Chapter 11

Spectral Expansions

> The mark of a mature, psychologically healthy mind is indeed the ability to live with uncertainty and ambiguity, but only as much as there really is.
>
> Julian Baggini

This chapter and its sequels consider several *spectral methods* for uncertainty quantification. At their core, these are orthogonal decomposition methods in which a random variable stochastic process (usually the solution of interest) over a probability space $(\Theta, \mathscr{F}, \mu)$ is expanded with respect to an appropriate orthogonal basis of $L^2(\Theta, \mu; \mathbb{R})$. This chapter lays the foundations by considering spectral expansions in general, starting with the *Karhunen–Loève biorthogonal decomposition*, and continuing with orthogonal polynomial bases for $L^2(\Theta, \mu; \mathbb{R})$ and the resulting *polynomial chaos decompositions*.

11.1 Karhunen–Loève Expansions

Fix a compact domain $\Omega \subseteq \mathbb{R}^d$ (which could be thought of as 'space', 'time', or a general parameter space) and a probability space $(\Theta, \mathscr{F}, \mu)$. The Karhunen–Loève expansion of a square-integrable stochastic process $U \colon \Omega \times \Theta \to \mathbb{R}$ is a particularly nice spectral decomposition, in that it decomposes U in a *biorthogonal* fashion, i.e. in terms of components that are both orthogonal over the parameter domain Ω and the probability space Θ.

To be more precise, consider a stochastic process $U \colon \Omega \times \Theta \to \mathbb{R}$ such that

- for all $x \in \Omega$, $U(x) \in L^2(\Theta, \mu; \mathbb{R})$;
- for all $x \in \Omega$, $\mathbb{E}_\mu[U(x)] = 0$;
- the *covariance function* $C_U(x, y) := \mathbb{E}_\mu[U(x)U(y)]$ is a well-defined continuous function of $x, y \in \Omega$.

Remark 11.1.

1. The condition that U is a zero-mean process is not a serious restriction; if U is not a zero-mean process, then simply consider $\tilde{U}$ defined by $\tilde{U}(x, \theta) := U(x, \theta) - \mathbb{E}_\mu[U(x)]$.
2. It is common in practice to see the covariance function interpreted as proving some information on the *correlation length* of the process U. That is, $C_U(x, y)$ depends only upon $\|x - y\|$ and, for some function $g \colon [0, \infty) \to [0, \infty)$, $C_U(x, y) = g(\|x - y\|)$. A typical such g is $g(r) = \exp(-r/r_0)$, and the constant r_0 encodes how similar values of U at nearby points of Ω are expected to be; when the correlation length r_0 is small, the field U has dissimilar values near to one another, and so is rough; when r_0 is large, the field U has only similar values near to one another, and so is more smooth.

Define the *covariance operator* of U, also denoted by $C_U \colon L^2(\Omega, \mathrm{d}x; \mathbb{R}) \to L^2(\Omega, \mathrm{d}x; \mathbb{R})$ by

$$(C_U f)(x) := \int_\Omega C_U(x, y) f(y)\, \mathrm{d}y.$$

Now let $\{\psi_n \mid n \in \mathbb{N}\}$ be an orthonormal basis of eigenvectors of $L^2(\Omega, \mathrm{d}x; \mathbb{R})$ with corresponding eigenvalues $\{\lambda_n \mid n \in \mathbb{N}\}$, i.e.

$$\int_\Omega C_U(x, y)\psi_n(y)\, \mathrm{d}y = \lambda_n \psi_n(x)$$

and

$$\int_\Omega \psi_m(x)\psi_n(x)\, \mathrm{d}x = \delta_{mn}.$$

Definition 11.2. Let $\mathcal{X}$ be a first-countable topological space. A function $K \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a *Mercer kernel* if

1. K is continuous;
2. K is symmetric, i.e. $K(x, x') = K(x', x)$ for all $x, x' \in \mathcal{X}$; and
3. K is positive semi-definite in the sense that, for all choices of finitely many points $x_1, \ldots, x_n \in \mathcal{X}$, the *Gram matrix*
$$G := \begin{bmatrix} K(x_1, x_1) & \cdots & K(x_1, x_n) \\ \vdots & \ddots & \vdots \\ K(x_n, x_1) & \cdots & K(x_n, x_n) \end{bmatrix}$$
is positive semi-definite, i.e. satisfies $\xi \cdot G\xi \geq 0$ for all $\xi \in \mathbb{R}^n$.

Theorem 11.3 (Mercer). *Let $\mathcal{X}$ be a first-countable topological space equipped with a complete Borel measure μ. Let $K \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a Mercer kernel. If $x \mapsto K(x, x)$ lies in $L^1(\mathcal{X}, \mu; \mathbb{R})$, then there is an orthonormal basis $\{\psi_n\}_{n \in \mathbb{N}}$ of $L^2(\mathcal{X}, \mu; \mathbb{R})$ consisting of eigenfunctions of the operator*

$$f \mapsto \int_{\mathcal{X}} K(\,\cdot\,, y) f(y)\, \mathrm{d}\mu(y)$$

with non-negative eigenvalues $\{\lambda_n\}_{n \in \mathbb{N}}$. Furthermore, the eigenfunctions corresponding to non-zero eigenvalues are continuous, and

$$K(x, y) = \sum_{n \in \mathbb{N}} \lambda_n \psi_n(x) \psi_n(y),$$

and this series converges absolutely, and uniformly over compact subsets of $\mathcal{X}$.

Theorem 11.4 (Karhunen–Loève). *Under the above assumptions on U, its covariance function and its covariance operator, U can be written as*

$$U = \sum_{n \in \mathbb{N}} Z_n \psi_n$$

where the $\{\psi_n\}_{n \in \mathbb{N}}$ are orthonormal eigenfunctions of the covariance operator C_U, the corresponding eigenvalues $\{\lambda_n\}_{n \in \mathbb{N}}$ are non-negative, the convergence of the series is in $L^2(\Theta, \mu; \mathbb{R})$ and uniform in $x \in \Omega$, with

$$Z_n = \int_\Omega U(x)\psi_n(x)\, \mathrm{d}x.$$

Furthermore, the random variables Z_n are centred, uncorrelated, and have variance λ_n:

$$\mathbb{E}_\mu[Z_n] = 0, \text{ and } \mathbb{E}_\mu[Z_m Z_n] = \lambda_n \delta_{mn}.$$

Proof. By Exercise 2.1, and since the covariance function C_U is continuous, C_U is a Mercer kernel. Hence, by Mercer's theorem, there is an orthonormal basis $\{\psi_n\}_{n\in\mathbb{N}}$ of $L^2(\Omega, \mathrm{d}x; \mathbb{R})$ consisting of eigenfunctions of the covariance operator with non-negative eigenvalues $\{\lambda_n\}_{n\in\mathbb{N}}$. In this basis, the covariance function has the representation

$$C_U(x, y) = \sum_{n\in\mathbb{N}} \lambda_n \psi_n(x)\psi_n(y).$$

Write the process U in terms of this basis as

$$U = \sum_{n\in\mathbb{N}} Z_n \psi_n,$$

where the coefficients Z_n are random variables given by orthogonal projection:

$$Z_n := \int_\Omega U(x)\psi_n(x)\,\mathrm{d}x.$$

Then

$$\mathbb{E}_\mu[Z_n] = \mathbb{E}_\mu\left[\int_\Omega U(x)\psi_n(x)\,\mathrm{d}x\right] = \int_\Omega \mathbb{E}[U(x)]\psi_n(x)\,\mathrm{d}x = 0.$$

and

$$\begin{aligned}
\mathbb{E}_\mu[Z_m Z_n] &= \mathbb{E}_\mu\left[\int_\Omega U(x)\psi_m(x)\,\mathrm{d}x \int_\Omega U(x)\psi_n(x)\,\mathrm{d}x\right] \\
&= \mathbb{E}_\mu\left[\int_\Omega\int_\Omega U(x)\psi_m(x)U(y)\psi_n(y)\,\mathrm{d}y\mathrm{d}x\right] \\
&= \int_\Omega\int_\Omega \mathbb{E}_\mu[U(x)U(y)]\psi_m(x)\psi_n(y)\,\mathrm{d}y\mathrm{d}x \\
&= \int_\Omega\int_\Omega C_U(x,y)\psi_m(x)\psi_n(y)\,\mathrm{d}y\mathrm{d}x \\
&= \int_\Omega \psi_m(x)\int_\Omega C_U(x,y)\psi_n(y)\,\mathrm{d}y\mathrm{d}x \\
&= \int_\Omega \psi_m(x)\lambda_n\psi_n(x)\,\mathrm{d}x \\
&= \lambda_n\delta_{mn}.
\end{aligned}$$

Let $S_N := \sum_{n=1}^N Z_n\psi_n \colon \Omega\times\Theta\to\mathbb{R}$. Then, for any $x\in\Omega$,

$$\begin{aligned}
&\mathbb{E}_\mu\left[|U(x) - S_N(x)|^2\right] \\
&\quad= \mathbb{E}_\mu[U(x)^2] + \mathbb{E}_\mu[S_N(x)^2] - 2\mathbb{E}_\mu[U(x)S_N(x)] \\
&\quad= C_U(x,x) + \mathbb{E}_\mu\left[\sum_{n=1}^N\sum_{m=1}^N Z_nZ_m\psi_m(x)\psi_n(x)\right] - 2\mathbb{E}_\mu\left[U(x)\sum_{n=1}^N Z_n\psi_n(x)\right] \\
&\quad= C_U(x,x) + \sum_{n=1}^N \lambda_n\psi_n(x)^2 - 2\mathbb{E}_\mu\left[\sum_{n=1}^N\int_\Omega U(x)U(y)\psi_n(y)\psi_n(x)\,\mathrm{d}y\right] \\
&\quad= C_U(x,x) + \sum_{n=1}^N \lambda_n\psi_n(x)^2 - 2\sum_{n=1}^N\int_\Omega C_U(x,y)\psi_n(y)\psi_n(x)\,\mathrm{d}y \\
&\quad= C_U(x,x) - \sum_{n=1}^N \lambda_n\psi_n(x)^2 \\
&\quad\to 0 \text{ as } N\to\infty, \text{ uniformly in } x, \text{ by Mercer's theorem.}
\end{aligned}$$

□

Among many possible decompositions of a random process, the Karhunen–Loève expansion is optimal in the sense that the mean-square error of any truncation of the expansion after finitely many terms is minimal. However, its utility is limited since the covariance function of the solution process is often not known a priori. Nevertheless, the Karhunen–Loève expansion provides an effective means of representing *input* random processes when their covariance structure is known, and provides a simple method for sampling Gaussian measures on Hilbert spaces, which is a necessary step in the implementation of the methods outlined in Chapter 6.

Example 11.5. Suppose that $C \colon \mathcal{H} \to \mathcal{H}$ is a self-adjoint, positive-definite, nuclear operator on a Hilbert space $\mathcal{H}$ and let $m \in \mathcal{H}$. Let $(\lambda_k, \psi_k)_{k \in \mathbb{N}}$ be a sequence of orthonormal eigenpairs for C, ordered by decreasing eigenvalue λ_k. Let $\Xi_1, \Xi_2, \ldots$ be independently distributed according to the standard Gaussian measure $\mathcal{N}(0, 1)$ on $\mathbb{R}$. Then, by the Karhunen–Loève theorem,

$$U := m + \sum_{k=1}^{\infty} \lambda_k^{1/2} \Xi_k \psi_k \tag{11.1}$$

is an $\mathcal{H}$-valued random variable with distribution $\mathcal{N}(m, C)$. Therefore, a finite sum of the form $m + \sum_{k=1}^{K} \lambda_k^{1/2} \Xi_k \psi_k$ for large K is a reasonable approximation to a $\mathcal{N}(m, C)$-distributed random variable; this is the procedure used to generate the sample paths in Figure 11.1.

Note that $\lambda_k^{1/2} \Xi_k$ has Lebesgue density on $\mathbb{R}$ proportional to $\exp(-|\xi_k|^2 / 2\lambda_k)$. Therefore, although Theorem 2.33 shows that the infinite product of Lebesgue measures on $\operatorname{span}\{\psi_k \mid k \in \mathbb{N}\}$ cannot define an infinite-dimensional Lebesgue measure on $\mathcal{H}$, $U - m$ defined by (11.1) may be said to have a 'formal Lebesgue density' proportional to

$$\begin{aligned} \prod_{k \in \mathbb{N}} \exp\left(-\frac{|\xi_k|^2}{2\lambda_k} \right) &= \exp\left(-\frac{1}{2} \sum_{k \in \mathbb{N}} \frac{|\xi_k|^2}{\lambda_k} \right) \\ &= \exp\left(-\frac{1}{2} \sum_{k \in \mathbb{N}} \frac{|\langle u - m, \psi_k \rangle_{\mathcal{H}}|^2}{\lambda_k} \right) \\ &= \exp\left(-\frac{1}{2} \left\| C^{-1/2}(u - m) \right\|_{\mathcal{H}}^2 \right) \end{aligned}$$

by Parseval's theorem and the eigenbasis representation of C. This formal derivation should make it intuitively reasonable that U is a Gaussian random variable on $\mathcal{H}$ with mean m and covariance operator C. For more general sampling schemes of this type, see the later remarks on the sampling of Besov measures.

Principal Component Analysis. As well as being useful for the analysis of random paths, surfaces, and so on, Karhunen–Loève expansions are also useful in the analysis of finite-dimensional random vectors and sample data:

Definition 11.6. A *principal component analysis* of an $\mathbb{R}^N$-valued random vector U is the Karhunen–Loève expansion of U seen as a stochastic process $U \colon \{1, \ldots, N\} \times \Omega \to \mathbb{R}$. It is also known as the *discrete Karhunen–Loève transform*, the *Hotelling transform*, and the *proper orthogonal decomposition*.

Principal component analysis is often applied to sample data, and is intimately related to the singular value decomposition:

Example 11.7. Let $X \in \mathbb{R}^{N \times M}$ be a matrix whose columns are M independent and identically distributed samples from some probability measure on $\mathbb{R}^N$, and assume without loss of generality that the samples have mean zero. The empirical covariance matrix of the samples is

$$\widehat{C} := \tfrac{1}{M^2} X X^{\top}.$$

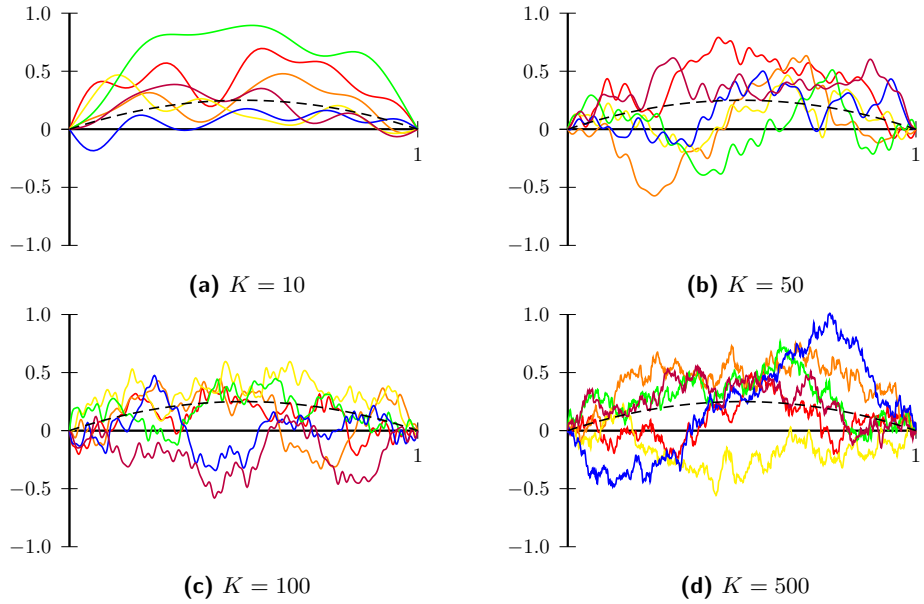


Figure 11.1: Approximate sample paths of the Gaussian distribution on $H_0^1([0,1])$ that has mean path $m(x) = x(1-x)$ and covariance operator $\left(-\frac{\mathrm{d}^2}{\mathrm{d}x^2}\right)^{-1}$. Along with the mean path (dashed), six sample paths are shown for truncated Karhunen–Loève expansions using $K \in \mathbb{N}$ terms.

The eigenvalues λ_n and eigenfunctions ψ_n of the Karhunen–Loève expansion are just the eigenvalues and eigenvectors of this matrix $\widehat{C}$. Let $\Lambda \in \mathbb{R}^{N\times N}$ be the diagonal matrix of the eigenvalues λ_n (which are non-negative, and are assumed to be in decreasing order) and $\Psi \in \mathbb{R}^{N\times N}$ the matrix of corresponding orthonormal eigenvectors, so that $\widehat{C}$ diagonalizes as

$$\widehat{C} = \Psi\Lambda\Psi^\top.$$

The principal component transform of the data X is $W := \Psi^\top X$; this is an orthogonal transformation of $\mathbb{R}^N$ that transforms X to a new coordinate system in which the greatest component-wise variance comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on.

On the other hand, taking the singular value decomposition of the data (normalized by the number of samples) yields

$$\tfrac{1}{M}X = U\Sigma V^\top,$$

where $U \in \mathbb{R}^{N\times N}$ and $V \in \mathbb{R}^{M\times M}$ are orthogonal and $\Sigma \in \mathbb{R}^{N\times M}$ is diagonal with decreasing non-negative diagonal entries (the singular values of $\frac{1}{M}X$). Then

$$\widehat{C} = U\Sigma V^\top (U\Sigma V^\top)^\top = U\Sigma V^\top V\Sigma^\top U^\top = U\Sigma^2 U^\top.$$

from which we see that $U = \Psi$ and $\Sigma^2 = \Lambda$. This is just another instance of the well-known relation that, for any matrix A, the eigenvalues of AA^* are the singular values of A and the right eigenvectors of AA^* are the left singular vectors of A; however, in this context, it also provides an alternative way to compute the principal component transform.

In fact, performing principal component analysis via the singular value decomposition is numerically preferable to forming and then diagonalizing the covariance matrix, since the formation of $XX^\top$ can cause a disastrous loss of precision; the classic example of this phenomenon is the

Läuchli matrix

$$\begin{bmatrix} 1 & \varepsilon & 0 & 0 \\ 1 & 0 & \varepsilon & 0 \\ 1 & 0 & 0 & \varepsilon \end{bmatrix} \quad (0 < \varepsilon \ll 1),$$

for which taking the singular value decomposition (e.g. by bidiagonalization followed by QR iteration) is stable, but forming and diagonalizing $XX^\top$ is unstable.

Karhunen–Loève Sampling of Non-Gaussian Besov Measures. The Karhunen–Loève approach to generating samples from Gaussian measures of known covariance operator, as in Example 11.5, can be extended to more general settings, in which a basis is prescribed a priori and (not necessarily Gaussian) random coefficients with a suitable decay rate are used. The choice of basis elements and the rate of decay of the coefficients together control the smoothness of the sample realizations; the mathematical hard work lies in showing that such random series do indeed converge to a well-defined limit, and thereby define a probability measure on the desired function space.

