International Journal for Uncertainty Quantification, 3 (1): 1–23 (2013)

POLYNOMIAL CHAOS FOR SEMIEXPLICIT DIFFERENTIAL ALGEBRAIC EQUATIONS OF INDEX 1

Roland Pulch

Bergische Universität Wuppertal, Fachbereich Mathematik und Naturwissenschaften, Lehrstuhl für Angewandte Mathematik und Numerische Mathematik, Gaußstr. 20, D-42119 Wuppertal, Germany; E-mail: pulch@math.uni-wuppertal.de

Original Manuscript Submitted: 03/28/2011; Final Draft Received: 09/08/2011

Mathematical modeling of technical applications often yields systems of differential algebraic equations. Uncertainties of physical parameters can be considered by the introduction of random variables. A corresponding uncertainty quantification requires one to solve the stochastic model. We focus on semiexplicit systems of nonlinear differential algebraic equations with index 1. The stochastic model is solved using the expansion of the generalised polynomial chaos. We investigate both the stochastic collocation technique and the stochastic Galerkin method to determine the unknown coefficient functions. In particular, we analyze the index of the larger coupled systems, which result from the stochastic Galerkin method. Numerical simulations of test examples are presented, where the two approaches are compared with respect to their efficiency.

KEY WORDS: *differential algebraic equation, index, polynomial chaos, stochastic collocation method, stochastic Galerkin method, uncertainty quantification*

1. INTRODUCTION

In technical applications, mathematical modeling of dynamical systems often results in differential algebraic equations (DAEs), i.e., a mixture of ordinary differential equations (ODEs) and algebraic equations. For example, network approaches yield large systems of DAEs corresponding to mechanical multibody dynamics or electric circuits [1–3]. Systems of DAEs exhibit qualitatively different properties than systems of ODEs. The index, which represents an integer (number), indicates the level of the differences between a particular system of DAEs and a general system of ODEs. Several concepts for the definition of an index exist. Numerical methods for initial value problems of ODEs are transferred into integrators for DAEs, where attention must be paid also in dependence on the index of the systems [4, 5].

We assume that uncertainties are inherent in some physical parameters of the dynamical system. Corresponding parameters are replaced by random variables to achieve an uncertainty quantification. The random-dependent system of DAEs can be resolved by a quasi Monte Carlo simulation, for example. Alternatively, we consider techniques based on the expansions of the generalized polynomial chaos (gPC) [6–8]. The unknown coefficient functions can be determined either via a stochastic collocation method or the stochastic Galerkin approach [9, 10]. Thereby, the stochastic Galerkin technique yields a larger coupled system of DAEs satisfied by an approximation of the coefficient functions.

The gPC expansions have already been applied for the simulation of systems of DAEs with random parameters in [11–13], where the focus is on periodic boundary value problems. In the case of linear systems of DAEs, the index of the coupled systems of the stochastic Galerkin method is analyzed in [14]. All index concepts are equivalent for linear DAEs. In this paper, we consider semiexplicit systems of nonlinear DAEs with a differential index 1. For

2152–5080/13/$35.00

semiexplicit DAEs, the differential index is 1 if and only if the perturbation index is 1 [15]. It is obvious that the index of the DAEs coincides within stochastic collocation methods. We examine the index of the larger coupled system of DAEs, which is obtained by the stochastic Galerkin technique.

The approach of the stochastic collocation and the stochastic Galerkin method are compared in this paper. On the one hand, the properties of the involved systems of DAEs are analyzed using the corresponding index. On the other hand, numerical simulations of initial value problems are performed to investigate the efficiency of each method, which is done by a comparison of both accuracy and computational effort.

The paper is organized as follows. We introduce the random-dependent systems of DAEs in Section 2. The stochastic collocation techniques and the stochastic Galerkin method are outlined. We analyze the index of the coupled systems from the Galerkin approach in Section 3. Numerical simulations of four test examples are illustrated in Section 4, where the efficiency of both gPC techniques is compared.

2. STOCHASTIC MODELING

In this section, we define the stochastic model and apply the expansions of the generalized polynomial chaos for the corresponding solutions.

2.1 Problem Definition

We consider dynamical systems of the form

$$A(\mathbf{p})\dot{\mathbf{x}}(t,\mathbf{p}) = \mathbf{f}[t,\mathbf{x}(t,\mathbf{p}),\mathbf{p}], \tag{1}$$

where parameters $\mathbf{p} = (p_1,\ldots,p_Q)$ with $\mathbf{p} \in \Pi \subseteq \mathbb{R}^Q$ are involved. The solution $\mathbf{x} : [t_0,t_1] \times \Pi \to \mathbb{R}^N$ depends on time as well as the parameters. In the case of a regular mass matrix $\{\det[A(\mathbf{p})] \neq 0\}$, the system (1) represents implicit ODEs. In case of a singular mass matrix $\{\det[A(\mathbf{p})] = 0\}$, we obtain a system (1) of DAEs. We consider initial value problems

$$\mathbf{x}(t_0,\mathbf{p}) = \mathbf{x}_0(\mathbf{p}), \tag{2}$$

where the initial values are allowed to depend on the parameters.

We assume that the parameters include some uncertainties. Consequently, we substitute the parameters by random variables

$$\mathbf{p} : \Omega \to \Pi, \quad \mathbf{p}(\omega) = [p_1(\omega),\ldots,p_Q(\omega)]$$

defined on some probability space $(\Omega,\mathcal{A},\mu)$. We apply independent distributions in this modeling, where a corresponding probability density function $\rho : \Pi \to \mathbb{R}$ is available. A random variable p can describe the perturbation of a physical parameter r, i.e.,

$$r(\omega) := \lambda p(\omega) + r_0 \tag{3}$$

with constants $\lambda, r_0 \in \mathbb{R}$. Thus, a standardized variable with $\langle p\rangle = 0$ and $\langle p^2\rangle = 1$ can be used in the modeling, whereas the information on λ, r_0 is included in the system (1). The solution $\mathbf{x}(t,\mathbf{p})$ of the dynamical system (1) becomes a random process, depending on time as well as the random parameters. We are interested in the key data of this random process, such as the expected value and the standard deviation, for example. More sophisticated data may also be resolved.

We define the function spaces as

$$L^k(\Pi,\rho) := \left\{ f : \Pi \to \mathbb{R} : \int_\Pi |f(\mathbf{p})|^k \rho(\mathbf{p})\, \mathrm{d}\mathbf{p} < \infty \right\}$$

for each integer k. Given a function $f \in L^1(\Pi,\rho)$, we apply the notation

$$\langle f(\mathbf{p})\rangle := \int_\Pi f(\mathbf{p})\rho(\mathbf{p})\, \mathrm{d}\mathbf{p} \tag{4}$$

for the corresponding expected value. For two functions $f, g \in L^2(\Pi, \rho)$, the expected value (4) implies the inner product

$$\langle f(\mathbf{p})g(\mathbf{p})\rangle = \int_\Pi f(\mathbf{p})g(\mathbf{p})\rho(\mathbf{p})\ \mathrm{d}\mathbf{p}. \tag{5}$$

We employ this notation also to vector- and matrix-valued functions by each component separately.

In this paper, we consider semiexplicit systems of DAEs, i.e., the dynamical system (1) becomes

$$\begin{aligned} \mathbf{y}'(t,\mathbf{p}) &= \mathbf{f}[t, \mathbf{y}(t,\mathbf{p}), \mathbf{z}(t,\mathbf{p}), \mathbf{p}] \\ \mathbf{0} &= \mathbf{g}[t, \mathbf{y}(t,\mathbf{p}), \mathbf{z}(t,\mathbf{p}), \mathbf{p}] \end{aligned} \tag{6}$$

with the differential variables $\mathbf{y} : [t_0, t_1] \times \Pi \to \mathbb{R}^{N_y}$ and the algebraic variables $\mathbf{z} : [t_0, t_1] \times \Pi \to \mathbb{R}^{N_z}$. The right-hand sides exhibit the dimensions $\mathbf{f} \in \mathbb{R}^{N_y}$ and $\mathbf{g} \in \mathbb{R}^{N_z}$. We assume that the right-hand sides are continuous or sufficiently smooth if required.

A system of DAEs features different properties than a system of ODEs. The level of these differences is characterized by the index of the system of DAEs, where several index concepts exist [5]. We focus on semiexplicit systems (6) of differential index 1 and perturbation index 1. In case of semiexplicit DAEs, the differential index is 1 if and only if the perturbation index is 1. An equivalent condition is that the Jacobian matrix $\partial\mathbf{g}/\partial\mathbf{z} \in \mathbb{R}^{N_z \times N_z}$ is regular, i.e.,

$$\det\left(\frac{\partial \mathbf{g}}{\partial \mathbf{z}}\right) \neq 0 \tag{7}$$

for all involved solutions and parameters in each time point $t \in [t_0, t_1]$. If (7) holds, then DAEs of the form (6) are also called Hessenberg DAEs of index 1. Dynamical systems (1) with a constant mass matrix $[A(\mathbf{p}) \equiv A_0]$ can be transformed directly into an equivalent semiexplicit system (6) of the same dimension ($N_y + N_z = N$) and the same index. Moreover, each system of the form (1) can be converted into a corresponding semiexplicit system (6) with $N_y = N_z = N$ using $\mathbf{y} := \mathbf{x}$ and $\mathbf{z} := \mathbf{y}'$, where a higher index appears in general.

We specify an initial value problem via

$$\mathbf{y}(t_0, \mathbf{p}) = \mathbf{y}_0(\mathbf{p}), \quad \mathbf{z}(t_0, \mathbf{p}) = \mathbf{z}_0(\mathbf{p}) \tag{8}$$

with predetermined parameter-dependent functions $\mathbf{y}_0, \mathbf{z}_0$. In the case of the semiexplicit systems (6) of index 1, the initial values (8) must satisfy the consistency condition

$$\mathbf{g}(t_0, \mathbf{y}_0(\mathbf{p}), \mathbf{z}_0(\mathbf{p}), \mathbf{p}) = \mathbf{0} \tag{9}$$

for each $\mathbf{p} \in \Pi$. Hence, the initial values $\mathbf{z}_0$ follow from the choice of the initial values $\mathbf{y}_0$ by the implicit function theorem. Even if the initial values $\mathbf{y}_0$ are independent of the parameters, the initial values $\mathbf{z}_0$ depend on the parameters if the function $\mathbf{g}$ does. Thus, we consider parameter-dependent initial values (8) in general. For semiexplicit DAEs of higher index, hidden consistency conditions exist in addition to the algebraic constraints (9).

2.2 Generalized Polynomial Chaos

Considering random parameters, the stochastic model [(6) and (8)] can be solved by a (quasi) Monte Carlo simulation, for example. Alternatively, we consider spectral methods based on the polynomial chaos [7, 8]. We assume finite second moments of the components of the differential and algebraic variables corresponding to a solution of the stochastic model [(6) and (8)]. It follows that the expansions:

$$\mathbf{y}[t, \mathbf{p}(\omega)] = \sum_{i=0}^{\infty} \mathbf{v}_i(t)\Phi_i[\mathbf{p}(\omega)], \quad \mathbf{z}[t, \mathbf{p}(\omega)] = \sum_{i=0}^{\infty} \mathbf{w}_i(t)\Phi_i[\mathbf{p}(\omega)] \tag{10}$$

converge with respect to the norm of $L^2(\Pi, \rho)$ for each $t \in [t_0, t_1]$. The series include orthogonal basis polynomials $\Phi_i : \Pi \to \mathbb{R}$. Thus, let $\langle \Phi_i \Phi_j \rangle = \delta_{ij}$ with the Kronecker delta symbol. The basis polynomials follow from the

probability distributions of the random parameters [16]. Thereby, the multivariate basis polynomials are the products of corresponding univariate basis polynomials. If all random parameters exhibit Gaussian distributions, then the traditional homogeneous polynomial chaos appears. In the case of non-Gaussian random parameters, we obtain the generalized polynomial chaos (gPC).

