Week 1: Stochastic Galerkin Method

Dhyan Laad
2024ADPS0875G

Preliminaries

Galerkin methods are a family of numerical techniques that are used to approximate solutions of continuous operator problems (such as differential equations, commonly partial differential equations) by converting them into algebraic systems that are easier to work with. The stochastic Galerkin method described here is used to remove uncertaintly introduced by a random variable in a fractional initial value problem of the form

$$\mathrm{D}^{\alpha}_{0,t}\, y(t,\xi) = a(\xi)y(t,\xi), \quad y(0,\xi) = y_0(\xi), \tag{1}$$

where $a : \Xi \to \mathbb{R}$ is a measurable function that depends on the parameter ξ defined on some subset Ξ of $\mathbb{R}$, which will be modelled as a random variable $\xi : \Omega \to \Xi$ on some probability space $(\Omega, \mathcal{F}, P)$.

In the case of the operator being the Caputo derivative, the solution to (1) is given by

$$y(t,\xi) = y_0(\xi)E_{\alpha}(a(\xi)t^{\alpha}) \tag{2}$$

where E_{α} is the one-parameter family of Mittag-Leffler functions:

$$E_{\alpha}(z) = \sum_{k=0}^{\infty} \frac{z^k}{\Gamma(\alpha k + 1)}.$$

Let the density of the probability measure P be ρ. We are interested in expressing (2) with a basis in the associated Hilbert space:

$$\mathcal{L}_2(\Xi, \rho) = \{f : \Xi \to \mathbb{R} : f \text{ is measurable and } \mathrm{E}(f^2) < \infty\}.$$

An orthnormal basis $\{\Phi_i\}$ of polynomials $\Phi_i : \Xi \to \mathbb{R}$ for $i \in \mathbb{N}_0$ can be constructed with the Gram-Schmidt process. Let $\Phi_0(x) = 1$, and $\deg(\Phi_i) = i$. It is possible to find a set of such orthogonal polynomials for a number of well-known probability distributions, although this is not always possible, and we continue under the assumption that the distribution may be modelled with such as a basis.

The representation of a function $f \in \mathcal{L}_2(\Xi, \rho)$ is called its general polynomial chaos (gPC) expansion:

$$f(\xi) = \sum_{i=0}^{\infty} f_i \Phi_i(\xi)$$

where $f_i = \langle f, \Phi_i \rangle$ are real coefficients. Applying this to (2), its gPC expansion would be

$$y(t,\xi) = \sum_{i=0}^{\infty} \hat{v}_i(t)\Phi_i(\xi), \tag{3}$$

with coefficient functions $\hat{v}_i : [0,\infty) \to \mathbb{R}$. Truncating (3) yields a finite approximation:

$$\hat{y}^{(n)}(t,\xi) = \sum_{i=0}^{n} \hat{v}_i(t)\Phi_i(\xi).$$

It holds that

$$\lim_{n\to\infty} \left\| y(t,\cdot) - \hat{y}^{(n)}(t,\cdot) \right\| = 0$$

on the Hilbert space for all $t \geq 0$.

Stochastic Galerkin Method

We now present the stochastic Galerkin method. The core idea is to approximate the unknown coefficient functions in the gPC expansion of (2). Start by defining a truncated approximation:

$$\tilde{y}^{(n)}(t,\xi) = \sum_{i=0}^{n} v_i(t)\Phi_i(\xi) \tag{4}$$

where each v_i approximates $\hat{v}_i$. Now inserting (4) into (1) would yield a residual function, which encapsulates the error. To minimize this error, we require the residual to be orthogonal to the subspace spanned by the basis $\{\Phi_0, \Phi_1, \ldots, \Phi_n\}$ in the Hilbert space. Mathematically, the inner product of the residual and each basis polynomial must evaluate to 0. This inner product has the effect of removing the stochastic component ξ from the system, and results in a coupled linear system of deterministic FDEs:

$$\mathrm{D}^{\alpha}_{0,t}\, \mathbf{v}(t) = A\mathbf{v}(t) \tag{5}$$

where $\mathbf{v}(t)$ is the vector of unknown approximated coefficients: $(v_0(t), v_1(t), \ldots, v_n(t))^{\top}$ and A is a symmetric matrix where the (i,j)th entry a_{ij} is $\langle a(\xi)\Phi_i, \Phi_j \rangle$. The system (5) can now be solved numerically with appropriate initial values.