One method for the construction of function spaces — and hence random functions — of desired smoothness is to use wavelets. Wavelet bases are particularly attractive because they allow for the representation of sharply localized features — e.g. the interface between two media with different material properties — in a way that globally smooth basis functions such as polynomials and the Fourier basis do not. Omitting several technicalities, a wavelet basis of $L^2(\mathbb{R}^d)$ or $L^2(\mathbb{T}^d)$ can be thought of as an orthonormal basis consisting of appropriately scaled and shifted copies of a single basic element that has some self-similarity. By controlling the rate of decay of the coefficients in a wavelet expansion, we obtain a family of function spaces — the *Besov spaces* — with three scales of smoothness, here denoted p, q and s. In what follows, for any function f on $\mathbb{R}^d$ or $\mathbb{T}^d$, define the scaled and shifted version $f_{j,k}$ of f for $j, k \in \mathbb{Z}$ by

$$f_{j,k}(x) := f(2^j x - k). \tag{11.2}$$

The starting point of a wavelet construction is a *scaling function* (also known as the *averaging function* or *father wavelet*) $\widetilde{\phi}\colon \mathbb{R} \to \mathbb{R}$ and a family of closed subspaces $\mathcal{V}_j \subseteq L^2(\mathbb{R})$, $j \in \mathbb{Z}$, called a *multiresolution analysis* of $L^2(\mathbb{R})$, satisfying

1. (nesting) for all $j \in \mathbb{Z}$, $\mathcal{V}_j \subseteq \mathcal{V}_{j+1}$;
2. (density and zero intersection) $\overline{\bigcup_{j\in\mathbb{Z}} \mathcal{V}_j} = L^2(\mathbb{R})$ and $\bigcap_{j\in\mathbb{Z}} \mathcal{V}_j = \{0\}$;
3. (scaling) for all $j, k \in \mathbb{Z}$, $f \in \mathcal{V}_0 \iff f_{j,k} \in \mathcal{V}_j$;
4. (translates of $\widetilde{\phi}$ generate $\mathcal{V}_0$) $\mathcal{V}_0 = \operatorname{span}\{\widetilde{\phi}_{0,k} \mid k \in \mathbb{Z}\}$;
5. (Riesz basis) there are finite positive constants A and B such that, for all sequences $(c_k)_{k\in\mathbb{Z}} \in \ell^2(\mathbb{Z})$,

$$A\|(c_k)\|_{\ell^2(\mathbb{Z})} \leq \left\| \sum_{k\in\mathbb{Z}} c_k \widetilde{\phi}_{0,k} \right\|_{L^2(\mathbb{R})} \leq B\|(c_k)\|_{\ell^2(\mathbb{Z})}.$$

Given such a scaling function $\widetilde{\phi}\colon \mathbb{R} \to \mathbb{R}$, the associated *mother wavelet* $\widetilde{\psi}\colon \mathbb{R} \to \mathbb{R}$ is defined as follows:

$$\begin{aligned} \text{if } \widetilde{\phi}(x) &= \sum_{k\in\mathbb{Z}} c_k \widetilde{\phi}(2x - k), \\ \text{then } \widetilde{\psi}(x) &= \sum_{k\in\mathbb{Z}} (-1)^k c_{k+1} \widetilde{\phi}(2x + k). \end{aligned}$$

It is the scaled and shifted copies of the mother wavelet $\widetilde{\psi}$ that will form the desired orthonormal basis of L^2.

Example 11.8. 1. The indicator function $\widetilde{\phi} = \mathbb{1}_{[0,1)}$ satisfies the self-similarity relation $\widetilde{\phi}(x) =$

$\widetilde{\phi}(2x) + \widetilde{\phi}(2x-1)$; the associated $\widetilde{\psi}$ given by

$$\widetilde{\psi}(x) = \widetilde{\phi}(2x) - \widetilde{\phi}(2x-1) = \begin{cases} 1, & \text{if } 0 \le x < \frac{1}{2}, \\ -1, & \text{if } \frac{1}{2} \le x < 1, \\ 0, & \text{otherwise.} \end{cases}$$

is called the *Haar wavelet*.

2. The B-spline scaling functions σ_r, $r \in \mathbb{N}_0$, are piecewise polynomial of degree r and globally $\mathcal{C}^{r-1}$, and are defined recursively by convolution:

$$\sigma_r := \begin{cases} \mathbb{1}_{[0,1)}, & \text{for } r = 0, \\ \sigma_{r-1} \star \sigma_0, & \text{for } r \in \mathbb{N}, \end{cases} \tag{11.3}$$

where

$$(f \star g)(x) := \int_{\mathbb{R}} f(y) g(x-y) \, \mathrm{d}y.$$

Here, the presentation focusses on Besov spaces of 1-periodic functions, i.e. functions on the unit circle $\mathbb{T} := \mathbb{R}/\mathbb{Z}$, and on the d-dimensional unit torus $\mathbb{T}^d := \mathbb{R}^d/\mathbb{Z}^d$. To this end, set

$$\phi(x) := \sum_{s \in \mathbb{Z}} \widetilde{\phi}(x+s) \quad \text{and} \quad \psi(x) := \sum_{s \in \mathbb{Z}} \widetilde{\psi}(x+s).$$

Scaled and translated versions of these functions are defined as usual by (11.2). Note that in the toroidal case the spaces $\mathcal{V}_j$ for $j < 0$ consist of constant functions, and that, for each scale $j \in \mathbb{N}_0$, $\phi \in \mathcal{V}_0$ has only 2^j distinct scaled translates $\phi_{j,k} \in \mathcal{V}_j$, i.e. those with $k = 0, \dots, 2^{j-1}$. Let

$$\mathcal{V}_j := \operatorname{span}\{\phi_{j,k} \mid k = 0, \dots, 2^j - 1\},$$
$$\mathcal{W}_j := \operatorname{span}\{\psi_{j,k} \mid k = 0, \dots, 2^j - 1\},$$

so that $\mathcal{W}_j$ is the orthogonal complement of $\mathcal{V}_j$ in $\mathcal{V}_{j+1}$ and

$$L^2(\mathbb{T}) = \overline{\bigcup_{j \in \mathbb{N}_0} \mathcal{V}_j} = \bigoplus_{j \in \mathbb{N}_0} \mathcal{W}_j$$

Indeed, if ψ has unit norm, then $2^{j/2}\psi_{j,k}$ also has unit norm, and

$$\{2^{j/2}\psi_{j,k} \mid k = 0, \dots, 2^j - 1\} \text{is an orthonormal basis of } \mathcal{W}_j, \text{ and}$$
$$\{2^{j/2}\psi_{j,k} \mid j \in \mathbb{N}_0, k = 0, \dots, 2^j - 1\} \text{ is an orthonormal basis of } L^2(\mathbb{T}),$$

a so-called *wavelet basis*.

To construct an analogous wavelet basis of $L^2(\mathbb{T}^d)$ for $d \ge 1$, proceed as follows: for $\nu \in \{0,1\}^d \setminus \{(0, \dots, 0)\}$, $j \in \mathbb{N}_0$, and $k \in \{0, \dots, 2^j - 1\}^d$, define the scaled and translated wavelet $\psi^\nu_{j,k} \colon \mathbb{T}^d \to \mathbb{R}$ by

$$\psi^\nu_{j,k}(x) := 2^{dj/2} \psi^{\nu_1}(2^j x_1 - k_1) \cdots \psi^{\nu_d}(2^j x_d - k_d)$$

where $\psi^0 = \phi$ and $\psi^1 = \psi$. The system

$$\left\{ \psi^\nu_{j,k} \,\middle|\, j \in \mathbb{N}_0, k \in \{0, \dots, 2^j - 1\}^d, \nu \in \{0,1\}^d \setminus \{(0, \dots, 0)\} \right\}$$

is an orthonormal wavelet basis of $L^2(\mathbb{T}^d)$.

The *Besov space* $B^s_{pq}(\mathbb{T}^d)$ can be characterized in terms of the summability of wavelet coefficients at the various scales:

Definition 11.9. Let $1 \leq p, q < \infty$ and let $s > 0$. The *Besov* (p, q, s) *norm* of a function $u = \sum_{j,k,\nu} u^\nu_{j,k} \psi^\nu_{j,k} \colon \mathbb{T}^d \to \mathbb{R}$ is defined by

$$\left\| \sum_{j \in \mathbb{N}_0} \sum_{\nu,k} u^\nu_{j,k} \psi^\nu_{j,k} \right\|_{B^s_{pq}(\mathbb{T}^d)} := \left\| j \mapsto 2^{js} 2^{jd(\frac{1}{2} - \frac{1}{p})} \left\| (k, \nu) \mapsto u^\nu_{j,k} \right\|_{\ell^p} \right\|_{\ell^q(\mathbb{N}_0)}$$
$$:= \left(\sum_{j \in \mathbb{N}_0} 2^{qjs} 2^{qjd(\frac{1}{2} - \frac{1}{p})} \left(\sum_{\nu,k} |u^\nu_{j,k}|^p \right)^{q/p} \right)^{1/q},$$

and the *Besov space* $B^s_{pq}(\mathbb{T}^d)$ is the space of functions for which this norm is finite.

Note that at each scale j, there are $(2^d - 1)2^{jd} = 2^{(j+1)d} - 2^{jd}$ wavelet coefficients. The indices j, k and ν can be combined into a single index $\ell \in \mathbb{N}$. First, $\ell = 1$ corresponds to the scaling function $\phi(x_1) \cdots \phi(x_d)$. The remaining numbering is done scale by scale; that is, we first number wavelets with $j = 0$, then wavelets with $j = 1$, and so on. Within each scale $j \in \mathbb{N}_0$, the $2^d - 1$ indices ν are ordered by thinking them as binary representation of integers, and an ordering of the 2^{jd} translations k can be chosen arbitrarily. With this renumbering,

$$\sum_{\ell=1}^{\infty} c_\ell \psi_\ell \in B^s_{pq}(\mathbb{T}^d) \iff 2^{js} 2^{jd(\frac{1}{2} - \frac{1}{p})} \left(\sum_{\ell=2^{jd}}^{2^{(j+1)d}-1} |c_\ell|^p \right)^{1/p} \in \ell^q(\mathbb{N}_0)$$

For $p = q$, since at scale j it holds that $2^{jd} \leq \ell < 2^{(j+1)d}$, an equivalent norm for $B^s_{pp}(\mathbb{T}^d)$ is

$$\left\| \sum_{\ell \in \mathbb{N}} u_\ell \psi_\ell \right\|_{B^s_{pp}(\mathbb{T}^d)} \simeq \left\| \sum_{\ell \in \mathbb{N}} u_\ell \psi_\ell \right\|_{X^{s,p}} := \left(\sum_{\ell=1}^{\infty} \ell^{(ps/d + p/2 - 1)} |u_\ell|^p \right)^{1/p};$$

in particular if the original scaling function and mother wavelet are r times differentiable with $r > s$, then B^s_{22} coincides with the Sobolev space H^s. This leads to a Karhunen–Loève-type sampling procedure for $B^s_{pp}(\mathbb{T}^d)$, as in Example 11.5: U defined by

$$U := \sum_{\ell \in \mathbb{N}} \ell^{-(\frac{s}{d} + \frac{1}{2} - \frac{1}{p})} \kappa^{-\frac{1}{p}} \Xi_\ell \psi_\ell, \tag{11.4}$$

where Ξ_ℓ are sampled independently and identically from the generalized Gaussian measure on $\mathbb{R}$ with Lebesgue density proportional to $\exp(-\frac{1}{2}|\xi_\ell|^p)$, can be said to have 'formal Lebesgue density' proportional to $\exp(-\frac{\kappa}{2}\|u\|^p_{B^s_{pp}})$, and is therefore a natural candidate for a 'typical' element of the Besov space $B^s_{pp}(\mathbb{T}^d)$. More generally, given *any* orthonormal basis $\{\psi_k \mid k \in \mathbb{N}\}$ of some Hilbert space, one can define a Banach subspace $X^{s,p}$ with norm

$$\left\| \sum_{\ell \in \mathbb{N}} u_\ell \psi_\ell \right\|_{X^{s,p}} := \left(\sum_{\ell=1}^{\infty} \ell^{(ps/d + p/2 - 1)} |u_\ell|^p \right)^{1/p}$$

and define a *Besov distributed random variable* U by (11.4).

It remains, however, to check that (11.4) not only defines a measure, but that it assigns unit probability mass to the Besov space from which it is desired to draw samples. It turns out that the question of whether or not $U \in X^{s,p}$ with probability one is closely related to having a Fernique theorem (q.v. Theorem 2.42) for Besov measures:

Theorem 11.10. *Let U be defined as in (11.4), with $1 \leq p < \infty$ and $s > 0$. Then*

$$\|U\|_{X^{t,p}} < \infty \text{ almost surely} \iff \mathbb{E}[\exp(\alpha \|U\|^p_{X^{t,p}})] < \infty \text{ for all } \alpha \in (0, \tfrac{\kappa}{2})$$
$$\iff t < s - \tfrac{d}{p}$$

Furthermore, for $p \geq 1$, $s > \frac{d}{p}$, and $t < s - \frac{d}{p}$, there is a constant r^* depending only on p, d, s, and t such that, for all $\alpha \in (0, \frac{\kappa}{2r^*})$,

$$\mathbb{E}[\exp(\alpha \|U\|_{\mathcal{C}^t})] < \infty.$$

11.2 Wiener–Hermite Polynomial Chaos

The next section will cover polynomial chaos (PC) expansions in greater generality, and this section serves as an introductory prelude. In this, the classical and notationally simplest setting, we consider expansions of a real-valued random variable U with respect to a single standard Gaussian random variable Ξ, using appropriate orthogonal polynomials of Ξ, i.e. the Hermite polynomials. This setting was pioneered by Norbert Wiener, and so it is known as the Wiener–Hermite polynomial chaos. The term 'chaos' is perhaps a bit confusing, and is not related to the use of the term in the study of dynamical systems; its original meaning, as used by Wiener, was something closer to what would nowadays be called a stochastic process:

> "Of all the forms of chaos occurring in physics, there is only one class which has been studied with anything approaching completeness. This is the class of types of chaos connected with the theory of Brownian motion." [180]

Let $\Xi \sim \gamma = \mathcal{N}(0, 1)$ be a standard Gaussian random variable, and let $\mathrm{He}_n(\xi) \in \mathbb{R}[\xi]$, for $n \in \mathbb{N}_0$, be the *Hermite polynomials*, the orthogonal polynomials for the standard Gaussian measure γ with the normalization

$$\int_{\mathbb{R}} \mathrm{He}_m(\xi)\mathrm{He}_n(\xi)\,\mathrm{d}\gamma(\xi) = n!\delta_{mn}.$$

By Weierstrass' theorem (Theorem 8.21) and the approximability of L^2 functions by continuous ones, the Hermite polynomials form a complete orthogonal basis of the Hilbert space $L^2(\mathbb{R}, \gamma; \mathbb{R})$ with the inner product

$$\langle U, V \rangle_{L^2(\gamma)} := \mathbb{E}[U(\Xi)V(\Xi)] \equiv \int_{\mathbb{R}} U(\xi)V(\xi)\,\mathrm{d}\gamma(\xi).$$

Definition 11.11. Let $U \in L^2(\mathbb{R}, \gamma; \mathbb{R})$ be a square-integrable real-valued random variable. The *Wiener–Hermite polynomial chaos expansion* of U with respect to the standard Gaussian Ξ is the expansion of U in the orthogonal basis $\{\mathrm{He}_n\}_{n\in\mathbb{N}_0}$, i.e.

$$U = \sum_{n\in\mathbb{N}_0} u_n \mathrm{He}_n(\Xi)$$

with scalar *Wiener–Hermite polynomial chaos coefficients* $\{u_n\}_{n\in\mathbb{N}_0} \subseteq \mathbb{R}$ given by

$$u_n = \frac{\langle U, \mathrm{He}_n \rangle_{L^2(\gamma)}}{\|\mathrm{He}_n\|^2_{L^2(\gamma)}} = \frac{1}{n!\sqrt{2\pi}} \int_{-\infty}^{\infty} U(\xi)\mathrm{He}_n(\xi)e^{-\xi^2/2}\,\mathrm{d}\xi.$$

Remark 11.12. From the perspective of sampling of random variables, this means that if we wish to draw a sample from the distribution of U and know the Wiener–Hermite coefficients $\{u_n\}_{n\in\mathbb{N}_0}$, it is enough to draw a sample ξ from the standard normal distribution and then evaluate the series $\sum_{n\in\mathbb{N}_0} u_n \mathrm{He}_n(\xi)$ at that ξ.

Note that, in particular, since $\mathrm{He}_0 \equiv 1$,

$$\mathbb{E}[U] = \langle \mathrm{He}_0, U \rangle_{L^2(\gamma)} = \sum_{n\in\mathbb{N}_0} u_n \langle \mathrm{He}_0, \mathrm{He}_n \rangle_{L^2(\gamma)} = u_0,$$

so the expected value of U is simply its 0^{th} PC coefficient. Similarly, its variance is a weighted sum of the squares of its PC coefficients:

$$\begin{aligned}
\mathbb{V}[U] &= \mathbb{E}\left[|U - \mathbb{E}[U]|^2\right] \\
&= \mathbb{E}\left[\left|\sum_{n\in\mathbb{N}} u_n \mathrm{He}_n\right|^2\right] && \text{since } \mathbb{E}[U] = u_0 \\
&= \sum_{m,n\in\mathbb{N}} u_m u_n \langle \mathrm{He}_m, \mathrm{He}_n\rangle_{L^2(\gamma)} \\
&= \sum_{n\in\mathbb{N}} u_n^2 \|\mathrm{He}_n\|_{L^2(\gamma)}^2 && \text{by Hermitian orthogonality} \\
&= \sum_{n\in\mathbb{N}} u_n^2 n!.
\end{aligned}$$

Example 11.13. Let $X \sim \mathcal{N}(m, \sigma^2)$ be a real-valued Gaussian random variable with mean $m \in \mathbb{R}$ and variance $\sigma^2 \geq 0$. Let $Y := e^X$; since $\log Y$ is normally distributed, the non-negative-valued random variable Y is said to be a *log-normal random variable*. As usual, let $\Xi \sim \mathcal{N}(0,1)$ be the standard Gaussian random variable; clearly $X \overset{\mathcal{L}}{=} m + \sigma\Xi$ and $Y \overset{\mathcal{L}}{=} e^m e^{\sigma\Xi}$. The Wiener–Hermite expansion of Y as $\sum_{k\in\mathbb{N}_0} y_k \mathrm{He}_k(\Xi)$ has coefficients

$$\begin{aligned}
y_k &= \frac{\langle e^{m+\sigma\Xi}, \mathrm{He}_k(\Xi)\rangle}{\|\mathrm{He}_k(\Xi)\|^2} \\
&= \frac{e^m}{k!}\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{\sigma\xi}\mathrm{He}_k(\xi)e^{-\xi^2/2}\,\mathrm{d}\xi \\
&= \frac{e^{m+\sigma^2/2}}{k!}\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} \mathrm{He}_k(\xi)e^{-(\xi-\sigma)^2/2}\,\mathrm{d}\xi \\
&= \frac{e^{m+\sigma^2/2}}{k!}\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} \mathrm{He}_k(w+\sigma)e^{-w^2/2}\,\mathrm{d}w.
\end{aligned}$$

The remaining Gaussian integral can be evaluated directly using the Cameron–Martin formula (Lemma 2.35), or else using the formula

$$\mathrm{He}_n(x+y) = \sum_{k=0}^{n}\binom{n}{k}x^{n-k}\mathrm{He}_k(y),$$

which follows from the derivative property $\mathrm{He}_n' = n\mathrm{He}_{n-1}$, with $x = \sigma$ and $y = w$: this formula yields that

$$y_k = \frac{e^{m+\sigma^2/2}}{k!}\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}}\sum_{j=0}^{k}\binom{k}{j}\sigma^{k-j}\mathrm{He}_j(w)e^{-w^2/2}\,\mathrm{d}w = \frac{e^{m+\sigma^2/2}\sigma^k}{k!}$$

since the orthogonality relation $\langle \mathrm{He}_m, \mathrm{He}_n\rangle_{L^2(\gamma)} = n!\delta_{mn}$ with $n = 0$ implies that every Hermite polynomial other than He_0 has mean 0 under standard Gaussian measure. That is,

$$Y = e^{m+\sigma^2/2}\sum_{k\in\mathbb{N}_0}\frac{\sigma^k}{k!}\mathrm{He}_k(\Xi). \tag{11.5}$$

The Wiener–Hermite expansion (11.5) reveals that $\mathbb{E}[Y] = e^{m+\sigma^2/2}$ and

$$\mathbb{V}[Y] = e^{2m+\sigma^2}\sum_{k\in\mathbb{N}}\left(\frac{\sigma^k}{k!}\right)^2\|\mathrm{He}_k\|_{L^2(\gamma)}^2 = e^{2m+\sigma^2}\left(e^{\sigma^2}-1\right).$$

Truncation of Wiener–Hermite Expansions. Of course, in practice, the series expansion $U = \sum_{n\in\mathbb{N}_0} u_n \mathrm{He}_n(\Xi)$ must be truncated after finitely many terms, and so it is natural to ask about the quality of the approximation

$$U \approx U^K := \sum_{n=0}^{K} u_n \mathrm{He}_n(\Xi).$$

Since the Hermite polynomials form a complete orthogonal basis for $L^2(\mathbb{R}, \gamma; \mathbb{R})$, the standard results about orthogonal approximations in Hilbert spaces apply. In particular, by Corollary 3.22, the truncation error $U - U^K$ is orthogonal to the space from which U^K was chosen, i.e.

$$\operatorname{span}\{\mathrm{He}_0, \mathrm{He}_1, \dots, \mathrm{He}_K\},$$

and tends to zero in mean square.

