

Slides for
Numerical methods for random partial differential
equations: hierarchical approximation and
machine learning approaches,
RWTH¹

Prof. Raúl Tempone

May 10, 2023

¹Partially based on course materials developed by F. Nobile and R. Tempone .

Numerical methods for random partial differential equations: hierarchical approximation and machine learning approaches. Summer 2023

Fundamentals (basics on probability theory and statistics, conditional expectation, Bayesian statistics, finite difference, and finite element methods, random fields and Gaussian random fields), Monte Carlo sampling for random PDEs, Multilevel Monte Carlo, Quasi-Monte Carlo, Multi-Index Monte Carlo for random PDEs, Bayesian inverse problems for PDEs, L^2 projection and discrete L^2 regression, sparse grids, stochastic collocation, Multi-index stochastic collocation, stochastic optimization with PDE connections, Bayesian optimal experimental design, machine learning.

Prerequisite: Familiarity with probability and stochastic processes, partial differential equations, numerical methods.

Instructors: Raúl Tempone (tempone@uq.rwth-aachen.de), Chiheb Ben Hammouda (benhammouda@uq.rwth-aachen.de), Sophia Wiechert (wiechert@uq.rwth-aachen.de) and Shyam Mohan Subbiah Pillai (subbiah@uq.rwth-aachen.de)

Teaching Assistants: Jonas Niessen (jonas.niessen1@rwth-aachen.de), Jan Rodriguez (jan.rodriguez@rwth-aachen.de), Veli Ünlü (Veli_Unlu@gmx.de), Leon Wilkosz (leon.wilkosz@rwth-aachen.de)

Admin details

- **Syllabus and tentative course outline** (meeting times, office hours, etc) ([Keep looking at the updated course outline document!](#))
- Classes:
 - ▶ Tuesdays, Wednesdays and Thursdays from 14:30 to 16:00 hrs.
 - ▶ Exercise sessions: Fridays from 14:30 to 16:00 hrs.
 - ▶ Tutorial sessions: Mondays from 16:00 to 17:30 hrs.
 - ▶ Zoom Access:
<https://rwth.zoom.us/j/99753798324?pwd=a0RLcXZMwmVPUnh6eVBCZW43LzRPZz09>
Meeting ID: 997 5379 8324
Passcode: 989173
- Groups, email list, order of groups for assignments.

Grading

The grading consists of two parts:

- 1- Homework presentations are carried out by *groups of students.*
 - ▶ 6 Homeworks:
 - ▶ Each group should hand in a written solution, and the homework can be improved only once and within one week after its hand in date.
 - ▶ One group will be assigned to give a presentation of its solution to the class.
 - ▶ Each student in the presenting group must be able to explain any part of the solution to the class.

Grading:

- 2- There will be a closed book, classroom, final exam.
Tentative list of questions will be provided.
- Numerical course grades will be determined according to the formula:

$$\text{Total Score} = (60 \times (\text{Final Exam})) + 40 \times (\text{Average Homework}) / 100$$

Option: You can dig deeper into the material related to the course with hands-on experience within our seminars. They will provide you with a "final project" opportunity.

Admin details

Web access: Link to all materials will be shared via Moodle.

Recorded material: Make sure you read carefully the course outline so you study the necessary recorded materials **before** the corresponding lecture. This will maximize your learning experience, allowing you to identify possible gaps in your existing knowledge and to come with prepared questions.

Alexander von Humboldt UQ Chair (Pontdriesch 14-16, Gebäude-1953, 1.OG)

Main focus: *Developing efficient numerical methods for solving forward and inverse problems plus related optimal control problems, involving stochastic differential equations.*

Chair's website:

<https://www.uq.rwth-aachen.de>

Alexander von Humboldt UQ Chair (Pontdriesch 14-16, Gebäude-1953, 1.OG)

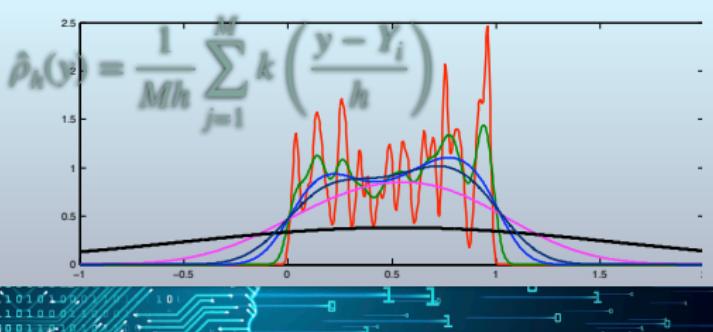
Some current real-world applications:

- ▶ Stochastic modeling (forecast and optimization) for power generation
- ▶ Stochastic forecasting for wear degradation
- ▶ Stochastic forecasting for fouling deposition
- ▶ Seismic source inversion
- ▶ Inference and model comparison for metallic fatigue
- ▶ Designing electrical impedance tomography experiments
- ▶ Computational finance for option pricing
- ▶ Crowd flow stochastic modeling

UQ Hybrid Seminar:

Recordings of several previous talks can be found on the Chair of Mathematics for UQ Youtube channel: <https://www.youtube.com/@chairofmathematicsforuq4149/featured>

Stochastic Numerics with applications in Simulation and Data Science



For whom?

This course is for CES, Mathematics, and Simulation Sciences master-students as well as everyone interested.

Content:

- Random variable generation
- Monte Carlo method: Error analysis
- Variance reduction techniques (antithetic variables, control variables, importance sampling)
- Large deviations and Rare events simulations
- Kernel density estimators
- Resampling techniques
- Simulation of stochastic processes: Gaussian fields and Kriging.
- Markov Chains
- Markov Chain Monte Carlo methods (Metropolis-Hastings, Gibbs sampler, Tempering)
- Bayesian Filters, Kalman Filters and data assimilation

Lecturers: Prof. Dr. Raul Tempone



Lectures: Tue 8:30 to 10:00,
Wed 10:30 - 12:00,
Thurs 10:30 - 12:00

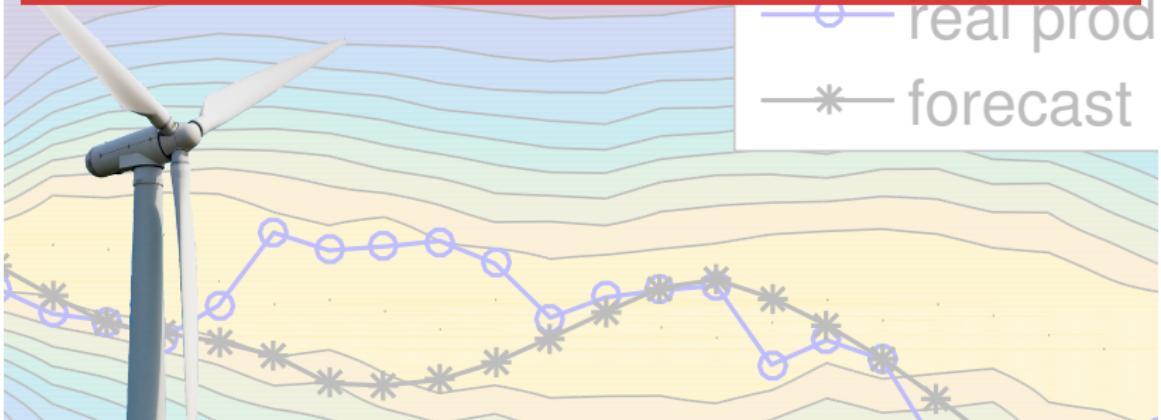
Übung: Fri, 10:30 - 12:00
Tutorials: Mon, 14:30 - 16:00

Where: online

Web: www.uq.rwth-aachen.de

Mathematics for Uncertainty Quantification

real prod
forecast



For whom?

This seminar is for CES, Mathematics, and Simulation Sciences master-students as well as everyone interested.

Content

Mathematical modeling and numerical simulation are central components of modern scientific research. A key challenge is to quantify uncertainty in model predictions. This seminar will explore research topics in the context of mathematical models and analysis for simulation techniques used in uncertainty quantification.

Lecturers:

Prof. Dr. Raúl Tempone



First Meeting:

Tue, 4 April, 2023
12:30–14:00

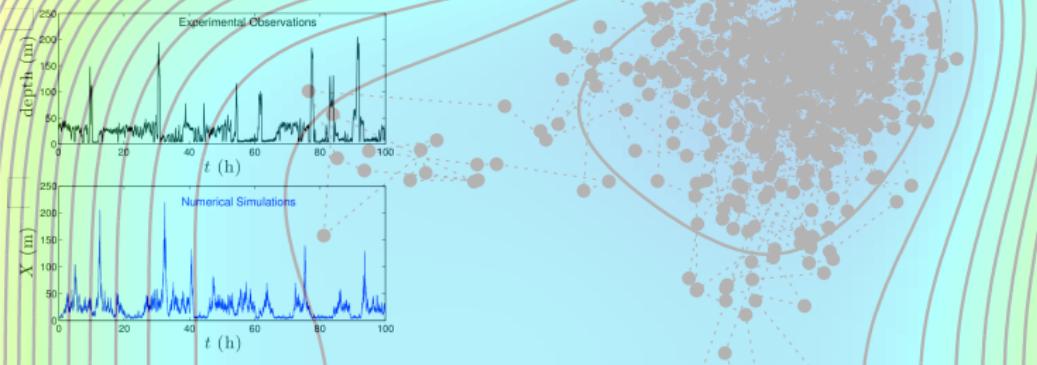
Where:

Online (see Moodle)

Web:

www.uq.rwth-aachen.de

Data Science under Uncertainty



For whom?

This seminar is for Data Science, Mathematics, Simulation Sciences, and CES master students as well as everyone interested.



Content

In this seminar we will cover data science applications subject to uncertainties. The focus will be on the mathematical and numerical analysis of stochastic tools used to treat these problems. For example, these tools include Markov chain Monte Carlo sampling methods, Data Assimilation techniques, optimal experimental design, model selection and validation, and statistical learning techniques such as clustering and support vector machines.

Lecturers: Prof. Dr. R. Tempone

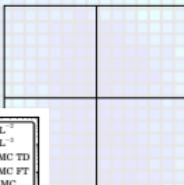
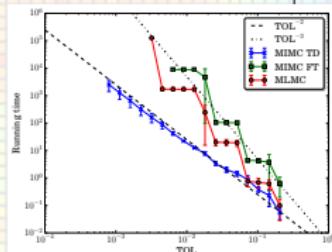


First Meeting: Thursday, April 6, 2023
12:30 – 14:00

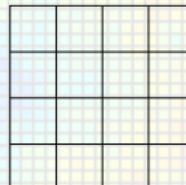
Where: Online (see Moodle)

Web: www.uq.rwth-aachen.de

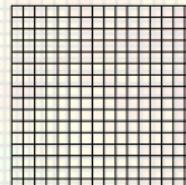
Multilevel Monte Carlo



h_0



h_I



h_L

$$\mathbf{E}[u(h_L)] = \mathbf{E}[u(h_0)] + \sum_{\ell=1}^L \mathbf{E}[u(h_\ell) - u(h_{\ell-1})]$$

For whom?

This seminar is for CES, Mathematics, and Simulation Sciences master-students as well as everyone interested.

Content

Computational efficiency of stochastic simulations is a key challenge in model-based scientific inquiry. This seminar will explore research topics for reducing computational cost in the context of Multilevel Monte Carlo sampling algorithms by addressing their construction and mathematical analysis and providing intuition on their appropriate use.

Lecturers: Prof. Dr. Raul Tempone



First Meeting: Wed 5 April
12:30 - 14:00

Where: Online (see Moodle)

Web: <https://www.uq.rwth-aachen.de>



Mathematics
for Uncertainty
Quantification

RWTH AACHEN
UNIVERSITY

The three seminars are easily integrated with this course!

Introduction

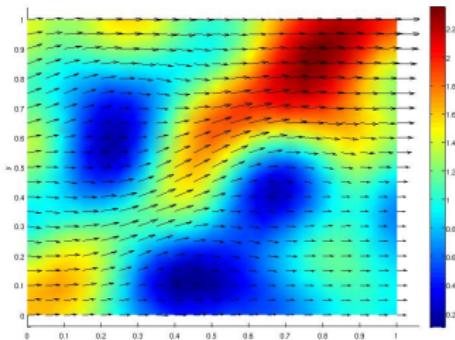
Mathematical models and computer simulations are widely used in engineering and science applications.

However, in many cases, the **parameters** in the model **are affected by uncertainty**, either because they are not perfectly known or because they are intrinsically variable.

Goal: devise effective ways to include and treat uncertainty in a mathematical model.

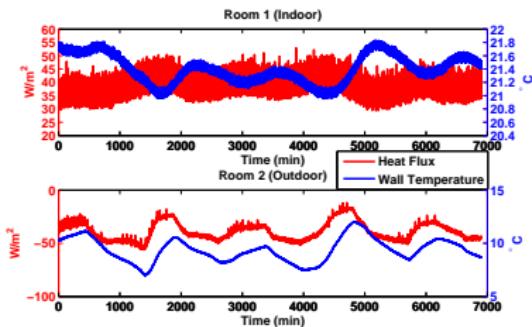
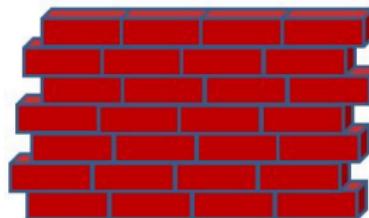
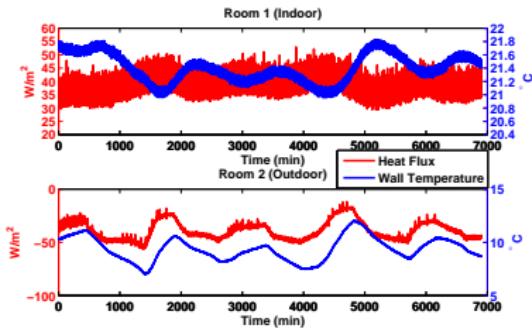
Examples

- ▶ Modeling of living tissues / biological fluids
- ▶ Subsurface modeling: groundwater flows, contaminant transport, earthquake simulations, ...
- ▶ Combustion problems / chemical reactions
- ▶ Weather forecast / climate modeling
- ▶ Molecular biology / protein dynamics / ...
- ▶ Finance

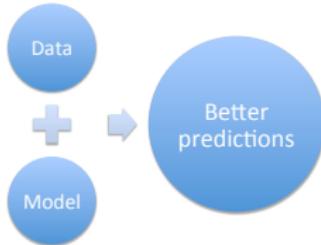


UQ Motivation

- ▶ Why should we care about Uncertainty Quantification?
- ▶ Because we want to improve our computational predictions and through them, better design/control/optimize our practical goals.
- ▶ The systematic ability to combine data with mathematical models is key to the step above.



$$\left\{ \begin{array}{l} \rho C \frac{\partial T}{\partial t} = - \frac{\partial}{\partial x} F, \\ T(0, t) = T_{int}(t), \quad t > 0 \\ T(x_0, t) = T_{ext}(t), \quad t > 0 \\ T(x, 0) = g(x), \quad x \in D, \\ F = - K \frac{\partial T}{\partial x}. \end{array} \right.$$



"Bayesian inferences of the thermal properties of a wall using temperature and heat flux measurements", by M. Iglesias, Z. Sawlan, M. Scavino, R. Tempone and C. Wood. International Journal of Heat and Mass Transfer 116, 417–43, 2018.

Why should UQ be based on rigorous mathematics?

- ▶ Mathematics helps us to describe our mathematical models and their relation with the data.
- ▶ Mathematics helps us to systematically build computable approximations to the models we use.
- ▶ Mathematics helps us to systematically combine data with computable approximations.
- ▶ Mathematics helps us to better design/control/optimize our goals.

Probabilistic approach

Probability theory provides an effective tool to describe and propagate uncertainty (although it is not the only one: we mention also *worst case scenario analysis, fuzzy logic, etc.*).

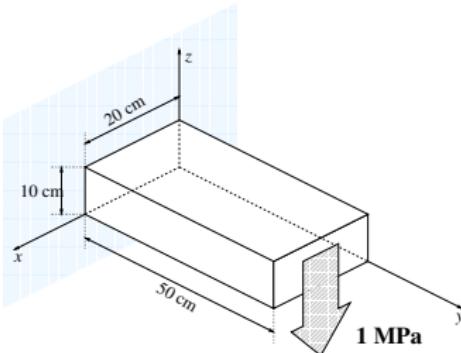
We will focus on mathematical models based on Partial Differential Equations (PDEs) whose input data (coefficients, forcing terms, boundary conditions, domain boundary,) are uncertain and described as random variables or random fields.

Therefore, the solution of the PDE is itself a random function,
 $u = u(\omega, x)$ (ω here denotes an elementary random event).

The main question we address in the course is how to effectively approximate the random function $u(\omega, x)$ or some (random) output Quantities of Interest $Q(u)$.

Linear elasticity with random elastic properties

Consider an elastic body, occupying the domain $D \subset \mathbb{R}^3$, with restricted displacement $\mathbf{u} = 0$ on a subset of its boundary, Σ_1 .



The (infinitesimal) displacement of the body $\mathbf{u} \in [H_{\Sigma_1}^1(D)]^3$ satisfies the equation

$$\int_D 2\mu \nabla_s \mathbf{u} : \nabla_s \mathbf{v} + \int_D \lambda \operatorname{div}(\mathbf{u}) \operatorname{div}(\mathbf{v}) = \int_{\Sigma_2} \mathbf{P} \cdot \mathbf{v}, \quad \forall \mathbf{v} \in [H_{\Sigma_1}^1(D)]^d$$

with $\nabla_s \mathbf{u} = \frac{\nabla \mathbf{u} + \nabla^T \mathbf{u}}{2}$, $\lambda = \frac{E\nu}{(1+\nu)(1-2\nu)}$, $\mu = \frac{E}{2(1+\nu)}$ (where E , ν are the Young modulus and Poisson ratio, resp.)

Assume in the previous problem that the Young modulus E , the Poisson ratio ν and the load vector $\mathbf{P} = (P_1, P_2, P_3)$ are uncertain and treated as random variables.

Vector of random parameters: $\mathbf{y} = (E, \nu, P_1, P_2, P_3)$, with $E > 0$ and $0 < \nu < \frac{1}{2}$.

Uncertainty model: \mathbf{y} takes values in $\Gamma \subset \mathbb{R}_+ \times (0, \frac{1}{2}) \times \mathbb{R}^3$ and has joint probability density function $\rho(\mathbf{y}) : \Gamma \rightarrow \mathbb{R}_+$.

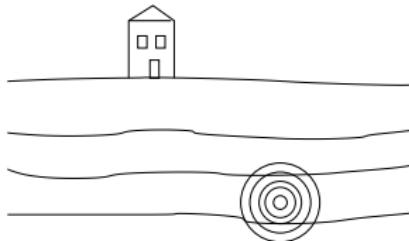
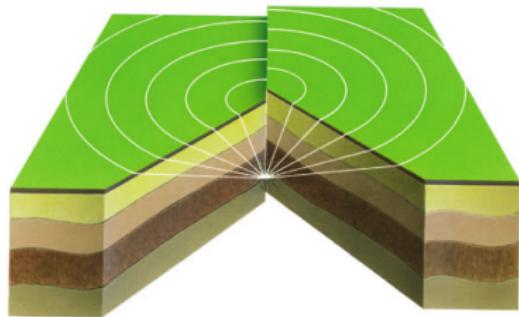
Then, for any outcome of the random vector $\mathbf{y} \in \Gamma$ there exists a unique solution $\mathbf{u} = \mathbf{u}(\mathbf{y}) \in V := [H_{\Sigma_1}^1(D)]^3$.

We can therefore define the random map $\mathbf{u}(\mathbf{y}) : \Gamma \rightarrow V$ that depends on $N = 5$ random parameters.

Seismic waves in random layered medium

Elastic waves in the ground can be well described by the linear elastodynamics equations

$$\begin{cases} \rho \frac{\partial^2 \mathbf{u}}{\partial t^2} - \operatorname{div} [2\mu \nabla_s \mathbf{u} + \lambda (\operatorname{div} \mathbf{u}) \mathbf{I}] = \mathbf{f}, & \text{in } D, \quad t \in (0, T] \\ + \text{ suitable initial and boundary conditions} \end{cases}$$



Typically, the medium is made of layers of different materials, whose mechanical properties ϱ , λ , μ are not perfectly known.

Vector of random parameters: $\mathbf{y} = (\varrho_1, \lambda_1, \mu_1, \dots, \varrho_N, \lambda_N, \mu_N)$ where $(\varrho_i, \lambda_i, \mu_i)$ are the density and Lame's constants in the i -th layer.

Uncertainty model: the parameters in two different layers are statistically independent and always positive. Joint probability density function:

$$\rho(\mathbf{y}) = \prod_{i=1}^N \rho_i(\varrho_i, \lambda_i, \mu_i), \quad \rho_i(\varrho_i, \lambda_i, \mu_i) : \mathbb{R}_+^3 \rightarrow \mathbb{R}_+$$

Other parameters could be uncertain as well, such as the position of the internal interfaces, the location and intensity of the source term, etc.

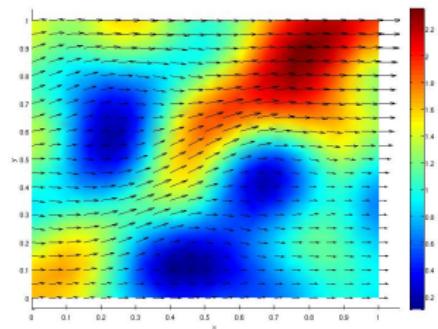
For any outcome of the random parameters \mathbf{y} , the problem admits a unique solution

$$\mathbf{u}(\mathbf{y}) \in V := L^2([0, T], \mathbf{H}^1(D)) \cap L^\infty([0, T], \mathbf{L}^2(D)).$$

We can therefore define the random map $\mathbf{y} \mapsto \mathbf{u}(\mathbf{y})$, $\mathbf{u} : \Gamma \rightarrow V$ that depends on $3N$ random parameters (N being the number of layers).

Groundwater flow in a random heterogeneous porous medium

According to Darcy's law, the pressure gradient ∇p and the fluid velocity \mathbf{u} in a porous medium follow a linear relation, that is



$$\begin{cases} \mathbf{u} = -k \nabla p & \text{in } D \\ \operatorname{div} \mathbf{u} = f & \\ + \text{boundary condns.} & \text{on } \partial D \end{cases}$$

The second equation (mass conservation) relates sinks and sources of flow to the velocity field.

In most aquifers, the macroscopic properties (porosity and permeability) of the ground are highly variable and never perfectly known.

They can be described as **random fields**, i.e. the permeability $k(x_i) > 0$ in each point $x_i \in D$ is a random variable and, taken n points x_1, \dots, x_n , the random variables $k(x_1), \dots, k(x_n)$ are in general correlated.

To guarantee positivity of the permeability field one often writes $k = e^{\mathcal{Y}}$.

Random parameters: log-permeability $\mathcal{Y}(x) = \log(k(x))$. It is actually a random field (collection of ∞ random variables)

Uncertainty model: given by the probability law of the random field.

How can we treat in practice the case of random fields?

We will see that any random field on a compact domain $D \subset \mathbb{R}^d$ with bounded variance can be expanded in series (e.g. Karhunen-Loève, Fourier, ...)

$$\mathcal{Y}(\omega, x) = \mathbb{E}[\mathcal{Y}](x) + \sum_{i=1}^{\infty} y_i(\omega) b_i(x)$$

where $\mathbf{y} = (y_1, y_2, \dots)$ is an infinite (countable) sequence of random variables. By suitable expansion one can make them uncorrelated and, sometimes, even independent.

For each outcome of the random sequence $\mathbf{y} \in \mathbb{R}^{\mathbb{N}}$, and hence each realization of the random field $\mathcal{Y}(x)$, the problem admits a unique solution $(p(\mathbf{y}), \mathbf{u}(\mathbf{y})) \in V := H^1(D) \times H(\text{div}, D)$

We can therefore define the random map $(p(\mathbf{y}), \mathbf{u}(\mathbf{y})) : \Gamma \subset \mathbb{R}^{\mathbb{N}} \rightarrow V$ that depends on an infinite countable number of random variables.

Uncertainty Quantification in Option Pricing

The Black-Scholes model for the value $f : (0, T) \times (0, \infty) \rightarrow \mathbb{R}$ of a European call option is the following linear parabolic partial differential equation

$$\begin{cases} \frac{\partial f}{\partial t} + rs \frac{\partial f}{\partial s} + \frac{\sigma^2 s^2}{2} \frac{\partial^2 f}{\partial s^2} = rf, & 0 < t < T, \\ f(s, T) = \max(s - K, 0), \end{cases} \quad (1)$$

where the constants r and σ denote the riskless interest rate and the volatility, respectively.

The volatility is typically estimated from either quoted option prices in the market or history matching from the evolution of the underlying and is often uncertain. Parametric uncertainty, in the context of derivative pricing, results in mis-pricing of contingent claims.

Goal: Quantify the impact of volatility uncertainty on option pricing under Black-Scholes framework.

Remark: The same mathematical problem appears in the modeling of flows in porous media with uncertain permeability.

Uncertainty Quantification in compressible aerodynamics

1. Uncertainties in **Operating Conditions**

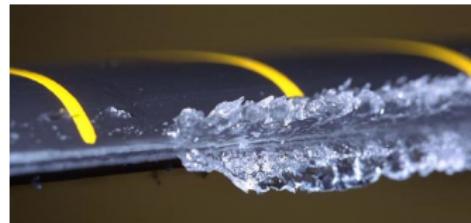
resulting e.g. from **atmospheric turbulence** during a flight.

- ▶ Angle of Attack
- ▶ Velocity (Mach)
- ▶ Density, Temperature



2. Uncertainties in **Geometry** due to manufacturing tolerances, icing, fatigue of the material

- ▶ Thickness
- ▶ Curvature
- ▶ Leading Edge radius

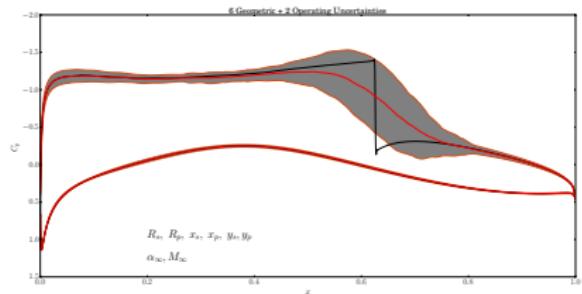
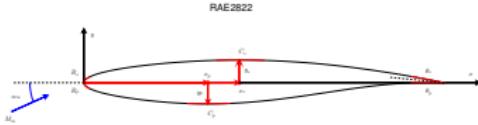


The fluid dynamics around an airfoil can be described by the Euler equations

$$\frac{\partial W}{\partial t} + \frac{\partial F_1(W)}{\partial x} + \frac{\partial F_2(W)}{\partial y} = 0$$

with **state vector** $W = [\rho, \rho u_1, \rho u_2, \rho e]$ representing the **density**, **momentum** and **energy** of the fluid.

If **y** denotes the **vector of unknown parameters** (**operational**: angle of attack, Mach number, inflow density and temperature; **geometrical**: thickness, curvature, etc.), treated as random variables, then, $W = W(y)$ is a **random solution**.



Probability Background

A probability space is a triple (Ω, \mathcal{F}, P) , where Ω is the set of outcomes, \mathcal{F} is the set of events and $P : \mathcal{F} \rightarrow [0, 1]$ is a set function that assigns probabilities to events satisfying certain rules.

Definition 2.1 (Measurable Space)

If Ω is a given non empty set, then a σ -algebra \mathcal{F} on Ω is a collection \mathcal{F} of subsets of Ω such that:

- (1) $\Omega \in \mathcal{F}$;
- (2) $F \in \mathcal{F} \Rightarrow F^c \in \mathcal{F}$, where $F^c = \Omega - F$ is the complement set of F in Ω ; and
- (3) $F_1, F_2, \dots \in \mathcal{F} \Rightarrow \bigcup_{i=1}^{+\infty} F_i \in \mathcal{F}$.

Exercise 2.1

Let \mathcal{F} be a σ -algebra. Prove that if $F_1, F_2, \dots \in \mathcal{F}$ then $\bigcap_{i=1}^{+\infty} F_i \in \mathcal{F}$.

Definition 2.2 (Probability Measure)

A probability measure on (Ω, \mathcal{F}) is a set function $P : \mathcal{F} \rightarrow [0, 1]$ such that:

- (1) $P(\emptyset) = 0$, $P(\Omega) = 1$; and
- (2) If $A_1, A_2, \dots \in \mathcal{F}$ are mutually disjoint sets then

$$P\left(\bigcup_{i=1}^{+\infty} A_i\right) = \sum_{i=1}^{+\infty} P(A_i).$$

Question 1: Give an example of a probability space and distinguish clearly the events $F \in \mathcal{F}$ from the outcomes $\omega \in \Omega$.

Question 2: Give an example of two different σ -algebras, $\mathcal{G} \subset \mathcal{F}$ for the same set of outcomes Ω . Can you give an intuitive interpretation of the relation $\mathcal{G} \subset \mathcal{F}$?

Question 3: Is the intersection of σ -algebras still a σ -algebra?

Question 4: What about the union of σ -algebras?

Definition 2.3 (generated σ -algebra)

Given a family of sets, $\{A_n\}$, there exists a unique σ -algebra, $\sigma(\{A_n\})$, s.t.

1. $\{A_n\} \subset \sigma(\{A_n\})$,
2. if \mathcal{F} is a σ -algebra,

$$\{A_n\} \subset \mathcal{F} \Rightarrow \sigma(\{A_n\}) \subset \mathcal{F}.$$

Definition 2.4 (Conditional Probability)

The conditional probability of A , given B with $P(B) > 0$, written as $P(A|B)$, is defined to be

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Observe that $P(\cdot|B)$ is a probability on the new sample space B . $P(A|B)$ is interpreted as the likelihood / probability that A occurs given knowledge that B has occurred.

Exercise 2.2

Can you identify the triplet (Ω, \mathcal{F}, P) for this new space?

Notation: For readability, sometimes we may write $P(A; B)$ instead of $P(A|B)$.

Definition 2.5 (Independence)

Two events $A, B \in \mathcal{F}$ are said to be independent if

$$P(A \cap B) = P(A)P(B).$$

Can you give intuitive meaning to the statement

$$P(A|B) > P(A)?$$

Proposition 2.1 (Law of total probability)

If B_1, B_2, \dots is sequence of events that is finite or countable infinite partition of Ω , i.e. $\bigcup_n B_n = \Omega$ and $B_n \cap B_m = \emptyset$ for $n \neq m$. Then, for any event A

$$\begin{aligned} P(A) &= \sum_n P(A \cap B_n) \\ &= \sum_n P(A|B_n)P(B_n) \end{aligned}$$

Definition 2.6

A random variable X , on the probability space (Ω, \mathcal{F}, P) , is a function

$$X : \Omega \rightarrow \mathbb{R}^d,$$

such that the inverse image

$$X^{-1}(A) \equiv \{\omega \in \Omega : X(\omega) \in A\} \in \mathcal{F},$$

for all open subsets A of \mathbb{R}^d .

Equivalently, we may say that X is an \mathcal{F} -measurable function and write $X \in \mathcal{F}$.



Shooting arrows, a motivating example

Observe:

Consider a \mathcal{F} -random variable X . If $\mathcal{F} \subset \mathcal{G}$ then X is also a \mathcal{G} -random variable.

Can you give an intuitive meaning to this?

Definition 2.7 (Discrete random variables)

A discrete random variable (r.v.) X takes on values in $S = \{x_1, x_2, \dots\}$, its probability mass function is defined by:

$$P_X(x_i) = P(X = x_i), i = 1, 2, \dots$$

Given a collection X_1, X_2, \dots, X_n of S -valued rvs, its joint probability mass function (pmf) is defined as

$$\begin{aligned} P_{(X_1, X_2, \dots, X_n)}(x_1, x_2, \dots, x_n) \\ = P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n). \end{aligned}$$

Example 2.8 (Discrete random variables)

Some classical examples:

1. *Bernoulli*(p), $X \sim Ber(p)$ if $X \in \{0, 1\}$, $0 \leq p \leq 1$, and

$$P(X = 1) = p = 1 - P(X = 0).$$

2. *Binomial*(n, p)

$X \sim Bin(n, p)$ if $X \in \{0, 1, \dots, n\}$, $0 \leq p \leq 1$, and

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}.$$

3. *Poisson*(λ),

$X \sim Poisson(\lambda)$ if $X \in \{0, 1, 2, \dots\}$, $0 \leq \lambda < \infty$, and

$$P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}.$$

Interesting fact:

$$\text{Binomial}(n, p) \rightarrow \text{Poisson}(\lambda)$$

as $n \rightarrow \infty$, $np \rightarrow \lambda$.

Definition 2.9 (Continuous random variables)

A random variable X taking values in \mathbb{R} is continuous with probability density function $f_X(\cdot)$ if for all $x \in \mathbb{R}$ we have

$$P(X \leq x) = \int_{-\infty}^x f_X(t)dt.$$

Example 2.10 (Continuous random variables)

Some classical examples:

- $\text{Uniform}(a, b)$, $X \sim U(a, b)$ if $X \in [a, b]$, and

$$P(X \leq x) = \frac{x-a}{b-a}, \quad a \leq x \leq b.$$

- $\text{Beta}(\alpha, \beta)$, if $X \in [A, B]$, $\alpha > 0, \beta > 0$, and

$$f_X(x) = \begin{cases} \frac{1}{B-A} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \left(\frac{x-A}{B-A}\right)^{\alpha-1} \left(\frac{B-x}{B-A}\right)^{\beta-1} & \text{if } A \leq x \leq B, \\ 0 & \text{otherwise} \end{cases}$$

where the **gamma function** $\Gamma(\cdot)$ is defined by

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx, \quad \alpha > 0,$$

$\Gamma(n) = (n-1)!$ for every positive integer n .

The **beta function** $B(\cdot, \cdot)$ is defined by

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1} (1-x)^{\beta-1} dx, \quad \alpha > 0, \beta > 0,$$

and is related to the gamma function through the identity

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}.$$

- ▶ *Gaussian*(μ, σ^2), $X \sim N(\mu, \sigma^2)$ if $X \in \mathbb{R}$, $\mu \in \mathbb{R}$, $\sigma > 0$, and

$$P(X \in B) = \int_B \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} dx.$$

- ▶ *Exponential*(λ), $X \sim Exp(\lambda)$ if $X \in (0, +\infty)$, $\lambda > 0$, and

$$P(0 \leq X \leq x) = \lambda \int_0^x e^{-\lambda x} dx.$$

- ▶ *Gamma*(α, β), if $X \in (0, +\infty)$, $\alpha > 0$, $\beta > 0$, and

$$f_X(x) = \begin{cases} \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\beta}} & \text{if } x > 0, \\ 0 & \text{otherwise.} \end{cases}$$

The probability density function (pdf) of $X \sim Exp(\lambda)$ is a special case of the pdf of $X \sim Gamma(\alpha, \beta)$ in which $\alpha = 1$ and $\beta = \frac{1}{\lambda}$.

Example 2.11

Consider a finite family of disjoint sets,

$$\{A_n\}_{n=1}^N$$

and let $\Omega \equiv \cup_{1 \leq n \leq N} A_n$, $\mathcal{F} \equiv \sigma(\{A_n\})$. What condition does

$$X : \Omega \rightarrow \mathbb{R}$$

have to satisfy in order to be a random variable in (Ω, \mathcal{F}) ?

With the help of index functions,

$$1_{A_n}(\omega) = \begin{cases} 1, & \text{if } \omega \in A_n \\ 0, & \text{otherwise,} \end{cases}$$

we can write X from the previous example as a staircase (piecewise constant) function:

$$X(\omega) = \sum_n 1_{A_n}(\omega) X_n.$$

Here $X_n \in \mathbb{R}$, $n = 1, \dots$

Observe: Given $X \in \mathcal{F}$ random variable, the minimal σ -algebra that still makes X a random variable is $\sigma(X) = \sigma(X^{-1}(B) : B \text{ open in } \mathbb{R})$.

Definition 2.12 (cdf)

Given a probability measure P , the cumulative (probability) distribution function of a random variable X , $F_X : \mathbb{R} \rightarrow [0, 1]$ is

$$F_X(x) \equiv P(X \leq x).$$

For continuous random variables we have

$$F'_X(x) = f_X(x).$$

Exercise 2.3

Graph the cdf of $X \sim U(0, 1)$.

Definition 2.13 (Independence)

Two random variables X, Y in \mathbb{R}^d are independent if for all open sets $A, B \subseteq \mathbb{R}^d$ we have that the events

$$X^{-1}(A) \text{ and } Y^{-1}(B) \text{ are independent .} \quad (2)$$

Definition 2.14 (Distribution)

Every random variable in (Ω, \mathcal{F}, P) induces a probability measure μ_X on \mathbb{R}^n , defined by

$$\mu_X(B) = P(X^{-1}(B)), \text{ for each open } B \subset \mathbb{R}^n.$$

μ_X is called the distribution of X and we can also think of the probability space

$$(\mathbb{R}^n, \mathcal{B}, \mu_X)$$

with

$$\mathcal{B} = \sigma(\{B \subset \mathbb{R}^n : B \text{ open}\}).$$

Definition 2.15 (Expectation)

Let $X : \Omega \rightarrow \mathbb{R}$ be a random variable and suppose that the density function

$$p'(x) = \frac{P(X \in dx)}{dx}$$

is integrable. The expectation of X is then defined by the integral

$$E[X] = \int_{-\infty}^{\infty} xp'(x)dx, \tag{3}$$

which also can be written as

$$E[X] = \int_{-\infty}^{\infty} xdp(x). \tag{4}$$

Obs:

The last integral,

$$E[X] = \int_{-\infty}^{\infty} x dp(x)$$

makes sense also in general when the density function is a measure (e.g. discrete r.v.), e.g. by successive approximation with random variables possessing integrable densities. A point mass, i.e. a Dirac delta measure, is an example of a measure.

Definition 2.16 (Variance)

$$\text{Var}[X] = E[(X - E[X])^2] = E[X^2] - (E[X])^2.$$

Proposition 2.2 (Law of total Expectation)

If B_1, B_2, \dots is sequence of events that is finite or countable infinite partition of Ω , i.e. $\bigcup_n B_n = \Omega$ and $B_n \cap B_m = \emptyset$ for $n \neq m$. Then, we have

$$E[X] = \sum_n E[X|B_n]P(B_n)$$

Exercise 2.4 (Examples of means and variances)

Verify the following:

1. $X \sim Ber(p)$: $E[X] = p$, $Var[X] = p(1 - p)$
2. $X \sim Bin(n, p)$: $E[X] = np$, $Var[X] = np(1 - p)$
3. $X \sim Poisson(\lambda)$: $E[X] = Var[X] = \lambda$.

Exercise 2.5 (Examples of means and variances)

Verify the following:

1. $X \sim Uni(a, b)$: $E[X] = \frac{a+b}{2}$, $Var[X] = \frac{(b-a)^2}{12}$
2. $X \sim Beta(\alpha, \beta)$: $E[X] = A + (B - A) \frac{\alpha}{\alpha+\beta}$, $Var[X] = \frac{(B-A)^2 \alpha \beta}{(\alpha+\beta)^2 (\alpha+\beta+1)}$
3. $X \sim N(\mu, \sigma^2)$: $E[X] = \mu$, $Var[X] = \sigma^2$
4. $X \sim Exp(\lambda)$: $E[X] = \frac{1}{\lambda}$, $Var[X] = \frac{1}{\lambda^2}$
5. $X \sim Gamma(\alpha, \beta)$: $E[X] = \alpha\beta$, $Var[X] = \alpha\beta^2$.

Example 2.17 (Law of total probability in the cont. case)

Let A an event and X a random variable with PDF $f(\cdot)$, then

$$P(A) = \int_{\mathbb{R}} P(A|X = x)f_X(x)dx$$

Remark 2.1

Let $Y = g(X)$, then

$$E[g(X)] = \int g(x)f_X(x)dx.$$

A similar formula holds for discrete random variables.

Exercise 2.6

Verify that the formula above holds for the case where

$$Y(\omega) = \sum_n y_n 1_{A_n}(\omega)$$

Vector valued discrete random variables

Given a collection X_1, X_2, \dots, X_n of S -valued rvs, its joint probability mass function (pmf) is

$$\begin{aligned} P_{(X_1, X_2, \dots, X_n)}(x_1, x_2, \dots, x_n) \\ = P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n). \end{aligned}$$

The *conditional pmf* of X , given $Y = y$, is then

$$P_{X|Y}(x|y) = \frac{P_{(X,Y)}(x,y)}{P_Y(y)}.$$

The collection of rvs X_1, X_2, \dots, X_n are independent if

$$\begin{aligned} P_{(X_1, X_2, \dots, X_n)}(x_1, x_2, \dots, x_n) \\ = P_{X_1}(x_1)P_{X_2}(x_2)\dots P_{X_n}(x_n), \text{ for all } (x_1, \dots, x_n) \in S^n. \end{aligned}$$

Vector valued continuous random variables

Given a collection X_1, X_2, \dots, X_n of real-valued continuous rvs its joint probability density function (pdf) is defined as the function $f_{(X_1, X_2, \dots, X_n)}(\cdot)$ satisfying

$$\begin{aligned} & P(X_1 \leq x_1, \dots, X_n \leq x_n) \\ &= \int_{-\infty}^{x_1} \dots \int_{-\infty}^{x_n} f_{(X_1, X_2, \dots, X_n)}(t_1, t_2, \dots, t_n) dt_1 \dots dt_n. \end{aligned}$$

The *marginal pdf of X_i* is the pdf of X_i alone, that is, for $i = 1$ we have

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f_{(X_1, X_2, \dots, X_n)}(x_1, t_2, \dots, t_n) dt_2 \dots dt_n.$$

The collection X_1, \dots, X_n is independent if

$$f_{(X_1, X_2, \dots, X_n)}(x_1, x_2, \dots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n)$$

for all $(x_1, \dots, x_n) \in \mathbb{R}^n$.

Finally, the *conditional pdf* of X , given $Y = y$, has the form

$$f_{X|Y}(x|y) = \frac{f_{(X,Y)}(x,y)}{f_Y(y)}.$$

Observe: $f_{X|Y}(\cdot|y)$ represents the randomness left in X once we observe Y .

Example 2.18 (Multivariate Gaussian distribution)

Let $\mathbf{X} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\mathbf{X}(\omega) \in \mathbb{R}^n$. Let $\boldsymbol{\mu} \in \mathbb{R}^n$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$, $\boldsymbol{\Sigma}$ symmetric, positive definite. The pdf of \mathbf{X} is

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{n/2} \sqrt{|\boldsymbol{\Sigma}|}} \exp\left(-\frac{(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})}{2}\right), \quad \mathbf{x} \in \mathbb{R}^n.$$

Observe: If $\boldsymbol{\Sigma}$ is not positive definite, then the $N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution has no density with respect to the Lebesgue measure on \mathbb{R}^n .

We have $E[\mathbf{X}] = \boldsymbol{\mu}$ (mean vector), and

$$E[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T] = \boldsymbol{\Sigma} \text{ (covariance matrix).}$$

Exercise 2.7 (Conditional Gaussian density)

Let $\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}_{n \times 1}$ be Gaussian with mean vector $\boldsymbol{\mu} = [\boldsymbol{\mu}_1^T, \boldsymbol{\mu}_2^T]^T_{n \times 1}$

and covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{12}^T & \boldsymbol{\Sigma}_{22} \end{bmatrix}.$$

What is the conditional distribution of \mathbf{X}_1 , given $\mathbf{X}_2 = \mathbf{x}_2$? Identify the conditional density function of \mathbf{X}_1 , given $\mathbf{X}_2 = \mathbf{x}_2$,

$$f_{\mathbf{X}_1|\mathbf{X}_2}(\mathbf{x}_1, \mathbf{x}_2).$$

Two steps:

1. Recognize that the conditional distribution of \mathbf{X}_1 , given $\mathbf{X}_2 = \mathbf{x}_2$, is Gaussian.
2. “Complete the square” to find the *conditional mean vector* and the *conditional covariance matrix* of \mathbf{X}_1 , given $\mathbf{X}_2 = \mathbf{x}_2$.

We obtain that the conditional mean vector of \mathbf{X}_1 , given $\mathbf{X}_2 = \mathbf{x}_2$, is

$$\begin{aligned} E[\mathbf{X}_1 | \mathbf{X}_2 = \mathbf{x}_2] &= E[\mathbf{X}_1] + Cov[\mathbf{X}_1, \mathbf{X}_2](Cov[\mathbf{X}_2, \mathbf{X}_2])^{-1}(\mathbf{x}_2 - E[\mathbf{X}_2]) \\ &= \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2), \end{aligned}$$

and that the conditional covariance matrix of \mathbf{X}_1 , given $\mathbf{X}_2 = \mathbf{x}_2$, is

$$\begin{aligned} Cov[\mathbf{X}_1 | \mathbf{X}_2 = \mathbf{x}_2] &= Cov[\mathbf{X}_1, \mathbf{X}_1] - Cov[\mathbf{X}_1, \mathbf{X}_2](Cov[\mathbf{X}_2, \mathbf{X}_2])^{-1}Cov[\mathbf{X}_2, \mathbf{X}_1] \\ &= \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}. \end{aligned}$$

Transformation of random variables: Let $\mathbf{X} = (X_1, X_2, \dots, X_d)'$ a random vector with joint PDF $f_{(X_1, X_2, \dots, X_d)}$. Let $\mathbf{Y} = \phi(\mathbf{X})$ where ϕ is an invertible function from \mathbb{R}^d to \mathbb{R}^d . Then, the joint PDF of the random vector $\mathbf{Y} = (Y_1, Y_2, \dots, Y_d)'$ is given by

$$f_{(Y_1, Y_2, \dots, Y_d)}(y_1, y_2, \dots, y_d) = \frac{f_{(X_1, X_2, \dots, X_d)}(x_1, x_2, \dots, x_d)}{|det(J_\phi)|(x_1, x_2, \dots, x_d)}$$

where $y = \phi(x)$, J_ϕ is the Jacobian matrix defined as $(J_\phi)_{ij} = \frac{\partial \phi_i}{\partial x_j}$ and $det(\cdot)$ denotes the determinant operator.

Exercise 2.8

Let X_1 and X_2 two i.i.d random variables. What is the PDF of $Y = X_1 X_2$?

Isserlis' theorem or Wick's theorem is a formula that allows one to compute higher-order moments of the multivariate normal distribution in terms of its covariance matrix.

Theorem 2.1 (Isserlis' theorem)

If (X_1, \dots, X_{2n}) is a zero-mean multivariate normal random vector, then

$$E[X_1 X_2 \cdots X_{2n}] = \sum \prod E[X_i X_j] = \sum \prod \text{Cov}(X_i, X_j),$$

$$E[X_1 X_2 \cdots X_{2n-1}] = 0,$$

where the notation $\sum \prod$ means summing over all distinct ways of partitioning X_1, \dots, X_{2n} into pairs X_i, X_j and each summand is the product of the n pairs.

Example 2.19

If (X_1, \dots, X_4) is a zero-mean multivariate normal random vector, then

$$E[X_1 X_2 X_3 X_4] =$$

$$E[X_1 X_2] E[X_3 X_4] + E[X_1 X_3] E[X_2 X_4] + E[X_1 X_4] E[X_2 X_3].$$

Exercise 2.9 (Sums of random variables)

Show that if X_1 and X_2 are continuous r.v. then

$$f_{X_1+X_2}(s) = (f_{X_1} * f_{X_2|X_1})(s) = \int_{-\infty}^{+\infty} f_{X_1}(x_1) f_{X_2|X_1}(s - x_1) dx_1.$$

Exercise 2.10 (Summing continuous and discrete r. vars.)

Let $X_1 \sim Ber(p)$ and $X_2 \sim U(0, 1)$. Assuming independence, graph the cdf of $X = X_1 + X_2$.

Exercise 2.11 (Variance of the sum)

Show that

$$\text{Var}[X_1 + X_2] = \text{Var}[X_1] + \text{Var}[X_2] + 2\text{Cov}[X_1, X_2]$$

with the covariance of X_1, X_2 defined as

$$\text{Cov}[X_1, X_2] = E[(X_1 - E[X_1])(X_2 - E[X_2])].$$

What happens if X_1 and X_2 are independent?

Lemma 2.20 (Borel-Cantelli)

Let E_1, E_2, \dots be a sequence of events in some probability space. If

$$\sum_{n=1}^{\infty} P(E_n) < \infty$$

Then the probability that infinitely many of them occur is 0, that is

$$P\left(\limsup_{n \rightarrow \infty} E_n\right) = 0$$

with

$$\limsup_{n \rightarrow \infty} E_n = \bigcap_{n=1}^{\infty} \bigcup_{k \geq n}^{\infty} E_k$$

Lemma 2.21 (Borel-Cantelli)

Suppose that E_1, E_2, \dots is a sequence of independent events in a probability space. If

$$\sum_n^{\infty} P(E_n) = \infty$$

then

$$P\left(\limsup_{n \rightarrow \infty} E_n\right) = 1$$

Exercise 2.12

Discuss the application of Lemma 1 and Lemma 2 when X_1, X_2, \dots is a sequence of independent random variables that take the values 0 or 1 with $P(X_n = 1) = p_n$ and $E_n = \{X_n = 1\}$.

Theorem 2.2 (Kolmogorov's two-series theorem)

Let X_1, X_2, \dots be independent random variables with expected values μ_n and variances σ_n^2 , such that $\sum_{n=1}^{\infty} \mu_n$ and $\sum_{n=1}^{\infty} \sigma_n^2$ converge in \mathbb{R} . Then $\sum_{n=1}^{\infty} X_n$ converges in \mathbb{R} almost surely.

Exercise 2.13

Let X_1, X_2, \dots be a sequence of independent random variables with mean zero and unit variance. Let $\epsilon > 0$, does $\sum_{n=1}^{\infty} \frac{X_n}{n^{1/2+\epsilon}}$ converge?

Theorem 2.3 (Lebesgue's Dominated Convergence)

Let f_n be a sequence of complex-valued measurable functions on a measure space (S, Σ, μ) . Suppose that the sequence converges pointwise to a function f and is dominated by some integrable function g

$$|f_n(x)| \leq g(x)$$

for all points $x \in S$. Then f is integrable and

$$\lim_{n \rightarrow \infty} \int_S f_n d\mu = \int_S \lim_{n \rightarrow \infty} f_n d\mu = \int_S f d\mu$$

Theorem 2.4 (Lebesgue's Monotone Convergence Theorem)

Let f_n be a sequence of nonnegative Lebesgue measurable functions defined on a Lebesgue measurable set E . Suppose that

$$0 \leq f_1(x) \leq f_2(x) \leq \cdots \leq f_n(x)$$

for all $x \in E$ and f_n convergence pointwise to f on E , then

$$\lim_{n \rightarrow \infty} \int_E f_n d\mu = \int_E \lim_{n \rightarrow \infty} f_n d\mu = \int_E f d\mu$$

Exercise 2.14 (Undominated but the result holds)

Consider $Y_n(\omega) = n\mathbf{1}_{(1/n+1, 1/n]}$ in the space (Ω, \mathcal{F}, P) with $\Omega = (0, 1]$.

Exercise 2.15

Let X_1, X_2, \dots is a sequence of non-negative random variables. Show that

$$E \left[\sum_{n=1}^{\infty} X_n \right] = \sum_{n=1}^{\infty} E [X_n]$$

Theorem 2.5 (Fubini)

Let X_1, X_2, \dots be a sequence of random variables for which $E [\sum_{n=1}^{\infty} |X_n|] = \sum_{n=1}^{\infty} E [|X_n|] < \infty$ then

$$E \left[\sum_{n=1}^{\infty} X_n \right] = \sum_{n=1}^{\infty} E [X_n]$$

Theorem 2.6 (Fubini)

If $\{X(t), t \geq 0\}$ is a process for which

$$E \left[\int_0^{\infty} |X(t)| dt \right] = \int_0^{\infty} E [|X(t)|] dt < \infty \text{ then}$$

$$E \left[\int_0^{\infty} X(t) dt \right] = \int_0^{\infty} E [X(t)] dt$$

Lemma 2.22 (Jensen's inequality)

Let $\phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a convex function:

Let $\lambda_1, \lambda_2, \dots, \lambda_N$ such that $0 \leq \lambda_i \leq 1, 1 \leq i \leq N$, and $\sum_{i=1}^N \lambda_i = 1$, then

$$\phi\left(\sum_{i=1}^N \lambda_i x_i\right) \leq \sum_{i=1}^N \lambda_i \phi(x_i).$$

Let (Ω, \mathcal{F}, P) be a probability space, and g a P -integrable real valued function, then

$$\phi\left(\int_{\Omega} g dP\right) \leq \int_{\Omega} \phi \circ g \, dP$$

Lemma 2.23 (Hölder's inequality)

Let (S, Γ, μ) be a measure space and let $p, q \in [1, \infty)$ with $1/p + 1/q = 1$. Then, for all measurable real- or complex-valued functions f and g on S ,

$$\int_S |fg| d\mu \leq \left(\int_S |f|^p d\mu \right)^{1/p} \left(\int_S |g|^q d\mu \right)^{1/q}$$

Lemma 2.24 (Young's inequality)

If a and b are nonnegative real numbers and p and q are real numbers greater than 1 such that $1/p + 1/q = 1$, then

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q} \tag{5}$$

The equality holds if and only if $a^p = b^q$.

Definition 2.25 (Hölder continuity)

A real or complex-valued function f on d -dimensional Euclidean space is Hölder continuous, if exists real constants $C > 0, \alpha > 0$, such that for any x and y in the domain of f

$$|f(x) - f(y)| \leq C \|x - y\|^\alpha.$$

α is called the exponent of the Hölder continuity.

Definition 2.26 (Lipschitz continuity)

A real or complex-valued function f on d -dimensional Euclidean space is Lipschitz continuous if there exists a real constant $K \geq 0$ such that, for any x and y in the domain of f

$$|f(x) - f(y)| \leq K \|x - y\|.$$

Observe that

For $0 < \alpha \leq 1$, and for functions over a closed and bounded non-trivial interval of the real line

$$\text{Lipschitz continuous} \subset \alpha\text{-Hölder continuous}$$

Strong and weak convergence

Depending on the application, we focus either on

- ▶ *strong convergence*, where approximation of the outcomes of $X(T)$ is relevant,
- ▶ or *weak convergence*, where only the distribution (law) of $X(T)$ needs to be approximated.

Definition 2.27

The sequence of random variables $\{Y_n\}_{n \in \mathbb{N}}$ converges *strongly* to the random variable Y if

$$\|Y - Y_n\|_{L_p^2(\Omega)} \equiv \sqrt{E[(Y - Y_n)^2]} \rightarrow 0$$

Obs: By Chebychev's inequality we have

$$P(|Y - Y_n| \geq \epsilon) \leq \frac{E[(Y - Y_n)^2]}{\epsilon^2} \rightarrow 0$$

for any fixed $\epsilon > 0$.

Definition 2.28

The sequence of random variables $\{Y_n\}_{n \in \mathbb{N}}$ converges *weakly* to the random variable Y if $E[g(Y)] - E[g(Y_n)] \rightarrow 0$, for all bounded continuous functions g .

Observe: strong convergence \Rightarrow weak convergence, but the converse is in general not true.

Proof of (\Rightarrow) for Lipschitz functions:

$$\begin{aligned}|E[g(X) - g(Y_n)]| &\leq E[|g(X) - g(Y_n)|] \\&\leq C_g E[|X - Y_n|] \\&\leq C_g \underbrace{\sqrt{E[|X - Y_n|^2]}}_{= \|X - Y_n\|_{L_P^2(\Omega)}} \rightarrow 0.\end{aligned}$$

Strong and weak convergence

Counterexample. Let random variables $\{Y_n\}_{n \in \mathbb{N}}$ be *iid* in (Ω, \mathcal{F}, P) , and $Y_n \sim N(0, 1)$, $n = 1, \dots$.

Verify that Y_n converges weakly but not strongly!

Observe: The previous estimate may not be optimal. There are cases where the weak error goes to zero much faster than the strong one.

Conditional expectation

Given a random variable X on a probability space (Ω, \mathcal{F}, P) one wants to define the expected value of X given *additional information*, i.e. knowing that certain events happened or a r.v. Y has been observed.

The conditional expectation of X , given the value of $Y = y$, with $P(Y = y) > 0$, is just the quantity

$$E[X|Y = y] = \sum_x x P_{X|Y}(x|y),$$

when X is a discrete random variable, and

$$E[X|Y = y] = \int x f_{X|Y}(x|y) dx,$$

when X is a continuous random variable.

Exercise 2.16

Verify that if X and Y are independent, $E[X|Y = y] = E[X]$ and that if $X = g(Y)$ for some function g then $E[X|Y = y] = g(y)$.

Observe:

In general we have $E[X|Y = y] = g(y)$ and in fact we can think of a random variable,

$$g(Y) = E[X|Y],$$

which satisfies the *best mean square approximation*:

$$E[(X - g(Y))^2] \leq E[(X - h(Y))^2]$$

for all functions h .

Observe:

The notion of conditional expectation of a random variable is in general defined by means of σ -algebras.

Definition 2.29 (Conditional expectation of a r.v. given a σ -algebra)

Let X be an integrable random variable on (Ω, \mathcal{F}, P) and \mathcal{G} be a σ -algebra contained in \mathcal{F} . The *conditional expectation of X given \mathcal{G}* is a random variable $E[X|\mathcal{G}]$ such that

- i) $E[X|\mathcal{G}]$ is measurable with respect to \mathcal{G} ;
- ii) $E(1_A E[X|\mathcal{G}]) = E[1_A X]$, for all $A \in \mathcal{G}$.

The random variable $E[X|\mathcal{G}]$ satisfying the two conditions above is essentially unique, that is the two conditions determine $E[X|\mathcal{G}]$ up to a set in \mathcal{G} of null P -measure.

Observe: If $A = \Omega$ in ii), then $E(E[X|\mathcal{G}]) = E(X)$.

Example 2.30 (Finite case)

Let $\mathcal{G} = \sigma(\{A_n\}_{n=1}^N)$ with $\{A_n\}_{n=1}^N$ a disjoint family.

Show that

$$E[X|\mathcal{G}] = \sum_n 1_{A_n} E[X|A_n],$$

with $E[X|A_n] = \frac{E[1_{A_n}X]}{P(A_n)}.$

Exercise 2.17

Let $X(\omega) \in \{1, 2, 3, 4, 5, 6\}$ be the outcome of the dice rolling experiment. Consider the event $A = \{\omega \in \Omega : X(\omega) \text{ is odd}\}$ and the sigma algebra $\mathcal{G} = \sigma(A)$. Compute explicitly $E[X|\mathcal{G}]$.

Some properties of the conditional expectation of random variables given a σ -algebra \mathcal{G}

Let X and Z be integrable random variables on (Ω, \mathcal{F}, P) and \mathcal{G} be a σ -algebra contained in \mathcal{F} .

1. Linearity: $E[\alpha X + Z | \mathcal{G}] = \alpha E[X | \mathcal{G}] + E[Z | \mathcal{G}]$, for any $\alpha \in \mathbb{R}$.
2. Monotonicity: If $X \geq Z$ a.s., then $E[X | \mathcal{G}] \geq E[Z | \mathcal{G}]$.
3. If $E(|XZ|) < \infty$ and Z is measurable with respect to \mathcal{G} , then $E[XZ | \mathcal{G}] = ZE[X | \mathcal{G}]$ and $E[Z | \mathcal{G}] = Z$.
4. If X is independent of \mathcal{G} then $E[X | \mathcal{G}] = E[X]$.

5. Tower property: if \mathcal{H} is another σ -algebra contained in $\mathcal{G} \subset \mathcal{F}$, then $E[E[X|\mathcal{G}]|\mathcal{H}] = E[E[X|\mathcal{H}]|\mathcal{G}] = E[X|\mathcal{H}]$.

Proof

The equality $E[E[X|\mathcal{H}]|\mathcal{G}] = E[X|\mathcal{H}]$ follows from the property $E[XZ|\mathcal{G}] = ZE[X|\mathcal{G}]$ with $X = 1$ and $Z = E[X|\mathcal{H}]$.

To verify that if X is an integrable random variable on (Ω, \mathcal{F}, P) and $\mathcal{H} \subset \mathcal{G} \subset \mathcal{F}$, then $E[E[X|\mathcal{G}]|\mathcal{H}] = E[X|\mathcal{H}]$ it is useful to observe that

$$E(1_A E[X|\mathcal{G}]) = E[1_A X], \text{ for all } A \in \mathcal{G}, \text{ and that}$$

$$E(1_A E[X|\mathcal{H}]) = E[1_A X], \text{ for all } A \in \mathcal{H}.$$

Since $\mathcal{H} \subset \mathcal{G}$, then $E(1_A E[X|\mathcal{H}]) = E(1_A E[X|\mathcal{G}])$, for all $A \in \mathcal{H}$, and we can conclude that $E[X|\mathcal{H}] = E[E[X|\mathcal{G}]|\mathcal{H}]$.

6. Best predictor (according to the mean squared error criterion)

$$E[(X - E[X|\mathcal{G}])^2] \leq E[(X - Z)^2],$$

for any \mathcal{G} – measurable Z with $E(Z^2) < \infty$.

7. Jensen's inequality for conditional expectations.

Let $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a convex function such that $E[\varphi(X)] < \infty$.

Then we have

$$E[\varphi(X)|\mathcal{G}] \geq \varphi(E[X|\mathcal{G}])$$

Lemma 2.31 (Jensen's inequality for conditional expectations)

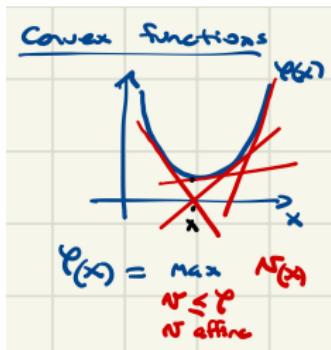
Proof: Since φ is convex we have

$$\varphi(x) = \max_{v \text{ affine s.t. } v \leq \varphi} v(x)$$

Therefore, for any v affine s.t. $v \leq \varphi$ we have

$$\begin{aligned} E[\varphi(X)|\mathcal{G}] &\geq E[v(X)|\mathcal{G}] \\ &\geq v(E[X|\mathcal{G}]). \end{aligned}$$

To conclude, take maximum over $v \leq \varphi$.



A direct consequence of Jensen's inequality, using $\varphi(x) = x^{2p}$, is the following:

Corollary 2.1 (Boundedness of conditional moments)

Let $p \geq 1$.

If $E[X^{2p}] < \infty$ then

$$(E[X|\mathcal{G}])^{2p} \leq E[X^{2p}|\mathcal{G}] < \infty.$$

Corollary 2.2 (L^{2p} -approximation)

Let $p \geq 1$.

If $E[X^{2p}] < \infty$ and $E[Y^{2p}] < \infty$ then

$$(E[X|\mathcal{G}] - E[Y|\mathcal{G}])^{2p} \leq E[(X - Y)^{2p}|\mathcal{G}]$$

and

$$E[(E[X|\mathcal{G}] - E[Y|\mathcal{G}])^{2p}] \leq E[(X - Y)^{2p}] < \infty,$$

or more concisely, $\|E[X|\mathcal{G}] - E[Y|\mathcal{G}]\|_{L_P^{2p}(\Omega)} \leq \|X - Y\|_{L_P^{2p}(\Omega)}$

The proof is again a direct application of Jensen's inequality in the multivariate case, using the convex function $\varphi(x, y) = (x - y)^{2p}$.

Lemma 2.32 (Conditional Hölder inequality)

Consider two random variables, X and Y on $(\Omega, \mathcal{F}, \mathbb{P})$. Let $\mathcal{G} \subset \mathcal{F}$ be a σ -algebra. Let $p, q > 1$ such that $1/p + 1/q = 1$. Then we have

$$E[|XY| | \mathcal{G}] \leq (E[|X|^p | \mathcal{G}])^{1/p} (E[|X|^q | \mathcal{G}])^{1/q}$$

Proof: Let $\delta > 0$, then define $A_\delta = (E[|X|^p | \mathcal{G}] + \delta)^{1/p} \geq \delta^{1/p}$ and $B_\delta = (E[|Y|^q | \mathcal{G}] + \delta)^{1/q} \geq \delta^{1/q}$.

Recall Young's inequality (5), $\frac{x^p}{p} + \frac{y^q}{q} - xy \geq 0$.

Taking $x = |X|/A_\delta$ and $y = |Y|/B_\delta$ yields

$$\frac{|XY|}{A_\delta B_\delta} \leq \frac{1}{p} \frac{|X|^p}{A_\delta^p} + \frac{1}{q} \frac{|Y|^q}{B_\delta^q} , \text{a.s.}$$

Now take conditional expectations, use monotonicity and the measurability of A_δ , B_δ yielding

$$\frac{\mathbb{E}(|XY||\mathcal{G})}{A_\delta B_\delta} \leq \frac{1}{p} \frac{A_0^p}{A_\delta^p} + \frac{1}{q} \frac{B_0^q}{B_\delta^q} \leq \frac{1}{p} + \frac{1}{q} = 1.$$

From here, we arrive at

$$\mathbb{E}(|XY||\mathcal{G}) \leq A_\delta B_\delta, \text{ for any } \delta > 0, \text{ a.s.}$$

Letting $\delta \rightarrow 0$ in the last inequality the rhs tends monotonically to $A_0 B_0$, and this finishes the proof.

Remark 2.2

If $h(y) = E[X|Y = y]$ and we let

$$Z(\omega) = h(Y(\omega)),$$

then

$$Z = E[X|\mathcal{G}]$$

where $\mathcal{G} = \sigma(Y)$.

The conditional expectation $E[X|Y = y]$ can be transferred to Ω by means of $Z(\omega) = h(Y(\omega))$.

Reciprocally, any conditional expectation $E[X|\mathcal{G}]$, \mathcal{G} a σ -algebra contained in \mathcal{F} , can be defined by means of a r.v. Y provided that $\mathcal{G} = \sigma(Y)$. Why?

The following characterizes the situation when one random variable is a function of another:

Lemma 2.33 (Doob-Dynkin)

If $X, Y : \Omega \rightarrow \mathbb{R}^n$ are two given functions, then Y is $\sigma(X)$ -measurable if and only if there exists a Borel measurable function $g : \mathbb{R}^n \rightarrow \mathbb{R}^n$ such that

$$Y = g(X).$$

Exercise 2.18 (Conditional Variance)

Let X be a square integrable random variable on (Ω, \mathcal{F}, P) , and define the conditional variance of X given Y as

$$\text{Var}[X|Y] = E[(X - E[X|Y])^2|Y].$$

Show that

$$\text{Var}[X] = E[\text{Var}[X|Y]] + \text{Var}[E[X|Y]],$$

which, in turn, implies that

$$\text{Var}[X] \geq \text{Var}[E[X|Y]].$$

Can you give an intuitive interpretation to this last fact?

Conditional Probability and the Prediction Problem

We will specifically look at the prediction problem as an application since conditional probability and prediction are intimately linked concepts. Two particular examples are covered in great detail:

- ▶ Best mean square prediction, where we will see how conditional expectation yields the best mean square predictor; and
- ▶ Affine prediction.

The Calculus-Based View of Conditional Expectation

Exercise 2.19

If Y and Z have a bivariate Gaussian distribution, then show that

$$E[Y|Z = z] = \mu_Y + \rho \frac{\sigma_Y}{\sigma_Z} (z - \mu_z).$$

Thus, the conditional expectation for the Gaussian random variable is affine in the conditioning variable Z .

Conditional Expectations and Prediction Theory

Problem: Given the distribution of a r.v. Y , compute \hat{Y} , the "best" prediction for Y .

Question: What do we mean by "best" prediction?

To give a meaning for "best", introduce a *loss function*

$$\ell : \mathbb{R} \rightarrow \mathbb{R}^+, \ell(0) = 0,$$

and then choose

$$\hat{Y}_\ell = \arg \min E[\ell(Y - \hat{Y})].$$

Typical choices:

1. squared error ($\ell(x) = x^2$) $\Rightarrow \hat{Y}_2 = E[Y]$.
2. absolute error ($\ell(x) = |x|$) $\Rightarrow \hat{Y}_1 = \text{median}[Y]$.
3. L^p error ($\ell(x) = |x|^p$).

Problem:

Given the joint distribution of a r.v. (Y, Z) , and observations of Z , compute \hat{Y}_2 , the "best" (in mean square sense) prediction for Y . Define

$$\mathcal{L} = \{g : \mathbb{R} \rightarrow \mathbb{R} : E[g^2(Z)] < \infty\}.$$

In other words, we seek $g^*(Z)$ s.t. it minimizes

$$\begin{aligned} g^* &= \arg \min_{g \in \mathcal{L}} E[(Y - g(Z))^2] \\ &= \arg \min_{g \in \mathcal{L}} E[E[(Y - g(Z))^2 | Z]]. \end{aligned} \tag{6}$$

Observe: In general g^* , the regression function, may be a nonlinear function.

Conditional Expectation as a Projection:

Let $L_P^2 = \{X \text{ r.v. in } (\Omega, \mathcal{F}, P) : E[X^2] < \infty\}$.

The regression problem can be formulated as a projection. Indeed, $\mathcal{L} \subset L_P^2$ is a linear subspace, and by introducing the L_P^2 inner product

$$(X_1, X_2)_2 = E[X_1 X_2]$$

and its induced norm $\|X\|_2 = (X, X)^{1/2}$,

we see that (6) is equivalent to

$$g^* = \arg \min_{g \in \mathcal{L}} \|Y - g(Z)\|_2.$$

The above minimizer is unique because L^2 is a Hilbert space and \mathcal{L} is a closed linear subspace of L^2 :

Theorem 2.7 (Hilbert Space Projection Theorem)

. Let \mathcal{L} be a closed linear subspace of a Hilbert space L^2 , and let $Y \in L^2$. There exists a unique $W^* \in \mathcal{L}$ that minimizes

$$\|Y - W\|_2 = (Y - W, Y - W)_2$$

over $W \in \mathcal{L}$. Such unique W^* (the orthogonal projection of Y onto \mathcal{L}) is characterized by

$$(Y - W^*, W) = 0, \text{ for } W \in \mathcal{L}. \quad (7)$$

In our case (7) reads

$$E[(Y - g^*(Z))g(Z)] = 0, \text{ for all } g \text{ with } E[g^2(Z)] < \infty.$$

Definition 2.34 (Conditional Expectation)

Suppose $Y \in L^2$. The conditional expectation of Y given Z is the r.v. $g^*(Z)$ satisfying

$$E[g^*(Z)g(Z)] = E[Yg(Z)], \text{ for all r.v. } g(Z) \in L^2. \quad (8)$$

We write $E[Y|Z] = g^*(Z)$.

Observe:

The definition of conditional expectation

- ▶ is consistent with the previous definition of conditional expectation in terms of σ -algebras, (and conditional densities where applicable);
- ▶ makes rigorous sense regardless of whether Z is a finite-dimensional random vector or a continuum of r.v.s.
- ▶ implies that $E[Y|Z]$ may be not so easy to compute exactly. . .

Example 2.35 (Polynomial regression)

Consider the subspace

$$\mathcal{L}_p = \{g : g \text{ is a polynomial with degree at most } p\}.$$

Then solve

$$g^* = \arg \min_{g \in \mathcal{L}_p} E[(Y - g(Z))^2 | Z].$$

Let $\{\varphi_n\}_{n=1}^{p+1}$ be a basis for \mathcal{L}_p . Then

$$g(Z) = \sum_{n=1}^{p+1} g_n \varphi_n(Z)$$

and (8) reads

$$\sum_{n=1}^{p+1} g_n E[\varphi_m(Z) \varphi_n(Z)] = E[\varphi_m(Z) Y], \text{ for } m = 1, \dots, p+1,$$

which is a linear system of the form

$$Mg = f.$$

Observe: The coefficients of M depend on the first p moments of Z , the rhs depends in general on more. Recall that higher order moments are difficult to estimate.

The condition of M may deteriorate with p unless φ_n are chosen orthogonal.

How many samples do we need to make the projection stable if we do it discretely?

Exercise 2.20

Suppose that T is a component lifetime that is exponential with parameter λ . To model manufacturing variability, we assume that λ is itself random. Assume that $\lambda \sim U(0, 1)$.

1. Show that $E[T|\lambda] = 1/\lambda$. Observe that $E[T|\lambda]$ is nonlinear in λ .
2. Compute $E[\lambda|T]$.

Affine Prediction

Here for computational reasons we restrict ourselves to \mathcal{L}_1 approximations of $E[Y|Z]$. Let $Z \in \mathbb{R}^M$ be multidimensional and write $g(Z) = a^T(Z - E[Z]) + b$. To find (a^*, b^*) we solve, for $m = 1, \dots, M$

$$E[(a^T(Z - E[Z]) + b)(Z_m - E[Z_m])] = E[Y(Z_m - E[Z_m])],$$

and

$$E[a^T(Z - E[Z]) + b] = E[Y].$$

This can be written in matrix form as

$$\begin{bmatrix} \text{Cov}[Z] & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} E[Y(Z - E[Z])] \\ E[Y] \end{bmatrix}$$

We then have

$$b^* = E[Y] \text{ and } a^* = \text{Cov}[Z]^{-1}E[Y(Z - E[Z])]$$

yielding

$$\begin{aligned} g_1^*(Z) &= E[Y(Z - E[Z])]^T \text{Cov}[Z]^{-1}(Z - E[Z]) + E[Y] \\ &= (E[YZ] - E[Y]E[Z])^T \text{Cov}[Z]^{-1}(Z - E[Z]) + E[Y] \end{aligned}$$

Remark 2.3

In the Gaussian case affine prediction is exact.

Example 2.36 (Filters)

Suppose $\{Z_t\}$ is now infinitely dimensional and t is time. We would like to predict Y_t based on observations $\{Z_s\}$, $s \leq t$. Propose and characterize an affine predictor.

Hint: assume $E[Y_t] = 0$ for all t and use the ansatz

$$\hat{Y}_t = \int_{-\infty}^t a(s, t) Z_s ds.$$

For best L^2 approximation, characterize the deterministic function $a(s, t)$ by minimising

$$\min_a E[(Y_t - \hat{Y}_t(a))^2].$$

Statistical Parameter Estimation: The Method of Maximum Likelihood

Problem: Estimate the parameter θ in the distribution of a random variable X using observed iid samples, $X_i, i = 1, \dots$

MLE Principle to follow: Estimate the parameter θ by the value $\hat{\theta}$ that maximizes the likelihood of observing the given sample.

For discrete iid random variables, the MLE reads

$$\hat{\theta} = \arg \max_{\theta \in \Theta} \prod_i P(X = x_i; \theta),$$

and for continuous iid random variables

$$\hat{\theta} = \arg \max_{\theta \in \Theta} \prod_i f_X(X = x_i; \theta).$$

Example 2.37

Estimate the mean μ of a normal random variable based on M iid samples.

Recall:

$$f_X(x; \mu, \sigma^2) = \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma}.$$

then the log-likelihood is proportional to

$$-\sum_{m=1}^M (x_m - \mu)^2,$$

and therefore $\hat{\mu} = \frac{\sum_{m=1}^M x_m}{M}$.

Exercise 2.21

Estimate the standard deviation σ of a normal random variable based on M iid samples and the estimator $\hat{\mu}$ for the mean.

Example 2.38

Estimate the parameter λ of an $Exp(\lambda)$ random variable based on M iid samples.

Recall, that for $\lambda > 0$

$$f_X(x; \lambda) = \lambda e^{-\lambda x}.$$

Censored estimation

Exercise 2.22

Estimate the parameter λ of an $\text{Exp}(\lambda)$ random variable based on M iid samples

BUT

if $X_i > T_{\text{end}}$ then the actual value of X_i is not observed!!

Question: Can you think of a practical example with this characteristics?

A poor boy's alternative to MLE? The method of moments (MoM)

Principle to follow: Find $\hat{\theta}$ that matches as many sample moments as possible.

In other words, $\hat{\theta}$ satisfies

$$E[X^k(\hat{\theta})] = \hat{E}[X^k], \quad k = 1, 2, \dots, k_{max} .$$

Example 2.39 (Normal estimation by MoM: mean)

Estimate the mean μ of a normal random variable based on M iid samples.

Exercise 2.23

Estimate the standard deviation σ of a normal random variable by the MoM based on M iid samples and the estimator $\hat{\mu}$ for the mean.

Cramér-Rao inequality

or what is the best unbiased estimator we could construct with given iid data.

Let $\{X_i\}_{i=1}^M$ be an iid sample from the pdf $f_X(\cdot; \theta)$, and $\hat{\theta}(X_1, \dots, X_M)$ an estimator for θ satisfying

1. $\text{Var}[\hat{\theta}] < \infty$;
2. $E[\hat{\theta}] = \theta$ for all $\theta \in \Theta$ (unbiased estimator).

Then, we have the lower bound

$$\frac{1}{ME \left[\left(\frac{d \log f_X(\cdot; \theta)}{d\theta} \right)^2 \right]} \leq \text{Var}[\hat{\theta}].$$

Observe:

The term

$$E \left[\left(\frac{d \log f_X(\cdot; \theta)}{d\theta} \right)^2 \right] = \int \left(\frac{d \log f_X(x; \theta)}{d\theta} \right)^2 f_X(x; \theta) dx$$

is called the *information number* or the *Fisher information* of the sample.
Can you provide some intuition to this naming?

Proof of Cramér-Rao inequality:

Since $\hat{\theta}$ is an unbiased estimator for θ we have, for all $\theta \in \Theta$ and for all M ,

$$\begin{aligned} 0 &= E[\theta - \hat{\theta}] \\ &= \int (\theta - \hat{\theta}(x_1, \dots, x_M)) \underbrace{\prod_{m=1}^M f_X(x_i; \theta)}_{=\rho_X(x_1, \dots, x_M; \theta)} dx_1 \dots dx_M. \end{aligned}$$

Then, taking derivatives in the above wrt θ gives

$$1 = E[(\hat{\theta} - \theta) \partial_\theta (\log \rho_X(x_1, \dots, x_M; \theta))] , \text{ for all } \theta \in \Theta.$$

Now use the Cauchy-Schwarz inequality

$$1 \leq \underbrace{E[(\theta - \hat{\theta})^2]}_{=Var[\hat{\theta}]} E[(\partial_\theta (\log \rho_X(\theta)))^2], \text{ for all } \theta.$$

To finish the proof, we need to simplify the term

$$\begin{aligned} E[(\partial_\theta (\log \rho_X(\theta)))^2] &= E \left[\left(\sum_{m=1}^M \partial_\theta \log f_X(X_m; \theta) \right)^2 \right] \\ &= M E \left[(\partial_\theta \log f_X(X; \theta))^2 \right] \\ &\quad + \frac{M(M-1)}{2} (E [\partial_\theta \log f_X(X; \theta)])^2. \end{aligned}$$

Conclude that

$$E[\partial_\theta \log f_X(X; \theta)] = 0$$

after differentiating

$$E[1] = \int 1 f_X(x; \theta) dx = 1$$

wrt θ .

Remark 2.4 (Extension to Quantities of Interest $\tau(\theta)$)

The Cramér-Rao inequality is easily extended to the case of an unbiased estimator $T(X_1, \dots, X_M)$ of a differentiable function of θ , say

$$\tau : \Theta \rightarrow \mathbb{R} :$$

$$\frac{(\tau'(\theta))^2}{ME \left[\left(\frac{d \log f_X(\cdot; \theta)}{d\theta} \right)^2 \right]} \leq Var(T(X_1, \dots, X_M)).$$

Remark 2.5 (Multidimensional θ)

In the multi-parameter case, say $\theta \in \Theta \subset \mathbb{R}^d$, under the assumption that $f(x_1, \dots, x_M; \theta_1, \dots, \theta_d) = f(\mathbf{x}; \theta)$ is differentiable wrt each θ_i , $i = 1, \dots, d$, we may consider the Fisher information matrix

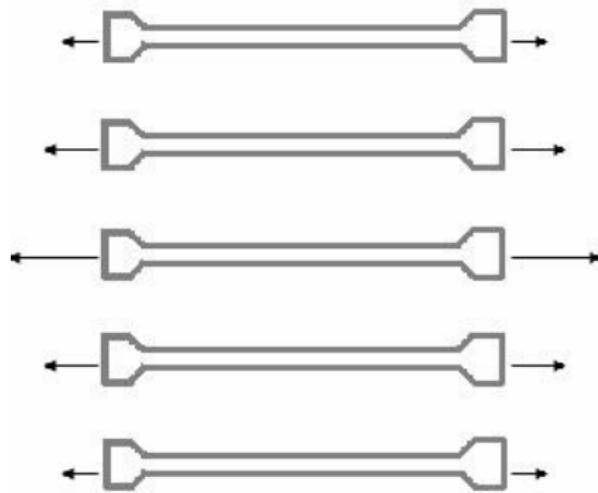
$$\mathcal{I}(\theta) := E \left[\frac{\partial \ell(\theta; \mathbf{X})}{\partial \theta} \frac{\partial \ell(\theta; \mathbf{X})}{\partial \theta}^T \right]$$

where $\ell(\theta; \mathbf{x})$ denotes the log-likelihood.

If $T(X_1, \dots, X_M)$ is an unbiased estimator of the differentiable function $\tau : \Theta \rightarrow \mathbb{R}$ then

$$\tau'(\theta)^T \mathcal{I}(\theta)^{-1} \tau'(\theta) \leq \text{Var}(T(X_1, \dots, X_M)).$$

Fatigue problem in one dimension



Goal: Calibrate a model to determine after how many reversals failure will occur.

S-N curves

Definition 2.40 (Fatigue: S-N curve)

The S-N curve describes the relation between stress amplitude and the corresponding number of cycles to failure.

When the number of cycle to failure is finite we say that the specimen has finite life.

Definition 2.41 (Fatigue limit)

If a specimen is exposed to a lower stress level than the fatigue limit then its life is infinity.

Goal: To compute the probability of not surviving a certain life.

About Failure

Remark 2.6 (Failure)

Depending on the application the definition of failure may be different. For instance, in one context we may say that we have a failure when a crack of a given length appears, while in other we may declare failure when the whole specimen breaks down.

Example 2.42 (Aircraft)

A flight critical component must be retired if there is a given percent chance that in the next flight operation a crack nucleation event will occur. Crack nucleation is defined as the first appearance of a 0.25 mm (0.01 inch) long crack.



4-28-1988 After 89,090 flight cycles on a 737-200, metal fatigue lets the top go in flight.

Available Experimental Data

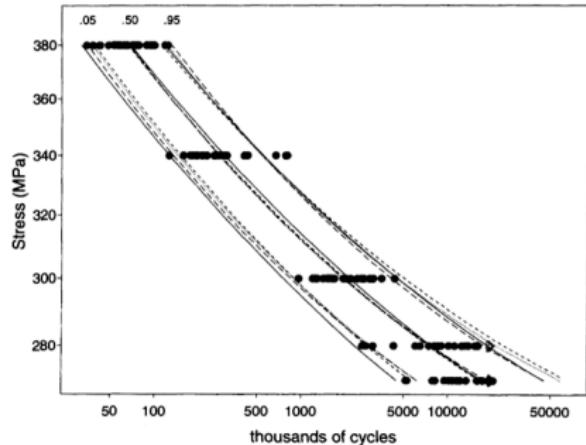


Figure 1. Log-Log S-N Plot for the Laminate Panel Data With ML Estimates of the .05, .50, and .95 Quantiles (●, failure; ▷, censored observation): —, Sev-Sev; - - -, Normal-Normal; - · - , Sev-Normal; - - - - , Normal-Sev.

The available experiment consists on exposing a specimen to a cyclic stress with a given stress amplitude until either failure is observed or we reach a prescribed maximum number of cycles,

$$n_{ro} = e^{y_{ro}}$$

without observing failure.

Observe: When not all units on test fail we have censored data!

Definition 2.43 (Censored and Uncensored Data)

The data corresponding to $n_i > n_{ro}$ are assigned the value n_{ro} and called censored data. The set of censored data is C while the set of uncensored data is U .

Therefore, the available data consists on a list of triplets:

$$\begin{aligned} & (U, \sigma_i, n_i), \text{ if } n_i \leq n_{ro} \\ & (C, \sigma_i, n_{ro}), \text{ if } n_i > n_{ro}. \end{aligned}$$

Likelihood Function for Reliability Data

Let $\rho(n; \sigma, \theta)$ be the PDF for the chosen life distribution model and

$$F(n; \sigma, \theta) = \mathbb{P}(N \leq n; \sigma, \theta) \quad (9)$$

the corresponding CDF. They depend on

- ▶ the imposed cyclic stress, σ , and
- ▶ θ , a vector of model parameters.

The corresponding likelihood function for θ is

$$L(\theta) = K \prod_{i \in C} \mathbb{P}(N > n_{ro}; \sigma_i, \theta) \times \prod_{i \in U} \rho(n_i; \sigma_i, \theta)$$

with K denoting a constant that plays no role when solving for θ in the Maximum Likelihood Estimator (MLE).

The MLE estimator for θ is

$$\begin{aligned}\hat{\theta} &= \arg \max_{\theta} L(\theta) \\ &= \arg \max_{\theta} \prod_{i \in C} \mathbb{P}(N > n_{ro}; \sigma_i, \theta) \times \prod_{i \in U} \rho(n_i; \sigma_i, \theta) \\ &= \arg \max_{\theta} \prod_{i \in C} (1 - F(n_{ro}; \sigma_i, \theta)) \times \prod_{i \in U} \rho(n_i; \sigma_i, \theta).\end{aligned}$$

Remark 2.7

With no censoring, the likelihood reduces to the product of the densities, each evaluated at a number of cycles to failure.

Remark 2.8

A completely similar approach can be done for cyclic strain imposed experiments and the fitting of $\mathbb{P}(N \leq n; \epsilon, \theta)$.

One possible model: lognormal distributed life

Let N be a lognormal random variable.

Then $y = \log(N)$ is a normal random variable with log life mean $\mu = \mu(\sigma; \theta)$ and log life standard deviation $\tau = \tau(\sigma; \theta)$.

We have

$$\begin{aligned}\mathbb{P}(N \leq n; \sigma, \theta) &= \mathbb{P}(\log(N) \leq \log(n); \sigma, \theta) \\ &= \Phi\left(\frac{\log(n) - \mu(\sigma; \theta)}{\tau(\sigma; \theta)}\right)\end{aligned}$$

Example 2.44

Let $x = \log(\sigma)$. Then take

$$\tau(\sigma; \theta) = \tau$$

and

$$\mu(\sigma; \theta) = \begin{cases} \frac{ax^2 + bx + c}{x - x_\infty}, & \text{if } x > x_\infty \\ +\infty, & \text{otherwise.} \end{cases}$$

We have the vector of 5 parameters, $\theta = (\tau, a, b, c, x_\infty)$.

We can estimate θ by a Maximum Likelihood approach.

Calibration of the lognormal life model

Denote $g(t; (\mu, \sigma)) = \frac{e^{-\frac{(t-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma}$.

The likelihood using experimental data is:

$$\begin{aligned} L(\theta) &= \prod_{i \in U} g(\log(n_i); (\mu(\sigma_i; \theta), \tau(\sigma_i; \theta))) \\ &\quad \times \prod_{i \in C} \left\{ 1 - \Phi \left(\frac{\log(n_{ri}) - \mu(\sigma_i; \theta)}{\tau(\sigma_i; \theta)} \right) \right\} \end{aligned}$$

and the Maximum Likelihood estimate is $\hat{\theta} = \arg \max_{\theta} L(\theta)$.

Another possible model

Now we follow [LL05]. Given a stress level $s = e^x$, the probability of not surviving a certain life $n_x = e^{y_x}$ is

$$\mathbb{P}(Y_x < y_x) = \Phi\left(\frac{y_x - \nu(x)}{\tau}\right) \Phi\left(\frac{s - \mu}{\sigma}\right) \quad (10)$$

where

- ▶ (μ, σ) are the mean and standard deviation of the fatigue limit distribution, which is assumed normal
- ▶ $(\nu(x), \tau^2)$ are the mean and variance of the logarithm of the finite life distribution.

Here $\Phi(t) = \int_{-\infty}^t \frac{e^{-u^2/2}}{\sqrt{2\pi}} du$ is the CDF of a $N(0, 1)$ random variable.

Calibration of [LL05] model

Question: How to obtain the parameters (μ, σ) , $(\nu(x), \tau^2)$ from observed data?

The mean logarithmic finite life $\nu(x)$ is approximated by

$$\nu(x) \approx \alpha + \beta x$$

so we need to determine 5 parameters.

Calibration of [LL05] model with MLE

Let the vector of parameters to determine be $\theta = (\nu(\cdot), \tau, \mu, \sigma)$ and

$$g(t; (\mu, \sigma)) = \frac{e^{-\frac{(t-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma}$$

Likelihood using experimental data ($y_i = \log(n_i)$, $x_i = \log(s_i)$):

$$\begin{aligned} L(\theta) &= \prod_{i \in U} \left\{ g(y_i; (\tau, \nu(x_i))) \Phi \left(\frac{s_i - \mu}{\sigma} \right) \right\} \\ &\quad \times \prod_{i \in C} \left\{ 1 - \Phi \left(\frac{y_{ro} - \nu(x_i)}{\tau} \right) \Phi \left(\frac{s_i - \mu}{\sigma} \right) \right\} \end{aligned}$$

and the Maximum Likelihood estimate is $\hat{\theta} = \arg \max_{\theta} L(\theta)$.

Drawing the quantiles for the S-N curves

Once the parameters $\theta = (\nu(\cdot), \tau, \mu, \sigma)$ are determined, we can use (10) to predict the life based on a stress level s and a given failure probability $0 < q < \Phi\left(\frac{s-\mu}{\sigma}\right)$.

We have

$$y(s, q) = \nu(\log(s)) + \tau \Phi^{-1} \left(\frac{q}{\Phi\left(\frac{s-\mu}{\sigma}\right)} \right)$$

with Φ^{-1} being the inverse function of Φ .

Other models

There are many other models for the S-N curve. See [PM99]² and its table partly reproduced below:

Some models for the relationship between applied stress and lifetime		
Model	Quantile estimate	Number of parameters
Little and Ekvall (1981)	$\log(\hat{y}_{ij}) = \hat{A} + \hat{B}x_j + z_{p_i}\hat{\sigma}$	3
Little and Ekvall (1981)	$\log(\hat{y}_{ij}) = \hat{A} + \hat{B}\log(x_j) + z_{p_i}\hat{\sigma}$	3
Bastenaire (1972)	$\log(\hat{y}_{ij}) = \hat{Y}_0 + \hat{A}\frac{\exp[-\hat{C}(x_j - \hat{X}_0)]}{x_j - \hat{X}_0} + z_{p_i}\hat{\sigma}$	5
Castillo et al. (1985)	$\log(\hat{y}_{ij}) = \log(\hat{Y}_0) + \frac{\hat{A}}{\log(x_j) - \log(\hat{X}_0)} + z_{p_i}\hat{\sigma}$	5
Pascual and Meeker (1999)	$\log(\hat{y}_{ij}) = F_W^{-1}(p_i; \log(x_j), \hat{\theta})$	5

where x_j is the j th stress level, y_{ij} is the i th smallest observation at stress x_j , \hat{y}_{ij} is the estimate of the $p_i = (i - .5)/15$ quantile, and z_{p_i} is the p_i quantile of the standard normal distribution.

²– Francis G. Pascual and William Q. Meeker, Estimating Fatigue Curves with the Random Fatigue-Limit Model (with discussion), *Technometrics*, **41** (4), November 1999, pp.277–302.

– Bayesian inference and model comparison for metallic fatigue data, by I. Babuška, Z. Sawlan, M. Scavino, B. Szabo and R. Tempone. Computer Methods in Applied Mechanics and Engineering, 2016.

Remark 2.9 (On the dependence wrt σ)

Let $\rho(n; \sigma, \theta)$ be the PDF for the chosen life distribution model, where σ denotes the cyclic stress. The actual dependence may be parametrized for instance in terms of σ_{\max} and σ_{\min} , the maximum and minimum stresses on each cycle, i.e.

$$\rho(n; \sigma, \theta) = \rho(n; \sigma_{\max}, \sigma_{\min}, \theta).$$

Assumption For the sake of exposition only, we now restrict ourselves to the case where $\sigma_{\min} = -\sigma_{\max}$ and write

$$\rho(n; \sigma, \theta) = \rho(n; \sigma_{\max}, \theta).$$

Error in variables problem, nuisance parameters and marginalized likelihood:

Estimate θ from observations (x_j, \hat{y}_j) corresponding to the model

$$\begin{aligned}\hat{y}_j &= g(\hat{x}_j; \theta) + \epsilon_{Y,j} \\ \hat{x}_j &= x_j + \epsilon_{X,j}.\end{aligned}$$

Here

\hat{y}_j is the observed output

$g(\cdot; \theta)$ is the model structure with parameter θ

$\epsilon_{Y,j}$ is the measurement error

x_j is the intended input

$\epsilon_{X,j}$ is the setup error

Error Model: Consider a simplified model, where all errors are independent and normal (possibly with different sizes), i.e.

$$\epsilon_{Y,j} \sim N(0, \sigma_{Y,j}^2)$$

$$\epsilon_{X,j} \sim N(0, \sigma_{X,j}^2)$$

Likelihood:

$$L(\theta, Data)$$

$$= \prod_{j=1}^M \rho_{\epsilon_{Y,j}|X,j}(\hat{y}_j - g(\hat{x}_j, \theta)) \rho_{\epsilon_{X,j}}(\epsilon_{X,j})$$

$$\propto \prod_{j=1}^M \exp\left(-\frac{(\hat{y}_j - g(\hat{x}_j, \theta))^2}{2\sigma_{Y,j}^2}\right) \exp\left(-\frac{(\epsilon_{X,j})^2}{2\sigma_{X,j}^2}\right)$$

$$\propto \prod_{j=1}^M \exp\left(-\frac{(\hat{y}_j - g(x_j + \epsilon_{X,j}, \theta))^2}{2\sigma_{Y,j}^2}\right) \exp\left(-\frac{(\epsilon_{X,j})^2}{2\sigma_{X,j}^2}\right)$$

Do you see any problem with this formula?

The values of $\epsilon_{X,j}$ **are not observable!!** They are *nuisance* parameters and we just *marginalize* the likelihood wrt them, computing

$$E[L(\theta, Data_{observed}) | (\hat{y}_j)_{j=1}^M] \\ \propto \prod_{j=1}^M E \left[\exp \left(-\frac{(\hat{y}_j - g(x_j + \sigma_{X,j}\epsilon, \theta))^2}{2\sigma_{Y,j}^2} \right) \middle| \hat{y}_j \right] \quad (11)$$

with $\epsilon \sim N(0, 1)$. Observe that

$$E \left[\exp \left(-\frac{(\hat{y}_j - g(x_j + \sigma_{X,j}\epsilon, \theta))^2}{2\sigma_{Y,j}^2} \right) \middle| \hat{y}_j \right] \\ = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \exp \left(-\frac{(\hat{y}_j - g(x_j + \sigma_{X,j}\epsilon, \theta))^2}{2\sigma_{Y,j}^2} \right) \exp(-\epsilon^2/2) d\epsilon$$

and this last integral can be approximated by numerical quadrature for instance.

To understand the effect of the error in X let us assume small noise and use the delta method, i.e. $\sigma_{X,j} << 1$. Then, denoting the residuals

$R_j(\theta) = \hat{y}_j - g(x_j, \theta)$ and using Taylor expansion,

$\hat{R}_j(\theta) = R_j(\theta) - g'(x_j, \theta)\sigma_{X,j}\epsilon + \mathcal{O}((\sigma_{X,j}\epsilon)^2)$, we have

$$\begin{aligned} & E \left[\exp \left(-\frac{(\hat{R}_j(\theta))^2}{2\sigma_{Y,j}^2} \right) \middle| \hat{y}_j \right] \\ &= \exp \left(-\frac{(R_j(\theta))^2}{2\sigma_{Y,j}^2} \right) \\ &\quad \times E \left[\exp \left(-\frac{2R_j(\theta)g'(x_j, \theta)\sigma_{X,j}\epsilon + \mathcal{O}(\sigma_{X,j}^2\epsilon^2)}{2\sigma_{Y,j}^2} \right) \middle| \hat{y}_j \right] \end{aligned}$$

Use that since $\epsilon \sim N(0, 1)$ then $E[e^{t\epsilon}] = e^{\frac{t^2}{2}}$ implying

$$\begin{aligned}
 & E \left[\exp \left(-\frac{(\hat{R}_j(\theta))^2}{2\sigma_{Y,j}^2} \right) \middle| \hat{y}_j \right] \\
 &= \exp \left(-\frac{(R_j(\theta))^2}{2\sigma_{Y,j}^2} \right) \\
 &\quad \times \exp \left(\frac{1}{2} \left(\frac{R_j(\theta)g'(x_j, \theta)\sigma_{X,j}}{\sigma_{Y,j}^2} + \dots \right)^2 \right) \\
 &= \exp \left(-\frac{(R_j(\theta))^2}{2\sigma_{Y,j}^2} \left(1 - \left(g'(x_j, \theta) \frac{\sigma_{X,j}}{\sigma_{Y,j}} \right)^2 \right) + \dots \right)
 \end{aligned}$$

Therefore the logarithm of the marginalized likelihood is proportional to

$$-\sum_{j=1}^M \frac{(R_j(\theta))^2}{2\sigma_{Y,j}^2} \left(1 - \left(g'(x_j, \theta) \frac{\sigma_{X,j}}{\sigma_{Y,j}}\right)^2\right) + \dots \quad (12)$$

Observe: that for our derivation with the Delta Method to make sense we need $\left(g'(x_j, \theta) \frac{\sigma_{X,j}}{\sigma_{Y,j}}\right)^2 \ll 1$.

Exercise 2.24

Compare (12) with the usual least squares. Interpret the effect of the correction term $\left(1 - \left(g'(x_j, \theta) \frac{\sigma_{X,j}}{\sigma_{Y,j}}\right)^2\right)$.

Exercise 2.25

Propose a computational method to use in (11) when the small noise approximation is not applicable.

Bayes Theorem

For any two events A and B there holds:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \quad P(B) > 0.$$

In general, for a disjoint family of events (B_n) s.t. $\Omega = \cup_n B_n$ we have
marginalization

$$\begin{aligned} P(A) &= \sum_n P(A \cap B_n) \\ &= \sum_n P(A|B_n)P(B_n). \end{aligned}$$

Now we will follow [Sivia], please read Ch. 1 and 2.

Bayes [1763], Laplace[1812], Jeffreys[1939], Cox[1946], Jaynes ...

Inductive vs. Deductive reasoning

Deductive

Given a cause, work out its consequences.

Derive many complicated and useful results as the logical consequence of a few well-defined axioms.

Inductive

Given that certain effects have been observed, what is (are) the underlying cause(s)?

Make the best inference based on the experimental data and any prior knowledge that we have available, reserving the right to revise our position if new information comes to light.

Probability: Richard Cox and the rules for consistent reasoning

Quantitative rules necessary for logical and consistent reasoning. Our belief in the occurrence of events is encoded in real numbers, following

1. Transitive: $b(A) < b(B)$ and $b(B) < b(C)$ THEN $b(A) < b(C)$
2. Complement: $b(A^c)$ is determined by $b(A)$
3. If $b(A)$ and $b(B/A)$ are given THEN $b(B)$ is determined.

Cox concluded

that $b(\cdot)$ can be mapped into a probability measure $P(\cdot) \geq 0$ satisfying the usual axioms! In particular

$$P(X|I) + P(X^c|I) = 1$$

and

$$P(X, Y|I) = P(X|Y, I) \times P(Y|I).$$

Observe: the above probabilities are conditional on I , the relevant background information at hand, because there is no such thing as an absolute probability!

Bayes Theorem revisited!

Replace X and Y by *hypothesis and data*:

$$P(\text{hypothesis} \mid \text{data}, I) \propto P(\text{data} \mid \text{hypothesis}, I) \\ \times P(\text{hypothesis} \mid I).$$

$P(\text{hypothesis} \mid I)$, is called the prior probability

$P(\text{data} \mid \text{hypothesis}, I)$, likelihood function,

$P(\text{hypothesis} \mid \text{data}, I)$, posterior probability

Observe: the normalization term $P(\text{data} \mid I)$ has been omitted in the above, since it plays no role for parameter estimation. BUT it does in other settings like model selection.

Example 2.45 (Bernoulli revisited)

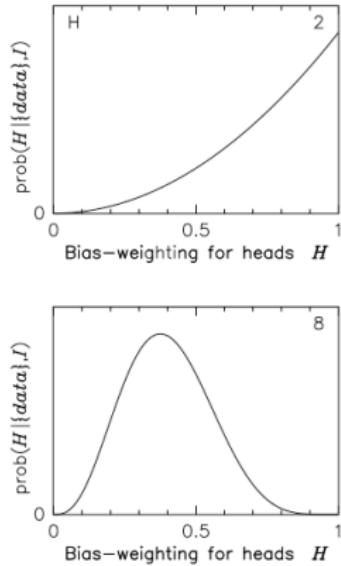
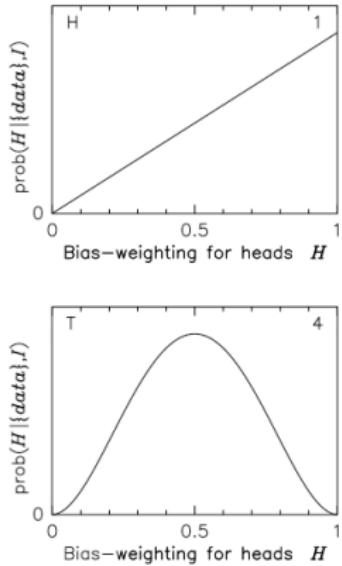
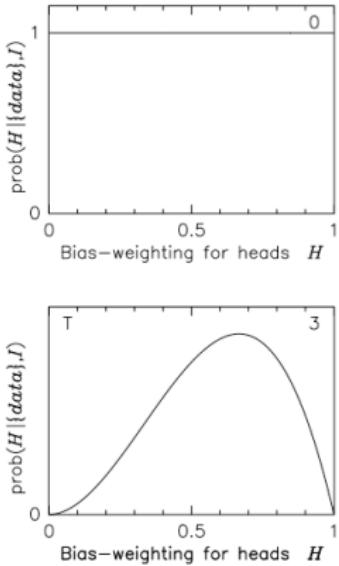
Given M iid data and some prior information, estimate p .

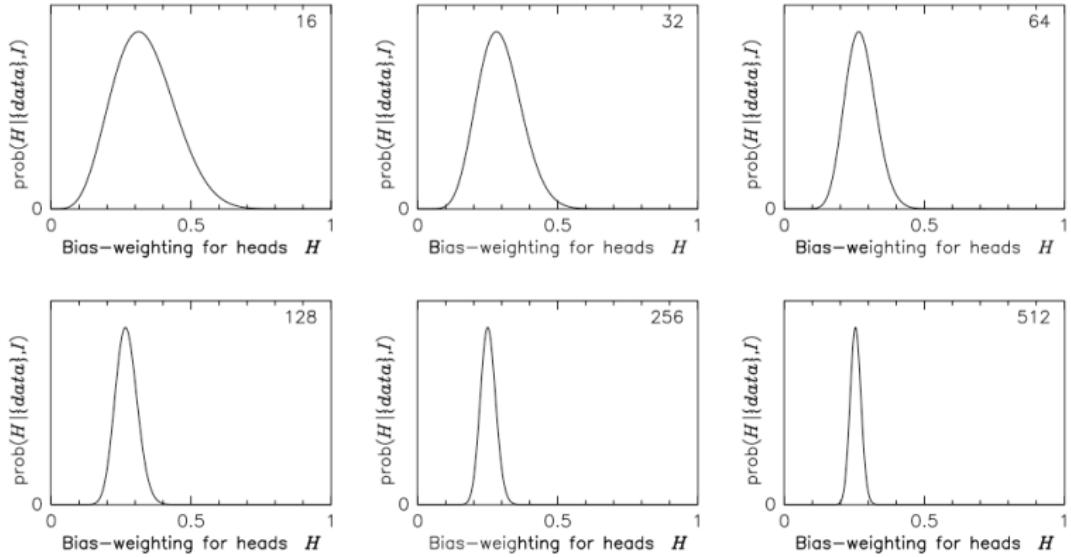
In the absence of information, we can take a prior $U(0, 1)$ for p .

What is the likelihood?

$$L(X; p) = p^{N_1} (1 - p)^{M - N_1}$$

with N_1 being the number of 1's observed. Therefore, the posterior for p is proportional to $L(X; p)$ in this case.





What if we use different priors?

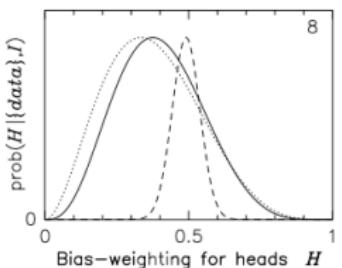
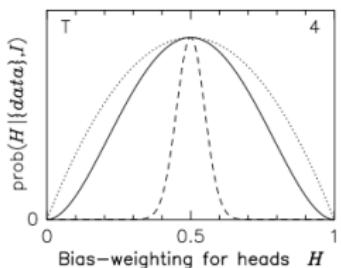
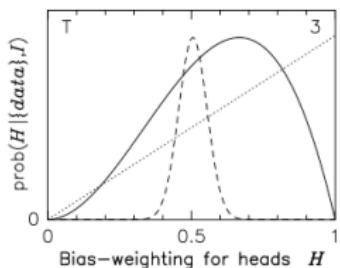
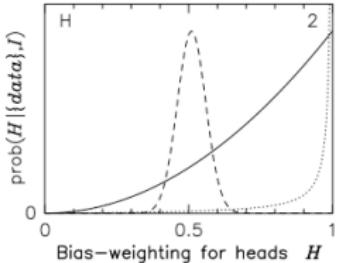
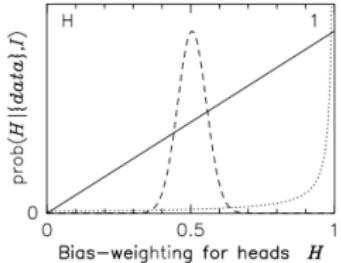
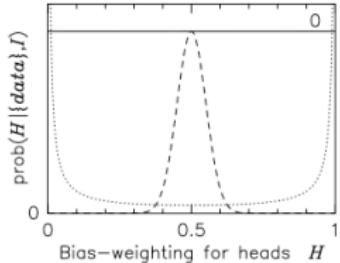
Exercise 2.26

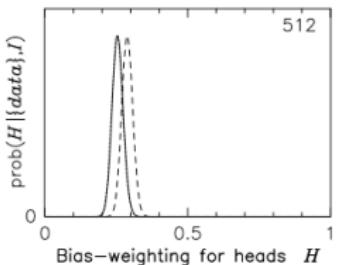
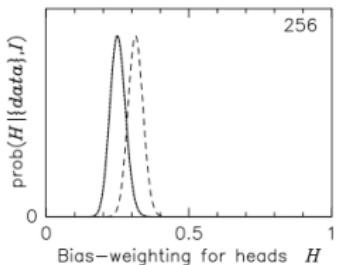
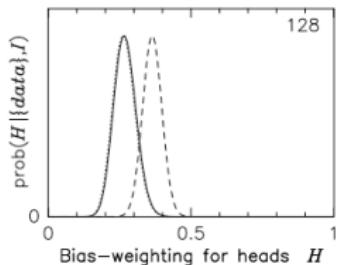
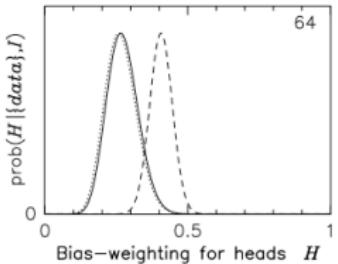
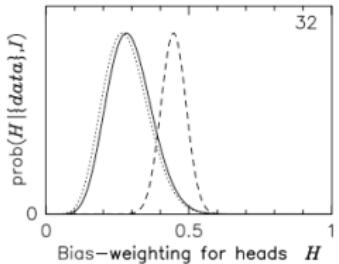
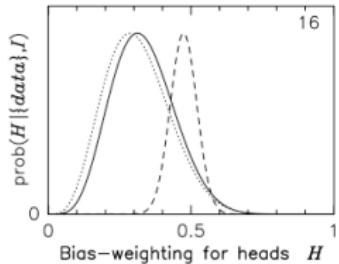
Propose priors to represent the statements

"this is a fair situation"

"this is an extreme case, not fair"

Observe: Using Bayesian statistics, our knowledge of the parameter is encoded by a pdf, not just a single value! Think of a multimodal posterior pdf, what is the value of MLE?





Question: Is there any difference in the results if we use our data in one step or sequentially?

Reliabilities: best estimates, error-bars and confidence intervals

If we know that our posterior is unimodal, we summarize the posterior with just two numbers: the best estimate and a measure of its reliability. Our best estimate is given by the maximum of the posterior pdf (maximum a posteriori (MAP) estimator).

It satisfies

$$\rho'_{posterior}(\theta^*) = 0,$$

and

$$\rho''_{posterior}(\theta^*) \leq 0.$$

Quick approximate error bar: expand the logarithm of the posterior pdf, say $L(\theta)$, around θ^* :

$$L(\theta) = L(\theta^*) + \frac{1}{2}L''(\theta^*)(\theta - \theta^*)^2 + \dots$$

Then, the posterior pdf for θ is approximated by

$$C \exp\left(\frac{1}{2}L''(\theta^*)(\theta - \theta^*)^2\right)$$

which is a Gaussian pdf!

An approximate error bar is given by the interval

$$\theta^* + \frac{C_\alpha}{\sqrt{-L''(\theta^*)}}[-1, 1].$$

with C_α being the confidence constant.

Example 2.46 (Gaussian revisited)

Let $X \sim N(\mu, \sigma^2)$. Given σ and the available data, estimate the value of μ using a Bayesian approach. Derive reliability estimates on μ .

We want to find

$$\rho(\mu | Data) \propto \rho(Data|\mu) \rho_{prior}(\mu)$$

Here

$$\rho(Data|\mu) \propto \exp\left(-\frac{1}{2} \sum_{m=1}^M (X_m - \mu)^2 / \sigma^2\right)$$

Assuming a Gaussian prior, $N(\mu_{prior}, \sigma_{prior}^2)$, we have

$$\rho_{prior}(\mu) \propto \exp\left(-\frac{1}{2} (\mu_{prior} - \mu)^2 / \sigma_{prior}^2\right)$$

In this context, the posterior distribution is still Gaussian, $N(\mu_{post}, \sigma_{post}^2)$ with (verify this!)

$$\mu_{post} = \frac{\mu_{prior} \frac{1}{\sigma_{prior}^2} + \frac{M}{\sigma^2} \bar{\mu}}{\frac{1}{\sigma_{prior}^2} + \frac{M}{\sigma^2}}$$

and

$$\sigma_{post}^2 = \left(\frac{1}{\sigma_{prior}^2} + \frac{M}{\sigma^2} \right)^{-1}$$

Data with different error bars

Consider the previous example, and now suppose that the data were obtained from several laboratories using equipment of varying sophistication. How should we then combine the evidence from observations of differing quality? Let us assume that the **additive measurement error** can still be modeled through a Gaussian pdf, so that the probability of the k th datum having a value x_k is

$$\rho(x_k | \mu, \sigma_k) = \frac{1}{\sigma_k \sqrt{2\pi}} \exp(-(x_k - \mu)^2 / (2\sigma_k^2))$$

Exercise 2.27 (Data with different error bars)

Obtain the MAP estimate of μ :

$$\hat{\mu} = \frac{\sum_k w_k x_k}{\sum_k w_k},$$

with $w_k = 1/\sigma_k^2$.

Moreover, show that the error bar is now proportional to

$$\frac{1}{\sqrt{\sum_k w_k}}.$$

Exercise 2.28

Propose a Bayesian estimation procedure for the Fatigue model.

Conjugate Distributions

Definition 2.47

If the posterior distributions $p(\theta|X)$ are in the same family as the prior probability distribution $p(\theta)$, the prior and posterior are then called *conjugate distributions*, and the prior is called a conjugate prior for the likelihood.

Example 2.48

The Gaussian family is conjugate to itself (self-conjugate) with respect to a Gaussian likelihood.

Hint: Recall Example 2.46.

Observe: A conjugate prior is an algebraic convenience, giving a closed-form expression for the posterior: *otherwise a difficult numerical integration may be necessary.*

Example 2.49

Consider $X \sim Ber(q)$ and the prior for $q \in [0, 1]$ being $Beta(\alpha, \beta)$ i.e.
 $p(q) = q^{\alpha-1}(1-q)^{\beta-1}$ with hyperparameters $0 < \alpha, \beta$.

What is the posterior pdf for X after M iid samples?

See more in

http://en.wikipedia.org/wiki/Conjugate_prior

http://www.johndcook.com/conjugate_prior_diagram.html

Bayesian approach versus Inverse Problems.

- ▶ **General Idea:** relate physical parameters m , which characterize a model, to selected observations d

$$d = G(m)$$

Classical approach: assume that exists a fixed m_{true} s.t.

$$d_{true} = G(m_{true})$$

We are given an actual noisy data set d ,

$$d = d_{true} + e$$

Then,

- ▶ **Inverse problem:** try to recover m_{true} given d .

Inverse problems - Bayesian approach - Principles

- ▶ All variables included in the model are random variables.
- ▶ The randomness describes our degree of information concerning their realizations.
- ▶ The degree of information is expressed in probability distributions.
- ▶ The solution of the inverse problem is a -posterior- probability distribution.

Here we can see the difference between the classical approach:

probability distributions vs. fixed values

Bayesian setting

We are measuring a quantity

$$d \in \mathbb{R}^d$$

in order to get information about another quantity

$$m \in \mathbb{R}^m$$

We relate these two quantities, using a model for their dependence.

We have 2 sources of uncertainty:

- ▶ the model may contain parameters that are not well known
- ▶ the measured quantity d always contains noise

Bayesian setting

We can write a model of the form

$$D = G(M, E)$$

where $G : \mathbb{R}^m \times \mathbb{R}^k \rightarrow \mathbb{R}^d$ and $E \in \mathbb{R}^k$ is a random vector containing all the poorly known parameters as well as the measurement noise.

A priori distributions

Assume that we have previous (before the sample is seen) knowledge about m coded as an *a priori* distribution, $P_p(m)$.

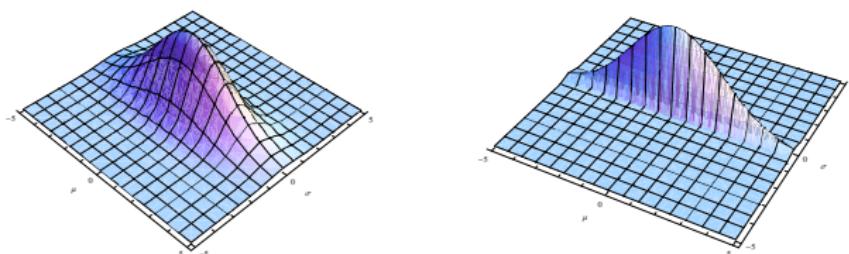
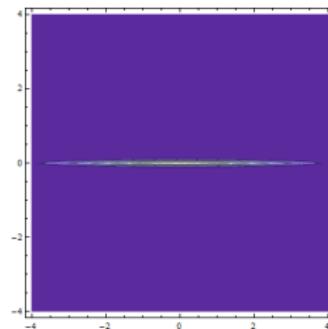
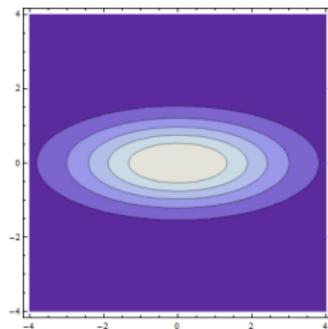


Figure: A priori information of m

Joint and Conditional distributions

Given that $M = m$, the conditional distribution of D is defined as

$$P(d|m) = \frac{P(d, m)}{P_p(m)}, \quad P_p(m) \neq 0$$

which is called the *likelihood* of the data. Finally, assume that the data d is given. The conditional distribution

$$P(m|d) = \frac{P(d, m)}{P(d)}, \quad P(d) = \int P(d, m)dm \neq 0$$

is called the *posterior* distribution.

Bayesian Inverse problem

The inverse problem in the Bayesian framework is

Given the data d , find $P(m|d)$

where

$$P(m|d) = \frac{P(d|m)P_p(m)}{P(d)}$$

which is called Bayes rule (update).

Bayesian Point Estimators

With the known posterior distribution, one can calculate *point* estimates.

Beware: We need to know that the posterior distribution is unimodal when doing this! Why?

1. Maximum a posteriori estimate (MAP): Given the data and the prior, what is the most probable value of m ?

$$m_{MAP} = \arg \max_m P(m|d)$$

possibly nonunique or nonexistent. (optimization problem).

2. Maximum likelihood (ML): Which value of m is most likely to produce the measured data d ?

$$m_{ML} = \arg \max_m P(d|m)$$

if exists.

3. Conditional mean: Mean of the posterior distribution.

$$m_{CM} = \mathbb{E}\{(m|d)\} = \int mP(m|d)dm.$$

Remark 2.10 (On choosing priors)

Choosing a prior distribution may be trickier than it seems. Have a look at

- ▶ *Chapter 5 from Sivia's book and*
- ▶ *"Penalising model component complexity: A principled, practical approach to constructing priors", by D. P. Simpson, H. Rue, T. G. Martins, A. Riebler, S. H. Sørbye. arXiv:1403.4630, 2015. In Journal of Statistical Science, Vol. 32(1), 2017.*

Bayesian setting

Exercise 2.29 (Connection to Tikhonov's regularization)

In a Bayesian setting, assume a Gaussian likelihood and a Gaussian prior for the parameters. Denoting the available iid data by (y_i) the data residuals of the model g are

$$r_i(\theta) = y_i - g(\theta).$$

Show that the max posterior likelihood estimator for the parameter θ solves

$$\hat{\theta} = \arg \min_{\theta} \sum_i r_i^T(\theta) C r_i(\theta) + (\theta - \mu_{prior})^T \Lambda (\theta - \mu_{prior}).$$

Identify the constant matrices C and Λ and the vector μ_{prior} . Make a connection with Tikhonov's regularization for inverse problems.

Finite Difference Method

We often wish to determine numerical approximations to partial differential equations, such as

- ▶ the solution $u: \overline{D} \rightarrow \mathbb{R}$ to an elliptic PDE such as

$$-\operatorname{div}(A \nabla u) = f \quad \text{in } D \subset \mathbb{R}^d ,$$

equipped with boundary conditions on ∂D for some source function $f: D \rightarrow \mathbb{R}$, where $A: D \rightarrow \mathbb{R}^{d \times d}$ is a suitable matrix valued function,

- ▶ or the solution $u: [0, T] \times \overline{D} \rightarrow \mathbb{R}$ that solves

$$\partial_t u - Lu = f \quad \text{in } (0, T) \times D , \quad T > 0 , \quad D \subset \mathbb{R}^d ,$$

equipped with an initial (or final) condition at time $t = 0$ ($t = T$) and boundary conditions on ∂D where L is an differential operator only acting on the “spatial” variable $x \in \mathbb{R}^d$.

Finite difference methods (FDM) are a simple and widely used tool for discretizing differential equations. In fact, the ideas apply to both temporal and spatial derivatives.

Objectives of this part:

1. introduce FDMs for approximating of (partial) differential equations
2. introduce theoretical concepts that offer a basic understanding of accuracy and stability for approximations of differential equations.
3. comment on the limitations of FDMs for spatial variables

We begin with finite differences for functions of one variable.

Definition 3.1

For some $n \in N$, let $\alpha = (\alpha_1, \dots, \alpha_n)^T$ and $\mathbf{h} = (h_1, \dots, h_n)^T \in \mathbb{R}^n$, and further let $f \in C^k(\mathbb{R})$ for $k \in N$. We then call

$$\mathfrak{D}_{\alpha}^{\mathbf{h}}[f](x) := \sum_{j=1}^n \alpha_j f(x + h_j)$$

finite difference of order p for the k -th derivative of f at $x \in \mathbb{R}$, if

$$\mathfrak{D}_{\alpha}^{\mathbf{h}}[f](x) = f^{(k)}(x) + \mathcal{O}(h^p)$$

with $h = \max_{j=1, \dots, n} |h_j|$.

Example 3.2

- The finite differences

$$\frac{f(x+h) - f(x)}{h} \quad \text{and} \quad \frac{f(x) - f(x-h)}{h}$$

are called *forward difference* and *backward difference*, resp. If $f \in C^2(\mathbb{R})$, then it follows from a Taylor expansion that

$$f(x+h) = f(x) + hf'(x) + \mathcal{O}(h^2) \Rightarrow \frac{f(x+h) - f(x)}{h} = f'(x) + \mathcal{O}(h).$$

Analogously, we find that

$$\frac{f(x) - f(x-h)}{h} = f'(x) + \mathcal{O}(h).$$

Both finite differences are therefore of first order for f' .

- The finite difference

$$\frac{f(x+h) - f(x-h)}{2h}$$

is called *central difference* for f' . It is of order 2, if $f \in C^3(\mathbb{R})$.

Example 3.2 (cont.)

- If $f \in C^4(\mathbb{R})$, then

$$f(x+h) = f(x) + hf'(x) + \frac{1}{2}h^2f''(x) + \frac{1}{6}h^3f'''(x) + \mathcal{O}(h^4),$$

$$f(x-h) = f(x) - hf'(x) + \frac{1}{2}h^2f''(x) - \frac{1}{6}h^3f'''(x) + \mathcal{O}(h^4).$$

Rearranging terms, we eventually find

$$\frac{f(x+h) - 2f(x) + f(x-h)}{h^2} = f''(x) + \mathcal{O}(h^2),$$

which is a finite difference of order 2 for f'' and is called *second order central difference*.

Remark 3.1

1. One can generalize the Taylor expansion approaches above and derive a generic ansatz for the construction of finite differences. Indeed, for $f \in C^{m+1}(\mathbb{R})$ we have

$$f(x + h_j) = \sum_{l=0}^m \frac{f^{(l)}(x)}{l!} (h_j)^l + \mathcal{O}((h_j)^{m+1}).$$

Replacing $f(x + h_j)$ in the definition of the finite difference operator \mathfrak{D}_α^h by the truncated Taylor expansion thus yields

$$\mathfrak{D}_\alpha^h[f](x) = \sum_{l=0}^m f^{(l)}(x) \sum_{j=1}^n \frac{1}{l!} \alpha_j (h_j)^l + \mathcal{O}\left(\sum_{j=1}^n \alpha_j (h_j)^{m+1}\right).$$

If $\alpha_j, h_j \in \mathbb{R}$, $j = 1, \dots, n$, satisfy the conditions

$$\sum_{j=1}^n \frac{1}{l!} \alpha_j (h_j)^l = \delta_{l,k} \quad \text{for } l = 0, \dots, m$$

and if $m \geq k$, then $\mathfrak{D}_\alpha^h[f](x)$ is a finite difference for the k -th derivative of f (Question: what is the order?).

