Advanced Uncertainty Quantification Notes

Zhengbo Zhou

Department of Mathematics, The University of Manchester, M13 9PL

Email: zhengbo.zhou@student.manchester.ac.uk

Contents

1   Representation of Second Order Random Fields

1.1   Essential Revision

Definition 1.1 (Probability space)**.** A probability space contains 3 components

$$(\Omega, \mathcal{F}, \mathbb{P}).$$

- Ω: a sample space (a set)

- $\mathcal{F}$: a σ-algebra (a collection of subsets of Ω)

- $\mathbb{P}$: a (probability) measure

Definition 1.2 (Real-valued random variable)**.** A random variable X is a function which defines on the probability space.

$$X : \Omega \to \mathbb{R}.$$

Example 1.3. Some real-valued random variable examples:

$$X \sim N(0, 1) \text{ (Normal)}, \quad X \sim U(-1, 1) \text{ (Uniform)}.$$

Definition 1.4 ($\mathbb{R}^N$-valued random varible)**.** A $\mathbb{R}^N$-valued random varible X is a function

$$X : \Omega \to \mathbb{R}^N \text{ where } X = (X_1, \dots, X_N)^T.$$

Example 1.5.

$$X \sim N(\boldsymbol{\mu}, C) \text{ (multivariate Normal)}$$

Definition 1.6 (Stochastic Process)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a **probability space** and let $\mathbf{T} \subset \mathbb{R}$. A real-valued *stochastic process* $f(t, \omega)$ can be viewed as

- A real-valued **function** $f : \mathbf{T} \times \Omega \to \mathbb{R}$.

- A collection of real-valued **random variables** for each $t \in \mathbf{T}$

$$\{f(t, \cdot), t \in \mathbf{T}\}.$$

- A collection of **realisations**

$$\{f(\cdot, \omega), \omega \in \Omega\}.$$

($\star$) We will write $f(t, \omega)$ to stress that a stochastic process is a *function*.

1.2   Motivating Example

Example 1.7 (Deterministic ODE)**.** Let $D = (0, 1)$ and consider the following boundary value problem (BVP):
    Find $u : \overline{D} \to \mathbb{R}$ such that

$$-\frac{\mathrm{d}}{\mathrm{d}x}\left({\color{red}a(x)}\frac{\mathrm{d}u(x)}{\mathrm{d}x} \right) = 1, \quad x \in D,$$

together with

$$u(0) = \alpha, \quad u(1) = \beta.$$

This is a simple model of heat diffusion in a wire where

- x denotes **spatial position**.

- $u(x)$ represents the temperature.

- $a(x)$ is the thermal conductivity (or diffusion coefficient).

If $a = 1$, and $\alpha = \beta = 0$, then $u(x) = x(1 - x)/2$ (analytic solution is available).

Random diffusion coefficient

Suppose a is not a simple function of x, but a *stochastic process* (also known as a *random field*).

Let $(\Omega, \mathcal{F}, \mathbb{R})$ be a probability space, and let $\xi_1, \xi_2, \ldots, \xi_{50} : \Omega \to \mathbb{R}$ be **independent**. Now define

$$a(x, \omega) = 1 + \sum_{i=1}^{50} \frac{4}{i^2 \pi^2} \cos(i\pi x)\xi_i(\omega), \quad x \in D, \quad \omega \in \Omega. \tag{1.1}$$

In particular, suppose $\xi_1, \xi_2, \ldots, \xi_{50} \sim U(-1, 1)$. We could then generate 20 realisations of $a(\cdot, \omega_j)$ shown in figure 1 .
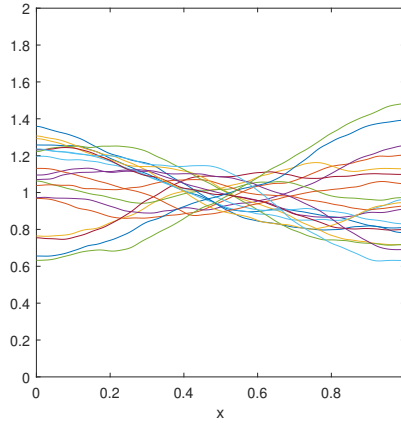
FIGURE 1: **Twenty realisations** $a(\cdot, \omega_j)$ of the diffusion coefficient for $j = 1, 2, \ldots, 20$.

Stochastic ODE Model

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $D = (0, 1)$ (**the spatial domain**).

Find $u : \overline{D} \times \Omega \to \mathbb{R}$ such that $\mathbb{P}$-a.s,

$$-\frac{\mathrm{d}}{\mathrm{d}x}\left(a(x, \omega)\frac{\mathrm{d}u(x, \omega)}{\mathrm{d}x}\right) = 1, \quad x \in D,$$

and

$$u(0, \omega) = \alpha, \quad u(1, \omega) = \beta.$$

Question.

- Does a solution $u(x, \omega)$ exist? Is it unique?

- What kind of random input $a(x, \omega)$ is allowed?

- If a solution $u(x, \omega)$ exist, what are its properties?

- What *function space* does $u(x, \omega)$ belong to?

Forward UQ

Given an input random field $a(x, \omega)$, estimate statistical quantity of interest associated with $u(x, \omega)$.

We can use *Monte Carlo* with FD approximation, to estimate the **mean**

$$\mathbb{E}[u(x, \omega)] \approx \frac{1}{M} \sum_{j=1}^{M} u(x, \omega_j) \approx \frac{1}{M} \sum_{j=1}^{M} u_h^j(x) =: \mu_{M,h}.$$

And similarly the variance

$$\mathbb{V}[u(x, \omega)] \approx \frac{1}{M-1} \sum_{j=1}^{M} (u_h^j(x) - \mu_{M,h})^2.$$

Question. How can we **generate realisations** of the random input $a(x, \omega)$. (Discussed in section 2)

1.3 Random Variables

In UQ1, we met $\mathbb{R}$-valued and $\mathbb{R}^N$-valued random variables,

$$X : \Omega \to \mathbb{R}, \quad X : \Omega \to \mathbb{R}^N,$$

we will make this concept **more general**.

Definition 1.8 (Measure Space). Let Ψ be a set with a σ-algebra $\mathcal{G}$. We call $(\Psi, \mathcal{G})$ a *measurable space* and each $G \in \mathcal{G}$ is a *measurable set*.

Given a *measure* π on $(\Psi, \mathcal{G})$, we call $(\Psi, \mathcal{G}, \pi)$ a *measure space*.

Example 1.9. A probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is a measure space, with

$$\mathbb{P}(\Omega) = 1.$$

Definition 1.10 (Borel sigma-algebra). The *Borel sigma-algebra* $\mathcal{B}(D)$ is defined to be the **smallest** sigma-algebra containing all **open subsets** of D.

Example 1.11. Let $\Psi = D$, where $D \subset \mathbb{R}$. $(D, \mathcal{B}(D))$ is a measurable space.

For each $(a, b) \in \mathcal{B}(D)$, we can assign a 'size', using *Lebesgue measure*,

$$Leb(a, b) := |b - a|.$$

and $(D, \mathcal{B}(D), Leb)$ is a *measure space*.

The measure of a Borel-set can be represented as an **integral**

$$Leb(a, b) = \int_a^b 1 \, \mathrm{dLeb}(x) = \int_a^b 1 \mathrm{d}x.$$

Definition 1.12 (Random Variable). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $(\Psi, \mathcal{G})$ be a measurable space. $X : \Omega \to \Psi$ is a Ψ-valued random variable, if it is $\mathcal{F}$-*measurable*. That is,

$$\forall G \in \mathcal{G} \quad X^{-1}(G) = \{\omega \in \Omega | X(\omega) \in G\} \in \mathcal{F}. \tag{1.2}$$

Definition 1.13 (Probability distribution)**.** The **probability distribution** of X is the *measure* $\mathbb{P}_X$ on $(\Psi, \mathcal{G})$ defined by

$$\mathbb{P}_X(G) := \mathbb{P}(X^{-1}(G)), \quad \forall G \in \mathcal{G},$$

and $(\Psi, \mathcal{G}, \mathbb{P}_X)$ is also a **probability space**.

Continuous Random Variables

We will only be dealing with random variables where Ψ is an *infinite set* like $\mathbb{R}$, or a set of functions on a spatial domain D.

Measurability is an important concept that allows us to define integration.

For continuous random variables $X : \Omega \to \Psi$,

$$\mathbb{P}(\{\omega \in \Omega | X(\omega) \in G\}) = \mathbb{P}_X(G) = \int_G 1 \mathrm{d}\mathbb{P}_X(x). \tag{1.3}$$

And the expectation:

$$\mathbb{E}[X] := \int_\Omega X(\omega)\mathrm{d}\mathbb{P}(\omega) = \int_\Psi x \,\mathrm{d}\mathbb{P}_X(x). \tag{1.4}$$

Real-valued Random Variable

If we know the *probability density function* $\rho(x)$ associated with $\mathbb{P}_X$, we can re-write the integrals w.r.t. $\mathbb{P}_X$ as standard integrals w.r.t. **Lebesgue measure**.

Probability from equation (3.35):

$$\mathbb{P}(\{\omega \in \Omega | a < X(\omega) < b\}) = \mathbb{P}_X(a, b) = \int_a^b 1 \mathrm{d}\mathbb{P}_X(x) = \int_a^b \rho(x)\mathrm{d}x. \tag{1.5}$$

Expectation from equation (3.36)

$$\mathbb{E}[X] := \int_\Omega X(\omega)\mathrm{d}\mathbb{P}(\omega) = \int_\mathbb{R} x \,\mathrm{d}\mathbb{P}_X(x) = \int_\mathbb{R} x\rho(x)\mathrm{d}x. \tag{1.6}$$

1.4   Random Fields

Definition 1.14 (Random Field)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $D \subset \mathbb{R}^d$. A real-valued *random field* $u : D \times \Omega \to \mathbb{R}$ is a real-valued random variable for each $\boldsymbol{x} \in D$.

If $d = 1$ and $D \in \mathbb{R}$, then we call a random field (RF) a *stochastic process*.

Three interpretations:

- A real-valued **function** $u : D \times \Omega \to \mathbb{R}$.

- A collection of real-valued **random variables** for each $\boldsymbol{x} \in D$

$$\{u(\boldsymbol{x}, \cdot), \, \boldsymbol{x} \in D\}.$$

- A collection of **realisations** $\{u(\cdot, \omega), \omega \in \Omega\}$.

Example 1.15 ($d = 1$)**.** Let $D = [0, 2\pi]$ and consider

$$a(x, \omega) = \cos(x)\xi_1(\omega) + \sin(x)\xi_2(\omega), \quad x \in D, \quad \xi_1, \xi_2 \sim N(0, 1) \, iid.$$
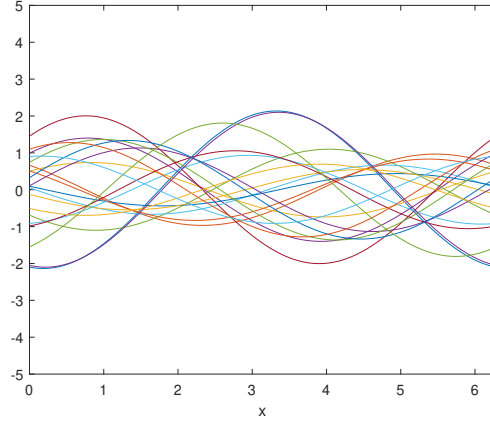
FIGURE 2: Twenty *realisations* $a(x, \omega_j)$, $j = 1, 2, \ldots, 20$ are plotted above.

Question. What kind of functions are the realisations?

Square-integrable functions

Realisations of the RF in this example are **continuous functions** of $x \in D$. They are also **differentiable**. *This is not true in general! (Brownian motion).*
   We will work with RFs with realisations which are at least *square-integrable*.

Definition 1.16 ($L^2(D)$ space). Let $D \subset \mathbb{R}^d$. $L^2(D)$ is the set of Borel-measurable functions $u : D \to \mathbb{R}$ such that

$$\|u\|_{L^2(D)} := \left(\int_D |u(x)|^2 \mathrm{d}x \right)^{1/2} < \infty. \tag{1.7}$$

This is a *Hilbert Space*, which is equipped with an *inner-product*

$$\langle u, v \rangle := \int_D u(x)v(x)\mathrm{d}x, \tag{1.8}$$

where $\|u\|_{L^2(D)} = \sqrt{\langle u, u \rangle}$.

Example 1.17 (Example 1.15 revisit). Let $D = [0, 2\pi]$ and consider

$$a(x, \omega) = \cos(x)\xi_1(\omega) + \sin(x)\xi_2(\omega), \quad x \in D, \quad \xi_1, \xi_2 \sim N(0, 1) \; iid.$$

For each $x \in D$, $a(x, \cdot)$ is a **random variable**. For any $x \in D$,

$$\mathbb{E}[a(x, \cdot)] = \cos(x)\,\mathbb{E}[\xi_1] + \sin(x)\,\mathbb{E}[\xi_2] = 0 + 0 = 0, \tag{1.9}$$

so it $a(x, \omega)$ clearly **mean zero**.

Finite Second Moment

Definition 1.18 ($L^2(\Omega)$ space). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. $L^2(\Omega)$ is the set of $\mathcal{F}$-measurable functions $X : \Omega \to \mathbb{R}$ (i.e., real-valued random variables) such that

$$\|X\|_{L^2(\Omega)} := \mathbb{E}[X^2]^{1/2} < \infty. \tag{1.10}$$

This is also a *Hilbert Space*, which is equipped with the *inner-product*

$$\langle X, Y \rangle := \mathbb{E}[XY]. \tag{1.11}$$

Recall that

$$\mathbb{E}[X^2] := \int_{\Omega} X(\omega)^2 d\mathbb{P}(\omega) = \int_{\mathbb{R}} x^2 d\mathbb{P}_X(x) = \int_{\mathbb{R}} x^2 \rho(x) dx. \tag{1.12}$$

Random variables in $L^2(\Omega)$ have *finite second moment*.

1.5   Second-Order Random Fields

Definition 1.19 (Second-Order RF). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $D \subset \mathbb{R}^d$. A real-valued RF $a(\boldsymbol{x}, \omega)$ is *second order* if $a(\boldsymbol{x}, \cdot) \in L^2(\Omega) \quad \forall \boldsymbol{x} \in D$.

Such RFs have well-defined **mean** and **covariance** functions.
The *mean* function $\mu : D \to \mathbb{R}$ is defined by

$$\mu(\boldsymbol{x}) := \mathbb{E}[a(\boldsymbol{x}, \omega)] = \int_{\Omega} a(\boldsymbol{x}, \omega) d\mathbb{P}(\omega). \tag{1.13}$$

The *covariance* function $C : D \times D \to \mathbb{R}$ is defined by

$$C(\boldsymbol{x}, \boldsymbol{y}) := \text{Cov}(a(\boldsymbol{x}, \omega), a(\boldsymbol{y}, \omega)), \quad \boldsymbol{x}, \boldsymbol{y} \in D. \tag{1.14}$$

Using the definition of covariance

$$C(\boldsymbol{x}, \boldsymbol{y}) = \mathbb{E}[(a(\boldsymbol{x}, \omega) - \mu(\boldsymbol{x}))(a(\boldsymbol{y}, \omega) - \mu(\boldsymbol{y}))]. \tag{1.15}$$

Given the covariance function $C(\boldsymbol{x}, \boldsymbol{y})$, the *variance* of a RF is

$$\mathbb{V}[a(\boldsymbol{x}, \omega)] := C(\boldsymbol{x}, \boldsymbol{x}) = \mathbb{E}[(a(\boldsymbol{x}, \omega) - \mu(\boldsymbol{x}))^2]. \tag{1.16}$$

Valid Covariance functions

Important: Not all functions $C(\boldsymbol{x}, \boldsymbol{y})$ are valid covariance functions.

Theorem 1.20 (Covariance Functions). *Let $D \in \mathbb{R}^d$ and let $a(\boldsymbol{x}, \omega)$ be a second-order RF. The covariance function $C : D \times D \to \mathbb{R}$ is* symmetric, *i.e.,*

$$C(\boldsymbol{x}, \boldsymbol{y}) = C(\boldsymbol{y}, \boldsymbol{x}) \quad \boldsymbol{x}, \boldsymbol{y} \in D$$

and non-negative definite. *That is, for any $N \in \mathbb{N}$ and any constants $\alpha_1, \dots, \alpha_N \in \mathbb{R}$, and any points $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N \in D$,*

$$\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i C(\boldsymbol{x}_i, \boldsymbol{x}_j) \alpha_j \geq 0.$$

Example 1.21 (Covariance Functions). Often, $C(\boldsymbol{x}, \boldsymbol{y})$ is a function of the *distance* between a pair of points $\boldsymbol{x}, \boldsymbol{y} \in D$.
Some examples for one-dimensional domains $D \subset \mathbb{R}$:

- The *exponential* covariance function is

$$C(x, y) = \sigma^2 \exp\left(\frac{-|x - y|}{\ell}\right)$$

Here, $C(x, x) = \sigma^2$ is the (constant) variance and ℓ is the **correlation length**.

- The *Gaussian* covariance function ($d = 1$) is

$$C(x, y) = \sigma^2 \exp\left(\frac{-|x - y|^2}{\ell}\right).$$

Smoothness

The smoothness of the **realisations** of a RF depends on the smoothness of the **covariance function**.
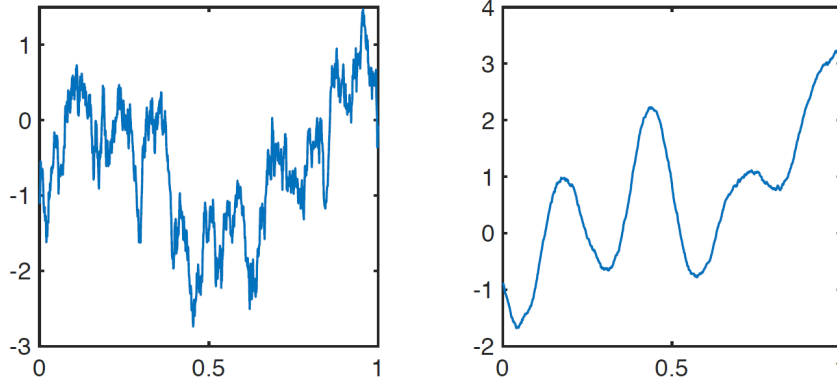
FIGURE 3: A **realisation** of mean zero RF with (Left) the *exponential* covariance function and (Right) the *Gaussian* covariance function with $\sigma = 1$, $\ell = 0.1$.

1.6   Gaussian Random Fields

To specify a RF, it is not enough to specify the domain D, the mean function $\mu(\boldsymbol{x})$ and covariance function $C(\boldsymbol{x}, \boldsymbol{y})$. We also have to specify the *probability distribution*.

Definition 1.22 (Gaussian RF). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $D \subset \mathbb{R}^d$. A real-vallued RF $a(\boldsymbol{x}, \omega)$ with mean $\mu(\boldsymbol{x})$ and covariance $C(\boldsymbol{x}, \boldsymbol{y})$ is *Gaussian* if

$$\forall \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \in D$$

(for any $N \in \mathbb{N}$),

$$\boldsymbol{a}(\omega) = [a(\boldsymbol{x}_1, \omega), \ldots, a(\boldsymbol{x}_N, \omega)]^T \sim N(\boldsymbol{\mu}, C_N)$$

(i.e., **multivariate Normal**) with *mean vector*

$$\boldsymbol{\mu} \in \mathbb{R}^N, \quad [\boldsymbol{\mu}]_i = \mu(\boldsymbol{x}_i), \quad i = 1, \ldots, N,$$

and *covariance matrix*

$$C_N \in \mathbb{R}^{N \times N}, \quad [C_N]_{i,j} = C(\boldsymbol{x}_i, \boldsymbol{x}_j), \quad i, j = 1, \ldots, N.$$

Example 1.23 (Brownian motion). Some Gaussian RFs have special name.

Let $\mathbf{T} \subset \mathbb{R}^+$. *Brownian Motion* $W : \mathbf{T} \times \Omega \to \mathbb{R}$ is a Gaussian stochastic process $W(t, \omega)$ with continuous realisations with

$$\mu(t) = 0, \quad C(s, t) = \min\{s, t\}.$$

Example 1.24 (Brownian Bridge). The *Brownian Bridge* is a version of Brownian motion on a bounded domain that also satisfies **boundary conditions**.

Let $\mathbf{T} = [0, 1]$. The **Brownian Bridge** $B : \mathbf{T} \times \Omega \to \mathbb{R}$ is a Gaussian stochastic process $B(t, \omega)$ satisfying

$$B(0, \omega) = 0, \quad B(1, \omega) = 0$$

almost surely, with

$$\mu(t) = 0, \quad C(s, t) = \min\{s, t\} - st.$$

1.7   Intergal Operators and Covariance Functions

Square-integrable functions on $D \times D$

We previously defined the space $L^2(D)$ of *square-integrable* functions $u : D \to \mathbb{R}$.

