# Multilevel Monte Carlo Methods for UQ

## Part I

Robert Scheichl
r.scheichl@uni-heidelberg.de

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

Institut für Angewandte Mathematik, Universität Heidelberg

Short Course, **Penn State University**, 21 & 23 April 2021
https://katana.iwr.uni-heidelberg.de/teaching/pennstate2021/



## Collaborators

- J Charrier (Marseille)
- KA Cliffe † (Nottingham)
- T Cui (Monash)
- G Detommaso (Amazon Berlin)
- T Dodwell (Exeter)
- C Fox (Otago)
- A Gilbert (UNSW Sydney)
- M Giles (Oxford)
- IG Graham (Bath)
- Ch Ketelsen (Boulder)
- F Kuo (UNSW Sydney)
- J Lang (Darmstadt)
- MB Lykkegaard (Exeter)
- E Müller (Bath)
- Ch Schwab (ETH Zürich)
- T Shardlow (Bath)
- D Silvester (Manchester)
- IH Sloan (UNSW Sydney)
- A Stuart (Caltech)
- AL Teckentrup (Edinburgh)
- E Ullmann (TU Munich)
- ...

# 1. What is Uncertainty Quantification?

## Motivation

- **Mathematical modelling**, e.g., in the form of differential equations, is essential to understand, optimise, control and predict physical, biological and engineering processes.
- **Numerical methods** are central in solving these often very complex mathematical models.
- These methods have reached a high level of maturity & sophistication.
- Many excellent PDE software packages exist to model complex problems efficiently and robustly.
- **But** models have input data that are typically not known precisely
  (parameters, source term, domain shape, boundary conditions, etc...)
- **It is of great importance to determine these parameters, their influence on the solution & uncertainties due to their variability.**

To find (and analyse) efficient numerical methods for those tasks is still a **very active field** of research. Here, I will present the promising family of

**Multilevel Monte Carlo methods.**



## What is Uncertainty Quantification (UQ)?

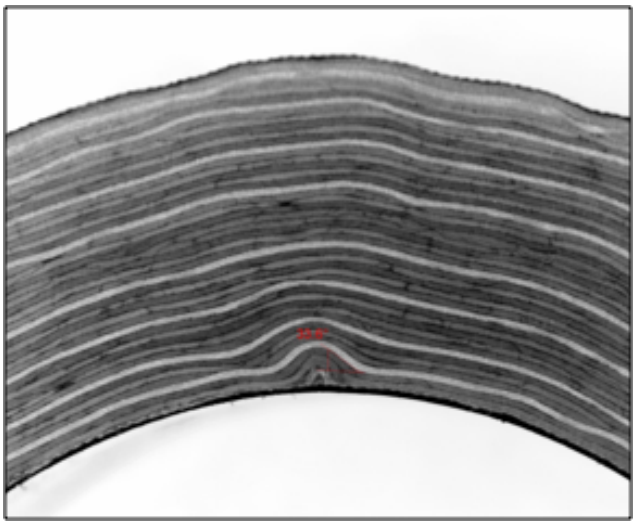
### Uncertainty in Modern Life

Many aspects of modern life involve uncertainty:

- **Society:** military, finance, insurance industry, elections
- **Environment:** weather, climate, seismic, subsurface geophysics
- **Engineering:** automobiles, aircraft, structures, materials
- **Biology:** health, medicine, pharmaceuticals, gene expression, cancer research
- **Physics:** quantum physics, radioactive decay

# What is Uncertainty Quantification (UQ)?

## Examples

Source: GKN Aerospace

Performance "knock-down" factors through wrinkling defects in carbon fibre composite aeroplane wing

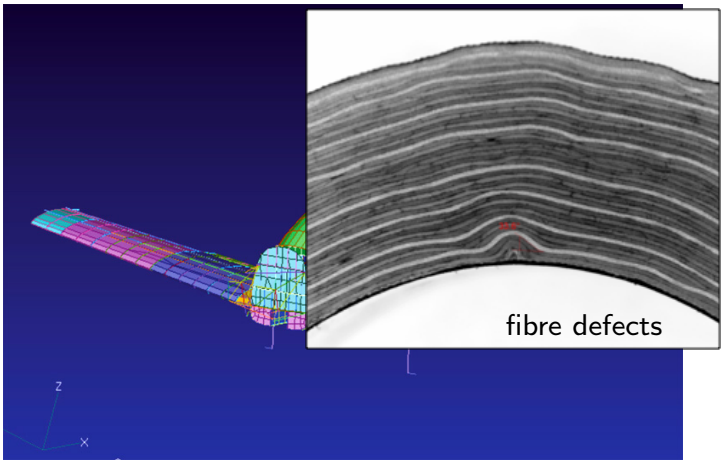


# Modelled via PDEs with Random Coefficients

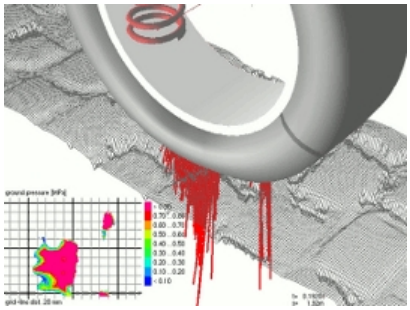
- **Structural Mechanics** (e.g. composites, tires or bone):

$$\nabla \cdot \left( \overline{\overline{C}}(x,\omega) : \frac{1}{2} \left[ \nabla \mathbf{u} + \nabla \mathbf{u}^T \right] \right) + \mathbf{F}(x,\omega) \;=\; 0 \quad \text{in} \quad \Omega(\omega)$$

subject to uncertain BCs




contact on rough surface



## The "Fruit Fly" of UQ

A popular **model problem** in the UQ community is the steady-state diffusion problem with uncertain coefficient function a(x):

$$-\nabla\cdot(a\nabla u) = f \qquad \text{on domain } D \subset \mathbb{R}^d. \quad \text{(an elliptic PDE)}$$

Rather than in the solution $u$, typically we are interested in a functional $Q$ of the solution $u$, known as quantity of interest (QoI):

$$\text{e.g.} \qquad Q(u) = u(\mathbf{x}_0) \quad \text{or} \quad Q(u) = \tfrac{1}{|D_0|}\int_{D_0} u(\mathbf{x})\,\mathbf{dx}.$$

In what way might uncertainty in the coefficient $a$ be addressed?