Remark (cont.)

Example: $m = 1, n = 2, h_1 = 0, h_2 = h$ yields the conditions

$$\begin{aligned}l = 0 : \quad \alpha_1 + \alpha_2 &= 0, \\l = 1 : \quad \alpha_2 h &= 1.\end{aligned}$$

The solution is $\alpha_1 = -\frac{1}{h}$ and $\alpha_2 = \frac{1}{h}$, which is the forward finite difference.

- Usually one takes h_j equidistant with $m \in N_0$ steps “before” of x (i.e., left of it) and $n \in N_0$ steps “after” x , so that

$$\mathbf{h} = (-mh, -(m-1)h, \dots, -h, 0, h, \dots, (n-1)h, nh)^T \in \mathbb{R}^{n+m+1}$$

for some $h \in (0, \infty)$. In that case we find

$$\mathfrak{D}_\alpha[f](x) \equiv \mathfrak{D}_\alpha^h[f](x) := \sum_{j=-m}^n \alpha_j f(x + jh).$$

If $m = n$, then the finite difference is called central difference, while if $m = 0$ or $n = 0$ are called one-sided difference. Sometimes one calls $m = 0$ right-sided difference and $n = 0$ left-sided difference.

Exercise 3.1

Show that the following expressions hold under suitable regularity requirements on f and state these requirements for each finite difference:

$$\frac{1}{2h} (f(x - 2h) - 4f(x - h) + 3f(x)) = f'(x) + \mathcal{O}(h^2),$$

$$\frac{1}{h^2} \left(\frac{7}{54} f(x-2h) + \frac{81}{110} f(x-\frac{2}{3}h) - \frac{640}{297} f(x+\frac{1}{4}h) + \frac{58}{45} f(x+h) \right) = f''(x) + \mathcal{O}(h).$$

Remark 3.2

Although we will use an elliptic PDE as a running example for the application of an FDM in what follows, the idea of FDMs is not limited to these problems. In fact, the underlying principle of FDMs (i.e., an appropriate Taylor expansion) is very general and can be readily applied to discretizing temporal derivatives in PDEs or in ODEs. FDMs tailored to temporal variables are also called time marching (or time stepping) schemes.

FDM for a PDE

We now use the FDM to approximate the solution to a PDE. We illustrate the idea exemplary for the Poisson problem

$$-\Delta u = f, \quad \text{in } D := (0, 1)^2 \subset \mathbb{R}^2$$

with Dirichlet boundary conditions $u = g$ on ∂D . Specifically, we discretize the PDE by using the point-wise approximations $u_{i,j} \approx u(x_i, y_j)$ for the unknowns on a uniform Cartesian grid

$$x_i = i h, \quad y_j = j h, \quad i, j = 0, 1, \dots, N+1, \quad h = \frac{1}{N+1} \quad \text{for } N \in \mathbb{N}.$$

We will use the second order central difference for each spatial dimension separately. That is, for every interior point $(x, y) \in D_h := \{(x_i, y_j) : 1 \leq i, j \leq N\}$ we

$$\frac{u(x+h, y) - 2u(x, y) + u(x-h, y)}{h^2} = \partial_{xx} u(x, y) + \mathcal{O}(h^2),$$

for $h > 0$ sufficiently small. We approximate $\partial_{yy} u$ analogously.

Eventually we thus obtain the difference relation

$$(\Delta_h u)(x, y) := \frac{1}{h^2} [u(x-h, y) + u(x+h, y) + u(x, y-h) + u(x, y+h) - 4u(x, y)] ,$$

which approximates the Laplace operator in every interior point $(x, y) \in D_h$, in the sense that

$$(\Delta_h u)(x, y) = (\Delta u)(x, y) + \mathcal{O}(h^2) .$$

The complete difference formula is sometimes summarized via the so-called *stencil*, which for the discrete second order Laplace operator reads

$$[-\Delta_h]_\xi = \frac{1}{h^2} \begin{bmatrix} -1 & & \\ -1 & 4 & -1 \\ & -1 & \end{bmatrix} , \quad \xi \in D_h . \quad (13)$$

We also introduce the notation \overline{D}_h as the union of grid points in D_h and on ∂D (i.e., the grid of \overline{D} not the completion of D_h !).

While the continuous Poisson problem entails finding $u \in C^2(D) \cap C(\overline{D})$ such that

$$\begin{cases} -\Delta u = & \text{in } D \\ u = g & \text{on } \partial D \end{cases},$$

the discretized Poisson problem is characterized by the values at grid points $\xi \in \overline{D}_h$. To formulate it, we first need the following definition.

Definition 3.3

The space

$$l^2(D_h) := \left\{ u: D_h \rightarrow \mathbb{R} \mid \sum_{\xi \in D_h} u(\xi)^2 < \infty \right\}$$

contains all (square summable) *grid functions* defined on the grid D_h .

The space $l^2(\overline{D}_h)$ is defined analogously.

The discretized Poisson problem then reads: find $u_h \in l^2(\overline{D}_h)$, such that

$$\begin{cases} -(\Delta_h u_h)(\xi) = f(\xi) & \text{for } \xi \in D_h \\ u_h(\xi) = g(\xi) & \text{for } \xi \in \overline{D}_h \setminus D_h. \end{cases}$$

The mapping $u_h \mapsto \Delta_h u_h$ is linear in u_h . Consequently, the discretized Poisson problem constitutes a linear system of equations and can be written as matrix vector product

$$A_h \mathbf{u}_h = b_h ,$$

where the vector $\mathbf{u}_h \in \mathbb{R}^{N^2}$ contains the unknowns at the interior grid points.

The particular shape of the matrix $A_h \in \mathbb{R}^{N^2 \times N^2}$ and the right-hand side $b_h \in \mathbb{R}^{N^2}$ depends of the particular labeling of grid points. A commonly used convention is row-wise labeling.

Exercise 3.2

Write down explicitly the matrix A_h and the vector b_h for the discretized Poisson equation above.

Convergence and Accuracy

Linear PDEs of the (abstract) form

$$Lu = f$$

will lead to linear systems of equations of the form

$$L_h \mathbf{u}_h = f_h$$

when discretizing with the FDM. This is also true for many time-dependent PDEs, where the FDM is applied to both temporal and spatial variables, possibly with different orders.

After discretizing a continuous equation the key **questions** are:

- ▶ Does the solution of the discretized system converge to the solutions of the continuous problem?
- ▶ Are there any constraints on the choice of finite difference mesh/step sizes **h** to achieve the convergence?
- ▶ If it converges, what is the speed of such convergence with respect to the mesh/step sizes **h**?

Assuming that no other approximation errors (e.g., discretization of boundaries or truncation of domains) are present, the usual idea to prove convergence is based on the following formal decomposition:

$$\|u|_{\bar{D}_h} - u_h\| \leq \|L_h^{-1}\| \|L_h(u - u|_{\bar{D}_h})\| \leq \|L_h^{-1}\| \left(\|f - f_h\| + \|L_h u - f|_{D_h}\| \right).$$

That is, we see that the discretized solution converges, if the continuous problem has a unique solution and if

- ▶ the approximate solution operators $\|L_h^{-1}\|$ are uniformly bounded in h (**stability**),
- ▶ $f_h \rightarrow f$ and $L_h u \rightarrow Lu = f$ as the mesh size $h \rightarrow 0$ (**consistency**).

In other words: **stability and consistency imply convergence**. Moreover, the estimate also directly implies that if a stable FDM is *consistent of order* $p > 0$, in the sense that if

$$\max\{\|f - f_h\|, \|L_h u - f|_{D_h}\|\} \leq Ch^p,$$

then the FDM converges with the same order.

Notice however, that these statements are norm-dependent!

Lax Equivalence Theorem

With a bit of more work, one can even show that the implication stated on the previous slide is actually an equivalence. That particular result is typically called *Lax equivalence theorem*, which is a fundamental theorem in the analysis of finite difference methods for the numerical solution of PDEs. In its classical form it is stated for linear initial value problems and states that a consistent FDM for a well-posed linear initial value problem is convergent if and only if it is stable.

Here, “well posed” means that the equation is solvable for data in a suitable function space and that the solution operator is bounded, that is, the solution depends continuously on the data.

We will first formally state the result without being mathematically precise with function spaces and norms. Then we discuss examples for which we also address the norms and functions spaces.

The ingredients of Lax Equivalence Theorem 3.1 below are:

- (0) an exact solution u , satisfying the *linear well posed equation* $Lu = f$, and an approximation u_h , obtained from $L_h u_h = f_h$;
- (1) *stability*, the approximate solution operators $\|L_h^{-1}\|$ are uniformly bounded in h and the exact solution operator $\|L^{-1}\|$ is bounded;
- (2) *consistency*, $f_h \rightarrow f$ and $L_h u \rightarrow Lu$ as the mesh size $h \rightarrow 0$; and
- (3) *convergence*, $u_h \rightarrow u$ as the mesh size $h \rightarrow 0$.

Theorem 3.1 (Lax Equivalence)

The combination of stability and consistency is equivalent to convergence.

Proof idea.

To verify the convergence, we consider the identity

$$u - u_h = L_h^{-1} [L_h u - L_h u_h] \stackrel{\text{Step}(0)}{=} L_h^{-1} [(L_h u - Lu) + (f - f_h)].$$

Stability implies that L_h^{-1} is bounded and consistency implies that $L_h u - Lu \rightarrow 0$ and $f - f_h \rightarrow 0$, and consequently the convergence holds

$$\lim_{h \rightarrow 0} (u - u_h) = \lim_{h \rightarrow 0} L_h^{-1} [(L_h u - Lu) + (f - f_h)] = 0 .$$

Clearly, consistency is necessary for convergence. Examples indicate that also stability is necessary. □

Exercise 3.3 (Convergence of discrete Poisson solution)

Consider the FDM for the Poisson problem derived above. A natural norm to study the discretized Poisson problem in $l^2(D_h)$ is given by

$$\langle u, v \rangle_h := \sum_{\xi \in D_h} u(\xi)v(\xi) , \quad \|u\|_2 := \sqrt{\langle u, u \rangle_h} ;$$

analogously for $l^2(\overline{D}_h)$.

1. Explain why the discrete Laplace operator $\Delta_h: l^2(\overline{D}_h) \rightarrow l^2(D_h)$ given by

$$u \mapsto (D_h \rightarrow \mathbb{R}, \xi \mapsto (\Delta_h u)(\xi)) ,$$

cannot be invertible. Hint: what is the dimension of the spaces?

2. (Optional) Show that the discrete Laplace operator restricted to the subspace

$$l_0^2(\overline{D}_h) := \{v_h \in l^2(\overline{D}_h): v_h(\xi) = 0 \text{ for all } \xi \in \overline{D}_h \setminus D_h\} \subset l^2(\overline{D}_h) ,$$

that is $\Delta_h: l_0^2(\overline{D}_h) \rightarrow l^2(D_h)$, is invertible.

Exercise (cont.)

3. Show that the FDM for the Poisson problem is first order consistent in the $\|\cdot\|_2$ norm. Detail also the regularity assumption on the exact solution.
4. Show that the FDM is stable in the $\|\cdot\|_2$ norm and give a reasonable upper bound. Hint: you can use that the eigenvalues of the discrete Laplace operator are given by

$$\lambda_{\nu,\mu} = \frac{4}{h^2} (\sin^2(\frac{1}{2}\pi\nu h) + \sin^2(\frac{1}{2}\pi\mu h)) ,$$

for $1 \leq \nu, \mu \leq N$.

5. (Optional) The FDM is also convergent in the sup-norm given by $\|u\|_\infty := \max_{\xi \in D_h} |u(\xi)|$ for any $u \in l^2(D_h)$; analogously for $l^2(\overline{D}_h)$. Do you expect a faster or slower rate of convergence? Explain your arguments.

Example 3.4 (Forward Euler for initial value problem)

Consider the forward FDM (Euler method) for the ODE

$$\begin{aligned} u'(t) &= Au(t) \quad 0 < t < 1, \\ u(0) &= u_0. \end{aligned} \tag{14}$$

Verify the stability and consistency conditions in the Lax Equivalence Theorem.

Solution: For a given partition, $0 = t_0 < t_1 < \dots < t_N = 1$, with $\Delta t = t_{n+1} - t_n$, let

$$\begin{aligned} u_{n+1} &\equiv (I + \Delta t A)u_n \\ &= G^n u_0 \quad \text{where } G = (I + \Delta t A). \end{aligned}$$

Then:

1. Stability means $|G^n| + |H^n| \leq e^{Kn\Delta t}$ for some K , where $H = e^{\Delta t A}$ and $|\cdot|$ denotes the matrix norm $|F| \equiv \sup_{\{v \in \mathbb{R}^n : |v| \leq 1\}} |Fv|$ with the Euclidean norm $|w| \equiv \sqrt{\sum_i w_i^2}$ in \mathbb{R}^n .
2. Consistency means $|(G - H)v| \leq C(\Delta t)^{p+1}$, where $H = e^{\Delta t A}$ and p is the order of accuracy. In other words, the consistency error $(G - H)v$ is the local approximation error after one time step with the same initial data v .

This stability and consistency imply the convergence

$$\begin{aligned} | u_n - u(n\Delta t) | &= | (G^n - H^n)u_0 | \\ &= | (G^{n-1} + G^{n-2}H + \dots + GH^{n-2} + H^{n-1})(G - H)u_0 | \\ &\leq | G^{n-1} + G^{n-2}H + \dots + GH^{n-2} + H^{n-1}| |(G - H)u_0| \\ &\leq C(\Delta t)^{p+1} n | u_0 | e^{Kn\Delta t} \\ &\leq C'(\Delta t)^p, \end{aligned}$$

with the convergence rate $\mathcal{O}(\Delta t^p)$. For example, $p = 1$ in case of the Euler method and $p = 2$ in case of the trapezoidal method.

Exercise 3.4 (Heat equation)

Consider the heat equation

$$\begin{aligned} u_t &= u_{xx} \quad t > 0, x \in \mathbb{R} \\ u(0) &= u_0, x \in \mathbb{R}. \end{aligned} \tag{15}$$

Verify the stability and consistency conditions in Lax Equivalence Theorem.

Solution: First, apply the Fourier transform to equation (15),

$$\hat{u}_t = -\omega^2 \hat{u}$$

so that

$$\hat{u}(t, \omega) = e^{-t\omega^2} \hat{u}_0(\omega).$$

Therefore $\hat{H} = e^{-\Delta t \omega^2}$ is the exact solution operator for one time step, i.e. $\hat{u}(t + \Delta t) = \hat{H}\hat{u}(t)$. Consider the difference approximation of (15)

$$\frac{u_{n+1,i} - u_{n,i}}{\Delta t} = \frac{u_{n,i+1} - 2u_{n,i} + u_{n,i-1}}{\Delta x^2},$$

which shows

$$u_{n+1,i} = u_{n,i} \left(1 - \frac{2\Delta t}{\Delta x^2} \right) + \frac{\Delta t}{\Delta x^2} (u_{n,i+1} + u_{n,i-1}),$$

where $u_{n,i} \simeq u(n\Delta t, i\Delta x)$. Apply the Fourier transform to obtain

$$\begin{aligned}\hat{u}_{n+1} &= \left[\left(1 - \frac{2\Delta t}{\Delta x^2} \right) + \frac{\Delta t}{\Delta x^2} (e^{j\Delta x \omega} + e^{-j\Delta x \omega}) \right] \hat{u}_n \\ &= \left[1 - 2\frac{\Delta t}{\Delta x^2} + 2\frac{\Delta t}{\Delta x^2} \cos(\Delta x \omega) \right] \hat{u}_n \\ &= \hat{G} \hat{u}_n \quad (\text{Let } \hat{G} \equiv 1 - 2\frac{\Delta t}{\Delta x^2} + 2\frac{\Delta t}{\Delta x^2} \cos(\Delta x \omega)) \\ &= \hat{G}^{n+1} \hat{u}_0.\end{aligned}$$

1. We have

$$\begin{aligned} 2\pi \|u_n\|_{L^2}^2 &= \|\hat{u}_n\|_{L^2}^2 \quad (\text{by Parseval's formula}) \\ &= \|\hat{G}^n \hat{u}_0\|_{L^2}^2 \\ &\leq \sup_{\omega} |\hat{G}^n|^2 \|\hat{u}_0\|_{L^2}^2. \end{aligned}$$

Therefore the condition

$$\|\hat{G}^n\|_{L^\infty} \leq e^{Kn\Delta t} \tag{16}$$

implies L^2 -stability.

2. We have

$$2\pi \|u_1 - u(\Delta t)\|_{L^2}^2 = \|\hat{G} \hat{u}_0 - \hat{H} \hat{u}_0\|_{L^2}^2,$$

where u_1 is the approximate solution after one time step. Let $\lambda \equiv \frac{\Delta t}{\Delta x^2}$, then we obtain

$$\begin{aligned} |(\hat{G} - \hat{H})\hat{u}_0| &= \left| \left(1 - 2\lambda + 2\lambda \cos \Delta x \omega - e^{-\Delta t \omega^2} \right) \hat{u}_0 \right| \\ &= \mathcal{O}(\Delta t^2) \omega^4 |\hat{u}_0|, \end{aligned}$$

since for $0 \leq \Delta t \omega^2 \equiv x \leq 1$

$$\begin{aligned} & |1 - 2\lambda + 2\lambda \cos \sqrt{x/\lambda} - e^{-x}| \\ &= \left(1 - 2\lambda + 2\lambda \left(1 - \frac{x}{2\lambda} + \mathcal{O}(x^2) \right) - (1 - x + \mathcal{O}(x^2)) \right) \\ &\leq Cx^2 = C(\Delta t)^2 \omega^4, \end{aligned}$$

and for $1 < \Delta t \omega^2 = x$

$$|1 - 2\lambda + 2\lambda \cos \sqrt{x/\lambda} - e^{-x}| \leq C = C \frac{(\Delta t)^2 \omega^4}{x^2} \leq C(\Delta t)^2 \omega^4.$$

Therefore the consistency condition reduces to

$$\| (\hat{G} - \hat{H}) \hat{u}_0 \| \leq \| K \Delta t^2 \omega^4 \hat{u}_0 \| \leq K \Delta t^2 \| \partial_{xxxx} u_0 \|_{L^2}. \quad (17)$$

3. The stability (16) holds if

$$\| \hat{G} \|_{L^\infty} \equiv \sup_{\omega} |\hat{G}(\omega)| = \max_{\omega} |1 - 2\lambda + 2\lambda \cos \Delta x \omega| \leq 1, \quad (18)$$

which requires

$$\lambda = \frac{\Delta t}{\Delta x^2} \leq \frac{1}{2}.$$

The L^2 -stability condition (18) is called the von Neuman stability condition.

4. Convergence follows by the estimates (17), (18) and $\|\hat{H}\|_{L^\infty} \leq 1$

$$\begin{aligned} 2\pi \| u_n - u(n\Delta t) \|_{L^2}^2 &= \| (\hat{G}^n - \hat{H}^n)\hat{u}_0 \|_{L^2}^2 \\ &= \| (\hat{G}^{n-1} + \hat{G}^{n-2}\hat{H} + \dots + \hat{H}^{n-1})(\hat{G} - \hat{H})\hat{u}_0 \|_{L^2}^2 \\ &\leq \| \hat{G}^{n-1} + \hat{G}^{n-2}\hat{H} + \dots + \hat{H}^{n-1} \|_{L^\infty}^2 \| (\hat{G} - \hat{H})\hat{u}_0 \|_{L^2}^2 \\ &\leq (Kn(\Delta t)^2)^2 \leq (KT\Delta t)^2, \end{aligned}$$

and consequently the convergence rate is $\mathcal{O}(\Delta t)$.

Limitations for FDMs for spatial variables

There are several limitations and shortcoming when using FDMs in spatial variables:

- ▶ The extension to non-Cartesian domains may be difficult: points close to the boundary require special treatment.
- ▶ Different types of boundary conditions, e.g., Neumann conditions, are not straightforward. For example, the condition $n \cdot \nabla u = g_N$ on ∂D would need to be discretized with appropriate special finite differences and would require an equation for the boundary values of $u_{i,j}$.
- ▶ Higher order of consistency $\mathcal{O}(h^p)$ with $p > 2$ requires a larger stencil, so that \mathbf{A}_h becomes denser. This also implies even more problems close to the boundary.
- ▶ The convergence theory requires very strong conditions on the regularity (i.e., smoothness) of the exact solution. For example, even for simple Poisson problems on $D = (0, 1)^2$ we require $u \in C^4(D)$.
- ▶ Non-smooth boundaries (in particular re-entrant corners) may only have an exact solution $u \notin C^1(D)$.

- Discretizing a self-adjoint operator with the FDM may not yield a symmetric system matrix A_h with small stencil.

Example 3.5

Consider

$$-(a(x)u'(x))' = f(x) \quad \text{for } x \in (0, 1).$$

There are various natural ways of discretizing the PDE via the FDM:

1. If, before discretizing, we rewrite the equation, then we loose symmetry. Indeed, we obtain

$$-(a(x)u'(x))' = -a'(x)u'(x) - a(x)u''(x),$$

which is an diffusion advection equation with space dependent coefficients. For these type of problems one has to consider the sign of $a'(x)$ and use an “upwind” direction with an one-sided finite difference, which yields a non-symmetric matrix.

Example 3.5 (cont.)

- Let's naively discretize sequentially from the outside to the inside

$$\begin{aligned} -(a(x)u'(x))' &\approx -\frac{1}{2h} (a(x+h)u'(x+h) - a(x-h)u'(x-h)) \\ &\approx -\frac{1}{2h} \left(a(x+h) \frac{u(x+2h) - u(x)}{2h} - a(x-h) \frac{u(x) - u(x-2h)}{2h} \right) \end{aligned}$$

Although this gives a symmetric matrix, it results in a 5-point stencil (in 1d!) that even leaves a few points out (irregular). Moreover, we know that 5-point stencils may require extra care close to the boundary.

- It is better to discretize on shifted grids:

$$\begin{aligned} -(a(x)u'(x))' &\approx -\frac{1}{h} (a(x+\frac{h}{2})u'(x+\frac{h}{2}) - a(x-\frac{h}{2})u'(x-\frac{h}{2})) \\ &\approx -\frac{1}{h} \left(a(x+\frac{h}{2}) \frac{u(x+h) - u(x)}{h} - a(x-\frac{h}{2}) \frac{u(x) - u(x-h)}{h} \right), \end{aligned}$$

which gives a 3-point stencil upon evaluating the pace dependent coefficient a in the shifted grid $x \pm \frac{h}{2}$ instead of $x \pm h$ and yields a symmetric matrix.

Finite Difference Methods, further reading

- ▶ *Numerical treatment of partial differential equations.* Grossmann, Ross, Stynes. Springer, 2007.
- ▶ *Analysis of Finite Difference Schemes.* Jovanović, Süli. Springer. 2014.
- ▶ *Finite Difference Methods for Ordinary and Partial Differential Equations.* LeVeque. SIAM, 2007.

Finite Element Method

The Finite Element Method (FEM) allows to remedy many shortcoming of the FDM for spatial variables. As a matter of fact, the FEM provides a general and efficient computational framework for solving elliptic and parabolic PDEs with computational simplicity and efficiency, allowing for construction of stable higher order discretizations.

Objectives of this part:

1. introduce basics idea of the FEM
2. address a few theoretical aspects of underlying variational problems for elliptic PDEs
3. discuss a basic adaptive approximation and error estimation procedure

Remark 4.1

We focus our FEM discussion on elliptic problems here. However, based on a semi-discretization (“method of lines”) it also offers a convenient for many time dependent differential equations:

$$\partial_t u = Lu \quad \rightsquigarrow \quad \dot{\mathbf{u}} = L_h \mathbf{u}$$

Motivation

Consider a one dimensional elliptic model problem for given coefficient functions $a, f, r : (0, 1) \rightarrow \mathbb{R}$, and solution $u : [0, 1] \rightarrow \mathbb{R}$:

$$\begin{aligned} (-au')' + ru &= f && \text{on } (0, 1) \\ u(x) &= 0 && \text{for } x \in \{0, 1\};, \end{aligned} \tag{19}$$

where $a > 0$ and $r \geq 0$.

The existence (and regularity) of a classical solution depends on the regularity of the coefficient functions. For example, if $a \notin C^1(0, 1)$, then (19) cannot support a classical solution. To remedy this, we will introduce a weaker (variational) solution concept.

The fundamental theoretical result for well-posedness, in the sense of existence, uniqueness, and continuous dependence, on which all results for general linear elliptic differential equations are based, is the

Lax–Milgram theorem. We will describe this essential result later on.

We will see that its stability properties, based on so called energy estimates, is automatically satisfied for finite element methods in contrast to finite difference methods.

Eventually, our goal is to find an approximation u_h of PDEs such as (19) that satisfy

$$\|u - u_h\| \leq \text{TOL},$$

for a given tolerance TOL using few degrees of freedom, assuming the FEM converges.

Tuning the FEM by using a mesh-size such that the tolerance criterion is met at minimal computational cost, requires knowledge of a-priori **convergence rates**.

Alternatively, the tuning can be carried out by means of an **adaptive finite element approximation**. In general, adaptive procedures are based on:

- (1) an automatic mesh generator,
- (2) a numerical method (e.g., the FEM or FDM),
- (3) a refinement criteria (e.g., a posteriori error estimation), and
- (4) a solution algorithm (e.g., a multigrid solver).

The basic FEM workflow

The derivation of a FEM can be divided into following basic building blocks:

- (1) variational formulation in an infinite dimensional space V ,
- (2) variational formulation in a finite dimensional subspace, $V_h \subset V$,
- (3) choice of a basis for V_h , and
- (4) solution of the discrete system of equations.

Remark 4.2

- ▶ Point (2) introduces the actual discretization of the PDE; this step is sometimes called Ritz–Galerkin method. The condition $V_h \subset V$ yields conforming methods, which is often convenient.
- ▶ Point (3) then makes the discretization precise. In particular, for the FEM one uses basis functions that, as the name suggests, have finite support on a small region of the domain D .
- ▶ Many variants and extensions to these four building blocks exist in practice.

FEM building blocks for the 1d example in Eqn. (19)

Step 1. *Variational formulation in an infinite dimensional space, V .*

Let $D := (0, 1)$ and consider the following Hilbert space V as

$$V = \left\{ v \in L^2(D) : \int_D (v^2(x) + (v'(x))^2) dx < \infty, \begin{array}{l} \\ v(0) = v(1) = 0 \end{array} \right\}.$$

Remark 4.3

- ▶ The space V defined above is, of course, the space $H_0^1(D)$.
- ▶ Notice that the assignment $v|_{\partial D} = 0$ is well-defined here, since $V = H_0^1(D) \subset H^1(D) \subset C(D)$ in view of the Sobolev embedding theorem (i.e., trace operator is not needed here).

Multiply equation (19) by $v \in C_0^\infty(D)$ and integrate by parts to get

$$\begin{aligned}\int_0^1 fv \, dx &= \int_0^1 ((-au')' + ru)v \, dx \\ &= [-au'v]_0^1 + \int_0^1 (au'v' + ruv) \, dx \\ &= \int_0^1 (au'v' + ruv) \, dx.\end{aligned}\tag{20}$$

Therefore the variational formulation of (19) is to find $u \in V$ such that

$$A(u, v) = L(v) \quad \forall v \in V, \quad (21)$$

where

$$\begin{aligned} A(u, v) &= \int_0^1 (au'v' + ruv) \, dx, \\ L(v) &= \int_0^1 fv \, dx. \end{aligned}$$

This completes **Step 1.**

Remark 4.4

The integration by parts in (20) shows that a smooth solution of equation (19) satisfies the variational formulation (21). For a solution of the variational formulation (21) to also be a solution of the equation (19), we need additional conditions on the regularity of the functions a , r and f so that u'' is continuous. Integration by parts yields, as in (20),

$$0 = \int_0^1 (au'v' + ruv - fv) \, dx = \int_0^1 (-(au')' + ru - f)v \, dx.$$

Since this holds for all $v \in C_0^\infty(D)$, it implies that

$$-(au')' + ru - f = 0,$$

almost everywhere in D , provided $-(au')' + ru - f$ is continuous.

Definition 4.1

A solution to the variational formulation (21) is called a *weak solution* for (19). If it also satisfies (19) point-wise then is called a *strong solution*.

Exercise 4.1

Consider the equation find $u \in V$ such that

$$A(u, v) = L(v) \quad \forall v \in V,$$

where

$$\begin{aligned} A(u, v) &= \int_0^1 u' v' \, dx, \\ L(v) &= \int_0^1 f v \, dx = v(1/2). \end{aligned}$$

In other words, $f(x) = \delta(1/2 - x)$ is a delta distribution. Find u and show that it is a weak solution for

$$\begin{aligned} -u''(x) &= \delta(x - 1/2), \quad x \in (0, 1) \\ u(0) &= u(1) = 0. \end{aligned}$$

Is u a strong solution? Motivate your answer.

Step 2. *Variational formulation in the finite dimensional subspace:*
 $V_h \subset V$.

To “discretize” the problem, the *Ritz-Galerkin method* replaces V by a finite dimensional subspace

$$V_h \subset V \quad \text{such that} \quad V_h \uparrow V \quad \text{as } h \rightarrow 0$$

and with $\dim(V_h) = N \equiv N_h \rightarrow \infty$ as $h \rightarrow 0$. Consider a basis in V_h , such that

$$V_h = \text{span}\{\phi_1, \dots, \phi_N\},$$

that is, every u_h can be expressed in that basis

$$u_h = \sum_{k=1}^N \xi_k \phi_k \quad \text{for some } \xi_k \in \mathbb{R}.$$

As $V_h \subset V$ it is thus natural to consider the variational formulation restricted to V_h :

$$\text{Find } u_h \in V_h \quad \text{such that } A(u_h, v) = L(v) \quad \text{for all } v \in V_h. \quad (22)$$

To determine the unknown coefficients (ξ_1, \dots, ξ_N) , we also use the basis representations of the test functions, $v = \sum_{k=1}^N \eta_k \phi_k \in V_h$. Hence u_h is a solution of (21) if and only if

$$\begin{aligned} A(u_h, v) &= L(v) \quad \text{for all } v \in V_h \\ \Leftrightarrow \quad \sum_{j=1}^N \xi_j A(\phi_j, \phi_i) &= L(\phi_i) \quad \text{for } i = 1, 2, \dots, N, \end{aligned}$$

which gives a system of linear equations for the unknowns
 $\xi := (\xi_1, \dots, \xi_N)^T \in \mathbb{R}^N$

$$\tilde{A}\xi = \tilde{L}, \tag{23}$$

where

$$\begin{aligned} \tilde{A}_{ij} &= A(\phi_j, \phi_i), \\ \tilde{L}_i &= L(\phi_i). \end{aligned}$$

The $N \times N$ matrix \tilde{A} is called the *stiffness matrix* and the vector $\tilde{L} \in \mathbb{R}^N$ is called the *load vector*.

Important, general observations for Step 2.

- ▶ As $V_h \subset V$ is finite dimensional, it is a closed subspace and therefore a Hilbert space endowed with the (same) inner product of V ; this will be important for the well-posedness analysis later.
- ▶ The discrete variational formulation restricted to a finite dimensional space means that instead of using *all* $v \in V$ it is sufficient to test with the basis elements ϕ_j , $j = 1, 2, \dots, N$.
- ▶ The unknown solution $u_h \in V_h$ of the discrete problem is uniquely determined by the N real values ξ_k (the unknowns), once the basis functions are fixed the basis is fixed.
- ▶ The procedure as well as the comments above are in abstract terms and therefore general. That is, the derivation does not depend on any properties of the 1d toy problem. In fact, **Step 2.** is abstract and thus problem independent.

Step 3. Choose a basis for V_h .

First, divide the interval $(0, 1)$ into $0 = x_0 < x_1 < \dots < x_{N+1} = 1$, i.e. generate the mesh. On that mesh, let us introduce the basis functions $\phi_i \in V_h$, for $i = 1, \dots, N$, defined by the Lagrange basis

$$\phi_i(x_j) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j. \end{cases} \quad (24)$$

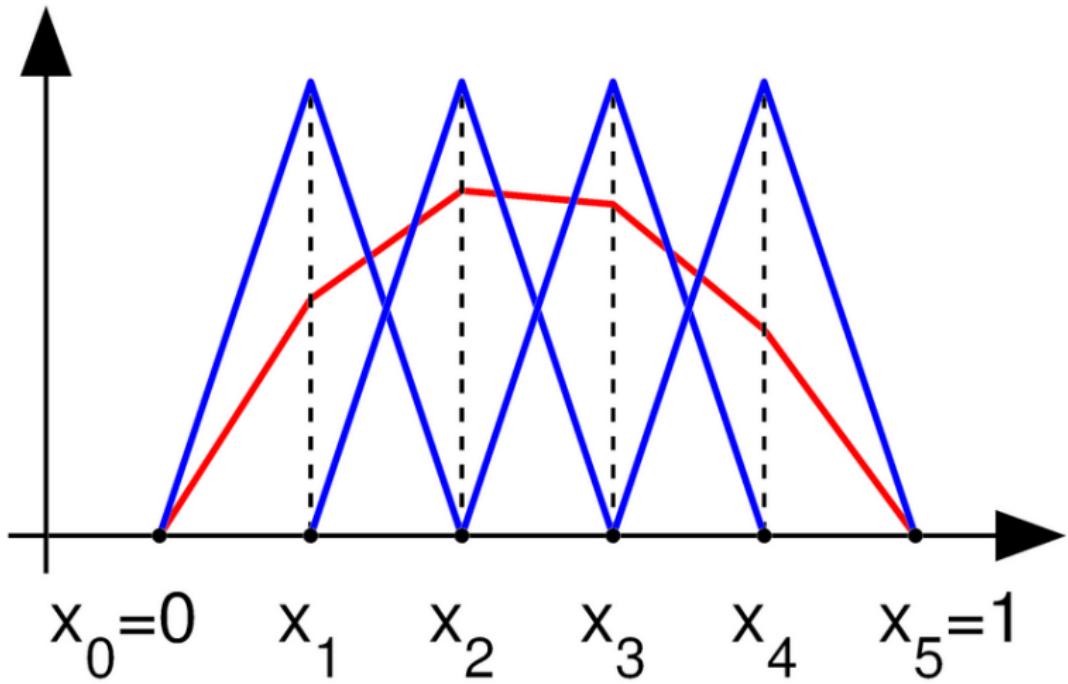
That is $V_h = \text{span}\{\phi_1, \dots, \phi_N\}$ is the space of globally continuous, piece-wise linear functions on the mesh with zero boundary conditions

$$V_h = \{v \in V : v(x) |_{(x_i, x_{i+1})} = c_i x + d_i, \\ \text{i.e. } v \text{ is linear on } (x_i, x_{i+1}), i = 0, \dots, N \\ \text{and } v \text{ is continuous on } (0, 1)\}.$$

A function $v \in V_h$ now has representation with fixed basis of the form

$$v(x) = \sum_{i=1}^N v_i \phi_i(x),$$

where the unknowns $\xi_i := v_i = v(x_i)$ are the *nodal values*. That is, $v \in V_h$ can be written in a unique way as a linear combination of the basis functions ϕ_i .



Example of a piece-wise linear function in 1 dimension and the basis functions.

- ▶ Based on this basis, the function u_h solving (22) is a finite element solution of the equation (19). Other finite element solutions are obtained from alternative finite dimensional subspaces, e.g., based on piece-wise quadratic approximation.

Step 4. *Solve the discrete problem (22).*

Using the basis functions ϕ_i , for $i = 1, \dots, N$ from Step 3, we have

$$u_h(x) = \sum_{i=1}^N \xi_i \phi_i(x),$$

where $\xi = (\xi_1, \dots, \xi_N)^T \in \mathbb{R}^N$. In Step 2, we have seen that $\xi \in \mathbb{R}^N$ solves the linear system

$$\tilde{A}\xi = \tilde{L},$$

where

$$\tilde{A}_{ij} = A(\phi_j, \phi_i), \quad \text{and} \quad \tilde{L}_i = L(\phi_i).$$

This system of equations needs to be solved, which requires (sparse) numerical linear algebra.

Exercise 4.2

Show that for the 1d model problem (19) the stiffness matrix \tilde{A} based on the basis functions above is symmetric positive definite.

Exercise 4.3

Consider the equation find $u \in V$ such that

$$A(u, v) = L(v) \quad \forall v \in V,$$

$$\begin{aligned} A(u, v) &= \int_0^1 u' v' \, dx, \\ L(v) &= \int_0^1 fv \, dx. \end{aligned}$$

Pose the FEM equations corresponding to piece-wise linear basis functions on a uniform mesh with mesh size $h = \Delta x$. Write explicitly the stiffness matrix \tilde{A} and the load vector \tilde{L} . Compare the resulting equations with a finite difference discretization.

Comments on d -dimensional domains

While the theoretical properties of variational problems are, of course, problem dependent, the general FEM workflow is generic. In fact, some steps are problem independent. However the dimensionality of the spatial domain affects some aspects. Clearly the construction of basis functions (even piece-wise linear ones) has to be adapted. But there are also theoretical considerations that go into defining the functions space V . These are a consequence of the following.

Theorem 4.2 (Sobolev embeddings)

Let $D \subset \mathbb{R}^d$ be a Lipschitz domain. If $m, k \in \mathbb{N}_0$ satisfy

$$m - \frac{d}{2} > k ,$$

then $C^\infty(D) \subset H^m(D) \subset C^k(D)$. That is, every function $u \in H^m(D)$ corresponds (almost everywhere) to a function from $C^k(D)$. Moreover the following embedding $\|u\|_{C^\infty(D)} \leq c \|u\|_{H^m(D)}$ holds.

For (unique) solutions of PDEs on bounded domains $D \subset \mathbb{R}^d$, we need to prescribe conditions on the boundary $\partial D \subset \mathbb{R}^{d-1}$. However, it is a-priori not clear if even point evaluations on ∂D for $u \in L^2(D)$ are well-defined, because ∂D is a null-set of D .

Theorem 4.3 (Trace theorem)

Let $D \subset \mathbb{R}^d$ be a Lipschitz domain and $u \in H^1(D)$. The restriction $v : \partial D \rightarrow \mathbb{R}$ of u to the boundary ∂D of D is called trace of u . It holds that $v \in L^2(\partial D)$ and $\|v\|_{L^2(\partial D)} \leq c \|u\|_{H^1(D)}$.

For $m \geq 1$ we write

$$H_0^m(D) := \{u \in H^m(D), u|_{\partial D} = 0 \text{ (in the trace sense)}\}.$$

Remark 4.5

Inhomogeneous Dirichlet boundary conditions can be considered similarly. Alternatively, they can be transformed into homogeneous ones via lifting, if possible, or imposed weakly via Robin conditions.

The FEM workflow for an elliptic 2d example

Consider the following two dimensional problem,

$$\begin{aligned}-\operatorname{div}(k \nabla u) + r u &= f \quad \text{in } D \subset \mathbb{R}^2 \\ u &= g_1 \quad \text{on } \Gamma_1 \\ \frac{\partial u}{\partial n} &= g_2 \quad \text{on } \Gamma_2,\end{aligned}\tag{25}$$

where $\partial D = \Gamma = \Gamma_1 \cup \Gamma_2$ and $\Gamma_1 \cap \Gamma_2 = \emptyset$.

Step 1. Variational formulation in the infinite dimensional space.

Let

$$V_g = \left\{ v(x) : \int_D (v^2(x) + |\nabla v(x)|^2) dx < \infty, v|_{\Gamma_1} = g \right\},$$

for an appropriate function $g: \Gamma_1 \rightarrow \mathbb{R}$ so that the trace exists. Take a function $v \in V_0$, i.e. $v = 0$ on Γ_1 , integrating (25) by parts, we find

$$\begin{aligned}
\int_D fv \, dx &= - \int_D \operatorname{div}(k \nabla u) v \, dx + \int_D ruv \, dx \\
&= \int_D k \nabla u \cdot \nabla v \, dx - \int_{\Gamma_1} k \frac{\partial u}{\partial n} v \, ds - \int_{\Gamma_2} k \frac{\partial u}{\partial n} v \, ds + \int_D ruv \, dx \\
&= \int_D k \nabla u \cdot \nabla v \, dx - \int_{\Gamma_2} kg_2 v \, ds + \int_D ruv \, dx.
\end{aligned}$$

Consequently, the variational formulation for the model problem (25) is to find $u \in V_{g_1}$ such that

$$A(u, v) = L(v) \quad \forall v \in V_0, \tag{26}$$

where

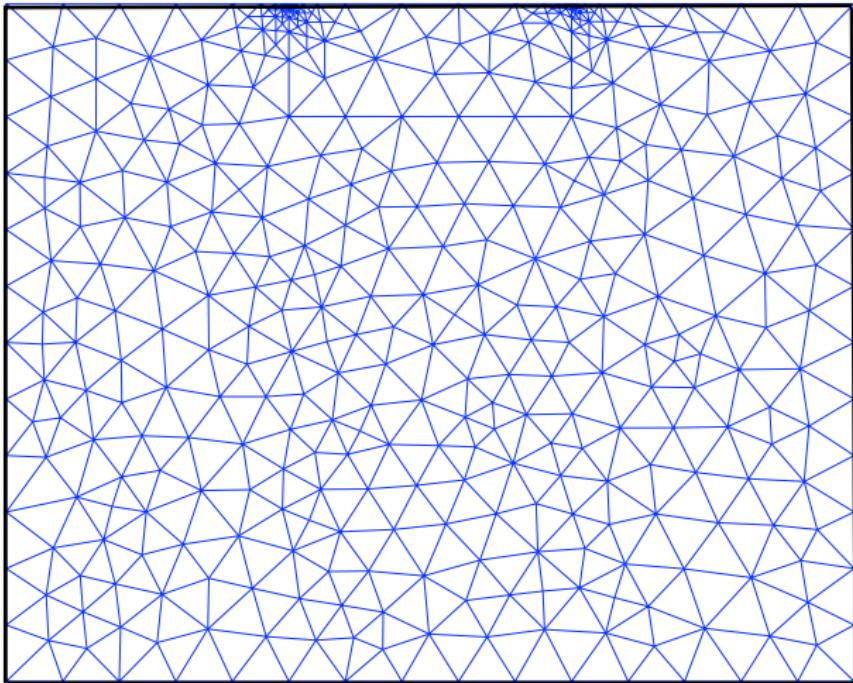
$$\begin{aligned}
A(u, v) &= \int_D (k \nabla u \cdot \nabla v + ruv) \, dx, \\
L(v) &= \int_D fv \, dx + \int_{\Gamma_2} kg_2 v \, ds.
\end{aligned}$$

Steps 2 and 3. Variational formulation in the finite dimensional space and Finite Element basis functions.

Assume for simplicity that D is a polygonal domain which can be divided into a triangular mesh $T_h = \{K_1, \dots, K_N\}$ of non overlapping triangles K_i and let

$$h = \max_i(\text{length of longest side of } K_i) .$$

Assume also that the boundary function g_1 is continuous and that its restriction to each edge $K_i \cap \Gamma_1$ is a linear function.



Example of a triangular mesh.

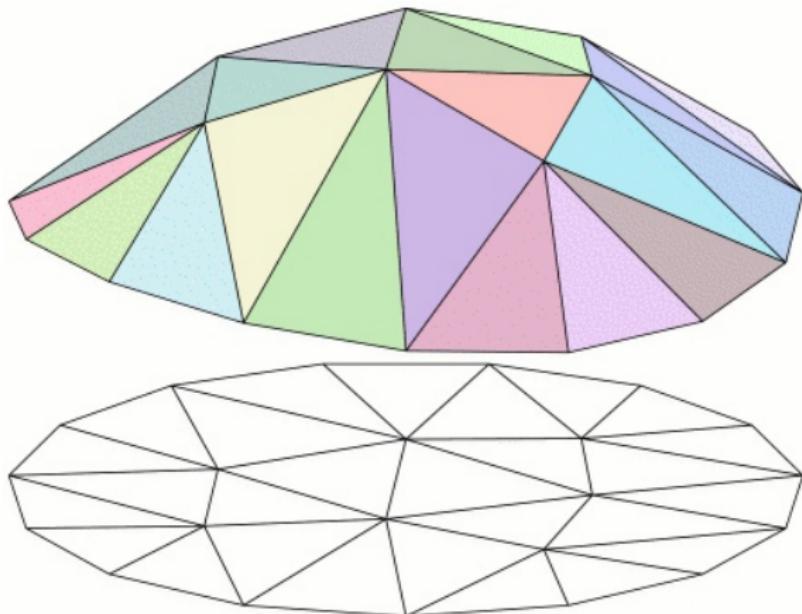
Define

$$V_0^h = \{v \in V_0 : v|_{K_i} \text{ is linear } \forall K_i \in T_h, \\ v \text{ is continuous on } D\},$$

$$V_{g_1}^h = \{v \in V_{g_1} : v|_{K_i} \text{ is linear } \forall K_i \in T_h, \\ v \text{ is continuous on } D\},$$

and the finite element method is to find $u_h \in V_{g_1}^h$ such that

$$A(u_h, v) = L(v), \quad \forall v \in V_0^h. \tag{27}$$



Example of a piece-wise linear function in 2 dimensions.

To construct a FEM basis of V_0^h , we choose, as in the one dimensional problem, the basis $\phi_j \in V_0^h$ as the Lagrange basis

$$\phi_j(x_i) = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \quad j = 1, 2, \dots, N,$$

where x_i , $i = 1, \dots, N$, are the vertices of the triangulation.

Step 4. Solve the discrete system.

Let

$$u_h(x) = \sum_{i=1}^N \xi_i \phi_i(x), \quad \text{and} \quad \xi_i = u_h(x_i).$$

Then (27) can be written in matrix form,

$$\tilde{A}\xi = \tilde{L}, \quad \text{where} \quad \tilde{A}_{ji} = A(\phi_i, \phi_j) \quad \text{and} \quad \tilde{L}_j = L(\phi_j).$$

Well-posedness of the variational formulation

While the FEM workflow based on the building blocks above provides a systematic procedure to derive a linear system of equations from a linear PDE, it immediately raises the following questions:

1. In **Step 1** of the workflow we change the PDE formulation to a variational formulation. Under what conditions is the variational formulation well posed?
2. In **Step 2** we introduce a variational formulation in a subspace. What can we say about its well posedness?
3. If both variational formulations are well posed, how accurate is the approximation?

We will address these questions next, starting with the well-posedness.

Theorem 4.4 (Lax–Milgram)

Let V be a Hilbert-space with norm $\|\cdot\|_V$ and inner product $(\cdot, \cdot)_V$, and assume that $A: V \times V \rightarrow \mathbb{R}$ is a bilinear form and $L: V \rightarrow \mathbb{R}$ is a linear functional. Suppose that

- (1) A is coercive (V -elliptic), i.e.

$$\exists \gamma > 0 \text{ such that } A(v, v) \geq \gamma \|v\|_V^2 \quad \forall v \in V;$$

- (2) A is continuous, i.e. $\exists C_1 \in \mathbb{R}$ such that $|A(v, w)| \leq C_1 \|v\|_V \|w\|_V$;

- (3) L is continuous, i.e. $\exists C_2 \in \mathbb{R}$ such that $|L(v)| \leq C_2 \|v\|_V \quad \forall v \in V$.

Then there is a unique function $u \in V$ such that $A(u, v) = L(v)$

$\forall v \in V$, which satisfies the stability estimate

$$\|u\|_V \leq \frac{1}{\gamma} \|L\|_V.$$

Remark 4.6

Notice that the well-posedness of the continuous variational problem in view of the Lax–Milgram Thm. implies directly the well-posedness of the discretized variational formulation for any conforming FEM.

Proof.

The existence proof of a unique weak solution follows from Riesz representation theorem, which is technical for general bilinear forms. Here, we only show the stability estimate. Without loss of generality we can assume that $u \neq 0$ solves the variational formulation. Together with the coercivity condition we then find

$$\gamma \|u\|_V^2 \leq A(u, u) = L(u) \quad \Leftrightarrow \quad \gamma \|u\|_V \leq \frac{|L(u)|}{\|u\|_V},$$

from which the claim follows. □

Theorem 4.5

Suppose A is symmetric, i.e. $A(u, v) = A(v, u)$ $\forall u, v \in V$, then
(Variational problem) \iff (Minimization problem) with

- (Var) Find $u \in V$ such that $A(u, v) = L(v)$ $\forall v \in V$,
(Min) Find $u \in V$ such that $F(u) \leq F(v)$ $\forall v \in V$,

where $F(w) := \frac{1}{2}A(w, w) - L(w)$ for all $w \in V$.

Proof.

Take $\epsilon \in \mathbb{R}$. Then (\Rightarrow)

$$\begin{aligned} F(u + \epsilon w) &= \frac{1}{2}A(u + \epsilon w, u + \epsilon w) - L(u + \epsilon w) \\ &= \left(\frac{1}{2}A(u, u) - L(u) \right) + \epsilon A(u, w) - \epsilon L(w) + \frac{1}{2}\epsilon^2 A(w, w) \\ &\geq \frac{1}{2}A(u, u) - L(u) = F(u). \end{aligned}$$

Conversely (\Leftarrow), let $g(\epsilon) = F(u + \epsilon w)$, where $g : \mathbb{R} \rightarrow \mathbb{R}$. Then

$$0 = g'(0) = 0 \cdot A(w, w) + A(u, w) - L(w) = A(u, w) - L(w).$$



Example 4.6

Determine conditions for the functions k, r and $f : D \rightarrow \mathbb{R}$ such that the assumptions in the Lax-Milgram theorem are satisfied for the following elliptic partial differential equation in $D \subset \mathbb{R}^2$

$$\begin{aligned}-\operatorname{div}(k \nabla u) + ru &= f \quad \text{in } D \\ u &= 0 \quad \text{on } \partial D.\end{aligned}$$

Solution. This problem satisfies (Var) with

$$V = \{v : \int_D (v^2(x) + |\nabla v(x)|^2) dx < \infty, \text{ and } v|_{\partial D} = 0\},$$

$$\begin{aligned}A(u, v) &= \int_D (k \nabla u \cdot \nabla v + ruv) dx, \\ L(v) &= \int_D fv dx, \\ \|v\|_V^2 &= \int_D (v^2(x) + |\nabla v|^2) dx.\end{aligned}$$

Consequently V is a Hilbert space and A is symmetric and continuous provided k and r are uniformly bounded.

The V -ellipticity follows by

$$\begin{aligned} A(v, v) &= \int_D (k|\nabla v|^2 + rv^2) \, dx \\ &\geq \alpha \int_D (v^2(x) + |\nabla v|^2) \, dx \\ &= \alpha \|v\|_{H^1}^2, \end{aligned}$$

provided $\alpha = \inf_{x \in D}(k(x), r(x)) > 0$.

The continuity of A is a consequence of

$$\begin{aligned} A(v, w) &\leq \max(\|k\|_{L^\infty}, \|r\|_{L^\infty}) \int_D (|\nabla v||\nabla w| + |v||w|) dx \\ &\leq \max(\|k\|_{L^\infty}, \|r\|_{L^\infty}) \|v\|_{H^1} \|w\|_{H^1}, \end{aligned}$$

provided $\max(\|k\|_{L^\infty}, \|r\|_{L^\infty}) = C < \infty$.

Finally, the functional L is continuous, since

$$|L(v)| \leq \|f\|_{L^2} \|v\|_{L^2} \leq \|f\|_{L^2} \|v\|_V,$$

which means that we may take $\Lambda = \|f\|_{L^2}$ provided we assume that $f \in L^2(D)$. Therefore the problem satisfies the Lax-Milgram theorem.

Example 4.7

Verify that the assumption of the Lax-Milgram theorem are satisfied for the following problem,

$$\begin{aligned}-\Delta u &= f \quad \text{in } D, \\ u &= 0 \quad \text{on } \partial D.\end{aligned}$$

Solution. This problem satisfies (Var) with

$$\begin{aligned}V = H_0^1 &= \{v \in H^1 : v|_{\partial D} = 0\}, \\ H^1 &= \{v : \int_D (v^2(x) + |\nabla v(x)|^2) dx < \infty\},\end{aligned}$$

$$\begin{aligned}A(u, v) &= \int_D \nabla u \nabla v \, dx, \\ L(v) &= \int_D fv \, dx.\end{aligned}$$

To verify the V-ellipticity, we use the *Poincaré inequality*, i.e. there is a constant C such that

$$v \in H_0^1 \Rightarrow \int_D v^2 \, dx \leq C \int_D |\nabla u|^2 \, dx. \quad (28)$$

Idea of Proof for Poincaré inequality: In one dimension and $D = (0, 1)$, the inequality (28) takes the form

$$\int_0^1 v^2(x) \, dx \leq \int_0^1 (v'(x))^2 \, dx, \quad (29)$$

provided $v(0) = 0$. Since

$$v(x) = v(0) + \int_0^x v'(s) \, ds = \int_0^x v'(s) \, ds,$$

and by Cauchy's inequality

$$\begin{aligned} v^2(x) &= \left(\int_0^x v'(s) \, ds \right)^2 \leq x \int_0^x v'(s)^2 \, ds \\ &\leq \int_0^1 v'(s)^2 \, ds \quad \text{since } x \in (0, 1). \end{aligned}$$

Generalization to higher dimensions in a component-wise manner.

The V-ellipticity of A eventually follows from (28) and

$$\begin{aligned}A(v, v) &= \int_D |\nabla v|^2 \, dx \\&= \frac{1}{2} \left(\int_D |\nabla v|^2 \, dx + \int_D |\nabla v|^2 \, dx \right) \\&\geq \frac{1}{2} \left(\frac{1}{C} \int_D v^2 \, dx + \int_D |\nabla v|^2 \, dx \right) \\&\geq \frac{\min\{1, C^{-1}\}}{2} \|v\|_{H^1(D)}^2 \quad \forall v \in H_0^1(D).\end{aligned}$$

The other conditions can be proved similarly as in the previous example. Therefore this problem satisfies the Lax-Milgram theorem.

Approximate input data

The classic stability estimate in the Lax–Milgram Thm. shows the solution's continuous dependence on the input L . One can extend this stability statement to also account for A as an input. In fact, one can use this to derive an estimate that reflects perturbations in input data.

Specifically, let u and \tilde{u} be the unique elements of V that solve the variational problems

$$A(u, v) = L(v) \quad \forall v \in V \quad \text{and} \quad \tilde{A}(\tilde{u}, v) = \tilde{L}(v) \quad \forall v \in V$$

resp. Moreover, let $\tilde{\gamma}$ denote the coercivity constant of \tilde{A} . Then

$$\begin{aligned}\tilde{\gamma}\|u - \tilde{u}\|_V^2 &\leq \tilde{A}(u - \tilde{u}, u - \tilde{u}) = \tilde{A}(u, u - \tilde{u}) - \tilde{L}(u - \tilde{u}) \\ &= \tilde{A}(u, u - \tilde{u}) - A(u, u - \tilde{u}) + L(u - \tilde{u}) - \tilde{L}(u - \tilde{u}) \\ &\leq \|\tilde{A}(u, \cdot) - A(u, \cdot)\|_{V'} \|u - \tilde{u}\|_V + \|\tilde{L} - L\|_{V'} \|u - \tilde{u}\|_V ,\end{aligned}$$

so that

$$\|u - \tilde{u}\|_V \leq \frac{1}{\tilde{\gamma}} \left(\|\tilde{A}(u, \cdot) - A(u, \cdot)\|_{V'} + \|\tilde{L} - L\|_{V'} \right) .$$

The previous estimate will be useful to quantify errors when solving the variational problem with approximate input data, such as truncations. This will be particularly useful when dealing with random fields affecting PDEs.

Exercise 4.4

Consider

$$\begin{aligned} A(u, v) &= \int_D k \nabla u \cdot \nabla v \, dx \quad \text{and} \quad L(v) = \int_D f v \, dx \\ \tilde{A}(u, v) &= \int_D \tilde{k} \nabla u \cdot \nabla v \, dx \quad \text{and} \quad \tilde{L}(v) = \int_D \tilde{f} v \, dx \end{aligned}$$

with $V = H_0^1(D)$ and assume that the corresponding variational problems are well-posed with unique solutions u and \tilde{u} , resp. Suppose further that $k, \tilde{k} \in L^\infty(D)$ and $f, \tilde{f} \in L^2(D)$. Derive an upper bound on $\|u - \tilde{u}\|_V$ that is independent of u , but may depend on $\|k - \tilde{k}\|_{L^\infty(D)}$ and $\|f - \tilde{f}\|_{L^2(D)}$.

Theorem 4.8 (Galerkin Orthogonality)

Let the hypotheses of the Lax–Milgram Thm. 4.4 hold with $V_h \subset V$. Then both the continuous and the discrete variational formulation have a unique solution $u \in V$ and $u_h \in V_h$, respectively. Moreover they satisfy the Galerkin orthogonality property:

$$A(u - u_h, v_h) = 0 \quad \text{for all } v_h \in V_h.$$

Proof.

Uniqueness is clear. We have $A(u, v) = L(v)$ for all $v \in V$, but in particular also $A(u, v_h) = L(v_h)$, because $v_h \in V_h \subset V$. In the discrete setting we have $A(u_h, v_h) = L(v_h)$ for all $v_h \in V_h$ and the claim follows. □

Remark 4.7

A symmetric, coercive, bilinear form A can be used to define a (special) scalar product $(u, v)_A := A(u, v)$. The Galerkin orthogonality is interpreted in this sense. That is, the error of the approximation $u - u_h$ is “orthogonal” to all elements of V_h measured in the scalar product $(\cdot, \cdot)_A$. The approximation u_h is called a Galerkin-orthogonal projection of the exact u into the space V_h and gives the best-approximation estimate

$$\|u - u_h\|_A \leq \min_{v_h \in V_h} \|u - v_h\|_A$$

The following important result then offers an “quasi” optimal best approximation in the usual norm of V , without symmetry assumptions.

Lemma 4.9 (Céa)

Let the hypotheses of the Lax–Milgram Thm. 4.4 hold with $V_h \subset V$. Then both the continuous and the discrete variational formulation has a unique solution $u \in V$ and $u_h \in V_h$, respectively. The discretization error $u - u_h$ satisfies

$$\|u - u_h\|_V \leq \frac{C}{\gamma} \inf_{v_h \in V_h} \|u - v_h\|_V ,$$

where γ and C are the coercivity constant and continuity constant of the bilinear form, respectively.

Proof.

First, we show that $A(u - u_h, u - u_h) = A(u - u_h, u - v_h)$ for all $v_h \in V_h$. Indeed, it follows from the Galerkin orthogonality that $A(u - u_h, u_h) = 0$ because $u_h \in V_h$. We thus find

$$\begin{aligned} A(u - u_h, u - u_h) &= A(u - u_h, u) - A(u - u_h, u_h) \\ &= A(u - u_h, u) - A(u - u_h, v_h) = A(u - u_h, u - v_h). \end{aligned}$$

Next, we use coercivity and the continuity of the bilinear form A to deduce that

$$\begin{aligned} \gamma \|u - u_h\|_V^2 &\leq |A(u - u_h, u - u_h)| = |A(u - u_h, u - v_h)| \\ &\leq C \|u - u_h\|_V \|u - v_h\|_V, \end{aligned}$$

from which the claim follows. □

FEM error analysis

We will now briefly discuss a priori and a posteriori error estimates for finite element methods, where

$$\begin{aligned}\|u - u_h\| &\leq E_1(h, u, f) \quad \text{is an a priori error estimate,} \\ \|u - u_h\| &\leq E_2(h, u_h, f) \quad \text{is an a posteriori error estimate.}\end{aligned}$$

In view of Céa's lemma estimating the FEM approximation error $\|u - u_h\|$ entails studying the best approximation error of $V_h \subset V$.

The expected rate of convergence depends on both the regularity of the exact solution (i.e., on the problem) but also on regularity properties of the mesh itself. Roughly speaking, if h is the mesh-size of a “regular” mesh, then one expects to observe

$$\|u - u_N\|_{H^1(D)} \leq c h^{\min(p+1,m)-1} \left\| D^{\min(p+1,m)} u \right\|_{L^2(D)}.$$

for the approximation of order $p \geq 1$ finite elements (degree of local polynomial FEM basis) of the solution $u \in H^m(D)$ to a second order elliptic PDE (so that $H^1(D)$ is the natural space).

It is possible to obtain better rates when measuring the approximation error in other norms, e.g., in $L^2(D)$; cf. Aubin–Nitsche trick.

In what will follow, we will exemplify the derivation of both a priori and a posteriori error estimates for the 1d model problem.

Quantifying the best approximation error

The approximation property of the space V_h can be characterized by

Lemma 4.10

Suppose V_h is the piece-wise linear finite element space (25), which discretizes the functions in V , defined on $(0, 1)$, with the interpolant $\pi : V \rightarrow V_h$ defined by

$$\pi v(x) = \sum_{i=1}^N v(x_i) \phi_i(x), \quad (30)$$

where $\{\phi_i\}$ is the basis (24) of V_h . Then

$$\|v - \pi v\|'_{L^2(0,1)} \leq \sqrt{\int_0^1 h^2 v''(x)^2 dx} \leq Ch, \quad (31)$$

$$\|v - \pi v\|_{L^2(0,1)} \leq \sqrt{\int_0^1 h^4 v''(x)^2 dx} \leq Ch^2,$$

where $h = \max_i (x_{i+1} - x_i)$.

Proof: Take $v \in V$ and consider first (31) on an interval (x_i, x_{i+1}) . By the mean value theorem, for each $x \in (x_i, x_{i+1})$ there is a $\xi \in (x_i, x_{i+1})$ such that $v'(\xi) = (\pi v)'(x)$. Therefore

$$v'(x) - (\pi v)'(x) = v'(x) - v'(\xi) = \int_{\xi}^x v''(s)ds,$$

provided that v is suff. regular (e.g., $v \in H^2(D)$), so that

$$\begin{aligned} \int_{x_i}^{x_{i+1}} |v'(x) - (\pi v)'(x)|^2 dx &= \int_{x_i}^{x_{i+1}} \left(\int_{\xi}^x v''(s)ds \right)^2 dx \\ &\leq \int_{x_i}^{x_{i+1}} |x - \xi| \int_{\xi}^x (v''(s))^2 ds dx \\ &\leq h^2 \int_{x_i}^{x_{i+1}} (v''(s))^2 ds, \end{aligned} \tag{32}$$

which after summation of the intervals proves (31).

Next, we have

$$v(x) - \pi v(x) = \int_{x_i}^x (v - \pi v)'(s) ds,$$

so by (32)

$$\begin{aligned} \int_{x_i}^{x_{i+1}} |v(x) - \pi v(x)|^2 dx &= \int_{x_i}^{x_{i+1}} \left(\int_{x_i}^x (v - \pi v)'(s) ds \right)^2 dx \\ &\leq \int_{x_i}^{x_{i+1}} |x - x_i| \int_{x_i}^x ((v - \pi v)')^2(s) ds dx \\ &\leq h^4 \int_{x_i}^{x_{i+1}} (v''(s))^2 ds, \end{aligned}$$

which after summation of the intervals proves the lemma. □

An a priori estimate in 1d

Our derivation of the a priori error estimate

$$\|u - u_h\|_V \leq Ch,$$

where u and u_h satisfy (21) and (22), respectively, uses Lemma 4.10 and a combination of the following four steps:

(1) error representation based on the *ellipticity*

$$\alpha \int_D (v^2(x) + (v'(x))^2) dx \leq A(v, v) = \int_D (a(v')^2 + rv^2) dx,$$

where $\alpha = \inf_{x \in (0,1)} (a(x), r(x)) > 0$,

(2) the *Galerkin orthogonality*

$$A(u - u_h, v) = 0 \quad \forall v \in V_h,$$

obtained by $V_h \subset V$ and subtraction of the two equations

$$A(u, v) = L(v) \quad \forall v \in V \quad \text{by (21)},$$

$$A(u_h, v) = L(v) \quad \forall v \in V_h \quad \text{by (22)},$$

(3) the *continuity*

$$|A(v, w)| \leq C \|v\|_V \|w\|_V \quad \forall v, w \in V,$$

where $C \leq \sup_{x \in (0,1)} (a(x), r(x))$, and

(4) the *interpolation estimates*

$$\begin{aligned}\|(v - \pi v)'\|_{L^2} &\leq Ch, \\ \|v - \pi v\|_{L^2} &\leq Ch^2,\end{aligned}\tag{33}$$

where $h = \max (x_{i+1} - x_i)$.

To start the derivation of an a priori estimate let $e \equiv u - u_h$. Then by Cauchy's inequality

$$\begin{aligned}A(e, e) &= A(e, u - \pi u + \pi u - u_h) \\ &= A(e, u - \pi u) + A(e, \pi u - u_h) \\ &\stackrel{\text{Step 2}}{=} A(e, u - \pi u) \\ &\leq \sqrt{A(e, e)} \sqrt{A(u - \pi u, u - \pi u)},\end{aligned}$$

so that by division of $\sqrt{A(e, e)}$,

$$\begin{aligned}\sqrt{A(e, e)} &\leq \sqrt{A(u - \pi u, u - \pi u)} \\ &\stackrel{\text{Step 3}}{\leq} C \|u - \pi u\|_V \\ &\equiv C \sqrt{\|u - \pi u\|_{L^2}^2 + (u - \pi u)' \cdot (u - \pi u)} \\ &\stackrel{\text{Step 4}}{\leq} Ch.\end{aligned}$$

Therefore, by Step 1

$$\alpha \|e\|_V^2 \leq A(e, e) \leq Ch^2,$$

which implies the a priori estimate

$$\|e\|_V \leq Ch, \text{ where } C = K(u).$$

An a posteriori error estimate for the 1d model problem

We consider a version of the 1d model problem (19), namely,

$$\begin{cases} -(au')' + ru = f & \text{in } (0, 1), \\ u(0) = u(1) = 0. \end{cases}$$

Then we show the following a posteriori error estimate:

$$\begin{aligned} \sqrt{A(u - u_h, u - u_h)} &\leq C \|a^{-\frac{1}{2}}(f - ru_h + a'u'_h)h\|_{L^2} \\ &\equiv E(h, u_h, f). \end{aligned} \tag{34}$$

Proof. Let $e = u - u_h$ and let $\pi e \in V_h$ be the nodal interpolant of e .

We have

$$\begin{aligned} A(e, e) &= A(e, e - \pi e) && \text{(by orthogonality)} \\ &= A(u, e - \pi e) - A(u_h, e - \pi e). \end{aligned}$$

Using the notation $(f, v) \equiv \int_0^1 fv \, dx$, we obtain by integration by parts

$$\begin{aligned}
A(e, e) &= (f, e - \pi e) \\
&\quad - \sum_{i=1}^N \int_{x_i}^{x_{i+1}} (au'_h(e - \pi e)' + ru_h(e - \pi e)) \, dx \\
&= (f - ru_h, e - \pi e) \\
&\quad - \sum_{i=1}^N \left\{ [au'_h(e - \pi e)]_{x_i}^{x_{i+1}} - \int_{x_i}^{x_{i+1}} (au'_h)'(e - \pi e) \, dx \right\} \\
&= (f - ru_h + a'u'_h, e - \pi e) \quad (\text{since } u''_h|_{(x_i, x_{i+1})} = 0, (e - \pi e)(x_i) = 0) \\
&\leq \|a^{-\frac{1}{2}} h(f - ru_h + a'u'_h)\|_{L^2} \|a^{\frac{1}{2}} h^{-1}(e - \pi e)\|_{L^2}.
\end{aligned}$$

Lemma 4.11 implies

$$\sqrt{A(e, e)} \leq C \|a^{-\frac{1}{2}} h(f - ru_h + a' u'_h)\|_{L^2},$$

which also shows that

$$\|e\|_V \leq Ch,$$

where $C = K'(u_h)$.

Lemma 4.11

There is a constant C , independent of u and u_h , such that,

$$\|a^{\frac{1}{2}} h^{-1}(e - \pi e)\|_{L^2} \leq C \sqrt{\int_0^1 a e' e' \ dx} \leq C \sqrt{A(e, e)}$$

Exercise 4.5

Use the interpolation estimates in Lemma 4.10 to prove Lemma 4.11.

Hint: Recall that

$$\int_{x_i}^{x_{i+1}} |v(x) - \pi v(x)|^2 dx \leq h^2 \int_{x_i}^{x_{i+1}} ((v - \pi v)')^2(x) dx$$

We formulate an adaptive algorithm based on the a posteriori error estimate (34) as follows:

- (1) Choose an initial coarse mesh T_{h_0} with mesh size h_0 .
- (2) Compute the corresponding FEM solution u_{h_i} in V_{h_i} .
- (3) Given a computed solution u_{h_i} in V_{h_i} , with the mesh size h_i ,

stop if $E(h_i, u_{h_i}, f) \leq TOL$
go to step 4 if $E(h_i, u_{h_i}, f) > TOL$.

- (4) Determine a new mesh $T_{h_{i+1}}$ with mesh size h_{i+1} such that

$$E(h_{i+1}, u_{h_i}, f) \cong TOL,$$

by letting the error contribution for all elements be approximately constant, i.e.

$$\|a^{-\frac{1}{2}} h(f - ru_h - a' u'_h)\|_{L^2(x_i, x_{i+1})} \cong C, \quad i = 1, \dots, N,$$

then go to Step 2.

Finite Element Methods, further reading

- ▶ *Numerical treatment of partial differential equations.* Grossmann, Ross, Stynes. Springer, 2007.
- ▶ *A Primer on PDEs: Models, Methods, Simulations.* Salsa, Vegni, Zaretti, Zunino. Springer. 2013.
- ▶ *Elliptic Differential Equations: Theory and Numerical Treatment.* Hackbusch. Springer 2017.
- ▶ *Partial Differential Equations.* Evans. AMS, 2010.

Bayesian Inference, Linear time dependent PDE

The following example is based on:

- "A hierarchical Bayesian setting for an inverse problem in linear parabolic PDEs with noisy boundary conditions", by Fabrizio Ruggeri, Zaid Sawlan, Marco Scavino, Raúl Tempone, *arXiv:1501.04739* . In *Bayesian Analysis*, Volume 12, Number 2 (2017), 407-433.

We develop a hierarchical Bayesian setting to infer unknown parameters of linear parabolic PDEs, as an example of linear time dependent PDEs, under the assumption that noisy measurements are available in the interior of a domain of interest and for the unknown boundary conditions.

Goal: Solve the inverse problem assuming that the boundary conditions are modeled by given distributions.

Formulation of the problem

Consider the deterministic one-dimensional parabolic initial-boundary value problem:

$$\begin{cases} \partial_t T + L_{\theta} T = 0, & x \in (x_L, x_R), 0 < t \leq t_N < \infty \\ T(x_L, t) = T_L(t), & t \in [0, t_N] \\ T(x_R, t) = T_R(t), & t \in [0, t_N] \\ T(x, 0) = g(x), & x \in (x_L, x_R), \end{cases} \quad (35)$$

where L_{θ} is a linear second-order partial differential operator that takes the form

$$L_{\theta} T = -\partial_x(a(x)\partial_x T) + b(x)\partial_x T + c(x)T,$$

$\theta(x) = (a(x), b(x), c(x))^{tr}$, and the partial differential operator $\partial_t + L_{\theta}$ is parabolic.

Our main objective is to provide a Bayesian solution to an inverse problem for θ , where we assume that

- i θ is unknown, while the initial condition g in the initial-boundary value problem is known;
- ii θ is allowed to vary with the spatial variable x .

Available Data: Noisy readings of the function $T(x, t)$ at the $I + 1$ spatial locations, including the boundaries, $x_L = x_0, x_1, x_2, \dots, x_{I-1}, x_I = x_R$, at each of the N times t_1, t_2, \dots, t_N , are assumed available.

Statistical setting

Let $\mathbf{Y}_n := (Y_{0,n}, \dots, Y_{I,n})^{tr}$ denote the vector of observed readings at time t_n , and assume a statistical model with an additive Gaussian experimental noise ϵ_n :

$$\mathbf{Y}_n = \begin{bmatrix} T_L(t_n) \\ T(x_1, t_n) \\ \vdots \\ T(x_{I-1}, t_n) \\ T_R(t_n) \end{bmatrix} + \epsilon_n, \quad (36)$$

where $\epsilon_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}_{I+1}, \sigma^2 \mathbf{I}_{I+1})$ for some measurement error variance $\sigma^2 > 0$.

Let

$$\mathbf{Y}_n^I := (Y_{1,n}, \dots, Y_{I-1,n})^{tr}$$

denote the vector of observed data at the **interior locations** x_1, x_2, \dots, x_{I-1} and let

$$\mathbf{Y}_n^B := (Y_{L,n}, Y_{R,n})^{tr}$$

be the vector of observed data at the **boundary locations** x_0, x_I at time t_n .

Local Auxiliary Problem

Consider a local time problem, defined between consecutive measurement times, i.e.

$$\begin{cases} \partial_t T + L_\theta T = 0, & x \in (x_L, x_R), t_{n-1} < t \leq t_n, \\ T(x_L, t) = T_L(t), & t \in [t_{n-1}, t_n], \\ T(x_R, t) = T_R(t), & t \in [t_{n-1}, t_n], \\ T(x, t_{n-1}) = \widehat{T}(x, t_{n-1}), & x \in (x_L, x_R), \end{cases} \quad (37)$$

whose exact solution, denoted by $\widehat{T}(\cdot, t_n)$, depends only on the parameter θ , the initial condition

$$\widehat{T}(\cdot, t_{n-1})$$

and the Dirichlet boundary conditions

$$\{T_L(t), T_R(t)\}_{t \in (t_{n-1}, t_n)}.$$

Lemma 4.12 (One Step Interior Data Likelihood)

The conditional probability density function (pdf) of the *interior* data \mathbf{Y}_n^I is given by

$$\begin{aligned} \rho(\mathbf{Y}_n^I & \mid \theta, \hat{T}(\cdot, t_{n-1}), \{T_L(t), T_R(t)\}_{t \in (t_{n-1}, t_n)}) \\ &= \frac{1}{(\sqrt{2\pi}\sigma)^{I-1}} \exp\left(-\frac{1}{2\sigma^2} \|\mathbf{R}_{t_n}\|_{\ell^2}^2\right), \end{aligned} \quad (38)$$

where $\mathbf{R}_{t_n} := (\hat{T}(x_1, t_n) - Y_{1,n}, \dots, \hat{T}(x_{I-1}, t_n) - Y_{I-1,n})^{tr}$ denotes the *data residual vector* at time $t = t_n$.

Dirichlet Boundary Data modeling

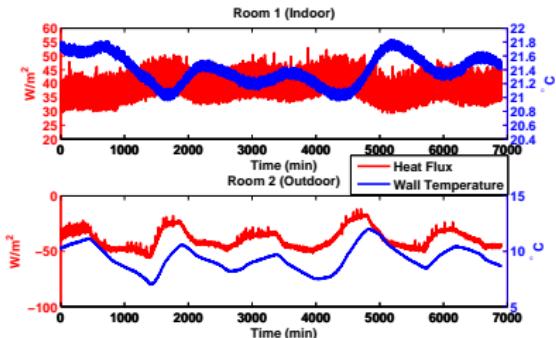


Figure: Temperature and heat flux measurements.

Assume now that the Dirichlet boundary condition functions, T_L and T_R , are well approximated on $[0, T]$ by some finite dimensional subspace, for instance splines or piecewise linear continuous functions. The functions T_L and T_R are found by least squares on the boundary data, see Figure ??.

Let LBC_n denote the time nodes that determine the local boundary conditions $\{T_{L,n-1}, T_{L,n}, T_{R,n-1}, T_{R,n}\}$.

Lemma 4.13 (All Interior and Boundary Data Likelihood)

The joint likelihood function of θ and the boundary parameters $\{LBC_n\}_{n=1,\dots,N}$ is given by

$$\begin{aligned} \rho(\mathbf{Y}_1, \dots, \mathbf{Y}_N | \boldsymbol{\theta}, \{LBC_n\}_{n=1,\dots,N}) &= \prod_{n=1}^N \frac{1}{(\sqrt{2\pi}\sigma)^{l-1}} \exp\left(-\frac{1}{2\sigma^2} \|\mathbf{R}_{t_n}\|_{\ell^2}^2\right) \\ &\quad \times \frac{1}{2\pi\sigma^2} \exp\left(-\frac{1}{2\sigma^2} (T_{L,n} - Y_{L,n})^2\right) \\ &\quad \exp\left(-\frac{1}{2\sigma^2} (T_{R,n} - Y_{R,n})^2\right). \end{aligned} \quad (39)$$

Notation: The boundary (unknown) temperature values are denoted as $\mathbf{T}_L = (T_{L,1}, \dots, T_{L,N})^{tr}$, $\mathbf{T}_R = (T_{R,1}, \dots, T_{R,N})^{tr}$, and the respective boundary (known) temperature recorded data are $\mathbf{Y}_L = (Y_{L,1}, \dots, Y_{L,N})^{tr}$ and $\mathbf{Y}_R = (Y_{R,1}, \dots, Y_{R,N})^{tr}$.

The joint likelihood function (39) (assuming additive Gaussian noise in the Dirichlet boundary conditions) can be written as

$$\rho(\mathbf{Y}_1, \dots, \mathbf{Y}_N | \boldsymbol{\theta}, \mathbf{T}_L, \mathbf{T}_R) \quad (40)$$

$$= (\sqrt{2\pi}\sigma)^{-N(I+1)} \exp\left(-\frac{1}{2\sigma^2} \sum_{n=1}^N \|\mathbf{R}_{t_n}\|_{\ell^2}^2\right) \\ \times \exp\left(-\frac{1}{2\sigma^2} \left[\|\mathbf{T}_L - \mathbf{Y}_L\|_{\ell^2}^2 + \|\mathbf{T}_R - \mathbf{Y}_R\|_{\ell^2}^2 \right]\right). \quad (41)$$

Observe: The values of \mathbf{T}_L and \mathbf{T}_R are not known, so we treat them as nuisance parameters, marginalizing them out from the likelihood. Their distribution is characterized using the boundary data.

The marginal likelihood of θ

Assume that the **nuisance parameters** \mathbf{T}_L and \mathbf{T}_R are independent Gaussian distributed:

$$\mathbf{T}_L \sim \mathcal{N}(\boldsymbol{\mu}_L, \sigma_p^2 \mathbf{I}_N), \quad \mathbf{T}_R \sim \mathcal{N}(\boldsymbol{\mu}_R, \sigma_p^2 \mathbf{I}_N). \quad (42)$$

Using (41), the **marginal likelihood** of θ is given by

$$\rho(\mathbf{Y}_1, \dots, \mathbf{Y}_N | \theta) \quad (43)$$

$$\begin{aligned} &= (\sqrt{2\pi}\sigma)^{-N(I+1)} (\sqrt{2\pi}\sigma_p)^{-2N} \int_{\mathcal{T}_R} \int_{\mathcal{T}_L} \exp \left(-\frac{1}{2\sigma^2} \sum_{n=1}^N \|\mathbf{R}_{t_n}\|_{\ell^2}^2 \right) \\ &\quad \times \exp \left(-\frac{1}{2\sigma^2} \|\mathbf{T}_L - \mathbf{Y}_L\|_{\ell^2}^2 - \frac{1}{2\sigma^2} \|\mathbf{T}_R - \mathbf{Y}_R\|_{\ell^2}^2 \right) \end{aligned}$$

$$\times \exp \left(-\frac{1}{2\sigma_p^2} \|\mathbf{T}_L - \boldsymbol{\mu}_L\|_{\ell^2}^2 - \frac{1}{2\sigma_p^2} \|\mathbf{T}_R - \boldsymbol{\mu}_R\|_{\ell^2}^2 \right) d\mathbf{T}_L d\mathbf{T}_R, \quad (44)$$

where \mathbf{R}_{t_n} is approximated by

$$\tilde{\mathbf{R}}_{t_n} = \left(\mathbf{B}^n(\theta) \mathbf{T}_0 - \mathbf{Y}_n^I \right) + A_{L,n}(\theta) \mathbf{T}_L + A_{R,n}(\theta) \mathbf{T}_R. \quad (45)$$

Observe that given θ we can compute analytically the integral in (44).

Exercise 4.6 (Solution Operators)

Motivate (45), identifying the matrices $\mathbf{B}(\theta)$, $A_{L,n}(\theta)$ and $A_{R,n}(\theta)$.

Example - inference for thermal diffusivity

Consider the heat equation (one-dim diffusion equation for $T(x, t)$):

$$\begin{cases} \partial_t T - \partial_x (\theta(x) \partial_x T) = 0, & x \in (x_L, x_R), 0 < t \leq t_N < \infty \\ T(0, t) = T_L(t), & t \in [0, t_N] \\ T(1, t) = T_R(t), & t \in [0, t_N] \\ T(x, 0) = g(x), & x \in (x_L, x_R). \end{cases} \quad (46)$$

Goal

To infer the thermal diffusivity, $\theta(x)$, an unknown parameter that measures the heat propagation through a material, using a Bayesian approach, when the temperature is measured at $I + 1$ locations, $x_0 = x_L, x_1, x_2, \dots, x_{I-1}, x_I = x_R$, at each of the N times, t_1, t_2, \dots, t_N . Clearly, this problem is a special case of (35) where $L_\theta = -\partial_x (\theta(x)\partial_x T)$ and $\theta(x) > 0$.

Assumption

Consider a lognormal prior $\log \theta \sim \mathcal{N}(\nu, \tau)$, where $\nu \in \mathbb{R}$ and $\tau > 0$. Assume noisy boundary measurements and a Gaussian distribution for the nuisance boundary parameters as in (42). The non-normalized posterior density for θ is given by

$$\begin{aligned}\rho_{\nu, \tau}(\theta | \mathbf{Y}_1, \dots, \mathbf{Y}_N) \\ \propto \frac{1}{\sqrt{2\pi\theta\tau}} \exp\left(-\frac{(\log \theta - \nu)^2}{2\tau^2}\right) \rho(\mathbf{Y}_1, \dots, \mathbf{Y}_N | \theta),\end{aligned}$$

where $\rho(\mathbf{Y}_1, \dots, \mathbf{Y}_N | \theta)$ is the marginal likelihood of θ .

To assess the performance of our method we use a synthetic dataset, and assume that θ is a lognormal random variable with parameters $\nu = \tau = 0.1$.

Bayesian inference for thermal diffusivity

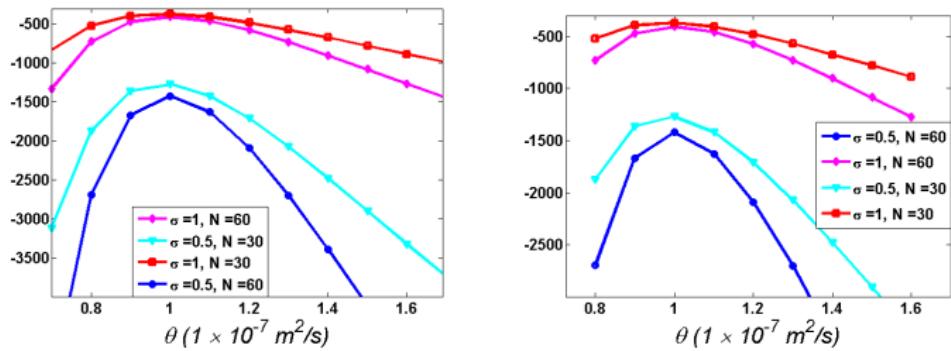


Figure: Example: Comparison between log-likelihoods (on the left) and log-posteriors (on the right) for θ using different numbers of observations, N , and different values of σ .

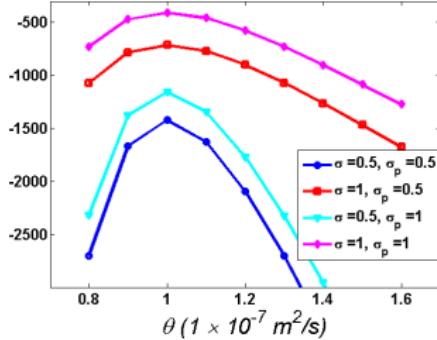
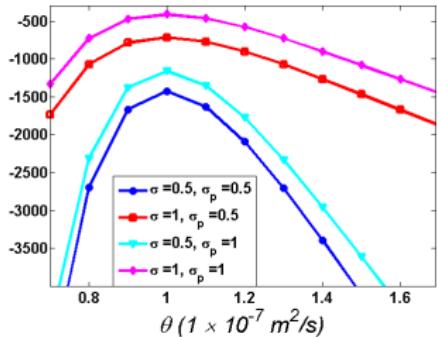


Figure: Example: Comparison between log-likelihoods (on the left) and log-posteriors (on the right) for θ using different values of σ and σ_p , with $N = 60$.

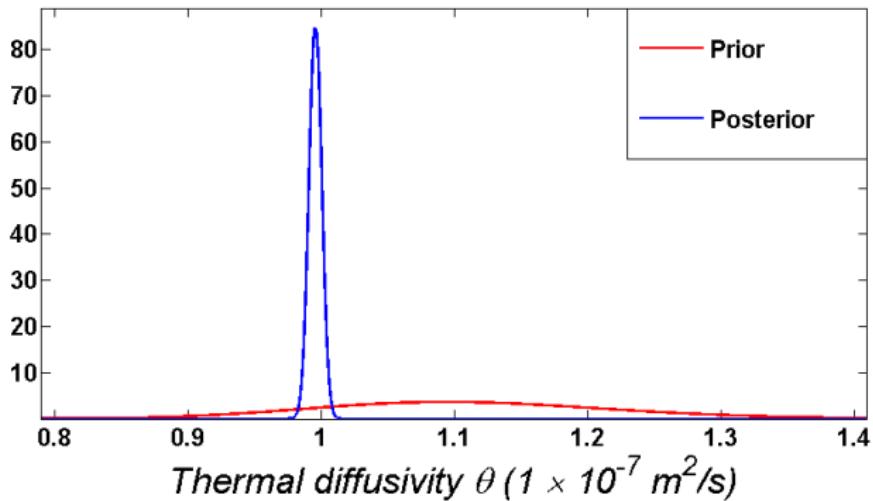


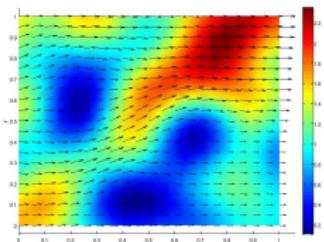
Figure: Example: Lognormal prior and approximated Gaussian posterior densities for θ where $\sigma_p = \sigma = 0.5$ and $N = 60$.

References:

- ▶ "Ensemble-marginalized Kalman filter for linear time-dependent PDEs with noisy boundary conditions: Application to heat transfer in building walls", by M. Iglesias, Z. Sawlan, M. Scavino, R. Tempone and C. Wood. *Inverse Problems* 34, no. 7, 075008, 26 pp, 2018.
- ▶ "Bayesian inferences of the thermal properties of a wall using temperature and heat flux measurements", by M. Iglesias, Z. Sawlan, M. Scavino, R. Tempone and C. Wood. *arXiv:1608.03855*, August 2016. *International Journal of Heat and Mass Transfer* 116, 417–43, 2018.
- ▶ "A hierarchical Bayesian setting for an inverse problem in linear parabolic PDEs with noisy boundary conditions", by F. Ruggeri, Z. Sawlan, M. Scavino, R. Tempone. *Bayesian Anal.* 12, no. 2, 407–433, 2017.

Modeling spatial uncertainties in PDE: random fields

To include uncertainties into PDE models, it is usually necessary to incorporate “random functions” into the PDE description to allow for spatially varying uncertainties.



$$\begin{cases} \mathbf{u} = -\mathbf{k} \nabla p & \text{in } D \\ \operatorname{div} \mathbf{u} = f & \\ + \text{boundary condns.} & \text{on } \partial D \end{cases}$$

These so-called random fields cannot be arbitrary but need to satisfy certain (problem dependent) statistical and spatial regularity properties.

Objectives of this part:

1. introduce random fields and discuss their theoretical properties
2. discuss approaches for sampling random fields

Random fields

Definition 5.1

Let $D \subset \mathbb{R}^d$ be a physical domain. A (real-valued) random field $a : D \times \Omega \rightarrow \mathbb{R}$ is a collection of random variables $\{a(x) : x \in D\}$ on a probability space (Ω, \mathcal{A}, P) . That is, $a(\cdot, \omega)$ is a function for a.a. $\omega \in \Omega$.

Definition 5.2

Let x_1, \dots, x_n be n points in the physical domain D . Then

$$F_n(z_1, \dots, z_n; x_1, \dots, x_n) := P(a(x_1, \omega) \leq z_1, \dots, a(x_n, \omega) \leq z_n) \quad (*)$$

is called the finite dimensional distribution (FDD) of order n .

Theorem 5.3 (Kolmogorov's extension theorem)

Given the collections of all FDDs $F_n : \mathbb{R}^n \times D^{\times n} \rightarrow [0, 1]$, satisfying a conditions below, then there exists a probability space (Ω, \mathcal{A}, P) and a random field $a : D \times \Omega \rightarrow \mathbb{R}$ satisfying (*).

- ▶ Consistency condition: for $m < n$:

$$F_m(z_1, \dots, z_m; x_1, \dots, x_m) = F_n(z_1, \dots, z_m, \infty, \dots, \infty; x_1, \dots, x_m, \dots, x_n)$$

- ▶ Symmetry condition: for any permutation π_1, \dots, π_n of the indices $1, \dots, n$:

$$F_n(z_{\pi_1}, \dots, z_{\pi_n}; x_{\pi_1}, \dots, x_{\pi_n}) = F_n(z_1, \dots, z_n; x_1, \dots, x_n)$$

We now list a few basic definitions for random fields:

- **Expected value function:** $\bar{a}(x) := \mathbb{E}[a(x, \cdot)]$, i.e., $\bar{a} : D \rightarrow \mathbb{R}$.
- **Covariance function:** $\text{Cov}_a : D \times D \rightarrow \mathbb{R}$, given by

$$\text{Cov}_a(x_1, x_2) := \mathbb{E}[(a(x_1, \cdot) - \bar{a}(x_1))(a(x_2, \cdot) - \bar{a}(x_2))].$$

- **Variance function:** $\text{Var}_a : D \rightarrow \mathbb{R}$, with $\text{Var}_a(x) := \text{Cov}_a(x, x)$

Definition 5.4

We say that $a(x, \omega)$ is a **second order random field**, if $\text{Var}_a(x) < \infty$, for all $x \in D$.

Properties of the covariance function:

- **bounded:** $\text{Cov}_a(x_1, x_2) \leq \sqrt{\text{Var}_a(x_1)} \sqrt{\text{Var}_a(x_2)}$.
- **symmetric:** $\text{Cov}_a(x_1, x_2) = \text{Cov}_a(x_2, x_1)$
- **semi-positive definite:** for any $\xi_1, \dots, \xi_N \in D$, the matrix $C_{ij} = \text{Cov}_a(\xi_i, \xi_j)$ is semi-positive definite.

Indeed, let $\alpha = (\alpha_1, \dots, \alpha_N) \in \mathbb{R}^N$ and $a'(x, \omega) = a(x, \omega) - \mathbb{E}[a](x)$.

$$\alpha^T C \alpha = \sum_{i,j} \alpha_i \alpha_j \text{Cov}_a(\xi_i, \xi_j) = \sum_{i,j} \alpha_i \alpha_j \mathbb{E}[a'(\xi_i, \cdot) a'(\xi_j, \cdot)] = \mathbb{E}\left[\left(\sum_i \alpha_i a'(\xi_i, \cdot)\right)^2\right] \geq 0$$

Definition 5.5

A second order random field $a : \mathbb{R}^d \times \Omega \rightarrow \mathbb{R}$ is said to be **stationary** if its law is invariant under translation: $a(x + h, \cdot) \sim a(x, \cdot) \forall h \in \mathbb{R}^d$.

For a stationary field, we find that

- ▶ $\bar{a}(x) \equiv \mu \in \mathbb{R}$ (indep. of x) and
- ▶ $Cov_a(x_1, x_2) = Cov_a(0, x_2 - x_1) = \widetilde{Cov}_a(x_2 - x_1)$.

Definition 5.6

A second order random field is said to be **weakly stationary** if $\mathbb{E}[a]$ is constant and $Cov_a(x_1, x_2) = \widetilde{Cov}_a(x_1 - x_2)$.

Definition 5.7

A weakly stationary random field is said **isotropic** if the covariance function depends only on $\|x_1 - x_2\|$, for all $x_1, x_2 \in D$.

If not stated otherwise, we will, from now on, always consider second order random fields.

Mean square continuity

Definition 5.8

A random field $a : D \times \Omega \rightarrow \mathbb{R}$ is said to be **mean square continuous** at $\bar{x} \in D$ if

$$\lim_{x \rightarrow \bar{x}} \mathbb{E}[(a(x) - a(\bar{x}))^2] = 0.$$

Theorem 5.9

A centered (zero mean) random field $a : D \times \Omega \rightarrow \mathbb{R}$ is mean square continuous at $\bar{x} \in D$ iff its covariance function $\text{Cov}_a(x_1, x_2)$ is continuous at $x_1 = x_2 = \bar{x}$.

Proof.

1) let $r(x_1, x_2) = \text{Cov}_a(x_1, x_2)$ be continuous at $x_1 = x_2 = \bar{x}$, Then

$$\mathbb{E}[(a(x) - a(\bar{x}))^2] = \mathbb{E}[a(x)^2 - 2a(x)a(\bar{x}) + a(\bar{x})^2] = r(x, x) - 2r(x, \bar{x}) + r(\bar{x}, \bar{x}) \xrightarrow{x \rightarrow \bar{x}} 0$$

2) let $a(x, \omega)$ be mean square continuous at \bar{x} . Then

$$r(x_1, x_2) - r(\bar{x}, \bar{x}) = \mathbb{E}[a(x_1)a(x_2) - a(\bar{x})^2 \pm a(\bar{x})a(x_2)]$$

$$\leq (\mathbb{E}[(a(x_1) - a(\bar{x}))^2]\mathbb{E}[a(x_2)^2])^{\frac{1}{2}} + (\mathbb{E}[(a(x_2) - a(\bar{x}))^2]\mathbb{E}[a(\bar{x})^2])^{\frac{1}{2}} \xrightarrow{(x_1, x_2) \rightarrow (\bar{x}, \bar{x})} 0$$



Spectral measure

If $a : \mathbb{R}^d \times \Omega \rightarrow \mathbb{R}$ is a **weakly stationary mean square continuous** random field, its covariance $r(x_1 - x_2) = \text{Cov}_a(x_1, x_2)$ has the representation

$$r(\tau) = \int_{\mathbb{R}^d} e^{2\pi i s \cdot \tau} d\mu(s)$$

for some positive finite measure μ . (Bochner's theorem).

If, moreover, μ has a density $S(s)$: $\mu(ds) = S(s)ds$, then S is called the **spectral density** and corresponds to the Fourier transform of r .

In particular, the Fourier transform of a covariance function of a weakly stationary random field is always non negative.

For a description of the use of the spectral density to approximate samples of a stationary Gaussian random field, please click [here](#).

- "Analysis of continuous spectral method for sampling stationary Gaussian random fields", by Jocelyne Erhel, Mestapha Oumouni, Géraldine Pichot and Franck Schoefs. 2019. fffhal-02109037.

Mean square differentiability

Definition 5.10

A random field $a : D \times \Omega \rightarrow \mathbb{R}$ is **mean square differentiable** at $\bar{x} \in D$ if

$$\frac{\partial}{\partial x} a(\bar{x}, \cdot) := \lim_{h \rightarrow 0} \frac{a(\bar{x} + h, \cdot) - a(\bar{x}, \cdot)}{h} \quad \text{exists in mean square sense.}$$

Theorem 5.11

A centered random field $a : D \times \Omega \rightarrow \mathbb{R}$ is mean square differentiable at $\bar{x} \in D$ iff $\frac{\partial^2}{\partial x_1 \partial x_2} \text{Cov}_a(x_1, x_2)$ exists and is finite at $x_1 = x_2 = \bar{x}$.

More generally, a random field $a : D \times \Omega \rightarrow \mathbb{R}$ is k -times differentiable at \bar{x} if the $2k$ partial derivative of the covariance exists and is finite at $x_1 = x_2 = \bar{x}$.

Exercise 5.1

1. Prove the Theorem (i.e., $k = 1$) for $D \subset \mathbb{R}$.

Hint: Let $(X_h)_{h \in \mathbb{R}} \subset \mathbb{R}$. Then $\mathbb{E}(X_h X_k) \rightarrow C$ as $h, k \rightarrow 0$ iff $\mathbb{E}(|X_h - X|^2) \rightarrow 0$ as $h \rightarrow 0$ from some random variable X .

2. (Optional) Prove the hint given in the previous part.

Sample path continuity

Definition 5.12

A random field $a : D \times \Omega \rightarrow \mathbb{R}$ is said to be sample path (or almost surely) continuous at $\bar{x} \in D$ if

$$P\left(\omega : \lim_{x \rightarrow \bar{x}} a(x, \omega) = a(\bar{x}, \omega)\right) = 1.$$

Theorem 5.13 (Kolmogorov's theorem)

Given a random field $a : D \times \Omega \rightarrow \mathbb{R}$, with $D \subset \mathbb{R}^d$ compact, if there exist positive constants p, β, K such that

$$\mathbb{E}\left[\left(\frac{|a(x_1, \cdot) - a(x_2, \cdot)|}{|x_1 - x_2|^\beta}\right)^p\right] \leq K, \quad \forall x_1, x_2 \in D,$$

then, $a \in C^{0,\alpha}(D)$ almost surely, for all $0 \leq \alpha < \beta - \frac{d}{p}$.

Series expansion of Random Fields: Karhunen-Loève

Let $D \subset \mathbb{R}^d$ be compact and $\text{Cov}_a : D \times D \rightarrow \mathbb{R}$ continuous.

Theorem 5.14

There exists a sequence of values $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_k \geq \dots \geq 0$, with $\lim_{k \rightarrow \infty} \lambda_k = 0$ and a corresponding sequence of functions $b_i : D \rightarrow \mathbb{R}$, $i = 1, 2, \dots$ such that

$$\int_D \text{Cov}_a(x, y) b_i(y) dy = \lambda_i b_i(x), \quad \text{and} \int_D b_i(x) b_j(x) dx = \delta_{ij},$$

for all $x \in D$.

Define, now, the sequence of random variables $y_i(\omega)$, $i = 1, 2, \dots$

$$y_i(\omega) = \frac{1}{\sqrt{\lambda_i}} \int_D (a(x, \omega) - \mathbb{E}[a](x)) b_i(x) dx$$

which are uncorrelated with zero mean and unit variance.

Then, the random field $a(x, \omega)$ can be represented as the infinite series

Karhunen-Loève expansion: $a(x, \omega) = \mathbb{E}[a](x) + \sum_{i=1}^{\infty} \sqrt{\lambda_i} b_i(x) y_i(\omega)$

The Karhunen-Loève expansion is a consequence of Mercer's theorem (see below). Moreover, under the above assumptions (D compact and Cov_a continuous) it holds that

$$\lim_{N \rightarrow \infty} \sup_{x \in D} \mathbb{E} \left[\left(a(x, \cdot) - \sum_{n=1}^N \sqrt{\lambda_n} b_n(x) y_n(\cdot) \right)^2 \right] = 0.$$

Moreover, the KL expansion is the best **N -terms approximation** in terms of variance

$$\{y_n, b_n\}_{n=1}^N = \underset{\substack{(\xi_n, \psi_n) \\ \int_D \psi_n \psi_m = \delta_{nm}}}{\arg \min} \mathbb{E} \left[\int_D \left(a(x, \cdot) - \mathbb{E}[a](x) - \sum_{n=1}^N \xi_n(\cdot) \psi_n(x) \right)^2 dx \right]$$

The convergence rate of the N -term truncation

$$a_N(x, \omega) = \mathbb{E}[a](x) + \sum_{n=1}^N \sqrt{\lambda_n} b_n(x) y_n(\omega)$$

depends on the **decay of the eigenvalues λ_n** , which, in turn, depends on the smoothness of the Covariance function.

Estimates on the decay of the KL eigenvalues can be found in [Schwab et al 05].

Mercer's theorem

Let $D \subset \mathbb{R}^d$ be a compact domain and $K : D \times D \rightarrow \mathbb{R}$ a Mercer's kernel

- ▶ K is symmetric: $K(x, y) = K(y, x)$
- ▶ K is continuous
- ▶ K is semi-positive definite

Define, moreover, the compact operator $T_K : L^2(D) \rightarrow L^2(D)$

$$T_K f(x) = \int_D K(x, y) f(y) dy, \quad \forall f \in L^2(D).$$

Theorem 5.15 (Mercer's theorem)

Under the above assumptions on D and K , there is an orthonormal basis $\{b_i\}_i$ of $L^2(D)$ consisting of eigenfunctions of T_K such that the corresponding sequence of eigenvalues $\{\lambda_i\}_i$ is non-negative. The eigenfunctions corresponding to non-zero eigenvalues are continuous in D and

$$K(x, y) = \sum_{i=1}^{\infty} \lambda_i b_i(x) b_i(y)$$

where the convergence is absolute and uniform, that is

$$\lim_{n \rightarrow \infty} \sup_{x, y \in D} \left| K(x, y) - \sum_{i=1}^n \lambda_i b_i(x) b_i(y) \right| = 0.$$

Proof of Mercer's theorem

We consider only the case of a (strictly) positive kernel K . The general case is left as an exercise.

Thanks to the symmetry, continuity and positivity of K and the fact that D is compact, the operator T_K is compact. Hence by the spectral theorem, there exists an orthonormal basis of eigenfunctions $\{b_i\}_i$ with corresponding eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots > 0$, with $\lim_{n \rightarrow \infty} \lambda_n = 0$.

Moreover, the eigenfunctions are continuous.

Next, let $K_N(x, y) = \sum_{i=1}^N \lambda_i b_i(x) b_i(y)$ and show its uniform convergence. Since

$$\begin{aligned}\sum_{i=1}^N \lambda_i |b_i(x)b_i(y)| &\leq \left(\sum_{i=1}^N \lambda_i |b_i(x)|^2 \right)^{\frac{1}{2}} \left(\sum_{i=1}^N \lambda_i |b_i(y)|^2 \right)^{\frac{1}{2}} \\ &\leq \max_{z \in D} \sum_{i=1}^N \lambda_i |b_i(z)|^2 = \max_{z \in D} K_N(z, z),\end{aligned}$$

we have that K_N increases with N , so that

$$\max_{z \in D} K_N(z, z) \leq \max_{z \in D} \lim_N K_N(z, z) \leq \max_{z \in D} K(z, z) < \infty.$$

Proof of Mercer's theorem (cont.)

Observation: The last inequality holds because, for all $f \in L^2(D)$, we have, thanks to the Spectral Theorem,

$$(f, T_{K_N} f)_{L^2(D)} = \sum_{n=1}^N \lambda_n(f, b_n)_{L^2(D)}^2 \leq \sum_{n=1}^{\infty} \lambda_n(f, b_n)_{L^2(D)}^2 = (f, T_K f). \quad (47)$$

Then, if we have that there exists $z_0 \in D$ such that

$$K(z_0, z_0) < K_N(z_0, z_0)$$

by the continuity of K we can pick f in (47) sufficiently concentrated around z_0 and obtain a contradiction.

In conclusion: We have showed the uniform convergence $K_N \rightarrow K_o$ on $D \times D$. Observe that, for all $f, g \in L^2(D)$ we have

$$(f, T_{K_o} g)_{L^2(D)} = \lim_N (f, T_{K_N} g)_{L^2(D)} = (f, T_K g)_{L^2(D)}.$$

Since K_o and K define the same operator and the map $K \mapsto T_K$ is injective, then $K = K_o$ and the theorem is proved. □

Proof of Karhunen-Loève expansion

Let $D \subset \mathbb{R}^d$ be compact and $a : D \times \Omega \rightarrow \mathbb{R}$ a random field with continuous covariance Cov_a . Then, Cov_a is a Mercer's kernel (symmetric, continuous and semi-positive definite) and by Mercer's theorem we get

$$\text{Cov}_a(x, y) = \sum_{i=1}^{\infty} \lambda_i b_i(x) b_i(y) \quad \rightsquigarrow \quad \text{Var}_a(x) = \sum_{i=1}^{\infty} \lambda_i b_i(x)^2$$

Define now the truncated Karhunen-Loève expansion

$$a_N(x, \omega) = \mathbb{E}[a](x) + \sum_{i=1}^N \sqrt{\lambda_i} y_i(\omega) b_i(x).$$

$$\text{with } y_i(\omega) = \frac{1}{\sqrt{\lambda_i}} \int_D (a(x, \omega) - \mathbb{E}[a](x)) b_i(x) dx.$$

Observation: $E[y_i y_j] = \delta_{ij}$ and $\text{Var}_{a_N}(x) = \sum_{i=1}^N \lambda_i b_i(x)^2$.

Proof of Karhunen-Loève expansion (cont.)

Now, set $a'(x, \omega) = a(x, \omega) - \mathbb{E}[a](x)$ and observe that

$$\begin{aligned}\mathbb{E}[a'(x, \cdot)a'_N(x, \cdot)] &= \sum_{i=1}^N \sqrt{\lambda_i} \mathbb{E}[a'(x, \cdot)y_i] b_i(x) \\ &= \sum_{i=1}^N \mathbb{E}[\int_D a'(x, \cdot)a'(z, \cdot)b_i(z) dz] b_i(x) \\ &= \sum_{i=1}^N b_i(x) \int_D \text{Cov}_a(x, z) b_i(z) dz = \sum_{i=1}^N \lambda_i b_i(x)^2 = \text{Var}_{a_N}(x)\end{aligned}$$

Therefore

$$\mathbb{E}[(a(x, \cdot) - a_N(x, \cdot))^2] = \text{Var}_a(x) - \text{Var}_{a_N}(x) = \sum_{i=N+1}^{\infty} \lambda_i b_i(x)^2 \xrightarrow{N \rightarrow \infty} 0,$$

uniformly in $x \in D$.

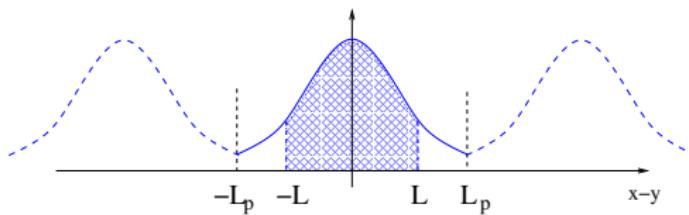
□

Sampling of random fields: Fourier expansion method

Knowledge of the eigen pairs (λ_i, b_i) of a random field's integral operator defined by its covariance allows to generate samples (realizations) of the random field via the (truncated) Karhunen-Loève expansion. If only the covariance (and the mean) is known then instead of first solving the eigen value problem, one can resort to an approximate sampling based on the Fourier transform. We will illustrate the idea here for $D = \mathbb{R}$ and leave the methodological extensions for \mathbb{R}^d as an exercise.

Let $a : \mathbb{R} \times \Omega \rightarrow \mathbb{R}$ a weakly stationary random field with covariance $\text{Cov}_a(x, y) = \text{Cov}_a(x - y)$. We aim at finding a representation by Fourier series of a restricted to the interval $[0, L]$.

Idea: Restrict the covariance to the interval $[-L_p, L_p]$, with $L_p \geq L$ and replicate it periodically.



Then, the periodicized covariance $\text{Cov}_a^\#$ can be expanded in Fourier series

$$\text{Cov}_a^\#(x - y) = \sum_{n=0}^{\infty} a_n^2 \cos\left(\frac{n\pi(x - y)}{L_p}\right)$$

The random field a admits the **exact representation** in $[0, L] \subset [0, L_p]$

$$a(\omega, x) = E[a](x) + \sum_{n=0}^{\infty} a_n \left(y_n(\omega) \cos\left(\frac{n\pi x}{L_p}\right) + z_n(\omega) \sin\left(\frac{n\pi x}{L_p}\right) \right)$$

with

- ▶ $E[y_n] = E[z_n] = 0$
- ▶ $\text{Var}[y_n] = \text{Var}[z_n] = 1$
- ▶ $\{y_n, z_n\}_n$ uncorrelated.

This technique is known as **circulant embedding** [Dietrich-Newsam '97, Wood-Chan '94]. The Fourier coeffs. can be efficiently computed by FFT.

Warning: by extending the covariance function periodically one may introduce (artificial) discontinuities at $x - y = nL_p$ which reflect into a slow decay of the Fourier coefficients.

Advise: take L_p much larger than the characteristic correlation length so that $\text{Cov}_a(L_p) \approx 0$. But not too large!

Gaussian random fields

Of particular importance in some applications are the following class of random fields.

Definition 5.16

A random field $a : D \times \Omega \rightarrow \mathbb{R}$ is **Gaussian** if all finite dimensional distributions are Gaussian.

- ▶ That is, for a Gaussian random field (GRF) a and any $x_1, \dots, x_n \in D$ the random vector $\mathbf{z}(\omega) = (a(x_1, \omega), \dots, a(x_n, \omega))$ has a multivariate Gaussian distribution with mean $\boldsymbol{\mu} = [\mathbb{E}[a](x_1), \dots, \mathbb{E}[a](x_n)]$ and covariance matrix $C = \{\text{Cov}_a(x_i, x_j)\}_{i,j=1}^n$. Its probability density function thus reads

$$\rho(\mathbf{z}) = \frac{\exp\left\{-\frac{1}{2}(\mathbf{z} - \boldsymbol{\mu})^T C^{-1}(\mathbf{z} - \boldsymbol{\mu})\right\}}{(2\pi)^{d/2} \sqrt{\det(C)}}$$

provided Cov_a is strictly positive definite.

A GRF is defined uniquely by its mean and covariance function.

- ▶ For a centered Gaussian random field, mean square integrability implies integrability in $L^p(\Omega, \mathcal{A}, P)$ for any $p > 0$. Indeed let $z(x, \omega) = a(x, \omega) / \text{std}_a(x) \sim N(0, 1)$. Then

$$\mathbb{E}[|a(x, \cdot)|^p]^{\frac{1}{p}} = \text{std}_a(x) \mathbb{E}[|z(x, \omega)|^p]^{\frac{1}{p}} = c_p \mathbb{E}[a(x, \omega)^2]^{\frac{1}{2}}$$

where $c_p = \mathbb{E}[|z|^p]^{\frac{1}{p}} < \infty$, for any p since $z \sim N(0, 1)$.

- ▶ As a consequence, for GRFs mean square differentiability implies almost sure sample path continuity! More generally, mean square Hölder continuity $C^{0,\beta}(D)$ implies almost sure sample path Hölder continuity $C^{0,\alpha}(D)$ for any $0 \leq \alpha < \beta$. **Hint:** for any $\alpha < \beta$ take $p > d/(\beta - \alpha)$ and apply Kolmogorov's theorem.
- ▶ Consider the Karhunen-Loève expansion of a GRF:

$$a(x, \omega) = \mathbb{E}[a](x) + \sum_{i=1}^{\infty} \sqrt{\lambda_i} b_i(x) y_i(\omega) .$$

Then, y_i are independent $N(0, 1)$ random variables. Remember that uncorrelated Gaussian random variables are independent. The same holds for the Fourier expansion of stationary GRFs.

Examples of Covariance models

Let σ^2 be the variance of a GRF and l_c its correlation length.

- Exponential:

$$\text{Cov}_a(\|x - y\|) = \sigma^2 e^{-\frac{\|x-y\|}{l_c}}$$

Smoothness of the Gaussian field: almost surely $C^{0,\alpha}$, $\alpha < \frac{1}{2}$ (same regularity as a Brownian motion)

- Squared exponential (Gaussian covariance):

$$\text{Cov}_a(\|x - y\|) = \sigma^2 e^{-\frac{\|x-y\|^2}{2l_c^2}}$$

Smoothness of the Gaussian field: analytic almost surely

- Matérn covariance:

$$\text{Cov}_a(\|x - y\|) = \sigma^2 \frac{1}{\Gamma(\nu) 2^{\nu-1}} \left(\sqrt{2\nu} \frac{\|x - y\|}{l_c} \right)^\nu K_\nu \left(\sqrt{2\nu} \frac{\|x - y\|}{l_c} \right)^\nu$$

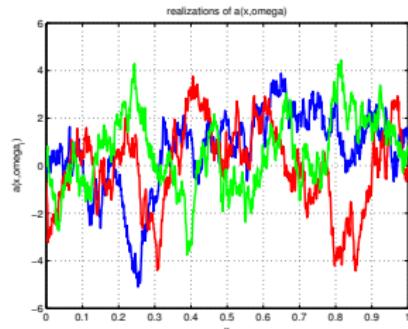
where Γ is the gamma function, K_ν is the modified Bessel function of the second kind, ν is a smoothness parameter.

Let $\nu = s + \beta$, with $s \in \mathbb{N}$ and $\beta \in (0, 1]$.

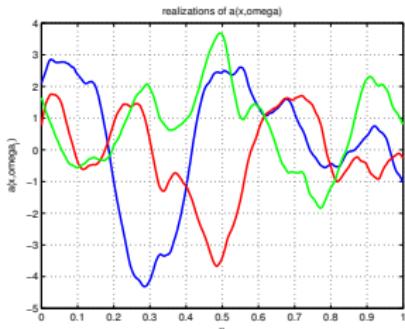
Smoothness of the GRF: almost surely $C^{s,\alpha}$ for any $\alpha < \beta$.

One recovers the exponential covariance for $\nu = \frac{1}{2}$, and the squared exponential covariance for $\nu \rightarrow \infty$.

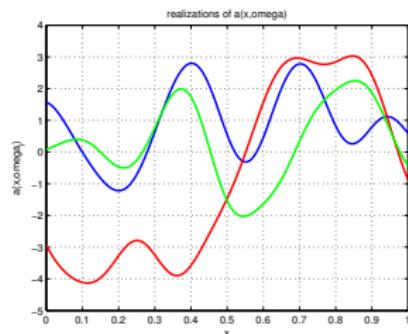
Comparison of Covariance models



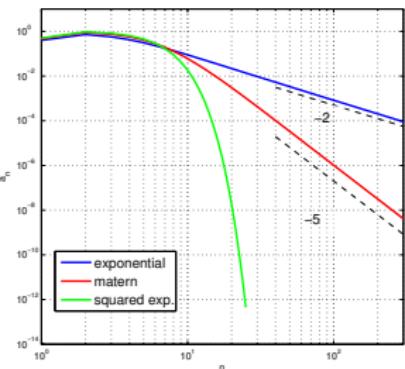
exponential covariance
 $\sigma^2 = 2, l_c = 0.1$



Matérn covariance
 $\sigma^2 = 2, l_c = 0.1, \nu = 2$



squared exp. covariance
 $\sigma^2 = 2, l_c = 0.1$



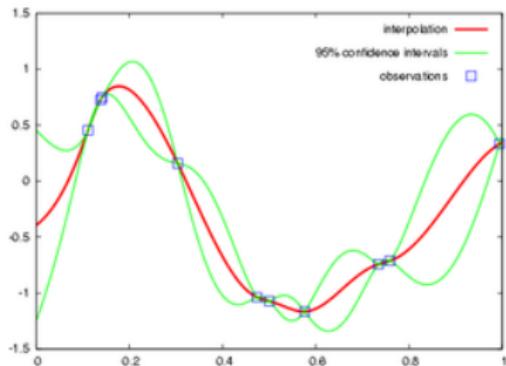
Decay of Fourier coeffs

Exercise 5.2 (Sampling GRF)

Given a finite set of points $x_1, \dots, x_N \in D$, use the definition of a GRF to interpret the values $z(x_1), \dots, z(x_N) \in D$ as a Gaussian vector. Use this to sample $z(x_1), \dots, z(x_N) \in D$.

Exercise 5.3 (Sampling conditioned GRFs: simple kriging)

Given a GRF z with known mean and covariance functions, we would like to compute the conditional distribution of z w.r.t. to measurements taken at a finite number of locations $x_1, \dots, x_M \in D$. Suppose $D \subset \mathbb{R}$.



One-dimensional data interpolation by kriging, with confidence intervals. Squares indicate the location of the data. The kriging interpolation is in red. The confidence intervals are in green.
(source: wikipedia+kriging)

Exercise (cont.)

Find an explicit expression for the distribution the conditioned value, $z(x)|_{(z(x_1), \dots, z(x_M))}$ for $x \in D$. What happens when $x \rightarrow x_1$? What is the influence of the covariance in the interpolation?

Hint: Observe that $(z(x_1), z(x_2)) = (Z_1, Z_2)$ is Gaussian with mean $(\mu(x_1), \mu(x_2))$ and covariance C . Then, the conditional distribution of Z_1 given Z_2 is Gaussian with a mean that is affine in Z_2 . We have

$$E[Z_1|Z_2] = \mu(x_1) + C(x_1, x_2)C(x_2, x_2)^{-1}(Z_2 - \mu(x_2))$$

and

$$\text{Cov}[Z_1|Z_2, Z_1|Z_2] = C(x_1, x_1) - C(x_1, x_2)C(x_2, x_2)^{-1}C(x_2, x_1).$$

Sampling of Gaussian random fields

For a review on different methods, please click [here](#) and read at the paper

- Liu, Y., Li, J., Sun, S., & Yu, B. (2019). Advances in Gaussian random field generation: a review. *Computational Geosciences*

Reminder of Gaussian random vectors

Definition 5.17 (Gaussian random vector)

A vector valued r.v. $Z \in \mathbb{R}^d$ is Gaussian with mean $\mu \in \mathbb{R}^d$ and covariance $C \in \mathbb{R}^{d \times d}$ if for any $\theta \in \mathbb{R}^d$

$$E[\exp(i\theta^T Z)] = \exp\left(i\theta^T \mu - \frac{\theta^T C \theta}{2}\right)$$

If C is strictly positive definite, then the PDF of Z is

$$\rho_Z(z) = \frac{\exp\left(-\frac{1}{2}(z - \mu)^T C^{-1}(z - \mu)\right)}{(2\pi)^{d/2} \sqrt{\det(C)}}$$

Observe: If C is singular with rank $k < d$, then $C = B\hat{C}B^T$ with $0 < \hat{C} \in \mathbb{R}^{k \times k}$ and a full rank matrix $B \in \mathbb{R}^{d \times k}$. Consequently, Z has the same distribution as an affine function of a lower dimensional Gaussian r.v. $\hat{Z} = BX + \mu$, where $X \sim N(0, \hat{C})$.

Clearly, \hat{Z} is Gaussian (why?) with mean μ and covariance

$$C_{\hat{Z}} = E[(BX)(BX)^T] = B\hat{C}B^T = C.$$

Exercise 5.4 (Properties of Gaussian random vectors)

Recall (or prove) the following properties:

1. The family of Gaussian r.vs is closed under affine operations.
2. If the covariance C_Z is diagonal, then the components of Z are independent r.vs, i.e. being uncorrelated implies independence when Z is Gaussian.
3. Any \mathbb{R}^d -valued Gaussian r.v can be written as an affine transformation of a vector with independent standard Gaussian components.
4. Sums of independent Gaussian r.vs are Gaussian.
5. Suppose $Z = (Z_1, Z_2)^T$ is Gaussian with mean μ and covariance C . Then, the conditional distribution of Z_1 given Z_2 is Gaussian with a mean that is affine in Z_2 . In particular, the best affine predictor coincides with the best non-linear predictor (i.e., the conditional expectation), with: We have

$$E[Z_1|Z_2] = E[Z_1] + \text{Cov}[Z_1, Z_2](\text{Cov}[Z_2, Z_2])^{-1}(Z_2 - E[Z_2])$$

and

$$\text{Cov}[Z_1|Z_2, Z_1|Z_2] = \text{Cov}[Z_1, Z_1] - \text{Cov}[Z_1, Z_2](\text{Cov}[Z_2, Z_2])^{-1}\text{Cov}[Z_2, Z_1].$$

Random Fields, further reading

- ▶ *The geometry of random fields.* Adler. SIAM, 2009.
- ▶ *Stochastic Calculus: Applications in Science and Engineering.* Grigoriu. Birkhäuser, 200.
- ▶ *Gaussian Measures.* Bogachev. AMS, 1998.
- ▶ *Advances in Gaussian random field generation: a review.* Liu et al. Comput. Geosci. 23:1011–104, 2019.

Monte Carlo Statistical Error

Goal: Approximate the expected value, $E[Y]$, by a sample average of M iid samples

$$\frac{\sum_{j=1}^M Y(\omega_j)}{M}$$

and choose M sufficiently large to control the statistical error,

$$E[Y] - \frac{\sum_{j=1}^M Y(\omega_j)}{M}.$$

For M independent samples of Y denote sample average $\mathcal{A}(Y; M)$, and sample standard deviation $\mathcal{S}(Y; M)$ of Y by

$$\begin{aligned}\mathcal{A}(Y; M) &\equiv \frac{1}{M} \sum_{j=1}^M Y(\omega_j) \\ \mathcal{S}(Y; M) &\equiv \left[\mathcal{A}(Y^2; M) - (\mathcal{A}(Y; M))^2 \right]^{1/2}.\end{aligned}$$

Let $\sigma_Y \equiv \{E[\|Y - E[Y]\|^2]\}^{1/2}$

Exercise 6.1

Compute the integral $I = \int_{[0,1]^d} f(x)dx$ by the Monte Carlo method, where we assume $f(x) : [0, 1]^d \rightarrow \mathbf{R}$.

We have

$$\begin{aligned} I &= \int_{[0,1]^d} f(x) dx \\ &= \int_{[0,1]^d} f(x)p(x) dx \text{ (where } p \text{ is the uniform pdf)} \\ &= E[f(X)] \text{ (where } X \text{ is uniformly distributed in } [0, 1]^d) \\ &\simeq \sum_{j=1}^M \frac{f(X(\omega_j))}{M} \\ &\equiv I_M, \end{aligned}$$

The values $\{X(\omega_j)\}$ are sampled uniformly in the cube $[0, 1]^d$, by sampling the components $x_i(\omega_n)$ independently and uniformly on the interval $[0, 1]$.

Remark 6.1 (Random number generators)

One can generate approximate random numbers, so called pseudo random numbers, see the lecture notes. By using transformations, one can also generate more complicated distributions in terms of simpler ones.

Sampling with MATLAB:

Standard uniform distribution:

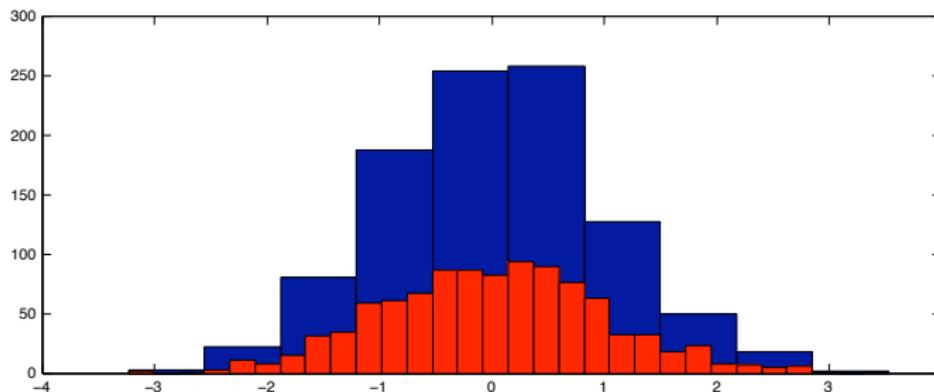
```
x = rand(M,N);
```

Standard normal distribution:

```
x = randn(M,N);
```

Example 6.1

```
x = randn(1000,1);  
hist(x);  
hist(x,30);
```



Example 6.2 (Inverse transform sampling)

Let Y be a given real valued random variable with

$$P(Y \leq x) = F_Y(x).$$

Suppose that we want to sample iid from Y and that we can cheaply compute $F_Y^{-1}(u)$, $u \in [0, 1]$.