The coefficient functions $\mathbf{v}_i : [t_0, t_1] \to \mathbb{R}^{N_y}$ and $\mathbf{w}_i : [t_0, t_1] \to \mathbb{R}^{N_z}$ are unknown a priori. These time-dependent functions satisfy the equations

$$\mathbf{v}_i(t) = \langle \mathbf{y}(t, \mathbf{p}) \Phi_i(\mathbf{p}) \rangle, \quad \mathbf{w}_i(t) = \langle \mathbf{z}(t, \mathbf{p}) \Phi_i(\mathbf{p}) \rangle. \tag{11}$$

Assuming $\Phi_0 \equiv 1$, it follows $\mathbf{v}_0 = \langle \mathbf{y} \rangle$ and $\mathbf{w}_0 = \langle \mathbf{z} \rangle$.

In practice, the gPC expansions (10) have to be truncated. The resulting finite approximations read

$$\mathbf{y}^{(M)}(t, \mathbf{p}) := \sum_{i=0}^{M} \mathbf{v}_i(t) \Phi_i(\mathbf{p}), \quad \mathbf{z}^{(M)}(t, \mathbf{p}) := \sum_{i=0}^{M} \mathbf{w}_i(t) \Phi_i(\mathbf{p}) \tag{12}$$

for some integer M. Often all basis polynomials up to a certain degree are chosen in the finite sums. We can also apply different bases or different orders M in the differential part and the algebraic part, respectively.

The coefficients in Eq. (12) yield approximations of the expected value and the variance of the random process. Nevertheless, more sophisticated quantities are also reproduced approximatively by this approach. For example, a truncated series (12) represents a surrogate model, which can be used to compute failure probabilities (cf. [12, 17]).

2.3 Stochastic Collocation Techniques

We want to determine approximations of the coefficient functions involved in the truncated gPC expansion (12). Because of the property (11), the unknown coefficient functions represent evaluations of probabilistic integrals. Thus, we achieve an approximation of the coefficient functions by a quadrature formula. We choose grid points $\mathbf{p}^{(1)}, \ldots, \mathbf{p}^{(K)} \in \Pi$ in the domain of the parameters. It follows the approximations

$$\mathbf{v}_i(t) \doteq \sum_{k=1}^{K} \omega_k \Phi_i(\mathbf{p}^{(k)}) \mathbf{y}(t, \mathbf{p}^{(k)}), \quad \mathbf{w}_i(t) \doteq \sum_{k=1}^{K} \omega_k \Phi_i(\mathbf{p}^{(k)}) \mathbf{z}(t, \mathbf{p}^{(k)}) \tag{13}$$

with weights $\omega_1, \ldots, \omega_K \in \mathbb{R}$.

For small numbers Q of parameters, a multivariate Gaussian quadrature can be employed straightforward, because the grids are tensor products of the nodes of the corresponding univariate Gaussian quadratures. For medium-sized Q, sparse grids should be preferred. In the case of large numbers Q of parameters, Monte Carlo simulations with pseudo random numbers or quasi Monte Carlo methods are applied. Examples for two random parameters ($Q = 2$) with independent standardized Gaussian distributions are shown in Fig. 1.

Each technique of the form (13) is called a stochastic collocation [9, 10, 18]. The nodes $\mathbf{p}^{(1)}, \ldots, \mathbf{p}^{(K)}$ can be seen as collocation points. In each method of this type, we have to solve K initial value problems (6) and (8) of the original systems of DAEs. Thereby, the numerical methods constructed for the deterministic initial value problems of the DAEs are applied directly. Hence, the stochastic collocation approach is also called the nonintrusive method.

2.4 Stochastic Galerkin Method

Inserting the truncated gPC expansions (12) into the semiexplicit system (6) yields the residuals

$$\begin{aligned} \mathbf{r}_y(t, \mathbf{p}) &:= \mathbf{y}^{(M)\prime}(t, \mathbf{p}) - \mathbf{f}[t, \mathbf{y}^{(M)}(t, \mathbf{p}), \mathbf{z}^{(M)}(t, \mathbf{p}), \mathbf{p}], \\ \mathbf{r}_z(t, \mathbf{p}) &:= \mathbf{g}[t, \mathbf{y}^{(M)}(t, \mathbf{p}), \mathbf{z}^{(M)}(t, \mathbf{p}), \mathbf{p}]. \end{aligned} \tag{14}$$

We want to determine the coefficient functions such that the residuals become small in some sense. The Galerkin method requires the residuals to be orthogonal with respect to the space of the applied basis polynomials, i.e.,

$$\langle \mathbf{r}_{y,z}(t, \mathbf{p}) \Phi_l(\mathbf{p}) \rangle = \mathbf{0} \tag{15}$$

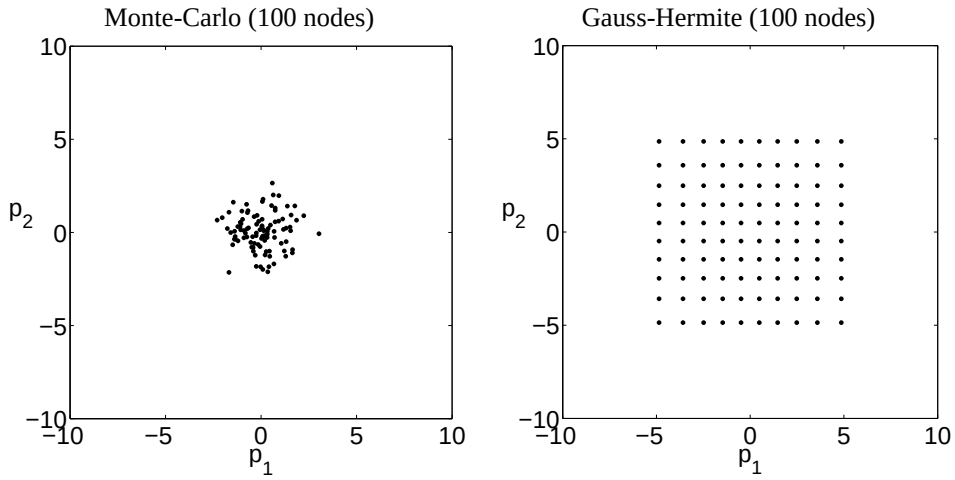

FIG. 1: Examples for grids in stochastic collocation techniques with two independent Gaussian random variables.

for $l = 0, 1, \ldots, M$ and each $t \in [t_0, t_1]$. Inserting the residuals (14) into the inner products (15), basic calculations lead to a larger coupled system, where the unknowns represent an approximation of the coefficient functions.

Definition 1. The coupled system of the stochastic Galerkin method corresponding to the semiexplicit systems (6) reads

$$\mathbf{v}_l'(t) = \left\langle \Phi_l(\mathbf{p}) \cdot \mathbf{f}\left(t, \sum_{i=0}^{M} \mathbf{v}_i(t)\Phi_i(\mathbf{p}), \sum_{i=0}^{M} \mathbf{w}_i(t)\Phi_i(\mathbf{p}), \mathbf{p}\right)\right\rangle \tag{16}$$

$$\mathbf{0} = \left\langle \Phi_l(\mathbf{p}) \cdot \mathbf{g}\left(t, \sum_{i=0}^{M} \mathbf{v}_i(t)\Phi_i(\mathbf{p}), \sum_{i=0}^{M} \mathbf{w}_i(t)\Phi_i(\mathbf{p}), \mathbf{p}\right)\right\rangle \tag{17}$$

for $l = 0, 1, \ldots, M$.

Although an exact solution of (16) and (17) is not identical to the exact coefficients in (10), we apply the same symbols for convenience. The coupled system of (16) and (17) represents a semiexplicit system of DAEs again due to the orthogonality of the basis polynomials.

To obtain initial values for the coupled system of (16) and (17), the original initial values (8) can be expanded in the gPC. It follows:

$$\mathbf{v}_l(t_0) = \langle \mathbf{y}_0(\mathbf{p})\Phi_l(\mathbf{p})\rangle, \quad \mathbf{w}_l(t_0) = \langle \mathbf{z}_0(\mathbf{p})\Phi_l(\mathbf{p})\rangle \tag{18}$$

for $l = 0, 1, \ldots, M$. However, because approximations of the type (12) are used, it holds $\mathbf{g} \neq \mathbf{0}$ at the solution of the coupled system in general. Hence, the algebraic constraints (17) are not satisfied exactly [i.e., the straightforward choice (18) of the initial values is inconsistent]. Alternatively, just the differential variables $\mathbf{v}_l(t_0)$ are computed via (18). We determine the algebraic variables $\mathbf{w}_l(t_0)$ by solving the $(M+1)N_z$ algebraic equations (17). Because of this construction, the initial values are consistent provided that the semiexplicit system of (16) and (17) also exhibits the index 1.

Solving the coupled system of (16) and (17) is also called the intrusive method. Given the original semiexplicit systems (6) of some index $I \geq 1$ for all parameters, we can solve the coupled system of (16) and (17) using the same numerical methods provided that the coupled system inherits the index I. If the coupled system exhibits a larger index, then disadvantages appear within the numerical simulation in comparison to the solution of the original systems.

3. INDEX ANALYSIS

Let $S := \text{supp}(\rho) = \overline{\{\mathbf{p} : \rho(\mathbf{p}) \neq 0\}} \subseteq \Pi \subseteq \mathbb{R}^Q$. We assume that a unique solution of the initial value problems (6) and (8) exists for each $\mathbf{p} \in S$. Consequently, the Jacobian matrix $\partial\mathbf{g}/\partial\mathbf{z}$ depends just on the choice of $\mathbf{p}$ for fixed $t \in [t_0, t_1]$. The analysis of the index can be done for each t separately. The index-1 property is achieved for all $t \in [t_0, t_1]$ provided that the criteria hold for each $t \in [t_0, t_1]$. Hence, we consider just a fixed $t \in [t_0, t_1]$ in the following.

3.1 Jacobian Matrix in Coupled System

For the algebraic part (17) of the coupled system from the stochastic Galerkin method, we use the abbreviation

$$\mathbf{G}_l := \left\langle \Phi_l(\mathbf{p}) \cdot \mathbf{g}\left(t, \sum_{i=0}^{M} \mathbf{v}_i(t)\Phi_i(\mathbf{p}), \sum_{i=0}^{M} \mathbf{w}_i(t)\Phi_i(\mathbf{p}), \mathbf{p}\right)\right\rangle = \mathbf{0}$$

for $l = 0, 1, \ldots, M$. Consequently, the Jacobian matrix, which determines the index-1 property of the coupled system, reads

$$\mathcal{G} := \left(\frac{\partial \mathbf{G}_l}{\partial \mathbf{w}_k}\right)_{l,k} \in \mathbb{R}^{(M+1)N_z \times (M+1)N_z}. \tag{19}$$

This matrix consists of the minors

$$\frac{\partial \mathbf{G}_l}{\partial \mathbf{w}_k} = \left\langle \Phi_l(\mathbf{p})\Phi_k(\mathbf{p}) \frac{\partial \mathbf{g}}{\partial \mathbf{z}}\left(t, \sum_{i=0}^{M} \mathbf{v}_i(t)\Phi_i(\mathbf{p}), \sum_{i=0}^{M} \mathbf{w}_i(t)\Phi_i(\mathbf{p}), \mathbf{p}\right)\right\rangle \tag{20}$$

for $l, k = 0, 1, \ldots, M$. The derivation of the formula (20) is based on the assumption that the differentiation and the probabilistic integration can be interchanged. Thus, we assume that the entries of the Jacobian matrix $\partial\mathbf{g}/\partial\mathbf{z}$ are continuous in the closed domain S. Let the probability density function ρ also be continuous in S. If S is bounded, then it follows that the differentiation and the integration can be interchanged. In the case of unbounded S, further integrability conditions are required to guarantee that Eq. (20) holds.

The condition $\det(\mathcal{G}) \neq 0$ is equivalent to the index-1 property of the coupled system of (16) and (17). In contrast, the coupled system exhibits an index at least 2 for $\det(\mathcal{G}) = 0$. An increase of the index represents a crucial drawback. Initial value problems of index 1 are well posed with respect to the dependence on perturbed data. Initial value problems of index larger than 1 are, strictly speaking, ill-posed with respect to the dependence on perturbed data, because the time derivative of the perturbation enters the problem [5].

Now we ask if the coupled system of (16) and (17) is of index 1 provided that the original systems (6) exhibit the index 1. As a first minor result, we obtain the following conclusion.