- Worst case analysis: Could calculate *uncertainty interval*

$$\mathcal{I} = \Big[\inf_{\|a-a_0\|<\varepsilon} Q(u(a)), \sup_{\|a-a_0\|<\varepsilon} Q(u(a))\Big].$$



## Probabilistic Modelling of Uncertainty

**But**, in general, some coefficients with $\|a - a_0\| < \varepsilon$ are more likely than others

$$\Rightarrow \textbf{ Probabilistic approach}$$

- Introduce probability measure on $S := \{a : \|a - a_0\| < \varepsilon\}$.
- Then $Q(u(\cdot))$ (as measurable mapping on $S$) induces probability measure for the QoI ("Uncertainty Propagation") Today
- **Big issue:** choice of distribution, information too subjective?
  (Some classical guidelines: Laplace's principle, maximum entropy,...)
- Choosing distribution based on (output) is starting point for *Bayesian inference* ("Uncertainty Quantification") Friday
  (Here the choice of distribution on $S$, the "prior", becomes less important, although classical, frequentist statisticians still consider it too subjective!)

## Learning from Sparse and Noisy Data

**Particulary important** when studying complex physical or biological systems where only **very sparse and noisy data** is available, but good mathematical models exist to describe the system.

**Examples:**

- Atmospheric, ocean or subsurface flow
- Cardiovascular system or tracer diffusion in brain imaging
- Structural mechanics of composite materials or bones

**Machine Learning** and **Neural Networks** alone will **not be sufficient!**

Need to add **mathematical modelling** and **numerical analysis** to the toolkit – **New Challenges!**



# 2. Computational Challenges

# Computational Challenges

in simulating PDEs with highly heterogeneous random coefficients

$$-\nabla \cdot (a(\mathbf{x},\omega)\nabla u(\mathbf{x},\omega)) = f(\mathbf{x},\omega), \quad \mathbf{x} \in D \subset \mathbb{R}^d, \ \omega \in \Omega \text{ (prob. space)}$$

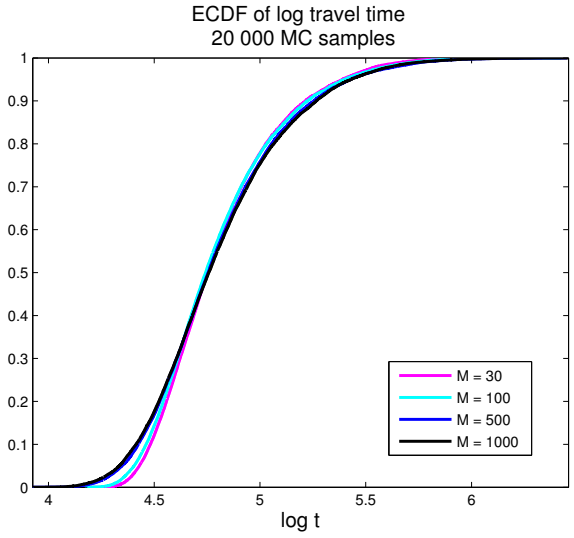
- **Sampling** from random field $\log a(\mathbf{x},\omega)$ (correlated Gaussian):
  - truncated Karhunen-Loève expansion of $\log a$
  - matrix factorisation, e.g. circulant embedding (FFT)
  - via pseudodifferential "precision" operator (PDE solves)
- **High-Dimensional Quadrature** The central focus of this course!
  - Monte Carlo, Quasi-Monte Carlo, Multilevel Monte Carlo
  - Sparse Grids & stochastic Galerkin/collocation
- **Solve** large number of **multiscale** deterministic PDEs:
  - Efficient discretisation & FE error analysis (mesh size $h$)
  - Multigrid Methods, AMG, DD Methods Big focus here at Penn State!



# Why is it Computationally so Challenging?

- Low regularity (global): $a \in C^\eta$, $\eta < \nu < 1$ (Hölder) $\Longrightarrow$ **fine** FE mesh $h \ll 1$
- Large variance $\sigma^2$ & exponential $\Longrightarrow$ **high** contrast $a_{\max}/a_{\min} > 10^6$
- Small correlation length $\lambda$

  $\Longrightarrow$ **multiscale** + **high stochastic dimension** $s \gg 1$




**Source: Ernst et al, 2014** ($s = M$)

## Standard Monte Carlo Quadrature

$$\underset{\text{random input}}{\mathbf{X}(\omega) \in \mathbb{R}^s} \overset{\text{FE Model}(h)}{\longrightarrow} \underset{\text{state vector}}{\mathbf{U}(\omega) \in \mathbb{R}^{M_h}} \overset{\text{Output}}{\longrightarrow} \underset{\text{quantity of interest}}{Q_{h,s}(\omega) \in \mathbb{R}}$$

- **Here:** $\mathbf{X}$ multivariate Gaussian for KL expansion; $\mathbf{U}$ numerical PDE solution; $Q_{h,s}$ a (non)linear functional of $\mathbf{U}$
- Real QoI $Q(\omega)$ inaccessible (exact PDE), but we can assume $\mathbb{E}[Q_{h,s}] \overset{h\to 0,\ s\to\infty}{\longrightarrow} \mathbb{E}[Q]$ and $|\mathbb{E}[Q_{h,s} - Q]| = \mathcal{O}(h^\alpha) + \mathcal{O}(s^{-\alpha'})$
- **Standard Monte Carlo** estimator for $\mathbb{E}[Q]$: More detail below!