Lemma 11.14. *The truncation error $U - U^K$ is orthogonal to the subspace*

$$\operatorname{span}\{\mathrm{He}_0, \mathrm{He}_1, \dots, \mathrm{He}_K\}$$

of $L^2(\mathbb{R}, \mathrm{d}\gamma; \mathbb{R})$. Furthermore, $\lim_{K\to\infty} U^K = U$ in $L^2(\mathbb{R}, \gamma; \mathbb{R})$.

Proof. Let $V := \sum_{m=0}^{K} v_m \mathrm{He}_m$ be any element of the subspace of $L^2(\mathbb{R}, \gamma; \mathbb{R})$ spanned by the Hermite polynomials of degree at most K. Then

$$\begin{aligned}
\langle U - U^K, V\rangle_{L^2(\gamma)} &= \left\langle \left(\sum_{n>K} u_n \mathrm{He}_n\right), \left(\sum_{m=0}^{K} v_m \mathrm{He}_m\right)\right\rangle \\
&= \sum_{\substack{n>K \\ m\in\{0,\dots,K\}}} u_n v_m \langle \mathrm{He}_n, \mathrm{He}_m\rangle \\
&= 0.
\end{aligned}$$

Hence, by Pythagoras' theorem,

$$\|U\|^2_{L^2(\gamma)} = \|U^K\|^2_{L^2(\gamma)} + \|U - U^K\|^2_{L^2(\gamma)},$$

and hence $\|U - U^K\|_{L^2(\gamma)} \to 0$ as $K \to \infty$. □

11.3 Generalized Polynomial Chaos Expansions

The ideas of polynomial chaos can be generalized well beyond the setting in which the elementary random variable Ξ used to generate the orthogonal decomposition is a standard Gaussian random variable, or even a vector $\Xi = (\Xi_1, \dots, \Xi_d)$ of mutually orthogonal Gaussian random variables. Such expansions are referred to as *generalized polynomial chaos* (gPC) expansions.

Let $\Xi = (\Xi_1, \dots, \Xi_d)$ be an $\mathbb{R}^d$-valued random variable with independent (and hence L^2-orthogonal) components, called the *stochastic germ*. Let the measurable rectangle $\Theta = \Theta_1 \times \cdots \times \Theta_d \subseteq \mathbb{R}^d$ be the support (i.e. range) of Ξ. Denote by $\mu = \mu_1 \otimes \cdots \otimes \mu_d$ the distribution of Ξ on Θ. The objective is to express any function (random variable, random vector, or even random field) $U \in L^2(\Theta, \mu)$ in terms of elementary μ-orthogonal functions of the stochastic germ Ξ.

As usual, let $\mathbb{R}[\xi_1, \dots, \xi_d]$ denote the ring of all polynomials in $\xi_1, \dots, \xi_d$ with real coefficients, and let $\mathbb{R}[\xi_1, \dots, \xi_d]_{\le p}$ denote those polynomials of total degree at most $p \in \mathbb{N}_0$. Let $\Gamma_p \subseteq \mathbb{R}[\xi_1, \dots, \xi_d]_{\le p}$ be a collection of polynomials that are mutually orthogonal, orthogonal to $\mathbb{R}[\xi_1, \dots, \xi_d]_{\le p-1}$, and span $\mathbb{R}[\xi_1, \dots, \xi_d]_{=p}$, and let $\tilde{\Gamma}_p := \operatorname{span} \Gamma_p$. This yields the orthogonal decomposition

$$L^2(\Theta, \mu; \mathbb{R}) = \bigoplus_{p\in\mathbb{N}_0} \tilde{\Gamma}_p.$$

It is important to note that there is a lack of uniqueness in these basis polynomials whenever $d \geq 2$: each choice of ordering of multi-indices $\alpha \in \mathbb{N}_0^d$ can yield a different orthogonal basis of $L^2(\Theta, \mu)$ when the Gram–Schmidt procedure is applied to the monomials ξ^α.

Note that (as usual, assuming separability) the L^2 space over the product probability space $(\Theta, \mathscr{F}, \mu)$ is isomorphic to the Hilbert space tensor product of the L^2 spaces over the marginal probability spaces:

$$L^2(\Theta_1 \times \cdots \times \Theta_d, \mu_1 \otimes \cdots \otimes \mu_d; \mathbb{R}) = \bigotimes_{i=1}^{d} L^2(\Theta_i, \mu_i; \mathbb{R});$$

hence, as in Theorem 8.26, an orthogonal system of multivariate polynomials for $L^2(\Theta, \mu; \mathbb{R})$ can be found by taking products of univariate orthogonal polynomials for the marginal spaces $L^2(\Theta_i, \mu_i; \mathbb{R})$. A *generalized polynomial chaos* (gPC) expansion of a random variable or stochastic process U is simply the expansion of U with respect to such a complete orthogonal polynomial basis of $L^2(\Theta, \mu)$.

Example 11.15. Let $\Xi = (\Xi_1, \Xi_2)$ be such that Ξ_1 and Ξ_2 are independent (and hence orthogonal) and such that Ξ_1 is a standard Gaussian random variable and Ξ_2 is uniformly distributed on $[-1, 1]$. Hence, the univariate orthogonal polynomials for Ξ_1 are the Hermite polynomials He_n and the univariate orthogonal polynomials for Ξ_2 are the Legendre polynomials Le_n. Thus, by Theorem 8.26, a system of orthogonal polynomials for Ξ up to total degree 3 is

$$\begin{aligned}
\Gamma_0 &= \{1\}, \\
\Gamma_1 &= \{\mathrm{He}_1(\xi_1), \mathrm{Le}_1(\xi_2)\} \\
&= \{\xi_1, \xi_2\}, \\
\Gamma_2 &= \{\mathrm{He}_2(\xi_1), \mathrm{He}_1(\xi_1)\mathrm{Le}_1(\xi_2), \mathrm{Le}_2(\xi_2)\} \\
&= \{\xi_1^2 - 1, \xi_1\xi_2, \tfrac{1}{2}(3\xi_2^2 - 1)\}, \\
\Gamma_3 &= \{\mathrm{He}_3(\xi_1), \mathrm{He}_2(\xi_1)\mathrm{Le}_1(\xi_2), \mathrm{He}_1(\xi_1)\mathrm{Le}_2(\xi_2), \mathrm{Le}_3(\xi_2)\} \\
&= \{\xi_1^3 - 3\xi_1, \xi_1^2\xi_2 - \xi_2, \tfrac{1}{2}(3\xi_1\xi_2^2 - \xi_1), \tfrac{1}{2}(5\xi_2^3 - 3\xi_2)\}.
\end{aligned}$$

Remark 11.16. To simplify the notation in what follows, the following conventions will be observed:

1. To simplify expectations, inner products and norms, $\langle \cdot \rangle_\mu$ or simply $\langle \cdot \rangle$ will denote integration (i.e. expectation) with respect to the probability measure μ, so that the $L^2(\mu)$ inner product is simply $\langle X, Y \rangle_{L^2(\mu)} = \langle XY \rangle_\mu$.
2. Rather than have the orthogonal basis polynomials be indexed by multi-indices $\alpha \in \mathbb{N}_0^d$, or have two scalar indices, one for the degree p and one within each set Γ_p, it is convenient to order the basis polynomials using a single scalar index $k \in \mathbb{N}_0$. It is common in practice to take $\Psi_0 = 1$ and to have the polynomial degree be (weakly) increasing with respect to the new index k. So, to continue Example 11.15, one could use the graded lexicographic ordering on $\alpha \in \mathbb{N}_0^2$ so that $\Psi_0(\xi) = 1$ and

$$\begin{aligned}
&\Psi_1(\xi) = \xi_1, && \Psi_2(\xi) = \xi_2, && \Psi_3(\xi) = \xi_1^2 - 1, \\
&\Psi_4(\xi) = \xi_1\xi_2, && \Psi_5(\xi) = \tfrac{1}{2}(3\xi_2^2 - 1), && \Psi_6(\xi) = \xi_1^3 - 3\xi_1, \\
&\Psi_7(\xi) = \xi_1^2\xi_2 - \xi_2, && \Psi_8(\xi) = \tfrac{1}{2}(3\xi_1\xi_2^2 - \xi_1), && \Psi_9(\xi) = \tfrac{1}{2}(5\xi_2^3 - 3\xi_2).
\end{aligned}$$

3. By abuse of notation, Ψ_k will stand for both a polynomial function (which is a deterministic function from $\mathbb{R}^d$ to $\mathbb{R}$) and for the real-valued random variable that is the composition of that polynomial with the stochastic germ Ξ (which is a function from an abstract probability space to $\mathbb{R}$).

Truncation of gPC Expansions. Suppose that a gPC expansion of the form $U = \sum_{k \in \mathbb{N}_0} u_k \Psi_k$ is truncated, i.e. we consider

$$U^K = \sum_{k=0}^{K} u_k \Psi_k.$$

It is an easy exercise to show that the truncation error $U - U^K$ is orthogonal to $\operatorname{span}\{\Psi_0, \dots, \Psi_K\}$. It is also worth considering how many terms there are in such a truncated gPC expansion. Suppose that the stochastic germ Ξ has dimension d (i.e. has d independent components), and we work only with polynomials of total degree at most p. The total number of coefficients in the truncated expansion U^K is

$$K + 1 = \frac{(d+p)!}{d!p!}.$$

That is, the total number of gPC coefficients that must be calculated grows combinatorially as a function of the number of input random variables and the degree of polynomial approximation. Such rapid growth limits the usefulness of gPC expansions for practical applications where d and p are much greater than the order of 10 or so.

Remark 11.17. It is possible to adapt the notion of a gPC expansion to the situation of dependent random variables, but there are some complications. In summary, suppose that $\Xi = (\Xi_1, \dots, \Xi_d)$, taking values in $\Theta = \Theta_1 \times \cdots \times \Theta_d$, has joint law μ, which is not necessarily a product measure. Nevertheless, let μ_i denote the marginal law of Ξ_i, i.e.

$$\mu_i(E_i) := \mu(\Theta_1 \times \cdots \times \Theta_{i-1} \times E_i \times \Theta_{i+1} \times \cdots \times \Theta_d).$$

To simplify matter further, assume that μ (resp. μ_i) has Lebesgue density ρ (resp. ρ_i). Now let $\phi_p^{(i)}(\xi_i) \in \mathbb{R}[\xi_i]$, $p \in \mathbb{N}_0$, be univariate orthogonal polynomials for μ_i. The *chaos function* associated to a multi-index $\alpha \in \mathbb{N}_0^d$ is defined to be

$$\Psi_\alpha(\xi) := \sqrt{\frac{\rho_1(\xi_1) \dots \rho_d(\xi_d)}{\rho(\xi)}} \phi_{\alpha_1}^{(1)}(\xi_1) \dots \phi_{\alpha_d}^{(d)}(\xi_d).$$

It can be shown that the family $\{\Psi_\alpha \mid \alpha \in \mathbb{N}_0^d\}$ is a complete orthonormal basis for $L^2(\Theta, \mu; \mathbb{R})$, so we have the usual series expansion $U = \sum_\alpha u_\alpha \Psi_\alpha$. Note, however, that with the exception of $\Psi_0 = 1$, the functions Ψ_α are not polynomials. Nevertheless, we still have the usual properties that truncation error is orthogonal the the approximation subspace, and

$$\mathbb{E}_\mu[U] = u_0, \quad \mathbb{V}_\mu[U] = \sum_{\alpha \neq 0} u_\alpha^2 \langle \Psi_\alpha^2 \rangle_\mu.$$

Expansions of Random Variables. Consider a real-valued random variable U, which we expand in terms of a stochastic germ Ξ as

$$U^K(\Xi) = \sum_{k \in \mathbb{N}_0} u_k \Psi_k(\Xi),$$

where the basis functions Ψ_k are orthogonal with respect to the law of Ξ, and with the usual convention that $\Psi_0 = 1$. A first, easy, observation is that

$$\mathbb{E}[U] = \langle \Psi_0 U \rangle = \sum_{k \in \mathbb{N}_0} u_k \langle \Psi_0 \Psi_k \rangle = u_0,$$

so the expected value of U is simply its 0^{th} gPC coefficient. Similarly, its variance is a weighted sum of the squares of its gPC coefficients:

$$\begin{aligned} \mathbb{E}\left[|U - \mathbb{E}[U]|^2\right] &= \mathbb{E}\left[\left|\sum_{k \in \mathbb{N}_0} u_k \Psi_k\right|^2\right] \\ &= \sum_{k,\ell \in \mathbb{N}} u_k u_\ell \langle \Psi_k \Psi_\ell \rangle \\ &= \sum_{k \in \mathbb{N}} u_k^2 \langle \Psi_k^2 \rangle. \end{aligned}$$

Similar remarks apply to any truncation $U^K = \sum_{k=1}^K u_k \Psi_k$ of the gPC expansion of U. In view of the expression for the variance, the gPC coefficients can be used as sensitivity indices. That is, a natural measure of how strongly U depends upon $\Psi_k(\Xi)$ is

$$\frac{u_k^2 \langle \Psi_k^2 \rangle}{\sum_{\ell \geq 1} u_\ell^2 \langle \Psi_\ell^2 \rangle}.$$

Expansions of Random Vectors. Similarly, if $U_1, \ldots, U_n$ are (not necessarily independent) real-valued random variables, then the $\mathbb{R}^n$-valued random variable $\boldsymbol{U} = [U_1, \ldots, U_n]^\top$ with the U_i as its components can be given a (possibly truncated) expansion

$$\boldsymbol{U}(\xi) = \sum_{k \in \mathbb{N}_0} \boldsymbol{u}_k \Psi_k(\xi),$$

with vector-valued gPC coefficients $\boldsymbol{u}_k = [u_{1,k}, \ldots, u_{n,k}]^\top \in \mathbb{R}^n$ for each $k \in \mathbb{N}_0$. As before,

$$\mathbb{E}[\boldsymbol{U}] = \langle \Psi_0 \boldsymbol{U} \rangle = \sum_{k \in \mathbb{N}_0} \boldsymbol{u}_k \langle \Psi_0 \Psi_k \rangle = \boldsymbol{u}_0 \in \mathbb{R}^n$$

and the covariance matrix $C \in \mathbb{R}^{n \times n}$ of U is given by

$$C = \sum_{k \in \mathbb{N}} \boldsymbol{u}_k \boldsymbol{u}_k^\top \langle \Psi_k^2 \rangle$$

i.e. its components are $C_{ij} = \sum_{k \in \mathbb{N}} u_{i,k} u_{j,k} \langle \Psi_k^2 \rangle$.

Expansions of Stochastic Processes. Consider now a stochastic process U, i.e. a function $U \colon \Theta \times \Omega \to \mathbb{R}$. Suppose that U is square integrable in the sense that, for each $x \in \Omega$, $U(\cdot, x) \in L^2(\Theta, \mu)$ is a real-valued random variable, and, for each $\theta \in \Theta$, $U(\theta, \cdot) \in L^2(\Omega, \mathrm{d}x)$ is a scalar field on the domain Ω. Recall that

$$L^2(\Theta, \mu; \mathbb{R}) \otimes L^2(\Omega, \mathrm{d}x; \mathbb{R}) \cong L^2(\Theta \times \Omega, \mu \otimes \mathrm{d}x; \mathbb{R}) \cong L^2\big(\Theta, \mu; L^2(\Omega, \mathrm{d}x)\big),$$

so U can be equivalently viewed as a linear combination of products of $\mathbb{R}$-valued random variables with deterministic scalar fields, or as a function on $\Theta \times \Omega$, or as a field-valued random variable. As usual, take $\{\Psi_k \mid k \in \mathbb{N}_0\}$ to be an orthogonal polynomial basis of $L^2(\Theta, \mu; \mathbb{R})$, ordered (weakly) by total degree, with $\Psi_0 = 1$. A gPC expansion of the random field U is an L^2-convergent expansion of the form

$$U(x, \xi) = \sum_{k \in \mathbb{N}_0} u_k(x) \Psi_k(\xi).$$

The functions $u_k \colon \Omega \to \mathbb{R}$ are called the *stochastic modes* of the process U. The stochastic mode $u_0 \colon \Omega \to \mathbb{R}$ is the *mean field* of U:

$$\mathbb{E}[U(x)] = u_0(x).$$

The variance of the field at $x \in \Omega$ is

$$\mathbb{V}[U(x)] = \sum_{k \in \mathbb{N}} u_k^2 \langle \Psi_k^2 \rangle,$$

whereas, for two points $x, y \in \Omega$,

$$\begin{aligned} \mathbb{E}[U(x)U(y)] &= \left\langle \sum_{k \in \mathbb{N}_0} u_k(x) \Psi_k(\xi) \sum_{\ell \in \mathbb{N}_0} u_\ell(y) \Psi_\ell(\xi) \right\rangle \\ &= \sum_{k \in \mathbb{N}_0} u_k(x) u_k(y) \langle \Psi_k^2 \rangle \end{aligned}$$

and so the covariance function of U is given by

$$C_U(x, y) = \sum_{k \in \mathbb{N}} u_k(x) u_k(y) \langle \Psi_k^2 \rangle.$$

The previous remarks about gPC expansions of vector-valued random variables are a special case of these remarks, namely $\Omega = \{1, \ldots, n\}$. At least when $\dim \Omega$ is low, it is very common to see the behaviour of a stochastic field U (or its truncation U^K) summarized by plots of the mean field and the variance field, as well as a few 'typical' sample realizations. The visualization of high-dimensional data is a subject unto itself, with many ingenious uses of shading, colour, transparency, videos, and user interaction tools.

Changes of gPC Basis. It is possible to change between representations of a stochastic quantity U with respect to gPC bases $\{\Psi_k \mid k \in \mathbb{N}_0\}$ and $\{\Phi_k \mid k \in \mathbb{N}_0\}$ generated by measures μ and ν respectively. Obviously, for such changes of basis to work in both directions, μ and ν must at least have the same support. Suppose that

$$U = \sum_{k \in \mathbb{N}_0} u_k \Psi_k = \sum_{k \in \mathbb{N}_0} v_k \Phi_k.$$

Then, taking the $L^2(\nu)$-inner product of this equation with Φ_ℓ,

$$\langle U \Phi_\ell \rangle_\nu = \sum_{k \in \mathbb{N}_0} u_k \langle \Psi_k \Phi_\ell \rangle_\nu = v_\ell \langle \Psi_\ell^2 \rangle_\nu,$$

provided that $\Psi_k \Phi_\ell \in L^2(\nu)$ for all $k \in \mathbb{N}_0$, i.e.

$$v_\ell = \sum_{k \in \mathbb{N}_0} \frac{u_k \langle \Psi_k \Phi_\ell \rangle_\nu}{\langle \Psi_\ell^2 \rangle_\nu}.$$

Similarly, taking the $L^2(\mu)$-inner product of this equation with Ψ_ℓ yields that, provided that $\Phi_k \Psi_\ell \in L^2(\mu)$ for all $k \in \mathbb{N}_0$,

$$u_\ell = \sum_{k \in \mathbb{N}_0} \frac{v_k \langle \Phi_k \Psi_\ell \rangle_\mu}{\langle \Psi_\ell^2 \rangle_\mu}.$$

11.4 Wavelet Expansions

Recall from the earlier discussion of Gibbs' phenomenon in Chapter 8 that expansions of non-smooth functions in terms of smooth basis functions such as polynomials, while guaranteed to be convergent in the L^2 sense, can have poor pointwise convergence properties. However, to remedy such problems, one can consider spectral expansions in term of orthogonal bases of functions in $L^2(\Theta, \mu; \mathbb{R})$ that are no longer polynomials: a classic example of such a construction is the use of *wavelets*, which were developed to resolve the same problem in harmonic analysis and its applications. This section considers, by way of example, orthogonal decomposition of random variables using Haar wavelets, the so-called *Wiener–Haar expansion*.

