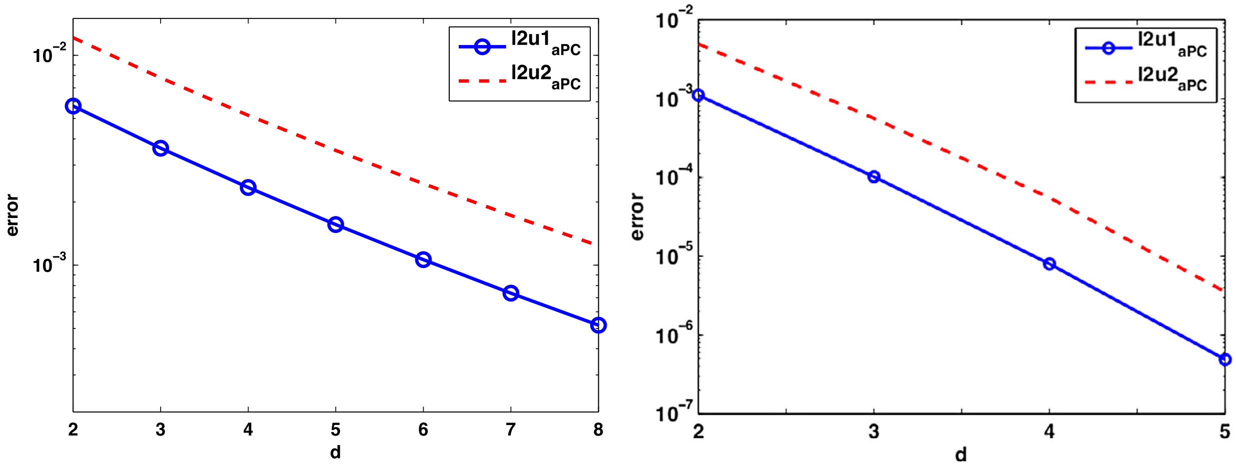

**Fig. 10.** $p$-convergence of PCM with respect to errors defined in Eqs. (60) and (61) for the KdV equation with $t = 1$. $a = 1$, $x_0 = -5$ and $\sigma = 0.2$, with 200 Fourier collocation points on the spatial domain $[-30, 30]$. Left: $\xi \sim$ Pois(10); Right: $\xi \sim Bino(n = 5, p = 1/2)$. aPC stands for arbitrary Polynomial Chaos, which is Polynomial Chaos with respect to arbitrary measure. Orthogonal polynomials are generated by Fischer's method.