Then, take U to be uniform distributed in $[0, 1]$ and let

$$Y(\omega) = F_Y^{-1}(U(\omega)).$$

We then have

$$P(Y \leq x) = P(F_Y^{-1}(U) \leq x) = P(U \leq F_Y(x)) = F_Y(x)$$

as we wanted!

Acceptance-rejection sampling

It generates sampling values from an arbitrary pdf $\rho_Y(x)$ by using an auxiliary pdf $\rho_X(x)$.

Assumptions:

- (i) It is simple to sample from ρ_X ,
- (ii) There exists $0 < \epsilon \leq 1$ s.t.

$$\epsilon \frac{\rho_Y(x)}{\rho_X(x)} \leq 1, \text{ for all } x.$$

Idea:

Rejection sampling is usually used in cases where the form of ρ_Y makes sampling difficult.

Instead of sampling directly from ρ_Y , we use samples from ρ_X .
These samples from ρ_X are probabilistically accepted or rejected.

Acceptance-rejection sampling

The steps below generate a single realization of Y with pdf ρ_Y .

Step 1 Set $k = 1$

Step 2 Sample two independent random variables:
 X_k from ρ_X and $U_k \sim U(0, 1)$.

Step 3 If $U_k \leq \epsilon \frac{\rho_Y(X_k)}{\rho_X(X_k)}$ then accept $Y = X_k$ be a sample from ρ_Y .

Otherwise reject X_k , increment k by 1 and go to Step 2.

Let us see that Y sampled by acceptance-rejection has indeed density ρ_Y
We have the *acceptance probability*

$$\begin{aligned} P\left(U_k \leq \epsilon \frac{\rho_Y(X_k)}{\rho_X(X_k)}\right) &= \int \int_0^{\epsilon \frac{\rho_Y(x)}{\rho_X(x)}} du \rho_X(x) dx \\ &= \epsilon \int \frac{\rho_Y(x)}{\rho_X(x)} \rho_X(x) dx \\ &= \epsilon \int \rho_Y(x) dx \\ &= \epsilon \end{aligned}$$

Let $K(\omega)$ be the first value of k for which X_k is accepted as a realization of Y . We want to show that X_K has the desired density, ρ_Y .

Consider an open set B

$$\begin{aligned}
 P(X_K \in B) &= \sum_{k \geq 1} P(X_k \in B, K = k) \\
 &= \sum_{k \geq 1} \underbrace{P\left(X_k \in B, U_k \leq \epsilon \frac{\rho_Y(X_k)}{\rho_X(X_k)}\right)}_{\text{does not depend on } k} \prod_{m=1}^{k-1} \underbrace{P\left(U_m > \epsilon \frac{\rho_Y(X_m)}{\rho_X(X_m)}\right)}_{=1-\epsilon} \\
 &= P\left(X_k \in B, U_k \leq \epsilon \frac{\rho_Y(X_k)}{\rho_X(X_k)}\right) \underbrace{\sum_{k \geq 1} (1-\epsilon)^{k-1}}_{=1/\epsilon}
 \end{aligned}$$

To finish compute

$$\begin{aligned} P\left(X_k \in B, U_k \leq \epsilon \frac{\rho_Y(X_k)}{\rho_X(X_k)}\right) &= \int_B \int_0^{\epsilon \frac{\rho_Y(x)}{\rho_X(x)}} du \rho_X(x) dx \\ &= \epsilon \int_B \frac{\rho_Y(x)}{\rho_X(x)} \rho_X(x) dx \\ &= \epsilon \int_B \rho_Y(x) dx \end{aligned}$$

which implies

$$P(X_K \in B) = \int_B \rho_Y(x) dx$$

as we claimed.

Remark 6.2 (Acceptance-rejection cost)

Compute the expected number of samples per accepted ones:

$$\begin{aligned} E[K] &= \sum_{k \geq 1} k P(K = k) \\ &= \sum_{k \geq 1} k (1 - \epsilon)^{k-1} \epsilon \\ &= 1/\epsilon \end{aligned}$$

Can you interpret this result?

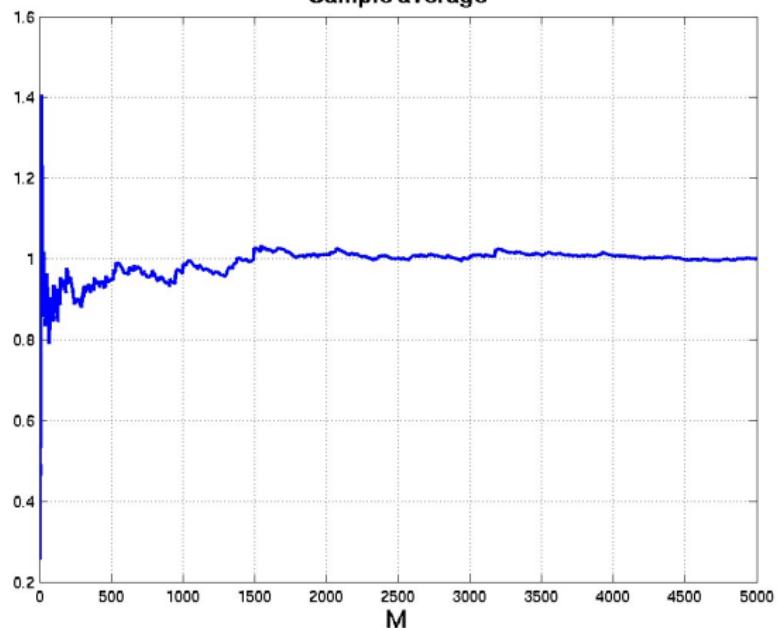
Monte Carlo: Numerical example

Consider the computation of the integral

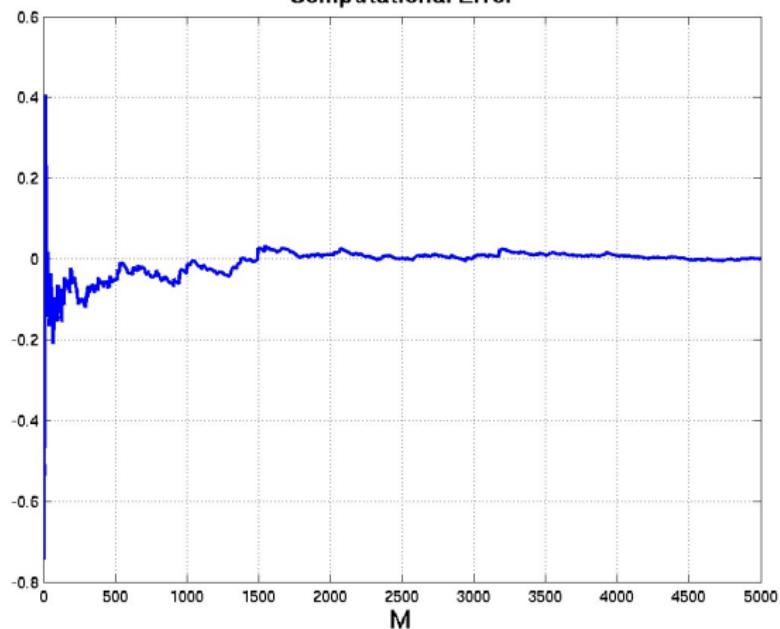
$$1 = \int_{[0,1]^N} \exp\left(\sum_{n=1}^N x_n\right) dx_1 \dots dx_N / (e - 1)^N$$

```
M = 1e6; % Max. number of realizations
N = 20; % Dimension of the problem
u = rand(M,N); f = exp(sum(u'));
run_aver = cumsum(f)./(((1:M)').*(exp(1)-1)^N);
plot(1:M, run_aver),
figure, plot(1:M, run_aver), xlabel 'M'
figure,plot(1:M,(run_aver-1)), xlabel 'M'
figure,semilogy(1:M,abs(run_aver-1)), xlabel 'M',
```

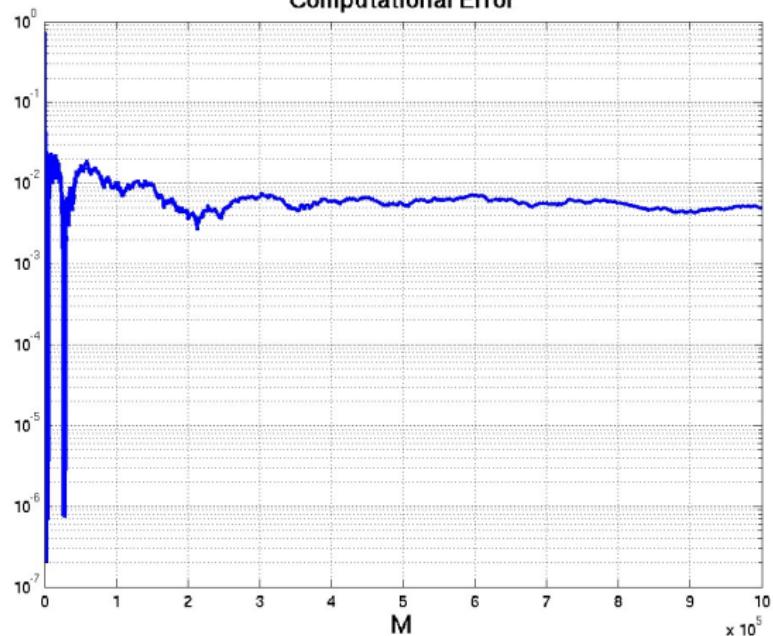
Sample average



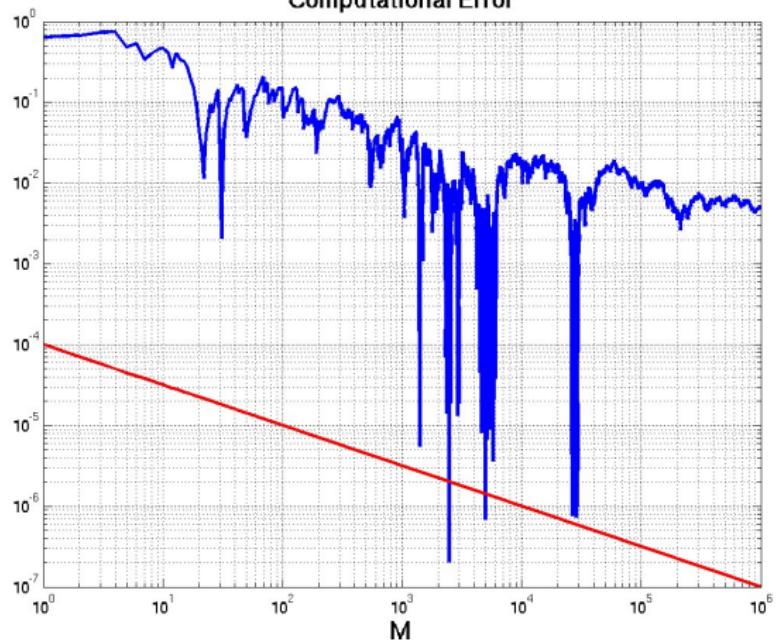
Computational Error



Computational Error



Computational Error



Monte Carlo error analysis

The Central Limit Theorem is the fundamental result to understand the statistical error of Monte Carlo methods.

Theorem 6.1 (The Central Limit Theorem)

Assume ξ_j , $j = 1, 2, 3, \dots$ are independent, identically distributed (i.i.d) and $E[\xi_j] = 0$, $E[\xi_j^2] = 1$. Then

$$\sum_{j=1}^M \frac{\xi_j}{\sqrt{M}} \rightharpoonup \nu, \quad \text{as } M \rightarrow \infty, \tag{48}$$

where ν is $N(0, 1)$ and \rightharpoonup denotes convergence of the distributions, also called weak convergence, i.e. the convergence (48) means

$E[g(\sum_{j=1}^M \xi_j / \sqrt{M})] \rightarrow E[g(\nu)]$ for all bounded and continuous functions g .

Characteristic function

Let X be a r.v. then

$$f(t) = E[e^{itX}]$$

is called the *characteristic function* of X . This function identifies completely the distribution of X , namely

Theorem 6.2

Two distributions having the same characteristic function are identical

Example: Consider a standard normal distribution, $X \sim N(0, 1)$. Then

$$f(t) = E[e^{itX}] = e^{-\frac{t^2}{2}}$$

In fact, we have inversion formulas closely related to the Fourier transform³

Theorem 6.3

Let x_1, x_2 be continuity points of F_X . Then

$$F(x_2) - F(x_1) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \frac{e^{-itx_2} - e^{-itx_1}}{-it} f(t) dt$$

³See [Petrov]

Proof. Consider the characteristic function $f(t) = E[e^{it\xi_j}]$. Then its derivatives satisfy

$$f^{(m)}(t) = E[i^m \xi_j^m e^{it\xi_j}]. \quad (49)$$

For the sample average of the ξ_j vars we have

$$\begin{aligned} E[e^{it \sum_{j=1}^M \xi_j / \sqrt{M}}] &= f\left(\frac{t}{\sqrt{M}}\right)^M \\ &= (f(0) + \frac{t}{\sqrt{M}} f'(0) + \frac{1}{2} \frac{t^2}{M} f''(0) + o\left(\frac{t^2}{M}\right))^M. \end{aligned}$$

The representation (49) implies

$$\begin{aligned}f(0) &= E[1] = 1, \\f'(0) &= iE[\xi_n] = 0, \\f''(0) &= -E[\xi_n^2] = -1.\end{aligned}$$

Therefore

$$\begin{aligned}E[e^{it \sum_{j=1}^M \xi_j / \sqrt{M}}] &= \left(1 - \frac{t^2}{2M} + o\left(\frac{t^2}{M}\right)\right)^M \\&\rightarrow e^{-t^2/2}, \quad \text{as } M \rightarrow \infty \\&= \int_{\mathbb{R}} \frac{e^{itx} e^{-x^2/2}}{\sqrt{2\pi}} dx,\end{aligned}\tag{50}$$

and we conclude that the Fourier transform of the pdf

(i.e. the characteristic function) of $\sum_{j=1}^M \xi_j / \sqrt{M}$ converges to the Fourier transform of the standard normal distribution. Therefore, denoting by $F(g)$ the complex conjugate of the Fourier Transform of g , we have

$$\begin{aligned}
E[g(\sum_{j=1}^M \xi_j / \sqrt{M})] &= \int_{\mathbb{R}} g(x) \rho_{\sum_{j=1}^M \xi_j / \sqrt{M}}(x) dx \\
&\stackrel{\text{Parseval}}{=} \frac{1}{2\pi} \int_{\mathbb{R}} E[e^{it \sum_{j=1}^M \xi_j / \sqrt{M}}] F(g)(t) dt \\
&\rightarrow \frac{1}{2\pi} \int_{\mathbb{R}} e^{-t^2/2} F(g)(t) dt \\
&\stackrel{\text{Parseval}}{=} E[g(\nu)].
\end{aligned}$$

Exercise 6.2

What is the error of $I_M - I$ in Example 6.1?

Let the error ϵ_M be defined by

$$\begin{aligned}\epsilon_M &= \sum_{j=1}^M \frac{f(x_j)}{M} - \int_{[0,1]^d} f(x) dx \\ &= \sum_{j=1}^M \frac{f(x_j) - E[f(x)]}{M}.\end{aligned}$$

By the Central Limit Theorem, $\sqrt{M}\epsilon_M \rightarrow \sigma\nu$, where ν is $N(0, 1)$ and

$$\begin{aligned}\sigma^2 &= \int_{[0,1]^d} f^2(x) dx - \left(\int_{[0,1]^d} f(x) dx \right)^2 \\ &= \int_{[0,1]^d} \left(f(x) - \int_{[0,1]^d} f(x) dx \right)^2 dx.\end{aligned}$$

In practice, σ^2 is approximated by

$$\hat{\sigma}^2 = \frac{1}{M-1} \sum_{j=1}^M \left(f(x_j) - \sum_{m=1}^M \frac{f(x_m)}{M} \right)^2.$$

More explicitly, the CLT states that

$$\sqrt{M}\epsilon_M \rightharpoonup \sigma\nu,$$

where ν is $N(0, 1)$.

This implies that for any set $B = (-C, C)$

$$P(\sqrt{M}\epsilon_M \in B) \rightarrow P(N(0, 1) \in B)$$

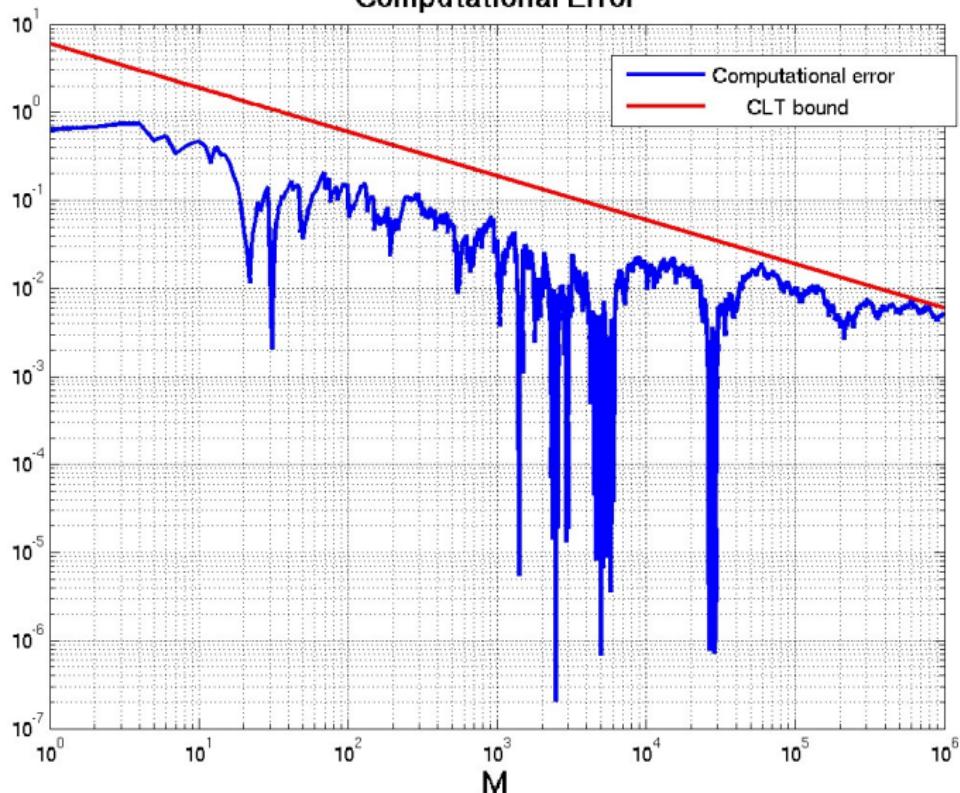
Given a constant, $0 < \alpha \ll 1$, choose $C = C_\alpha$ s.t. the following confidence level constraint is satisfied

$$P(N(0, 1) \in B) = P(|N(0, 1)| \leq C_\alpha) = \int_{|x| \leq C_\alpha} \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} dx = 1 - \alpha$$

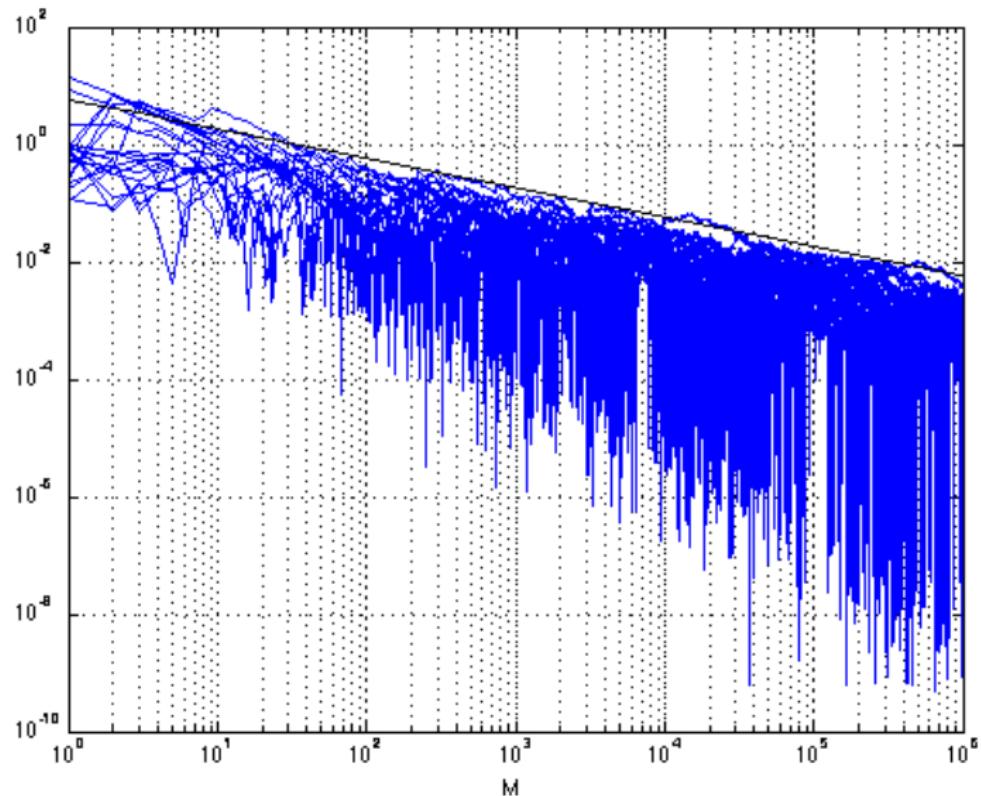
We now have

$$P\left(|\epsilon_M| \leq \frac{C_\alpha \sigma}{\sqrt{M}}\right) \approx 1 - \alpha, \text{ for large } M$$

Computational Error



Approximate error bound: $c_o \hat{\sigma} / \sqrt{M}$. Here $c_o = 3$.



Approximate error bound: $c_o \hat{\sigma} / \sqrt{M}$. Here $c_o = 3$. Showing 20 runs of Monte Carlo Sampling.

Exercise 6.3 (Monte Carlo vs. deterministic quadrature)

Consider again the approximation of $I := \int_{[0,1]^d} f(x)dx$ with $f \in C^\infty([0, 1]^d)$.

1. Discuss the amount of computational work needed to approximate I via Monte Carlo for a given precision TOL and a given confidence level α .
2. Consider a one dimensional quadrature rule⁴ with order $p > 0$ and nodes (s_n) and weights (t_n) of the form

$$\int_{[a,b]} g(x)dx = (b-a) \sum_{n=1}^N g(a + s_n(b-a))t_n + \mathcal{O}(|g^{(p)}|_\infty |b-a|^p).$$

Propose a composite quadrature rule for the $[0, 1]$ integration problem. Discuss accuracy versus numerical work.

3. Propose a full tensor product quadrature rule for the $[0, 1]^d$ integration problem. Discuss accuracy versus numerical work in terms of the problem dimension, d . Compare your results with the ones corresponding to Monte Carlo.

⁴see http://en.wikipedia.org/wiki/Numerical_integration

Theorem 6.4 (Law of the iterated logarithm)

Assume ξ_j , $j = 1, 2, 3, \dots$ are independent, identically distributed (i.i.d) and $E[\xi_j] = 0$, $E[\xi_j^2] = 1$. Then

$$\limsup_{M \rightarrow \infty} \frac{|\sum_{m=1}^M \xi_m|}{\sqrt{2M \log \log M}} = 1, \text{ a.s.}$$

The original statement (1924) of the law of the iterated logarithm is due to A. Y. Khinchin. Another statement was given by A.N. Kolmogorov (1929). See (Theorem 3.52 in Breiman's book).

Exercise 6.4

Reproduce the numerical results for the computation of I_M , including now the law of the iterated logarithm result.

The Delta Method and Small Noise Approximations

Consider a random variable $X = \mu + \sigma Z$, where Z is such that $E[Z] = 0$, $\text{Var}[Z] = 1$, i.e. $E[X] = \mu$ and $\text{Var}[X] = \sigma^2$.

Idea: To approximate the distribution of $g(X)$, with g smooth and $\sigma \ll 1$, use Taylor expansions around $\mu = E[X]$.

Then

$$\begin{aligned} g(X) - g(E[X]) &= g'(E[X])(X - E[X]) \\ &\quad + \frac{g''(E[X])}{2}(X - E[X])^2 + o((X - E[X])^2) \\ &= g'(\mu)\sigma Z + \frac{g''(\mu)}{2}\sigma^2 Z^2 + o(\sigma^2 Z^2) \end{aligned}$$

Assume $g'(E[X]) \neq 0$, and write

$$Y = \frac{g(X) - g(E[X])}{\sigma g'(\mu)} = Z + \frac{g''(\mu)}{2g'(\mu)}\sigma Z^2 + o(\sigma Z^2)$$

We conclude that up to second order, the moments of Y are well approximated by

$$E[Y] = \frac{g''(\mu)}{2g'(\mu)}\sigma + o(\sigma)$$

$$E[Y^2] = 1 + \frac{g''(\mu)}{g'(\mu)}\sigma E[Z^3] + o(\sigma).$$

and this has implications on the moments of $g(X)$, the ones that are the goal of the computation.

Exercise 6.5

What happens if $g'(\mu) = 0$?

Let us look at the characteristic function of $g(X)$.

$$\begin{aligned}E[e^{it(g(X) - E[g(X)])}] &= E[1 + it(g(X) - E[g(X)]) - \frac{t^2}{2}(g(X) - E[g(X)])^2] + \dots \\&= 1 - \frac{t^2}{2}E[(g(X) - E[g(X)])^2] + \dots \\&= 1 - \frac{t^2}{2}(\sigma g'(\mu))^2 E[Y^2] + \dots \\&= 1 - \frac{t^2}{2}(\sigma g'(\mu))^2 \left(1 + \frac{g''(\mu)}{g'(\mu)}\sigma E[Z^3]\right) + \dots \\&= 1 - \frac{t^2}{2}(\sigma g'(\mu))^2 + \dots\end{aligned}$$

The previous characteristic function computation justifies the approximation *in distribution sense* for $\sigma|g'(\mu)| \ll 1$

$$g(X) \approx g(\mu) + \sigma g'(\mu)N(0, 1)$$

Exercise 6.6

Verify that if $g'(\mu) \neq 0$ and $Z \sim N(0, 1)$ then we have a CLT approximation, i.e. Y becomes normal as $\sigma \rightarrow 0$.

Exercise 6.7

Compare the above discussion for small noise with the log posterior expansion and gaussian approximation we did for approximate error bars of parameter uncertainty in Bayesian Statistics.

Question: How fast is the convergence in the Central Limit Theorem? Consider the scaled random variable

$$Z_M \equiv \frac{\sqrt{M}}{\sigma_Y} (\mathcal{A}(Y; M) - E[Y])$$

with cumulative distribution function

$$F_{Z_M}(x) \equiv P(Z_M \leq x), \quad \forall x \in \mathbb{R}.$$

Theorem 6.5 (Berry–Esseen)

Assume

$$\lambda \equiv \frac{(E[|Y - E[Y]|^3])^{1/3}}{\sigma_Y} < +\infty,$$

then we have a uniform estimate in the central limit theorem

$$|F_{Z_M}(x) - \Phi(x)| \leq \frac{C_{BE} \lambda^3}{(1 + |x|)^3 \sqrt{M}}$$

Here Φ is the distribution function of $N(0, 1)$,

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \exp\left(-\frac{s^2}{2}\right) ds. \quad (51)$$

and $C_{BE} = 30.51175\dots$

By the Berry–Esseen thm., the statistical error

$$\mathcal{E}_S(Y; M) \equiv E[Y] - A(Y; M)$$

satisfies, $\forall c_0 > 0$,

$$P\left(\left[|\mathcal{E}_S(Y; M)| \leq c_0 \frac{\sigma_Y}{\sqrt{M}}\right]\right) \geq 2\Phi(c_0) - 1 - 2 \frac{C_{BE} \lambda^3}{(1 + c_0)^3 \sqrt{M}}.$$

In practice choose $c_0 \geq 1.65$, $\Rightarrow 1 > 2\Phi(c_0) - 1 \geq 0.901$ and the event

$$|\mathcal{E}_S(Y; M)| \leq E_S(Y; M) \equiv c_0 \frac{\mathcal{S}(Y; M)}{\sqrt{M}} \quad (52)$$

has probability close to one, which involves the additional step to approximate σ_Y by $\mathcal{S}(Y; M)$. Thus, in the computations $E_S(Y; M)$ is a good approximation of the statistical error $\mathcal{E}_S(Y; M)$.

Numerical Example:

Taking $c_0 = 3$ yields $2\Phi(c_0) - 1 = 0.9973\dots$ and

$$P \left(\left[|\mathcal{E}_S(Y; M)| \leq 3 \frac{\sigma_Y}{\sqrt{M}} \right] \right) \geq 0.9973 - 0.4766 \frac{\lambda^3}{\sqrt{M}}.$$

In particular, if Y is a uniform random variable, then

$$\lambda^3 = \frac{1}{4} 3^{1.5} = 1.299\dots$$

and we have the bound

$$P \left(\left[|\mathcal{E}_S(Y; M)| \leq 3 \frac{\sigma_Y}{\sqrt{M}} \right] \right) \geq 0.9973 - \frac{0.6196}{\sqrt{M}}.$$

Obs: the last term on the right will determine the confidence level for $M \leq 5 \times 10^4$

Example 6.3 (CLT approximation of cdf)

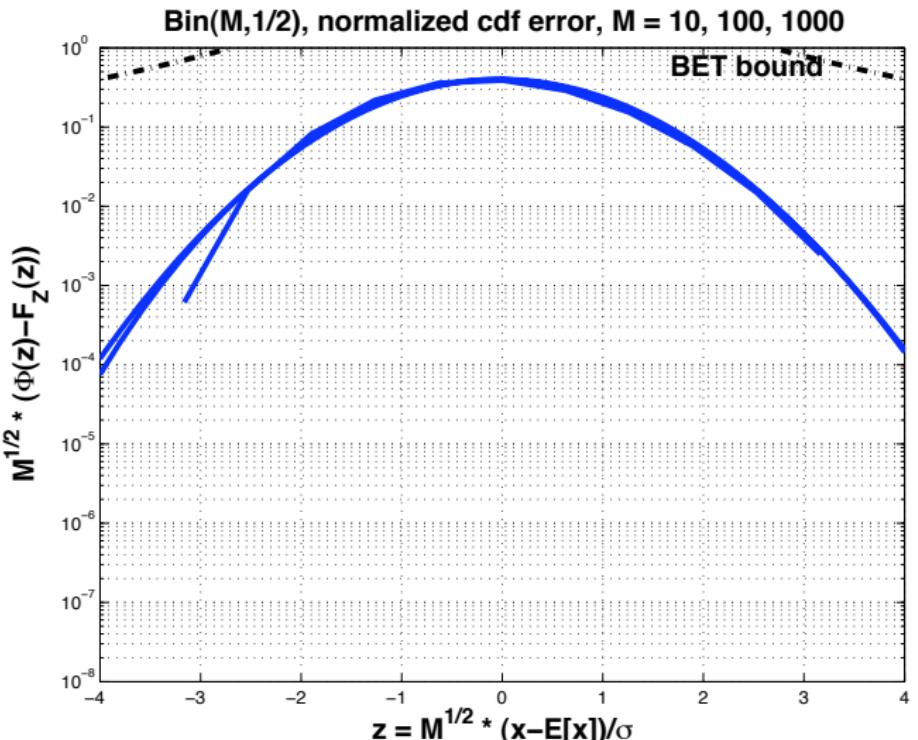
Consider a Binomial r.v. with parameter $p = 1/2$,

$$X = \sum_{i=1}^M Y_i$$

and Y_i iid Bernoulli r.vars., $\sigma^2 = p(1 - p)$. Let

$$Z = \frac{(X - Mp)}{\sigma\sqrt{M}},$$

then we compare its cdf (computed exactly) vs. the CLT approximation, $\Phi(z)$. We do it for several values of $M \dots$



Numerical results from sampling with Bernoulli random variables

Observe that on then previous figure:

$$\log \left(\sqrt{M} |F_{Z_M}(z) - \Phi(z)| \right) \approx c_1 - c_2 z^2$$

or simply

$$|F_{Z_M}(z) - \Phi(z)| \approx \frac{C e^{-c_2 z^2}}{\sqrt{M}}.$$

Compare the above with the B.-E. Theorem (assuming only bounded 3:rd moment)

$$|F_{Z_M}(x) - \Phi(x)| \leq \frac{C_{BE} \lambda^3}{(1 + |x|)^3 \sqrt{M}}.$$

Note that the binomial distribution has finite exponential moments, i.e.

$$E [\exp (\Theta Y_i)] < \infty, \forall \Theta.$$

Edgeworth Expansion

$$F_{Z_M}(x) - \Phi(x) = \frac{\lambda^3}{6\sqrt{M}}(1-x^2)\varphi(x) + o(1/\sqrt{M})$$

Compare with

$$|F_{Z_M}(x) - \Phi(x)| \leq \frac{C_{BE} \lambda^3}{(1+|x|)^3 \sqrt{M}}$$

Here $\varphi = \Phi'$, Φ is the distribution function of $N(0, 1)$,

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \exp\left(-\frac{s^2}{2}\right) ds. \quad (53)$$

and $C_{BE} = 30.51175\dots$

Exercise 6.8

Reproduce the computations in Example 6.3, adding the first term of the Edgeworth expansion to the plot. Comment your results.

Adaptive Monte Carlo

For a given $\text{TOL}_S > 0$, the goal is to find M such that $E_S(Y; M) \leq \text{TOL}_S$. The following algorithm adaptively finds the number of realizations M to compute the sample average $\mathcal{A}(Y; M)$ as an approximation to $E[Y]$. With probability close to one, depending on c_0 , the statistical error in the approximation is then bounded by TOL_S .

routine Monte-Carlo(TOL_S , Y , M_0 ; EY)

 Set the batch counter $m = 1$, $M[1] = M_0$ and $E_S[1] = 2 TOL_S$.

Do while ($E_S[m] > TOL_S$)

 Compute $M[m]$ new samples of Y , along with the sample average $EY \equiv \mathcal{A}(Y; M[m])$, the sample variance $S[m] \equiv S(Y; M[m])$ and

 the deviation $E_S[m + 1] \equiv E_S(Y; M[m])$.

 Compute $M[m + 1]$ by

 change_M ($M[m]$, $S[m]$, TOL_S ; $M[m + 1]$).

 Increase m by 1.

end-do

end of Monte-Carlo

routine change_ M (M_{in} , S_{in} , TOL_S ; M_{out})

$$\begin{aligned} M^* &= \min \left\{ \text{integer part} \left(\frac{c_0 S_{\text{in}}}{\text{TOL}_S} \right)^2, \text{MCH} \times M_{\text{in}} \right\} \\ n &= \text{integer part} (\log_2 M^*) + 1 \\ M_{\text{out}} &= 2^n. \end{aligned} \tag{54}$$

end of change_ M

Remark 6.3 (Parameters for change M)

Here, M_0 is a given initial value for M , and $MCH > 1$ is a positive integer parameter introduced to avoid a large new number of realizations in the next batch due to a possibly inaccurate sample standard deviation $S[m]$. Indeed, $M[m + 1]$ cannot be greater than $MCH \times M[m]$.

We will use $MCH = 2$ in the next example:

Numerical Example: Adaptive MC, TOL = 1e - 2 to approximate

$$1 = \int_{[0,1]^N} \exp\left(\sum_{n=1}^N x_n\right) dx_1 \dots dx_N / (e - 1)^N,$$

$N = 20$.

M	Sample E	Sample std Error est.	Comp.	Error
200	1.2526	4.4003e+00	9.3e-01	2.5e-01
400	1.0411	2.1068e+00	3.1e-01	4.1e-02
800	0.9889	1.8730e+00	2.0e-01	-1.1e-02
1600	1.0699	2.3410e+00	1.7e-01	7.0e-02
3200	1.0324	1.9087e+00	1.0e-01	3.2e-02
6400	1.0764	2.1659e+00	8.1e-02	7.6e-02
12800	1.0212	2.0788e+00	5.5e-02	2.1e-02
25600	9.9549	1.9036e+00	3.6e-02	-4.5e-03
51200	1.0104	1.8946e+00	2.5e-02	1.0e-02
102400	0.98721	1.8248e+00	1.7e-02	-1.3e-02
204800	0.99375	1.9103e+00	1.3e-02	-6.6e-03
409600	0.99611	1.9320e+00	9.1e-03	-3.9e-03

Question: Can you compute the confidence level corresponding to the above computations as a function of M using the BE Theorem?
More information and a related adaptive Monte Carlo algorithm can be found here:

“On non-asymptotic optimal stopping criteria in Monte Carlo Simulations”, by C. Bayer, H. Hoel, E. Von Schwerin, and R. Tempone. SIAM Journal on Scientific Computing (SISC), 36(2), pp. A869–A885, 2014.

Other computations with Monte Carlo:

Example 6.4 (Medians and quantiles)

The median of a continuous random variable Y is defined as the smallest value of m_Y for which: $P(Y \leq m_Y) = 0.5$. Note that m_Y is defined as the root of an equation and is not of the form $E[g(Y)]$ for some given function g .

Question: How do you choose M in this case?

Example 6.5 (Optimization)

Suppose that we want to solve

$$\min_{\alpha \in A} E[g(Y, \alpha)]$$

where Y is a given random vector and α is a deterministic vector of "control variables".

Since in general the expectation will not be computable in closed form, we may approximate

$$E[g(Y, \alpha)] \approx \frac{1}{M} \sum_{j=1}^M g(Y(\omega_j), \alpha)$$

with $Y(\omega_j)$ iid samples from Y .

Question: How do you choose M in this case?

Monte Carlo Sampling for PDEs

For simplicity, let $\mathbf{y} = (y_1, \dots, y_N) \in \Gamma \subset \mathbb{R}^N$ be a random vector with density $\rho : \Gamma \rightarrow \mathbb{R}_+$, $u : \Gamma \rightarrow V$ a Hilbert-space valued function satisfying $u \in L^2_\rho(\Gamma; V)$, and $Q : V \rightarrow \mathbb{R}$ a continuous functional on V (possibly non-linear), s.t. $\mathbb{E}[|Q(u(\mathbf{y}))|^p] < \infty$ for p sufficiently large.

Goal: Compute $\mathbb{E}[Q(u(\mathbf{y}))]$

Classic Monte Carlo approach:

Approximate expectations by sample averages: Let $\{\mathbf{y}(\omega_m)\}_{m=1}^M$ a sample of M i.i.d. replica of \mathbf{y} :

$$\mathbb{E}[Q(u(\mathbf{y}))] \approx \frac{1}{M} \sum_{m=1}^M Q(u(\mathbf{y}(\omega_m)))$$

Here for each $\mathbf{y}(\omega_m)$ we have to solve for $u(\mathbf{y}(\omega_m))$ through the PDE and evaluate the q.o.i. $Q(u(\mathbf{y}(\omega_m)))$.

Pros and cons of Monte Carlo

- Pros** Re-usability of deterministic legacy codes, convergence rate $M^{-0.5}$ resilient w.r.t. length of \mathbf{y} and regularity of $u(\cdot)$.
- Cons** Slow convergence, does not exploit possibly available regularity.

Monte Carlo error analysis

Suppose that the PDE is discretized by some numerical method (e.g., FDM, FEM, spectral methods, ...), so that in practice we compute discrete solutions $u_h(\mathbf{y}(\omega_m)) \approx u(\mathbf{y}(\omega_m))$, $m = 1, \dots, M$. Then our estimator is

$$\mathbb{E}[Q(u(\mathbf{y}))] \approx \mathbb{E}[Q(u_h(\mathbf{y}))] \approx \frac{1}{M} \sum_{m=1}^M Q(u_h(\mathbf{y}(\omega_m)))$$

Error splitting:

$$\begin{aligned} & \mathbb{E}[Q(u(\mathbf{y}))] - \frac{1}{M} \sum_{m=1}^M Q(u_h(\mathbf{y}(\omega_m))) \\ &= \underbrace{\mathbb{E}[Q(u(\mathbf{y})) - Q(u_h(\mathbf{y}))]}_{\text{discretization error } \mathcal{E}^Q(h) \text{ (bias)}} + \underbrace{\mathbb{E}[Q(u_h(\mathbf{y}))] - \frac{1}{M} \sum_{m=1}^M Q(u_h(\mathbf{y}(\omega_m)))}_{\text{statistical error } \mathcal{E}_h^Q(M)} \end{aligned}$$

Monte Carlo error analysis. Discretization Error

Let us consider a functional $Q : V \rightarrow \mathbb{R}$, with $Q(0) = 0$, that is **globally Lipschitz**, i.e.

$$\exists C_Q > 0 \text{ s.t. } |Q(u) - Q(v)| \leq C_Q \|u - v\|_V, \quad \forall u, v \in V.$$

Assumption 7.1

There exists $\alpha > 0$ and $0 < C_u \in L_p^p(\Gamma)$ for some $p \geq 1$ such that

$$\|u(\mathbf{y}) - u_h(\mathbf{y})\|_V \leq C_u(\mathbf{y}) h^\alpha, \quad \text{for } 0 < h < h_0, \quad (55)$$

and a.a. $\mathbf{y} \in \Gamma$.

Then

$$\begin{aligned} |\mathcal{E}^Q(h)| &= |\mathbb{E}[Q(u(\mathbf{y})) - Q(u_h(\mathbf{y}))]| = \left| \int_{\Gamma} (Q(u(\mathbf{y})) - Q(u_h(\mathbf{y}))) \rho(\mathbf{y}) d\mathbf{y} \right| \\ &\leq C_Q \int_{\Gamma} \|u(\mathbf{y}) - u_h(\mathbf{y})\|_V \rho(\mathbf{y}) d\mathbf{y} \leq C_Q h^\alpha \int_{\Gamma} C_u(\mathbf{y}) \rho(\mathbf{y}) d\mathbf{y} \\ &\leq C_Q \|C_u\|_{L_p^p} h^\alpha \end{aligned}$$

Usually $C_u(\mathbf{y})$ depends on some stronger norm $\|u(\mathbf{y})\|_W$. For instance, for P1 finite elements and $V = H^1(D)$, then one has $\alpha = \min\{s-1, 1\}$ and $W = H^s(D)$, $s > 1$.

Example – elliptic PDE with random coefficient

Consider the following elliptic problem with uniformly bounded random coefficients

$$\begin{cases} -\operatorname{div}(a(x, \mathbf{y}) \nabla u(x, \mathbf{y})) = f(x), & x \in D, \\ u(x, \mathbf{y}) = 0, & x \in \partial D \end{cases}, \quad \forall \mathbf{y} \in \Gamma \subset \mathbb{R}^N$$

where $D \subset \mathbb{R}^d$ is a convex, Lipschitz domain and $f \in L^2(D)$.

Random coefficient model:

- ▶ $a(x, \mathbf{y}) = \bar{a} + \sum_{i=1}^N \sqrt{\lambda_i} y_i b_i(x)$ (here N could be “large” Question: can it be ∞ ?)
- ▶ $y_i \sim \mathcal{U}(-\sqrt{3}, \sqrt{3})$ i.i.d. (so $\Gamma = [-\sqrt{3}, \sqrt{3}]^N$)
- ▶ $b_i \in C^\infty(D)$
- ▶ $\sum_{i=1}^N \sqrt{3\lambda_i} \|b_i\|_\infty \leq \delta \bar{a}, \quad 0 < \delta < 1.$

we have uniform coercivity and continuity, namely

$$\implies (1 - \delta) \bar{a} \leq a(x, \mathbf{y}) \leq (1 + \delta) \bar{a}, \quad \forall \mathbf{y} \in \Gamma$$

Denote $H_0^1(D) = \{v \in H^1(D), v = 0 \text{ on } \partial D\}$ endowed with the norm $\|v\|_{H_0^1(D)} = \|\nabla v\|_{L^2(D)}$.

Under the previous assumptions, there exists a unique solution $u(\mathbf{y}) \in H_0^1(D)$, such that

$$\|u(\mathbf{y})\|_{H_0^1(D)} \leq \frac{\textcolor{red}{C}_P \|f\|_{L^2(D)}}{(1 - \delta)\bar{a}}, \text{ for all } \mathbf{y} \in \Gamma$$

where C_P is the Poincaré constant, i.e.

$$\|v\|_{L^2(D)} \leq \textcolor{red}{C}_P \|v\|_{H_0^1(D)}, \text{ for all } v \in H_0^1(D).$$

Moreover, if $\|\nabla a(\cdot, \mathbf{y})\|_{L^\infty(D)} \leq C_a, \forall \mathbf{y} \in \Gamma$, then

$$\exists C_2 > 0 \text{ s.t. } \|u(\mathbf{y})\|_{H^2(D)} \leq C_2, \quad \forall \mathbf{y} \in \Gamma.$$

That is, we have a *uniform bound in \mathbf{y}* on the H^2 -norm of the solution.

We consider now a piece-wise linear finite element approximation:
 for any $\mathbf{y} \in \Gamma$, find $u_h(\mathbf{y}) \in V_h$ s.t.

$$\int_D a(\cdot, \mathbf{y}) \nabla u_h(\mathbf{y}) \cdot \nabla v_h(\mathbf{y}) = \int_D f v_h, \quad \forall v_h \in V_h$$

where $V_h \subset H_0^1(D)$ is the space of continuous piece-wise linear functions on a triangulation \mathcal{T}_h , vanishing on ∂D .

The discrete solution $u_h(\mathbf{y})$ satisfies the same bound as the continuous one, namely $\|u_h(\mathbf{y})\|_{H_0^1(D)} \leq \frac{C_P \|f\|_{L^2(D)}}{(1-\delta)\bar{\alpha}}$, for all $\mathbf{y} \in \Gamma$

Then

$$\|u(\mathbf{y}) - u_h(\mathbf{y})\|_{L^2(D)} + h \|\nabla u(\mathbf{y}) - \nabla u_h(\mathbf{y})\|_{L^2(D)} \leq C |u(\mathbf{y})|_{H^2(D)} h^2 \quad (56)$$

Since the H^2 bound is uniform in \mathbf{y} , all moments are bounded and (55) holds:

$$\mathbb{E}[\|u(\mathbf{y}) - u_h(\mathbf{y})\|_{L^2(D)}^p]^{\frac{1}{p}} + h \mathbb{E}[\|\nabla u(\mathbf{y}) - \nabla u_h(\mathbf{y})\|_{L^2(D)}^p]^{\frac{1}{p}} \leq Ch^2, \quad \forall p \geq 1$$

Now, given a Lipschitz functional $Q : H_0^1(D) \rightarrow \mathbb{R}$, with $Q(0) = 0$, it holds

$$\mathbb{E}[|Q(u(\mathbf{y}))|^p]^{\frac{1}{p}} \leq \mathbb{E}[C_Q^p \|u(\mathbf{y})\|_{H_0^1(D)}^p]^{\frac{1}{p}} \leq \frac{C_Q C_P \|f\|_{L^2(D)}}{(1 - \delta)\bar{a}}, \quad \forall p \geq 1$$

and the same for the finite element approximation

$$\mathbb{E}[|Q(u_h(\mathbf{y}))|^p]^{\frac{1}{p}} \leq \mathbb{E}[C_Q^p \|u_h(\mathbf{y})\|_{H_0^1(D)}^p]^{\frac{1}{p}} \leq \frac{C_Q C_P \|f\|_{L^2(D)}}{(1 - \delta)\bar{a}}, \quad \forall p \geq 1$$

hence all moments of $Q(u)$ and $Q(u_h)$ are bounded.

For the **discretization error** in the Monte Carlo method we have

$$\mathbb{E}[Q(u(\mathbf{y})) - Q(u_h(\mathbf{y}))] \leq C_Q \mathbb{E}[\|u(\mathbf{y}) - u_h(\mathbf{y})\|_{H_0^1(D)}] \leq C_Q C_u h$$

Remark 7.1 (Aubin–Nitsche's duality trick)

For functionals that are uniformly Lipschitz in $L^2(D)$ sense, it is usually possible to improve the convergence rate when approximating the QoI by duality arguments (Aubin–Nitsche argument)

$$|\mathbb{E}[Q(u(\mathbf{y})) - Q(u_h(\mathbf{y}))]| \leq Ch^{1+\beta}, \quad 0 \leq \beta \leq 1.$$

where $\beta = 1$ if u is bounded uniformly in $H^2(D)$ and D is convex polygonal.

Monte Carlo error analysis. Statistical Error

Let us analyze the error assuming that Q is a globally Lipschitz functional. Observe now that we have the **statistical error**

$$\begin{aligned}\mathcal{E}_h^Q(M) &= \mathbb{E}[Q(u_h(\mathbf{y}))] - \frac{1}{M} \sum_{m=1}^M Q(u_h(\mathbf{y}(\omega_m))) \\ &= \frac{1}{M} \sum_{m=1}^M \left\{ \mathbb{E}[Q(u_h(\mathbf{y}))] - Q(u_h(\mathbf{y}(\omega_m))) \right\}\end{aligned}$$

and $\mathbb{E}[\mathcal{E}_h^Q(M)] = 0$.

Then

$$\text{Var}[\mathcal{E}_h^Q(M)] = \frac{1}{M} \text{Var}[Q(u_h(\mathbf{y}))]$$

and we can estimate

$$\begin{aligned}\text{Var}[Q(u_h(\mathbf{y}))] &\leq \mathbb{E}[Q(u_h(\mathbf{y}))]^2 \\ &\leq C_Q^2 \|u_h\|_{L_\rho^2(\Gamma; V)}^2 \\ &\leq 2C_Q^2 (\|u\|_{L_\rho^2(\Gamma; V)}^2 + \|u - u_h\|_{L_\rho^2(\Gamma; V)}^2) \\ &\leq 2C_Q^2 (\|u\|_{L_\rho^2(\Gamma; V)}^2 + h^{2\alpha} \|C_u\|_{L_\rho^2(\Gamma)}^2) < \infty.\end{aligned}$$

We concluded that for Lipschitz functionals Q we have

$$\text{Var}[\mathcal{E}_h^Q(M)] \leq \frac{C}{M} \rightarrow 0 \quad \text{as } M \rightarrow \infty$$

with constant C uniformly bounded w.r.t. $h < h_0$.

Combining this with the bias estimate, we obtain the mean square error (MSE) estimate

$$\mathbb{E} \left[\left(\mathbb{E}[Q(u)] - \frac{1}{M} \sum_{m=1}^M Q(u_h(\mathbf{y}(\omega_m))) \right)^2 \right] = (\text{bias})^2 + \text{Var} \leq C_1 h^{2\alpha} + \frac{C_2}{M}$$

Remark 7.2

- ▶ Since $\text{Var}[Q(u_h(\mathbf{y}))] \leq C$ we can apply the law of large numbers, law of iterated logarithms and the Central Limit Theorem.
- ▶ In particular, the previous result implies, via the Chebyshev inequality, convergence in probability, that is, for any given $\epsilon > 0$

$$P \left(\left| \mathcal{E}_h^Q(M) \right| > \epsilon \right) \leq \frac{\text{Var}[\mathcal{E}_h^Q(M)]}{\epsilon^2} \leq \frac{C}{M \epsilon^2} \rightarrow 0 \quad \text{as } M \rightarrow \infty$$

The random linear elliptic PDE example revisited

Consider again the model problem

$$\begin{cases} -\operatorname{div}(a(x, \mathbf{y}) \nabla u(x, \mathbf{y})) = f(x), & x \in D, \\ u(x, \mathbf{y}) = 0, & x \in \partial D \end{cases}, \quad \forall \mathbf{y} \in \Gamma := [-\sqrt{3}, \sqrt{3}]^N$$

with $a(x, \mathbf{y}) = \bar{a} + \sum_{i=1}^N \sqrt{\lambda_i} y_i b_i(x)$, $y_i \sim \mathcal{U}(-\sqrt{3}, \sqrt{3})$ i.i.d.,

$\sum_{i=1}^N \sqrt{3\lambda_i} \|b_i\|_\infty \leq \delta \bar{a}$ for some $0 < \delta < 1$, and its approximation $u_h(\mathbf{y})$ by piece-wise linear finite elements.

We have seen that $\|u_h(\mathbf{y})\|_{H^1(D)} \leq C$, $\implies |Q(u_h(\mathbf{y}))| \leq C_Q C$ for any globally Lipschitz functional $Q : H_0^1(D) \rightarrow \mathbb{R}$ with $Q(0) = 0$ and $\forall \mathbf{y} \in \Gamma$.

We have therefore a control on the third moment of $Q(u_h(\mathbf{y}))$ and we can apply the Berry-Esseen estimate⁵ to conclude that the statistical error

$$\mathcal{E}_h^Q(M) = \mathbb{E}[Q(u_h)] - \frac{1}{M} \sum_{m=1}^M Q(u_h(\mathbf{y}(\omega_m)))$$

satisfies $P\left(|\mathcal{E}_h^Q(M)| \leq c_0 \frac{\operatorname{std}[Q(u_h)]}{\sqrt{M}}\right) \geq 2\Phi(c_0) - 1 - 2 \frac{C_{BE} \lambda^3}{(1 + c_0)^3 \sqrt{M}}$.

with $\lambda^3 = \mathbb{E}[|Q(u_h) - \mathbb{E}(Q(u_h))|^3] / \operatorname{std}[Q(u_h)]^3$.

⁵Observe that we can use the Edgeworth Expansion as well.

Monte Carlo complexity analysis: work vs. error

Recall the error splitting:

$$\mathbb{E}[Q(u(\mathbf{y}))] - \frac{1}{M} \sum_{m=1}^M Q(u_h(\mathbf{y}(\omega_m))) = \mathcal{E}^Q(h) + \mathcal{E}_h^Q(M),$$

with

$$|\mathcal{E}^Q(h)| = \underbrace{|E[Q(u(\mathbf{y})) - Q(u_h(\mathbf{y}))]|}_{\text{discretization error}} \leq Ch^\alpha$$

$$|\mathcal{E}_h^Q(M)| = \underbrace{|E[Q(u_h(\mathbf{y}))] - \frac{1}{M} \sum_{m=1}^M Q(u_h(\mathbf{y}(\omega_m)))|}_{\text{statistical error}} \lesssim c_0 \frac{\text{std}[Q(u_h)]}{\sqrt{M}}$$

The last approximation is motivated in probability by the Central Limit Theorem, i.e.

$$P(\sqrt{M}|\mathcal{E}_h^Q(M)| \leq c_0 \text{std}[Q(u_h)]) \rightarrow 2\Phi(c_0) - 1 \quad \text{as } M \rightarrow \infty.$$

Let us assume now that the expected (average) computational work to solve for each $u(\mathbf{y}(\omega_m))$ is $\mathcal{O}(h^{-d\gamma})$, with $\gamma \geq 1$ related to the quality of our solver and d related to the dimension of our problem.

Example 7.1

Consider again the random linear elliptic PDE with piecewise linear continuous FEM approximations. There we have $1 \leq \gamma \leq 3$ and d being the dimension of our domain $D \subset \mathbb{R}^d$.

Summing up, we have therefore the following estimates

$$\text{Total work : } W \propto M h^{-d\gamma}$$

$$\text{Total error : } |\mathcal{E}^Q(h)| + |\mathcal{E}_h^Q(M)| \leq C_1 h^\alpha + C_2 M^{-1/2}$$

we will use them to choose M and h optimally, yielding the optimal work for the MCS approximation.

We want now to choose optimally h and M . Here we minimize the computational work subject to an accuracy constraint, i.e. we solve

$$\begin{cases} \min_{h,M} M h^{-d\gamma} \\ \text{s.t. } C_1 h^\alpha + \frac{C_2}{\sqrt{M}} \leq \text{TOL} \end{cases}$$

The Lagrangian of the above problem is

$$\mathcal{L}(M, h, \lambda) = M h^{-d\gamma} + \lambda(C_1 h^\alpha + \frac{C_2}{\sqrt{M}} - \text{TOL}).$$

Enforcing $\partial_M \mathcal{L}(M, h, \lambda) = \partial_h \mathcal{L}(M, h, \lambda) = 0$ yields $\frac{1}{\sqrt{M}} = \frac{2\alpha C_1}{d\gamma C_2} h^\alpha$ and using the accuracy constraint we have $h^\alpha = \text{TOL} \frac{1}{C_1(1+2\alpha/(d\gamma))}$.

We can interpret the above as a tolerance splitting into statistical and space discretization tolerances, $\text{TOL} = \text{TOL}_S + \text{TOL}_h$, such that

$$\text{TOL}_h = \frac{\text{TOL}}{(1+2\alpha/(d\gamma))} \quad \text{and} \quad \text{TOL}_S = \text{TOL} \left(1 - \frac{1}{(1+2\alpha/(d\gamma))}\right).$$

The resulting complexity (error versus computational work) is then

$$W \propto \text{TOL}^{-(2+d\gamma/\alpha)}$$

Example 7.2

Consider again the random linear elliptic PDE with piecewise linear continuous FEM approximations. There we have $1 \leq \gamma \leq 3$ and $d = 3$ being the dimension of our domain $D \subset \mathbb{R}^d$. Suppose that we have optimal regularity, so $\alpha = 2$.

Then the exponent in the complexity of MC FEM will be dominated by the discretization cost whenever

$$4/3 \leq \gamma,$$

which will likely happen unless one uses an optimal solver like Multigrid. Observe that a naive Gaussian Elimination solver yields $\gamma = 3$ and a MC FEM complexity of

$$\text{TOL}^{-6.5}$$

Example 7.3 (Failure Probability)

Consider again the previous example but now, our goal is to compute a failure probability,

$$P(Q(u(y)) \geq K).$$

Observe that we can still write the above as an expected value but now the functional is not continuous anymore and we cannot apply the previous complexity discussion. Further tricks may be needed to improve the situation. Moreover, if K is large the failure probability will be small and we may need to use variance reduction techniques.

Variance reduction ⁶

Idea: Since the Monte Carlo Error,

$$|E[Y] - \frac{1}{M} \sum_{j=1}^M Y(\omega_j)|$$

is approximately bounded by

$$C_\alpha \sqrt{\frac{Var[Y]}{M}}$$

then we introduce techniques to "reduce" $Var[Y]$ while keeping $E[Y]$ unchanged. For these techniques to be efficient, we need to use particular features of Y ...

⁶See for instance *Monte Carlo methods in financial engineering*, by P. Glasserman

Control Variates

Suppose that we want to compute $E[Y]$. Instead of sampling just Y_j we also sample an auxiliary r.v., X_j **for which we know $E[X]$** .

Then, for a given β , we consider the unbiased estimator

$$V_M = \frac{1}{M} \sum_{j=1}^M Y_j - \beta \frac{1}{M} \sum_{j=1}^M (X_j - E[X])$$

Questions:

- Is there a way to choose optimally β to minimize $\text{Var}[V_M]$?
- Does the strategy reduce the computational effort?

Now we compute

$$\begin{aligned} \text{Var}[V_M](\beta) &= \frac{1}{M} \text{Var}[Y - \beta X] \\ &= \frac{1}{M} \{ \text{Var}[Y] + \beta^2 \text{Var}[X] - 2\beta \text{Cov}[Y, X] \} \end{aligned}$$

and we minimize over β , yielding

$$\beta^* = \frac{\text{Cov}[Y, X]}{\text{Var}[X]}$$

and

$$\text{Var}[V_M](\beta^*) = \frac{1}{M} \text{Var}[Y] \left(1 - \frac{(\text{Cov}[Y, X])^2}{\text{Var}[X] \text{Var}[Y]} \right) < \frac{1}{M} \text{Var}[Y]$$

Observe:

- As long as X is correlated with Y the above procedure reduces the variance ...
- In practice, we can approximate β^* by using sample covariances and variances (at least for large M)

Exercise 8.1

Generalize control variates to the X vector valued (multivariate) case.

Does control variates pay in terms of computational work?

Assume that the work to generate the pair (X_j, Y_j) is $(1 + \delta)$ times the work to generate Y_j , $\delta > 0$.

Method	# samples	Computat. Work
MC	$\left(\frac{C}{\epsilon}\right)^2 Var[Y]$	$\left(\frac{C}{\epsilon}\right)^2 Var[Y]$
MC + Cont.Var.	$\left(\frac{C}{\epsilon}\right)^2 Var[V(\beta^*)]$	$(1 + \delta) \left(\frac{C}{\epsilon}\right)^2 Var[V(\beta^*)]$

We see that the strategy only pays if

$$Var[Y] > (1 + \delta) Var[V(\beta^*)]$$

i.e.

$$1 > (1 + \delta)(1 - (\rho_{XY})^2) \text{ iff } (\rho_{XY})^2 > \delta/(1 + \delta).$$

Obs: As said before, we need to use some knowledge of Y to find a sufficiently correlated X with known $E[X]!!$

Question: Can we use the structure of the SDEs for this purpose?

Antithetic Variates

Let $Y = g(X)$ and s.t. X has a symmetric distribution around its mean.

Assume $E[X] = 0$.

Then X and $-X$ are identically distributed, yielding

$$E[g(X)] = E[g(-X)]$$

and

$$E[Y] = E\left[\frac{g(X) + g(-X)}{2}\right].$$

We then use the unbiased estimator

$$V_M = \frac{1}{M} \sum_{j=1}^M \frac{g(X_j) + g(-X_j)}{2}$$

Again we may ask if it pays to do so in terms of computational work ...

We have

$$\text{Var}[V_M] = \frac{\text{Var}[g(X) + g(-X)]}{4M}.$$

Then, assuming that computing the pair $(g(X), g(-X))$ takes double the work for $g(X)$, we need

$$2\text{Var}[V_1] \leq \text{Var}[Y]$$

i.e.

$$\text{Var}[g(X) + g(-X)] \leq 2\text{Var}[g(X)]$$

or just

$$\text{Cov}[g(X), g(-X)] < 0.$$

Obs: If g is linear then $\text{Var}[g(X) + g(-X)] = 0$ so we expect the method to work for functions that are close to linear

On the other hand, for $g(X) = X^2$,

$$\text{Cov}[g(X), g(-X)] = \text{Var}[g(X)] > 0$$

and it is worse to apply antithetic variates than to use standard Monte Carlo!

Importance Sampling

Idea: Change the probability measure to reduce the variance!!

Let ρ_X, ρ_Z be pdfs s.t. $\frac{\rho_X(x)}{\hat{\rho}_Z(x)} < C$. Then

$$\begin{aligned} E[Y] &= E[g(X)] \\ &= \int_{\mathbb{R}} g(x) \rho_X(x) dx \\ &= \int_{\mathbb{R}} g(x) \underbrace{\frac{\rho_X(x)}{\hat{\rho}_Z(x)}}_{=\hat{g}(x)} \hat{\rho}_Z(x) dx \\ &= E[\hat{g}(Z)] \end{aligned}$$

However, $E[(g(X))^2]$ and $E[(\hat{g}(Z))^2]$ may be different!!

Example:

Let $g > 0$ and choose

$$\hat{\rho} = Cg\rho,$$

with the normalizing constant

$$C = \left(\int_{\mathbb{R}} g\rho \right)^{-1}$$

Then

$$E[(\hat{g}(Z))^2] = \int_{\mathbb{R}} \frac{g^2\rho}{Cg\rho} \rho = \left(\int_{\mathbb{R}} g\rho \right)^2 = E[\hat{g}(Z)]^2$$

and we have a zero variance estimator!!

This approach is not practical because we need to know $C = (\int_{\mathbb{R}} g\rho)^{-1}$ which is equivalent to solve the original problem!

It indicates though that $\hat{\rho}$ has to follow the product $g\rho$ as much as possible. If that happens, \hat{g} will be "flat" and with little variance!!

Exercise 8.2

Think of the case where g is nonzero for $x \in A$ given set. How should you choose $\hat{\rho}$?

Girsanov's Theorem:

Let under \mathbb{P} , X solve for $s > 0$,

$$dX_s = a(X_s)ds + b(X_s)dW_s$$

Consider a change of measure from \mathbb{P} to \mathbb{Q}^α , so under \mathbb{Q}^α the new dynamics for X become

$$dX_s^\alpha = (a(X_s^\alpha) + b(X_s^\alpha)\alpha(X_s^\alpha, s))ds + b(X_s^\alpha)dW_s$$

Then, we have

$$E^{\mathbb{P}}[g(X)] = E^{\mathbb{Q}^\alpha}[g(X^\alpha)] \frac{d\mathbb{P}}{d\mathbb{Q}^\alpha}$$

with

$$\frac{d\mathbb{P}}{d\mathbb{Q}^\alpha} = \exp \left(- \int_0^T \alpha(X_s^\alpha, s) dW_s - \frac{1}{2} \int_0^T |\alpha(X_s^\alpha, s)|^2 ds \right)$$

Log-normal random variables

Definition 8.1

Let $X \sim N(\mu, \sigma^2)$ be Gaussian, then $Y = e^X$ is called *log-normal*. Observe that $Y \geq 0$ and

$$E[Y] = e^{\mu + \frac{\sigma^2}{2}}$$

$$\text{Var}[Y] = (e^{\sigma^2} - 1)e^{2\mu + \sigma^2}$$

$$E[Y^k] = e^{k\mu + \frac{(k\sigma)^2}{2}}$$

Exercise 8.3

Let X_n be iid and consider

$$Z_N(\mu) = \left(\prod_{1 \leq n \leq N} \frac{X_n}{\mu} \right)^{1/\sqrt{N}}.$$

Can you find a constant μ s.t. $Z_N(\mu)$ converges as $N \rightarrow \infty$?

Exercise 8.4

Consider the computation of

$$E[\max(S(T) - K, 0)]$$

with $S(T) = S(0) \exp((r - \sigma^2/2)T + \sigma W(T))$, $W(T) \sim N(0, T)$ and $S(0) = S_0 \ll K$. Propose the use of importance sampling based on an exponential distribution $X \sim K + \text{Exp}(\lambda)$. Motivate an optimal choice for the parameter λ .

Note: The above stock actually follows a stochastic differential equation,
 $dS_t/S_t = rdt + \sigma dW_t$.

Example 8.2 (Shift-Dilation technique)

A simple and effective biasing technique employs translation of the density function to place much of its probability mass in the rare event region. Here the simulation density $\hat{\rho}$ is given by

$$\hat{\rho}(x) = \frac{1}{\hat{\sigma}} \rho \left(\frac{x - \hat{\mu}}{\hat{\sigma}} \right).$$

Ideally, the parameters $(\hat{\mu}, \hat{\sigma})$ should be chosen to minimize the resulting variance, i.e.

$$(\hat{\mu}, \hat{\sigma})^* = \arg \min_{(\hat{\mu}, \hat{\sigma})} \int (g\rho)^2(x) \frac{1}{\hat{\rho}(x)} dx$$

Observe The previous optimization problem may be too difficult and or computationally expensive. A simpler approach is to take

$$\hat{\mu}^* = \arg \max_x (g\rho)^2(x).$$

The optimization problem above may be solved numerically. Can you motivate intuitively this choice? Comment on the numerical effort needed to solve each optimization problem.

For the standard deviation, we need to estimate a reasonable value for $\hat{\sigma}$ that mimics the standard deviation of $|g|\rho$.

Exercise 8.5

Consider again Exercise 8.4. Propose a shift-dilation variance reduction.

Theorem 8.1 (Optimal Importance Sampling)

The optimal density that minimizes the variance (or equivalently the second moment) under the new measure, namely

$$\min_{\|\rho_Z\|_{L^1}=1} E[(\hat{g}(Z))^2] = \int \left(\frac{g^2 \rho_X^2}{\rho_Z} \right) (x) dx$$

satisfies

$$\rho_Z^* \propto |g| \rho_X$$

proof We observe that this problem is strictly convex with a linear constraint, so the minimizer is unique and only first order conditions are needed to characterize the optimum. The Lagrangian is given as

$$L = \int \left(\frac{g^2 \rho_X^2}{\rho_Z} \right) (x) dx + \lambda (\|\rho_Z\|_{L^1} - 1)$$

By taking variations in the Lagrangian with respect to the density ρ_Z and λ and equate the partial derivatives to zero, the result follows.

Corollary 8.1

The minimal variance achieved by importance sampling is given by

$$\begin{aligned} \text{Var}[\hat{g}(Z)] &= \left(\int |g(x)|\rho_X(x)dx \right)^2 \\ &\times \left(1 - \left(\int_{g>0} \rho_Z^*(x)dx - \int_{g<0} \rho_Z^*(x)dx \right)^2 \right) \end{aligned}$$

Proof: Recall that the optimal density is $\rho_Z^*(x) = c|g(x)|\rho_X(x)$ with $c = (\int |g(x)|\rho_X(x)dx)^{-1}$. Hence, we get

$$\hat{g}(x) = \frac{\text{sign}(g(x))}{c}$$

Thus, the minimal variance is

$$\text{Var}[\hat{g}(Z)] = \frac{1}{c^2} \left(1 - \left(\int \text{sign}(g(x))\rho_Z^*(x)dx \right)^2 \right)$$

and hence the proof is concluded.

Observe: The previous result shows us that when g has a constant sign, we can achieve zero variance with optimal importance sampling.

Numerical Example, Variance reduction: Ex 5.13

Look at J. Carlsson's implementation

`uppg5_13.m`

Uses antithetic variates and control variates.

Consider the computation of a call option on an index Z ,

$$\pi_t = e^{-r(T-t)} E[\max(Z(T) - K, 0)], \quad (57)$$

where Z is the average of d stocks,

$$Z(t) \equiv \frac{1}{d} \sum_{i=1}^d S_i(t)$$

and

$$dS_i(t) = rS_i(t)dt + \sigma_i S_i(t)dW_i(t), \quad i = 1, \dots, d$$

with volatilities

$$\sigma_i \equiv 0.2 * (2 + \sin(i)) \quad i = 1, \dots, d.$$

The correlation between Wiener processes is given by

$$E[dW_i(t)dW_{i'}(t)] = \exp(-2|i - i'|/d))dt \quad 1 \leq i, i' \leq d.$$

The goal of this exercise is to experiment with two different variance reduction techniques, namely the antithetic variates and the control variates.

From now on we take $d = 10$, $r = 0.04$ and $T = 0.5$ in the example above.

For the application of control variates to (57) use the geometric average

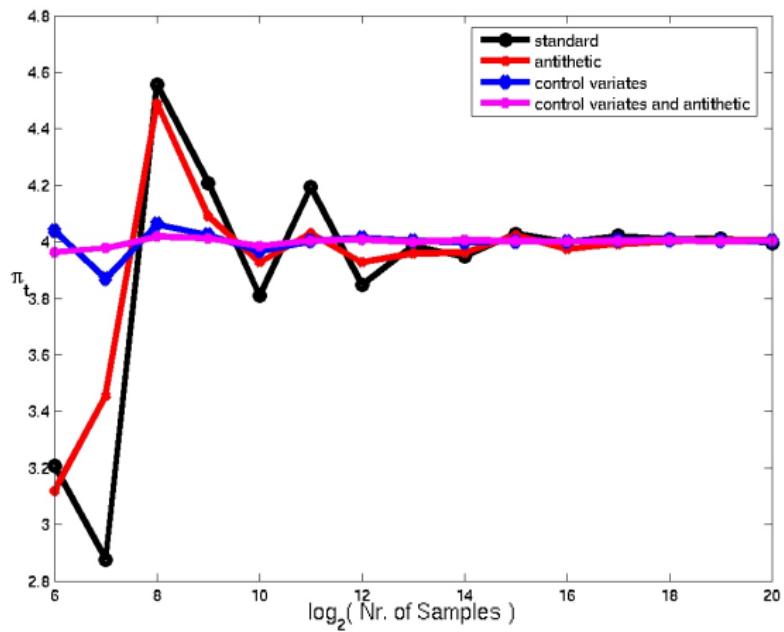
$$\hat{Z}(t) \equiv \left\{ \prod_{i=1}^d S_i(t) \right\}^{\frac{1}{d}},$$

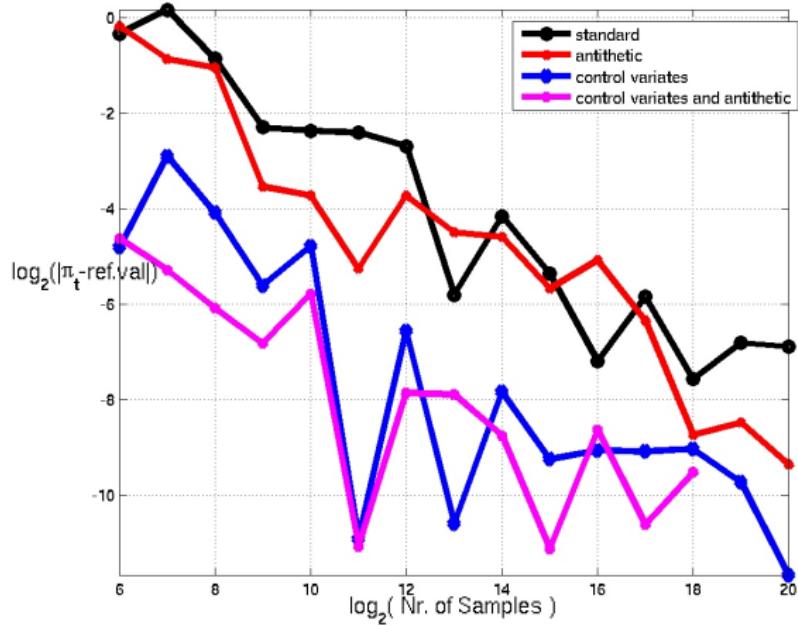
compute

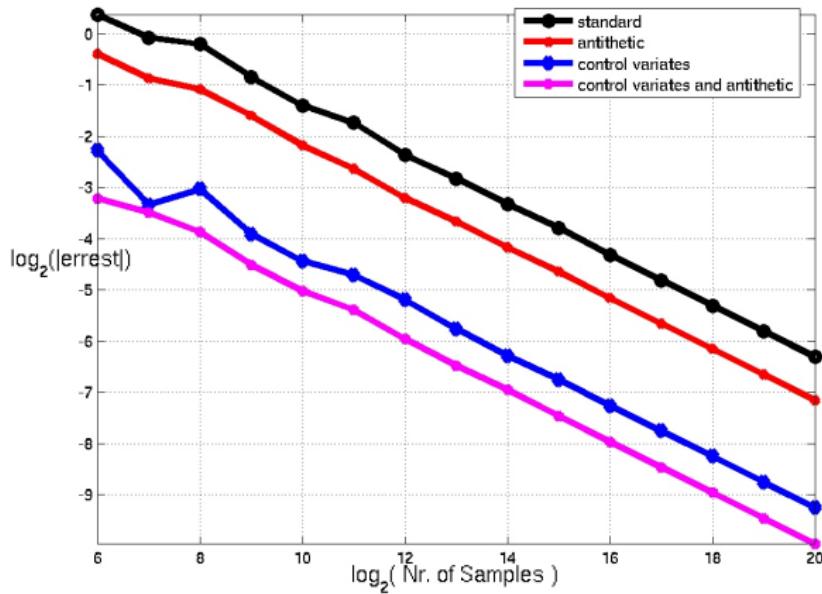
$$\hat{\pi}_t = e^{-r(T-t)} E[\max(\hat{Z}(T) - K, 0)]$$

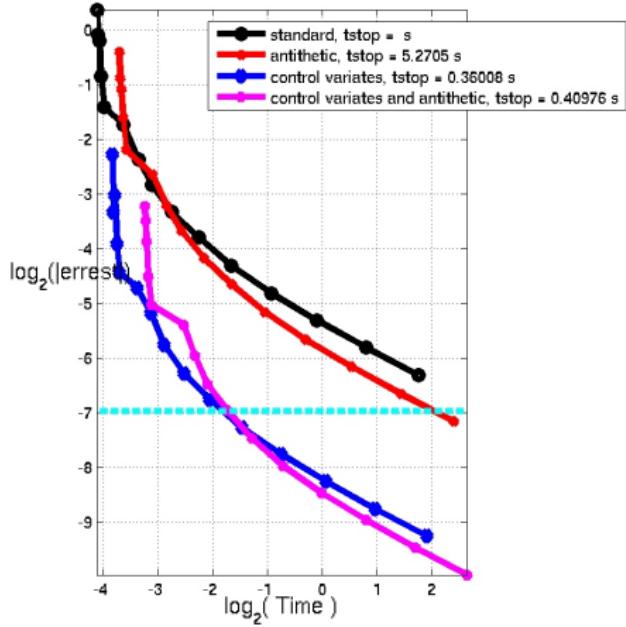
exactly (hint: \hat{Z} has a log-normal distribution, find a way to apply Black-Scholes formula). Then approximate

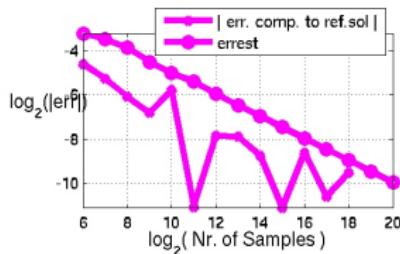
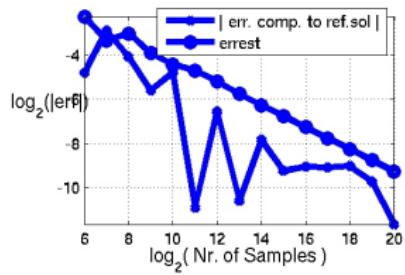
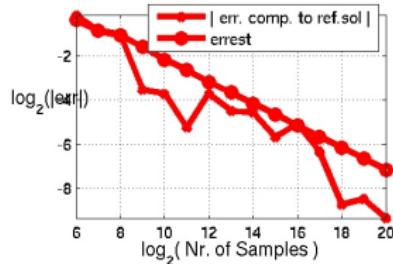
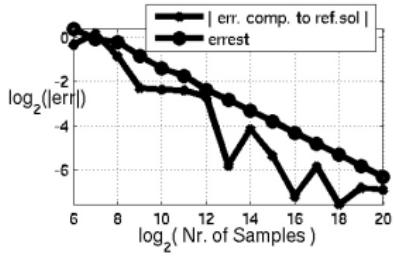
$$\begin{aligned} \pi_t \approx \hat{\pi}_t + \frac{e^{-r(T-t)}}{M} \sum_{j=1}^M & \left\{ \max(Z(W(T, \omega_j)) - K, 0) \right. \\ & \left. - \max(\hat{Z}(W(T, \omega_j)) - K, 0) \right\}. \end{aligned}$$











Can you improve the use of Control Variates for this example?

Propose an importance sampling technique for this example.

Use conditional expectations to reduce variance as much as you can!

Combining Control Variates with Importance Sampling

Let the Residual be $R = Y - X$ and recall the control variate for the approximation of $\mathbb{E}[Y]$,

$$\mathbb{E}[Y] \approx \frac{1}{M} \sum_{m=1}^M R_m + \mathbb{E}[X]$$

with R_m iid with density ρ_R that depends of the distribution of (X, Y) .

Now our problem is to approximate the expectation of the Residual, $E[R]$, in an efficient way.

Furthermore, suppose that $\mathbb{P}(R \neq 0) \ll 1$, then we need to use importance sampling.

Exercise 8.6

Write ρ_R in terms of the distribution of (X, Y) .

Recall that the optimal IS pdf for the computation of $E[R]$ satisfies

$$\rho^*(r) \propto |r|\rho_R(r).$$

Since sampling from ρ^* may be too difficult, the challenge is to find a suitable approximation, $\hat{\rho} \approx \rho^*$ and then use

$$\mathbb{E}[Y] \approx \frac{1}{M} \sum_{m=1}^M \hat{R}_m L(\hat{R}_m) + \mathbb{E}[X]$$

with \hat{R}_m iid with density $\hat{\rho}$ and the likelihood ratio $L(r) = \left(\frac{\rho_R}{\hat{\rho}}\right)(r)$.

Remark 8.1

The optimal IS pdf for R depends on the chosen control variate, X . Even if $Y \geq 0$ a.s. usually R does change sign, so the optimal minimal variance is not zero, even if the original minimal IS variance for Y was zero.

Remark 8.2

We here make the discussion in terms of the sampling densities for the residual, R . In practice it may be simpler to work with the distribution of (X, Y) directly and make the change of measure there.

Exercise 8.7

Discuss the computational cost versus error in the combination of IS and Control Variates.

Exercise 8.8

Provide an example where the use of combined IS and control variate yields substantial computational gains. Hint: Consider (57) for a large value of K.

Exercise 8.9

Discuss the case where the random variables involved are discrete.

Exercise 8.10

Discuss the case where both X, Y are deterministic functions of another random variable, W.

Generalized Control Variates

Suppose that we want to compute $E[Y]$. Instead of sampling just Y_j we also sample an auxiliary r.v., X_j . **We here assume that we do not know $E[X]$ anymore.**

Then, for a given β , we consider the unbiased estimator

$$A(M_1, M_2) = \frac{1}{M_1} \sum_{j=1}^{M_1} (Y_j - \beta X_j) + \frac{\beta}{M_2} \sum_{j'=1}^{M_2} X_{j'}$$

The two ensembles of realizations in the sums above are assumed to be independent, and each of them is composed of iid samples. To simplify the notation, let $Z = Z(\beta) = Y - \beta X$ and $\hat{X} = \beta X$.

Naturally, due to independence, we have that

$$\text{Var}[A(M_1, M_2)] = \frac{\text{Var}[Z]}{M_1} + \frac{\text{Var}[\hat{X}]}{M_2}.$$

Now we compute the optimal work to produce an approximation error of order TOL, recalling that the work is

$$\text{Work}[A(M_1, M_2)] = M_1 W_Z + M_2 W_X.$$

Here $W_Z = (1 + \theta)W_Y$ and $W_X = \theta W_Y$ are the costs to generate a single realization of Z and X , respectively. W_Y is the cost to generate a single realization of Y and $0 < \theta < 1$ is the relative cost of a single realization of X .

Minimizing $Work[A(M_1, M_2)]$ with the constraint
 $Var[A(M_1, M_2)] \leq TOL^2$
yields the optimal work

$$Work^*[A] = \frac{W_Y Var[Y]}{TOL^2} \left(\sqrt{(1 + \theta)\epsilon} + \sqrt{\theta\eta} \right)^2$$

where

$$\epsilon = \frac{Var[Z]}{Var[Y]}$$

and

$$\eta = \frac{\beta^2 Var[X]}{Var[Y]}.$$

Therefore, this generalized control variates approach is more efficient than plain Monte Carlo (based on sampling Y directly) whenever

$$\sqrt{(1 + \theta)\epsilon} + \sqrt{\theta\eta} < 1.$$

Example 8.3 (Two level MC)

Consider the case where $\beta = 1$, $Y = g_h$ and $X = g_{2h}$. Here g_h is assumed to be a numerical approximation based on a mesh size h . Assume that $\text{Var}[Z] = \mathcal{O}(h^p)$ with $p > 0$ and apply the theory above to conclude on the possible benefits of using g_{2h} as a control variate for g_h . In particular, conclude that as $h \rightarrow 0$, the asymptotic reduction in the computational work is θ . Motivate in this case the use of $\beta = 1$.

Now we generalize the analysis of the two level Monte Carlo to a multilevel Monte Carlo setting using a hierarchy of L levels as control variate to level L .

Multilevel Monte Carlo (MLMC)

(Heinrich, 1998) and (Giles, 2008)

For the approximation of $E[S]$, take an integer $\beta > 1$ and for each $\ell = 1, 2, \dots$ use a discretization $S_\ell \approx S$, with $h_\ell = h_0/\beta^\ell$. Recall the standard MLMC difference operator (between consecutive discretizations)

$$\tilde{\Delta}S_\ell = \begin{cases} S_0 & \text{if } \ell = 0, \\ S_\ell - S_{\ell-1} & \text{if } \ell > 0. \end{cases}$$

We use a telescopic identity to represent the desired Quantity of Interest,

$$E[S] \approx E[S_{L,1}] = \sum_{\ell=0}^L E[\tilde{\Delta}S_\ell].$$

Then, using MC to approximate each level independently, the MLMC estimator can be written as

$$\mathcal{A}_{\text{MLMC}} = \sum_{\ell=0}^L \frac{1}{M_\ell} \sum_{m=1}^{M_\ell} \tilde{\Delta}S_\ell(\omega_{\ell,m}).$$

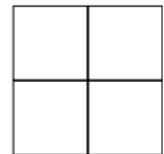
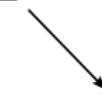
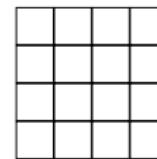
Hierarchical Variance reduction: MLMC

Recall: With Monte Carlo we have to satisfy

$$\text{Var}[A_{MC}] = \frac{1}{M_L} \text{Var}[S_L] \approx \frac{1}{M_L} \text{Var}[S] \leq \text{TOL}^2.$$

Main point: MLMC reduces the variance of the deepest level using samples on coarser (**less expensive**) levels!

$$\begin{aligned}\text{Var}[A_{MLMC}] &= \frac{1}{M_0} \text{Var}[S_0] \\ &+ \sum_{\ell=1}^L \frac{1}{M_\ell} \text{Var}[\tilde{\Delta}S_\ell] \leq \text{TOL}^2.\end{aligned}$$



Observe: Level 0 in MLMC is usually determined by *both* stability and accuracy, i.e.

$$\text{Var}[\tilde{\Delta}S_1] \ll \text{Var}[S_0] \approx \text{Var}[S] < \infty.$$

Classical assumptions for MLMC

For every ℓ , we assume the following:

Assumption 1 (Bias): $|E[S - S_\ell]| \lesssim \beta^{-w\ell},$

Assumption 2 (Variance): $V_\ell = \text{Var}[\tilde{\Delta}S_\ell] \lesssim \beta^{-s\ell},$

Assumption 3 (Work): $W_\ell = \text{Work}(\tilde{\Delta}S_\ell) \lesssim \beta^{d\gamma\ell},$

for positive constants γ, w and $s \leq 2w$.

Example: A smooth linear elliptic PDE example approximated with Multilinear piecewise cont. FEM yields: $2w = s = 4, 1 \leq \gamma \leq 3.$

Work of MLMC: $\text{Work(MLMC)} = \sum_{\ell=0}^L M_\ell W_\ell$

Choose the samples $(M_\ell)_{\ell=0}^L$ optimally so $\text{Var}[A_{\text{MLMC}}] \lesssim \text{TOL}^2$.

Optimal Work of MLMC: $\text{Work(MLMC)} \lesssim \text{TOL}^{-2} \left(\sum_{\ell=0}^L \sqrt{V_\ell W_\ell} \right)^2$

MLMC Computational Complexity

Choose the number of levels L (TOL) to bound the bias

$$|E[S - S_L]| \lesssim \beta^{-L^w} \leq CTOL \quad \Rightarrow \quad L \geq \frac{\log(TOL^{-1}) - \log(C)}{w \log(\beta)},$$

Then the optimal work satisfies (Giles et al., 2008, 2011):

$$\text{Work(MLMC)} = \begin{cases} \mathcal{O}(TOL^{-2}), & s > d\gamma, \\ \mathcal{O}\left(TOL^{-2} (\log(TOL^{-1}))^2\right), & s = d\gamma, \\ \mathcal{O}\left(TOL^{-(2+\frac{(d\gamma-s)}{w})}\right), & s < d\gamma. \end{cases}$$

Recall: $\text{Work(MC)} = \mathcal{O}\left(TOL^{-(2+\frac{d\gamma}{w})}\right)$.

A generalization of MLMC is Multi-index Monte Carlo (MIMC), which yields even better complexity provided additional regularity.

Multi Level Monte Carlo for random PDEs: main idea

Consider again, for simplicity, the setting

- ▶ a **random input** (vector) $\mathbf{y}(\omega) = (y_1(\omega), \dots, y_N(\omega))$ with density $\rho : \Gamma \rightarrow \mathbb{R}_+$, $\Gamma \subset \mathbb{R}^N$
- ▶ a **Hilbert-space valued function** $u : \Gamma \rightarrow V$, $u \in L_p^2(\Gamma; V)$, which is the solution of a differential problem
- ▶ a **Lipschitz functional** $Q : V \rightarrow \mathbb{R}$, with $Q(0) = 0$.

Goal: Compute $\mathbb{E}[Q(u(\mathbf{y}(\omega)))]$.

In practice the solution $u(\mathbf{y}(\omega))$ can not be computed exactly and we will approximate the differential problem by some discretization scheme.

Let h be the **discretization parameter** and let $u_h(\mathbf{y}(\omega))$ be the approximate solution, such that

$$u_h(\mathbf{y}(\omega)) \rightarrow u(\mathbf{y}(\omega)) \quad \text{as } h \rightarrow 0 ,$$

in a suitable sense (to be defined below).

Instead of computing $\mathbb{E}[Q(u(\mathbf{y}(\omega)))]$ we can thus only hope to obtain

$$\mathbb{E}[Q(u_h(\mathbf{y}(\omega)))] \approx \mathbb{E}[Q(u(\mathbf{y}(\omega)))]$$

for h sufficiently small. In view of the generalized control variate principle, we know that the estimator

$$\frac{1}{M_0} \sum_{j=1}^{M_0} Q(u_{2h}(\mathbf{y}(\omega_j^0))) + \frac{1}{M_1} \sum_{j=1}^{M_1} [Q(u_h(\mathbf{y}(\omega_j^1))) - Q(u_{2h}(\mathbf{y}(\omega_j^1)))]$$

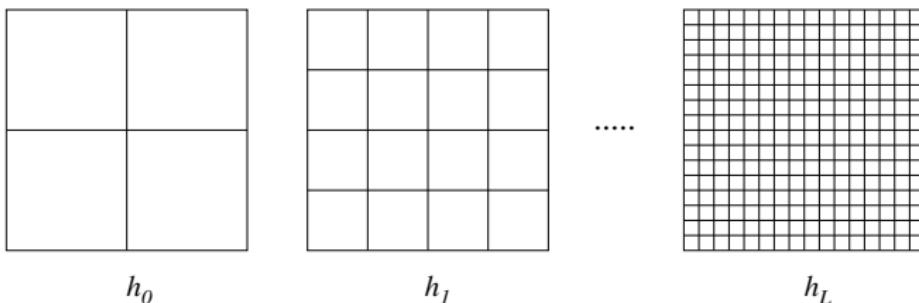
offers an approximation of $\mathbb{E}[Q(u_h(\mathbf{y}(\omega)))]$ with reduced variance. The Multi Level Monte Carlo methods generalizes this idea further.

Objectives of this part:

1. introduce MLMC methods and provide a complexity analysis
2. discuss theoretical properties and extensions

The Multi Level Monte Carlo Method

Introduce a sequence of finer and finer discretizations $h_0 > h_1 > \dots, h_L$ and assume that mesh size h_L achieves the desired target accuracy.



Notation:

- $g(\omega) = Q(u(\mathbf{y}(\omega)))$, (random) output quantity of interest
- $g_\ell(\omega) = Q(u_{h_\ell}(\mathbf{y}(\omega)))$, $\ell = 0, \dots, L$.

Goal: compute $\mathbb{E}[g_L]$

Idea: write the expectation as a telescopic sum

$$\mathbb{E}[g_L] = \mathbb{E}[g_0] + \sum_{\ell=1}^L \mathbb{E}[g_\ell - g_{\ell-1}]$$

and sample **independently** each term on the right hand side with MC.

Definition 9.1 (MLMC estimator)

The *MLMC estimator* of $\mathbb{E}[g]$ is given by

$$\mathcal{A} := \frac{1}{M_0} \sum_{j=1}^{M_0} g_0(\omega_j^0) + \sum_{\ell=1}^L \frac{1}{M_\ell} \sum_{j=1}^{M_\ell} (g_\ell - g_{\ell-1})(\omega_j^\ell),$$

with $\{\omega_j^\ell, j = 1, \dots, M_\ell\}$ independent samples of iid replica on each level.

Observation: $\mathbb{E}[\mathcal{A}] = \mathbb{E}[g_L]$.

Let's asses the quality of the MLMC estimator \mathcal{A} as an approximation of $\mathbb{E}[g]$ in therms of the **Mean Square Error** (MSE):

$$\mathbb{E}[(\mathcal{A} - \mathbb{E}[g])^2] = \underbrace{\mathbb{E}[(\mathcal{A} - \mathbb{E}[g_L])^2]}_{\text{Variance (statistical error)}} + \underbrace{(\mathbb{E}[g_L] - g)^2}_{\text{Bias (discret. error)}}$$

We identify two key error contributions.

Bias of the estimator. This term is directly related to the discretization error h , hence the discretization of the differential equation.

Variance of the estimator. Thanks to the independent samples among levels:

$$\text{Var}[\mathcal{A}] = \mathbb{E}[(\mathcal{A} - \mathbb{E}[g_L])^2] = \frac{\text{Var}[g_0]}{M_0} + \sum_{\ell=1}^L \frac{\text{Var}[g_\ell - g_{\ell-1}]}{M_\ell}$$

Key point: Since $\text{Var}[g_\ell - g_{\ell-1}]$ gets smaller and smaller for ℓ large, one can take M_ℓ smaller and smaller \rightsquigarrow Only few replica on the fine grid h_L .

Optimal choice of M_ℓ

Let

- ▶ C_0 : cost of generating one realization of g_0
- ▶ C_ℓ : cost of generating one realization of $g_\ell - g_{\ell-1}$, $\ell > 0$
- ▶ $V_0 = \text{Var}[g_0]$
- ▶ $V_\ell = \text{Var}[g_\ell - g_{\ell-1}]$, $\ell > 0$

Then

$$\text{Total work: } W_{MLMC} = \sum_{\ell=0}^L M_\ell C_\ell, \quad \text{Total variance: } \text{Var}[\mathcal{A}] = \sum_{\ell=0}^L \frac{V_\ell}{M_\ell}.$$

Problem: Find optimal $\{M_\ell\}$ to minimize the cost at a fixed variance level

$$\min_{\{M_\ell\}} \sum_{\ell=0}^L M_\ell C_\ell \quad \text{subject to} \quad \sum_{\ell=0}^L M_\ell^{-1} V_\ell \leq \text{TOL}^2$$

Solution: If we replace M_ℓ by continuous variables, the opt. solution is

$$M_\ell = \text{TOL}^{-2} \sqrt{\frac{V_\ell}{C_\ell}} \sum_{j=0}^L \sqrt{V_j C_j}$$

Proof.

Define the Lagrangian function

$$\mathcal{L}(M_0, \dots, M_L, \lambda) := \sum_{\ell=0}^L M_\ell C_\ell - \lambda \left(\text{TOL}^2 - \sum_{j=0}^L \frac{V_j}{M_j} \right).$$

Therefore

$$\frac{\partial \mathcal{L}}{\partial M_\ell} = C_\ell - \lambda \frac{V_\ell}{M_\ell^2} = 0 \quad \Rightarrow \quad M_\ell = \sqrt{\lambda \frac{V_\ell}{C_\ell}}.$$

Substituting into the constraint gives

$$\sum_{j=0}^L \sqrt{\frac{V_j C_j}{\lambda}} = \text{TOL}^2, \quad \Rightarrow \quad \sqrt{\lambda} = \text{TOL}^{-2} \sum_{j=0}^L \sqrt{V_j C_j}$$



In practice, one should take the ceiling of the real value M_ℓ (important if $M_\ell < 1$). That is, we have:

- ▶ Optimal sample sizes: $M_\ell = \left\lceil \text{TOL}^{-2} \sqrt{\frac{V_\ell}{C_\ell}} \sum_{j=0}^L \sqrt{V_j C_j} \right\rceil$
- ▶ Optimal work: $W_{MLMC} \leq \text{TOL}^{-2} \left(\sum_{j=0}^L \sqrt{V_j C_j} \right)^2 + \sum_{\ell=0}^L C_\ell$

for the MLMC estimator.

Complexity analysis (error vs. cost)

To analyze the complexity of the MLMC estimator, we make the following assumptions (see also [Giles 2008], [Cliffe et al. 2011]) for now.

Assumption 9.1

For a problem in $D \subset \mathbb{R}^d$ (d -dimensional), assume

1. $h_\ell = h_0 \delta^\ell$, $0 < \delta < 1$ (geometric meshes)
2. $|\mathbb{E}[g - g_\ell]| \leq C_w h_\ell^\alpha$ (weak rate of conv.)
3. $\mathbb{E}[(g - g_\ell)^2] \leq C_s h_\ell^\beta$ (strong rate of conv.)
4. $C_\ell = C_c h_\ell^{-d\gamma}$ ($\gamma = 3$ for direct solver and full matrix; $\gamma \approx 1$ for optimal solver with linear complexity)

Notice that from 3. it follows that

$$5. V_\ell \leq C_v h_{\ell-1}^\beta, \text{ with } C_v = 2C_s(\delta^\beta + 1).$$

Indeed:

$$\begin{aligned} V_\ell &= \text{Var}[g_\ell - g_{\ell-1}] \leq \mathbb{E}[(g_\ell - g_{\ell-1})^2] \\ &\leq 2\mathbb{E}[(g - g_\ell)^2] + 2\mathbb{E}[(g - g_{\ell-1})^2] \leq 2C_s(\delta^\beta + 1)h_{\ell-1}^\beta \end{aligned}$$

Moreover one always has $\beta \leq 2\alpha$ (typically $\beta = 2\alpha$ for PDEs with random coefficients). Indeed by Cauchy-Schwarz inequality

$$\mathbb{E}[g - g_\ell] \leq \mathbb{E}[(g - g_\ell)^2]^{\frac{1}{2}} \leq C_s h_\ell^{\frac{\beta}{2}}, \quad \text{hence } \alpha \geq \frac{\beta}{2}.$$

Theorem 9.2 (MLMC Complexity, [Cliffe et al. 2011])

Under the assumptions 1-4 above, if $2\alpha \geq \min(\beta, d\gamma)$, the computational work required to approximate $\mathbb{E}[g]$ with MLMC with accuracy $0 < \text{TOL} < 1/e$ in mean square sense, that is $\mathbb{E}[(\mathcal{A} - \mathbb{E}[g])^2] \leq \text{TOL}^2$ is bounded as follows:

$$W_{MLMC} \leq C \begin{cases} \text{TOL}^{-2}, & \text{for } \beta > d\gamma, \\ \text{TOL}^{-2} \log^2(\text{TOL}), & \text{for } \beta = d\gamma, \\ \text{TOL}^{-2-(d\gamma-\beta)/\alpha}, & \text{for } \beta < d\gamma, \end{cases}$$

Remark: recall standard MC has corresponding complexity of

$$W_{MC} \propto \text{TOL}^{-2-d\gamma/\alpha}.$$

Proof: We enforce the error constraint $MSE \leq TOL^2$ as

$$\text{Bias constraint: } |\mathbb{E}[g - g_L]|^2 \leq \frac{1}{2} TOL^2, \quad \text{Var. constraint: } \text{Var}[\mathcal{A}] \leq \frac{1}{2} TOL^2$$

From the Bias constraint we get

$$L(TOL) \equiv L = \left\lceil \frac{\log(\sqrt{2}C_w h_0^\alpha TOL^{-1})}{\alpha \log \delta^{-1}} \right\rceil \sim \log_\delta TOL^{\frac{1}{\alpha}}.$$

Setting $\tilde{C}_v = C_v h_0^\beta$ and $\tilde{C}_c = C_c h_0^{-d\gamma}$, the total work is:

$$W_{MLMC} \leq TOL^{-2} \left(\sum_{j=1}^L \sqrt{\tilde{C}_v \tilde{C}_c} \delta^{j \frac{\beta-d\gamma}{2}} \right)^2 + \sum_{j=0}^L C_j$$

Next, we consider three cases:

- ▶ Case $\beta > d\gamma$:

$$W_{MLMC} \leq \text{TOL}^{-2} \left(\sum_{j=1}^{\infty} \sqrt{\tilde{C}_v \tilde{C}_c} \delta^j \right)^2 \leq C \text{TOL}^{-2}$$

- ▶ Case $\beta = d\gamma$: $W_{MLMC} \leq C \text{TOL}^{-2} L \leq C \text{TOL}^{-2} (\log \text{TOL}^{-1})^2$

- ▶ Case $\beta < d\gamma$: $W_{MLMC} \leq C \text{TOL}^{-2} \delta^{L \frac{\beta - d\gamma}{2}} \leq C \text{TOL}^{-2 - \frac{d\gamma - \beta}{\alpha}}$

If, moreover, $2\alpha \geq \min(\beta, d\gamma)$ then the term $\sum_{j=0}^L C_j$ is of order $\mathcal{O}(\text{TOL}^{-2})$ and therefore negligible with respect to the other one. □

Exercise 9.1

Check that $\sum_{j=0}^L C_j = \mathcal{O}(\text{TOL}^{-2})$ for $2\alpha \geq \min(\beta, d\gamma)$ and under the assumptions above.

Important (shocking ?) comments on the MLMC rates

Let us focus on two particular cases:

- ▶ Fastest convergence rate, $\beta > d\gamma$.

Here the convergence rate is TOL^{-2} , which is the same of Monte Carlo sampling when the cost to sample each realization is *fixed*.

This means that we do not see the effect of the discretization in the rates!

- ▶ Smooth noise, $\beta = 2\alpha$ and $\beta < d\gamma$.

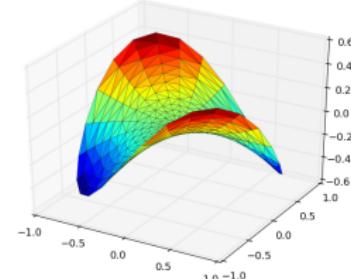
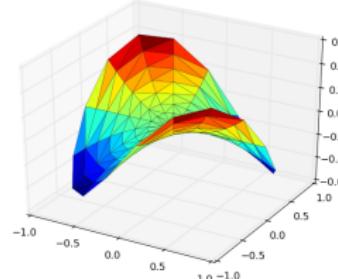
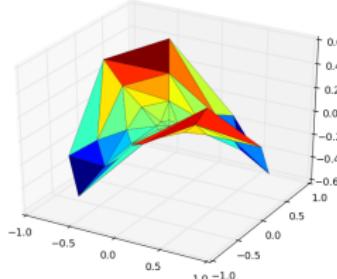
Here the resulting convergence rate is $\text{TOL}^{-d\gamma/\alpha}$, which is the complexity of solving just **one** realization in the deepest level!

An example: random elliptic PDE

Our goal is to compute $E[g]$ where $g = g(u)$. Here g is either a bounded linear functional or a Lipschitz functional with respect to u , and u solves a random PDE. Let us consider a random elliptic PDE:

$$\begin{aligned} -\nabla \cdot (a(\mathbf{x}; \omega) \nabla u(\mathbf{x}; \omega)) &= f(\mathbf{x}; \omega) && \text{for } \mathbf{x} \in D = [0, 1]^d, \\ u(\mathbf{x}; \omega) &= 0 && \text{for } \mathbf{x} \in \partial D, \end{aligned}$$

for sufficiently regular (random) coefficients (i.e., random fields) a and f , such that the PDE is well-posed (a.s.). Suppose we can approximate the solution to the random PDE by a numerical method with prescribed “mesh-size” h for (almost) all $\omega \in \Omega$.



As quantity of interest we consider the linear observable

$$g(u) := \int_D \kappa(\mathbf{x}) u(\mathbf{x}) d\mathbf{x},$$

for a sufficiently regular function κ .

Following the standard MLMC approach, we introduce a hierarchy of $L + 1$ meshes defined by decreasing mesh sizes $\{h_\ell\}_{\ell=0}^L$ and we denote the approximation of g using mesh size h_ℓ by g_ℓ . We then write the MLMC estimator as

$$\mathcal{A} := \frac{1}{M_0} \sum_{j=1}^{M_0} g_0(\omega_j^0) + \sum_{\ell=1}^L \frac{1}{M_\ell} \sum_{j=1}^{M_\ell} g_\ell(\omega_j^\ell) - g_{\ell-1}(\omega_j^\ell),$$

We assume the following error and cost models:

$$|E[g_\ell - g]| \approx C_w h_\ell^\alpha, \quad (59a)$$

$$\text{Var}[g_\ell - g_{\ell-1}] := V_\ell \approx C_v h_{\ell-1}^\beta, \quad (59b)$$

$$\text{Cost per sample of level } \ell := C_\ell \approx C_c h_\ell^{-d\gamma}. \quad (59c)$$

for some positive constants $C_c, C_v, C_w, \alpha, \beta, d$, and γ .

Examples for the rates:

- ▶ **Error models:** In our example for a PDE with smooth random coefficients and for piece-wise linear or piece-wise bilinear continuous finite element approximations we expect $\alpha = 2$ and $\beta = 2\alpha = 4$.
- ▶ **Cost model:** We expect $\gamma = 3$ for a naive Gaussian elimination implementation. For our PDE example, using an *Iterative* solver has $\gamma \approx 1$ while using *Direct* solver has $\gamma \approx 1.5$.

Define:

$$\chi = \frac{\beta}{d\gamma} \quad \text{and} \quad \eta = \frac{\alpha}{d\gamma},$$

then we expect the MLMC complexity of

$$W_{MLMC} \leq C \begin{cases} \text{TOL}^{-2}, & \text{for } \chi > 1, \\ \text{TOL}^{-2} \log^2(\text{TOL}), & \text{for } \chi = 1, \\ \text{TOL}^{-2-(1-\chi)/\eta}, & \text{for } \chi < 1. \end{cases} \quad (60)$$

For our PDE example:

solver	$d = 1$	$d = 2$	$d = 3$
direct	$\chi \approx 2.67, \eta \approx 1.34$	$\chi \approx 1.33, \eta \approx 0.67$	$\chi \approx 0.89, \eta \approx 0.44$
iterative	$\chi \approx 4.00, \eta \approx 2.00$	$\chi \approx 2.00, \eta \approx 1.00$	$\chi \approx 1.33, \eta \approx 0.67$

A CLT for the MLMC method

As the MC estimator, also the MLMC estimator satisfies a central limit theorem. The key ingredient is the classic Lindeberg–Feller Theorem, which is stated below.

Theorem 9.3 (Lindeberg–Feller CLT)

For each n , let $X_{n,m}$, for $1 \leq n \leq m$, be independent random variables (not necessarily identical). Denote

$$a_n = \sum_{m=1}^n X_{n,m}, \quad Y_{n,m} = X_{n,m} - E[X_{n,m}], \quad s_n^2 = \sum_{m=1}^n E[Y_{n,m}^2].$$

Suppose the following Lindeberg condition is satisfied for all $\epsilon > 0$:

$$\lim_{n \rightarrow \infty} s_n^{-2} \sum_{m=1}^n E[Y_{n,m}^2 \mathbf{1}_{|Y_{n,m}| > \epsilon s_n}] = 0. \quad (61)$$

Then,

$$\lim_{n \rightarrow \infty} P\left[\frac{a_n - E[a_n]}{s_n} \leq z\right] = \Phi(z),$$

Lyapunov's condition for the Lindeberg Feller CLT

Lyapunov's condition: There exist $\tau > 0$ such that

$$\lim_{n \rightarrow \infty} \frac{\sum_{m=1}^n E[|Y_{n,m}|^{2+\tau}]}{s_n^{2+\tau}} = 0$$

Exercise 9.2

Show that Lyapunov's condition implies the Lindeberg condition (61).

Hint: Use Markov's inequality to bound

$$E[Y_{n,m}^2 \mathbf{1}_{|Y_{n,m}| > \epsilon s_n}] \leq \frac{E[|Y_{n,m}|^{2+\tau} \mathbf{1}_{|Y_{n,m}| > \epsilon s_n}]}{(\epsilon s_n)^\tau} \leq \frac{E[|Y_{n,m}|^{2+\tau}]}{(\epsilon s_n)^\tau}$$

Exercise 9.3

Let

$$(X_m)_{m \geq 1}$$

be a sequence of independent Bernoulli random variables, each with parameter $0 < p_m < 1$. Then consider the sequence of random variables,

$$Z_n = \sum_{m=1}^n (X_m - p_m).$$

We want to know if

$$\frac{Z_n}{\sqrt{\text{Var } Z_n}}$$

converges to a normal random variable. Show conditions for asymptotic normality, verifying that Lyapunov's condition holds, for this case.

Exercise 9.4

Verify, under the standard CLT assumptions, that is (X_n) iid with zero mean and finite variance $\sigma^2 > 0$, that the Lindeberg-Feller CLT also applies. Hint: verify that (61) holds observing that in this case it reduces to show that, for any fixed $\epsilon > 0$,

$$\lim_{n \rightarrow \infty} E[X_1^2 \mathbf{1}_{\{|X_1| > \epsilon \sigma \sqrt{n}\}}] = 0$$

Now, recalling that $E[X_1^2] = \sigma^2 < \infty$ use dominated convergence to show the limit above.

Lemma 9.4 (CLT for MLMC)

Consider the MLMC estimator \mathcal{A} given by

$$\mathcal{A} = \sum_{\ell=0}^L \sum_{m=1}^{M_\ell} \frac{G_\ell(\omega_{\ell,m})}{M_\ell},$$

where $G_\ell(\omega_{\ell,m})$ denote as usual i.i.d. samples of the random variable $G_\ell = g_\ell - g_{\ell-1}$. The family of random variables, $(G_\ell)_{\ell \geq 0}$, is also assumed independent. Denote $Y_\ell = |G_\ell - \mathbb{E}[G_\ell]|$ and assume the following Lyapunov condition

$$C_1 \delta^{\tilde{\beta}\ell} \leq \mathbb{E}[Y_\ell^2] \quad \text{for all } \ell \geq 0, \quad (62a)$$

$$\mathbb{E}[Y_\ell^{2+\eta}] \leq C_2 \delta^{\tau\ell} \quad \text{for all } \ell \geq 0, \quad (62b)$$

for some $0 < \delta < 1$ and strictly positive constants $C_1, C_2, \tilde{\beta}, \eta$ and τ . Choose the number of samples on each level M_ℓ to satisfy, for $\beta > 0$ and a strictly positive sequence $\{H_\ell\}_{\ell \geq 0}$

$$M_\ell \geq \delta^{\beta\ell} \text{TOL}^{-2} H_\ell^{-1} \left(\sum_{\ell=0}^L H_\ell \right) \quad \text{for all } \ell \geq 0. \quad (63)$$

Lemma 9.4 (cont.)

Moreover, choose the number of levels L to satisfy

$$L \leq \max \left(0, \frac{c \log (\text{TOL})}{\log \delta} + C \right) \quad (64)$$

for some constants C , and $c > 0$. Finally, denoting

$$p = (1 + \eta/2)\tilde{\beta} + (\eta/2)\beta - \tau,$$

if we have that either $p > 0$ or $c < \delta/p$, then

$$\lim_{\text{TOL} \rightarrow 0} P \left[\frac{\mathcal{A} - E[\mathcal{A}]}{\sqrt{\text{Var}[\mathcal{A}]}} \leq z \right] = \Phi(z). \quad (65)$$

Statement as in

- ▶ “*A Continuation Multilevel Monte Carlo*”. N. Collier, A.-L Haji-Ali, F. Nobile, E. von Schwerin and R. Tempone. BIT Numerical Mathematics **55**, 399–432 (2015).

A recent work on relaxing the (technical) uniform integrability condition:

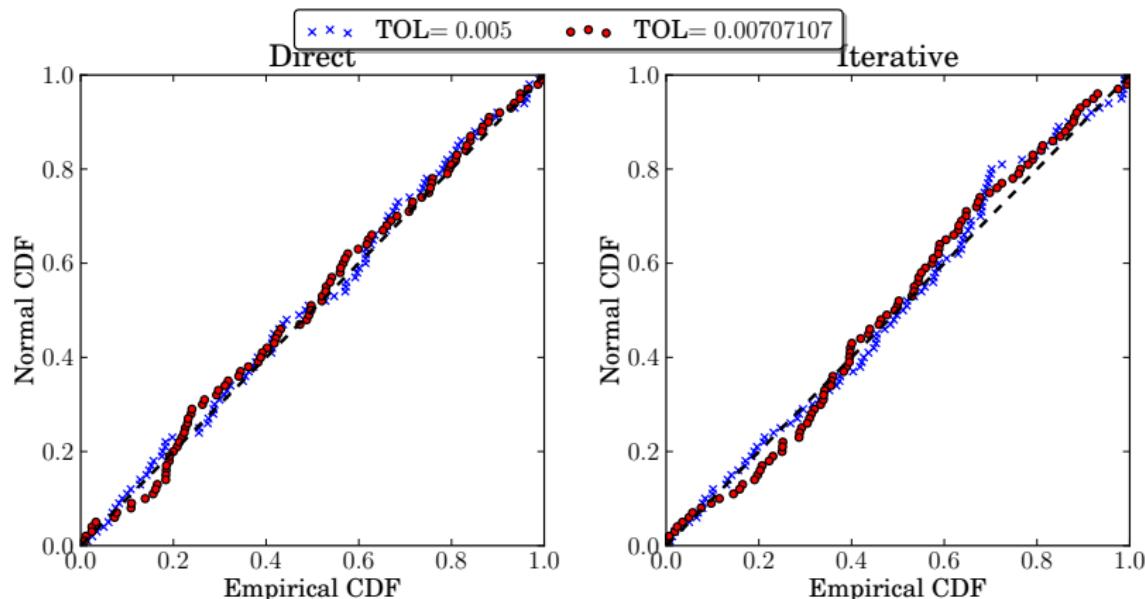
- ▶ “*Central limit theorems for multilevel Monte Carlo methods*”. H. Hoel and S. Krumscheid. J. of Complexity **54**, 101407 (2019).

QQ Plot: Experimental verification of normality

A Q–Q (quantile-quantile) plot is a plot of the quantiles of two distributions against each other, or a plot based on estimates of the quantiles. The pattern of points in the plot is used to compare the two distributions.

For values $0 < u < 1$, we show in the plot the couples $(F^{-1}(u), \Phi^{-1}(u))$. the closer the plot is to the diagonal $x = y$, the closer the normal cdf Φ to the empirical cdf F .

QQ Plot: Experimental verification of normality



Normalized empirical cumulative distribution function (CDF) of MLMC normalized statistical error for different tolerances versus the standard normal CDF. Notice that, the standard normal CDF is a good approximation of the CDF of the MLMC statistical errors.

Error splitting: probability of failure

Using the asymptotic normality of the MLMC estimator, cf. (65), instead of controlling the MSE we can instead aim for

$$P[|\mathcal{A} - E[g]| \leq TOL] \geq 1 - \delta.$$

Along this line, we split the MLMC error, requiring separately

$$\begin{aligned} Bias &= |E[\mathcal{A}] - E[g]| \approx C_w h_L^\alpha \leq (1 - \theta)TOL, \\ P[|\mathcal{A} - E[\mathcal{A}]| \leq \theta TOL] &\geq 1 - \delta. \end{aligned}$$

The last probability inequality is approximated, thanks to the CLT by a simpler, variance constraint

$$Var[\mathcal{A}] \approx \frac{V_0}{M_0} + \sum_{\ell=1}^L \frac{V_\ell}{M_\ell} \leq \left(\frac{\theta TOL}{c_\delta} \right)^2,$$

to guarantee a total error smaller than TOL with probability $\geq 1 - \delta$, for confidence parameter c_δ such that $\Phi(c_\delta) = 1 - \delta/2$.

Here, the splitting between bias and statistical contributions, $\theta \in (0, 1)$ is a free parameter.

Observe:

- ▶ The Bias constraint determines the maximum level L to use
- ▶ Minimization of work under variance constraint leads to optimal sample sizes $\{M_\ell\}_{\ell=0}^L$

To find optimal L and $\{M_\ell\}_\ell$ one needs good estimates of the rates (α, β, γ) , constants (C_w, C_v, C_c) and variances $\{V_\ell\}_{\ell=0}^L$.

Given a hierarchy $\{h_\ell\}_{\ell=0}^L$ and the previous models with certain estimates, it is easy to find some optimal parameters

- ▶ Splitting parameter is given in terms of h_L as $\theta = 1 - \frac{C_w h_L^\alpha}{\text{TOL}}$.
- ▶ Optimal number of samples in \mathbb{R}

$$M_\ell^* = \left(\frac{c_\delta}{\theta \text{TOL}} \right)^2 \sqrt{\frac{V_\ell}{C_\ell}} \left(\sum_{\ell=0}^L \sqrt{V_\ell C_\ell} \right). \quad (66)$$

- ▶ Of course, $M_\ell \in \mathbb{N}$. We can take

$$M_\ell = \lceil M_\ell^* \rceil \leq M_\ell^* + 1.$$

$$\text{Total work} = \sum_{\ell=0}^L M_\ell C_\ell \leq \sum_{\ell=0}^L M_\ell^* C_\ell + \sum_{\ell=0}^L W_\ell.$$

However, under certain conditions (Recall $2\alpha \geq \min(\beta, d\gamma)$) the first term dominates the second one as $\text{TOL} \rightarrow 0$. Hence, we can consider $M_\ell \approx M_\ell^*$ to get the same work estimate, approximately.

- ▶ Optimal $L \in \mathbb{N}$ can be found with brute-force optimization in some limited range.

An adaptive “continuation” algorithm (CMLMC)

Instead of attacking our MLMC problem head on, we can use a sequence of related smaller problems to learn the parameters in our problem.

We proposed the Continuation Multi Level Monte Carlo (CMLMC) algorithm for weak approximation of stochastic models. The CMLMC algorithm solves the given approximation problem for a sequence of decreasing tolerances, ending when the required error tolerance is satisfied. CMLMC assumes discretization hierarchies that are defined a priori for each level and are geometrically refined across levels. The actual choice of computational work across levels is based on parametric models for the average cost per sample and the corresponding weak and strong errors. These parameters are calibrated using Bayesian estimation, taking particular notice of the deepest levels of the discretization hierarchy, where only few realizations are available to produce the estimates. The resulting CMLMC estimator exhibits a non-trivial splitting between bias and statistical contributions. We also show the asymptotic normality of the statistical error in the MLMC estimator and justify in this way our error estimate that allows prescribing both required accuracy and confidence in the final result.

An adaptive “continuation” algorithm (CMLMC)

We use an adaptive continuation MLMC algorithm that, given a hierarchy, solves the given approximation problem for a sequence of decreasing tolerances, ending with the desired one.

Choose a sequence of decreasing tolerances:

$\text{TOL}_0 > \text{TOL}_1 > \dots > \text{TOL}_K = \text{TOL}$ and an initial guess of the rates $(\alpha^{(0)}, \beta^{(0)}, \gamma^{(0)})$, constants $(C_w^{(0)}, C_v^{(0)}, C_c^{(0)})$ and variances $\{V_\ell\}_{\ell=0}^{L^{(0)}}$,

FOR $k = 1, \dots, K$

Based on rates $(\alpha^{(k-1)}, \beta^{(k-1)}, \gamma^{(k-1)})$, constants $(C_w^{(k-1)}, C_v^{(k-1)}, C_c^{(k-1)})$ and variances $\{V_\ell\}_{\ell=0}^{L^{(k-1)}}$

► compute optimal

$$(L^{(k)}, \theta^{(k)}) = \arg \min_{\substack{\theta \in (0,1) \\ L^{(k-1)} \leq L \leq L_{\max}}} \text{Work}(L, \theta), \quad \text{s.t. } C_w h_L^\alpha \leq (1 - \theta) \text{TOL}_k$$

► compute optimal $\{M_\ell^{(k)}\}_{\ell=0}^{L^{(k)}}$ to satisfy TOL_k .

► run MLMC with $L^{(k)}$, $\{M_\ell^{(k)}\}_{\ell=0}^{L^{(k)}}$

► update rates $(\alpha^{(k)}, \beta^{(k)}, \gamma^{(k)})$, constants $(C_w^{(k)}, C_v^{(k)}, C_c^{(k)})$ and variances $\{V_\ell\}_{\ell=0}^{L^{(k)}}$ based on the new simulations performed

► $k = k + 1$

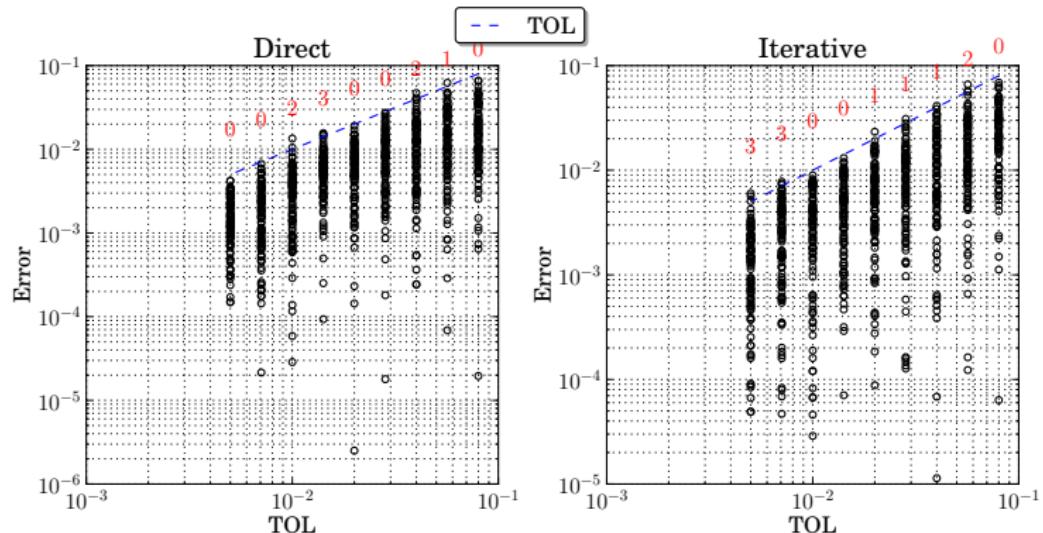
end FOR

A critical issue is how to estimate variances V_ℓ on fine levels where very few simulations are run. One possibility is to use a Bayesian update: prior model $V_\ell \approx C_v h_\ell^\beta$.