Definition 11.18. The *Haar scaling function* is $\phi(x) := \mathbb{1}_{[0,1)}(x)$. For $j \in \mathbb{N}_0$ and $k \in \{0, \ldots, 2^j - 1\}$, let $\phi_{j,k}(x) := 2^{j/2} \phi(2^j x - k)$ and

$$\mathcal{V}_j := \operatorname{span}\{\phi_{j,0}, \ldots, \phi_{j,2^j-1}\}.$$

The *Haar function* (or *Haar mother wavelet*) $\psi \colon [0, 1] \to \mathbb{R}$ is defined by

$$\psi(x) := \begin{cases} 1, & \text{if } 0 \leq x < \frac{1}{2}, \\ -1, & \text{if } \frac{1}{2} \leq x < 1, \\ 0, & \text{otherwise.} \end{cases}$$

The *Haar wavelet family* is the collection of scaled and shifted versions $\psi_{j,k}$ of the mother wavelet ψ defined by

$$\psi_{j,k}(x) := 2^{j/2}\psi(2^j x - k) \quad \text{for } j \in \mathbb{N}_0 \text{ and } k \in \{0, \dots, 2^j - 1\}.$$

The spaces $\mathcal{V}_j$ form an increasing family of subspaces of $L^2([0,1], \mathrm{d}x; \mathbb{R})$, with the index j representing the level of 'detail' permissible in a function $f \in \mathcal{V}_j$: more concretely, $\mathcal{V}_j$ is the set of functions on $[0,1]$ that are constant on each half-open interval $[2^{-j}k, 2^{-j}(k+1))$. A straightforward calculation from the above definition yields the following:

Lemma 11.19. *For all $j, j' \in \mathbb{N}_0$, $k \in \{0, \dots, 2^j - 1\}$ and $k' \in \{0, \dots, 2^{j'} - 1\}$,*

$$\int_0^1 \psi_{j,k}(x)\,\mathrm{d}x = 0, \quad \textit{and}$$
$$\int_0^1 \psi_{j,k}(x)\psi_{j',k'}(x)\,\mathrm{d}x = \delta_{jj'}\delta_{kk'}.$$

Hence, $\{1\} \cup \{\psi_{j,k} \mid j \in \mathbb{N}_0, k \in \{0, 1, \dots, 2^j - 1\}\}$ is a complete orthonormal basis of $L^2([0,1], \mathrm{d}x; \mathbb{R})$. If $\mathcal{W}_j$ denotes the orthogonal complement of $\mathcal{V}_j$ in $\mathcal{V}_{j+1}$, then

$$\mathcal{W}_j = \operatorname{span}\{\psi_{j,0}, \dots, \psi_{j,2^j-1}\}, \quad \textit{and}$$
$$L^2([0,1], \mathrm{d}x; \mathbb{R}) = \bigoplus_{j \in \mathbb{N}_0} \mathcal{W}_j.$$

Consider a stochastic germ $\Xi \sim \mu \in \mathcal{M}_1(\mathbb{R})$ with cumulative distribution function $F_\Xi \colon \mathbb{R} \to [0,1]$. For simplicity, suppose that F_Ξ is continuous and strictly increasing, so that F_Ξ is differentiable (with $F_\Xi' = \frac{\mathrm{d}\mu}{\mathrm{d}x} = \rho_\Xi$) almost everywhere, and also invertible. We wish to write a random variable $U \in L^2(\mathbb{R}, \mu; \mathbb{R})$, in particular one that may be a non-smooth function of Ξ, as

$$\begin{aligned} U(\xi) &= u_0 + \sum_{j \in \mathbb{N}_0} \sum_{k=0}^{2^j - 1} u_{j,k} \psi_{j,k}(F_\Xi(\xi)) \\ &= u_0 + \sum_{j \in \mathbb{N}_0} \sum_{k=0}^{2^j - 1} u_{j,k} W_{j,k}(\xi); \end{aligned}$$

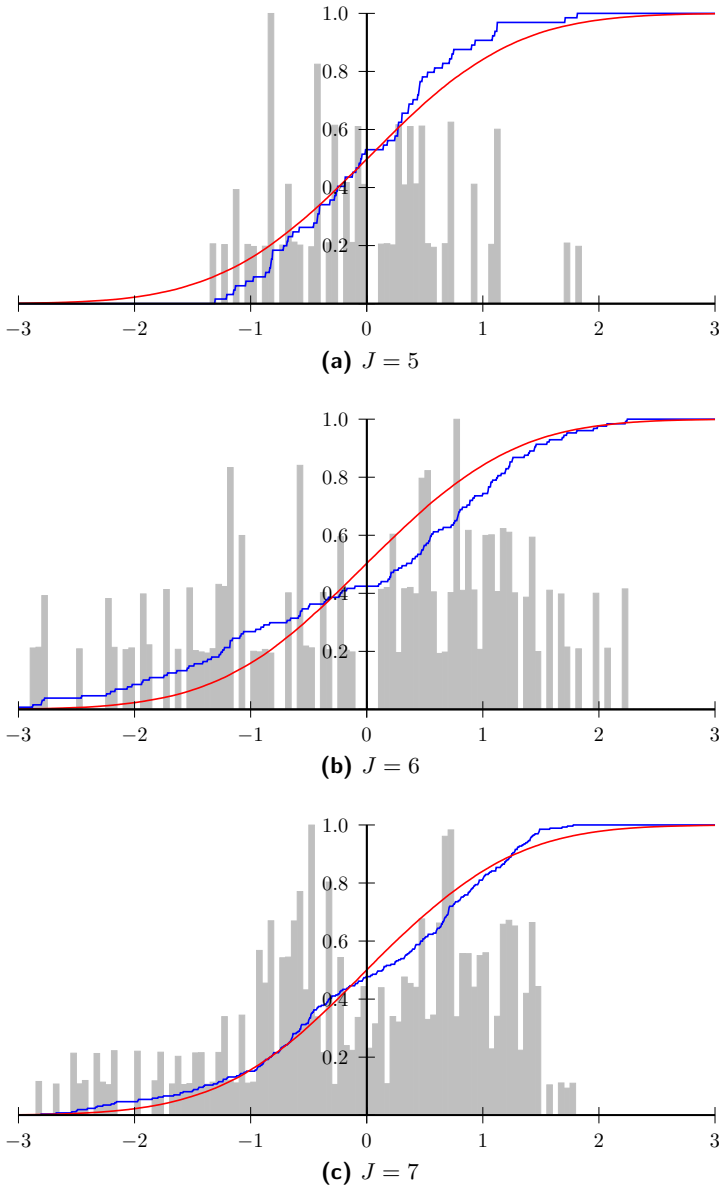
such an expansion will be called a *Wiener–Haar* expansion of U. See Figure 11.2 for an illustration comparing the cumulative distribution function of a truncated Wiener–Haar expansion to that of a standard Gaussian, showing the 'clumping' of probability mass that is to be expected of Wiener–Haar wavelet expansions but not of Wiener–Hermite polynomial chaos expansions.

Note that, by a straightforward change of variables $x = F_\Xi(\xi)$:

$$\begin{aligned} \int_{\mathbb{R}} W_{j,k}(\xi) W_{j',k'}(\xi)\,\mathrm{d}\mu(\xi) &= \int_{\mathbb{R}} W_{j,k}(\xi) W_{j',k'}(\xi) \rho_\Xi(\xi)\,\mathrm{d}\xi \\ &= \int_0^1 \psi_{j,k}(x)\psi_{j',k'}(x)\,\mathrm{d}x \\ &= \delta_{jj'}\delta_{kk'}, \end{aligned}$$

so the family $\{W_{j,k} \mid j \in \mathbb{N}_0, k \in \{0, \dots, 2^j - 1\}\}$ forms a complete orthonormal basis for $L^2(\mathbb{R}, \mu; \mathbb{R})$. Hence, the Wiener–Haar coefficients are determined by

$$\begin{aligned} u_{j,k} = \langle U W_{j,k} \rangle &= \int_{\mathbb{R}} U(\xi) W_{j,k}(\xi) \rho_\Xi(\xi)\,\mathrm{d}\xi \\ &= \int_0^1 U(F_\Xi^{-1}(x))\psi_{j,k}(x)\,\mathrm{d}x. \end{aligned}$$



(a) $J = 5$

(b) $J = 6$

(c) $J = 7$

Figure 11.2: The cumulative distribution function (blue) and binned peak-normalized probability density function (grey) of a random variable U (10^5 independent samples) with truncated Wiener–Haar expansion $U = \sum_{j=0}^{J} \sum_{k=0}^{2^j-1} u_{j,k} W_{j,k}(\Xi)$, where $\Xi \sim \mathcal{N}(0,1)$. The coefficients $u_{j,k}$ were sampled independently from $u_{j,k} \sim 2^{-j}\mathcal{N}(0,1)$. The cumulative distribution function of a standard Gaussian is shown in red for comparison. Note that the (sample) law of U is more localized than would be possible with a polynomial expansion, and indeed has regions of zero probability mass.

As in the case of a gPC expansion, the usual expressions for the mean and variance of U hold:

$$\mathbb{E}[U] = u_0 \quad \text{and} \quad \mathbb{V}[U] = \sum_{j \in \mathbb{N}_0} \sum_{k=0}^{2^j - 1} |u_{j,k}|^2.$$

Comparison of Wavelet and gPC Expansions. Despite the formal similarities of the corresponding expansions, there are differences between wavelet and gPC spectral expansions. For gPC expansions, the globally-smooth orthogonal polynomials used as the basis elements have the property that expansions of smooth functions / random variables enjoy a fast convergence rate, as in Theorem 8.24; no such connection between smoothness and convergence rate is to be expected for Wiener–Haar expansions, in which the basis functions are non-smooth. However, in cases in which U shows a localized sharp variation or a discontinuity, a Wiener–Haar expansion may be more efficient than a gPC expansion, since the convergence rate of the latter would be impaired by Gibbs-type phenomena. Another distinctive feature of the Wiener–Haar expansion concerns products of piecewise constant processes. For instance, for $f, g \in \mathcal{V}_j$ the product fg is again an element of $\mathcal{V}_j$; it is not true that the product of two polynomials of degree at most n is again a polynomial of degree at most n. Therefore, for problems with strong dependence upon high-degree / high-detail features, or with multiplicative structure, Wiener–Haar expansions may be more appropriate than gPC expansions.

Bibliography

Spectral expansions in general are covered in Chapter 2 of the monograph of Le Maître & Knio [102], and Chapter 5 of the book of Xiu [182].

The Karhunen–Loève expansion bears the names of Karhunen [86] and Loève [111], but KL-type series expansions of stochastic processes were considered earlier by Kosambi [93]. Lemma 11.14, that the truncation error in a PC expansion is orthogonal to the approximation subspace, is nowadays a simple corollary of standard results in Hilbert spaces, but is an observation that appears to have first first been made in the stochastic context by Cameron & Martin [25]. The application of Wiener–Hermite PC expansions to engineering systems was popularized by Ghanem & Spanos [64]; the extension to gPC and the connection with the Askey scheme is due to Xiu & Karniadakis [183].

The extension of gPC expansions to arbitrary dependency among the components of the stochastic germ, as in Remark 11.17, is due to Soize & Ghanem [154]. The orthogonal decomposition properties of the Haar basis were first noted by Haar [72]. The book of Meyer [117] provides a thorough introduction to wavelets in general. Wavelet bases for UQ, which can better resolve locally non-smooth features of random fields, are discussed in Chapter 8 of the book of Le Maître & Knio [102] and in articles of Le Maître & al. [103, 104, 105]. Wavelets are also used in the sampling of Besov measures, as in the articles of Dashti & al. [40] and Lassas & al. [99], and Theorem 11.10 is synthesized from results in those two papers.

Exercises

Exercise 11.1. Consider the negative Laplacian operator $\mathcal{L} := -\frac{\mathrm{d}^2}{\mathrm{d}x^2}$ acting on real-valued functions on the interval $[0, 1]$, with zero boundary conditions. Show that the eigenvalues μ_n and normalized eigenfunctions ψ_n of $\mathcal{L}$ are

$$\mu_n = (\pi n)^2, \quad \psi_n(x) = \sqrt{2}\sin(\pi n x).$$

Hence show that $C := \mathcal{L}^{-1}$ has the same eigenfunctions with eigenvalues $\lambda_n = (\pi n)^{-2}$. Hence, using the Karhunen–Loève theorem, generate figures similar to Figure 11.1 for your choice of mean field $m \colon [0, 1] \to \mathbb{R}$.

Exercise 11.2. Do the analogue of Exercise 11.2 for the negative Laplacian operator $\mathcal{L} := -\frac{\mathrm{d}^2}{\mathrm{d}x^2} - \frac{\mathrm{d}^2}{\mathrm{d}y^2}$ acting on real-valued functions on the square $[0,1]^2$, again with zero boundary conditions.

Exercise 11.3. Show that the eigenvalues λ_n and eigenfunctions e_n of the exponential covariance function $C(x,y) = \exp(-|x-y|/a)$ on $[-b,b]$ are given by

$$\lambda_n = \begin{cases} \frac{2a}{1+a^2 w_n^2}, & \text{if } n \in 2\mathbb{Z}, \\ \frac{2a}{1+a^2 v_n^2}, & \text{if } n \in 2\mathbb{Z}+1, \end{cases}$$

$$e_n(x) = \begin{cases} \sin(w_n x) \Big/ \sqrt{b - \frac{\sin(2w_n b)}{2w_n}}, & \text{if } n \in 2\mathbb{Z}, \\ \cos(v_n x) \Big/ \sqrt{b + \frac{\sin(2v_n b)}{2v_n}}, & \text{if } n \in 2\mathbb{Z}+1, \end{cases}$$

where w_n and v_n solve the transcendental equations

$$\begin{cases} a w_n + \tan(w_n b) = 0, & \text{for } n \in 2\mathbb{Z}, \\ 1 - a v_n \tan(v_n b) = 0, & \text{for } n \in 2\mathbb{Z}+1. \end{cases}$$

Hence, using the Karhunen–Loève theorem, generate sample paths from the Gaussian measure with covariance kernel C and your choice of mean path.

Exercise 11.4 (Karhunen–Loève-type sampling of Besov measures)**.** Let $\mathbb{T}^d := \mathbb{R}^d/\mathbb{Z}^d$ denote the the d-dimensional unit torus. Let $\{\psi_\ell \mid \ell \in \mathbb{N}\}$ be an orthonormal basis for $L^2(\mathbb{T}^d, \mathrm{d}x; \mathbb{R})$. Let $q \in [1,\infty)$ and $s \in (0,\infty)$, and define a new norm $\|\cdot\|_{X^{s,q}}$ on series $u = \sum_{\ell\in\mathbb{N}} u_\ell \psi_\ell$ by

$$\left\| \sum_{\ell\in\mathbb{N}} u_\ell \psi_\ell \right\|_{X^{s,q}} := \left(\sum_{\ell\in\mathbb{N}} \ell^{\frac{sq}{d}+\frac{q}{2}-1} |u_\ell|^q \right)^{1/q}.$$

Show that $\|\cdot\|_{X^{s,q}}$ is indeed a norm and that the set of u with $\|u\|_{X^{s,q}}$ finite forms a Banach space. Now, for $q \in [1,\infty)$, $s > 0$, and $\kappa > 0$, define a random function U by

$$U(x) := \sum_{\ell\in\mathbb{N}} \ell^{-(\frac{s}{d}+\frac{1}{2}-\frac{1}{q})} \kappa^{-\frac{1}{q}} \Xi_\ell \psi_\ell(x)$$

where Ξ_ℓ are sampled independently and identically from the generalized Gaussian measure on $\mathbb{R}$ with Lebesgue density proportional to $\exp(-\frac{1}{2}|\xi|^q)$. By treating the above construction as an infinite product measure and considering the product of the densities $\exp(-\frac{1}{2}|\xi_\ell|^q)$, show formally that U has 'Lebesgue density' proportional to $\exp(-\frac{\kappa}{2}\|u\|^q_{X^{s,q}})$.

Generate sample realizations of U and investigate the effect of the various parameters q, s, and κ. It may be useful to know that samples from the probability measure $\frac{\beta^{1/2}}{2\Gamma(1+\frac{1}{q})} \exp(-\beta^{q/2}|x-m|^q)\,\mathrm{d}x$ can be generated as $m + \beta^{-1/2} S |Y|^{1/q}$ where S is uniformly distributed in $\{-1,+1\}$ and Y is distributed according to the Gamma distribution with parameter q, which has Lebesgue density $qe^{-qx}\mathbb{1}_{[0,\infty)}(x)$

Chapter 12

Stochastic Galerkin Methods

> Not to be absolutely certain is, I think, one of the essential things in rationality.
>
> *Am I an Atheist or an Agnostic?*
> Bertrand Russell

The previous chapter considered spectral expansions of square-integrable random variables, random vectors and random fields of the form

$$U = \sum_{k \in \mathbb{N}_0} u_k \Psi_k,$$

where $U \in L^2(\Theta, \mu; \mathcal{V})$ and $\{\Psi_k \mid k \in \mathbb{N}_0\}$ is an orthogonal basis for $L^2(\Theta, \mu; \mathbb{R})$. However, beyond the standard Hilbert space orthogonal projection relation

$$u_k = \frac{\langle U \Psi_k \rangle}{\langle \Psi_k^2 \rangle},$$

we know very little about how to solve for the stochastic modes $u_k \in \mathcal{V}$. For example, if U is the solution to a stochastic version of some problem such as an ODE or PDE (e.g. with randomized coefficients), how are the coefficients u_k related to solutions of the orginal deterministic problem? This chapter and the next one focus on the determination of stochastic modes by two classes of methods, the *intrusive* and the *non-intrusive*.

This chapter considers intrusive spectral methods, and in particular Galerkin methods. Galerkin methods use the formalism of *weak solutions*, as expressed in terms of inner products and commonly used in PDE theory, to form systems of governing equations for the solution's stochastic modes, which are generally coupled together. This stands in contrast to the non-intrusive approaches of the next chapter, which rely on individual realizations to determine the stochastic model response to random inputs.

Suppose that the model relationship between some input data d and the output (solution) u can be expressed formally as

$$\mathcal{M}(u; d) = 0, \tag{12.1}$$

an equality in some topological vector space $\mathcal{V}$. A *weak interpretation* of this model relationship is that, for some collection of *test functions* $\mathcal{T} \subseteq \mathcal{V}'$,

$$\langle \tau \,|\, \mathcal{M}(u; d) \rangle = 0 \quad \text{for all } \tau \in \mathcal{T}. \tag{12.2}$$

Although it is clear that (12.1) $\implies$ (12.2), the converse implication is not generally true, which is why (12.2) is known as a 'weak' interpretation of (12.1). The weak formulation (12.2) is very

attractive both for theory and for practical implementation: in particular, the requirement that (12.2) should hold only for τ in some basis of a finite-dimensional test space $\mathcal{T}$ lies at the foundation of many numerical methods.

Example 12.1. A good model for this kind of set-up is an elliptic boundary value problem on, say, a bounded, connected domain $\Omega \subseteq \mathbb{R}^n$ with smooth boundary $\partial\Omega$:

$$\begin{aligned} -\nabla \cdot (\kappa(x) \nabla u(x)) &= f(x) && \text{for } x \in \Omega, \\ u(x) &= 0 && \text{for } x \in \partial\Omega. \end{aligned} \tag{12.3}$$

In this case, the input data d are typically the forcing term $f \colon \Omega \to \mathbb{R}$ and the permeability field $\kappa \colon \Omega \to \mathbb{R}^{n \times n}$; in some cases, the domain Ω itself might depend upon d, but this introduces additional complications that will not be considered in this chapter. For a PDE such as this, solutions u are typically sought in the Sobolev space $H_0^1(\Omega)$ of L^2 functions that have a weak derivative that itself lies in L^2, and that vanish on $\partial\Omega$ in the sense of trace. Moreover, it is usual to seek *weak solutions*, i.e. $u \in H_0^1(\Omega)$ for which the inner product of (12.3) with any $v \in H_0^1(\Omega)$ is an equality of scalars. That is, integrating by parts, we seek $u \in H_0^1(\Omega)$ such that

$$\langle \kappa \nabla u, \nabla v \rangle_{L^2(\Omega)} = \langle f, v \rangle_{L^2(\Omega)} \quad \text{for all } v \in H_0^1(\Omega). \tag{12.4}$$

On expressing this problem in a chosen basis of $H_0^1(\Omega)$, the column vector $\boldsymbol{u}$ of coefficients of u in this basis turn out to satisfy a matrix-vector equation (i.e. a system of simultaneous linear equations) of the form $\boldsymbol{a}\boldsymbol{u} = \boldsymbol{b}$ for some matrix $\boldsymbol{a}$ determined by the permeability field κ and a column vector $\boldsymbol{b}$ determined by the forcing term f.