Definition 1.25 ($L^2(D \times D)$ space). Let $D \in \mathbb{R}^d$. A function $G : D \times D \to \mathbb{R}$ belongs to $L^2(D \times D)$ if

$$\|G\|_{L^2(D \times D)} := \left(\int_D \int_D G(\boldsymbol{x}, \boldsymbol{y})^2 \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y} \right)^{1/2} < \infty. \tag{1.17}$$

Definition 1.26 (Integral Operator with Kernel G). Let $D \subset \mathbb{R}^d$. Given a 'kernel' $G \in L^2(D \times D)$, the *integral operator* $\mathcal{L} : L^2(D) \to L^2(D)$ associated with G is defined by

$$\mathcal{L}f(\boldsymbol{x}) := \int_D G(\boldsymbol{x}, \boldsymbol{y})f(\boldsymbol{y})\mathrm{d}\boldsymbol{y}, \quad f \in L^2(D). \tag{1.18}$$

Such operators are: *bounded, linear* and *compact* on $L^2(D)$.

Let $H = L^2(D)$.

Definition 1.27 (Linear operator). We say $\mathcal{L} : H \to H$ is *linear* on H if

- $\mathcal{L}(f + g) = \mathcal{L}(f) + \mathcal{L}(g), \quad \forall f, g \in H$

- $\mathcal{L}(\alpha f) = \alpha \mathcal{L}(f), \quad \forall \alpha \in \mathbb{R}$ and $\forall f \in H$.

Definition 1.28 (Bounded operator). We say $\mathcal{L} : H \to H$ is *bounded* on H if there exist a constant $K > 0$ such that

$$\|\mathcal{L}f\|_H \leq K\|f\|_H, \quad \forall f \in H.$$

[colback=green!40!white]

Theorem 1.29 (Symmetric operator). *If the kernel G is* *symmetric*

$$G(\boldsymbol{x}, \boldsymbol{y}) = G(\boldsymbol{y}, \boldsymbol{x}), \quad \boldsymbol{x}, \boldsymbol{y} \in D,$$

then the operator $\mathcal{L}$ in definition *1.26* *is also* *symmetric**, i.e.,*

$$\langle \mathcal{L}f, g \rangle = \langle f, \mathcal{L}g \rangle, \quad \forall f, g \in L^2(D),$$

where $\langle \cdot, \cdot \rangle$ is the inner-product on $L^2(D)$.

Recall. Covariance functions $C(\boldsymbol{x}, \boldsymbol{y})$ are always symmetric. Given a covariance function $C(\boldsymbol{x}, \boldsymbol{y})$ that belongs to $L^2(D \times D)$, we can use definition 1.26 to construct an *integral operator*

$$\mathcal{C} : L^2(D) \to L^2(D)$$

through the relation

$$\mathcal{C}f(\boldsymbol{x}) := \int_D C(\boldsymbol{x}, \boldsymbol{y})f(\boldsymbol{y})\mathrm{d}\boldsymbol{y}, \quad f \in L^2(D),$$

this is *linear, bounded, compact* and *symmetric*.

Eigenvalues & *Eigenfunctions*

There are infinitely many *eigenvalues* λ and *eigenfunctions* $\phi(\neq 0)$ satisfying

$$\mathcal{C}\phi(\boldsymbol{x}) = \lambda\phi(\boldsymbol{x}).$$

Using the definition of $\mathcal{C}$, we have

$$\int_D C(\boldsymbol{x}, \boldsymbol{y})\phi(\boldsymbol{y})\mathrm{d}\boldsymbol{y} = \lambda\phi(\boldsymbol{x}).$$

Example 1.30 (Brownian Bridge)**.** Let $\mathbf{T} = [0, 1]$. Recall that the *Brownian Bridge* $B : \mathbf{T} \times \Omega \to \mathbb{R}$ is Gaussian stochastic process with covariance function

$$C(s, t) = \min\{s, t\} - st.$$

The associated integral operator is defined via

$$\mathcal{C}f(t) := \int_0^1 C(s, t)f(s)\mathrm{d}s = \int_0^1 (\min\{s, t\} - st)f(s)\mathrm{d}s, \quad f \in L^2(\mathbf{T}).$$

The associated eigenvalues and eigenfunctions satisfies $\mathcal{C}\phi(t) = \lambda\phi(t)$, or

$$\int_0^1 (\min\{s, t\} - st)\phi(s) = \lambda\phi(t). \tag{1.19}$$

Notice that

$$\mathcal{C}f(t) = \int_0^1 (\min\{s, t\} - st)f(s)\mathrm{d}s = \int_0^t (s - st)f(s)\mathrm{d}s + \int_t^1 (t - st)f(s)\mathrm{d}s$$
$$= (1 - t)\int_0^t sf(s)\mathrm{d}s + t\int_t^1 (1 - s)f(s)\mathrm{d}s. \tag{1.20}$$

Thus the eigenfunction ϕ and eigenvalues λ can be determined by solving

$$(1 - t)\int_0^t s\phi(s)\mathrm{d}s + t\int_t^1 (1 - s)\phi(s)\mathrm{d}s = \lambda\phi(t)$$

Notice that $\phi(0) = 0, \phi(1) = 0$. Differentiate w.r.t. t, we have

$$t\phi(t) - t^2\phi(t) - \int_0^t s\phi(s)\mathrm{d}s + \int_t^1 (1 - s)\phi(s)\mathrm{d}s - t(1 - t)\phi(t) = \lambda\phi'(t)$$

$$-\int_0^t s\phi(s)\mathrm{d}s + \int_t^1 (1 - s)\phi(s)\mathrm{d}s = \lambda\phi'(t)$$

$$-t\phi(t) + (-(1 - t)\phi(t)) = \lambda\phi''(t) \tag{1.21}$$

$$\lambda\phi''(t) + \phi(t) = 0$$

$$\phi(t) + \frac{1}{\lambda}\phi(t) = 0 \Leftrightarrow \underbrace{\phi''(t) + \nu\phi(t) = 0}_{\text{Second Order ODE with } \nu = 1/\lambda}$$

We then have the general solution:

$$\phi(t) = A\cos(\omega t) + B\sin(\omega t), \quad \phi(0) = \phi(1) = 0.$$

This system will gives

$$\phi_i(t) = B_i \sin(i\pi t), \quad \lambda_i = \frac{1}{\omega^2} = \frac{1}{i^2\pi^2}, \quad i \in \mathbb{Z}^+.$$

We could normalize this function by acquiring $\|\phi_i\|_{L^2(0,1)} = 1$, and we have $B_i = \sqrt{2}$.

Theorem 1.31 (Hilbert-Schmidt Spectral Theorem). *Let $H = L^2(D)$, and $\mathcal{L} : H \to H$ be a bounded linear operator that is symmetric and compact. Denote the eigenvalue of $\mathcal{L}$ by λ_i, ordered such that $|\lambda_i| \le |\lambda_{i+1}|$, and denoted the corresponding eigenfunctions be $\phi_i \in L^2(D)$. Then*

- *The eigenvalues λ_i are real and $\lambda_i \to 0$ as $i \to \infty$.*

- *The eigenfunctions ϕ_j can be chosen to form an* orthonormal *basis for the range of $\mathcal{L}(L^2(D))$.*

- *For any $u \in L^2(D)$,*

$$\mathcal{L}u = \sum_{i=1}^{\infty} \lambda_i \langle u, \phi_i \rangle_{L^2(D)} \phi_i.$$

1.8   Karhuren-Loève expansion

Definition 1.32 ($L^2(\Omega, L^2(D))$ space). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $D \in \mathbb{R}^d$. $L^2(\Omega, L^2(D))$ is the space of $L^2(D)$-valued random variables $a : \Omega \to L^2(D)$ satisfying

$$\|a\|_{L^2(\Omega, L^2(D))} := \mathbb{E}[\|a(\boldsymbol{x}, \omega)\|^2_{L^2(D)}]^{1/2} \le \infty.$$

We can interpret such functions as RF: If $a(\boldsymbol{x}, \omega) \in L^2(\Omega, L^2(D))$.

- For each ω, we have a realisation $a(\cdot, \omega) \in L^2(D)$.

- For each $\boldsymbol{x}$, we have a random variable $a(\boldsymbol{x}, \cdot) \in L^2(\Omega)$.

Series Expansion

Suppose $a(\boldsymbol{x}, \omega) \in L^2(\Omega, L^2(D))$. If we know a *basis* $\{\phi_i(\boldsymbol{x}), i = 1, 2, \ldots,\}$ for $L^2(D)$, then any *realisation* of $a(\boldsymbol{x}, \omega)$ can be represented as a linear combination of those functions $a(\boldsymbol{x}, \omega) = \sum_{i=1}^{\infty} a_i(\omega)\phi_i(\boldsymbol{x})$, where the coefficients are random variables. The KL expansion uses the eigenfunctions ϕ_i of the integral operator associated with the covariance function as the chosen basis.

Theorem 1.33 (Karhuren-Loève expansion). *Suppose we have a second order random field such that*

$$a(\boldsymbol{x}, \omega) \in L^2(\Omega, L^2(D)).$$

Then

$$a(\boldsymbol{x}, \omega) = \underbrace{\mu(\boldsymbol{x})}_{\text{Not random}} + \sum_{i=1}^{\infty} \sqrt{\lambda_i} \; \underbrace{\xi_i(\omega)}_{\text{Random variables}} \phi_i(\boldsymbol{x}). \tag{1.22}$$

- $\mu(\boldsymbol{x})$ *is the mean function.*

- $\{\lambda_i, \phi_i(\boldsymbol{x})\}$ *are the eigenvalues and normalized eigenfunctions of the integal operator* $\mathcal{C}$ *associated with the covariance function where* $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0.$

- *The random variables are*

$$\xi(\omega) := \frac{1}{\sqrt{\lambda_i}} \langle a(\boldsymbol{x}, \omega) - \mu(\boldsymbol{x}), \phi_i(\boldsymbol{x}) \rangle,$$

  and they have mean zero, variance one, and are uncorrelated.

- *The series converges in the* $L^2(\Omega, L^2(D))$ *sense.*

Remark 1.34. The distribution of the random variables ξ_i appearing in KL expansion depends on the distribution of the RF $a(\boldsymbol{x}, \omega)$.

- If $a(\boldsymbol{x}, \omega)$ is a *Gaussian RF*, then the ξ_i are also Gaussian.

- Uncorrelated Gaussian random variables are independent.

Convergence

Note that the function in $L^2(D)$ does not need to be continuous. When we write

$$a(\boldsymbol{x}, \omega) = \mu(\boldsymbol{x}) + \sum_{i=1}^{\infty} \sqrt{\lambda_i} \xi_i(\omega) \phi_i(\boldsymbol{x}),$$

this does not mean that the series converges to $a(\boldsymbol{x}, \omega)$ for all $\boldsymbol{x} \in D$. But we could have converges in 'the $L^2(\Omega, L^2(D))$ sense', which means that if we define

$$a_J(\boldsymbol{x}, \omega) := \mu(\boldsymbol{x}) + \sum_{i=1}^{J} \sqrt{\lambda_i} \xi_i(\omega) \phi_i(\boldsymbol{x}),$$

then

$$\|a(\boldsymbol{x}, \omega) - a_J(\boldsymbol{x}, \omega)\|_{L^2(\Omega, L^2(D))} \to 0 \quad \text{as} \quad J \to \infty.$$

1.9   Truncated KL expansion

What we defined previously

$$a_J(\boldsymbol{x}, \omega) := \mu(\boldsymbol{x}) + \sum_{i=1}^{J} \sqrt{\lambda_i} \xi_i(\omega) \phi_i(\boldsymbol{x}),$$

is the *truncated expansion* which is often used as an approximation to do numerical approximations.

Example 1.35 (Browinan Bridge)**.** Let $\mathbf{T} = [0, 1]$, recall that the Browian Bridge $B : \mathbf{T} \times \Omega \to \mathbb{R}$ is a Gaussian RF (stochastic process) with $\mu(t) = 0$, $C(s, t) = \min\{s, t\} - st$. The eigenvalues and *normalised* eigenfunctions of the integral operator are

$$\lambda_i = \frac{1}{i^2 \pi^2}, \quad \phi_i(t) = \sqrt{2} \sin(i\pi t), \quad i = 1, 2, \dots.$$

Using the definition of KL expansion:

$$B(t, \omega) = \sum_{i=1}^{\infty} \frac{\sqrt{2}}{i\pi} \xi_i(\omega) \sin(i\pi t), \quad \xi_i \sim N(0, 1) iid.$$

we could define the truncated KL expansion for $B(t, \omega)$,

$$B_J(t, \omega) = \sum_{i=1}^{J} \frac{\sqrt{2}}{i\pi} \xi_i(\omega) \sin(i\pi t), \quad \xi_i \sim N(0, 1) iid.$$

From figure 4, we could see the series is converging.
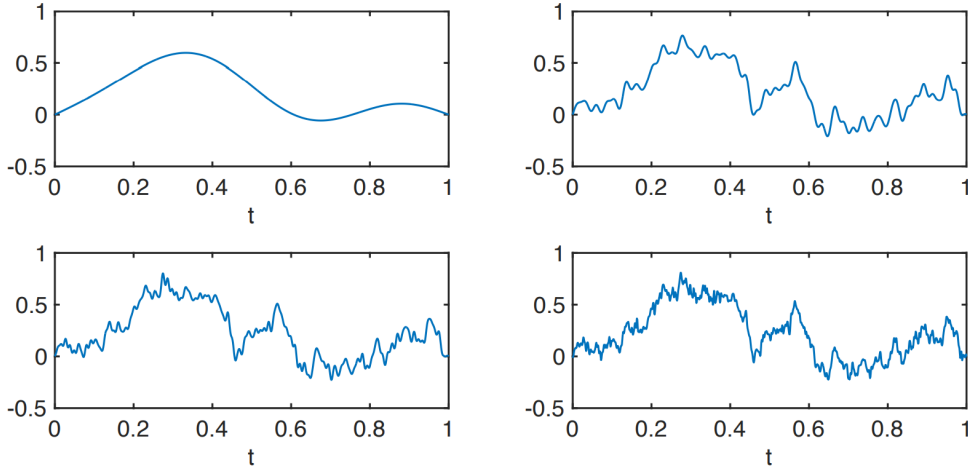
FIGURE 4: One realisation $B_J(t, \omega^*)$ of the truncated KL expansion of the standard Brownian bridge $B(t, \omega)$ defined in Example 1.30 for $J = 5, 50, 100, 500$ (left to right, top to bottom).

Error in Truncated KL expansion

Theorem 1.36. *If $a(\boldsymbol{x}, \omega) \in L^2(\Omega, L^2(D))$, and $a_J(\boldsymbol{x}, \omega)$ is the truncated approximation to $a(\boldsymbol{x}, \omega)$. Then*

$$\mathbb{E}[\|a(\boldsymbol{x}, \omega) - a_J(\boldsymbol{x}, \omega)\|^2_{L^2(D)}] = \sum_{i=J+1}^{\infty} \lambda_i$$

Theorem 1.37 (Mercer's Theorem). *Let $a(\boldsymbol{x}, \omega) \in L^2(\Omega, L^2(D))$ be a second-order RF with KL expansion:*

$$a(\boldsymbol{x}, \omega) = \mu(\boldsymbol{x}) + \sum_{i=1}^{\infty} \sqrt{\lambda_i} \xi_i(\omega) \phi_i(\boldsymbol{x}).$$

If $D \subset \mathbb{R}^d$ is bounded and closed and if $C(\boldsymbol{x}, \boldsymbol{y})$ is continuous, then

- *The eigenfunctions $\phi_i(\boldsymbol{x})$ are continuous.*

- *The covariance function can be expressed as*

$$C(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{\infty} \lambda_i \phi_i(\boldsymbol{x}) \phi_j(\boldsymbol{y})$$

where the series converges uniformly. *That is :*

$$\|C - \widehat{C}\|_\infty = \sup_{x,y \in D} |C(x,y) - \widehat{C}(x,y)| \to 0, \quad as \ \ J \to \infty,$$

where $\widehat{C}(x,y)$ is the truncated convariance function approximation.

Error in Variance

It can be shown that $\widehat{C}(x,y)$ is covariance function of the truncated RF

$$a_J(x,\omega) = \mu(x) + \sum_{i=1}^{J} \sqrt{\lambda_i}\xi_i(\omega)\phi_i(x).$$

Using this and the Mercer's theorem, gives for any $x \in D$,

$$\begin{aligned}
\mathbb{V}[a(x,\omega)] &= C(x,x) = \sum_{i=1}^{\infty} \lambda_i\phi_i(x)^2 \\
\mathbb{V}[a_J(x,\omega)] &= \widehat{C}(x,x) = \sum_{i=1}^{J} \lambda_i\phi_i(x)^2
\end{aligned} \tag{1.23}$$

Hence

$$\mathbb{V}[a(x,\omega)] - \mathbb{V}[a_J(x,\omega)] = \sum_{i=J+1}^{\infty} \lambda_i\phi_i(x)^2 \geq 0 \to \mathbb{V}[a(x,\omega)] \geq \mathbb{V}[a_J(x,\omega)].$$

We always *underestimate* the variance.

A Practical Error Estimator

We could define

$$E_J := \frac{\int_D \mathbb{V}[a(x,\omega) - a_J(x,\omega)]\mathrm{d}x}{\int_D \mathbb{V}[a(x,\omega)]\mathrm{d}x} \tag{1.24}$$

If the RF has constant variance $\mathbb{V}[a(x,\omega)] = \sigma^2$, then we have

$$E_J := \frac{\sigma^2|D| - \sum_{i=1}^{J} \lambda_i}{\sigma^2|D|,} \tag{1.25}$$

which is computable. If we know the eigenvalues, we can find J such that $E_J \leq \texttt{tol}$.

2   Numerical Methods for Generating RFs

Recall. The KL expansion of a random field $a : D \times \Omega \to \mathbb{R}$ is

$$a(\boldsymbol{x}, \omega) = \mu(\boldsymbol{x}) + \sum_{i=1}^{\infty} \sqrt{\lambda_i} \xi_i(\omega) \phi_i(\boldsymbol{x})$$

this is a function of $\boldsymbol{x} \in D$.

Key Point. When solving ODEs/PDEs with RFs inputs numerically, we often only need access to samples at a finite set of grid points $x_1, x_2, \ldots, x_N$. (see example 2.1)

Example 2.1. Let $D = (0, 1)$, find $u : \overline{D} \to \mathbb{R}$ such that

$$-\frac{\mathrm{d}}{\mathrm{d}x}\left(a(x)\frac{\mathrm{d}u(x)}{\mathrm{d}x}\right) = 1 \quad x \in D.$$

with $u(0) = u(1) = 0$. We could approximate the solution using a *finite difference method*.

- Divide the interval $[0, 1]$ into N subintervals (or elements) of width $h = 1/N$ and label the vertices of these elements as $x_0, x_1, ..., x_N$. Then we have

$$-\frac{d}{dx}\left(a\,(x_i)\,\frac{du\,(x_i)}{dx}\right) = 1, \quad i = 1, \ldots, N - 1 \tag{2.1}$$

  with $u(x_0) = u(x_N) = 0$.

- Using a centered finite difference scheme:

$$\frac{\mathrm{d}u}{\mathrm{d}x}(x_i) \approx \frac{u(x_i + \frac{h}{2}) - u(x_i - \frac{h}{2})}{h} = \frac{u(x_{i+1/2}) - u(x_{i-1/2})}{h} \tag{2.2}$$

  where $x_{i+1/2}$ is the midpoint of $[x_i, x_{i+1}]$ and similarly for $x_{i-1/2}$. Approximating the derivatives in (2.1) in this way and replacing $u\,(x_i)$ by U_i, where U_i denotes the resulting approximation to $u\,(x_i)$ gives a linear system of $N - 1$ equations

$$-\left(\frac{a\,(x_{i-1/2})}{h^2}\right)U_{i-1} + \left(\frac{a\,(x_{i-1/2}) + a\,(x_{i+1/2})}{h^2}\right)U_i - \left(\frac{a\,(x_{i+1/2})}{h^2}\right)U_{i+1} = 1 \tag{2.3}$$

  for $i = 1, \ldots, N - 1$, together with $U_0 = 0$ and $U_N = 0$. If we consider a stochastic model where $a(x, \omega)$ is a stochastic process, then to perform a Monte Carlo simulation, we will need to generate samples of the random vector

$$\mathbf{a}(\omega) = \left[a\,(x_{1/2}, \omega), a\,(x_{3/2}, \omega), \ldots, a\,(x_{N-(1/2)}, \omega)\right]^T \tag{2.4}$$

  corresponding to $a(x, \omega)$ evaluated at the mid-points of the elements in the grid.

2.1   Generating Realisations of Gaussian RFs

Let $D \subset \mathbb{R}^d$ be a domain ($d = 1, 2$), and let $z(\boldsymbol{x}, \omega)$ be an associated second-order RF with mean function $\mu(\boldsymbol{x})$ and covariance function $C(\boldsymbol{x}, \boldsymbol{y})$.

Random fields to random vectors

Let $N \in \mathbb{N}$, for any $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N \in D$, we can define

$$\boldsymbol{Z}(\omega) := [z(\boldsymbol{x}_1, \omega), z(\boldsymbol{x}_2, \omega), \ldots, z(\boldsymbol{x}_N, \omega)]^T \tag{2.5}$$

This is an $\mathbb{R}^N$-valued random variable $\boldsymbol{Z} : \Omega \to \mathbb{R}^N$.