Theorem 1. *If the matrix $\partial\mathbf{g}/\partial\mathbf{z}$ does not depend on $\mathbf{y}, \mathbf{z}, \mathbf{p}$ and the semiexplicit system (6) is of index 1, then the coupled system of (16) and (17) inherits the index 1.*

Proof. The differential variables and the algebraic variables depend on the parameters. Due to the assumptions, the matrix $\partial\mathbf{g}/\partial\mathbf{z}$ does not exhibit a dependence on the parameters at all. Now the matrix $\partial\mathbf{g}/\partial\mathbf{z}$ depends only on time. It follows

$$\left\langle \Phi_l(\mathbf{p})\Phi_k(\mathbf{p}) \frac{\partial \mathbf{g}}{\partial \mathbf{z}} \right\rangle = \langle \Phi_l(\mathbf{p})\Phi_k(\mathbf{p}) \rangle \frac{\partial \mathbf{g}}{\partial \mathbf{z}}.$$

Since the system of basis polynomials is orthonormal, we obtain

$$\mathcal{G} = I_{M+1} \otimes \frac{\partial \mathbf{g}}{\partial \mathbf{z}}$$

with the identity matrix I_{M+1} and the Kronecker product of matrices. Hence the matrix $\mathcal{G}$ is regular if and only if the matrix $\partial\mathbf{g}/\partial\mathbf{z}$ is regular. □

Even if the algebraic constraints in Eq. (6) described by $\mathbf{g}$ include $\mathbf{y}, \mathbf{z}, \mathbf{p}$, the matrix $\partial\mathbf{g}/\partial\mathbf{z}$ is independent of $\mathbf{y}, \mathbf{z}, \mathbf{p}$ in some cases (see the example in Section 4.3). However, the assumption of Theorem 1 is often not given in practice.

In the following examinations, we will assume one of the following two properties.

Condition 1. The matrix $\partial\mathbf{g}/\partial\mathbf{z}$ corresponding to the semiexplicit system (6) is regular for all $\mathbf{p} \in S = \text{supp}(\rho)$.

Condition 2. The matrix $\partial\mathbf{g}/\partial\mathbf{z}$ corresponding to the semiexplicit system (6) is regular for almost all $\mathbf{p}$ (i.e., all $\mathbf{p} \in R \subseteq S$ for some measurable set R where $S \backslash R$ has probability zero).

Conditions 1 and 2 represent the index-1 property of the original systems for all parameters and almost all parameters, respectively. In a stochastic collocation method, Condition 1 guarantees that all involved semiexplicit systems (6) are of index 1. However, systems (6) of larger index may appear if just the Condition 2 is valid.

3.2 Counterexample

Neither Conditions 1 nor 2 is sufficient for the regularity of the matrix $\mathcal{G}$ of the coupled system of (16) and (17). A corresponding counterexample exists already in the case $N_z = 1$, and a single parameter $Q = 1$. This counterexample can be embedded straightforward into examples with $N_z > 1$ and/or $Q > 1$.

We define the algebraic part

$$g(t, \mathbf{y}, z, p) := p \cdot z + u(\mathbf{y})$$

with an arbitrary function $u : \mathbb{R}^{N_y} \to \mathbb{R}$. Since it holds

$$\frac{\partial g}{\partial z} = p,$$

a corresponding system (6) is of index 1 for $p \neq 0$ and of index at least 2 for $p = 0$. The matrix of the corresponding coupled system of (16) and (17) consists of the entries

$$\mathcal{G} = [\langle p\Phi_l(p)\Phi_k(p)\rangle]_{l,k=0,1,\ldots,M}.$$

Because of the orthogonality of the basis polynomials, the matrix $\mathcal{G}$ is tridiagonal. We choose a symmetric probability density $\rho(p)$ around the critical point $p = 0$. Condition 2 is always satisfied in this case. For example, we can apply a Gaussian distribution with mean value $p = 0$. However, the diagonal entries of $\mathcal{G}$ become zero. It follows that the matrix $\mathcal{G}$ has the property

$$\det(\mathcal{G}) \begin{cases} = 0 & \text{for even } M, \\ \neq 0 & \text{for odd } M. \end{cases}$$

We recognize that there is no integer M_0 such that the coupled system is of index 1 for all $M > M_0$. Hence, an improvement of the accuracy of the gPC by increasing M does not omit this behavior.

Moreover, we can choose a uniform distribution corresponding to the domain $S = [-b, -a] \cup [a, b]$ for some $0 < a < b$. It follows a symmetric probability density function again. Now the stronger Condition 1 is satisfied. Yet the matrix $\mathcal{G}$ is singular for even M again. The critical point $\partial g/\partial z = p = 0$ is not within the support of the probability density function. However, this critical point is situated in the convex hull of the support. We will reconsider this quality in Section 3.5.

The above counterexample also indicates that problems with respect to the index may appear in a stochastic collocation if Condition 2 is satisfied but not Condition 1. We suppose a Gaussian distribution with mean value $p = 0$. If we apply a Gauss-Hermite quadrature in the stochastic collocation method, then the critical point $p = 0$ is a node of this quadrature scheme in case of an odd number of nodes. It follows that a semiexplicit system (6) of index at least 2 must be solved in the stochastic collocation, although almost all systems exhibit the index 1.

3.3 Limit Case of Small Variances

In this subsection, we apply a slightly different notation for the semiexplicit system (6) using physical parameters in the form (3). It follows the system:

$$\begin{aligned} \mathbf{y}'_\lambda(t,\mathbf{p}) &= \mathbf{f}[t,\mathbf{y}_\lambda(t,\mathbf{p}),\mathbf{z}_\lambda(t,\mathbf{p}),\lambda\mathbf{p}+\mathbf{r}_0] \\ \mathbf{0} &= \mathbf{g}[t,\mathbf{y}_\lambda(t,\mathbf{p}),\mathbf{z}_\lambda(t,\mathbf{p}),\lambda\mathbf{p}+\mathbf{r}_0] \end{aligned} \tag{21}$$

depending on some $\lambda \in \mathbb{R}$. Similar modification have to be done within the initial values (8). For $\lambda = 0$, the system (21) becomes deterministic and involves the constant reference parameters $\mathbf{r}_0$. We assume that the index of the system of DAEs is 1 in case of the reference parameters. For $\lambda \neq 0$, a random perturbation appears. Using fixed random variables $\mathbf{p}$, the variance of the physical input parameters vanishes in the limit case $\lambda \to 0$.

The stochastic Galerkin method yields a corresponding coupled system with a matrix $\mathcal{G}_\lambda$ now, cf. (19). We obtain the following result in the limit of small variances, where the Kronecker product of matrices and the identity matrix I_{M+1} is involved. A subscript zero refers to the case $\lambda = 0$ and not to the initial values at t_0 now.

Theorem 2. *Assume that all functions within the matrix $\partial\mathbf{g}/\partial\mathbf{z}$ are in $L^2(\Pi,\rho)$ and Lipschitz-continuous with respect to $\mathbf{y},\mathbf{z}$ as well as the parameters. The matrix $\mathcal{G}_\lambda$ in the stochastic Galerkin method for the system (21) satisfies*

$$\lim_{\lambda\to 0} \mathcal{G}_\lambda = I_{M+1} \otimes \frac{\partial \mathbf{g}}{\partial \mathbf{z}}[t,\mathbf{y}_0(t,\mathbf{p}),\mathbf{z}_0(t,\mathbf{p}),\mathbf{r}_0] \tag{22}$$

provided that it holds

$$\lim_{\lambda\to 0} \left\langle \left\| \mathbf{y}_\lambda^{(M)}(t,\mathbf{p}) - \mathbf{y}_0(t,\mathbf{p}) \right\|^2 \right\rangle = 0, \quad \lim_{\lambda\to 0} \left\langle \left\| \mathbf{z}_\lambda^{(M)}(t,\mathbf{p}) - \mathbf{z}_0(t,\mathbf{p}) \right\|^2 \right\rangle = 0 \tag{23}$$

in an arbitrary vector norm $\|\cdot\|$.

Proof. In the following, we apply the abbreviation

$$F(t) := \frac{\partial \mathbf{g}}{\partial \mathbf{z}}(t,\mathbf{y}_0(t,\mathbf{p}),\mathbf{z}_0(t,\mathbf{p}),\mathbf{r}_0).$$

We rearrange the minors (20) of the matrix $\mathcal{G}_\lambda$ to

$$\frac{\partial \mathbf{G}_l}{\partial \mathbf{w}_k} = \langle \Phi_l(\mathbf{p})\Phi_k(\mathbf{p})F(t)\rangle + \left\langle \Phi_l(\mathbf{p})\Phi_k(\mathbf{p}) \left\{ \frac{\partial \mathbf{g}}{\partial \mathbf{z}}[t,\mathbf{y}_\lambda^{(M)}(t,\mathbf{p}),\mathbf{z}_\lambda^{(M)}(t,\mathbf{p}),\lambda\mathbf{p}+\mathbf{r}_0] - F(t) \right\} \right\rangle.$$

Using $\langle \Phi_l\Phi_k\rangle = \delta_{lk}$, we write the complete matrix in the form

$$\mathcal{G}(t) = I_{M+1} \otimes F(t) + \mathcal{R}(t).$$

Let $D = (d_{ij}) \in \mathbb{R}^{N_z \times N_z}$ be the matrix consisting of the differences $\partial\mathbf{g}/\partial\mathbf{z} - F$, which is independent of l,k. The Cauchy-Schwarz inequality yields

$$|\langle \Phi_l\Phi_k d_{ij}\rangle| \le \sqrt{\langle \Phi_l^2\Phi_k^2\rangle} \cdot \sqrt{\langle d_{ij}^2\rangle} \le \sqrt{\max_{i=0,1,\ldots,M} \langle \Phi_i^4\rangle} \cdot \sqrt{\langle d_{ij}^2\rangle}.$$

Without loss of generality, we apply the Euclidean vector norm $\|\cdot\|_2$ and the consistent Frobenius matrix norm $\|\cdot\|_*$. The Lipschitz-continuity of the functions in $\partial\mathbf{g}/\partial\mathbf{z}$ allows for the conclusion

$$\|\langle \Phi_l\Phi_k D\rangle\|_*^2 \le C_M \left\langle \left\| \mathbf{y}_\lambda^{(M)} - \mathbf{y}_0 \right\|_2^2 + \left\| \mathbf{z}_\lambda^{(M)} - \mathbf{z}_0 \right\|_2^2 + \|\lambda\mathbf{p}\|_2^2 \right\rangle$$

with a constant $C_M > 0$. Using the assumptions of the theorem, it follows:

$$\lim_{\lambda \to 0} \|\langle \Phi_l \Phi_k D \rangle\|_* = 0$$

for all $l, k = 0, 1, \ldots, M$. Thus, we obtain

$$\lim_{\lambda \to 0} \mathcal{R}(t) = 0,$$

which implies the Eq. (22). □

We motivate the assumption (23) further. It holds

$$\|\mathbf{y}_\lambda^{(M)}(t, \mathbf{p}) - \mathbf{y}_0(t, \mathbf{p})\| \leq \|\mathbf{y}_\lambda^{(M)}(t, \mathbf{p}) - \mathbf{y}_\lambda(t, \mathbf{p})\| + \|\mathbf{y}_\lambda(t, \mathbf{p}) - \mathbf{y}_0(t, \mathbf{p})\|.$$

The first term on the right-hand side represents the error of the stochastic Galerkin method. This error tends to zero in the norm $L^2(\Pi, \rho)$ [see Eq. (23)] for $M \to \infty$ provided that the stochastic Galerkin method is convergent. However, the constant C_M, which appears in the proof of Theorem 2, depends on M. Hence, we do not achieve a condition uniformly for all $M \geq M_0$ with some sufficiently large M_0. The second term on the right-hand side converges to zero for $\lambda \to 0$ if the solution depends continuously on the physical parameters. The same discussion applies to the algebraic part $\mathbf{z}$.

Concerning the index of the corresponding systems of DAEs, we achieve the following result.

Corollary 1. If the system (21) exhibits index 1 in case of the reference physical parameter $\mathbf{r}_0$ ($\lambda = 0$), then the coupled system of (16) and (17) inherits the index 1 for a sufficiently small variance of the input random variables provided that the assumptions of Theorem 2 are satisfied.

The results of Theorem 2 and Corollary 1 can be generalized straightforward to the case of DAEs in Hessenberg form [5], with higher index, because the index is characterized by the regularity of specific matrices.

The above conclusions do not contradict the results from Section 3.2. The counterexample requires to choose a symmetric distribution around $p = 0$. Hence, we must select $r_0 = 0$ in this case, where the index-1 assumption is violated for the corresponding system (21).