$$\hat{Q}^{\mathrm{MC}} := \frac{1}{N} \sum_{i=1}^{N} Q_{h,s}^{(i)}$$

where $\{Q_{h,s}^{(i)}\}_{i=1}^N$ are i.i.d. samples computed with FE Model($h$)

- Cost per sample is $\mathcal{O}(M_h^\gamma)$ (optimal: $\gamma = 1$, e.g. multigrid)



## Standard Monte Carlo Quadrature

- Convergence of plain vanilla MC **(mean square error)**:

$$\underbrace{\mathbb{E}\big[(\hat{Q}^{\mathrm{MC}} - \mathbb{E}[Q])^2\big]}_{=:\ \mathrm{MSE}} = \mathbb{V}[\hat{Q}^{\mathrm{MC}}] + \big(\mathbb{E}[\hat{Q}^{\mathrm{MC}}] - \mathbb{E}[Q]\big)^2$$

$$= \underbrace{\frac{\mathbb{V}[Q_{h,s}]}{N}}_{\text{sampling error}} + \underbrace{\big(\mathbb{E}[Q_{h,s} - Q]\big)^2}_{\text{model error ("bias")}}$$

- **Typical:** $\alpha = 1 \ \Rightarrow$

  $\mathrm{MSE} = \mathcal{O}(N^{-1}) + \mathcal{O}(h^2) \le \mathrm{TOL}^2$ and so $h \sim \mathrm{TOL}$ and $N \sim \mathrm{TOL}^{-2}$

- Using optimal PDE solver: $\mathrm{Cost} = \mathcal{O}(Nh^{-d}) = \mathcal{O}(\mathrm{TOL}^{-(d+2)})$

  (e.g. for $\mathrm{TOL} = 10^{-3}$: $h \sim 10^{-3}$, $N \sim 10^6$ and $\mathrm{Cost} = \mathcal{O}(10^{12})$ in 2D!!)

Quickly becomes **prohibitively expensive !**

## Numerical Experiment with Standard Monte Carlo

$D = (0,1)^2$, $Q = \| - a\frac{\partial u}{\partial x_1}\|_{L^1(D)}$, sampling via truncated KL expansion, and using mixed FEs and the AMG solver `amg1r5` [Ruge, Stüben, 1992]

- Numerically observed FE-error: $\approx \mathcal{O}(h^{3/4}) \implies \alpha \approx 3/4$.
- Numerically observed cost/sample: $\approx \mathcal{O}(h^{-2}) \implies \gamma \approx 1$.
- **Total cost** to get RMSE $\mathcal{O}(\mathrm{TOL})$: $\approx \mathcal{O}(\mathrm{TOL}^{-14/3})$

To get error reduction by a factor 2 $\longrightarrow$ Cost grows by a **factor 25!**

**Case 1:** $\sigma^2 = 1$, $\lambda = 0.3$, $\nu = 0.5$

| TOL | $h^{-1}$ | $N$ | Cost |
|---|---|---|---|
| 0.01 | 129 | $1.4 \times 10^4$ | 21 min |
| 0.002 | 1025 | $3.5 \times 10^5$ | 30 days |

**Case 2:** $\sigma^2 = 3$, $\lambda = 0.1$, $\nu = 0.5$

| TOL | $h^{-1}$ | $N$ | Cost |
|---|---|---|---|
| 0.01 | 513 | $8.5 \times 10^3$ | 4 h |
| 0.002 | **Prohibitively large!!** | | |

(actual numbers & CPU times on a cluster of 2GHz Intel T7300 processors)



## Alternatives

- **Polynomial quadrature: stochastic Galerkin/collocation methods**
  - Cost grows very fast with dimension $s$ & polynomial order $q$
    $\rightarrow$ #stochastic DOFs $N_{\mathrm{SC}} = \mathcal{O}\left(\frac{(s+q)!}{s!q!}\right)$ **(faster than exponential!)**
  - Lower number with **sparse grids** (Smolyak), but **still exponential** growth with $s$!

The **"Curse of Dimensionality"**

  - Anisotropic sparse grids or adaptive best $N$-term approximation can be dimension independent with **sufficient smoothness**! [Zech, 2018]
- **Monte Carlo type methods**
  - Convergence of plain vanilla Monte Carlo is **always dimension independent** ! (No smoothness needed!) **BUT** (as shown) order is **way too slow**: $\mathcal{O}(N^{-1/2})$!
  - **Quasi-Monte Carlo** can also be dimension independent and (almost) $\mathcal{O}(N^{-1})$! **But** requires also **(some) smoothness** !

**Focus here instead:** Use a multilevel hierarchy of numerical models **!!**

# 3. Convergence & Complexity of Basic Monte Carlo



## Basic Monte Carlo Simulation – Convergence Results

- Given a sequence $\{X_k\}$ of i.i.d. copies of a given random variable $X$, basic MC simulation uses the estimator

$$\mathbb{E}[X] \approx \frac{S_N}{N}, \qquad S_N = X_1 + \cdots + X_N.$$

- **Strong Law of Large Numbers**: $\frac{S_N}{N} \to \mathbb{E}[X]$ a.s.
- Also, for any measurable function $f$: $\frac{1}{N}\sum_{k=1}^{N} f(X_k) \to \mathbb{E}[f(X)]$ a.s.
- **Central Limit Theorem**: If $\mathbb{E}[X] = \mu$ and $\mathbf{Var}[X] = \sigma^2$, then

$$\mathbb{E}[S_N] = N\mu, \quad \mathbf{Var}[S_N] = N\sigma^2 \quad \text{and} \quad S_N^* = \frac{S_N - N\mu}{\sqrt{N}\sigma} \to \mathrm{N}(0,1),$$

i.e. the estimate is unbiased, the standard error is $\sigma N^{-1/2}$ and the distribution of the normalised RV $S_N^*$ becomes Gaussian as $N \to \infty$.