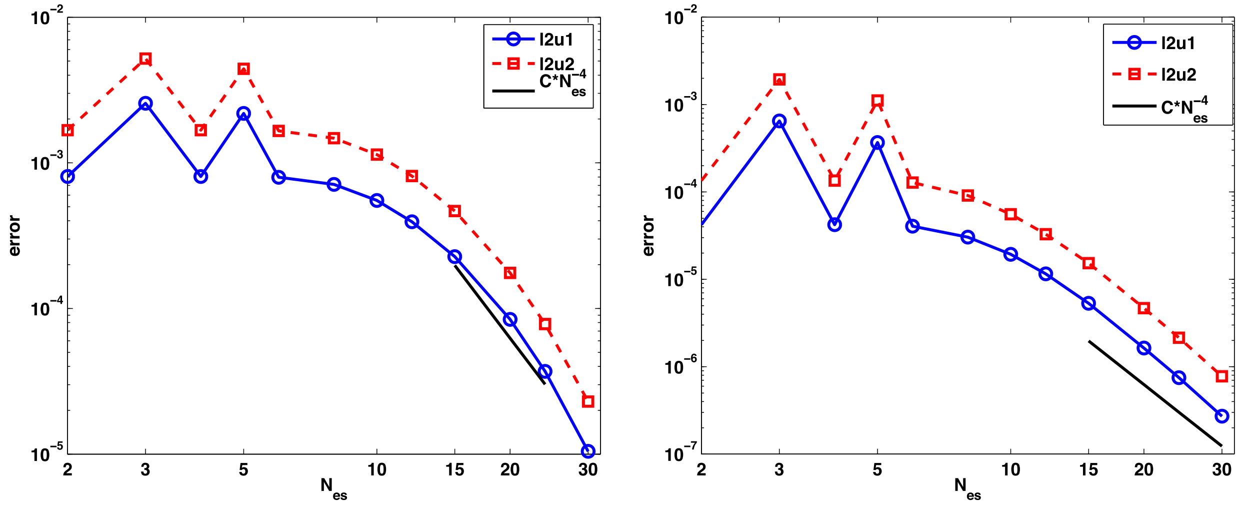

**Fig. 11.** $h$-convergence of ME-PCM with respect to errors defined in Eqs. (60) and (61) for the KdV equation with $t = 1.05$, $a = 1$, $x_0 = -5$, $\sigma = 0.2$, and $\xi \sim Bino(n = 120, p = 1/2)$, with 200 Fourier collocation points on the spatial domain $[-30, 30]$, where two collocation points are employed in each element. Orthogonal polynomials are generated by the Fischer method (left) and the Stieltjes method (right).

We solve Eq. (57) by PCM with collocation points generated by Fischer's method. The $p$-convergence is demonstrated in Fig. 10 for distributions Pois(10) and $Bino(n = 5, p = 1/2)$, respectively, with respect to errors defined in Eqs. (60) and (61). For the $h$-convergence of ME-PCM we examine the distribution $Bino(n = 120, p = 1/2)$, where each element contains the same number of discrete data points. Furthermore, in each element we employ two Gauss quadrature points for the gPC approximation. We see in Fig. 11 that the desired $h$-convergence rate $N_{es}^{-4}$ is obtained for both Stieltjes and Fischer method. We note that all five methods exhibit the same convergence rate and the same error level except the Fischer method, which exhibits errors by two orders of magnitude larger. To explain this, we refer to Fig. 1, which shows that if the number of points is large, the orthogonality condition in Fischer's method suffers from the round-off errors.

We now consider the adaptive ME-PCM, where the local variance criterion for adaptivity is employed (see Appendix B for more details). A five-element adaptive decomposition of the parametric space for the distribution $\xi \sim$ Pois(40) is given in Fig. 12. We see that in the region of small probability, the element size is large while in the region of high probability, the element size is much smaller. We then examine the effectiveness of adaptivity. Consider a uniform mesh and an adapted one, which have the same number of elements and the same number of collocation points within each element. In Fig. 13, we plot the $p$-convergence behavior of ME-PCM given by the uniform and adapted meshes. We see that although both meshes yield exponential convergence, the adapted mesh results in a better accuracy especially when the number of elements is relatively small. In other words, for a certain accuracy, the adapted ME-PCM can be more efficient than ME-PCM on a uniform mesh.