- It has *mean vector*

$$\boldsymbol{\mu} := \mathbb{E}[\boldsymbol{Z}], \text{ where } [\mu]_i = \mu(\boldsymbol{x}_i), \quad i = 1, 2, \ldots, N.$$

- It has *Covariance matrix*

$$C_N := \mathbb{E}[(\boldsymbol{Z} - \boldsymbol{\mu})(\boldsymbol{Z} - \boldsymbol{\mu})^T] \in \mathbb{R}^{N \times N},$$

   where entries are $[C_N]_{ij} = C(\boldsymbol{x}_i, \boldsymbol{x}_j), i, j = 1, \ldots, N$.

By theorem 1.20, the $N \times N$ covariance matrix C_N of $\boldsymbol{Z}$ is also symmetric and non-negative definite (or positive semidefinite) matrix, that is $\boldsymbol{v}^T C_N \boldsymbol{v} \geq 0$, for all $\boldsymbol{v} \in \mathbb{R}^N$. From definition 1.22, we knwo that if $z(\boldsymbol{x}, \omega)$ is a Gaussian RF, then for any $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$, $\boldsymbol{Z} \sim N(\boldsymbol{\mu}, C_N)$. That is, $\boldsymbol{Z}$ is always a *multivariate normal* random variable. Hence, if we need to generate realisations of a Gaussian RF at a set of grid points, then we only need to generate samples of random vectors $\boldsymbol{Z}$ with distribution $\boldsymbol{Z} \sim N(\boldsymbol{\mu}, C_N)$.

Standard approach is to use a *matrix fractorisation* of C_N. Suppose we can find an $N \times N$ matrix V such that

$$C_N = VV^T.$$

We can create a sample of $\boldsymbol{Z} \sim N(\boldsymbol{\mu}, C_N)$ as follows:

Theorem 2.2 (Matrix Factorization Method)**.**

- *Generate a sample $\boldsymbol{\xi} \sim N(\boldsymbol{0}, I_N)$, where $\boldsymbol{\xi} = [\xi_1, \xi_2, \ldots, \xi_N]^T$, $\xi_i \sim N(0, 1)$ iid.*

- *Factorize $C_N = VV^T$.*

- *Form $\boldsymbol{Z} = \boldsymbol{\mu} + V\boldsymbol{\xi}$, and then $\boldsymbol{Z} \sim N(\boldsymbol{\mu}, C_N)$.*

2.2   Cholesky Factorization

Theorem 2.3 (Cholesky factorisation)**.** *If C_N is an $N \times N$ symmetric and positive definite matrix, i.e.,*

$$\boldsymbol{v}^T C_N \boldsymbol{v} > 0 \quad \forall \boldsymbol{v} \in \mathbb{R}^N \setminus \{\boldsymbol{0}\},$$

then there exist an $N \times N$ lower triangular L such that

$$C_N = LL^T. \tag{2.6}$$

Potential disadvantages for Cholesky method

- If C_N is not strictly positive definite, or has small eigenvalues very close to zero, then the method can fail.

- Since covariance matrices are usually dense, it can be very expensive:

   - Cost of Cholesky factorisation: $O(N^3)$,
   - Cost of Multiplication with L: $O(N^2)$.

If we need M samples, then the cost is

$$O(N^3) + M \cdot O(N^2) \quad \text{where } N \text{ is the number of grid points} \tag{2.7}$$

2.3   Eigenvalue Decomposition

If C_N is real valued, symmetric and non-negative definite. We can factorise it as

$$C_N = U\Lambda U^T$$

where

- Λ is the $N \times N$ diagonal matrix of eigenvalues, ordered such that

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N \geq 0.$$

- U is a matrix whose columns are eigenvectors

$$U = [\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_N] \in \mathbb{R}^{N \times N},$$

and we assume these are normalised such that

$$u_i^T u_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \Rightarrow U^T U = I_N. \tag{2.8}$$

We could then manipulate the factorisation as

$$C_N = U\Lambda U^T = U\Lambda^{1/2}\Lambda^{1/2}U^T = U\Lambda^{1/2}(U\Lambda^{1/2})^T \tag{2.9}$$

Here $\Lambda^{1/2}$ is the $N \times N$ diagonal matrix with diagonal entries,

$$\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_N}.$$

This is well-defined since C_N is non-negative definite. We can now use matrix factorisation method with $V = U\Lambda^{1/2}$, and this gives $\boldsymbol{Z} = \boldsymbol{\mu} + U\Lambda^{1/2}\boldsymbol{\xi}$ where $\boldsymbol{\xi} \sim N(\boldsymbol{0}, I_N)$.

2.4   Truncated Eigenvalue Decomposition

Note that

$$\boldsymbol{Z} = \boldsymbol{\mu} + U\Lambda^{1/2}\boldsymbol{\xi} = \boldsymbol{\mu} + \sum_{i=1}^{N} \sqrt{\lambda_i}\boldsymbol{u}_i\xi_i.$$

Recall. N is the number of grid points, and if we compare the cost with Cholesky method

- The cost of computing U and Λ is similar to Cholesky, $O(N^3)$.

- The cost of a multiplication with $V = U\Lambda^{1/2}$ is also $O(N^2)$.

However, We could make savings if we truncated the sum after $n < N$ terms.
Important. We assume that $\lambda_1 \leq \cdots \leq \lambda_N$. Let $n \in \mathbb{Z}$ with $n < N$ and define

$$\widehat{\boldsymbol{Z}} = \boldsymbol{\mu} + \sum_{i=1}^{n} \sqrt{\lambda_i}\boldsymbol{u}_i\xi_i, \quad \xi \sim N(0,1), \; iid. \tag{2.10}$$

We can also write

$$\widehat{\boldsymbol{Z}} = \boldsymbol{\mu} + U_n\Lambda_n^{1/2}\xi_n, \tag{2.11}$$

where $U_n = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n]$, $\Lambda_n = \text{diag}(\lambda_1, \ldots, \lambda_n)$, and $\boldsymbol{\xi}_n = [\xi_1, \ldots, \xi_n]^T$.

Key Question. How good $\widehat{\boldsymbol{Z}}$ as an *approximation* to Z?

The random vector $\widehat{\boldsymbol{Z}}$ is Gaussian but the covariance matrix is not C_N. And it can be shown that

$$\widehat{\boldsymbol{Z}} \sim N(\boldsymbol{\mu}, \widehat{C}_n), \quad \text{where } \widehat{C}_n = U_n \Lambda_n U_n^T \tag{2.12}$$

and we have the equality:

$$\frac{\|C_N - \widehat{C}_n\|_2}{\|C_N\|_2} = \frac{\lambda_{n+1}}{\lambda_1} \tag{2.13}$$

And also by considering the mean-square sense:

$$\mathbb{E}[\|\boldsymbol{Z} - \widehat{\boldsymbol{Z}}\|_2^2] = \sum_{i=n+1}^{N} \lambda_i. \tag{2.14}$$

2.5 Toeplitz and Circular Matrix

Definition 2.4. A *second order* RF is called *stationary* if the mean $\mu(\boldsymbol{x})$ is a constant and the covariance function

$$C(\boldsymbol{x}, \boldsymbol{y}) = c(\boldsymbol{x} - \boldsymbol{y}), \quad \boldsymbol{x}, \boldsymbol{y} \in D \tag{2.15}$$

for some function $c : D \to \mathbb{R}$ called the stationary covariance.

Example 2.5. A mean zero RF with exponential covariance function

$$C(\boldsymbol{x}, \boldsymbol{y}) = \sigma^2 \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{y}\|_2}{\ell}\right), \quad \boldsymbol{x}, \boldsymbol{y} \in D \subset \mathbb{R}^2 \tag{2.16}$$

is stationary since $C(\boldsymbol{x}, \boldsymbol{y}) = c(\boldsymbol{x} - \boldsymbol{y})$ where

$$c(\boldsymbol{x}) = \sigma^2 \exp\left(-\frac{\|\boldsymbol{x}\|_2}{\ell}\right) \tag{2.17}$$

Definition 2.6 (Toeplitz matrix)**.** An $N \times N$ real-valued matrix C is toeplitz if the (i, j)th entry

$$[C]_{i,j} = c_{i-j}$$

for some real numbers $c_{1-N}, \ldots, c_{N-1}$ In order words, the toeplitz matrix is constant along the diagonal and off diagonals:

$$C = \begin{pmatrix} c_0 & c_{-1} & \cdots & c_{2-N} & c_{1-N} \\ c_1 & c_0 & c_{-1} & \cdots & c_{2-N} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ c_{N-2} & \ddots & c_1 & c_0 & c_{-1} \\ c_{N-1} & c_{N-2} & \cdots & c_1 & c_0 \end{pmatrix} \tag{2.18}$$

We only need to store the first column and row to reconstruct the whole toeplitz matrix.

NOTE. A *symmetric* toeplitz matrix is fully defined by its first column.

$$\boldsymbol{c}_1 = [c_0, c_1, \ldots, c_{N-1}]^T \in \mathbb{R}^N.$$

Example 2.7. Let $D = [0, 1]$, and let $z(x, \omega)$ be a mean-zero Gaussian RF with stationary covariance

$$c(x) = e^{-|x|/\ell}. \tag{2.19}$$

Divide D into 100 uniform intervals and choose the grid points

$$x_i = \frac{i - 1}{100}, \quad i = 1, 2, \ldots, 101.$$

The plot of covariance matrix is shown in figure 5.
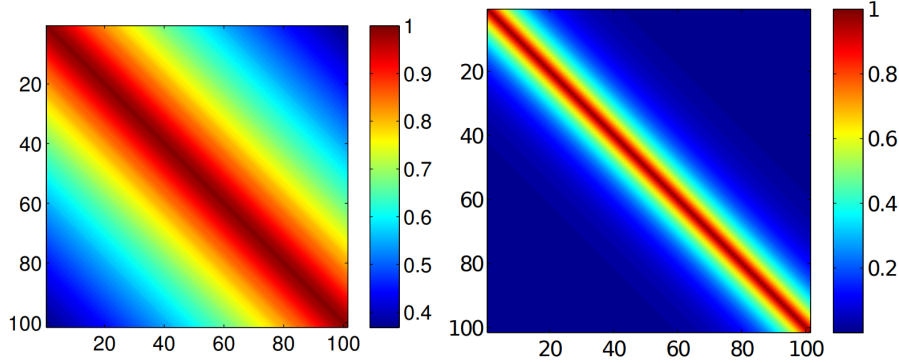
FIGURE 5: MATLAB imagesc plots of the covariance matrix C_N associated with a mean zero Gaussian process with exponential covariance function with $\ell = 1$ (left) and $\ell = 1/10$ (right), sampled at $N = 101$ equally spaced points in $D = [0, 1]$.

Definition 2.8 (Circulant matrix). An $N \times N$ real-valued Topelitz matrix C is circulant if each column is a circular (clockwise) shift of the preceding (previous) column.

$$C = \begin{pmatrix} c_0 & c_{N-1} & \cdots & c_2 & c_1 \\ c_1 & c_0 & c_{N-1} & \cdots & c_2 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ c_{N-2} & \ddots & c_1 & c_0 & c_{N-1} \\ c_{N-1} & c_{N-2} & \cdots & c_1 & c_0 \end{pmatrix} \tag{2.20}$$

Example 2.9. Example of Circulant matrix:

$$C = \begin{pmatrix} 1 & 4 & 3 & 2 \\ 2 & 1 & 4 & 3 \\ 3 & 2 & 1 & 4 \\ 4 & 3 & 2 & 1 \end{pmatrix}$$

is circulant.

Definition 2.10 (Hermitian vector). A vector $\boldsymbol{v} = [v_0, v_1, \ldots, v_{N-1}]^T \in \mathbb{C}^N$ is said to be Hermitian if

$$v_0 \in \mathbb{R}, \quad \text{and } v_j = \overline{v_{N-j}}, \quad j = 1, 2, \ldots, N - 1. \tag{2.21}$$

Remark 2.11. Again, circulant matrix is uniquely determined by the first column

$$\boldsymbol{c}_1 = [c_0, c_1, \ldots, c_{N-1}]^T \in \mathbb{R}^N.$$

If the matrix is also symmetric, the first column has a special structure.

$$c_i = c_j \ \text{ if } \ i + j = N, \text{ for } i, j = 1, 2, \ldots, N - 1. \tag{2.22}$$

i.e., the first column is real Hermitian vector.

Theorem 2.12 (Topelitz Covariance Matrix)**.** *Let $D = [a, b]$. If*

- $z(x, \omega)$ *be a stationary stochastic process on D with* stationary *covariance function $c(x)$.*

- *Suppose we evaluate $z(x, \omega)$ at N uniformly spaced points $x_1, \ldots, x_N$.*

Then the associated covariance matrix C_N is always Toeplitz.

Proof. (Note the points are labelled $x_1, \ldots, x_N$ here). Since the points are uniformly spaced on $[a, b]$, we have $x_i = a + (i - 1)h, i = 1, 2, \ldots, N$, where $h = (b - a)/(N - 1)$. (If there are N points, there are $N - 1$ intervals). Then $x_i - x_j = (i - j)h$. The $(i, j)th$ entry of the covariance matrix is

$$[C_N]_{i,j} = \underbrace{C\left(x_i, x_j\right) = c\left(x_i - x_j\right)}_{\text{Stationary covariance function}} = c((i - j)h)$$

where $c(x)$ is the stationary covariance function. Hence $[C_N]_{i,j}$ is constant for all pairs of point x_i, x_j for which $(i - j)$ is constant. This means that C_N is Toeplitz (by definition 2.6), since all entries along a fixed diagonal have the same value of $i - j$. $\qquad\square$

2.6   Discrete Fourier Transform

Definition 2.13 (Discrete Fourier Transform (DFT))**.** The DFT of a vector $\boldsymbol{v} \in \mathbb{C}^N$ is the vector $\widehat{\boldsymbol{v}} \in \mathbb{C}^N$ with entries

$$\widehat{v}_k := \sum_{j=1}^{N} v_j \omega^{(k-1)(j-1)}, \quad k = 1, 2, \ldots, N. \tag{2.23}$$

where $\omega := e^{-2\pi i/N}$. (This is the Nth unity, and distinguish ω with the iid. random normal distribution samples)

Equivalently, we could write the definition as

$$\widehat{\boldsymbol{v}} := \sqrt{N}W\boldsymbol{v} \in \mathbb{C}^N, \tag{2.24}$$

where $W \in \mathbb{C}^{N \times N}$ is the *Fourier matrix* has entries

$$[W]_{k,j} = \frac{1}{\sqrt{N}}\omega^{(k-1)(j-1)}, \quad k, j = 1, \ldots, N. \tag{2.25}$$

NOTE.

- The MATLAB command `fft(v)` will perform a DFT on the vector $\boldsymbol{v}$.

- This important algorithm will reduce the cost of

$$(\mathbb{C}^{N \times N} \times \mathbb{C}^N) = O(N^2) to O(N \log N)$$

Definition 2.14 (Inverse Discrete Fourier Transform (IDFT)). The IDFT of $\widehat{v} \in C^N$ is the vector

$$v := \frac{1}{\sqrt{N}} W^* \widehat{v} \in \mathbb{C}^N \tag{2.26}$$

where W^* is the *Hermitian transpose* of W. That is, W^* satisfies

$$[W^*]_{k,j} = \overline{W_{j,k}}, \quad k, j = 1, 2, \ldots, N.$$

NOTE.

- The MATLAB command `ifft(v)` will perform a IDFT on the vector v.

- This important algorithm will reduce the cost of

$$(\mathbb{C}^{N \times N} \times \mathbb{C}^N) = O(N^2) \to O(N \log N).$$

- The DFT and IDFT of a *Hermitian vector* is always *real-valued*.

Factorization of Circulant matrix

Theorem 2.15. *Let $C \in \mathbb{R}^{N \times N}$ be a circulant matrix with first column c_1, then*

$$C = W \Lambda W^* \tag{2.27}$$

- $W \in \mathbb{C}^{N \times N}$ *is the $N \times N$ Fourier matrix.*

- Λ *is the $N \times N$ diagonal matrix of eigenvalues with diagonal entries*

$$d = \sqrt{N} W^* c_1. \tag{2.28}$$

- We can compute the eigenvalues using the IDFT of c_1 in MATLAB d=N*`ifft`(c1).

- If C is symmetric, and circulant, by remark 2.11, then c_1 is a Hermitian vector, hence $d \in \mathbb{R}^N$, i.e., the eigenvalues are all real.

- If C is a valid covariance matrix as well as circulant, i.e., the matrix is *symmetric, non-negative definite (positive semi-definite), and circulant*, then by theorem 2.15 we have

$$C = W \Lambda W^* = W \Lambda^{1/2} \Lambda^{1/2} W^* = (W \Lambda^{1/2})(W \Lambda^{1/2})^* \tag{2.29}$$

Problem. Suppose C_N is $N \times N$ covariance matrix that is also circulant, then theorem 2.15 and equation (2.29) give

$$C_N = V V^*, \quad \text{where } V := W \Lambda^{1/2}. \tag{2.30}$$

If we form $Z = \mu + W \Lambda^{1/2} \xi$, where $\xi \sim N(0, I_N)$, what's the distribution of Z?

$$\mathbb{E}[Z] = \mu,$$
$$\text{Cov}[Z] = \text{Cov}[(Z - \mu)(Z - \mu)^T]$$
$$= \text{Cov}[W \Lambda^{1/2} \xi \xi^T \Lambda^{1/2} W^T] \tag{2.31}$$
$$= W \Lambda^{1/2} \text{Cov}[\xi \xi^T] \Lambda^{1/2} W^T$$
$$= W \Lambda^{1/2} I_N \Lambda^{1/2} W^T = W \Lambda W^T \neq W \Lambda W^*.$$

2.7 Complex-Valued Random Variables

Given 2 $\mathbb{R}^N$-valued random variables

$$X_1, X_2 : \Omega \to \mathbb{R}^N. \tag{2.32}$$

We can always create a $\mathbb{C}^N$-valued random variable via

$$X = X_1 + iX_2 \tag{2.33}$$

where $X_1 = \mathfrak{Re}(X)$, $X_2 = \mathfrak{Im}(X)$.

Definition 2.16 (Complex R.V.s). If $X = [X_1, X_2, \ldots, X_N]^T$ is a $\mathbb{C}^N$-valued r.v., then

- The *mean vector* $\boldsymbol{\mu} \in \mathbb{C}^N$ where

$$[\boldsymbol{\mu}]_j = \mathbb{E}[X_j], \quad j = 1, 2, \ldots, N. \tag{2.34}$$

- The *covariance matrix $C_X \in \mathbb{C}^{N \times N}$* where

$$C_X = \mathbb{E}[(X - \boldsymbol{\mu})(X - \boldsymbol{\mu})^*] \in \mathbb{C}^{N \times N} \tag{2.35}$$

where $*$ denotes the Hermitian transpose. That is,

$$[C_X]_{k,j} = \mathrm{Cov}(X_k, X_j) = \mathbb{E}[(X_k - \mathbb{E}[X_k])\overline{(X_j - \mathbb{E}[X_j])}]. \tag{2.36}$$

Example 2.17. Let X_1, X_2 be $\mathbb{R}^N$ random variables with *mean zero*. Define $X = X_1 + iX_2$. Then

$$\begin{aligned} \mathbb{E}[X] &= \mathbb{E}[X_1] + i\,\mathbb{E}[X_2] = \mathbf{0} \\ C_X &= \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^*] = \mathbb{E}[XX^*] \end{aligned} \tag{2.37}$$

Theorem 2.18 (Uncorrelated $\to$ Real-valued). *Suppose we have a complex valued random variable X. If $\mathfrak{Re}(X)$ and $\mathfrak{Im}(X)$ are uncorrelated and mean zero, then the corresponding covariance C_X is $\mathbb{R}$-valued.*

Proof. For complex valued random variable $X = X_1 + iX_2$, the covariance matrix is

$$\begin{aligned} C_X &= \mathbb{E}\left[(X_1 + iX_2)(X_1 + iX_2)^*\right] = \mathbb{E}\left[(X_1 + iX_2)(X_1 - iX_2)^\top\right] \\ &= \mathbb{E}\left[X_1 X_1^\top + X_2 X_2^\top\right] + i\,\mathbb{E}\left[-X_1 X_2^\top + X_2 X_1^\top\right] \end{aligned}$$

Since X_1 and X_2 are assumed to be uncorrelated, $\mathrm{Cov}(X_1, X_2) = \mathbb{E}\left[X_1 X_2^T\right] = \mathbf{0} = \mathbb{E}\left[X_2 X_1^T\right]$. The second term is zero and we have

$$C_X = \mathbb{E}\left[X_1 X_1^T + X_2 X_2^T\right] = \mathbb{E}[X_1 X_1^T] + \mathbb{E}[X_2 X_2^T] = \mathrm{Cov}(X_1) + \mathrm{Cov}(X_2),$$

which is the sum of the covariance matrices of X_1 and X_2 and, in particular, is real-valued. $\square$

Lemma 2.19 (Real-valued $\to$ Uncorrelated). *Let $X = X_1 + iX_2$ where X_1 and X_2 are $\mathbb{R}^N$-valued random variables with mean zero.*

- *If $\mathbb{E}[XX^*]$ and $\mathbb{E}[XX^T]$ are real-valued, then X_1 and X_2 are uncorrelated.*

- *If in addition,* $\mathbb{E}[XX^T] = 0$, *then the covariance matrices of X_1, and X_2 satisfy*

$$C_{X_1} = C_{X_2} = \frac{\mathbb{E}[XX^*]}{2} = \frac{C_X}{2}. \tag{2.38}$$

Proof. Let $X = X_1 + iX_2$, and we have $\overline{X} = X_1 - iX_2$. Hence we could rewrite

$$X_1 = \frac{X + \overline{X}}{2}, \quad X_2 = \frac{X - \overline{X}}{2i} \tag{2.39}$$

then we could evaluate

$$
\begin{aligned}
\mathrm{Cov}(X_1, X_2) &= \mathbb{E}[X_1 X_2^T] \\
&= \frac{1}{4i} \mathbb{E}[(X + \overline{X})(X - \overline{X})^T] \\
&= \frac{1}{4i} \mathbb{E}[XX^T - XX^* + \overline{X}X^T - \overline{X}X^*] \\
&= \frac{1}{4i}(\mathbb{E}[XX^T] - \mathbb{E}[XX^*] + \mathbb{E}[\overline{X}X^T] - \mathbb{E}[\overline{X}X^*])
\end{aligned}
\tag{2.40}
$$

By assumptions that $\mathbb{E}[XX^*]$ and $\mathbb{E}[XX^T]$ are real valued, then we could have

$$
\begin{aligned}
\mathbb{E}[XX^*] &= \overline{\mathbb{E}[XX^*]} = \mathbb{E}[\overline{X}X^T], \\
\mathbb{E}[XX^T] &= \overline{\mathbb{E}[XX^T]} = \mathbb{E}[\overline{X}X^*].
\end{aligned}
\tag{2.41}
$$

then we have the cancellation, and we have $\mathrm{Cov}(X_1, X_2) = 0$ as required.

$$
\begin{aligned}
C_{X_1} = \mathbb{E}[X_1 X_1^T] &= \frac{1}{4} \mathbb{E}[(X + \overline{X})(X + \overline{X})^T] \\
&= \frac{1}{4} \mathbb{E}[XX^T + XX^* + \overline{X}X^T + \overline{X}X^*] \\
&= \frac{1}{2} \mathbb{E}[XX^*] = \frac{C_X}{2}. \quad \text{By } \mathbb{E}[XX^T] = \mathbb{E}[\overline{X}X^*] = 0 \\
&\qquad\qquad\qquad\qquad\quad \text{and equation (2.41)}
\end{aligned}
\tag{2.42}
$$

$\square$

Definition 2.20 (Complex Gaussian Distribution). An $\mathbb{C}^N$ random variable $X = X_1 + iX_2$ (X_1 and X_2 are real valued random variables) follows the complex Gaussian distribution

$$X \sim CN(\mathbf{0}, C_X) \tag{2.43}$$

with real-valued covariance matrix $C_X \in \mathbb{R}^{N \times N}$ if

- X_1 and X_2 are independent

- $X_1, X_2 \sim N\left(\mathbf{0}, \frac{C_X}{2}\right)$.