Similar adaptive principles have, for example, been used in

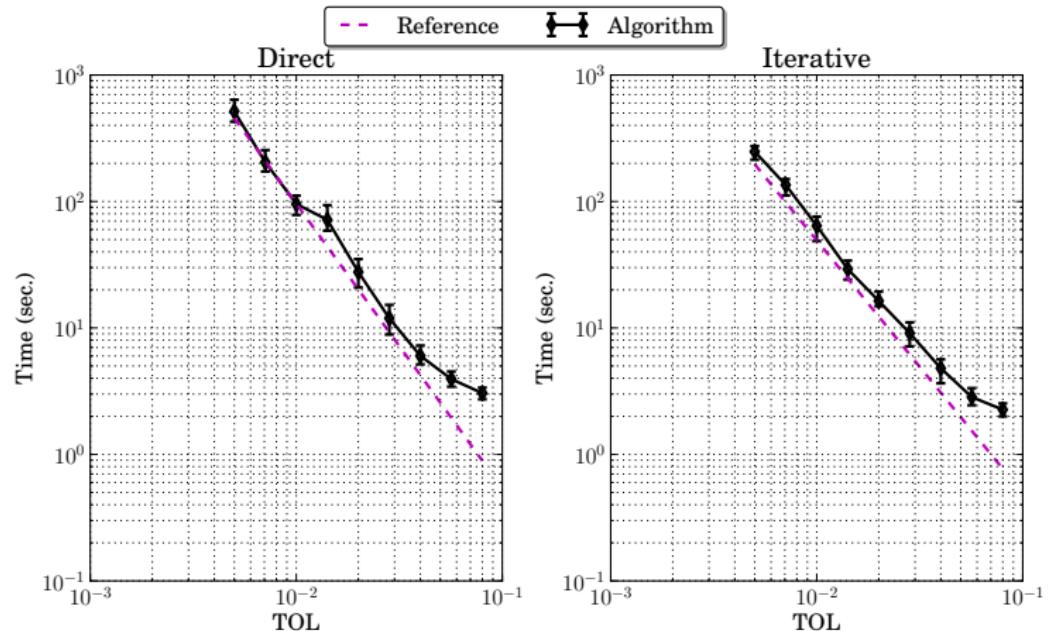
- ▶ Adaptive Multilevel Monte Carlo Simulation, by H. Hoel, E. von Schwerin, A. Szepessy, Anders and R. Tempone, , Numerical Analysis of Multiscale Computations, **82**, Lect. Notes Comput. Sci. Eng., (2012),
- ▶ Implementation and Analysis of an Adaptive Multi Level Monte Carlo Algorithm, by H. Hoel, E. von Schwerin, A. Szepessy, Anders and R. Tempone, Monte Carlo Methods and Applications **20**:1–41(2014).

Error plot of CMLMC

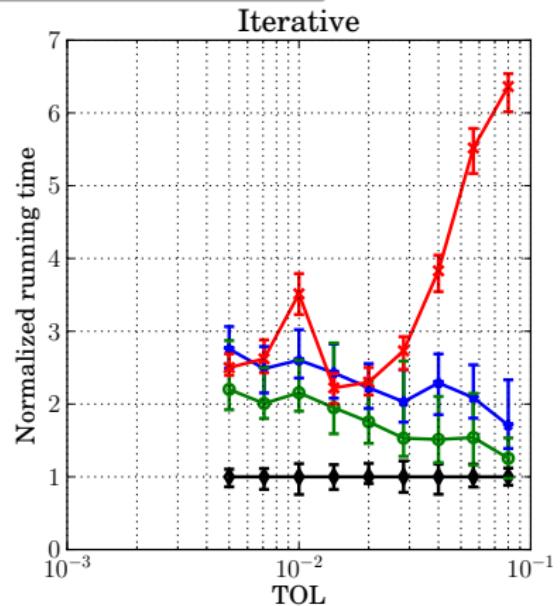
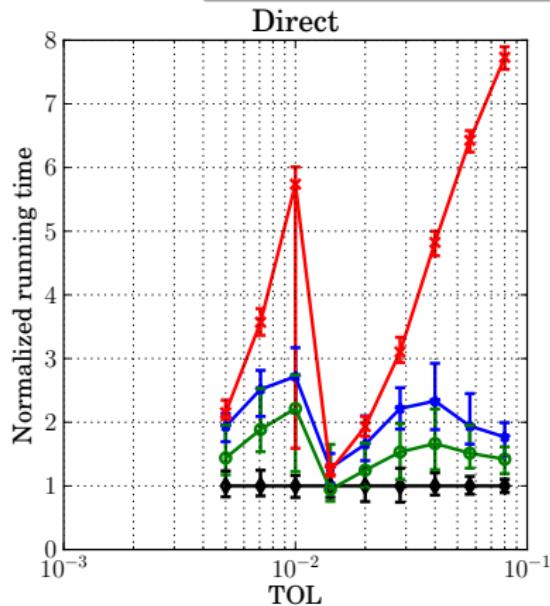


The CMLMC algorithm was run with confidence constant $C_\alpha = 2$ so that the statistical error bound holds with 95% confidence.

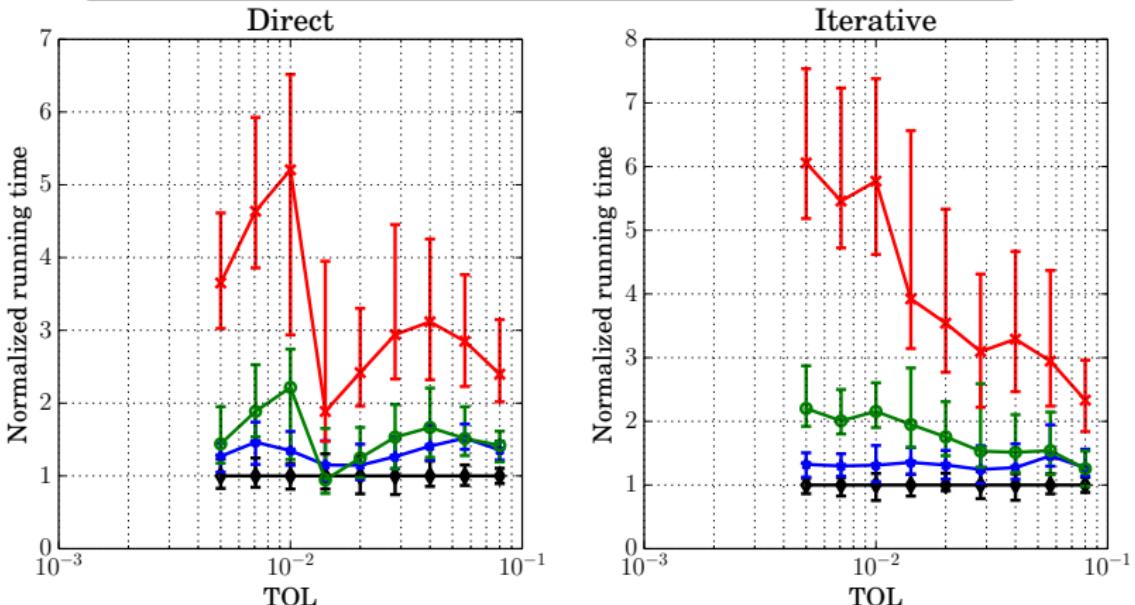
Total work of CMLMC



Reference lines are $TOL^{-2.25}$ and TOL^{-2} , respectively. This is consistent with the MLMC complexity estimate (60).



Improvement in running time due to better choice of splitting parameter,
 θ .



Reusing samples in CMLMC does not significantly improve running, since the work is dominated by the work of the last CMLMC iteration.

MLMC: Beyond geometric hierarchies

The previous result is based on geometric discretizations, i.e. $h_\ell = h_0 \delta^\ell$.

Questions:

- ▶ Can we construct non-geometric discretization MLMC hierarchies?
- ▶ Do they reduce the computational work of MLMC even further?

This leads to

Problem 9.5 (Optimization of computational work)

Given $L \in \mathbb{N}$ and $\theta \in (0, 1)$, find $\mathbf{H} = (\{h_\ell\}_{\ell=0}^L, \{M_\ell\}_{\ell=0}^L) \in \mathbb{R}_+^{L+1} \times \mathbb{R}_+^{L+1}$ such that

$$\text{Work}(\mathbf{H}) = \sum_{\ell=0}^L \frac{M_\ell}{h_\ell^{d\gamma}},$$

is minimized while satisfying the constraints

$$C_w h_L^\alpha \leq (1 - \theta) \text{TOL}, \quad \frac{V_0}{M_0} + C_v \sum_{\ell=1}^L \frac{h_{\ell-1}^\beta}{M_\ell} \leq \left(\frac{\theta \text{TOL}}{c_\delta} \right)^2,$$

for a confidence parameter, c_δ , such that $\Phi(c_\delta) = 1 - \frac{\delta}{2}$; here, $0 < \delta \ll 1$ and Φ is standard normal CDF.

This problem has been studied in

- ▶ “*Optimization of mesh hierarchies for Multilevel Monte Carlo samplers*”, by A.-L Haji-Ali, F. Nobile, E. von Schwerin and R. Tempone. *Stochastics and Partial Differential Equations Analysis and Computations* **4**:76–112 (2016).

One can show that geometric hierarchies are near-optimal and that the splitting parameter plays a non trivial role.

Moreover, it is possible to derive the computational complexity with known rates and constants depending on $\chi = 1$ or $\chi \neq 1$, at least asymptotically. Knowing the multiplicative constants in the complexity allows us to decide on each case if further tricks may be effective.

MLMC, further extensions

MLMC methods constitute an active research areas both on theoretical/methodological advances as well as focusing on applications and high-performance computing. As a matter of fact, we will revisit the MLMC idea multiple time in the context of this course.

However, there are various extensions that go beyond the scope of course. If you are keen to further deepening your knowledge on these techniques, attend the MLMC seminar or chat with us concerning thesis or semester projects.

Recording of related lecture on adaptive strategies for MLMC:
<https://www.youtube.com/watch?v=CZSESGi34kI>

Large Deviations

Theory for estimating probabilities of rare events, deep in the distribution tails.

Example 10.1 (Insurance, H. Cramér)

An insurance company has to face random claims with payoff

$$S_N = \sum_{n=1}^N X_n.$$

Here X_n are modeled as independent random variables. What should be the corresponding insurance premium qN so that

$$P(S_N > qN) < \epsilon,$$

for some *given* $\epsilon \ll 1$?

Observe that

$$P(S_N > qN) = P\left(\frac{1}{N}S_N - \mu > q - \mu\right)$$

so we are interested in estimating the probability of $\mathcal{O}(1)$ deviations of the sample average from the mean.

A note on small deviations.

Remember CLT and BET: Assume ξ_j , $j = 1, 2, 3, \dots$ are independent, identically distributed (i.i.d) and $E[\xi_j] = 0$, $E[\xi_j^2] = 1$. Then

$$\sum_{j=1}^J \frac{\xi_j}{\sqrt{J}} \rightarrow \nu \sim N(0, 1), \quad (67)$$

This result estimates the probability of (small) deviations from the mean of size $\mathcal{O}(1/\sqrt{J})$, namely

$$P\left(\left|\sum_{j=1}^J \frac{\xi_j}{J}\right| \geq \frac{C}{\sqrt{J}}\right) = P\left(\underbrace{\left|\sum_{j=1}^J \frac{\xi_j}{\sqrt{J}}\right|}_{\approx \nu} \geq C\right) \approx 2(1 - \Phi(C))$$

However, the approximate value itself decays exponentially fast as $C \rightarrow \infty$.

Will the approximation error dominate the probability to estimate if we are interested in much larger, say $C/\sqrt{J} = \mathcal{O}(1)$, deviations from the mean as in Example 10.1?

They will be very rare, can we estimate their probability?

Example: Sums of iid Gaussian random variables

-Let $\xi_i \sim N(0, 1)$ iid. Then

$$\begin{aligned} P\left(\left|\sum_{j=1}^J \frac{\xi_j}{J}\right| \geq \text{TOL}\right) &= P(|\xi_1| \geq \text{TOL}\sqrt{J}) \\ &= 2(1 - \Phi(\text{TOL}\sqrt{J})) \quad \text{Note equality} \\ &\approx 2\text{TOL}\sqrt{J} \Phi'(\text{TOL}\sqrt{J}) \\ &= \text{TOL}\sqrt{\frac{2J}{\pi}} \exp\left(-\frac{\text{TOL}^2 J}{2}\right) \end{aligned}$$

Therefore, we conclude that

$$\lim_{J \rightarrow \infty} \frac{\log\left(P\left(\left|\sum_{j=1}^J \frac{\xi_j}{J}\right| \geq \text{TOL}\right)\right)}{J} = -\frac{\text{TOL}^2}{2}$$

and we see that the probability of $\mathcal{O}(1)$ deviations from the mean decreases **exponentially** with the number of samples J .

Observe: The actual constant in such exponential convergence depends on the ξ distribution.

Idea:

To get tight bounds, use high order (exponential) moments as in the following Lemma.

Lemma 10.2

Let X be a random variable and assume there is a function $f > 0$ which is monotone increasing such that $E[f(X)] < \infty$. Then we have

$$P(X \geq a) \leq \frac{E[f(X)]}{f(a)}$$

Proof.

$$\begin{aligned} P(X \geq a) &\leq \int_a^{+\infty} dP(x) = \int_a^{+\infty} \frac{f(x)}{f(x)} dP(x) \\ &\leq \frac{1}{f(a)} \int_a^{+\infty} f(x) dP(x) \\ &\leq \frac{E[f(X)]}{f(a)}. \end{aligned}$$

From the previous lemma we can conclude for instance

$$P(X \geq a) \leq \min \left\{ \inf_{p>1} \frac{E[X^p]}{a^p}, \inf_{\theta>0} \frac{E[\exp(\theta X)]}{\exp(\theta a)} \right\}$$

In LDT, we use bounded exponential moments as follows.

Lemma 10.3 (Chernoff bound)

Assume that ξ_j , $j = 1, 2, 3, \dots$ are independent, identically distributed (i.i.d) and that $E[e^{\theta\xi_1}] < \infty$ for some $\theta > 0$.

Then for all $x \in \mathbb{R}$ we have

$$\begin{aligned} P\left(\frac{1}{J} \sum_{j=1}^J \xi_j > x\right) &\leq \frac{E[e^{\theta\xi_1}]^J}{e^{J\theta x}} \\ &\leq e^{-J(\theta x - \log(E[e^{\theta\xi_1}]))}. \end{aligned}$$

Proof.

We just rewrite the desired probability as an expected value,

$$\begin{aligned} P\left(\sum_{j=1}^M \frac{\xi_j}{M} > x\right) &= P\left(\sum_{j=1}^M \xi_j > Mx\right) \\ &= E[\mathbf{1}_{\{\sum_{j=1}^M \xi_j > Mx\}}] \\ &\leq E[e^{-\theta Mx} e^{\theta \sum_{j=1}^M \xi_j} \mathbf{1}_{\{\sum_{j=1}^M \xi_j > Mx\}}] \\ &\leq e^{-\theta Mx} E[e^{\theta \sum_{j=1}^M \xi_j} \mathbf{1}_{\{\sum_{j=1}^M \xi_j > Mx\}}] \\ &\leq e^{-\theta Mx} E[e^{\theta \sum_{j=1}^M \xi_j}] \\ &\leq e^{-\theta Mx} E[e^{\theta \xi_1}]^M. \end{aligned}$$

Observe:

The previous result gives a faster decay than a Markov type inequality,

$$P\left(\sum_{j=1}^M \frac{\xi_j - E[\xi]}{M} > x\right) \leq \frac{\text{Var}[\xi]}{Mx^2},$$

and it implies that the tail probability of the sample average converges exponentially wrt M .

Observe: The Chernoff bound is not an asymptotic result. It is named after Herman Chernoff but due to Herman Rubin.

Exercise 10.1

Generalize the Chernoff bound to independent, but not identically distributed random variables with bounded exponential moments.

The assumption on bounded exponential moments may be replaced by a strict bound on the random variables averaged.

Theorem 10.1 (Hoeffding's inequality)

If ξ_i are independent random variables satisfying $a_i < \xi_i < b_i$. Then, for $\mu = \frac{1}{M} \sum_{j=1}^M E[\xi_j]$ and $t > 0$ we have

$$P\left(\sum_{j=1}^M \frac{\xi_j}{M} - \mu > t\right) \leq \exp\left(-\frac{2M t^2}{\frac{1}{M} \sum_{j=1}^M (a_j - b_j)^2}\right)$$

Reference: "Probability Inequalities for Sums of Bounded Random Variables", by Wassily Hoeffding. Journal of the American Statistical Association, Vol. 58, No. 301 (Mar., 1963), pp. 13- 30.

Remark 10.1

We can further improve the bound by taking infimum of the rhs over all possible positive values of θ s.t. $E[e^{\theta\xi}] < \infty$.

This yields the Chernoff bound

$$P\left(\sum_{j=1}^M \frac{\xi_j}{M} > x\right) \leq e^{-MI(x)}, \quad (68)$$

where the rate function, $I(x)$, is given by the Fenchel-Legendre transform of the log moment generating function, $\Lambda(\theta) \equiv \log(E[e^{\theta\xi_1}])$, i.e.

$$I(x) = \sup_{\theta \in \mathbb{R}} \{\theta x - \Lambda(\theta)\}.$$

Exercise 10.2

Let $Z = X + Y$, with X and Y independent random variables. Show that

$$I_Z(z) \leq \inf_{\{(x,y):x+y=z\}} \{I_X(x) + I_Y(y)\}.$$

For asymptotic results, we have the following Theorem:

Theorem 10.2 (Cramér-Chernoff)

Assume that ξ_j , $j = 1, \dots, J$ are independent, identically distributed (i.i.d) rvs and $E[\xi_j] = 0$.

Assume that the moment generating function,

$$M(\theta) := E[e^{\theta\xi_1}] < \infty$$

in some neighborhood of $\theta = 0$.

Let $TOL > 0$ s.t. $P(\xi_1 > TOL) > 0$.

Then $I(TOL) > 0$ and

$$\lim_{J \rightarrow \infty} \frac{1}{J} \log \left(P \left(\sum_{j=1}^J \frac{\xi_j}{J} > TOL \right) \right) = -I(TOL). \quad (69)$$

Remark 10.2

The theorem works also for deviations of $\frac{1}{J} \sum_{j=1}^J \xi_j$ below its mean. Assuming $P(\xi_1 < -\text{TOL}) > 0$, apply (69) to the rvs $-\xi_j$, $j = 1, 2, 3, \dots$, then

$$\lim_J \frac{1}{J} \log \left(P \left(\sum_{j=1}^J \frac{\xi_j}{J} < -\text{TOL} \right) \right) = -I(-\text{TOL}).$$

Proof of Thm. 10.2.

STEP 1: Check that

$$0 < I(\text{TOL}) = \sup_{\theta \in \mathbb{R}} \{\theta \text{TOL} - \Lambda(\theta)\}.$$

(*Why is it important?*)

Observe that, by definition we have

$$\Lambda(\theta) = \log(M(\theta)) = \log(E[e^{\theta \xi_1}])$$

and thus, using a Taylor expansion around $\theta = 0$,

$$\begin{aligned}\text{TOL}\theta - \Lambda(\theta) &= \log\left(\frac{e^{\text{TOL}\theta}}{M(\theta)}\right) \\ &= \log\left(\frac{1 + \text{TOL}\theta + o(\theta)}{M(0) + M'(0)\theta + \frac{M''(0)}{2}\theta^2 + o(\theta^2)}\right).\end{aligned}$$

Moreover, since $M(\theta) = E[e^{\theta \xi_1}]$ satisfies
 $M(0) = 1$, $M'(0) = 0$, $M''(0) = \text{Var}[\xi_1]$
we have

$$\begin{aligned}\text{TOL}\theta - \Lambda(\theta) &= \log\left(\frac{e^{\text{TOL}\theta}}{M(\theta)}\right) \\ &= \log\left(\frac{1 + \text{TOL}\theta + o(\theta)}{1 + \frac{\text{Var}[\xi_1]}{2}\theta^2 + o(\theta^2)}\right).\end{aligned}$$

Finally, observe that for θ sufficiently small,

$$1 + \text{TOL}\theta + o(\theta) > 1 + \frac{\text{Var}[\xi_1]\theta^2}{2} + o(\theta^2).$$

Conclude that $0 < I(\text{TOL}) := \sup_{\theta \in \mathbb{R}} \{\text{TOL}\theta - \Lambda(\theta)\}$.

STEP 2:

To prove (69) we will first show that there exist a lower bound,
 $L(TOL, J)$, s.t.

$$L(TOL, J) \leq \frac{1}{J} \log \left(P \left(\sum_{j=1}^J \frac{\xi_j}{J} > TOL \right) \right) \leq -I(TOL).$$

Then we will show that $L(TOL, J) \rightarrow -I(TOL)$ when $J \rightarrow \infty$.

Upper bound: apply the Chernoff bound (68).

Lower bound with $L(TOL, J)$:

Assume that the supremum is attained, namely that there exists an optimizer $\eta = \eta(TOL) \in (-\Theta, \Theta)$, with $\Theta > 0$, such that

$$I(TOL) = TOL\eta - \Lambda(\eta) = \sup_{\theta \in (-\Theta, \Theta)} \{\theta TOL - \Lambda(\theta)\}.$$

Also, for the same reason we have that

$$\Lambda'(\eta) = TOL \text{ and that } \Lambda''(\eta) > 0.$$

Now, given $\xi_1 \sim F$, consider the *exponential change of (marginal) distribution*

$$d\tilde{F}(u) = \frac{e^{\eta u}}{M(\eta)} dF(u). \quad (70)$$

Denote by

$$\tilde{\xi}_j, \quad j = 1, \dots, J$$

the *independent* rvs with marginal distribution \tilde{F} .

Now, due to (70), we can write the new moment generating function as

$$\tilde{M}(\theta) := M_{\tilde{\xi}_1}(\theta) = \int_{-\infty}^{\infty} e^{\theta u} d\tilde{F}(u) = \frac{M(\theta + \eta)}{M(\eta)}.$$

Let

$$\tilde{S}_J = \tilde{\xi}_1 + \cdots + \tilde{\xi}_J.$$

Due to the convolution formula for the distribution of a sum of iid random variables we have, for \tilde{S}_J ,

$$d\tilde{F}_J(s) = \frac{e^{\eta s}}{M(\eta)^J} dF_J(s) \quad (71)$$

where F_J is the distribution function of the sum

$$S_J = \xi_1 + \cdots + \xi_J.$$

Exercise: Verify (71).

Observe that due to the iid nature of $\{\xi_j\}$ and $\{\tilde{\xi}_j\}$, we have

$$M_{\tilde{S}_J}(\theta) = \left(\frac{M(\theta + \eta)}{M(\eta)} \right)^J = \frac{1}{(M(\eta))^J} \int_{-\infty}^{\infty} e^{(\theta + \eta)u} dF_J(u).$$

Also, observe that (see Exercise 10.3),

$$E[\tilde{\xi}_j] = \Lambda'(\eta) = \text{TOL}$$

and

$$\text{Var}[\tilde{\xi}_j] = \Lambda''(\eta) > 0.$$

Let us now construct the lower bound for

$$P\left(\sum_{j=1}^J \frac{\xi_j}{J} > \text{TOL}\right) = P(S_J > J\text{TOL}).$$

Let $\text{TOL}_2 > \text{TOL}$, so

$$\begin{aligned} P(S_J > J\text{TOL}) &= \int_{J\text{TOL}}^{\infty} dF_J(u) \\ &= \int_{J\text{TOL}}^{\infty} (M(\eta))^J e^{-\eta u} d\tilde{F}_J(u) \\ &\geq (M(\eta))^J \int_{J\text{TOL}}^{J\text{TOL}_2} e^{-\eta J\text{TOL}_2} d\tilde{F}_J(u) \\ &\geq e^{-J(\eta J\text{TOL}_2 - \Lambda(\eta))} P(J\text{TOL} < \tilde{S}_J < J\text{TOL}_2). \end{aligned} \tag{72}$$

Finally, rewrite (72) as

$$\begin{aligned} \frac{1}{J} \log P(S_J > JTOL) \\ \geq L(TOL, J) = -(\eta TOL_2 - \Lambda(\eta)) \\ + \frac{1}{J} \log P(JTOL < \tilde{S}_J < JTOL_2). \end{aligned}$$

and choose $TOL_2(J) \rightarrow TOL$ s.t.

$$\frac{1}{J} \log P(JTOL < \tilde{S}_J < JTOL_2) \rightarrow 0$$

so $L(TOL, J) \rightarrow -I(TOL)$ as $J \rightarrow \infty$.

About TOL_2 :

We can take $\text{TOL}_2 = \text{TOL} + f(J)$ with $f(J) \rightarrow 0$ sufficiently slow.
Indeed, we need to guarantee that

$$\frac{1}{J} \log P(0 < \tilde{S}_J - JTOL < Jf(J)) \rightarrow 0$$

To conclude, observe that if $\sqrt{J}f(J) \rightarrow \infty$ we have by the CLT (recall that $E[\tilde{S}_J] = JTOL$) that

$$P\left(0 < \frac{\tilde{S}_J - JTOL}{\sqrt{J}} < \sqrt{J}f(J)\right) \rightarrow \underbrace{P(N(0, \text{Var}[\tilde{\xi}_1]) \in (0, +\infty))}_{=1/2}$$

Exercise 10.3

Show that due to the measure change,

$$E[\tilde{\xi}_j] = \Lambda'(\eta) = \text{TOL}$$

and

$$\text{Var}[\tilde{\xi}_j] = \Lambda''(\eta) > 0.$$

Hint: Recall that

$$d\tilde{F}(u) = \frac{e^{\eta u}}{M(\eta)} dF(u)$$

and thus

$$\tilde{M}(\theta) = \frac{M(\theta + \eta)}{M(\eta)}.$$

To conclude, use the fact that

$$E[\tilde{\xi}_j] = \tilde{M}'(0), \text{ and } E[\tilde{\xi}_j^2] = \tilde{M}''(0).$$

More general case:

See the book by Grimmett and Stirzaker⁷, p.205, for the proof of the lower bound when $\sup_{\theta \in \mathbb{R}} \{TOL\theta - \Lambda(\theta)\}$ is not achieved strictly within the domain of convergence of $M(\theta)$.

⁷Geoffrey R. Grimmett and David R. Stirzaker (2001). *Probability and Random Processes*, 3rd edition, Oxford University Press.

Computational examples with different distributions

Example: Uniform random variables

Let ξ_i be iid $U(-\sqrt{3}, \sqrt{3})$. Then

$$E[e^{\theta\xi}] = \frac{\sinh(\theta\sqrt{3})}{\theta\sqrt{3}},$$

$$I(x) = \sup_{\theta} \{\theta x - \log(\sinh(\theta\sqrt{3})) + \log(\theta\sqrt{3})\}.$$

Since $\log(\sinh(\theta\sqrt{3})) \sim \theta\sqrt{3}$ for large θ we obtain

$$I(x) = +\infty \text{ for } x > \sqrt{3}$$

This is in complete agreement with the obvious fact

$$P\left(\sum_{j=1}^J \xi_j > J\sqrt{3}\right) = 0.$$

A naive application of the CLT gives instead

$$P\left(\sum_{j=1}^J \xi_i > J\sqrt{3}\right) \approx \frac{\Phi'(\sqrt{3J})}{\sqrt{3J}}$$

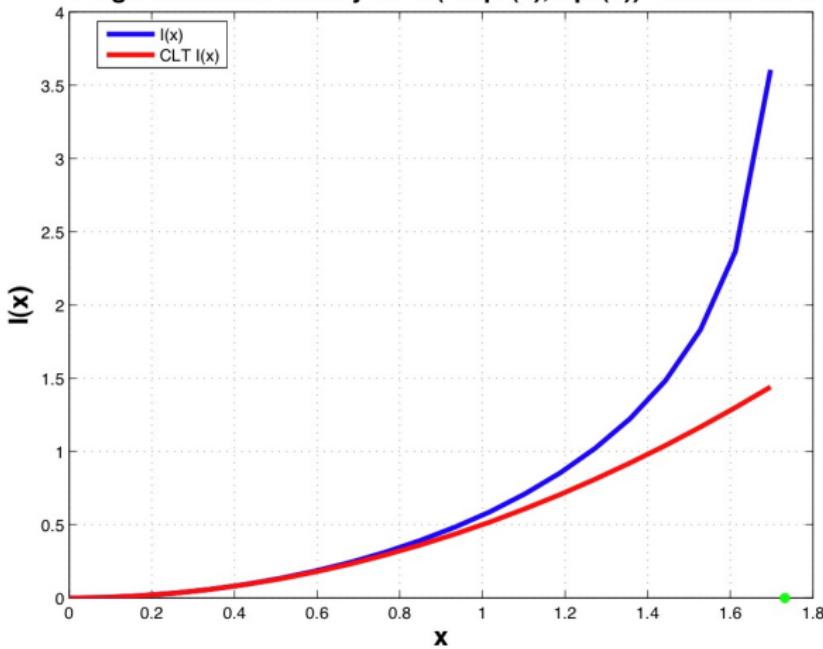
which converges exponentially with J but is still strictly positive!!

This example shows that the LDT gives a sharper estimation of the tail probability than the CLT!

Let us compare the LDT rate for this uniform r.v.s. with the one provided by the CLT,

$$I_{CLT}(x) = \frac{x^2}{2}$$

Large Deviation Theory for $U(-\sqrt{3}, \sqrt{3})$ random vars.



Example: Exponential random variables

Let ξ_i be iid $Exp(1)$, i.e. $\rho_\xi(x) = \exp(-x) 1_{[0,+\infty)}(x)$, $E[\xi_i] = 1$,
 $\text{Var}([\cdot]\xi) = 1$.

Then we have bounded exponential moments,

$$E[e^{\theta\xi}] = \frac{1}{1-\theta}, \text{ for } \theta < 1$$

and

$$\Lambda(\theta) = \log(E[e^{\theta\xi}]) = -\log(1-\theta).$$

We now compute, for the auxiliary mean zero random variables

$$X \equiv \xi - 1,$$

the rate function

$$\begin{aligned} I(x) &= \sup_{\theta} \{\theta x - (\Lambda(\theta) - \theta)\} \\ &= \sup_{\theta} \{\theta x + \log(1 - \theta) + \theta\} \\ &= \begin{cases} x - \log(x + 1), & \text{for } x > -1 \\ +\infty, & \text{otherwise.} \end{cases} \end{aligned}$$

The rate explosion $I(x) = +\infty$ for $x \leq -1$ agrees with the obvious fact

$$P\left(\sum_{j=1}^J \xi_j \leq 0\right) = 0.$$

Now we compare again with the CLT, keeping in mind that x is a deviation from the mean. Then the CLT rate function is

$$I_{CLT}(x) = \frac{x^2}{2}$$

For

$$|x| \ll 1$$

we expect good agreement with the LDT prediction.

Large Deviation Theory for $\text{Exp}(1)$ random vars.

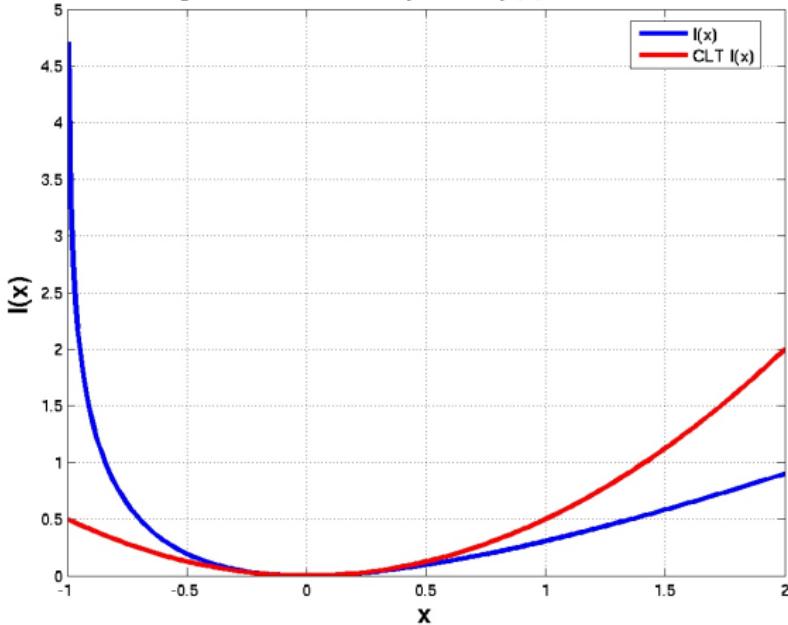


Figure: $I(x)$ and $I_{\text{CLT}}(x) = x^2$. $X \sim \text{Exp}(1) - 1$ takes values in $(-1, +\infty)$.

Large Deviations and Small Deviations

Question:

What does LDT say about small deviations of the order $x = \frac{y}{\sqrt{J}}$, where y is a given constant?

We have

$$I\left(\frac{y}{\sqrt{J}}\right) = I(0) + I'(0)\frac{y}{\sqrt{J}} + I''(0)\frac{y^2}{2J} + \mathcal{O}\left(\frac{y^3}{J^{3/2}}\right)$$

Let us investigate the behavior of $I(x)$ near zero. We have

$$I(x) = \sup_{\theta} \{x\theta - \Lambda(\theta)\}$$

First we observe that

$$E[e^{\theta\xi}] = 1 + \theta E[\xi] + \frac{\theta^2}{2} E[\xi^2] + \mathcal{O}(\theta^3)$$

and

$$\Lambda(\theta) = \log(E[e^{\theta\xi}]) = \frac{\theta^2}{2} Var[\xi] + \mathcal{O}(\theta^3)$$

so

$$\begin{aligned} I(x) &= \sup_{\theta} \left\{ x\theta - \frac{\theta^2}{2} Var[\xi] + \mathcal{O}(\theta^3) \right\} \\ &= \frac{x^2}{2Var[\xi]} + \mathcal{O}(x^3) \end{aligned}$$

In conclusion, we have from the combination of the Chernoff bound with $x = y/\sqrt{J}$ and the estimate $I(x) = \frac{x^2}{2\text{Var}[\xi]} + \mathcal{O}(x^3)$:

$$\begin{aligned} P\left(\sum_{j=1}^J \frac{\xi_j}{\sqrt{J}} > y\right) &\leq e^{-JI(y/\sqrt{J})} \\ &\leq e^{-\frac{y^2}{2\text{Var}[\xi]} + \mathcal{O}(y^3/\sqrt{J})} \end{aligned}$$

which is a similar (but not as sharp) to the type of estimates one can produce with the Central Limit Theorem!

Exercise 10.4 (Cramér insurance revisited)

Consider again Example 10.1. An insurance company has to face random claims with payoff

$$S_N = \sum_{n=1}^N X_n.$$

Here X_n are modeled as iid positive random variables with mean μ and bounded exponential moments. Estimate, in terms of μ and the rate function I the corresponding insurance premium

$$qN$$

so that, for some given $\epsilon \ll 1$, we have

$$P(S_N > qN) < \epsilon.$$

What happens, for fixed ϵ , with the difference

$$0 < q - \mu$$

as N gets larger?

Generalizing the Cramér-Chernoff's theorem: vector case

Reference: An introduction to large deviations / Vitalii Konarovskyi / Lecture Notes, 2019.

Let $X \in \mathbb{R}^d$. As before, we define, for $\theta \in \mathbb{R}^d$,

$$\Lambda_X(\theta) = \log E[\exp(\theta \cdot X)]$$

and its Legendre transform, the rate function

$$I_X(x) = \sup_{\theta \in \mathbb{R}^d} \{\Lambda_X(\theta) - \theta \cdot x\}$$

Exercise 10.5

Let X, Y be independent random vectors and their concatenation, $Z = [X, Y]$. Show that

$$\Lambda_Z([\theta_1, \theta_2]) = \Lambda_X(\theta_1) + \Lambda_Y(\theta_2)$$

and that the corresponding rate function satisfies

$$I_Z([x, y]) = I_X(x) + I_Y(y).$$

Exercise 10.6

For any random vector $X \in \mathbb{R}^d$ and a non-singular matrix $A \in \mathbb{R}^{d \times d}$, show that

$$\Lambda_{AX}(\theta) = \Lambda_X(A\theta)$$

and the rate function satisfies

$$I_{AX}(x) = I_X(A^{-1}x).$$

Exercise 10.7

Compute the rate function for a Gaussian vector $N(0, \Sigma)$.

Theorem 10.3 (iid vector Chernoff's theorem)

Let $(X_i)_{i \geq 1}$ be a sequence of independent identically distributed random vectors in \mathbb{R}^d with cumulant generating function Λ and let $S_N = \sum_{n=1}^N X_n$. If Λ is finite in a neighborhood of 0 then the family S_N satisfies the large deviation principle with good rate function I , that is, for every Borel set A

$$\begin{aligned}-\inf_{x \in A^\circ} I(x) &\leq \liminf_{N \rightarrow +\infty} \frac{1}{N} \log P \left(\frac{1}{N} S_N \in A \right) \\ &\leq \limsup_{N \rightarrow +\infty} \frac{1}{N} \log P \left(\frac{1}{N} S_N \in A \right) \\ &\leq -\inf_{x \in \bar{A}} I(x)\end{aligned}$$

Generalizing the Chernoff's bound: matrix case

A Chernoff's bound similar to the scalar case holds also for symmetric and positive definite (spd) matrices.

Theorem 10.4

Matrix Chernoff's bound (for i.i.d. random matrices) [J. Tropp, FoCM 2011]

Let $X_1, \dots, X_J \in \mathbb{R}^{d \times d}$ be i.i.d. spd random matrices s.t. $\lambda_{\max}(X_j) \leq R$ almost surely. Let $\mu_{\min} = \lambda_{\min}(E[X_j])$, $\mu_{\max} = \lambda_{\max}(E[X_j])$ and $\bar{X} = \frac{1}{J} \sum_{j=1}^J X_j$. Then

$$P(\lambda_{\max}(\bar{X}) \geq (1 + \delta)\mu_{\max}) \leq d \exp \left\{ -\frac{J\mu_{\max}\tilde{\beta}_{\delta}}{R} \right\}, \quad \delta \geq 0$$

$$P(\lambda_{\min}(\bar{X}) \leq (1 - \delta)\mu_{\min}) \leq d \exp \left\{ -\frac{J\mu_{\min}\beta_{\delta}}{R} \right\}, \quad \delta \in [0, 1),$$

with

$$\tilde{\beta}_{\delta} = (1 + \delta) \log(1 + \delta) - \delta$$

and

$$\beta_{\delta} = \delta + (1 - \delta) \log(1 - \delta)$$

Remark 10.3 (Special case)

Let $E[X_j] = \text{Identity}$ in $\mathbb{R}^{d \times d}$. In this case, $E[\bar{X}] = \text{Identity}$, $\mu_{\max} = \mu_{\min} = 1$ and the following bound can then be obtained

$$P(\|\bar{X} - I\| \geq \delta) \leq 2d \exp \left\{ -\frac{J\beta_\delta}{R} \right\}$$

For more related results, see

https://en.wikipedia.org/wiki/Matrix_Chernoff_bound

Remark 10.4 (L^2 -discrete regression)

Matrix LDT results are useful to analyze the stability of the L^2 -discrete regression problem, connecting the number of data with the dimension of the projection subspace.

Large Deviation Theory connection to importance sampling⁸

Motivation: sampling of rare events

Consider the Monte Carlo computation of $E[X]$, where $X \sim Ber(p)$ and $p << 1$.

We have that $E[X] = \mu = p$ and $Var[X] = \sigma^2 = p(1 - p) \approx p$. Therefore, the *relative error* in the Monte Carlo computation is approximately bounded in probability by

$$C_\alpha \frac{\sigma}{\mu\sqrt{J}} \approx C_\alpha \frac{1}{\sqrt{pJ}}.$$

As a consequence, we need the number of samples to be $J \gg \frac{1}{p}$ to control the relative error.

Example: Consider the following Bernoulli random variable

$$X = 1_{\{A_N > \text{TOL}\}}$$

with

$$A_N = \frac{S_N}{N} = \frac{1}{N} \sum_{n=1}^N \xi_n$$

and $\{\xi_n\}$ an iid family with mean zero, unit variance and finite exponential moments. Then, the Cramér-Chernoff theorem in LDT tells us that

$$p_N = P(A_N > \text{TOL}) \approx \exp(-NI(\text{TOL}))$$

and therefore Monte Carlo needs

$$J(N) \gg \frac{1}{p_N} \approx \exp(NI(\text{TOL}))$$

samples to control the relative sampling error!!

This large number of samples, $J(N)$, is clearly unfeasible and we need an importance sampling technique to reduce the MC variance.

Importance sampling by exponential twisting

⁸Huyen Pham (2007), Some Applications and Methods of Large Deviations in Finance and Insurance, In: *Paris-Princeton Lectures on Mathematical Finance 2004*, Lecture Notes in Mathematics, Springer.

Following on our previous example, we want to estimate

$$\begin{aligned} p_N &= P(A_N > \text{TOL}) = P(S_N > N \text{TOL}) \\ &\approx \exp(-N I(\text{TOL})). \end{aligned} \tag{73}$$

We have seen that an exponential change of measure is used in proving large deviations results, cf. (70).

Given the real random variable X with pdf ρ_X , consider a new distribution on \mathbb{R} with pdf

$$\rho_\theta(x) := \exp(\theta x - \Lambda(\theta)) \rho_X(x). \tag{74}$$

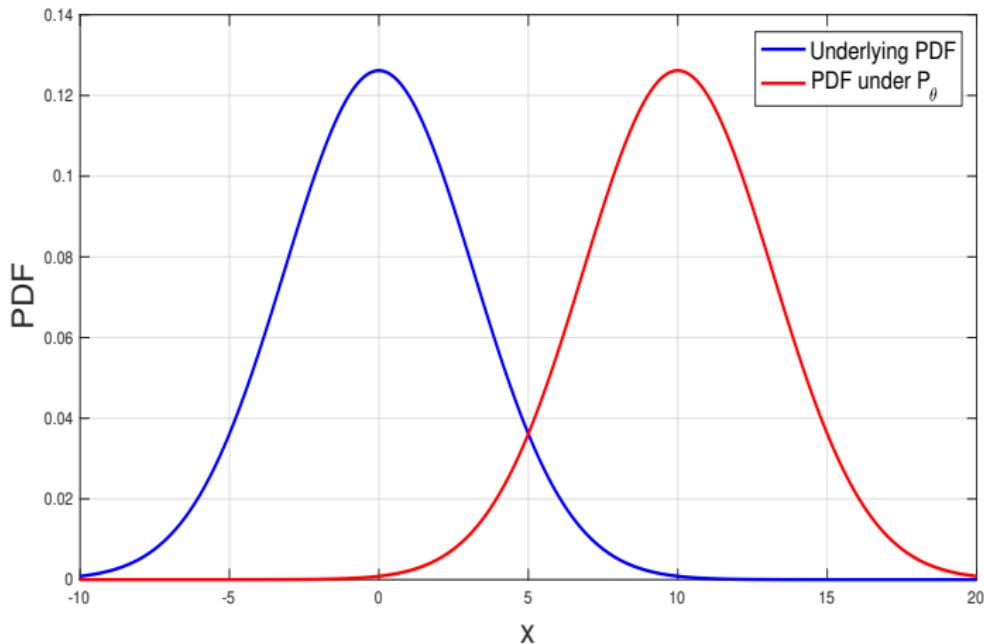
Then, given a sequence X_1, \dots, X_N of real iid random variables with pdf ρ_X , consider the new probability measure P_θ such that the X 's are iid with marginal ρ_θ . Denote by E_θ the expectation with respect to P_θ . We have, due to iid,

$$\begin{aligned} E[f(X_1, \dots, X_N)] \\ = E_\theta \left[f(X_1, \dots, X_N) \exp \left(-\theta \sum_{n=1}^N X_n + N\Lambda(\theta) \right) \right], \end{aligned}$$

for all integers $N \geq 1$.

Applying the exponential change of measure (74) to the computation of p_N yields

$$\begin{aligned} p_N &= E \left[\mathbf{1}_{\left\{ \frac{s_N}{N} \geq \text{TOL} \right\}} \right] \\ &= E_\theta \left[\exp(-\theta S_N + N\Lambda(\theta)) \mathbf{1}_{\left\{ \frac{s_N}{N} \geq \text{TOL} \right\}} \right]. \end{aligned} \tag{75}$$



The underlying and the exponential twisting PDFs of S_N with $N = 10$, $TOL = 1$ and $\xi_i \sim \mathcal{N}(0, 1)$. Under P_θ , $S_N \sim \mathcal{N}(NTOL, N)$

Consider the random variable

$$Y_N = \exp(-\theta S_N + N\Lambda(\theta)) \mathbf{1}_{\left\{ \frac{S_N}{N} \geq \text{TOL} \right\}}$$

Recall (75). An unbiased *importance sampling (IP) estimator* for p_N is

$$Z_J = \frac{1}{J} \sum_{j=1}^J Y_{N,j},$$

with the iid samples of Y_N taken from the probability measure P_θ .

The second moment of Y_N under P_θ satisfies

$$\begin{aligned} E_\theta [Y_N^2] &= E_\theta \left[\left(\exp(-\theta S_N + N\Lambda(\theta)) \mathbf{1}_{\{\frac{s_N}{N} \geq \text{TOL}\}} \right)^2 \right] \\ &= E_\theta \left[e^{-2\theta S_N + 2N\Lambda(\theta)} \mathbf{1}_{\{\frac{s_N}{N} \geq \text{TOL}\}} \right] \\ &\leq e^{-2N(\theta \text{TOL} - \Lambda(\theta))}. \end{aligned} \tag{76}$$

By the Cramér-Chernoff theorem we have

$$E_\theta[Z_J] = E_\theta[Y_N] = p_N = P\left(\frac{S_N}{N} \geq \text{TOL}\right) \approx e^{-NI(\text{TOL})}.$$

Also, we have that, by (76),

$$\begin{aligned} \text{Var}_\theta[Y_N] &= E_\theta[Y_N^2] - p_N^2 \\ &\leq e^{-2N(\theta\text{TOL} - \Lambda(\theta))} - p_N^2 \end{aligned} \tag{77}$$

Therefore, by minimizing the upper bound above wrt θ , we see that the fastest possible exponential rate of decay of the second moment of the IP estimator is $2I(\text{TOL})$.

Recall from (73) that $p_N \approx \exp(-NI(\text{TOL}))$. Choosing $\theta = \theta_{\text{TOL}}$ to minimize the upper bound of $\text{Var}_\theta[Y_N]$ in (77), namely

$$I(\text{TOL}) = \theta_{\text{TOL}} \text{TOL} - \Lambda(\theta_{\text{TOL}}),$$

we achieve that rate and obtain an asymptotically optimal IP estimator:

$$\lim_{N \rightarrow \infty} \frac{1}{N} \log p_N^2 = \lim_{N \rightarrow \infty} \frac{1}{N} \log E_{\theta_{\text{TOL}}} [Y_N^2].$$

Interpret this result, recall that we want to have control on $\frac{E_{\theta_{\text{TOL}}}[Y_N^2]}{p_N^2}$ so the number of samples J does not grow too fast with N .

Large Deviations, PDE connections

Consider a linear PDE, whose solution u depends linearly on a random forcing term f via a *deterministic* solution operator \mathcal{S} , i.e.

$$u = \mathcal{S}f$$

Consider now a linear expansion for the random field

$$f = f_0 + \sum_{n \geq 1} f_n Y_n$$

Here (f_n) are deterministic functions and (Y_n) are zero mean independent random variables, not necessarily identically distributed. We assume that the r.v.s. Y_n have exponential moments bounded.

We want to estimate the following probability

$$P(Q(u) > K)$$

with Q being a linear bounded functional.

Example 10.4

Let us consider a random elliptic PDE:

$$\begin{aligned} -\nabla \cdot (a(\mathbf{x}) \nabla u(\mathbf{x}; \omega)) &= f(\mathbf{x}; \omega) && \text{for } \mathbf{x} \in D = [0, 1]^d, \\ u(\mathbf{x}; \omega) &= 0 && \text{for } \mathbf{x} \in \partial D, \end{aligned}$$

for sufficiently regular deterministic coefficient a and a random forcing f , such that the PDE is well-posed (a.s.). Given $v \in \mathbb{R}^d$, we can consider the linear bounded functional

$$Q(u) = \int_{\Sigma} a(\mathbf{x}) \nabla u(\mathbf{x}) \cdot v \, d\mathbf{x}.$$

with the open set $\Sigma \subset D$.

If we introduce a sequence of auxiliary deterministic solutions

$$u_n = \mathcal{S}f_n$$

and corresponding real valued functional evaluations,

$$q_n = Q(u_n)$$

then we can write, thanks to linearity,

$$Q(u) = q_0 + \sum_{n \geq 1} q_n Y_n$$

and thus, using the standard LDT arguments,

$$\begin{aligned} P(Q(u) > K) &= P(q_0 + \sum_{n \geq 1} q_n Y_n > K) \\ &\leq \frac{E[\exp(\theta \sum_{n \geq 1} q_n Y_n)]}{\exp(\theta(K - q_0))} \\ &\leq \frac{\exp(\sum_{n \geq 1} \Lambda_{Y_n}(\theta q_n))}{\exp(\theta(K - q_0))}. \end{aligned}$$

In conclusion, we can bound

$$P(Q(u) > K) \leq \exp(-I_Q(K - q_0))$$

with the rate function

$$I_Q(x) = \sup_{\theta} \left\{ \theta x - \sum_{n \geq 1} \Lambda_{Y_n}(\theta q_n) \right\}$$

Observe:

The previous argument holds for both time dependent and time independent linear PDEs!

Linear Elliptic PDE with random diffusion coefficient

Consider again Example 10.4 but now assume that the forcing term f is deterministic and the diffusion coefficient a can be approximated by the (random) expansion

$$\log(a) = b_0 + \sum_{n \geq 1} b_n Y_n$$

Here (b_n) are deterministic functions and (Y_n) are zero mean independent random variables, not necessarily identically distributed. We assume that the r.vars. $|Y_n|$ have exponential moments bounded. We want to estimate the following probability

$$P(Q(u) > K)$$

with Q being a linear bounded functional.

We now estimate

$$|Q(u)| \leq \frac{\|Q\|_{H^{-1}(D)} C_D \|f\|_{L^2(D)}}{a_{\min}}$$

with the random variable

$$a_{\min}(\omega) = \min_{x \in D} a(x, \omega).$$

Define the constant

$$\tilde{q}_0 = \frac{\|Q\|_{H^{-1}(D)} C_D \|f\|_{L^2(D)}}{\exp(\min_{x \in D} b_0(x))}$$

and further bound

$$|Q(u)| \leq \tilde{q}_0 \exp \left(\sum_{n \geq 1} \|b_n\|_\infty |Y_n| \right)$$

yielding

$$P(Q(u) > K) \leq P \left(\exp \left(\sum_{n \geq 1} \|b_n\|_\infty |Y_n| \right) > K/\tilde{q}_0 \right)$$

To conclude, we use LDT estimates on

$$P(Q(u) > K) \leq P \left(\sum_{n \geq 1} \|b_n\|_\infty |Y_n| > \log(K/\tilde{q}_0) \right)$$

To this end, introduce the centered random variables

$$\xi_n = |Y_n| - E[|Y_n|]$$

and then

$$P(Q(u) > K) \leq \exp \left(-I \left(\log(K/\tilde{q}_0) - \sum_{n \geq 1} \|b_n\|_\infty E[|Y_n|] \right) \right)$$

with

$$I(x) = \sup_{\theta} \left\{ \theta x - \sum_{n \geq 1} \Lambda_{\xi_n}(\theta \|b_n\|_\infty) \right\}$$

Exercise 10.8

Consider Example 10.4 but now assume that both the forcing term f and the diffusion coefficient a are random. Making suitable assumptions, estimate

$$P(Q(u) > K).$$

Quasi Monte Carlo Methods

Quasi Monte Carlo: Motivation

Let $\mathbf{y} = (y_1, \dots, y_N)$ be a vector of independent uniform random variables in $\Gamma = [0, 1]^N$, $u: \Gamma \rightarrow V$ a given Hilbert-valued function and $Q: V \rightarrow \mathbb{R}$ a functional on V . We are interested in computing

$$I_N := \mathbb{E}[Q(u)] = \int_{\Gamma} Q(u(\mathbf{y})) \rho(\mathbf{y}) d\mathbf{y}, \quad \Gamma = [0, 1]^N, \quad \rho(\mathbf{y}) \equiv 1$$

A Monte Carlo approximation uses M random points $\xi_1, \dots, \xi_M \in \Gamma$, sampled independently of the density ρ , and approximates I_N as

$$I_{N,M} = \frac{1}{M} \sum_{i=1}^M Q(u(\xi_i)), \quad (*)$$

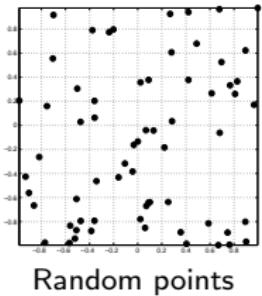
achieves an error of

$$|I_N - I_{N,M}| \leq c_{1-\delta/2} \frac{\sqrt{\text{Var}[Q(u)]}}{\sqrt{M}}$$

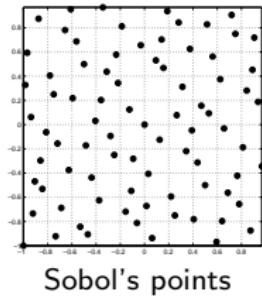
with asymptotic confidence $1 - \delta$.

Observe that the MC estimator (*) can be seen as a quadrature formula with equal weights M^{-1} .

Question: while keeping the approximation form (*) with equal weights, is there a better distribution of points than the purely random one?



Random points



Sobol's points

Observation: sample of uniformly distributed random variables does not seem to *cover* the hypercube “uniformly”

Idea of Quasi Monte Carlo sampling: consider **purely deterministic** point distributions to improve upon the rate $1/\sqrt{M}$ by using “better” space-filling, while keeping the simple structure (*).

Quasi Monte Carlo methods and Discrepancy

Let $P = \{\xi_1, \dots, \xi_M\}$ be a set of points $\xi_i \in [0, 1]^N$ and $f : [0, 1]^N \rightarrow \mathbb{R}$ a continuous function. A **Quasi Monte Carlo (QMC)** method to approximate $I_N(f) = \int_{[0,1]^N} f(\mathbf{y}) d\mathbf{y}$ is an *equal weight cubature formula* of the form

$$I_{N,M}(f) = \frac{1}{M} \sum_{i=1}^M f(\xi_i).$$

The key concept in the analysis of QMC methods is the one of **discrepancy**. **Notation:** for $\mathbf{x} \in [0, 1]^N$ let $[\mathbf{0}, \mathbf{x}] := [0, x_1] \times \dots \times [0, x_N]$. Then

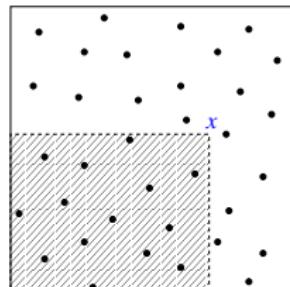
$$\text{Vol}([\mathbf{0}, \mathbf{x}]) \approx \widehat{\text{Vol}}_P([\mathbf{0}, \mathbf{x}]) := \frac{\# \text{ points in } [\mathbf{0}, \mathbf{x}]}{M}$$

for a given point set $P = \{\xi_1, \dots, \xi_M\}$.

Local **discrepancy function** $\Delta_P : [0, 1]^N \rightarrow [-1, 1]$

$$\Delta_P(\mathbf{x}) := \widehat{\text{Vol}}_P([\mathbf{0}, \mathbf{x}]) - \text{Vol}([\mathbf{0}, \mathbf{x}])$$

$$= \frac{1}{M} \sum_{i=1}^M 1_{[\mathbf{0}, \mathbf{x}]}(\xi_i) - \prod_{i=1}^N x_i$$



Intuition suggests that a good QMC method should be built based on a point set that has a “small” local discrepancy function. Possible measures for “small” include:

- ▶ Star discrepancy (or uniform discrepancy):

$$\Delta_{P,N}^* := \|\Delta_P\|_{L^\infty([0,1]^N)} \equiv \sup_{\mathbf{x} \in [0,1]^N} |\Delta_P(\mathbf{x})|$$

- ▶ L^q discrepancy ($1 \leq q < \infty$):

$$\Delta_{P,N,q} := \|\Delta_P\|_{L^q([0,1]^N)} \equiv \left(\int_{[0,1]^N} |\Delta_P(\mathbf{x})|^q d\mathbf{x} \right)^{1/q}$$

The importance of discrepancy: error representation

The reason why the discrepancy plays an important role in the study of QMC quadrature formulas follows from the famous **Koksma–Hlawka inequality**, which provides an upper bound on the QMC integration error.

We begin with $N = 1$. We will need the following.

Lemma 11.1 (Zaremba's identity)

Let $f : [0, 1] \rightarrow \mathbb{R}$ be an absolutely continuous function with integrable derivative and let $P = \{\xi_1, \dots, \xi_M\}$ be any point set in $[0, 1]$. Then

$$\begin{aligned} \int_0^1 f(y) dy - \frac{1}{M} \sum_{i=1}^M f(\xi_i) &= \int_0^1 f'(y) \Delta_P(y) dy \\ &= \int_0^1 f'(y) \Delta_P(y) dy - \Delta_P(1)f(1). \end{aligned}$$

Proof.

We use that for all $y \in [0, 1]$:

$$f(y) = f(1) - \int_y^1 f'(x) dx = f(1)1_{[0,1]}(y) - \int_0^1 f'(x) \underbrace{1_{[y,1]}(x)}_{=1_{[0,x]}(y)} dx$$

to find

$$\begin{aligned} \int_0^1 f(y) dy - \frac{1}{M} \sum_{i=1}^M f(\xi_i) &= f(1) \underbrace{\left(\int_0^1 1_{[0,1]}(y) dy - \frac{1}{M} \sum_{i=1}^M 1_{[0,1]}(\xi_i) \right)}_{\equiv -\Delta_P(1)=0} \\ &\quad - \int_0^1 f'(x) \underbrace{\left(\int_0^x 1 dy - \frac{1}{M} \sum_{i=1}^M 1_{[0,x]}(\xi_i) \right)}_{=-\Delta_P(x)} dx \\ &= \int_0^1 f'(x) \Delta_P(x) dx . \end{aligned}$$

□

Hence, the error in the QMC integration is bounded by

$$\left| \frac{1}{M} \sum_{i=1}^M f(\xi_i) - \int_{[0,1]} f(y) dy \right| \leq \|\Delta_P\|_{L^p(0,1)} \|f'\|_{L^q(0,1)}, \quad \text{with } \frac{1}{p} + \frac{1}{q} = 1,$$

and $p, q \in [1, \infty]$ in view of Hölder's inequality. The preceding bound is sometimes called **Koksma–Hlawka** inequality.

In particular:

$$\left| \frac{1}{M} \sum_{i=1}^M f(\xi_i) - \int_{[0,1]} f(y) dy \right| \leq \Delta_P^* \|f'\|_{L^1(0,1)},$$

provided that f' is integrable.

Remark: the norm $\|f'\|_{L^1}$ can be replaced by $\|f\|_{TV}$, the total variation norm of f in the sense of Hardy and Krause.

QMC error representation in N dimensions

The previous analysis extends with some care to the multi-dimensional setting. We introduce the following **notation**:

- ▶ let $[1 : N]$ denote the set of subsets of $\{1, 2, \dots, N\}$, including the empty set,
- ▶ for $\alpha \in [1 : N]$, let $\#(\alpha) \equiv |\alpha|$ denote the cardinality of α ,
- ▶ for $\mathbf{y} \in [0, 1]^N$ and $\alpha \in [1 : N]$, let \mathbf{y}_α denote the vector $\{\mathbf{y}_j, \text{ with } j \in \alpha\}$ and $(\mathbf{y}_\alpha, 1)$ the vector identical to \mathbf{y} for components included in α but with components set to 1 for those not included in α .

Lemma 11.2 (Hlawka's identity)

Let $f: [0, 1]^N \rightarrow \mathbb{R}$ be an integrable function with mixed first order derivatives and let $P = \{\xi_1, \dots, \xi_M\}$ be any point set in $[0, 1]^N$. Then

$$\frac{1}{M} \sum_{i=1}^M f(\xi_i) - \int_{[0, 1]^N} f(\mathbf{y}) d\mathbf{y} = \sum_{\alpha \in [1:N]} (-1)^{\#(\alpha)} \int_{[0, 1]^{\#(\alpha)}} \frac{\partial^{\#(\alpha)} f}{\partial \mathbf{y}_\alpha} (\mathbf{y}_\alpha, 1) \Delta_P(\mathbf{y}_\alpha, 1) d\mathbf{y}_\alpha ,$$

where $\frac{\partial^{\#(\alpha)} f}{\partial \mathbf{y}_\alpha}$ is a mixed first order derivative.

Proof Idea.

By induction on N , one first proves the identity

$$f(\mathbf{x}) = \sum_{\alpha \in [1:N]} (-1)^{\#(\alpha)} \int_{[\mathbf{x}_\alpha, 1]} \frac{\partial^{\#(\alpha)} f}{\partial \mathbf{y}_\alpha}(\mathbf{y}_\alpha, 1) \Delta_P(\mathbf{y}_\alpha, 1) d\mathbf{y}_\alpha \quad \mathbf{x} \in [0, 1]^N,$$

which generalizes the $N = 1$ identity $f(x_1) = f(1) - \int_{x_1}^1 \frac{\partial f}{\partial x_1}(y) dy$ used in the proof of Lemma 11.1. The rest of the proof then is analogous to the one of Zaremba's identity. \square

Define now the function space $V_{1,p} = \{f : [0, 1]^N \rightarrow \mathbb{R}, \|f\|_{1,p} < +\infty\}$, with norm

$$\|f\|_{1,p} = \left(\sum_{\alpha \in [1:N]} \int_{[0, 1]^{\#(\alpha)}} \left| \frac{\partial^{\#(\alpha)} f}{\partial \mathbf{y}_\alpha}(\mathbf{y}_\alpha, 1) \right|^p d\mathbf{y}_\alpha \right)^{\frac{1}{p}}$$

For $p = 2$, $V_{1,2}$ is a Hilbert space isomorphic to $H_{mix}^1([0, 1]^N)$. It measures **mixed first derivatives** of the function f .

Excursion: the $H_{mix}^k(\Gamma)$ space

Space of functions with k mixed square integrable derivatives: let $\beta = (\beta_1, \dots, \beta_n) \in \mathbb{N}^N$ be a multi-index and the seminorms

$$|v|_{H_{mix}^k(\Gamma, V)}^2 = \sum_{|\beta|=k} \int_{\Gamma} \left| \frac{\partial^k v(\mathbf{y})}{\partial y_1^{\beta_1} \cdots \partial y_n^{\beta_n}} \right|^2 \rho(\mathbf{y}) d\mathbf{y}.$$

The Hilbert space

$$H_{mix}^k(\Gamma, V) = \{v \in L^2_{\rho}(\Gamma, V) : \sum_{s=0}^k |v|_{H_{mix}^s(\Gamma, V)}^2 < +\infty\}$$

is endowed with the norm

$$\|v\|_{H_{mix}^k(\Gamma, V)}^2 = \sum_{s=0}^k |v|_{H_{mix}^s(\Gamma, V)}^2 .$$

Koksma–Hlawka inequality

From the Hlawka's identity in Lemma 11.2, the multidimensional version of the **Koksma–Hlawka inequality** follows. Indeed, let $1 \leq p < \infty$. Then, Hölder's inequality yields

$$\left| \int_{[0,1]^N} f(\mathbf{y}) d\mathbf{y} - \frac{1}{M} \sum_{i=1}^M f(\boldsymbol{\xi}_i) \right| \leq \|\Delta_P\|_{1,q} \|f\|_{1,p}, \quad \text{with } \frac{1}{p} + \frac{1}{q} = 1$$

The Quasi Monte Carlo error is bounded by the sum of the q -norm of the local discrepancies in all dimensions $1, \dots, N$, provided the function has p -integrable mixed first derivatives.

Sometimes only the special case $p = 1$ (and $q = \infty$) is known as Koksma–Hlawka inequality:

$$\left| \int_{[0,1]^N} f(\mathbf{y}) d\mathbf{y} - \frac{1}{M} \sum_{i=1}^M f(\boldsymbol{\xi}_i) \right| \leq \Delta_P^* \|f\|_{1,1}.$$

Remark 11.1

Other equivalent norms (but the equivalence constant depends on $N!$) can be considered as well, and analogous results can be obtained. See [Dick-Kuo-Sloan '13] for an extensive discussion. Examples include:

- ▶ Anchored Sobolev space: for $\mathbf{c} \in [0, 1]$

$$\|f\|_{1,2} = \left(\sum_{\alpha \in [1:N]} \int_{[0,1]^{\#(\alpha)}} \left| \frac{\partial^{\#(\alpha)} f}{\partial \mathbf{y}_\alpha}(\mathbf{y}_\alpha, \mathbf{c}) \right|^2 d\mathbf{y}_\alpha \right)^{\frac{1}{2}}$$

- ▶ Unanchored Sobolev space:

$$\|f\|_{1,2} = \left(\sum_{\alpha \in [1:N]} \int_{[0,1]^{\#(\alpha)}} \left(\int_{[0,1]^{N-\#(\alpha)}} \frac{\partial^{\#(\alpha)} f}{\partial \mathbf{y}_\alpha}(\mathbf{y}_\alpha, \mathbf{y}_{-\alpha}) d\mathbf{y}_{-\alpha} \right)^2 d\mathbf{y}_\alpha \right)^{\frac{1}{2}}$$

where $\mathbf{y}_{-\alpha}$ denotes the vector $\{y_j, j \notin \alpha\}$.

Low discrepancy sequences

Definition 11.3

A sequence (ξ_1, ξ_2, \dots) , $\xi_i \in [0, 1]^N$ is a **low discrepancy sequence** if the set of points $P_M = (\xi_1, \dots, \xi_M)$ satisfies

$$\Delta_{P_M, N}^* \equiv \Delta_{P_M}^* \leq C_N \frac{(\log M)^N}{M}.$$

QMC methods use equal weight cubature formulae with low discrepancy sequences of points.

Low order projections: let $\alpha \subset \{1, \dots, N\}$ a subset of indices, with $\#(\alpha) = s$. The projected sequence $P_M^\alpha = (\xi_1^\alpha, \dots, \xi_M^\alpha)$ onto the α directions has discrepancy satisfying

$$\Delta_{P_M^\alpha, s}^* \leq \Delta_{P_M, N}^*.$$

Proof:
$$\Delta_{P_M^\alpha, s}^* = \sup_{x^\alpha \in [0, 1]^s} \left(\frac{1}{M} \sum_{i=1}^M 1_{[0, x^\alpha]}(\xi_i^\alpha) - \prod_{i \in \alpha} x_i \right)$$

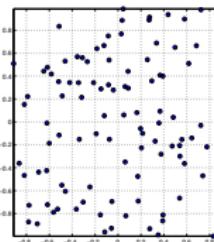
$$= \sup_{y = (x^\alpha, 1) \in [0, 1]^N} \left(\frac{1}{M} \sum_{i=1}^M 1_{[0, y]}(\xi_i) - \prod_{i=1}^N y_i \right) \leq \Delta_{P_M, N}^*.$$

Implication: if $\Delta_{P_M, N}^* \sim (\log M)^N / M$, the discrepancy of any projection on dimension $s < N$ has to behave the same way or better!

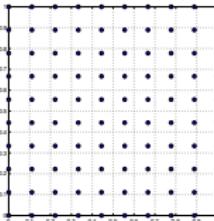
Exercise 11.1

Let $N \in \mathbb{N}$. How does the star discrepancy scale in M for:

- ▶ M random i.i.d. uniformly distributed points in $[0, 1]^N$



- ▶ uniform grid of $m = M^{1/N}$ points per edge



Low discrepancy sequences: Examples

Several low discrepancy sequences exist (\rightsquigarrow Matlab/Python codes):
Halton, Hammersley, Sobol, Faure, Niederreiter, ...; cf. [Niederreiter '92]

The simplest example: Halton sequence. For integers i and $b \geq 2$, we first define the *radical inverse function* $\phi_b(i)$ as follows:

$$\text{if } i = \sum_{n=1}^{\infty} i_n b^{n-1}, \quad i_n \in \{0, 1, \dots, b-1\}, \quad \text{then} \quad \phi_b(i) = \sum_{n=1}^{\infty} i_n b^{-n},$$

so, in base $b = 10$, the radical inverse of $i = 15421$ is $\phi_{10}(i) = 0.12451$.
Then, if p_1, \dots, p_N denote the first N prime numbers, the Halton sequence $P = \{\xi_1, \xi_2, \dots\}$ is given by

$$\xi_i = (\phi_{p_1}(i), \phi_{p_2}(i), \dots, \phi_{p_N}(i))$$

and its star-discrepancy satisfies

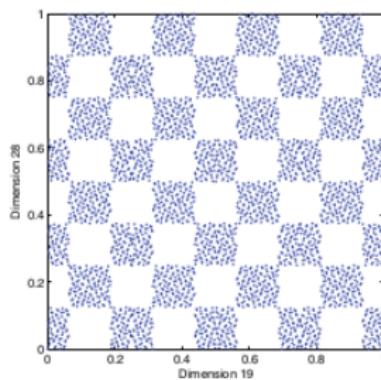
$$\Delta_N^* := \sup_{\mathbf{x} \in [0,1]^N} |\Delta_P(\mathbf{x})| = O\left(\frac{(\log M)^N}{M}\right).$$

Other (better) seq. such as Hammersley, Sobol, Faure, etc., even have

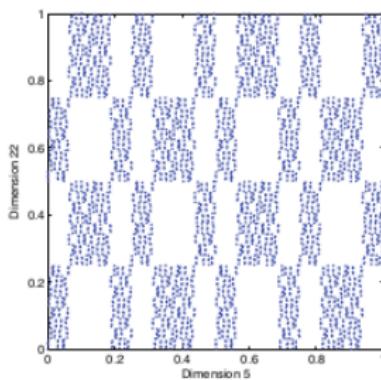
$$\text{a star-discrepancy } \Delta_N^* = O\left(\frac{(\log M)^{N-1}}{M}\right).$$

Having low discrepancy in high dimension is difficult!

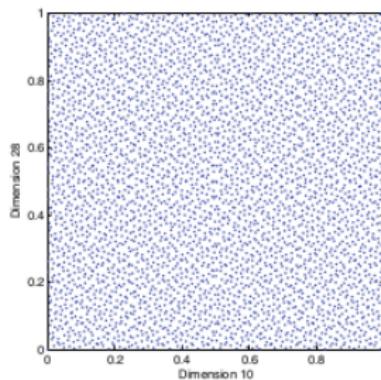
(a) The $(19, 28)$ -projection (t -value is 6)



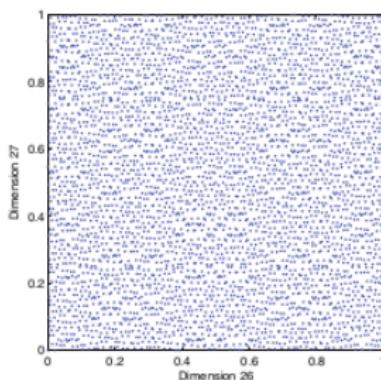
(b) The $(5, 22)$ -projection (t -value is 7)



(c) The $(10, 28)$ -projection (t -value is 1)



(d) The $(26, 27)$ -projection (t -value is 2)



from [Joe-Kuo, 2008]:
two dimensional
projections of 4096 Sobol'
points in dimension 28.

Lattice point set rules

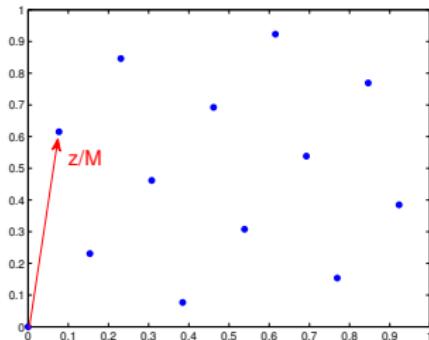
Let M be the number of points we want to use and $\mathbf{z} \in \mathbb{N}^N$ be an integer “generating” vector whose components have no factor in common with M :

$$\text{rank-one lattice rule} \quad \xi_i = \left\{ \frac{i\mathbf{z}}{M} \right\}, \quad i = 0, \dots, M - 1,$$

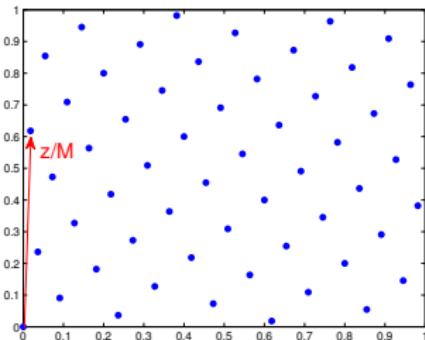
where $\{\cdot\}$ denotes the fractional part (of each component of the vector).

The lattice rule quality depends heavily on the choice of the vector \mathbf{z} .

In 2D a good choice is $M = \text{fibonacci}(k)$ and $\mathbf{z} = (1, \text{fibonacci}(k - 1))$.



$M = 13, \mathbf{z} = (1, 8)$



$M = 55, \mathbf{z} = (1, 34)$

Lattice rules

In higher dimension, there is no obvious generalization of the Fibonacci lattice. Let $\gcd(x, y)$ be the *greatest common divisor* of x and y .

Each component z_i of the vector \mathbf{z} should be looked for in the set

$$\mathbb{U}_M := \{z \in \mathbb{Z} : 1 \leq z \leq M - 1, \gcd(z, M) = 1\}$$

so that each one-dimensional projection of the lattice has M distinct points. If M is prime, there are $|\mathbb{U}_M| = M - 1$ possible choices for each component z_i .

Component-by-Component (CBC) construction.

- ▶ Set $z_1 = 1$
- ▶ for $i = 2, \dots, N$
 - ▶ with z_1, \dots, z_{i-1} held fixed, choose $z_i \in \mathbb{U}_M$ to minimize a desired error criterion in dimension i .

end

Randomized QMC: randomly shifted lattice rules

To estimate the error in a Quasi Monte Carlo computation in practice, the following strategy can be adopted. Let $\boldsymbol{\eta} \sim U([0, 1]^N)$ be a uniformly distributed random vector (shift) and

$$I_{N,M}(f, \boldsymbol{\eta}) = \frac{1}{M} \sum_{i=1}^M f(\{\boldsymbol{\xi}_i + \boldsymbol{\eta}\})$$

the QMC formula with the random shift $\boldsymbol{\eta}$.

Observe that the shift preserves the lattice structure and $\mathbb{E}[f(\{\boldsymbol{\xi}_i + \boldsymbol{\eta}\})] = I_N(f)$, for all $i = 1 \dots, M$.

Since we no longer have a bias error, we need to understand the resulting variance in the randomly shifted rules.

Take s iid shifts $\eta_1, \dots, \eta_s \sim U([0, 1]^N)$. Then

- ▶ Estimate of the integral

$$I_{N,M}(f) = \frac{1}{s} \sum_{j=1}^s I_{N,M}(f, \eta_j) = \frac{1}{sM} \sum_{j=1}^s \sum_{i=1}^M f(\{\xi_i + \eta_j\})$$

- ▶ Estimate of the (randomized) QMC error based on the η -sample standard deviation, (observe that there is no bias any more!) i.e.

$$e_{N,M}(f) = \frac{c_o}{\sqrt{s}} \sqrt{\frac{1}{(s-1)} \sum_{j=1}^s (I_{N,M}(f, \eta_j) - I_{N,M}(f))^2}$$

The analysis of randomized QMC goes through the reinterpretation of the induced Sobolev space as a *Reproducing Kernel Hilbert Space* (RKHS) and a representation of the worst mean square error in terms of the Kernel function; see [Dick-Kuo-Sloan '13] for details.

Quasi Monte Carlo Methods: infinite dimensional
approximation

Infinite dimensional approximation

We have seen that the QMC error can be bounded by

$$\left| \int_{[0,1]^N} f(\mathbf{y}) d\mathbf{y} - \frac{1}{M} \sum_{i=1}^M f(\boldsymbol{\xi}_i) \right| \leq \|\Delta_P\|_{1,q} \|f\|_{1,p} \quad (79)$$

Moreover, good low discrepancy sequences (Sobol, Faure, lattice rules, ...) have a star-discrepancy $\Delta_{P,N}^* = \|\Delta_P\|_{1,\infty} = O(\log(M)^N/M)$.

Is there hope for an infinite dimensional approximation $N \rightarrow \infty$?

Note: if all variables have the same importance in (79), we cannot take the limit $N \rightarrow \infty$.

However, infinite dimensional approximation can be achieved if the function f belongs to a **weighted space**. Let $\{\gamma_\alpha\}_{\alpha \in [1:N]}$ be a sequence of positive weights:

- ▶ Anchored weighted Sobolev space

$$\|f\|_{1,2,\gamma} = \left(\sum_{\alpha \in [1:N]} \frac{1}{\gamma_\alpha} \int_{[0,1]^{\#(\alpha)}} \left| \frac{\partial^{\#(\alpha)} f}{\partial \mathbf{y}_\alpha} (\mathbf{y}_\alpha, \mathbf{c}) \right|^2 d\mathbf{y}_\alpha \right)^{\frac{1}{2}}$$

- ▶ Analogous definition for the unanchored version.

We immediately obtain the error representation in the weighted spaces as

$$\left| \int_{[0,1]^N} f(\mathbf{y}) d\mathbf{y} - \frac{1}{M} \sum_{i=1}^M f(\boldsymbol{\xi}_i) \right| \leq \|\Delta_P\|_{1,2,\frac{1}{\gamma}} \|f\|_{1,2,\gamma},$$

from Zaremba's and Hlawka's identity.

The weights γ should be chosen such that

- ▶ $\|f\|_{1,2,\gamma} < \infty$
- ▶ $\|\Delta_P\|_{1,2,\frac{1}{\gamma}}$ as small as possible

How can/should we choose the weights in practice?

Infinite dimensional approximation in weighted space

► Product weights [Sloan-Woźniakowski]

$$\gamma_{\alpha} = \prod_{i \in \alpha} \gamma_i, \quad \sum_{i=1}^{\infty} \gamma_i < +\infty$$

Each variable is weighted differently. The intuition behind is that the variables y_i for $i \rightarrow \infty$ become “unimportant” in the computation of the integral.

► Order dependent weights [Dick et al. '06]

$$\gamma_{\alpha} = \beta_{|\alpha|}, \quad \text{with } \beta_n \xrightarrow{n \rightarrow \infty} 0$$

Here we think that the terms involving many variables at the same time become “unimportant” in the computation of the integral as the number of involved variables increases.

► Product and Order Dependent (POD) weights [Kuo et al. '12]

$$\gamma_{\alpha} = \beta_{|\alpha|} \prod_{i \in \alpha} \gamma_i, \quad \text{with } \beta_n \xrightarrow{n \rightarrow \infty} 0$$

Efficient constructions of points by the CBC (Component-by-component) techniques corresponding to γ weights have been proposed in [Kuo-Joe, 03].

An asymptotic convergence result for QMC

Theorem 11.4 (asymp. QMC error [Dick-Kuo-Sloan '13, Thm. 5.10])

The error given by a shifted averaged lattice rule with CBC optimal construction of the generating vector \mathbf{z} satisfies

$$e_{N,M}(f) = O\left(\frac{\log \log M}{M}\right)^{\frac{1}{2\lambda}}, \quad \lambda \in \left(\frac{1}{2}, 1\right]$$

provided the function f belongs to the (anchored or unanchored) weighted space with weights γ_α such that

$$\sum_{\emptyset \neq \alpha \in [1:N]} \gamma_\alpha^\lambda \left(\frac{2\zeta(2\lambda)}{(2\pi^2)^\lambda} + \beta^\lambda \right)^{\#(\alpha)} < +\infty$$

where $\zeta(\cdot)$ is the Riemann zeta function (observe that $\zeta(2\lambda) \rightarrow \infty$ as $\lambda \rightarrow \frac{1}{2}$) and $\beta = 0$ for the unanchored space and $\beta = c^2 - c + \frac{1}{3}$ for the anchored space.

This result shows that a limit rate $\sim M^{-1}$ is possible even in infinite dimension if the integrand function belongs to a suitable weighted space, with sufficiently fast decaying weights.

Example: elliptic PDE with random diffusivity coefficient

We consider once again the model problem

$$\begin{cases} -\operatorname{div}(a(x, \mathbf{y}) \nabla u(x, \mathbf{y})) = f(x), & x \in D, \\ u(x, \mathbf{y}) = 0, & x \in \partial D, \end{cases} \quad \forall \mathbf{y} \in \Gamma := [-\sqrt{3}, \sqrt{3}]^N$$

with $a(x, \mathbf{y}) = \bar{a} + \sum_{i=1}^N \sqrt{\lambda_i} y_i b_i(x)$ (here, N could be ∞),

$y_i \sim \mathcal{U}(-\sqrt{3}, \sqrt{3})$ i.i.d., $\sum_{i=1}^N \sqrt{3\lambda_i} \|b_i\|_\infty \leq \delta \bar{a}$ for some $0 < \delta < 1$, so that $\|\nabla u(\mathbf{y})\|_{L^2(D)} \leq C_u := \frac{C_p}{(1-\delta)\bar{a}} \|f\|_{L^2(D)}$.

Moreover, setting $\beta_i = \frac{\sqrt{\lambda_i} \|b_i\|_\infty}{(1-\delta)\bar{a}}$, we assume there exists $0 < \bar{p} \leq 1$ s.t.

$$\sum_{i=1}^N \beta_i^p \leq C, \quad \forall \bar{p} \leq p \leq 1, \quad \text{with } C \text{ independent of } N \quad (80)$$

Consider a linear functional $Q(u)$: will the random variable $\psi(\mathbf{y}) = Q(u(\mathbf{y}))$ belong to a certain weighted space $W_{1,2,\gamma}$ for some suitable choice of weights γ_α , that is, is it true that

$$\|\psi\|_{1,2,\gamma} = \left(\sum_{\alpha \in [1:N]} \frac{1}{\gamma_\alpha} \int_{[-\sqrt{3}, \sqrt{3}]^{\#(\alpha)}} \left| \frac{\partial^{\#(\alpha)}}{\partial \mathbf{y}_\alpha} \psi(\mathbf{y}_\alpha, \mathbf{0}) \right|^2 \frac{1}{(2\sqrt{3})^{\#(\alpha)}} d\mathbf{y}_\alpha \right)^{\frac{1}{2}} < \infty ?$$

Observe that

$$\frac{\partial^{\#(\alpha)} \psi(\mathbf{y})}{\partial \mathbf{y}_\alpha} = Q \left(\frac{\partial^{\#(\alpha)} u(\mathbf{y})}{\partial \mathbf{y}_\alpha} \right) \leq \|Q\|_{V'} \left\| \frac{\partial^{\#(\alpha)} u(\mathbf{y})}{\partial \mathbf{y}_\alpha} \right\|_V$$

so we have to check if u belongs to the space $W_{1,2,\gamma}(\Gamma; V)$, where $V = H_0^1(D)$ which we endow with the norm $\|v\|_V = \|\nabla v\|_{L^2(D)}$.

That is, we need sufficient regularity of the PDE solution wrt the stochastic variable.

Procedure: We (formally) differentiate the equation with respect to y_i at $\mathbf{y} = 0$:

i) first derivative ∂_{y_i} . For all $x \in D$ we have

$$-\operatorname{div}(a(x, \mathbf{0}) \nabla \partial_{y_i} u(x, \mathbf{0})) = \operatorname{div}(\partial_{y_i} a(x, \mathbf{0}) \nabla u(x, \mathbf{0})) = \operatorname{div}(\sqrt{\lambda_i} b_i(x) \nabla u(x, \mathbf{0})).$$

Multiplying by $\partial_{y_i} u(x, \mathbf{0})$ and integrating by parts in x yields

$$(1 - \delta) \bar{a} \|\nabla \partial_{y_i} u(\cdot, \mathbf{0})\|_{L^2(D)}^2 \leq \sqrt{\lambda_i} \|b_i\|_\infty \|\nabla u(\cdot, \mathbf{0})\|_{L^2(D)} \|\nabla \partial_{y_i} u(\cdot, \mathbf{0})\|_{L^2(D)}$$

so that $\|\partial_{y_i} u(\cdot, \mathbf{0})\|_V \leq \beta_i C_u$

ii) second mixed derivative $\partial_{y_i y_j}$. For all $x \in D$ we have

$$\begin{aligned} -\operatorname{div}(a(x, \mathbf{0}) \nabla \partial_{y_i y_j}^2 u(x, \mathbf{0})) &= \operatorname{div}(\sqrt{\lambda_j} b_j(x) \nabla \partial_{y_i} u(x, \mathbf{0})) \\ &\quad + \operatorname{div}(\sqrt{\lambda_i} b_i(x) \nabla \partial_{y_j} u(x, \mathbf{0})). \end{aligned}$$

so that

$$\|\partial_{y_i y_j}^2 u(\cdot, \mathbf{0})\|_V \leq \beta_j \|\partial_{y_i} u(\cdot, \mathbf{0})\|_V + \beta_i \|\partial_{y_j} u(\cdot, \mathbf{0})\|_V \leq 2\beta_i\beta_j C_u$$

Iterating the procedure, yields that for any $\alpha \in [1 : N]$ we have

$$\|\partial_{y_\alpha}^{\#(\alpha)} u(\cdot, \mathbf{0})\|_V \leq C_u \#(\alpha)! \prod_{j \in \alpha} \beta_j.$$

Furthermore, following [Kuo-Schwab-Sloan '11], this in turn implies that

$$u \in W_{1,2,\gamma}(\Gamma; V) \text{ with weights: } \gamma_\alpha = \left(\#(\alpha)! \prod_{j \in \alpha} \beta_j \right)^{2-p} \quad (*)$$

with p as in (80).

Proof of (*)

$$\begin{aligned}
\|u\|_{W_{1,2,\gamma}(\Gamma;V)}^2 &\leq \sum_{\alpha \in [1:N]} \frac{1}{\gamma_\alpha} C_u^2 (\#(\alpha)!)^2 \prod_{j \in \alpha} \beta_j^2 = C_u^2 \sum_{\alpha \in [1:N]} (\#(\alpha)!)^p \prod_{j \in \alpha} \beta_j^p \\
&\leq C_u^2 \sum_{\alpha \in [1:N]} (\#(\alpha)!)^p \prod_{j \in \alpha} \beta_j^{p^2} \prod_{j \in \alpha} \beta_j^{p(1-p)} \quad [\text{next: Hölder with } \frac{1}{p}, \frac{1}{1-p}] \\
&\leq C_u^2 \left(\sum_{\alpha \in [1:N]} (\#(\alpha)! \prod_{j \in \alpha} \beta_j^p)^p \right)^{\frac{1}{p}} \left(\sum_{\alpha \in [1:N]} \prod_{j \in \alpha} \beta_j^p \right)^{1-p} \\
&\leq C_u^2 \left(\sum_{\tilde{\alpha} \in \mathbb{N}^N} \frac{|\tilde{\alpha}|!}{\tilde{\alpha}!} \prod_j \beta_j^{p\tilde{\alpha}_j} \right)^p \left(\sum_{\alpha \in [1:N]} \prod_{j \in \alpha} \beta_j^p \right)^{1-p} \\
&= C_u^2 \left(\sum_j \beta_j^p \right)^p \prod_j (1 + \beta_j^p)^{1-p} < +\infty
\end{aligned}$$

□

Finally, for this choice of weights, the quantity

$$\sum_{\emptyset \neq \alpha \in [1:N]} \gamma_\alpha^\lambda \left(\frac{2\zeta(2\lambda)}{(2\pi^2)^\lambda} + \beta^\lambda \right)^{\#(\alpha)}$$

in the asymptotic QMC convergence Thm. 11.4 is bounded for any $\lambda = \max\{\frac{p}{2-p}, \frac{1}{2-2\delta}\}$, $\delta > 0$ (here $\beta = 1/12$ and with p as in (80)).

Proof: Indeed, set $\sigma(\lambda) = \frac{2\zeta(2\lambda)}{(2\pi^2)^\lambda} + 12^{-\lambda}$. Then, since $p \leq (2-p)\lambda \leq 1$

$$\begin{aligned} \sum_{\alpha \in [1:N]} \gamma_\alpha^\lambda \sigma(\lambda)^{\#(\alpha)} &= \sum_{\alpha \in [1:N]} (\#(\alpha)!)^{(2-p)\lambda} \prod_{j \in \alpha} \sigma(\lambda) \beta_j^{(2-p)\lambda} \\ &\leq \sum_{\alpha \in [1:N]} \#(\alpha)! \prod_{j \in \alpha} \sigma(\lambda) \beta_j^{(2-p)\lambda} \\ &\leq \sigma(\lambda) \exp \left(\sum_j \beta_j^{(2-p)\lambda} \right) < \infty , \quad \text{since } (2-p)\lambda \geq p \end{aligned}$$

□

Finally, we conclude with the general convergence theorem for QMC applied to the random elliptic PDE.

Theorem 11.5 (QMC convergence [Kuo-Schwab-Sloan '11])

The Quasi Monte Carlo method with CBC construction of the lattice rule and weights () converges at rate*

$$O(M^{-1+\delta}) \text{ for } p < \frac{2}{3-2\delta}, \quad O(M^{\frac{1}{2}-\frac{1}{p}}) \text{ for } \frac{2}{3-2\delta} < p \leq 1$$

with p as in (87) and for any $\delta > 0$ (with a constant that blows up when $\delta \rightarrow 0$).

QMC complexity analysis

So far the discussion had been ignored spatial discretizations. Now we also discretize the problem in space by finite elements and assume that

- ▶ $\forall \mathbf{y} \in \Gamma$, the discretization error $|Q(u(\mathbf{y})) - Q(u_h(\mathbf{y}))| \leq C_1 h^\alpha$
- ▶ The computational work to solve for each $u_h(\xi_j)$ is $O(h^{-d\beta})$.

As in Monte Carlo, we can therefore choose M and h to

$$\begin{cases} \min_{h,M} M h^{-d\beta} \\ \text{s.t. } C_1 h^\alpha + C_2 M^{-\gamma} \leq \text{TOL} \end{cases}$$

with $\gamma = 1 - \delta$ if $p < \frac{2}{3 - 2\delta}$ and $\gamma = \frac{1}{p} - \frac{1}{2}$ if $\frac{2}{3 - 2\delta} < p \leq 1$.

which leads to a resulting complexity $O(\text{TOL}^{-\frac{1}{\gamma} - \frac{d\beta}{\alpha}})$. Hence,

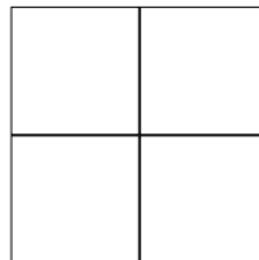
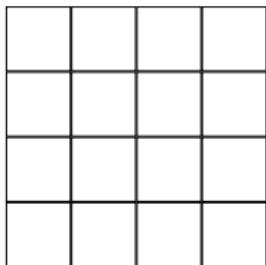
$$W \propto \text{TOL}^{-\frac{1}{1-\delta} - \frac{d\beta}{\alpha}}, \quad \text{for } p < \frac{2}{3 - 2\delta}$$

$$W \propto \text{TOL}^{-\frac{2p}{2-p} - \frac{d\beta}{\alpha}}, \quad \text{for } \frac{2}{3 - 2\delta} < p \leq 1$$

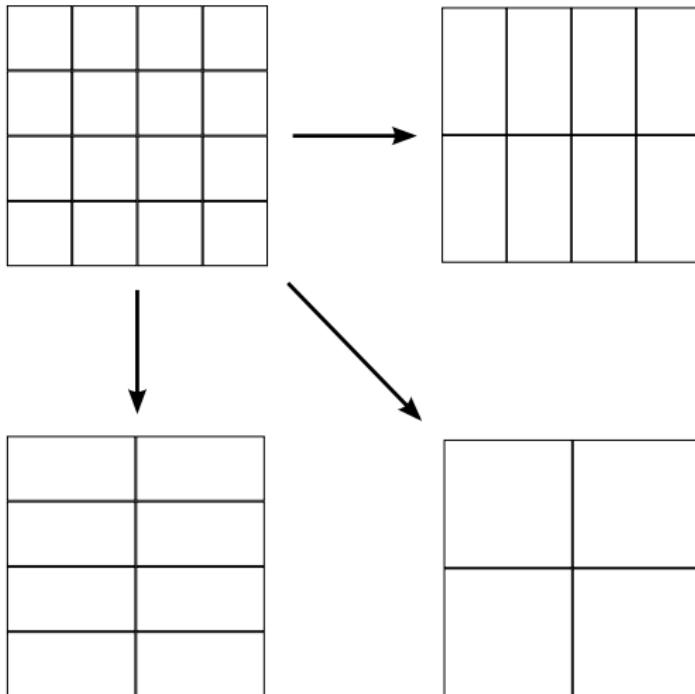
Extensions of Quasi Monte Carlo

- ▶ Multilevel QMC, e.g., [Kuo et al, 2013]
- ▶ Higher order QMC, which aim to achieve a rate of convergence faster than linear,
- ▶ lattice rules that are extensible in number of points or dimension,
- ▶ more general distributions and unbounded domains, etc.

Variance reduction: MLMC



Variance reduction: further potential



MIMC Estimator

Consider discretization parameters possibly different in each direction

$$h_{i,\alpha_i} = h_{i,0} \beta_i^{-\alpha_i}$$

with $\beta_i > 1$. For a multi-index $\alpha = (\alpha_i)_{i=1}^d \in \mathbb{N}^d$, we denote by S_α the approximation of S calculated using a discretization defined by α .

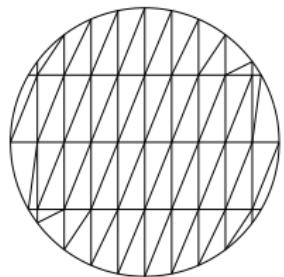
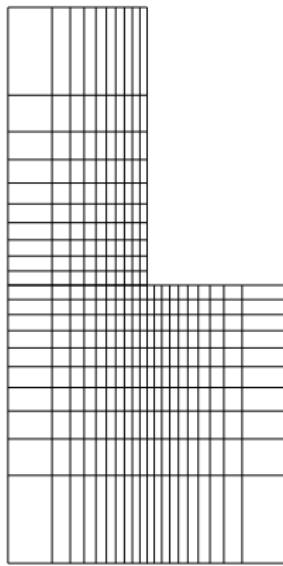
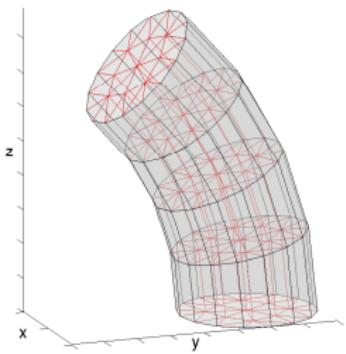
For $i = 1, \dots, d$, define the first order difference operators

$$\Delta_i S_\alpha = \begin{cases} S_\alpha & \text{if } \alpha_i = 0, \\ S_\alpha - S_{\alpha - e_i} & \text{if } \alpha_i > 0, \end{cases}$$

and construct the first order mixed difference

$$\Delta S_\alpha = (\otimes_{i=1}^d \Delta_i) S_\alpha = \sum_{j \in \{0,1\}^d} (-1)^{|j|} S_{\alpha-j}$$

with $|j| = \sum_{i=1}^d j_i$. Requires 2^d evaluations of S on different grids.



Left: Tensor domain, cylinder.

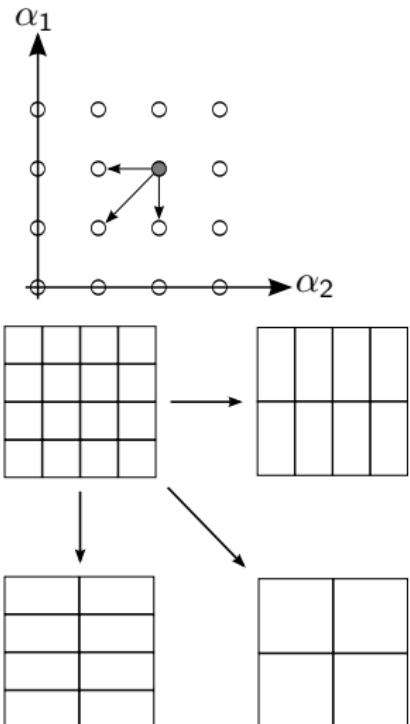
Center: Non-tensor domain immersed in a tensor box.

Right: Non-tensor domain with a structured mesh.

Example: Computing S_α in $d = 2$

For $\alpha = (\alpha_1, \alpha_2)$, we have

$$\begin{aligned}\Delta S_{(\alpha_1, \alpha_2)} &= \Delta_2(\Delta_1 S_{(\alpha_1, \alpha_2)}) \\&= \Delta_2(S_{\alpha_1, \alpha_2} - S_{\alpha_1-1, \alpha_2}) \\&= S_{\alpha_1, \alpha_2} - S_{\alpha_1-1, \alpha_2} \\&\quad - S_{\alpha_1, \alpha_2-1} + S_{\alpha_1-1, \alpha_2-1}.\end{aligned}$$



MIMC Estimator

Then, assuming that

$$E[S_\alpha] \rightarrow E[S] \quad \text{as} \quad \min_{1 \leq i \leq d} \alpha_i \rightarrow \infty,$$

it is not difficult to see that we can represent

$$E[S] = \sum_{\alpha \in \mathbb{N}^d} E[\Delta S_\alpha]$$

and then approximate by truncation

$$E[S] \approx \sum_{\alpha \in \mathcal{I}} E[\Delta S_\alpha]$$

where $\mathcal{I} \subset \mathbb{N}^d$ is a *properly chosen* index set.

As in MLMC, approximating each term by independent MC samplers, the MIMC estimator can be written as

$$\mathcal{A}_{\text{MIMC}} = \sum_{\alpha \in \mathcal{I}} \frac{1}{M_\alpha} \sum_{m=1}^{M_\alpha} \Delta S_\alpha(\omega_{\alpha,m})$$

with *properly chosen* sample sizes $(M_\alpha)_{\alpha \in \mathcal{I}}$.

Example: On mixed differences

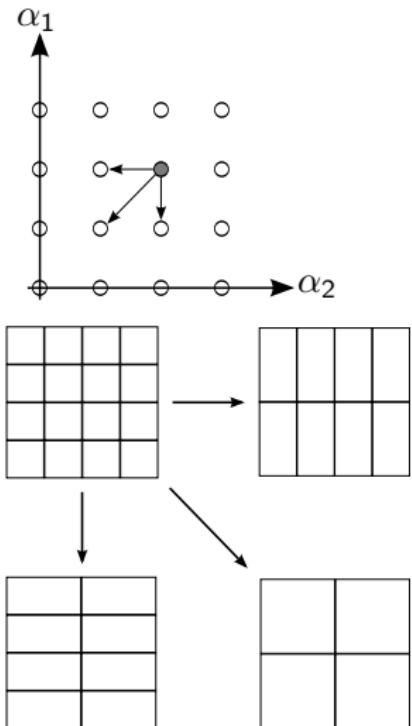
Consider $d = 2$. In this case, letting $\alpha = (\alpha_1, \alpha_2)$, we have

$$\begin{aligned}\Delta S_{(\alpha_1, \alpha_2)} &= \Delta_2(\Delta_1 S_{(\alpha_1, \alpha_2)}) \\ &= \Delta_2(S_{\alpha_1, \alpha_2} - S_{\alpha_1-1, \alpha_2}) \\ &= S_{\alpha_1, \alpha_2} - S_{\alpha_1-1, \alpha_2} \\ &\quad - S_{\alpha_1, \alpha_2-1} + S_{\alpha_1-1, \alpha_2-1}.\end{aligned}$$

Notice that in general, ΔS_α requires 2^d evaluations of S at different discretization parameters, the largest work of which corresponds precisely to the index appearing in ΔS_α , namely α .

More precisely, for $\beta_i = \beta$, and $\gamma_i = \gamma$, $\forall i$,

$$Cost(\Delta S_\alpha) = (1 + \beta^{-\gamma})^d Cost(S_\alpha)$$



Our objective is to build an estimator $\mathcal{A} = \mathcal{A}_{\text{MIMC}}$ where

$$P(|\mathcal{A} - E[\mathcal{A}]| \leq \text{TOL}) \geq 1 - \epsilon \quad (81)$$

for a given accuracy TOL and a given confidence level determined by $0 < \epsilon \ll 1$. We instead impose the following, more restrictive, two constraints:

Bias constraint: $|E[\mathcal{A} - \mathcal{S}]| \leq (1 - \theta)\text{TOL}, \quad (82)$

Statistical constraint: $P(|\mathcal{A} - E[\mathcal{A}]| \leq \theta\text{TOL}) \geq 1 - \epsilon. \quad (83)$

For a given fixed $\theta \in (0, 1)$. Moreover, motivated by the asymptotic normality of the estimator, \mathcal{A} , we approximate (83) by

$$\text{Var}[\mathcal{A}] \leq \left(\frac{\theta\text{TOL}}{C_\epsilon} \right)^2. \quad (84)$$

Here, $0 < C_\epsilon$ is such that $\Phi(C_\epsilon) = 1 - \frac{\epsilon}{2}$, where Φ is the cumulative distribution function of a standard normal random variable.

Given variance and work estimates we can already optimize for the optimal number of samples $M_\alpha^* \in \mathbb{R}$ to satisfy the variance constraint
 (84)

$$M_\alpha^* = \left(\frac{C_\epsilon}{\theta \text{TOL}} \right)^2 \sqrt{\frac{V_\alpha}{W_\alpha}} \left(\sum_{\alpha \in \mathcal{I}} \sqrt{V_\alpha W_\alpha} \right).$$

Taking $M_\alpha^* \leq M_\alpha \leq M_\alpha^* + 1$ such that $M_\alpha \in \mathbb{N}$ and substituting in the total work gives

$$\text{Work}(\mathcal{I}) \leq \left(\frac{C_\epsilon}{\theta \text{TOL}} \right)^2 \left(\sum_{\alpha \in \mathcal{I}} \sqrt{V_\alpha W_\alpha} \right)^2 + \underbrace{\sum_{\alpha \in \mathcal{I}} W_\alpha}_{\text{Min. cost of } \mathcal{I}}.$$

Observe: The upper bound for the work now depends on the index set, \mathcal{I} , only.

Choosing the Index set \mathcal{I} – Tensor Product sets

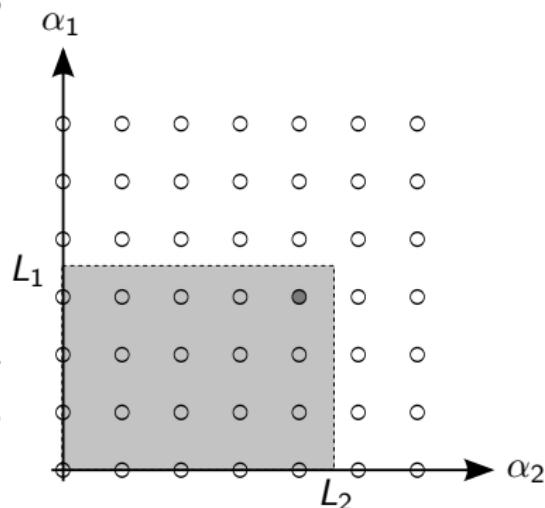
An obvious (although naive) choice of \mathcal{I} is the Full Tensor index-set

$$\mathcal{I}(\mathbf{L}) = \{\boldsymbol{\alpha} \in \mathbb{N}^d : \alpha_i \leq L_i \text{ for } i \in \{1 \cdots d\}\}$$

for some $\mathbf{L} \in \mathbb{R}^d$.

It turns out, unsurprisingly, that Full Tensor (FT) index-sets impose restrictive conditions on the weak rates w_i and yield sub-optimal complexity rates.

Remark: The $\text{Bias} = |\mathbb{E}[S - \sum_{\boldsymbol{\alpha} \in \mathcal{I}_{\mathbf{L}}} S_{\boldsymbol{\alpha}}]| = |\mathbb{E}[S - \mathcal{S}_{\mathbf{L}}]|$ corresponds to the discretization error on a (possibly anisotropic) full tensor grid of level $\mathbf{L} = (L_1, \dots, L_d)$.



MIMC general analysis framework

Question: How do we find optimal index set \mathcal{I} for MIMC?

Denote by $E_\alpha = |\mathbb{E}[\Delta S_\alpha]|$ the bias contributions. We need to solve

$$\min_{\mathcal{I} \subset \mathbb{N}^d} \text{Work}(\mathcal{I}) \quad \text{such that Bias} = \sum_{\alpha \notin \mathcal{I}} E_\alpha \leq (1 - \theta) \text{TOL},$$

Assumption: MIMC work is *not* dominated by the work to compute a single sample corresponding to each $\alpha \in \mathcal{I}$.

Then, minimizing equivalently $\sqrt{\text{Work}(\mathcal{I})} \propto \sum_{\alpha \in \mathcal{I}} \sqrt{V_\alpha W_\alpha}$, the previous min problem can be recast into a knapsack problem with profits defined for each multi-index α .

The corresponding α profit is

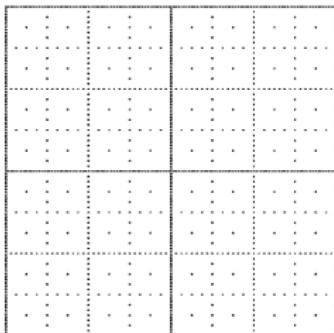
$$\mathcal{P}_\alpha = \frac{\text{Bias contribution}}{\text{Work contribution}} = \frac{E_\alpha}{\sqrt{V_\alpha W_\alpha}}$$

Bias for MIMC

What is the Bias in a MIMC approximation?

$$\text{Bias} = E \left[S - \sum_{\alpha \in \mathcal{I}} S_\alpha \right] = \sum_{\alpha \notin \mathcal{I}} E[S_\alpha]$$

It corresponds to the discretization error on a **sparse grid** by the so called **Combination Technique** [Griebel-Schneider-Zenger '91]



(from [Burgartz-Griebel, Acta Num. '04])

MIMC general analysis framework

Define the total error associated with an index-set \mathcal{I} as

$$\mathfrak{E}(\mathcal{I}) = \sum_{\alpha \notin \mathcal{I}} E_\alpha$$

and the corresponding total work estimate as

$$\mathfrak{W}(\mathcal{I}) = \sum_{\alpha \in \mathcal{I}} \sqrt{V_\alpha W_\alpha}.$$

Then we can show the following optimality result with respect to $\mathfrak{E}(\mathcal{I})$ and $\mathfrak{W}(\mathcal{I})$, namely:

Lemma 12.1 (Optimal profit sets)

The index-set

$$\mathcal{I}(\nu) = \{\alpha \in \mathbb{N}^d : \mathcal{P}_\alpha \geq \nu\}$$

for $\mathcal{P}_\alpha = \frac{E_\alpha}{\sqrt{V_\alpha W_\alpha}}$ is optimal in the sense that any other index-set, $\tilde{\mathcal{I}}$, with smaller work, $\mathfrak{W}(\tilde{\mathcal{I}}) < \mathfrak{W}(\mathcal{I}(\nu))$, leads to a larger error, $\mathfrak{E}(\tilde{\mathcal{I}}) > \mathfrak{E}(\mathcal{I}(\nu))$.

MIMC general analysis framework

Once the shape of \mathcal{I} is determined, we find $\mathcal{I}(\text{TOL})$ to be the minimum set of the family that satisfies the bias constraint

$$\mathfrak{E}(\mathcal{I}(\text{TOL})) = \sum_{\alpha \notin \mathcal{I}(\text{TOL})} E_\alpha \leq (1 - \theta)\text{TOL}$$

yielding the corresponding computational work

$$\left(\frac{C_\epsilon}{\theta \text{TOL}} \right)^2 \left(\sum_{\alpha \in \mathcal{I}(\text{TOL})} \sqrt{V_\alpha W_\alpha} \right)^2 \lesssim \text{TOL}^{-(2+p)}$$

with $p \geq 0$ and possibly some multiplicative log factors in the above estimate. To get sharper complexity results we need particular, problem dependent, assumptions, as we see next.

Particular Assumptions for MIMC

For every α , we assume the following

Assumption 1 (Bias) : $E_\alpha = |\mathbb{E}[\Delta S_\alpha]| \lesssim \prod_{i=1}^d \beta_i^{-\alpha_i w_i}$

Assumption 2 (Variance) : $V_\alpha = \text{Var}[\Delta S_\alpha] \lesssim \prod_{i=1}^d \beta_i^{-\alpha_i s_i}$,

Assumption 3 (Work) : $W_\alpha = \text{Work}(\Delta S_\alpha) \lesssim \prod_{i=1}^d \beta_i^{\alpha_i \gamma_i}$,

For positive constants $\gamma_i, w_i, s_i \leq 2w_i$ and for $i = 1 \dots d$.

$$\text{Work(MIMC)} = \sum_{\alpha \in \mathcal{I}} M_\alpha W_\alpha \lesssim \sum_{\alpha \in \mathcal{I}} M_\alpha \left(\prod_{i=1}^d \beta_i^{\alpha_i \gamma_i} \right).$$

Remark on product rate particular assumptions

- ▶ The **Assumptions 1 & 2** in MIMC that the rates for each α are *products of 1D rates*

$$E_\alpha \propto \prod_{i=1}^d \beta_i^{-\alpha_i w_i}, \quad V_\alpha \propto \prod_{i=1}^d \beta_i^{-\alpha_i s_i}$$

are stronger than the corresponding assumptions in MLMC.

- ▶ They imply existence of **mixed derivatives** of the solution of the PDE (and possibly the solution of the adjoint problem associated to the functional Ψ), as opposed to standard derivatives for MLMC.

Defining the optimal index-set for MIMC

Under **Assumptions 1-3** (assuming that they hold as sharp estimates and not just as upper bounds) we have

$$\mathcal{P}_{\boldsymbol{\alpha}} = \prod_{i=1}^d \beta_i^{-(w_i + \frac{\gamma_i - s_i}{2})\alpha_i} = e^{-\sum_{i=1}^d \log(\beta_i)(w_i + \frac{\gamma_i - s_i}{2})\alpha_i} = e^{-C_{\delta} \sum_{i=1}^d \delta_i \alpha_i}$$

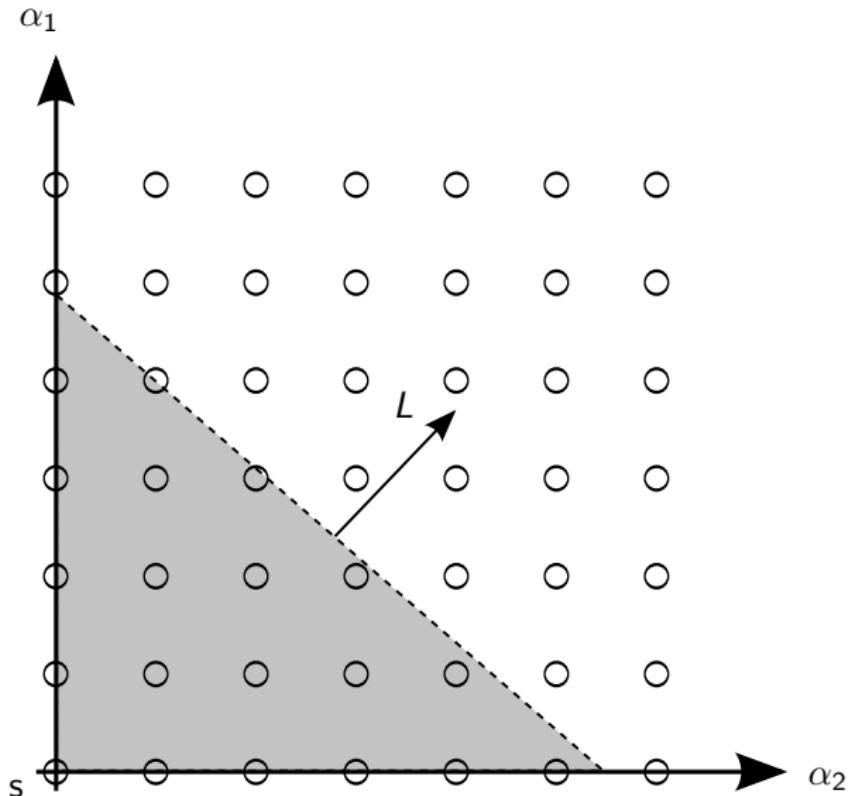
$$\text{with } \delta_i = \frac{\log(\beta_i)(w_i + \frac{\gamma_i - s_i}{2})}{C_{\delta}}, \quad \text{and} \quad C_{\delta} = \sum_{j=1}^d \log(\beta_j)(w_j + \frac{\gamma_j - s_j}{2}).$$

Observe that $0 < \delta_i \leq 1$, since $s_i \leq 2w_i$ and $\gamma_i > 0$. Moreover,

$$\sum_{i=1}^d \delta_i = 1.$$

Then, for any $L \in \mathbb{R}$, the optimal index-set can be written as

$$\mathcal{I}_{\boldsymbol{\delta}}(L) = \{\boldsymbol{\alpha} \in \mathbb{N}^d : \boldsymbol{\alpha} \cdot \boldsymbol{\delta} = \sum_{i=1}^d \alpha_i \delta_i \leq L\}. \quad (85)$$



MIMC work estimate

$$\eta = \min_{i \in \{1 \cdots d\}} \frac{\log(\beta_i) w_i}{\delta_i}, \quad \zeta = \max_{i \in \{1 \cdots d\}} \frac{\gamma_i - s_i}{2w_i}, \quad \mathfrak{z} = \#\{i \in \{1 \cdots d\} : \frac{\gamma_i - s_i}{2w_i} = \zeta\}.$$

Theorem 12.2 (Work estimate with optimal weights)

Let the total-degree index set $\mathcal{I}_\delta(L)$ be given by (85), taking

$$L = \frac{1}{\eta} \left(\log(\text{TOL}^{-1}) + (\mathfrak{z} - 1) \log \left(\frac{1}{\eta} \log(\text{TOL}^{-1}) \right) + C \right).$$

Under Assumptions 1-3, the bias constraint in (82) is satisfied asymptotically and the total work, $W(\mathcal{I}_\delta)$, of the MIMC estimator, \mathcal{A} , subject to the variance constraint (84) satisfies:

$$\limsup_{\text{TOL} \downarrow 0} \frac{W(\mathcal{I}_\delta)}{\text{TOL}^{-2-2\max(0,\zeta)} (\log(\text{TOL}^{-1}))^{\mathfrak{p}}} < \infty,$$

where $0 \leq \mathfrak{p} \leq 3d + 2(d-1)\zeta$ is known and depends on $d, \gamma, \mathbf{w}, \mathbf{s}$ and β .

Powers of the logarithmic term

$$\xi = \min_{i \in \{1 \cdots d\}} \frac{2w_i - s_i}{\gamma_i}, \quad d_2 = \#\{i \in \{1 \cdots d\} : \gamma_i = s_i\},$$
$$\zeta = \max_{i \in \{1 \cdots d\}} \frac{\gamma_i - s_i}{2w_i}, \quad \mathfrak{z} = \#\{i \in \{1 \cdots d\} : \frac{\gamma_i - s_i}{2w_i} = \zeta\}.$$

Cases for \mathfrak{p} :

- A) if $\zeta \leq 0$ and $\zeta < \xi$,
or $\zeta = \xi = 0$ then $\mathfrak{p} = 2d_2$.
- B) if $\zeta > 0$ and $\xi > 0$ then $\mathfrak{p} = 2(\mathfrak{z} - 1)(\zeta + 1)$.
- C-D) if $\zeta \geq 0$ and $\xi = 0$ then $\mathfrak{p} = d - 1 + 2(\mathfrak{z} - 1)(\zeta + 1)$.

Fully Isotropic Case: Rough noise case

Assume $w_i = w$, $s_i = s < 2w$, $\beta_i = \beta$ and $\gamma_i = \gamma$ for all $i \in \{1 \cdots d\}$.
Then the optimal work is

$$\text{Work(MC)} = \mathcal{O}\left(\text{TOL}^{-2 - \frac{d\gamma}{w}}\right).$$

$$\text{Work(MLMC)} = \begin{cases} \mathcal{O}(\text{TOL}^{-2}), & s > d\gamma, \\ \mathcal{O}\left(\text{TOL}^{-2} (\log(\text{TOL}^{-1}))^2\right), & s = d\gamma, \\ \mathcal{O}\left(\text{TOL}^{-\left(2 + \frac{d\gamma-s}{w}\right)}\right), & s < d\gamma. \end{cases}$$

$$\text{Work(MIMC)} = \begin{cases} \mathcal{O}(\text{TOL}^{-2}), & s > \gamma, \\ \mathcal{O}\left(\text{TOL}^{-2} (\log(\text{TOL}^{-1}))^{2d}\right), & s = \gamma, \\ \mathcal{O}\left(\text{TOL}^{-\left(2 + \frac{\gamma-s}{w}\right)} \log(\text{TOL}^{-1})^{(d-1)\frac{\gamma-s}{w}}\right), & s < \gamma. \end{cases}$$

Fully Isotropic Case: Smooth noise case

Assume $w_i = w$, $s_i = 2w$, $\beta_i = \beta$ and $\gamma_i = \gamma$ for all $i \in \{1 \cdots d\}$ and $d \geq 3$. Then the optimal work is

$$\text{Work(MC)} = \mathcal{O}\left(\text{TOL}^{-2 - \frac{d\gamma}{w}}\right).$$

$$\text{Work(MLMC)} = \begin{cases} \mathcal{O}(\text{TOL}^{-2}), & 2w > d\gamma, \\ \mathcal{O}\left(\text{TOL}^{-2} (\log(\text{TOL}^{-1}))^2\right), & 2w = d\gamma, \\ \mathcal{O}\left(\text{TOL}^{-\frac{d\gamma}{w}}\right), & 2w < d\gamma. \end{cases}$$

$$\text{Work(MIMC)} = \begin{cases} \mathcal{O}(\text{TOL}^{-2}), & 2w > \gamma, \\ \mathcal{O}\left(\text{TOL}^{-2} (\log(\text{TOL}^{-1}))^{3(d-1)}\right), & 2w = \gamma, \\ \mathcal{O}\left(\text{TOL}^{-\frac{\gamma}{w}} (\log(\text{TOL}^{-1}))^{(d-1)(1+\gamma/w)}\right), & 2w < \gamma, \end{cases}$$

Up to a multiplicative logarithmic term, Work(MIMC) is the same as solving just a **one dimensional** deterministic problem.

MIMC: Case with a single worst direction

Recall $\zeta = \max_{i \in \{1 \dots d\}} \frac{\gamma_i - s_i}{2w_i}$ and $\mathfrak{z} = \#\{i \in \{1 \dots d\} : \frac{\gamma_i - s_i}{2w_i} = \zeta\}$.

In the special case when $\zeta > 0$ and $\mathfrak{z} = 1$, i.e. when the directions are dominated by a single “worst” direction with the maximum difference between the work rate and the rate of variance convergence. In this case, the value of L becomes

$$L = \frac{1}{\eta} (\log(\text{TOL}^{-1}) + \log(C))$$

and MIMC with a TD index-set achieves a better rate for the computational complexity, namely $\mathcal{O}(\text{TOL}^{2-2\zeta})$. In other words, the logarithmic term disappears from the computational complexity.

Observe: TD-MIMC with a single worst direction has the same rate of computational complexity as a **one-dimensional** MLMC along that single direction.

Problem description

We test our methods on a three-dimensional, linear elliptic PDE with variable, smooth, stochastic coefficients. The problem is isotropic and we have

$$\gamma_i = 2,$$

$$w_i = 2,$$

§ and

$$s_i = 4$$

as $\text{TOL} \rightarrow 0$.

Problem description

$$\begin{aligned}-\nabla \cdot (a(x; \omega) \nabla u(x; \omega)) &= 1 && \text{for } x \in (0, 1)^3, \\ u(x; \omega) &= 0 && \text{for } x \in \partial(0, 1)^3,\end{aligned}$$

$$\text{where } a(x; \omega) = 1 + \exp \left(2Y_1 \Phi_{121}(x) + 2Y_2 \Phi_{877}(x) \right).$$

Here, Y_1 and Y_2 are i.i.d. uniform random variables in the range $[-1, 1]$. We also take

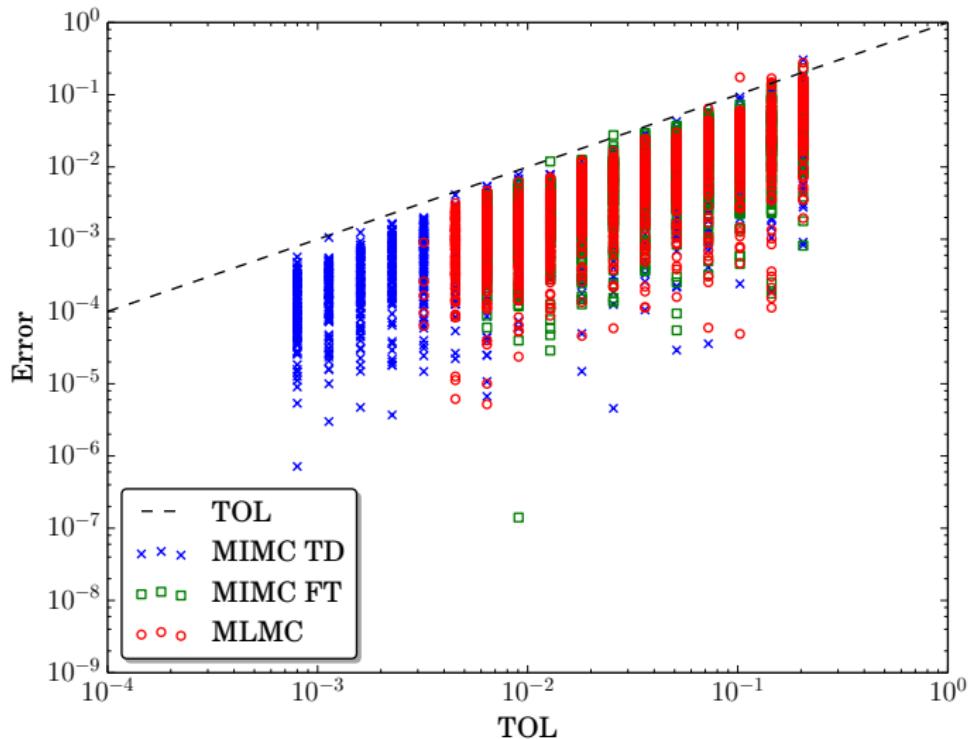
$$\begin{aligned}\Phi_{ijk}(x) &= \phi_i(x_1) \phi_j(x_2) \phi_k(x_3), \\ \text{and } \phi_i(x) &= \begin{cases} \cos \left(\frac{i}{2} \pi x \right) & i \text{ is even,} \\ \sin \left(\frac{i+1}{2} \pi x \right) & i \text{ is odd,} \end{cases}\end{aligned}$$

Finally, the quantity of interest, S , is

$$S = 100 \left(2\pi\sigma^2 \right)^{\frac{-3}{2}} \int_{\mathcal{D}} \exp \left(-\frac{\|x - x_0\|_2^2}{2\sigma^2} \right) u(x) dx,$$

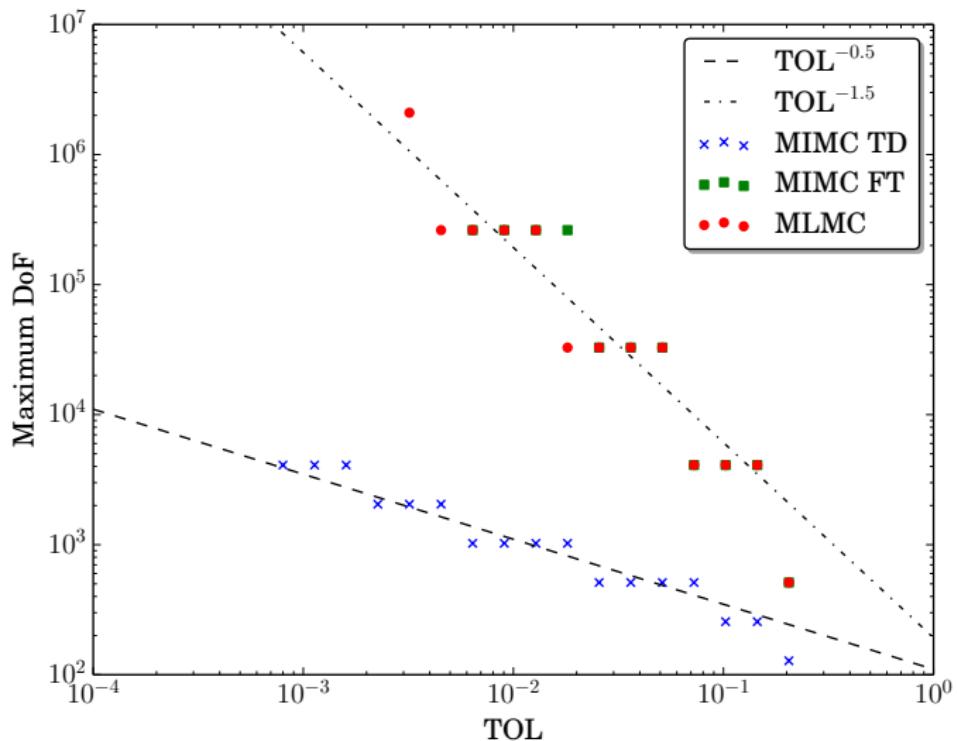
and the selected parameters are $\sigma = 0.04$ and $x_0 = [0.5, 0.2, 0.6]$.