The concept applied in this subsection implies only an asymptotic statement. We do not obtain a criterion on the index for fixed $\lambda \neq 0$. Hence, further investigations are performed in the following Sections 3.4 and 3.5.

3.4 Dependence on Sign of Eigenvalues

A criterion for the index-1 property can be obtained by demanding that the signs of the eigenvalues of the Jacobian matrix do not change. We investigate the scalar case ($N_z = 1$) first.

Theorem 3. *If it holds $\partial g/\partial z > 0$ for almost all $\mathbf{p}$ or $\partial g/\partial z < 0$ for almost all $\mathbf{p}$, then the matrix $\mathcal{G}$ from (19) is positive or negative definite, respectively.*

Proof. Let $\mathbf{u} = (u_0, u_1, \ldots, u_M)^\top \in \mathbb{R}^{M+1}$ and $\mathbf{u} \neq 0$. It follows

$$\mathbf{u}^\top \mathcal{G} \mathbf{u} = \sum_{l,k=0}^{M} u_l u_k \left\langle \Phi_l \Phi_k \frac{\partial g}{\partial z} \right\rangle = \left\langle \left(\sum_{l,k=0}^{M} u_l u_k \Phi_l \Phi_k \right) \frac{\partial g}{\partial z} \right\rangle = \left\langle \left(\sum_{l=0}^{M} u_l \Phi_l \right)^2 \frac{\partial g}{\partial z} \right\rangle.$$

The latter probabilistic integral includes a non-negative polynomial. Because of $\mathbf{u} \neq 0$ and the linear independence of the basis polynomials, this polynomial is not identical to zero. Thus, the number of zeros of the polynomial is finite. We obtain $\mathbf{u}^\top \mathcal{G} \mathbf{u} > 0$ for $\partial g/\partial z > 0$ and $\mathbf{u}^\top \mathcal{G} \mathbf{u} < 0$ for $\partial g/\partial z < 0$. □

In both cases of the theorem, the coupled system of (16) and (17) inherits the index-1 property from the original systems (6). Moreover, the property is independent of the choice of the subset of orthogonal basis polynomials. Theorem 3 does not contradict the results of Section 3.2. The counterexample applies a symmetric probability distribution around the critical point $\partial g/\partial z = p = 0$. Hence, both positive and negative values appear and the assumptions of Theorem 3 are not satisfied. The results of the scalar case can be generalized to the multidimensional case $N_z > 1$ under additional assumptions.

Theorem 4. *Let the matrix $\partial\mathbf{g}/\partial\mathbf{z}$ be real diagonalisable in the form*

$$\frac{\partial \mathbf{g}}{\partial \mathbf{z}} = U(t) D(t,\mathbf{p}) U(t)^{-1} \tag{24}$$

with a regular matrix $U(t) \in \mathbb{R}^{N_z \times N_z}$ and a diagonal matrix $D(t,\mathbf{p})$, including the entries $\lambda_i(t,\mathbf{p})$ for $i = 1,\ldots,N_z$. If each eigenvalue λ_i is either positive or negative for almost all $\mathbf{p}$, then the matrix $\mathcal{G}$ from (19) is regular.

Proof. The minors of the matrix $\mathcal{G}$ can be written as

$$\frac{\partial \mathbf{G}_l}{\partial \mathbf{w}_k} = \left\langle \Phi_l(\mathbf{p})\Phi_k(\mathbf{p}) \frac{\partial \mathbf{g}}{\partial \mathbf{z}} \right\rangle = \left\langle \Phi_l(\mathbf{p})\Phi_k(\mathbf{p}) U(t) D(t,\mathbf{p}) U(t)^{-1} \right\rangle = U(t) \left\langle \Phi_l(\mathbf{p})\Phi_k(\mathbf{p}) D(t,\mathbf{p}) \right\rangle U(t)^{-1}$$

for $l,k = 0,1,\ldots,M$. We obtain the similarity transformation

$$\mathcal{G}(t) = [I_{M+1} \otimes U(t)] \hat{\mathcal{G}}(t) [I_{M+1} \otimes U(t)^{-1}].$$

The matrix $\hat{\mathcal{G}}(t)$ consists of the minors

$$(\hat{\mathcal{G}}(t))_{l,k} = \left\langle \Phi_l(\mathbf{p})\Phi_k(\mathbf{p}) D(t,\mathbf{p}) \right\rangle .$$

Let $B(\mathbf{p}) := [\Phi_l(\mathbf{p})\Phi_k(\mathbf{p})] \in \mathbb{R}^{(M+1)\times(M+1)}$. A specific permutation matrix $\mathcal{P} \in \mathbb{R}^{(M+1)N_z \times (M+1)N_z}$ independent of t exists such that the transformed matrix exhibits a block diagonal structure with the minors

$$[\mathcal{P} \hat{\mathcal{G}}(t) \mathcal{P}]_{i,i} = \langle \lambda_i(t,\mathbf{p}) B(\mathbf{p}) \rangle$$

for $i = 1,\ldots,N_z$. The entries of each diagonal block are the same as in the matrix of the coupled system for the case $N_z = 1$ with λ_i instead of $\partial g/\partial z$. Theorem 3 implies that each diagonal block is positive or negative definite provided that the eigenvalues λ_i do not change their signs. Because the regularity of a matrix is invariant under permutations of rows and columns, it follows $\det[\hat{\mathcal{G}}(t)] \neq 0$ and, thus, $\det[\mathcal{G}(t)] \neq 0$. □

The assumptions of Theorem 4 are relatively strong. First, the Jacobian matrix $\partial\mathbf{g}/\partial\mathbf{z}$ has to be real diagonalizable, which excludes matrices with complex eigenvalues, for example. Second, the transformation matrices U are assumed to be independent of the parameters $\mathbf{p}$. Nevertheless, both properties are given in case of a diagonal matrix $\partial\mathbf{g}/\partial\mathbf{z}$. Concerning the index-1 property, we obtain the following implication.

Corollary 2. Let the assumptions of Theorem 4 be fulfilled. Moreover, let the matrix $\partial\mathbf{g}/\partial\mathbf{z}$ depend continuously on the parameters. If the domain $S = \mathrm{supp}(\rho)$ is path connected, then Condition 1 implies an index of 1 for the coupled system of (16) and (17).

Proof. The assumption (24) yields

$$U(t)^{-1} \frac{\partial \mathbf{g}}{\partial \mathbf{z}}(t,\mathbf{p}) U(t) = D(t,\mathbf{p}).$$

Hence, the eigenvalues depend continuously on the parameters. Condition 1 implies that each eigenvalue is nonzero for all $\mathbf{p} \in S$. It follows that an eigenvalue does not change its sign. Otherwise, we obtain a path between two points in S, where an eigenvalue zero appears on the path. Now, Theorem 4 yields the index-1 property. □

The counterexample of Section 3.2, where Condition 1 is satisfied, does not involve a path connected domain S. Furthermore, a convex domain is always path connected. We will retrieve this property in Section 3.5.

3.5 Criterion from Numerical Range

For a matrix $C \in \mathbb{C}^{N\times N}$, the numerical range

$$W(C) := \left\{ \mathbf{u}^* C \mathbf{u} : \mathbf{u} \in \mathbb{C}^N,\ \mathbf{u}^*\mathbf{u} = 1 \right\} \subset \mathbb{C}$$

represents a closed and convex set. The spectrum of C is a subset of $W(C)$. We define the numerical range of our random Jacobian matrix as

$$\widehat{W}\left(\frac{\partial \mathbf{g}}{\partial \mathbf{z}}\right) := \left\{ \mathbf{u}^* \frac{\partial \mathbf{g}}{\partial \mathbf{z}} \mathbf{u} : \mathbf{u} \in \mathbb{C}^{N_z},\ \mathbf{u}^*\mathbf{u} = 1,\ \mathbf{p} \in S \right\} = \bigcup_{\mathbf{p}\in S} W\left(\frac{\partial \mathbf{g}}{\partial \mathbf{z}}\right). \tag{25}$$

This set is not necessarily closed or convex within $\mathbb{C}$. However, the numerical range (25) is larger than required for our purposes. We apply an essential numerical range introduced in [19].

Definition 2. Let $B(z,\varepsilon)$ be the ball of radius ε centered at $z \in \mathbb{C}$ and

$$A(z,\varepsilon) := \left\{ \mathbf{p} \in S : B(z,\varepsilon) \cap W\left(\frac{\partial \mathbf{g}}{\partial \mathbf{z}}\right) \neq \emptyset \right\}.$$

The essential numerical range of the random Jacobian matrix is

$$\widetilde{W}\left(\frac{\partial \mathbf{g}}{\partial \mathbf{z}}\right) := \left\{ z \in \mathbb{C} : \int_{A(z,\varepsilon)} \rho(\mathbf{p})\, d\mathbf{p} > 0 \ \text{ for all } \varepsilon > 0 \right\}.$$

In comparison to the numerical range (25), the definition of the essential numerical range yields

$$\widetilde{W}\left(\frac{\partial \mathbf{g}}{\partial \mathbf{z}}\right) \subseteq \overline{\widehat{W}\left(\frac{\partial \mathbf{g}}{\partial \mathbf{z}}\right)}.$$

It follows:

$$\text{conv}\left[\widetilde{W}\left(\frac{\partial \mathbf{g}}{\partial \mathbf{z}}\right)\right] \subseteq \text{conv}\left[\overline{\widehat{W}\left(\frac{\partial \mathbf{g}}{\partial \mathbf{z}}\right)}\right] \tag{26}$$

for the convex hull of the sets.

Theorem 5. *If* $0 \notin \text{conv}\left[\widetilde{W}\left(\partial \mathbf{g}/\partial \mathbf{z}\right)\right]$ *holds, then the matrix* $\mathcal{G}$ *from (19) is regular for an arbitrary choice of an orthogonal basis.*

Proof. Theorem 2 in [19] implies that

$$\text{spect}(\mathcal{G}) \subset \text{conv}\left[\widetilde{W}\left(\frac{\partial \mathbf{g}}{\partial \mathbf{z}}\right)\right]$$

holds for the spectrum of the gPC matrix $\mathcal{G}$. If $0 \notin \text{conv}\left[\widetilde{W}\left(\partial \mathbf{g}/\partial \mathbf{z}\right)\right]$ is satisfied; then it follows $0 \notin \text{spect}(\mathcal{G})$ and the matrix $\mathcal{G}$ is regular. □

The essential numerical range may be difficult to determine. Applying the inclusion (26), we obtain the following criterion, which is often easier to verify provided that it is fulfilled by the problem.

Corollary 3. If $0 \notin \text{conv}\left[\overline{\widehat{W}\left(\partial \mathbf{g}/\partial \mathbf{z}\right)}\right]$ holds, then the matrix $\mathcal{G}$ from (19) is regular for an arbitrary choice of an orthogonal basis.

We consider problems, where the original semiexplicit systems exhibit the index 1 for all $\mathbf{p} \in S$ (see Condition 1). We conclude that zero is not in the spectrum of $\partial \mathbf{g}/\partial \mathbf{z}$ for all $\mathbf{p} \in S$. Yet this condition is not sufficient to guarantee $0 \notin \widehat{W}\left(\partial \mathbf{g}/\partial \mathbf{z}\right)$ in the case $N_z > 1$.

For $N_z = 1$, we obtain more potential for conclusions. The numerical range (25) simplifies to

$$\widehat{W}\left(\frac{\partial g}{\partial z}\right) = \left\{ \frac{\partial g}{\partial z} : \mathbf{p} \in S \right\} \subseteq \mathbb{R}. \tag{27}$$

Consequently, Condition 1 of the original semiexplicit systems (6) implies the property $0 \notin \widehat{W}\,(\partial g/\partial z)$. Because of Theorem 5, we must consider the convex hull of the numerical range. However, the condition $0 \notin \text{conv}\left[\widehat{W}\,(\partial g/\partial z)\right]$ is equivalent to $\partial g/\partial z \geq \eta > 0$ or $\partial g/\partial z \leq \eta < 0$ provided that the partial derivative depends continuously on the parameters. Because the assumption can be weakened using the essential numerical range, the criterion becomes $\partial g/\partial z > 0$ or $\partial g/\partial z < 0$. Therefore, we achieve the same conclusion as in Section 3.4.