Example 2.21. Let $X_1, X_2 \sim N(\mathbf{0}, I_N)$ be independent and form $\boldsymbol{\xi} = X_1 + iX_2$, then we have

$$\boldsymbol{\xi} \sim CN(\mathbf{0}, 2I_N). \tag{2.44}$$

2.8   Circulant Covariance Matrix

Suppose C_N is an $N \times N$ real-valued covariance matrix that is

- a valid covariance matrix

- circulant

Recall (Usual factorisation method (based on theorem 2.2)).

- Generate a sample of $\boldsymbol{\xi} \sim N(\mathbf{0}, I_N)$.

- Compute Λ (the eigenvalues of C_N).

- Form $\mathbf{Z} = W\Lambda^{1/2}\boldsymbol{\xi}$.

We will introduce a new method for generating independent samples from

$$N(\mathbf{0}, C_N).$$

However, from equation (2.31), we recognize that the method we introduced in theorem 2.2 will not gives the desired distribution $N(\mathbf{0}, C_N)$. In particular, the resulting $\mathbf{Z}$ is complex-valued.

Modified Factorisation Method

Corollary 2.22. *For circulant symmetric and non-negative definite matrix C_N, we could modify the theorem 2.2*

1. *Generate a sample of $\boldsymbol{\xi} \sim CN(\mathbf{0}, 2I_N)$.*

2. *Compute Λ (the eigenvalues of C_N).*

3. *Form $\mathbf{Z} = W\Lambda^{1/2}\boldsymbol{\xi}$.*

Then $\mathbf{Z} \sim CN(\mathbf{0}, 2C_N)$.

Since C_N is circulant, hence step (2) can be done using IDFT on $\boldsymbol{c}_1$, and step (3) can be done using DFT. Hence the resulting cost will be $O(N \log N)$.

Theorem 2.23. *Using the method in corollary 2.22, we have*

$$\mathfrak{Re}(\mathbf{Z}), \mathfrak{Im}(\mathbf{Z}) \sim N(\mathbf{0}, C_N) \text{ and independent.} \tag{2.45}$$

Proof. It is easy to see that $\mathbf{Z}$ is a linear transformation of a multivariate Gaussian random variable $\boldsymbol{\xi}$, hence $\mathbf{Z}$ is a multivariate Gaussian random variable.

$$\mathbb{E}[\mathbf{Z}] = W\Lambda^{1/2}\mathbb{E}[\boldsymbol{\xi}] = \mathbf{0}. \tag{2.46}$$

Hence if we define $\mathbf{Z} = \mathfrak{Re}(\mathbf{Z}) + i\mathfrak{Im}(\mathbf{Z}) = \mathbf{Z}_1 + i\mathbf{Z}_2$, then $\mathbf{Z}_1$ and $\mathbf{Z}_2$ both mean $\mathbf{0}$.

$$\begin{aligned}
\mathbb{E}[\mathbf{Z}\mathbf{Z}^*] &= \mathbb{E}[W\Lambda^{1/2}\boldsymbol{\xi}\boldsymbol{\xi}^*\Lambda^{1/2}W^*] \\
&= W\Lambda^{1/2}\mathbb{E}[\boldsymbol{\xi}\boldsymbol{\xi}^*]\Lambda^{1/2}W^* \\
&= 2W\Lambda W^* = 2C_N \in \mathbb{R}^N
\end{aligned} \tag{2.47}$$

Since $\boldsymbol{\xi} \sim CN(\mathbf{0}, 2I_N)$, hence by definition 2.20 if we write $\boldsymbol{\xi} = \boldsymbol{\xi}_1 + i\boldsymbol{\xi}_2$, then $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ are independent ($\mathbb{E}[\boldsymbol{\xi}_1\boldsymbol{\xi}_2^T] = \mathbb{E}[\boldsymbol{\xi}_2\boldsymbol{\xi}_1^T] = \mathbf{0}$) and distributed as $N(\mathbf{0}, I_N)$.

$$
\begin{aligned}
\mathbb{E}[ZZ^T] &= \mathbb{E}[W\Lambda^{1/2}\boldsymbol{\xi}\boldsymbol{\xi}^T\Lambda^{1/2}W^T] \\
&= W\Lambda^{1/2}\,\mathbb{E}[\boldsymbol{\xi}\boldsymbol{\xi}^T]\Lambda^{1/2}W^T \\
\mathbb{E}[\boldsymbol{\xi}\boldsymbol{\xi}^T] &= \mathbb{E}[(\boldsymbol{\xi}_1 + i\boldsymbol{\xi}_2)(\boldsymbol{\xi}_1 + i\boldsymbol{\xi}_2)^T] \\
&= \mathbb{E}[\boldsymbol{\xi}_1\boldsymbol{\xi}_1^T] + i\,\mathbb{E}[\boldsymbol{\xi}_1\boldsymbol{\xi}_2^T] + i\,\mathbb{E}[\boldsymbol{\xi}_2\boldsymbol{\xi}_1^T] - \mathbb{E}[\boldsymbol{\xi}_2\boldsymbol{\xi}_2^T] = \mathbf{0} \\
\mathbb{E}[ZZ^T] &= \mathbf{0} \in \mathbb{R}^N.
\end{aligned}
\tag{2.48}
$$

By lemma 2.19, Z_1 and Z_2 are mean zero, $\mathbb{E}[ZZ^*], \mathbb{E}[ZZ^T] \in \mathbb{R}^N$, the we could conclude that

$$
Z_1, Z_2 \sim N(\mathbf{0}, C_N), \quad \text{and independent.} \tag{2.49}
$$

$\square$

2.9 Circulant Embedding Method

Recall.

- We now have the method for generating a pair of samples from $N(\mathbf{0}, C_N)$ with C_N is circulant (from corollary 2.22 and theorem 2.23).

- Unfortunately, covariance matrices are *not* usually circulant.

  Let $D \subset \mathbb{R}$ and let $z(x, \omega)$ be a mean-zero Gaussian RF, if

- the sample points $x_1, x_2, \ldots, x_N$ are uniformly spaced.

- the covariance function is stationary: $C(x, y) = c(x - y)$.

Then $Z = [z(x_1, \omega), \ldots, z(x_N, \omega)]^T \sim N(\mathbf{0}, C_N)$ where C_N is always Toeplitz (theorem 2.12).

Circulant Embedding

Any symmetric Toeplitz matrix C_N can be embedded inside a *larger* symmetric circulant matrix $\widetilde{C}$ of the form

$$
\widetilde{C} = \left(\begin{array}{c|c} C_N & A^T \\ \hline A & B \end{array} \right) \tag{2.50}
$$

Example 2.24 ($N = 4$). For

$$
C_N = \begin{pmatrix} 5 & 2 & 3 & 4 \\ 2 & 5 & 2 & 3 \\ 3 & 2 & 5 & 2 \\ 4 & 3 & 2 & 5 \end{pmatrix} \tag{2.51}
$$

Clearly, this matrix is symmetric and Toeplitz. This can be embedded into a larger symmetric circulant matrix.

$$
\tilde{C} = \left(\begin{array}{cccc|cc} 5 & 2 & 3 & 4 & 3 & 2 \\ 2 & 5 & 2 & 3 & 4 & 3 \\ 3 & 2 & 5 & 2 & 3 & 4 \\ 4 & 3 & 2 & 5 & 2 & 3 \\ \hline 3 & 4 & 3 & 2 & 5 & 2 \\ 2 & 3 & 4 & 3 & 2 & 5 \end{array} \right) \tag{2.52}
$$

If we focus on the first column of $\widetilde{C}$, $\widetilde{c}_1 = [5, 2, 3, 4, 3, 2]^T$ which is a Hermitian vector.

Minimal circulant extension

Definition 2.25. Let C_N be an $N \times N$ symmetric Toeplitz matrix with first column

$$c_1 = [c_0, c_1, \ldots, c_{N-2}, c_{N-1}]^T. \tag{2.53}$$

The *minimal circulant extension* $\widetilde{C}$ is the symmetric and circulant matrix of size $P \times P$ where $P = 2N - 2$ with first column

$$\widetilde{c_1} = [c_0, \underbrace{c_1, \ldots, c_{N-2}}_{\substack{\text{take this part,} \\ \text{flip the order}}}, c_{N-1} | c_{N-2}, \ldots, c_1]^T. \tag{2.54}$$

That is,

$$\widetilde{C} = \left(\begin{array}{c|c} C_N & A^T \\ \hline A & B \end{array} \right) \tag{2.55}$$

where $A \in \mathbb{R}^{(N-2) \times N}$, and $B \in \mathbb{R}^{(N-2) \times (N-2)}$.

Recovering Samples

If the extension matrix is a *valid covariance matrix*, then we can generate samples from

$$\widetilde{Z} \sim N(\mathbf{0}, \widetilde{C}) \quad \widetilde{C} \text{ from equation } (2.55)$$

using the DFT-method. Such samples have the structure

$$\widetilde{Z} = \left(\frac{Z}{Y} \right) \quad Z \in \mathbb{R}^N, \, Y \in \mathbb{R}^{N-2}$$

and it can be shown that $Z \sim N(\mathbf{0}, C_N)$.

Proof. Write $\widetilde{Z} = [Z, Y]^\top$ where Z is the $\mathbb{R}^N$-valued random variable of interest. If $\widetilde{Z} \sim N(\mathbf{0}, \widetilde{C})$ then clearly Z has mean vector zero. The covariance matrix of Z is $\mathbb{E}\left[ZZ^T\right]$ and the covariance matrix of $\widetilde{Z}$ is $\widetilde{C} = \mathbb{E}\left[\widetilde{Z}\widetilde{Z}^T\right]$. We have

$$\widetilde{C} = \left(\begin{array}{cc} \mathbb{E}[ZZ^T] & \mathbb{E}[ZY^T] \\ \mathbb{E}[YZ^T] & \mathbb{E}[YY^T] \end{array} \right) \underbrace{=}_{\text{equation } (2.55)} \left(\begin{array}{cc} C_N & B^T \\ B & D \end{array} \right) \tag{2.56}$$

Hence the covariance matrix of Z, $\mathbb{E}[ZZ^T]$, is C_N. Finally, Z is Gaussian because it is a linear transformation of the Gaussian random variable $\widetilde{Z}$. $\qquad \square$

Theorem 2.26 (Circulant Embedding Method)**.** *To draw a pair of independent samples from $N(\mathbf{0}, C_N)$ when C_N is Toeplitz:*

- *Construct first column c_1 of C_N*

- *Create first column $\widetilde{c}_1$ of the minimal circulant extension $\widetilde{C}$*

- *Obtain two independent samples from $N(\mathbf{0}, \widetilde{C})$ (apply IDFT & DFT from corollary 2.22, and using theorem 2.23)*

- *Extract two independent samples from $N(\mathbf{0}, C_N)$*

2.10   Approximate Circulant Embedding

Recall. Any symmetric Toeplitz matrix C_N can be embedded inside a *larger* symmetric circulant matrix $\widetilde{C}$ of the form

$$\widetilde{C} = \left(\begin{array}{c|c} C_N & A^T \\ \hline A & B \end{array} \right) \tag{2.57}$$

If $\widetilde{C}$ is *non-negative definite*, we can generate samples from $N(\mathbf{0}, \widetilde{C})$ and easily recover samples from $N(\mathbf{0}, C_N)$ using theorem 2.26.

However, the method is not working if we have a invalid (have negative eigenvalues) $\widetilde{C}$.

Matrix Splitting

If we have

$$\widetilde{C} = W \Lambda W^*.$$

We can split Λ into two order diagonal matrices,

$$\Lambda_+ - \Lambda_- \tag{2.58}$$

Let $\boldsymbol{d}$ be the vector containing the eigenvalues of $\widetilde{C}$ and define

$$[\Lambda_+]_{i,i} = \begin{cases} d_i & \text{if } d_i \geq 0 \\ 0 & \text{otherwise} \end{cases}, \qquad [\Lambda_-]_{i,i} = \begin{cases} -d_i & \text{if } d_i < 0 \\ 0 & \text{otherwise} \end{cases} \tag{2.59}$$

Approximate Circulant Embedding

We now have the *matrix splitting*:

$$\widetilde{C} = W \Lambda_+ W^* - W \Lambda_- W^* := \widetilde{C}_{pos} - \widetilde{C}_{neg} \approx \widetilde{C}_{pos}, \tag{2.60}$$

if the negative eigenvalues are small.

- The matrix $\widetilde{C}_{pos}$ is circulant and *non-negative definite*.

- We can apply the method (corollary 2.22) with $\widetilde{C}_{pos}$ in place of $\widetilde{C}$:

$$\widehat{\mathbf{Z}} = W \Lambda_+^{1/2} \boldsymbol{\xi}, \quad \text{where } \boldsymbol{\xi} \sim CN(\mathbf{0}, 2 I_P). \tag{2.61}$$

It can be shown that

$$\widehat{\mathbf{Z}} \sim CN(\mathbf{0}, 2\widetilde{C}_{pos}), \qquad \mathfrak{Re}(\widehat{\mathbf{Z}}), \mathfrak{Im}(\widehat{\mathbf{Z}}) \sim N(\mathbf{0}, \widetilde{C}_{pos}) \tag{2.62}$$

Proof. $\mathfrak{Re}(\widehat{\mathbf{Z}})$ and $\mathfrak{Im}(\widehat{\mathbf{Z}})$ are both linear transformation of Gaussian $\mathfrak{Re}(\boldsymbol{\xi})$ and $\mathfrak{Im}(\boldsymbol{\xi})$. Hence $\mathfrak{Re}(\widehat{\mathbf{Z}})$ and $\mathfrak{Im}(\widehat{\mathbf{Z}})$ are Gaussian.

$$\begin{aligned} \mathbb{E}[\widehat{\mathbf{Z}}\widehat{\mathbf{Z}}^*] &= \mathbb{E}[W \Lambda_+^{1/2} \boldsymbol{\xi} \boldsymbol{\xi}^* \Lambda_+^{1/2} W^*] \\ &= W \Lambda_+^{1/2} (2 I_P) \Lambda_+^{1/2} W^* \\ &= 2 W \Lambda_+ W^* = 2\widetilde{C}_{pos} \in \mathbb{R}^{P \times P} \end{aligned} \tag{2.63}$$

$$\mathbb{E}[\widehat{\mathbf{Z}}\widehat{\mathbf{Z}}] = \mathbf{0} \quad \text{By proof of theorem 2.23}$$

By lemma 2.19, $\mathfrak{Re}(\widehat{\mathbf{Z}})$ and $\mathfrak{Im}(\widehat{\mathbf{Z}})$ are independent and

$$\mathfrak{Re}(\widehat{\mathbf{Z}}), \mathfrak{Im}(\widehat{\mathbf{Z}}) \sim N(\mathbf{0}, \widetilde{C}_{pos})$$

$\square$

Recovering Approximate Samples

Let $\mathbf{Z}_1$ and $\mathbf{Z}_2$ denote the first N components of these random vectors in equation (2.62). Then

$$\mathbf{Z}_1, \mathbf{Z}_2 \sim N(\mathbf{0}, \widehat{C}) \tag{2.64}$$

where $\widehat{C}$ is the $N \times N$ leading block of $\widetilde{C}_{pos}$.

$$\widetilde{C}_{pos} = \left(\begin{array}{c|c} \widehat{C} & A^T \\ \hline A & B \end{array} \right) \tag{2.65}$$

However, since $\widehat{C} \neq C_N$, we do note have samples from the target distribution $N(\mathbf{0}, C_N)$.

Theorem 2.27. *It can be shown that,*

$$\|C_N - \widehat{C}\|_2 \leq \rho\left(\Lambda_-\right)$$

where $\rho(\cdot)$ denotes the spectral radius.

Proof. (Not examinable).

$$\begin{aligned} \|\widetilde{C} - \widetilde{C}_{pos}\|_2 = \|(\widetilde{C}_{pos} - \widetilde{C}_{neg}) - \widetilde{C}_{pos}\|_2 = \| - \widetilde{C}_{neg}\|_2 = \rho(\widetilde{C}_{neg}) = \rho(\Lambda_-) \\ \text{By Cauchy interlacing theorem} \quad \|C_N - \widehat{C}\|_2 \leq \|\widetilde{C} - \widetilde{C}_{pos}\|_2 = \rho(\Lambda_-) \end{aligned} \tag{2.66}$$

$\square$

3   Sampling Methods for Forward UQ in ODEs

3.1   (Revision) Monte Carlo Estimate of Mean

Let $X : \Omega \to \mathbb{R}$ be a real-valued random variable with

$$\mathbb{E}[X] = \mu, \quad \mathbb{V}[X] = \sigma^2 < \infty. \tag{3.1}$$

Let $X_1, \ldots, X_M$ be independent samples of X, i.e.,

$$X_i = X(\omega_i), \quad i = 1, 2, \ldots, M. \tag{3.2}$$

The standard MC estimate for μ is

$$\mu_M := \frac{1}{M} \sum_{i=1}^{M} X_i. \tag{3.3}$$

Recall.

- μ_M is a *random variable*, and so is the error $|\mu_M - \mu|$.

- μ_M is an *unbiased* estimate for μ since

$$\mathbb{E}[\mu_M] = \frac{1}{M}(M\mu) = \mu. \tag{3.4}$$

Mean-Square Error

$$\mu \approx \mu_M := \frac{1}{M} \sum_{i=1}^{M} X_i. \tag{3.5}$$

Definition 3.1 (Mean-Square Error)**.** The *mean-square error* of the estimator μ_M is

$$\text{MSE}(\mu_M) = \mathbb{E}[(\mu_M - \mu)^2] = \|\mu_M - \mu\|_{L^2(\Omega)}^2 \tag{3.6}$$

Proposition 3.2. *Since μ_M is an unbiased estimator,*

$$\text{MSE}(\mu_M) = \mathbb{E}[(\mu_M - \mathbb{E}[\mu_M])^2] = \mathbb{V}[\mu_M] = \frac{1}{M^2}(M\sigma^2) = \frac{\sigma^2}{M}. \tag{3.7}$$

The root mean-square error (RMSE) *is*

$$\textbf{RMSE}(\mu_M) = \sqrt{\text{MSE}(\mu_M)} = \|\mu_M - \mu\|_{L^2(\Omega)} = \frac{\sigma}{\sqrt{M}} \tag{3.8}$$

Theorem 3.3 (Chebyshev Inequality)**.** *For any $\epsilon > 0$, we have*

$$\mathbb{P}(|\mu_M - \mu| \geq M^{-1/2+\epsilon}) \leq \sigma^2 M^{-2\epsilon}. \tag{3.9}$$

As $M \to \infty$, the probability of the error being larger than $O(M^{-1/2+\epsilon})$ converges to zero.