(if $\mathbf{Var}[X] < \infty$ then the **normalised RV** $X^* := \frac{X - \mathbb{E}[X]}{\sqrt{\mathbf{Var}[X]}}$ has $\mathbb{E}[X^*] = 0$, $\mathbf{Var}[X^*] = 1$)

## Basic Monte Carlo Simulation – Convergence Results

- Mean square convergence:
$$\mathbb{E}\left[\left(\frac{S_N}{N}-\mu\right)^2\right] = \textbf{Var}\,\frac{S_N}{N} = \frac{\sigma^2}{N} \to 0.$$
- *Chebyshev's Inequality* implies, for any $\epsilon > 0$:
$$\mathbb{P}\left\{\left|\frac{S_N}{N}-\mu\right| > N^{-1/2+\epsilon}\right\} \le \frac{\sigma^2}{N^{2\epsilon}},$$
(i.e. the probability of the error being $> N^{-1/2+\epsilon}$ converges to zero as $N \to \infty$)
- If $\rho := \mathbb{E}\left[|X-\mu|^3\right] < \infty$, then the *Berry-Esseen Inequality* gives
$$\left|\mathbb{P}\{S_N^* \le x\} - \Phi(x)\right| \le \frac{\rho}{2\sigma^3\sqrt{N}},$$
where $\Phi$ denotes *cumulative density function (CDF)* of $\mathrm{N}(0,1)$.
- Using Berry-Esseen, the asymptotic 95% confidence interval for $S_N/N$ is
$$0.95 - \frac{\rho}{\sigma^3\sqrt{N}} \le \mathbb{P}\left\{\mu \in \left[\frac{S_N}{N} - \frac{1.96\sigma}{\sqrt{N}}, \frac{S_N}{N} + \frac{1.96\sigma}{\sqrt{N}}\right]\right\} \le 0.95 + \frac{\rho}{\sigma^3\sqrt{N}}$$



# 4. A Simple ODE Example

## Predator-Prey Dynamical System

Now apply Monte Carlo in a UQ application. Consider the **Lotka-Volterra** (or **predator-prey**) model of the dynamics of two interacting populations

$$\dot{\mathbf{u}} = \begin{bmatrix} \dot{u}_1 \\ \dot{u}_2 \end{bmatrix} = \begin{bmatrix} \theta_1 u_1 \ - \ \theta_{12} u_1 u_2 \\ \theta_{21} u_1 u_2 \ - \ \theta_2 u_2 \end{bmatrix} = \mathbf{f}(\mathbf{u}), \quad \mathbf{u}(0) = \mathbf{u}_0,$$

where $u_1$ is the number of *prey*, $u_2$ is the number of *predator* and $\theta_1, \theta_2, \theta_{12}, \theta_{21} \geq 0$ are *parameters* describing the interaction of the two species.

- For simplicity, assume that

  $$\theta_1 = \theta_2 = \theta_{12} = \theta_{21} = 1$$

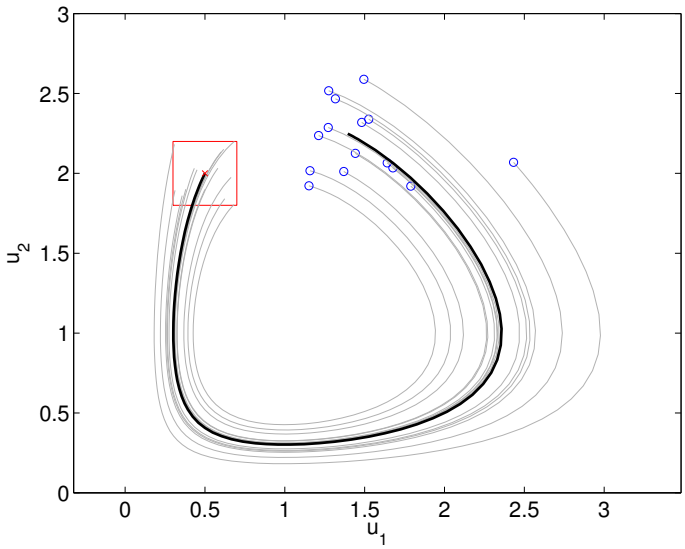
  and only the vector of initial conditions $\mathbf{u}_0$ is **uncertain**.
- We model it as a (uniform) random vector $\mathbf{u}_0 \sim \mathrm{U}(\Gamma)$, where $\Gamma$ denotes the square

  $$\Gamma = \overline{\mathbf{u}}_0 + [-\delta, \delta]^2.$$

**Goal:** estimate $\mathbb{E}[u_1(T)]$ at time $T > 0$ using the Monte Carlo method.



## Predator-Prey Dynamical System – Sample Trajectories




Population dynamics problem (with $\theta_1 = \theta_2 = \theta_{12} = \theta_{21} = 1$) integrated over $[0, T]$ with $\overline{\mathbf{u}}_0 = [0.5, 2]^\mathsf{T}$, $\delta = 0.2$ and $T = 6$. Unperturbed trajectory (black) alongside 15 perturbed trajectories. For the unperturbed trajectory $u_1(T) = 1.3942$.

## Modelling Epidemics like COVID-19

- Obviously there are arbitrarily many variations to this simple UQ problem (the distribution of $\mathbf{u}_0$ may be more complicated, the interaction parameters may also be uncertain, there may be more species, or the quantity of interest may be something more complicated) ... **in particular** ...
- A special case of the **Lotka-Volterra** model is the simplest and most widely used model for the spread of diseases (such as COVID-19):

### SIR model

$$\begin{aligned} \dot{S} &= -\tfrac{\beta}{N} SI \\ \dot{I} &= \tfrac{\beta}{N} SI - \gamma I \\ \dot{R} &= \gamma I \end{aligned}$$

where population of $N$ individuals is divided into categories *susceptibles* ($S$), *infecteous* ($I$) and *recovered* ($R$). The total number $N = S + I + R$ of individuals is assumed to be constant (i.e. birth/death processes negligible)

- For constant $N$, this problem can be reduced to solving the first two ODEs, which is the Lotka-Volterra system with $\theta_1 = 0$, $\theta_{12} = \theta_{21} = \beta/N$, $\theta_2 = \gamma$.