The assumptions in Theorem 5 and Corollary 3 are relatively strong because neither Conditions 1 nor 2 are sufficient to guarantee these requirements, in general. Nevertheless, we have to demand this criterion due to the counterexample in Section 3.2 because it holds

$$0 \notin \overline{\widehat{W}\left(\frac{\partial g}{\partial z}\right)}, \qquad 0 \in \text{conv}\left[\overline{\widehat{W}\left(\frac{\partial g}{\partial z}\right)}\right]$$

in the case of a uniform distribution within $S = [-b, -a] \cup [a, b]$ for some $0 < a < b$, for example.

4. TEST EXAMPLES

We now apply the stochastic collocation techniques as well as the stochastic Galerkin method to four problems.

4.1 Benchmark System

As a benchmark, we employ a simple test example with $N_y = N_z = 1$, i.e.,

$$\begin{aligned} y'(t) &= -(1 + \sigma_1 p_1) y(t) + z(t)^3 \\ 0 &= -(1 + \sigma_2 p_2) y(t) - z(t)^3 \end{aligned} \tag{28}$$

including two parameters p_1, p_2. It holds

$$\frac{\partial g}{\partial z} = -3z(t)^2 < 0 \quad \text{for all} \;\; z(t) \neq 0. \tag{29}$$

Thus, the index of the system (28) is one for arbitrary parameters provided that $z \neq 0$. We specify the consistent initial values

$$y(0) = 1, \quad z(0) = -\sqrt[3]{1 + \sigma_2 p_2}. \tag{30}$$

The exact solution of the initial value problem (28) and (30) reads

$$y(t) = \exp[-(2 + \sigma_1 p_1 + \sigma_2 p_2)t], \quad z(t) = -\sqrt[3]{(1 + \sigma_2 p_2) y(t)}. \tag{31}$$

If we arrange distributions of the random parameters with a bounded domain S and sufficiently small values p_1, p_2, then the problem satisfies Condition 1. Alternatively, we choose two independent Gaussian random variables with mean zero and unit variance for p_1, p_2 now. Hence, only Condition 2 holds. Both increasing and decreasing exponential functions y appear, where the probability of a decreasing process is much higher. Figure 2 shows the maximum variances for $t \in [0, 10]$ of the solution of (28) with respect to the variances $\sigma_1^2 = \sigma_2^2$ in the input parameters, which are computed by a quadrature using Eq. (31).

Although the problem fullfils Condition 2 only, Theorem 3 guarantees that the coupled system of the stochastic Galerkin approach exhibits index 1 due to the property (29). Because Gaussian distributions are considered, the gPC applies the Hermite polynomials. We include all two-variate polynomials up to degree 3 in the truncated series (12), which results in 10 basis polynomials ($M = 9$). On the one hand, a stochastic collocation yields approximations of the coefficient functions based on a two-dimensional Gauss-Hermite quadrature with a grid of size 7×7. On the other hand, the stochastic Galerkin method requires to solve the coupled system of (16) and (17), where the probabilistic integrals in the right-hand sides are discretized using a Gauss-Hermite quadrature on a grid of size 7×7 again. Because

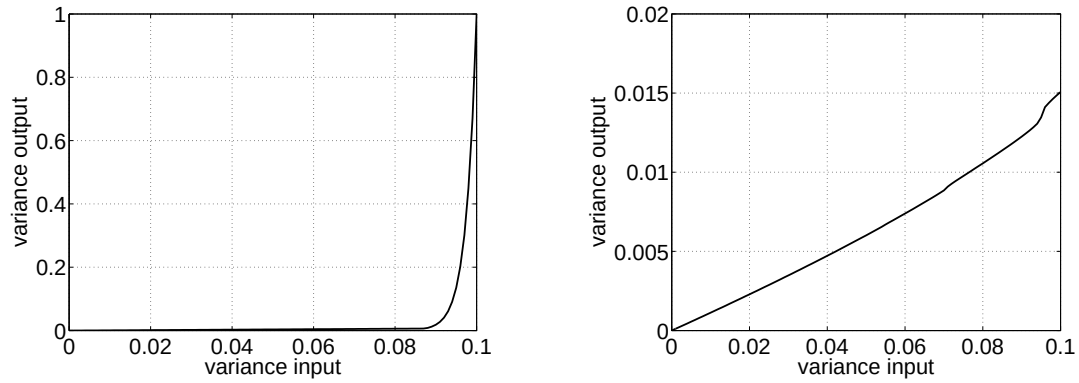

FIG. 2: Maximum variance for $t \in [0, 10]$ of differential component y (left) and algebraic component z (right) for different variances in the input parameters.

the right-hand side of (28) consists of polynomials, the evaluations of the probabilistic integrals are exact except for roundoff errors in the stochastic Galerkin technique.

We choose $\sigma_1 = \sigma_2 = 0.1$ in (28) now. We compute the solutions within the time interval $[t_0, t_1] = [0, 10]$. In each method, the backward differentiation formula (BDF) of second order solves the systems of DAEs [5]. Thereby, we apply a constant step size $\Delta t = 0.02$.

Because both time integration and discretization in probability space employ the same schemes and grids, the number of right-hand side evaluations of the original systems (28) is identical in both stochastic collocation and stochastic Galerkin method. Only the linear algebra part of the Newton iterations within the implicit time integration causes a computational overhead in the stochastic Galerkin technique. The computations have been performed in the software package MATLAB. The CPU time of the collocation and the Galerkin method was 4.8 and 22.9, respectively.

We illustrate the results of the stochastic collocation. Figure 3 shows the expected values and the standard deviation of the stochastic processes. Furthermore, Fig. 4 depicts the coefficient functions of the gPC expansions for higher degrees. Note that some coefficient functions coincide for the differential component y.

To compare the accuracy of the stochastic collocation and the stochastic Galerkin method, we compute a reference solution using the exact solutions (31) in a stochastic collocation based on a Monte Carlo simulation with $K = 10^7$ samples. Figure 5 visualizes the maximum differences of the approximations with respect to the reference solution. It follows that the accuracy of both approaches coincides for all computed coefficient functions. The approximations of both methods are nearly the same due to the identical discretizations. Although numerical errors of the approximations

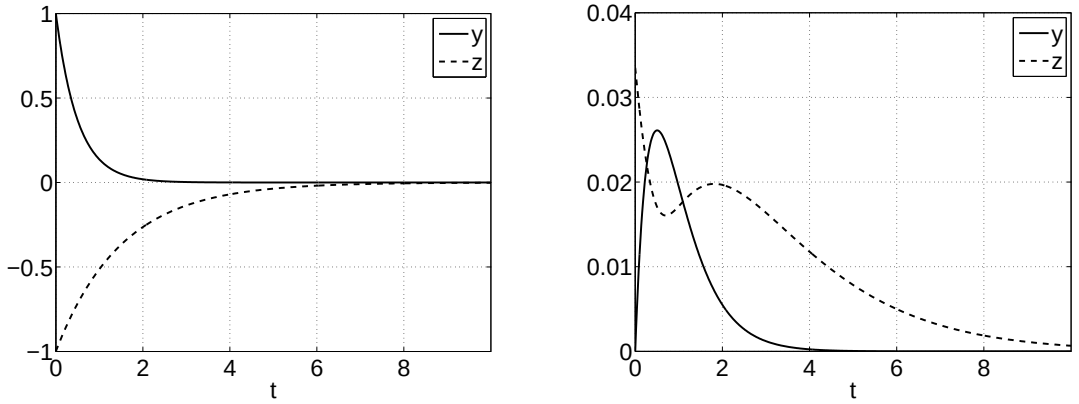

FIG. 3: Expected values (left) and standard deviation (right) for differential component y and algebraic component z.

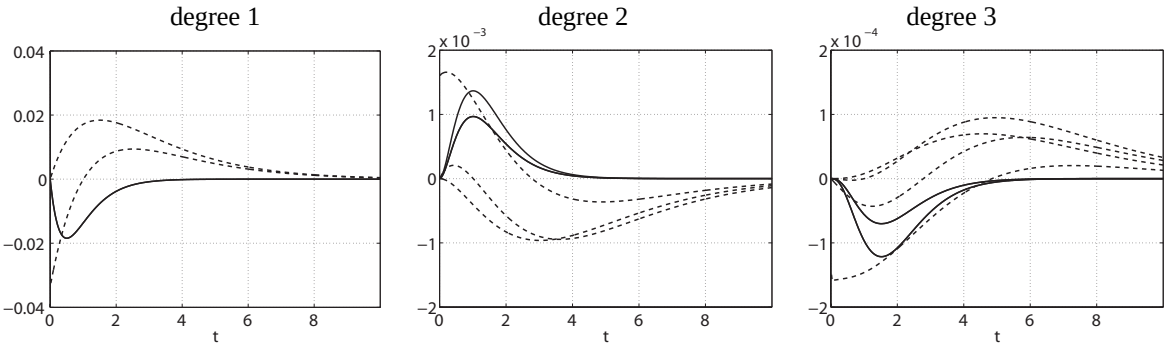

FIG. 4: Coefficient functions in gPC expansion for differential component y (solid lines) and algebraic component z (dashed lines).

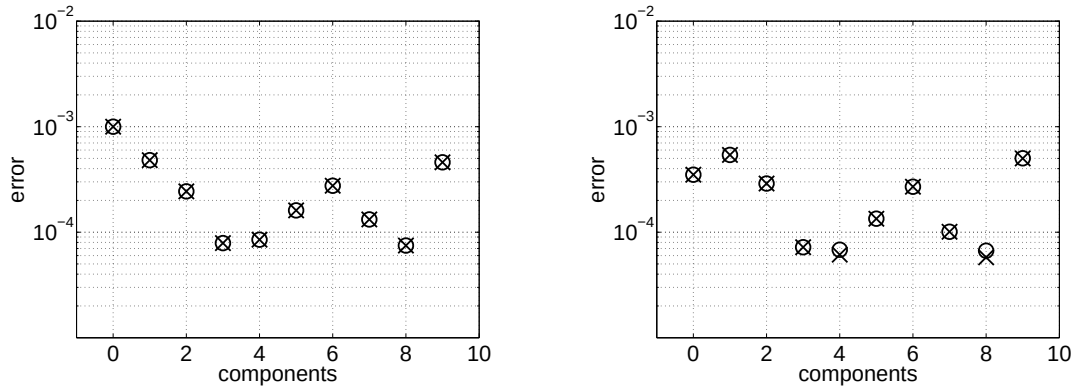

FIG. 5: Maximum differences in reference solution versus st. Galerkin (circles) and reference solution versus st. collocation (crosses) for coefficient functions of differential component (left) and algebraic component (right)—semilog. scale.

of probabilistic integrals are avoided within the stochastic Galerkin method, the stochastic collocation method is not outperformed with respect to accuracy.

Finally, we analyze the speed of convergence of the gPC expansions. For the basis $(\Phi_i)_{i\in\mathbb{N}}$, let $\mathcal{J}(R)$ be the set of all indices i with polynomials Φ_i of degree less or equal R. Using the norm of $L^2(\Pi,\rho)$, we examine the difference between partial sums in the style of a Cauchy sequence, i.e.,

$$\left\| \sum_{i\in\mathcal{J}(R)} v_i(t)\Phi_i(\mathbf{p}) - \sum_{i\in\mathcal{J}(R-1)} v_i(t)\Phi_i(\mathbf{p}) \right\| = \sqrt{\sum_{i\in\mathcal{J}(R)\setminus\mathcal{J}(R-1)} v_i(t)^2} \tag{32}$$

depending on $R = 1, 2, 3, \ldots$ for the differential part and the algebraic part of (28). Stochastic collocation including Gauss-Hermite quadrature yields the coefficient functions. The maximum norms (32) for $t \in [0, 10]$ are shown in Fig. 6. We recognize an exponential convergence of the gPC expansions, which is typical for smooth functions in $C^\infty(\Pi)$.