In this chapter, after reviewing basic Lax–Milgram theory and Galerkin projection for problems like (12.3)–(12.4), we consider the situation in which the input data d are uncertain and are described as a random variable $D(\xi)$. Then the solution is also a random variable $U(\xi)$ and the model relationship becomes

$$\mathcal{M}(U(\xi); D(\xi)) = 0.$$

Again, this equation is usually interpreted in a weak sense in a suitable space of random variables. If D and U are expanded in some gPC basis, it is natural to ask how the gPC coefficients of U with respect to this gPC basis and a chosen basis of the 'deterministic space' $H_0^1(\Omega)$ are related to one another. It will turn out that, like in the standard deterministic setting, this problem can be written in the form of a matrix-vector equation $\boldsymbol{A}\boldsymbol{U} = \boldsymbol{B}$ related to, but more complicated than, the deterministic problem $\boldsymbol{a}\boldsymbol{u} = \boldsymbol{b}$.

12.1 Lax–Milgram Theory and Galerkin Projection

Let $\mathcal{H}$ be a real Hilbert space equipped with a bilinear form $a \colon \mathcal{H} \times \mathcal{H} \to \mathbb{R}$. Given $f \in \mathcal{H}'$ (i.e. a continuous linear functional $f \colon \mathcal{H} \to \mathbb{R}$), the associated *weak problem* is:

$$\text{find } u \in \mathcal{H} \text{ such that } a(u, v) = \langle f \,|\, v \rangle \text{ for all } v \in \mathcal{H}. \tag{12.5}$$

Example 12.2. Let $\Omega \subseteq \mathbb{R}^n$ be a bounded, connected domain. Let a matrix-valued function $\kappa \colon \Omega \to \mathbb{R}^{n \times n}$ and a scalar-valued function $f \colon \Omega \to \mathbb{R}$ be given, and consider the elliptic problem (12.3). The appropriate bilinear form $a(\cdot, \cdot)$ is defined by

$$a(u, v) := \langle -\nabla \cdot (\kappa \nabla u), v \rangle_{L^2(\Omega)} = \langle \kappa \nabla u, \nabla v \rangle_{L^2(\Omega)},$$

where the second equality follows from integration by parts when u, v are smooth functions that vanish on $\partial\Omega$; such functions form a dense subset of the Sobolev space $H_0^1(\Omega)$. This short calculation motivates two important developments in the treatment of the PDE (12.3). First, even though the original formulation (12.3) seems to require the solution u to have two orders of differentiability, the last line of the above calculation makes sense even if u and v have only one order

of (weak) differentiability, and so we restrict attention to $H_0^1(\Omega)$. Second, we declare $u \in H_0^1(\Omega)$ to be a *weak solution* of (12.3) if the $L^2(\Omega)$ inner product of (12.3) with any $v \in H_0^1(\Omega)$ holds as an equality of real numbers, i.e. if

$$-\int_\Omega \nabla \cdot (\kappa(x)\nabla u(x))v(x)\,\mathrm{d}x = \int_\Omega f(x)v(x)\,\mathrm{d}x$$

i.e. if

$$a(u,v) = \langle f, v\rangle_{L^2(\Omega)} \quad \text{for all } v \in H_0^1(\Omega).$$

The existence and uniqueness of solutions problems like (12.5), under appropriate conditions on a (which of course are inherited from appropriate conditions on κ), is ensured by the Lax–Milgram theorem, which generalizes the Riesz representation theorem that any Hilbert space is isomorphic to its dual space.

Theorem 12.3 (Lax–Milgram). *Let a be a bilinear form on a Hilbert space $\mathcal{H}$, i.e. $a \in \mathcal{H}' \otimes \mathcal{H}'$, such that*

1. *(boundedness) there exists a constant $C > 0$ such that, for all $u, v \in \mathcal{H}$, $|a(u,v)| \leq C\|u\|\|v\|$; and*
2. *(coercivity) there exists a constant $c > 0$ such that, for all $v \in \mathcal{H}$, $|a(v,v)| \geq c\|v\|^2$.*

Then, for all $f \in \mathcal{H}'$, there exists a unique $u \in \mathcal{H}$ such that, for all $v \in \mathcal{H}$, $a(u,v) = \langle f \,|\, v\rangle$. Furthermore, u satisfies the estimate $\|u\|_{\mathcal{H}} \leq c^{-1}\|f\|_{\mathcal{H}'}$.

Proof. For each $u \in \mathcal{H}$, $v \mapsto a(u,v)$ is a bounded linear functional on $\mathcal{H}$. So, by the Riesz representation theorem (Theorem 3.11), given $u \in \mathcal{H}$, there is a unique $w \in \mathcal{H}$ such that $\langle w, \cdot\,\rangle = a(u, \cdot\,)$. Define $Au := w$. This defines a well-defined function $A\colon \mathcal{H} \to \mathcal{H}$, the properties of which we now check:

- A is linear. Take $\alpha_1, \alpha_2 \in \mathbb{R}$ and $u_1, u_2 \in \mathcal{H}$:

$$\begin{aligned}\langle A(\alpha_1 u_1 + \alpha_2 u_2), v\rangle &= a(\alpha_1 u_1 + \alpha_2 u_2, v)\\ &= \alpha_1 a(u_1, v) + \alpha_2 a(u_2, v)\\ &= \alpha_1\langle Au_1, v\rangle + \alpha_2\langle Au_2, v\rangle\\ &= \langle \alpha_1 Au_1 + \alpha_2 Au_2, v\rangle.\end{aligned}$$

- A is a bounded (i.e. continuous) map, since, for any $u \in \mathcal{H}$,

$$\|Au\|^2 = \langle Au, Au\rangle = a(u, Au) \leq C\|u\|\|Au\|,$$

  so $\|Au\| \leq C\|u\|$.

- A is injective, since, for any $u \in \mathcal{H}$,

$$\|Au\|\|u\| \geq |\langle Au, u\rangle| = |a(u,u)| \geq c\|u\|^2,$$

  so $Au = 0 \implies u = 0$.

- The range of A, $\mathcal{R}(A) \subseteq \mathcal{H}$, is closed. Consider a convergent sequence $(v_n)_{n\in\mathbb{N}}$ in $\mathcal{R}(A)$ that converges to some $v \in \mathcal{H}$. Choose $u_n \in \mathcal{H}$ such that $Au_n = v_n$ for each $n \in \mathbb{N}$. The sequence $(Au_n)_{n\in\mathbb{N}}$ is Cauchy, so

$$\begin{aligned}\|Au_n - Au_m\|\|u_n - u_m\| &\geq |\langle Au_n - Au_m, u_n - u_m\rangle|\\ &= |a(u_n - u_m, u_n - u_m)|\\ &\geq c\|u_n - u_m\|^2.\end{aligned}$$

  So $c\|u_n - u_m\| \leq \|v_n - v_m\| \to 0$. So $(u_n)_{n\in\mathbb{N}}$ is Cauchy and converges to some $u \in \mathcal{H}$. So $v_n = Au_n \to Au = v$ by the continuity (boundedness) of A, so $v \in \mathcal{R}(A)$, and so $\mathcal{R}(A)$ is closed.

- Finally, A is surjective. Since $\mathcal{H}$ is Hilbert and $\mathcal{R}(A)$ is closed, if $\mathcal{R}(A) \neq \mathcal{H}$, then there must exist some non-zero $s \in \mathcal{H}$ such that $s \perp \mathcal{R}(A)$. But then

$$c\|s\|^2 \leq a(s,s) = \langle s, As \rangle = 0,$$

so $s = 0$, a contradiction.

So, take $f \in \mathcal{H}'$. By the Riesz representation theorem, there is a unique $w \in \mathcal{H}$ such that $\langle w, v \rangle = \langle f \,|\, v \rangle$ for all $v \in \mathcal{H}$. The equation $Au = w$ has a unique solution u since A is invertible. So $\langle Au, v \rangle = \langle f \,|\, v \rangle$ for all $v \in \mathcal{H}$. But $\langle Au, v \rangle = a(u,v)$. So there is a unique $u \in \mathcal{H}$ such that $a(u,v) = \langle f \,|\, v \rangle$.

The proof of the estimate $\|u\|_{\mathcal{H}} \leq c^{-1}\|f\|_{\mathcal{H}'}$ is left as an exercise (Exercise 12.1). □

Galerkin Projection. Now consider the problem of finding a good approximation to u in a prescribed subspace $\mathcal{V}_M \subseteq \mathcal{H}$ of finite dimension[12.1] — as we must necessarily do when working discretely on a computer. We could, of course, consider the optimal approximation to u in $\mathcal{V}_M$, namely the orthogonal projection of u onto $\mathcal{V}_M$. However, since u is not known a priori, and in any case cannot be stored to arbitrary precision on a computer, this 'optimal' approximation is not much use in practice.

An alternative approach to approximating u is Galerkin projection: we seek a *Galerkin solution* $u_\Gamma = u_\Gamma^{(M)} \in \mathcal{V}_M$, an approximation to the exact solution u, such that

$$a(u_\Gamma, v^{(M)}) = \langle f \,|\, v^{(M)} \rangle \quad \text{for all } v^{(M)} \in \mathcal{V}_M. \tag{12.6}$$

Note that if the hypotheses of the Lax–Milgram theorem are satisfied on the full space $\mathcal{H}$, then they are certainly satisfied on the subspace $\mathcal{V}_M$, thereby ensuring the existence and uniqueness of solutions to the Galerkin problem. Note well, though, that existence of a unique Galerkin solution for each $M \in \mathbb{N}_0$ does *not* imply the existence of a unique weak solution (nor even multiple weak solutions) to the full problem; for this, one typically needs to show that the Galerkin approximations are uniformly bounded and appeal to a Sobolev embedding theorem to extract a convergent subsequence.

Example 12.4. 1. The *Fourier basis* $\{e_k\}_{k \in \mathbb{Z}}$ of the space $L^2_{\text{per}}([0, 2\pi], \mathrm{d}x; \mathbb{C})$ of complex-valued 2π-periodic functions on $[0, 2\pi]$ is defined by

$$e_k(x) = \frac{1}{\sqrt{2\pi}} \exp(ikx).$$

For Galerkin projection, one can use the $(2M+1)$-dimensional subspace

$$\mathcal{V}_M := \operatorname{span}\{e_{-M}, \ldots, e_{-1}, e_0, e_1, \ldots, e_M\}$$

of functions that are band-limited to contain frequencies at most M. In case of real-valued functions, one can use the functions

$$\begin{aligned} x &\mapsto \cos(kx), && \text{for } k \in \mathbb{N}_0, \\ x &\mapsto \sin(kx), && \text{for } k \in \mathbb{N}. \end{aligned}$$

2. Fix a partition $a = x_0 < x_1 < \cdots < x_M = b$ of a compact interval $[a, b] \subsetneq \mathbb{R}$ and consider the associated *tent functions* defined by

$$\phi_m(x) := \begin{cases} 0, & \text{if } x \leq a \text{ or } x \leq x_{m-1}; \\ \dfrac{x - x_{m-1}}{x_m - x_{m-1}}, & \text{if } x_{m-1} \leq x \leq x_m; \\ \dfrac{x_{m+1} - x}{x_{m+1} - x_m}, & \text{if } x_m \leq x \leq x_{m+1}; \\ 0, & \text{if } x \geq b \text{ or } x \geq x_{m+1}. \end{cases}$$

[12.1] Usually, but not always, the convention will be that $\dim \mathcal{V}_M = M$; sometimes, alternative conventions will be followed.

The function ϕ_m takes the value 1 at x_m and decays linearly to 0 along the two line segments adjacent to x_m. The $(M+1)$-dimensional vector space $\mathcal{V}_M := \operatorname{span}\{\phi_0, \dots, \phi_M\}$ consists of all continuous functions on $[a, b]$ that are piecewise affine on the partition, i.e. have constant derivative on each of the open intervals (x_{m-1}, x_m). The space $\widetilde{\mathcal{V}}_M := \operatorname{span}\{\phi_1, \dots, \phi_{M-1}\}$ consists of the continuous functions that piecewise affine on the partition and take the value 0 at a and b; hence, $\widetilde{\mathcal{V}}_M$ is one good choice for a finite-dimensional space to approximate the Sobolev space $H_0^1([a, b])$. More generally, one could consider tent functions associated to any simplicial mesh in $\mathbb{R}^n$.

Another viewpoint on the Galerkin solution is to see u_Γ as the projection $P\tilde{u}$ of some $\tilde{u} \in \mathcal{H}$, where $P\colon \mathcal{H} \to \mathcal{V}_M$ denotes projection (truncation), and the adjoint operator P^* is the inclusion map in the other direction. Suppose for simplicity that the operator A corresponding to the bilinear form a, as constructed in the proof of the Lax–Milgram theorem, is a self-adjoint operator. If we were to try to minimize the A-weighted norm of the residual, i.e.

$$\text{find } \tilde{u} \in \mathcal{H} \text{ to minimize } \|P\tilde{u} - u\|_A,$$

then Corollary 4.24 says that $\tilde{u}$ satisfies the normal equations

$$\begin{aligned} & P^*AP\tilde{u} = P^*Au \\ \text{i.e.} \qquad & P^*Au^{(M)} = P^*f, \end{aligned}$$

and the weak interpretation of this equation in $\mathcal{H}'$ is that it should hold as an equality of scalars whenever it is tested against any $v \in \mathcal{H} \cong \mathcal{H}''$,

$$\begin{aligned} \text{i.e.} \qquad & \langle v \,|\, P^*Au_\Gamma \rangle = \langle v \,|\, P^*f \rangle && \text{for all } v \in \mathcal{H}, \\ \text{i.e.} \qquad & \langle Pv \,|\, Au_\Gamma \rangle = \langle Pv \,|\, f \rangle && \text{for all } v \in \mathcal{H}, \\ \text{i.e.} \qquad & \langle v^{(M)} \,|\, Au_\Gamma \rangle = \langle v^{(M)} \,|\, f \rangle && \text{for all } v^{(M)} \in \mathcal{V}_M. \end{aligned}$$

Writing these dual pairings as inner products in $\mathcal{H}$ yields that the weak form of the normal equations is

$$\langle Au_\Gamma, v^{(M)} \rangle = \langle f, v^{(M)} \rangle \quad \text{for all } v^{(M)} \in \mathcal{V}_M,$$

and since $\langle Au_\Gamma, v^{(M)} \rangle = a(u_\Gamma, v^{(M)})$, this is exactly the Galerkin problem (12.6) for u_Γ. That is, the Galerkin problem (12.6) for u_Γ is the weak formulation of the variational problem of minimizing the norm of the difference between the approximate solution and the true one, with the norm being weighted by the operator corresponding to the bilinear form a.

From this variational characterization of the Galerkin solution, it follows immediately that the error $u - u_\Gamma$ is a-orthogonal to the approximation subspace $\mathcal{V}_M$: for any choice of $v^{(M)} \in \mathcal{V}_M \subseteq \mathcal{H}$,

$$\begin{aligned} a(u - u_\Gamma, v^{(M)}) &= a(u, v^{(M)}) - a(u_\Gamma, v^{(M)}) \\ &= \langle f \,|\, v^{(M)} \rangle - \langle f \,|\, v^{(M)} \rangle \\ &= 0. \end{aligned}$$

However, note well that u_Γ is generally *not* the optimal approximation of u from the subspace $\mathcal{V}_M$ with respect to the original Hilbert norm $\|\cdot\|$ on $\mathcal{H}$, i.e.

$$\left\|u - u_\Gamma\right\| \neq \inf\left\{ \left\|u - v^{(M)}\right\| \,\middle|\, v^{(M)} \in \mathcal{V}_M \right\}.$$

The optimal approximation of u from $\mathcal{V}_M$ is the orthogonal projection of u onto $\mathcal{V}_M$; if $\mathcal{H}$ has an orthonormal basis $\{e_n\}$ and $u = \sum_{n \in \mathbb{N}} u^n e_n$, then the optimal approximation of u in $\mathcal{V}_M = \operatorname{span}\{e_1, \dots, e_M\}$ is $\sum_{n=1}^{M} u^n e_n$, but this is not generally the same as the Galerkin solution u_Γ. However, the next result, Céa's lemma, shows that u_Γ is a quasi-optimal approximation to u (note that the ratio C/c is always at least 1):

Lemma 12.5 (Céa's lemma)**.** *Let a, c and C be as in the statement of the Lax–Milgram theorem. Then the weak solution $u \in \mathcal{H}$ and the Galerkin solution $u_\Gamma \in \mathcal{V}_M$ satisfy*

$$\|u - u_\Gamma\| \leq \frac{C}{c} \inf \left\{ \|u - v^{(M)}\| \,\middle|\, v^{(M)} \in \mathcal{V}_M \right\}.$$

Proof. Exercise 12.3. □

Matrix Form. It is helpful to cast the Galerkin problem in the form of a matrix-vector equation by expressing it in terms of a basis $\{\phi_1, \ldots, \phi_M\}$ of $\mathcal{V}_M$. Then $u = u_\Gamma$ solves the Galerkin problem if and only if

$$a(u, \phi_m) = \langle f, \phi_m \rangle \text{ for } m \in \{1, \ldots, M\}.$$

Now expand u in this basis as $u = \sum_{m=1}^M u_m \phi_m$ and insert this into the previous equation:

$$a\left(\sum_{m=1}^M u_m \phi_m, \phi_i\right) = \sum_{m=1}^M u_m a(\phi_m, \phi_i) = \langle f, \phi_i \rangle \text{ for } i \in \{1, \ldots, M\}.$$

That is, the column vector $\boldsymbol{u} := [u_1, \ldots, u_M]^\top \in \mathbb{R}^M$ of coefficients of u in the basis $\{\phi_1, \ldots, \phi_M\}$ solves the matrix-vector equation

$$\boldsymbol{a}\boldsymbol{u} = \boldsymbol{b} := \begin{bmatrix} \langle f, \phi_1 \rangle \\ \vdots \\ \langle f, \phi_M \rangle \end{bmatrix} \tag{12.7}$$

where the matrix

$$\boldsymbol{a} := \begin{bmatrix} a(\phi_1, \phi_1) & \ldots & a(\phi_M, \phi_1) \\ \vdots & \ddots & \vdots \\ a(\phi_1, \phi_M) & \ldots & a(\phi_M, \phi_M) \end{bmatrix} \in \mathbb{R}^{M \times M}$$

is the *Gram matrix* of the bilinear form a, and is of course a symmetric matrix whenever a is a symmetric bilinear form.

Remark 12.6. In practice the matrix-vector equation $\boldsymbol{a}\boldsymbol{u} = \boldsymbol{b}$ is *never* solved by explicitly inverting the Gram matrix $\boldsymbol{a}$ to obtain the coefficients u_m via $\boldsymbol{u} = \boldsymbol{a}^{-1}\boldsymbol{b}$. Even a relatively naïve solution using a Cholesky factorization of the Gram matrix and forward and backward substitution would be cheaper and more numerically stable than an explicit inversion. Indeed, in many situations the Gram matrix is sparse, and so solution methods that take advantage of that sparsity are used; furthermore, for large systems, the methods used are often iterative rather than direct.

12.2 Stochastic Galerkin Projection

Stochastic Lax–Milgram Theory. The next step is to build appropriate Lax–Milgram theory and Galerkin projection for stochastic problems, for which a good prototype is

$$\begin{aligned} -\nabla \cdot (\kappa(\theta, x) \nabla u(\theta, x)) &= f(\theta, x) && \text{for } x \in \Omega, \\ u(x) &= 0 && \text{for } x \in \partial\Omega, \end{aligned}$$

with θ being drawn from some probability space $(\Theta, \mathscr{F}, \mu)$. To that end, we introduce a stochastic space $\mathcal{S}$, which in the following will be $L^2(\Theta, \mu; \mathbb{R})$. We retain also a Hilbert space $\mathcal{V}$ in which the deterministic solution $u(\theta)$ is sought for each $\theta \in \Theta$; implicitly, $\mathcal{V}$ is independent of the problem data, or rather of θ. Thus, the space in which the stochastic solution U is sought is the tensor product Hilbert space $\mathcal{H} := \mathcal{V} \otimes \mathcal{S}$, which is isomorphic to the space $L^2(\Theta, \mu; \mathcal{V})$ of square-integrable $\mathcal{V}$-valued random variables.