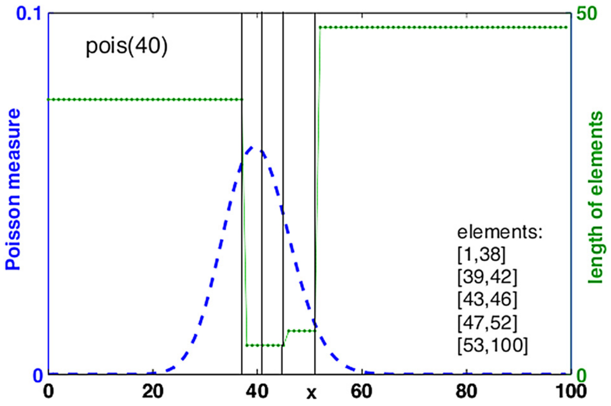

**Fig. 12.** Adapted mesh with five elements with respect to Pois(40) distribution.

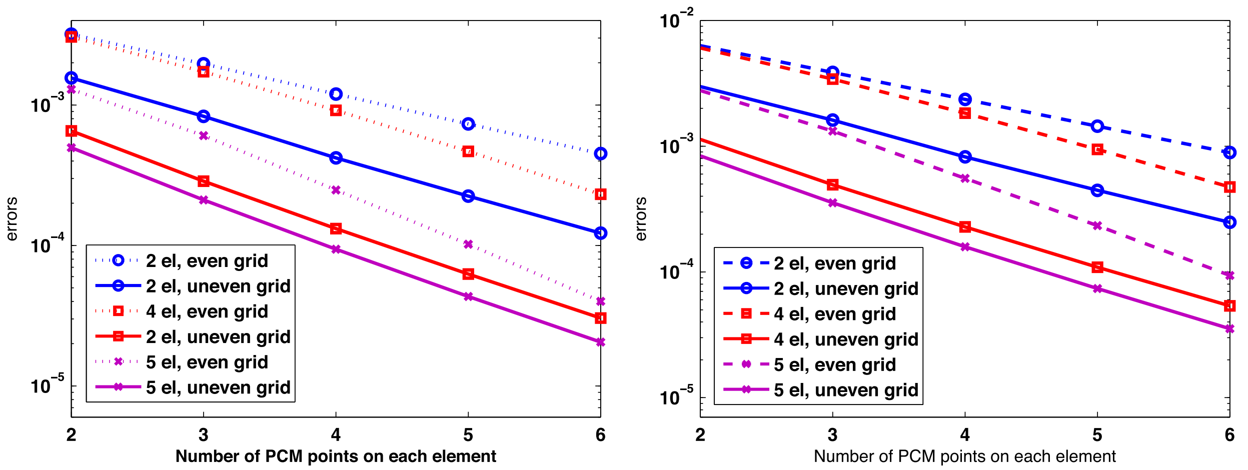

**Fig. 13.** $p$-convergence of ME-PCM on a uniform mesh and an adapted mesh with respect to errors defined in Eqs. (60) and (61) for the KdV equation with $t = 1$, $a = 1$, $x_0 = -5$, $\sigma = 0.2$, and $\xi \sim \text{Pois}(40)$, with 200 Fourier collocation points on the spatial domain $[-30, 30]$. Left: Errors of the mean. Right: Errors of the second moment. Orthogonal polynomials are generated by the Nowak method.

#### 4.2.1. Stochastic excitation given by two discrete RVs

We now use sparse grids to study the KdV equation subject to stochastic excitation:

$$u_t + 6uu_x + u_{xxx} = \sigma_1\xi_1 + \sigma_2\xi_2, \quad x \in \mathbb{R}, \tag{62}$$

with the same initial condition given by Eq. (58), where $\xi_1$ and $\xi_2$ are two i.i.d. random variables associated with a discrete measure.

In Fig. 14, we plot the convergence behavior of sparse grids and tensor product grids for problem (62), where the discrete measure is chosen as $Bino(10, 1/2)$. We see that with respect to the *total number* $r(k)$ collocation points an algebraic-like convergence is obtained with the rate slower than tensor product grid with respect to the total number of PCM collocation points, in lower dimension, consistent with the results in Fig. 7. Specifically the error line for $l2u1$ and $l2u2$ become flat mainly due to the fact that the numerical errors from spatial discretization and temporal integration for the deterministic KdV equation become dominant when $r(k)$ is relatively large.

#### 4.2.2. Stochastic excitation given by a discrete RV and a continuous RV

We still consider Eq. (62), where we only require the independence between $\xi_1$ and $\xi_2$, and assume that $\xi_1 \sim Bino(10, 1/2)$ is a discrete RV and $\xi_2 \sim \mathcal{N}(0, 1)$ is a continuous RV.

In Fig. 15, we plot the convergence behavior of sparse grids and tensor product grids for the KdV equation subject to hybrid (discrete/continuous) random inputs. Similar phenomena are observed as in the previous case where both RVs are discrete. An algebraic-like convergence rate with respect to the total number of grid points is obtained, which is slower than convergence from PCM on tensor product grids in lower dimension, in agreement with the results in Fig. 7. This numerical example demonstrates that the sparse grids approach can be applied to deal with hybrid (discrete/continuous) random inputs when the solution is smooth enough.