4.2 Linear Oscillator

Mathematical modeling of electric circuits typically results in systems of DAEs [2, 3]. We consider an electromagnetic oscillator consisting of a capacitance C, an inductance L, and a resistance R in parallel. A particular modeling yields the linear semiexplicit system

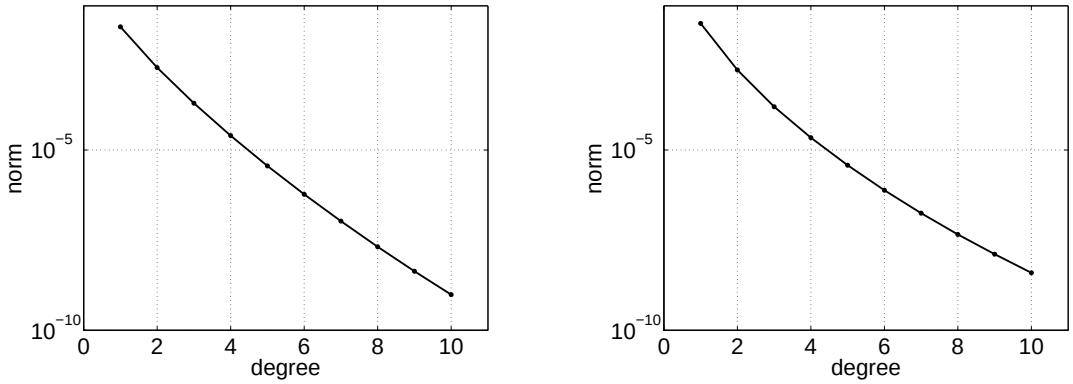

FIG. 6: Convergence of gPC expansions for differential part y (left) and algebraic part z (right)—semilog. scale.

$$\begin{aligned} u'(t) &= \frac{1}{C}\imath_C(t) \\ \imath_L'(t) &= \frac{1}{L}u(t) \\ 0 &= R\,\imath_R(t) - u(t) \\ 0 &= \imath_C(t) + \imath_L(t) + \imath_R(t) \end{aligned} \tag{33}$$

for the unknown node voltage u and the branch currents $\imath_L, \imath_R, \imath_C$. The index of the system (33) is equal to one for $R \neq 0$. Physically reasonable parameters are $C, L, R > 0$.

We change the resistance R into a random variable now. The solution of (33) depends nonlinear on this random parameter. The index of the coupled system of (16) and (17) from the stochastic Galerkin method can be investigated via the criterion from Theorem 5. Let the ordering of the unknowns be $\mathbf{y} = (u, \imath_L)^\top$ and $\mathbf{z} = (\imath_R, \imath_C)^\top$. The numerical range of the corresponding Jacobian matrices $\partial\mathbf{g}/\partial\mathbf{z}$ follows from

$$\mathbf{u}^* \frac{\partial \mathbf{g}}{\partial \mathbf{z}} \mathbf{u} = \begin{pmatrix} \overline{u_1} & \overline{u_2} \end{pmatrix} \begin{pmatrix} R & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = R|u_1|^2 + u_1\overline{u_2} + |u_2|^2$$

with $|u_1|^2 + |u_2|^2 = 1$. Assuming $R \geq 1$ for each realization of the random variable, it holds $\mathrm{Re}\,(\mathbf{u}^* \partial\mathbf{g}/\partial\mathbf{z}\mathbf{u}) \geq 1/2$ for the real part. Thus, the numerical range (25) satisfies

$$\widehat{W}\left(\frac{\partial \mathbf{g}}{\partial \mathbf{z}}\right) \subseteq \left\{ z \in \mathbb{C} : \mathrm{Re}(z) \geq \frac{1}{2} \right\}.$$

Because this half-plane is closed as well as convex and does not include zero, it follows the regularity of the Jacobian matrix $\mathcal{G}$ by Corollary 3. Hence, the coupled system of (16) and (17) is also of index 1.

We arrange the physical parameters $C = 10^{-9}$ F, $L = 10^{-6}$ H, and $R = 10^2\ \Omega$. The corresponding solutions of initial value problems of the system (33) represent damped oscillations. We apply the random resistance

$$\widetilde{R}(p) := R(1 + 0.2p)$$

with a uniformly distributed random variable $p \in [-1, 1]$. The initial values

$$u(0) = 0\ \mathrm{V}, \quad \imath_L(0) = 0.1\ \mathrm{A}, \quad \imath_R(0) = 0\ \mathrm{A}, \quad \imath_C(0) = -0.1\ \mathrm{A} \tag{34}$$

are a consistent choice for arbitrary parameters. Numerical simulations are performed within the time interval $[t_0, t_1] = [0\mathrm{s}, 10^{-6}\mathrm{s}]$. The BDF scheme of second order is used for the integrations with equidistant step size $\Delta t = 10^{-9}\mathrm{s}$.

The gPC expansions are based on the Legendre polynomials now. We include all polynomials up to degree 5 in the truncated series (12) (i.e., $M = 5$). The stochastic collocation technique applies a Gauss-Legendre quadrature with $K = 4$ nodes. Because the system of DAEs (33) is linear, the right-hand side of the coupled system of (16) and (17) is also linear in the stochastic Galerkin method. The matrix of the right-hand side consists of time-independent probabilistic integrals (4), which are calculated just once before starting the time integration. Moreover, a Gauss-Legendre quadrature yields the exact probabilistic integrals except for roundoff errors.

The expected values and standard deviations of the solution resulting from the stochastic Galerkin method are illustrated in Figs. 7 and 8, respectively. Because the solutions of the system (33) represent damped oscillations, the variance decreases in time.

The corresponding computational effort of both methods is nearly the same with CPU times of 0.08 for stochastic collocation and 0.06 for stochastic Galerkin. For a comparison of accuracy, we compute a reference solution based on stochastic collocation using the midpoint rule with $K = 10^3$ equidistant nodes. The corresponding time integration applies the step size $\Delta t = 0.5 \times 10^{-9}$ s. The resulting errors are shown in Fig. 9. The stochastic Galerkin method achieves a better accuracy for coefficients of a high degree. Increasing the accuracy of the stochastic collocation requires more nodes in the Gaussian quadrature, which causes a higher computational effort. Hence, the stochastic Galerkin technique is more efficient in this test example.

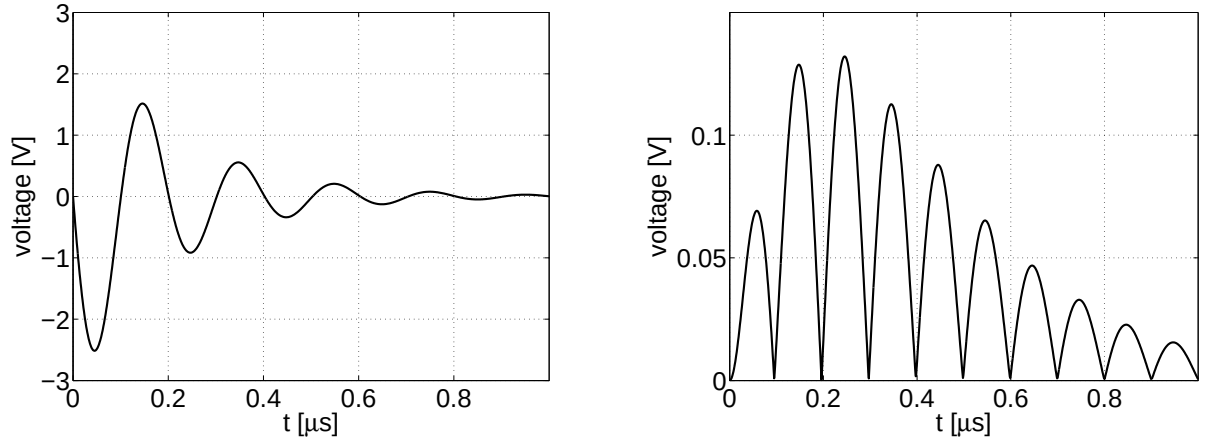

FIG. 7: Expected values (left) and standard deviation (right) of node voltage u in linear oscillator.

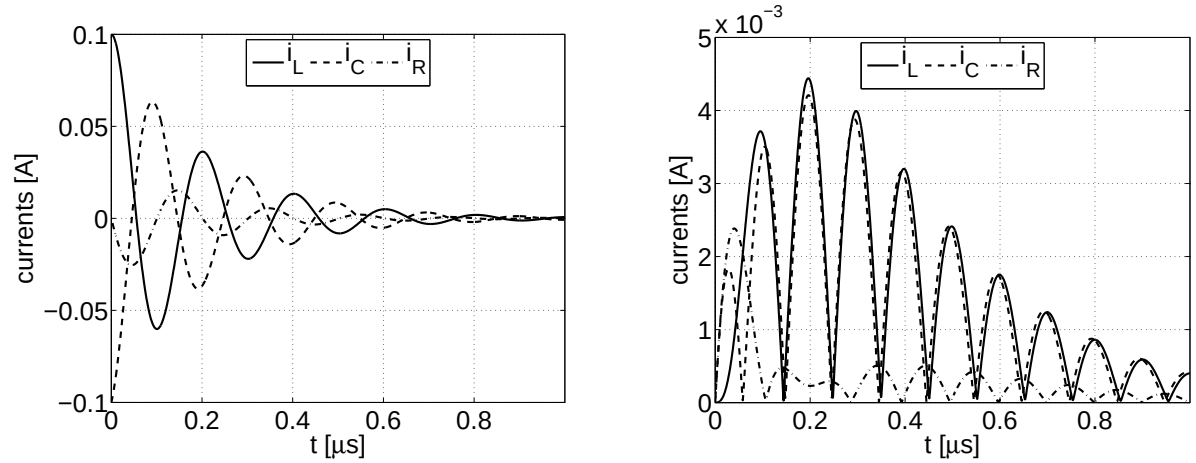

FIG. 8: Expected values (left) and standard deviation (right) of branch currents in linear oscillator.

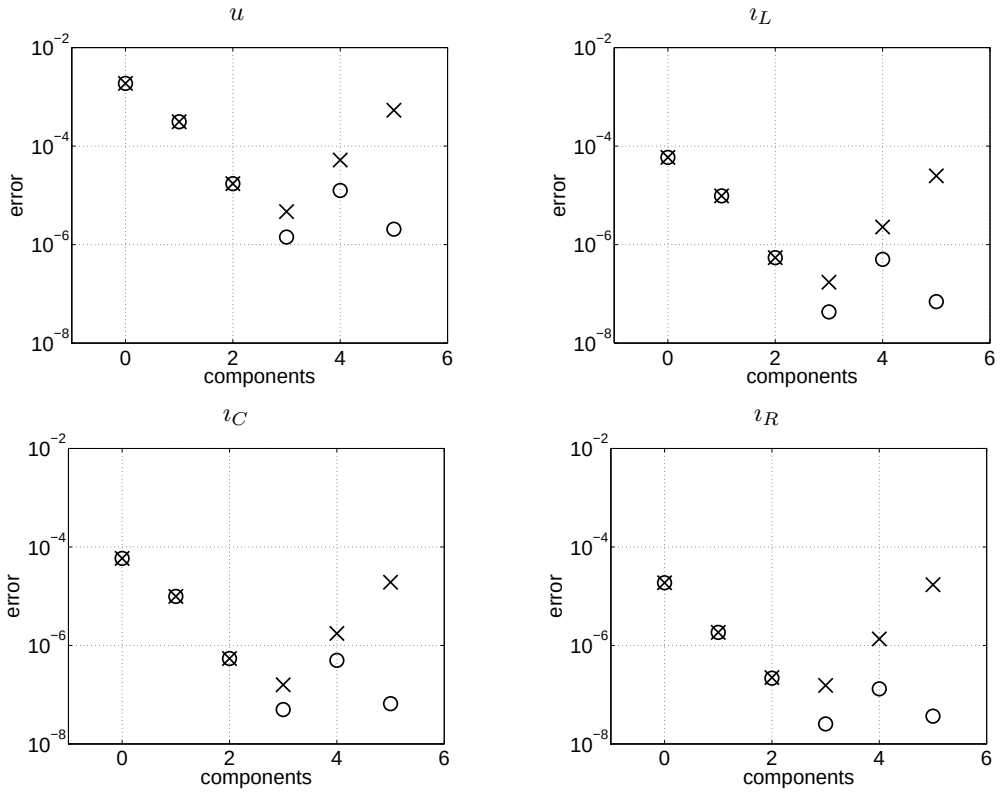

FIG. 9: Maximum differences in reference solution versus st. Galerkin (circles) and reference solution versus st. collocation (crosses) for coefficient functions of gPC expansion for the solution of the linear oscillator—semilog. scale.