## Modelling Epidemics like COVID-19

The current situation, including lock-down measures, is modelled more accurately with the **SEIR model**, which includes a category *exposed* ($E$):

### SEIR model

$$\begin{aligned} \dot{S} &= \mu(N - S) - \tfrac{\beta}{N} SI \\ \dot{E} &= \tfrac{\beta}{N} SI - (\mu + \alpha)E \\ \dot{I} &= \alpha E - (\gamma + \mu)I \\ \dot{R} &= \gamma I - \mu R \end{aligned}$$

One of my postdocs, Tobias Siems, has been collaborating with the *Heidelberg Institute for Global Health* and the *regional health authority* last year to model the **Rhein-Neckar-Kreis** with **SEIR** (and a further refinement **SEIHR**) in order to predict case numbers and the resulting need for hospital beds.

## Explicit Euler Discretisation

- Denote by $\mathbf{u}_M = \mathbf{u}_M(\omega)$ the explicit Euler approximation after $M = M_h$ time steps of length $h = \frac{T}{M}$, starting with initial data $\mathbf{u}_0 = \mathbf{u}_0(\omega)$, i.e.

$$\mathbf{u}_j = \mathbf{u}_{j-1} + h\mathbf{f}(\mathbf{u}_{j-1}), \quad j = 1, \dots, M_h.$$

- Explicit Euler has consistency order 1 and thus there exists a constant $K > 0$ sucht that the discretisation error can be bounded by

$$\|\mathbf{u}(T) - \mathbf{u}_{M_h}\| \leq K\, h\,.$$

- Define the quantity of interest (QoI) $Q = \phi(\mathbf{u}(T)) = u_1(T)$ for $\mathbf{u} = [u_1, u_2]^\mathsf{T}$ and estimate $\mathbb{E}[Q_h]$ using the MC method just described with $Q_h = \phi(\mathbf{u}_{M_h})$.
- The QoI $\phi$ is Lipschitz-continuous with constant $L = 1$, such that also

$$|\mathbb{E}[Q] - \mathbb{E}[Q_h]| = |\mathbb{E}[Q - Q_h]| \leq K\, h. \tag{4.1}$$

- Denote the Monte Carlo estimator for $\mathbb{E}[Q_h]$ by

$$\widehat{Q}_h := \widehat{Q}_{h,N} = \frac{1}{N}\sum_{k=1}^{N} Q_h^{(k)} \quad N \text{ i.i.d. samples } Q_h^{(1)}, \dots, Q_h^{(N)} \text{ of } Q_h.$$

  Expect better approximations for $N$ large and $h$ small.



## Bias-Variance Decomposition – Balancing Error Contributions

### Lemma 4.1 (Bias-Variance Decomposition)

*The mean square error (MSE) can be expanded*

$$\mathbb{E}\left[\left(\mathbb{E}[Q] - \widehat{Q}_h\right)^2\right] = \left(\mathbb{E}[Q - Q_h]\right)^2 + \frac{\mathbf{Var}[Q_h]}{N}$$

*Proof.* Demonstrated on tablet.

*Hint:* Note that $\mathbb{E}[Q]$ is constant and only $\widehat{Q}_h$ is actually random. □

Thus, using the bias error bound above and the fact that, for $h$ sufficiently small, $\mathbf{Var}[Q_h] \leq \sigma^2_{\text{bnd}} \leq 1.1\,\mathbf{Var}[Q]$ (independently of $h$), we get the following bound:

$$\text{MSE} := \mathbb{E}\left[\left(\mathbb{E}[Q] - \widehat{Q}_h\right)^2\right] \;\leq\; K^2h^2 + \sigma^2_{\text{bnd}} N^{-1} \tag{4.2}$$

## Balancing Discretisation and Sampling Error (in probability)

Using above convergence results, error can also be bounded in probability:

- Error with $N$ samples and $M = T/h$ time steps:

$$e_{h,N} := |\mathbb{E}[Q] - \widehat{Q}_h| \leq \underbrace{|\mathbb{E}[Q] - \mathbb{E}[Q_h]|}_{\text{discretisation error}} + \underbrace{|\mathbb{E}[Q_h] - \widehat{Q}_h|}_{\text{Monte Carlo error}}$$

- Using the asymptotic 95% confidence interval for the MC error on Slide 21 (with $\mathbf{Var}[Q_h] = \sigma_h^2 \leq \sigma_{\text{bnd}}^2$) we get

$$\mathbb{P}\left\{\left|\mathbb{E}[Q_h] - \widehat{Q}_h\right| \leq \frac{1.96\sigma_h}{\sqrt{N}}\right\} > 0.95 + \mathcal{O}(N^{-1/2})$$

- Combining this with discretisation error in (4.1) (using triangle inequality):

$$\mathbb{P}\Big\{e_{h,N} \leq K\,h + 1.96\sigma_h N^{-1/2}\Big\} > 0.95 + \mathcal{O}(N^{-1/2}). \qquad (4.3)$$



## Monte Carlo Complexity for Predator-Prey Problem

Finally noting that the cost in each time step is 8 FLOPs, the total cost for the MC estimator is

$$\text{Cost}(\widehat{Q}_h) = 8M_hN = 8T\,h^{-1}N \quad \text{(FLOPs)} \qquad (4.4)$$

and we have the following complexity result:

### Proposition 4.2 (Monte Carlo Complexity)

*The* **total cost** *to compute a* **standard Monte Carlo** *estimator for* $\mathbb{E}[u_1(T)]$ *for the* **predator-prey** *model with* **explicit Euler** *time discretisation, such that* $MSE < \varepsilon^2$ *or* $\mathbb{P}\{e_{h,N} < \varepsilon\} > \theta$ *for any* $\theta \in (0,1)$, *satisfies*

$$\text{Cost}(\widehat{Q}_h) \;=\; \mathcal{O}(\varepsilon^{-3}).$$

*Proof.* (only the proof for in probability) A sufficient condition for $e_{h,N} < \varepsilon$ is

$$K\,h = \varepsilon/2 \quad \text{and} \quad 1.96\sigma_h N^{-1/2} = \varepsilon/2 \quad \text{(balancing the two terms)}.$$