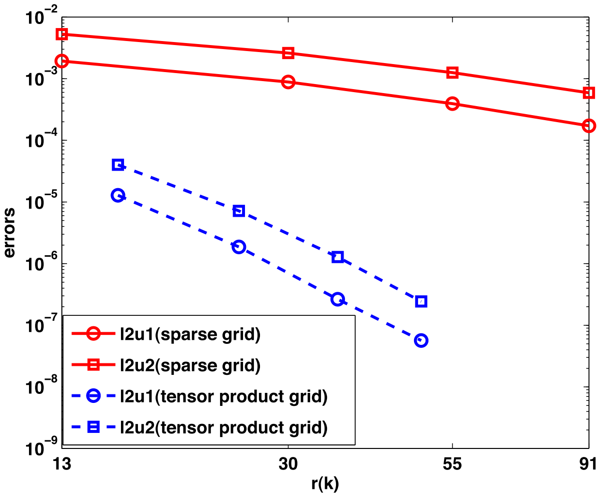

**Fig. 14.** $\xi_1, \xi_2 \sim Bino(10, 1/2)$: convergence of sparse grids and tensor product grids with respect to errors defined in Eqs. (60) and (61) for problem (62), where $t = 1$, $a = 1$, $x_0 = -5$, and $\sigma_1 = \sigma_2 = 0.2$, with 200 Fourier collocation points on the spatial domain $[-30, 30]$. Orthogonal polynomials are generated by the Lanczos method.

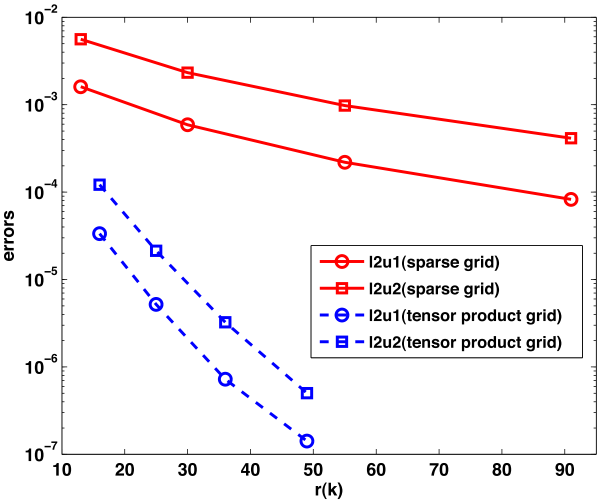

**Fig. 15.** $\xi_1 \sim Bino(10, 1/2)$ and $\xi_2 \sim \mathcal{N}(0, 1)$: convergence of sparse grids and tensor product grids with respect to errors defined in Eqs. (60) and (61) for problem (62), where $t = 1$, $a = 1$, $x_0 = -5$, and $\sigma_1 = \sigma_2 = 0.2$, with 200 Fourier collocation points on the spatial domain $[-30, 30]$. Orthogonal polynomials are generated by Lanczos method.

### 4.2.3. *Stochastic excitation given by eight discrete RVs*

We finally examine a higher-dimensional case:

$$u_t + 6uu_x + u_{xxx} = \sum_{i=1}^{8} \sigma_i \xi_i, \quad x \in \mathbb{R} \tag{63}$$

with the initial condition given in Eq. (58), where the stochastic excitation is subject to eight i.i.d. discrete RVs of the same Binomial distribution $Bino(5, 1/2)$.

We plot the convergence behavior of sparse grids and tensor product grids for problem (63) in Fig. 16. We see that as the number of dimensions increases, the rate of algebraic-like convergence from PCM with sparse grids and tensor product grids both becomes slower. However, with higher dimensional randomness, the sparse grids outperform the tensor product grids in terms of accuracy.

## 5. Summary

In this work we presented a multi-element probabilistic collocation method (ME-PCM) for discrete measures, where we focus on the $h$-convergence with respect to the number of elements and the convergence behavior of the associated sparse