Convergence Analysis

We impose three assumptions before analyzing the convergence of the approximations on an interval $[0,T]$ for $T > 0$.

A1. For order $\alpha \in (0,1)$, there is a fractional derivative with respect to $\mathrm{D}^{\alpha}_{0,t}$ for each term of the series (3). Furthermore,

$$\sum_{i=0}^{\infty} \hat{v}_i(t)\Phi_t(\xi) \quad \text{and} \quad \sum_{i=0}^{\infty} (\mathrm{D}^{\alpha}_{0,t}\, \hat{v}_i(t))\Phi_i(\xi)$$

are uniformly convergent on $[0,T]$ for every $\xi \in \Xi$.

A2. The function a as described in (1) is essentially bounded.

A3. In the series (3), the coefficient functions $\hat{v}_i(t)$ satisfy

$$\lim_{n\to\infty}(n+1)\max_{t\in[0,T]}\sum_{i=n+1}^{\infty}|\hat{v}_i(t)|=0.$$

The last axiom is to ensure the convergence of the series is at least greater than that of the harmonic series.

Now for some preliminary results.

Lemma 1. *If the function $a:\Xi\to\mathbb{R}$ is measurable and essentially bounded by a constant $C_a>0$, then the entries of the matrix A from (5): a_{ij} satisfy $|a_{ij}|\leq C_a$ for all $i,j\in 1:n$.*

Lemma 2. *If the function $a:\Xi\to\mathbb{R}$ is measurable and essentially bounded by $a_{min}\leq a\leq a_{max}$ where a_{min} and a_{max} are constants, then each eigenvalue λ of the matrix A as described in (5) satisfies $\lambda\in[a_{min},a_{max}]$.*

What follows is an outline of the logic of the proof to the central theorem.

Theorem 3. *Consider the linear FDE (1) with the Caputo derivative operator. Let $y(t,\xi)$ be the solution to the equation whose exact gPC expansion is (3), and $\tilde{y}^{(n)}(t,\xi)$ be the $(n+1)$th term gPC solution (4) generated by the Galerkin system (5). If the assumptions A1, A2, and A3 hold on an interval $[0,T]$ where $T>0$, then*

$$\lim_{n\to\infty}\max_{t\in[0,T]}\left\|y(t,\cdot)-\tilde{y}^{(n)}(t,\cdot)\right\|_{\mathcal{L}_2(\Xi,\rho)}=0.$$

The error is firstly split in two:

$$\left\|y(t,\cdot)-\tilde{y}^{(n)}(t,\cdot)\right\|_{\mathcal{L}_2(\Xi,\rho)}\leq\left\|y(t,\cdot)-\hat{y}^{(n)}(t,\cdot)\right\|_{\mathcal{L}_2(\Xi,\rho)}+\left\|\hat{y}^{(n)}(t,\cdot)-\tilde{y}^{(n)}(t,\cdot)\right\|_{\mathcal{L}_2(\Xi,\rho)}.$$

Since the first term on the right has been shown to converge to 0 by construction, the problem is now reduced to showing that

$$\lim_{n\to\infty}\left\|\hat{y}^{(n)}(t,\cdot)-\tilde{y}^{(n)}(t,\cdot)\right\|_{\mathcal{L}_2(\Xi,\rho)}=0$$

for all $t\in[0,T]$.

Let $\mathbf{v}^{(n)}(t)$ be the vector of unknown approximate coefficient functions as described in (5), and $\hat{\mathbf{v}}^{(n)}(t)$ be the analogous vector of true coefficient functions. Then using Parseval's identity we have

$$\left\|\hat{y}^{(n)}(t,\cdot)-\tilde{y}^{(n)}(t,\cdot)\right\|^2_{\mathcal{L}(\Xi,\rho)}=\left\|\hat{\mathbf{v}}^{(n)}-\mathbf{v}^{(n)}(t)\right\|_2^2=\sum_{i=0}^{n}(\hat{v}_i(t)-v_i(t))^2.$$