Numerical test: Computational Errors



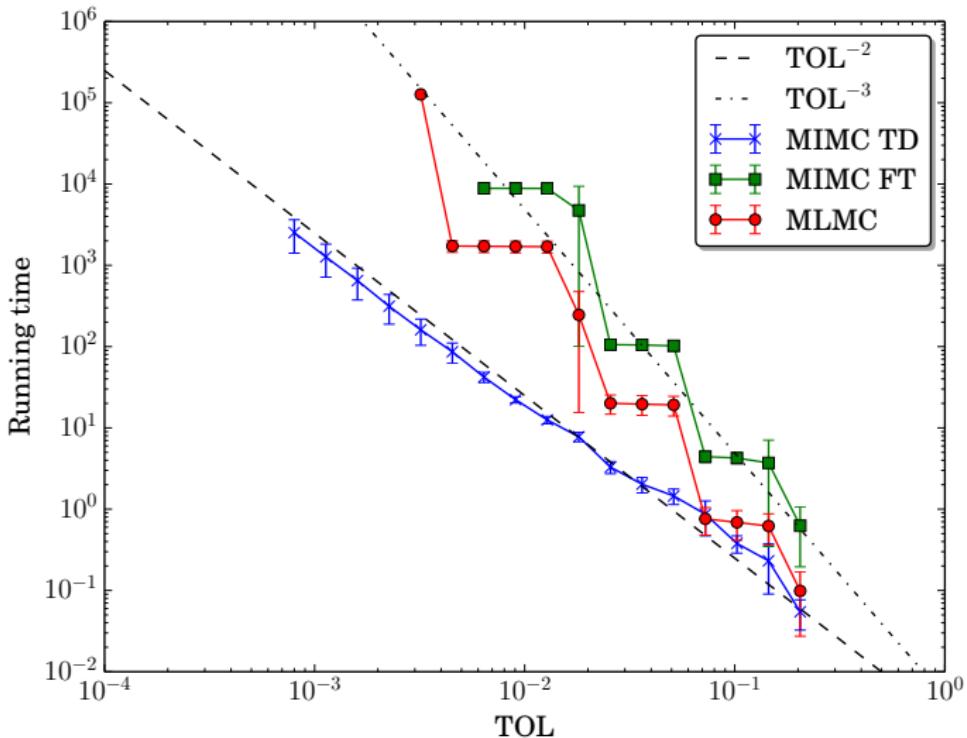
Several runs
for different
TOL values.
Error is
satisfied in
probability but
not
over-killed.

Numerical test: Maximum degrees of freedom



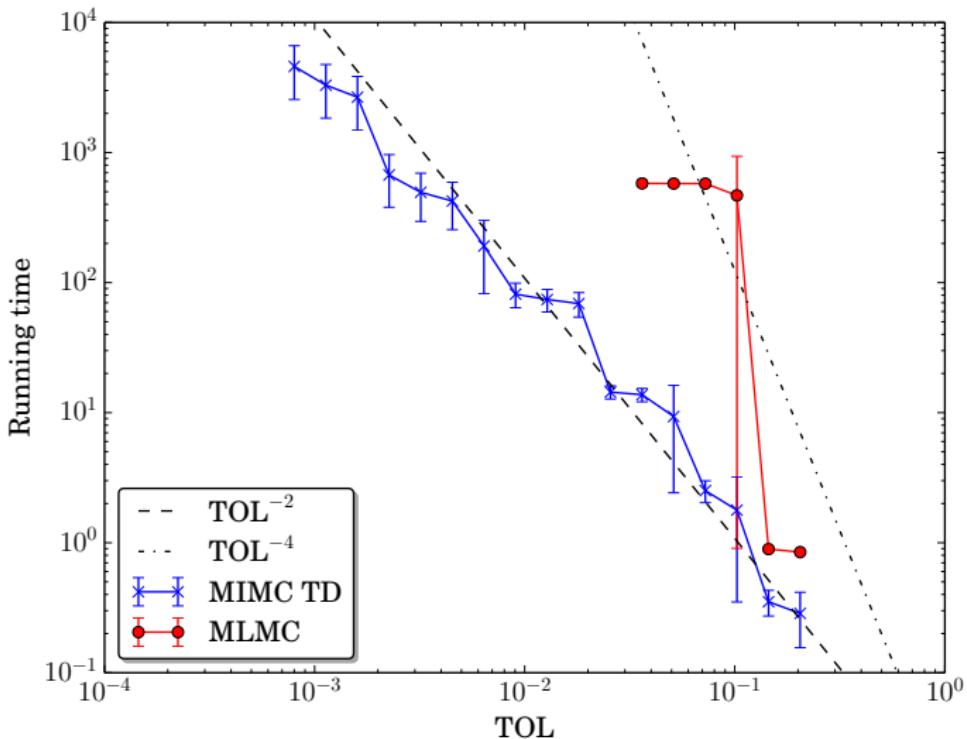
Maximum number of degrees of freedom of a sample PDE solve for different TOL values. This is an indication of required memory.

Numerical test: Running time, 3D problem



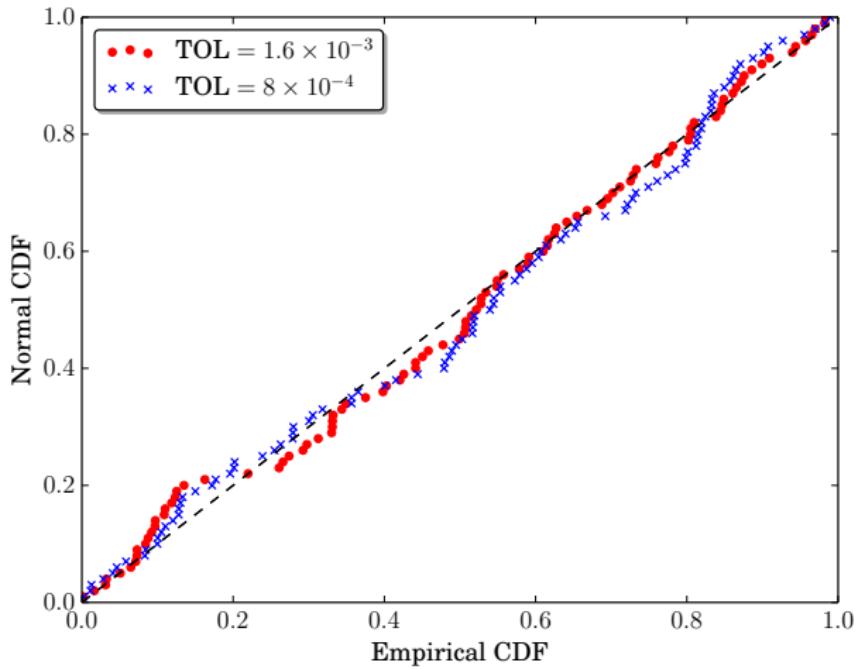
Recall that
the work
complexity of
MC is
 $\mathcal{O}(\text{TOL}^{-5})$

Numerical test: Running time, 4D problem



A similar PDE problem with $d=4$.
The work complexity of MC is $\mathcal{O}(TOL^{-6})$

Numerical test: QQ-plot



Numerical verification of asymptotic normality of the MIMC estimator. A corresponding statement and proof of the normality of an MIMC estimator can be found in (Haji-Ali et al. 2014).

MIMC Conclusions and Extra Points

- ▶ MIMC is a generalization of MLMC and performs better, especially in higher dimensions, provided mixed regularity between discretization parameters.
- ▶ MIMC general analysis framework, identifying optimal index-set through profit thresholding. Each *particular* set of regularity assumptions yield its optimal index-set and related complexity.
- ▶ A MIMC direction does not have to be a spatial dimension. It can represent any form of discretization parameter!
Example: 1-DIM Stochastic Particle Systems, MIMC brings complexity down from $\mathcal{O}(\text{TOL}^{-4})$ to $\mathcal{O}(\text{TOL}^{-2} \log (\text{TOL}^{-1})^2)$.
"A study of Monte Carlo methods for weak approximations of stochastic particle systems in the mean-field", by A. L. Haji Ali and R. T. May 2016.
- ▶ Observe, connection to Ensemble Kalman Filter (EnKF): ML-MIMC can compute other statistics, for instance the covariance.
 - "Multi-index ensemble Kalman filtering", by H. Hoel, G. Shaimerdenova and R. Tempone. arXiv:2104.07263, April 2021.

Conclusions and Extra Points

Recall that assuming

$$\mathbb{E}[S_\alpha] \rightarrow \mathbb{E}[S] \quad \text{as} \quad \alpha_i \rightarrow \infty \quad \text{for all } i = 1, \dots, d,$$

we approximated

$$\mathbb{E}[S] = \sum_{\alpha \in \mathbb{N}^d} \mathbb{E}[\Delta S_\alpha] \approx \sum_{\alpha \in \mathcal{I}} \mathbb{E}[\Delta S_\alpha]$$

where $\mathcal{I} \subset \mathbb{N}^d$ is a *properly chosen* index set.

Idea: Use **Other Quadratures used in the approximation of the Hierarchical differences $\mathbb{E}[\Delta S_\alpha]$.**

- ▶ Quasi Monte Carlo sampling, yielding Multilevel Quasi Monte Carlo (MLQMC) and Multiindex Quasi Monte Carlo (MIQMC).
- ▶ Sparse grids, yielding Multilevel Stochastic Collocation (MLSC) and Multiindex Stochastic Collocation (MISC).

Observe: The choice of the optimal index set $\mathcal{I}(\text{TOL})$ depends on the quadrature type!

Conclusions and Extra Points: Hierarchical Unbiased methods

Idea: Use randomization and importance sampling over levels, yielding unbiased versions of Multilevel and Multiindex Monte Carlo. (McLeish 2011, Rhee & Glynn 2012, 2015)

Suppose that p_ℓ is a probability mass on \mathbb{N}^+ , such that $p_\ell > 0$ for all $\ell \geq 0$. Let

$$R \sim p_\ell$$

be a positive random integer independent of the differences ΔS_ℓ . Define

$$Z = \frac{\Delta S_R}{p_R} \text{ (single-term estimator)}$$

Then we have

$$E[Z] = \sum_{\ell \geq 0} E[Z|R=\ell]P(R=\ell) = \sum_{\ell \geq 0} E[\Delta S_\ell] = E[S] \text{ (Unbiased!)}$$

Conclusions and Extra Points: Hierarchical Unbiased methods

Now pick the pmf sequence $p_\ell > 0$, with $\sum_{\ell \geq 0} p_\ell = 1$, to minimize the expected computational work subject to a variance upperbound:

$$\text{Var}[Z] \leq \sum_{\ell \geq 0} E[Z^2|R = \ell]P(R = \ell) = \sum_{\ell \geq 0} \frac{E[(\Delta S_\ell)^2]}{p_\ell}$$

The expected computational work **per sample** is

$$E[Work(Z)] = \sum_{\ell \geq 0} E[Work(Z)|R = \ell]P(R = \ell) = \sum_{\ell \geq 0} Work_\ell p_\ell$$

The optimal importance sampling probabilities then become, for some constant λ to be determined and $V_\ell = E[(\Delta S_\ell)^2]$,

$$p_\ell^* \propto \sqrt{\frac{V_\ell}{Work_\ell + \lambda}}. \quad (86)$$

Conclusions and Extra Points: Unbiased Hierarchical methods

Exercise 12.1 (Optimal Importance Sampling for Unbiased MLMC methods)

Show (86), using the formulation

$$\min_{M > 0, (p_\ell) > 0} M \sum_{\ell \geq 0} \text{Work}_\ell p_\ell$$

s.t.

$$\sum_{\ell \geq 0} p_\ell = 1$$

$$\sum_{\ell \geq 0} \frac{E[(\Delta S_\ell)^2]}{p_\ell} \leq M \text{TOL}^2$$

for some given $\text{TOL} > 0$.

Conclusions and Extra Points: Unbiased methods

Remark 12.1 (Optimal Tail Distributions)

According to (86), the tail distribution of p_ℓ^* , since $W_{\text{Work}} = W_\ell \rightarrow \infty$, becomes

$$p_\ell^* \underset{\sim}{\propto} \sqrt{\frac{V_\ell}{W_\ell}}, \text{ as } \ell \rightarrow \infty.$$

With the standard MLMC assumptions, this means

$$p_\ell^* \underset{\sim}{\propto} \exp(-(s + w)\ell/2), \text{ as } \ell \rightarrow \infty.$$

This is a geometric distribution.

Conclusions and Extra Points: Unbiased methods

Remark 12.2 (Biased versus Unbiased)

*Observe that for the unbiased MLMC to work, we need to have $\text{Var}[Z] < \infty$. Under the standard MLMC assumptions, the last condition means that $s > \gamma$ (strong rate larger than the work rate). In such a case, the expected computational work of a sample from the unbiased method is $\mathcal{O}(1)$ and the resulting total work $\mathcal{O}(\text{TOL}^{-2})$. On the other hand, we already know that in that case, the total computational work of the **biased** MLMC is $\mathcal{O}(\text{TOL}^{-2})$ and that the biased MLMC can be used also in the case where $s \leq \gamma$.*

Exercise 12.2 (Biased versus Unbiased, computing with Infinite Variance)

When the unbiased MLMC has $\text{Var}[Z] = +\infty$, we can think of the following regularization. Given a desired accuracy TOL, we truncate the number of levels $L(\text{TOL})$ s.t. the resulting bias is less than TOL and then work with the randomized levels below $L(\text{TOL})$. The resulting biased, randomized levels MLMC has finite variance (although this variance blows up as $L(\text{TOL}) \rightarrow \infty$). Compare the resulting complexity with the standard MLMC.

Conclusions and Extra Points

Exercise 12.3 (Unbiased MIMC)

Extend the unbiased MLMC computations into the MIMC case. Which are the main changes?

Remark 12.3 (Infinite number of input random variables)

We can work with infinite number of input random variables, for instance given by a KL expansion, provided that we have some prior information on the available anisotropy in your problem.

The idea in this case is to extend the optimal multiindex-set corresponding to thresholded profits to sequences, namely

$$\mathcal{I}(\nu) = \{\boldsymbol{\alpha} \in \mathbb{N}^d : \mathcal{P}_{\boldsymbol{\alpha}} \geq \nu\}$$

Click here for a recorded lecture on Unbiased Multi Level Monte Carlo.
(Matti Vihola, 2020)

References

1. "IGA-based Multi-Index Stochastic Collocation for random PDEs on arbitrary domains", by J. Beck, L .Tamellini and R. Tempone. *arXiv:1810.01661*, Computer Methods in Applied Mechanics and Engineering 351, 330–350, 2019.
2. "Multilevel and Multi-index Monte Carlo methods for the McKean-Vlasov equation" by A. L. Haji Ali and R. Tempone. *arXiv:1610.09934* , October 2016. Vol. 28, Issue 4, pp 923–935 Statistics and Computing, 2018.
3. "Multi-Index Stochastic Collocation convergence rates for random PDEs with parametric regularity", by A. L. Haji Ali, F. Nobile, L. Tamellini and R. Tempone. Foundations of Computational Mathematics, Vol. 16(6), Pages 1555-1605, 2016.
4. "Multi-Index Stochastic Collocation for random PDEs", by A. L. Haji Ali, F. Nobile, L. Tamellini and R. Tempone. Computers and Mathematics with Applications, Vol. 306, pp. 95–122, July 2016.
5. "Multi Index Monte Carlo: When Sparsity Meets Sampling" , by A.-L. Haji-Ali, F. Nobile, and R. Tempone. Numerische Mathematik, Vol. 132(4), Pages 767–806, 2016.

Additional references

1. "Unbiased multi-index Monte Carlo", by D. Crisan, P. Del Moral, J. Houssineau and A. Jasra. Stochastic Analysis and Applications, 36:2, 257-273. (2018)
2. "A Multi-Index Quasi-Monte Carlo Algorithm for Lognormal Diffusion Problems". by Robbe, P.; Nuyens, D.; Vandewalle, S. SIAM Journal on Scientific Computing. 39 (5): A1811–C392. (2017).
3. "Enhanced Multi-Index Monte Carlo by means of Multiple Semi-Coarsened Multigrid for Anisotropic Diffusion Problems", by Pieterjan Robbe, Dirk Nuyens, Stefan Vandewalle, arXiv:1907.12334, (2019).
4. Robbe P., Nuyens D., Vandewalle S. (2018) A Dimension-Adaptive Multi-Index Monte Carlo Method Applied to a Model of a Heat Exchanger. In: Monte Carlo and Quasi-Monte Carlo Methods. MCQMC 2016. Springer Proceedings in Mathematics & Statistics, vol 241. Springer, Cham.

L^2 projection – generalized polynomial chaos

Polynomial approximation, motivation

Revisit elliptic PDE with random diffusivity coefficient

We consider once again the model problem

$$\begin{cases} -\operatorname{div}(a(x, \mathbf{y}) \nabla u(x, \mathbf{y})) = f(x), & x \in D, \\ u(x, \mathbf{y}) = 0, & x \in \partial D \end{cases}, \quad \forall \mathbf{y} \in \Gamma := [-\sqrt{3}, \sqrt{3}]^N$$

with $a(x, \mathbf{y}) = \bar{a} + \sum_{i=1}^N \sqrt{\lambda_i} y_i b_i(x)$ (here N could be ∞),

$y_i \sim \mathcal{U}(-\sqrt{3}, \sqrt{3})$ i.i.d., $\sum_{i=1}^N \sqrt{3\lambda_i} \|b_i\|_\infty \leq \delta \bar{a}$ for some $0 < \delta < 1$, so that $\|\nabla u(\mathbf{y})\|_{L^2(D)} \leq C_u := \frac{C_p}{(1-\delta)\bar{a}} \|f\|_{L^2(D)}$.

Moreover, setting $\beta_i = \frac{\sqrt{\lambda_i} \|b_i\|_\infty}{(1-\delta)\bar{a}}$, we assume that exists $0 < \bar{p} \leq 1$ s.t.

$$\sum_{i=1}^N \beta_i^p \leq C_\beta, \quad \forall \bar{p} \leq p \leq 1, \quad \text{with } C \text{ independent of } N \quad (87)$$

Consider a linear functional $Q(u)$. From the QMC Section we know that $Q(u) \in W_{1,2,\gamma}$ for certain weights γ .

But how smooth is the dependence of $\psi(\mathbf{y}) = Q(u(\mathbf{y}))$ with respect to \mathbf{y} really?

For a multi-index $\alpha \in \mathbb{N}^N$, $\alpha = (\alpha_1, \dots, \alpha_N)$, we use the **notation**:

$$\partial_{\mathbf{y}}^\alpha \psi = \frac{\partial^{|\alpha|} \psi}{\partial y_1^{\alpha_1} \cdots \partial y_N^{\alpha_N}}, \quad \alpha! = \prod_{n=1}^N \alpha_n!, \quad \text{and} \quad \mathbf{y}^\alpha = \prod_{n=1}^N y_n^{\alpha_n}.$$

As we have that

$$|\partial_{\mathbf{y}}^\alpha \psi(\mathbf{y})| = |Q(\partial_{\mathbf{y}}^\alpha u(\mathbf{y}))| \leq \|Q\|_{V'} \| \partial_{\mathbf{y}}^\alpha u(\mathbf{y}) \|_V,$$

we have to understand how smoothly the **solution map**

$$\mathbf{y} \mapsto u(\mathbf{y}) \in V$$

depends on \mathbf{y} , where $V = H_0^1(D)$ and $\|v\|_V = \|\nabla v\|_{L^2(D)}$ for the toy problem.

Analogously to the approach for the QMC method, we (formally) differentiate the equation with respect to y_i at $\mathbf{y} = \bar{\mathbf{y}}$ fixed:

i) first derivative ∂_{y_i} . For all $x \in D$ we have

$$-\operatorname{div}(a(x, \bar{\mathbf{y}}) \nabla \partial_{y_i} u(x, \bar{\mathbf{y}})) = \operatorname{div}(\partial_{y_i} a(x, \bar{\mathbf{y}}) \nabla u(x, \bar{\mathbf{y}})) = \operatorname{div}(\sqrt{\lambda_i} b_i(x) \nabla u(x, \bar{\mathbf{y}})).$$

Multiplying by $\partial_{y_i} u(x, \bar{\mathbf{y}})$ and integrating by parts in x yields

$$(1 - \delta) \bar{a} \|\nabla \partial_{y_i} u(\cdot, \bar{\mathbf{y}})\|_{L^2(D)}^2 \leq \sqrt{\lambda_i} \|b_i\|_\infty \|\nabla u(\cdot, \bar{\mathbf{y}})\|_{L^2(D)} \|\nabla \partial_{y_i} u(\cdot, \bar{\mathbf{y}})\|_{L^2(D)}$$

so that $\|\partial_{y_i} u(\cdot, \bar{\mathbf{y}})\|_V \leq \beta_i C_u, \quad \forall \bar{\mathbf{y}} \in \Gamma.$

ii) second mixed derivative $\partial_{y_i y_j}$. For all $x \in D$ we have

$$-\operatorname{div}(a(x, \bar{\mathbf{y}}) \nabla \partial_{y_i y_j}^2 u(x, \bar{\mathbf{y}})) = \operatorname{div}(\sqrt{\lambda_j} b_j(x) \nabla \partial_{y_i} u(x, \bar{\mathbf{y}})) + \operatorname{div}(\sqrt{\lambda_i} b_i(x) \nabla \partial_{y_j} u(x, \bar{\mathbf{y}})).$$

so that

$$\|\partial_{y_i y_j}^2 u(\cdot, \bar{\mathbf{y}})\|_V \leq \beta_j \|\partial_{y_i} u(\cdot, \bar{\mathbf{y}})\|_V + \beta_i \|\partial_{y_j} u(\cdot, \bar{\mathbf{y}})\|_V \leq 2\beta_i \beta_j C_u, \quad \forall \bar{\mathbf{y}} \in \Gamma.$$

Iterating the procedure, yields that for any $\alpha \in \mathbb{N}^N$:

$$\|\partial_{\mathbf{y}_\alpha}^\alpha u(\cdot, \bar{\mathbf{y}})\|_V \leq C_u |\alpha|! \prod_{n=1}^N \beta_j^{\alpha_j}.$$

We see from the argument above that the map $\mathbf{y} \mapsto u(\mathbf{y}) \in H_0^1(D)$ is C^∞ . Moreover, the multivariate Taylor series

$$\sum_{\alpha \in \mathbb{N}^N} \frac{1}{\alpha!} \partial_{\mathbf{y}}^\alpha u(\mathbf{y}) (\mathbf{y} - \bar{\mathbf{y}})^\alpha$$

converges absolutely (in $V = H_0^1(D)$) for any

$$\mathbf{y} \in \{\mathbf{x} \in \mathbb{R}^N : \sum_{n=1}^N \beta_n (x_n - \bar{y}_n) < 1\}.$$

Replacing \mathbf{y} with $\mathbf{z} \in \mathbb{C}^N$ we conclude that the map $\mathbf{z} \mapsto u(\mathbf{z}) \in H_0^1(D; \mathbb{C})$ is analytic in the complex region

$$\Sigma = \{\mathbf{z} \in \mathbb{C}^N : \mathbf{z} = \mathbf{y} + i\mathbf{w}, \quad \mathbf{y} \in \Gamma \subset \mathbb{R}^N, \text{ and } \sum_{n=1}^N \beta_n |w_n| < 1\}$$

Since we have assumed the sequence $\{\beta_j\}_j$ to be ℓ^p summable, we conclude that $\Sigma \supset \prod_{n=1}^N \tilde{\Sigma}_n$ with

$$\tilde{\Sigma}_n = \{z_n \in \mathbb{C} : z_n = y_n + i w_n, \quad y_n \in [-\sqrt{3}, \sqrt{3}], \quad |w_n| < \frac{\beta_n^{p-1}}{C_\beta}\}$$

- ▶ The previous example shows that the solution map $\mathbf{y} \mapsto u(\mathbf{y})$ is analytic in a certain region of the complex plane \mathbb{C}^N containing the parameter space Γ .
- ▶ Can we exploit such (infinite) regularity to build more effective approximation strategies than Monte Carlo?

Idea: Build a **surrogate model** for either the solution, $u(y)$, or the quantity of interest, $Q(u)(y)$, directly. Then use the surrogate model, which is cheap to evaluate, in Monte Carlo computations or other type of sample-based approaches.

Polynomial surrogate models

- ▶ Consider again a Hilbert-space valued function $u : \Gamma \rightarrow V$ where $\mathbf{y} \in \Gamma \subset \mathbb{R}^N$ is a random vector with joint probability density function $\rho : \Gamma \rightarrow \mathbb{R}_+$.
- ▶ Alternative to Monte Carlo type sampling methods, we can also consider **deterministic techniques** to approximate the map $\mathbf{y} \mapsto u(\mathbf{y})$.
- ▶ When u is the solution of a PDE depending on random parameters \mathbf{y} , the map $\mathbf{y} \mapsto u(\mathbf{y})$ can *sometimes* be very smooth and can thus be well approximated by global polynomials.

Once a **suitable polynomial approximation** $u_\Lambda(\mathbf{y})$ of $u(\mathbf{y})$ is available, statistical moments of $u_\Lambda(\mathbf{y})$ can be easily computed

$$\mathbb{E}[u] \approx \mathbb{E}[u_\Lambda], \quad \text{Var}[u] \approx \mathbb{E}[u_\Lambda^2] - \mathbb{E}[u_\Lambda]^2, \quad \dots$$

Other quantities such as failure probabilities are less straightforward to compute

$$p_f = P[Q(u) > q_{cr}] \approx P[Q(u_\Lambda) > q_{cr}] = \int_{\Gamma} \mathbb{1}_{\{Q(u_\Lambda(\mathbf{y})) > q_{cr}\}} \rho(\mathbf{y}) d\mathbf{y}.$$

The last integral can be computed by Monte Carlo type sampling methods. Sampling $u_\Lambda(\mathbf{y})$ will be much cheaper than sampling $u(\mathbf{y})$.

To make this surrogate idea precise, we will first need to discuss some **approximation theoretical basics** using polynomial, in particular their L^2 best approximation property. Only afterwards we can combine it with further discretizations to make it practical.

Objectives of this part:

1. review of $N = 1$ dimensional (orthonormal) polynomial approximation results
2. discuss extensions to for multidimensional polynomial approximations

Uniform distribution and Legendre polynomials

Let $y \sim \mathcal{U}[-1, 1]$ be a uniform random variable. Set $\Gamma = [-1, 1]$ and $\rho(y) = \frac{1}{2}$, $\forall y \in \Gamma$. Then for $v \in L^2_\rho(\Gamma)$ we have

$$\|v\|_{L^2_\rho(\Gamma)}^2 = \int_{-1}^1 v(y)^2 \rho(y) dy = \int_{-1}^1 v(y)^2 \frac{1}{2} dy = \frac{1}{2} \|v\|_{L^2(\Gamma)}^2.$$

Recall that the sequence of Legendre polynomials forms a complete orthogonal basis of $L^2([-1, 1])$.

Reminder: standard Legendre polynomials

The standard Legendre polynomials $\{L_p\}_p$ satisfy the condition

$$L_p(-1) = (-1)^p, \quad L_p(1) = 1, \quad \forall p \geq 0$$

Rodriguez formula

$$L_p(y) = \frac{(-1)^p}{p!2^p} \frac{d^p}{dy^p} [(1 - y^2)^p], \quad y \in [-1, 1], \quad p \geq 0$$

Legendre differential equation

$$((1 - y^2)L'_p(y))' + p(p + 1)L_p(y) = 0, \quad y \in [-1, 1] \quad (88)$$

Three term recurrence relation: $L_0(y) = 1, L_1(y) = y,$

$$(p + 1)L_{p+1}(y) = (2p + 1)yL_p(y) - pL_{p-1}(y), \quad p \geq 1.$$

Orthogonality relations

$$\int_{-1}^1 L_p(y)L_q(y)dy = \frac{2}{2p + 1}\delta_{pq},$$

$$\int_{-1}^1 (1 - y^2)^k L_p^{(k)}(y)L_q^{(k)}(y)dy = \frac{2}{2p + 1} \frac{(p + k)!}{(p - k)!} \delta_{pq}.$$

Orthonormal Legendre polynomials w.r.t. $\rho(y)$

Let now

$$\psi_p(y) = \sqrt{2p+1} L_p(y).$$

Then $\{\psi_p\}_{p=1}^{\infty}$ is an orthonormal basis of $L_p^2(\Gamma)$ and every $u \in L_p^2(\Gamma, V)$ can be expanded in Legendre series with $u_p \in V$ coefficients, namely

$$u(y) = \sum_{p=0}^{\infty} \hat{u}_p \psi_p(y), \quad \text{with } \|u\|_{L_p^2(\Gamma, V)}^2 = \sum_{p=0}^{\infty} \|\hat{u}_p\|_V^2.$$

Let $P_w = \text{span}\{y^p, p = 0, \dots, w\} = \text{span}\{\psi_p, p = 0, \dots, w\}$ be the space of polynomials of degree at most w and $u_w(y) = \sum_{p=0}^w \hat{u}_p \psi_p(y)$ the corresponding truncated series. Then by Parseval's identity the best $L_p^2(\Gamma, V)$ approximation error satisfies

$$\min_{v \in P_w \otimes V} \|u - v\|_{L_p^2(\Gamma, V)}^2 = \|u - u_w\|_{L_p^2(\Gamma, V)}^2 = \sum_{p=w+1}^{\infty} \|\hat{u}_p\|_V^2.$$

In the above formula, we used the **tensor space**

$$\mathbb{P}_w \otimes V = \{v \in L_p^2(\Gamma, V) : v(y) = \sum_{p=0}^w \hat{v}_p \psi_p(y), \hat{v}_p \in V\}.$$

Best approximation error – finite regularity

Decay of the Legendre coefficients

How big is the best approximation error $\sum_{p=w+1}^{\infty} \|\hat{u}_p\|_V^2$?

Answer: it depends on the decay of the Legendre coefficients and hence on the regularity of the u .

Let $\Gamma = [-1, 1]$ and consider

$$H^k(\Gamma, V) = \{y \in \Gamma \mapsto v(y) \in V : \partial_y^s v \in L_p^2(\Gamma, V), \forall s = 0, \dots, k\}$$

Lemma 13.1

Let $u \in H^k(\Gamma, V)$. Then the **Legendre coefficients decay** as

$$\|\hat{u}_p\|_V \leq c(k)p^{-k}|u|_{H^k(\Gamma, V)}, \quad (89)$$

and the **best approximation error** can be bounded by

$$\|u - u_w\|_{L_p^2(\Gamma, V)} \leq c(k)(w+1)^{-k}\|u\|_{H^k(\Gamma, V)}. \quad (90)$$

Proof of Lemma

Proof of (89): The previous result is based on the following argument. For simplicity, let $u : \Gamma = [-1, 1] \rightarrow \mathbb{R}$. Let \mathcal{A} be the second order operator

$$\mathcal{A}u = ((1 - y^2)u')'$$

which is self-adjoint in L_p^2 , that is $(\mathcal{A}u, \phi)_{L_p^2} = (u, \mathcal{A}\phi)_{L_p^2}$, and, thanks to (88), has ψ_p as eigenfunctions. Moreover, we can bound $\|\mathcal{A}^m u\|_{L_p^2(\Gamma, V)} \leq c(2m)\|u\|_{H^{2m}(\Gamma, V)}$. Then for $p > 0$ and using (88),

$$\begin{aligned} |\hat{u}_p| &= |(u, \psi_p)| = \frac{1}{p(p+1)} |(u, \mathcal{A}\psi_p)| = \frac{1}{p(p+1)} |(\mathcal{A}u, \psi_p)| = \dots \\ &= \left(\frac{1}{p(p+1)} \right)^m |(\mathcal{A}^m u, \psi_p)| \leq (p(p+1))^{-m} c(2m) \|u\|_{H^{2m}(\Gamma, V)}. \end{aligned}$$

Taking $k = m/2$, the claim (89) follows.

Proof of (90): We now directly compute

$$\begin{aligned} \sum_{p>w} \|\hat{u}_p\|_V^2 &\leq \sum_{p>w} [p(p+1)]^{-2m} (\mathcal{A}^m u, \psi_p)^2 \\ &\leq [(w+1)(w+2)]^{-2m} \sum_{p>w} (\mathcal{A}^m u, \psi_p)^2 \\ &\leq (w+1)^{-4m} \|\mathcal{A}^m u\|_{L_p^2(\Gamma, V)}^2 \leq (w+1)^{-4m} c(2m)^2 \|u\|_{H^{2m}(\Gamma, V)}^2. \end{aligned}$$

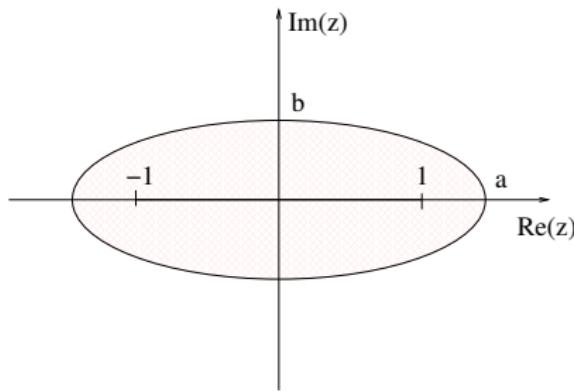


Best approximation error – analytic regularity

Define the ellipse in the complex plane $[-1, 1] \subset \mathcal{E}_r \subset \mathbb{C}$, $r > 1$ (called *Bernstein ellipse*),

$$\mathcal{E}_r = \{z \in \mathbb{C}, |\Re(z)| \leq \frac{r+r^{-1}}{2} \cos \phi, |\Im(z)| \leq \frac{r-r^{-1}}{2} \cos \phi, \phi \in [0, 2\pi]\}$$

with foci ± 1 and semi-axis $a = \frac{r+r^{-1}}{2}$, $b = \frac{r-r^{-1}}{2}$ so that $r = a + b$.



Lemma 13.2

Assume that $y \mapsto u(y)$ admits an analytic extension to the complex plane in an ellipse \mathcal{E}_r , $u : \mathcal{E}_r \rightarrow V$, and let

$$\|u\|_{A,r} := \sup_{z \in \mathcal{E}_r} \|u(z)\|_V < \infty.$$

Then the Legendre coefficients decay as

$$\|\hat{u}_p\|_V \leq c(r)(2p+1)^{\frac{1}{2}} r^{-p} \|u\|_{A,r}, \quad c(r) = \frac{l(\mathcal{E}_r)}{4(r-1)}$$

where $l(\mathcal{E}_r)$ is the length of the ellipse \mathcal{E}_r . Moreover, the best approximation error decays (exponentially) as

$$\|u - u_w\|_{L_p^2(\Gamma, V)}^2 \leq c(r, w) r^{-w} \|u\|_{A,r}, \quad \text{with } c(r, w) = O(\sqrt{w}).$$

For the decay of the Legendre coefficients, see, e.g., [Davis '63].

Sketch of the Proof

Notice that

$$\begin{aligned}\hat{u}_p &= \frac{\sqrt{2p+1}}{2} \int_{-1}^1 u(y) L_p(y) dy \\ &= \frac{\sqrt{2p+1}}{2} \frac{(-1)^p}{p! 2^p} \int_{-1}^1 u(y) \frac{d^p}{dy^p} [(1-y^2)^p] dy \\ &= \frac{\sqrt{2p+1}}{p! 2^{p+1}} \int_{-1}^1 (1-y^2)^p \frac{d^p u}{dy^p}(y) dy.\end{aligned}$$

Then extend u to complex domain and use Cauchy's integral formula to represent derivatives

$$\frac{d^p u}{dy^p}(y) = \frac{p!}{2\pi i} \int_{\partial \mathcal{E}_r} \frac{u(z)}{(z-y)^{p+1}} dz,$$

where the integral is taken along the boundary of the Bernstein ellipse \mathcal{E}_r . We thus find

$$\|\hat{u}_p\|_V \leq \frac{\sqrt{2p+1}}{2^{p+2}\pi} \|u\|_{\mathcal{A},r} \underbrace{\int_{-1}^1 \int_{\partial \mathcal{E}_r} \frac{(1-y^2)^p}{|z-y|^{p+1}} dz dy}_{\sim r^{-p}}$$

Finally

$$\|u - u_w\|_{L_p^2(\Gamma, V)} = \sum_{p>w} \|\hat{u}_p\|_V^2 \leq O(\sqrt{w}) r^{-w} \|u\|_{\mathcal{A},r}.$$

□

Best approximation error – analytic regularity II

The previous result shows that the convergence becomes faster and faster as the solution becomes smoother. It is worth looking at the case of an **analytic function** $u : \Gamma \rightarrow \mathbb{R}$.

Sketch of the approach. Recall that

$$\begin{aligned} u_p &= \frac{\sqrt{2p+1}}{2} \int_{-1}^1 u(y) L_p(y) dy \\ &= \frac{\sqrt{2p+1}}{2} \frac{(-1)^p}{p! 2^p} \int_{-1}^1 u(y) \frac{d^p}{dy^p} [(1-y^2)^p] dy \\ &= \frac{\sqrt{2p+1}}{p! 2^{p+1}} \int_{-1}^1 (1-y^2)^p \frac{d^p u}{dy^p}(y) dy \end{aligned}$$

Extend u to complex, use Cauchy formula to represent derivatives

$$\frac{d^p u}{dy^p}(y) = \frac{p!}{2\pi i} \int_{\gamma_y} \frac{u(z)}{(z-y)^{p+1}} dz$$

with γ_y positively oriented circumference centered at $y \in [-1, 1]$, radius $R(y)$, and all singularities from u are exterior to γ_y .

Then we can estimate

$$\left| \frac{d^p u}{dy^p}(y) \right| \leq \frac{p!}{\{R(y)\}^p} \sup_{z \in \gamma_y} |u(z)|$$

Consider the case of a function with an order 1 pole at $\xi \in \mathbb{R}^-$, let $\delta = \text{dist}(\xi, [-1, 1]) > 0$ and choose $\tau \in (0, 1)$.

Take $R(y) = y + 1 + (1 - \tau)\delta$ s.t. $\sup_{z \in \gamma_y} |u(z)| \leq \frac{C_u}{\tau\delta}$

Then

$$|u_p| \leq \frac{C_u}{\tau\delta} \frac{\sqrt{2p+1}}{2^p} \int_{-1}^1 \left(\frac{1-y^2}{y+1+(1-\tau)\delta} \right)^p dy$$

so it remains to estimate the integral above using [Gui, Babuska 86], i.e.

$$\int_{-1}^1 \left(\frac{1-t^2}{t+1+(1-\tau)\delta} \right)^p dt \leq \frac{(2r)^p}{2p+1} \sqrt{\frac{\pi p}{2}} \left(\sqrt{1-r^2} + \mathcal{O}\left(\frac{1}{p^{1/3}}\right) \right)$$

with $r \equiv \frac{1}{1+\delta(1-\tau)+\sqrt{\delta^2(1-\tau)^2+2\delta(1-\tau)}}$, $0 < r < 1$.

Finally, we obtain the following exponential decay:

$$|u_p| \leq \frac{C_u}{\tau\delta} r^p \left(\sqrt{1-r^2} + \mathcal{O}\left(\frac{1}{p^{1/3}}\right) \right).$$

Assume that $u(y)$ admits and analytic extension to the complex plane in an ellipse \mathcal{E}_r , $u(z) : \mathcal{E}_r \rightarrow V$ and let

$$\|u\|_{A,r} := \sup_{z \in \mathcal{E}_r} \|u(z)\|_V < \infty$$

Then, the Legendre coefficients decay as [see e.g. Davis '63]

$$\|u_p\|_V \leq c(r)(2p+1)^{\frac{1}{2}} r^{-p} \|u\|_{A,r}, \quad c(r) = \frac{l(\mathcal{E}_r)}{4(r-1)}$$

where $l(\mathcal{E}_r)$ is the length of the ellipse \mathcal{E}_r , and

$$\begin{aligned} \|u - u_w\|_{L_p^2(\Gamma, V)}^2 &= \sum_{p=w+1}^{\infty} \|u_p\|_V^2 \leq \sum_{p=w+1}^{\infty} c^2(r) \|u\|_r^2 (2p+1) r^{-2p} \\ &\leq \tilde{c}(r, w) \tilde{r}^{-2w} \|u\|_{A,r}^2, \quad \text{with } \tilde{c}(r, w) = \frac{c^2(r)}{r^2-1} (2w+1 + \frac{2r^2}{r^2-1}) \end{aligned}$$

Roughly speaking, the convergence is exponential: $\|u - u_w\| \sim \tilde{r}^{-w}$.

Remark. If $\|u\|_{A,r} < \infty$ then the sequence $\{\|u_p\|_V^2\}_{p=k}^{\infty}$ is τ summable for any $0 < \tau \leq 1$.

An analogous result holds for the $L^\infty(\Gamma, V)$ -norm

$$\|u - u_w\|_{L^\infty(\Gamma, V)} = \sum_{p=w+1}^{\infty} \sqrt{2p+1} \|u_p\|_V \leq \hat{c}(r, w) \tilde{r}^{-w} \|u\|_{A,r}$$

Non-uniform bounded random variable with $\rho_{max} < +\infty$

We extend now the previous results to the case of a non-uniform bounded random variable $y \in \Gamma \subset \mathbb{R}$, with $\Gamma = [Y_{min}, Y_{max}]$ bounded and with density $\rho : \Gamma \rightarrow \mathbb{R}_+$ bounded. Let $\rho_{max} = \sup_{y \in \Gamma} \rho(y)$.

Let $\{\psi_p\}_{p=0}^{\infty}$ be the sequence of orthonormal polynomials w.r.t ρ . Under the above assumptions, it forms a complete basis of $L^2_{\rho}(\Gamma)$. Therefore, any $u \in L^2_{\rho}(\Gamma, V)$ can be expanded in series $u(y) = \sum_{p=0}^{\infty} u_p \psi_p(y)$.

We set now

$$u_w(y) = \sum_{p=0}^w u_p \psi_p(y)$$

and investigate the best approximation error

$$\|u - u_w\|_{L^2_{\rho}(\Gamma, V)} = \min_{v \in \mathbb{P}_w \otimes V} \|u - v\|_{L^2_{\rho}(\Gamma, V)}$$

We map the interval Γ into $[-1, 1]$. Let $\hat{y} = \frac{|\Gamma|}{2}y + \frac{y_{\max} + y_{\min}}{2}$ and $\tilde{u}(\hat{y}) = u\left(\frac{2}{|\Gamma|}\left(\hat{y} - \frac{y_{\max} + y_{\min}}{2}\right)\right)$. Clearly the smoothness properties of $\tilde{u}(\hat{y})$ and $u(y)$ are the same. If we denote by $\mu(\hat{y}) = \frac{1}{2}$ the uniform density on $[-1, 1]$, then

$$\|u\|_{L_p^2(\Gamma, V)}^2 = \int_{Y_{\min}}^{Y_{\max}} u^2(y) \rho(y) dy = \int_{-1}^1 \tilde{u}^2(\hat{y}) 2\tilde{\rho}(\hat{y}) \mu(\hat{y}) d\hat{y} \leq \frac{2\rho_{\max}}{|\Gamma|} \|\tilde{u}\|_{L_\mu^2([-1,1], V)}^2$$

It follows

$$\|u - u_w\|_{L_p^2(\Gamma, V)} \leq \sqrt{\frac{2\rho_{\max}}{|\Gamma|}} \min_{v \in \mathbb{P}_w([-1,1]) \otimes V} \|\tilde{u} - v\|_{L_\mu^2([-1,1], V)}.$$

- if u has k weighted square integrable derivatives then

$$\|u - u_w\|_{L_p^2(\Gamma, V)} \leq c(k, \rho_{\max}, |\Gamma|, u) w^{-k}$$

- if \tilde{u} admits an analytic extension in an ellipse $\mathcal{E}_r \subset \mathbb{C}$, $r > 1$, then

$$\|u - u_w\|_{L_p^2(\Gamma, V)} \leq c(r, \rho_{\max}, |\Gamma|, u) r^{-w}$$

Other common distributions

For many common probability distributions, the corresponding orthogonal polynomials have properties similar to the Legendre ones and lead to spectral / exponential convergence of the best approximation error.

beta	Jacobi
exponential	Laguerre
Gamma	Generalized Laguerre
Normal	Hermite

For the case of unbounded Γ , for example with Gaussian random variables, the previous analysis has to be slightly modified.

Remark 13.1

The problem of determining polynomials on the real line that are orthogonal w.r.t. a given density is closely related to the Hamburger moment problem, which is the problem of ensuring the existence of a Borel measure (e.g., one with Lebesgue density) such that its moments match the values of a given sequence. For example, in the context of orthogonal polynomials given by a three-term recursion the well-posedness of the associated Hamburger moment problem has been studied in [Chihara '89].

Normal distribution – Hermite polynomials

The polynomials orthogonal with respect to the Gaussian distribution $\rho(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2}$ are the **Hermite polynomials $\{H_p\}_p$** , which satisfy

$$H_p(y) = y^p + \sum_{k=0}^{p-1} \alpha_{p,k} y^k$$

Analogue of Rodriguez formula

$$H_p(y) = (-1)^p e^{y^2/2} \frac{d^p}{dy^p} [e^{-y^2/2}], \quad y \in \mathbb{R}, \quad p \geq 0$$

Hermite differential equation

$$\left(e^{-y^2/2} H'_p(y) \right)' = -p e^{-y^2/2} H_p(y), \quad y \in \mathbb{R}$$

Three term recurrence relation: $H_0(y) = 1, H_1(y) = y,$

$$H_{p+1}(y) = yH_p(y) - pH_{p-1}(y), \quad p \geq 1.$$

Orthogonality relations

$$\int_{\mathbb{R}} H_p(y) H_q(y) \rho(y) dy = p! \delta_{pq}$$

Chebyshev distribution and polynomials

The Chebyshev distribution is given by $\rho(y) = \frac{1}{\pi\sqrt{1-y^2}}$, $y \in (-1, 1)$, and the corresponding orthogonal polynomials are the **Chebyshev polynomials**

$$T_p(y) = \cos(p\theta), \quad \text{with } \theta = \arccos y, \quad p \geq 0.$$

Rodriguez formula

$$T_p(y) = \frac{(-1)^p}{p!2^p} \sqrt{1-y^2} \frac{d^p}{dy^p} \left(\frac{(1-y^2)^p}{\sqrt{1-y^2}} \right), \quad y \in [-1, 1], \quad p \geq 0$$

Chebyshev differential equation

$$\left(\sqrt{1-y^2} T'_p(y) \right)' + \frac{p^2}{\sqrt{1-y^2}} T_p(y) = 0, \quad y \in [-1, 1]$$

Three term recurrence relation: $T_0(y) = 1$, $T_1(y) = y$,

$$T_{p+1}(y) = 2yT_p(y) - T_{p-1}(y), \quad p \geq 1.$$

Orthogonality relations

$$\int_{-1}^1 T_p(y) T_q(y) \rho(y) dy = \frac{1}{2} \delta_{pq}, \quad p \geq 1, \quad \int_{-1}^1 T_0(y)^2 \rho(y) dy = 1$$

L^2 projection – generalized polynomial chaos
Part II: Multidimensional polynomial approximation

Multidimensional polynomial approximation

Now we consider the case of a random vector $\mathbf{y} = (y_1, \dots, y_N)$ with joint density $\rho : \Gamma \subset \mathbb{R}^N \rightarrow \mathbb{R}_+$ and a function $u(\mathbf{y}) \in L^2_\rho(\Gamma, V)$.

Let us focus on the **uniform case**: $\Gamma = [-1, 1]^N$ and $\rho(y) = \frac{1}{2^N}$.

Orthonormal basis: Let $\{\psi_p\}_{p=0}^\infty$ be the sequence of 1D Legendre polynomials orthonormal w.r.t. the weight $\rho(y) = \frac{1}{2}$,
 $\mathbf{p} = (p_1, \dots, p_n) \in \mathbb{N}^N$ a multi-index and let

$$\psi_{\mathbf{p}}(\mathbf{y}) = \prod_{n=1}^N \psi_{p_n}(y_n)$$

Then, $\{\psi_{\mathbf{p}}\}_{\mathbf{p} \in \mathbb{N}^N}$ is an orthonormal basis of $L^2_\rho(\Gamma)$.

Polynomial space: Recall that in 1D we have considered the space $\mathbb{P}_w([-1, 1]) = \text{span}\{\psi_p(y), \ p = 0, \dots, w, \ y \in [-1, 1]\}$.

What is the proper choice of polynomial subspaces for approximation in N dimensions?

Tensor product polynomial space

Notation: For a multi-index \mathbf{p} and $\alpha = (\alpha_1, \dots, \alpha_N) \in \mathbb{R}^N$:

- ▶ $|\mathbf{p}|_q = \left(\sum_{n=1}^N p_n^q \right)^{\frac{1}{q}}$ and $|\mathbf{p}|_\infty = \max_{1 \leq n \leq N} p_n$
- ▶ $\mathbf{p}^\alpha = \prod_{n=1}^N p_n^{\alpha_n}$, $\mathbf{p}! = \prod_{n=1}^N p_n!$
- ▶ $\mathbf{p} \leq k$ means $p_n \leq k$, $\forall n = 1, \dots, N$.

A natural choice of polynomial space is a **tensor product space**

$$\mathbb{P}_{TP(w)}([-1, 1]^N) = \bigotimes_{n=1}^N \mathbb{P}_w([-1, 1]) = \text{span}\{\psi_{\mathbf{p}}, \quad |\mathbf{p}|_\infty \leq w\}$$

i.e. the space of multivariate polynomials of degree at most w *in each variable separately*.

Dimension of the space: $M = \dim(\mathbb{P}_{TP(w)}(\Gamma)) = (w+1)^N$.

Truncated Legendre expansion: Given $u \in L^2_\rho(\Gamma, V)$, let

$$u_{TP(w)}(\mathbf{y}) = \sum_{\mathbf{p} \leq w} \hat{u}_{\mathbf{p}} \psi_{\mathbf{p}}(\mathbf{y}), \quad \psi_{\mathbf{p}} \text{ basis of } \mathbb{P}_{TP(w)}(\Gamma).$$

Then by Parseval we have $\|u - u_{TP(w)}\|_{L^2_\rho(\Gamma, V)}^2 = \sum_{|\mathbf{p}|_\infty > w} \|\hat{u}_{\mathbf{p}}\|_V^2$.

Best approximation error – finite regularity

Analogously to the 1D case, let

$$H^k(\Gamma, V) = \{\mathbf{y} \in \Gamma \mapsto v(\mathbf{y}) \in V : \partial_y^\alpha v \in L_\rho^2(\Gamma, V), \forall |\alpha| \leq k\}$$

and denote by Π_n^w the $L_\rho^2(\Gamma, V)$ orthogonal projection onto \mathbb{P}_w in the variable y_n .

Best approximation error:

$$\begin{aligned}\|u - u_{TP(w)}\|_{L_\rho^2(\Gamma, V)} &\equiv \|u - u_{TP(w)}\| = \|u - \Pi_1^w \Pi_2^w \cdots \Pi_N^w u\| \\ &\leq \|u - \Pi_1^w u\| + \|\Pi_1^w(u - \Pi_2^w \cdots \Pi_N^w u)\| + \dots \\ &\quad + \|\Pi_1^w \cdots \Pi_{N-1}^w(u - \Pi_N^w u)\| \\ &\leq \sum_{n=1}^N \|u - \Pi_n^w u\| \leq Nc(k)w^{-k}\|u\|_{H^k(\Gamma, V)}\end{aligned}$$

by Lemma 13.1.

Observe that this is the same estimate as in the one dimensional case!

Best approximation error – finite regularity II

Space of functions with k weighted square integrable derivatives: let
 $\alpha = (\alpha_1, \dots, \alpha_n) \in \mathbb{N}^N$

$$|v|_{\mathcal{H}^k(\Gamma, V)} = \sum_{|\alpha|=k} \int_{\Gamma} \prod_{n=1}^N (1 - y_n^2)^{\alpha_n} \left| \frac{\partial^k v(\mathbf{y})}{\partial y_1^{\alpha_1} \cdots \partial y_n^{\alpha_n}} \right|^2 \rho(\mathbf{y}) d\mathbf{y}$$

$$\mathcal{H}^k(\Gamma, V) = \{v \in L^2_{\rho}(\Gamma, V) : \sum_{s=0}^k |v|_{\mathcal{H}^s(\Gamma, V)} < +\infty\}$$

Best approximation error: for $w \geq k - 1$

$$\begin{aligned} \|u - u_{TP(w)}\|_{L^2_{\rho}(\Gamma, V)}^2 &= \sum_{|p|_{\infty} > w} \|u_p\|_V^2 \leq \sum_{n=1}^N \sum_{p_n > w} \|u_p\|_V^2 \\ &\leq (w+1)^{-2k} \left(\sum_{n=1}^N \sum_{p_n \geq w+1} p_n^{2k} \|u_p\|_V^2 \right) \\ &\leq N \left(\frac{e}{2} \right)^{2k} (w+1)^{-2k} |u|_{\mathcal{H}^k(\Gamma, V)}^2 \end{aligned}$$

Observe that this is the same estimate as in the one dimensional case.

However, if we express the convergence rate in terms of the number of degrees of freedom (Legendre coeffs) $M = \dim(\mathbb{P}_{TP(w)}(\Gamma)) = (w + 1)^N$ we have

$$\|u - u_{TP(w)}\|_{L_p^2(\Gamma, V)} \leq N c(k) M^{-\frac{k}{N}} |u|_{\mathcal{H}^k(\Gamma, V)}, \quad \text{with } c(k) = (e/2)^k.$$

Similar results hold for an **analytic function** in a tensor poly-ellipse $\mathcal{E}_r \otimes \dots \otimes \mathcal{E}_r$, $r > 1$:

$$\|u - u_{TP(w)}\|_{L_p^2(\Gamma, V)} \leq N c(r) r^{-w} \|u\|_{A,r} \leq N c(r) r^{-\sqrt[N]{M}} \|u\|_{A,r}$$

Conclusions:

1. The dimension of the space grows exponentially fast in the number of input random variables N . Even for moderately large N , the space is just too large for any practical computation.

Example: Take $w = 1$ and $N = 30$, then $M = 2^{30} \approx 10^9$.

2. The convergence rate heavily deteriorates with N !!! Even for a very smooth function the convergence will be very slow in large N .

Example: Take $k = 10$ and $N = 30$, then

$\|u - u_{TP(w)}\|_{L_p^2(\Gamma, V)} \sim M^{-\frac{1}{3}}$, hence slower than Monte Carlo!

This effect is called the **curse of dimensionality**

A useful concept from Information Based Complexity

The concept of curse of dimensionality has been formalized in the field of Information Based Complexity (see e.g. [Novak, Woźniakowski ...])

Consider the problem of approximating or integrating a function (as the one previously analyzed) in dimension N .

Let ϵ be a prescribed tolerance and $M(\epsilon, N)$ be the minimum number of degrees of freedom needed to achieve that tolerance. Depending on the scaling of $M(\epsilon, N)$ w.r.t. ϵ and N , the problem is classified by one of the following cases

- ▶ strongly polynomially tractable
- ▶ weakly polynomially tractable
- ▶ weakly tractable
- ▶ intractable (or affected by the curse of dimensionality)

From the estimate $\|u - u_{TP(w)}\|_{L_p^2(\Gamma, V)} \leq M^{-\frac{k}{N}}$ we conclude that the problem is **intractable** in high dimensions.

General multivariate polynomial space

Question: Is there a better choice of polynomial space that makes the approximation problem tractable in high dimensions?

Let

- ▶ $\mathbf{p} = (p_1, \dots, p_N) \in \mathbb{N}^N$ be a multi-index
- ▶ $\Lambda(w) \subset \mathbb{N}^N$ an index set such that $\max_{\mathbf{p} \in \Lambda} \max_{n=1, \dots, N} p_n = w$.
- ▶ for convenience we take only **monotone** (equivalently **downward closed**) sets, which satisfy

$$\mathbf{p} \in \Lambda(w) \implies \mathbf{p} - \mathbf{e}_j \in \Lambda(w), \forall j = 1, \dots, N \text{ s.t. } p_j - 1 \geq 0.$$

Definition 13.3 (General multivariate polynomial space)

Recall that $\mathbf{y} \mapsto \psi_{\mathbf{p}}(\mathbf{y}) = \prod_{n=1}^N \psi_{p_n}(y_n)$. The space

$$\mathbb{P}_{\Lambda(w)}(\Gamma) = \text{span}\{\psi_{\mathbf{p}}, \mathbf{p} = (p_1, \dots, p_N) \in \Lambda(w)\}$$

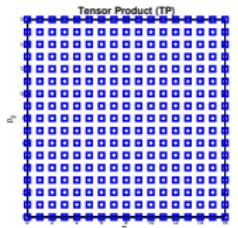
is called **general multivariate polynomial space**.

Observe that the maximum polynomial degree in each variable y_n is at most w .

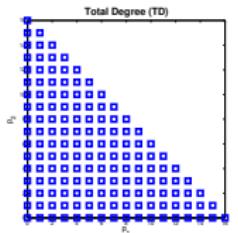
Examples of (general multivariate) polynomial spaces

Let us consider a bi-dimensional problem $\mathbf{y} = (y_1, y_2)$ and $w = 16$.

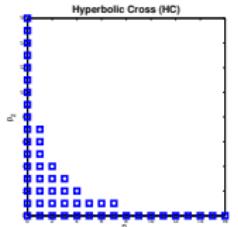
Tensor product (TP): $\max_n p_n \leq w$



Total degree (TD): $\sum_{n=1}^N p_n \leq w$



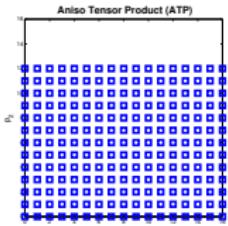
Hyperbolic cross (HC): $\prod_{n=1}^N (p_n + 1) \leq w + 1$



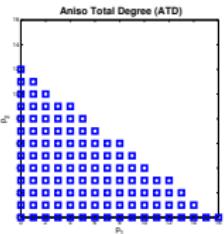
Anisotropic variants

Consider a weight vector $\alpha = (\alpha_1, \dots, \alpha_N) \in \mathbb{R}_+^N$, with $\alpha_{\min} = 1$.

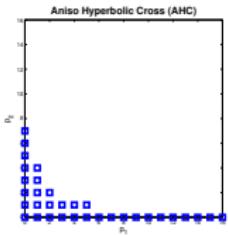
Tensor product (ATP): $\max_n \alpha_n p_n \leq w$



Total degree (ATD): $\sum_{n=1}^N \alpha_n p_n \leq w$



Hyperbolic cross (AHC): $\prod_{n=1}^N (p_n + 1)^{\alpha_n} \leq w + 1$



Best approximation space – general guideline

Given a function $u \in L^2_\rho(\Gamma, V)$ with Legendre coefficients $\{\hat{u}_\mathbf{p}\}_\mathbf{p}$, we may ask the question: what is the best polynomial space \mathbb{P}_Λ of dimension $M = \#(\Lambda)$ that leads to the minimal approximation error, so that

$$\min_{\#(\Lambda)=M} \|u - u_\Lambda\|_{L^2_\rho(\Gamma, V)}^2 = \min_{\#(\Lambda)=M} \sum_{\mathbf{p} \notin \Lambda} \|\hat{u}_\mathbf{p}\|_V^2 ?$$

This question has a simple (almost tautological) solution:

$$\Lambda_{opt}(M) = \{\mathbf{p} \in \mathbb{N}^N \text{ corresponding to the } M \text{ largest coefficients } \|\hat{u}_\mathbf{p}\|_V\} .$$

We call $u_{\Lambda_{opt}(M)}$ the best M-term approximation.

Mixed derivatives and Hyperbolic cross space

Let us consider the space of functions with k mixed square integrable derivatives:

$$H_{mix}^k(\Gamma, V) = \{ \mathbf{y} \in \Gamma \mapsto v(\mathbf{y}) \in V : \partial_{\mathbf{y}}^{\alpha} v \in L_p^2(\Gamma, V), \forall \alpha \leq k \}.$$

Reproducing the 1D result, the Legendre coefficients decay as

$$\|\hat{u}_{\mathbf{p}}\|_V \leq c(k, N) \prod_{n=1}^N (p_n + 1)^{-k} \|u\|_{H_{mix}^k(\Gamma, V)}$$

and moreover $\sum_{\mathbf{p} \in \mathbb{N}^N} \prod_{n=1}^N (p_n + 1)^{2k} \|\hat{u}_{\mathbf{p}}\|_V^2 \leq 2c(k, N)^2 \|u\|_{H_{mix}^k(\Gamma, V)}^2$.

Then, the best polynomial space (the one collecting the largest estimates of the coefficients) is the Hyperbolic Cross space

$$\mathbb{P}_{HC(w)}(\Gamma) = \text{span}\{\psi_{\mathbf{p}}, \prod_{n=1}^N (p_n + 1) \leq w + 1\}.$$

Consequently, we find that the best approximation error satisfies

$$\begin{aligned}\|u - u_{HC(w)}\|_{L^2_\rho(\Gamma, V)}^2 &= \sum_{\prod_n(p_n+1) > w+1} \|\hat{u}_p\|_V^2 \\ &\leq (w+1)^{-2k} \sum_{\prod_n(p_n+1) > w+1} \prod_{n=1}^N (p_n+1)^{2k} \|\hat{u}_p\|_V^2 \\ &\leq 2c(k, N)^2 (w+1)^{-2k} \|u\|_{H_{mix}^k(\Gamma, V)}^2\end{aligned}$$

Moreover, the dimension of the $HC(w)$ space: (see [Burgartz-Griebel '04])

$$M \leq (w+1)(1 + \log(w+1))^{N-1}$$

Hence,

$$\|u - u_{HC(w)}\|_{L^2_\rho(\Gamma, V)} \leq 2c(k, N)(1 + \log(w+1))^{k(N-1)} M^{-k} \|u\|_{H_{mix}^k(\Gamma, V)}$$

The curse of dimensionality is greatly reduced. However, the price to pay is a required control on k -th order mixed derivatives. Compare this with QMC methods, that require $k = 1$ mixed derivatives.

Analytic functions and Total Degree polynomial space

Let us consider a function that is analytic in a polyellipse
 $\mathcal{E}_r \otimes \dots \otimes \mathcal{E}_r \subset \mathbb{C}^N$, $r > 1$.

By repeating the 1D argument in each variable y_n , the Legendre coefficients decay as

$$\|\hat{u}_{\mathbf{p}}\|_V \leq \|u\|_{A,r} \prod_{n=1}^N c(r) \sqrt{(2p_n + 1)} r^{-p_n} \sim r^{-|\mathbf{p}|_1}$$

Hence

$$\|\hat{u}_{\mathbf{p}}\|_V \geq \epsilon \implies |\mathbf{p}|_1 \leq \frac{\log(\epsilon^{-1})}{\log(r)} \implies \text{TD space!}$$

Hence, the best polynomial space is the Total degree space:

$$\mathbb{P}_{TD(w)}(\Gamma) = \text{span}\{\psi_{\mathbf{p}}, \ |\mathbf{p}|_1 \leq w\}.$$

Analytic functions and Total Degree polynomial space

Best approximation error

$$\|u - u_{TD(w)}\|_{L^2_\rho(\Gamma, V)} = \sum_{|\mathbf{p}|_1 > w} \|\hat{u}_{\mathbf{p}}\|_V^2 \leq \frac{1}{r-1} r^{-w} \|u\|_{A,r}$$

Let us consider the polynomial space

$$\mathbb{P}_{TD(w)}(\Gamma) = \text{span}\{\psi_{\mathbf{p}}, \ |\mathbf{p}|_1 \leq w\}$$

which corresponds to the space of **polynomials of total degree at most w** , and the truncated Legendre expansion $u_{TD(w)}(\mathbf{y}) = \sum_{|\mathbf{p}|_1 \leq w} u_{\mathbf{p}} \psi_{\mathbf{p}}(\mathbf{y})$.
The dimension M of the polynomial space satisfies for $w \geq N$

$$M = \binom{w+N}{N} = \prod_{n=1}^N \left(1 + \frac{w}{n}\right) \leq \frac{(N+w)^N}{N!} \leq \left(\frac{e(w+N)}{N}\right)^N$$

so we get the final estimate (see [Beck-Nobile-Tamellini-Tempone, CAMWA '13] for a sharper one)

$$\|u - u_{TD(w)}\|_{L^2_\rho(\Gamma, V)} \leq \frac{1}{r-1} r^{-N(\sqrt[N]{M}/e-1)} \|u\|_{A,r}, \quad w \geq 0$$

(Compare with TP estimate $\|u - u_{TP(w)}\|_{L^2_\rho(\Gamma, V)} \leq N c(r) r^{-\sqrt[N]{M}} \|u\|_{A,r}$)

Optimality of TD space for analytic functions

We argue that the approximation in the total degree space is nearly the best that one can do for an analytic function in a polydisk (or polyellipse).

Indeed, the Legendre coefficients decay as

$$\|u_{\mathbf{p}}\|_V \leq \|u\|_{A,r} \prod_{n=1}^N c(r) \sqrt{(2p_n + 1)} r^{-p_n} \sim r^{-|\mathbf{p}|_1}$$

where we have neglected the algebraic term $\prod_{n=1}^N \sqrt{(2p_n + 1)}$.

The best approximation in a generic polynomial space $\mathbb{P}_{\Lambda(w)}$ satisfies

$$\|u - u_{\Lambda(w)}\|_{L_p^2(\Gamma, V)}^2 = \sum_{\mathbf{p} \notin \Lambda(w)} \|u_{\mathbf{p}}\|_V^2$$

The **Best M -term approximation** corresponds to the index set $\Lambda(w)$ that contains the M largest coefficients.

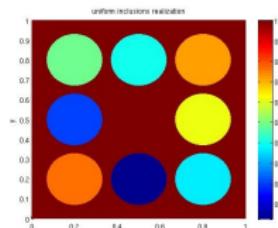
In the case of an analytic function in a polyellipse we have

$$\|u_{\mathbf{p}}\|_V \geq \epsilon \implies |\mathbf{p}|_1 \leq \frac{\log(\epsilon^{-1})}{\log(r)} \implies \text{TD space!}$$

Random PDE example – non overlapping inclusions

We consider, once again, the model problem

$$\begin{cases} -\operatorname{div}(a(x, \mathbf{y}) \nabla u(x, \mathbf{y})) = f(x), & x \in D, \\ u(x, \mathbf{y}) = 0, & x \in \partial D \end{cases}$$



Denoting by 1_{D_i} the characteristic (or indicator) function of each inclusion,

$$a(x, \mathbf{y}) = \bar{a} + \sum_{i=1}^N (\sigma y_i + \bar{y}) 1_{D_i}(x), \quad y_i \sim \mathcal{U}(-1, 1) \text{ i.i.d.}$$

Provided that $\sigma \leq \delta \bar{a} + \bar{y}$ the problem is well posed with uniform stability estimate $\|\nabla u(\mathbf{y})\|_{L^2(D)} \leq C_u := \frac{C_p}{(1-\delta)\bar{a}} \|f\|_{L^2(D)}$.

We have already seen that we can differentiate the equation w.r.t. \mathbf{y} as many times as we want, so $u(\mathbf{y}) \in C^\infty(\Gamma, V)$.

We now show that u is an analytic function of \mathbf{y} in a polydisk $D_r \otimes \dots \otimes D_r \subset \mathbb{C}^N$ that contains $[-1, 1]^N$.

Non overlapping random inclusion problem: Analyticity

The (usual) idea to prove analyticity is to study the equation in the complex domain.

Replace each real variable y_j with a complex variable $z_j = y_j + iw_j \in \mathbb{C}$.

Denoting with $\mathbf{z} = (z_1, \dots, z_N) \in \mathbb{C}^N$, the diffusion coefficient

$$a(x, \mathbf{z}) = \bar{a} + \sum_{i=1}^N (\sigma y_i + \bar{y} + i\sigma w_i) 1_{D_i}(x)$$

is complex valued, as is the solution $u(\mathbf{z}) = u_R(\mathbf{z}) + iu_I(\mathbf{z})$.

By the Lax-Milgram Thm. (for complex Hilbert spaces) one can show that the problem has a unique solution $u(\mathbf{z}) \in V := H_0^1(D, \mathbb{C})$ for a.a. suitable $\mathbf{z} \in \mathbb{C}^N$, provided that $\alpha(\mathbf{z}) := \min_{x \in D} \Re a(x, \mathbf{z}) > 0$. Moreover

$$\|u(\mathbf{z})\|_V := \|\nabla u(\mathbf{z})\|_{L^2(D, \mathbb{C})} \leq \frac{C_P \|f\|_{L^2(D, \mathbb{R})}}{\alpha(\mathbf{z})} .$$

On the other hand, inside the inclusion D_i we find that

$$\Re a(x, \mathbf{z}) = \bar{a} + \sigma \Re z_i + \bar{y} > 0 \implies \Re z_i > -\frac{\bar{a} + \bar{y}}{\sigma} .$$

Let us define the set

$$\Sigma_\delta := \{\mathbf{z} \in \mathbb{C}^N : \min_{i=1,\dots,N} \Re z_i \geq -(\delta \bar{a} + \bar{y})/\sigma\} \subset \mathbb{C}^N.$$

The PDE problem is well posed for any $\mathbf{z} \in \Sigma_\delta$. Moreover, by differentiating the equation, one can show that the solution $\mathbf{z} \mapsto u(\mathbf{z})$ is complex differentiable in Σ_δ .⁹

That is, the solution $u : \Sigma_\delta \rightarrow V := H_0^1(D, \mathbb{C})$ is analytic and

$$\max_{\mathbf{z} \in \Sigma_\delta} \|u(\mathbf{z})\|_V \leq \frac{C_P}{(1-\delta)\bar{a}} \|f\|_{L^2(D, \mathbb{R})}$$

Moreover, the set Σ_δ contains the polydisk

$$D_r \otimes \dots \otimes D_r = \{\mathbf{z} \in \mathbb{C}^N : |z_i| \leq r\}, \quad \text{with } r = \frac{\delta \bar{a} + \bar{y}}{\sigma} > 1.$$

Consequently, the “best” polynomial approximation space is the TD space for this problem.

⁹One can write the problems for $\partial_{\Re z_i} u_R$, $\partial_{\Im z_i} u_R$, $\partial_{\Re z_i} u_I$, $\partial_{\Im z_i} u_I$ and check that the Cauchy-Riemann conditions are satisfied in Σ_δ ; see e.g., [Nobile-Tempone IJNME '09].

Non overlapping random inclusion problem: Numerics

This example is taken from [Beck-Nobile-Tamellini-Tempone CAMWA '13], with $D = [0, 1]^2$, $\bar{a} = 1$, $\bar{y} = -0.595$, $\sigma = 0.395$, $f(x) = 1_F(x)$ with $F = [0.4, 0.6]^2$ and computed quantity of interest $Q(u) = \int_F u(x) dx$.

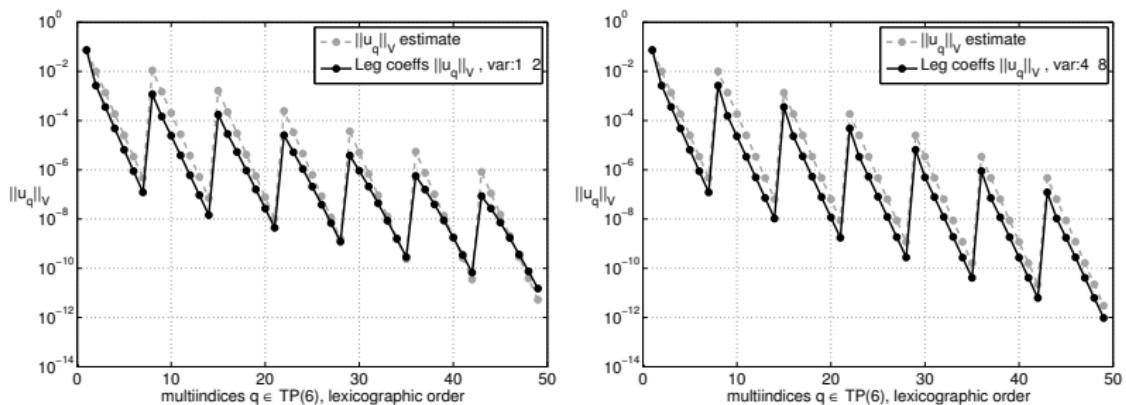


Figure: Comparison between some coefficients of the Legendre expansion of $Q(u)$ (computed with a highly accurate Galerkin approximation in HC space at level $w = 56$) and the corresponding bound $\|u_p\| \sim r^{-|p|_1}$ with r suitably tuned. The multiindices corresponding to the coefficients shown in the plots are nonzero only in $y_1 - y_2$ (left) and $y_4 - y_8$ (right) and ordered in lexicographic order.

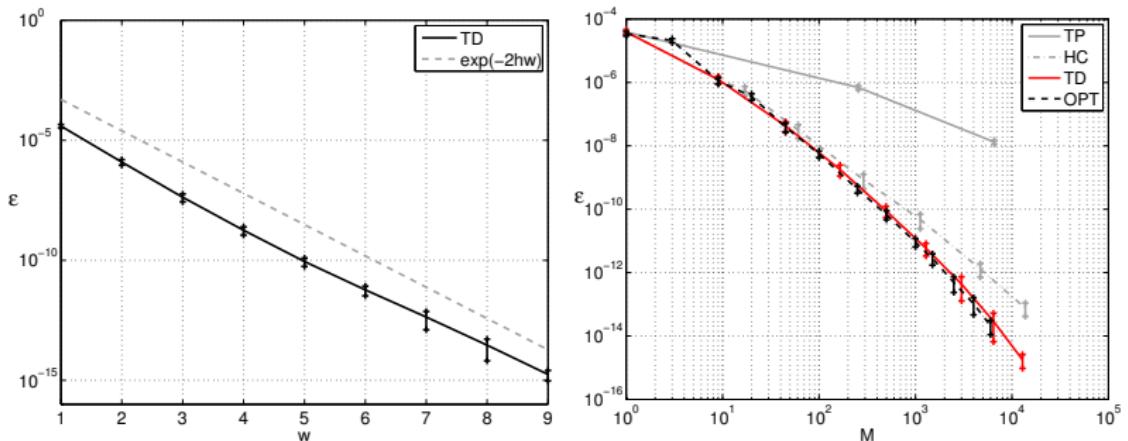


Figure: Left: convergence of the error $\|Q(u) - Q(u_{TD(w)})\|_{L_p^2(\Gamma)}^2$ with respect to w for the quasi-optimal TD polynomial approximation. Right: Convergence of the same error in terms of the dimension of the polynomial space, for the TD approximation, as well as Tensor Product (TP), Hyperbolic cross (HC) and best M -terms (OPT) approximations.

Non-linear approximation

Recall that the best polynomial space \mathbb{P}_Λ of dimension $M = \#(\Lambda)$ that leads to the minimal approximation error, i.e.,

$$\min_{\#(\Lambda)=M} \|u - u_\Lambda\|_{L^2_\rho(\Gamma, V)}^2 = \min_{\#(\Lambda)=M} \sum_{\mathbf{p} \notin \Lambda} \|\hat{u}_\mathbf{p}\|_V^2$$

is simply characterized by

$$\Lambda_{opt}(M) = \{\mathbf{p} \in \mathbb{N}^N \text{ corresponding to the } M \text{ largest coefficients } \|\hat{u}_\mathbf{p}\|_V\}.$$

We call $u_{\Lambda_{opt}(M)}$ the best M-term approximation.

The analysis of the best M-term approximation is strictly related to summability properties of the sequence $\{\|\hat{u}_\mathbf{p}\|_V\}_{\mathbf{p}}$ for some $0 < \tau \leq 2$.¹⁰

Denote
$$\{\|\hat{u}_\mathbf{p}\|_V\}_{I^\tau} = \left(\sum_{\mathbf{p} \in \mathbb{N}^N} \|\hat{u}_\mathbf{p}\|_V^\tau \right)^{\frac{1}{\tau}}.$$

¹⁰Observe that we already guaranteed summability for $\tau = 2$ before.

Best M-term approximation

Theorem 13.4 (Best M -term approximation error)

If the sequence of Legendre coefficients is τ summable for some $0 < \tau < 2$, then

$$\|u - u_{\Lambda_{opt}(M)}\|_{L_p^2(\Gamma, V)} \leq M^{\frac{1}{2} - \frac{1}{\tau}} \{\|\hat{u}_p\|_V\}_{I^\tau}. \quad (**)$$

The proof of this result relies on the following simple Lemma.

Lemma 13.5 (Stechkin's Lemma)

Let $0 \leq p \leq q$ and $\{a_j\}_{j=1}^\infty$ a positive and decreasing sequence. Then

$$\left(\sum_{j>M}^{\infty} a_j^q \right)^{\frac{1}{q}} \leq M^{\frac{1}{q} - \frac{1}{p}} \left(\sum_{j=1}^{\infty} a_j^p \right)^{\frac{1}{p}}. \quad (*)$$

Indeed, to prove $(**)$, apply $(*)$ with $q = 2$, $p = \tau$ and $\{a_j\}_j$ being the ordered sequence of $\{\|\hat{u}_p\|_V\}_p$.

Proof of Stechkin's Lemma.

Since $q \geq p$, a_j^{q-p} does not increase, and we find

$$\sum_{j>M}^{\infty} a_j^q \leq a_M^{q-p} \sum_{j>M}^{\infty} a_j^p \leq a_M^{q-p} \sum_{j=1}^{\infty} a_j^p .$$

Moreover,

$$Ma_M^p \leq \sum_{j=1}^M a_j^p \leq \sum_{j=1}^{\infty} a_j^p .$$

Thus, (*) follows from substituting the last inequality into the previous one. □

Infinite dimensional approximation

Consider now a function u that depends on $N = \infty$ random variables

$$u = u(y_1, y_2, \dots, y_n, \dots).$$

The previous result will allow us to decide whether or not such a function can be effectively approximated by polynomials.

- ▶ The results that we have seen so far in finite dimension, are all affected to some extent by the curse of dimensionality. In particular, we cannot let $N \rightarrow \infty$. This is because all variables have the same importance and the more we add, the more difficult is the approximation problem.
- ▶ The only hope for infinite dimensional approximation is that the function $u = u(y_1, y_2, \dots, y_n, \dots)$ depends less and less on the “tail” random variables, i.e., those y_n , with $n \gg 1$.
- ▶ This will indeed be the case when considering functions of random fields properly expanded in series.

Remark 13.2

These considerations should, yet again, be compared with our discussion on the QMC method for $N = \infty$, where we introduced suitable weights.

Infinite dimensional analytic functions

We consider the following case:

$u \in L^2_\rho(\Gamma, V)$ admits an analytic extension in any polyellipse \mathcal{E}_r with $\sum_{n=1}^{\infty} \gamma_n r_n \leq 1$, with $\sum_{n=1}^{\infty} \gamma_n \leq 1$.

This case arises in the analysis of elliptic PDEs with stochastic coefficients (see slides below).

Let

$$\mathcal{A}_\gamma = \bigcup_{\sum_{n=1}^{\infty} \gamma_n r_n \leq 1} \mathcal{E}_r \quad \text{and} \quad \|u\|_{A,\gamma} = \max_{z \in \mathcal{A}_\gamma} \|u(z)\|_V.$$

Theorem 13.6 ([Cohen-Devore-Schwab, 2010])

Assume that $\|\{\gamma_n\}_n\|_{\ell^\tau} := (\sum_{n=1}^{\infty} \gamma_n^\tau)^{\frac{1}{\tau}} < \infty$ for $0 < \tau \leq 2$. Then, the best M -term approximation satisfies

$$\|u - u_{\Lambda_{opt}(M)}\|_{L^2_\rho(\Gamma, V)} \leq C(\tau) \|\{\gamma_n\}_n\|_{\ell^\tau} M^{\frac{1}{2} - \frac{1}{\tau}} \|u\|_{A,\gamma}$$

and

$$\|u - u_{\Lambda_{opt}(M)}\|_{L^\infty(\Gamma, V)} \leq \tilde{C}(\tau) \|\{\gamma_n\}_n\|_{\ell^\tau} M^{1 - \frac{1}{\tau}} \|u\|_{A,\gamma}$$

Hint of the proof

We prove the result under the **more stringent assumption** that $\sum_{n=1}^{\infty} \gamma_n \leq \frac{1}{3e\pi+2}$.

Let $\mathbf{p} = (p_1, p_2, \dots) \in \mathbb{N}^{\mathbb{N}}$ a finitely supported multi-index, with $|\mathbf{p}| = \sum_n p_n$. The corresponding Legendre coefficients satisfy, for any \mathbf{r} s.t. $\sum_{n=1}^{\infty} \gamma_n r_n \leq 1$,

$$\|u_{\mathbf{p}}\|_V \leq \|u\|_{A,\gamma} \prod_{n \in \text{supp}(\mathbf{p})} \frac{\pi r_n}{2(r_n - 1)} \sqrt{2p_n + 1} \left(\frac{1}{r_n} \right)^{p_n}$$

We now take

$$r_n = 1 + \kappa + \frac{\kappa p_n \sum_j \gamma_j}{|\mathbf{p}| \gamma_n}, \quad n = 1, 2, \dots, \quad \text{with } \kappa = \frac{1 - \sum_n \gamma_n}{2 \sum_n \gamma_n}$$

Observe that $\mathbf{r} \in \mathcal{A}_{\gamma}$ since

$$\sum_n \gamma_n r_n = \sum_n \gamma_n (1 + \kappa) + \kappa \sum_n \gamma_n \frac{p_n \sum_j \gamma_j}{|\mathbf{p}| \gamma_n} = \sum_n \gamma_n (1 + 2\kappa) = 1$$

Moreover, since $r_n \geq 1 + \kappa > 1$ and $r/(r - 1)$ is a decreasing function, we have,

$$\frac{\pi r_i}{2(r_i - 1)} \sqrt{2p_i + 1} \leq \frac{\pi(1 + \kappa)}{2\kappa} \sqrt{2p_i + 1} \leq C_k^{p_i}, \quad C_{\kappa} := \frac{3\pi(1 + \kappa)}{2\kappa}$$

hence, using the bound (from Stirling approximation) $n! \leq n^n \leq n! e^n$,

$$\frac{\|u_{\mathbf{p}}\|_V}{\|u\|_{A,\gamma}} \leq \prod_{n \in \text{supp}(\mathbf{p})} \left(\frac{C_{\kappa}}{r_n} \right)^{p_i} \leq \prod_{n \in \text{supp}(\mathbf{p})} \left(\frac{C_k \gamma_n |\mathbf{p}|}{\kappa p_n \sum_j \gamma_j} \right)^{p_n} \leq \frac{|\mathbf{p}|!}{\mathbf{p}!} \prod_{n \in \text{supp}(\mathbf{p})} \left(\frac{e C_k \gamma_n}{\kappa \sum_j \gamma_j} \right)^{p_n}$$

We now check the τ -summability of the Legendre coefficients. Set

$$\beta_n = \frac{eC_k\gamma_n}{\kappa \sum_j \gamma_j} = \frac{3e\pi(1+\kappa)\gamma_n}{2\kappa^2 \sum_j \gamma_j}.$$

$$\sum_{\mathbf{p} \in \mathbb{N}^{\mathbb{N}}} \|u_{\mathbf{p}}\|_V^\tau \leq \|u\|_{A,\gamma}^\tau \sum_{\mathbf{p} \in \mathbb{N}^{\mathbb{N}}} \left(\frac{|\mathbf{p}|!}{\mathbf{p}!} \right)^\tau \prod_{n \in \text{supp}(\mathbf{p})} \beta_n^{\tau p_i} < \infty \quad ?$$

Lemma 13.7 ([Cohen-Devore-Schwab FoCM '10, Th. 7.2])

The sequence $\{ \frac{|\mathbf{p}|!}{\mathbf{p}!} \prod_n \beta_n \}_{\mathbf{p} \in \mathbb{N}^{\mathbb{N}}} \text{ is } \tau\text{-summable iff the sequence } \{\beta_n\}_{n \in \mathbb{N}} \text{ is } \tau\text{-summable and } \sum_{n=1}^{\infty} |\beta_n| < 1$. Moreover, denoting with $B_\tau = (\sum_n |\beta_n|^\tau)^{\frac{1}{\tau}}$,

$$\sum_{\mathbf{p} \in \mathbb{N}^{\mathbb{N}}} \left(\frac{|\mathbf{p}|!}{\mathbf{p}!} \right)^\tau \prod_{n \in \text{supp}(\mathbf{p})} \beta_n^{\tau p_i} \leq \frac{2}{1 - B_1} \exp \left\{ \frac{2(1-\tau)(J(B_1) + B_\tau^\tau)}{\tau^2(1-B_1)} \right\}$$

with $J(B_1)$ such that $\sum_{j > J(B_1)} |\beta_j|^\tau \leq (1 - B_1)/2$.

Now, the summability of $\{\beta_n\}_n$ is the same as the summability of $\{\gamma_n\}_n$.

On the other hand, $\sum_{n=1}^{\infty} \beta_n \leq 1 \implies \frac{3e\pi(1+k)}{2k^2} \leq 1$ which holds for instance for $k \geq (3e\pi + 1)/2$ which implies $\sum_{n=1}^{\infty} \gamma_n \leq 1/(3e\pi + 2)$.

Remark. To avoid the constraint $\sum_{n=1}^{\infty} \gamma_n \leq 1/(3e\pi + 2)$, one has to use the choice $r_n = 1 + \kappa + \frac{\kappa p_n \sum_j \gamma_j}{|\mathbf{p}| \gamma_n}$ only for $n > J$ give precise estimate and $r_n = 1 + \kappa$ for $1 \leq n \leq J$, with J suitably chosen.

Example – elliptic PDE with random diffusivity coefficient

Let's consider our running toy problem again. That is,

$$\begin{cases} -\operatorname{div}(a(x, \mathbf{y}) \nabla u(x, \mathbf{y})) = f(x), & x \in D, \\ u(x, \mathbf{y}) = 0, & x \in \partial D \end{cases}, \quad \forall \mathbf{y} \in \Gamma := [-\sqrt{3}, \sqrt{3}]^{N=\infty}$$

with $a(x, \mathbf{y}) = \bar{a} + \sum_{i=1}^{\infty} \sqrt{3\lambda_i} y_i b_i(x)$, $y_i \sim \mathcal{U}(-1, 1)$ i.i.d.,

$\sum_{i=1}^{\infty} \sqrt{3\lambda_i} \|b_i\|_{\infty} \leq \delta \bar{a}$ for some $0 < \delta < 1$, so that

$$\|\nabla u(\mathbf{y})\|_{L^2(D)} \leq C_u := \frac{C_P}{(1-\delta)\bar{a}} \|f\|_{L^2(D)}.$$

Theorem 13.8

Let $\gamma_i = \frac{\sqrt{3\lambda_i} \|b_i\|_{\infty}}{\delta \bar{a}}$. The solution $\mathbf{y} \mapsto u(\mathbf{y})$ admits an analytic continuation in any poly-ellipse \mathcal{E}_r with $\sum_{n=1}^{\infty} \gamma_n r_n \leq 1$, with $\sum_{n=1}^{\infty} \gamma_n \leq 1$.

Therefore, the best M-term approximation converges at rate

$$\|u - u_{\Lambda_{opt}(M)}\|_{L^2(\Gamma, V)} \leq C(u, \gamma, \tau) M^{\frac{1}{2} - \frac{1}{\tau}}$$

provided that $\sum_{n=1}^{\infty} (\sqrt{\lambda_n} \|b_n\|_{\infty})^{\tau} < \infty$.

Proof.

As before, we analyze the equation in the complex domain. Let us replace each variable y_j with a complex variable $z_j = y_j + iw_j$. As we have seen before, the problem is well posed for any \mathbf{z} such that $\min_{x \in D} \Re(a(x, \mathbf{z})) > 0$. Moreover, the solution will be complex differentiable and hence analytic.

Next, $\sum_{i=1}^{\infty} \gamma_i |\Re(z_i)| \leq 1$ implies that

$$\min_{x \in D} \Re(a(x, \mathbf{z})) \geq \bar{a} - \sum_{i=1}^{\infty} \sqrt{3\lambda_i} \|\beta_i\|_{\infty} |\Re(z_i)| = \bar{a} - \delta \bar{a} \sum_{i=1}^{\infty} \gamma_i |\Re(z_i)| \geq (1-\delta) \bar{a},$$

so that $\mathbf{z} \mapsto u(\mathbf{z})$ is analytic in the region

$$\Sigma = \{\mathbf{z} \in \mathbb{C}^N, \sum_{i=1}^{\infty} \gamma_i |\Re(z_i)| \leq 1\} \text{ and}$$

$$\max_{\mathbf{z} \in \Sigma} \|u(\mathbf{z})\|_{A, \gamma} \leq \frac{c_p}{(1-\delta)\bar{a}} \|f\|_{L^2(D)}.$$

It is easy to show that the set Σ contains any polyellipse \mathcal{E}_r such that $\sum_{n=1}^{\infty} \gamma_n r_n \leq 1$. Indeed,

$$\mathbf{z} \in \mathcal{E}_r \Rightarrow |\Re(z_i)| \leq \frac{r_i + r_i^{-1}}{2} \leq r_i, \quad \forall i \geq 1$$

$$\Rightarrow \sum_{i=1}^{\infty} \gamma_i |\Re(z_i)| \leq \sum_{i=1}^{\infty} \gamma_i r_i \leq 1.$$

Optimal polynomial space

In the previous analysis, we have derived the following bound for the Legendre coefficients

$$\|u_{\mathbf{p}}\|_V \leq \|u\|_{A,\gamma} \frac{|\mathbf{p}|!}{\mathbf{p}!} \prod_{n \in \text{supp}(\mathbf{p})} (\alpha \gamma_n)^{p_n} = \|u\|_{A,\gamma} \frac{|\mathbf{p}|!}{\mathbf{p}!} \prod_{n \in \text{supp}(\mathbf{p})} e^{\sum_n p_n \log(\alpha \gamma_n)}$$

for a suitable coefficient α .

This shows that the optimal set $\Lambda_{opt}(M)$ contains the multi-indices

$$\frac{|\mathbf{p}|!}{\mathbf{p}!} \prod_{n \in \text{supp}(\mathbf{p})} e^{\sum_n p_n \log(\alpha \gamma_n)} \geq \epsilon$$

This estimate has been used in [Beck-Nobile-Tamellini-Tempone, M3AS '12] and provides accurate results (i.e., the approximation space built upon these estimates is close to the best M-term approximation).

A simple numerical test

We consider the 1D problem

$$\begin{cases} -(a(x, \mathbf{y})u(x, \mathbf{y})')' = 1 & x \in D = (0, 1), \mathbf{y} \in \Gamma \\ u(0, \mathbf{y}) = u(1, \mathbf{y}) = 0, & \mathbf{y} \in \Gamma \end{cases}$$

with $a(\mathbf{x}, \mathbf{y}) = 1 + 0.1y_1 + 0.5y_2$, $y_1, y_2 \sim \mathcal{U}(-1, 1)$ and compute $Q(u) = u(\frac{1}{2})$.

We compare:

► (Aniso) TD space: $\Lambda(w) = \left\{ \mathbf{p} \in \mathbb{N}^N : \sum_n g_n p_n \leq w \right\}$.

► (Iso) TD space: $\Lambda(w) = \left\{ \mathbf{p} \in \mathbb{N}^N : \sum_n p_n \leq w \right\}$.

► (Aniso) TD-FC space:

$$\Lambda(w) = \left\{ \mathbf{p} \in \mathbb{N}^N : \sum_{n=1}^N g_n p_n - \log \frac{|\mathbf{p}|!}{\mathbf{p}!} \leq w \right\}.$$

The rates g_n have been estimated numerically by inexpensive 1D analyses.

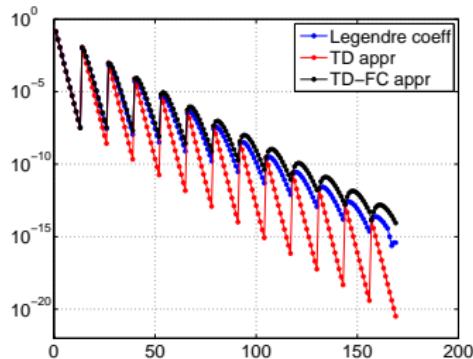


Figure: Legendre coeffs of $Q(u)$ in lexicographic order, with TD and TD-FC estimates

The Legendre coefficients have been computed with a sufficiently high level sparse grids.

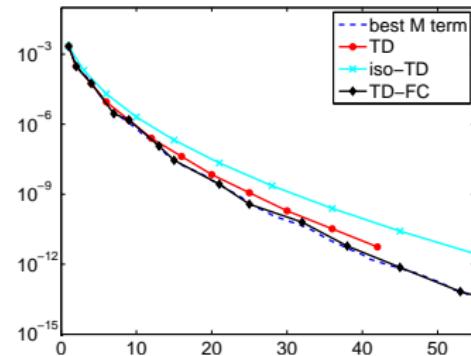
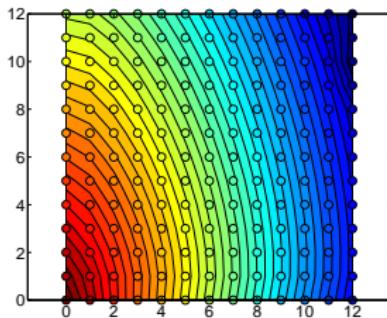
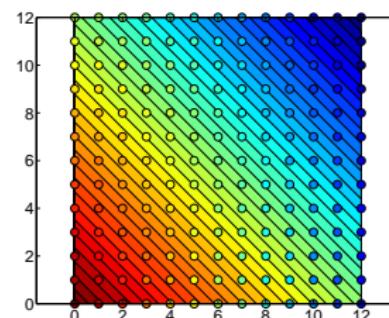


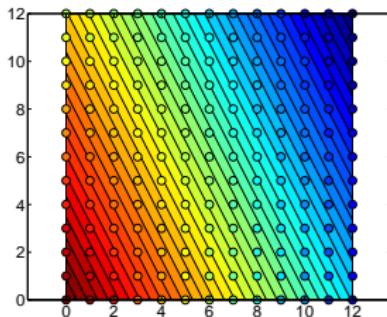
Figure: Convergence plot for $\|Q(u) - Q(u_\Lambda)\|_{L_\rho^2(\Gamma)}^2$ w.r.t. $M = \#\Lambda$



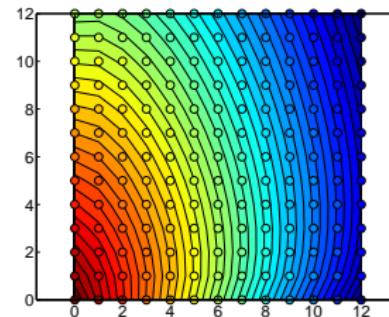
“true” Legendre coeffs.



iso-TD estimate.



aniso TD estimate



TD-FC estimate.

Discrete L^2 projection using random evaluations:
computable approximations to L^2 projections

Why is this useful?

1. L^2 is a popular loss function, usually motivated by Maximum Likelihood and Gaussian, additive noise assumptions.
2. This loss function is commonplace in machine learning as well.

How to make L^2 approximations computable

- ▶ The estimates seen so far on L^2 approximation are only theoretical as the “exact” L^2 projection can not be computed.
- ▶ we look now at **computable** methods to obtain polynomial approximations, hopefully close enough to the L^2 projection (which is the best approximation in the L^2 sense)
- ▶ In particular, we look at computable methods based on point evaluations of the map $u(\mathbf{y})$ on a set of **random** points
(see next chapter for approximations based on deterministic points and interpolation-type methods)

Setting: Consider again a Hilbert-valued function $u(\mathbf{y}) : \Gamma \rightarrow V$, $u \in L^2_\rho(\Gamma; V)$, where $\mathbf{y} \in \Gamma \subset \mathbb{R}^N$ is a random vector with joint probability density function $\rho : \Gamma \rightarrow \mathbb{R}_+$.

1. Generate M random i.i.d. points $\mathbf{y}(\omega_k) \sim \rho(\mathbf{y})d\mathbf{y}$, $k = 1, \dots, M$
2. Compute the corresponding solutions $u^{(k)} = u(\mathbf{y}(\omega_k)) \in V$, $k = 1, \dots, M$ (this implies solving M PDEs)
3. Construct a suitable approximation $\Pi_\Lambda^M u \in \mathbb{P}_\Lambda(\Gamma) \otimes V$ based on the point evaluations

How to make L^2 approximations computable

Let $\{\psi_p\}_{p \in \Lambda}$ be an orthonormal basis of a given subspace \mathbb{P}_Λ w.r.t the weight ρ .

Then, the exact L^2 projection $\Pi_\Lambda u$ of u in $\mathbb{P}_\Lambda(\Gamma) \otimes V$ (best approximation) is

$$\Pi_\Lambda u = \sum_{p \in \Lambda} \mathbb{E}[u\psi_p]\psi_p = \arg \min_{v \in \mathbb{P}_\Lambda(\Gamma) \otimes V} \mathbb{E}[\|u - v\|_V^2]$$

How to compute an approximated L^2 projection using the evaluations $u^{(k)}$ on random points?

General Idea: Approximate the expectations above, $\mathbb{E}[v]$, using Monte Carlo:

$$\mathbb{E}_M[v] = \frac{1}{M} \sum_{i=1}^M v(\mathbf{y}(\omega_i))$$

In which of the two formulas should we use Monte Carlo Sampling?

First idea (bad): use MCS on Fourier coefficients

We construct an approximation as

$$\Pi_{\Lambda}^M u = \sum_{\mathbf{p} \in \Lambda} \mathbb{E}_M[u\psi_{\mathbf{p}}] \psi_{\mathbf{p}} = \sum_{\mathbf{p} \in \Lambda} \left(\frac{1}{M} \sum_{i=1}^M u(\mathbf{y}(\omega_i)) \psi_{\mathbf{p}}(\mathbf{y}(\omega_i)) \right) \psi_{\mathbf{p}}$$

Notation: $u_{\mathbf{p}} := \mathbb{E}[u\psi_{\mathbf{p}}]$ (exact Fourier coeff.)

$u_{\mathbf{p}}^M := \mathbb{E}_M[u\psi_{\mathbf{p}}]$ (approximate Fourier coeff.)

Error analysis: (here $u : \Gamma \rightarrow \mathbb{R}$ is a scalar function for simplicity)

$$\begin{aligned} \|u - \Pi_{\Lambda}^M u\|_{L_p^2(\Gamma; \mathbb{R})}^2 &= \left\| \sum_{\mathbf{q} \in \mathbb{N}^N} u_{\mathbf{q}} \psi_{\mathbf{q}} - \sum_{\mathbf{p} \in \Lambda} u_{\mathbf{p}}^M \psi_{\mathbf{p}} \right\|_{L_p^2(\Gamma; \mathbb{R})}^2 \\ &= \underbrace{\sum_{\mathbf{q} \notin \Lambda} |u_{\mathbf{q}}|^2}_{L^2 \text{ projection error}} + \sum_{\mathbf{p} \in \Lambda} \underbrace{|u_{\mathbf{p}} - u_{\mathbf{p}}^M|^2}_{\text{Monte Carlo error} \sim O(M^{-1})} \end{aligned}$$

Observe now that

$$\mathbb{E}[u_{\mathbf{p}}^M] = u_{\mathbf{p}} \quad \text{and} \quad \mathbb{E}[(u_{\mathbf{p}}^M - u_{\mathbf{p}})^2] = \frac{\text{Var}(u\psi_{\mathbf{p}})}{M} \leq \frac{\mathbb{E}[u^2\psi_{\mathbf{p}}^2]}{M}$$

so that

$$M \sum_{\mathbf{p} \in \Lambda} \mathbb{E}[(u_{\mathbf{p}}^M - u_{\mathbf{p}})^2] \leq \sum_{\mathbf{p} \in \Lambda} \mathbb{E}[u^2\psi_{\mathbf{p}}^2] \leq \|u\|_{L^2_\rho(\Gamma; \mathbb{R})}^2 \sup_{\mathbf{y} \in \Gamma} \left(\sum_{\mathbf{p} \in \Lambda} |\psi_{\mathbf{p}}(\mathbf{y})|^2 \right)$$

Define now the quantity

$$K(\Lambda) = \sup_{\mathbf{y} \in \Gamma} \left(\sum_{\mathbf{p} \in \Lambda} |\psi_{\mathbf{p}}(\mathbf{y})|^2 \right)$$

Then,

$$\mathbb{E}[\|u - \Pi_{\Lambda}^M u\|_{L^2_\rho(\Gamma; \mathbb{R})}^2] \leq \inf_{v \in \mathbb{P}_{\Lambda}(\Gamma)} \|u - v\|_{L^2_\rho(\Gamma; \mathbb{R})}^2 + \frac{K(\Lambda)}{M} \|u\|_{L^2_\rho(\Gamma; \mathbb{R})}^2$$

Alternative estimate if $u \in L^\infty(\Gamma; \mathbb{R})$:

$$\mathbb{E}[\|u - \Pi_{\Lambda}^M u\|_{L^2_\rho(\Gamma; \mathbb{R})}^2] \leq \inf_{v \in \mathbb{P}_{\Lambda}(\Gamma)} \|u - v\|_{L^2_\rho(\Gamma; \mathbb{R})}^2 + \frac{|\Lambda|}{M} \|u\|_{L^\infty(\Gamma; \mathbb{R})}^2$$

Even for smooth functions the convergence is only $O(\sqrt{|\Lambda|/M}) \dots$

Second idea (good): Discrete least squares approximation

(see e.g. [Hosder-Walters et al. 2010, Blatman-Sudret 2008, Burkardt-Eldred 2009, Eldred 2011, Yan-Guo-Xiu 2012, Cohen-Davenport-Leviatan 2013, Migliorati et al 2011-2014])

$$\Pi_{\Lambda}^M u = \arg \min_{v \in \mathbb{P}_{\Lambda}(\Gamma) \otimes V} \frac{1}{M} \sum_{k=1}^M \|u^{(k)} - v(\mathbf{y}(\omega_k))\|_V^2$$

Two relevant questions

- ▶ What is the accuracy of the random discrete least square approximation?
- ▶ For a given set Λ , how many samples should one use?

Discrete L^2 projection

Notation

- continuous norm: $\|v\|_{L^2_\rho(\Gamma, V)}^2 = \mathbb{E}[\|v\|_V^2] = \int_\Gamma \|v(\mathbf{y})\|_V^2 \rho(\mathbf{y}) d\mathbf{y}$
- discrete norm: $\|v\|_{M, V}^2 = \mathbb{E}_M[\|v\|_V^2] = \frac{1}{M} \sum_{i=1}^M \|v(\mathbf{y}(\omega_i))\|_V^2$

We have:

$$\|u - \Pi_\Lambda^M u\|_{M, V}^2 + \|v - \Pi_\Lambda^M u\|_{M, V}^2 = \|u - v\|_{M, V}^2, \quad \forall v \in \mathbb{P}_\Lambda(\Gamma) \otimes V$$

$$\|u - \Pi_\Lambda^M u\|_{M, V} \leq \|u - v\|_{M, V}, \quad \forall v \in \mathbb{P}_\Lambda(\Gamma) \otimes V$$

$$\|v - \Pi_\Lambda^M u\|_{M, V} \leq \|u - v\|_{M, V}, \quad \forall v \in \mathbb{P}_\Lambda(\Gamma) \otimes V$$

Algebraic formulation: Approximation of QoI

Observe that the problem can be posed directly in terms of a Quantity of Interest given by $\varphi(\mathbf{y}) = Q(u(\mathbf{y}))$, where Q is a globally Lipschitz functional. In such a case, we simply find a real valued discrete least square approximation, $\Pi_{\Lambda}^M \varphi \in \mathbb{P}_{\Lambda}(\Gamma)$, which satisfies

$$\Pi_{\Lambda}^M \varphi = \arg \min_{v \in \mathbb{P}_{\Lambda}(\Gamma)} \frac{1}{M} \sum_{i=1}^M |\varphi^{(i)} - v(\mathbf{y}(\omega_i))|^2$$

For an orthonormal basis $\{\psi_{\mathbf{p}}\}_{\mathbf{p} \in \Lambda}$ of $\mathbb{P}_{\Lambda}(\Gamma)$, define the

Design matrix : $D \in \mathbb{R}^{|\Lambda| \times M}$, $D_{i\mathbf{p}} = \psi_{\mathbf{p}}(\mathbf{y}(\omega_i))$, $\mathbf{p} \in \Lambda, 1 \leq i \leq M$.

Then, expanding $\Pi_{\Lambda}^M \varphi$ onto the basis

$$\Pi_{\Lambda}^M \varphi(\mathbf{y}) = \sum_{\mathbf{p} \in \Lambda} c_{\mathbf{p}} \psi_{\mathbf{p}}(\mathbf{y}),$$

the vector $\mathbf{c} = \{c_{\mathbf{p}}\}$ of Fourier coefficients satisfies the normal equations

$$(D^T D) \mathbf{c} = (D^T) \varphi$$

with $(\varphi)_i = \varphi(\mathbf{y}(\omega_i))$.

On the normal equations (still in the scalar case)

Rewrite the normal equations as

$$G\mathbf{c} = J\varphi, \quad \text{with } G = \frac{1}{M}D^T D, \quad J = \frac{1}{M}D^T$$

- ▶ G is symmetric and (semi)-positive definite.
- ▶ The stability of the discrete least squares is related to $\|G^{-1}\|$.
- ▶ Let $\mathbf{v} \in \mathbb{P}_\Lambda(\Gamma)$, $\mathbf{v} = \sum_{\mathbf{p} \in \Lambda} v_{\mathbf{p}} \psi_{\mathbf{p}}$ and $\mathbf{v} = \{v_{\mathbf{p}}\} \in \mathbb{R}^{\#\Lambda}$. Then

$$\mathbf{v}^T G \mathbf{v} = \frac{1}{M} \|D\mathbf{v}\|^2 = \frac{1}{M} \sum_{i=1}^M \left(\sum_{\mathbf{p} \in \Lambda} \psi_{\mathbf{p}}(\mathbf{y}(\omega_i)) v_{\mathbf{p}} \right)^2 = \|\mathbf{v}\|_{M,\mathbb{R}}^2$$

$$\text{and } \mathbf{v}^T \mathbf{v} = \sum_{\mathbf{p} \in \Lambda} v_{\mathbf{p}}^2 = \|\mathbf{v}\|_{L_\rho^2(\Gamma, \mathbb{R})}^2$$

- ▶ It follows that

$$\|G\| = \sup_{\mathbf{v} \in \mathbb{R}^{\#\Lambda}} \frac{\mathbf{v}^T G \mathbf{v}}{\mathbf{v}^T \mathbf{v}} = \sup_{\mathbf{v} \in \mathbb{P}_\Lambda(\Gamma)} \frac{\|\mathbf{v}\|_{M,\mathbb{R}}^2}{\|\mathbf{v}\|_{L_\rho^2(\Gamma, \mathbb{R})}^2},$$

$$\|G^{-1}\| = \left(\inf_{\mathbf{v} \in \mathbb{R}^{\#\Lambda}} \frac{\mathbf{v}^T G \mathbf{v}}{\mathbf{v}^T \mathbf{v}} \right)^{-1} = \sup_{\mathbf{v} \in \mathbb{P}_\Lambda(\Gamma)} \frac{\|\mathbf{v}\|_{L_\rho^2(\Gamma, \mathbb{R})}^2}{\|\mathbf{v}\|_{M,\mathbb{R}}^2}$$

Algebraic formulation: Approximation of u [Optional]

The formulation is analogous in the case of a Hilbert space valued function $u(\mathbf{y}) : \Gamma \rightarrow V$.

Expand $\Pi_{\Lambda}^M u$ on the basis:

$$\Pi_{\Lambda}^M u = \sum_{\mathbf{p} \in \Lambda} c_{\mathbf{p}} \psi_{\mathbf{p}}, \quad \text{with functional coeffs } c_{\mathbf{p}} = c_{\mathbf{p}}(x) \in V.$$

Then the vector of functions $\mathbf{c}(x) = \{c_{\mathbf{p}}(x)\}_{\mathbf{p} \in \Lambda}$ satisfies the normal equations

$$D^T D \mathbf{c}(x) = D^T \mathbf{u}(x), \quad \forall x \in D$$

with $\mathbf{u}_i(x) = u(x, \mathbf{y}(\omega_i))$.

Similarly as in the scalar case, given $v = \sum_{\mathbf{p} \in \Lambda} v_{\mathbf{p}} \psi_{\mathbf{p}}$ and $\mathbf{v} = \{v_{\mathbf{p}}\}$,

$$(G\mathbf{v}, \mathbf{v})_{V^{\# \Lambda}} = \|v\|_{M,V}^2, \quad (\mathbf{v}, \mathbf{v})_{V^{\# \Lambda}} = \|v\|_{L_{\rho}^2(\Gamma; V)}^2$$

and

$$\|G\| = \sup_{v \in \mathbb{P}_{\Lambda}(\Gamma) \otimes V} \frac{\|v\|_{M,V}^2}{\|v\|_{L_{\rho}^2(\Gamma; V)}^2}, \quad \|G^{-1}\| = \sup_{v \in \mathbb{P}_{\Lambda}(\Gamma) \otimes V} \frac{\|v\|_{L_{\rho}^2(\Gamma; V)}^2}{\|v\|_{M,V}^2}$$

Error analysis

Stability of the discrete least squares projection implies convergence!

Theorem [Migliorati-Nobile-von Schwerin-Tempone '11]

1. $\|G\|, \|G^{-1}\| \rightarrow 1$ almost surely when $M \rightarrow \infty$ (i.e. $G \rightarrow I$)
2. $\|u - \Pi_{\Lambda}^M u\|_{L_p^2(\Gamma, V)} \leq (1 + \|G^{-1}\|) \inf_{v \in \mathbb{P}_{\Lambda}(\Gamma) \otimes V} \|u - v\|_{L^{\infty}(\Gamma, V)}$

Proof: for any $v \in \mathbb{P}_{\Lambda} \otimes V$:

$$\begin{aligned} \|u - \Pi_{\Lambda}^M u\|_{L_p^2(\Gamma, V)} &\leq \|u - v\|_{L_p^2(\Gamma, V)} + \|v - \Pi_{\Lambda}^M u\|_{L_p^2(\Gamma, V)} , \text{ Triangular ineq.} \\ &\leq \|u - v\|_{L_p^2(\Gamma, V)} + \frac{\|v - \Pi_{\Lambda}^M u\|_{L_p^2(\Gamma, V)}}{\|v - \Pi_{\Lambda}^M u\|_{M, V}} \|v - \Pi_{\Lambda}^M u\|_{M, V} \\ &\leq \|u - v\|_{L_p^2(\Gamma, V)} + \|G^{-1}\| \|v - \Pi_{\Lambda}^M u\|_{M, V} , \text{ Operator norm} \\ &\leq \|u - v\|_{L_p^2(\Gamma, V)} + \|G^{-1}\| \|u - v\|_{M, V} , \text{ Optimality of discrete projection} \end{aligned}$$

Remark: $\|G^{-1}\|$ is a random quantity (G is a random matrix since we are sampling the evaluation points with MC). Therefore we need to understand with what probability $\|G^{-1}\|$ will be sufficiently small.

A simple 1D setting: quasi uniform random variable

A first estimate of $\|G^{-1}\|$ can be obtained in the monovariate case using order statistics of uniformly distributed points in $[-1, 1]$.

- ▶ $\mathbb{P}_\Lambda = \mathbb{P}_{\textcolor{blue}{w}}$ = space of polynomials of degree at most $\textcolor{blue}{w}$
- ▶ **Quasi uniform distribution:** Γ is a bounded interval and $0 < \rho_{min} \leq \rho(\mathbf{y}) \leq \rho_{max} < \infty$, for all $\mathbf{y} \in \Gamma$.

Theorem [Migliorati-Nobile-von Schwerin-Tempone FoCM '14]

For any $\alpha \in (0, 1)$, let M be such that

$$\frac{2M\rho_{min}}{3\log((M+1)/\alpha)} \geq \frac{8\sqrt{3}}{|\Gamma|} \textcolor{blue}{w}^2 \quad (*)$$

Then, it holds

$$P \left(\|u - \Pi_w^M u\|_{L_\rho^2(\Gamma, V)} \leq \left(1 + \sqrt{\frac{3\rho_{max}}{\rho_{min}} \log \frac{M+1}{\alpha}} \right) \inf_{v \in \mathbb{P}_{\textcolor{blue}{w}} \otimes V} \|u - v\|_{L^\infty(\Gamma, V)} \right) \geq 1 - \alpha.$$

Hint of the proof [Optional]

First step: Estimate the maximum distance between any two consecutive y -sample points $\{Y_j\}$. Let $F(y) = \int_{-\infty}^y \rho(z)dz$ be the cumulative distribution,

$U_i = F^{-1}(y(\omega_i))$ the samples mapped onto $[0, 1]$, which are uniformly distributed, and $U_{(i)}$ the order statistics: $0 = U_{(0)} \leq U_{(1)} \leq \dots \leq U_{(M)} \leq U_{(M+1)} = 1$.

Then, $\Delta U_i = U_{(i+1)} - U_{(i)} \sim \text{beta}(1, M)$ and

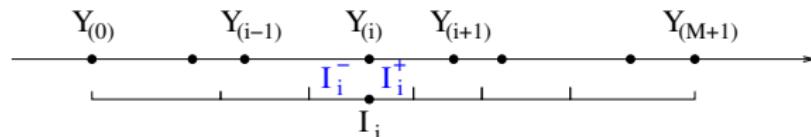
$$P\left(\max_{i=0, \dots, M} \Delta U_i > \delta\right) \leq \sum_{i=1}^M P(\Delta U_i > \delta) = (M+1)(1-\delta)^M$$

so that for any $\alpha \in (0, 1)$, $P\left(\max_{i=0, \dots, M} \Delta U_i > \frac{\log((M+1)/\alpha)}{M}\right) \leq \alpha$ and

$$P\left(\max_{i=0, \dots, M} (Y_{(i+1)} - Y_{(i)}) > \frac{\log((M+1)/\alpha)}{M\rho_{\min}}\right) \leq \alpha$$

Second step: Introduce a (random) covering of Γ . We define

$$I_i = \left[\frac{Y_{(i-1)} + Y_{(i)}}{2}, \frac{Y_{(i)} + Y_{(i+1)}}{2}\right]$$



Clearly

$$\max_i |I_i| \leq \frac{3}{2} \max_i \Delta Y_{(i)}, \quad \text{and} \quad P\left(\max_{i=0, \dots, M} |I_i| > \frac{3 \log((M+1)/\alpha)}{2M\rho_{\min}}\right) \leq \alpha$$

Hint of the proof

Third step: Prove the equivalence of norms $\|v\|_M \sim \|v\|_{L_p^2}$ on \mathbb{P}_w .

$$\begin{aligned}\|v\|_{L^2}^2 &= \sum_{j=1}^M \int_{I_j} v(y)^2 dy \quad \left[\text{use } \int_{I_j} v(y)^2 dy \leq |I_j| (v(y_j)^2 + 2\|v\|_{I_j} \|v'\|_{I_j}) \right] \\ &\leq \sum_{j=1}^M |I_j| v(y_j)^2 + 2 \sum_{j=1}^M |I_j| \|v\|_{I_j} \|v'\|_{I_j} \quad \left[\text{setting } I_{max} = \max_i |I_i| \right] \\ &\leq M I_{max} \|v\|_M^2 + 2 I_{max} \sum_{j=1}^M \|v\|_{I_j} \|v'\|_{I_j} \quad \left[\text{disc. Cauchy-Schwarz ineq.} \right] \\ &\leq M I_{max} \|v\|_M^2 + 2 I_{max} \|v\|_{L^2} \|v'\|_{L^2} \quad \left[\text{inv.ineq. with constant } \frac{2\sqrt{3}}{|\Gamma|} w^2 \right] \\ &\leq M I_{max} \|v\|_M^2 + I_{max} \frac{4\sqrt{3}}{|\Gamma|} w^2 \|v\|_{L^2}^2.\end{aligned}$$

Therefore, under the condition (*) of the theorem and the previous result:

$$\|G^{-1}\| := \sup_{v \in \mathbb{P}_w} \frac{\|v\|_{L_p^2}^2}{\|v\|_M^2} \leq \frac{M I_{max} \rho_{max}}{1 - 4 I_{max} \sqrt{3} w^2 / |\Gamma|} \leq \frac{3 \rho_{max} \log((M+1)/\alpha)}{\rho_{min}}$$

with probability smaller than α .

General theory

[Cohen-Davenport-Leviatan '12], [Chkifa-Cohen-Migliorati-Nobile-Tempone '14]

Goal: obtain conditions under which

$$\|G - I\| \leq \delta, \quad \delta \in (0, 1)$$

Observe that this bound implies

$$\|G\| \leq 1 + \delta, \quad \|G^{-1}\| \leq \frac{1}{1 - \delta}, \quad \text{cond}(G) \leq \frac{1 + \delta}{1 - \delta}$$

and the equivalence of norms on $\mathbb{P}_\Lambda(\Gamma) \otimes V$

$$(1 - \delta)\|v\|_{L_p^2(\Gamma; V)}^2 \leq \|v\|_{M, V}^2 \leq (1 + \delta)\|v\|_{L_p^2(\Gamma; V)}^2, \quad \forall v \in \mathbb{P}_\Lambda \otimes V$$

The previous condition is analogous to the Restricted Isometry Property (RIP) in compressed sensing, see [Candès-Tao '06, Rauhut-Ward '12, ...]

More on the matrix G

Recall $G = \frac{1}{M} D^T D$. Hence

$$G_{\mathbf{pq}} = \frac{1}{M} \sum_{i=1}^M D_{i\mathbf{p}} D_{i\mathbf{q}} = \frac{1}{M} \sum_{i=1}^M G_{\mathbf{pq}}^{(i)}, \quad \text{with } G_{\mathbf{pq}}^{(i)} = \psi_{\mathbf{q}}(\mathbf{y}(\omega_i)) \psi_{\mathbf{p}}(\mathbf{y}(\omega_i))$$

Remarks:

- ▶ The matrices $G^{(i)}$, $i = 1, \dots, M$ are i.i.d.
- ▶ $\mathbb{E}[G^{(i)}] = I$. Indeed $\mathbb{E}[G_{\mathbf{pq}}^{(i)}] = \mathbb{E}[\psi_{\mathbf{p}} \psi_{\mathbf{q}}] = \delta_{\mathbf{pq}}$. **remember, $(\psi_{\mathbf{p}})$ onb**
- ▶ G is the sample average of i.i.d. random matrices

$$G = \frac{1}{M} \sum_{i=1}^M G^{(i)} \quad \text{and } \mathbb{E}[G] = I$$

- ▶ Results on $\|G - I\| = \|G - \mathbb{E}[G]\|$ can be obtained from concentration of measure results (LDT) for sums of independent matrices.

A uniform bound on $G^{(i)}$

Define

$$K(\Lambda) = \sup_{\mathbf{y} \in \Gamma} \left(\sum_{\mathbf{p} \in \Lambda} |\psi_{\mathbf{p}}(\mathbf{y})|^2 \right) = \sup_{v \in \mathbb{P}_{\Lambda}} \frac{\|v\|_{L^{\infty}(\Gamma)}^2}{\|v\|_{L_{\rho}^2(\Gamma)}^2}$$

Remark: the constant $K(\Lambda)$ can be defined starting from [any](#) orthonormal basis of \mathbb{P}_{Λ} , $\{\psi_{\mathbf{p}}\}_{\mathbf{p} \in \Lambda}$, and is actually a property of the polynomial space \mathbb{P}_{Λ} and the sampling measure $\rho(\mathbf{y})d\mathbf{y}$.

To see this, we have by Cauchy-Schwarz inequality for any

$$v = \sum_{\mathbf{p} \in \Lambda} v_{\mathbf{p}} \psi_{\mathbf{p}} \in \mathbb{P}_{\Lambda}(\Gamma) \otimes V$$

$$\|v\|_{L^{\infty}(\Gamma, V)}^2 = \sup_{\mathbf{y} \in \Gamma} \left| \sum_{\mathbf{p} \in \Lambda} v_{\mathbf{p}} \psi_{\mathbf{p}}(\mathbf{y}) \right|^2 \leq \sup_{\mathbf{y} \in \Gamma} \sum_{\mathbf{p} \in \Lambda} |\psi_{\mathbf{p}}(\mathbf{y})|^2 \sum_{\mathbf{q} \in \Lambda} |v_{\mathbf{q}}|^2 \leq K(\Lambda) \|v\|_{L_{\rho}^2(\Gamma, V)}^2,$$

with equality when $\mathbf{y} = \mathbf{y}^*$ is the point of Γ where the supremum in $K(\Lambda)$ is attained and the function v is defined by the coefficients $v_{\mathbf{p}} = \psi_{\mathbf{p}}(\mathbf{y}^*)$.

Then, [we have the following uniform bound:](#)

$$\|G^{(i)}\| = \sup_{v \in \mathbb{P}_{\Lambda}(\Gamma) \otimes V} \frac{\|v(y(\omega_i))\|_V^2}{\|v\|_{L_{\rho}^2(\Gamma, V)}^2} \leq \sup_{v \in \mathbb{P}_{\Lambda}(\Gamma) \otimes V} \frac{\|v\|_{L^{\infty}(\Gamma, V)}^2}{\|v\|_{L_{\rho}^2(\Gamma, V)}^2} = K(\Lambda)$$

Chernoff's bound (scalar case)

Chernoff's bound (for i.i.d. random variables)

Let X_1, \dots, X_M be i.i.d. random variables s.t. $P(X_i \in [0, R]) = 1$.