This leads to

$$h = \tfrac{1}{2}K\varepsilon \quad \text{and} \quad N = 3.92^2\sigma_h^2\varepsilon^{-2} \quad \Rightarrow \quad \text{Cost}(\widehat{Q}_h) \leq \frac{256T\sigma_{\text{bnd}}^2}{K}\,\varepsilon^{-3}.$$

□

# 5. The Multilevel Monte Carlo Method



## History

- The **multilevel Monte Carlo** method is a powerful "new" variance reduction technique (especially for UQ applications).
- First ideas for high-dimensional quadrature by [Heinrich, 2000].
- Independently discovered and popularised by [Giles, 2007] in the context of **stochastic DEs in mathematical finance**.
- First papers in the context of UQ:
  - [Barth, Schwab, Zollinger, 2011] & [Cliffe, Giles, **RS**, Teckentrup, 2011]
- Stochastic simulation of discrete state systems (biology, chemistry)
  - [Anderson, Higham, 2012]
- ...

**Goal:** Estimate $\mathbb{E}[Q]$ of an inacccessible RV $Q$ (e.g. derived from the solution of DE)

**Assume:** Access to sequence of approximations $Q_h \approx Q$, parametrised by $h$, such that $\lim_{h\to 0} Q_h = Q$ (#time steps, #grid points, ...)

**Idea:** Reduce variance by a clever use of the hierarchy of approximations.

## Abstract Complexity Result for Standard MC

Recall from Lemma 4.1 that mean square error (MSE) for standard MC estimator $\widehat{Q}_{h,N}$ (using samples from the approximation $Q_h$ instead of $Q$) expands as

$$\mathbb{E}\left[\left(\widehat{Q}_{h,N} - \mathbb{E}[Q]\right)^2\right] = \left(\mathbb{E}[Q_h - Q]\right)^2 + \frac{\mathbf{Var}[Q_h]}{N}.$$

Thus, we can derive an abstract version of Proposition 4.2 (with identical proof):

### Theorem 5.1 (Abstract Complexity Theorem for standard MC)

*Assume that there exist constants $\alpha, \gamma > 0$, such that*

$$|\mathbb{E}[Q_h - Q]| = \mathcal{O}(h^\alpha), \quad \text{as } h \to 0, \tag{5.1}$$

$$\mathit{Cost}(Q_h^{(k)}) = \mathcal{O}(h^{-\gamma}), \quad \text{as } h \to 0, \tag{5.2}$$

*where $\mathit{Cost}(Q_h^{(k)})$ denotes the cost to compute a sample from approximation $Q_h$.*

*Then, for any $\varepsilon > 0$ and $\theta \in (0,1)$, the* **total cost** *to compute a standard Monte Carlo estimator for $\mathbb{E}[Q]$, such that $\mathit{MSE} < \varepsilon^2$ or $\mathbb{P}\{e_{h,N} < \varepsilon\} > \theta$, satisfies*

$$\mathit{Cost}(\widehat{Q}_{h,N}) = \mathcal{O}(\varepsilon^{-2-\gamma/\alpha}).$$



## Multilevel Estimator

- **Key idea:** use samples of $Q_h$ on a hierarchy of different **levels**, i.e., for different values $h_0, \ldots, h_L$ of the discretization parameter, and decompose

$$\mathbb{E}[Q_{h_L}] = \mathbb{E}[Q_{h_0}] + \sum_{\ell=1}^{L} \mathbb{E}\left[Q_{h_\ell} - Q_{h_{\ell-1}}\right] =: \sum_{\ell=0}^{L} \mathbb{E}[Y_\ell],$$

- For simplicity, we will often choose $h_{\ell-1} = m h_\ell$, $\ell = 1, \ldots, L$, for some $m \in \mathbb{N} \setminus \{1\}$ and $h_0 > 0$) (uniform grid refinement).
- Given estimators $\{\widehat{Y}_\ell\}_{\ell=0}^L$ for $\mathbb{E}[Y_\ell]$, we refer to

$$\widehat{Q}_L^{\mathsf{ML}} := \sum_{\ell=0}^{L} \widehat{Y}_\ell$$

as a **multilevel estimator** for $Q$.

- Different variants of this multilevel estimator now arise from different choices of the level estimators, e.g. standard Monte Carlo, quasi-Monte Carlo, etc ...

## Multilevel Monte Carlo (MLMC) Estimator

- If each $\widehat{Y}_\ell$ is itself a standard Monte Carlo estimator, i.e.,

$$\widehat{Y}_0 = \widehat{Y}_{0,N_0} := \frac{1}{N_0} \sum_{k=1}^{N_0} Q_{h_0}^{(k)}$$

and

$$\widehat{Y}_\ell = \widehat{Y}_{\ell,N_\ell} := \frac{1}{N_\ell} \sum_{k=1}^{N_\ell} \left( Q_{h_\ell}^{(k)} - Q_{h_{\ell-1}}^{(k)} \right), \qquad \ell = 1, \ldots, L,$$

one obtains the multilevel Monte Carlo estimator and $\widehat{Q}_L^{\mathsf{ML}}$ is **unbiased**.

- If all expectations $\mathbb{E}[Y_\ell]$ are sampled independently (not neccessary), then

$$\mathbf{Var}\, \widehat{Q}_L^{\mathsf{ML}} = \sum_{\ell=0}^{L} \mathbf{Var}\, \widehat{Y}_\ell.$$

and the associated MSE has the standard decomposition

$$\mathbb{E}\left[ \left( \widehat{Q}_{L,\{N_\ell\}}^{\mathsf{ML}} - \mathbb{E}[Q] \right)^2 \right] = \mathbb{E}\left[ Q_{h_L} - Q \right]^2 + \sum_{\ell=0}^{L} \frac{\mathbf{Var}\, Y_\ell}{N_\ell}$$

into bias and sample variance (shown identically to Lemma 4.1 for standard MC).