Informally, we often say that the MC error is $O(M^{-1/2})$.

Estimating σ^2

We can estimate σ^2 by

$$\sigma_M^2 = \frac{1}{M-1} \sum_{i=1}^{M} (X_i - \mu_M)^2. \tag{3.10}$$

This is also an *unbiased* estimator.

3.2   IVPs & Explicit Euler Method

We consider initial value problems (IVPs) of the form:

$$\frac{\mathrm{d}\boldsymbol{u}(t)}{\mathrm{d}t} = \boldsymbol{f}(\boldsymbol{u}(t)), \quad 0 < t < T, \tag{3.11}$$

with vector-valued solution

$$\boldsymbol{u}(t) = [u_1(t), \dots, u_d(t)]^T$$

with inital condition $\boldsymbol{u}(0) = \boldsymbol{u}_0$.

Example 3.4 (Population Model)**.** We introduce the following *test problem* ($d = 2$ case)

$$\frac{\mathrm{d}}{\mathrm{d}t} \begin{pmatrix} u_1(t) \\ u_2(t) \end{pmatrix} = \begin{pmatrix} u_1(t)\,(1 - u_2(t)) \\ u_2(t)\,(u_1(t) - 1) \end{pmatrix}, \quad \begin{pmatrix} u_1(0) \\ u_2(0) \end{pmatrix} = \begin{pmatrix} u_{1,0} \\ u_{2,0} \end{pmatrix} \tag{3.12}$$

A typical solution to this problem shown in figure 6.
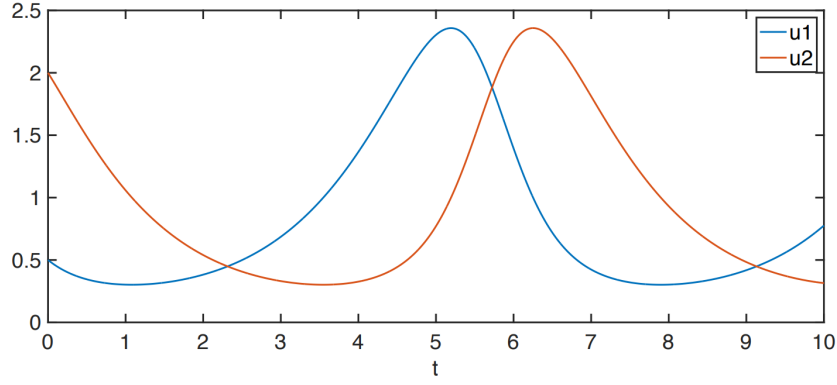
FIGURE 6: Explicit Euler approximation to the solution $\boldsymbol{u}(t) = [u_1(t), u_2(t)]^T$ to Example 3.4 computed with $T = 10$, $N = 10^5$ and initial condition $\boldsymbol{u}_0 = [0.5, 2]^T$.

Explicit Euler Method

For

$$\frac{\mathrm{d}\boldsymbol{u}(t)}{\mathrm{d}t} = \boldsymbol{f}(\boldsymbol{u}(t)), \quad 0 < t < T, \quad \boldsymbol{u}(0) = \boldsymbol{u}_0. \tag{3.13}$$

Time-stepping methods estimate the solution at a set of grid points t_n.

$$\boldsymbol{u}_n \approx \boldsymbol{u}(t_n)$$

Theorem 3.5 (Explict Euler Method).

- *Divide $[0, T]$ into N intervals, and define the step size*

$$\Delta t \coloneqq T/N.$$

- *Choose the* grid point $t_n = n\Delta t$, $n = 0, 1, \ldots, N$.

- *Starting with $\boldsymbol{u}_0$, compute*

$$\boldsymbol{u}_{n+1} = \boldsymbol{u}_n + \Delta t \, \boldsymbol{f}(\boldsymbol{u}_n), \quad n = 1, 2, \ldots, N. \tag{3.14}$$

What can we say about the *error* $\|\boldsymbol{u}(t_n) - \boldsymbol{u}_n\|_2$ at time t_n?

Theorem 3.6. *Suppose $\boldsymbol{f} : \mathbb{R}^d \to \mathbb{R}^d$ satisfies the Lipschitz condition*

$$\|\boldsymbol{f}(\boldsymbol{v}) - \boldsymbol{f}(\boldsymbol{w})\|_2 \leq L\|\boldsymbol{x} - \boldsymbol{w}\|_2, \quad \forall \boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^d. \tag{3.15}$$

for some $L > 0$ and that the true solution $\boldsymbol{u}(t)$ is twice continuosly differentiable. *Then, for any $\boldsymbol{u}_0 \in \mathbb{R}^d$, there exist a constant $K > 0$ independent of Δt such that*

$$\max_{0 \leq t_n \leq T} \|\boldsymbol{u}(t_n) - \boldsymbol{u}_n\|_2 \leq K\Delta t. \tag{3.16}$$

Hence the explicit Euler error is $O(\Delta t)$.

3.3   IVP & Uncertain Initial Condition

Example 3.7 (Population Model).

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} u_1(t) \\ u_2(t) \end{pmatrix} = \begin{pmatrix} u_1(t)\,(1 - u_2(t)) \\ u_2(t)\,(u_1(t) - 1) \end{pmatrix}, \qquad \begin{pmatrix} u_1(0) \\ u_2(0) \end{pmatrix} = \begin{pmatrix} u_{1,0} \\ u_{2,0} \end{pmatrix} \tag{3.17}$$

If we model $u_{1,0}$ and $u_{2,0}$ as *random variables*, then the ODE solution $\boldsymbol{u} : [0, T] \times \Omega \to \mathbb{R}^2$ is a vector-valued *stochastic process*

$$\boldsymbol{u}(t, \omega) = \begin{pmatrix} u_1(t, \omega) \\ u_2(t, \omega) \end{pmatrix}. \tag{3.18}$$

For this example, we modelled as

$$u_{1,0} \sim U(\overline{u}_1 - \tau, \overline{u}_1 + \tau), \qquad u_{2,0} \sim U(\overline{u}_2 - \tau, \overline{u}_2 + \tau). \tag{3.19}$$

and we choose: $\overline{u}_1 = 0.5$, $\overline{u}_2 = 2$, and $\tau = 0.1$. For each sample $\boldsymbol{u}_0^i = u_0(\omega_i)$ of the initial condition, we can apply the explicit Euler method (with N intervals) to approximate

$$\boldsymbol{u}_n^i \approx \boldsymbol{u}(t_n, \omega_i), \quad n = 1, 2, \ldots, N, \tag{3.20}$$

at the grid points t_n on $[0, T]$. If we choose $T = 6$ and $N = 10^6$, we have the figure 7.
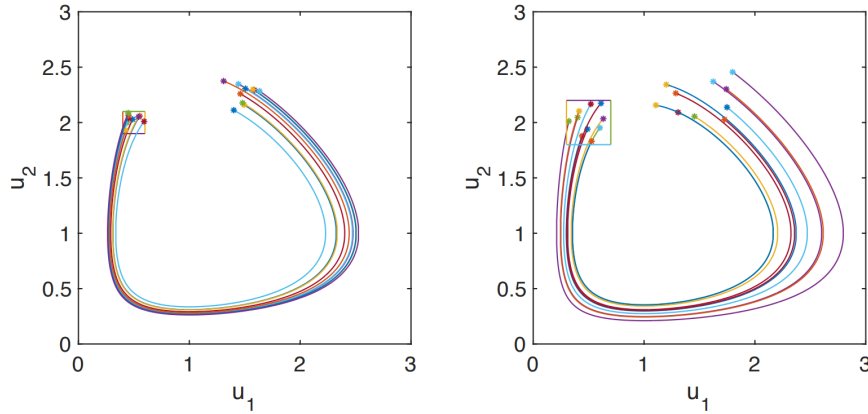
FIGURE 7: Ten samples of the explicit Euler approximation to the solution to Example 3.7 computed with $T = 6$, and $N = 10^6$ and $\tau = 0.1$ (left) and $\tau = 0.2$ (right).

Estimating Mean of Quantity of Intest

Suppose our quantity of interest is

$$X := u_2(T, \omega). \tag{3.21}$$

Key Question. How can we estimate $\mathbb{E}[X]$ efficiently and accurately?

But we don't have access to samples of the exact solution $X_i := u_2(T, \omega_i)$. For a fixed N, we can define the random variable

$$X_N = u_{2,N}, \quad \text{where } \boldsymbol{u}_N(\omega) = \begin{pmatrix} u_{1,N}(\omega) \\ u_{2,N}(\omega) \end{pmatrix} \tag{3.22}$$

is the *explicit Euler* approximation at $t_N = T$.

Let X_N^i denote a sample of X_N and define the Monte Carlo estimate

$$\mu_{M,N} := \frac{1}{M} \sum_{i=1}^{M} X_N^i \approx \mathbb{E}[X_N] \approx \mathbb{E}[X]. \tag{3.23}$$

If we use $\mu_{M,N}$ to estimate $\mathbb{E}[X]$, then the error depends on:

- the number of *samples M*,

- the number of *intervals N* (equivalently, the *step size Δt*).

3.4   Standard Monte Carlo: Cost & Accuracy

Let X be a $\mathbb{R}$-valued random variable with mean

$$\mathbb{E}[X] = \mu. \tag{3.24}$$

Let Equation (3.23) be the standard MC estimate.

Question. How do we choose M, N so that we can guarantee

$$\mathbf{RMSE}(\mu_{M,N}) \leq \epsilon. \tag{3.25}$$

and what will be the corresponding cost?

Theorem 3.8. *Let $\mu_{M,N} := \frac{1}{M} \sum_{i=1}^{M} X_N^i \approx \mu$, then*

$$\mathrm{MSE}(\mu_{M,N}) = \mathbb{E}[(\mu_{M,N} - \mu)^2] = \underbrace{M^{-1}\mathbb{V}[X_N]}_{\text{MC Samping Error}} + \underbrace{(\mathbb{E}[X_N - X])^2}_{\text{Bias term } \mathbb{E}[X_N] \neq \mathbb{E}[X]} \tag{3.26}$$

Proof.

$$\begin{aligned}
\mathrm{MSE}(\mu_{M,N}) &= \mathbb{E}[\mu_{M,N} - 2\mu\mu_{M,N} + \mu^2] \\
&= \mathbb{E}[\mu_{M,N}^2] - 2\mu\,\mathbb{E}[\mu_{M,N}] + \mu^2 \\
&= \mathbb{E}[\mu_{M,N}^2] - \mathbb{E}[\mu_{M,N}]^2 + \mathbb{E}[\mu_{M,N}]^2 - 2\mu\,\mathbb{E}[\mu_{M,N}] + \mu^2 \\
&= \mathbb{V}[\mu_{M,N}] + (\mathbb{E}[\mu_{M,N}] - \mu)^2 \\
&= \mathbb{V}[\mu_{M,N}] + (\mathbb{E}[\mu_{M,N} - X])^2 \quad (\mu = \mathbb{E}[X]).
\end{aligned}$$

$$\mathbb{V}[\mu_{M,N}] = \mathbb{V}\Big[\frac{1}{M} \sum_{i=1}^{M} X_N^i\Big] \tag{3.27}$$

$$= \frac{1}{M^2} \sum_{i=1}^{M} \mathbb{V}[X_N^i] = \frac{1}{M}\mathbb{V}[X_N].$$

$$\mathbb{E}[\mu_{M,N}] = \frac{1}{M} \sum_{i=1}^{M} \mathbb{E}[X_N^i] = \mathbb{E}[X_N]$$

$$\mathrm{MSE}(\mu_{M,N}) = M^{-1}\mathbb{V}[X_N] + (\mathbb{E}[X_N - X])^2 \quad \text{as required.}$$

$\square$

Bias : Accuracy of X_N

To investigate the *bias*, we need information about the quality of the approximation, $X_N \approx X$, and how the error depends on N.

Assumption 3.9. *We assume that $\exists C$ and $\alpha > 0$ such that*

$$|\mathbb{E}[X_N - X]| \leq CN^{-\alpha}, \tag{3.28}$$

where C is independent of N.

- *X_N converges in mean to X as $N \to \infty$,*

- *α is the rate of convergence.*

Cost to Controlling the MSE

To obtain **RMSE**$(\mu_{M,N}) \leq \epsilon$, where ϵ is a fixed tolerance, we need MSE$(\mu_{M,N}) \leq \epsilon^2$. From Theorem 3.8, this is satisfied if

$$M^{-1}\mathbb{V}[X_N] \leq \frac{\epsilon^2}{2} \quad \& \quad (\mathbb{E}[X_N - X])^2 \leq \frac{\epsilon^2}{2} \tag{3.29}$$

Hence we could choose

- $M \geq 2\mathbb{V}[X_N]\epsilon^{-2} \to M = O(\epsilon^{-2})$.

- $N \geq (2C^2)^{1/2\alpha}\epsilon^{-1/\alpha} \to N = O(\epsilon^{-1/\alpha})$. (From Assumption 3.9)

Theorem 3.10 (Cost of MC estimator)**.** *If the cost of generating one sample of X_N is $O(N^\beta)$, then the cost to compute $\mu_{M,N}$ is*

$$C_{MC} = M \times O(N^\beta). \tag{3.30}$$

The cost to deliver an estimate such that **RMSE**$(\mu_{M,N}) \leq \epsilon$ *is*

$$C_{MC} = O(\epsilon^{-2}) \times O(\epsilon^{-\beta/\alpha}) = O(\epsilon^{-2-\beta/\alpha}). \tag{3.31}$$

3.5   Multilevel Monte Carlo : Key idea

Recap: Let X be real-valued r.v. with mean $\mathbb{E}[X] = \mu$. Suppose $X_N \approx X$ is another r.v. associated with a *discretization parameter N* and the standard Monte Carlo estimate is:

$$\mu_{M,N} := \frac{1}{M}\sum_{i=1}^{M}X_N^i \approx \mu$$

with the *mean-square error*

$$\text{MSE}(\mu_{M,N}) = M^{-1}\underbrace{\mathbb{V}[X_N]}_{\text{Sampling error}} + (\underbrace{\mathbb{E}[X_N - X]}_{\text{bias due to approximation}})^2$$

But we have some key disadvantages:

1. Error depends on $\mathbb{V}[X_N]$.

2. Cost to achieve **RMSE** $\leq \epsilon$ often too high.

Sequence of Approximation

Suppose we have access to samples of a sequence of r.v.s $X_{N_0}, X_{N_1}, \ldots, X_{N_L}$ that each approximate X, where N_0 is fixed and

$$N_\ell = sN_{\ell-1}, \quad \ell = 1, 2, \ldots, L,$$

for some $s > 1$. (We ensure that $N_0 < N_1 < \cdots < N_L$)

Here ℓ is the *level* number of the approximation X_{N_ℓ}.

Assumption 3.11.

- X_{N_ℓ} *is a more* accurate *approxoimation of X as* $\ell \to \infty$.

- *Samples of* X_{N_ℓ} *Become more* expensive *to generate as* $\ell \to \infty$.

MLMC : Basic Ideas

In the MLMC method, we use a *decreasing* number of samples of X_{N_ℓ} to estimate $\mu = \mathbb{E}[X]$ as the level number ℓ *increases*.

Recall. Samples of X_{N_0} are cheaper to compute and samples of X_{N_ℓ} are the most *expensive* to compute.

3.6   MLMC Estimator : Definition & Properties

The standard MC estimator of $\mu = \mathbb{E}[X]$ associated with level L is

$$\mu_{M,N_L} := \frac{1}{M} \sum_{i=1}^{M} X_{N_L}^i \approx \mu.$$

Telescoping Sum

First observe that $\mathbb{E}[X_{N_L}]$ can be decomposed as follows:

$$\mathbb{E}[X_{N_L}] = \mathbb{E}[X_{N_0}] + \sum_{\ell=1}^{L} \left\{ \mathbb{E}[X_{N_\ell}] - \mathbb{E}[X_{N_{\ell-1}}] \right\},$$

using linearity of $\mathbb{E}[\cdot]$, we have

$$\mathbb{E}[X_{N_L}] = \mathbb{E}[X_{N_0}] + \sum_{\ell=1}^{L} \underbrace{\mathbb{E}[X_{N_\ell} - X_{N_{\ell-1}}]}_{\text{Correction terms}}.$$

By defining

$$Y_0 = X_{N_0} \text{ and } Y_\ell := X_{N_\ell} - X_{N_{\ell-1}} \text{ for } \ell = 1, 2, \ldots, L.$$

We could write

$$\mathbb{E}[X_{N_L}] = \mathbb{E}[Y_0] + \sum_{\ell=1}^{L} \mathbb{E}[Y_\ell] \to \text{The Telescoping Sum}$$

MLMC Estimator

Idea: Estimate the terms associated with each level *separately* using the standard MC method with M_ℓ samples.

1. On *level* 0, define

$$\widehat{\mu_0} := \frac{1}{M_0} \sum_{i=1}^{M_0} Y_0^i,$$

   using M_0 *independent* samples of $Y_0 = X_{N_0}$.

2. On *level* $1, 2, \ldots, L$, define

$$\widehat{\mu_\ell} := \frac{1}{M_\ell} \sum_{i=1}^{M_\ell} Y_\ell^i,$$

   using M_ℓ *independent* samples of $Y_\ell := X_{N_\ell} - X_{N_{\ell-1}}$.

NOTE. To stress that samples should be *independent* for each level ℓ, we can also use the alternative notation:

$$\widehat{\mu_\ell} := \frac{1}{M_\ell} \sum_{i_\ell=1}^{M_\ell} Y_{N_\ell}^{i_\ell}.$$

Using the definition of Y_{N_ℓ} gives,

$$Y_{N_\ell}^i = X_{N_\ell}^i - X_{N_{\ell-1}}^i = X_{N_\ell}(\omega_i) - X_{N_{\ell-1}}(\omega_i)$$

To compute $\widehat{\mu_\ell}$, $X_{N_\ell}^i$ and $X_{N_{\ell-1}}^i$ correspond to the same input sample.
Combing the estimates for each term, the final MLMC estimator is

$$\widehat{\mu}_{ML} := \sum_{\ell=0}^{L} \widehat{\mu_\ell}$$

Question.