Let $\mu = \mathbb{E}[X_i]$ and $\bar{X} = \frac{1}{M} \sum_{i=1}^M X_i$. Then

$$P(\bar{X} \geq (1 + \delta)\mu) \leq \exp \left\{ -\frac{M\mu\tilde{\beta}_\delta}{R} \right\}, \quad \tilde{\beta}_\delta = (1 + \delta) \log(1 + \delta) - \delta \approx \delta^2$$

$$P(\bar{X} \leq (1 - \delta)\mu) \leq \exp \left\{ -\frac{M\mu\beta_\delta}{R} \right\}, \quad \beta_\delta = \delta + (1 - \delta) \log(1 - \delta) \approx \delta^2$$

Remarks

- ▶ The standard Chernoff's bound does not assume identical distribution. (we used only for convenience)
- ▶ Chernoff's bound says that the probability that the sample mean deviates from the true mean by more than a factor δ is exponentially decaying in M (under the assumption that the random variables are independent and uniformly bounded).

Proof

First remark that for a uniformly bounded random variable $X_i \in [0, R]$ it holds

$$\mathbb{E}[e^{tX_i}] \leq \mathbb{E}[1 + (e^{tR} - 1) \frac{X_i}{R}] = 1 + (e^{tR} - 1) \frac{\mu}{R} \leq \exp\{(e^{tR} - 1)\mu/R\}$$

Moreover, for any random variable

$$P(X \geq a) = P(e^{tX} \geq e^{ta}) \leq \frac{\mathbb{E}[e^{tX}]}{e^{ta}}$$

Then

$$\begin{aligned} P(\bar{X} \geq (1 + \delta)\mu) &= P\left(\sum_{i=1}^M X_i \geq (1 + \delta)M\mu\right) = P(e^{t \sum_i X_i} \geq e^{t(1+\delta)M\mu}) \\ &\leq \frac{\mathbb{E}[e^{t \sum_i X_i}]}{e^{t(1+\delta)M\mu}} = \frac{\prod_i \mathbb{E}[e^{tX_i}]}{e^{t(1+\delta)M\mu}} \leq \frac{\exp\{M(e^{tR} - 1)\mu/R\}}{\exp\{tM\mu(1 + \delta)\}} \end{aligned}$$

Take now $t = \log(1 + \delta)/R$ so that

$$P(\bar{X} \geq (1 + \delta)\mu) \leq \exp\{-M \frac{\mu}{R} ((1 + \delta) \log(1 + \delta) - \delta)\}$$

The bound for $P(\bar{X} \leq (1 - \delta)\mu)$ can be obtained in a similar way using that $P(X \leq a) \leq e^{ta} \mathbb{E}[e^{-tX}]$ and choosing $t = -\log(1 - \delta)/R$.

Chernoff's bound (matrix case)

A Chernoff's bound similar to the scalar case holds also for symmetric and positive definite (spd) matrices.

Matrix Chernoff's bound (for i.i.d. random matrices) [J. Tropp, FoCM 2011]

Let $X_1, \dots, X_M \in \mathbb{R}^{d \times d}$ be i.i.d. spd random matrices s.t. $\lambda_{\max}(X_i) \leq R$ almost surely. Let $\mu_{\min} = \lambda_{\min}(\mathbb{E}[X_i])$, $\mu_{\max} = \lambda_{\max}(\mathbb{E}[X_i])$ and $\bar{X} = \frac{1}{M} \sum_{i=1}^M X_i$. Then

$$P(\lambda_{\max}(\bar{X}) \geq (1 + \delta)\mu_{\max}) \leq d \exp \left\{ -\frac{M\mu_{\max}\tilde{\beta}_{\delta}}{R} \right\}, \quad \delta \geq 0$$

$$P(\lambda_{\min}(\bar{X}) \leq (1 - \delta)\mu_{\min}) \leq d \exp \left\{ -\frac{M\mu_{\max}\beta_{\delta}}{R} \right\}, \quad \delta \in [0, 1],$$

with $\tilde{\beta}_{\delta} = (1 + \delta) \log(1 + \delta) - \delta$ and $\beta_{\delta} = \delta + (1 - \delta) \log(1 - \delta)$

Special case: $\mathbb{E}[X_i] = I$. In this case, $\mathbb{E}[\bar{X}] = I$, $\mu_{\max} = \mu_{\min} = 1$ and the following bound can then be obtained

$$P(\|\bar{X} - I\| \geq \delta) \leq 2d \exp \left\{ -\frac{M\beta_{\delta}}{R} \right\}$$

Concentration of measure result

Theorem [Cohen-Davenport-Leviatan '13]

Introduce the event

$$\begin{aligned}\Omega_+^M(\delta) &:= \{\|G - I\| \leq \delta\} \\ &= \{(1 - \delta)\|\mathbf{v}\|_{L_\rho^2(\Gamma, V)}^2 \leq \|\mathbf{v}\|_M^2, \mathbf{v} \leq (1 + \delta)\|\mathbf{v}\|_{L_\rho^2(\Gamma, V)}^2, \forall \mathbf{v} \in \mathbb{P}_\Lambda\}.\end{aligned}$$

For any $\delta, \gamma > 0$ and M satisfying

$$K(\Lambda) \leq \frac{\beta_\delta}{1 + \gamma} \frac{M}{\log M}, \quad \beta_\delta = \delta + (1 - \delta) \log(1 - \delta) \quad (91)$$

we have that $P(\Omega_+^M(\delta)) \geq 1 - 2M^{-\gamma}$.

Proof: apply previous Matrix Chernoff's bound on $G = \frac{1}{M} \sum_{i=1}^M G^{(i)}$ with $\mathbb{E}[G^{(i)}] = I$ and $\lambda_{\max}(G^{(i)}) = \|G^{(i)}\| \leq K(\Lambda)$.

Then, under condition (91)

$$P(\Omega_-^M(\delta)) \leq 2(\#\Lambda) e^{-\frac{M\beta_\delta}{K(\Lambda)}} \leq 2M e^{-(1+\gamma) \log M} \leq 2M^{-\gamma}.$$

Convergence in Probability

From the stability of the random projection one can derive optimality results either in expectation or probability

Theorem [Chkifa-Cohen-Migliorati-Nobile-Tempone '14], [Migliorati-Nobile-Tempone '15]

For any $\alpha, \delta \in (0, 1)$, under the condition $\frac{M}{\log M + \log(2/\alpha)} \geq \frac{K(\Lambda)}{\beta_\delta}$, it holds with probability greater than $1 - \alpha$

$$\|u - \Pi_\Lambda^M u\|_{L_p^2(\Gamma, V)} \leq \left(1 + \sqrt{\frac{1}{1-\delta}}\right) \inf_{v \in \mathbb{P}_\Lambda \otimes V} \|u - v\|_{L^\infty(\Gamma, V)}$$

Proof: Under the above condition $P(\Omega_+^M(\delta)) \geq 1 - \alpha$ (take $\gamma = \frac{\log(2/\alpha)}{\log(M)}$ in (91)). Given any draw in $\Omega_+^M(\delta)$, we have for any $v \in \mathbb{P}_\Lambda$

$$\begin{aligned}\|u - \Pi_\Lambda^M u\|_{L_p^2(\Gamma, V)} &\leq \|u - v\|_{L_p^2(\Gamma, V)} + \|v - \Pi_\Lambda^M u\|_{L_p^2(\Gamma, V)} \\ &\leq \|u - v\|_{L_p^2(\Gamma, V)} + \sqrt{(1-\delta)^{-1}} \|v - \Pi_\Lambda^M u\|_{M, V}\end{aligned}$$

By the orthogonality identity $\|u - v\|_{M, V}^2 = \|u - \Pi_\Lambda^M u\|_{M, V}^2 + \|\Pi_\Lambda^M u - v\|_{M, V}^2$, we deduce

$$\begin{aligned}\|u - \Pi_\Lambda^M u\|_{L_p^2(\Gamma, V)} &\leq \|u - v\|_{L_p^2(\Gamma, V)} + \sqrt{(1-\delta)^{-1}} \|u - v\|_{M, V} \\ &\leq \left(1 + \sqrt{(1-\delta)^{-1}}\right) \|u - v\|_{L^\infty(\Gamma, V)}.\end{aligned}$$

Example

Consider the case $N = 1$, $y \sim \mathcal{U}([-1, 1])$ and expansion in Legendre polynomials up to degree w .

Since $|\psi_p(y)| \leq \sqrt{2p+1}$ we have

$$\begin{aligned} K(\Lambda) &= \sup_{y \in [-1, 1]} \sum_{p=0}^w |\psi_p(y)|^2 \leq \sum_{p=0}^w (2p+1) = w(w+1) + w + 1 \\ &= (w+1)^2 = \#\Lambda^2. \end{aligned}$$

Hence, for any $\alpha, \delta \in (0, 1)$, under the condition

$$\frac{M}{\log M + \log(2/\alpha)} \geq \frac{(w+1)^2}{\beta_\delta},$$

i.e. $M \sim w^2$, we have with probability greater than $1 - \alpha$

$$\|u - \Pi_\Lambda^M u\|_{L_p^2(\Gamma, V)} \leq \left(1 + \sqrt{\frac{1}{1-\delta}}\right) \inf_{v \in \mathbb{P}_\Lambda \otimes V} \|u - v\|_{L^\infty(\Gamma, V)}$$

Exercise 14.1 (Periodic functions, approximation with trigonometric functions)

Let u be periodic in $[0, 1]$. Consider the orthonormal basis $\{\exp(i2\pi kx)\}_{k \in \mathbb{Z}}$ corresponding to the uniform density.

1. Compute $K(\Lambda)$, corresponding to $k \in \Lambda$.
2. Consider $\Lambda = \{k \in \mathbb{Z} : |k| \leq w\}$. What is the corresponding approximation result for C^p functions using discrete L^2 projection?

Exercise 14.2 (Sampling with another density)

Consider the approximation of u in L^2_ρ sense. Let $\hat{\rho}$ be s.t.

$$\sup_y \frac{\rho}{\hat{\rho}} = C < \infty.$$

Show that, for any v we have

$$\|v\|_{L^2_\rho} \leq \sqrt{C} \|v\|_{L^2_{\hat{\rho}}}$$

and use this fact to estimate the resulting L^2_ρ error of a discrete projection based on $\hat{\rho}$ samples.

Convergence in expectation

Assume $\|u\|_{L^\infty(\Gamma, V)} \leq \tau$ and define the truncation operator

$$T_\tau : V \rightarrow V, \quad T_\tau(v) = \begin{cases} v & \text{if } \|v\|_V \leq \tau \\ \frac{\tau}{\|v\|_V} v, & \text{if } \|v\|_V > \tau \end{cases}$$

Theorem [Cohen-Davenport-Leviatan '13], [Chkifa-Cohen-Migliorati-Nobile-Tempone '14]

For any $\delta \in (0, 1)$ and any $\gamma > 0$, under the condition $\frac{M}{\log M} \geq (1 + \gamma) \frac{K(\Lambda)}{\beta_\delta}$, it holds

$$\mathbb{E}(\|u - T_\tau \circ \Pi_\Lambda^M u\|_{L_p^2(\Gamma, V)}^2) \leq C \inf_{v \in \mathbb{P}_\Lambda \otimes V} \|u - v\|_{L_p^2(\Gamma, V)}^2 + 8\tau^2 M^{-\gamma}$$

$$\text{with } C = 1 + \frac{4\beta_\delta}{(1+\gamma)\log M} \xrightarrow{M \rightarrow \infty} 1.$$

Observe: Here in discrete L^2 projections, Monte Carlo is used to achieve stability! Accuracy comes from the approximation properties of the subspace P_Λ .

Case of noisy observations

Let us consider the case of a QoI $\varphi(\mathbf{y}) = Q(u(\mathbf{y}))$ and noisy observations

$$\varphi^{(k)} = \varphi(\mathbf{y}_k) + \eta_k$$

with η_k i.i.d. and

$$\mathbb{E}[\eta_k | \mathbf{y}_k] = \bar{\eta}(\mathbf{y}_k) \in L^2_\rho(\Gamma) \quad (\text{offset})$$

$$\sup_{\mathbf{y}_k \in \Gamma} \text{Var}(\eta_k | \mathbf{y}_k) = \sigma^2 < \infty \quad (\text{variance})$$

The *offset* could model any deterministic source of error due e.g. to numerical discretization.

The *fluctuations* $\tilde{\eta}_k = \eta_k - \bar{\eta}(\mathbf{y}_k)$ model random measurement errors.

We will also consider the case of *bounded* noise

$$|\tilde{\eta}_k| \leq \tilde{\eta}_{\max}, \quad \|\bar{\eta}\|_{L^\infty(\Gamma)} < \infty.$$

Convergence in expectation

Theorem [Chkifa-Cohen-Migliorati-Nobile-Tempone '14], [Migliorati-Nobile-Tempone '15]

Assume $\|\varphi\|_{L^\infty(\Gamma)} \leq \tau$. For any $\delta \in (0, 1)$ and any $\gamma > 0$, under the condition $\frac{M}{\log M} \geq (1 + \gamma) \frac{K(\Lambda)}{\beta_\delta}$, it holds

$$\begin{aligned}\mathbb{E}(\|\varphi - T_\tau \circ \Pi_\Lambda^M \varphi\|_{L_\rho^2(\Gamma)}^2) &\leq C_1 \underbrace{\inf_{v \in \mathbb{P}_\Lambda} \|\varphi - v\|_{L_\rho^2(\Gamma)}^2}_{\text{best approx. error in } L^2} \\ &\quad + \frac{2}{(1 - \delta)^2} \left(\underbrace{\frac{\#\Lambda}{M} \sigma^2}_{\text{noise variance}} + C_2 \underbrace{\|\bar{\eta}\|_{L_\rho^2(\Gamma)}^2}_{\text{noise offset}} \right) \\ &\quad + \underbrace{8\tau^2 M^{-\gamma}}_{\text{prob. bad events}}\end{aligned}$$

with $C_1, C_2 \xrightarrow{M \rightarrow \infty} 1$.

Convergence in Probability – bounded noise

Theorem [Migliorati-Nobile-Tempone '15]

In the **bounded noise** case, for any $\alpha, \delta \in (0, 1)$, under the condition $\frac{M}{\log M + \log(3/\alpha)} \geq \frac{K(\Lambda)}{\beta_\delta}$, it holds with probability greater than $1 - \alpha$

$$\|\varphi - \Pi_\Lambda^M \varphi\|_{L_p^2(\Gamma)}^2 \leq (1 + \frac{2}{1 - \delta}) \underbrace{\inf_{v \in \mathbb{P}_\Lambda} \|\varphi - v\|_{L_p^\infty(\Gamma)}^2}_{\text{best approx. error in } L^\infty} \\ + \frac{4(1 + \delta)}{(1 - \delta)^2} \left(2 \underbrace{\frac{\#\Lambda \log(3M\alpha^{-1})}{M} \tilde{\eta}_{\max}^2}_{\text{bounded noise}} + \underbrace{\|\bar{\eta}\|_{L_p^\infty(\Gamma)}^2}_{\text{noise offset}} \right)$$

Observe: The lowest value we can have for $K(\Lambda)$ is $\#(\Lambda)$. In that case, the number of samples M needs to be slightly larger than $\#(\Lambda)$ to produce a stable discrete projection. Still, to reduce the approximation error in the presence of unbiased noise we need $\#(\Lambda)/M \rightarrow 0$, which is a stronger requirement.

Question: How can one optimize the choice of the discretization parameters in this context?

Case of uniform random variables in $[-1, 1]^N$

The discrete L^2 projection is stable and optimally convergent under the condition

$$K(\Lambda) := \sup_{\mathbf{y} \in \Gamma} \left(\sum_{\mathbf{p} \in \Lambda} |\psi_{\mathbf{p}}(\mathbf{y})|^2 \right) \leq \frac{\beta_\delta}{1 + \gamma} \frac{M}{\log M}$$

where β_δ is defined in (91). Recall that for Legendre polynomials we have: $|\psi_{\mathbf{p}}(\mathbf{y})| \leq \prod_{n=1}^N \sqrt{2p_n + 1}$, for all $\mathbf{y} \in [-1, 1]^N$.

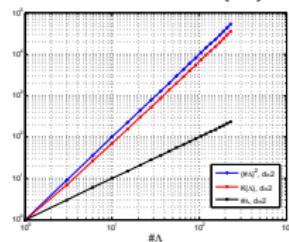
Theorem [Chkifa-Cohen-Migliorati-Nobile-Tempone '13]

For any set $\Lambda \subset \mathbb{N}^N$ monotone it holds $(\#\Lambda) \leq K_L(\Lambda) \leq (\#\Lambda)^2$. Here K_L refers to Legendre polynomials. Hence, the discrete L^2 projection over \mathbb{P}_Λ is stable and optimally convergent under the (sufficient) condition

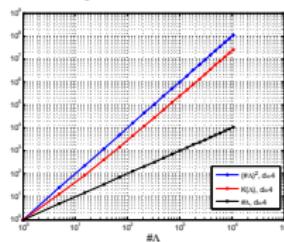
$$\frac{1 + \gamma}{\beta_\delta} (\#\Lambda)^2 \leq \frac{M}{\log M}$$

Case of uniform random variables in $[-1, 1]^N$

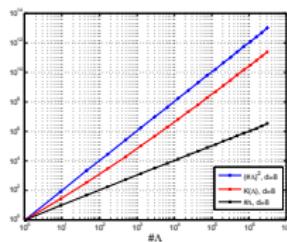
- ▶ For specific index sets Λ the bound for $K(\Lambda)$ can be improved.
- ▶ For instance for the Total Degree polynomial space of degree w the bound $K(\Lambda) \leq (\#\Lambda)^2$ is very conservative



dimension $N = 2$



dimension $N = 4$



dimension $N = 8$

- ▶ The bound for $K(\Lambda)$ heavily depends on the underlying distribution. For instance, using the same polynomial subspace we have

$$\text{Chebyshev distribution} \implies K(\Lambda) \leq \min\{(\#\Lambda)^{\frac{\log 3}{\log 2}}, 2^N \#\Lambda\}$$

$$\text{Beta distribution with } \theta_1, \theta_2 \in \mathbb{N} \implies K(\Lambda) \leq (\#\Lambda)^{2 \max\{\theta_1, \theta_2\} + 2}$$

On the $K(\Lambda)$ bound

We expect this bound to be pessimistic for monotone sets that have shapes very different from rectangles. Consider again Legendre polynomials. For instance, let $w \geq 1$ and consider the monotone set

$$S_{w,N} := \{\mathbf{p} \in \mathbb{N}_0^N : |\mathbf{p}| \leq w\},$$

where $|\mathbf{p}| := \sum_{j=1}^N p_j$, associated to the polynomial space $\mathbb{P}_{S_{w,N}}$ of **total degree** w in dimension N .

By the inequality of arithmetic and geometric means, one has for any $\mathbf{p} \in S_{w,N}$

$$\prod_{1 \leq j \leq N} (2p_j + 1) \leq \left(\frac{1}{N} \sum_{1 \leq j \leq N} (2p_j + 1) \right)^N = \left(\frac{2|\mathbf{p}|}{N} + 1 \right)^N \leq \left(\frac{2w}{N} + 1 \right)^N.$$

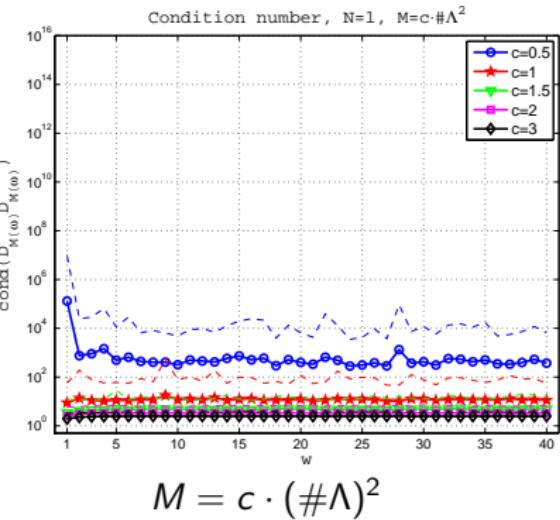
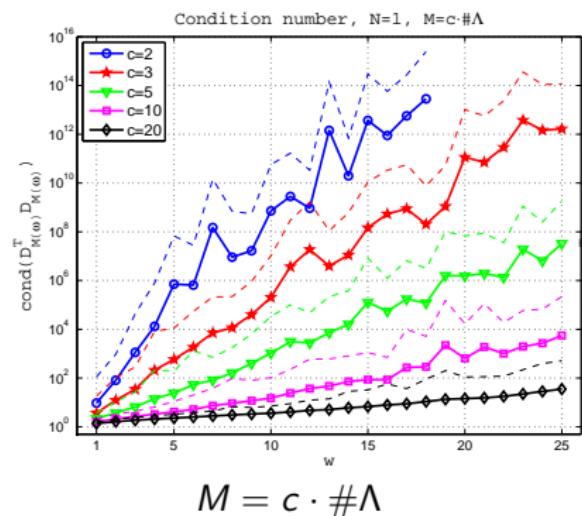
Therefore

$$K(S_{w,N}) \leq \left(\frac{2w}{N} + 1 \right)^N \#(S_{w,N}),$$

and $\left(\frac{2w}{N} + 1 \right)^N \approx e^{2w}$ is very small compared to $\#(S_{w,N}) = \binom{N+w}{w}$ for large values of N .

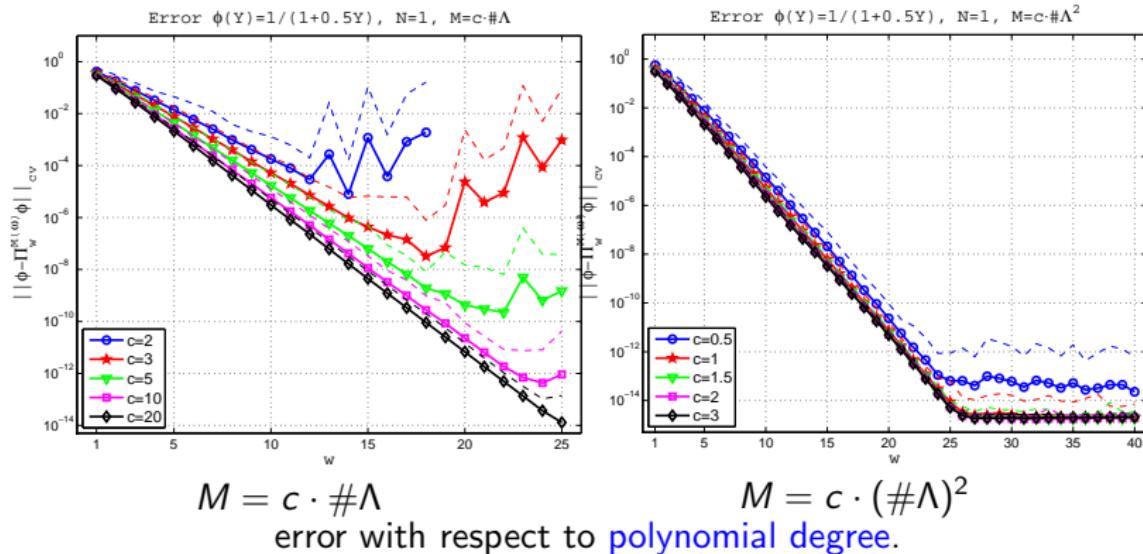
Some numerical examples – 1D function

Condition number of $D^T D$



Some numerical examples – 1D function

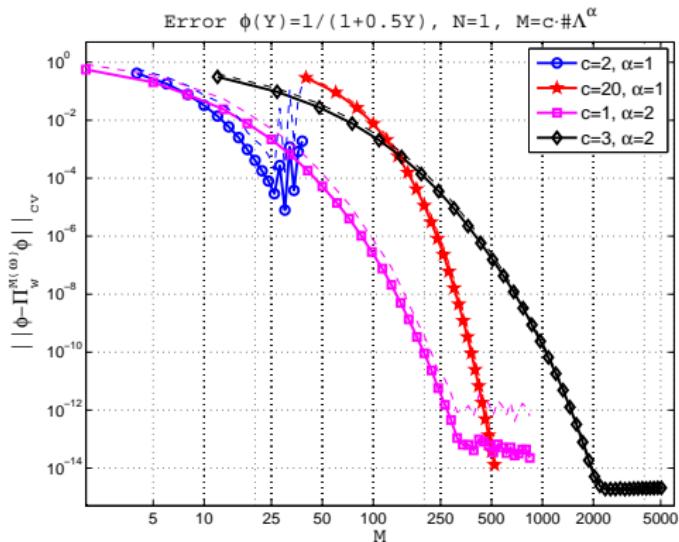
Approximation of the meromorphic function $\phi(y) = \frac{1}{1+0.5y}$



How do we measure the error in this case? We can use cross-validation since we have access to the exact function! Use M_1 samples to do the regression and M_2 to sample the regression error.

Some numerical examples – 1D function

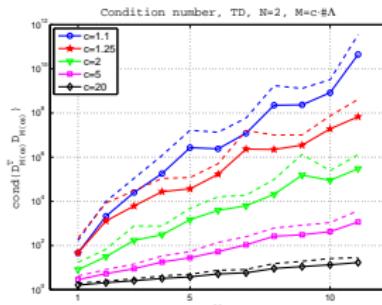
Approximation of the meromorphic function $\phi(y) = \frac{1}{1+0.5y}$



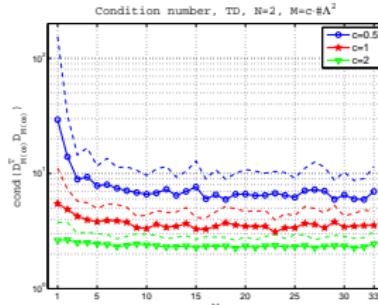
error with respect to **total number of sampling points**.

Some numerical examples

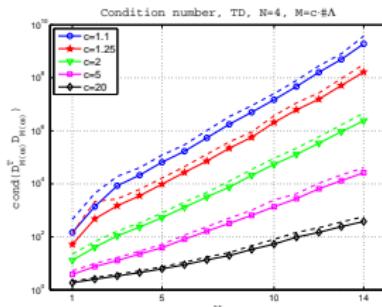
Condition number of $D^T D$ – multiD – Total Degree poly. space



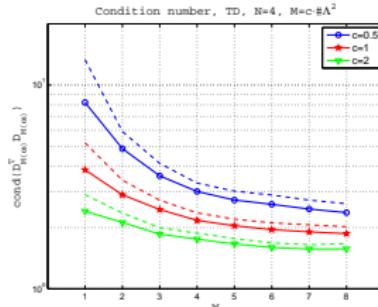
$$M = c \cdot \#\Lambda$$



$$M = c \cdot (\#\Lambda)^2$$



$$M = c \cdot \#\Lambda$$



$$M = c \cdot (\#\Lambda)^2$$

Exercise 14.3 (Periodic functions, approximation with trigonometric functions)

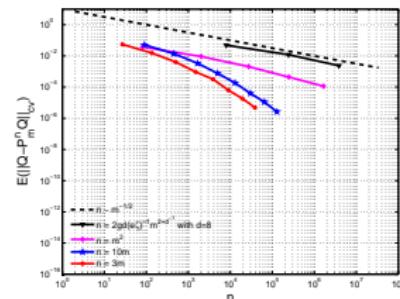
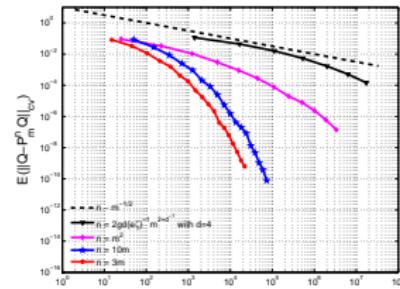
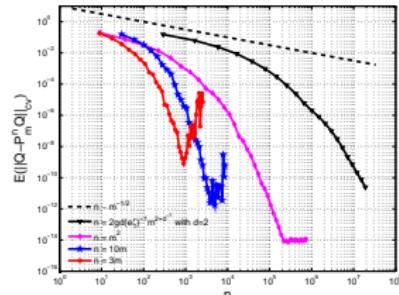
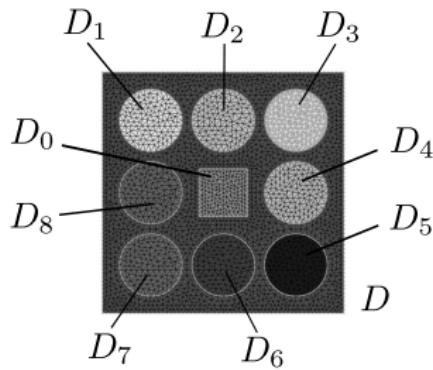
Let u be periodic in $[0, 1]^N$. Consider the orthonormal basis $\{\exp(i2\pi k \cdot x)\}_{k \in \mathbb{Z}^N}$ corresponding to the uniform density.

1. Compute $K(\Lambda)$, corresponding to $\Lambda = \{k \in \mathbb{Z}^N; |k|_1 \leq w\}$.
2. What is the corresponding approximation result for C^p functions using discrete L^2 projection?

Linear Elliptic PDE with N random inclusions

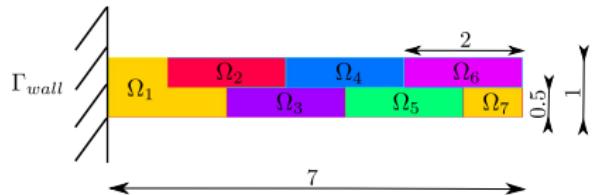
Here $Q(U)$ is the integral of the solution, we use zero Dirichlet bc, the rhs is 100 in D_0 and zero elsewhere. We derived the theoretical bound

$$\mathbb{E}(\|u - T_\tau \circ \Pi_\Lambda^M u\|_{L_p^2(\Gamma, H_0^1(D))}^2) \leq c_1 e^{-c_2 NM^{\frac{1}{1+2N}}}$$



In the plots, $n = \#$, Samples, $m = P_\Lambda$
Dimension

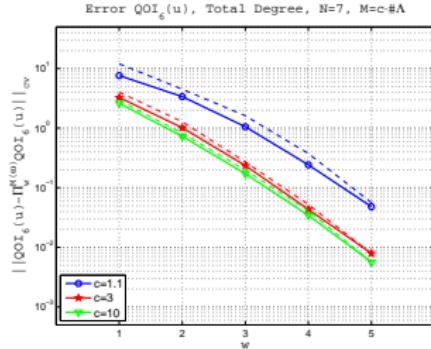
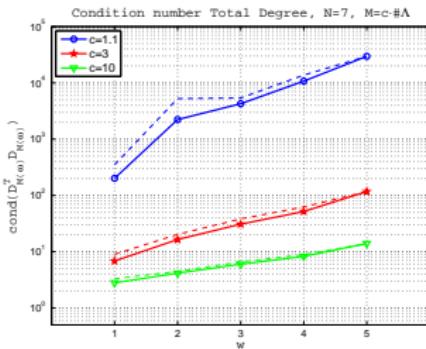
Cantilever beam



- ▶ linear elasticity equations
- ▶ Young modulus uncertain in each brick:

$$E_i = e^{7+Y_i}, \quad \text{in } \Omega_i, \quad Y_i \sim \mathcal{U}([-1, 1]), \text{iid}$$

- ▶ Uncertainty analysis on maximum vertical displacement.



Improvements on the quadratic relation

Improvements can be obtained by using **importance sampling** with a different distribution $\hat{\rho}$. Let us consider the **weighted least squares** approximation

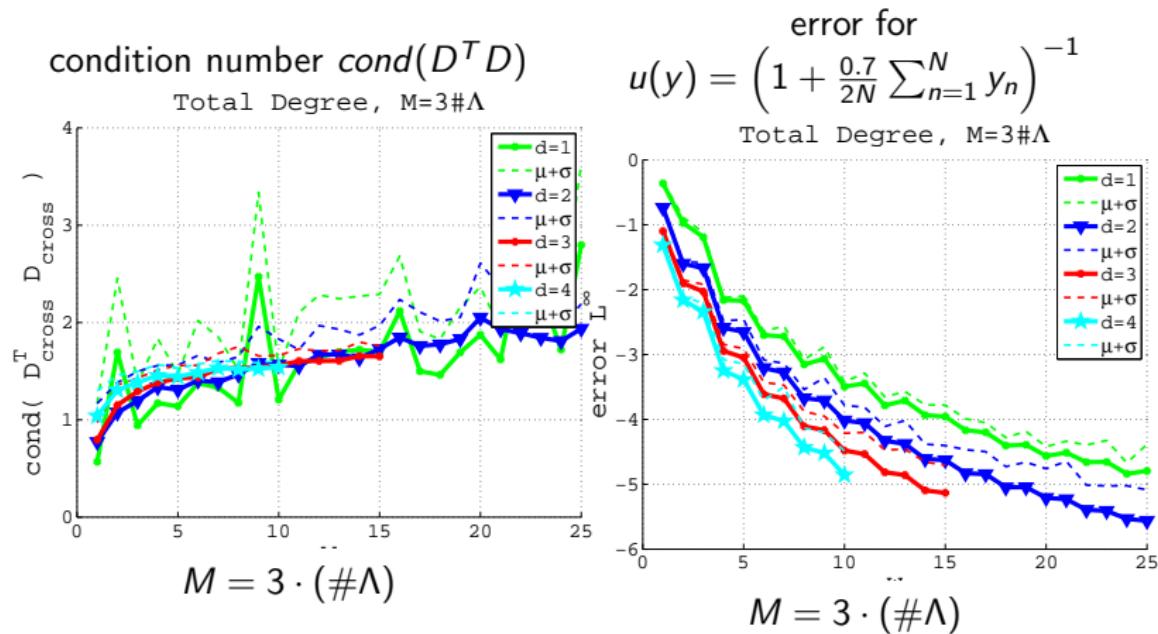
$$\hat{u}_{\Lambda, M} = \arg \min_{v \in \mathbb{P}_{\Lambda}(\Gamma) \otimes V} \frac{1}{M} \sum_{k=1}^M \frac{\rho(\mathbf{y}^{(k)})}{\hat{\rho}(\mathbf{y}^{(k)})} \|u^{(k)} - v(\mathbf{y}^{(k)})\|_V^2$$

where the sample $\{\mathbf{y}^{(k)}\}_k$ is drawn from the distribution $\hat{\rho}(\mathbf{y})d\mathbf{y}$.

- ▶ $\rho(\mathbf{y}) = \hat{\rho}(\mathbf{y}) = \text{Chebyshev distribution in } [-1, 1]^N$, then the relation $M \propto \min\{2^N(\#\Lambda), (\#\Lambda)^{\frac{\log(3)}{\log(2)}}\}$ is enough to guarantee optimal convergence [Chkifa-Cohen-Migliorati-N.-Tempone '14]
- ▶ $\rho(\mathbf{y})=\text{uniform}$ and $\hat{\rho}(\mathbf{y})=\text{Chebyshev distribution in } [-1, 1]^N$, then, the relation $M \propto 2^N(\#\Lambda)$ guarantees optimal convergence [Rauhut-Ward '12]. However, the constant depends on N [Yan-Guo-Xiu '12].
- ▶ $\rho(\mathbf{y})=\text{Gaussian}$: still unclear. Numerically, the situation seems to be worse. Improvements suggested in [Tang-Zhou '14]

Numerical example with Chebyshev preconditioning

Expansion in Legendre polynomials ($\rho(\mathbf{y})$ =uniform) and samples from Chebyshev distribution ($\hat{\rho}(\mathbf{y})$ =Chebyshev)



Adaptive construction of polynomial spaces

$\{\Lambda_k\}_{k \geq 0}$ sequence of downward closed multi-index sets, with $\Lambda_0 = \{\mathbf{0}\}$.
The sequence is adaptively computed by means of greedy algorithms based on the random discrete L^2 projection.

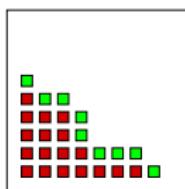
Definitions:

- Margin $\mathcal{M}(\Lambda)$ associated to a multi-index set Λ :

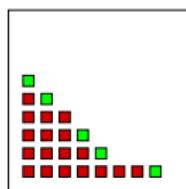
$$\mathcal{M}(\Lambda) = \{\mathbf{p} : \mathbf{p} \notin \Lambda \text{ and } \exists j > 0 : \mathbf{p} - \mathbf{e}_j \in \Lambda\}$$

- Reduced margin $\mathcal{R}(\Lambda)$ associated to a multi-index set Λ :

$$\mathcal{R}(\Lambda) = \{\mathbf{p} : \mathbf{p} \notin \Lambda \text{ and } \forall j = 1, \dots, d : p_j \neq 0 \Rightarrow \mathbf{p} - \mathbf{e}_j \in \Lambda\}$$



set Λ and its Margin



set Λ and its Reduced margin

The Dörfler marking

Idea proposed by W. Dörfler in 1996 for Adaptive Finite Elements.

Given a multi-index set Λ , a subset $R \subseteq \mathcal{R}(\Lambda)$, a (continuous) function $e : R \rightarrow \mathbb{R}$ and a parameter $\theta \in (0, 1]$, we define a procedure

$$\text{Dörfler} = \text{Dörfler}(R, e, \theta)$$

that computes a set $F \subseteq R \subseteq \mathcal{R}(\Lambda)$ of minimal cardinality such that

$$\sum_{\mathbf{p} \in F} e(\mathbf{p})^2 \geq \theta \sum_{\mathbf{p} \in R} e(\mathbf{p})^2.$$

In practice, for any $\mathbf{p} \in R$, the error indicator $e(\mathbf{p})$ will be either an estimator of the coefficient c_p of the function u expanded over the Legendre basis or the projected residual on the \mathbf{p} -th Legendre basis function.

This corresponds to choose a fraction θ of the energy associated with the (estimates of the) coefficients in the set R .

Orthogonal Matching Pursuit with Dörfler marking

Algorithm Orthogonal Matching Pursuit with Dörfler marking

Set $r_0 = u(\mathbf{y})$, $u_0 \equiv 0$ and $\Lambda_0 = \{\mathbf{0}\}$,

for $k = 1, \dots, k_{max}$ **do**

$F_1 = \text{Dörfler}(\mathcal{R}(\Lambda_{k-1}), \{|(r_{k-1}, \psi_p)_{M,V}| \}_p, \theta_1)$

$\tilde{\Lambda}_k = \Lambda_{k-1} \cup F_1$

$u_k = \arg \min_{v \in \mathbb{P}_{\tilde{\Lambda}_k}} \|u - v\|_{M,V}, \quad u_k = \sum_{p \in \tilde{\Lambda}_k} c_p^{(k)} \psi_p$

$F_2 = \text{Dörfler}(F_1, \{c_p^{(k)}\}_p, \theta_2)$

$\Lambda_k = \Lambda_{k-1} \cup F_2$

$r_k = u - u_k|_{\Lambda_k}$

end for

$\theta_1 \in (0, 1)$ and $\theta_2 = 1$: Dörfler marking only with the correlations.

$\theta_1 = 1$ and $\theta_2 \in (0, 1)$: Dörfler marking only with the random discrete L^2 projection on $\Lambda_{k-1} \cup \mathcal{R}(\Lambda_k)$.

Some remarks and open issues

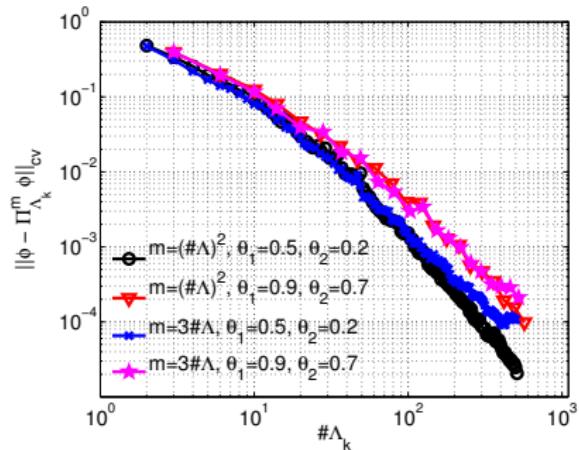
- ▶ The first Dörfler marking performs a screening of the reduced margin, to avoid an L^2 discrete minimization over a too large polynomial space.
- ▶ At each step the correlations $\{|(r_{k-1}, \psi_{\mathbf{p}})_{M,V}| : \mathbf{p} \in \mathcal{R}(\Lambda_k)\}$ are mutually uncoupled and cheap to compute, but might provide only a rough estimate of the coefficients (depending on the choice of M_k).
- ▶ The second Dörfler marking performs a selection based on the more accurate estimates of the coefficients coming from the L^2 projection.
- ▶ At each step the adaptive algorithm remains stable and accurate by choosing $M_k \propto (\#\Lambda_k)^2$ (consequence of the theory in the first part).
- ▶ The adaptive algorithm generates a sequence $\{\Lambda_k\}_{k \geq 0}$ of only quasi best N -term sets.
- ▶ Rate of convergence? Choice of θ_1, θ_2 ? What if $M_k \propto \#\Lambda_k$?

A numerical test

Approximation of a meromorphic function (16-variables)

$$\phi(\mathbf{y}) = \frac{1}{1 - \gamma \cdot \mathbf{y}}, \quad \mathbf{y} \sim \mathcal{U}([-1, 1]^{16})$$

$$\gamma = 0.3 * (1, 5 \cdot 10^{-1}, 10^{-1}, 5 \cdot 10^{-2}, \dots, 5 \cdot 10^{-8})$$



Extension to Multilevel discrete least squares approximation

Beware of the notation changes ...

PDEs with random parameters

Consider a differential problem

$$\mathcal{L}(u; \mathbf{y}) = \mathcal{G} \quad (*)$$

depending on a set of random parameters $\mathbf{y} = (y_1, \dots, y_N) \in \Gamma \subset \mathbb{R}^N$ with joint probability measure μ on Γ .

We assume that $(*)$ is well posed, with unique solution $u(\mathbf{y})$, in some suitable function space V , and we focus on a Quantity of Interest $Q : V \rightarrow \mathbb{R}$.

Goal: approximate the whole response function

$$\mathbf{y} \mapsto f(\mathbf{y}) := Q(u(\mathbf{y})) : \Gamma \rightarrow \mathbb{R}$$

by multivariate polynomials.

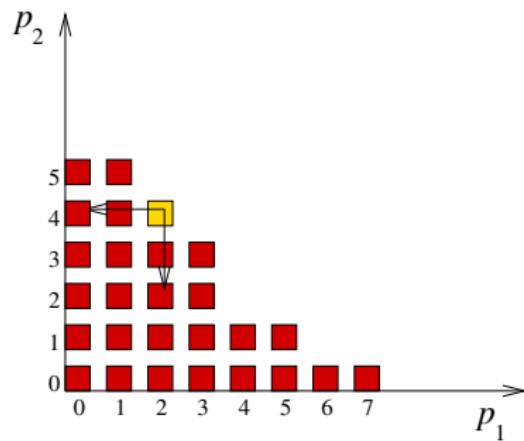
Possibly derive approximated statistics as $\mathbb{E}[f]$, $\text{Var}[f]$, or even solve a related inverse problem.

Polynomial approximation on downward closed sets

Assume $f \in L^2_\mu(\Gamma)$. We seek an approximation of f in a finite dimensional polynomial subspace

$$V_{\Lambda} = \text{span} \left\{ \prod_{n=1}^N y_n^{p_n}, \quad \text{with } \mathbf{p} = (p_1, \dots, p_N) \in \Lambda \right\}$$

with $\Lambda \subset \mathbb{N}^N$ a downward closed index set.



Definition. An index set Λ is downward closed if

$$\mathbf{p} \in \Lambda \text{ and } \mathbf{q} \leq \mathbf{p} \implies \mathbf{q} \in \Lambda.$$

Weighted discrete least squares approximation

1. Sample independently M points $(\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(M)}) \in \Gamma^M$ from a distribution $\nu \ll \mu$, with density $\rho = \frac{d\nu}{d\mu}$;
2. define the weight function $w(\mathbf{y}) = \frac{1}{\rho(\mathbf{y})}$;
3. find weighted discrete least squares approximation on V_Λ

$$\hat{\Pi}_M f = \arg \min_{v \in V_\Lambda} \|f - v\|_M \quad \text{with} \quad \|g\|_M^2 = \frac{1}{M} \sum_{j=1}^M w(\mathbf{y}^{(j)}) g(\mathbf{y}^{(j)})^2.$$

Here: $\mathbb{E} [\|g\|_M^2] = \int_{\Gamma} w(\mathbf{y}) g(\mathbf{y})^2 \nu(d\mathbf{y}) = \int_{\Gamma} g(\mathbf{y})^2 \mu(d\mathbf{y}) = \|g\|_{L_\mu^2}^2$.

Algebraic system: let $\{\phi_j\}_{j=1}^{|\Lambda|}$ be a basis of V_Λ , orthonormal w.r.t. μ , and $\hat{\Pi}_M f(\mathbf{y}) = \sum_{j=1}^{|\Lambda|} c_j \phi_j(\mathbf{y})$. Then, $\mathbf{c} = (c_1, \dots, c_{|\Lambda|})^T$ satisfies

$$G\mathbf{c} = \hat{\mathbf{f}}, \quad G_{i,j} = (\phi_i, \phi_j)_M, \quad \hat{f}_i = (f, \phi_i)_M.$$

Optimality of discrete least squares approximation

Theorem 15.1 ([Cohen-Migliorati 2017][Cohen-Davenport-Leviatan 2013])

For arbitrary $r > 0$ define

$$\kappa_r := \frac{1/2(1 - \log 2)}{1 + r} \quad \text{and} \quad K_{\Lambda, w} := \sup_{y \in \Gamma} \left(w(y) \sum_{j=1}^{|\Lambda|} \phi_i(y)^2 \right).$$

If

$$\frac{M}{\log M} \geq \frac{K_{\Lambda, w}}{\kappa_r},$$

then

- ▶ $P(\|G - I\| \leq \frac{1}{2}) > 1 - 2M^{-r}.$
- ▶ $\|f - \hat{\Pi}_M f\|_{L^2_\mu} \leq (1 + \sqrt{2}) \inf_{v \in V_\Lambda} \|f - v\|_{L^\infty_{\sqrt{w}}} \text{ with prob. } > 1 - 2M^{-r}.$
- ▶ $\mathbb{E} \left[\|f - \hat{\Pi}_M^c f\|_{L^2_\mu}^2 \right] \leq C_M \inf_{v \in V_\Lambda} \|f - v\|_{L^2_\mu}^2 + 2\|f\|_{L^2_\mu}^2 M^{-r}$

where $\hat{\Pi}_M^c f = \hat{\Pi}_M f \cdot \mathbf{1}_{\{\|G - I\| \leq \frac{1}{2}\}}$ and $C_M = \left(1 + \frac{4\kappa_r}{\log M}\right) \xrightarrow{M \rightarrow \infty} 1.$

Sufficient number of points - uniform measure

- ▶ Uniform measure: $\mu = \mathcal{U} \left(\prod_{i=1}^N \Gamma_i \right)$

[Chkifa-Cohen-Migliorati-Nobile-Tempone 2015] When sampling from the same distribution ($\nu = \mu$ and $w = 1$), then

$$|\Lambda| \leq K_{\Lambda,1} \leq |\Lambda|^2.$$

Hence, (unweighted) discrete least square is stable and optimally convergent under the condition

$$\frac{M}{\log M} \geq \frac{|\Lambda|^2}{\kappa_r} \quad (\text{quadratic proportionality}).$$

Sufficient number of points - optimal measure

- ▶ [Cohen-Migliorati 2017] For arbitrary μ , when sampling from the optimal measure

$$\frac{d\nu^*}{d\mu}(\mathbf{y}) = \rho^*(\mathbf{y}) = \frac{1}{|\Lambda|} \sum_{j=1}^{|\Lambda|} \phi_j(\mathbf{y})^2 \implies K_{\Lambda, w^*} = |\Lambda|.$$

Hence, weighted discrete least squares stable and optimal with

$$\frac{M}{\log M} \geq \frac{|\Lambda|}{\kappa_r} \quad (\text{linear proportionality}).$$

- ▶ Sampling algorithms from the optimal distribution are available (marginalization [Cohen-Migliorati 2017], acceptance rejection [HajiAli-Nobile-Tempone-Wolters, 2017])
- ▶ The optimal distribution depends on Λ . Not good for adaptive algorithms!

Sufficient number of points - Chebyshev measure

- ▶ Alternatively, for uniform measure μ (or more generally a product measure $\mu = \otimes_{j=1}^N \mu_j$, with μ_j doubling measure, i.e. $\mu_j(2I) = L\mu_j(I)$) one can sample from the arcsin (Chebyshev) distribution.

$$K_{\Lambda,w} \leq C^N |\Lambda|, \quad \frac{M}{\log M} \geq \frac{C^N}{\kappa_r} |\Lambda|.$$

Still linear scaling but with a constant exponentially dependent on N .
Advantage: the sampling measure does not depend on Λ . Good for adaptivity.

Multilevel least squares approximation

In practice $f(\mathbf{y}) = Q(u(\mathbf{y}))$ can not be evaluated exactly as it requires the solution of a differential equation.

- We introduce a sequence of approximations f_{n_ℓ} , $n_\ell \in \mathbb{N}$ with increasing cost, s.t.

$$\lim_{\ell \rightarrow \infty} \|f - f_{n_\ell}\|_{L^2_\mu} = 0,$$

(or possibly a stronger norm)

- Similarly, we introduce a sequence of nested downward closed sets

$$\Lambda_{m_0} \subset \Lambda_{m_1} \subset \dots \subset \Lambda_{m_k} \subset \dots$$

such that

$$\lim_{k \rightarrow \infty} \inf_{v \in V_{\Lambda_{m_k}}} \|f - v\|_{L^2_\mu} = 0.$$

Correspondingly, for each Λ_{m_k} we introduce a weighted discrete least squares projector $\hat{\Pi}_{M_k}$ using $\frac{M_k}{\log M_k} = O(|\Lambda_{m_k}|)$ random points.

Multilevel least squares approximation

Multilevel formula: given maximum level $L \in \mathbb{N}$

$$\begin{aligned} S_L f &= \sum_{k+\ell \leq L} (\hat{\Pi}_{M_k} - \hat{\Pi}_{M_{k-1}})(f_{n_\ell} - f_{n_{\ell-1}}) \\ &= \sum_{\ell=0}^L \hat{\Pi}_{M_{L-\ell}}(f_{n_\ell} - f_{n_{\ell-1}}). \end{aligned}$$

- ▶ In the multilevel formula one might consider more general index sets $(k, \ell) \in \mathcal{I} \subset \mathbb{R}^2$. However, one can always recast to $k + \ell \leq L$ by properly choosing $\{n_\ell\}$ and $\{m_k\}$.
- ▶ **Question:** How to properly choose $\{n_\ell\}$ and $\{m_k\}\?$
- ▶ **Issue:** Since the least squares projection is random, we have to ensure that it is stable and optimally convergent on all levels. (Need union bound on failure probabilities)

Assumptions for ML

- ▶ For the Multilevel algorithm to be effective, we have to rely on certain “mixed regularity”.
- ▶ Let $(F, \|\cdot\|_F) \hookrightarrow (L^2_\mu, \|\cdot\|_{L^2_\mu})$ be a normed vector space of “smooth” functions (e.g. Hölder / Sobolev / analytic regularity).

Assumptions for ML

- ▶ **Assumption 1 (regularity):** $f, f_{n_\ell} \in F$ for all $\ell \in \mathbb{N}$
- ▶ **Assumption 2 (PDE discretization):** the sequence $\{f_{n_\ell}\}$ is s.t.

$$\|f - f_{n_\ell}\|_{L^2_\mu} \lesssim n_\ell^{-\beta_w}, \quad \|f - f_{n_\ell}\|_F \lesssim n_\ell^{-\beta_s}$$

and, for a single $\mathbf{y} \in \Gamma$, the cost of computing $f_{n_\ell}(\mathbf{y})$ is

$$\text{Work}(f_{n_\ell}) \lesssim n_\ell^\gamma.$$

- ▶ **Assumption 3 (polynomial approximability):** the sequence $\{\Lambda_{m_k}\}$ is s.t.

$$\dim(V_{\Lambda_{m_k}}) = |\Lambda_{m_k}| \lesssim m_k^\sigma,$$

$$\inf_{v \in V_{\Lambda_{m_k}}} \|f - v\|_{L^\infty_{\sqrt{w}}} \lesssim m_k^{-\alpha_p} \|f\|_F, \quad \forall f \in F,$$

(Alternatively $\inf_{v \in V_{\Lambda_{m_k}}} \|f - v\|_{L^2_\mu} \lesssim m_k^{-\alpha_e} \|f\|_F, \quad \forall f \in F$).

Tuning the ML least squares algorithm

We now choose

$$n_\ell = C \exp\left(\frac{\ell}{\gamma + \beta_s}\right), \quad \ell = 0, \dots, L \quad (\text{space discretization})$$

$$m_k = C \exp\left(\frac{k}{\sigma + \alpha_p}\right), \quad k = 0, \dots, L \quad (\text{Polynomial approx.})$$

$$\frac{m_k^\sigma}{\kappa_L} \leq \frac{M_k}{\log M_k} \leq \frac{2m_k^\sigma}{\kappa_L}, \quad k = 0, \dots, L \quad (\text{sample size with } r = L)$$

By taking $r = L$ we guarantee that

$$P\left(\exists k : \|G_k - I\| > \frac{1}{2}\right) \leq \sum_{k=0}^L P\left(\|G_k - I\| > \frac{1}{2}\right) \lesssim L^{-L}.$$

Complexity result

Theorem 15.2 ([HajiAli-Nobile-Tempone-Wolffers 2017])

Given $\epsilon > 0$ and $\beta_s = \beta_w$, we can choose $L \in \mathbb{N}$ such that

$$\|f - S_L f\|_{L^2_\mu} \leq \epsilon, \quad \text{with prob. } \geq 1 - C\epsilon^{\log |\log \epsilon|},$$

$$Work(S_L f) \lesssim \epsilon^{-\lambda} |\log \epsilon|^t \log |\log \epsilon|,$$

with

$$\lambda = \max(\sigma/\alpha_p, \gamma/\beta_s),$$

$$t = \begin{cases} 2 & \text{if } \gamma/\beta_s < \sigma/\alpha_p, \\ 3 + \sigma/\alpha_p & \text{if } \gamma/\beta_s = \sigma/\alpha_p, \\ 1 & \text{if } \gamma/\beta_s > \sigma/\alpha_p. \end{cases}$$

Proof

Analogous result holds in expectation with α_p replaced by α_e .

Improved complexity in the case $\gamma/\beta_s > \sigma/\alpha$

In the case $\gamma/\beta_s > \sigma/\alpha$ and $\beta_w > \beta_s$ the complexity can be improved by taking

$$m_k = C \exp \left(\frac{k}{\sigma + \alpha_p} + \frac{L(\beta_w - \beta_s)}{\alpha(\gamma + \beta_s)} \right).$$

In this case the complexity result becomes

$$\|f - S_L f\|_{L^2_\mu} \leq \epsilon, \quad \text{with prob. } \geq 1 - C\epsilon^{\log |\log \epsilon|},$$

$$\text{Work}(S_L f) \lesssim \epsilon^{-\lambda} |\log \epsilon|^t \log |\log \epsilon|,$$

with $t = 1$ and

$$\lambda = \frac{\gamma}{\beta_w} + \left(1 - \frac{\beta_s}{\beta_w}\right) \frac{\sigma}{\alpha_p}$$

which always improves the single level rate $\lambda_{SL} = \frac{\gamma}{\beta_w} + \frac{\sigma}{\alpha_p}$.

Application to random elliptic PDEs

Consider

$$\begin{cases} -\operatorname{div}(a(\mathbf{x}, \mathbf{y}) \nabla u(\mathbf{x}, \mathbf{y})) = g, & \text{for } \mathbf{x} \in D \subset \mathbb{R}^d \\ u(\mathbf{x}, \mathbf{y}) = 0, & \text{for } \mathbf{x} \in \partial D \end{cases}$$

with $\mathbf{y} \in \Gamma = [-1, 1]^N$ and Q linear bounded functional in $L^2(D)$ (e.g. $Q(u) = \int_D u$).

Goal: approximate $f(\mathbf{y}) = Q(u(\mathbf{y}))$.

Assumptions:

- ▶ $0 < a_{min} \leq a(\mathbf{x}, \mathbf{y}) \leq a_{max}, \quad \forall (\mathbf{x}, \mathbf{y}) \in D \times \Gamma.$
- ▶ g and D sufficiently smooth.

Application to random elliptic PDEs

Proposition 15.1

Let u_n be a finite element approximation of order $r \geq 1$ with maximal element diameter $h = n^{-1}$ and $f_n(\mathbf{y}) = Q(u_n(\mathbf{y}))$.

If $a \in C^{r,s}(D \times \Gamma) = \{v : D \times \Gamma \rightarrow \mathbb{R} : \|\partial_x^{\mathbf{r}} \partial_y^{\mathbf{s}} v\|_{C^0(D \times \Gamma)} < \infty, \forall |\mathbf{r}|_1 \leq r, |\mathbf{s}|_1 \leq s\}$, then

$$\|f - f_n\|_{C^p(\Gamma)} \lesssim n^{-(r+1)}, \quad \forall p = 0, \dots, s.$$

ML least squares complexity – mixed regularity

Consider the coefficient

$$a(\mathbf{x}, \mathbf{y}) = 1 + \|\mathbf{x}\|_2^r + \|\mathbf{y}\|_2^s \in C^{r-1,1}(D) \otimes C^{s-1,1}(\Gamma).$$

- ▶ smoother space: $F = C^{s-1,1}(\Gamma)$;
- ▶ spatial approximation: continuous finite elements of degree r ,
 - ▶ error:
 $\|f - f_n\|_{L_\mu^2} = \mathcal{O}(n^{-(r+1)}) = \|f - f_n\|_{C^{s-1,1}} \implies \beta_w = \beta_s = r + 1$;
 - ▶ cost: $\text{Work}(f_n) = n^d$ with optimal solver $\implies \gamma = d$;
- ▶ polynomial approximation: $V_{\Lambda_m} = \mathbb{P}_m$ = polynomial space of total degree m ,
 - ▶ error: $\|f - \Pi_{\mathbb{P}_m} f\|_{L^\infty} = \mathcal{O}(m^{-s}) \implies \alpha_p = s$;
 - ▶ cost: $\dim(V_{\Lambda_m}) = \binom{m+N}{N} \lesssim m^N \implies \sigma = N$.

ML least squares complexity – mixed regularity

- ▶ Complexity of single level method

$$\text{Work}_{\text{SL}} = \mathcal{O} \left(\epsilon^{-\frac{d}{r+1} - \frac{N}{s}} \log \epsilon^{-1} \right).$$

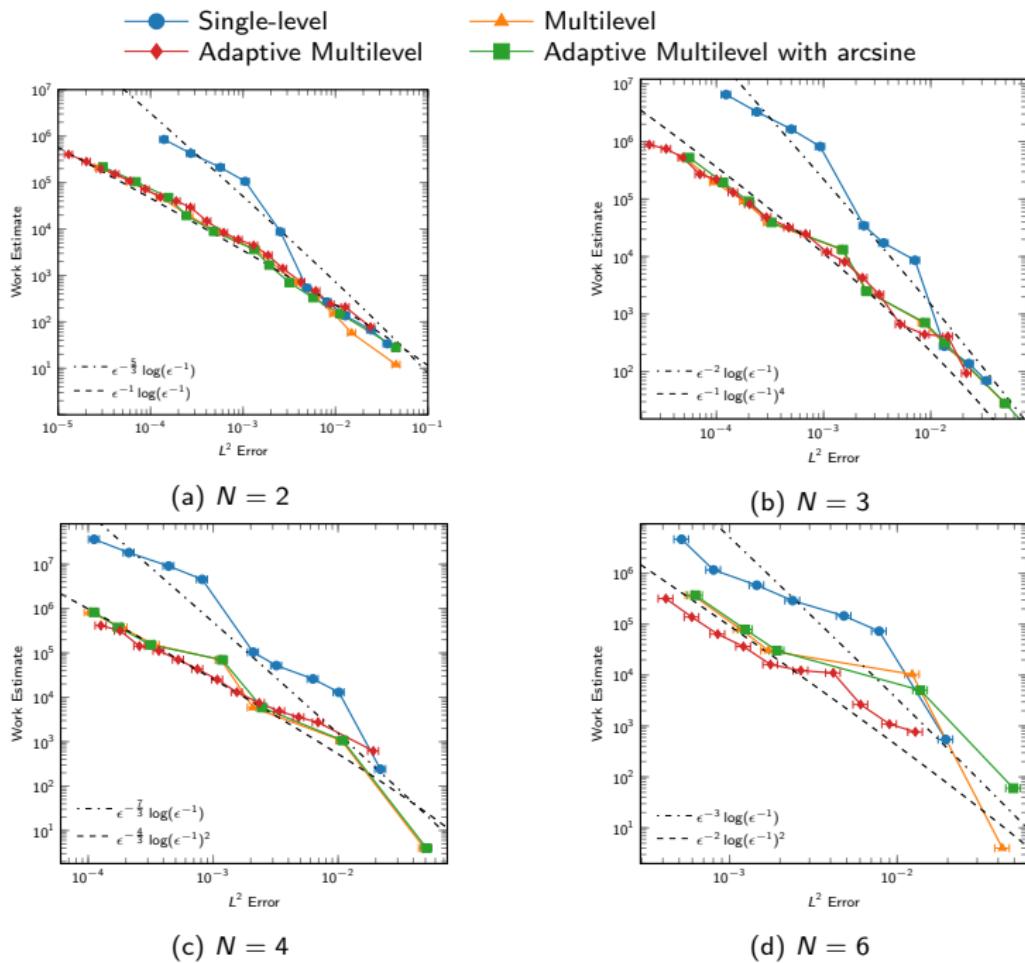
- ▶ Complexity of multilevel method

$$\text{Work}_{\text{ML}} = \mathcal{O} \left(\epsilon^{-\max\left\{\frac{d}{r+1}, \frac{N}{s}\right\}} (\log \epsilon^{-1})^t \right),$$

with

$$t = \begin{cases} 1, & \text{if } \frac{d}{r+1} > \frac{N}{s}, \\ 3 + \frac{d}{r+1}, & \text{if } \frac{d}{r+1} = \frac{N}{s}, \\ 2, & \text{if } \frac{d}{r+1} < \frac{N}{s}. \end{cases}$$

- ▶ In our experiments: $d = 2, r = 1, s = 3$ and $N = 2, 3, 4, 6$.



Conclusions

- ▶ We have derived a multilevel discrete least squares method for polynomial approximation of an output quantity of interest of a random PDE.
- ▶ The method uses the classical “Combination technique” and sparsifies sequences of polynomial approximations, obtained by weighted discrete least squares, and sequences of spatial discretizations of the underlying PDE.
- ▶ In particular, we have proposed a way to select the number of sample points on each level to guarantee the overall stability and accuracy of the ML formula with high probability.
- ▶ Complexity analysis carries over to infinite dimensional problems (different choice of polynomial spaces).

References

-  [A.-L. Haji-Ali, F. Nobile, R. Tempone, S. Wolfers,](#)
Multilevel weighted least squares polynomial approximation, arXiv:1707.00026.
-  [A.-L. Haji-Ali, F. Nobile, L. Tamellini and R. Tempone.](#)
Multi-Index Stochastic Collocation convergence rates for random PDEs with parametric regularity, FoCM 16(2016) 1555-1605.
-  [A.-L. Haji-Ali, F. Nobile, L. Tamellini and R. Tempone.](#)
Multi-Index Stochastic Collocation for random PDEs, CMAME 306(2016) 95–122.
-  [A.-L. Haji-Ali, F. Nobile, R. Tempone,](#)
Multi index Monte Carlo: when sparsity meets sampling, Numer. Math. 132(2016) 767–806, published online.
-  [A. Chkifa, A. Cohen, G. Migliorati, F. Nobile, and R. Tempone](#)
Discrete least squares polynomial approximation with random evaluations – application to parametric and stochastic elliptic PDEs, ESAIM: M2AN, vol. 49, num. 3, p. 815-837, 2015.

Sketch of the proof

- ▶ Bound on M_k : use that $\sqrt{M_k} \leq \frac{M_k}{\log M_k} \leq \frac{2m_k^\sigma}{\kappa_L}$ and $\kappa_L \approx 1/(L+1)$

$$\begin{aligned} M_k &\leq \frac{2}{\kappa_L} m_k^\sigma \log M_k \lesssim (L+1) e^{\frac{k\sigma}{\sigma+\alpha_p}} \\ &\lesssim (L+1) \log(L+1) e^{\frac{k\sigma}{\sigma+\alpha_p}} (k+1) \end{aligned}$$

- ▶ Bound on total work:

$$\begin{aligned} \text{Work}(S_L f) &\lesssim \sum_{\ell=0}^L M_{L-\ell} n_\ell^\gamma \\ &\lesssim (L+1) \log(L+1) e^{\frac{L\sigma}{\sigma-\alpha_p}} \sum_{\ell=0}^L \exp \left\{ -I \left(\frac{\sigma}{\sigma-\alpha_p} - \frac{\gamma}{\gamma+\beta_s} \right) \right\} (L-\ell+1) \end{aligned}$$

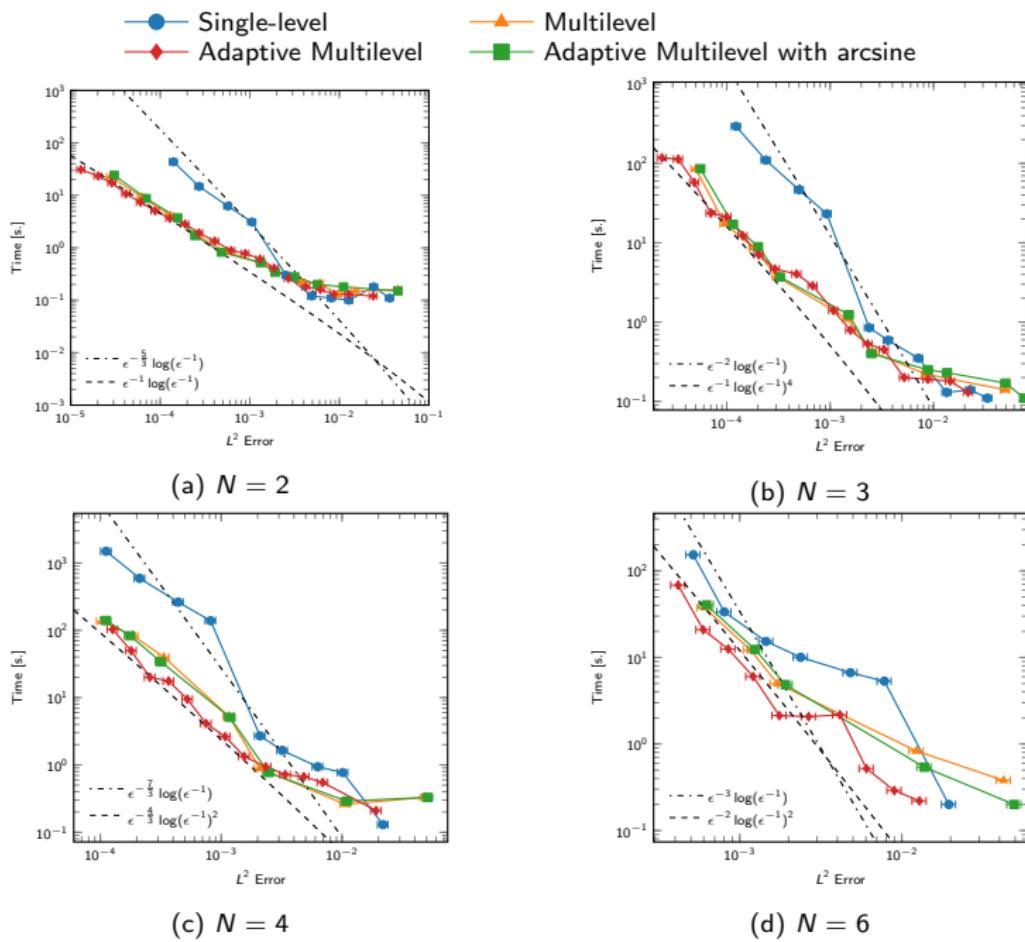
hence, distinguish three cases $\gamma/\beta_s <, =, > \sigma/\alpha_p$

Sketch of the proof

- ▶ Bound on the error in probability:

$$\begin{aligned}\|f - S_L f\|_{L^2_\mu} &= \|f - f_L + \sum_{\ell=0}^L (Id - \hat{\Pi}_{M_{L-\ell}})(f_\ell - f_{\ell-1})\|_{L^2_\mu} \\ &\leq \|f - f_L\|_{L^2_\mu} + \sum_{\ell=0}^L \|Id - \hat{\Pi}_{M_{L-\ell}}\|_{F \rightarrow L^2_\mu} \|f_\ell - f_{\ell-1}\|_F \\ &\lesssim e^{-\frac{L\beta_w}{\gamma+\beta_s}} + e^{-\frac{L\alpha}{\sigma+\alpha}} \sum_{\ell=0}^L \exp \left\{ \ell \left(\frac{\alpha}{\sigma+\alpha_p} - \frac{\beta_s}{\gamma+\beta_s} \right) \right\}\end{aligned}$$

Again split the three cases $\gamma/\beta_s <, =, > \sigma/\alpha_p$ and notice that the first term $e^{-\frac{L\beta_w}{\gamma+\beta_s}}$ is always negligible as $\beta_w > \beta_s$.



Polynomial Interpolation on Sparse Grids

Polynomial approximation by interpolation: Sparse Grids

An alternative approach to obtain a suitable polynomial surrogate for $\mathbf{y} \mapsto Q(u)(\mathbf{y})$, which can be viewed somewhat complementary to the L^2 projection method discussed above, is based on interpolation.

Consider again a Hilbert-valued function $u: \Gamma \rightarrow V$, $u \in L_p^2(\Gamma; V)$, where $\mathbf{y} \in \Gamma \subset \mathbb{R}^N$ is a random vector with joint probability density function $\rho: \Gamma \rightarrow \mathbb{R}_+$.

Interpolation then entails:

1. Carefully choose M “good” interpolation points $\{\mathbf{y}^k\}_{k=1}^M$.
2. Compute the corresponding solutions $u^{(k)} = u(\mathbf{y}^k) \in V$, $k = 1, \dots, M$; this implies solving M PDEs.
3. Construct a suitable approximation $\Pi_{\Lambda}^M u \in \mathbb{P}_{\Lambda}(\Gamma) \otimes V$ by using standard polynomial interpolation techniques.

Objectives of this part:

1. review of $N = 1$ dimensional polynomial interpolation
2. discuss extensions to interpolation on multidimensional ([sparse](#)) grids

Review 1D interpolation – single random variable

Given m interpolation points $y^1, \dots, y^m \in \Gamma \subset \mathbb{R}$, define the interpolant operator $\mathcal{U}^m : C^0(\Gamma, V) \rightarrow \mathbb{P}_{m-1}(\Gamma) \otimes V$ by

$$\mathcal{U}^m(u)(y) = \sum_{j=1}^m u(y^j) l_j(y), \quad \text{with } l_j(y) = \prod_{k \neq j} \frac{y - y^k}{y^j - y^k}, \quad u(y^j) \in V.$$

Good interpolation points are:

- ▶ **Gauss points** (zeros of orthogonal polynomials w.r.t. density ρ).
Best suited for approximation in $L^2_\rho(\Gamma)$.
- ▶ For Γ bounded, **Chebyshev** or **Clenshaw-Curtis points** (extrema of Chebyshev polynomials): best suited for approximation in $L^\infty(\Gamma)$.

They both lead to spectral convergence.

Other choices are equally possible (not discussed in this course): Leja points, Gauss-Patterson (nested Gaussian sequences), ...

Error analysis

Let us analyze the error $\|u - \mathcal{U}^m u\|_*$ where the norm $\|\cdot\|_*$ will be either $L_\rho^2(\Gamma; V)$ or $L^\infty(\Gamma; V)$. Moreover, we define the operator norm

$$\|\mathcal{U}^m\|_* = \sup_{v \in C^0(\Gamma, V)} \frac{\|\mathcal{U}^m v\|_*}{\|v\|_{L^\infty(\Gamma; V)}}.$$

Since \mathcal{U}^m is a linear operator that is exact on polynomials of degree less than m , we immediately have the following result¹¹

$$\|u - \mathcal{U}^m u\|_* \leq (1 + \|\mathcal{U}^m\|_*) \inf_{v \in \mathbb{P}_{m-1}(\Gamma) \otimes V} \|u - v\|_{L^\infty(\Gamma; V)}$$

Proof.

Indeed for any $v \in \mathbb{P}_{m-1}(\Gamma) \otimes V$ we have

$$\begin{aligned}\|u - \mathcal{U}^m u\|_* &\leq \|u - v\|_* + \|v - \mathcal{U}^m u\|_* \\&\leq \|u - v\|_* + \|\mathcal{U}^m(v - u)\|_* \\&\leq \|u - v\|_* + \|\mathcal{U}^m\|_* \|v - u\|_{L^\infty(\Gamma; V)} \\&\leq (1 + \|\mathcal{U}^m\|_*) \|u - v\|_{L^\infty(\Gamma; V)}\end{aligned}$$

□

¹¹Compare this to the results for the discrete least squares projection!

Operator norm in L^∞ : Lebesgue constant

The previous result shows that the interpolation error $\|u - \mathcal{U}^m u\|_*$ is (roughly) proportional to the best approximation error in $L^\infty(\Gamma; V)$ as long as the operator norm $\|\mathcal{U}^m\|_*$ is well behaving (i.e., finite).

The operator norm $\|\mathcal{U}^m\|_{L^\infty(\Gamma; V)}$ is traditionally called **Lebesgue constant** and is characterized by

$$\mathbb{L}(m) := \|\mathcal{U}^m\|_\infty = \sup_{y \in \Gamma} \sum_{j=1}^m |l_j(y)|.$$

Exercise 16.1

Proof the identity above.

For $\Gamma = [-1, 1]$, Chebyshev, Clenshaw-Curtis and Gauss-Legendre points have well behaving Lebesgue constant $\mathbb{L}(m) \sim \log m$.

Other interpolation points might have a much faster growth of the Lebesgue constant, e.g.,

- ▶ Leja points: polynomial growth,
- ▶ uniform points: exponential growth!

Operator norm in L^2

It clearly holds

$$\|\mathcal{U}^m\|_2 \leq \|\mathcal{U}^m\|_\infty$$

so the previous bounds can be applied in this case as well.

However, for Gauss abscissas and bounded random variables, it even holds that

$$\|\mathcal{U}^m\|_2 = 1 .$$

Proof: We recall that the Gaussian quadrature with m points,

$\int_{\Gamma} v \rho dy \approx \sum_{j=1}^m v(y^j) \alpha_j$, has degree of exactness $2m - 1$. Hence the Lagrange basis functions $\{l_j\}_{j=1}^m$ built on Gauss points are mutually orthogonal since

$$\int_{\Gamma} l_\ell(y) l_k(y) \rho(y) dy = \sum_{j=1}^m l_\ell(y^j) l_k(y^j) \alpha_j = \delta_{\ell,k} \alpha_k .$$

As $\sum_{k=1}^m \alpha_k = 1$, for any $v \in C^0(\Gamma; V)$ we thus find

$$\begin{aligned} \|\mathcal{U}^m v\|_{L_\rho^2(\Gamma; V)}^2 &= \int_{\Gamma} \|\mathcal{U}^m v(y)\|_V^2 \rho(y) dy = \int_{\Gamma} \left(\sum_{\ell=1}^m v(y^\ell) l_\ell(y), \sum_{k=1}^m v(y^k) l_k(y) \right)_V \rho(y) dy \\ &= \sum_{\ell,k=1}^m (v(y^\ell), v(y^k))_V \underbrace{\int_{\Gamma} l_\ell(y) l_k(y) \rho(y) dy}_{\delta_{\ell,k} \alpha_k} = \sum_{k=1}^m \alpha_k \|v(y^k)\|_V^2 \leq \|v\|_{L^\infty(\Gamma; V)}^2 \end{aligned}$$

Higher dimensions: Tensor Product interpolation

Let $u = u(y_1, \dots, y_N) \in V$ be a function of N variables $(y_1, \dots, y_N) \in \Gamma$ and consider the **tensor product interpolation formula** that uses m_i points in the variable y_i , $i = 1, \dots, N$. Denote $\mathbf{m} = (m_1, \dots, m_N)$. Introduce the tensor interpolation as

$$\begin{aligned}\mathcal{I}_{\mathbf{m}}^N[u](y) &= (\mathcal{U}^{m_1} \otimes \cdots \otimes \mathcal{U}^{m_N})[u](\mathbf{y}) \\ &= \sum_{j_1=1}^{m_1} \cdots \sum_{j_N=1}^{m_N} u(y^{j_1}, \dots, y^{j_N}) \cdot (I_{j_1}(y_1) \otimes \cdots \otimes I_{j_N}(y_N)) .\end{aligned}$$

- The interpolation $\mathcal{I}_{\mathbf{m}}^N[u](y)$ requires $M = \prod_{i=1}^N m_i$ function eval.
- To $\mathcal{I}_{\mathbf{m}}^N$ one can associate a corresponding **interpolatory quadrature**:

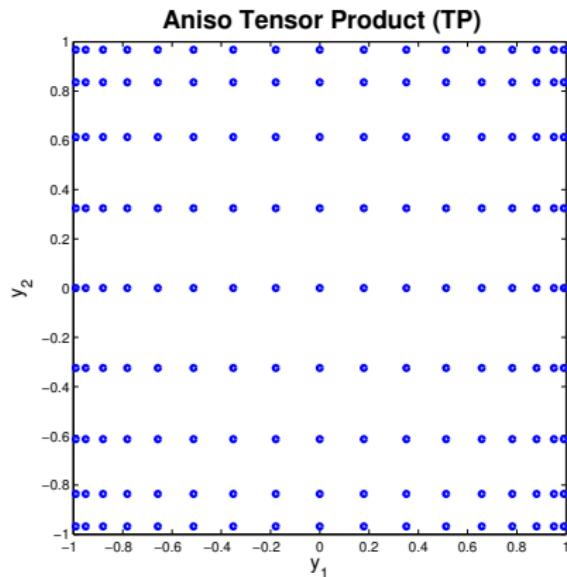
$$E_{\mathbf{m}}^N[u] = \sum_{j_1=1}^{m_1} \cdots \sum_{j_N=1}^{m_N} u(y^{j_1}, \dots, y^{j_N}) \alpha_{j_1} \cdots \alpha_{j_N}$$

- **Error analysis:** Observe that $\|\mathcal{I}_{\mathbf{m}}^N\|_* = \prod_{n=1}^N \|\mathcal{U}^{m_n}\|_*$ so that

$$\|u - \mathcal{I}_{\mathbf{m}}^N u\|_{L^*(\Gamma; V)} \leq (1 + \prod_{n=1}^N \|\mathcal{U}^{m_n}\|_*) \inf_{v \in \mathbb{P}_{TP(\mathbf{m}-1)} \otimes V} \|u - v\|_{L^\infty(\Gamma, V)}$$

Example: full anisotropic Tensor Grid in dimension 2

17×9 Clenshaw-Curtis points



Convergence for tensor grid of Gauss points

Theorem 16.1 ([Babuška, Nobile, Tempone. SINUM '07])

Let $u: \Gamma^N \rightarrow V$, with $\Gamma = \prod_{n=1}^N \Gamma_n \subset \mathbb{R}^N$, be a function *analytic* in each variable in the region of the complex domain

$u(z_n, \cdot)$ analytic in $\Sigma_n \equiv \{z \in \mathbb{C} : \text{dist}(z, \Gamma_n) \leq \mathcal{T}_n\}$.

Then, the interpolation on a *tensor grid* using (m_1, \dots, m_n) Gauss points in each direction yields

i) Γ_n bounded

$$\|u - \mathcal{I}_{\mathbf{m}}^N u\| \leq C \sum_{n=1}^N e^{-g_n m_n}, \quad \text{with } g_n = \log \left\{ \frac{2\mathcal{T}_n}{|\Gamma_n|} + \sqrt{1 + \frac{4\mathcal{T}_n^2}{|\Gamma_n|^2}} \right\}$$

ii) Γ_n unbounded, $\hat{\rho}_n \approx e^{-(\delta_n y_n)^2}$ at infinity

$$\|u - \mathcal{I}_{\mathbf{m}}^N u\| \leq C \sum_{n=1}^N \sqrt{m_n} e^{-g_n \sqrt{m_n}}, \quad \text{with } g_n = \sqrt{2\mathcal{T}_n}/\delta_n$$

where the error being measured in the $L_p^2(\Gamma^N, V)$ norm.

Curse of dimensionality

For large N , Tensor Product interpolation is impractical !

For a tensor product interpolation of degree p in each variable,

$$\text{\#dofs: } M = (p + 1)^N \quad \text{curse of dimensionality}$$

Exponential convergence w.r.t. polynomial degree p does not imply exponential convergence w.r.t. M

$$\text{error} \sim e^{-g p} \quad \implies \quad \text{error} \sim e^{-g M^{1/N}} \sim M^{-\frac{g}{N}} \text{ for } N \text{ large}$$

Hierarchical interpolation - 1D case

Let $m(i) : \mathbb{N} \rightarrow \mathbb{N}$ a strictly increasing function with $m(1) = 1$;

- ▶ i : level of interpolation
- ▶ $m(i)$: number of interpolation points used at level i

Hierarchical construction: example for level 3 interpolation $\mathcal{U}^{m(3)}$:

$$\mathcal{U}^{m(3)}(u) = \underbrace{\mathcal{U}^{m(1)}(u)}_{\Delta^{m(1)}} + (\underbrace{\mathcal{U}^{m(2)} - \mathcal{U}^{m(1)}}_{\Delta^{m(2)}})(u) + (\underbrace{\mathcal{U}^{m(3)} - \mathcal{U}^{m(2)}}_{\Delta^{m(3)}})(u)$$

The incremental formula uses

- ▶ $m(3)$ evaluations if nested sequences of points are used (as Clenshaw-Curtis or Leja points)
- ▶ $\sim \sum_{i=1}^3 m(i)$ evaluations if non-nested sequences of points are used (as Gauss points) – not very attractive in 1D.

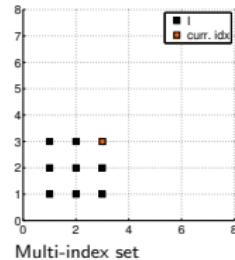
Clearly $u = \sum_{i=1}^{\infty} \Delta^{m(i)}(u)$.

Hierarchical interpolation - 2D case

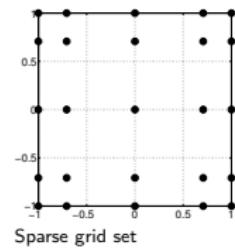
Hierarchical construction of bivariate interpolation $\mathcal{U}^{m(3)} \otimes \mathcal{U}^{m(3)}(u) =: \mathcal{U}^{m(3,3)}(u)$.

Bivariate increments related to 1D interpolation. Introduce double difference

$$\Delta^{m(i,j)}(u) = (\mathcal{U}^{m(i)} - \mathcal{U}^{m(i-1)}) \otimes (\mathcal{U}^{m(j)} - \mathcal{U}^{m(j-1)})(u)$$



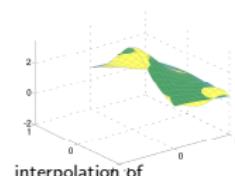
E.g., nested Clenshaw-Curtis points: $m(1) = 1$, $m(2) = 3$, $m(3) = 5$.



$$\begin{aligned}\mathcal{U}^{m(3,3)}(u) &= \Delta^{m(1,1)}(u) + \Delta^{m(2,1)}(u) + \Delta^{m(1,2)}(u) \\ &\quad + \Delta^{m(2,2)}(u) + \Delta^{m(1,3)}(u) + \Delta^{m(3,1)}(u) \\ &\quad + \Delta^{m(3,2)}(u) + \Delta^{m(2,3)}(u) + \Delta^{m(3,3)}(u)\end{aligned}$$

Also here $u = \sum_{i,j=1}^{\infty} \Delta^{m(i,j)}(u)$ by construction.

NB: we don't need to form the full tensor grid



$$f(y) = \frac{1}{y_1^2 + y_2^2 + 0.3}$$

Sparse grids – General construction in N dimensions

Let $\mathbf{i} = [i_1, \dots, i_N] \in \mathbb{N}_+^N$ and $m(i) : \mathbb{N}_+ \rightarrow \mathbb{N}_+$ an increasing function

- 1D polynomial interpolant operators: $\mathcal{U}_n^{m(i_n)}$ on $m(i_n)$ abscissas.
- Detail operator: $\Delta_n^{m(i_n)} = \mathcal{U}_n^{m(i_n)} - \mathcal{U}_n^{m(i_n-1)}$, $\mathcal{U}_n^{m(0)} = 0$.
- Hierarchical surplus: $\Delta^{m(\mathbf{i})} = \bigotimes_{n=1}^N \Delta_n^{m(i_n)}$.
- Sparse grid approximation: on an index set $\mathcal{I} \subset \mathbb{N}^N$

$$\mathcal{S}_{\mathcal{I}} u = \sum_{\mathbf{i} \in \mathcal{I}} \Delta^{m(\mathbf{i})}[u] \quad (92)$$

Assumption: The set \mathcal{I} is downward closed:

$$\mathbf{i} \in \mathcal{I} \Rightarrow \mathbf{i} - \mathbf{e}_n \in \mathcal{I}, \quad n = 1, \dots, N.$$

This construction can be applied also in infinite dimension ($N = \infty$).
Indeed, from the previous assumption it follows that if $\#\mathcal{I} < \infty$

$$\exists \bar{N} : \forall \mathbf{i} \in \mathcal{I}, \quad i_n = 1 \quad \forall n > \bar{N}.$$

Equivalent formulation

We can rewrite the sparse approximation in terms of the interpolation operators, namely

$$\mathcal{S}_{\mathcal{I}}^m[u] = \sum_{\mathbf{i} \in \mathcal{I}} c(\mathbf{i}) \left(\mathcal{U}_1^{m(i_1)} \otimes \cdots \otimes \mathcal{U}_N^{m(i_N)} \right) [u]. \quad (93)$$

$$\text{with } c(\mathbf{i}) = \sum_{\substack{\mathbf{j} \in \{0,1\}^N \\ (\mathbf{i} + \mathbf{j}) \in \mathcal{I}}} (-1)^{|\mathbf{j}|_1}$$

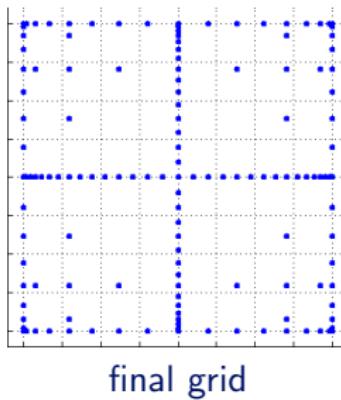
- ▶ linear combination of full tensor grids, each with relatively low number of points (!).
- ▶ Given \mathbf{i} , if $(\mathbf{i} + \mathbf{j}) \in \mathcal{I}$ for all $\mathbf{j} \in \{0,1\}^N$ then $c(\mathbf{i}) = 0$.
- ▶ The collection of all tensor grids $\mathcal{H}^{m(\mathbf{i})}$, with $\mathbf{i} \in \mathcal{I}$, $c(\mathbf{i}) \neq 0$, forms the sparse grid $\mathcal{H}_{\mathcal{I}}^m$

Remark: The approximation $\mathcal{S}_{\mathcal{I}}^m(u)$ is interpolatory only if nested points are used !

Example – classical Smolyak sparse grid

- ▶ Uses Clenshaw-Curtis abscissas in $[-1,1]$, i.e., the extrema of Chebyshev polynomials,
- ▶ $m(i) = 2^{i-1} + 1, \ m(1) = 1 \implies$ nested grids !
- ▶ $\mathcal{I}(w) \equiv \{\mathbf{i} \in \mathbb{N}^N : \sum_{n=1}^N (i_n - 1) \leq w\}$

$N = 2$ isotropic sparse grid: $p = 5$ (\Rightarrow max. pol. degree 32)



Exactness of the sparse grid approximation

Theorem 16.2 ([Beck et al. 2011])

Consider the Polynomial space $\mathbb{P}_{\Lambda(\mathcal{I}, m)}(\Gamma)$ with

$$\Lambda(\mathcal{I}, m) = \{\mathbf{p} \in \mathbb{N}_0^N : \mathbf{p} < m(\mathbf{i}) \text{ for some } \mathbf{i} \in \mathcal{I}\}.$$

Then, $S_{\mathcal{I}}^m[u] \in \mathbb{P}_{\Lambda(\mathcal{I}, m)}(\Gamma)$ for any $u(\mathbf{y}) \in C^0(\Gamma; V)$. Moreover, the approximation $S_{\mathcal{I}}^m$ is exact in $\mathbb{P}_{\Lambda(\mathcal{I}, m)}$.

Proof: The first claim is immediate using the representation (93). For the second one (i.e., the exactness), we need to show that $S_{\mathcal{I}}^m[\mathbf{y}^{\mathbf{p}}] = \mathbf{y}^{\mathbf{p}}$ for any $\mathbf{p} \in \Lambda(\mathcal{I}, m)$. Notice,

$$S_{\mathcal{I}}^m[\mathbf{y}^{\mathbf{p}}] = \sum_{\mathbf{i} \in \mathcal{I}} \Delta^{m(\mathbf{i})}[\mathbf{y}^{\mathbf{p}}] = \sum_{\mathbf{i} \in \mathcal{I}} \bigotimes_{n=1}^N \Delta^{m(i_n)}[y_n^{p_n}] = \sum_{\mathbf{i} \in \mathcal{I}} \bigotimes_{n=1}^N (\mathcal{U}^{m(i_n)} - \mathcal{U}^{m(i_n-1)})[y_n^{p_n}].$$

Now, if $m(i_n - 1) \geq p_n + 1$ then $(\mathcal{U}^{m(i_n)} - \mathcal{U}^{m(i_n-1)})[y_n^{p_n}] = 0$ by the exactness of \mathcal{U}^m . Therefore, setting \bar{i}_n such that $m(\bar{i}_n - 1) \leq p_n$ and $m(\bar{i}_n) > p_n$, then $\bar{\mathbf{i}} = (\bar{i}_1, \dots, \bar{i}_N) \in \mathcal{I}$ whenever $\mathbf{p} \in \Lambda(\mathcal{I}, m)$ and

$$\begin{aligned} S_{\mathcal{I}}^m[\mathbf{y}^{\mathbf{p}}] &= \sum_{\mathbf{i} \in \mathcal{I}, m(\mathbf{i}-1) \leq \mathbf{p}} \bigotimes_{n=1}^N (\mathcal{U}^{m(i_n)} - \mathcal{U}^{m(i_n-1)})[y_n^{p_n}] \\ &= \bigotimes_{n=1}^N \sum_{i_n=1}^{\bar{i}_n} (\mathcal{U}^{m(i_n)} - \mathcal{U}^{m(i_n-1)})[y_n^{p_n}] = \bigotimes_{n=1}^N \mathcal{U}^{m(\bar{i}_n)}[y_n^{p_n}] = \mathbf{y}^{\mathbf{p}}. \end{aligned}$$

Remarks

- ▶ In general the number of points in the sparse grid space $\mathcal{H}_{\mathcal{I}}^m$ is larger than the dimension of the corresponding polynomial space:

$$\#\mathcal{H}_{\mathcal{I}}^m > \dim(\mathbb{P}_{\Lambda(\mathcal{I},m)}) !$$

- ▶ We have $\#\mathcal{H}_{\mathcal{I}}^m = \dim(\mathbb{P}_{\Lambda(\mathcal{I},m)})$, if either
 - ▶ \mathcal{I} is a TP set or,
 - ▶ \mathcal{I} is arbitrary and nested sequences of points are used.

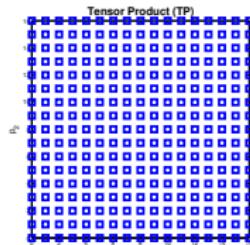
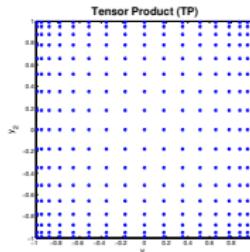
Choice of the index set \mathcal{I}

Approx. space	SC: $m, \mathcal{I}(w)$
Tensor Product (TP)	$m(i) = i$ $\mathcal{I}(w) \equiv \{\mathbf{i} \in \mathbb{N}^N : \max_n(i_n - 1) \leq w\}$
Total Degree (TD)	$m(i) = i$ $\mathcal{I}(w) \equiv \{\mathbf{i} \in \mathbb{N}^N : \sum_n(i_n - 1) \leq w\}$
Hyperbolic Cross (HC)	$m(i) = i$ $\mathcal{I}(w) \equiv \{\mathbf{i} \in \mathbb{N}^N : \prod_n(i_n) \leq w + 1\}$
Smolyak (SM)	$m(i) = \begin{cases} 2^{i-1} + 1, & i > 1 \\ 1, & i = 1 \end{cases}$ $\mathcal{I}(w) \equiv \{\mathbf{i} \in \mathbb{N}^N : \sum_n(i_n - 1) \leq w\}$

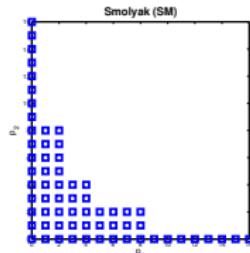
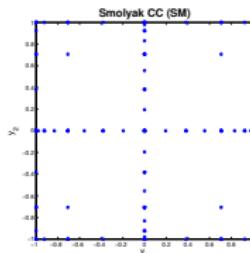
- ▶ we can build a sparse grid corresponding to any polynomial space \mathbb{P}_Λ
- ▶ The corresponding anisotropic versions are straightforward
- ▶ The last choice (SM) is the most popular and corresponds to the Smolyak construction

Examples of sparse grids

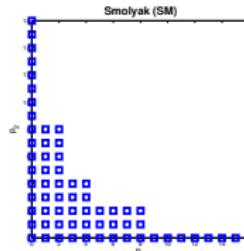
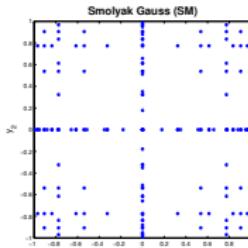
$N = 2$, max. polynomial degree $p = 16$



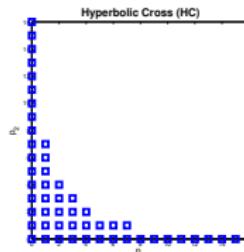
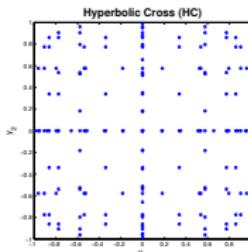
tensor product (**TP**) space



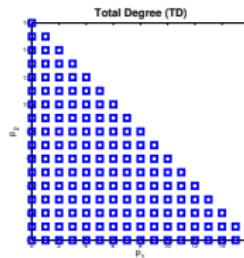
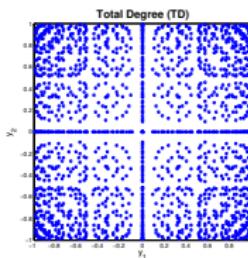
“Smolyak” (**SM**) space (nested points)



"Smolyak" (**SM**) space (non-nested pts)

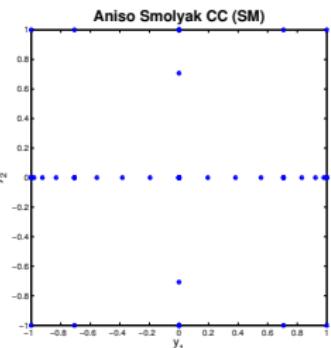
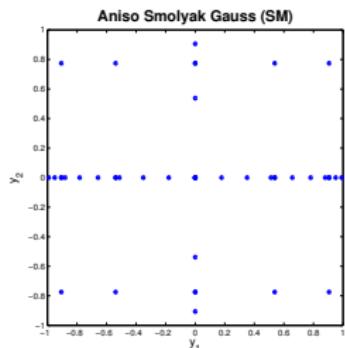
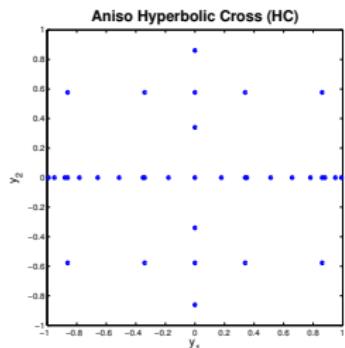
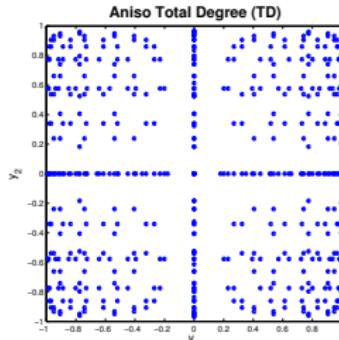
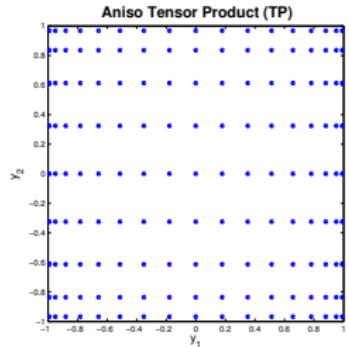


hyperbolic cross (**HC**) space



total degree (**TD**) space

Examples of anisotropic sparse grids: $\alpha = (1, 2)$



Sparse grid quadrature

Starting from the sparse grid approximation $S_{\mathcal{I}}^m[u]$ one can build a corresponding sparse grid quadrature formula

$$\mathcal{Q}_{\mathcal{I}}^m[u] = \int_{\Gamma} S_{\mathcal{I}}^m[u](\mathbf{y}) \rho(\mathbf{y}) d\mathbf{y} = \sum_{\mathbf{y}_\ell \in \mathcal{H}_{\mathcal{I}}^m} u(\mathbf{y}_\ell) \alpha_\ell$$

where the density $\rho(\mathbf{y}) = \prod_{n=1}^N \rho_n(y_n)$ is the one associated to the interpolation points used.

Remarks

- ▶ In general, the quadrature formula might have negative weights α_ℓ !
- ▶ If \mathcal{I} is a *TP* set and Gauss points are used, the quadrature reduces to the tensor product Gaussian quadrature and the weights are all positive.
- ▶ The quadrature formula will always be exact in $\mathbb{P}_{\Lambda(\mathcal{I}, m)}$.
- ▶ If nested points are used, the quadrature (interpolation) formula is unisolvant on $\mathbb{P}_{\Lambda(\mathcal{I}, m)}$; however, in general will have only exactness in $\mathbb{P}_{\Lambda(\mathcal{I}, m)}^m$ (Gaussian formulae are not nested!)
- ▶ If \mathcal{I} is *TD(w)* and Gauss points are used, the quadrature formula is Gaussian, i.e. exact in $TD(2w + 1)$; however, the number of points used is much larger than $\dim(TD(w))$ (a factor 2^w larger) .

Polynomial Interpolation on Sparse Grids: Error estimation and adaptivity

On error estimation for sparse grids interpolation

A general way to derive an error estimate for a sparse grid approximation $\mathcal{S}_{\mathcal{I}(w)}^m[u]$ for a given sequence $\mathcal{I}(w)$, $w = 0, 1, \dots$ of index sets consists of the following ingredients:

► Error estimation

$$\begin{aligned}\mathcal{E}(w) &= \|u - \mathcal{S}_{\mathcal{I}(w)}^m u\|_{L_p^2(\Gamma; V)} = \left\| \sum_{\mathbf{i} \notin \mathcal{I}(w)} \Delta^{m(\mathbf{i})}[u] \right\|_{L_p^2(\Gamma; V)} \\ &\leq \sum_{\mathbf{i} \notin \mathcal{I}(w)} \|\Delta^{m(\mathbf{i})}[u]\|_{L_p^2(\Gamma; V)} \leq \mathcal{E}_{ub}(w)\end{aligned}$$

► Total work

$$\mathcal{W}(w) = \sum_{\mathbf{i} \in \mathcal{I}(w)} \text{Work}(\Delta^{m(\mathbf{i})}(u)) \geq \mathcal{W}_{lb}(w)$$

Then, assuming that the function \mathcal{W}_{lb} is strictly increasing, a bound can be obtained for the error versus total work

$$\mathcal{E} \leq \mathcal{E}_{up}(\mathcal{W}_{lb}^{-1}(\mathcal{W})) .$$

(Quasi) optimal sparse grids

$$\mathcal{S}_{\mathcal{I}} u = \sum_{\mathbf{i} \in \mathcal{I}} \Delta^{m(\mathbf{i})}[u] \implies \|u - \mathcal{S}_{\mathcal{I}} u\| = \left\| \sum_{\mathbf{i} \notin \mathcal{I}} \Delta^{m(\mathbf{i})}[u] \right\| \leq \sum_{\mathbf{i} \notin \mathcal{I}} \|\Delta^{m(\mathbf{i})}[u]\|$$

One can use a [knapsack problem](#) approach¹² to select the best \mathcal{I} . That is, for each multiindex \mathbf{i} one considers

- ▶ [estimated error contribution](#) (how much error decreases if \mathbf{i} is added to \mathcal{I})

$$\Delta E(\mathbf{i}) \geq \|\Delta^{m(\mathbf{i})}[u]\|$$

- ▶ [estimated work contribution](#) (how much the work, i.e. number of evaluations, increases if \mathbf{i} is added to \mathcal{I})

$$\Delta W(\mathbf{i}) \quad \text{such that} \quad \sum_{\mathbf{i} \in \mathcal{I}} \Delta W(\mathbf{i}) \geq W(\mathcal{I})$$

where $W(\mathcal{I})$ is the total number of points in the sparse grid

¹²see, e.g., [Griebel-Knapek '09, Gerstner-Griebel '03, Bungartz-Griebel '04]

Then we estimate the profit of each \mathbf{i} as

$$P(\mathbf{i}) = \frac{\Delta E(\mathbf{i})}{\Delta W(\mathbf{i})}$$

and build the sparse grid using the set \mathcal{I}_M of the M indices with the largest estimated profit:

$$\mathcal{I}_M := \{\mathbf{i} \in \mathbb{N}^N \mid P(\mathbf{i}) \geq P_M^{ord}\}$$

where $\{P_j^{ord}\}_j$ is the ordered sequence of profits.

If the set \mathcal{I}_M is not downward closed (lower), take the smallest lower set $\tilde{\mathcal{I}}_M \supset \mathcal{I}_M$. This is equivalent to consider the modified profits $\tilde{P}(\mathbf{i}) = \max_{\mathbf{j} > \mathbf{i}} P(\mathbf{j})$; see ,e.g., [Chkifa-Cohen-DeVore-Schwab M2AN '13].

A priori constructions and error estimates

For some problems, sharp estimates are available for the error contributions $\Delta E(\mathbf{i})$; e.g., [Beck-Nobile-Tamellini-Tempone M3AS 2012, Nobile-Tamellini-Tempone 2014]).

This allows to

- ▶ construct a priori sequences of “quasi optimal index sets \mathcal{I} ”
- ▶ obtain error estimates

Theorem 16.3 ([Nobile-Tamellini-Tempone, 2014])

Let $\mathcal{S}_{\tilde{\mathcal{I}}_M} u$ be the quasi-optimal sparse grid approximation and $W_{\tilde{\mathcal{I}}_M}$ the total number of points in the sparse grid.

If $C(\tau) := \left(\sum_{\mathbf{i} \in \mathbb{N}^N} \tilde{P}(\mathbf{i})^\tau \Delta W(\mathbf{i}) \right)^{\frac{1}{\tau}} < \infty$ for some $\tau < 1$, then

$$\|u - \mathcal{S}_{\tilde{\mathcal{I}}_M} u\|_{L_\rho^2(\Gamma, V)} \leq C(\tau) W_{\tilde{\mathcal{I}}_M}^{1 - \frac{1}{\tau}} .$$

Proof.

The proof uses Stechkin's lemma 13.5, that is given a non-negative decreasing sequence $\{a_k\}_k$

$$\sum_{k=N+1}^{\infty} a_k \leq N^{1-\frac{1}{\tau}} \left(\sum_{k=1}^{\infty} a_k^{\tau} \right)^{\frac{1}{\tau}}, \quad 0 < \tau < 1.$$

In particular, this is applied to the ordered repeated sequence of profits

$$\{\hat{P}_k\}_k = \underbrace{\{\tilde{P}_1, \dots, \tilde{P}_1\}}_{\Delta W_1 \text{ times}}, \underbrace{\{\tilde{P}_2, \dots, \tilde{P}_2\}, \dots}_{\Delta W_2 \text{ times}}.$$

Moreover, let $W_M = \sum_{j=1}^M \Delta W_j$ and observe that $W_M \geq W_{\tilde{\mathcal{I}}_M}$. Then

$$\begin{aligned} \|u - \mathcal{S}_{\tilde{\mathcal{I}}_M} u\|_{L_p^2(\Gamma, V)} &\leq \sum_{k=M+1}^{\infty} \Delta E_k \leq \sum_{k=W_M+1}^{\infty} \hat{P}_k \quad [\text{apply Stechkin}] \\ &\leq \|\{\hat{P}_k\}_k\|_{l^\tau} W_M^{1-\frac{1}{\tau}} \leq C(\tau) W_{\tilde{\mathcal{I}}_M}^{1-\frac{1}{\tau}}. \end{aligned}$$



How to estimate ΔW and ΔE ?

- ▶ $\Delta W(\mathbf{i})$: number of newly added points in $\bigotimes_{n=1}^N \Delta_n^{m(i_n)}$.

Count all points in $\mathcal{U}_1^{m(i_1)} \otimes \cdots \otimes \mathcal{U}_N^{m(i_N)}$ (non-nested case) or just the extra points added (nested case), that is

$$\Delta W(\mathbf{i}) = \begin{cases} \prod_{n=1}^N (m(i_n) - m(i_n - 1)) & \text{nested points} \\ \prod_{n=1}^N m(i_n) & \text{non-nested points} \end{cases}$$

- ▶ $\Delta E(\mathbf{i})$: use suitable basis expansion $u = \sum_{\mathbf{p}} u_{\mathbf{p}} \psi_{\mathbf{p}}$ (e.g., Legendre, Chebyshev, ...) and relate $\Delta E(\mathbf{i})$ with the decay of the coefficients $u_{\mathbf{p}}$. For example, Chebyshev expansion ($\|\psi_{\mathbf{p}}\|_{\infty} = 1$) yields

$$\Delta E(\mathbf{i}) \leq 2 \mathbb{L}_{m(\mathbf{i})} \sum_{\mathbf{p} \geq m(\mathbf{i}-1)} \|u_{\mathbf{p}}\|_V .$$

Proof:

$$\begin{aligned} \Delta E(\mathbf{i}) &= \|\Delta^{m(\mathbf{i})}[u]\|_{V \otimes L^2_{\rho}} = \left\| \sum_{\mathbf{p}} u_{\mathbf{p}} \Delta^{m(\mathbf{i})} \psi_{\mathbf{p}} \right\|_{V \otimes L^2_{\rho}} \\ &\leq \sum_{\mathbf{p} \geq m(\mathbf{i}-1)} \|u_{\mathbf{p}}\|_V \|\Delta^{m(\mathbf{i})} \psi_{\mathbf{p}}\|_{L^2_{\rho}} \leq 2 \prod_{n=1}^N \mathbb{L}_{m(i_n)} \sum_{\mathbf{p} \geq m(\mathbf{i}-1)} \|u_{\mathbf{p}}\|_V \end{aligned}$$

where $\mathbb{L}_{m(\mathbf{i})} = \prod_{n=1}^N \mathbb{L}_{m(i_n)}$ and $\mathbb{L}_{m(i_n)} := \|\mathcal{U}_n^{m(i_n)}\|_{\mathcal{L}(C^0, L^2_{\rho})}$ is the Lebesgue constant from C^0 to L^2_{ρ} .

Application to a class of analytic functions

Let \mathbf{y} be independent and uniformly distributed in $[-1, 1]^N$, and let $u: [-1, 1]^N \rightarrow (V, \mathbb{R})$ be an analytic function that admits and holomorphic extension $u(\mathbf{z}): \Sigma \subset \mathbb{C}^N \rightarrow (V, \mathbb{C})$ in a polydisk¹³

$$\Sigma = \{\mathbf{z} \in \mathbb{C}^N : |z_n| \leq e^{g_n}, g_n > 0, n = 1, \dots, N\}.$$

Theorem 16.4 ([Nobile-Tamellini-Tempone, 2014])

Let $\mathcal{S}_{\tilde{\mathcal{I}}_M} u$ be the quasi-optimal sparse grid approximation and $W_{\tilde{\mathcal{I}}_M}$ the total number of points in the sparse grid. Then the profits $\tilde{P}(\mathbf{i})$ are τ -summable for any $0 < \tau \leq 1$. Moreover, for

- nested Clenshaw-Curtis points ($m(i) \sim 2^i$)

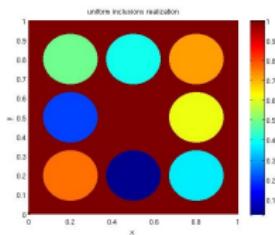
$$\|u - \mathcal{S}_{\tilde{\mathcal{I}}_M} u\|_{L_p^2(\Gamma, V)} \leq C_1(N, g) \exp\{-NC_2(g) W_{\tilde{\mathcal{I}}_M}^{\frac{1}{N}}\}$$

- non-nested Legendre points ($m(i) = i$)

$$\|u - \mathcal{S}_{\tilde{\mathcal{I}}_M} u\|_{L_p^2(\Gamma, V)} \leq \tilde{C}_1(N, g) \exp\{-N\tilde{C}_2(g) W_{\tilde{\mathcal{I}}_M}^{\frac{1}{2N}}\}$$

¹³Sharper estimates can be obtained if u is holomorphic in a polyellipse

Example: elliptic equation with random inclusions



- $\operatorname{div}(a \nabla u) = f$, in D , $u = 0$ on ∂D
- ▶ $a(\mathbf{y}, x) = a_0 + \sum_{n=1}^N y_n \mathbb{1}_{D_n}(x)$.
- ▶ $y_i \sim \mathcal{U}[a_i, b_i]$, independent
- ▶ $\Gamma = \prod_{i=1}^N [a_i, b_i]$
- ▶ $u(\mathbf{y}) : \Gamma \rightarrow H_0^1(D)$

- ▶ The solution $u : \Gamma \rightarrow V \equiv H_0^1(D)$ is analytic in a polydisk in the complex plane \mathbb{C}^N containing Γ and the previous theorem applies
- ▶ The profits can be estimated as

$$P(\mathbf{i}) = \frac{\Delta E(\mathbf{i})}{\Delta W(\mathbf{i})} \propto \prod_{n=1}^N \frac{\mathbb{L}_{m(i_n)} e^{-g_{i_n} m(i_n - 1)}}{\Delta W(i_n)}$$

where \mathbb{L}_n is the Lebesgue constant on the 1D interpolation scheme with n points and the rates g_i can be numerically estimated.¹⁴

¹⁴see [Beck-N.-Tamellini-Tempone M3AS '12, Cohen-DeVore-Schwab FoCM '10]

Adaptive sparse grids construction [Gerstner-Griebel '03, Klimke, PhD '06]

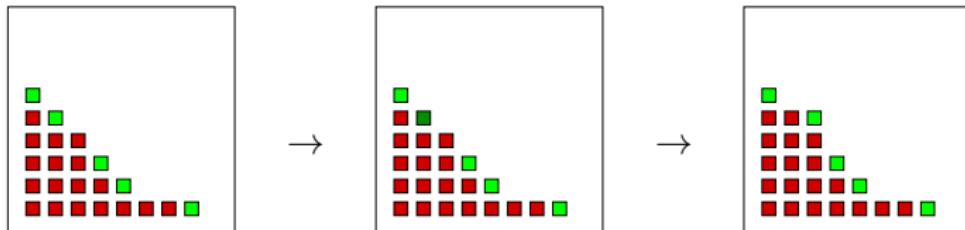
Definition 16.5

The **reduced margin** $\mathcal{R}(\mathcal{I})$ associated to a multi-index set \mathcal{I} is given by

$$\mathcal{R}(\mathcal{I}) = \{\mathbf{i} : \mathbf{i} \notin \mathcal{I} \text{ and } \forall k = 1, \dots, N : i_k \neq 1 \Rightarrow \mathbf{i} - \mathbf{e}_k \in \mathcal{I}\}.$$

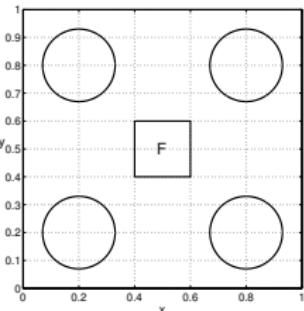
Adaptive algorithm: given the current index set \mathcal{I}_k and its reduced margin $\mathcal{R}_k = \mathcal{R}(\mathcal{I}_k)$, then repeatedly perform:

1. compute $P(\mathbf{i})$ for all $\mathbf{i} \in \mathcal{R}_k$ and find $\mathbf{j} = \arg \min_{\mathbf{i} \in \mathcal{R}_k} P(\mathbf{i})$
2. set $\mathcal{I}_{k+1} = \mathcal{I}_k \cup \{\mathbf{j}\}$ and compute the new reduced margin $\mathcal{R}_{k+1} = \mathcal{R}(\mathcal{I}_{k+1})$

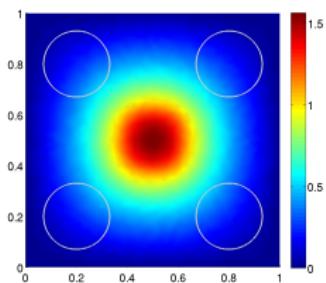


Remark: the index set always remains downward closed.

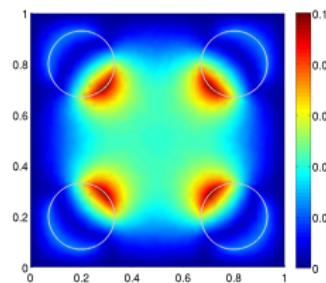
Isotropic test case – 4 random inclusions



- ▶ Conductivity coefficient: matrix $k=1$
circular inclusions: $k|_{\Omega_i} \sim \mathcal{U}(0.01, 1.99)$ →
4 iid uniform random variables
- ▶ forcing term $f = 100\mathbb{1}_F$
- ▶ zero boundary conditions
- ▶ quantity of interest $\psi(u) = \int_F u$



mean



std

convergence plot for $\|\psi(u) - \mathcal{S}_{\mathcal{I}}\psi(u)\|_{L_p^2(\Gamma)}$ versus # pts sparse grid

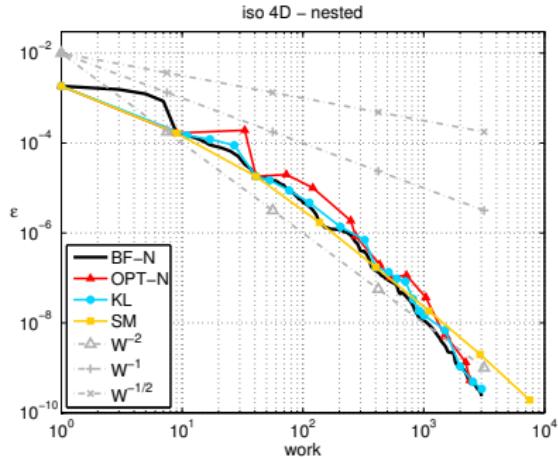


Figure: Results for the isotropic problem. (nested) CC points

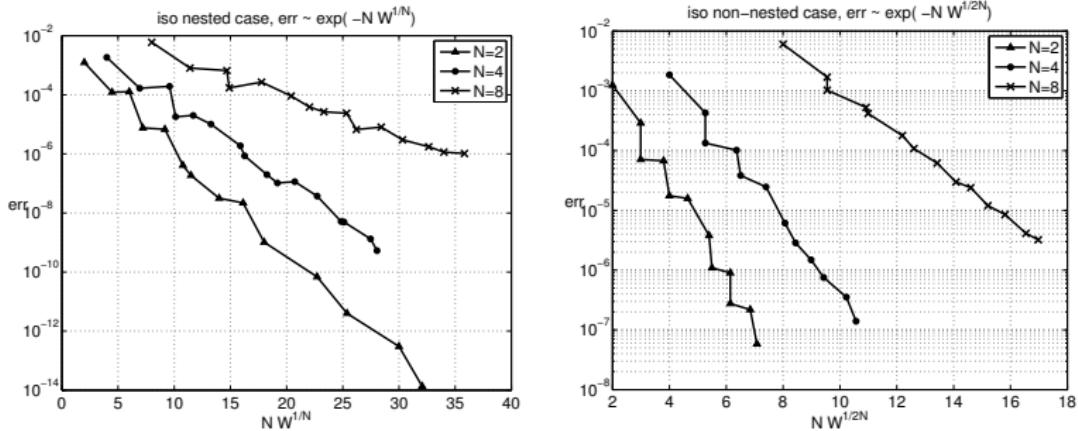
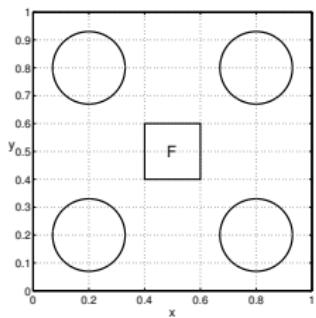
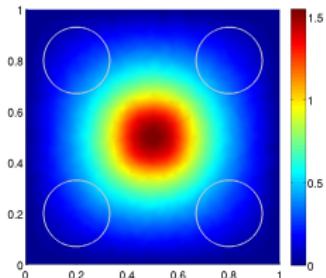


Figure: Results for the isotropic problem. Optimal sparse grids and their predicted convergence rates. **Top:** Nested CC points. **Bottom:** Non-nested Gauss Legendre points.

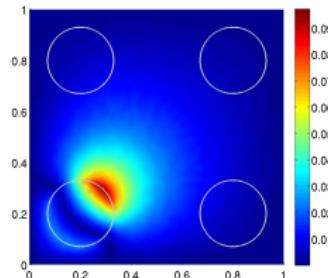
Anisotropic test case – 4 random inclusions



- ▶ Conductivity coefficient: matrix $k=1$
circular inclusions:
 $k|_{\Omega_i} \sim 1 + \gamma_i \mathcal{U}(-0.99, 0.99) \rightarrow 4$ iid
uniform random variables
- ▶ $\gamma_{1,2,3,4} = 1, 0.06, 0.0035, 0.0002$
- ▶ forcing term $f = 100\mathbb{1}_F$, zero b.cs
- ▶ quantity of interest $\psi(u) = \int_F u$



mean



std

convergence plot for $\|\psi(u) - \mathcal{S}_{\mathcal{I}}\psi(u)\|_{L_p^2(\Gamma)}$ versus # pts sparse grid

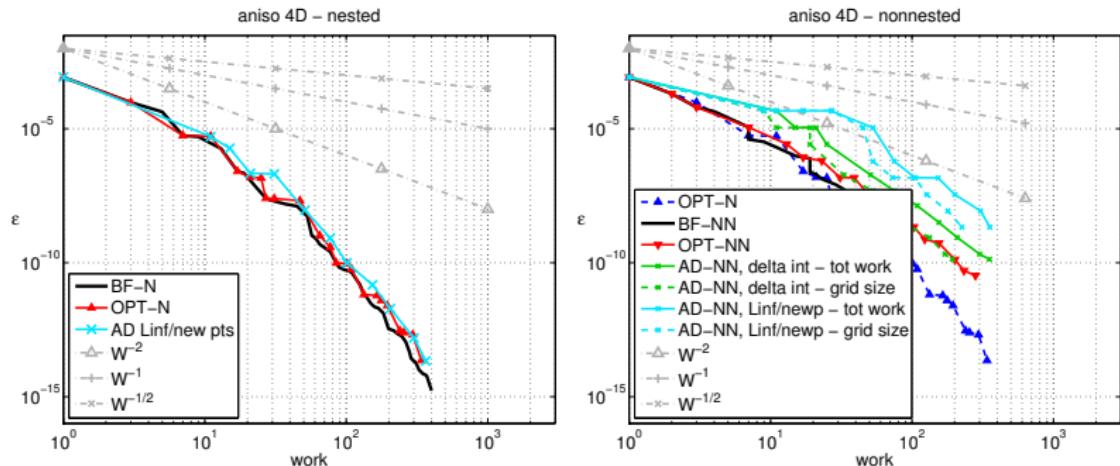
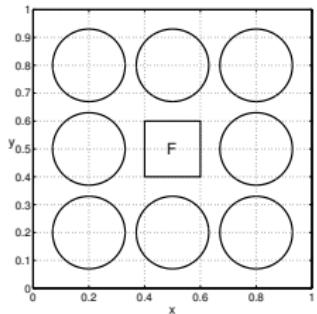
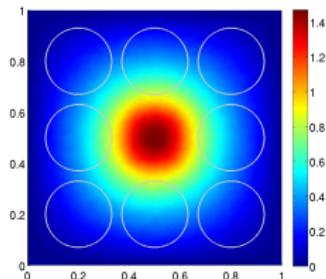


Figure: Results for the anisotropic problem. **Left:** (nested) CC points **Right:** (non-nested) Gauss-Legendre points

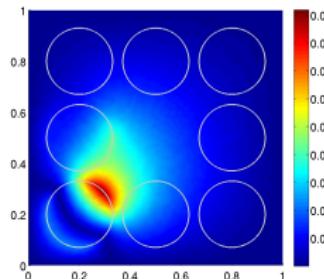
Anisotropic test case – 8 random inclusions



- ▶ Conductivity coefficient: matrix $k=1$ circular inclusions:
 $k|_{\Omega_i} \sim 1 + \gamma_i \mathcal{U}(-0.99, 0.99) \rightarrow 8$ iid uniform random variables
- ▶ $\gamma_1, \dots, 8 = 1, 0.25, 0.06, 0.015, 0.0035, 0.0009, 0.0002, 0.00005$
- ▶ forcing term $f = 100\mathbb{1}_F$, zero b.cs
- ▶ quantity of interest $\psi(u) = \int_F u$



mean



std

Anisotropic test case – 8 random inclusions

convergence plot for $\|\psi(u) - \mathcal{S}_{\mathcal{I}}\psi(u)\|_{L_p^2(\Gamma)}$ versus # pts sparse grid

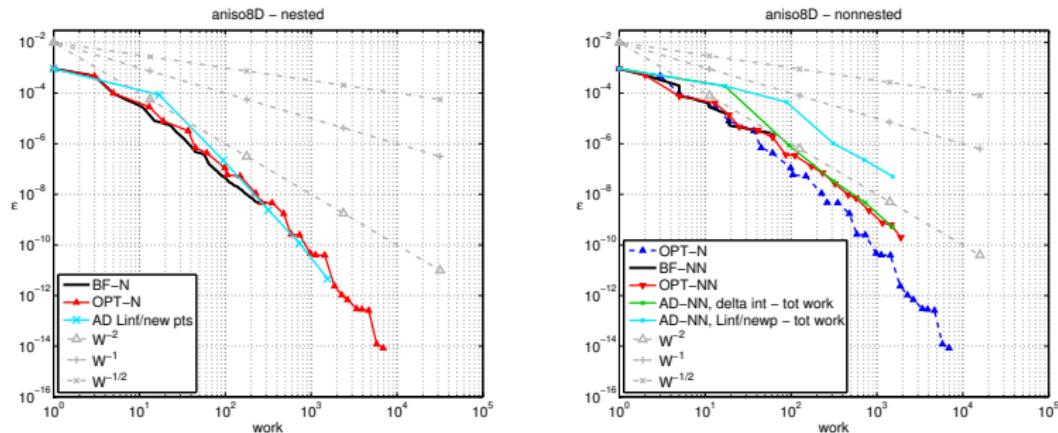


Figure: Results for the anisotropic problem. **Left:** (nested) CC points **Right:** (non-nested) Gauss-Legendre points

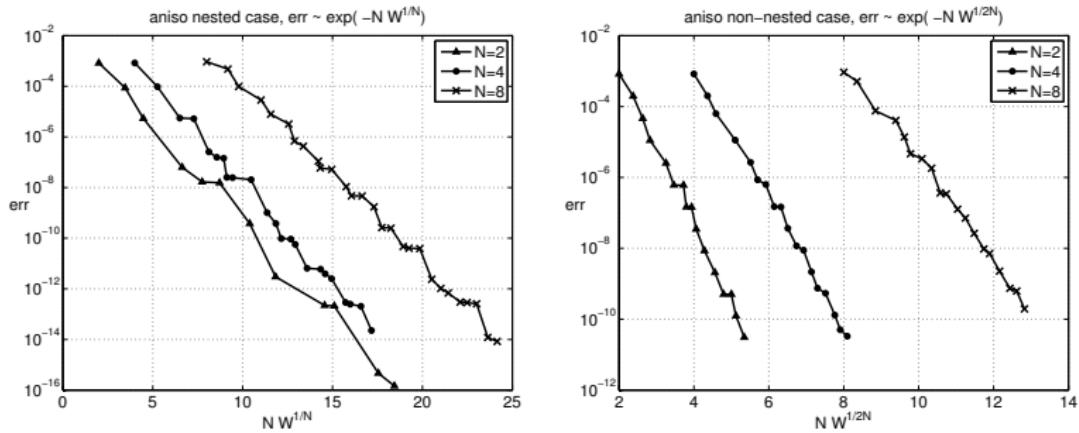


Figure: Results for the anisotropic problem. Optimal sparse grids and their predicted convergence rates. **Top:** Nested CC points. **Bottom:** Non-nested Gauss Legendre points.