In terms of bilinear forms, the setup is that of a bilinear-form-on-$\mathcal{V}$-valued random variable A and a $\mathcal{V}'$-valued random variable F. Define a bilinear form α on $\mathcal{H}$ by

$$\alpha(X, Y) := \mathbb{E}_\mu[A(X, Y)] \equiv \int_\Theta A(\theta)\big(X(\theta), Y(\theta)\big)\, \mathrm{d}\mu(\theta)$$

and, similarly, a linear functional β on $\mathcal{H}$ by

$$\langle \beta \,|\, Y \rangle := \mathbb{E}_\mu[\langle F \,|\, Y \rangle_{\mathcal{V}}].$$

Clearly, if α satisfies the boundedness and coercivity assumptions of the Lax–Milgram theorem on $\mathcal{H}$, then, for every $F \in L^2(\Theta, \mu; \mathcal{V}')$, there is a unique weak solution $U \in L^2(\Theta, \mu; \mathcal{V})$ satisfying

$$\alpha(U, Y) = \langle \beta \,|\, Y \rangle \text{ for all } Y \in L^2(\Theta, \mu; \mathcal{V}).$$

A sufficient, but not necessary, condition for α to satisfy the hypotheses of the Lax–Milgram theorem on $\mathcal{H}$ is for $A(\theta)$ to satisfy those hypotheses uniformly in θ on $\mathcal{V}$:

Theorem 12.7 (Stochastic Lax–Milgram theorem). *Let $(\Theta, \mathscr{F}, \mu)$ be a probability space, and let A be a random variable on Θ, taking values in the space of bilinear forms on a Hilbert space $\mathcal{V}$, and satisfying the hypotheses of the deterministic Lax–Milgram theorem (Theorem 12.3) uniformly with respect to $\theta \in \Theta$. Define a bilinear form α and a linear functional β on $L^2(\Theta, \mu; \mathcal{V})$ by*

$$\begin{aligned} \alpha(X, Y) &:= \mathbb{E}_\mu[A(X, Y)], \\ \langle \beta \,|\, Y \rangle &:= \mathbb{E}_\mu[\langle F \,|\, Y \rangle_{\mathcal{V}}]. \end{aligned}$$

Then, for every $F \in L^2(\Theta, \mu; \mathcal{V}')$, there is a unique $U \in L^2(\Theta, \mu; \mathcal{V})$ such that

$$\alpha(U, V) = \langle \beta \,|\, V \rangle \quad \textit{for all } V \in L^2(\Theta, \mu; \mathcal{V}).$$

Proof. Suppose that $A(\theta)$ satisfies the boundedness assumption with constant $C(\theta)$ and the coercivity assumption with constant $c(\theta)$. By hypothesis,

$$\begin{aligned} C' &:= \sup_{\theta \in \Theta} C(\theta) \quad \text{and} \\ c' &:= \inf_{\theta \in \Theta} c(\theta) \end{aligned}$$

are both strictly positive and finite. Then α satisfies, for all $X, Y \in \mathcal{H}$,

$$\begin{aligned} \alpha(X, Y) &= \mathbb{E}_\mu[A(X, Y)] \\ &\leq \mathbb{E}_\mu\big[C \|X\|_{\mathcal{V}} \|Y\|_{\mathcal{V}}\big] \\ &\leq C' \mathbb{E}_\mu\big[\|X\|_{\mathcal{V}}^2\big]^{1/2} \mathbb{E}_\mu\big[\|Y\|_{\mathcal{V}}^2\big]^{1/2} \\ &= C' \|X\|_{\mathcal{H}} \|Y\|_{\mathcal{H}}, \end{aligned}$$

and

$$\begin{aligned} \alpha(X, X) &= \mathbb{E}_\mu[A(X, X)] \\ &\geq \mathbb{E}_\mu\big[c \|X\|_{\mathcal{V}}^2\big] \\ &\geq c' \|X\|_{\mathcal{H}}^2. \end{aligned}$$

Hence, by the deterministic Lax–Milgram theorem applied to the bilinear form α on the Hilbert space $\mathcal{H}$, for every $F \in L^2(\Theta, \mu; \mathcal{V}')$, there exists a unique $U \in L^2(\Theta, \mu; \mathcal{V})$ such that

$$\alpha(U, V) = \langle \beta \,|\, V \rangle \text{ for all } V \in L^2(\Theta, \mu; \mathcal{V}).$$

□

Remark 12.8. Note, however, that uniform boundedness and coercivity of A are not necessary for α to be bounded and coercive. For example, the constants $c(\theta)$ and $C(\theta)$ may degenerate to 0 or ∞ as θ approaches certain points of the sample space Θ. Provided that these degeneracies are integrable and yield positive and finite expected values, this will not ruin the boundedness and coercivity of α. Indeed, there may be an arbitrarily large (but μ-measure zero) set of θ for which there is no weak solution $u(\theta)$ to the deterministic problem

$$A(\theta)(u(\theta), v) = \langle F(\theta) \,|\, v \rangle \text{ for all } v \in \mathcal{V}.$$

Stochastic Galerkin Projection. Let $\mathcal{V}_M$ be a finite-dimensional subspace of $\mathcal{V}$, with basis $\{\phi_1, \dots, \phi_M\}$. As indicated above, take the stochastic space $\mathcal{S}$ to be $L^2(\Theta, \mu; \mathbb{K})$, which we assume to be equipped with an orthogonal decomposition such as a gPC decomposition. Let $\mathcal{S}_K$ be a finite-dimensional subspace of $\mathcal{S}$, for example the span of a system of orthogonal polynomials up to degree K. The Galerkin projection of the stochastic problem on $\mathcal{H}$ is to find

$$U = U_\Gamma^{M,K} = \sum_{\substack{m=1,\dots m \\ k=0,\dots,K}} u_{mk} \phi_m \otimes \Psi_k \in \mathcal{V}_M \otimes \mathcal{S}_K$$

such that

$$\alpha(U, V) = \langle \beta \,|\, V \rangle \text{ for all } V \in \mathcal{V}_M \otimes \mathcal{S}_K.$$

In particular, it suffices to find U that satisfies this condition for each basis element $V = \phi_n \otimes \Psi_\ell$ of $\mathcal{V}_M \otimes \mathcal{S}_K$. Recall that $\phi_n \otimes \Psi_\ell$ is the function $(\theta, x) \mapsto \phi_n(x)\Psi_\ell(\theta)$.

Matrix Form. Let $\boldsymbol{\alpha} \in \mathbb{R}^{M(K+1) \times M(K+1)}$ be the Gram matrix of the bilinear form α with respect to the basis $\{\phi_m \otimes \Psi_k \mid m = 1, \dots M; k = 0, \dots, K\}$ of $\mathcal{V}_M \otimes \mathcal{S}_K$. As before, the Galerkin problem is equivalent to the matrix-vector equation

$$\boldsymbol{\alpha} \boldsymbol{U} = \boldsymbol{\beta},$$

where $\boldsymbol{U} \in \mathbb{R}^{M(K+1)}$ is the column vector comprised of the coefficients u_{mk} and $\boldsymbol{\beta} \in \mathbb{R}^{M(K+1)}$ has components $\langle \beta \,|\, \phi_m \otimes \Psi_k \rangle$. An obvious question is how the Gram matrix $\boldsymbol{\alpha}$ is related to the $\mathbb{R}^{M \times M}$-valued random variable $\boldsymbol{A}$ that is the Gram matrix of the random bilinear form A.

Suppose that, for each fixed $\theta \in \Theta$, the deterministic problem, discretized and written in matrix-vector form in the basis $\{\phi_1, \dots, \phi_M\}$ of $\mathcal{V}_M$, is

$$\boldsymbol{A}(\theta)\boldsymbol{U}(\theta) = \boldsymbol{B}(\theta).$$

Here, the Galerkin solution is $U(\theta) \in \mathcal{V}_M$ and $\boldsymbol{U}(\theta) \in \mathbb{R}^M$ is the column vector of coefficients of $U(\theta)$ with respect to $\{\phi_1, \dots, \phi_M\}$. Write the Galerkin solution $U \in \mathcal{V}_M \otimes \mathcal{S}_K$ as $U = \sum_{k=0}^K u_k \Psi_k$, and further write $\boldsymbol{u}_k \in \mathbb{R}^M$ for the column vector corresponding to the stochastic mode $u_k \in \mathcal{V}_M$ in the basis $\{\phi_1, \dots, \phi_M\}$, so that $\boldsymbol{U} = \sum_{k=0}^K \boldsymbol{u}_k \Psi_k$. Galerkin projection — more specifically, testing the equation $\boldsymbol{A}\boldsymbol{U} = \boldsymbol{B}$ against Ψ_k — reveals that

$$\sum_{j=0}^K \langle \Psi_k \boldsymbol{A} \Psi_j \rangle \boldsymbol{u}_j = \langle \boldsymbol{B} \Psi_k \rangle \text{ for each } k \in \{0, \dots, K\}.$$

This is equivalent to the (large!) block system

$$\begin{bmatrix} \langle \boldsymbol{A} \rangle_{00} & \dots & \langle \boldsymbol{A} \rangle_{0K} \\ \vdots & \ddots & \vdots \\ \langle \boldsymbol{A} \rangle_{K0} & \dots & \langle \boldsymbol{A} \rangle_{KK} \end{bmatrix} \begin{bmatrix} \boldsymbol{u}_0 \\ \vdots \\ \boldsymbol{u}_K \end{bmatrix} = \begin{bmatrix} \langle \boldsymbol{B} \Psi_0 \rangle \\ \vdots \\ \langle \boldsymbol{B} \Psi_K \rangle \end{bmatrix}, \tag{12.8}$$

where, for $0 \leq i, j \leq K$,

$$\langle \boldsymbol{A} \rangle_{ij} := \langle \Psi_i \boldsymbol{A} \Psi_j \rangle \in \mathbb{R}^{M \times M}.$$

Note that, in general, the stochastic modes u_j of the solution U (and, indeed the coefficients u_{jm} of the stochastic modes in the deterministic basis $\{\phi_1, \ldots, \phi_M\}$) are all coupled together through the matrix on the left-hand side of (12.8). This can be a limitation of stochastic Galerkin methods, and will be remarked upon later.

Remark 12.9. Note well that the entries $\langle \boldsymbol{B} \Psi_k \rangle$ on the right-hand side of (12.8) are *not* the stochastic modes $\boldsymbol{b}_k \in \mathbb{R}^M$ of $\boldsymbol{B}$, since they have not been normalized by $\langle \Psi_k^2 \rangle$.

Example 12.10. As a special case, suppose that the random data have no impact on the differential operator and affect only the right-hand side $B = \sum_{k \in \mathbb{N}_0} b_k \Psi_k$. In this case the random bilinear form $\theta \mapsto A(\theta)(\cdot, \cdot)$ is identically equal to one bilinear form $a(\cdot, \cdot)$, so the random Gram matrix $\boldsymbol{A}$ is a deterministic matrix $\boldsymbol{a}$, and so the blocks $\langle \boldsymbol{A} \rangle_{ij}$ in (12.8) are given by

$$\langle \boldsymbol{A} \rangle_{ij} := \langle \Psi_i \boldsymbol{a} \Psi_j \rangle = \boldsymbol{a} \langle \Psi_i \Psi_j \rangle = \boldsymbol{a} \delta_{ij} \langle \Psi_i^2 \rangle.$$

Hence, the stochastic Galerkin system, in its matrix form (12.8), becomes the block-diagonal system

$$\begin{bmatrix} \boldsymbol{a} & \boldsymbol{0} & \ldots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{a}\langle \Psi_1^2 \rangle & \ddots & \vdots \\ \vdots & \ddots & \ddots & \boldsymbol{0} \\ \boldsymbol{0} & \ldots & \boldsymbol{0} & \boldsymbol{a}\langle \Psi_K^2 \rangle \end{bmatrix} \begin{bmatrix} \boldsymbol{u}_0 \\ \boldsymbol{u}_1 \\ \vdots \\ \boldsymbol{u}_K \end{bmatrix} = \begin{bmatrix} \langle \boldsymbol{B} \Psi_0 \rangle \\ \langle \boldsymbol{B} \Psi_1 \rangle \\ \vdots \\ \langle \boldsymbol{B} \Psi_K \rangle \end{bmatrix}.$$

Thus, in this case, the stochastic modes u_j decouple and are given by

$$\boldsymbol{u}_k = \boldsymbol{a}^{-1} \frac{\langle \boldsymbol{B} \Psi_k \rangle}{\langle \Psi_k^2 \rangle} = \boldsymbol{a}^{-1} \boldsymbol{b}_k.$$

Thus, in this case, the any pre-existing solver for the deterministic problem $\boldsymbol{a}\boldsymbol{u} = \boldsymbol{b}$ can simply be re-used 'as is' $K + 1$ times with $\boldsymbol{b} = \boldsymbol{b}_k$ for $k = 0, \ldots, K$ to obtain the Galerkin solution of the stochastic problem.

The Galerkin Multiplication Tensor. In contrast to Example 12.10, in which the differential operator is deterministic, we can consider the case in which the random Gram matrix $\boldsymbol{A}$ has a (truncated) gPC expansion

$$\boldsymbol{A} = \sum_{k=0}^{K} \boldsymbol{A}_k \Psi_k$$

with coefficient matrices

$$\boldsymbol{A}_k = \frac{\langle \boldsymbol{A} \Psi_k \rangle}{\langle \Psi_k^2 \rangle} \in \mathbb{R}^{M \times M}.$$

In this case, the blocks $\langle \boldsymbol{A} \rangle_{kj}$ in (12.8) are given by

$$\langle \boldsymbol{A} \rangle_{kj} = \langle \Psi_k \boldsymbol{A} \Psi_j \rangle = \sum_{i=0}^{K} \boldsymbol{A}_i \langle \Psi_i \Psi_j \Psi_k \rangle.$$

Hence, the Galerkin block system (12.8) is equivalent to

$$\begin{bmatrix} \widetilde{\langle \boldsymbol{A} \rangle}_{00} & \ldots & \widetilde{\langle \boldsymbol{A} \rangle}_{0K} \\ \vdots & \ddots & \vdots \\ \widetilde{\langle \boldsymbol{A} \rangle}_{K0} & \ldots & \widetilde{\langle \boldsymbol{A} \rangle}_{KK} \end{bmatrix} \begin{bmatrix} \boldsymbol{u}_0 \\ \vdots \\ \boldsymbol{u}_K \end{bmatrix} = \begin{bmatrix} \boldsymbol{b}_0 \\ \vdots \\ \boldsymbol{b}_K \end{bmatrix}, \tag{12.9}$$

where $\boldsymbol{b}_k = \frac{\langle \boldsymbol{B}\Psi_k \rangle}{\langle \Psi_k^2 \rangle} \in \mathbb{R}^M$ is the column vector of coefficients of the k^{th} stochastic mode b_k of B in the basis $\{\phi_1, \dots, \phi_M\}$ of $\mathcal{V}_M$, and

$$\widetilde{\langle \boldsymbol{A} \rangle}_{kj} := \sum_{i=0}^{K} \boldsymbol{A}_i M_{ijk},$$
$$M_{ijk} := \frac{\langle \Psi_i \Psi_j \Psi_k \rangle}{\langle \Psi_k^2 \rangle}.$$

Note that while the matrix in (12.8) is block-symmetric, since clearly $\langle \boldsymbol{A} \rangle_{ij} := \langle \Psi_i \boldsymbol{A} \Psi_j \rangle = \langle \boldsymbol{A} \rangle_{ji}$, the matrix in (12.9) is *not* block-symmetric, since the k^{th} block row is normalized by $\langle \Psi_k^2 \rangle$, and in general the normalizing factors for each block row will be distinct. Nevertheless, formulation (12.9) has some advantages: the properly-normalized stochastic modes of $\boldsymbol{A}$, $\boldsymbol{U}$ and $\boldsymbol{B}$ appear throughout, and the tensor M_{ijk} is a central object that recurs in the study of many derived quantities of Galerkin approximations, e.g. the treatment of nonlinearities in the next section.

Definition 12.11. Let $\{\Psi_k\}_{k \in \mathbb{N}_0}$ be a system of orthogonal polynomials for a measure μ. The associated *multiplication tensor*[12.2] (or *Galerkin tensor*) is the rank-3 tensor M_{ijk}, $(i, j, k) \in \mathbb{N}_0^3$, defined by

$$M_{ijk} := \frac{\langle \Psi_i \Psi_j \Psi_k \rangle}{\langle \Psi_k \Psi_k \rangle}.$$

By mild abuse of notation, we also write M_{ijk} for the finite-dimensional rank-3 tensor defined by the same formula for $0 \leq i, j, k \leq K$.

Observe that the multiplication tensor M_{ijk} is symmetric in the first two indices (i.e. $M_{ijk} = M_{jik}$). Furthermore, since $\{\Psi_k\}_{k \in \mathbb{N}_0}$ is a system of μ-orthogonal polynomials, many of the entries of M_{ijk} are zero, leading to sparsity for the matrix problem: for example,

$$\widetilde{\langle \boldsymbol{A} \rangle}_{00} = \sum_{k=0}^{K} \boldsymbol{A}_k M_{k00} = \boldsymbol{A}_0.$$

Note that the multiplication tensor is determined entirely by the gPC basis $\{\Psi_k\}_{k \in \mathbb{N}_0}$ and the measure μ, and so while there is a significant computational cost associated to evaluating its entries, this is a one-time cost: the multiplication tensor can be pre-computed, stored, and then used for many different problems, i.e. many As and Bs. In a few special cases, the multiplication tensor can be calculated in closed form: see e.g. Exercise 12.5. In other cases, it is necessary to resort to numerical integration; note, however, that since Ψ_k is a polynomial, so is $\Psi_i \Psi_j \Psi_k$, and hence the multiplication tensor can be evaluated numerically but exactly by Gauss quadrature once the orthogonal polynomials of sufficiently high degree and their zeros have been identified.

Example 12.12. Consider random variables $Y, B \in L^2(\Theta, \mu; \mathbb{R})$ and the random linear first-order ordinary differential equation

$$\frac{\mathrm{d}U(t)}{\mathrm{d}t} = -YU(t), \quad U(0) = B, \tag{12.10}$$

for $U \colon [0, T] \times \Theta \to \mathbb{R}$. This ODE can be used as a simple model for the amount of radiation emitted by a sample with decay constant Y, i.e. half-life $\frac{1}{Y} \log 2$; the initial level of radiation emission at time $t = 0$ is B. Let $\{\Psi_k\}_{k \in \mathbb{N}_0}$ be an orthogonal basis for $L^2(\Theta, \mu; \mathbb{R})$ with the usual convention that $\Psi_0 = 1$. Suppose that our knowledge about Y and B is encoded in the

[12.2] Careful readers will also note that M_{ijk} is covariant in the indices i and j and contravariant in the index k; therefore, if this text were following standard tensor algebra notation and writing vectors as $\sum_k u^k \Psi_k$, then the multiplication tensor would be denoted M_{ij}^k. In terms of the dual basis $\{\Psi^k \mid k \in \mathbb{N}_0\}$ defined by $\langle \Psi^k \mid \Psi_\ell \rangle = \delta_\ell^k$, $M_{ij}^k = \langle \Psi^k \mid \Psi_i \Psi_j \rangle$.