1. How to choose L? (No. of levels)

2. How to choose $M_0, M_1, \ldots, M_L$? (No. of samples at each level)

Theorem 3.12 (Mean and Variance of MLMC estimator)**.**

$$\mathbb{E}[\widehat{\mu}_{ML}] = \mathbb{E}[\widehat{\mu_L}] = \mathbb{E}[X_{N_L}],$$

$$\mathbb{V}[\widehat{\mu}_{ML}] = \sum_{\ell=0}^{L} M_\ell^{-1} \mathbb{V}[Y_\ell] \tag{3.32}$$

Proof.

$$\mathbb{E}[\widehat{\mu}_{ML}] = \mathbb{E}[\sum_{\ell=0}^{L} \widehat{\mu_\ell}] = \sum_{\ell=0}^{L} \mathbb{E}[\widehat{\mu_\ell}].$$

We evaluate these separately

$$\mathbb{E}[\widehat{\mu}_0] = \mathbb{E}[\frac{1}{M_0} \sum_{i=1}^{M_0} X_{N_0}^i] = \frac{1}{M_0} \sum_{i=1}^{M_0} \mathbb{E}[X_{N_0}^i] = \mathbb{E}[X_{N_0}].$$

$$\mathbb{E}[\widehat{\mu}_\ell] = \cdots = \mathbb{E}[X_{N_\ell}] - \mathbb{E}[X_{N_{\ell-1}}].$$

Substituted into the first equation:

$$\mathbb{E}[\widehat{\mu}_{ML}] = \mathbb{E}[X_{N_0}] + \mathbb{E}[X_{N_1}] - \mathbb{E}[X_{N_0}] + \cdots + \mathbb{E}[X_{N_L}] - \mathbb{E}[X_{N_{L-1}}] = \mathbb{E}[X_{N_L}]$$

$$\mathbb{V}[\widehat{\mu}_{ML}] = \mathbb{V}[\sum_{\ell=0}^{L} \widehat{\mu}_\ell] = \sum_{\ell=0}^{L} \mathbb{V}[\widehat{\mu}_\ell].$$

For $\ell = 0, 1, 2 \ldots, L$, we always have

$$\mathbb{V}[\widehat{\mu}_\ell] = \mathbb{V}[\frac{1}{M_\ell} \sum_{i=1}^{M_\ell} Y_\ell] = \frac{1}{M_\ell^2} \sum_{i=1}^{M_\ell} \mathbb{V}[Y_\ell] = M_\ell^{-1} \mathbb{V}[Y_\ell]$$

Substituted into the first equation

$$\mathbb{V}[\widehat{\mu}_{ML}] = \sum_{\ell=0}^{L} \mathbb{V}[\widehat{\mu}_\ell] = \sum_{\ell=0}^{L} M_\ell^{-1} \mathbb{V}[Y_\ell].$$

$\square$

3.7 MLMC : Mean Square Error

Goal: Compute an estimator $\widehat{\mu}_{ML}$ such that

$$\mathbf{RMSE}(\widehat{\mu}_{ML}) \leq \epsilon \iff \mathrm{MSE}(\widehat{\mu}_{ML}) \leq \epsilon^2$$

Theorem 3.13 (Mean Square Error of MLMC estimator)**.** *The mean-square error of the MLMC estimator* $\widehat{\mu}_{ML}$ *satisfies*

$$\mathrm{MSE}(\widehat{\mu}_{ML}) = \sum_{\ell=0}^{L} M_\ell^{-1} \mathbb{V}[Y_\ell] + (\mathbb{E}[X_{N_L} - X])^2$$

Proof.

$$\begin{aligned}
\mathrm{MSE}(\widehat{\mu}_{ML}) &= \mathbb{E}\left[(\widehat{\mu}_{ML} - \mu)^2\right] \\
&= \mathbb{E}\left[((\widehat{\mu}_{ML} - \mathbb{E}\left[\widehat{\mu}_{ML}\right]) + (\mathbb{E}\left[\widehat{\mu}_{ML}\right] - \mu))^2\right] \\
&= \mathbb{E}\left[\left(\underbrace{\widehat{\mu}_{ML} - \mathbb{E}\left[\widehat{\mu}_{ML}\right]}_{\text{R.V. with mean }0}\right)^2\right] + \mathbb{E}\left[\left(\underbrace{\mathbb{E}\left[\widehat{\mu}_{ML}\right] - \mu}_{\text{real number}}\right)^2\right] \\
&\qquad - 2 \cdot \underbrace{\mathbb{E}\left[(\widehat{\mu}_{ML} - \mathbb{E}\left[\widehat{\mu}_{ML}\right])\left(\mathbb{E}\left[\widehat{\mu}_{ML}\right] - \mu\right)\right]}_{=0} \\
&= \mathbb{E}\left[(\widehat{\mu}_{ML} - \mathbb{E}\left[\widehat{\mu}_{ML}\right])^2\right] + (\mathbb{E}\left[\widehat{\mu}_{ML}\right] - \mu)^2 \\
&= \mathbb{V}[\widehat{\mu}_{ML}] + \left(\mathbb{E}\left[X_{N_L}\right] - \mathbb{E}[X]\right)^2 \\
&= \mathbb{V}[\widehat{\mu}_{ML}] + \left(\mathbb{E}\left[X_{N_L} - X\right]\right)^2 \\
&= \sum_{\ell=0}^{L} M_\ell^{-1}\mathbb{V}[Y_\ell] + (\mathbb{E}[X_{N_L} - X])^2
\end{aligned}$$

$\square$

Controlling the MSE : Bias

Goal to achieve $\mathrm{MSE}(\widehat{\mu}_{ML}) \le \epsilon^2$, and this is satisfied if

$$\left(\mathbb{E}[X_{N_L} - X]\right)^2 \le \frac{\epsilon^2}{2} \quad \& \quad \left(\sum_{\ell=0}^{L} M_\ell^{-1}\mathbb{V}[Y_\ell]\right) \le \frac{\epsilon^2}{2}$$

Assumption 3.14. *We assume that there exist C and $\alpha > 0$ such that*

$$\left|\mathbb{E}\left[X_{N_L} - X\right]\right| \le CN_L^{-\alpha},$$

where C is independent of N_L.

Using this assumption, we obtain

$$\left(\mathbb{E}\left[X_{N_L} - X\right]\right)^2 \le \frac{\epsilon^2}{2},$$

if we choose

$$N_L > (2C^2)^{1/2\alpha}\epsilon^{-1/\alpha}$$

Assuming that N_0 and s are given, and

$$N_\ell = sN_{\ell-1} \quad \text{for } \ell = 1, 2, \ldots, L.$$

How many levels L are needed to control the bias?

3.8 Variance of Correction Terms

Recall: The mean square error of MLMC estimator $\widehat{\mu}_{ML}$ satisfies

$$\text{MSE}(\widehat{\mu}_{ML}) = \sum_{\ell=0}^{L} M_\ell^{-1} \mathbb{V}[Y_\ell] + \underbrace{(\mathbb{E}[X_{N_L} - X])^2}_{\text{bias term}}$$

The bias term can be satisfied if we choose L properly by section 1.3.1. How we could deal with

$$\left(\sum_{\ell=0}^{L} M_\ell^{-1} \mathbb{V}[Y_\ell] \right) \leq \frac{\epsilon^2}{2}.$$

If we have a *sequence* of approximations $X_{N_0}, X_{N_1}, \ldots, X_{N_L}$ to X and $Y_0 = X_{N_0}$, and $Y_\ell := X_{N_\ell} - X_{N_{\ell-1}}$ for $\ell = 1, 2, \ldots, L$.

Assumption 3.15. *The sequence* $X_{N_0}, X_{N_1}, \ldots, X_{N_L}$ *converges to X in the* mean-square *sense, i.e.*

$$\mathbb{E}\left[(X_{N_\ell} - X)^2 \right] = \left\| X_{N_\ell} - X \right\|_{L^2(\Omega)}^2 \to 0 \quad as \quad \ell \to \infty.$$

Lemma 3.16. *If the assumption 3.15 is satisfied, then*

$$\mathbb{V}[Y_\ell] \to 0 \quad as \quad \ell \to \infty$$

Proof. For $\ell = 1, 2, \ldots$, (not include 0).

$$\begin{aligned}
\mathbb{V}[Y_\ell] &= \mathbb{E}[Y_\ell^2] - (\mathbb{E}[Y_\ell])^2 \\
&\leq \mathbb{E}[Y_\ell^2] \\
&= \|Y_\ell\|_{L^2(\Omega)}^2 \\
&= \left\| X_{N_\ell} - X_{N_{\ell-1}} \right\|_{L^2(\Omega)}^2 \\
&= \left\| X_{N_\ell} - X + X - X_{N_{\ell-1}} \right\|_{L^2(\Omega)}^2 \\
&\leq \left(\underbrace{\left\| X_{N_\ell} - X \right\|_{L^2(\Omega)}}_{\to 0 \text{ as } \ell \to \infty} + \underbrace{\left\| X - X_{N_{\ell-1}} \right\|_{L^2(\Omega)}}_{\to 0 \text{ as } (\ell-1) \to \infty} \right)^2 \\
\Rightarrow \mathbb{V}[Y_\ell] &\to 0 \quad \text{as} \quad \ell \to \infty
\end{aligned}$$

$\square$

Mean Square Error

We need to choose $M_0, M_1, \ldots, M_L$ such that

$$M_0^{-1} \mathbb{V}[X_{N_0}] + \sum_{\ell=1}^{L} M_\ell^{-1} \mathbb{V}[Y_\ell] \leq \frac{\epsilon^2}{2}.$$

Simple way: There are $L + 1$ terms in total. The condition is satisfied if

$$M_0^{-1} \mathbb{V}[X_{N_0}] \leq \frac{\epsilon^2}{2(L+1)} \quad \& \quad M_\ell^{-1} \mathbb{V}[Y_\ell] \leq \frac{\epsilon^2}{2(L+1)}.$$

That is, if we choose the *number of samples* such that

$$M_0 \geq 2(L+1)\mathbb{V}[X_{N_0}]\epsilon^{-2} \quad \& \quad M_\ell \geq 2(L+1)\mathbb{V}[Y_\ell]\epsilon^{-2}$$

Key Conclusion: If $\mathbb{V}[Y_\ell] \to 0$ as $\ell \to \infty$, the number of samples M_ℓ we need to choose to control the MSE also decrease as $\ell \to \infty$.

3.9 MLMC : Accuracy and Cost

Recall. MLMC estimator for $\mu = \mathbb{E}[X]$ is

$$\widehat{\mu}_{ML} := \sum_{\ell=0}^{L} \widehat{\mu}_\ell \quad \text{where } \widehat{\mu}_\ell = \frac{1}{M_\ell}\sum_{i=1}^{M_\ell} Y_\ell^i$$

where we define the following:

$$Y_0 := X_{N_0}, \quad Y_\ell := X_{N_\ell} - X_{N_{\ell-1}}, \quad \ell = 1, 2, \ldots, L.$$

Question. What is the *cost* of computing $\widehat{\mu}_{ML}$ in such a way that **RMSE**$(\widehat{\mu}_{ML}) \leq \epsilon$

MLMC Cost

Let C_{N_ℓ} denote the cost of generate *one* sample of Y_ℓ, $\ell = 0, 1, \ldots, L$. Hence the total cost of computing the samples needed for $\widehat{\mu}_{ML}$ is

$$C_{MLMC} = \sum_{\ell=0}^{L} M_\ell \times C_{N_\ell}. \tag{3.33}$$

Example 3.17 (Population Model). The cost of computing one sample of X_{N_ℓ} using the explicit Euler method with N_ℓ intervals can be *characteristised by the number of intervals.*

$$C_{N_0} = N_0, \quad C_{N_\ell} = N_\ell + N_{\ell-1} = (1 + s^{-1})N_\ell = (1 + s^{-1})s^\ell N_0. \tag{3.34}$$

Hence,

$$C_{MLMC} = M_0 N_0 + \sum_{\ell=1}^{L} M_\ell (1 + s^{-1})s^\ell N_0. \tag{3.35}$$

The cost dependents on s, N_0 and L and M_ℓ.

Controlling the MSE

Recall (Theorerm 3.13). The mean-square error of $\widehat{\mu}_{ML}$ satisfies

$$\text{MSE}(\widehat{\mu}_{ML}) = \sum_{\ell=0}^{L} M_\ell^{-1}\mathbb{V}[Y_\ell] + (\mathbb{E}[X_{N_\ell} - X])^2.$$

To satisfy $\text{MSE}(\widehat{\mu}_{ML}) \leq \epsilon^2$, we could choose L (no. of levels) and M_ℓ (no. of samples at each levels) such that

$$(\mathbb{E}[X_{N_\ell} - X])^2 \leq \frac{\epsilon^2}{2} \quad \& \quad \sum_{\ell=0}^{L} M_\ell^{-1}\mathbb{V}[Y_\ell] \leq \frac{\epsilon^2}{2}. \tag{3.36}$$

<div align="center">***Number of Samples.***</div>

If we choose the number of samples so that

$$M_0 \geq 2(L+1)\mathbb{V}[X_{N_0}]\epsilon^{-2} \quad \& \quad M_\ell \geq 2(L+1)\mathbb{V}[Y_\ell]\epsilon^{-2}. \tag{3.37}$$

to get integers:

$$M_0 = \lceil 2(L+1)\mathbb{V}[X_{N_0}]\epsilon^{-2}\rceil \quad \& \quad \lceil M_\ell = 2(L+1)\mathbb{V}[Y_\ell]\epsilon^{-2}\rceil.$$

Example 3.18 (Population Model)**.** We have proved that $\mathbb{V}[Y_\ell] \leq C_s T^2 N_\ell^{-2}$, hence we could choose $M_\ell = \lceil 2(L+1)C_s T^2 N_\ell^{-2}\epsilon^{-2}\rceil$

<div align="center">***Number of Levels.***</div>

Assumption 3.19. *There exist C and $\alpha > 0$ such that*

$$|\mathbb{E}[X_{N_\ell} - X]| \leq CN_L^{-\alpha},$$

where C is independent of L.

Then we obtain $(\mathbb{E}[X_{N_\ell} - X])^2 \leq \epsilon^2/2$ if we choose $N_L \geq (2C^2)^{1/2\alpha}\epsilon^{-1/\alpha}$.

Example 3.20 (Population Model ($\alpha = 1$))**.** Since $N_L = s^L N_0$, we show that $L = O(|\log(\epsilon)|)$. Then the *total cost of the Population model* will be

$$C_{MLMC} := \sum_{\ell=0}^{L} M_\ell \times C_{N_\ell} = M_0 N_0 + \sum_{\ell=1}^{L} M_\ell(1 + s^{-1})N_\ell.$$

We obtain $\text{MSE}(\widehat{\mu}_{ML}) \leq \epsilon^2$ if we choose:

- $L = O(|\log(\epsilon)|)$.

- $M_0 = \lceil 2(L+1)\mathbb{V}[X_{N_0}]\epsilon^{-2}\rceil$.

- $M_\ell = \lceil 2(L+1)C_s T^2 N_\ell^{-2}\epsilon^{-2}\rceil$.

Since M_ℓ decays to zero more quickly than C_{N_ℓ} increase as $\ell \to \infty$, the dominant term is

$$M_0 N_0 = \lceil 2(L+1)\mathbb{V}[X_{N_0}]\epsilon^{-2}\rceil N_0 = O(\epsilon^{-2}|\log(\epsilon)|) \tag{3.38}$$

4   Stochastic Galerkin Approximation

- Sampling method like Monte Carlo are called *non-intrusive* as we can implement them by running existing code for deterministic model.

<div align="center">Generating Samples $\to$ Code $\to$ Solution Samples $\to$ Monte Carlo</div>

- In section 4, different kind of approach which is called *intrusive* method. (Don't need to sample)

Key Idea.

- If the model's solution u can be expressed as a function of J (i.e. finitely many) real-valued random variables

$$\xi_1, \xi_2, \ldots, \xi_J : \Omega \to \mathbb{R}.$$

We can construct approximations that are functions of those variables.

Example 4.1 (Simple ODE model with $J = 1$)**.** Deterministic IVP, find $u : [0,T] \to \mathbb{R}$ such that

$$\frac{\mathrm{d}u(t)}{\mathrm{d}t} = -\alpha u(t), \quad u(0) = u_0,$$

and we have the exact solution: $u(t) = u_0 e^{-\alpha t}$.

- If $\alpha > 0$, then $u(t) \to 0$ as $t \to \infty$ (decaying process).

- If $\alpha < 0$, then $u(t) \to \infty$ as $t \to \infty$ (growth process).

Question. How varying in α will varying in solution u?

Example 4.2 (IVP with Random α)**.** Assume u_0 is given but α is *uncertain*. Let $\alpha : \Omega \to \mathbb{R}$ be a random variable associated with a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, then we have a stochastic IVP: Find $u : [0,T] \times \Omega \to \mathbb{R}$ such that $\mathbb{P}$-a.s.,

$$\frac{\mathrm{d}u(t,\omega)}{\mathrm{d}t} = -\alpha(\omega)u(t,\omega), \quad u(0,\omega) = u_0. \tag{4.1}$$

The exact solution: $u(t,\omega) = u_0 e^{-\alpha(\omega)t}$, for $t \in [0,T]$, $\omega \in \Omega$.

Specific Choice of α

Suppose we make a specific choice

$$\alpha = \alpha_0 + \alpha_1 \xi, \quad \xi \sim U(-\sqrt{3}, \sqrt{3}). \tag{4.2}$$

Then α is uniform random variable with

$$\mathbb{E}[\alpha] = \alpha_0, \quad \mathbb{V}[\alpha] = \alpha_1^2.$$

The exact solution can be expressed as:

$$u(t,\xi) = u_0 e^{-(\alpha_0 + \alpha_1 \xi)t}, \quad t \in [0,T], \ \xi \in [-\sqrt{3}, \sqrt{3}].$$

- Since α (the input) is a function of ξ, then so does the solution u.

Reformulated IVP

If we define the 'Parameter domain' $\Gamma = [-\sqrt{3}, \sqrt{3}]$, then we can rewrite the IVP as follows:
Find $u : [0,T] \times \Gamma \to \mathbb{R}$ such that for each $\xi \in \Gamma$,

$$\frac{\mathrm{d}u(t,\xi)}{\mathrm{d}t} = -(\alpha_0 + \alpha_1 \xi)u(t,\xi), \quad u(0,\xi) = u_0. \tag{4.3}$$

Key Point. The solution to the reformulated IVP is a function of ξ. It can be proved that:

$$\mathbb{E}[u(t,\xi)] = \frac{u_0 e^{-\alpha_0 t}}{t\alpha_1 \sqrt{3}} \sinh(\alpha_1 \sqrt{3}t).$$

$$\mathbb{V}[u(t,\xi)] = u_0^2 e^{-2\alpha_0 t} \left(\frac{1}{2t\alpha_1 \sqrt{3}} \sinh(2\alpha\sqrt{3}t) - \frac{1}{3t^2\alpha_1^2} \sinh^2(2\alpha_1\sqrt{3}t) \right). \tag{4.4}$$

Investigation. How do these quantities behaved as $t \to \infty$? How does the behaviour in time depends on α_0 and α_1?

4.1 Polynomial Based Approximation

- Suppose we have a model (PDE/ODE) whose inputs and solution u can be expressed as functions of J real-valued random variables,

$$\xi_1, \xi_2, \ldots, \xi_J : \Omega \to \mathbb{R}.$$

IDEA. Construct approximations to the model's solution of the form

$$u(\boldsymbol{x}, t, \boldsymbol{\xi}) \approx \sum_{i=1}^{N} \underbrace{u_i(\boldsymbol{x}, t)}_{\text{Coefficients}} \underbrace{\psi_i(\boldsymbol{\xi})}_{\substack{\text{functions in } J \\ \text{random variables}}}, \tag{4.5}$$

where $\psi_i(\boldsymbol{\xi})$ are *polynomials* in $\boldsymbol{\xi} = [\xi_1, \xi_2, \ldots, \xi_J]^T$.

Example 4.3 (Recap: IVP Example 4.2 ($J = 1$)). Let $\alpha = \alpha_0 + \alpha_1 \xi$ where $\xi \sim U(-\sqrt{3}, \sqrt{3})$. Define $\Gamma = [-\sqrt{3}, \sqrt{3}]$ and consider the following IVP.

Find $u : [0, T] \times \Gamma \to \mathbb{R}$ such that for each $\xi \in \Gamma$

$$\frac{\mathrm{d}u(t, \xi)}{\mathrm{d}t} = -\alpha(\xi)u(t, \xi), \quad u(0, \xi) = u_0.$$

- The solution u is a function of t & ξ.

Example 4.4 (Population Model ($J = 2$)). Let the component of initial condition be ξ_1 and ξ_2. Define $\boldsymbol{\xi} = [\xi_1, \xi_2]^T$, $\Gamma = \Gamma_1 \times \Gamma_2$, where $\xi_1 \in \Gamma_1$, $\xi_2 \in \Gamma_2$. Hence the population model becomes

Find $\boldsymbol{u} : [0, T] \times \Gamma \to \mathbb{R}^2$ such that for each $\boldsymbol{\xi} \in \Gamma$,

$$\frac{\mathrm{d}}{\mathrm{d}t} \begin{pmatrix} u_1(t, \boldsymbol{\xi}) \\ u_2(t, \boldsymbol{\xi}) \end{pmatrix} = \begin{pmatrix} u_1(t, \boldsymbol{\xi})(1 - u_2(t, \boldsymbol{\xi})) \\ u_2(t, \boldsymbol{\xi})(u_1(t, \boldsymbol{\xi}) - 1) \end{pmatrix}, \quad 0 < t < T,$$

with initial condition $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ (4.6)

$$\begin{pmatrix} u_1(0, \boldsymbol{\xi}) \\ u_2(0, \boldsymbol{\xi}) \end{pmatrix} = \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}$$

- the (vector-valued) ODE solution is a functino of ξ_1 and ξ_2 and t,

$$\boldsymbol{u}(t, \boldsymbol{\xi}) = \begin{pmatrix} u_1(t, \xi_1, \xi_2) \\ u_2(t, \xi_1, \xi_2) \end{pmatrix}$$

Observed from the graph 8

- By getting samples of ξ_1, ξ_2, we get a point on surface but not the whole surface.

- Approximate the surface somehow using a polynomial of ξ_1 and ξ_2.

Polynomial Chaos Approximation

- Consider a model whose solution u depends on t (time) and a set of J real-valued random variables $\boldsymbol{\xi} \in \Gamma$, that is, $u = u(t, \boldsymbol{\xi})$.
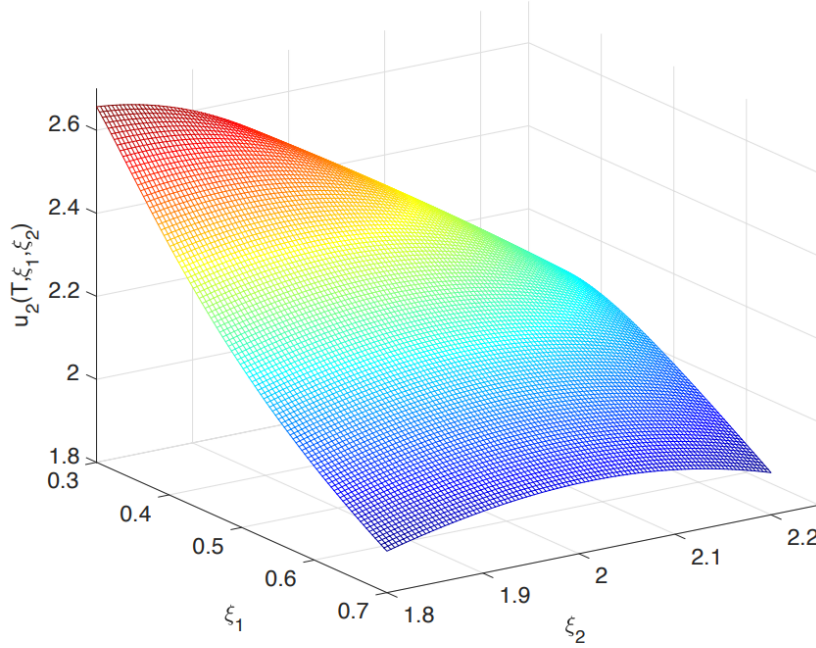
FIGURE 8: An approximation of the response surface $u_2(T, \xi_1, \xi_2)$ for the predator-prey problem from Example 4.4 with $T = 6$, where $\xi_1 \in \Gamma_1 = [u_1 - \tau, u_1 + \tau]$ and $\xi_2 \in \Gamma_2 = [u_2 - \tau, u_2 + \tau]$ with $u_1 = 0.5$, $u_2 = 2$ and $\tau = 0.2$.