Infinite dimensional polynomial approximations are possible

A theoretical result on polynomial approximation [Cohen-DeVore-Schwab '11]

Consider the diffusion equation with coefficient

$$a(\omega, x) = \bar{a}(x) + \sum_{n=1}^{\infty} \sqrt{\lambda_n} y_n(\omega) b_n(x), \text{ with } y_n \sim \mathcal{U}(-\sqrt{3}, \sqrt{3}) \text{ i.i.d}$$

and $\sum_{n=1}^{\infty} \sqrt{3\lambda_n} \|b_n\|_{\infty} < \min_x \bar{a}(x)$, $\forall x \in D$ so that $a(\omega, x) > 0$ almost surely for all $x \in D$.

If $\sum_{n=1}^{\infty} (\sqrt{\lambda_n} \|b_n\|_{L^\infty(D)})^\tau < +\infty$ for some $\tau < 1$, then, there exists a M-terms polynomial approximation u_{Λ_M} such that

$$\|u - u_{\Lambda_M}\|_{L_p^2(\Gamma) \otimes V} \leq C(\tau) M^{\frac{1}{2} - \frac{1}{\tau}}$$

- ▶ This result holds in infinite dimension.
- ▶ Some extensions to collocation methods are available
[Chkifa-Cohen-Schwab '13], [Schillings-Schwab '13]

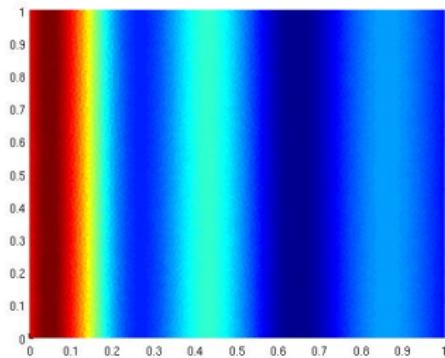
Profit estimate for random fields

When working with random fields instead of inclusion problems a good estimate of the profits is (see [Beck-N.-Tamellini-Tempone M3AS '12])

$$P(\mathbf{i}) \propto \left(\sum_{n=1}^N m(i_n - 1) \right)! \prod_{n=1}^N \frac{\mathbb{L}_{m(i_n)} e^{-g_{i_n} m(i_n - 1)}}{m(i_n - 1)! \Delta W(i_n)}$$

Again, the rates g_i can be estimated numerically.

Numerical test - 1D stationary lognormal field



$$L = 1, D = [0, L]^2.$$

$$\begin{cases} -\nabla \cdot a(\mathbf{y}, \mathbf{x}) \nabla u(\mathbf{y}, \mathbf{x}) = 0 \\ u = 1 \text{ on } x = 0, h = 0 \text{ on } x = 1 \\ \text{no flux otherwise} \end{cases}$$

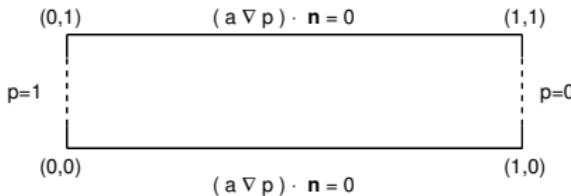
$$a(\mathbf{x}, \mathbf{y}) = e^{\gamma(\mathbf{x}, \mathbf{y})}, \mu_\gamma(\mathbf{x}) = 0,$$
$$\text{Cov}_\gamma(\mathbf{x}, \mathbf{x}') = \sigma^2 e^{-\frac{|\mathbf{x}_1 - \mathbf{x}'_1|^2}{LC^2}}$$

We approximate γ as

$$\gamma(\mathbf{y}, \mathbf{x}) \approx \mu(\mathbf{x}) + \sigma a_0 y_0 + \sigma \sum_{k=1}^K a_k \left[y_{2k-1} \cos\left(\frac{\pi}{L} k \mathbf{x}_1\right) + y_{2k} \sin\left(\frac{\pi}{L} k \mathbf{x}_1\right) \right]$$

with $y_i \sim \mathcal{N}(0, 1)$, i.i.d.

$$\text{Given the Fourier series } \sigma^2 e^{-\frac{|z|^2}{LC^2}} = \sum_{k=0}^{\infty} c_k \cos\left(\frac{\pi}{L} kz\right), a_k = \sqrt{c_k}.$$



- ▶ Quantity of interest: effective permeability $\mathbb{E}[\Phi(u)]$, with

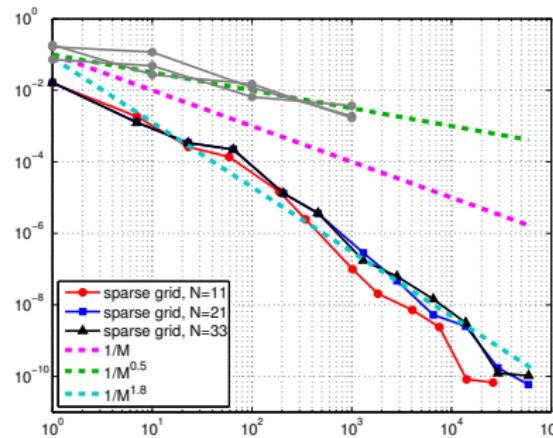
$$\Phi(u) := \int_{\mathcal{D}} k(\cdot, \mathbf{x}) \partial_{\mathbf{n}} u(\cdot, \mathbf{x}) d\mathbf{x}, \quad \mathcal{D} = \{\mathbf{x} = (x_1, x_2) \in \overline{D} : x_1 = L\}$$

- ▶ Convergence: $|\mathbb{E}[\Phi(S_{\tilde{\mathcal{I}}_M} u)] - \mathbb{E}[\Phi(u)]|$
- ▶ We compare Monte Carlo estimate with quasi-optimal sparse grids based on Gauss-Hermite-Patterson points (nested Gauss-Hermite)
- ▶ Estimate of the profits

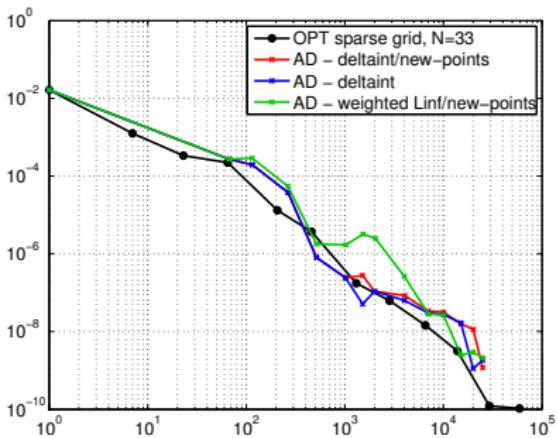
$$P(\mathbf{i}) \sim \prod_{n=1}^N \mathbb{L}_{m(i_n)} \frac{e^{-g_n m(i_n - 1)}}{\Delta W(i_n) \sqrt{m(i_n - 1)!}}$$

with rates g_n numerically estimated.

Correlation length: $LC = 0.2$, Std: $\sigma = 0.3$ (c.o.v. $\sim 30\%$)



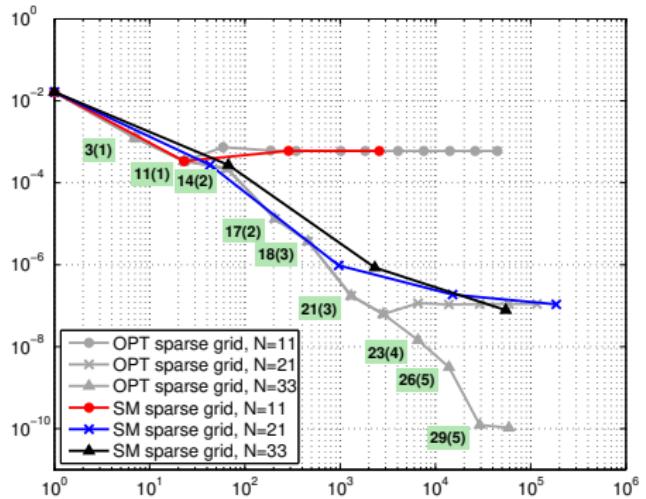
(a) Convergence of quasi-optimal a-priori sparse grid approx.



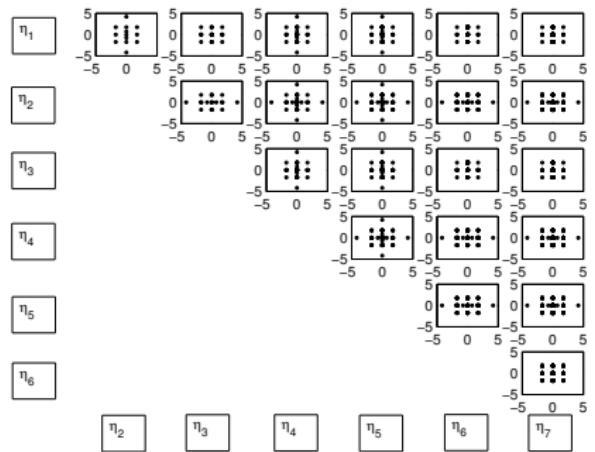
(b) Convergence of adaptive algorithm ($N = 33$ r.vs.)

- ▶ The quasi optimal / adaptive constructions automatically add new variables when needed.
- ▶ No need to truncate a-priori the random field
- ▶ Reference:

"A quasi-optimal sparse grids procedure for groundwater flows" by J. Beck, F. Nobile, L. Tamellini and R. Tempone. LNCSE, Springer, 2014.



Quasi optimal sparse grid, green boxes: number of active variables (max. number of active variables at the same time)



Several 2D projections of the quasi optimal sparse grid

Sparse Grids Matlab Kit

- ▶ <https://sites.google.com/view/sparse-grids-kit/home>
- ▶ <https://arxiv.org/abs/2203.09314>

Hierarchical approximations

Consider a quantity of interest $S = Q(u(\mathbf{y}))$. So far we have separately discussed the following hierarchical approximations:

- ▶ **Stochastic approximation** by tensor grid interpolation-quadrature:

For an interpolation level $\beta = (\beta_1, \dots, \beta_N)$, denote

$S_\beta = \mathbb{E}[\mathcal{U}^{m(\beta)}[S]]$ and **Hierarchical surplus**

$$\Delta[S_\beta] = \bigotimes_{n=1}^N \Delta_j S_\beta, \quad \Delta_j S_\beta = \begin{cases} S_\beta - S_{\beta - e_j}, & \text{if } \beta_j > 1, \\ S_\beta & \text{if } \beta_j = 1. \end{cases}$$

- ▶ **Deterministic discretization**: for an approximation level $\alpha = (\alpha_1, \dots, \alpha_d)$, corresponding to discretization parameters h_{i,α_i} (e.g. $h_{i,\alpha_i} = h_{i,0} 2^{-\alpha_i}$), denote $S^\alpha(\mathbf{y}) = Q(u_\alpha(\mathbf{y}))$ the approximate QoI and **Hierarchical surplus**

$$\Delta[S^\alpha] = \bigotimes_{i=1}^d \Delta_i S^\alpha, \quad \Delta_i S^\alpha = \begin{cases} S^\alpha - S^{\alpha - e_i}, & \text{if } \alpha_i > 1, \\ S^\alpha & \text{if } \alpha_i = 1, \end{cases}$$

Multi Index Stochastic Collocation (MISC)

Idea: build a generalized sparse construction in both approximation multi-indices

- ▶ β (stochastic directions) and
- ▶ α (space directions).

⇒ Multi Index Stochastic Collocation (MISC) method

Our MISC discussion is primarily based on:

[MISC1] A.-L. Haji-Ali, F. Nobile, L. Tamellini and R. Tempone. “*Multi-Index Stochastic Collocation for random PDEs*”, Computer Methods in Applied Mechanics and Engineering, Vol. 306, pp. 95–122, 2016.

[MISC2] A.-L. Haji-Ali, F. Nobile, L. Tamellini and R. Tempone. “*Multi-Index Stochastic Collocation convergence rates for random PDEs with parametric regularity*”, Foundations of Computational Mathematics, Vol. 16, pp. 1555–1605, 2016.

MISC difference operators

Denote by $S_{\alpha,\beta}$ a hierarchical approximation based on the multi-indices $\alpha \in \mathbb{N}^N$ and $\beta \in \mathbb{N}^d$, defined as

$$S_{\alpha,\beta} := E \left[\mathcal{U}^{m(\beta)} [S^\alpha] \right] = E \left[\mathcal{U}^{m(\beta)} [Q(u_\alpha)] \right].$$

We (separately) define the **Delta operators** along the *stochastic* and *deterministic* dimensions

$$\Delta_i^d S_{\alpha,\beta} = \begin{cases} S_{\alpha,\beta} - S_{\alpha-\mathbf{e}_i,\beta}, & \text{if } \alpha_i > 1, \\ S_{\alpha,\beta} & \text{if } \alpha_i = 1, \end{cases}$$
$$\Delta_j^s S_{\alpha,\beta} = \begin{cases} S_{\alpha,\beta} - S_{\alpha,\beta-\mathbf{e}_j}, & \text{if } \beta_j > 1, \\ S_{\alpha,\beta} & \text{if } \beta_j = 1. \end{cases}$$

Moreover, we introduce the **mixed difference operators**:

deterministic difference :

$$\Delta^d [S_{\alpha,\beta}] = \bigotimes_{i=1}^d \Delta_i^d S_{\alpha,\beta}$$

stochastic difference :

$$\Delta^s [S_{\alpha,\beta}] = \bigotimes_{j=1}^N \Delta_j^s S_{\alpha,\beta}$$

MISC estimator

Definition 17.1

The Multi Index Stochastic Collocation (MISC) estimator for $E[S]$ is defined as

$$\mathcal{A}_{\text{MISC}}(\mathcal{I}) = \sum_{(\alpha, \beta) \in \mathcal{I}} \Delta^s \left(\Delta^d [S_{\alpha, \beta}] \right)$$

for some index set $\mathcal{I} \in \mathbb{N}^{d+N}$.

Relevant questions:

- ▶ What is the optimal choice for \mathcal{I} ?
 - ~~ Can be found computationally using knapsack optimization theory.
- ▶ Can we say something about the resulting rate of work complexity using the optimal \mathcal{I} ?

MISC Analysis: Quasi-optimal index sets

The “best” index sets are given by a knapsack problem.¹⁵

Error representation and approximation

$$|E[S] - \mathcal{A}_{\text{MISC}}(\mathcal{I})| = \left| \sum_{(\alpha, \beta) \notin \mathcal{I}} \Delta^s (\Delta^d [S_{\alpha, \beta}]) \right| \leq \sum_{(\alpha, \beta) \notin \mathcal{I}} |\Delta^s \Delta^d [S_{\alpha, \beta}]|$$

Define

- ▶ Error contribution: $\Delta E_{\alpha, \beta} \approx |\Delta^s (\Delta^d [S_{\alpha, \beta}])|$
- ▶ Work contribution: $\Delta W_{\alpha, \beta}$

“Binary” **Knapsack problem** to determine tuples (α, β) that contribute the most (w.r.t. benefit-to-cost ratio) to the overall error:

$$\max_{x_{\alpha, \beta}} \sum_{(\alpha, \beta) \in \mathbb{N}^{d+N}} x_{\alpha, \beta} \Delta E_{\alpha, \beta} \quad \text{s.t.} \quad \sum_{(\alpha, \beta) \in \mathbb{N}^{d+N}} x_{\alpha, \beta} \Delta W_{\alpha, \beta} \leq W_{\max}, \quad x_{\alpha, \beta} \in \{0, 1\}$$

¹⁵[Griebel-Knapek '09, Gerstner-Griebel '03, Bungartz-Griebel '04];
[Nobile-Tamellini-Tempone '16] considers weighted summability of profits for conv. rates.

The Knapsack problem is, in general, computationally intractable, but its relaxation $x_{\alpha,\beta} \in [0, 1]$ has a simple solution (*Dantzig algorithm*):

- ▶ Define Profit: $P_{\alpha,\beta} = \frac{\Delta E_{\alpha,\beta}}{\Delta W_{\alpha,\beta}}$
- ▶ Sort $P_{\alpha,\beta}$ by decreasing profit
- ▶ Add (α, β) to \mathcal{I} according to such order until W_{max} is reached

Quasi-optimal index sets:

$$\mathcal{I} = \mathcal{I}(\epsilon) = \{(\alpha, \beta) \in \mathbb{N}^{d+N} : P_{\alpha,\beta} \geq \epsilon\}$$

for a given profit threshold $\epsilon > 0$.

Remark 17.1

This construction is valid also for $N = \infty$ random parameters (distributed fields) as long as $P_{\alpha,\beta} \geq \epsilon$ implies that $\exists j_\epsilon \in \mathbb{N}$ such that $\beta_k = 1, \forall k > j_\epsilon$. Here it is crucial that level to nodes function fulfills $m(1) = 1$ so that $\Delta W_{\alpha,\beta} \propto \prod_{n=1}^{\infty} m(\beta_n) < \infty$.

Complexity analysis – N finite

Refinement schemes:

- ▶ Stochastic Collocation: $m(\beta_j) \sim 2^{\beta_j - 1}$.
- ▶ deterministic directions: $h_{i,\alpha_i} = h_{i,0} 2^{-\alpha_i}$

Assumption 17.1 (Product rates)

There exist strictly positive constants Q_W , C_{work} , and rates w_i , γ_i for $i = 1 \dots d$ and g_j for $j = 1 \dots N$, such that

$$\Delta E_{\alpha,\beta} \leq Q_W \left(\prod_{j=1}^N e^{-g_j m(\beta_j)} \right) \left(\prod_{i=1}^d (h_{i,\alpha_i})^{\tilde{w}_i} \right)$$

$$\Delta W_{\alpha,\beta} \leq C_{\text{work}} \left(\prod_{j=1}^N 2^{\beta_j} \right) \left(\prod_{i=1}^d (h_{i,\alpha_i})^{-\tilde{\gamma}_i} \right).$$

In other words:

- ▶ **Exponential convergence** in the stochastic variables
- ▶ **Algebraic convergence** in the deterministic variables

Implication of previous assumptions

The assumption on product rates

$$\Delta E_{\alpha, \beta} \leq Q_W \prod_{j=1}^N e^{-g_j m(\beta_j)} \prod_{i=1}^d e^{-w_i \alpha_i}$$

implies:

- ▶ some mixed Sobolev regularity in the deterministic variables,
- ▶ “mixed analytic” regularity (analyticity in poly-ellipses) in the stochastic variables,
- ▶ mixed deterministic / stochastic regularity.

Example 17.2

These assumptions are fulfilled, e.g., for the elliptic problem (on regular domain)

$$-\nabla \cdot (a(x, y) \nabla u(x, y)) = f(x)$$

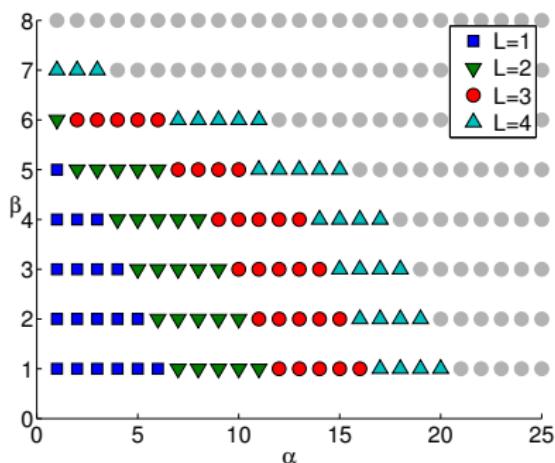
with finite number of random variables $a(x, y) = \exp \left(\sum_{j=1}^N \psi_j(x) y_j \right)$
with sufficiently regular ψ_j , and equipped with suitable boundary conditions.

Corresponding optimal index sets

Under the previous assumptions, the optimal index-sets have the form:

$$\mathcal{I}(L) = \{(\alpha, \beta) \in \mathbb{N}^{d+N} : \sum_{i=1}^d (w_i + \gamma_i)\alpha_i + \sum_{j=1}^N (\log(2)\beta_j + g_j 2^{\beta_j}) \leq L\},$$

with $L = -\log(\epsilon)$.



For example, index sets for diffusion problem

$$-\nabla \cdot (a(\mathbf{y}) \nabla u) = f$$

in 1D ($d=1$) and with 1 random variable ($N=1$)

MISC error estimates

Theorem 17.3 (Error estimate with optimal sets, [MISC1])

Under previous particular assumptions on the error and work contributions, for any $W \equiv W_{\max}$ sufficiently large, there exists an index-set $\mathcal{I}(W) \equiv \mathcal{I}(L(W))$ such that

$$\text{Work}[\mathcal{A}_{\text{MISC}}(\mathcal{I}(W))] \leq W,$$

$$\text{Error}[\mathcal{A}_{\text{MISC}}(\mathcal{I}(W))] \leq \mathcal{C}_E(N) W^{-\zeta} (\log(W))^{(\zeta+1)(\mathfrak{z}-1)}.$$

where $\zeta = \min_{i=1,\dots,d} \frac{w_i}{\gamma_i}$ and $\mathfrak{z} = \#\{i = 1, \dots, d : \frac{w_i}{\gamma_i} = \zeta\}$.

Remark 17.2

- ▶ Up to logarithmic terms, asymptotically $\text{Error} \sim W^{-\zeta}$: complexity of a 1D deterministic problem (the worst one). The complexity of the Stochastic Collocation is not seen.
- ▶ The asymptotic rate is independent of the number N of random variables (but the constant \mathcal{C}_E is not!)

Numerical example

Problem: $-\nabla \cdot (a(\mathbf{y}) \nabla u) = 1$, in $D = [0, 1]^d$, $u = 0$, on ∂D

$$a(\mathbf{y}) = e^{\sum_{n=1}^N \sqrt{\lambda_n} \psi_n y_n}, \quad \psi_n: \text{Fourier modes}, \quad y_n \stackrel{\text{iid}}{\sim} \mathcal{U}([-1, 1]), \quad \lambda_n \sim e^{-n}$$

Clenshaw-Curtis collocation points; finite difference discretization in space.

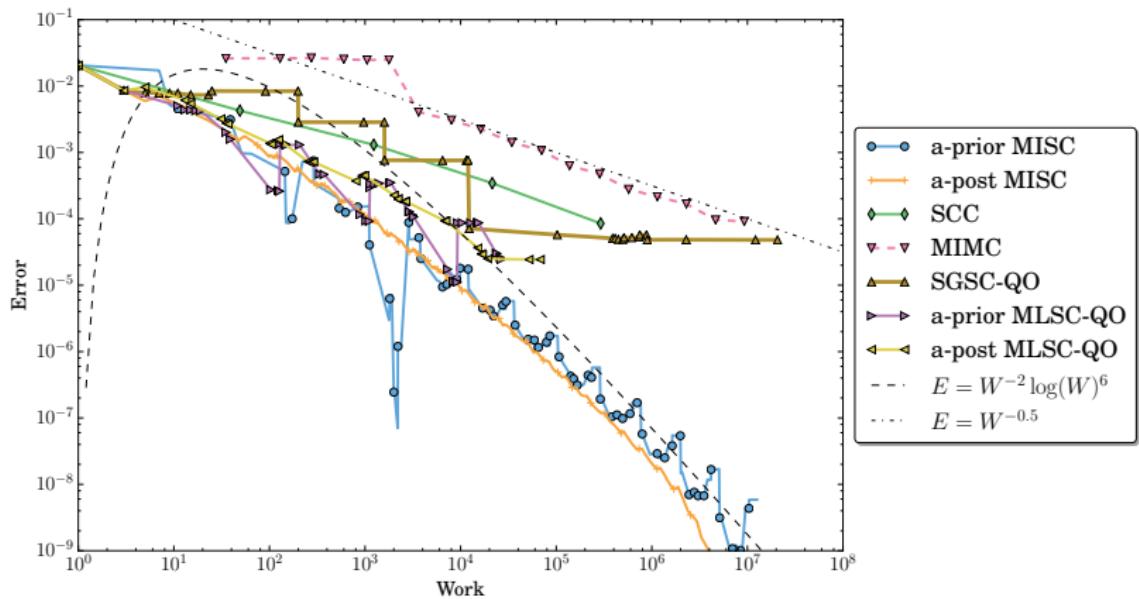
Alternative methods:

- ▶ **MISC** with **a-priori** and **a-posteriori** construction of optimal set
- ▶ **MLSC** (uniform refinement in space; only one discretization param.)
- ▶ **SGSC**: Single level SC (with optimal balance of spatial and stochastic errors)
- ▶ Stochastic Composite Collocation Method (**SCC**) [Bieri 2011]. Corresponds to the choice

$$\mathcal{I} = \{(\alpha, \beta) : \alpha + \sum_{n=1}^N \beta_n \leq L\}$$

- ▶ Multi-Index Monte Carlo (**MIMC**)

Test case in dimension $d = 3$ with $N = 10$ random variables



Infinite dimensional case ($N = \infty$) – analytic model problem

We apply MISC to the following problem on a hypercube domain $D \subset \mathbb{R}^d$

$$\begin{aligned}-\nabla \cdot (\textcolor{red}{a}(\mathbf{x}, \mathbf{y}) \nabla u(\mathbf{x}, \mathbf{y})) &= f(\mathbf{x}) \quad \text{in } D \\ u(\mathbf{x}, \mathbf{y}) &= 0 \quad \text{on } \partial D,\end{aligned}$$

where

$$\textcolor{red}{a}(\mathbf{x}, \mathbf{y}) = e^{\kappa(\mathbf{x}, \mathbf{y})}, \text{ with } \kappa(\mathbf{x}, \mathbf{y}) = \sum_{j \in \mathbb{N}_+} \psi_j(\mathbf{x}) y_j.$$

Here, $y_j \stackrel{\text{iid}}{\sim} \mathcal{U}([-1, 1])$.

QoI: $S = Q(u)$

Goal: compute $\mathbb{E}[S] = \mathbb{E}[Q(u)]$

- ▶ Stochastic collocation on Clenshaw-Curtis points
- ▶ (sparse) piecewise linear finite elements in space.

Polynomial approximation (and Stochastic Collocation) in the random variables only achieves algebraic convergence.

Consider the Sparse Grid Stochastic Collocation estimator $\mathcal{S}_{\mathcal{I}}[u]$ for $u(\mathbf{y})$ using Clenshaw-Curtis points (no spatial discretization).

Theorem 17.4 ([MISC2])

There exists a sequence of (quasi)-optimal sparse grids with cardinality W_ℓ and index sets \mathcal{I}_ℓ such that

$$\|u - \mathcal{S}_{\mathcal{I}_\ell}[u]\|_{L^2(\Omega; H^1(D))} \leq CW_\ell^{-k_0}, \quad k_0 = \frac{1}{p_0} - 2$$

where p_0 is the summability exponent of the sequence $\{\|\psi_j\|_{C^0(D)}\}_j$, i.e. $p_0 \in (0, \frac{1}{2}]$ s.t. $\sum_{j=1}^{\infty} \|\psi_j\|_{C^0(D)}^{p_0} < \infty$

Remark 17.3

- ▶ The rate k_0 can be improved to $k_0 = \frac{1}{p_0} - 1$ using the arguments in [Cohen-Devore-Schwab 2011].
- ▶ Better summability exponent p_0 can be obtained by looking at pointwise summability, $p_0 : \sum_{j=1}^{\infty} |\psi_j(x)|^{p_0} < \infty, \forall x \in \bar{D}$; see [Bachmayr-Cohen-Devore-Migliorati '15].

However, when applying MISC, for the mixed (spatial/parameter) differences to have product structure

$$\Delta E_{\alpha, \beta} \leq Q_W \left(\prod_{j=1}^{\infty} e^{-g_j m(\beta_j)} \right) \left(\prod_{i=1}^d e^{-w_i \alpha_i} \right)$$

we need to estimate the convergence of the sparse grid in a stronger norm H^{1+s} , $1 < s \leq d$:

$$\|u - S_{\mathcal{I}_\ell}[u]\|_{L^2(\Omega; H^{1+s}(D))} \leq CW_\ell^{-k_s}, \quad k_s = \frac{1}{p_s} - 2$$

where p_s is the summability exponent of the sequence $\{\|\psi_j\|_{C^s(D)}\}_j$.

- ▶ The maximum regularity achievable depends on the smoothness of the covariance function,
- ▶ the larger s , the larger p_s (smaller summability).

Assumption 17.2

- ▶ There exist $s_{max} \in \mathbb{N}_+$ and an increasing sequence $p_0 < p_1 < \dots < p_{s_{max}}$ with $p_{s_{max}} < \frac{1}{2}$ such that

$$\sum_{j \geq 1} \|\psi_j\|_{C^s(D)}^{p_s} \leq \kappa_s^{p_s} < \infty, \quad s = 0, 1, \dots, s_{max}.$$

with κ_s sufficiently small.

- ▶ D and f are such that $u(\mathbf{y}) \in H^{1+s_{max}}(D)$ for any $\mathbf{y} \in [-1, 1]^{\mathbb{N}}$.

Under these assumptions there exist $\{g_j(s)\}_{j \in \mathbb{N}}$ with $\sum_{j \geq 1} e^{-p_s g_j(s)} < \infty$ for $s = 0, \dots, s_{max}$ such that

$$\Delta E_{\alpha, \beta} \leq Q_W \min_{s=0, \dots, s_{max}} \left(\prod_{j=1}^{\infty} e^{-g_j(s) m(\beta_j)} \right) \left(\prod_{i=1}^d e^{-\min\{1, \frac{s}{d}\} \alpha_i} \right).$$

Theorem 17.5 (MISC convergence, [MISC2])

Under technical assumptions the profit-based MISC estimator built using Stochastic Collocation over Clenshaw-Curtis points and piecewise multilinear finite elements for solving the deterministic problems, we have, for any $\delta > 0$,

$$\text{Error}[\mathcal{A}_{\text{MISC}}(\mathcal{I})] \leq \tilde{C}_P(\delta) \text{Work}[\mathcal{A}_{\text{MISC}}(\mathcal{I})]^{-r_{\text{MISC}}+\delta}.$$

The rate r_{MISC} is as follows:

Case 1 if $\min\left\{\frac{1}{\gamma}, \frac{s_{\max}}{\gamma d}\right\} \leq \frac{1}{p_{s_{\max}}} - 2$, then $r_{\text{MISC}} = \min\left\{\frac{1}{\gamma}, \frac{s_{\max}}{\gamma d}\right\}$,

Case 2 if $\min\left\{\frac{1}{\gamma}, \frac{s_{\max}}{\gamma d}\right\} \geq \frac{1}{p_{s_{\max}}} - 2$, then

$$r_{\text{MISC}} = \left(\frac{1}{p_0} - 2 \right) \max_{s=1, \dots, s_{\max}} \left(\frac{1}{\min\left\{\frac{1}{\gamma}, \frac{s}{\gamma d}\right\}} \left(\frac{1}{p_0} - \frac{1}{p_s} \right) + 1 \right)^{-1}.$$

Ideas for proofs in [MISC2]

- ▶ Given the sequences

$$b_{0,j} = \|\psi_j\|_{L^\infty(D)}, \quad j \geq 1, \tag{94}$$

$$b_{s,j} = \max_{s \in \mathbb{N}^d : |s| \leq s} \|D^s \psi_j\|_{L^\infty(D)}, \quad j \geq 1, \tag{95}$$

we assume that there exist $0 < p_0 \leq p_s < \frac{1}{2}$ such that
 $\{b_{0,j}\}_{j \in \mathbb{N}_+} \in \ell^{p_0}$ and $\{b_{s,j}\}_{j \in \mathbb{N}_+} \in \ell^{p_s}$,

- ▶ Shift theorem: From regularity of a and f to regularity of $u \in H^{1+s}(D) \Rightarrow u \in \mathcal{H}_{mix}^{1+q}(D)$, for $0 < q < s/d$.
- ▶ Extend holomorphically $u(\cdot, z) \in H^{1+r}(D)$ on polyellipse $z \in \Sigma_r$ (use p_r summability of b_r) to get stochastic rates and estimates for Δ .
- ▶ Use weighted summability of knapsack profits to prove convergence rates.

Example: log uniform field with parametric regularity

Here, the regularity of $\kappa = \log(a)$ is determined through $\nu > 0$

$$\kappa(\mathbf{x}, \mathbf{y}) = \sum_{\mathbf{k} \in \mathbb{N}^d} A_{\mathbf{k}} \sum_{\ell \in \{0,1\}^d} y_{\mathbf{k}, \ell} \prod_{j=1}^d \left(\cos \left(\frac{\pi}{L} k_j x_j \right) \right)^{\ell_j} \left(\sin \left(\frac{\pi}{L} k_j x_j \right) \right)^{1-\ell_j},$$

where the coefficients $A_{\mathbf{k}}$ are taken as

$$A_{\mathbf{k}} = \left(\sqrt{3} \right) 2^{\frac{|\mathbf{k}|_0}{2}} (1 + |\mathbf{k}|^2)^{-\frac{\nu+d/2}{2}},$$

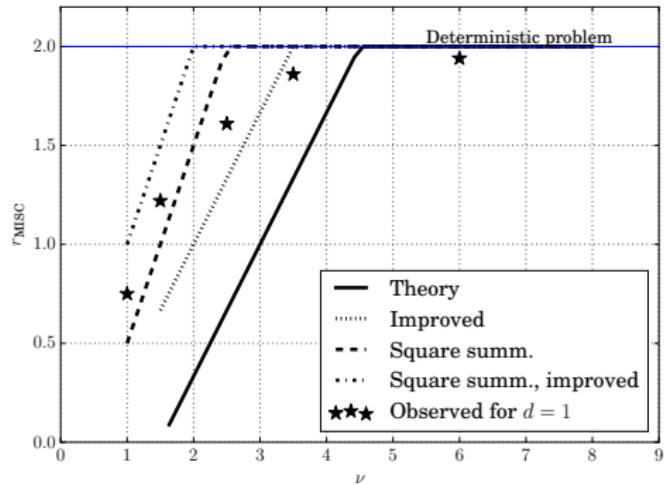
same decay as Matérn covariance of parameter ν . We have

$$p_0 > \left(\frac{\nu}{d} + \frac{1}{2} \right)^{-1} \text{ and } p_s > \left(\frac{\nu-s}{d} + \frac{1}{2} \right)^{-1}.$$

Applying the theorem with optimal s we obtain

$$r_{MISC} = \min \left\{ \frac{1}{\gamma}, \left(\frac{\nu}{d} - \frac{3}{2} \right) \frac{1}{1+\gamma} \right\}.$$

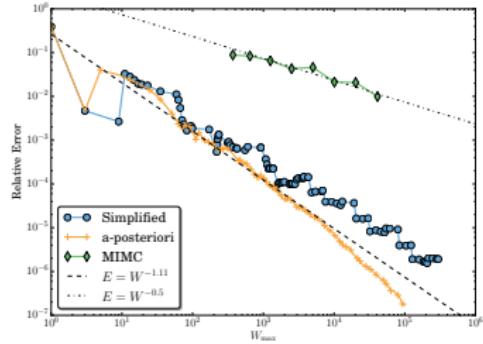
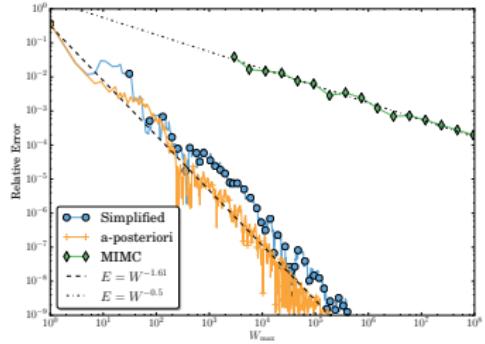
Application of the theorem



$$\text{Error} \propto \text{Work}^{-r_{MIMC}(\nu, d)}$$

A similar analysis shows the corresponding ν -dependent convergence rates of MIMC but based on ℓ^2 summability of b_s and Fernique type of results.

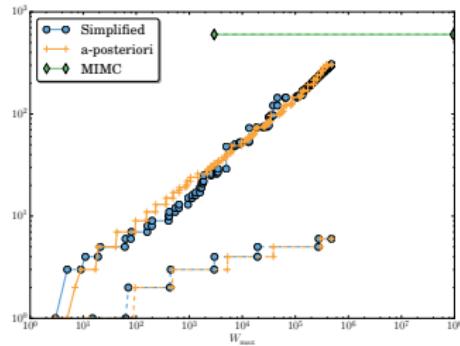
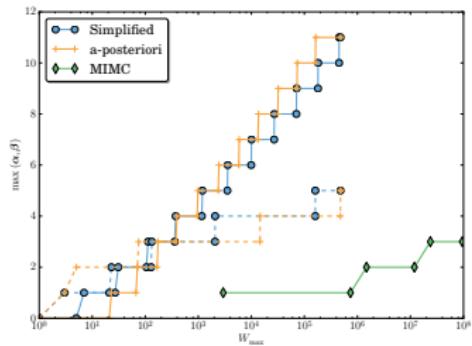
Numerical results



Left: $d = 1, \nu = 2.5$. Right: $d = 3, \nu = 4.5$.

$$\text{Error} \propto \text{Work}^{-r_{MISC}(\nu, d)}$$

Numerical results – 1D



$d = 1$ and $\nu = 2.5$. Extreme values of α and β included in the MISC set \mathcal{I} .

Specifically, *left-solid* is the maximum space discretization level,

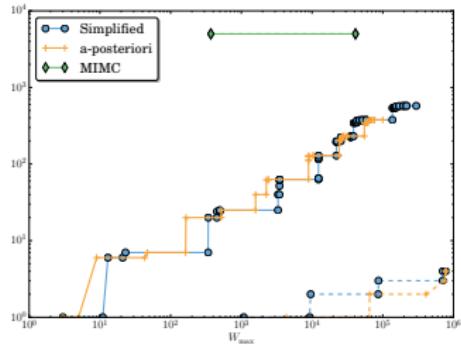
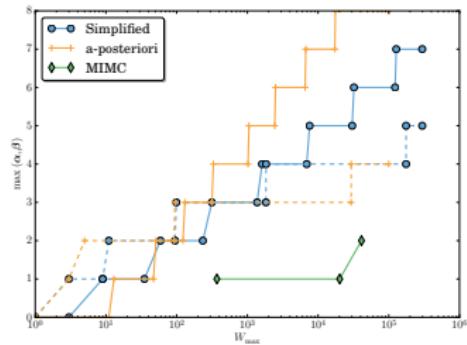
$\max_{(\alpha, \beta) \in \mathcal{I}} (\max(\alpha))$, *left-dashed* is the maximum quadrature level,

$\max_{(\alpha, \beta) \in \mathcal{I}} (\max(\beta))$, *right-solid* is the index of the last activated random variable,

$\max_{(\alpha, \beta) \in \mathcal{I}} (\max_{\beta_j > 1} j)$, and *right-dashed* is the maximum number of jointly

activated variables, $\max_{(\alpha, \beta) \in \mathcal{I}} (|\beta - 1|_0)$.

Numerical results – 3D



$d = 3$ and $\nu = 4.5$. Extreme values of α and β included in the MISC set \mathcal{I} .

Specifically, *left-solid* is the maximum space discretization level,

$\max_{(\alpha, \beta) \in \mathcal{I}} (\max(\alpha))$, *left-dashed* is the maximum quadrature level,

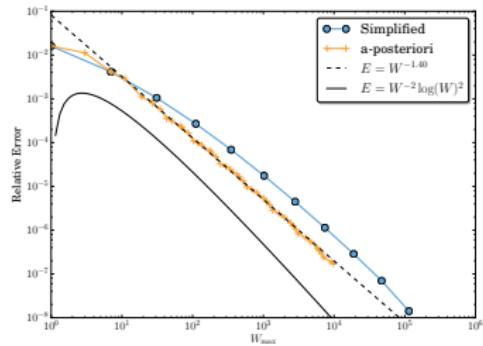
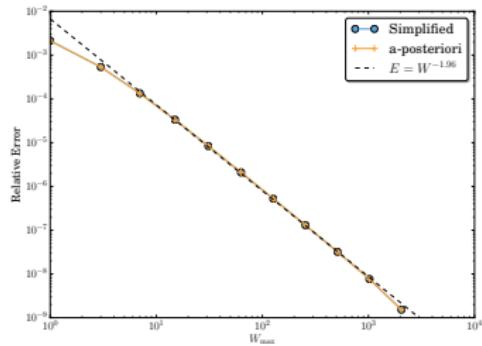
$\max_{(\alpha, \beta) \in \mathcal{I}} (\max(\beta))$, *right-solid* is the index of the last activated random variable,

$\max_{(\alpha, \beta) \in \mathcal{I}} (\max_{\beta_j > 1} j)$, and *right-dashed* is the maximum number of **jointly**

activated variables, $\max_{(\alpha, \beta) \in \mathcal{I}} (|\beta - 1|_0)$.

Deterministic runs, numerical results

These plots shows the non-asymptotic effect of the logarithmic factor for $d > 1$ (as discussed in [Thm. MISC1]) on the linear convergence fit in log-log scale.



Left: $d = 1$. Right: $d = 3$.

Stochastic Galerkin

Stochastic Galerkin: Motivation

We consider the model problem

$$\begin{cases} -\operatorname{div}(a(\mathbf{y}, x)\nabla u(\mathbf{y}, x)) = f(\mathbf{y}, x), & x \in D, \\ u(\mathbf{y}, x) = 0, & x \in \partial D \end{cases}, \quad \forall \mathbf{y} \in \Gamma := \prod_{n=1}^N \Gamma_n$$

with

- ▶ $\mathbf{y} = (y_1, \dots, y_N)$ random vector with independent components and density $\rho(\mathbf{y}) = \prod_{n=1}^N \rho_n(y_n)$,
- ▶ $0 < a_{\min} \leq a(\mathbf{y}, x) \leq a_{\max}$ for a.e. $x \in D$ and ρ -a.e. $\mathbf{y} \in \Gamma$.
- ▶ $f \in L^2_\rho(\Gamma, L^2(D))$.

Then, the solution is $u \in L^2_\rho(\Gamma, H_0^1(D)) \sim L^2_\rho(\Gamma) \otimes H_0^1(D)$.

Global weak (mean) formulation: Find $u \in L^2_\rho(\Gamma) \otimes H_0^1(D)$ s.t.

$$\mathbb{E}\left[\int_D a(\mathbf{y}, x)\nabla u(\mathbf{y}, x) \cdot \nabla v(\mathbf{y}, x) dx\right] = \mathbb{E}\left[\int_D f(\mathbf{y}, x)v(\mathbf{y}, x)dx\right]$$

for all $v \in L^2_\rho(\Gamma) \otimes H_0^1(D)$.

Let $\mathcal{V} := L^2_\rho(\Gamma) \otimes H_0^1(D)$ and define $A: \mathcal{V} \times \mathcal{V} \rightarrow \mathbb{R}$ and $b: \mathcal{V} \rightarrow \mathbb{R}$ by

$$A(u, v) := \mathbb{E} \left[\int_D a \nabla_x u \cdot \nabla_x v \, dx \right] \quad \text{and} \quad b(v) := \mathbb{E} \left[\int_D f v \, dx \right]$$

resp. Then the global weak formulation can be stated as: find $u \in \mathcal{V}$ such that

$$A(u, v) = b(v) \quad \forall v \in \mathcal{V}.$$

That is, well-posedness can be studied with classic Lax–Milgram Thm. 4.4.

Exercise 18.1

Does the variational problem have a unique solution? Explain your answer.

To discretize the problem, we use the [Galerkin method](#). That is, we need to identify a finite dimensional subspace of \mathcal{V} .

Approach: consider *separately* finite dimensional subspaces in $H_0^1(D)$ (i.e., discretization in physical space) and in $L^2_\rho(\Gamma)$ (i.e., discretization of stochastic space), respectively, to obtain finite dimensional subspace of $\mathcal{V} = L^2_\rho(\Gamma) \otimes H_0^1(D)$.

Finite element approximation in physical space

- ▶ Let \mathcal{T}_h be a triangulation of D and $H_h(D) \subset H_0^1(D)$ a suitable Finite Element space.
- ▶ Let $\{\phi_i\}_{i=1}^{N_h}$ be a basis for $H_h(D)$.

We seek a solution of the form $u(\mathbf{y}, x) = \sum_{i=1}^{N_h} u_i(\mathbf{y}) \phi_i(x)$

Observation: each DoF (nodal value) $u_i = u_i(\mathbf{y})$ is a random variable!

Semi-discrete Finite Element approximation: Find

$u_h \in L_\rho^2(\Gamma) \otimes H_h(D)$ s.t.

$$\mathbb{E}\left[\int_D a(\mathbf{y}, x) \nabla u_h(\mathbf{y}, x) \cdot \nabla v_h(\mathbf{y}, x) dx\right] = \mathbb{E}\left[\int_D f(\mathbf{y}, x) v_h(\mathbf{y}, x) dx\right]$$

for all $v_h \in L_\rho^2(\Gamma) \otimes H_h(D)$.

Stochastic domain: polynomial approximation

- ▶ Let $\Lambda(w) \subset \mathbb{N}^N$ be an index set of cardinality M_w and consider the multivariate polynomial space

$$\mathbb{P}_{\Lambda(w)}(\Gamma) = \text{span} \left\{ \prod_{n=1}^N y_n^{\mathbf{p}_n}, \quad \text{with } \mathbf{p} \in \Lambda(w) \right\}$$

- ▶ Let $\{\psi_i\}_{i=1}^{M_w}$ be a basis of $\mathbb{P}_{\Lambda(w)}(\Gamma)$ (e.g., multivariate Legendre, Hermite, ..., orthogonal polynomials)

We seek for a **fully discrete** solution $u_{h,w} \in \mathbb{P}_{\Lambda(w)}(\Gamma) \otimes H_h(D) \subset \mathcal{V}$:

$$\begin{aligned} u_{h,w}(\mathbf{y}, \mathbf{x}) &= \sum_{j=1}^{N_h} \sum_{i=1}^{M_w} u_{ij} \psi_i(\mathbf{y}) \phi_j(\mathbf{x}) \\ &= \sum_{j=1}^{N_h} u_p^{(j)}(\mathbf{y}) \phi_j(\mathbf{x}), \quad \text{with } u_p^{(j)} \in \mathbb{P}_{\Lambda(w)}(\Gamma^N) \\ &= \sum_{i=1}^{M_w} u_h^{(i)}(\mathbf{x}) \psi_i(\mathbf{y}), \quad \text{with } u_h^{(i)} \in H_h(D) \end{aligned}$$

Stochastic Galerkin approximation

One way to find an approximation $u_{h,w} \in \mathbb{P}_{\Lambda(w)}(\Gamma) \otimes H_h(D)$ is by Galerkin projection,¹⁶ since $\mathbb{P}_{\Lambda(w)}(\Gamma) \otimes H_h(D) \subset L^2(\Gamma) \otimes H_0^1(D)$.

Definition 18.1

If $u_{h,w}^{SG} \in \mathbb{P}_{\Lambda(w)}(\Gamma) \otimes H_h(D)$ is such that

$$\mathbb{E} \left[\int_D a(\mathbf{y}, x) \nabla u_{h,w}^{SG}(\mathbf{y}, x) \cdot \nabla v(\mathbf{y}, x) dx \right] = \mathbb{E} \left[\int_D f(\mathbf{y}, x) v(\mathbf{y}, x) dx \right]$$

for all $v \in \mathbb{P}_{\Lambda(w)}(\Gamma) \otimes H_h(D)$, then $u_{h,w}^{SG}$ is called **stochastic Galerkin approximation** of u .

The error analysis then can be carried out using techniques similar to the case of deterministic variational problems.

¹⁶[Ghanem-Spanos, Karniadakis-Xiu et al., Matthies-Keese, Schwab-Todor et al., Knio-Le Maître et al., Babuska et al.,...]

Error analysis – Galerkin optimality

Since the global bilinear form $A(u, v) = \mathbb{E}[\int_D a \nabla u \cdot \nabla v \, dx]$ is continuous and coercive in $\mathcal{V} = L^2_\rho(\Gamma) \otimes H_0^1(D)$, standard analysis (Ceà's lemma) applies, leading to the (stochastic) Galerkin optimality

$$\|u - u_{h,w}\|_{\mathcal{V}} \leq \frac{C_P}{a_{min}} \inf_{v \in \mathbb{P}_{\Lambda(w)} \otimes H_h} \|u - v\|_{\mathcal{V}},$$

where $u_{h,w} \equiv u_{h,w}^{SG}$ is the stochastic Galerkin approximation.

Error splitting:

$$\|u - u_{h,w}\|_{\mathcal{V}} \leq \frac{C_P}{a_{min}} \left(\underbrace{\inf_{v \in \mathbb{P}_{\Lambda(w)} \otimes H_0^1} \|u - v\|_{\mathcal{V}}}_{\text{pol. approx. in probability}} + \underbrace{\inf_{v \in L^2_\rho \otimes H_h} \|u - v\|_{\mathcal{V}}}_{\text{FE approx. in physical space}} \right)$$

- ▶ The red term corresponds to the L^2 -projection error on the polynomial space $\mathbb{P}_{\Lambda(w)}$. We have studied this term in the previous parts.
- ▶ The blue term is a standard Finite Element error term.

Algebraic formulation

Expand on the basis $\{\phi_i(x)\} \times \{\psi_j(\mathbf{y})\}$ and use orthogonal polynomials $\{\psi_j\}$ with respect to $\rho(\mathbf{y})$. Then

$$\begin{aligned} \sum_{l=1}^{M_w} \sum_{k=1}^{N_h} u_{lk} \int_D \mathbb{E}[a(\mathbf{y}, x) \psi_l(\mathbf{y}) \psi_i(\mathbf{y})] \nabla \phi_k(x) \cdot \nabla \phi_j(x) dx \\ = \int_D \mathbb{E}[f(\mathbf{y}, x) \psi_i(\mathbf{y})] \phi_j(x) dx \end{aligned}$$

for all $j = 1, \dots, N_h$, $i = 1, \dots, M_w$.

- ▶ This is a **huge, fully coupled** linear system in the $M_w \times N_h$ unknowns u_{lk} .
- ▶ Need to compute in every quadrature point x_s a term of the form

$$c_{il}(x_s) = \mathbb{E}[a(\cdot, x_s) \psi_i \psi_l] = \int_{\Gamma} a(\mathbf{y}, x_s) \psi_i(\mathbf{y}) \psi_l(\mathbf{y}) \rho(\mathbf{y}) d\mathbf{y} \quad (96)$$

High dimensional integration (!) can be computational prohibitive

A special case: separable random fields

A case where (96) becomes tractable is when a and f have a **separable form**, in the sense that

$$a(\mathbf{y}, x) = \sum_{n=1}^{K_a} b_n^a(x) g_n^a(\mathbf{y}), \quad f(\mathbf{y}, x) = \sum_{n=1}^{K_f} b_n^f(x) g_n^f(\mathbf{y}).$$

In that case

$$\begin{aligned} & \sum_{l=1}^{M_w} \sum_{k=1}^{N_h} u_{lk} \int_D \mathbb{E} [a(\mathbf{y}, x) \psi_l(\mathbf{y}) \psi_i(\mathbf{y})] \nabla \phi_k(x) \cdot \nabla \phi_j(x) dx \\ &= \sum_{l=1}^{M_w} \sum_{k=1}^{N_h} u_{lk} \underbrace{\sum_{n=1}^{K_a} \mathbb{E} [g_n^a(\mathbf{y}) \psi_l(\mathbf{y}) \psi_i(\mathbf{y})]}_{\text{stochastic matrices } G_{il}^{(n)}} \underbrace{\int_D b_n^a(x) \nabla \phi_k(x) \cdot \nabla \phi_j(x) dx}_{\text{deterministic stiffness mat. } K_{jk}^{(n)}} \\ &= \sum_{l=1}^{M_w} \sum_{k=1}^{N_h} \sum_{n=1}^{K_a} G_{il}^{(n)} K_{jk}^{(n)} u_{lk} \end{aligned}$$

Similarly

$$\int_D \mathbb{E} [f(\mathbf{y}, x) \psi_i(\mathbf{y})] \phi_j(x) dx = \sum_{n=1}^{K_f} \underbrace{\mathbb{E} [g_n^f(\mathbf{y}) \psi_i(\mathbf{y})]}_{\text{stochastic r.h.s. } \mathbf{S}_i^{(n)}} \underbrace{\int_D b_n^f(x) \phi_j(x) dx}_{\text{deterministic r.h.s. } \mathbf{F}_j^{(n)}}$$

Tensor formulation:

$$\left(\sum_{n=1}^{K_a} G^{(n)} \otimes K^{(n)} \right) \mathbf{U} = \sum_{n=1}^{K_f} \mathbf{S}^{(n)} \otimes \mathbf{F}^{(n)}$$

with $\mathbf{U} = \{u_{lk}\}$.

Remember that the discrete solution can be thought as a linear combination of M_w deterministic finite element functions:

$$u_{h,w}(\mathbf{y}, x) = \sum_{l=1}^{M_w} u_h^{(l)}(x) \psi_l(\mathbf{y}), \quad \text{with } u_h^{(l)}(x) = \sum_{k=1}^{N_h} u_{lk} \phi_k(x).$$

Define the vectors of nodal values of such finite element functions:

$$\mathbf{U}^{(l)} = [u_{l1}, u_{l2}, \dots, u_{l,N_h}]^T.$$

Equivalent matrix formulation

$$\sum_{n=1}^{K_a} \begin{bmatrix} G_{11}^{(n)} K^{(n)} & G_{12}^{(n)} K^{(n)} & \dots & G_{1M_w}^{(n)} K^{(n)} \\ G_{21}^{(n)} K^{(n)} & \ddots & & \\ \vdots & & & \\ \vdots & & & \\ G_{M_w 1}^{(n)} K^{(n)} & \dots & & G_{M_w M_w}^{(n)} K^{(n)} \end{bmatrix} \begin{bmatrix} \mathbf{U}^{(1)} \\ \mathbf{U}^{(2)} \\ \vdots \\ \vdots \\ \mathbf{U}^{(M_w)} \end{bmatrix} = \sum_{n=1}^{K_f} \begin{bmatrix} S_1^{(n)} \mathbf{F}^{(n)} \\ S_2^{(n)} \mathbf{F}^{(n)} \\ \vdots \\ \vdots \\ S_{M_w}^{(n)} \mathbf{F}^{(n)} \end{bmatrix}$$

Remarks

- ▶ The system has dimension $N_h \cdot M_w \times N_h \cdot M_w$.
- ▶ Each block of the matrix is a whole deterministic FE stiffness matrix (which is sparse, however), multiplied by the coeff.
$$G_{il}^{(n)} = \mathbb{E}[g_n^a \psi_l \psi_i].$$
- ▶ One will never assemble the whole SGFEM matrix! Rather, we store only the stiffness matrices $K^{(n)}$ and the stochastic matrices $G^{(n)}$. Both can be precomputed.
- ▶ In general (non-linear problems and/or non-linear expansions of $a(\mathbf{y}, \mathbf{x})$) the coefficients $G_{il}^{(n)}$ will not be zero. Therefore, all the blocks of the **SGFEM matrix will be full!**
However, in special cases (see later) the SGFEM matrix turns out to be block sparse.

Remarks II

- ▶ For the problem analyzed, the SGFEM matrix is symmetric and positive definite. We can solve the problem with the Preconditioned Conjugate Gradient method.
- ▶ Commonly used preconditioner (see e.g. [Ghanem-Pellissetti, Powell-Elman, Powell-Ullmann, ...]): **mean-based** preconditioner

$$P = G^{(0)} \otimes K^{(0)}$$

with $K_{jk}^{(0)} = \int_D \mathbb{E}[a] \nabla \phi_k(x) \cdot \nabla \phi_j(x) dx$ and $G_{ii}^{(0)} = \mathbb{E}[\psi_i \psi_i]$.

- ▶ The preconditioner is **block diagonal** if an orthogonal basis $\{\psi_i\}$ is used. In this case, **the computation of the preconditioned residual in the PCG algorithm implies the solution of M_w uncoupled deterministic problems.**

Example 1 – monovariate linear expansion

- ▶ One uniform random variable, linear expansion:
 $a(y, x) = b_0 + b_1(x)y$, $y \sim \mathcal{U}(-\sqrt{3}, \sqrt{3})$.
- ▶ Basis of Legendre orthogonal polynomials up to degree w .

Tensor equation: $(\underbrace{\mathbb{I} \otimes K^{(0)} + G^{(1)} \otimes K^{(1)}}_A) \mathbf{U} = \mathbf{F}$

with $K_{jk}^{(n)} = \int_D b_n \nabla \phi_k \cdot \nabla \phi_j$ and $G_{ii}^{(1)} = \mathbb{E}[y \psi_i \psi_i]$.

- ▶ The matrix $G^{(1)}$ is **tridiagonal** (easy to check using the three term recurrence of Legendre polynomials).
- ▶ A result from [Powell-Elman '08]: the condition number of the preconditioned matrix $P^{-1}A$ is independent of w .

$$\text{cond}(P^{-1}A) \leq \frac{1+\tau}{1-\tau}, \quad \text{with } \tau = \frac{\|b_1\|_{L^\infty(D)}}{b_0} \sqrt{3},$$

Example 2 – multivariate linear expansion

- ▶ $a(y, x) = b_0 + \sum_{n=1}^N b_n(x)y_n$, $y \sim \mathcal{U}(-\sqrt{3}, \sqrt{3})$.
- ▶ Polynomial space: **total degree w** ; Legendre basis.

Tensor equation: $(\mathbb{I} \otimes K^{(0)} + \sum_{n=1}^N G^{(n)} \otimes K^{(n)})\mathbf{U} = F$
with $G_{ii}^{(n)} = \mathbb{E}[y_n \psi_i(\mathbf{y}) \psi_i(\mathbf{y})]$; highly sparse matrices!



$$N = 4$$
$$w = 2$$



$$N = 4$$
$$w = 3$$

The result on the condition number applies in this case too with
 $\tau = \frac{\sqrt{3}}{b_0} \sum_{n=1}^N \|b_n\|_{L^\infty(D)}$.

Example 3 – multivariate exponential expansion

- ▶ $a(\mathbf{y}, x) = a_{min} + e^{b_0(x) + \sum_{n=1}^N b_n(x)y_n}$, $y \sim \mathcal{N}(0, 1)$.
- ▶ Polynomial space: **total degree w** ; Hermite basis.

In this case, the diffusion coefficient is not in separable form:

$$a(y, x) = a_{min} + e^{b_0} \prod_{n=1}^N e^{b_n(x)y_n}$$

Idea: expand it in Hermite polynomials (Polynomial chaos expansion)

$$a(\mathbf{y}, x) = a_{min} + e^{b_0(x)} \sum_{\mathbf{s} \in \mathbb{N}^N} \left(\prod_{n=1}^N \frac{b_n(x)^{s_n}}{\sqrt{s_n!}} \right) \psi_{\mathbf{s}}(\mathbf{y})$$

with $\mathbf{s} = (s_1, \dots, s_N)$ a multi-index and $\psi_{\mathbf{s}}(\mathbf{y}) = \prod_{n=1}^N \psi_{s_n}(y_n)$
multivariate Hermite basis.

Example 3 – matrix formulation

$$u(\mathbf{y}, x) = \sum_{|\mathbf{p}| \leq w} u_h^{\mathbf{p}}(x) \psi_{\mathbf{p}}(\mathbf{y}) = \sum_{|\mathbf{p}| \leq w} \sum_{j=1}^{N_h} u_{\mathbf{p}j} \phi_j(x) \psi_{\mathbf{p}}(\mathbf{y})$$

$$\implies \left(\mathbb{I} \otimes K^{(0)} + \sum_{\mathbf{s} \in \mathbb{N}^N} G^{(\mathbf{s})} \otimes K^{(\mathbf{s})} \right) \mathbf{U} = F$$

with $G_{\mathbf{pq}}^{(\mathbf{s})} = \mathbb{E}[\psi_{\mathbf{s}} \psi_{\mathbf{p}} \psi_{\mathbf{q}}]$

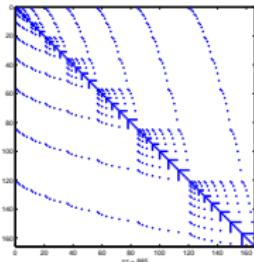
$$K_{jk}^{(\mathbf{s})} = \int_D e^{b_0(x)} \left(\prod_{n=1}^N \frac{b_n(x)^{s_n}}{\sqrt{s_n!}} \right) \nabla \phi_j(x) \cdot \nabla \phi_k(x) dx$$

- ▶ The algebraic system contains an infinite sum of matrices!
- ▶ Observe however that $\mathbb{E}[\psi_s \psi_p \psi_q] \neq 0$ only if $s \leq p + q$ thanks to the orthogonality of the Hermite polynomials. Therefore, we can restrict the infinite sum to the TD set $|s| \leq 2w$, i.e.

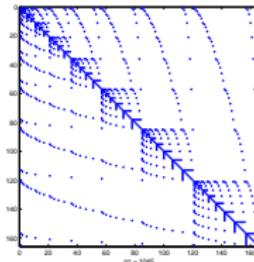
$$\sum_{s \in \mathbb{N}^N} G^{(s)} \otimes K^{(s)} = \sum_{|s| \leq 2w} G^{(s)} \otimes K^{(s)}$$

(see [Matthies et al])

- ▶ We can still use a mean-based preconditioner. However, in this case the condition number of the preconditioned matrix is not independent of w any more. Effective preconditioning is still an open question (see [Powell-Ullmann 2010]).
- ▶ The resulting matrix is **denser** than in the case of linear expansion but still sparse



8 uniform r.vs, TD(3)



8 log-normal r.vs, TD(3)

References

The part dedicated to the deterministic optimization of this lecture is based on the books by Nocedal & Wright [?] and Nesterov [?]. The stochastic counterpart is based on the lecture notes by Gower [?] and the work by Nemirovski et al. [?].

Optimization problem formulation I

In optimization, one seeks to minimize an objective function F that depends on a real vector variable \mathbf{x} . The mathematical formulation is stated as

$$\text{find } \mathbf{x}^* = \arg \min_{\mathbf{x} \in \Xi} F(\mathbf{x}), \quad (97)$$

where $\Xi \subset \mathbf{bR}^d$ is a feasible set in dimension d , $\mathbf{x} \in \Xi \subset \mathbf{bR}^d$ is a real vector with d components and the objective function $F: \Xi \mapsto \mathbf{bR}$. We focus on the case where Ξ is an Euclidian space and F is continuous and smooth on Ξ .

Solution:

- A point \mathbf{x}^* is a global minimizer if $F(\mathbf{x}^*) \leq F(\mathbf{x}) \quad \forall \mathbf{x} \in \Xi$.
- A point \mathbf{x}^* is a **(strict)** local minimizer if there is a neighborhood \mathcal{N} of \mathbf{x}^* such that $F(\mathbf{x}^*) \leq (<)F(\mathbf{x}) \quad \forall \mathbf{x} \in \mathcal{N}$.

Theorem 19.1 (First-Order Necessary Conditions)

If \mathbf{x}^* is a local minimizer and F is continuously differentiable in an open neighborhood of \mathbf{x}^* , then $\nabla F(\mathbf{x}^*) = \mathbf{0}$.

Theorem 19.2 (Second-Order Necessary Conditions)

Optimization problem formulation II

If \mathbf{x}^* is a local minimizer of F and $\nabla^2 F$ exists and is continuous in an open neighborhood of \mathbf{x}^* , then $\nabla F(\mathbf{x}^*) = \mathbf{0}$ and $\nabla^2 F(\mathbf{x}^*)$ is positive semidefinite.

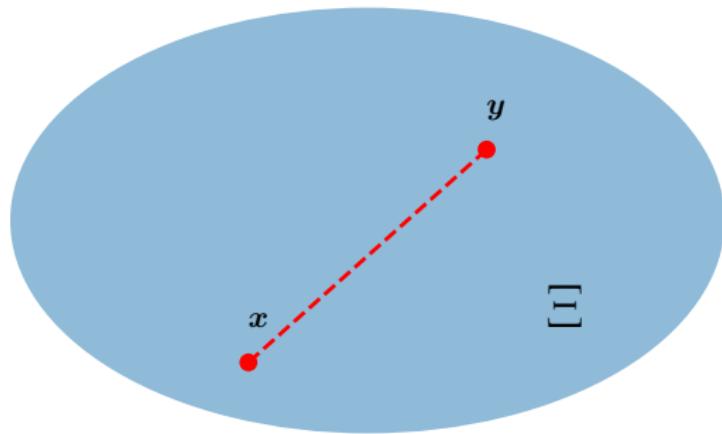
Theorem 19.3 (Second-Order Sufficient Conditions)

Suppose that $\nabla^2 F$ is continuous in an open neighborhood of \mathbf{x}^* and that $\nabla F(\mathbf{x}^*) = \mathbf{0}$ and $\nabla^2 F(\mathbf{x}^*)$ is positive definite. Then \mathbf{x}^* is a strict local minimizer of F .

Convexity I

The search space Ξ is convex if and only if

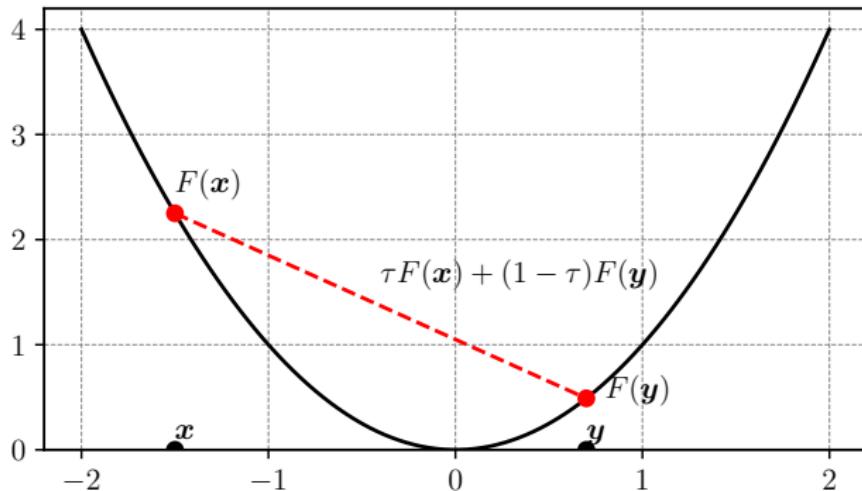
$$\tau \mathbf{x} + (1 - \tau) \mathbf{y} \in \Xi, \quad \forall \mathbf{x}, \mathbf{y} \in \Xi \quad \tau \in [0, 1]. \quad (98)$$



A function F is convex on Ξ if

$$F(\tau \mathbf{x} + (1 - \tau) \mathbf{y}) \leq \tau F(\mathbf{x}) + (1 - \tau) F(\mathbf{y}), \quad \forall \mathbf{x}, \mathbf{y} \in \Xi \quad \tau \in [0, 1]. \quad (99)$$

Convexity II



In addition, if F is differentiable, dividing by τ , and taking the limit $\tau \rightarrow 0$,

$$\lim_{\tau \rightarrow 0} \frac{F(\mathbf{y} + \tau(\mathbf{x} - \mathbf{y})) - F(\mathbf{y})}{\tau} \leq F(\mathbf{x}) - F(\mathbf{y}), \quad \forall \mathbf{x}, \mathbf{y} \in \Xi \quad \tau \in [0, 1], \quad (100)$$

thus,

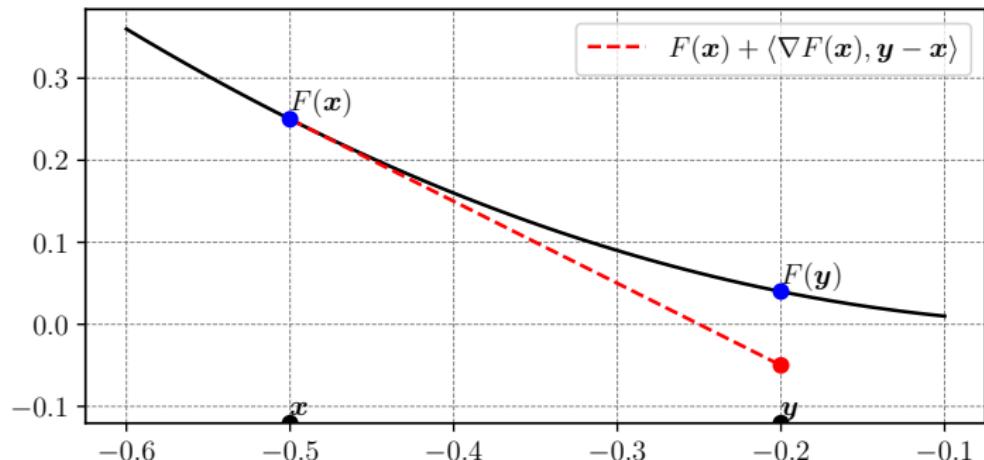
Convexity III

Assumption 19.1 (Convexity)

If F is convex and differentiable on Ξ , then

$$F(\mathbf{y}) \geq F(\mathbf{x}) + \langle \nabla F(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle, \quad \forall \mathbf{x}, \mathbf{y} \in \Xi. \quad (101)$$

Convexity IV



Convexity V

As a consequence of convexity, we have that the Hessian of F must be positive semi-definite.

$$\nabla^2 F(\mathbf{x}) \succeq 0 \quad (\text{positive semi-definite } \forall \mathbf{x} \in \Xi). \quad (102)$$

Convexity of Ξ and of F on Ξ implies that any local minimizer is also a global minimizer.

Smoothness I

Assumption 19.2 (L -smooth)

A differentiable function F is L -smooth if its gradient is Lipschitz continuous.

$$\|\nabla F(\mathbf{x}) - \nabla F(\mathbf{y})\| \leq L \|\mathbf{x} - \mathbf{y}\|, \quad \forall \mathbf{x}, \mathbf{y} \in \Xi. \quad (103)$$

With F twice differentiable,

$$\nabla F(\mathbf{x}) - \nabla F(\mathbf{x} + \tau\omega) = \int_0^\tau \nabla^2 F(\mathbf{x} + t\omega)\omega dt, \quad (104)$$

and L -smooth

$$\left\| \int_0^\tau \nabla^2 F(\mathbf{x} + t\omega)\omega dt \right\| \leq L\tau \|\omega\|. \quad (105)$$

Smoothness II

Without loss of generality, take $\|\omega\| = 1$, divide by τ , and take the limit when $\tau \rightarrow 0$ in (104), that is,

$$\lim_{\tau \rightarrow 0} \tau^{-1} \left\| \int_0^\tau \nabla^2 F(\mathbf{x} + t\omega) \omega \, dt \right\| \leq L \quad (106)$$

with $\omega \in \mathbf{b}R_{\text{unit}}^d$, expression (106) is equivalent to

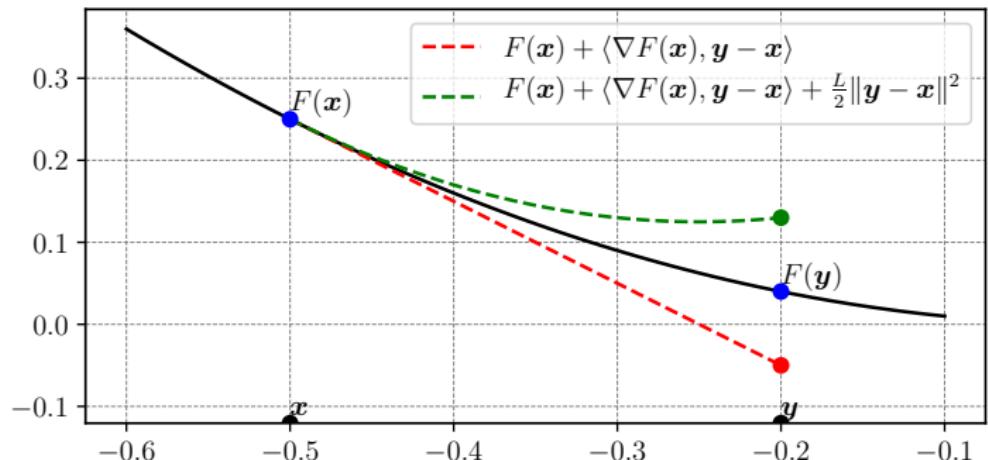
$$\nabla^2 F(\mathbf{x}) \preceq L \quad (107)$$

Thus, with an uniform bound over Hessian, the Taylor expansion of F yields

Remark 19.1 (Convex and L -smooth)

$$F(\mathbf{y}) \leq F(\mathbf{x}) + \langle \nabla F(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2} \|\mathbf{y} - \mathbf{x}\|^2, \quad \forall \mathbf{x}, \mathbf{y} \in \Xi \quad (108)$$

Smoothness III



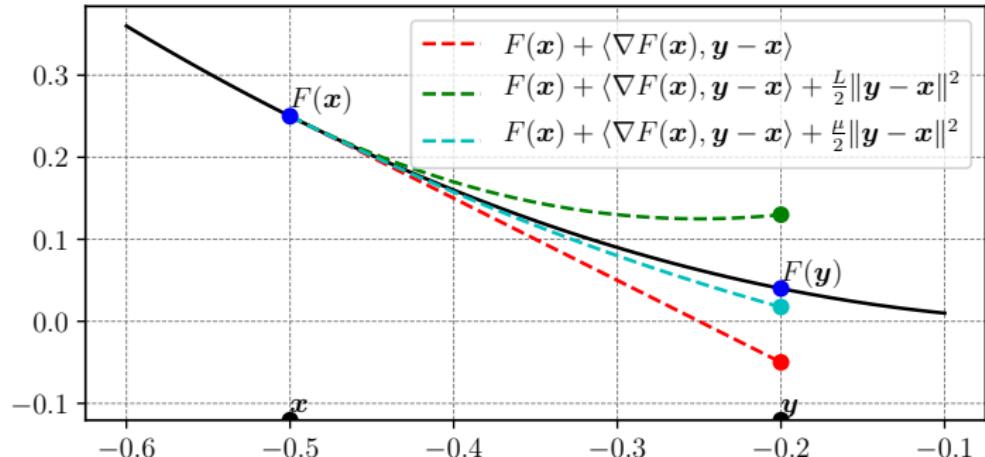
μ -strongly convex |

Assumption 19.3 (Strongly convex)

If F is μ -convex, then $\mu \preceq \nabla^2 F(\mathbf{x}) \quad \forall \mathbf{x} \in \Xi$, thus we have the lower bound

$$F(\mathbf{y}) \geq F(\mathbf{x}) + \langle \nabla F(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\mu}{2} \|\mathbf{y} - \mathbf{x}\|^2, \quad \forall \mathbf{x}, \mathbf{y} \in \Xi. \quad (109)$$

μ -strongly convex II



Steepest Descent (SD) I

To approximate the solution of (97), the Steepest Descent (SD) is constructed to solve the equivalent gradient flow problem

$$\dot{\mathbf{x}} = -\nabla F(\mathbf{x}), \quad (110)$$

where the dot represents the time derivative. Thus, the SD updating rule may be obtained by a Forward Euler discretization, that is,

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta \nabla F(\mathbf{x}_k), \quad (111)$$

where η is the step-size.

Note that the stability condition of the Forward Euler discretization is $2L^{-1}$, that is, $\eta < 2L^{-1}$.

Theorem 19.4 (Convergence of SD for L -smooth convex problems)

Steepest Descent (SD) II

Let F be convex and L -smooth and let \mathbf{x}_k be the sequence of iterates generated by the steepest descent method with $\eta = 1/L$. It follows that, at iteration k ,

$$F(\mathbf{x}_k) - F(\mathbf{x}^*) \leq \frac{2L \|\mathbf{x}_0 - \mathbf{x}^*\|^2}{k-1}. \quad (112)$$

Steepest Descent (SD) III

For the proof of this Theorem, we will use

$$\langle \nabla F(\mathbf{x}) - \nabla F(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \frac{1}{L} \|\nabla F(\mathbf{x}) - \nabla F(\mathbf{y})\|^2 \quad (\text{Co-coercivity}) \quad (113)$$

(114)

From L -smoothness and convexity of F ,

$$F(\mathbf{y}) - F(\mathbf{x}) \leq \langle \nabla F(\mathbf{y}), \mathbf{y} - \mathbf{x} \rangle - \frac{L}{2} \|\mathbf{x} - \mathbf{y}\|^2 \quad (115)$$

$$F(\mathbf{y}) - F(\mathbf{x}) \stackrel{(L\text{-smooth})}{\leq} \langle \nabla F(\mathbf{y}), \mathbf{y} - \mathbf{x} \rangle - \frac{1}{2L} \|\nabla F(\mathbf{x}) - \nabla F(\mathbf{y})\|^2 \quad (116)$$

$$F(\mathbf{x}) - F(\mathbf{y}) \leq \langle \nabla F(\mathbf{x}), \mathbf{x} - \mathbf{y} \rangle - \frac{1}{2L} \|\nabla F(\mathbf{y}) - \nabla F(\mathbf{x})\|^2. \quad (117)$$

Summing (116) and (117) we arrive at co-coercivity of the gradient.

Steepest Descent (SD) IV

Let's look at the squared distance to the optimum at \mathbf{x}_{k+1} ,

$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 = \|\mathbf{x}_k - \mathbf{x}^* - \eta \nabla F(\mathbf{x}_k)\|^2, \quad (118)$$

$$= \|\mathbf{x}_k - \mathbf{x}^*\|^2 - 2\eta \langle \mathbf{x}_k - \mathbf{x}^*, \nabla F(\mathbf{x}_k) \rangle + \eta^2 \|\nabla F(\mathbf{x}_k)\|^2 \quad (119)$$

$$\stackrel{\text{(Co-coercivity)}}{\leq} \|\mathbf{x}_k - \mathbf{x}^*\|^2 - 2\eta \frac{1}{L} \|\nabla F(\mathbf{x}_k)\|^2 + \eta^2 \|\nabla F(\mathbf{x}_k)\|^2 \quad (120)$$

$$\stackrel{\eta^* = 1/L}{=} \|\mathbf{x}_k - \mathbf{x}^*\|^2 - \frac{1}{L^2} \|\nabla F(\mathbf{x}_k)\|^2, \quad (121)$$

proving that the distance to the optimum decreases monotonically. From L -smoothness,

$$F(\mathbf{x}_{k+1}) \leq F(\mathbf{x}_k) + \langle \nabla F(\mathbf{x}_k), \mathbf{x}_{k+1} - \mathbf{x}_k \rangle + \frac{L}{2} \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \quad (122)$$

$$\stackrel{\text{s.d.}}{=} F(\mathbf{x}_k) + \langle \nabla F(\mathbf{x}_k), -\frac{1}{L} \nabla F(\mathbf{x}_k) \rangle + \frac{L}{2} \left\| -\frac{1}{L} \nabla F(\mathbf{x}_k) \right\|^2 \quad (123)$$

$$= F(\mathbf{x}_k) - \frac{1}{2L} \|\nabla F(\mathbf{x}_k)\|^2. \quad (124)$$

Steepest Descent (SD) V

Subtracting the optimum value we get a bound on the optimality gap,

$$F(\mathbf{x}_{k+1}) - F(\mathbf{x}^*) \leq F(\mathbf{x}_k) - F(\mathbf{x}^*) - \frac{1}{2L} \|\nabla F(\mathbf{x}_k)\|^2 \quad (125)$$

From convexity,

$$F(\mathbf{x}_k) - F(\mathbf{x}^*) \leq \langle \nabla F(\mathbf{x}_k), \mathbf{x}_k - \mathbf{x}^* \rangle \quad (126)$$

$$\stackrel{\text{C.S.}}{\leq} \|\nabla F(\mathbf{x}_k)\| \|\mathbf{x}_k - \mathbf{x}^*\| \quad (127)$$

$$\leq \|\nabla F(\mathbf{x}_k)\| \|\mathbf{x}_0 - \mathbf{x}^*\|, \quad (128)$$

thus,

$$\|\nabla F(\mathbf{x}_k)\| \geq \frac{F(\mathbf{x}_k) - F(\mathbf{x}^*)}{\|\mathbf{x}_0 - \mathbf{x}^*\|}. \quad (129)$$

Substituting on (125),

$$F(\mathbf{x}_{k+1}) - F(\mathbf{x}^*) \leq F(\mathbf{x}_k) - F(\mathbf{x}^*) - \frac{1}{2L} \|\nabla F(\mathbf{x}_k)\|^2 \quad (130)$$

$$\leq F(\mathbf{x}_k) - F(\mathbf{x}^*) - \frac{1}{2L} \frac{(F(\mathbf{x}_k) - F(\mathbf{x}^*))^2}{\|\mathbf{x}_0 - \mathbf{x}^*\|^2}. \quad (131)$$

Steepest Descent (SD) VI

Defining $\Delta_k := F(\mathbf{x}_k) - F(\mathbf{x}^*)$ and $\beta := (2L \|\mathbf{x}_0 - \mathbf{x}^*\|^2)^{-1}$,

$$\Delta_{k+1} \leq \Delta_k - \beta \Delta_k^2 \quad (132)$$

$$\beta \frac{\Delta_k}{\Delta_{k+1}} \leq \Delta_{k+1}^{-1} - \Delta_k^{-1} \quad (133)$$

$$\beta \leq \Delta_{k+1}^{-1} - \Delta_k^{-1}. \quad (134)$$

Summing up,

$$(k-1)\beta \leq \Delta_k^{-1} - \Delta_0^{-1} \quad (135)$$

$$\leq \Delta_k^{-1}, \quad (136)$$

thus,

$$F(\mathbf{x}_k) - F(\mathbf{x}^*) \leq \frac{2L \|\mathbf{x}_0 - \mathbf{x}^*\|^2}{k-1}. \quad (137)$$

(Jupyter notebook `sd_conv.ipynb`)



Steepest Descent - strongly-convex case I

Theorem 19.5 (Convergence of SD for L -smooth strongly-convex functions)

Let F be L -smooth and strongly-convex with constant μ . The sequence of iterates \mathbf{x}_k generated by the steepest descent method using step-size $\eta = \frac{2}{\mu+L}$ satisfies

$$\|\mathbf{x}_k - \mathbf{x}^*\|^2 \leq \left(\frac{L-\mu}{L+\mu} \right)^{2k} \|\mathbf{x}_0 - \mathbf{x}^*\|^2, \quad (138)$$

or, equivalently,

$$F(\mathbf{x}_k) - F(\mathbf{x}^*) \leq \frac{L}{2} \left(\frac{L-\mu}{L+\mu} \right)^{2k} \|\mathbf{x}_0 - \mathbf{x}^*\|^2. \quad (139)$$

First, we need to prove that, if F is μ -convex and L -smooth, for all $\mathbf{x}, \mathbf{y} \in \Xi$,

$$\langle \nabla F(\mathbf{x}) - \nabla F(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \frac{\mu L}{\mu + L} \|\mathbf{x} - \mathbf{y}\|^2 + \frac{1}{\mu + L} \|\nabla F(\mathbf{x}) - \nabla F(\mathbf{y})\|^2. \quad (140)$$

Steepest Descent - strongly-convex case II

Let's start by defining the convex and $(L - \mu)$ -smooth function
 $\phi(\mathbf{x}) = F(\mathbf{x}) - \mu/2 \|\mathbf{x}\|^2$. Then, from co-coercivity of the gradient of ϕ ,

$$\langle \nabla \phi(\mathbf{x}) - \nabla \phi(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \frac{1}{L - \mu} \|\nabla \phi(\mathbf{x}) - \nabla \phi(\mathbf{y})\|^2 \quad (141)$$

$$\begin{aligned} \langle \nabla F(\mathbf{x}) - \nabla F(\mathbf{y}) - \mu(\mathbf{x} - \mathbf{y}), \mathbf{x} - \mathbf{y} \rangle &\geq \frac{1}{L - \mu} \|\nabla F(\mathbf{x}) - \nabla F(\mathbf{y})\|^2 \\ &\quad - 2\mu \langle \nabla F(\mathbf{x}) - \nabla F(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \mu^2 \|\mathbf{x} - \mathbf{y}\|^2 \end{aligned} \quad (142)$$

$$\frac{L + \mu}{L - \mu} \langle \nabla F(\mathbf{x}) - \nabla F(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \frac{L\mu}{L - \mu} \|\mathbf{x} - \mathbf{y}\|^2 + \frac{1}{L - \mu} \|\nabla F(\mathbf{x}) - \nabla F(\mathbf{y})\|^2, \quad (143)$$

from where we conclude our proof. □

Steepest Descent - strongly-convex case III

The distance to the optimum squared is

$$\begin{aligned}\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 &= \|\mathbf{x}_k - \mathbf{x}^* - \eta \nabla F(\mathbf{x}_k)\|^2, \\ &= \|\mathbf{x}_k - \mathbf{x}^*\|^2 - 2\eta \langle \mathbf{x}_k - \mathbf{x}^*, \nabla F(\mathbf{x}_k) \rangle + \eta^2 \|\nabla F(\mathbf{x}_k)\|^2 \\ &\stackrel{(140)}{\leq} \left(1 - \frac{2\eta\mu L}{\mu + L}\right) \|\mathbf{x}_k - \mathbf{x}^*\|^2 + \eta \left(\eta - \frac{2}{\mu + L}\right) \|\nabla F(\mathbf{x}_k)\|^2 \\ &\stackrel{\eta^* = 2/L + \mu}{=} \left(\frac{L - \mu}{L + \mu}\right)^2 \|\mathbf{x}_k - \mathbf{x}^*\|^2 \\ &\stackrel{\text{sum}}{=} \left(\frac{L - \mu}{L + \mu}\right)^{2k} \|\mathbf{x}_0 - \mathbf{x}^*\|^2.\end{aligned}$$

From convexity and L -smoothness of F ,

$$F(\mathbf{x}_k) \leq F(\mathbf{x}^*) + \frac{L}{2} \|\mathbf{x}_k - \mathbf{x}^*\|^2, \quad (144)$$

hence,

$$F(\mathbf{x}_k) - F(\mathbf{x}^*) \leq \frac{L}{2} \left(\frac{L - \mu}{L + \mu}\right)^{2k} \|\mathbf{x}_0 - \mathbf{x}^*\|^2. \quad \square \quad (145)$$

(Jupyter notebook `sd_conv.ipynb`)

Stochastic Gradient Descent (SGD) I

In stochastic optimization, one seeks to minimize an objective function $F := \mathbf{b}E[f(\mathbf{x}, \boldsymbol{\theta})] = \int f(\mathbf{x}, \boldsymbol{\theta})\rho(\boldsymbol{\theta}) d\boldsymbol{\theta}$ that depends on a real vector variable \mathbf{x} and a random vector variable $\boldsymbol{\theta}$. The mathematical formulation is stated as,

$$\text{find } \mathbf{x}^* = \arg \min_{\mathbf{x} \in \Xi} \mathbf{b}E[f(\mathbf{x}, \boldsymbol{\theta})], \quad (146)$$

where $\Xi \subset \mathbf{b}R^d$ is a feasible set in dimension d and $\mathbf{x} \in \Xi \subset \mathbf{b}R^d$ is a real vector with d components and $\boldsymbol{\theta} \in \Theta \subset \mathbf{b}R^r$ is a measurable function. Additionally, $f: \Xi \times \Theta \mapsto \mathbf{b}R$ is a smooth function with respect to \mathbf{x} . The Stochastic Gradient Descent SGD is used to solve stochastic optimization problems. SGD has update

Sample $\boldsymbol{\theta}_k \sim \rho$ where ρ is the probability distribution of $\boldsymbol{\theta}$,

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta_k \mathbf{v}(\mathbf{x}_k, \boldsymbol{\theta}_k), \quad (147)$$

where

- ▶ \mathbf{v} is an unbiased estimator of the mean gradient of f at \mathbf{x} ,

Stochastic Gradient Descent (SGD) II

- ▶ η_k is a decreasing stepsize, e.g., $\frac{1}{L\sqrt{k+1}}$, $\frac{a}{c+\sqrt{k+1}}$.

In the classical setting, the estimator v is defined as follows.

$v(x_k) := \nabla f(x_k, \theta_k)$. However, to reduce the statistical error, a minibatch sampling of size b may be used, that is,

$v := \frac{1}{b} \sum_{i=1}^b \nabla f(x, \theta_i)$, where $\{\theta_i\}_{i=1}^b$ is the minibatch sample.

Stochastic Gradient Descent - convex case I

Theorem 19.6

Let $f(\mathbf{x}, \boldsymbol{\theta})$ be a function which is L -smooth and convex on \mathbf{x} .

Additionally, let $\mathbf{v}_k := \mathbf{v}(\mathbf{x}_k, \boldsymbol{\theta})$ be an unbiased estimator of the mean gradient of F . Assume that exists a $\sigma > 0$ such that $\mathbf{b}E[\|\mathbf{v}_k\|^2 | \mathbf{x}_k] \leq \sigma^2$ for all k and that $\|\mathbf{x}\| \leq r$. For a stepsize $\eta_k = \frac{\sqrt{2}r}{\sigma\sqrt{k+1}}$, the optimality gap satisfies

$$\mathbf{b}E[F(\bar{\mathbf{x}}_k)] - F(\mathbf{x}^*) \leq \frac{2\sqrt{2}r\sigma}{\sqrt{k+1}} \quad (148)$$

for $\bar{\mathbf{x}}_k = \frac{1}{k+1} \sum_{i=0}^k \mathbf{x}_i$ (Polyak–Ruppert averaging).

Stochastic Gradient Descent - convex case II

Similarly to the steepest descent analysis,

$$\begin{aligned}\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 &= \|\mathbf{x}_k - \mathbf{x}^* - \eta_k \mathbf{v}_k\|^2 \\ &= \|\mathbf{x}_k - \mathbf{x}^*\|^2 - 2\eta_k \langle \mathbf{v}_k, \mathbf{x}_k - \mathbf{x}^* \rangle + \eta_k^2 \|\mathbf{v}_k\|^2,\end{aligned}\quad (149)$$

Taking conditional expectation on \mathbf{x}_k

$$\begin{aligned}\mathbf{b}E[\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 | \mathbf{x}_k] &= \|\mathbf{x}_k - \mathbf{x}^*\|^2 - 2\eta_k \langle \nabla F(\mathbf{x}_k), \mathbf{x}_k - \mathbf{x}^* \rangle + \eta_k^2 \mathbf{b}E[\|\mathbf{v}_k\|^2 | \mathbf{x}_k] \\ &\leq \|\mathbf{x}_k - \mathbf{x}^*\|^2 - 2\eta_k \langle \nabla F(\mathbf{x}_k), \mathbf{x}_k - \mathbf{x}^* \rangle + \eta_k^2 \sigma^2.\end{aligned}\quad (150)$$

$$\leq \|\mathbf{x}_k - \mathbf{x}^*\|^2 - 2\eta_k (F(\mathbf{x}_k) - F(\mathbf{x}^*)) + \eta_k^2 \sigma^2. \quad (151)$$

Taking total expectation and rearranging,

$$\mathbf{b}E[F(\mathbf{x}_k)] - F(\mathbf{x}^*) \leq \frac{1}{2\eta_k} (\mathbf{b}E[\|\mathbf{x}_k - \mathbf{x}^*\|^2] - \mathbf{b}E[\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2]) + \frac{\eta_k \sigma^2}{2}. \quad (152)$$

Stochastic Gradient Descent - convex case III

Summing up all the iterations and assuming that η is a decreasing sequence,

$$\begin{aligned} \sum_{i=0}^k (\mathbf{b}E[F(\mathbf{x}_i)] - F(\mathbf{x}^*)) &\leq \frac{1}{2\eta_0} \|\mathbf{x}_0 - \mathbf{x}^*\|^2 + \frac{1}{2} \sum_{i=0}^{k-1} \left(\frac{1}{\eta_{i+1}} - \frac{1}{\eta_i} \right) \mathbf{b}E[\|\mathbf{x}_{i+1} - \mathbf{x}^*\|^2] \\ &\quad - \frac{1}{2\eta_{k+1}} \mathbf{b}E[\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2] + \frac{\sigma^2}{2} \sum_{i=1}^k \eta_i \end{aligned} \tag{153}$$

$$\|\mathbf{x}\| \leq \frac{2}{\eta_0} r^2 + \sum_{i=0}^{k-1} \left(\frac{1}{\eta_{i+1}} - \frac{1}{\eta_i} \right) 2r^2 + \frac{\sigma^2}{2} \sum_{i=1}^k \eta_i. \tag{154}$$

Now, letting $\bar{\mathbf{x}}_k = \frac{1}{k+1} \sum_{i=0}^k \mathbf{x}_i$, from Jensen's inequality,

$$\begin{aligned} \mathbf{b}E[F(\bar{\mathbf{x}}_k)] - F(\mathbf{x}^*) &\leq \frac{1}{k+1} \sum_{i=0}^k (\mathbf{b}E[F(\mathbf{x})] - F(\mathbf{x}^*)), \\ &\leq \frac{2r^2}{(k+1)\eta_k} + \frac{\sigma^2}{2(k+1)} \sum_{i=0}^k \eta_i. \end{aligned} \tag{155}$$

Stochastic Gradient Descent - convex case IV

Substituting the stepsize $\eta_k = \frac{\eta_0}{\sqrt{k+1}}$, we arrive at

$$\begin{aligned}\mathbf{b}E[F(\bar{x}_k)] - F(x^*) &\leq \frac{2r^2}{\eta_0\sqrt{k+1}} + \frac{\eta_0\sigma^2}{2(k+1)} \sum_{i=0}^k \frac{1}{\sqrt{k+1}}, \\ &\leq \frac{1}{\sqrt{k+1}} \left(\frac{2r^2}{\eta_0} + \eta_0\sigma^2 \right).\end{aligned}\tag{156}$$

The initial stepsize η_0 that minimizes the RHS of (156) is given by

$$\eta_0 = \frac{\sqrt{2}r}{\sigma}.\tag{157}$$

Substituting the stepsize η_0 in (156),

$$\mathbf{b}E[F(\bar{x}_k)] - F(x^*) \leq \frac{2\sqrt{2}r\sigma}{\sqrt{k+1}},\tag{158}$$

we conclude our proof. □

(Jupyter notebook sgd.ipynb)

Stochastic Gradient Descent - strongly-convex case I

In the strongly-convex case, can linear convergence be recovered?

Theorem 19.7 (Linear convergence for SGD on L-smooth and strongly-convex problems with adaptive sample-size)

Let F be L -smooth and strongly-convex with constant μ . For an unbiased estimator of the gradient $v_k := \frac{1}{b} \sum_{i=1}^b \nabla f(x_k, \theta_i)$ with total variation $\frac{\text{tr}(\Sigma_k)}{b_k}$, the stochastic gradient method generates a sequence $x_k - x^*$ that converges linearly in the L^2 sense, thus

$$\mathbf{E}[F(x_k)] - F(x^*) \leq \frac{L}{2} \left(\frac{\left(\frac{L-\mu}{L+\mu} \right)^2 + \epsilon^2}{1 + \epsilon^2} \right)^k \|x_0 - x^*\|^2, \quad (159)$$

for a batch-size $b_k = \frac{\text{tr}(\Sigma_k)}{\epsilon^2 \|\nabla F(x_k)\|^2}$, step-size $\eta = \frac{2}{(L+\mu)(1+\epsilon^2)}$, and confidence parameter $\epsilon > 0$.

Stochastic Gradient Descent - strongly-convex case II

For the minibatch gradient estimator with sample-size b ,

$$\mathbf{b}E[\|\mathbf{v}_k\|^2 | \mathbf{x}_k] = \mathbf{b}E[\|\mathbf{v}_k - \nabla F(\mathbf{x}_k) + \nabla F(\mathbf{x}_k)\|^2 | \mathbf{x}_k] \quad (160)$$

$$= \mathbb{E}[\|\mathbf{v}_k - \nabla F(\mathbf{x}_k)\|^2 + 2 \langle \mathbf{v}_k - \nabla F(\mathbf{x}_k), \nabla F(\mathbf{x}_k) \rangle + \|\nabla F(\mathbf{x}_k)\|^2 | \mathbf{x}_k] \quad (161)$$

$$= \frac{\text{tr}(\Sigma_k)}{b} + \|\nabla F(\mathbf{x}_k)\|^2, \quad (162)$$

with $\Sigma_k = \mathbf{b}E[(\nabla f(\mathbf{x}_k, \theta) - \nabla F(\mathbf{x}_k))(\nabla f(\mathbf{x}_k, \theta) - \nabla F(\mathbf{x}_k))^T]$.

Stochastic Gradient Descent - strongly-convex case III

$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 = \|\mathbf{x}_k - \eta \mathbf{v}_k - \mathbf{x}^*\|^2 \quad (163)$$

$$= \|\mathbf{x}_k - \mathbf{x}^*\|^2 - 2\eta \langle \mathbf{x}_k - \mathbf{x}^*, \mathbf{v}_k \rangle + \eta^2 \|\mathbf{v}_k\|^2 \quad (164)$$

Taking expectation conditioned on \mathbf{x}_k

$$\mathbf{b}E[\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 | \mathbf{x}_k] = \|\mathbf{x}_k - \mathbf{x}^*\|^2 - 2\eta \langle \mathbf{x}_k - \mathbf{x}^*, \nabla F(\mathbf{x}_k) \rangle + \eta^2 \|\nabla F(\mathbf{x}_k)\|^2 + \eta^2 \frac{\text{tr}(\Sigma_k)}{b} \quad (165)$$

We can adaptively increase b in order to control the gradient noise,

$$b_k := \frac{\text{tr}(\Sigma_k)}{\epsilon^2 \|\nabla F(\mathbf{x}_k)\|^2}. \quad (166)$$

Stochastic Gradient Descent - strongly-convex case IV

$$\mathbf{b}E[\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 | \mathbf{x}_k] = \|\mathbf{x}_k - \mathbf{x}^*\|^2 - 2\eta \langle \mathbf{x}_k - \mathbf{x}^*, \nabla F(\mathbf{x}_k) \rangle + \eta^2(1 + \epsilon^2) \|\nabla F(\mathbf{x}_k)\|^2 \quad (167)$$

$$\leq \left(1 - \frac{2\eta\mu L}{L + \mu}\right) \|\mathbf{x}_k - \mathbf{x}^*\|^2 + \eta \left(-\frac{2}{L + \mu} + \eta(1 + \epsilon^2)\right) \|\nabla F(\mathbf{x}_k)\|^2 \quad (168)$$

$$\stackrel{\eta^*}{=} \left(\left(\frac{L - \mu}{L + \mu}\right)^2 + \epsilon^2\right) \frac{1}{1 + \epsilon^2} \|\mathbf{x}_k - \mathbf{x}^*\|^2. \quad (169)$$

The inequality in the second line is obtained from the same identity used before,

$$\langle \nabla F(\mathbf{x}) - \nabla F(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \frac{\mu L}{\mu + L} \|\mathbf{x} - \mathbf{y}\|^2 + \frac{1}{\mu + L} \|\nabla F(\mathbf{x}) - \nabla F(\mathbf{y})\|^2. \quad (170)$$

Stochastic Gradient Descent - strongly-convex case V

Taking total expectation and summing up the iterations,

$$\mathbf{b}E[\|\mathbf{x}_k - \mathbf{x}^*\|^2] \leq \left(\frac{\left(\frac{L-\mu}{L+\mu}\right)^2 + \epsilon^2}{1 + \epsilon^2} \right)^k \|\mathbf{x}_0 - \mathbf{x}^*\|^2. \quad (171)$$

From convexity and L -smoothness of F ,

$$F(\mathbf{x}_k) \leq F(\mathbf{x}^*) + \frac{L}{2} \|\mathbf{x}_k - \mathbf{x}^*\|^2, \quad (172)$$

$$\mathbf{b}E[F(\mathbf{x}_k)] - F(\mathbf{x}^*) \leq \frac{L}{2} \left(\frac{\left(\frac{L-\mu}{L+\mu}\right)^2 + \epsilon^2}{1 + \epsilon^2} \right)^k \|\mathbf{x}_0 - \mathbf{x}^*\|^2. \quad (173)$$

□

Stochastic Gradient Descent - strongly-convex case VI

From L -smoothness,

$$\|\nabla F(\mathbf{x}_k)\|^2 \leq L^2 \|\mathbf{x}_k - \mathbf{x}^*\|^2, \quad (174)$$

thus,

$$\mathbf{E}[\|\nabla F(\mathbf{x}_k)\|^2] \leq L^2 \underbrace{\left(\frac{\left(\frac{L-\mu}{L+\mu} \right)^2 + \epsilon^2}{1 + \epsilon^2} \right)^k}_{:=c} \|\mathbf{x}_0 - \mathbf{x}^*\|^2. \quad (175)$$

thus,

$$b_k = \mathcal{O}(c^{-k}), \quad (176)$$

and

$$k = - \left(\frac{\log(b_k) - \log(C)}{\log(c)} \right). \quad (177)$$

Stochastic Gradient Descent - strongly-convex case VII

Then,

$$\mathbf{b}E[F(\mathbf{x}_k)] - F(\mathbf{x}^*) \leq \frac{L}{2} c^k \|\mathbf{x}_0 - \mathbf{x}^*\|^2. \quad (178)$$

$$= \frac{L}{2} c^{\frac{-\log(b_k) + \log(C)}{\log(c)}} \|\mathbf{x}_0 - \mathbf{x}^*\|^2. \quad (179)$$

$$= \frac{L}{2} \frac{C}{b_k} \|\mathbf{x}_0 - \mathbf{x}^*\|^2, \quad (180)$$

thus,

$$\mathbf{b}E[F(\mathbf{x}_k)] - F(\mathbf{x}^*) = \mathcal{O}(b_k^{-1}). \quad (181)$$

(Jupyter notebook `sgd_adpt.ipynb`)

There are many other SGD variations, e.g., SVRG, SAGA, SAG, SARAH, SPIDER.

Recapitulating

L -smooth	Convex	μ -convex	stochastic	opt. gap
✓	✓	✗	✗	$\mathcal{O}(k^{-1})$
✓	✓	✓	✗	$\mathcal{O}(c^k)$
✓	✓	✗	✓	$\mathcal{O}(k^{-\frac{1}{2}})$
✓	✓	✓	✓	$\mathcal{O}(c^k)^{17}$

¹⁷using adaptivity.

Finite sum minimization I

In stochastic optimization, one seeks to solve

$$\text{find } \mathbf{x}^* = \arg \min_{\mathbf{x} \in \Xi} \mathbf{b}E[f(\mathbf{x}, \theta)]. \quad (182)$$

What if we can't sample θ from ρ and instead we have a finite-sized sample $\Theta := \{\theta_\alpha\}_{\alpha=1}^N$?

In this case, we can model our optimization problem as

$$\text{find } \mathbf{x}^* = \arg \min_{\mathbf{x} \in \Xi} \frac{1}{N} \sum_{\alpha=1}^N f(\mathbf{x}, \theta_\alpha). \quad (183)$$

Variance controlled methods I

Stochastic average gradient (SAG)

SAG [?] is an stochastic optimization method for minimizing the average of a finite number of functions.

- ▶ Each iteration a θ_α is randomly chosen from Θ and the objective function gradient is evaluated at θ_α .
- ▶ A memory of the last gradient evaluated for each θ_α is kept.
- ▶ The average of the memory is used as a gradient mean estimate.

Variance controlled methods II

Stochastic average gradient (SAG)

Let N be the number of functions, such that $F(\mathbf{x}) = \frac{1}{N} \sum_{\alpha=1}^N f(\mathbf{x}, \theta_\alpha)$. Then, for a set $\{G_\alpha\}_{\alpha=1}^N$ containing the gradients for each function,

$$\text{Sample } \alpha \sim \mathcal{U}[1, N] \quad (184)$$

$$G_\alpha \leftarrow \nabla f(\mathbf{x}_k, \theta_\alpha) \quad (185)$$

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta \frac{1}{N} \sum_{\alpha=1}^N G_\alpha. \quad (186)$$

G_α is a set of gradients that grows with k . SAG update is biased, that is, $\mathbf{b}E[G] \neq \nabla F$ at k ; however, the variance is greatly reduced and the bias decreases as optimization proceeds.

Control variates in stochastic optimization

Variance controlled methods III

Control variates is a technique for estimating the expected value of a random variable X , given that it is highly correlated with another random variable Y with known expected value:

$$Z = X - Y + \mathbf{b}E[Y], \quad \mathbf{b}E[Z] = \mathbf{b}E[X], \quad \mathbf{b}V[Z] \leq \mathbf{b}V[X]. \quad (187)$$

If (component-wise) control variates are used to estimate the gradient $\nabla f(\mathbf{x}_k, \boldsymbol{\theta})$ given some already known $\nabla f(\mathbf{x}_0, \boldsymbol{\theta})$,

$$\widehat{\nabla F(\mathbf{x}_k)} = \nabla f(\mathbf{x}_k, \boldsymbol{\theta}_\alpha) - \nabla f(\mathbf{x}_0, \boldsymbol{\theta}_\alpha) + \frac{1}{N} \sum_{\beta=1}^N \nabla f(\mathbf{x}_0, \boldsymbol{\theta}_\beta). \quad (188)$$

Variance controlled methods IV

Stochastic variance reduction gradient (SVRG)

The SVRG method [?] for the minimization of a finite sum of strongly-convex functions. Then, the gradient is updated each iteration, and the optimization step is

$$\text{Sample } \alpha \sim \mathcal{U}[1, N] \quad (189)$$

$$\widehat{\nabla F(\mathbf{x}_k)} = \nabla f(\mathbf{x}_k, \boldsymbol{\theta}_\alpha) - \nabla f(\mathbf{x}_0, \boldsymbol{\theta}_\alpha) + \frac{1}{N} \sum_{\beta=1}^N \nabla f(\mathbf{x}_0, \boldsymbol{\theta}_\beta) \quad (190)$$

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta \widehat{\nabla F(\mathbf{x}_k)}. \quad (191)$$

After a given number of iterations, the full gradient is reevaluated and $\mathbf{x}_0 = \mathbf{x}_k$.

Variance controlled methods V

SAGA

SAGA [?] can be seen as a combination of SAG and SVRG. It keeps a memory of the gradients like SAG, and performs control variate with respect to it, as in SVRG.

Variance controlled methods VI

SAGA

Each iteration,

$$\text{Sample } \alpha \sim \mathcal{U}[1, N] \quad (192)$$

$$\widehat{\nabla F(\mathbf{x}_k)} = \nabla f(\mathbf{x}_k, \boldsymbol{\theta}_\alpha) - \mathcal{G}_\alpha + \frac{1}{N} \sum_{\beta=1}^N \mathcal{G}_\beta \quad (193)$$

$$\mathcal{G}_\alpha \leftarrow \nabla f(\mathbf{x}_k, \boldsymbol{\theta}_\alpha) \quad (194)$$

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta \widehat{\nabla F(\mathbf{x}_k)}. \quad (195)$$

SAGA does not require the evaluation of the true gradient as SVRG, but it is a biased method as SAG.

Costs

Variance controlled methods VII

The table below presents the number of gradients to achieve an tol -precision of the average of N functions with conditioning number $\kappa = L/\mu$.

	Strongly convex	Convex
SD	$\mathcal{O}(N\kappa \log(tol^{-1}))$	$\mathcal{O}(N/tol)$
SGD	$\mathcal{O}(tol^{-\frac{1}{2}})$	$\mathcal{O}(tol^{-\frac{1}{2}})$
SVRG	$\mathcal{O}((N + \kappa) \log(tol^{-1}))$	$\mathcal{O}(N + (\sqrt{N}/tol))$
SAG	$\mathcal{O}((N + \kappa) \log(tol^{-1}))$	—
SAGA	$\mathcal{O}((N + \kappa) \log(tol^{-1}))$	$\mathcal{O}(N + (N/tol))$

Multi-iteration Stochastic Optimizers (MICE) I

The MICE algorithm [?] is a control variates method for estimating gradients for optimization methods in a hierarchical fashion. MICE constructs a hierarchy of previous optimization iterations and samples gradients in such a way as to achieve a desired precision with as few gradient evaluations as possible.

Multi-iteration Stochastic Optimizers (MICE) II

Given an index set \mathcal{L}_k such that $k \in \mathcal{L}_k \subset \{0, \dots, k\}$, we define the MICE estimator as

$$\nabla \mathcal{F}_k := \sum_{\ell \in \mathcal{L}_k} \frac{1}{M_{\ell,k}} \sum_{\alpha \in \mathcal{I}_{\ell,k}} \Delta_{\ell,k,\alpha}, \quad (196)$$

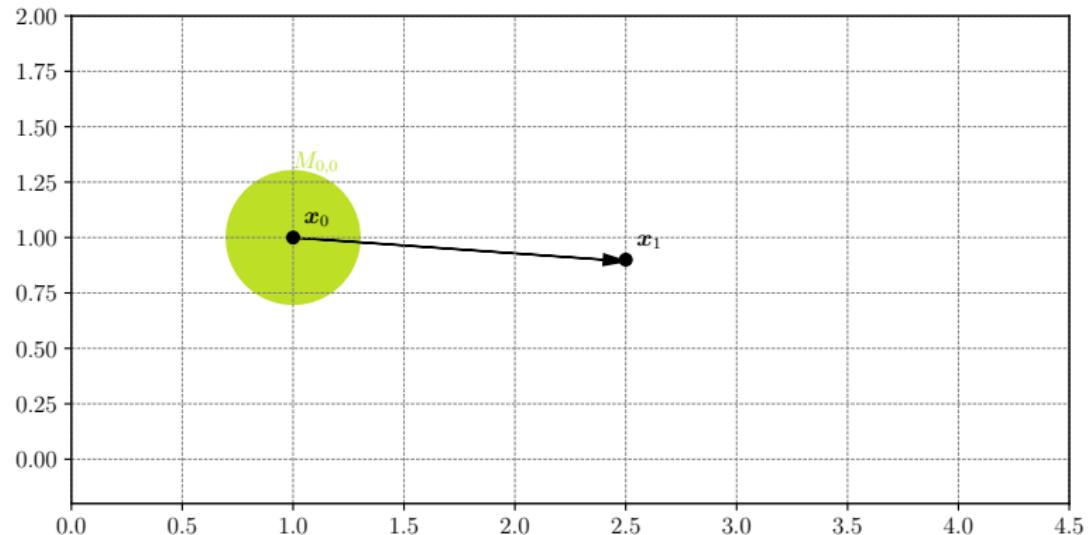
where

$$\Delta_{\ell,k,\alpha} := \nabla f(\mathbf{x}_\ell, \boldsymbol{\theta}_\alpha) - \nabla f(\mathbf{x}_{p_k(\ell)}, \boldsymbol{\theta}_\alpha), \quad (197)$$

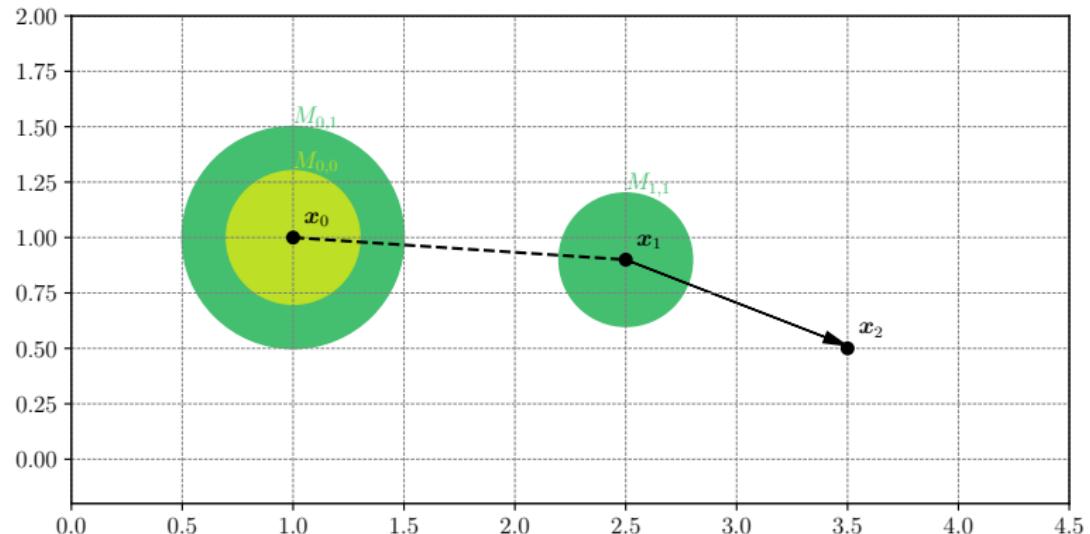
and \mathcal{I} are sets of indices α sampled for each iteration ℓ . In a simple case where we keep all iterations in \mathcal{L}_k , this simplifies to

$$\nabla \mathcal{F}_k =: \sum_{\ell=1}^k \frac{1}{M_{\ell,k}} \sum_{\alpha \in \mathcal{I}_{\ell,k}} (\nabla f(\mathbf{x}_\ell, \boldsymbol{\theta}_\alpha) - \nabla f(\mathbf{x}_{\ell-1}, \boldsymbol{\theta}_\alpha)) + \frac{1}{M_{0,k}} \sum_{\alpha \in \mathcal{I}_{0,k}} \nabla f(\mathbf{x}_0, \boldsymbol{\theta}_\alpha). \quad (198)$$

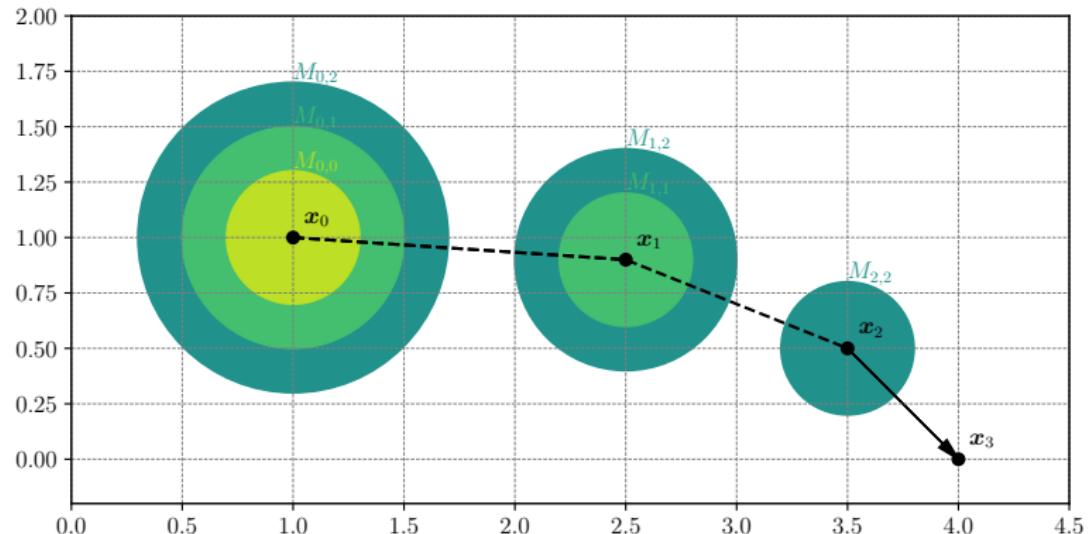
Multi-iteration Stochastic Optimizers (MICE) III



Multi-iteration Stochastic Optimizers (MICE) IV



Multi-iteration Stochastic Optimizers (MICE) V



Multi-iteration Stochastic Optimizers (MICE) VI

To find the optimal sample sizes for each $\ell \in \mathcal{L}_k$, we need the variance of $\Delta_{\ell,k}$,

$$V_{\ell,k}^{(i)} := \mathbf{b} V \left[\Delta_{\ell,k}^{(i)} \middle| \{\mathbf{x}_{\ell'}\}_{\ell' \in \mathcal{L}_k} \right], \quad 1 \leq i \leq d \quad (199)$$

$$V_{\ell,k} := \sum_{i=1}^d V_{\ell,k}^{(i)}. \quad (200)$$

The squared statistical error of the estimator is approximated as

$$(\mathcal{E}_{\text{stat}}(k))^2 \approx \sum_{\ell \in \mathcal{L}_k} \frac{V_{\ell,k}}{M_{\ell,k}}. \quad (201)$$

We want to find all $M_{\ell,k}$ that minimize the total gradient sampling cost while satisfying

$$(\mathcal{E}_{\text{stat}}(k))^2 \leq \epsilon^2 \|\nabla F(\mathbf{x}_k)\|^2. \quad (202)$$

Multi-iteration Stochastic Optimizers (MICE) VII

Then, we calculate the sample-sizes $M_{\ell,k}$ as

$$M_{\ell,k}^* = \left\lceil \frac{1}{\epsilon^2 \|\nabla F(\mathbf{x}_k)\|^2} \left(\sum_{\ell' \in \mathcal{L}_k} \left(V_{\ell',k} (1 + \mathbb{1}_{\underline{\mathcal{L}}_k}(\ell')) \right)^{1/2} \right) \left(\frac{V_{\ell,k}}{(1 + \mathbb{1}_{\underline{\mathcal{L}}_k}(\ell))} \right)^{1/2} \right\rceil, \quad (203)$$

where

$$\mathbb{1}_{\underline{\mathcal{L}}_k}(\ell) := \begin{cases} 0 & \text{if } \ell = \min \mathcal{L}_k \\ 1 & \text{otherwise} \end{cases}. \quad (204)$$

Multi-iteration Stochastic Optimizers (MICE) VIII

Every iteration we add k to the hierarchy, $\mathcal{L}_k \leftarrow \mathcal{L}_{k-1} \cup \{k\}$. To improve the hierarchy, each iteration we if check the following operations would reduce the total gradient sampling cost.

Restart If the cost of perming a step, $\sum_{\ell \in \mathcal{L}_k} M_{\ell,k} - M_{\ell,k-1}$, is larger than the cost of restarting, $\frac{\text{tr}(\Sigma_k)}{\epsilon^2 \|\nabla F(\mathbf{x})\|^2}$, we restart the hierarchy, i.e., we set $\mathcal{L}_k \leftarrow \{k\}$.

Dropping We check if it is advantageous to drop $k-1$ from the hierarchy. If the update cost dropping $k-1$ reduces, we set $\mathcal{L}_k \leftarrow \mathcal{L}_k \setminus \{k-1\}$.

Clipping We check if it is advantageous to clip the hierarchy at some iteration $\ell \in \mathcal{L}_k$, i.e., $\mathcal{L}_k \leftarrow \mathcal{L}_k \setminus \{\ell'\}_{\ell' < \ell}$.

Multi-iteration Stochastic Optimizers (MICE) IX

If we control the statistical error of the gradient with MICE, we can achieve the same linear convergence for the optimality gap for SGD-MICE on L -smooth and strongly-convex problems than the one proven before,

$$\mathbf{b}E[F(\mathbf{x}_k)] - F(\mathbf{x}^*) \leq \frac{L}{2} \left(\frac{\left(\frac{L-\mu}{L+\mu} \right)^2 + \epsilon^2}{1 + \epsilon^2} \right)^k \|\mathbf{x}_0 - \mathbf{x}^*\|^2. \quad (205)$$

Adam - Adaptive moments method I

The Adam algorithm [?] is a stochastic gradient method that keeps memory of the gradients first and second moments (component-wise). For parameters β_1, β_2 , and ε ,

$$\mathbf{m}_k \leftarrow \beta_1 \mathbf{m}_{k-1} + (1 - \beta_1) \nabla f(\mathbf{x}_k, \theta_\alpha) \quad (206)$$

$$\mathbf{v}_k \leftarrow \beta_2 \mathbf{v}_{k-1} + (1 - \beta_2) (\nabla f(\mathbf{x}_k, \theta_\alpha))^2 \quad (207)$$

$$\hat{\mathbf{m}}_k \leftarrow \frac{\mathbf{m}_k}{1 - \beta_1^k} \quad (208)$$

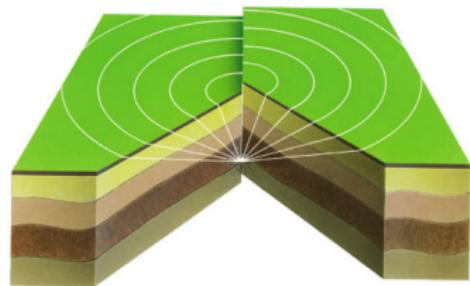
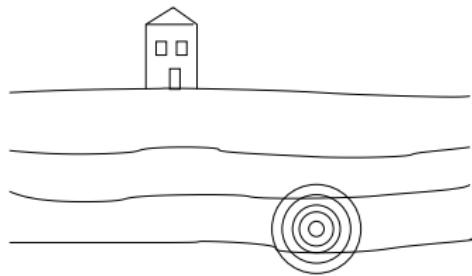
$$\hat{\mathbf{v}}_k \leftarrow \frac{\mathbf{v}_k}{1 - \beta_2^k} \quad (209)$$

$$\mathbf{x}_{k+1} \leftarrow \mathbf{x}_k - \eta \frac{\hat{\mathbf{m}}_k}{\sqrt{\hat{\mathbf{v}}_k} + \varepsilon}. \quad (210)$$

Adam is available for most machine learning libraries, like Tensorflow and PyTorch.