Define error functions $e_i(t)=\hat{v}_i(t)-v_i(t)$ for $i\in 0:n$, and $\mathbf{e}^{(n)}(t)=(e_1(t),e_2(t),\ldots,e_n(t))^\top$. Therefore, we must show that

$$\lim_{n\to\infty}\left\|\mathbf{e}^{(n)}(t)\right\|_2=0.$$

From the original differential equation, the series representation of the solution,

$$\mathrm{D}^{\alpha}_{0,t}\, y(t,\xi) = a(\xi)y(t,\xi) \Rightarrow \mathrm{D}^{\alpha}_{0,t}\left(\sum_{k=0}^{\infty} \hat{v}_k(t)\Phi_k(\xi)\right) = a(\xi)\left(\sum_{k=0}^{\infty} \hat{v}_k(t)\Phi_k(\xi)\right).$$

Utilizing A1, we may take the derivative operator inside the summation, and taking the inner product of both sides with an arbitrary basis polynomial Φ_i, we have

$$\mathrm{D}^{\alpha}_{0,t}\, \hat{v}_i = \sum_{k=0}^{\infty} a_{ki}\hat{v}_k(t). \tag{6}$$

Taking the derivative of the error function e_i and substituting (5) and (6) into it:

$$\mathrm{D}^{\alpha}_{0,t}\, e_i(t) = \sum_{k=0}^{n} a_{ik}e_k(t) + \underbrace{\sum_{k=n+1}^{\infty} a_{ik}\hat{v}_k(t)}_{\hat{R}_i(t)}$$

for $i \in 0:n$. Let $\hat{\mathbf{R}}(t) = (\hat{R}_0, \hat{R}_1, \ldots, \hat{R}_n)^{\top}$. Then the above equations may be represented as a system

$$\mathrm{D}^{\alpha}_{0,t}\, \mathbf{e}(t) = A\mathbf{e}(t) + \hat{\mathbf{R}}(t).$$

By the spectral decomposition theorem, there exists an orthogonal matric P and a diagonal matrix Λ of eigenvalues λ_i for $i \in 0:n$ such that $A = P\Lambda P^{\top}$. Define

$$d_i(t) = \sum_{k=0}^{n} p_{ik}e_k(t)$$

for every $i \in 0:n$, where p_{ik} is the (i,k)th term of the matrix P. This can also be represented as a system:

$$\mathbf{d}(t) = P^{\top}\mathbf{e}(t). \tag{7}$$

Applying the derivative operator on both sides gives us

$$\mathrm{D}^{\alpha}_{0,t}\, \mathbf{d}(t) = \Lambda\mathbf{d}(t) + \mathbf{R}(t).$$

where $\mathbf{R}(t) = (R_0(t), R_1(t), \ldots, R_n(t))^{\top}$ is the residual vector for

$$R_i(t) = \sum_{\ell=0}^{n}\sum_{k=n+1}^{\infty} p_{\ell i}a_{\ell k}\hat{v}_k(t).$$

This decouples the equations

$$\mathrm{D}^{\alpha}_{0,t}\, d_i(t) = \lambda_i d_i(t) + R_i(t)$$

for $i \in 0:n$. The initial values are $d_i(0) = 0$, and hence the solution to the decoupled equations are given by

$$d_i(t) = \int_0^t (t-\tau)^{\alpha-1}E_{\alpha,\alpha}(\lambda_i(t-\tau)^{\alpha})R_i(\tau)\, d\tau\,.$$

Using A2 and Young's convolution inequality, it is possible to bound the functions

$$|d_i(t)| \leq C\frac{T^{\alpha+1}}{\alpha}\max_{t\in[0,T]}|R_i(t)|$$

where

$$C = \max\{|E_{\alpha,\alpha}(-C_\alpha T^\alpha)|, |E_{\alpha,\alpha}(C_a T^\alpha)|\}$$

uniformly for all $t \in [0,T]$ and $i \in 0:n$. It is also possible to now bound the residual vector in the infinity norm:

$$\|\mathbf{R}(t)\|_\infty \leq C_a\sqrt{n+1}\max_{t\in[0,T]}\sum_{k=n+1}^{\infty}|\hat{v}_k(t)|$$

uniformly for $t \in [0,T]$.

From (7), we have

$$\|\mathbf{e}(t)\|_2 \leq \|P\|_2\|\mathbf{d}(t)\|_2 \leq CC_a\frac{T^{\alpha+1}}{\alpha}(n+1)\max_{t\in[0,T]}\sum_{k=n+1}^{\infty}|\hat{v}_k(t)|.$$

From A3 we can now conclude that

$$\lim_{n\to\infty}\max_{t\in[0,T]}\|\mathbf{e}(t)\|_2 = 0.$$