## MLMC variance reduction

- Choose the discretisation parameter $h_L$ on the highest level and the numbers of samples $(N_\ell)_{\ell=0}^L$ again to balance the terms in the MSE.
- The bias term is the same as for the standard MC estimator if $h_L = h$, so that under Assumption (5.1), this leads again to a choice of $h_L = \mathcal{O}(\varepsilon^{1/\alpha})$.
- **But why** do we get variance reduction or lower cost for the same variance?
- **Two reasons:**

1. As we coarsen the problem, the cost per sample **decays** rapidly from level to level under Assumption (5.2); by a factor $m^\gamma$ if $h_{\ell-1}/h_\ell = m$.
2. Since $Q_h \to Q$, then $\mathbf{Var}[Y_\ell] = \mathbf{Var}[Q_{h_\ell} - Q_{h_{\ell-1}}] \to 0$ as $\ell \to \infty$, allowing for smaller and smaller sample sizes $N_\ell$ on higher and higher levels.

## Optimal Sample Sizes

- The cost of the MLMC estimator is

$$\text{Cost}(\widehat{Q}^{\text{ML}}_{L,\{N_\ell\}}) = \sum_{\ell=0}^{L} N_\ell \mathcal{C}_\ell, \qquad \mathcal{C}_\ell := \text{Cost}(Y_\ell^{(k)}).$$

- Treating the $N_\ell$ as continuous variables, the cost of the MLMC estimator can be minimised for a fixed variance

$$\sum_{\ell=0}^{L} \frac{\mathbf{Var}\, Y_\ell}{N_\ell} = \frac{\varepsilon^2}{2}$$

- The solution to this constrained minimisation problem is

$$N_\ell = \gamma \sqrt{\mathbf{Var}[Y_\ell]/\mathcal{C}_\ell} \tag{5.3}$$

with the constant $\gamma$ chosen such that the total variance is $\frac{\varepsilon^2}{2}$, leading to the constant $\gamma = \frac{2}{\varepsilon^2} \sum_\ell \sqrt{\mathcal{C}_\ell \mathbf{Var}[Y_\ell]}$.

Demonstrated on tablet.



## Cost Comparison MLMC vs. Standard MC

- Thus the total cost on level $\ell$ is proportional to $\sqrt{\mathcal{C}_\ell \mathbf{Var}[Y_\ell]}$ and therefore

$$\text{Cost}(\widehat{Q}^{\text{ML}}_{L,\{N_\ell\}}) \leq \frac{2}{\varepsilon^2} \left( \sum_{\ell=0}^{L} \sqrt{\mathcal{C}_\ell \mathbf{Var}[Y_\ell]} \right)^2$$

- For comparison, standard MC has $\text{Cost}(\widehat{Q}_{h_L,N}) = \frac{2}{\varepsilon^2} \mathcal{C}_L \mathbf{Var}[Q_{M_L}]$.
- If $\mathbf{Var}[Y_\ell]$ decays faster than $\mathcal{C}_\ell$ increases, the cost on level $\ell = 0$ dominates. Since $\mathbf{Var}[Q_{h_0}] \approx \mathbf{Var}[Q_{h_L}]$, the cost ratio of MLMC to MC estimation is then approximately

$$\mathcal{C}_0/\mathcal{C}_L \eqsim \left(m^{-\gamma}\right)^L$$

- If $\mathcal{C}_\ell$ increases faster than $\mathbf{Var}[Y_\ell]$ decays, the cost on level $\ell = L$ dominates. Then the cost ratio is approximately

$$\mathbf{Var}[Y_L]/\,\mathbf{Var}[Q_{h_L}] \eqsim \varepsilon^2$$

(provided $\mathbb{E}\left[(Q - Q_L)^2\right] \eqsim (\mathbb{E}\left[Q - Q_L\right])^2$, which is problem dependent).

## General Multilevel Monte Carlo Complexity Theorem

### Theorem 5.2

*Let* $\varepsilon < \exp(-1)$ *and assume that there are constants* $\alpha, \beta, \gamma > 0$ *such that* $\alpha \geq \frac{1}{2}\min\{\beta, \gamma\}$ *and, for all* $\ell = 0, \ldots, L$,

*(M1)* $|\mathbb{E}[Q_{h_\ell}] - \mathbb{E}[Q]| = \mathcal{O}(h_\ell^\alpha)$,

*(M2)* $\mathbf{Var}[Y_\ell] = \mathcal{O}(h_\ell^\beta)$,

*(M3)* $\mathcal{C}_\ell = \mathcal{O}(h_\ell^{-\gamma})$.

*Then there are* $L$ *and* $\{N_\ell\}_{\ell=0}^L$ *such that* $\mathbb{E}\left[\left(\widehat{Q}^{ML}_{L,\{N_\ell\}} - \mathbb{E}[Q]\right)^2\right] \leq \varepsilon^2$ *and*

$$\mathrm{Cost}\left(\widehat{Q}^{ML}_{L,\{N_\ell\}}\right) = \begin{cases} \mathcal{O}\left(\varepsilon^{-2}\right), & \text{if } \beta > \gamma, \\ \mathcal{O}\left(\varepsilon^{-2}\,|\log\varepsilon|^2\right), & \text{if } \beta = \gamma, \\ \mathcal{O}\left(\varepsilon^{-2-(\gamma-\beta)/\alpha}\right), & \text{if } \beta < \gamma. \end{cases}$$

*Proof.* Demonstrated on tablet.

[Giles, 2007] for special case of SDEs with $\alpha = \gamma = 1$.