4.3 Nonlinear Oscillator

Now we consider a nonlinear electric network, where a nonlinear resistance replaces the linear resistance in the previous example (33). It follows the system of DAEs:

$$\begin{aligned} u'(t) &= \frac{1}{C}\imath_C(t) \\ \imath_L'(t) &= \frac{1}{L}u(t) \\ 0 &= \imath_R(t) - (G_0 - G_\infty)U_0 \tanh\left[\frac{u(t)}{U_0}\right] - G_\infty u(t) \\ 0 &= \imath_C(t) + \imath_L(t) + \imath_R(t). \end{aligned} \tag{35}$$

The index of the semiexplicit system (35) is always equal to 1. We arrange the physical parameters

$$C = 10^{-9}\ \text{F},\quad L = 10^{-6}\ \text{H},\quad U_0 = 1\ \text{V},\quad G_0 = -0.1\ \text{A/V},\quad G_\infty = 0.25\ \text{A/V}.$$

The solutions of initial value problems tend to a periodic limit cycle. Therefore this example has been applied in [13] for an uncertainty quantification of periodic boundary value problems. Alternatively, we consider the initial value problem (34), which represents a consistent choice for arbitrary parameters.

In a stochastic modeling, we replace the two deterministic parameters C and G_∞ by the random parameters

$$\widetilde{C}(p_1) := C(1 + 0.01p_1), \quad \widetilde{G}_\infty(p_2) := G_\infty(1 + 0.01p_2)$$

with independent random variables p_1, p_2 both uniformly distributed in the interval $[-1, 1]$. Hence, it holds $\widetilde{C}, \widetilde{G}_\infty > 0$ for all realizations of the random variables, which is necessary to achieve physically reasonable systems (35). Both parameters $\widetilde{C}$ and $\widetilde{G}_\infty$ influence the frequency of the corresponding solutions.

In the gPC methods, we use all two-variate basis polynomials up to degree 3 in the truncated series (12) (i.e., $M = 9$). Although the algebraic part of (35) includes a random parameter, this example satisfies the assumptions of Theorem 1. Hence, the coupled system of (16) and (17) inherits the index-1 property in the stochastic Galerkin approach. Gauss-Legendre quadrature yields approximations of probabilistic integrals using a grid of size 4×4 in both stochastic collocation and stochastic Galerkin techniques. Numerical simulations are performed within the time interval $[t_0, t_1] = [0 \text{ s}, 10^{-6} \text{ s}]$ again. Time integration applies the BDF scheme of second order with constant step size $\Delta t = 10^{-9}$ s.

Figures 10 and 11 illustrate the expected values and standard deviations of the components of the random process, which are reconstructed by the solutions from the stochastic collocation method. Since solutions of the original system (35) corresponding to different parameters exhibit different frequencies, the variances increase in time. However, the variance does not tend to infinity but to a periodic state, indicating high uncertainties. Nevertheless, the order of the truncated gPC expansions (12) must be increased for larger times to guarantee sufficiently accurate approximations (i.e., larger values M must be applied).

Using the software package MATLAB, the CPU times of the stochastic collocation and the stochastic Galerkin method were 5.4 and 17.5, respectively. To compare the accuracy, we compute a reference solution via stochastic collocation, including a quadrature with midpoint rule on a grid of size 100×100 and time integration with step size

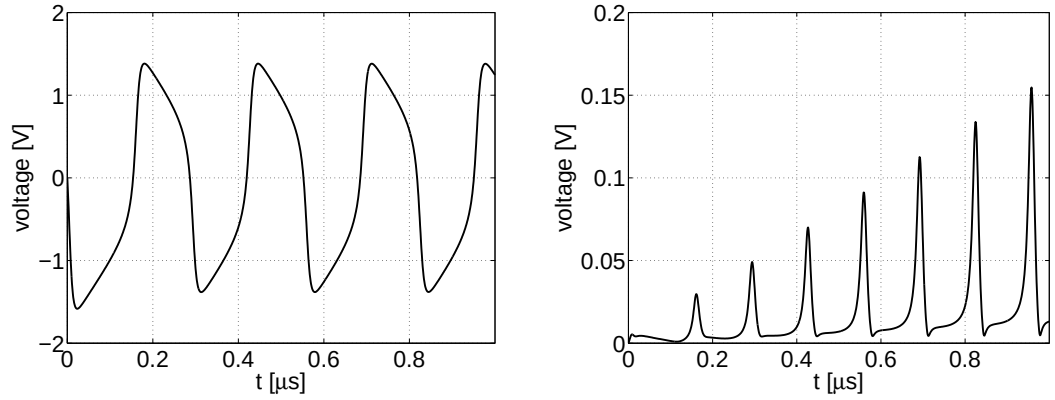

FIG. 10: Expected values (left) and standard deviation (right) of node voltage u in nonlinear oscillator.

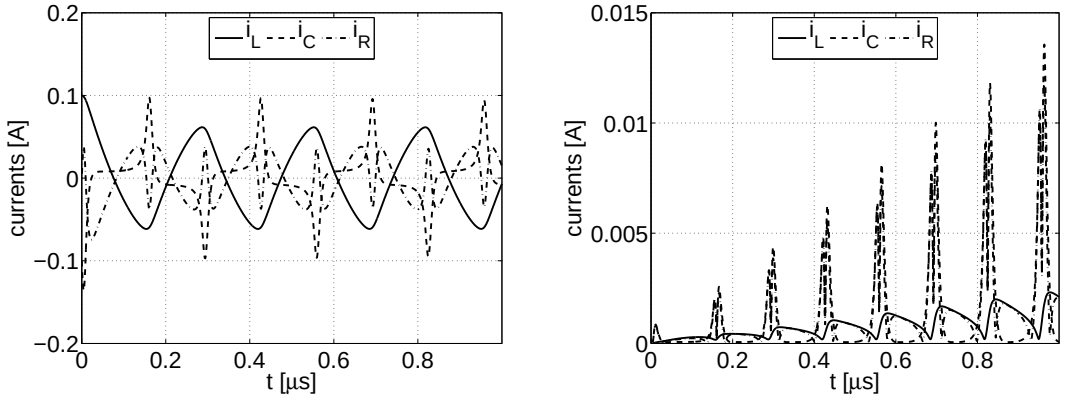

FIG. 11: Expected values (left) and standard deviation (right) of branch currents in nonlinear oscillator.

$\Delta t = 0.5 \times 10^{-9}$ s. Figure 12 shows the maximum differences of the approximations with respect to the reference solution. We observe a good agreement of the achieved accuracies within the stochastic collocation and the stochastic Galerkin technique, which is caused by the application of the same discretization schemes again. However, the computational effort of the stochastic Galerkin method is significantly larger due to the linear algebra part.

4.4 Transistor Modulator

Finally, we consider the electric circuit of a transistor modulator, which was introduced in [20]. A mathematical modeling yields a nonlinear semiexplicit system of DAEs for three node voltages u_1, u_2, u_3 and a branch current $\imath$, i.e.,

$$\begin{aligned} u_1' &= \frac{1}{C}\left(-\frac{1}{R_L}u_1 - \imath - I_0\left\{\exp[\delta(u_1 - U_{\mathrm{cc}})] - 1\right\} + \alpha I_0\left[\exp(\delta u_3) - 1\right]\right) \\ \imath' &= \frac{1}{L}u_1 \\ 0 &= -I_0\left\{\exp[\delta(u_3 + U_{\mathrm{in2}})] - 1\right\} + \alpha I_0\left\{\exp[\delta(U_{\mathrm{in2}} - U_{\mathrm{cc}})] - 1\right\} - I_0\left[\exp(\delta u_3) - 1\right] \\ &\quad + \alpha I_0\left\{\exp[\delta(u_1 - U_{\mathrm{cc}})] - 1\right\} - I_0\left\{\exp[\delta(u_3 + U_{\mathrm{in1}} - U_{\mathrm{op}})] - 1\right\} + \alpha I_0\left[\exp(\delta u_2) - 1\right] \\ 0 &= -I_0\left[\exp(\delta u_2) - 1\right] + \alpha I_0\left\{\exp[\delta(u_3 + U_{\mathrm{in1}} - U_{\mathrm{op}})] - 1\right\} + \frac{1}{R_B}\left[U_{\mathrm{ss}} + U_{\mathrm{in1}} - U_{\mathrm{op}} - u_2\right]. \end{aligned} \tag{36}$$

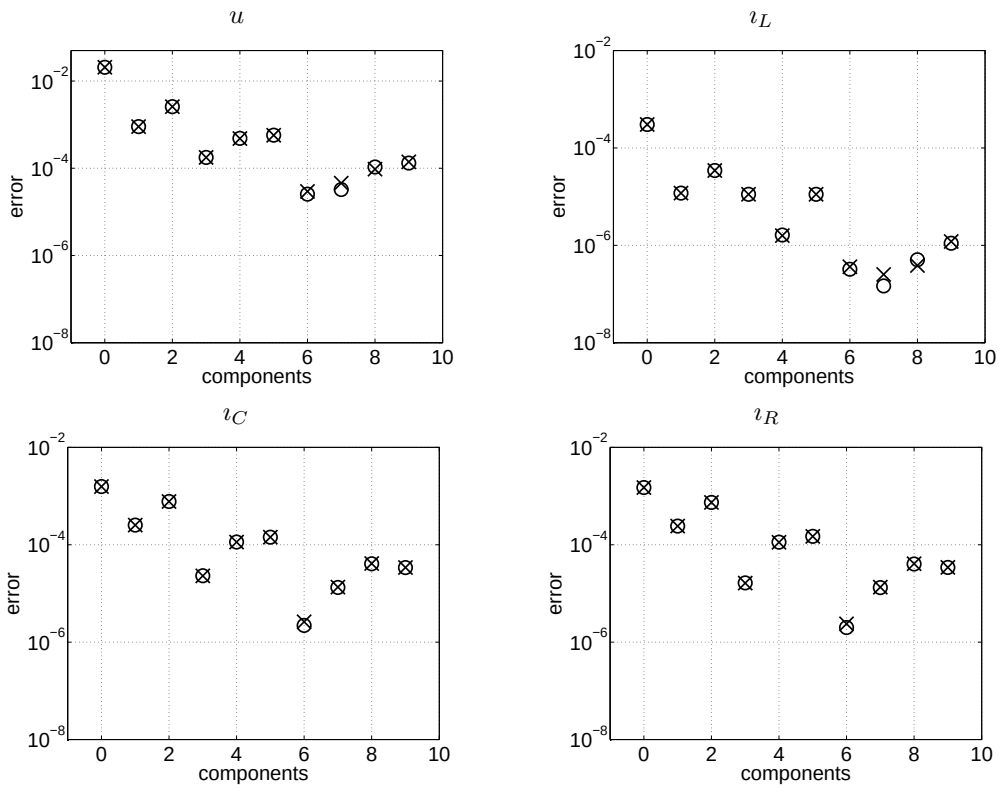

FIG. 12: Maximum differences in reference solution versus st. Galerkin (circles) and reference solution versus st. collocation (crosses) for coefficient functions of gPC expansion for the solution of the nonlinear oscillator—semilog. scale.

The node voltage u_1 represents the output signal. The index of the system (36) is one in case of physically reasonable parameters. We apply the constants

$$U_{\mathrm{cc}} = U_{\mathrm{ss}} = 10 \text{ V},\ U_{\mathrm{op}} = 5 \text{ V},\ R_L = 2 \times 10^3\ \Omega,\ R_B = 15 \times 10^3\ \Omega,$$
$$C = 5 \times 10^{-10} \text{ F},\ L = 2 \times 10^{-3} \text{ H},\ I_0 = 10^{-8} \text{ A},\ \delta = 40 \text{ V}^{-1},\ \alpha = 0.99.$$

The input signals are chosen as harmonic oscillations

$$U_{\mathrm{in1}}(t) = A_1 \cos\left(\frac{2\pi}{T_1} t\right), \qquad U_{\mathrm{in2}}(t) = A_2 \cos\left(\frac{2\pi}{T_2} t\right)$$

with amplitudes $A_1 = 4$ V, $A_2 = 0.1$ V and periods $T_1 = 10^{-4}$ s, $T_2 = 10^{-5}$ s.

Now we replace the operating voltage by a random parameter

$$\widetilde{U}_{\mathrm{op}}(p) := U_{\mathrm{op}}(1 + 0.1p)$$

with a uniformly distributed random variable $p \in [-1, 1]$. This random parameter appears in the algebraic part of (36) only. For the stochastic Galerkin method, the index-1 property does not follow directly from the sufficient criteria in Section 3 due to the sophisticated structure of the right-hand side of the system (36). Nevertheless, the regularity of the corresponding Jacobian matrix $\mathcal{G}$ can be confirmed during a time integration by observing the condition number.