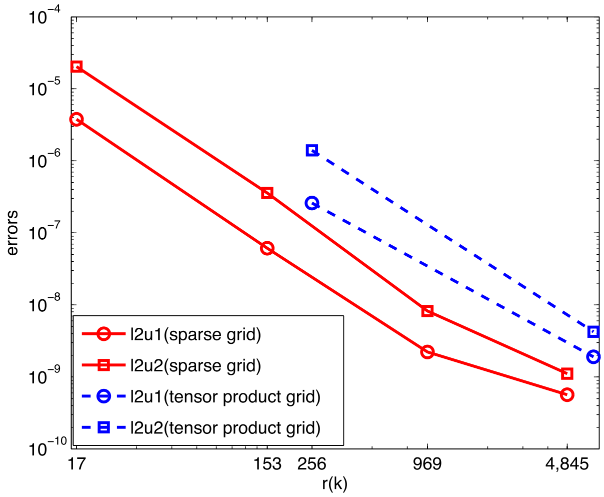

**Fig. 16.** Convergence of sparse grids and tensor product grids with respect to errors defined in Eqs. (60) and (61) for problem (63), where $t = 0.5$, $a = 0.5$, $x_0 = -5$, $\sigma_i = 0.1$ and $\xi_i \sim Bino(5, 1/2)$, $i = 1, 2, ..., 8$, with 300 Fourier collocation points on the spatial domain $[-50, 50]$. Orthogonal polynomials are generated by Lanczos method.

grids based on the one-dimensional Gauss quadrature rule. We first compared five methods of constructing orthogonal polynomials for discrete measures. From numerical experiments in Section 2.7, we conclude that the Stieltjes, Modified Chebyshev, and Lanczos methods generate polynomials that exhibit the best orthogonality among the five methods. For computational cost, we conclude that Stieltjes method has the least computational cost in the case that we have examined.

The relation between $h$-convergence and the degree of exactness given by a certain quadrature rule was discussed for ME-PCM with respect to discrete measures. The $h$-convergence rate $O(N_{es}^{-(m+1)})$ was demonstrated numerically by performing numerical integration of GENZ functions. For moderate-dimensional discrete random inputs, we have demonstrated that non-nested sparse grids based on the Gauss quadrature rule can also be effective. In lower dimensions, PCM on sparse grids is less efficient than on tensor product grids in integration of GENZ functions, however in higher dimensions, sparse grids are more efficient than tensor product grids. In particular, it appears that the convergence behavior is not sensitive to hybrid (discrete/continuous) random inputs.

We have also considered the numerical solution of the reaction equation and the KdV equation subject to stochastic excitation. For the one-dimensional discrete random inputs, we have demonstrated the $h$- and $p$-convergence of ME-PCM. In particular, an adaptive procedure was established using the local variance criterion.

In this work, we focus on the convergence behavior of ME-PCM for arbitrary discrete measures by performing numerical experiments on given random variables. In the future, we would like to generalize and apply our algorithms to study stochastic problems associated with discrete random processes, such as discrete Levy processes.

## Acknowledgements

This research was sponsored by the U.S. Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-12-2-0023 to University of Utah. The second author acknowledges support from NSF Grant DMS-1115632.

## Appendix A. KdV solver

### A.1. Derivation of Eq. (59) with stochastic transformation

The exact solution for the $m$-th moment of solution can be performed by a simple stochastic transformation:

$$W(t;\omega) = \int_0^t \sigma\xi d\tau = \sigma\xi t, \tag{A.1}$$

$$U(x,t;\omega) = u(x,t) - W(t;\omega) = u(x,t) - \sigma\xi t, \tag{A.2}$$

$$X = x - 6\int_0^t W(\tau;\omega)d\tau = x - 3\sigma\xi t^2, \tag{A.3}$$

such that

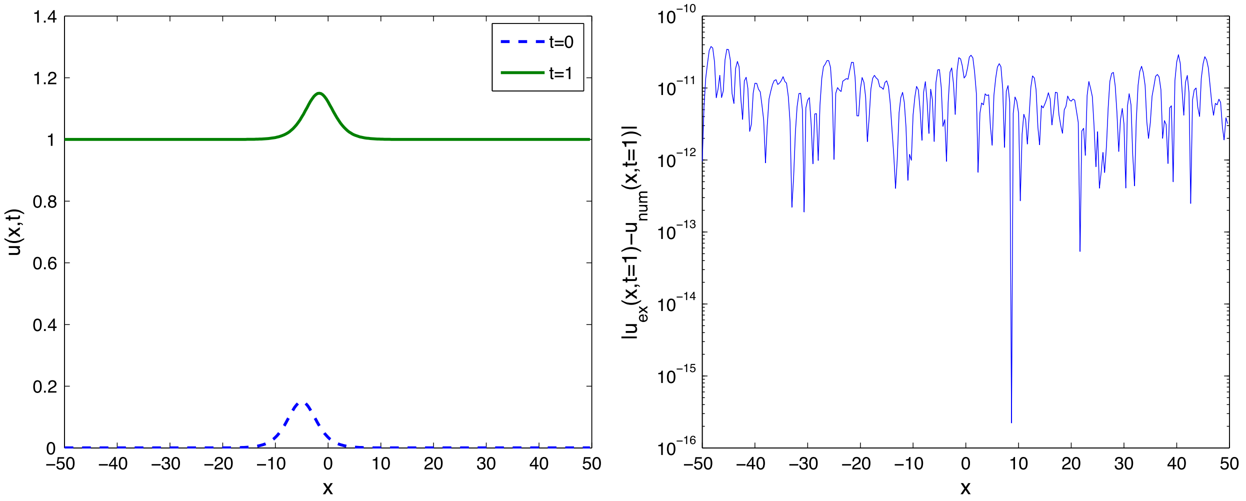

**Fig. 17.** Left: exact solution of the KdV equation (A.6) at time $t = 0, 1$. Right: the pointwise error for the soliton at time $t = 1$.

$$\frac{\partial U}{\partial t} + 6U\frac{\partial U}{\partial X} + \frac{\partial^3 U}{\partial X^3} = 0, \tag{A.4}$$

which has an exact solution

$$U(X,t) = \frac{a}{2}\operatorname{sech}^2\left(\frac{\sqrt{a}}{2}(X - x_0 - at)\right). \tag{A.5}$$

### *A.2. Numerical solver for the deterministic KdV equation*

On each collocation point for the RV $\xi$ we run a deterministic solver of the KdV equation with the Fourier-collocation discretization in physical space, and time splitting scheme like this: we first compute third-order Adams–Bashforth scheme for $6uu_x$ term and then Crank–Nicolson scheme for $u_{xxx}$ in time. We test the accuracy of the deterministic solver using the following problem:

$$u_t + 6uu_x + u_{xxx} = 1 \tag{A.6}$$

with the initial condition:

$$u(x,0) = \frac{a}{2}\operatorname{sech}^2\left(\frac{\sqrt{a}}{2}(x - x_0)\right), \tag{A.7}$$

where $a = 0.3$, $x_0 = -5$, and $t = 1$, and the time step is $1.25 \times 10^{-5}$. For the spatial discretization, we use 300 Fourier collocation points on an interval $[-50, 50]$. The point-wise numerical error is plotted in Fig. 17.

## Appendix B. Local variance criterion

Here we explain the local variance criterion that we adopted for the adaptive ME-PCM. First, let us define the local variance. For any RV $\xi$ with a probability measure $\mu(d\xi)$ on the parametric space $\xi \in \Gamma$, we consider a decomposition of $\Gamma = \bigcup_i^{N_e} B_i$ such that $B_i \cap B_j = \emptyset$, $\forall i \neq j$. On the element $B_i$, we can calculate the local variance $\sigma_i^2$ with respect to the associated conditional measure as $\mu(d\xi)/\int_{B_i}\mu(d\xi)$. We then consider an adaptive decomposition of the parametric space for ME-PCM such that the quantity $\sigma_i^2 \Pr(\xi \in B_i)$ in each element is nearly uniform. Here for the numerical experiments in Fig. 12, we typically minimized the quantity $\sum_{i=1}^{Ne}\sigma_i^2 \Pr(\xi \in B_i)$. In other words, given a discrete measure and number of elements $N_e$, we try all possible $\{B_i, i = 1..N_e\}$ to divide $\Gamma$ until the sum $\sum_{i=1}^{Ne}\sigma_i^2 \Pr(\xi \in B_i)$ is minimized. We found that the size of the element is balanced by the local oscillations and the probability of $\xi \in B_i$ (see more details in [12]).

## References

[1] R.A. Adams, Sobolev Spaces, Academic Press, Boston, MA, 1975.
[2] R. Askey, J. Wilson, Some Basic Hypergeometric Polynomials That Generalize Jacobi Polynomials, Mem. Am. Math. Soc., AMS, Providence, RI, 1985, pp. 319.
[3] K.E. Atkinson, An Introduction to Numerical Analysis, John Wiley and Sons, Inc., 1989.
[4] D. Boley, G.H. Golub, A survey of matrix inverse eigenvalue problems, Inverse Problems 3 (1987) 595–622.