- We will look for approximations of the form

$$u(t, \boldsymbol{\xi}) \approx \sum_{i=1}^{N} u_i(t)\psi_1(\boldsymbol{\xi}), \qquad (\star) \tag{4.7}$$

  where $\{\psi_1(\boldsymbol{\xi}), \psi_2(\boldsymbol{\xi}), \ldots, \psi_N(\boldsymbol{\xi})\}$ are *polynomials* in $\xi_1, \xi_2, \ldots, \xi_J$.

- When we choose the polynomials to be *orthogonal* with respect to the *probability measure* associated with the input random variables i.e.

$$\mathbb{E}[\psi_i(\boldsymbol{\xi})\psi_j(\boldsymbol{\xi})] = 0, \quad \text{if } i \neq j, \tag{4.8}$$

  then we call $(\star)$ a *polynomial chaos* approximation.

Change of Variables

Recall. If $\boldsymbol{\xi} : \Omega \to \Gamma \subset \mathbb{R}^J$ and if we know the associated *joint pdf* $\rho(\boldsymbol{y})$, then

$$\mathbb{E}[f(\boldsymbol{\xi})] = \int_{\Omega} f(\boldsymbol{\xi}(\omega)) \mathrm{d}\mathbb{P}(\omega) = \int_{\Gamma} f(\boldsymbol{y})\rho(\boldsymbol{y}) \mathrm{d}\boldsymbol{y}. \tag{4.9}$$

That is, we do a change of variables in the integral from $\omega \in \Omega$ to $\boldsymbol{y} \in \Gamma$.

So equivalently, we need polynomials that satisfy

$$\mathbb{E}[\psi_i(\boldsymbol{\xi})\psi_j(\boldsymbol{\xi})] = \int_{\Gamma} \psi_i(\boldsymbol{y})\psi_j(\boldsymbol{y})\rho(\boldsymbol{y}) \mathrm{d}\boldsymbol{y} = 0, \quad \text{when } i \neq j. \tag{4.10}$$

Surrogate Model

$$u(t, \boldsymbol{\xi}) \approx \widehat{u}(t, \boldsymbol{\xi}) := \sum_{i=1}^{N} u_i(t) \psi_i(\boldsymbol{\xi}). \tag{4.11}$$

- Approximation like $\widehat{u}(t, \boldsymbol{\xi})$ are often called *surrogates*. They can be *cheaply evaluated* for *any* choice of the random input $\boldsymbol{\xi}$ without solving the underlying ODE/PDE model again.

- Once the polynomials $\{\psi_1(\boldsymbol{\xi}), \psi_2(\boldsymbol{\xi}), \ldots, \psi_N(\boldsymbol{\xi})\}$ have been chosen, there are several strategies for finding appropriate coefficients $u_i(t)$.

- We will look at an 'intrusive' approach known as *Stochastic Galerkin Approximation*.

Exercise 4.5. We need polynomials that satisfy

$$\mathbb{E}[\psi_i(\boldsymbol{\xi})\psi_j(\boldsymbol{\xi})] = 0, \quad \text{when } i \neq j$$

Consider the following questions:

- If the random variables $\xi_1, \xi_2, \ldots, \xi_J$ are *independent*, what can we say about the joint pdf $\rho(\boldsymbol{y})$ of $\boldsymbol{\xi}$?

- What families of orthogonal polynomials (in one variable) do you already know? In what sense are they 'orthogonal'?

4.2 Univariate Orthogonal Polynomials

Let $\xi : \Omega \to \Gamma \subset \mathbb{R}$ be a real-valued r.v. with pdf $\rho(y)$. We first need to define a weighted inner product.

Weighted Inner Product

To define such an inner product, we need

- An interval $\Gamma \in \mathbb{R}$

- A weight function $w : \Gamma \to [0, \infty)$

Definition 4.6. A weight function $w : \Gamma \to [0, \infty)$ is continuous and non-negative on Γ and

$$\int_{\Gamma} w(y) \mathrm{d}y > 0. \tag{4.12}$$

Based on this, we define the *weighted L^2* inner product

$$\langle u, v \rangle_w := \int_{\Gamma} w(y) u(y) v(y) \mathrm{d}y. \tag{4.13}$$

Weighted L2 Space

Definition 4.7 ($L_w^2(\Gamma)$)**.** Let $\Gamma \subset \mathbb{R}$ and let $w(y)$ be a weighted function. The weighted L^2 space is

$$L_w^2(\Gamma) := \{v : \Gamma \to \mathbb{R}, \|v\|_{L_w^2(\Gamma)} < \infty\} \tag{4.14}$$

where the norm is defined by $\|v\|_{L_w^2(\Gamma)}^2 := (v, v)_w$

Orthogonal Polynomial

Let ϕ_0, ϕ_1, ϕ_2,. . . be a set of polynomials on Γ (where ϕ_i has degree i). We say that ϕ_i and ϕ_j are *orthogonal* on Γ with respect to $(\cdot, \cdot)_w$ if

$$(\phi_i, \phi_j)_w = 0, \text{ when } i \neq j \tag{4.15}$$

In addition, we say that ϕ_i is normalised if

$$\|\phi_i\|_{L_w^2(\Gamma)} = 1 \tag{4.16}$$

4.3   Three-term Recurrence

Standard Three-term Recurrence

Theorem 4.8. *A family of orthogonal polynomials that satisfies*

$$\langle \psi_j, \psi_i \rangle_w = 0, \quad when\ i \neq j, \tag{4.17}$$

always satisfies a three-term recurrence *of the form:*

$$\psi_j(y) = (a_j y + b_j)\psi_{j-1}(y) - c_j \psi_{j-2}(y), \quad j = 1, 2, \ldots, \tag{4.18}$$

for some coefficients a_j, b_j and c_j.

Notice that, to initialize we specify $\psi_{-1} = 0$ and $\psi_0(y)$, therefore ψ_0 has *degree zero*. If we want *orthonormal* polynomials then

$$\|\psi_0\|_{L_w^2(\Gamma)} = 1 \quad \Leftrightarrow \quad \psi_0^2 \int_{\Gamma} w(y)\mathrm{d}y = 1. \tag{4.19}$$

If the weight function is a *probability density function* then

$$\int_{\Gamma} w(y)\mathrm{d}y = 1 \tag{4.20}$$

and we choose $\psi_0 = 1$. Therefore we have

$$\begin{aligned} \psi_{-1} &= 0, \quad \psi_0 = 1, \\ \psi_j(y) &= (a_j y + b_j)\psi_{j-1}(y) - c_j \psi_{j-2}(y) \quad j = 1, 2, \ldots. \end{aligned} \tag{4.21}$$

Simplified Three-term Recurrence

The standard recurrence formula (4.21) can be simplifies if

1. If the polynomials are orthonormal,

2. $\Gamma = [-\gamma, \gamma]$: the domain is symmetric about zero,

3. $w(y)$ is *even* on Γ.

Then the recurrence simplifies to

$$\frac{1}{a_j}\psi_j(y) = y\psi_{j-1}(y) - \frac{1}{a_{j-1}}\psi_{j-2}(y), \quad j = 1, 2, \ldots, \tag{4.22}$$

where $a_j \neq 0$.

Example 4.9. Suppose $\xi \sim N(0, 1)$. Then $\Gamma = \mathbb{R}$ and

$$w(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2}. \tag{4.23}$$

Example 4.10. Suppose $\xi \sim U(-\sqrt{3}, \sqrt{3})$, then $\Gamma = (-\sqrt{3}, \sqrt{3})$ and

$$w(y) = \frac{1}{2\sqrt{3}} \tag{4.24}$$

Rewrite the formula as the following

$$y\psi_{j-1}(y) = \frac{1}{a_j}\psi_j(y) + \frac{1}{a_{j-1}}\psi_{j-2}(y), \quad j = 1, 2, \ldots. \tag{4.25}$$

Collecting equations for $j = 1, 2, \ldots, k + 1$, gives a *linear system*:

$$y\begin{pmatrix} \psi_0(y) \\ \psi_1(y) \\ \vdots \\ \psi_k(y) \end{pmatrix} = T_{k+1}\begin{pmatrix} \psi_0(y) \\ \psi_1(y) \\ \vdots \\ \psi_k(y) \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ \psi_{k+1}(y)/a_{k+1} \end{pmatrix} \tag{4.26}$$

Theorem 4.11. *The matrix* $T_{k+1} \in \mathbb{R}^{(k+1)\times(k+1)}$ *in* (4.26) *is symmetric and* tri-diagonal

$$T_{k+1} = \begin{pmatrix} 0 & \frac{1}{a_1} & & & \\ \frac{1}{a_1} & 0 & \frac{1}{a_2} & & \\ & \ddots & \ddots & \ddots & \\ & & \frac{1}{a_{k-1}} & 0 & \frac{1}{a_k} \\ & & & \frac{1}{a_k} & 0 \end{pmatrix}. \tag{4.27}$$

The eigenvalues *of* T_{k+1} *are the zeros of the polynomial* $\psi_{k+1}(y)$.

4.4   Hermite and Legendre Polynomials

Hermite Polynomial

Let $H_{-1} = 0$ and $H_0(y) = 1$ and then define $H_j(y)$, $j = 1, 2, \ldots$, via

$$\sqrt{j}H_j(y) = yH_{j-1}(y) - \sqrt{j-1}H_{j-2}(y), \quad j = 1, 2, \ldots, \tag{4.28}$$

(namely $1/a_j = \sqrt{j}$.)

These are Hermite polynomials and they satisfy

$$\langle H_i, H_j \rangle_\rho = \int_{-\infty}^{\infty} \left(\frac{1}{\sqrt{2\pi}} e^{-y^2/2} \right) H_i(y)H_j(y)\mathrm{d}y = \delta_{i,j}. \tag{4.29}$$

Hence these are *orthonormal* on $\Gamma = \mathbb{R}$ with respect to the weight function (probability density function) associated with the distribution $N(0, 1)$.

Legendre Polynomials

Let $L_{-1} = 0$ and $L_0(y) = 1$ and then define $L_j(y)$, $j = 1, 2, \ldots$ via

$$\frac{1}{a_j} L_j(y) = y L_{j-1}(y) - \frac{1}{a_{j-1}} L_{j-2}(y), \quad j = 1, 2, \ldots \tag{4.30}$$

where

$$\frac{1}{a_j} = \frac{j\sqrt{3}}{\sqrt{2j-1}\sqrt{2j+1}}. \tag{4.31}$$

These are *Legendre* polynomials and they satisfy

$$\langle L_i, L_j \rangle_\rho = \int_{-\sqrt{3}}^{\sqrt{3}} \left(\frac{1}{2\sqrt{3}} \right) L_i(y) L_j(y) \mathrm{d}y = \delta_{i,j}. \tag{4.32}$$

They are orthonormal on $\Gamma = [-\sqrt{3}, \sqrt{3}]$ with respect to the weight function associated with the distribution $U(-\sqrt{3}, \sqrt{3})$.

4.5 Approximation using Orthogonal Polynomials

Recall. Let $\Gamma \in \mathbb{R}$ and $w(y)$ be a given weight function, and let

$$\{\psi_0(y), \psi_1(y), \psi_2(y), \ldots\} \tag{4.33}$$

be a family of polynomials that are orthonormal with respect to $\langle \cdot, \cdot \rangle_w$.

A function $f(y) \in L_w^2(\Gamma)$ where f is continuous, then

$$\|f\|_{L_w^2(\Gamma)} := \left(\int_\Gamma w(y) f(y)^2 \mathrm{d}y \right)^{1/2} < \infty. \tag{4.34}$$

To define an *approximation* to f

1. Choose $k \in \mathbb{N}_0$ and define the finite-dimensional subspace

$$S_k := \mathrm{span}\{\psi_0(y), \psi_1(y), \ldots, \psi_k(y)\} \subset L_w^2(\Gamma). \tag{4.35}$$

2. S_k is the set of polynomials in y on Γ of degree less or equal to k.

3. We look for an approximation $\widehat{f_k} \in S_k$.

Best Weighted L^2 approximation

Question. What is the *best* approximation $\widehat{f_k} \in S_k$ to f?

The best approximation $\widehat{f_k} \in S_k$ satisfies

$$\left\| f - \widehat{f_k} \right\|_{L_w^2(\Gamma)} \le \| f - \widehat{v_k} \|_{L_w^2(\Gamma)}, \quad \forall \widehat{v_k} \in S_k. \tag{4.36}$$

Since $\widehat{f_k} \in S_k$, we can write it as

$$\widehat{f_k}(y) = \sum_{i=0}^{k} \alpha_i \psi_i(y), \tag{4.37}$$

and we need to find the coefficients $\alpha_0, \ldots, \alpha_k$ that minimise

$$\left\| f - \widehat{f_k} \right\|_{L_w^2(\Gamma)}. \tag{4.38}$$

It can be easy show that the coefficient in the best approximation must be

$$\alpha_i = \langle f, \psi_i \rangle_w, \quad i = 0, 1, \ldots, k. \tag{4.39}$$

Proof. Proof of this is trivial by using multivariable calculus: by considering the minimised function as

$$\mathcal{S} = \left\|f - \widehat{f_k}\right\|^2_{L^2_w(\Gamma)} = \int_\Gamma w(y) \left(f(y) - \sum_{i=0}^{k} \alpha \psi_i(y)\right)^2 \mathrm{d}y \tag{4.40}$$

and this quantity is minimise if and only if

$$\frac{\partial \mathcal{S}}{\partial \alpha_j} = 0, \quad j = 0, 1, \ldots, k. \tag{4.41}$$

$\square$

4.6 Orthogonal Projection

Recall. Suppose $f(y)$ is continuous and belongs to $L^2_w(\Gamma)$ and define

$$S_k := \mathrm{span}\{\psi_0(y), \cdots, \psi_k(y)\} \subset L^2_w(\Gamma). \tag{4.42}$$

The best weighted L^2 approximation to f from S_k is

$$\widehat{f_k}(y) = \sum_{i=0}^{k} \langle f, \psi_i \rangle_w \psi_i(y). \tag{4.43}$$

This is known as the *orthogonal projection* of f onto S_k.

We will further discuss the error $\left\|f - \widehat{f_k}\right\|_{L^2_w(\Gamma)}$.

The orthogonal projection (best approximation) also satisfies

$$\langle f, \phi \rangle_w = \langle \widehat{f_k}, \phi \rangle_w, \quad \forall \phi \in S_k, \tag{4.44}$$

or equivalently,

$$\langle f - \widehat{f_k}, \phi \rangle_w = 0, \quad \forall \phi \in S_k. \tag{4.45}$$

That is the error is orthogonal to functions in S_k.

Best Approximation Error

In the case of Hermite and Legendre polynomials,

$$\{\psi_0(y), \psi_1(y), \ldots\} \tag{4.46}$$

provides an orthonormal basis for $L^2_w(\Gamma)$. Then every function in $L^2_w(\Gamma)$ can be written as a linear combination of the basis functions,

$$f(y) = \sum_{i=0}^{\infty} \alpha_i \psi_i(y) \rightarrow \langle f, \psi_j \rangle_w = \alpha_j. \tag{4.47}$$

Hence the error is

$$f(y) - \widehat{f_k}(y) = \sum_{i=k+1}^{\infty} \langle f, \psi_i \rangle_w \psi_i(y). \tag{4.48}$$

The important things is that, if the functions $\{\psi_0(y), \psi_1(y), \ldots\}$ provide a basis for the space $L^2_w(\Gamma)$, then we should able to represent f as an infinite series of linear combinations of the

basis functions. If we use the fact that these functions are orthonormal, then we can figure out what these coefficients are. Then the $\widehat{f}_k$ is just the truncated version of that representation.

Using the definition of the weighted L^2 norm and the orthonormality of the basis polynomials gives

$$
\begin{aligned}
\left\| f - \widehat{f}_k \right\|^2_{L^2_w(\Gamma)} &= \int_\Gamma w(y) \left(\sum_{i=k+1}^\infty \langle f, \psi_i \rangle_w \psi_i(y) \right)^2 dy \\
&= \sum_{i=k+1}^\infty |\langle f, \psi_i \rangle_w|^2 \, \|\psi_i(y)\|^2_{L^2_w(\Gamma)} \\
&= \sum_{i=k+1}^\infty |\langle f, \psi_i \rangle_w|^2.
\end{aligned}
\tag{4.49}
$$

Question. Does $\left\| f - \widehat{f}_k \right\|_{L^2_w(\Gamma)} \to 0$ as the polynomial degree $k \to \infty$?

The coefficients $\alpha_i = \langle f, \psi_i \rangle_w$ decay to zero as $i \to \infty$ at a rate that depends on the smoothness of f. The smoother f is, the faster the coefficients decay to zero and the faster

$$
\left\| f - \widehat{f}_k \right\|_{L^2_w(\Gamma)} \to 0, \quad \text{as } k \to \infty.
\tag{4.50}
$$

4.7   Stochastic Galerkin Approximation

Let $\alpha = \alpha_0 + \alpha_\xi$ where $\xi \sim U(-\sqrt{3}, \sqrt{3})$ and define $\Gamma := [-\sqrt{3}, \sqrt{3}]$. Find $u : [0, T] \times \Gamma \to \mathbb{R}$ such that for every $\xi \in \Gamma$,

$$
\frac{du(t, \xi)}{dt} = -\alpha(\xi)u(t, \xi), \quad u(0, \xi) = u_0.
\tag{4.51}
$$

Weak Formulation

To define the weak formulation, we preform the following steps:

1. Multiply both sides of the ODE by a test function $v(\xi)$,

$$
\frac{du(t, \xi)}{dt} v(\xi) = -\alpha(\xi)u(t, \xi)v(\xi).
\tag{4.52}
$$

2. Take the expectation of both sides

$$
\mathbb{E}\left[\frac{du(t, \xi)}{dt} v(\xi) \right] = -\mathbb{E}[\alpha(\xi)u(t, \xi)v(\xi)].
\tag{WF}
$$

The integral (WF) are well defined if

$$
\frac{du(t, \xi)}{dt}, \quad u(t, \xi), \quad v(\xi)
\tag{4.53}
$$

have finite second moments. That is, if they belongs to

$$
V := \left\{ v : \Gamma \to \mathbb{R}, \quad \mathbb{E}[v(\xi)^2] < \infty \right\}
\tag{4.54}
$$

Note that,

$$
\mathbb{E}[v(\xi)^2] = \int_\Gamma \rho(y)v(y)^2 dy,
\tag{4.55}
$$

so V is equivalent to the weighted space $L^2_\rho(\Gamma)$ defined previously.

Stochastic Galerkin(SG) Approximation

In the SG method, we choose a finite-dimensional subspace $S_k \subset V$.

We look for an approximation $\widehat{u}_k(t, \cdot) \in S_k$ that satisfies

$$\mathbb{E}\left[\frac{\mathrm{d}\widehat{u}_k(t,\xi)}{\mathrm{d}t}v(\xi)\right] = -\mathbb{E}[\alpha(\xi)\widehat{u}_k(t,\xi)v(\xi)], \quad \forall v \in S_k. \tag{WF1}$$

and the initial condition

$$\widehat{u}_k(0,\xi) = u_0, \quad \forall \xi \in \Gamma. \tag{WF2}$$

NOTE.

1. (WF1) with (WF2) is called the (finite dimensional) weak formulation,

2. The choice of space S_k is very important.

3. Both the approximation and the test functions belongs to S_k.

Rearranging (WF1) gives

$$\mathbb{E}\left[(\frac{\mathrm{d}\widehat{u}_k}{\mathrm{d}t} + \alpha\widehat{u}_k)v\right] = 0, \quad \forall v \in S_k \tag{4.56}$$

or equivalently,

$$\langle\frac{\mathrm{d}\widehat{u}_k}{\mathrm{d}t} + \alpha\widehat{u}_k, v\rangle_\rho = 0, \forall v \in S_k. \tag{4.57}$$

where $\frac{\mathrm{d}\widehat{u}_k}{\mathrm{d}t} + \alpha\widehat{u}_k$ is the *residue error* in the ODE.

How do we solve the (WF1) + (WF2)?

- Substitute $\widehat{u}_k(t,\xi) = \sum_{i=0}^k u_i(t)\psi_i(\xi)$.

- Choose test functions $v(\xi) = \psi_j(\xi)$, for $j = 0, 1, \ldots, k$.

This gives $k + 1$ equations,

$$\sum_{i=0}^k \frac{\mathrm{d}u_i(t)}{\mathrm{d}t} \mathbb{E}[\psi_i(\xi)\psi_j(\xi)] = -\sum_{i=0}^k u_i(t) \mathbb{E}[\alpha(\xi)\psi_i(\xi)\psi_j(\xi)], j = 0, 1, \ldots, k. \tag{4.58}$$

4.8 Multiple Random variables

We know how to implement the *stochastic Galerkin* method for models with *one* input random variable.

$$u(t,\xi) \approx \hat{u}_k(t,\xi) = \sum_{i=0}^k u_i(t)\psi_i(\xi). \tag{4.59}$$

Question. What if the model has J input random variables $\xi_1, \xi_2, \ldots, \xi_J$?