In the gPC expansions, we apply the Legendre polynomials up to degree 5. An initial value problem is considered in the time interval $[t_0, t_1] = [0 \text{ s}, 2{\times}10^{-4} \text{ s}]$, where a constant choice of the initial condition is used. The BDF method of second order discretizes the involved DAEs with an equidistant step size $\Delta t = 1.25 \times 10^{-8}$ s. In the stochastic collocation as well as the stochastic Galerkin approach, we approximate the probabilistic integrals by Gauss-Legendre quadrature on a grid of size 6×6.

Both stochastic collocation and stochastic Galerkin method produce appropriate approximations. The computational times of the stochastic collocation and the stochastic Galerkin method were 16.5 and 32.6, respectively. Each time step of the Galerkin technique is about twice as expensive as a step of the stochastic collocation method. Figures 13 and 14 depict the expected values and the standard deviations of the random processes obtained from the stochastic collocation technique.

Again, we calculate a reference solution by the stochastic collocation, including the midpoint rule with $K = 100$ equidistant nodes. The time integration applies half the step size as in the other simulations. The corresponding maximum differences between the solutions of the different methods are illustrated by Fig. 15. It follows that the accuracy coincides in the stochastic collocation and the stochastic Galerkin technique. Thus, the stochastic collocation is also more efficient for this test example.

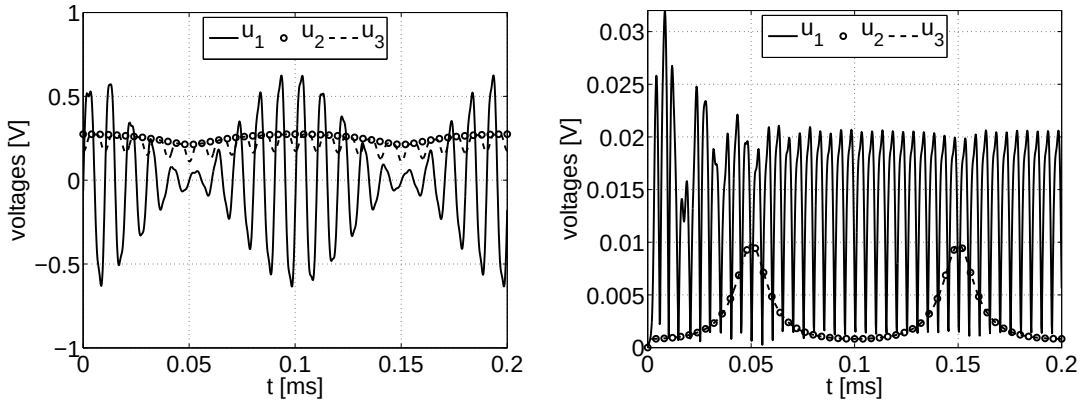

FIG. 13: Expected values (left) and standard deviation (right) of node voltages in transistor modulator.

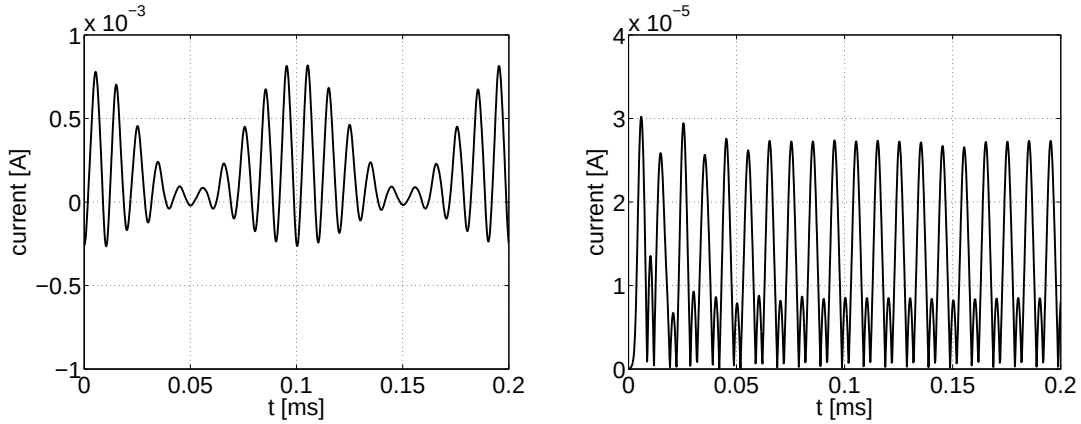

FIG. 14: Expected values (left) and standard deviation (right) of branch current $\imath$ in transistor modulator.

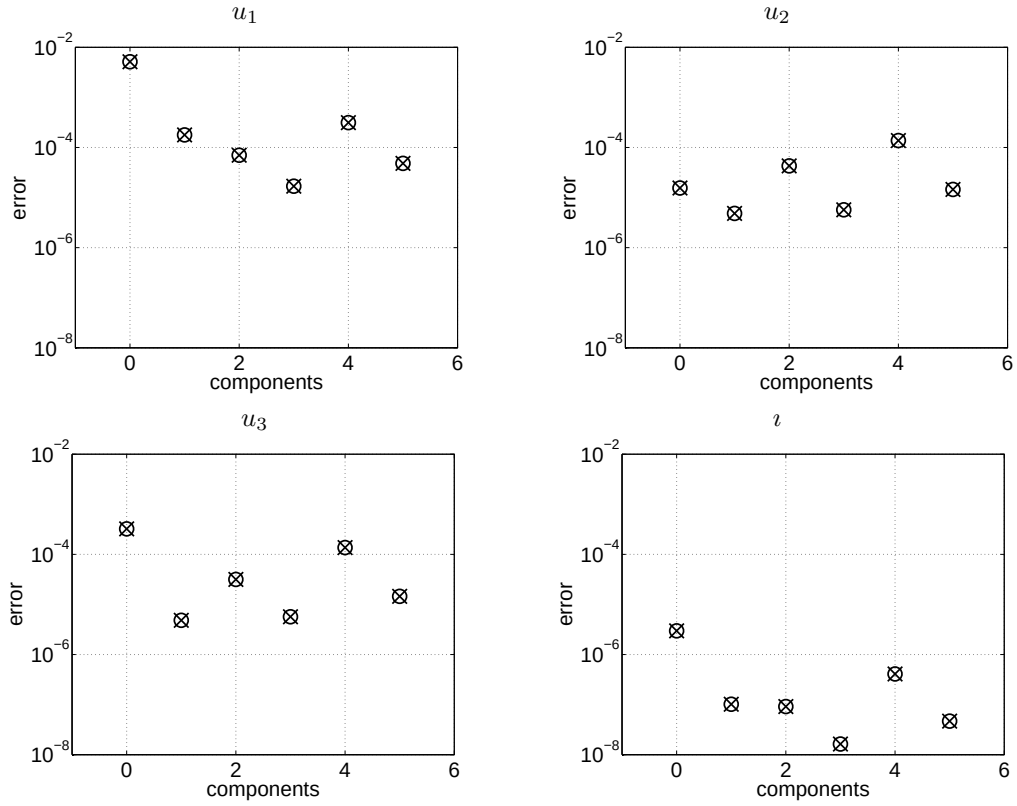

FIG. 15: Maximum differences in reference solution versus st. Galerkin (circles) and reference solution versus st. collocation (crosses) for coefficient functions of gPC expansion for the solution of the transistor modulator—semilog. scale.

5. CONCLUSIONS

The approach of the generalized polynomial chaos has been applied to semiexplicit systems of differential algebraic equations with index 1, which include random parameters. Either a stochastic collocation technique or a stochastic

Galerkin method can be used to compute the unknown coefficient functions of the expansions. The presented results indicate that a stochastic collocation should be preferred over the stochastic Galerkin approach. First, the theoretical investigations show that the index of the larger coupled system from the Galerkin method can increase in comparison to the original systems of index 1. Several sufficient conditions have been proven, which guarantee the index-1 property of the coupled system. Second, the numerical simulations illustrate that the Galerkin method requires a larger computational work due to the linear algebra part, whereas the accuracy is nearly the same in both techniques. An exception is given by linear time-invariant systems of differential algebraic equations, where probabilistic integrals must be calculated just once prior to the time integration in the stochastic Galerkin technique. A numerical simulation confirmed that the stochastic Galerkin method is more efficient in this case. For nonlinear problems, the differences with respect to the efficiency may also become small if the right-hand sides of the original systems are expensive to evaluate. In this case, the computational effort of the linear algebra part is negligible. We expect a similar behavior in both the theoretical properties and the numerical simulations for general systems of differential algebraic equations with a possibly higher index.

REFERENCES

1. Eich-Soellner, E. and Führer, C., *Numerical Methods in Multibody Dynamics*, Teubner, Stuttgart, 1998.
2. Günther, M. and Feldmann, U., CAD based electric circuit modeling in industry I: Mathematical structure and index of network equations, *Surv. Math. Ind.*, 8:97–129, 1999.
3. Kampowsky, W., Rentrop, P., and Schmitt. W., Classification and numerical simulation of electric circuits, *Surv. Math. Ind.*, 2:23–65, 1992.
4. Brenan, K. E., Campbell, S. L., and Petzold, L. R., *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations*, SIAM, Philadelphia, 1996.
5. Hairer, E. and Wanner, G., *Solving Ordinary Differential Equations. Vol. 2: Stiff and Differential-Algebraic Equations*, 2nd ed., Springer, Berlin, 1996.
6. Augustin, F., Gilg, A., Paffrath, M., Rentrop, P., and Wever, U., Polynomial chaos for the approximation of uncertainties: Chances and limits, *Eur. J. Appl. Math.*, 19:149–190, 2008.
7. Ghanem, R. G. and Spanos, P., *Stochastic Finite Elements: A Spectral Approach*, Springer, New York, 1991.
8. Xiu, D., *Numerical Methods for Stochastic Computations: A Spectral Method Approach*, Princeton University Press, Princeton, 2010.
9. Xiu, D. and Hesthaven, J. S., High order collocation methods for differential equations with random inputs, *SIAM J. Sci. Comput.*, 27(3):1118–1139, 2005.
10. Xiu, D., Fast numerical methods for stochastic computations: A review, *Comm. Comput. Phys.*, 5:242–272, 2009.
11. Pulch, R., Polynomial chaos for analysing periodic processes of differential algebraic equations with random parameters, *Proc. Appl. Math. Mech.*, 8:10069–10072, 2008.
12. Pulch, R., Polynomial chaos for the computation of failure probabilities in periodic problems, Scientific Computing in Electrical Engineering, Roos, J. and Costa L. (eds.), Mathematics in Industry Vol. 14, Springer, Berlin, pp. 191–198, 2010.
13. Pulch, R., Modelling and simulation of autonomous oscillators with random parameters, *Math. Comput. Simulat.*, 81:1128–1143, 2011.
14. Pulch, R., Polynomial chaos for linear differential algebraic equations with random parameters, *Int. J. Uncertainty Quantification*, 1(3):223–240, 2011.
15. Campbell, S. L. and Gear, C. W., The index of general nonlinear DAEs, *Numer. Math.*, 72:173–196, 1995.
16. Xiu, D. and Karniadakis, G. E., The Wiener-Askey polynomial chaos for stochastic differential equations, *SIAM J. Sci. Comput.*, 24(2):619–644, 2002.
17. Li, J. and Xiu, D., Evaluation of failure probability via surrogate models, *J. Comput. Phys.*, 229:8966–8980, 2010.
18. Loeven, A., Witteveen, J., and Bijl, H., Probabilistic collocation: An efficient non-intrusive approach for arbitrarily distributed parametric uncertainties, *Proceedings of 45th AIAA Aerospace Sciences Meeting and Exhibit*, Bragg, M. B. (ed.), pp. 1–14, 2007.

19. Sonday, B., Berry, R., Debusschere, B., and Najm, H., Eigenvalues of the Jacobian of a Galerkin-projected uncertain ODE system, *SIAM J. Sci. Comput.*, 33(3):1212–1233, 2011.
20. Chua, L. O. and Ushida, A., Algorithms for computing almost periodic steady-state response of nonlinear systems to multiple input frequencies, *IEEE Trans. Circuits Syst.*, 28(10):953–971, 1981.