[5] T.S. Chihara, An Introduction to Orthogonal Polynomials, Mathematics and Its Applications, vol. 13, Gordon and Breach Science Publishers, New York, 1978.

[6] J. Dalibard, Y. Castin, Wave-function approach to dissipative processes in quantum optics, Phys. Rev. Lett. 68 (1992) 580–583.

[7] R. Erban, S.J. Chapman, Stochastic modeling of reaction diffusion processes: algorithms for bimolecular reactions, Phys. Biol. 6 (4) (2009) 046001.

[8] L.C. Evans, Partial Differential Equations, AMS/Chelsea, 1998.

[9] J. Favard, Sur les polynomes de Tchebicheff, C. R. Acad. Sci., Paris 200 (1935) 2052–2053 (in French).

[10] H.J. Fischer, On the condition of orthogonal polynomials via modified moments, Z. Anal. Anwend. 15 (1996) 1–18.

[11] H.J. Fischer, On generating orthogonal polynomials for discrete measures, Z. Anal. Anwend. 17 (1998) 183–205.

[12] J. Foo, X. Wan, G.E. Karniadakis, A multi-element probabilistic collocation method for PDEs with parametric uncertainty: error analysis and applications, J. Comput. Phys. 227 (2008) 9572–9595.

[13] G.E. Forsythe, Generation and use of orthogonal polynomials for data-fitting with a digital computer, J. Soc. Ind. Appl. Math. 5 (1957) 74–88.

[14] W. Gautschi, On generating orthogonal polynomials, SIAM J. Sci. Stat. Comput. 3 (3) (1982) 289–317.

[15] A. Genz, A package for testing multiple integration subroutines, numerical integration: recent developments, Softw. Appl. (1987) 337–340.

[16] G.H. Golub, C. Van Loan, Matrix Computations, Johns Hopkins Univ. Press, 1983.

[17] G.H. Golub, J.H. Welsch, Calculation of Gauss quadrature rules, Math. Comput. 23 (106) (1969) 221–230.

[18] G.E. Karniadakis, S. Sherwin, Spectral/hp Element Methods for Computational Fluid Dynamics, second edition, Oxford University Press, 2005, pp. 597–598.

[19] G. Lin, L. Grinberg, G.E. Karniadakis, Numerical studies of the stochastic Korteweg–de Vries equation, J. Comput. Phys. 213 (2) (2006) 676–703.

[20] E. Novak, K. Ritter, High dimensional integration of smooth functions over cubes, Numer. Math. 75 (1996) 79–97.

[21] E. Novak, K. Ritter, Simple cubature formulas with high polynomial exactness, Constr. Approx. 15 (1999) 49–522.

[22] S. Oladyshkin, W. Nowak, Data-driven uncertainty quantification using the arbitrary polynomial chaos expansion, Reliab. Eng. Syst. Saf. 106 (2012) 179–190.

[23] S. Smolyak, Quadrature and interpolation formulas for tensor products of certain classes of functions, Sov. Math. Dokl. 4 (1963) 240–243.

[24] W. Trench, An algorithm for the inversion of finite Toeplitz matrices, SIAM J. Control Optim. 12 (1964) 512–522.

[25] E.I. Tzenova, Ivo J.B.F. Adan, V.G. Kulkarni, Fluid models with jumps, Stoch. Models 21 (1) (2005) 37–55.

[26] X. Wan, G.E. Karniadakis, An adaptive multi-element generalized polynomial chaos method for stochastic differential equations, J. Comput. Phys. 209 (2) (2005) 617–642.

[27] D. Xiu, J.S. Hesthaven, High-order collocation methods for differential equations with random inputs, SIAM J. Sci. Comput. 27 (3) (2005) 1118–1139.

[28] D. Xiu, G.E. Karniadakis, The Wiener–Askey polynomial chaos for stochastic differential equations, SIAM J. Sci. Comput. 24 (2002) 619–644.

[29] G. Yan, F.B. Hanson, Option pricing for a stochastic-volatility jump-diffusion model with log uniform jump-amplitudes, in: Proceedings of the 2006 American Control Conference, 2006, pp. 2989–2994.

[30] T.J. Ypma, Historical development of the Newton–Raphson method, SIAM Rev. 37 (4) (1995) 531–551.