Example 4.12 (Two random variables)**.**

$$\frac{\mathrm{d}u}{\mathrm{d}t} = -\alpha u, \quad u(0) = u_0. \tag{4.60}$$

Suppose α and u_0 are *both* uncertain.

- Choose

$$\alpha = a_0 + a_1\xi_1, \quad u_0 = b_0 + b_1\xi_2, \quad \xi_1, \xi_2 \sim U(-\sqrt{3}, \sqrt{3}), \tag{4.61}$$

  where ξ_1, ξ_2 are *independent*.

- Define $\boldsymbol{\xi} = [\xi_1, \xi_2]^T$ and $\Gamma = [-\sqrt{3}, \sqrt{3}] \times [-\sqrt{3}, \sqrt{3}]$.

- The IVP solution is a function of both ξ_1 and ξ_2:

$$u(t, \boldsymbol{\xi}) = (b_0 + b_1\xi_2)e^{-(a_0 + a_1\xi_1)t}, \quad t \in [0, T], \quad \boldsymbol{\xi} \in \Gamma \tag{4.62}$$

When we view both the inputs and the solution as *functions* of ξ_1 and ξ_2, we can reformulate the IVP as follows.

Find $u : [0, T] \times \Gamma \to \mathbb{R}$ such that for every $\boldsymbol{\xi} \in \Gamma$,

$$\frac{\mathrm{d}u(t, \boldsymbol{\xi})}{\mathrm{d}t} = -\alpha(\boldsymbol{\xi})u(t, \boldsymbol{\xi}), \quad u(0, \boldsymbol{\xi}) = u_0(\boldsymbol{\xi}) \tag{4.63}$$

where, here, we define

$$\alpha(\boldsymbol{\xi}) := a_0 + a_1\xi_1, \quad u_0(\boldsymbol{\xi}) := b_0 + b_1\xi_2. \tag{4.64}$$

Example 4.13. Let $\xi_1, \xi_2, \xi_3 \sim U(-\sqrt{3}, \sqrt{3})$ be independent and define

$$\boldsymbol{\xi} = [\xi_1, \xi_2, \xi_3]^T. \tag{4.65}$$

Suppose we have

$$\alpha(\boldsymbol{\xi}) = a_0 + \sum_{r=1}^{2} a_r\xi_r, \quad u_0(\boldsymbol{\xi}) := b_0 + b_1\xi_3. \tag{4.66}$$

The IVP solution $u = u(t, \boldsymbol{\xi})$ where $\boldsymbol{\xi} \in \Gamma := [-\sqrt{3}, \sqrt{3}]^3$.

Find $u : [0, T] \times \Gamma \to \mathbb{R}$ such that for every $\boldsymbol{\xi} \in \Gamma$,

$$\frac{\mathrm{d}u(t, \boldsymbol{\xi})}{\mathrm{d}t} = -\alpha(\boldsymbol{\xi})u(t, \boldsymbol{\xi}), \quad u(0, \boldsymbol{\xi}) = u_0(\boldsymbol{\xi}) \tag{4.67}$$

Example 4.14 (*J* Random variables)**.**

$$\frac{\mathrm{d}u}{\mathrm{d}t} = -\alpha u + f(t), \quad u(0) = u_0. \tag{4.68}$$

Suppose we model the forcing term $f(t)$ as a *stochastic process*.

- If we choose a (truncated) KL expansion, then we can write

$$f(t, \boldsymbol{\xi}) = \mu(t) + \sum_{r=1}^{J} \sqrt{\lambda_r}\phi_r(t)\xi_r. \tag{4.69}$$

  where $\boldsymbol{\xi} = [\xi_1, \ldots, \xi_J]^T$ is a vector of J random variables.

- The IVP solution $u = u(t, \boldsymbol{\xi})$ where $\boldsymbol{\xi} \in \Gamma \subset \mathbb{R}^J$.

- Find $u : [0, T] \times \Gamma \to \mathbb{R}$ such that for every $\boldsymbol{\xi} \in \Gamma$,

$$\frac{\mathrm{d}u(t, \boldsymbol{\xi})}{\mathrm{d}t} = -\alpha u(t, \boldsymbol{\xi}) + f(t, \boldsymbol{\xi}), \quad u(0, \boldsymbol{\xi}) = u_0. \tag{4.70}$$

Question. Can we apply the *SG method* to these problems?

$$u(t, \boldsymbol{\xi}) \approx \widehat{u}(t, \boldsymbol{\xi}) = \sum_{i=1}^{N} u_i(t) \psi_i(\boldsymbol{\xi}). \tag{4.71}$$

1. Choose a polynomial *approximation space*:

$$S = \mathrm{span}\{\{\}\psi_1(\boldsymbol{\xi}), \ldots, \psi_N(\boldsymbol{\xi})\} \tag{4.72}$$

2. Set up the *weak formulation* on S, and seek $\widehat{u}$ with $\widehat{u}(t, \cdot) \in S$.

3. Solve the resulting *system of ODEs* for the coefficients $u_i(t)$.

Multivariate Polynomial

$$u(t, \boldsymbol{\xi}) \approx \widehat{u}(t, \boldsymbol{\xi}) = \sum_{i=1}^{N} u_i(t) \psi_i(\boldsymbol{\xi}). \tag{4.73}$$

- S need to be a set of multivariate polynomials in $\xi_1, \xi_2, \ldots, \xi_J$.

- We will chose polynomials of *total degree* $\leq k$.

- We need orthonormal basis polynomials $\psi_i(\boldsymbol{\xi})$ that satisfy

$$\mathbb{E}[\psi_i(\boldsymbol{\xi})\psi_j(\boldsymbol{\xi})] = \int_{\Gamma} \rho(\boldsymbol{y})\psi_i(\boldsymbol{y})\psi_j(\boldsymbol{y})\mathrm{d}\boldsymbol{y} = \delta_{i,j}. \tag{4.74}$$

Key Difference.

- $\Gamma \subset \mathbb{R}^J$ is a J-dimensional domain.

- The weight function is the *joint pdf* on $\boldsymbol{\xi}$.

4.9 Multi-index Notation

$$\{\psi_0(y), \psi_1(y), \psi_2(y), \ldots\} \tag{4.75}$$

For *univariate* polynomials, we use the index i to denote the *degree*.

For *multivariate* polynomials in J variables, we need a different notation, because we need to specify a polynomial degree for each variable:

$$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_J). \tag{4.76}$$

Multivariate Polynomials

We can construct a *mutlivariate* polynomial in $\boldsymbol{y} = [y_1, y_2, \ldots, y_J]^T$ by multiplying *univariate* ones together.

1. Suppose $\{\psi_0(y), \psi_1(y), \ldots\}$ is a given set of *univariate* polynomials, where i denotes the polynomial degree.

2. Pick the degree α_r of the univariate polynomials to multiply:

$$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_J), \quad \alpha_r \in \mathbb{N}_0, \quad r = 1, \ldots, J. \tag{4.77}$$

3. The define the *multivariate* polynomial

$$\psi_{\boldsymbol{\alpha}}(\boldsymbol{y}) = \prod_{r=1}^{J} \psi_{\alpha_r}(y_r). \tag{4.78}$$

4. The *total degree* is: $\alpha_1 + \cdots + \alpha_J$.

Example 4.15. Recall that the *Legendre* polynomails of degrees $0, 1, 2, 3$ are

$$\psi_0(y) = 1, \quad \psi_1(y) = y, \quad \psi_2(y) = \frac{\sqrt{5}}{2}(y^2 - 1), \quad \psi_3(y) = \sqrt{\frac{7}{3}}\left(\frac{5}{6}y^3 - \frac{3}{2}y\right). \tag{4.79}$$

- Let $J = 3$ and choose the multi-index $\boldsymbol{\alpha} = (3, 0, 1)$.

- We can define a polynomial in $\boldsymbol{y} = [y_1, y_2, y_3]^T$ via:

$$\psi_{\boldsymbol{\alpha}}(\boldsymbol{y}) = \psi_3(y_1)\psi_0(y_2)\psi_1(y_3) = \sqrt{\frac{7}{3}}\left(\frac{5}{6}y_1^3 - \frac{3}{2}y_1\right)y_3 \tag{4.80}$$

- This has total degree 4.

NOTE. J indicates how many random variables we are dealing with, i.e. `length`(xi). The use of $\boldsymbol{\alpha}$ indicates the choice of polynomials, i.e. `length`(alpha) = J, `sum`(alpha) = degree.

Define the set of *all* multi-indices of length J as follows

$$Q = \{\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_J), \quad \alpha_r \in \mathbb{N}_0, \quad \text{for } r = 1, \ldots, J\}. \tag{4.81}$$

To construct a set of multivariate polynomials, we just need to choose

- a family of univariate polynomials $\{\psi_0(y), \psi_1(y), \psi_2(y), \ldots\}$

- a subset $\Lambda \subset Q$ of multi-indices.

Then we have

$$\left\{\psi_{\boldsymbol{\alpha}}(\boldsymbol{y}) = \prod_{r=1}^{J} \psi_{\alpha_r}(y_r), \quad \boldsymbol{\alpha} \in \Lambda\right\} \tag{4.82}$$

4.10   Multivariate Orthonormal Polynomials

Suppose we have a model with J random inputs $\xi_1, \ldots, \xi_J : \Omega \to \mathbb{R}$ that are independent and identically distributed and define

$$\boldsymbol{\xi} = [\xi_1, \ldots, \xi_J]^T \tag{4.83}$$

To apply the stochastic Galerkin method, we choose a polynomial approximation space of the form

$$S_\Lambda = \text{span}\{\{\}\psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi}), \ \boldsymbol{\alpha} \in \Lambda\} \tag{4.84}$$

where $\Lambda \subset Q$ is a chosen set of *multi-indices*.

Using multi-index notation, the SG approximation is written

$$\widehat{u}(t, \boldsymbol{\xi}) = \sum_{\boldsymbol{\alpha} \in \Lambda} u_{\boldsymbol{\alpha}}(t)\psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi}). \tag{4.85}$$

That is, $\widehat{u}(t, \cdot) \in S_\Lambda$ for all $t \in [0, T]$.

Orthonormal Basis Polynomials

We want the basis polynomials $\psi_\alpha(\boldsymbol{\xi})$ for S_Λ to satisfy

$$\mathbb{E}[\phi_\alpha(\boldsymbol{\xi})\phi_\beta(\boldsymbol{\xi})] = \begin{cases} 1 & \text{if } \boldsymbol{\alpha} = \boldsymbol{\beta} \\ 0 & \text{if } \boldsymbol{\alpha} \neq \boldsymbol{\beta} \end{cases} \tag{4.86}$$

where

$$\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_J), \quad \boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_J) \tag{4.87}$$

are any pair of multi-indices belonging to Λ.

Weight Inner-Product

If ϕ_α and ϕ_β are functions of $\boldsymbol{\xi}$, where $\boldsymbol{\xi} : \Omega \to \Gamma \subset \mathbb{R}^J$ then

$$\mathbb{E}[\phi_\alpha(\boldsymbol{\xi})\phi_\beta(\boldsymbol{\xi})] = \int_\Gamma \rho(\boldsymbol{y})\phi_\alpha(\boldsymbol{y})\phi_\beta(\boldsymbol{y})\mathrm{d}\boldsymbol{y}. \tag{4.88}$$

Given the set Γ and the pdf $\rho(\boldsymbol{y})$, we can define a weighted inner product

$$\langle \phi_\alpha, \phi_\beta \rangle_\rho := \int_\Gamma \rho(\boldsymbol{y})\phi_\alpha(\boldsymbol{y})\phi_\beta(\boldsymbol{y})\mathrm{d}\boldsymbol{y}. \tag{4.89}$$

We need to construct a set of multivariate polynomials that satisfy

$$\langle \phi_\alpha, \phi_\beta \rangle_\rho = \begin{cases} 1 & \text{if } \boldsymbol{\alpha} = \boldsymbol{\beta} \\ 0 & \text{if } \boldsymbol{\alpha} \neq \boldsymbol{\beta} \end{cases} \tag{4.90}$$

Joint Probability Density Function

Let $\xi_r \in \Gamma_r \subset \mathbb{R}$, for $r = 1, 2, \dots, J$. Then $\boldsymbol{\xi} \in \Gamma$, where

$$\Gamma = \Gamma_1 \times \Gamma_2 \times \cdots \times \Gamma_J \subset \mathbb{R}^J. \tag{4.91}$$

If the ξ_r are *independent*, then the joint pdf of $\boldsymbol{\xi}$ is seperable:

$$\rho(\boldsymbol{y}) = \rho_1(y_1)\rho_2(y_2) \cdots \rho_J(y_J), \tag{4.92}$$

where $\rho_r : \Gamma_r \to [0, \infty)$ is the pdf of ξ_r.

Example 4.16. If $\xi_r \sim U(-\sqrt{3}, \sqrt{3})$ and independent, for $r = 1, 2, \dots, J$. Then

$$\Gamma = [-\sqrt{3}, \sqrt{3}]^J, \quad \rho_i(y_i) = \frac{1}{2\sqrt{3}}, \quad \rho(\boldsymbol{y}) = \left(\frac{1}{2\sqrt{3}}\right)^J. \tag{4.93}$$

Univariate Polynomials

We construct orthonormal multivariate polynomials by multiplying together univariate ones that are orthonormal.

For each $\xi_r \in \Gamma_r$, $r = 1, 2 \dots, J$. Find the family $\{\phi_0^r, \phi_1^r, \dots\}$ of univariate polynomials that are orthonormal w.r.t. the weight ρ_r on Γ_r. That is,

$$\langle \phi_i^r, \phi_j^r \rangle = \int_{\Gamma_r} \rho_r(y_r)\phi_i^r(y_r)\phi_j^r(y_r)\mathrm{d}y_r = \delta_{i,j} \tag{4.94}$$

If the variables are identically distributed then Γ_r and ρ_r are the same for each $r = 1, 2, \dots, J$. So we only need *one* family

$$\{\phi_0, \phi_1, \dots\} \tag{4.95}$$

For example the Legendre (for uniform), Hermite (for Gaussian), etc.

Uni to Multi variate Polynomials

Now let $\Lambda \subset Q$ be any finite subset of multi-indices, and consider the set

$$\{\psi_\alpha(\boldsymbol{\xi}),\ \alpha \in \Lambda\}. \tag{4.96}$$

where we define

$$\psi_\alpha(\boldsymbol{\xi}) = \prod_{r=1}^{J} \phi_{\alpha_r}(\xi_r) \tag{4.97}$$

where $\{\phi_0, \phi_1, \dots\}$ are the univariate polynomials that are orthonormal with respect to weighted inner product associated with the distribution of ξ_r.

Theorem 4.17. *For $\{\phi_\alpha(\boldsymbol{\xi}), \alpha \in \Lambda\}$, if the random variables ξ_r are independent and identically distributed, the above construction gives*

$$\mathbb{E}[\phi_\alpha(\boldsymbol{\xi})\phi_\beta(\boldsymbol{\xi})] = \langle \phi_\alpha(\boldsymbol{\xi}), \phi_\beta(\boldsymbol{\xi}) \rangle_\rho = \begin{cases} 1 & \text{if } \alpha = \beta \\ 0 & \text{if } \alpha \neq \beta \end{cases} \tag{4.98}$$

for all $\alpha, \beta \in \Lambda$.

Proof. See Exercise 6. $\qquad\qquad\square$

4.11 Total Degree Polynomial

Which Set of Multi-indices?

Let k be fixed, and define

$$\Lambda_k = \left\{ \alpha = (\alpha_1, \alpha_2, \dots, \alpha_J);\ \alpha_r \in \mathbb{N}_0\ \forall r = 1, \dots, J;\ \sum_{r=1}^{J} \alpha_r \leq k \right\} \tag{4.99}$$

The associated SG approximation space is

$$S_k := \mathrm{span}\{\{\}\phi_\alpha(\boldsymbol{\xi}),\ \alpha \in \Lambda_k\}. \tag{4.100}$$

And the SG approximation is

$$\widehat{u}(t, \boldsymbol{\xi}) = \sum_{\alpha \in \Lambda_k} u_\alpha \phi_\alpha(\boldsymbol{\xi}) \tag{4.101}$$

Question. What space is Λ_k?

Answer: Polynomials in $\boldsymbol{\xi}$ of total degree $\leq k$.

Example 4.18 (Uniform r.v. $J = 2, k = 2$)**.** Let $\xi_1, \xi_2 \sim U(-\sqrt{3}, \sqrt{3})$ be independent and choose $k = 2$. We construct an orthonormal basis for S_2 (polynomials in xi_1 and ξ_2 of total degree ≤ 2) as follows: Using the definition

$$S_2 = \mathrm{span}\{\{\}\phi_\alpha(\boldsymbol{\xi}),\ \alpha \in \Lambda_2\}, \tag{4.102}$$

where

$$\Lambda_2 = \{\alpha = (\alpha_1, \alpha_2);\ \alpha_1, \alpha_2 \in \mathbb{N}_0;\ \alpha_1 + \alpha_2 \leq 2\} \tag{4.103}$$

$\alpha = (\alpha_1, \alpha_2)$	$\phi_{\alpha_1}(\xi_1)\phi_{\alpha_2}(\xi_2)$	$\phi_\alpha(\xi)$
$(0,0)$	$\phi_0(\xi_1)\phi_0(\xi_2)$	1
$(1,0)$	$\phi_1(\xi_1)\phi_0(\xi_2)$	ξ_1
$(0,1)$	$\phi_0(\xi_1)\phi_1(\xi_2)$	ξ_2
$(2,0)$	$\phi_2(\xi_1)\phi_0(\xi_2)$	$\frac{\sqrt{5}}{2}(\xi_1^2 - 1)$
$(1,1)$	$\phi_1(\xi_1)\phi_1(\xi_2)$	$\xi_1\xi_2$
$(0,2)$	$\phi_0(\xi_1)\phi_2(\xi_2)$	$\frac{\sqrt{5}}{2}(\xi_2^2 - 1)$

Question. How many multi-indices are there in the index Λ_2?

    Answer. Six.

$$\Lambda_2 = \{(0,0), (0,1), (1,0), (0,2), (1,1), (2,0)\}. \tag{4.104}$$

Using the univariate Legendre polynomials $\{\phi_0, \phi_1, \dots\}$, we construct the six orthonormal basis functions for S_2.

Polynomials of total degree $\leq k$

Question. What's the dimension of S_k?

$$dim(S_k) = |\Lambda_k|. \tag{4.105}$$

The no. of multi-indices of length J whose components sum to k or less is

$$N_k = \frac{(J + k)!}{J!k!} \tag{4.106}$$

This is the no. of ODEs we have to solve to find the coefficients $u_\alpha(t)$.

J	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$
1	2	3	4	5	6
5	6	21	56	126	252
10	11	66	286	1001	3003
20	21	231	1771	10626	53130

Mean and Variance of SG Approximation

The SG approximation associated with S_k has the form

$$\widehat{u}(t, \xi) = \sum_{\alpha \in \Lambda_k} u_\alpha \phi_\alpha(\xi). \tag{4.107}$$

Note that $\mathbf{0} = (0, \dots, 0)$ is always a multi-index in the set Λ_k. The associated basis polynomial is

$$\phi_{\mathbf{0}}(\xi) = \phi_0(\xi_1)\phi_0(\xi_2) \cdots \phi_0(\xi_J) = 1. \tag{4.108}$$

Take the expectation gives

$$\mathbb{E}[\widehat{u}(t, \boldsymbol{\xi})] = \sum_{\boldsymbol{a} \in \Lambda_k} u_{\alpha}(t) \, \mathbb{E}[\phi_{\alpha}(\boldsymbol{\xi})\phi_{\boldsymbol{0}}(\boldsymbol{\xi})]$$

$$= u_{\boldsymbol{0}}(t) \quad \text{(Orthonormality)}$$

(4.109)

Using the same trick, we could find the variance of the SG approximation,

$$\mathbb{V}[\widehat{u}(t, \boldsymbol{\xi})] = \mathbb{E}[\widehat{u}^2(t, \boldsymbol{\xi})] - u_{\boldsymbol{0}}^2(t)$$

$$= \mathbb{E}\left[\left\{\sum_{\alpha \in \Lambda_k} u_{\alpha}\phi_{\alpha}(\boldsymbol{\xi})\right\}\left\{\sum_{\beta \in \Lambda_k} u_{\beta}\phi_{\beta}(\boldsymbol{\xi})\right\}\right] - u_{\boldsymbol{0}}^2(t)$$

$$= \sum_{\alpha \in \Lambda_k}\sum_{\beta \in \Lambda_k} u_{\alpha}u_{\beta} \, \mathbb{E}\left[\phi_{\alpha}(\boldsymbol{\xi})\phi_{\beta}(\boldsymbol{\xi})\right] - u_{\boldsymbol{0}}^2(t)$$

$$= \sum_{\alpha \in \Lambda_k} u_{\alpha}^2(t) - u_{\boldsymbol{0}}^2(t) = \sum_{\alpha \in \Lambda_k \setminus \{\boldsymbol{0}\}} u_{\alpha}^2(t)$$

(4.110)