[Cliffe, Giles, **RS**, Teckentrup, 2011] for the general case. □



## Application to the Predator-Prey Problem

In the case of the predator-prey model problem we have already seen in (4.1) and (4.4) that (M1) and (M3) hold with $\alpha = 1$ and $\gamma = 1$, respectively.

Finally, it can be proved similarly to (M1) that (M2) holds with $\beta = 2$. Indeed

$$\begin{aligned} \mathbf{Var}[Y_\ell] = \mathbf{Var}[Q_{h_\ell} - Q_{h_{\ell-1}}] &\leq \mathbb{E}\left[\left(Q_{h_\ell} - Q_{h_{\ell-1}}\right)^2\right] \\ &\leq 2\left(\mathbb{E}\left[\left(Q - Q_{h_{\ell-1}}\right)^2\right] + \mathbb{E}\left[\left(Q - Q_{h_\ell}\right)^2\right]\right) \\ &\leq 2\left(K^2 h_{\ell-1}^2 + K^2 h_\ell^2\right) \\ &\leq \underbrace{2K^2(1+m^2)}_{\text{constant}}\, h_\ell^2 \,. \end{aligned}$$

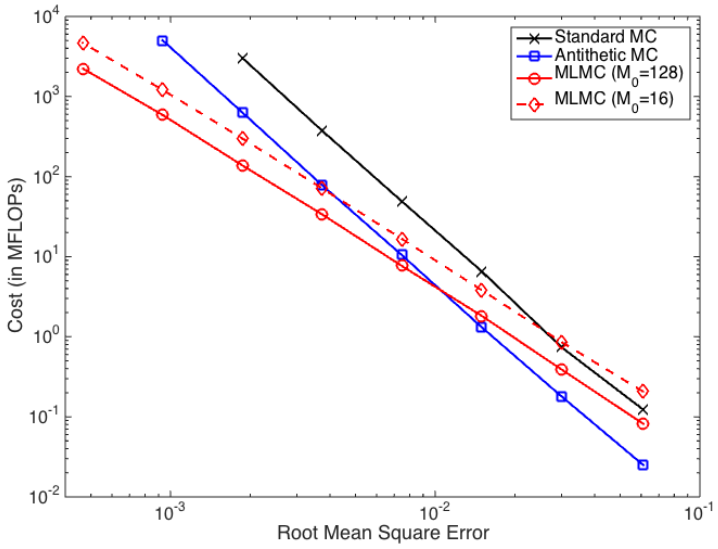
Thus, $\beta > \gamma$ and it follows from **Theorem 5.2** that

$$\mathrm{Cost}\left(\widehat{Q}^{\mathrm{ML}}_{L,\{N_\ell\}}\right) = \mathcal{O}\left(\varepsilon^{-2}\right).$$

Recall that for standard MC we had $\mathrm{Cost}(\widehat{Q}_{h,N}) = \mathcal{O}\left(\varepsilon^{-3}\right)$, so we gained a whole **order of magnitude**.

## Numerical Results – CPU time vs. Root Mean Square Error

Comparing standard MC & MLMC for the predator-prey model (and anithetic MC)




The cost of the standard MC method grows like $\mathcal{O}(\varepsilon^{-3})$, as predicted, while the cost for MLMC grows like $\mathcal{O}(\varepsilon^{-2})$ (actual cost depends on number of levels).

(We can also observe the variance reduction through antithetic sampling.)



## Adaptive MLMC Algorithm

- The following MLMC algorithm computes the optimal values of $L$ and $N_\ell$ adaptively using the sample averages $\widehat{Y}_{\ell,N_\ell}$ and sample variances

$$\mathrm{s}_\ell^2 := \frac{1}{N-1}\sum_{k=1}^{N_\ell}\left(Y_\ell^{(k)} - \widehat{Y}_{\ell,N_\ell}\right)^2 \quad \text{of} \quad Y_\ell\,.$$

- Sample variances can be used directly to estimate the MC error on each level.
- To bound the bias error, we assume there exists an $h^\star > 0$ such that the error decay in $|\mathbb{E}\left[Q_h - Q\right]|$ is monotonic for $h \le h^\star$ and satisfies

$$ch^\alpha \le |\mathbb{E}\left[Q_h - Q\right]| \le Ch^\alpha.$$

- This ensures that in the case $\frac{h_{\ell-1}}{h_\ell} = m$ (via inverse triangle inequality) DIY

$$|\mathbb{E}\left[Q_{h_\ell} - Q\right]| \le \tfrac{1}{rm^\alpha - 1}\widehat{Y}_\ell \qquad \text{for} \quad r = c/C.$$

- For the predator-prey problem $r = c/C \approx 1$ ($c$ can safely be chosen as $0.9$) and this gives a computable error estimator on level $L$ to determine whether $h_L$ is sufficiently small or whether $L$ needs to be increased.

# Adaptive MLMC Algorithm

## Adaptive MLMC Algorithm

1. Set $h_0$, $m$, $\varepsilon$, $L = 1$ and $N_0 = N_1 = N_{\text{Init}}$.
2. For $\ell = 0, \ldots, L$
   a. Compute new samples $Y_\ell^{(k)}$ on level $\ell$ until there are $N_\ell$.
   b. Compute $\widehat{Y}_\ell$ and $\mathrm{s}_\ell^2$, and estimate $\mathcal{C}_\ell$.
3. Update estimates for $N_\ell$ using the formula in (5.3) and if $\widehat{Y}_L > \frac{rm^\alpha - 1}{\sqrt{2}}\varepsilon$, increase $L \to L+1$ and set $N_L = N_{\text{Init}}$.
4. While $\widehat{Y}_L > \frac{rm^\alpha - 1}{\sqrt{2}}\varepsilon$ or $\sum_{\ell=0}^{L} s_\ell^2/N_\ell > \varepsilon^2/2$
   Return to 2.
5. Set $\widehat{Q}^{\text{ML}}_{L,\{N_\ell\}} = \sum_{\ell=0}^{L} \widehat{Y}_\ell$.