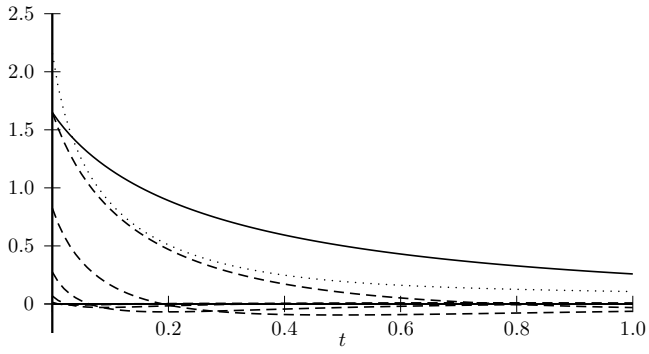


Figure 12.1: The degree-10 Hermite PC Galerkin solution to the random ODE (12.10), with log-normally distributed decay constant and initial condition. The mean of the solution is shown in black, the higher-degree Hermite coefficients dashed, and the standard deviation of the solution dotted. Note that, on these axes, only the coefficients of degree ≤ 5 are visible; the others are all of order 10^{-2} or smaller.

gPC expansions $Y = \sum_{k\in\mathbb{N}_0} y_k\Psi_k$, $B = \sum_{k\in\mathbb{N}_0} b_k\Psi_k$; the aim is to find the gPC expansion of $U(t) = \sum_{k\in\mathbb{N}_0} u_k(t)\Psi_k$. Projecting the evolution equation (12.10) onto the basis $\{\Psi_k\}_{k\in\mathbb{N}_0}$ yields

$$\left\langle \frac{\mathrm{d}U}{\mathrm{d}t}\Psi_k \right\rangle = -\langle YU\Psi_k\rangle \text{ for each } k \in \mathbb{N}_0.$$

Inserting the gPC expansions for Y and U into this yields, for every $k \in \mathbb{N}_0$,

$$\left\langle \sum_{j\in\mathbb{N}_0} \dot{u}_j(t)\Psi_j\Psi_k \right\rangle = -\left\langle \sum_{i\in\mathbb{N}_0} y_i\Psi_i \sum_{j\in\mathbb{N}_0} u_j(t)\Psi_i\Psi_k \right\rangle,$$

$$\text{i.e.} \qquad \dot{u}_k(t)\langle\Psi_k^2\rangle = -\sum_{i,j\in\mathbb{N}_0} y_i u_j(t)\langle\Psi_j\Psi_i\Psi_k\rangle,$$

$$\text{i.e.} \qquad \dot{u}_k(t) = -\sum_{i,j\in\mathbb{N}_0} M_{ijk} y_i u_j(t).$$

The coefficients u_k are a coupled system of countably many ordinary differential equations. If all the chaos expansions are truncated at order K, then all the above summations over $\mathbb{N}_0$ become summations over $\{0,\ldots,K\}$, yielding a coupled system of $K+1$ ordinary differential equations. In matrix-vector form, the vector $\boldsymbol{u}(t) \in \mathbb{R}^{K+1}$ of coefficients of the degree-K Galerkin solution $U_\Gamma(t)$ satisfies

$$\dot{\boldsymbol{u}}(t) = \boldsymbol{A}^\top \boldsymbol{u}(t), \quad \boldsymbol{u}(0) = \boldsymbol{b},$$

where the matrix $\boldsymbol{A} \in \mathbb{R}^{(K+1)\times(K+1)}$ has as its $(i,k)^{\text{th}}$ entry $-\sum_{j=0}^{K} M_{ijk}y_j$. See Figure 12.1 for an illustration of the evolution of the Galerkin solution to (12.10) in the Hermite basis when $\log Y, \log B \sim \mathcal{N}(0,1)$ are independent, so that, by Example 11.13, Y and B are log-normal and have Hermite PC coefficients $y_k = b_k = \sqrt{e}/k!$.

Example 12.13. Consider the simple harmonic oscillator equation

$$\ddot{U}(t) = -\Omega^2 U(t). \tag{12.11}$$

For simplicity, suppose that the initial conditions $U(0) = 1$ and $\dot{U}(0) = 0$ are known, but that Ω is stochastic. Let $\{\Psi_k\}_{k\in\mathbb{N}_0}$ be an orthogonal basis for $L^2(\Theta,\mu;\mathbb{R})$ with the usual convention that $\Psi_0 = 1$. Suppose that Ω has a gPC expansion $\Omega = \sum_{k\in\mathbb{N}_0} \omega_k\Psi_k$ and it is desired to find the gPC

expansion of U, i.e. $U(t) = \sum_{k\in\mathbb{N}_0} u_k(t)\Psi_k$. Note that the random variable $W = \Omega^2$ has a gPC expansion $W = \sum_{k\in\mathbb{N}_0} w_k\Psi_k$ with

$$w_k = \sum_{i,j\in\mathbb{N}_0} M_{ijk}\omega_i\omega_j.$$

Projecting the evolution equation (12.11) onto the basis $\{\Psi_k\}_{k\in\mathbb{N}_0}$ yields

$$\left\langle \ddot{U}(t)\Psi_k \right\rangle = -\left\langle WU(t)\Psi_k \right\rangle \text{ for each } k \in \mathbb{N}_0.$$

Inserting the chaos expansions for W and U into this yields, for every $k \in \mathbb{N}_0$,

$$\begin{aligned}
\left\langle \sum_{i\in\mathbb{N}_0} \ddot{u}_i(t)\Psi_i\Psi_k \right\rangle &= -\left\langle \sum_{j\in\mathbb{N}_0} w_j\Psi_j \sum_{i\in\mathbb{N}_0} u_i(t)\Psi_i\Psi_k \right\rangle, \\
\text{i.e.} \quad \ddot{u}_k(t)\langle\Psi_k^2\rangle &= -\sum_{i,j\in\mathbb{N}_0} w_j u_i(t)\langle\Psi_i\Psi_j\Psi_k\rangle, \\
\text{i.e.} \quad \ddot{u}_k(t) &= -\sum_{i,j\in\mathbb{N}_0} M_{ijk} w_j u_i(t).
\end{aligned}$$

If all these gPC expansions are truncated at order K, and $\boldsymbol{A} \in \mathbb{R}^{(K+1)\times(K+1)}$ is defined by

$$A_{ik} := \sum_{j=0}^{K} M_{ijk} w_j = \sum_{j,p,q=0}^{K} M_{ijk} M_{pqj} \omega_p \omega_q,$$

then the vector $\boldsymbol{u}(t)$ of coeffcients for the degree-K Galerkin solution $U_\Gamma(t)$ satisfies the vector oscillator equation

$$\ddot{\boldsymbol{u}}(t) = -\boldsymbol{A}^\top \boldsymbol{u}(t)$$

with the obvious initial conditions. See Figure 12.2 for illustrations of this Galerkin solution when the Hermite basis is used and Ω is log-normally distributed with $\log\Omega \sim \mathcal{N}(0,\sigma^2)$ for various values of $\sigma \geq 0$. Recall from Example 11.13 that the Hermite coefficients of such a log-normal Ω are $\omega_k = e^{\sigma^2/2}\sigma^k/k!$.

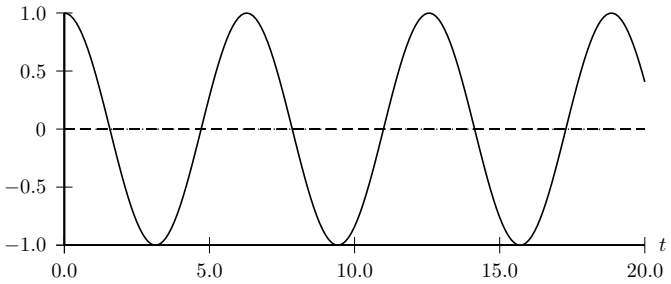
12.3 Nonlinearities

Nonlinearities of various types occur throughout practical problems, and their treatment is critical in the context of stochastic Galerkin methods, which require the projection of these nonlinearities onto the finite-dimensional solution spaces. For example, given the gPC expansion $U = \sum_{k\in\mathbb{N}_0} u_k\Psi_k$, how does one calculate the gPC coefficients of, say, U^2 or $\sqrt{U}$ in terms of those of U? More practically, given a truncated gPC expansion

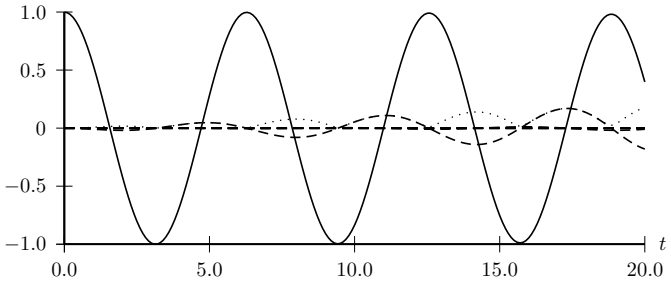
$$U(\xi) \approx U^K(\xi) = \sum_{k=0}^{K} u_k\Psi_k(\xi)$$

how does one calculate order-K coefficients of U^2 or $\sqrt{U}$? The first example, U^2, is a special case of taking the product of two gPC expansions, and can be resolved using the multiplication tensor M_{ijk} of the previous section.

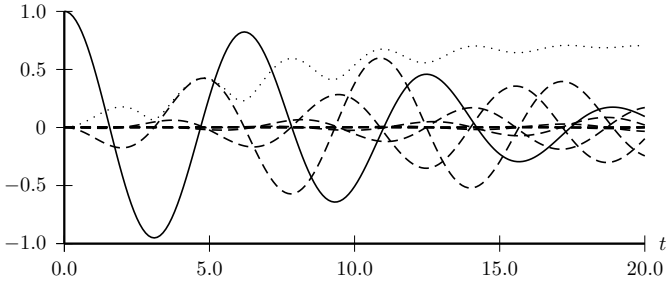
Galerkin Multiplication. The first, simplest, kind of nonlinearity to consider is the product of two or more random variables in terms of their gPC expansions. The natural question to ask is how to quickly compute the gPC coefficients of a product in terms of the gPC coefficients of the factors — particularly if expansions are truncated to finite precision. The case of a product of two random variables is quite illustrative:



(a) $\sigma = 0$, i.e. deterministic solution.



(b) $\sigma = \frac{1}{100}$.



(c) $\sigma = \frac{1}{10}$.

Figure 12.2: The degree-10 Hermite PC Galerkin solution to the simple harmonic oscillator equation $\ddot{U}(t) = -\Omega^2 U(t)$, with deterministic initial conditions $U(0) = 1$ and $\dot{U}(0) = 0$, and a log-normally distributed spring constant Ω, with $\log \Omega \sim \mathcal{N}(0, \sigma^2)$. The mean of the solution is shown in black, the higher-degree Hermite coefficients dashed, and the standard deviation of the solution dotted. Note that, in the case $\sigma = \frac{1}{10}$, the variance grows so quickly that accurate predictions of the system's state after just one or two cycles are essentially impossible.

Example 12.14. Suppose that $U = \sum_{k\in\mathbb{N}_0} u_k\Psi_k$ and $V = \sum_{k\in\mathbb{N}_0} v_k\Psi_k$ are random variables in $L^2(\Theta,\mu;\mathbb{R})$. Assuming that their product $W := UV$ is again a random variable in $L^2(\Theta,\mu;\mathbb{R})$,

$$W = \sum_{i,j\in\mathbb{N}_0} u_i v_j \Psi_i \Psi_j,$$

so the coefficients $\{w_k \mid k \in \mathbb{N}_0\}$ are given by

$$w_k = \frac{\langle W\Psi_k\rangle}{\langle\Psi_k^2\rangle} = \sum_{i,j\in\mathbb{N}_0} M_{ijk} u_i v_j.$$

It is this formula that motivates the name *multiplication tensor* for M_{ijk}. Now suppose that the expansions for U and V are truncated at order K, so that $U = \sum_{k=0}^K u_k\Psi_k$ and $V = \sum_{k=0}^K v_k\Psi_k$. Then their product $W := UV$ has the expansion

$$W = \sum_{i,j=0}^K u_i v_j \Psi_i \Psi_j.$$

Note that, while W is guaranteed to be in L^2, it is *not* necessarily in $\mathcal{S}_K$. Nevertheless, the truncated expansion $W \approx \sum_{i,j,k=0}^K M_{ijk} u_i v_j \Psi_k$ is the orthogonal projection of W onto $\mathcal{S}_K$, and hence the L^2-closest approximation of W in $\mathcal{S}_K$. It is called the *Galerkin product*, or *pseudo-spectral product*, of U and V, denoted $U *_K V$ or simply $U * V$ if it is not necessary to call attention to the order of the truncation.

The fact that multiplication of two random variables can be handled efficiently, albeit with some truncation error, in terms of their expansions in the gPC basis and the multiplication tensor is very useful: it adds to the list of reasons why one might wish to pre-computate and store the multiplication tensor of a basis for use in many problems.

However, outside the situation of binary products (and hence squares), Galerkin multiplication has undesirable features beyond simple truncation error. For example, suppose that we wish to multiply three random variables $U, V, W \in L^2(\Theta,\mu)$ in terms of their gPC expansions in a fashion similar to the Galerkin product above. First of all, it must be acknowledged that perhaps $Z := UVW \notin L^2(\Theta,\mu)$. Nevertheless, assuming that Z is, after all, square-integrable, a gPC expansion of the triple product is

$$Z = \sum_{m\in\mathbb{N}_0} z_m\Psi_m = \sum_{m\in\mathbb{N}_0} \left[\sum_{j,k,\ell\in\mathbb{N}_0} T_{jk\ell m} u_j v_k w_\ell \right] \Psi_m,$$

or an appropriate truncation of the same, where the rank-4 tensor $T_{jk\ell m}$ is defined by

$$T_{jk\ell m} := \frac{\langle \Psi_j\Psi_k\Psi_\ell\Psi_m\rangle}{\langle\Psi_m^2\rangle}.$$

This approach can be extended to higher-order multiplication. However, even with sparsity, computation and storage of these tensors — which have $(K+1)^d$ entries when working with products of d random variables to polynomial degree K — quickly becomes prohibitively expensive. Therefore, it is common to approximate the triple product in Galerkin fashion by two binary products, i.e.

$$UVW \approx U * (V * W).$$

Unfortunately, this approximation incurs additional truncation errors, since each binary multiplication discards the part orthogonal to $\mathcal{S}_K$; the terms that are discarded depend upon the order of approximate multiplication and truncation, and in general

$$U * (V * W) \neq V * (W * U) \neq W * (U * V).$$

As a result, higher-order Galerkin multiplication is generally not commutative.

Galerkin Inversion. Another common transformation that must be performed is the inversion of a random variable: given

$$U = \sum_{k\geq 0} u_k \Psi_k \approx \sum_{k=0}^{K} u_k \Psi_k$$

we seek a random variable $V = \sum_{k\geq 0} v_k \Psi_k \approx \sum_{k=0}^{K} v_k \Psi_k$ such that $U(\theta)V(\theta) = 1$ for almost every $\theta \in \Theta$. The weak interpretation of this desideratum is that $U * V = \Psi_0$. Since $U * V$ has as its k^{th} gPC coefficient $\sum_{i,j=0}^{K} M_{ijk} u_i v_j$, we arrive at the following matrix-vector equation for the gPC coefficients of V:

$$\begin{bmatrix} \sum_{i=0}^{K} M_{i00} u_i & \dots & \sum_{i=0}^{K} M_{iK0} u_i \\ \sum_{i=0}^{K} M_{i01} u_i & \dots & \sum_{i=0}^{K} M_{iK1} u_i \\ \vdots & \ddots & \vdots \\ \sum_{i=0}^{K} M_{i0K} u_i & \dots & \sum_{i=0}^{K} M_{iKK} u_i \end{bmatrix} \begin{bmatrix} v_0 \\ v_1 \\ \vdots \\ v_K \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{12.12}$$

Naturally, if $U(\theta) = 0$ for some θ, then $V(\theta)$ will be undefined for that θ. Furthermore, if $U \approx 0$ with 'too large' probability, then V may exist a.e. but fail to be in L^2. Hence, it is not surprising to learn that while (12.12) has a unique solution whenever the matrix on the left-hand is non-singular, the system becomes highly ill-conditioned as the amount of probability mass near $U = 0$ increases. Compare this with the Monte Carlo estimation of the (undefined) expected value of the reciprocal of a Gaussian random variable in Remark 9.15.

Similar ideas can be used to produce a Galerkin division algorithm for Galerkin gPC coefficients of U/V in terms of the gPC coefficients of U and V respectively; see Exercise 12.8.

More General Nonlinearities. More general nonlinearities can be treated by the methods outlined above if one knows the Taylor expansion of the nonlinearity. The standard words of warning about compounded truncation error all apply, as do warnings about slowly-convergent power series, which necessitate very high order approximation of random variables in order to accurately resolve nonlinearities even at low order.

Bibliography

Basic Lax–Milgram theory and Galerkin methods for PDEs can be found in any modern textbook on PDEs, such as those by Evans [51] (see Chapter 6) and Renardy & Rogers [138] (see Chapter 9). Such topics are also covered in the Warwick mathematics module MA4A2 Advanced PDEs. W

The book of Ghanem & Spanos [64] was influential in popularizing stochastic Galerkin methods in applications. The monograph of Xiu [182] provides a general introduction to spectral methods for uncertainty quantification, including Galerkin methods (Chapter 6), but is light on proofs. The book of Le Maître & Knio [102] covers Galerkin methods in Chapter 4, including an extensive treatment of nonlinearities in Section 4.5. Constantine & al. [35] present an interesting change of basis for the Galerkin system (12.8) from the usual basis representation to a nodal representation that enables easy comparison with the stochastic collocation methods of Chapter 13.

Exercises

Exercise 12.1. Let a be a bilinear form satisfying the hypotheses of the Lax–Milgram theorem. Given $f \in \mathcal{H}^*$, show that the unique u such that $a(u, v) = \langle f \,|\, v \rangle$ for all $v \in \mathcal{H}$ satisfies $\|u\|_{\mathcal{H}} \leq c^{-1} \|f\|_{\mathcal{H}'}$.

Exercise 12.2 (Lax–Milgram with two Hilbert spaces)**.** Let $\mathcal{U}$ and $\mathcal{V}$ be Hilbert spaces, and let $a \colon \mathcal{U} \times \mathcal{V} \to \mathbb{K}$ be a bilinear form such that there exist constants $0 < c \leq C < \infty$ such that, for all $u \in \mathcal{U}$ and $v \in \mathcal{V}$,

$$c\|u\|_{\mathcal{U}}\|v\|_{\mathcal{V}} \leq |a(u, v)| \leq C\|u\|_{\mathcal{U}}\|v\|_{\mathcal{V}}.$$

By following the steps in the proof of the usual Lax–Milgram theorem, show that, for all $f \in \mathcal{V}'$, there exists a unique $u \in \mathcal{U}$ such that, for all $v \in \mathcal{V}$, $a(u, v) = \langle f \,|\, v \rangle$, and show also that this u satisfies the estimate $\|u\|_{\mathcal{U}} \leq c^{-1} \|f\|_{\mathcal{V}'}$.

Exercise 12.3 (Céa's lemma)**.** Let a, c and C be as in the statement of the Lax–Milgram theorem. Show that the weak solution $u \in \mathcal{H}$ and the Galerkin solution $u_\Gamma \in \mathcal{V}_M$ satisfy

$$\left\| u - u_\Gamma \right\| \leq \frac{C}{c} \inf \left\{ \left\| u - v^{(M)} \right\| \,\middle|\, v^{(M)} \in \mathcal{V}_M \right\}.$$

Exercise 12.4. Consider a partition of the unit interval $[0, 1]$ into $N + 1$ equally spaced nodes

$$0 = x_0 < x_1 = h < x_2 = 2h < \cdots < x_N = 1,$$

where $h = \frac{1}{N} > 0$. For $n = 0, \ldots, N$, let

$$\phi_n(x) := \begin{cases} 0, & \text{if } x \leq 0 \text{ or } x \leq x_{n-1}; \\ (x - x_{n-1})/h, & \text{if } x_{n-1} \leq x \leq x_n; \\ (x_{n+1} - x)/h, & \text{if } x_n \leq x \leq x_{n+1}; \\ 0, & \text{if } x \geq 1 \text{ or } x \geq x_{n+1}. \end{cases}$$

What space of functions is spanned by $\phi_0, \ldots, \phi_N$? For these functions $\phi_0, \ldots, \phi_N$, calculate the Gram matrix for the bilinear form

$$a(u, v) := \int_0^1 u'(x) v'(x) \, \mathrm{d}x$$

corresponding to the Laplace operator. Determine also the vector components $\langle f, \phi_n \rangle$ in the Galerkin equation (12.7).

Exercise 12.5. Let $\gamma = \mathcal{N}(0, 1)$ be the standard Gaussian measure on $\mathbb{R}$, and let $\{\mathrm{He}_n\}_{n \in \mathbb{N}_0}$ be the associated orthogonal system of Hermite polynomials with $\langle \mathrm{He}_n^2 \rangle = n!$. Show that

$$\langle \mathrm{He}_i \mathrm{He}_j \mathrm{He}_k \rangle = \frac{i! j! k!}{(s - i)! (s - j)! (s - k)!}$$

whenever $2s = i + j + k$ is even, $i + j \geq k$, $j + k \geq i$, and $k + i \geq j$; and zero otherwise. Hence, show that the Galerkin multiplication tensor for the Hermite polynomials is

$$M_{ijk} = \begin{cases} \frac{i! j!}{(s-i)!(s-j)!(s-k)!}, & \text{if } 2s = i + j + k \in 2\mathbb{Z},\ i + j \geq k, \\ & j + k \geq i, \text{ and } k + i \geq j, \\ 0, & \text{otherwise.} \end{cases}$$

Exercise 12.6. Show that the multiplication tensor M_{ijk} is covariant in the indices i and j and contravariant in the index k. That is, if $\{\Psi_k \mid k \in \mathbb{N}_0\}$ and $\left\{\widetilde{\Psi}_k \middle| k \in \mathbb{N}_0\right\}$ are two orthogonal bases and A is the change-of-basis matrix in the sense that $\widetilde{\Psi}_j = \sum_i A_{ij} \Psi_i$, then the corresponding multiplication tensors M_{ijk} and $\widetilde{M}_{ijk}$ satisfy

$$\widetilde{M}_{ijk} = \sum_{m,n,p} A_{mi} A_{nj} (A^{-1})_{kp} M_{mnp}.$$

(Thus, the multiplication tensor is a $(2, 1)$-tensor and differential geometers would denote it by M_{ij}^k.)

Exercise 12.7. Show that, for fixed K, the Galerkin product satisfies for all $U, V, W \in \mathcal{S}_K$ and $\alpha, \beta \in \mathbb{R}$,

$$\begin{aligned} U * V &= V * U, \\ (\alpha U) * (\beta V) &= \alpha\beta (U * V), \\ (U + V) * W &= U * W + V * W. \end{aligned}$$

Exercise 12.8. Following the model of Galerkin inversion, formulate a method for calculating the Galerkin spectral coefficients of a degree-K Galerkin approximation to U/V given spectral expansions $U = \sum_{k\in\mathbb{N}_0} u_k\Psi_k$ and $V = \sum_{k\in\mathbb{N}_0} v_k\Psi_k$ that are truncated to degree K.

Chapter 13

Non-Intrusive Methods

> [W]hen people thought the Earth was flat, they were wrong. When people thought the Earth was spherical, they were wrong. But if *you* think that thinking the Earth is spherical is *just as wrong* as thinking the Earth is flat, then your view is wronger than both of them put together.
>
> *The Relativity of Wrong*
> Isaac Asimov

Chapter 12 considered a spectral approach to UQ, namely Galerkin expansion, that is mathematically very attractive in that it is a natural extension of the Galerkin methods that are commonly used for deterministic PDEs and (up to a constant) minimizes the stochastic residual, but has the severe disadvantage that the stochastic modes of the solution are coupled together by a large system such as (12.9). Hence, the Galerkin formalism is not suitable for situations in which deterministic solutions are slow and expensive to obtain, and the deterministic solution method cannot be modified. Many so-called *legacy codes* are not amenable to such *intrusive* methods of UQ.

In contrast, this chapter considers *non-intrusive* spectral methods for UQ. These are characterized by the feature that the solution $U(\theta)$ of the deterministic problem is a 'black box' that does not need to be modified for use in the spectral method, beyond being able to be evaluated at any desired point θ of the probability space $(\Theta, \mathcal{F}, \mu)$. Indeed, sometimes, it is necessary to go one step further than this and consider the case of *legacy data*, i.e. an archive of past input-output pairs $\{(\theta_n, U(\theta_n)) \mid n = 1, \ldots, N\}$, sampled according to a possibly unknown or sub-optimal strategy, that is provided 'as is' and that cannot be modified or extended at all: the reasons for such restrictions may range from financial or practical difficulties to legal and ethical concerns.

13.1 Pseudo-Spectral Methods

Consider a square-integrable stochastic process $U \colon \Theta \to \mathcal{V}$ taking values in a separable Hilbert space $\mathcal{V}$, with a spectral expansion

$$U = \sum_{k \in \mathbb{N}_0} u_k \Psi_k$$

of $U \in L^2(\Theta, \mu; \mathcal{V}) \cong \mathcal{V} \otimes L^2(\Theta, \mu; \mathbb{R})$ in terms of coefficients (stochastic modes) $u_k \in \mathcal{V}$ and an orthogonal basis $\{\Psi_k \mid k \in \mathbb{N}_0\}$ of $L^2(\Theta, \mu; \mathbb{R})$. As usual, the stochastic modes are given by

$$u_k = \frac{\langle U \Psi_k \rangle}{\langle \Psi_k^2 \rangle} = \frac{1}{\gamma_k} \int_\Theta U(\theta) \Psi_k(\theta) \, \mathrm{d}\mu(\theta).$$

If the normalization constants $\gamma_k := \langle \Psi_k^2 \rangle \equiv \|\Psi_k\|^2_{L^2(\mu)}$ are known ahead of time, then it remains only to approximate the integral with respect to μ of the product of U with each basis function Ψ_k. In some cases, the normalization constants must also be approximated.

Deterministic Quadrature. If the dimension of Θ is low and $U(\theta)$ is relatively smooth as a function of θ, then an appealing approach to the estimation of $\langle U\Psi_k \rangle$ is deterministic quadrature. For optimal polynomial accuracy, Gaussian quadrature (i.e. nodes at the roots of μ-orthogonal polynomials) may be used. In practice, nested quadrature rules such as Clenshaw–Curtis may be preferable since one does not wish to have to discard past solutions of U upon passing to a more accurate quadrature rule. For multi-dimensional domains of integration Θ, sparse quadrature rules may be used to partially alleviate the curse of dimension.

Note that, if the basis elements Ψ_k are polynomials, then the normalization constant $\gamma_k := \langle \Psi_k^2 \rangle$ can be evaluated numerically but with zero quadrature error by Gaussian quadrature with at least $(k+1)/2$ nodes.

Monte Carlo and Quasi-Monte Carlo Integration. If the dimension of Θ is high, or $U(\theta)$ is a non-smooth function of θ, then it is tempting to resort to Monte Carlo approximation of $\langle U\Psi_k \rangle$. This approach is also appealing because the calculation of the stochastic modes u_k can be written as a straightforward (but often large) matrix-matrix multiplication. The problem with Monte Carlo methods, as ever, is the slow convergence rate of $\sim (\text{number of samples})^{-1/2}$; quasi-Monte Carlo quadrature may be used to improve the convergence rate for smoother integrands.

Connection with Linear Least Squares. Consider for a moment the Monte Carlo approach to estimating both the stochastic modes u_k and the normalization constants γ_k. That is, suppose that N independent and identically μ-distributed samples $\theta_1, \dots, \theta_N$ are given, along with the corresponding values $U(\theta_n)$ of $U \in L^2(\Theta, \mu; \mathbb{R})$. Let

$$P := \begin{bmatrix} \boldsymbol{p}_0 & \cdots & \boldsymbol{p}_K \end{bmatrix} = \begin{bmatrix} \Psi_0(\theta_1) & \cdots & \Psi_K(\theta_1) \\ \vdots & \ddots & \vdots \\ \Psi_0(\theta_N) & \cdots & \Psi_K(\theta_N) \end{bmatrix} \in \mathbb{R}^{N \times (K+1)}$$

and let $\boldsymbol{d} := [U(\theta_1), \dots, U(\theta_N)]$ be the row vector of observed data. The Monte Carlo approximation to $\langle U\Psi_k \rangle$ is

$$\begin{aligned} \langle U\Psi_k \rangle &= \int_\Theta U(\theta)\Psi_k(\theta)\,\mathrm{d}\mu(\theta) \\ &\approx \frac{1}{N}\sum_{n=1}^{N} U(\theta_n)\Psi_k(\theta_n) \\ &= N^{-1}\boldsymbol{d}\boldsymbol{p}_k \\ &= N^{-1}\boldsymbol{p}_k^\top \boldsymbol{d}^\top. \end{aligned}$$

Thus, the column vector $\tilde{\boldsymbol{u}} = [\tilde{u}_0, \dots, \tilde{u}_K]^\top$ of approximate gPC coefficients for U satisfies

$$P^\top P\tilde{\boldsymbol{u}} = P^\top \boldsymbol{d}^\top \tag{13.1}$$

and $P^\top P$ is an approximation to $\mathrm{diag}(\gamma_0, \dots, \gamma_k)$. However, (13.1) are normal equations: $\tilde{\boldsymbol{u}}$ is the minimizer of $\|P\tilde{\boldsymbol{u}} - \boldsymbol{d}^\top\|_2$. In other words, the approximate stochastic model $\widetilde{U} := \sum_{k=0}^K \tilde{u}_k\Psi_k$ has the property that the sum of squared residuals with respect to the data,

$$\sum_{n=1}^{n} \left|\widetilde{U}(\theta_n) - U(\theta_n)\right|^2 \tag{13.2}$$

is minimal among all choices of coefficients $\tilde{u}_0, \dots, \tilde{u}_K$. Therefore, even when given $\{\theta_n\}_{n=1}^N$ that are not necessarily independent and identically μ-distributed, along with corresponding output values $\{U(\theta_n)\}_{n=1}^N$, it is common to construct approximate stochastic modes and hence a pseudo-spectral expansion $\tilde{U}$ by choosing $\tilde{u}_0, \dots \tilde{u}_k$ to minimize the sum of squared residuals (13.2), i.e. according to (13.1).

Conversely, one can engage in the *design of experiments* — i.e. the selection of $\{\theta_n\}_{n=1}^N$ — to optimize some derived quantity of the matrix P; common choices include

- A-optimality, in which the trace of $(P^\top P)^{-1}$ is minimized;
- D-optimality, in which the determinant of $P^\top P$ is maximized;
- E-optimality, in which the least singular value of $P^\top P$ is maximized; and
- G-optimality, in which the largest diagonal term in the orthogonal projection $P(P^\top P)^{-1}P^\top \in \mathbb{R}^{N\times N}$ is minimized.

Sources of Error. In practice, the following sources of error arise when computing pseudo-spectral expansions of this type:

1. *discretization error* comes about through the approximation of $\mathcal{V}$ by a finite-dimensional subspace $\mathcal{V}_M$, i.e. the approximation the stochastic modes u_k by a finite sum $u_k \approx \sum_{m=1}^M u_{km}\phi_m$, where $\{\phi_m \mid m \in \mathbb{N}\}$ is some basis for $\mathcal{V}$;
2. *truncation error* comes about through the truncation of the spectral expansion for U after finitely many terms, i.e. $U \approx \sum_{k=0}^K u_k \Psi_k$;
3. *quadrature error* comes about through the approximate nature of the numerical integration scheme used to find the stochastic modes.

13.2 Stochastic Collocation

Collocation methods for ordinary and partial differential equations are a form of polynomial interpolation. The idea is to find a low-dimensional object — usually a polynomial — that approximates the true solution to the differential equation by means of *exactly* satisfying the differential equation at a selected set of points, called *collocation points* or *collocation nodes*. An important feature of the collocation approach is that an approximation is constructed not on a pre-defined stochastic subspace, but instead uses interpolation, and hence both the approximation and the approximation space are implicitly prescribed by the collocation nodes. As the number of collocation nodes increases, the space in which the solution is sought becomes correspondingly larger.

Example 13.1 (Collocation for an ODE)**.** Consider for example the initial value problem

$$\begin{aligned} \dot{u}(t) &= f(t, u(t)), && \text{for } t \in [a, b] \\ u(a) &= u_a, \end{aligned}$$

to be solved on an interval of time $[a, b]$. Choose n points

$$a \le t_1 < t_2 < \cdots < t_n \le b,$$

called *collocation nodes*. Now find a polynomial $p(t) \in \mathbb{R}_{\le n}[t]$ so that the ODE

$$\dot{p}(t_k) = f(t_k, p(t_k))$$

is satisfied for $k = 1, \dots, n$, as is the initial condition $p(a) = u_a$. For example, if $n = 2$, $t_1 = a$ and $t_2 = b$, then the coefficients $c_2, c_1, c_0 \in \mathbb{R}$ of the polynomial approximation

$$p(t) = \sum_{k=0}^{2} c_k (t - a)^k,$$

which has derivative $\dot{p}(t) = 2c_2(t-a) + c_1$, are required to satisfy

$$\begin{aligned} \dot{p}(a) &= c_1 = f(a, p(a)) \\ \dot{p}(b) &= 2c_2(b-a) + c_1 = f(b, p(b)) \\ p(a) &= c_0 = u_a \end{aligned}$$

i.e.

$$p(t) = \frac{f(b, p(b)) - f(a, u_a)}{2(b-a)}(t-a)^2 + f(a, u_a)(t-a) + u_a.$$

The above equation implicitly defines the final value $p(b)$ of the collocation solution. This method is also known as the *trapezoidal rule* for ODEs, since the same solution is obtained by rewriting the differential equation as

$$u(t) = u(a) + \int_a^t f(s, u(s)) \, \mathrm{d}s$$

and approximating the integral on the right-hand side by the trapezoidal quadrature rule for integrals.

It should be made clear at the outset that there is nothing stochastic about 'stochastic collocation', just as there is nothing chaotic about 'polynomial chaos'. The meaning of the term 'stochastic' in this case is that the collocation principle is being applied across the 'stochastic space' (i.e. the probability space) of a stochastic process, rather than the space/time/space-time domain. That is, for a stochastic process U with known values $U(\theta_n)$ at known collocation points $\theta_1, \ldots, \theta_N \in \Theta$, we seek an approximation $\widetilde{U}$ such that

$$\widetilde{U}(\theta_n) = U(\theta_n) \quad \text{for } n = 1, \ldots, N.$$

There is, however, some flexibility in how to approximate $U\theta)$ for $\theta \neq \theta_1, \ldots, \theta_N$.

Example 13.2. Consider for example the random PDE

$$\begin{aligned} \mathcal{L}_\theta[U(x,\theta)] &= 0 && \text{for } x \in \Omega,\ \theta \in \Theta, \\ \mathcal{B}_\theta[U(x,\theta)] &= 0 && \text{for } x \in \partial\Omega,\ \theta \in \Theta, \end{aligned}$$

where, for μ-a.e. θ in some probability space $(\Theta, \mathscr{F}, \mu)$, the differential operator $\mathcal{L}_\theta$ and boundary operator $\mathcal{B}_\theta$ are well-defined and the PDE admits a unique solution $U(\,\cdot\,, \theta) \colon \Omega \to \mathbb{R}$. The solution $U \colon \Omega \times \Theta \to \mathbb{R}$ is then a stochastic process. We now let $\Theta_M := \{\theta_1, \ldots, \theta_M\} \subseteq \Theta$ be a finite set of prescribed collocation nodes. The collocation problem is to find a *collocation solution* $\widetilde{U}$, an approximation to the exact solution U, that satisfies

$$\begin{aligned} \mathcal{L}_{\theta_m}\big[\widetilde{U}(x, \theta_m)\big] &= 0 && \text{for } x \in \Omega, \\ \mathcal{B}_{\theta_m}\big[\widetilde{U}(x, \theta_m)\big] &= 0 && \text{for } x \in \partial\Omega, \end{aligned}$$

for $m = 1, \ldots, M$.

Interpolation Approach. An obvious first approach is to use interpolating polynomials when they are available. This is easiest when the stochastic space Θ is one-dimensional, in which case the Lagrange basis polynomials of a given nodal set are an attractive choice of interpolation basis. As always, though, care must be taken to use nodal sets that will not lead to Runge oscillations; if there is very little a priori information about the process U, then constructing a 'good' nodal set may be a matter of trial and error.

Given values $U(\theta_1), \ldots, U(\theta_N)$ of U at nodes $\theta_1, \ldots, \theta_N$ in a one-dimensional spaced Θ, the (Lagrange-form polynomial interpolation) collocation approximation $\widetilde{U}$ to U is given by

$$\widetilde{U}(\theta) = \sum_{n=1}^{N} U(\theta_n) \ell_n(\theta) = \sum_{n=1}^{N} U(\theta_n) \prod_{\substack{1 \leq k \leq N \\ k \neq n}} \frac{\theta - \theta_k}{\theta_n - \theta_k}.$$

Example 13.3. Consider the initial value problem

$$\dot{U}(t,\theta) = -e^{\theta}U(t,\theta), \quad U(0,\theta) = 1,$$

with $\theta \sim \mathcal{N}(0,1)$. Take the collocation nodes $\theta_1, \ldots, \theta_N \in \mathbb{R}$ to be the N roots of the Hermite polynomial He_N of degree N. The collocation solution $\widetilde{U}(\cdot, \theta_n)$ at each of the collocation nodes θ_n is the solution of the deterministic problem

$$\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{U}(t,\theta_n) = -e^{\theta_n}U(t,\theta_n), \quad \widetilde{U}(0,\theta_n) = 1,$$

i.e. $\widetilde{U}(t,\theta_n) = \exp(-e^{\theta_n}t)$. Away from the collocation nodes, $\widetilde{U}$ is defined by polynomial interpolation: for each t, $\widetilde{U}(t,\theta)$ is a polynomial in θ of degree at most N with prescribed values at the collocation nodes. Writing this interpolation in terms of Lagrange basis polynomials

$$\ell_n(\theta; \theta_1, \ldots \theta_N) := \prod_{\substack{1 \leq k \leq N \\ k \neq n}} \frac{\theta - \theta_k}{\theta_n - \theta_k}$$

yields

$$\widetilde{U}(t,\theta) = \sum_{n=1}^{N} U(t,\theta_n)\ell_n(\theta).$$

Extension of one-dimensional interpolation methods to the multi-dimensional case can be handled in a theoretically straightforward manner using tensor product grids, similar to the constructions used in quadrature. In tensor product constructions, both the grid of interpolation points and the interpolation polynomials are products of the associated one-dimensional objects. Thus, in a product space $\Theta = \Theta_1 \times \cdots \times \Theta_d$, we take nodes

$$\begin{gathered} \theta_1^1, \ldots, \theta_{N_1}^1 \in \Theta_1 \\ \vdots \\ \theta_1^d, \ldots, \theta_{N_d}^d \in \Theta_d \end{gathered}$$

and construct a product grid of nodes

$$\theta_{\boldsymbol{n}} := (\theta_{n_1}^1, \ldots, \theta_{n_d}^d) \in \Theta \text{ for } \boldsymbol{n} = (n_1, \ldots, n_d) \in \{1, \ldots, N_1\} \times \cdots \times \{1, \ldots, N_d\}.$$

The corresponding interpolation formula, in terms of Lagrange basis polynomials, is then

$$\widetilde{U}(\theta) = \sum_{\boldsymbol{n}=(1,\ldots,1)}^{(N_1,\ldots,N_d)} U(\theta_{\boldsymbol{n}}) \prod_{i=1}^{d} \ell_{n_i}\big(\theta^i; \theta_1^i, \ldots, \theta_{N_i}^i\big).$$

The problem with tensor product grids for interpolative collocation is the same as for tensor product quadrature: the curse of dimension, i.e. the large number of nodes needed to adequately resolve features of functions on high-dimensional spaces. The curse of dimension can be partially circumvented by using interpolation through sparse grids, e.g. those of Smolyak type.

Collocation for arbitrary unstructured sets of nodes — such as those that arise when inheriting an archive of 'legacy' data that cannot be modified or extended for whatever reason — is a notably tricky subject, essentially because it boils down to polynomial interpolation through an unstructured set of nodes. Even the existence of interpolating polynomials such as analogues of the Lagrange basis polynomials is not, in general, guaranteed.

Other Approximation Strategies. There are many other strategies for the construction of collocation solutions, especially in high dimension. Examples include splines, radial basis functions, and kriging.

Bibliography

The monograph of Xiu [182] provides a general introduction to spectral methods for uncertainty quantification, including collocation methods, but is light on proofs. A classic paper on interpolation using sparse grids is that of Barthelmann & al. [10]. The recent paper of Narayan & Xiu [121] presents a method for stochastic collocation on arbitrary sets of nodes using the framework of least orthogonal interpolation, following de Boor & Ron [41]. Non-intrusive methods for UQ, including pseudo-spectral expansions and stochastic collocation, are covered in Chapter 3 of Le Maître & Knio [102]. Buhmann [23] provides a general introduction to the theory and practical usage of radial basis functions.

Exercises

Exercise 13.1. Solve the stochastic oscillator equation of Example 12.13 using an interpolative collocation method.