Hyperbolic Problems

Wave equation in a random medium – Motivations in seismology applications



- ▶ Predictions of the effect of an earthquake should account for uncertainty in the underground wave speed
- ▶ Typical situation: stratified medium with discontinuous uncertain wave speed.

Problem setting

Let $(\Omega, \mathcal{F}, \mathbb{P})$ a complete probability space and $D \subset \mathbb{R}^d$ a physical domain.

$$\begin{cases} \partial_{tt} u(t, x, \omega) - \operatorname{div}(a^2(x, \omega) \nabla u(t, x, \omega)) = f(t, x), & \text{in } (0, T] \times D \times \Omega \\ u = 0, & \text{on } (0, T] \times \partial D \times \Omega, \\ u|_{t=0} = g_1, \quad \frac{\partial u}{\partial t} \Big|_{t=0} = g_2, & \text{in } D \times \Omega \end{cases}$$

with $a(x, \omega)$ random field s.t. $0 \leq a_{min} < a(x, \omega) < a_{max} < \infty$.

Technical assumptions: the forcing term f and initial data g_1, g_2 are compactly supported in D ; the final time T and the domain D are such that the wave does not reach the boundary ∂D .

This is equivalent to consider the Cauchy problem in R^d .

Uncertainty model: the medium is possibly made of several layers, each with a random wave speed

- ▶ $\{D_i\}_{i=1}^N$ non overlapping partition of D
- ▶ $a(x, \omega) = \sum_{n=1}^N y_n(\omega) b_n(x) \mathbb{1}_{D_n}(x)$, with b_n smooth
- ▶ $y_n(\omega)$ independent random variables with pdf $\varrho_n : \Gamma_n \rightarrow \mathbb{R}_+$ with Γ_n bounded. W.l.o.g take $\Gamma_n = [-1, 1]$.

Wave equation in a random medium

- ▶ The random wave speed $a(x, \omega)$ is parametrized by N independent random variables $a(x, \omega) = a(x, y_1(\omega), \dots, y_N(\omega))$.
- ▶ Therefore, the solution u is also a function of N random variables

$$u(t, x, \omega) = u(t, x, y_1(\omega), \dots, y_N(\omega))$$

Goal:

- ▶ approximate $u(\cdot, \cdot, \mathbf{y}) \approx u_w(\cdot, \cdot, \mathbf{y})$, $\forall \mathbf{y} \in \Gamma = \prod_{n=1}^N \Gamma_n$.
- ▶ Compute statistics of the solution or Quantities of Interest
 $\mathbb{E}[Q(u)] \approx \mathbb{E}[Q(u_w)]$ (such as: Arias intensity, spectral acceleration, peak ground acceleration, ...)

Crucial point: how smooth is the function $u(\cdot, \cdot, \mathbf{y})$ w.r.t the random parameters $\mathbf{y} = (y_1, \dots, y_N)$?

Example 1: single layer, finite regularity

$$\begin{cases} \partial_{tt} u - y^2 \partial_{xx} u = 0, & \text{in } \mathbb{R}, \ t > 0 \\ u(x, 0) = u_0(x), \quad \frac{\partial u}{\partial t}(x, 0) = 0, & \text{in } \mathbb{R} \end{cases}$$

D'Alambert formula: $u(x, t) = \frac{1}{2}u_0(x - yt) + \frac{1}{2}u_0(x + yt)$

$$\implies \frac{\partial^k u}{\partial y^k}(x, t) = \frac{(-t)^k}{2} \frac{\partial^k u_0}{\partial \xi^k} \Big|_{\xi=x-yt} + \frac{t^k}{2} \frac{\partial^k u_0}{\partial \xi^k} \Big|_{\xi=x+yt}$$

Remarks:

- $u_0(\cdot) \in C^k(\mathbb{R})$ w.r.t. $x \rightarrow u(y)|_{(x,t)} \in C^k(\Gamma)$ w.r.t y

In particular, a discontinuous initial datum implies discontinuous dependence on the parameter (wave speed)

- However, consider a functional $Q(y) = \int_{\mathbb{R}} u(y; x, T) \psi(x) dx$ with $\psi \in C^{2m}(\mathbb{R})$ with compact support. Then $Q(y) \in C^{k+2m}(\Gamma)$

$$\begin{aligned}\frac{\partial Q}{\partial y} &= \int_{\mathbb{R}} \frac{\partial u}{\partial y}(y; x, T) \psi(x) dx \\ &= \int_{\mathbb{R}} \left(-\frac{T}{2} u'_0(x - yT) + \frac{T}{2} u'_0(x + yT) \right) \psi(x) dx \\ &= \int_{\mathbb{R}} \left(\frac{T}{2} u_0(x - yT) - \frac{T}{2} u_0(x + yT) \right) \psi'(x) dx\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 Q}{\partial y^2} &= -T^2 \int_{\mathbb{R}} \left(\frac{1}{2} u'_0(x - yT) + \frac{1}{2} u'_0(x + yT) \right) \psi'(x) dx \\ &= T^2 \int_{\mathbb{R}} u(y; x, T) \psi''(x) dx < \infty\end{aligned}$$

$$\implies \frac{\partial^{k+2m}}{\partial y^{k+2m}} Q = T^{k+2m} \int_{\mathbb{R}} \frac{\partial^k u}{\partial y^k}(y; x, T) \psi^{(2m)}(x) dx < \infty$$

Example 2: single layer, infinite regularity

Assume as before

$$\begin{cases} \partial_{tt} u - y^2 \partial_{xx} u = 0, & \text{in } \mathbb{R}, t > 0 \\ u(x, 0) = u_0(x), \quad \frac{\partial u}{\partial t}(x, 0) = 0, & \text{in } \mathbb{R} \end{cases}$$

with u_0 analytic, vanishing at ∞ with all derivatives

$$\text{Fourier transform : } |\hat{u}_0(k)| \leq C e^{-\alpha|k|}$$

Question: Is $u(y)$ analytic?

We study the problem in the complex domain: $y \in \mathbb{C}$.

Fourier transform the problem

$$\partial_{tt} \hat{u}(t, k) + y^2 k^2 \hat{u}(t, k) = 0, \quad \hat{u}(0, k) = \hat{u}_0(k)$$

$$\text{Then: } \hat{u}(t, k) = \hat{u}_0(k) \frac{e^{iykt} + e^{-iykt}}{2} \implies |\hat{u}(t, k)| \leq |\hat{u}_0(k)| e^{| \operatorname{Im}(y) | |k| t} \leq C e^{-|k|(\alpha - | \operatorname{Im}(y) | t)}$$

$\rightsquigarrow u(t, x) \in H^1(D)$ and satisfies the Cauchy-Riemann conditions only for $t \operatorname{Im}(y) < \alpha$

$\rightsquigarrow u(t, \cdot, y)$ is H^1 -analytic in y in the region

$$\Sigma_t = \{y \in \mathbb{C} : \operatorname{Im}(y) < \alpha/t\}$$

Remarks:

- ▶ u is analytic in $y \in \mathbb{R}$ for all finite times
- ▶ The analyticity region shrinks in time around the real axis
- ▶ The y -dependence of the solution is more and more difficult to approximate for longer times (see [Wang-Karniadakis '06])

To have an analyticity region independent of t we need to require even more regularity to the initial datum

Gevrey regularity $|\hat{u}_0(k)| \leq C e^{-\alpha|k|^q}, \alpha > 0, q > 1.$

Then, $|\hat{u}(k)| \leq C e^{-|k|(\alpha|k|^{q-1} - |\operatorname{Im}(y)|t)}$ and the solution $u(t, \cdot) \in H^1(D)$ is analytic in any strip $\Sigma \equiv \{y \in \mathbb{C}, |\operatorname{Im}(y)| \leq r\}$ for any $r > 0$ and $t > 0$.

Example 2: compare with parabolic equation

Consider instead the parabolic equation

$$\begin{cases} \partial_t u - (D_0 + y^2) \partial_{xx} u = 0, & \text{in } \mathbb{R}, t > 0 \\ u(x, 0) = u_0(x), & \text{in } \mathbb{R} \end{cases}$$

and study the problem in the complex domain: $y \in \mathbb{C}$.

Fourier transform the problem

$$\partial_t \hat{u}(t, k) + (D_0 + y^2)k^2 \hat{u}(t, k) = 0, \quad \hat{u}(0, k) = \hat{u}_0(k)$$

Then:

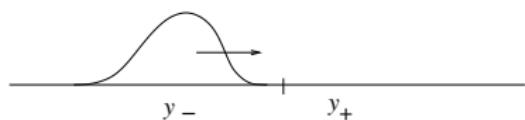
$$\hat{u}(t, k) = \hat{u}_0(k) e^{-(D_0 + y^2)k^2 t} \implies |\hat{u}(t, k)| \leq |\hat{u}_0(k)| e^{-(D_0 + \operatorname{Re}(y)^2 - \operatorname{Im}(y)^2)k^2 t}$$

and the solution $u(t, \cdot) \in H^1(D)$ is analytic in the region

$\Sigma \equiv \{y \in \mathbb{C}, \operatorname{Im}(y)^2 \leq D_0 + \operatorname{Re}(y)^2\}$ for any $t > 0$ and for any $u_0 \in L^2(\mathbb{R})$.

Example 3: two layers, smooth initial datum not crossing the interface

$$\partial_{tt}u - \partial_x(a\partial_x u) = 0, \quad \text{in } \mathbb{R}, \quad t > 0, \quad a(x, y) = \begin{cases} y_-, & x < 0 \\ y_+, & x > 0 \end{cases}$$



- ▶ single discontinuity interface
- ▶ right going initial wave:
given $g(x) \in C_0^\infty(0, \infty)$,

$$\begin{cases} u(0, x) = g(-x) \\ u_t(0, x) = y_- g'(-x) \end{cases}$$

Solution:

$$u(x, t) = \begin{cases} g(y_- t - x) + \frac{y_- - y_+}{y_- + y_+} g(y_+ t + x), & x < 0 \\ \frac{2y_-}{y_- + y_+} g\left(\frac{y_-}{y_+}(y_+ t - x)\right), & x > 0 \end{cases}$$

The solution is smooth w.r.t. $y_-, y_+ > 0$ since the initial datum is smooth and does not cross the interface

On the regularity of the solution

Consider the solution u as a parametric Banach valued function

$$u(\mathbf{y}) : \Gamma \rightarrow L^2(0, T; H_0^1(D)) \text{ or } u(\mathbf{y}) : \Gamma \rightarrow L^2([0, T] \times D)$$

- ▶ In general $u(\mathbf{y})$ is **not analytic** (contrary to elliptic and parabolic cases)
- ▶ In general $u(\mathbf{y})$ will have only finite regularity depending on the spatial smoothness of the initial and forcing data
- ▶ The smoothness of **functionals** $Q(u)$ can be considerably higher than the smoothness of u itself.
- ▶ Smoothness in the physical space and parameter space are strongly entangled.

Regularity results for smooth (single layered) coefficients

We first recall a result on the **space regularity** of the solution:

Shift Theorem (see e.g. [Hörmander 1994])

Assuming $a(\cdot, \mathbf{y}) \in C^\infty(D)$ a.s. and D large enough so that the wave never reaches the boundary in $t \in [0, T]$, if

$$f \in L^2((0, T), H^s(D)), \quad g_1 \in H^{s+1}(D), \quad g_2 \in H^s(D), \quad s \geq -1, \quad (211)$$

Then for $s \geq 0$

$$u(\cdot, \cdot, \mathbf{y}) \in C^0([0, T], H^{s+1}(D)) \cap C^1([0, T], H^s(D)) \cap H^2((0, T), H^{s-1}(D)),$$

and for $s = -1$

$$u(\cdot, \cdot, \mathbf{y}) \in C^0([0, T], H^{s+1}(D)) \cap C^1([0, T], H^s(D)).$$

Regularity results for smooth (single layered) coefficients

Concerning the regularity w.r.t. \mathbf{y} the following holds

Theorem [Motamed-Nobile-Tempone, '13, '14]

Under the assumptions of the previous theorem, for any $\mathbf{k} = (k_1, \dots, k_N)$,
 $0 \leq |\mathbf{k}| \leq s + 1$

$$\partial_{\mathbf{y}}^{\mathbf{k}} u(\cdot, \mathbf{y}) := \frac{\partial^{|\mathbf{k}|} u(\cdot, \mathbf{y})}{\partial^{k_1} y_1 \cdots \partial^{k_N} y_N} \in C^0([0, T]; H^{s+1-|\mathbf{k}|}(D))$$

Proof: By assumption, $f \in L^2((0, T); H^s(D))$ and $u(\cdot, \mathbf{y}) \in C^0([0, T], H^{s+1}(D))$ for any \mathbf{y} . We now differentiate the equation w.r.t. one variable y_n .

$$\partial_{tt} \partial_{y_n} u - \operatorname{div}(\mathbf{a}^2 \nabla \partial_{y_n} u) = \underbrace{\operatorname{div}(\mathbf{2a} \partial_{y_n} \mathbf{a} \nabla u)}_{\in L^2((0, T); H^{s-1})}, \quad \text{in } (0, T] \times D \times \Omega$$

and $\partial_{y_n} u|_{t=0} = \partial_t \partial_{y_n} u|_{t=0} = 0$.

Hence $\partial_{y_n} u$ satisfies the same problem as u but with a less regular forcing term $\tilde{f} \in L^2((0, T); H^{s-1}(D))$ and zero initial conditions.

Therefore $\partial_{y_n} u \in C^0([0, T], H^s(D))$.

Iterating the argument leads to the final result.

Regularity of linear quantities of interest

Consider the Q.o.I

$$Q(\mathbf{y}) = \int_0^T \int_D u(t, x, \mathbf{y}) \phi(t, x) dx dt + \int_D u(T, x, \mathbf{y}) \psi(x) dx$$

and assume

$$\phi \in L^2((0, T); H^\ell(D)), \quad \psi \in H^\ell(D) \quad (212)$$

complactly supported in D .

Then

Theorem [Motamed-Nobile-Tempone, '13, '14]

Under assumptions (211) and (212)

$$\partial_{\mathbf{y}}^{\mathbf{k}} Q \in L^\infty(\Gamma), \quad \forall \mathbf{k} = (k_1, \dots, k_N), \quad |\mathbf{k}| \leq s + \ell + 1$$

- ▶ The quantity of interest Q can be considerably smoother in \mathbf{y} than the solution itself if the functions ψ and ϕ are smooth.

Regularity of linear quantities of interest

Hint of the proof:

From the definition $Q(\mathbf{y}) = \int_0^T \int_D u(t, x, \mathbf{y}) \phi(t, x) dx dt + \int_D u(T, x, \mathbf{y}) \psi(x) dx$ we see that Q has at least $s + 1$ bounded derivatives.

Introduce the adjoint problem

$$\begin{cases} \partial_{tt}\varphi(t, x, \mathbf{y}) - \operatorname{div}(a^2(x, \mathbf{y}) \nabla \varphi(t, x, \mathbf{y})) = \phi(t, x), & \text{in } D \times \Omega, \quad t < T \\ \varphi|_{\partial D} = 0, \quad \varphi|_{t=T} = 0, \quad \partial_t \varphi|_{t=T} = -\psi \end{cases}$$

From the previous theorem, the solution φ is $\ell + 1$ times differentiable in $L^2(D)$ w.r.t. \mathbf{y} .

Then the Q.o.I can be equivalently expressed as

$$\begin{aligned} Q(\mathbf{y}) &= \int_0^T \int_D u(\mathbf{y}) (\partial_{tt}\varphi(\mathbf{y}) - \operatorname{div}(a^2(\mathbf{y}) \nabla \varphi(\mathbf{y}))) + \int_D u(\mathbf{y})|_{t=T} \psi \\ &= \int_0^T \int_D (-\partial_t u(\mathbf{y}) \partial_t \varphi(\mathbf{y}) + a^2(\mathbf{y}) \nabla u(\mathbf{y}) \cdot \nabla \varphi(\mathbf{y})) - \int_D g_1 \partial_t \varphi(\mathbf{y})|_{t=0} \\ &= \int_0^T \int_D \varphi(t, x, \mathbf{y}) f(t, x) dx dt - \int_D g_1(x) \partial_t \varphi(0, x, \mathbf{y}) dx + \int_D g_2(x) \varphi(0, x, \mathbf{y}) dx \end{aligned}$$

and we see that Q has at least $\ell + 1$ bounded derivative in \mathbf{y} .

A finer argument allows to say that indeed Q has $s + \ell + 1$ bounded derivatives.

General result for layered materials

- ▶ Wave speed in the form $a(x, \mathbf{y}(\omega)) = \sum_{n=1}^N y_n(\omega) b_n(x) \mathbb{1}_{D_n}(x)$ with $\{D_n\}_{n=1}^N$ non-overlapping partition of the domain **with smooth interfaces** and b_n smooth.
- ▶ minimal regularity assumptions on the data:

$$f \in L^2([0, T] \times D), \quad g_1 \in H_0^1(D), \quad g_2 \in L^2(D)$$

Theorem [Motamed-Nobile-Tempone '13]

The solution has **one bounded derivative** $\partial_{y_n} u \in L^2(0, T; L^2(D))$.

The solution will be **y-smoother** for smoother data **not intersecting any interface between the layers**.

Similarly, for a Quantity of interest

$$Q(\mathbf{y}) = \int_0^T \int_D u(t, x, \mathbf{y}) \phi(t, x) dx dt + \int_D u(T, x, \mathbf{y}) \psi(x) dx$$

if the functions ϕ, ψ are in $H^\ell(D)$ and are linear combinations of functions compactly supported in one layer (i.e. ψ and ϕ do not cross any interface), then

Q will have $\ell + 1$ bounded derivatives in \mathbf{y}

Stochastic collocation

We approximate the function $u(\mathbf{y})$ by stochastic collocation.

- ▶ Let $i \geq 1$ be the **level** of interpolation and $m(i) = 2^{i-1} + 1$ the number of interpolation points at level i .
- ▶ For each random variable consider **Gauss points** w.r.t. the weight ρ_n
- ▶ Take a sequence of 1D polynomial interpolant operators
 $\mathcal{U}_n^{m(i)} : C^0(\Gamma_n) \rightarrow \mathbb{P}_{m(i)-1}(\Gamma_n)$ with increasing number of points

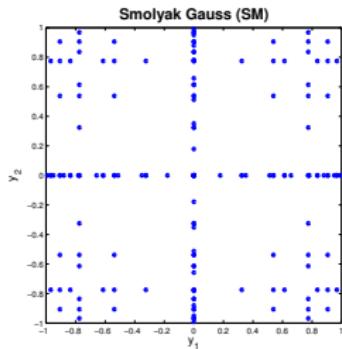
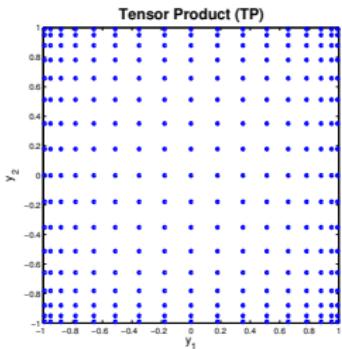
The i -th interpolant uses $m(i)$ abscissae $\vartheta_n^i = \left\{ y_{n,1}^{(i)}, \dots, y_{n,m_i}^{(i)} \right\}$.

Tensor grid approximation (isotropic)

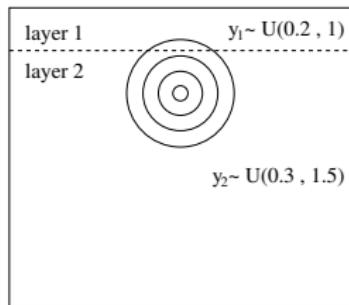
$$u_w^{TP}(t, x, \cdot) = \mathcal{U}_1^{m(w)} \otimes \cdots \otimes \mathcal{U}_N^{m(w)} u(t, x, \cdot)$$

Sparse grid approximation (isotropic)

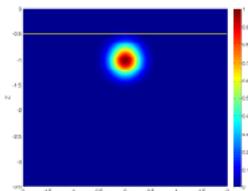
$$u_w^{SG}(t, x, \cdot) = \sum_{\sum_{n=1}^N (i_n - 1) \leq w} \bigotimes_{n=1}^N \left(\mathcal{U}_n^{m(i_n)} - \mathcal{U}_n^{m(i_n - 1)} \right) u(t, x, \cdot)$$



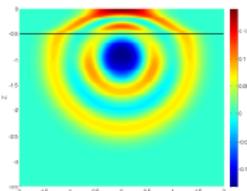
Numerical test – rough case



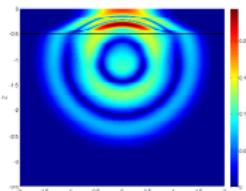
- ▶ wave equation in two-layers medium
- ▶ random wave speed in each layer
- ▶ smooth initial displacement crossing the interface
- ▶ approximating the solution u by Stochastic Collocation



initial solution

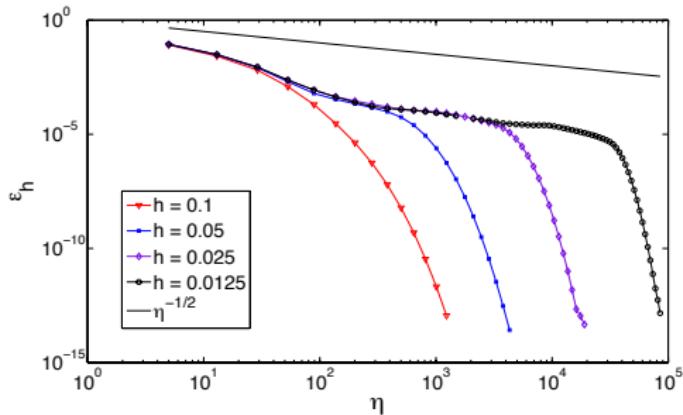


$\mathbb{E}[u](x, t = 1)$



$\text{std}[u](x, t = 1)$

- ▶ Isotropic Smolyak grid approximation on Gauss-Legendre abscissae
- ▶ Finite difference approximation in space + leapfrog in time
- ▶ Maximum error in the expected value at $t=1$ versus # of collocation points η .



- ▶ In the limit $h \rightarrow 0$ the discrete solution has low y -regularity \rightsquigarrow slow algebraic convergence
- ▶ When $hp \gg 1$ (where p is the max. polynomial degree used in each variable y_n), the convergence is exponential.

Smoothness of the space-time discretized solution

Assume $u(\mathbf{y})$ has low regularity in \mathbf{y}

Consider a finite difference (or finite element) approximation of the equation in space with discretization parameter h .

- The discrete solution $u_h(t, x, \mathbf{y})$ is always analytic with respect to \mathbf{y} for all $\mathbf{y} \in [-1, 1]^N$ (even in the case $a_{min} = 0$).

The Taylor series converges for any $\mathbf{y} \in [-1, 1]^N$ with radius

$$r \leq \begin{cases} \frac{1}{N} \frac{h\tilde{a}_{min}}{h\tilde{a}_{min} + Ct} & \tilde{a}_{min} = \min\{a_{min}, 1\}, \text{ with } a_{min} > 0 \\ \frac{1}{N} \frac{h^2}{h^2 + Ct^2} & \text{with } a_{min} = 0 \end{cases}$$

with C independent of \mathbf{y} .

Should we say more here?

- ▶ derivation of the analiticity result?
- ▶ convergence result?

or, should we avoit this discussion at all (regularity of the discrete solution and role of a_{min})?

Smoothness of the space-time discretized solution

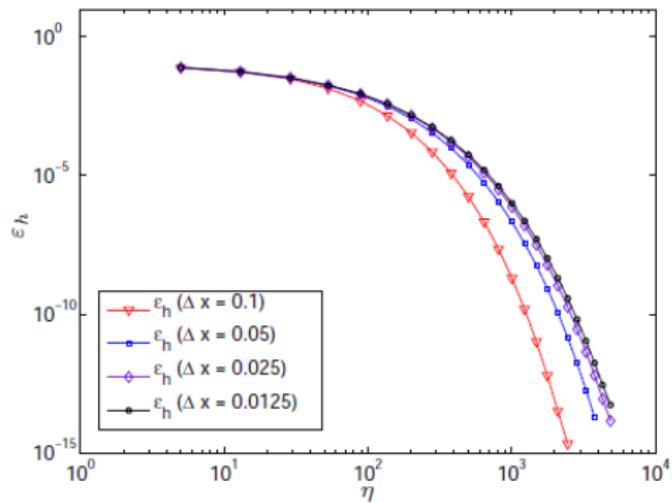
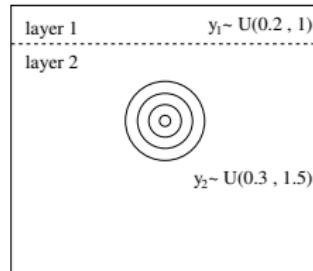
Consider a tensor product polynomial approximation if \mathbf{y} of degree p ($= m(w) - 1$). We have two regimes

- ▶ for $h^\alpha p \gg Ct$ \rightsquigarrow exponential convergence in p
- ▶ for $h^\alpha p \ll Ct$ \rightsquigarrow algebraic slow convergence in p due to small regularity of u w.r.t. \mathbf{y} .

with $\alpha = 1, 2$ depending on a_{min} .

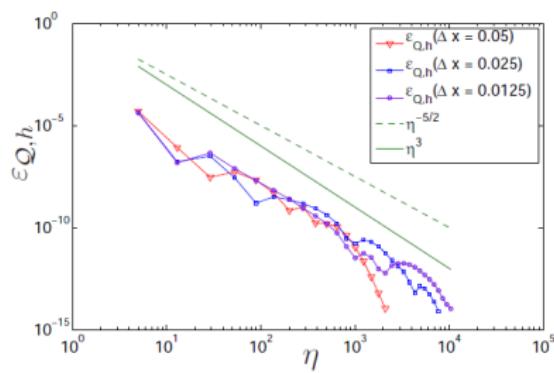
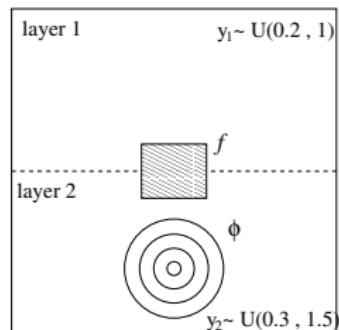
Numerical test – smooth case

- The initial displacement is smooth and confined in the second layer
- Maximum error in the expected value at $t=1$ versus # of collocation points \tilde{M} .
- Approx. of the solution itself.



Numerical test – rough solution; smooth Q.o.I.

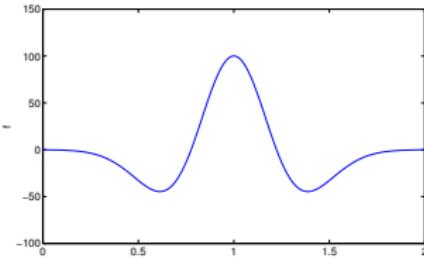
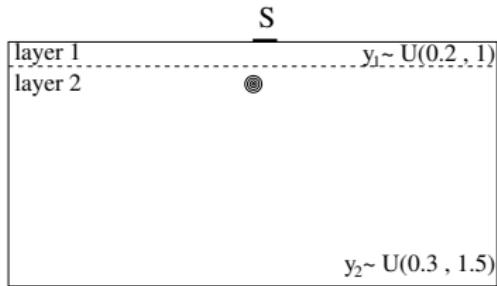
- ▶ Forcing term discontinuous and across the interface
- ▶ Q.o.I $Q(\mathbf{y}) = \int_0^1 \int_D u(t, x, \mathbf{y}) \phi(x) dx dt$ with $\phi \in C_0^\infty(D)$ only in one layer
- ▶ Error in expected value of Q versus # of collocation points \tilde{M} .



Arias intensity

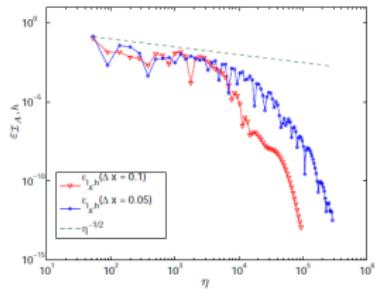
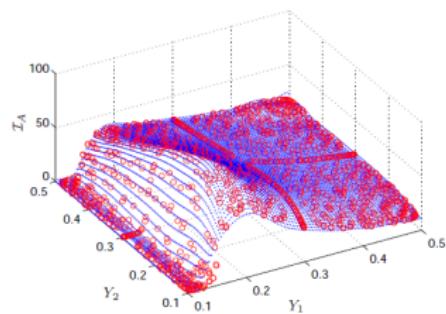
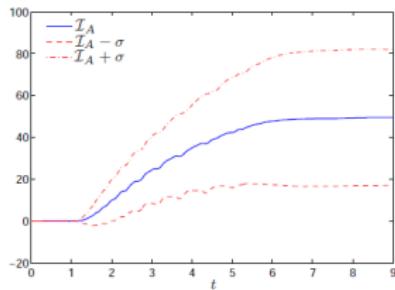
- ▶ Zero initial conditions
- ▶ Forcing term: Ricker wavelet
- ▶ Q.o.I: **Arias intensity** (surface acceleration)

$$\mathcal{I}_A(\mathbf{y}) = \int_0^T \int_S |\partial_{tt} u(t, x, \mathbf{y})|^2 dx dt$$



Ricker wavelet

Arias intensity



- ▶ the solution has low regularity (forcing term discontinuous in space)
- ▶ The functional is non-linear \rightsquigarrow the dual solution has also low regularity
- ▶ slow convergence rate on $Q(\mathbf{y})$

Fatigue life prediction

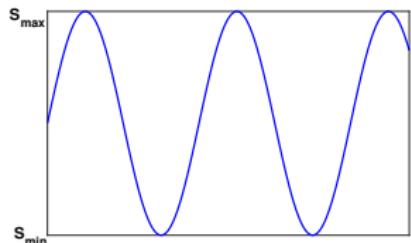
Fatigue is the cumulative damage and eventual failure of mechanical components subjected to cyclic loading.



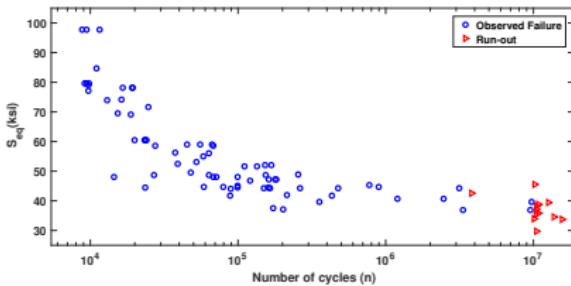
4-28-1988 After 89,090 flight cycles on a 737-200, metal fatigue lets the top go in flight.

Source: <http://anengineersaspect.blogspot.com/2010/04/the-22nd-anniversary-of-aloha-airlines.html>

Fatigue life prediction



Provide a systematic approach to calibrate and rank stress-life (S-N) models given a collection of records of fatigue experiments on unnotched dog bone specimens.



Fatigue data

- ▶ Data are available from fatigue experiments applied to specimens of 75S-T6 aluminum alloys [Grover et al., 1951]. The following data are recorded for each specimen:
 - ▶ the maximum stress, S_{max} , measured in ksi units.
 - ▶ the cycle ratio, R , defined as the minimum to maximum stress ratio.
 - ▶ the fatigue life, N , defined as the number of load cycles at which fatigue failure occurred.
 - ▶ a binary variable (0/1) to denote whether failure occurred or not (run-out).
- ▶ Fatigue data obtained for particular test ratios need to be generalized. To this purpose, we define the *equivalent stress*

$$S_{eq}^{(q)} = S_{max} (1 - R)^q,$$

where q is a fitting parameter [Walker, 1970].

- ▶ We consider fatigue-limit models and random fatigue-limit models with special treatment of the run-outs (right-censored data).

Model 1a

- The *fatigue life* N is modeled with a lognormal distribution [Pascual and Meeker, 1997]:

$$\log_{10}(N) \sim \mathcal{N}(\mu(S_{eq}), \tau)$$

where

$$\mu(S_{eq}) = \begin{cases} A_1 + A_2 \log_{10}(S_{eq} - A_3), & \text{if } S_{eq} > A_3 \\ +\infty, & \text{otherwise} \end{cases} \quad (213)$$

- The probability of survival after n cycles is

$$P(N > n) = 1 - \Phi \left(\frac{\log_{10}(n) - \mu(S_{eq})}{\tau} \right),$$

where Φ is the cumulative distribution function (cdf) of the standard normal distribution.

Model Ia

- Given the data $\mathbf{n} = (n_1, \dots, n_m)$, the *likelihood function* is given by

$$\prod_{i=1}^m \left[\frac{g(\log_{10}(n_i); \mu(S_{eq}), \tau)}{n_i \log(10)} \right]^{\delta_i} \left[1 - \Phi \left(\frac{\log_{10}(n_i) - \mu(S_{eq})}{\tau} \right) \right]^{1-\delta_i}$$

where $g(t; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{(t-\mu)^2}{2\sigma^2} \right\}$, and

$$\delta_i = \begin{cases} 1 & n_i \text{ is a failure} \\ 0 & n_i \text{ is a run-out.} \end{cases}$$

Model Ia

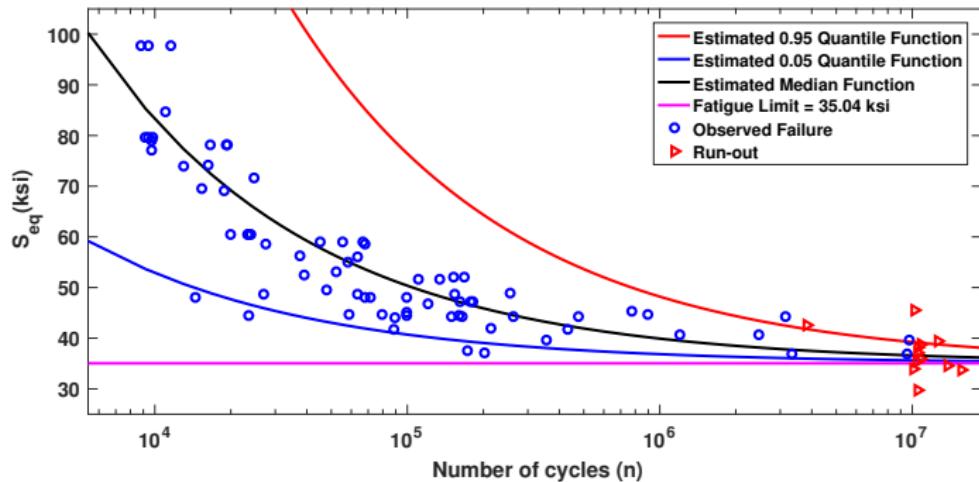


Figure: MLE: $A_1 = 7.38$, $A_2 = -2.01$, $A_3 = 35.04$, $q = 0.563$, $\tau = 0.527$.

Model Ib

We extend the **Model Ia** by allowing a non-constant standard deviation:

- ▶ $\mu(S_{eq}) = A_1 + A_2 \log_{10}(S_{eq} - A_3)$, if $S_{eq} > A_3$
- ▶ $\sigma(S_{eq}) = 10^{(B_1 + B_2 \log_{10}(S_{eq}))}$, if $S_{eq} > A_3$.

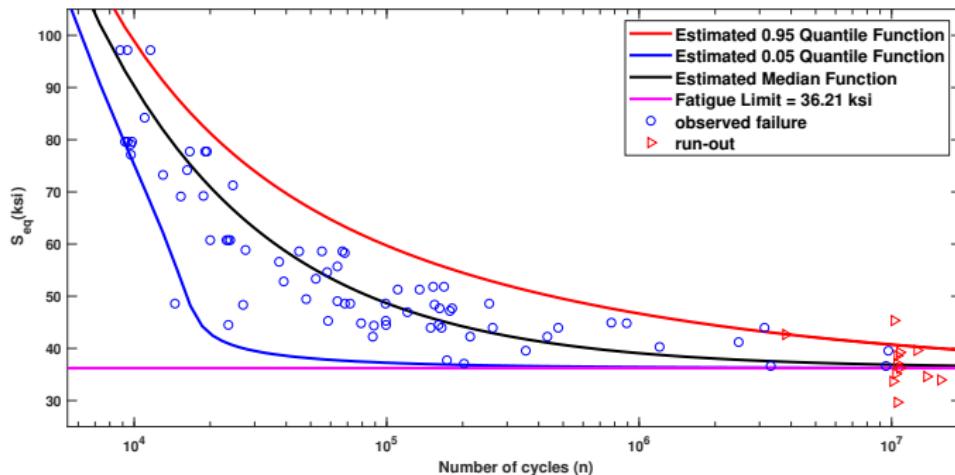


Figure: $A_1 = 6.72$, $A_2 = -1.57$, $A_3 = 36.21$, $q = 0.551$, $B_1 = 4.55$, $B_2 = -2.89$.

Model II |

We extend **Model Ia** to allow a random fatigue-limit parameter [Pascual and Meeker, 1999]:

- ▶ the pdf of $\log_{10}(A_3)$ is $\phi(t; \mu_f, \sigma_f)$,
- ▶ the pdf of $\log_{10}(N) | A_3$ is $\phi(t; \mu(S_{eq}), \tau)$,
- ▶ where $\phi(t; \mu, \sigma) = \frac{1}{\sigma} \exp \left\{ \left(\frac{t-\mu}{\sigma} \right) - \exp \left(\frac{t-\mu}{\sigma} \right) \right\}$ is the smallest extreme value (sev) pdf.
with location parameter μ and scale parameter σ .
- ▶ the pdf of $\log_{10}(N)$ is obtained by marginalizing A_3 :

$$f_{\log_{10}(N)}(u; \theta) = \int_{-\infty}^{\log_{10}(S_{eq})} \phi(u; \mu(S_{eq}), \tau) \phi_{\log_{10}(A_3)}(w; \mu_f, \sigma_f) dw,$$

where $\theta = (A_1, A_2, \mu_f, \sigma_f, q, \tau)$.

- ▶ Similarly, the cdf of $\log_{10}(N)$ is given by

$$F_{\log_{10}(N)}(u; \theta) = \int_{-\infty}^{\log_{10}(S_{eq})} \Phi \left(\frac{u - \mu(S_{eq})}{\tau} \right) \phi_{\log_{10}(A_3)}(w; \mu_f, \sigma_f) dw,$$

where Φ is the conditional cdf of $\log_{10}(N)$ given A_3 .

The functions $f_{\log_{10}(N)}$ and $F_{\log_{10}(N)}$ no longer have closed forms and must be numerically evaluated.

Model II

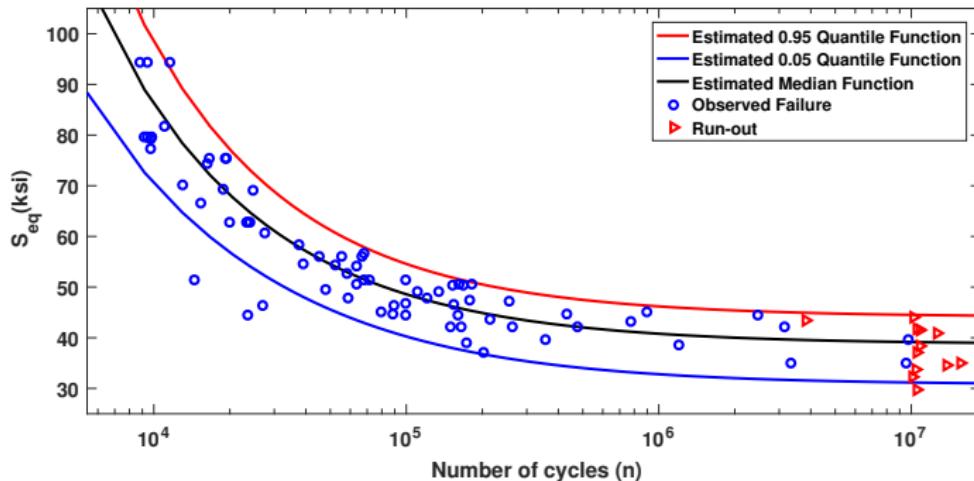


Figure: MLE:

$$A_1 = 6.51, A_2 = -1.47, \mu_f = 1.60, \sigma_f = 0.039, q = 0.489, \tau = 0.09 .$$

Unlike fatigue-limit models, the random fatigue-limit model has the property that each estimated quantile approaches a different horizontal asymptote.

Bayesian Inference

Assume θ is random with a *prior* density $\rho(\theta)$, then the conditional density of θ given \mathbf{Y}

$$\rho(\theta|\mathbf{Y}) = \frac{\mathcal{L}(\theta)\rho(\theta)}{\int_{\Theta} \mathcal{L}(\theta)\rho(\theta)d\theta} \quad (214)$$

called the *posterior* density is the basis for making Bayesian inference.

1. Prior probability density $\rho(\theta)$ that reflects prior experience,
2. Likelihood function that connects the observations and θ ,
3. Methods to explore the posterior probability density.

Markov Chain Monte Carlo (MCMC)

- ▶ We would like to estimate a complicated posterior density (target) $\rho(\theta|\mathbf{Y})$.
- ▶ MCMC methods generate a correlated sequence of samples $\{\theta_t\}_{t=1}^N$ using a proposal density $q(\theta_t|\theta_{t-1})$ that is simpler to sample from.
- ▶ Under certain conditions, the stationary distribution of the sampled sequence is the target posterior $\rho(\theta|\mathbf{Y})$ [Robert and Casella, 2009].

Algorithm Random walk Metropolis-Hastings algorithm

- 1: **set** an initial value for the chain: $\theta_c = \theta_0$ and **choose** δ
- 2: **draw** θ_p from $N(\theta_c, diag(\delta))$
- 3: **let** $H = \min\{1, \frac{\mathcal{L}(\theta_p)\rho(\theta_p)}{\mathcal{L}(\theta_c)\rho(\theta_c)}\}$ and **draw** r from $U(0, 1)$
- 4: **if** $H > r$ **then**
- 5: $\theta_c = \theta_p$
- 6: **end if**
- 7: **repeat** steps (2 to 4) **until** L samples are attained.

- ▶ Kernel density estimation (KDE) can be used to approximate the full target density.

Laplace Approximation

- If the posterior distribution of θ is unimodal, then it can be approximated by a Gaussian distribution

$$\rho(\theta|\mathbf{Y}) \approx \frac{1}{\sqrt{(2\pi)^k |\Sigma(\hat{\theta})|}} \exp \left\{ -\frac{1}{2} (\theta - \hat{\theta})' \Sigma(\hat{\theta})^{-1} (\theta - \hat{\theta}) \right\}$$

where $\hat{\theta}$ is the maximum a posteriori (MAP) probability estimate of θ and $\Sigma(\hat{\theta})$ is the inverse Hessian matrix of the negative log posterior evaluated at $\hat{\theta}$.

- Laplace method uses the Gaussian approximation to estimate integrals of the posterior distribution.

Bayesian approach – Model Ia I

Model I: $A_1 \sim U(5, 13)$, $A_2 \sim U(-5, 0)$, $A_3 \sim U(24, 40)$,
 $q \sim U(0.1, 0.9)$, $\tau \sim U(0.1, 1.5)$.

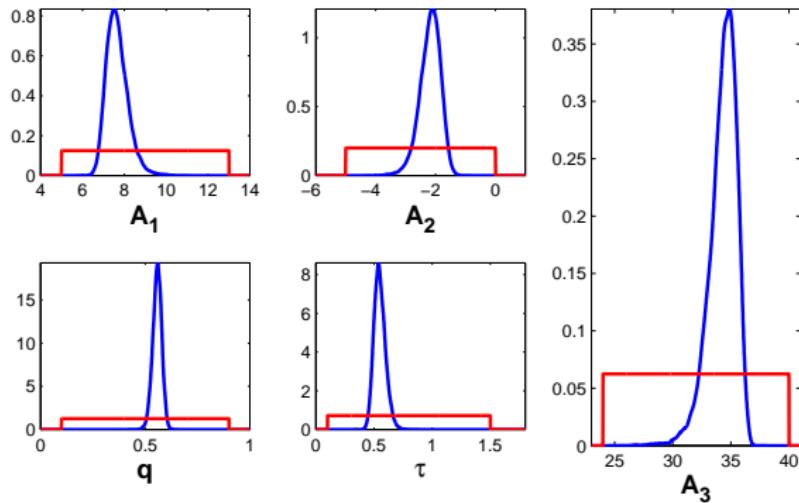


Figure: Prior densities (red line) and approximate marginal posterior densities (blue line) for A_1, A_2, q, τ and A_3 .

Bayesian approach – Model Ia II

Table: Correlation coefficients between each pair of parameters in Model Ia.

	A_1	A_2	A_3	q
A_2	-0.986	—	—	—
A_3	-0.908	0.863	—	—
q	-0.447	0.430	0.448	—
τ	-0.017	-0.018	0.018	0.060

Bayesian approach – Model Ib I

Model Ib: $A_1 \sim U(4, 10)$, $A_2 \sim U(-4, 0)$, $A_3 \sim U(30, 40)$,
 $q \sim U(0.1, 0.9)$, $B_1 \sim U(2, 7)$, $B_2 \sim U(-5, -1)$.

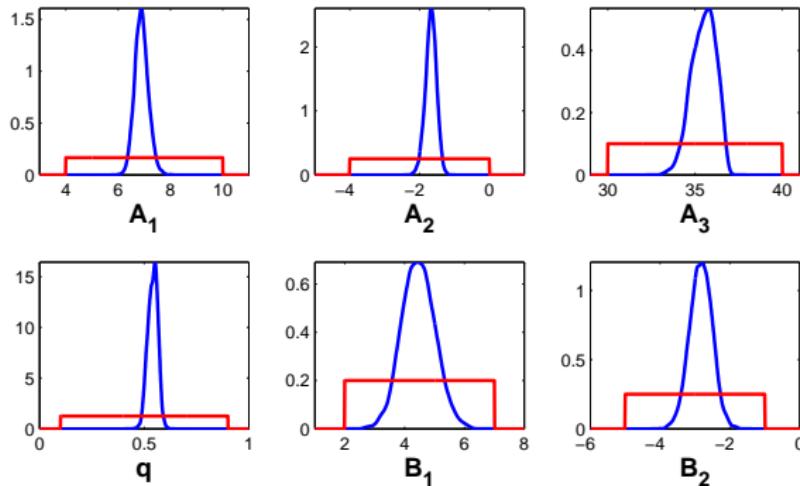


Figure: Prior densities (red line) and approximate marginal posterior densities (blue line) for A_1, A_2, q, B_1, B_2 and A_3 .

Bayesian approach – Model Ib II

Table: Correlation coefficients between each pair of parameters in Model Ib.

	A_1	A_2	A_3	q	B_1
A_2	-0.995	—	—	—	—
A_3	-0.671	0.653	—	—	—
q	-0.428	0.436	0.664	—	—
B_1	-0.272	0.279	0.001	-0.226	—
B_2	0.271	-0.278	0.004	0.235	-0.998

Bayesian approach – Model III

Model II: $A_1 \sim U(4, 10)$, $A_2 \sim U(-4, 0)$, $\mu_f \sim U(1.4, 1.8)$,
 $\sigma_f \sim U(0, 0.1)$, $q \sim U(0.1, 0.9)$, $\tau \sim U(0, 0.25)$.

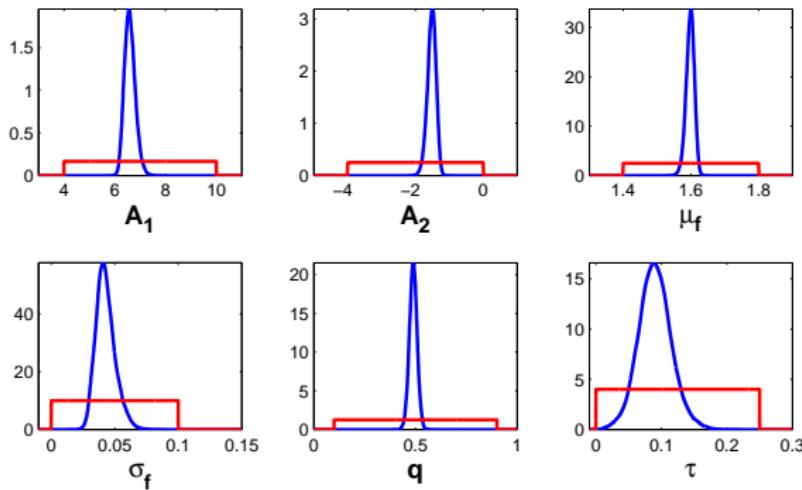


Figure: Prior densities (red line) and approximate marginal posterior densities (blue line) for A_1, A_2, q, τ, μ_f and σ_f .

Bayesian approach – Model II II

Table: Correlation coefficients between each pair of parameters in Model II.

	A_1	A_2	μ_f	σ_f	q
A_2	-0.987	—	—	—	—
μ_f	-0.788	0.722	—	—	—
σ_f	0.494	-0.452	-0.566	—	—
q	-0.067	0.111	-0.037	-0.158	—
τ	-0.019	0.031	0.080	-0.412	0.326

Model comparison

- ▶ **Akaike information criterion** $AIC = 2(k - \log \hat{\mathcal{L}})$, where $\hat{\mathcal{L}}$ is the maximum value of the likelihood function and k is the number of the parameters.
- ▶ **Bayes factor** of Model A against Model B:

$$F_{B,A} := \frac{\int \mathcal{L}_B(\theta_B) \rho_B(\theta_B) d\theta_B}{\int \mathcal{L}_A(\theta_A) \rho_A(\theta_A) d\theta_A},$$

- ▶ **Log pointwise predictive density**

$$lppd = \sum_{i=1}^m \log \left(\frac{1}{L} \sum_{s=1}^L \rho(Y_i | \theta_s) \right),$$

where $\{\theta_s\}_{s=1}^L$ are MCMC posterior samples of θ .

- ▶ **Deviance information criterion** $DIC = 2(p_{DIC} - \log \mathcal{L}(\bar{\theta}))$, where $\bar{\theta}$ is the posterior mean and

$$p_{DIC} = 2 \left(\log \mathcal{L}(\bar{\theta}) - \frac{1}{L} \sum_{s=1}^L \log \mathcal{L}(\theta_s) \right).$$

Model comparison

► K-fold cross-validation

Assume that the data are randomly partitioned into K disjoint subsets, $\{\mathbf{Y}_k\}_{k=1}^K$. Then, define

$\{\mathbf{Y}_{(-k)}\} = \{\mathbf{Y}_1, \dots, \mathbf{Y}_{k-1}, \mathbf{Y}_{k+1}, \dots, \mathbf{Y}_K\}$ to be a training set. For each training set, compute the corresponding posterior distribution, $p(\theta|\mathbf{Y}_{(-k)})$. Then, the log predictive density for $Y_i \in \mathbf{Y}_k$ is computed using the training set $\{\mathbf{Y}_{(-k)}\}$, that is:

$$lpd_i = \log \left(\frac{1}{L} \sum_{s=1}^L \rho(Y_i | \theta_{k,s}) \right),$$

where $\{\theta_{k,s}\}_{s=1}^L$ are the MCMC samples of the posterior $p(\theta|\mathbf{Y}_{(-k)})$. Finally, the expected log predictive density (elpd):

$$\text{elpd} = \sum_{i=1}^m lpd_i.$$

A model with a smaller absolute value criterion is a better (predictive) model.

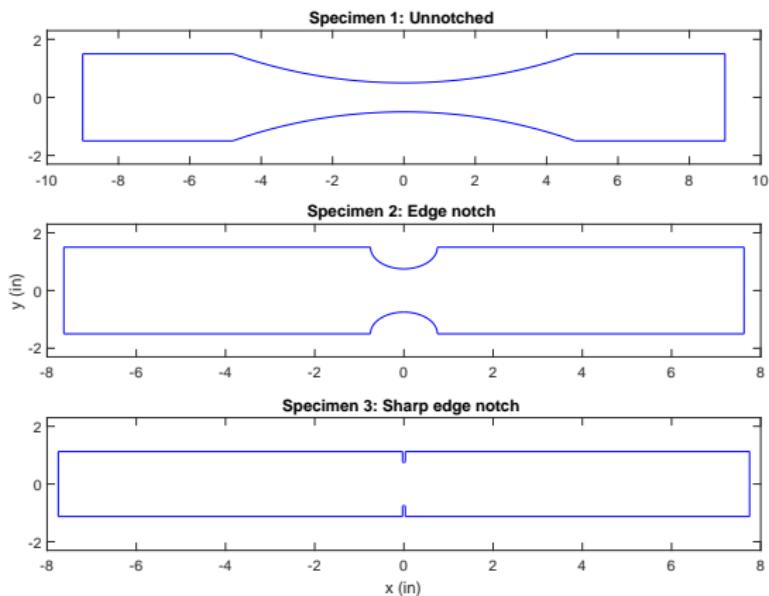
Model comparison

classical information criteria	Model Ia	Model Ib	Model II
maximum log-likelihood	-950	-920	-907
Akaike Information Criterion (AIC)	1910	1853	1827
Bayesian Information Criterion (BIC)	1923	1868	1841
AIC with correction	1911	1854	1828
Bayesian/predictive criteria	Model Ia	Model Ib	Model II
Log marginal likelihood (Laplace)	-963	-940	-924
Log marginal likelihood (Laplace-Metropolis)	-964	-938	-924
Log pointwise predictive density (lpd)	-950	-921	-908
Deviance information criterion (DIC)	1910	1852	1827
Watanabe-Akaike information criterion (WAIC)	1912	1854	1826
5-fold cross-validation (elpd)	-956	-928	-914

Model II is better than **Model Ia** and **Model Ib** under all types of information criteria and predictive criteria.

Fatigue life prediction: Notched specimens

We generalize the use of S-N models to notched specimens where the stress is not uniform. The spatial Poisson processes combines the S-N models with the (averaged) effective stress, $\sigma_{\text{eff}}^{\Delta}(x)$, which is computed after solving the linear elasticity equations using finite element methods.



Numerical computation of the stress field I

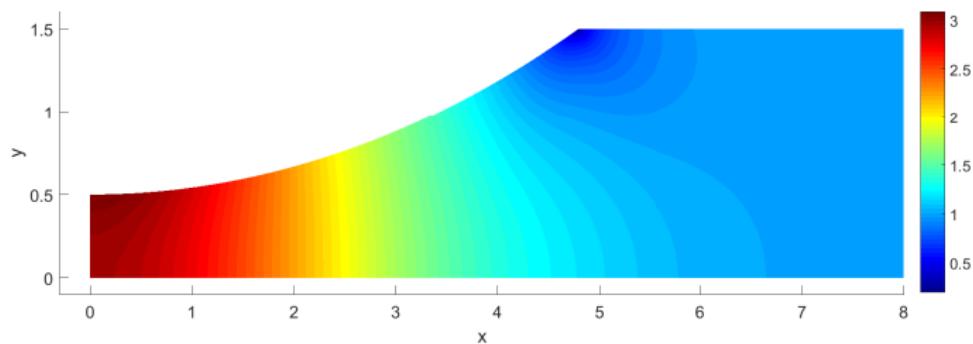


Figure: Distribution of σ_x for the unnotched specimen (Specimen 1).

Numerical computation of the stress field II

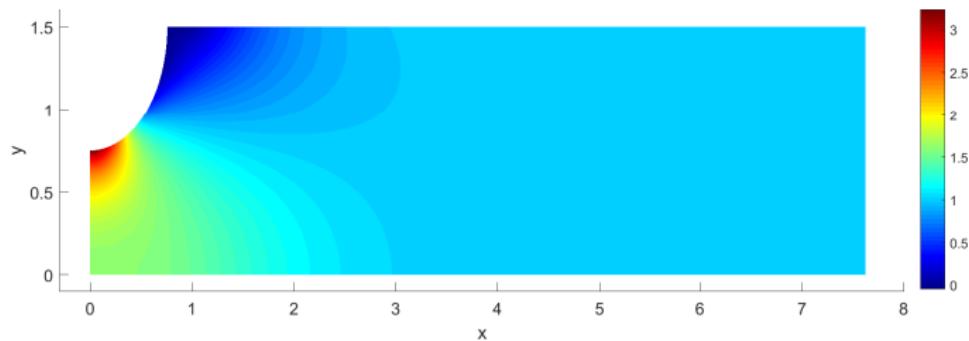


Figure: Distribution of σ_x for the notched specimen (Specimen 2).

Numerical computation of the stress field III

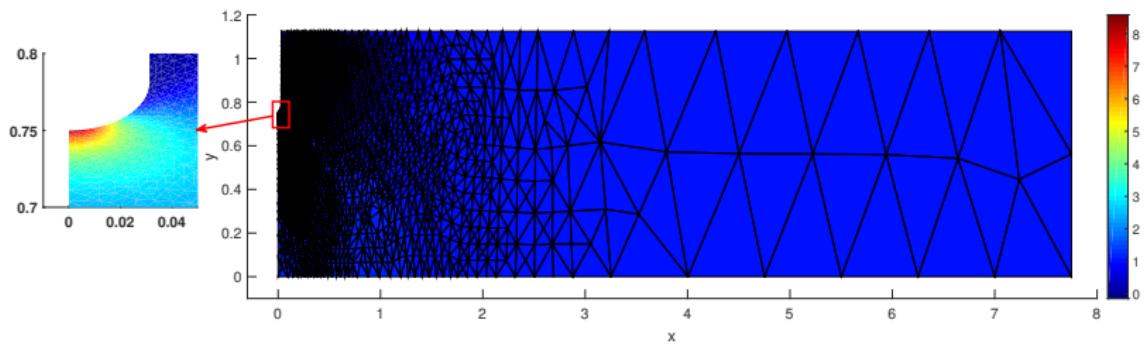


Figure: Distribution of σ_x for the notched specimen (Specimen 3).

Numerical computation of the stress field IV

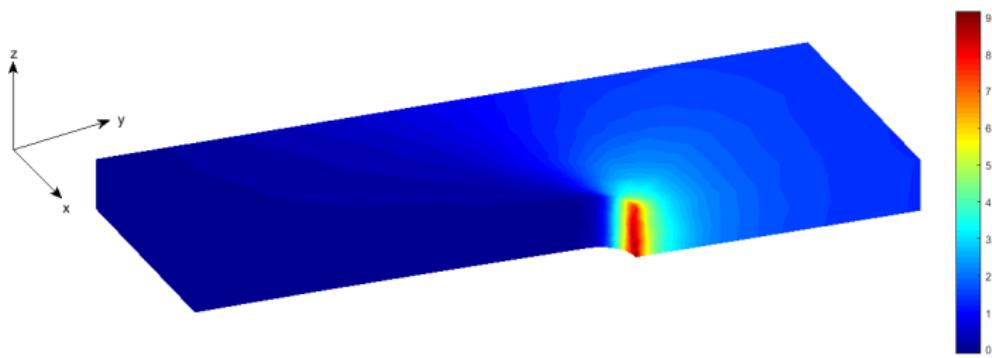


Figure: Three-dimensional σ_x of Specimen 3.

Numerical computation of the averaged effective stress

- ▶ The stress tensor field is defined by the linear elasticity theory. It is sufficient to compute the stress field once for each geometry.
- ▶ There are several proposals for the definition of effective stress. For example, we consider the maximum principal stress:

$$\sigma_{\text{eff}}(x, y) = \frac{1}{2}(\sigma_x + \sigma_y) + \sqrt{\left(\frac{\sigma_x - \sigma_y}{2}\right)^2 + \tau_{xy}^2}, \forall (x, y) \in D.$$

- ▶ The averaged effective stress field is obtained by averaging the effective stress locally:

$$\sigma_{\text{eff}}^{\Delta}(x) = \frac{1}{|B(x, \Delta) \cap D|} \int_{B(x, \Delta) \cap D} \sigma_{\text{eff}}(y) dy, \forall x \in D, \quad (215)$$

where $B(x, \Delta)$ is a cube (square) of length Δ centered at x .

Spatial Poisson Process I

Assumptions:

- (a) The (averaged) effective stress at $\mathbf{x} \in \partial D$ determines the crack formation at \mathbf{x} .
 - (b) Different cracks initiate independently.
- $\lambda(\mathbf{x}, n) \geq 0$: the intensity function which relates the spatial location \mathbf{x} to the number of cycles n .
- The intensity function depends on the effective stress, $\lambda(\mathbf{x}, n) = \eta(n; \sigma_{\text{eff}}^{\Delta}(\mathbf{x}))$, where η is a failure-rate function.
- $M_B(n)$: the number of crack initiations in the surface region $B \subset \partial D$, after performing n stress cycles

$$P(M_B(n) = m) = \frac{(\Lambda_B(n))^m}{m!} \exp(-\Lambda_B(n)), m = 0, 1, \dots$$

$$\text{where } \Lambda_B(n) = \int_0^n \int_B \lambda(\mathbf{x}, n') dS(\mathbf{x}) dn' \geq 0.$$

First crack initiation

- ▶ $N_{\partial D}$: the number of load cycles when the first crack initiates on ∂D .
- ▶ $\{N_{\partial D} > n\} \iff \{M_{\partial D}(n) = 0\}$.
- ▶ The survival probability after n cycles is

$$P(N_{\partial D} > n; \sigma_{\text{eff}}^{\Delta}) = \exp \left(- \int_0^n \int_{\partial D} \eta(n'; \sigma_{\text{eff}}^{\Delta}(\mathbf{x})) dS(\mathbf{x}) dn' \right)$$

- ▶ To parameterize the rate function $\eta(n; s)$, we relate it to a given S-N model:

$$\eta(n; s) = -\frac{1}{\gamma} \frac{\partial}{\partial n} \log (1 - F_{SN}(n; s, \theta)) = \frac{1}{\gamma} h_{SN}(n; s, \theta), \quad (216)$$

where $F_{SN}(n; s, \theta)$ and $h_{SN}(n; s, \theta)$ are the cdf and the hazard rate function for the S-N model, respectively, and γ is the highly stressed volume.

Highly stressed volume

- ▶ The highly stressed volume (area), γ , depends on the geometry.
- ▶ We let $\mathcal{A}_\beta = \{\mathbf{x} \in \partial D : \sigma_{\text{eff}}^1(\mathbf{x}) > \beta\}$, where $\sigma_{\text{eff}}^1(\mathbf{x})$ is the effective stress that corresponds to a unity traction. The highly stressed volume is given by

$$\gamma(\beta) = \int_{\partial D} \mathbf{1}_{\mathcal{A}_\beta}(\mathbf{x}) dS(\mathbf{x}).$$

- ▶ Thus, the spatial Poisson model is fully characterized by θ, β , and Δ , where θ depends on the selected S-N model.

Likelihood function

Under the assumption of independent experiments, the log-likelihood function is

$$\begin{aligned}\ell(\theta, \beta, \Delta) &= \sum_{i=1}^m [(1 - \delta_i) \log(P(N_{\partial D} > n_i; \sigma_{\text{eff},i}^\Delta)) + \delta_i \log(\rho^{\partial D}(n_i; \sigma_{\text{eff},i}^\Delta))] \\ &= \sum_{i=1}^m \left\{ \frac{1}{\gamma(\beta)} \int_{\partial D} \log(1 - F_{SN}(n_i; \sigma_{\text{eff},i}^\Delta(\mathbf{x}), \theta)) dS(\mathbf{x}) \right. \\ &\quad \left. + \delta_i \log \left(\frac{1}{\gamma(\beta)} \int_{\partial D} h_{SN}(n_i, \sigma_{\text{eff},i}^\Delta(\mathbf{x}), \theta) dS(\mathbf{x}) \right) \right\}. \quad (217)\end{aligned}$$

Calibration of the spatial Poisson model

Table: ML estimates of θ (from **Model Ia**) and β using (217) where $\Delta = 0$.

Data set/Specimen	A_1	A_2	A_3	q	τ	β	ℓ^*
1	5.88	-1.32	35.88	0.56	0.30	1.16	-938.90
2	6.00	-1.22	40.98	0.60	0.23	1.95	-391.45
3	7.62	-2.18	45.35	0.65	0.27	2.54	-301.82
1, 2 & 3	6.28	-1.47	35.99	0.57	0.38	1.83	-1650.05

Table: ML estimate of θ , β and Δ using (217).

Data set	A_1	A_2	A_3	q	τ	β	Δ	ℓ^*
1, 2 & 3	6.29	-1.47	35.99	0.57	0.35	1.83	0.0125	-1648.16

Survival functions I

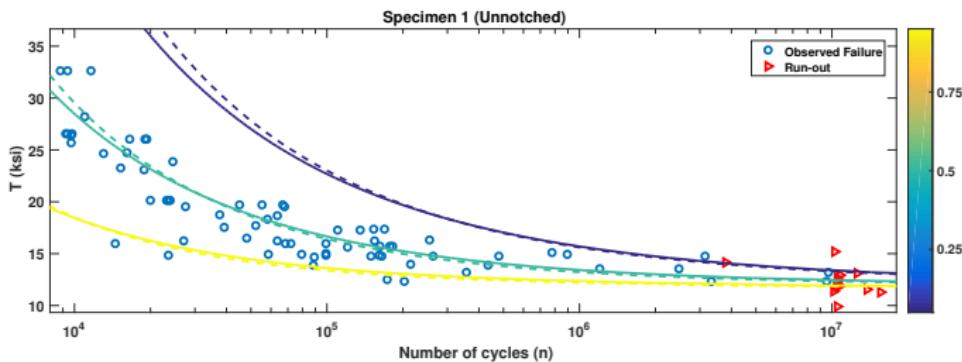


Figure: Contour plots of the survival-probability functions for specimen 1 computed with data set 1 ML estimates (dashed line) and pooled ML estimates (solid line); yellow is 0.95 probability, green is 0.5, and blue is 0.05.

Survival functions II

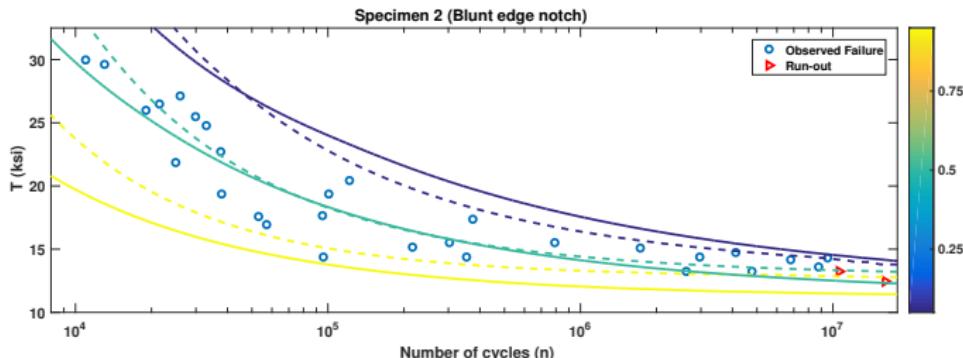


Figure: Contour plots of the survival-probability functions for specimen 2 computed with data set 2 ML estimates (dashed line) and pooled ML estimates (solid line); yellow is 0.95 probability, green is 0.5, and blue is 0.05.

Survival functions III

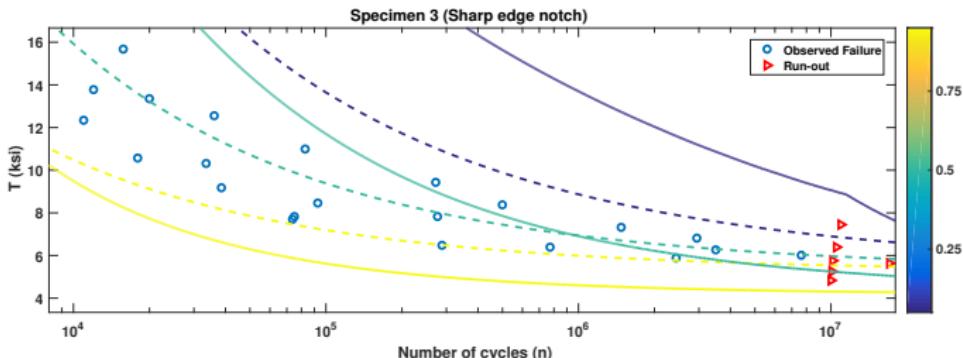


Figure: Contour plots of the survival-probability functions for specimen 3 computed with data set 3 ML estimates (dashed line) and pooled ML estimates (solid line); yellow is 0.95 probability, green is 0.5, and blue is 0.05.

Bayesian analysis

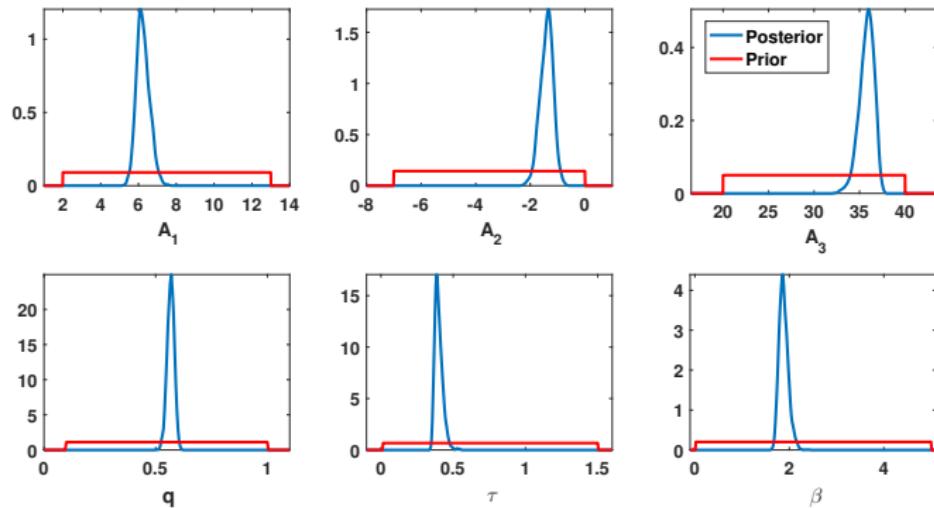


Figure: The estimated marginal posteriors for the parameters $A_1, A_2, A_3, q, \tau, \beta$, where $\Delta = 0$.

Bayesian analysis

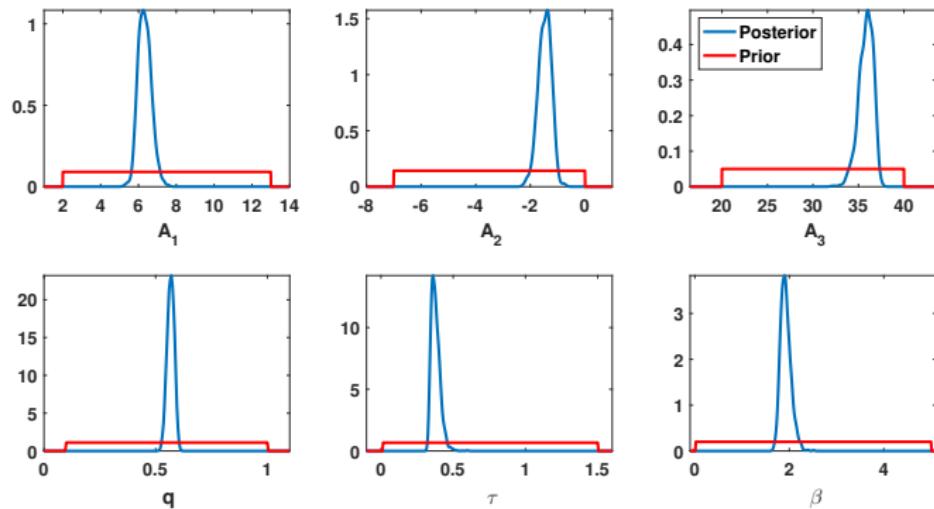


Figure: The estimated marginal posteriors for the parameters $A_1, A_2, A_3, q, \tau, \beta$ where $\Delta = 0.0125$.

Bayesian analysis

Table: Bayesian comparison between two different specifications of **Model Ia**.

Model Ia given data sets 1, 2 & 3 (2D)	$\Delta = 0$	$\Delta = 0.0125$
Log marginal likelihood	-1660.24	-1658.48
Deviance information criterion (DIC)	3311.33	3308.27

**Model Ia performs better with $\Delta = 0.0125$ than with $\Delta = 0$.
However, the small difference between the two cases suggests that
conclusions cannot be generalized to other S-N models.**

Survival functions I

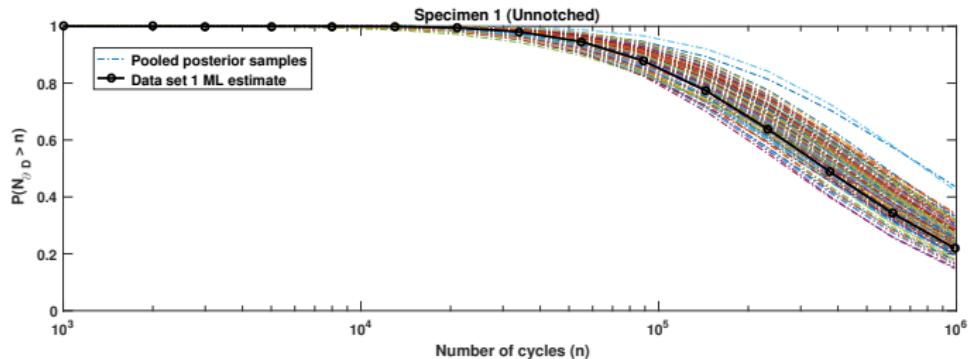


Figure: Survival-probability functions for specimen 1 when $S_{max} = 45$ ksi and $R = 0.1$, given different posterior samples of the parameters $A_1, A_2, A_3, q, \tau, \beta$.

Survival functions II

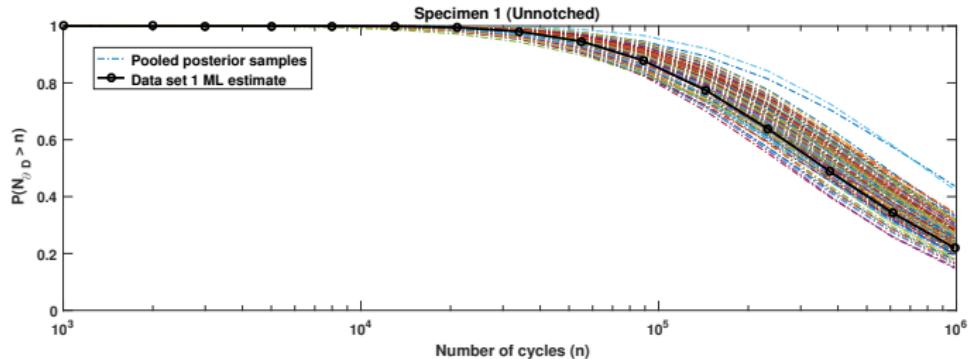


Figure: Survival-probability functions for specimen 1 when $S_{max} = 45$ ksi and $R = -0.1$, given different posterior samples of the parameters $A_1, A_2, A_3, q, \tau, \beta$.

Survival functions III

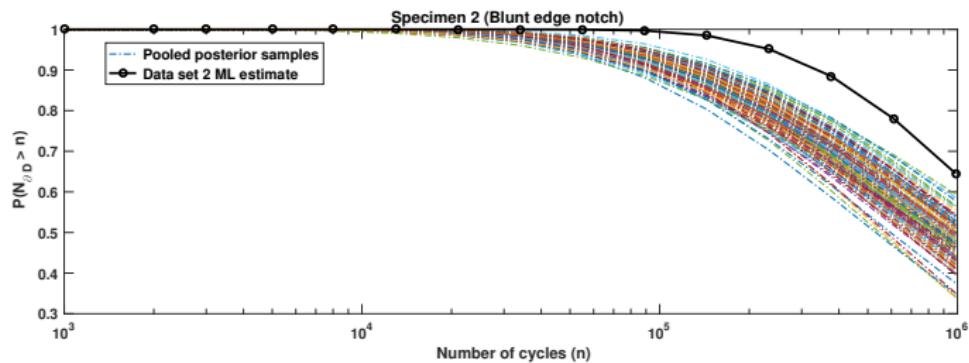


Figure: Survival-probability functions for specimen 2 when $S_{max} = 30$ ksi and $R = 0.1$, given different posterior samples of the parameters $A_1, A_2, A_3, q, \tau, \beta$.

Survival functions IV

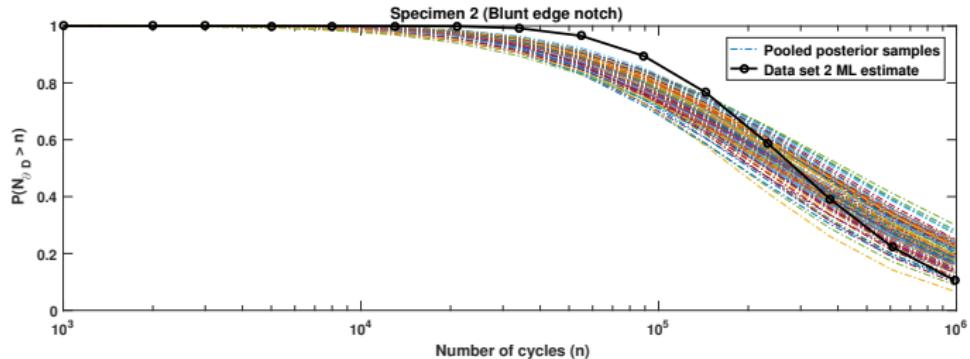


Figure: Survival-probability functions for specimen 2 when $S_{max} = 30$ ksi and $R = -0.1$, given different posterior samples of the parameters $A_1, A_2, A_3, q, \tau, \beta$.

Survival functions V

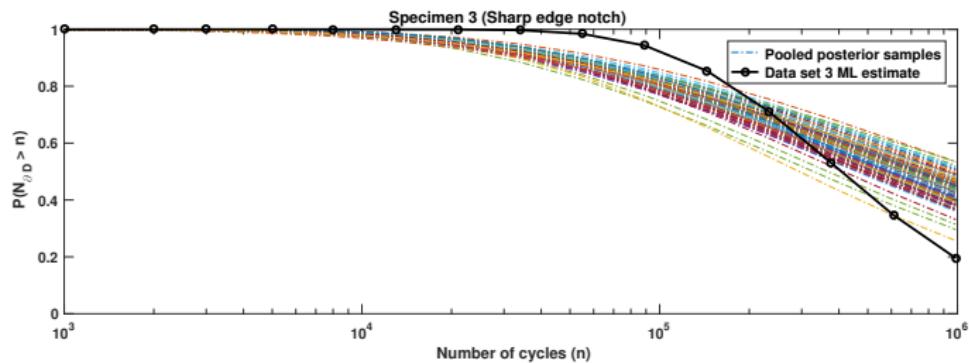


Figure: Survival-probability functions for specimen 3 when $S_{max} = 12$ ksi and $R = 0.1$, given different posterior samples of the parameters $A_1, A_2, A_3, q, \tau, \beta$.

Survival functions VI

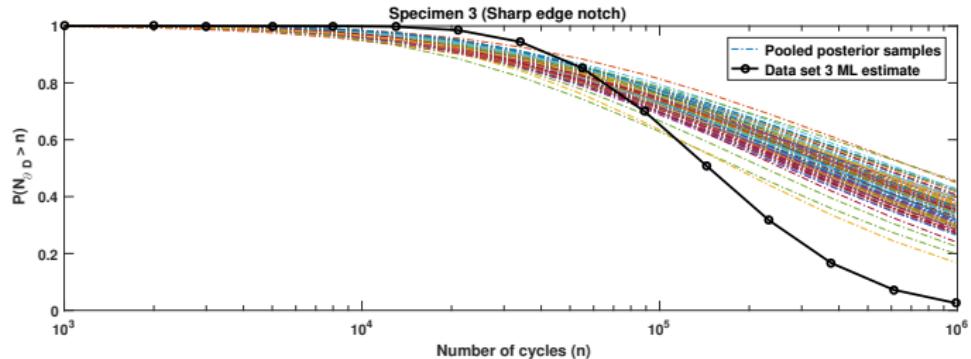


Figure: Survival-probability functions for specimen 3 when $S_{max} = 12$ ksi and $R = -0.1$, given different posterior samples of the parameters $A_1, A_2, A_3, q, \tau, \beta$.

Summary

- ▶ S-N models of various complexity are calibrated by means of the maximum likelihood method and Bayesian approach.
- ▶ Classical measures of fit and Bayesian predictive criteria were applied to compare and rank different models.
- ▶ The classical approach and the Bayesian approach for model comparison have provided evidence in favor of **Model II given the 75S-T6 data set.**
- ▶ The spatial Poisson process provides a systematic approach to generalize S-N models and calibrate fatigue experiments on different specimens without special treatments for notches.
- ▶ Given a sufficient number of fatigue experiments for specimens with diverse geometries, the proposed approach could be used to predict the life of any mechanical component made from the same material and having the same surface finish.

References

-  I. Babuška, Z. Sawlan, M. Scavino, B. Szabó, R. Tempone, Bayesian inference and model comparison for metallic fatigue data, Computer Methods in Applied Mechanics and Engineering 304 (2016): 171-196.
-  I. Babuška, Z. Sawlan, M. Scavino, B. Szabó, R. Tempone, Spatial Poisson processes for fatigue crack initiation, Computer Methods in Applied Mechanics and Engineering 345 (2019): 454-475.

References

-  Grover, H. J., Bishop, S. M., and Jackson, L. R. (March 1951).
Fatigue Strengths of Aircraft Materials. Axial-Load Fatigue Tests on Unnotched Sheet Specimens of 24S-T3 and 75S-T6 Aluminum Alloys and of SAE 4130 Steel NACA TN 2324.
National Advisory Committee on Aeronautics.
-  Pascual, F. G. and Meeker, W. Q. (1997).
Analysis of fatigue data with runouts based on a model with nonconstant standard deviation and a fatigue limit parameter.
Journal of Testing and Evaluation, 25:292–301.
-  Pascual, F. G. and Meeker, W. Q. (1999).
Estimating fatigue curves with the random fatigue-limit model.
Technometrics, 41(4):277–289.
-  Robert, C. and Casella, G. (2009).
Introducing Monte Carlo Methods with R.
Springer.
-  Walker, K. (1970).
The effect of stress ratio during crack propagation and fatigue for 2024-t3 and 7075-t6 aluminum.
Effects of environment and complex load history on fatigue life, 462:1–14.

Excursion: Gaussian measures

Recall that a random vector $x : (\Omega, \mathcal{F}, \mathbb{P}) \rightarrow \mathbb{R}^d$ is said to follow a (multivariate) **normal distribution** with mean $m \in \mathbb{R}^d$ and covariance $\Sigma \in \mathbb{R}^{d \times d}$ if

$$\mathbb{P}(x \in A) = \frac{1}{\det(\Sigma)^{\frac{1}{2}} (2\pi)^{\frac{d}{2}}} \int_A \exp \left(-\underbrace{\frac{(x - m)^\top \Sigma^{-1} (x - m)}{2}}_{\frac{1}{2} \|x - m\|_{\Sigma^{-1}}^2} \right) \lambda(dx) \quad (*)$$

where $\lambda : \mathcal{B}(\mathbb{R}^d) \rightarrow [0, \infty[$ denotes the d -dim. Lebesgue measure.

Notation: P pos. self-adj. operator on Hilbert space, then

$$\langle x, y \rangle_P := \langle x, Py \rangle \equiv \left\langle P^{\frac{1}{2}}x, P^{\frac{1}{2}}y \right\rangle, \quad \|x\|_P := \sqrt{\langle x, x \rangle_P}$$

Definition 22.1 (Gaussian measure)

Let $m \in \mathbb{R}^d$ and $C \in \mathbb{R}^{d \times d}$ be symmetric and positive definite. The **Gaussian measure** with mean m and covariance C , denoted as $\mathcal{N}(m, C)$, is defined by:

$$\mathcal{N}(m, C)(A) := \frac{1}{\det(C)^{\frac{1}{2}}(2\pi)^{\frac{d}{2}}} \int_A \exp\left(-\frac{1}{2}\|x - m\|_{C^{-1}}^2\right) \lambda(\mathrm{d}x)$$

for each $A \in \mathcal{B}(\mathbb{R}^d)$. The Gaussian measure $\gamma := \mathcal{N}(0, I)$ is called **standard Gaussian measure**.

Remark 22.1

- i A Dirac measure δ_m can be considered as a degenerate Gaussian measure on \mathbb{R} , namely one with covariance equal to zero.
- ii In terms of the Radon–Nikodym derivative, we thus have

$$\frac{d\mu}{d\lambda}(x) = \frac{1}{\det(C)^{\frac{1}{2}}(2\pi)^{\frac{d}{2}}} \exp\left(-\frac{1}{2}\|x - m\|_{C^{-1}}^2\right)$$

for the Gaussian measure $\mu := \mathcal{N}(m, C)$.

- iii That is, in the view of the push-forward measure (i.e., the measure induced by the distribution of x) $A \mapsto \mathbb{P}(x \in A) =: \mathbb{P}_x(A)$ in (*), we see the (unsurprising) fact that a normally distributed random vector induces a Gaussian measure.

Definition 22.2 (Non-degenerate Gaussian measure)

A non-degenerate Gaussian measure is a strictly positive probability measure on \mathbb{R}^d .

Remark 22.2

However, unlike the Lebesgue measure, it is not translation invariant.

Definition 22.3 (Push-forward measure)

Given a measurable function $f : \mathcal{X}_1 \rightarrow \mathcal{X}_2$ and a measure $\mu : \mathcal{B}(\mathcal{X}_1) \rightarrow [0, \infty[$, the push-forward measure is defined as:

$$(f)_*\mu(A) := \mu(f^{-1}(A))$$

for each $A \in \mathcal{B}(\mathcal{X}_2)$.

Lemma 22.4 (Cameron–Martin formula)

Let $\mu = \mathcal{N}(m, C)$ be a Gaussian measure on \mathbb{R}^d and $T_v : \mathbb{R}^d \rightarrow \mathbb{R}^d$ the translation map $T_v(x) := v + x$ for $v \in \mathbb{R}^d$. Then the push-forward measure $(T_v)_*\mu$ is equivalent to μ and $[(T_v)^*\mu](A) := \mu(\{T_v \in A\})$

$$\frac{d(T_v)_*\mu}{d\mu} = \exp\left(\langle v, x - m \rangle_{C^{-1}} - \frac{1}{2}\|v\|_{C^{-1}}^2\right).$$

Proof: direct ([Exercise](#)), as the claim means that for every integrable function f :

$$\int_{\mathbb{R}^d} f(x + v)\mu(dx) = \int_{\mathbb{R}^d} f(x) \exp\left(\langle v, x - m \rangle_{C^{-1}} - \frac{1}{2}\|v\|_{C^{-1}}^2\right) \mu(dx)$$



One can show that there is no analogue to the Lebesgue measure on infinite-dimensional vector spaces!

(think of what happens as $d \rightarrow \infty$!)

To define Gaussian measures for settings in which there may not be a Lebesgue measure with respect to which Radon–Nikodym derivatives can be computed, we use the easily verifiable **alternative characterization**:

Definition 22.5 (Gaussian measure in infinite dimensions)

A Borel measure μ on a normed space V is said to be a (non-degenerate) **Gaussian** if, for every continuous linear functional $\ell : V \rightarrow \mathbb{R}$, $\ell \neq 0$, the push-forward measure $\ell_*\mu$ is a (non-degenerate) Gaussian measure on \mathbb{R} . Equivalently, μ is Gaussian if, for every linear map $T : V \rightarrow \mathbb{R}^d$, $T_*\mu = \mathcal{N}(m_T, C_T)$ for some $m_T \in \mathbb{R}^d$ and some sym. pos. def $C_T \in \mathbb{R}^{d \times d}$.

Definition 22.6 (Mean and covariance in Banach spaces)

Let μ be a probability measure on a Banach space V . An element $m = m_\mu \in V$ is called **mean of μ** , if

$$\int_V \underbrace{\langle \ell, x - m_\mu \rangle}_{=\ell(x-m_\mu)=\ell(x)-\ell(m_\mu)} \mu(dx) = 0 \quad \forall \ell \in V',$$

so that $\mu(\ell) = \ell(m_\mu)$ for all $\ell \in V'$;

[i.e. $\int_V x \mu(dx) = m_\mu$ in the sense of a Pettis integral. The Pettis integral can be understood as the extension of the Lebesgue integral to vector-valued functions, i.e., x is Pettis integrable if $\ell(x)$ is Lebesgue integrable, see [Brooks, 1969]]. If $m_\mu = 0$, then μ is said to be **centred**. The **covariance operator** is a self-adjoint operator $C_\mu : V' \times V' \rightarrow \mathbb{K}$
 \mathbb{R} or \mathbb{C} defined by

$$C_\mu(\ell, k) := \int_V \langle k, x - m_\mu \rangle \overline{\langle \ell, x - m_\mu \rangle} \mu(dx).$$

It is sometimes convenient to abuse notation and write $C_\mu : V' \rightarrow V''$ for the operator defined by

$$\langle C_\mu k, \ell \rangle := C_\mu(k, \ell) \quad \forall \ell, k \in V'.$$

In the case that $V = H$ is a Hilbert space, it is common to use Riesz's representation theorem to identify H with H' and H'' . Hence, one can treat C_μ as a linear operator from H to itself. The inverse of C_μ is called precision operator of μ , provided it exists.

Remark 22.3

- i *The properties of the covariance operator of a Gaussian measure are closely related to the measure's non-degeneracy; see [Vakhania, 1975]*
- ii *Analogous to the finite dimensional case, Gaussian measures on Banach spaces can also equivalently be defined through their Fourier transforms.*

Also analogous to the finite dimensional case is that a Gaussian measure has well-defined moments of all orders.

Theorem 22.7 (Fernique, 1970)

Let μ be a centred Gaussian measure on a separable Banach space V . Then there exists $\alpha > 0$, such that

$$\int_V \exp(\alpha \|x\|^2) \mu(dx) < \infty.$$

It thus follows that $\int_V \|x\|^k \mu(dx) < \infty$ for all $k \geq 0$.

Remark 22.4 (cont.)

- iii If $V = H$ is a Hilbert space, then *Sazonov's Thm.* provides a characterization of self-adjoint operators on H that can be covariance operators of a Gaussian measure: pos., self-adj., and of trace class ($\|\sigma\|_{\ell^1} < \infty$, if $\sigma \equiv \sigma(n)$ is seq. of sing. values, and $\text{tr}(C_\mu) = \int_H \|x - m_\mu\|_H^2 \mu(dx) < \infty$).
- iv As not even translations lead to new measures that are abs.-cont. w.r.t. previous measures, an important object that is associated with a Gaussian measure μ in infinite dimensions, is its *Cameron–Martin space*.

Definition 22.8 (Cameron–Martin space)

Let μ be a Gaussian measure on a Hilbert space V . The subspace $H_\mu \subset V$ is a Hilbert space (when equipped with a suitable inner product defined through $C_\mu : V' \times V' \rightarrow \mathbb{K}$) containing the elements of V for which the (known) Cameron–Martin formula (see Lemma (22.4)) holds (with \mathbb{R}^d replaced by V suitably). This subspace H_μ is known as the Cameron–Martin space associated with μ .

Remark 22.5 (cont.)

- v An attractive point of view of Gaussian measures on Hilbert spaces is that random draws from such a measure are the same as draws from random series of the form

$$m + \sum_{k \in \mathbb{N}} \sqrt{\lambda_k} \xi_k \psi_k,$$

where $(\psi_k)_{k \in \mathbb{N}}$ are orthonormal eigenfunctions of C_μ and $(\lambda_k)_{k \in \mathbb{N}}$ are the associated (non negative) eigenvalues. Moreover, $(\xi_k)_{k \in \mathbb{N}}$ are i.i.d. $\mathcal{N}(0, 1)$ on \mathbb{R} .

~ \rightsquigarrow Karhunen–Loève expansion

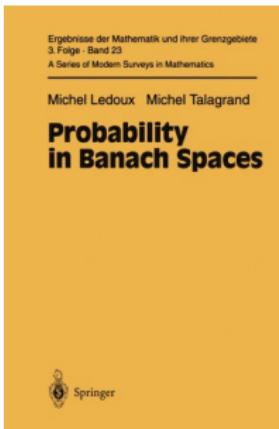
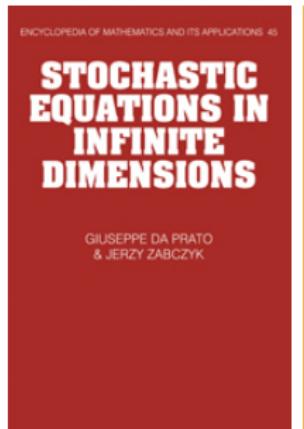
Theorem 22.9 (Feldman-Hájek)

Let μ and ν be two Gaussian measures on a locally convex measurable space $(\mathcal{X}, \mathcal{A})$. Then, either

- ▶ $\{A \in \mathcal{A} \mid \mu(A) = 0\} = \{A \in \mathcal{A} \mid \nu(A) = 0\}$, or
- ▶ there exist $A, B \in \mathcal{A}$, $A \cap B = \emptyset$, $A \cup B = \mathcal{X}$ such that $\mu = 0$ on all measurable subsets of B and $\nu = 0$ on all measurable subsets of A .

References:

- ▶ J.K. Brooks. Representations of weak and strong integrals in Banach spaces. *Proceedings of the National Academy of Sciences of the United States of America* 63, 1969, pp. 266–270
- ▶ Da Prato and Zabczyk, 1992: Chapter 2
- ▶ Ledoux and Talagrand, 1991
- ▶ A. Stuart. Inverse problems: A Bayesian perspective. *Acta Numerica*, 2010, 19, pp. 451–559
- ▶ M. Dashti and A. Stuart. The Bayesian approach to inverse problems. In: *Handbook of UQ*, Springer, 2017, pp. 311–428



Am Felsenweg 1969, pp. 451–559
© Cambridge University Press 2010
Printed in the United Kingdom

Inverse problems: A Bayesian perspective

A. M. Stuart
Mathematical Institute,
University of Warwick,
Coventry CV4 7AL,
UK
E-mail: a.m.stuart@warwick.ac.uk

The subject of inverse problems in differential equations is of enormous practical importance, and has been given substantial theoretical and computational treatment over the last few decades. This lecture notes highlight the mathematical structure of inverse problems while simultaneously being application oriented. The emphasis is on the Bayesian approach to inverse problems, which provides a natural way to incorporate the quantification of uncertainty and risk, something which is necessarily demanded in inverse problems arising in science and engineering. The Bayesian approach is also well suited to the construction of numerical methods for inverse problems, and can be used to analyse the performance of such methods. The Bayesian approach to inverse problems is a natural extension of the Bayesian framework for statistical inference, and is analogous to the frequentist approach to inverse problems. These notes are intended for researchers approaching inverse problems from a mathematical perspective, and are intended to provide a self-contained introduction to the Bayesian approach to inverse problems.

CONTENTS	PAGE
1. Introduction	521
2. The Bayesian framework	521
3. Bayesian inverse problems	521
4. Convolution operators	521
5. Approximation theory	521
6. Probability	521
References	521

The Bayesian Approach to Inverse Problems 10

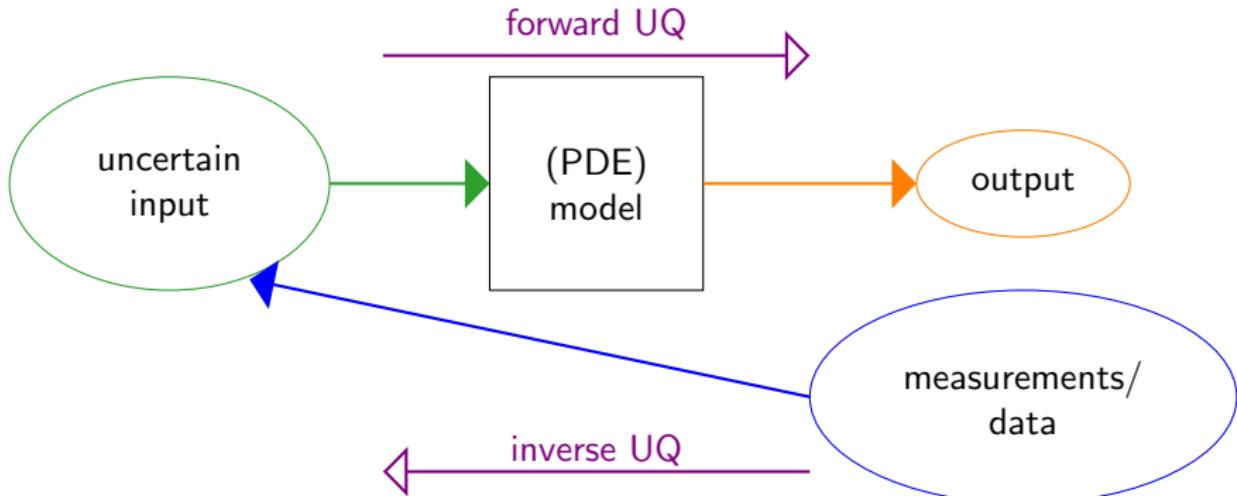
Masoumeh Dashti and Andrew M. Stuart

Abstract
These lecture notes highlight the mathematical and computational structure relating to the formulation and analysis of inverse problems via the Bayesian approach to inverse problems in a differential equation. The approach to inverse problems in the quantification of uncertainty within applications involving the finding of unknown parameters in a differential equation is first introduced and described first, along with some motivational examples. Then the development of probability theory and the Bayesian approach to inverse problems is introduced. These notes are an inferior set of functions to construct devices; these probability measures are as close as possible in the Bayesian approach to inverse problems. Finally, the Bayesian approach to inverse problems is applied to the solution of inverse problems using a common mathematical framework, and an application to the inverse problem of recovering the initial condition of a linear parabolic partial differential equation is provided. The notes conclude with a number of statistical and approximation results, and a brief discussion of the Bayesian approach to inverse problems, which are used to explain the Bayesian approach to inverse problems. These notes are intended for researchers approaching inverse problems from a mathematical perspective, and are intended to provide a self-contained introduction to the Bayesian approach to inverse problems.

M. Dashti (✉)
Department of Mathematics, University of Sussex, Brighton, UK
e-mail: m.dashti@sussex.ac.uk
A.M. Stuart
Mathematics Institute, University of Warwick, Coventry, UK
e-mail: a.m.stuart@warwick.ac.uk

Bayesian Inverse Problems

- ▶ So far we have mainly considered forward uncertainty quantification (UQ) problems, in the sense that



Question:

- ▶ What input distribution should one use?

For example: Consider $-\operatorname{div}(\exp(\gamma)\operatorname{grad} p) = f$ in $D \in \mathbb{R}^d$ equipped with boundary conditions. Given observations of p , what can we say about γ ?

Objectives:

- i describe inverse problem's perspective and their regularization
- ii Bayesian inversion (in Banach spaces)
- iii well-posedness and approximation
- iv practical considerations

A) main references:

- ▶ A. Stuart. Inverse problems: A Bayesian perspective. *Acta Numerica*, 2010, 19, pp. 451–559
- ▶ M. Dashti and A. Stuart. The Bayesian approach to inverse problems. In: *Handbook of UQ*, Springer, 2017, pp. 311–428

B) (alternative, common) discretize first approach:

- ▶ J. Kaipio and E. Somersalo. *Statistical and Computational inverse problems*. Springer 2004.

Inverse problems: A Bayesian perspective

A. M. Stuart
Mathematical Institute,
University of Oxford,
Oxford OX2 6EL, UK
e-mail: a.m.stuart@maths.ox.ac.uk

The subject of inverse problems is concerned with recovering information about objects from indirect measurements. Typically, one form of information is sought to recover from another which is more easily measured. Inverse problems arise in many fields of science and engineering, including medical imaging, non-destructive testing, geophysics, and signal processing. This subject offers a highly interdisciplinary area of research, with applications in a wide range of fields. Inverse problems are often ill-posed, meaning that small errors in the data can lead to large errors in the solution. It is therefore important to understand the mathematical properties of inverse problems in order to make them tractable for their application. Furthermore, the approach to computational inverse problems must take account of the fact that they are often applied to real-world computational resources in a noisy application sense. In this paper we introduce the Bayesian approach to inverse problems, and highlight its strengths and weaknesses.

We demonstrate that, when formulated in a Bayesian fashion, a wide range of inverse problems can be solved using standard statistical methods. We highlight a theory of ill-posedness which arises from this. The techniques discussed here are general enough to apply to a wide range of inverse problems, which we describe. We also review a range of algorithms which can be used to solve inverse problems, including MCMC methods, the Bayesian linearization approach, and the variational approach.

CONTENTS

1 Introduction	421
2 The Bayesian Framework	430
3 Data Assimilation	435
4 Bayesian Inference	438
5 Sparse Recovery	468
6 Variational Methods	471
References	484

The Bayesian Approach to Inverse Problems 10

Masoumeh Dashti and Andrew M. Stuart

Abstract
These lecture notes highlight the mathematical and computational analysis relating to the formulation and development of algorithms for the Bayesian approach to inverse problems in differential equations. The approach is for forward models which are linear or nonlinear, and the data is noisy. The field of mathematical analysis with this theme is developing rapidly. The Bayesian framework is developed, and the corresponding theory of probability measures on separable Banach spaces is undertaken, using a random measure over the space of states of a dynamical system. These probability measures are used to provide the Bayesian posterior distributions of unknown parameters. Regularity of draws from the priors is studied in the natural Sobolev or Hilbert spaces, and the resulting posterior distributions are shown to be Hölder continuous and the Kullback–Leibler consistency theorem is used to extend regularity considerations to the posterior distributions. The posterior distributions are then used in a prior setting, and here interpreted as finding conditions under which the posterior distributions are well approximated by a Gaussian distribution, using a formula for the Radon–Nikodym derivative in terms of the likelihood of the data. Having made this use of the posterior distributions, various variational approaches are used in the numerical solution setting. These approaches include self-adjoint, approximation theory, and the existence of maximum a posteriori (MAP) estimates. Numerical examples are given, including one on the infinite-dimensional space, including Matrix Chain Monte Carlo and sequential Monte Carlo methods, and measure preserving reversible stochastic

Applied
Mathematical
Sciences
160

Jari Kaipo
Erkki Somersalo

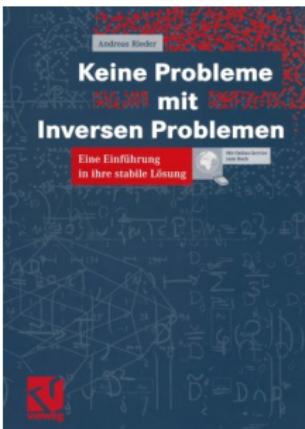
Statistical and Computational Inverse Problems

M. Dashti (✉)
Department of Mathematics, University of Sussex, Brighton, UK
e-mail: m.dashti@susx.ac.uk
A.M. Stuart
Mathematical Institute, University of Oxford, Oxford, UK
e-mail: a.m.stuart@maths.ox.ac.uk

© Springer International Publishing AG 2018
B. Hansen et al. (eds.), *Handbook of Uncertainty Quantification*,
DOI 10.1007/978-3-319-23442-7_10

111

Springer

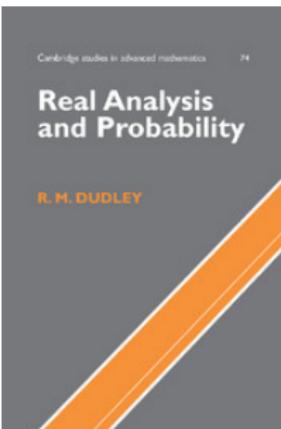


Cambridge studies in advanced mathematics

34

Real Analysis and Probability

R. M. DUDLEY

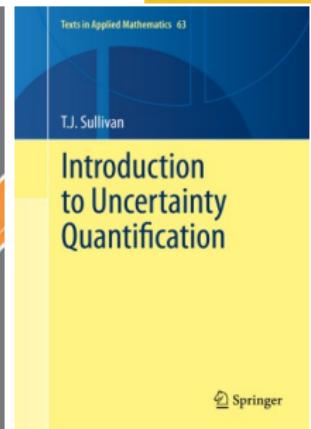


Tests in Applied Mathematics 63

T.J. Sullivan

Introduction to Uncertainty Quantification

Springer



Rank-Deficient and Discrete Ill-Posed Problems

Numerical Aspects
of Linear Inversion

Per Christian Hansen



1.) Inverse Problems and Regularization

- We begin with the **forward problem**, where some given **input u** for a mathematical model G is used to determine the corresponding **output y** through:

$$y = G(u), \quad (*)$$

where $G : U \rightarrow Y$ is called **observation operator**, so that $u \in U$ and $y \in Y$. Here, **U and Y are Banach spaces**. That is, given u and G , one finds ("computes") y .

Example: Given a realization of a random field γ , $G(u)$ with $u \equiv u(\gamma)$ is the value of a functional applied to the solution of a PDE. For instance, if $u \in H_0^1(D)$ solves $-\operatorname{div}(e^\gamma \operatorname{grad} u) = f$ in D , $u|_{\partial D} = 0$, and $G : H_0^1(D) \rightarrow \mathbb{R}^m$, then $U = H_0^1(D)$ and $Y = \mathbb{R}^m$.

In many applications, it is required to solve the **inverse problem**: given y and G , find u such that $(*)$ holds.

- ▶ Inverse problems are typically **ill-posed** (in the sense of Hadamard):
 - i there may be no solution,
 - ii the solution may not be unique,
 - iii the solution depends sensitively on y .

Indeed, often we do not observe $G(u)$ exactly, but **only a perturbed version** that is corrupted by **noise** η , for example, in an additive way:

$$y = G(u) + \eta. \tag{**}$$

Remark 22.6

- ▶ *We will focus on the additive noise case $(**)$ only. However, we emphasize that other noise models have been established in the literature, see [Kaipio & Somersalo; Chapt. 3.2]. The fundamental principle we will discuss here remains unaltered, although some theoretical results may need to be adapted for non-additive noise models.*

Example: The inverse problem framework (***) comprises the problem of **parameter estimation** (or **model calibration**), which we have discussed briefly at the beginning of the semester already. That is, the situation where the model $G \equiv G_\theta$ relating inputs to outputs depends on the same parameters $\theta \in \Theta \subset \mathbb{R}^p$, $p \geq 1$, and we seek to identify the parameter value θ such that $y_i = G_\theta(u_i)$ for each **input-output pair** (u_i, y_i) , $i = 1, \dots$. Also, this may be ill-posed, even without noise.

One popular approach to remedy the ill-posedness of problem (***) is to seek for a **least-squares solution**: find

$$\arg \min_{u \in U} \|y - G(u)\|_Y^2$$

for some norm $\|\cdot\|_Y$ on Y . However, the least-squares optimization problem, too, can be “difficult” to solve as there may be minimizing sequences that do not have a limit in U , it may have multiple minima, or it may possess a sensitive dependence on the observed data y .

These issues can (in particular the sensitive dependence) be somewhat ameliorated by solving a **regularized least-squares minimization** problem of the form

$$\arg \min_{u \in V} (\|y - G(u)\|_Y^2 + \|u - \bar{u}\|_V^2)$$

for some Banach space $V \subset U$ and given $\bar{u} \in V$, which has the effect to not fit the observed data too closely only.

The standard example of this classical regularization setting is [Tikhonov regularization](#): when U and Y are Hilbert spaces, we seek

$$\arg \min_{u \in U} \left(\|y - G(u)\|_Y^2 + \|R^{-\frac{1}{2}}(u - \bar{u})\|_U^2 \right)$$

for a given compact, positive, self-adjoint operator $R : U \rightarrow U$. The operator R and the state \bar{u} describe the structure of the regularization. This is an arguably arbitrary modelling assumption and can be viewed as a [practitioner's "prior belief of what the solution should look like"](#).

Even more generally, one might seek

$$\arg \min_{u \in U} \left(\|Q^{-\frac{1}{2}}(y - G(u))\|_Y^2 + \|R^{-\frac{1}{2}}(u - \bar{u})\|_U^2 \right)$$

for a given compact, positive, self-adjoint operator $Q : Y \rightarrow Y$ in order to weight the various components of y differently.

1.1 A probabilistic viewpoint

These least-squares approaches all seem to be somewhat *ad hoc*, however, in particular where the choice of regularization is concerned. **Adopting a probabilistic, in particular Bayesian, approach to inverse problems alleviates these difficulties.** In fact, if we think of u and y as random variables, then (**), i.e., $y = G(u) + \eta$, defines the **conditional random variable $y|u$** and one may define the “solution” of the inverse problem to be the conditioned random variable $u|y$. Using the statistical properties of the noise η we can thus **a priori specify the form of solutions that we believe to be more “likely”**, thereby allowing us to attach weights to multiple solutions that may explain the data. **This is the essence of the Bayesian approach to inverse problems**, which we will discuss in detail in the following. Notice that this means finding a distribution!

Remark 22.7

*In practice, we will often not have access to the true/exact observation operator $G : U \rightarrow Y$ but to a numerical approximation $G_h : U \rightarrow Y$ instead, where $h > 0$ denotes a discretization parameter. Using the approximate observation operator to gauge the likelihood of a certain input u , we suggest to write (**)) as:*

$$y = G_h(u) + \varepsilon_h + \eta,$$

where $\varepsilon_h \equiv \varepsilon_h(u) := G(u) - G_h(u)$. One could, in principle, combine the observation noise η and the discretization error ε_h into a single term. However, keeping them separate is usually more appropriate: unlike η , ε_h is typically not mean zero and depends on u . We will return to this aspect later.

1.2 Excursion on the central role of least-squares in elementary statistics:

To further motivate the more general discussion that will follow, let's first consider the following finite-dimensional linear problem. Suppose we want to "learn" a parameter vector $u \in \mathbb{R}^n$, which realizes the observation vector $y \in \mathbb{R}^m$ via

$$y = Au + \eta,$$

where:

- ▶ $A \in \mathbb{R}^{m \times n}$ is a known linear operator (i.e., a matrix)
- ▶ $\eta \in \mathbb{R}^m$ is a **noise vector** (not necessarily Gaussian) with mean zero and symmetric, positive definite covariance $Q := \mathbb{E}(\eta\eta^\top) \in \mathbb{R}^{m \times m}$
- ▶ η is **independent of u** .

A common approach in statistics then is to seek an estimator \hat{u} of u that satisfies:

- ▶ \hat{u} is a linear function of the data y , that is $\hat{u} = Ky$
 - ▶ \hat{u} is an unbiased estimator: $\mathbb{E}(\hat{u}) = u$
 - ▶ \hat{u} is the best estimator in that it minimizes a certain cost function.
- ~~~ BLUE: best linear unbiased estimator

The following celebrated Gauss-Markov Thm. states that there exists precisely one such estimator of u , and it is the solution to the weighted least-squares problem with weight Q^{-1} :

$$\hat{u} = \arg \min_{u \in \mathbb{R}^n} J(u), \quad J(u) := \frac{1}{2} \|Au - y\|_{Q^{-1}}^2$$

Exercise 22.1

Determine the explicit form of \hat{u} defined above, assuming that $A^\top Q^{-1} A$ is invertible and that A has rank $\text{rg}(A) = n$.

Solution: Note that

$$\begin{aligned} 2J(u) &= \|Q^{-\frac{1}{2}}(Au - y)\|_2^2 = \langle Au - y, Q^{-1}(Au - y) \rangle \\ &= \langle u, A^\top Q^{-1}Au \rangle - \langle y, Q^{-1}Au \rangle - \langle Au, Q^{-1}y \rangle \\ &\quad + \langle y, Q^{-1}y \rangle \\ &= \langle u, A^\top Q^{-1}Au \rangle - 2\langle y, Q^{-1}Au \rangle + \langle y, Q^{-1}y \rangle. \end{aligned}$$

By the usual matrix calculus, it follows that

$$\begin{aligned} \nabla J(u) &= \frac{1}{2} \left[A^\top Q^{-1}A + (A^\top Q^{-1}A)^\top \right] u - \frac{2}{2} (Q^{-1}A)^\top y \\ &= A^\top Q^{-1}Au - A^\top Q^{-1}y, \end{aligned}$$

whose critical point is $\hat{u} := (A^\top Q^{-1}A)^{-1} A^\top Q^{-1}y$, which is the minimum as $\nabla^2 J(u) = A^\top Q^{-1}A$ is pos.-def.

$$0 \leq \|Q^{-\frac{1}{2}}Ax\|_2^2 = x^\top A^\top Q^{-1}Ax = y^\top Q^{-1}y, \quad y := Ax$$

$$\begin{aligned} 0 = y^\top Q^{-1}y \Leftrightarrow 0 = y = Ax \Leftrightarrow \text{kernel}(A) = \{0\} \\ \Leftrightarrow \text{rg}(A) = n \end{aligned}$$

□

In fact, the Gauss-Markov Thm. holds for arbitrary Hilbert spaces. For simplicity, we will state the Thm. for the finite-dimensional setting only.

Theorem 22.10 (Gauss-Markov)

Let $A : \mathbb{R}^n \rightarrow \mathbb{R}^m$ be linear and $y = Au + \eta$ for $u \in \mathbb{R}^n$, where η is a centred \mathbb{R}^m -valued random variable with (sym.) positive-definite covariance $Q \in \mathbb{R}^{m \times m}$. Suppose that $\text{rg}(A) = n$ and that $A^\top Q^{-1} A$ is invertible. Then, among all unbiased linear estimators $K : \mathbb{R}^m \rightarrow \mathbb{R}^n$ that produce an estimate $\hat{u} = Ky$ of u given y , the one that minimizes both the mean-squared error $\mathbb{E} [\|\hat{u} - u\|_2^2]$ and the covariance $\mathbb{E} [(\hat{u} - u)(\hat{u} - u)^\top]$ (in the sense of pos.-def. operators) is given by

$$K = (A^\top Q^{-1} A)^{-1} A^\top Q^{-1},$$

so that $\mathbb{E}(\hat{u}) = u$ and $\mathbb{E} [(\hat{u} - u)(\hat{u} - u)^\top] = (A^\top Q^{-1} A)^{-1}$.

Proof: It is easily seen that $\hat{u} = Ky$ with K given above is an unbiased estimator. Indeed, we write

$$\begin{aligned}\hat{u} &= Ky = (A^\top Q^{-1} A)^{-1} A^\top Q^{-1} (Au + \eta) \\ &= u + (A^\top Q^{-1} A)^{-1} A^\top Q^{-1} \eta,\end{aligned}$$

so that $\mathbb{E}(\hat{u}) = u$ since $\mathbb{E}(\eta) = 0$. Moreover,

$$\begin{aligned}\mathbb{E}[(u - \hat{u})(u - \hat{u})^\top] &= K\mathbb{E}(\eta\eta^\top)K^\top \\ &= [(A^\top Q^{-1} A)^{-1} A^\top Q^{-1}] Q [Q^{-1} A (A^\top Q^{-1} A)^{-\top}] \\ &= (A^\top Q^{-1} A)^{-1} A^\top Q^{-1} A (A^\top Q^{-1} A)^{-1} \\ &= (A^\top Q^{-1} A)^{-1}\end{aligned}$$

as $Q = \mathbb{E}(\eta\eta^\top)$ is symmetric.

To assume uniqueness, let $L = K + D$ be any other linear unbiased estimator. Then

$$\begin{aligned} Ly &= Ky + Dy = u + K\eta + D(Au + \eta) \\ &= (I + DA)u + (K + D)\eta. \end{aligned}$$

We immediately see that this implies $DA = 0$, as it will not be unbiased otherwise. The corresponding covariance is:

$$\begin{aligned} \mathbb{E}[(Ly - u)(Ly - u)^\top] &= (K + D)Q(K + D)^\top \\ &= KQK^\top + \underbrace{DQK^\top}_{=(KQD^\top)^\top} + KQD^\top + DQD^\top. \end{aligned}$$

As $DA = 0$, we also find:

$$KQD^\top = (A^\top Q^{-1}A)^{-1} A^\top Q^{-1} QD^\top = (A^\top Q^{-1}A)^{-1} (DA)^\top = 0,$$

hence

$$\mathbb{E}[(Ly - u)(Ly - u)^\top] = KQK^\top + DQD^\top \geq KQK^\top$$

because DQD^\top is symmetric and positive semi-definite. □

Definition 22.11 (Moore-Penrose pseudo-inverse)

For $B \in \mathbb{R}^{m \times n}$, a matrix B^+ satisfying the following conditions

- ▶ $BB^+B = B$
- ▶ $B^+BB^+ = B^+$
- ▶ $(BB^+)^* = BB^+$
- ▶ $(B^+B)^* = B^+B$

is known as the **Moore-Penrose pseudo-inverse** of B .

Remark 22.8

The Moore-Penrose pseudo-inverse exists for any matrix B and is unique.
Furthermore, it solves the **linear least-squares problem**

$$B^+b = \arg \min_{x \in \mathbb{R}^n} \|b - Bx\|_2^2.$$

Remark 22.9

If $A^\top Q^{-1}A$ is not invertible, then it is common to use the estimator
 $\hat{u} = Ky$, where

$$K = (A^\top Q^{-1}A)^+ A^\top Q^{-1}.$$

1.3 Bayesian Interpretation of Regularization

Although the Gauss-Markov estimator offers a systematic approach to incorporate the observation noise (under minimal assumptions) through a probabilistic viewpoint, it is not ideal. Indeed, its characterization as the minimizer of a quadratic “cost” function means that the Gauss-Markov estimator is sensitive to large outliers in the data (i.e., components of y that differ greatly from the corresponding components of $A\hat{u}$). It thus seems natural to also regularize here, to not try to fit the observed data too closely, e.g., via

$$J(u) := \frac{1}{2} \|Au - y\|_{Q^{-1}}^2 + \frac{1}{2} \|u - \bar{u}\|_{R^{-1}}^2$$

for a fixed $\bar{u} \in \mathbb{R}^n$ and symmetric positive-definite “Tikhonov matrix” $R \in \mathbb{R}^{n \times n}$.

As mentioned before, this procedure seems somewhat ad hoc; however, it has a **natural Bayesian interpretation**. To illustrate this, we make the additional assumption that

$$\eta \sim \mathcal{N}(0, Q).$$

From a **Bayesian viewpoint**, the observation equation

$$y = Au + \eta$$

defines the conditional distribution $y|u$ as

$$\underbrace{(y - Au)|_u}_{=\eta} \sim \mathcal{N}(0, Q), \text{ or } y|u \sim \mathcal{N}(Au, Q).$$

Finding the minimizer of $u \mapsto \frac{1}{2} \|Au - y\|_{Q^{-1}}^2$, i.e., determining the **Gauss-Markov estimator**, therefore amounts to finding the **maximum likelihood estimator** of u given y . The **Bayesian interpretation** of the **additional regularization term** is then that $\mathcal{N}(\bar{u}, R)$ is a **prior distribution** for u ! In view of Bayes' rule, the resulting **posterior distribution** for $u|y$ has a Lebesgue density $\rho(u|y)$ with

$$\rho(u|y) = \frac{1}{Z(y)} \exp \left(-\frac{1}{2} \|Au - y\|_{Q^{-1}}^2 - \frac{1}{2} \|u - \bar{u}\|_{R^{-1}}^2 \right),$$

where

$$Z(y) = \int_{\mathbb{R}^n} \exp \left(-\frac{1}{2} \|Au - y\|_{Q^{-1}}^2 - \frac{1}{2} \|u - \bar{u}\|_{R^{-1}}^2 \right) du.$$

Exercise 22.2

Show that $u \mapsto \rho(u|y)$ is the (Lebesgue-) density of a (multivariate) Normal distribution. Also determine its mean and covariance.

Consequently, the solution of $\arg \min_{u \in \mathbb{R}^n} J(u)$, with the regularized cost $J(u)$ defined above, is equivalent to the **maximum a-posteriori (MAP) estimator** of u given y . However, the full posterior distribution of u given y contains even more information than just the MAP (point-) estimator.

Exercise 22.3

Describe how the precision matrix of the posterior distribution can be used to assess uncertainties in the MAP estimator.

2. Bayesian Inversion in Banach spaces

Recall that we are concerned with inverse problems in function spaces for differential equations. Naturally, such problems can (need to be) discretized (eventually) and it is tempting to consider the inverse problem in the finite-dimensional setting.

But this “*discretize first, invert then*” approach does not always yield a well-behaved procedure, e.g., due to:

- ▶ stable vs. unstable discretizations
- ▶ discrete problem has a solution while continuous one does not (e.g., inverse heat eqn., or control problem for wave eqn.)
 - ~~ ill-conditioning as dimension tends to infinity

Instead, the guiding principle for the Bayesian inverse problems will be to “[invert first, then discretize](#)”. This “avoid discretization until last possible moment” is enormously empowering throughout numerical analysis, as it will automatically propagate the well-posedness properties from the continuous level to the finite-dimensional one. This should be compared to the guiding principle of FEM and the theoretical well-posedness via Lax-Milgram (for linear problems).

In this part, we discuss Bayesian inversion in Banach spaces.

Example: Consider our “usual” toy example of an elliptic PDE, namely Darcy’s law:

$$-\operatorname{div}(\kappa \nabla p) = f \quad \text{in } D \subset \mathbb{R}^d$$

equipped with suitable boundary conditions on ∂D .

That is, the pressure field p induces a velocity $v = -\kappa \nabla p$ in a medium characterized by κ . For simplicity, suppose that $\kappa : D \rightarrow \mathbb{R}^+$. Let $u := \ln(\kappa)$. Then we consider the problem of determining $u \in U$ from observations

$$y_j = p(x_j) + \eta_j, \quad j = 1, \dots, m$$

at positions $x_1, \dots, x_m \in D$; or some other collection of functionals. This fits our abstract setting

$$y = G(x) + \eta, \quad \eta \sim \mathcal{N}(0, Q)$$

for $y \in Y = \mathbb{R}^m$, where the operator G is defined implicitly by the solution operator to the elliptic BVP. Finally, we know already conditions under which the PDE is well-posed, which tell us: $U = L^\infty(D)$. Recall that $L^\infty(D)$ is a Banach space.

As we work in Banach spaces for U , we will need to:

- i establish an appropriate rigorous statement of Bayes' rule for settings where no Lebesgue density exists;
- ii use a suitable prior measure on U (e.g., Gaussian or Besov measure)

For simplicity, here we will mainly consider Gaussian priors. However, it is important to note that even if the observation noise is Gaussian and if a "simple" Gaussian prior is used, the posterior distribution will, in general, not be Gaussian!

Remark 22.10

If you are not familiar with Gaussian measures, please consult the supplementary notes.

Infinite dimensional spaces

We begin with establishing the infinite-dimensional **analogue of Bayes' rule**:

Let U and Y be **separable Banach spaces**, equipped with the Borel σ -algebra, and suppose that the observation operator $G : U \rightarrow Y$ is measurable. Let the observation $y \in Y$ be given by

$$y = G(u) + \eta,$$

where the **noise** η is a **Y -valued** random variable. Moreover, we consider the pair (u, y) to be a **$U \times Y$ -valued** random variable. The Bayesian inversion then entails computing the conditional distribution $u|y$.

The joint random variable (u, y) is specified via

- a **prior measure**: $u \sim \mu_0$ on U
- b **noise measure**: $\eta \sim Q_0$ on Y , and η indep. of u .

The conditional random variable $y|u$ is then distributed according to the measure Q_u , which is the translate of Q_0 by $G(u)$. Assume that $Q_u \ll Q_0$ $u - \mu_0$ a.s. Then

$$\frac{dQ_u}{dQ_0}(y) = C(y) \exp(-\Phi(u; y))$$

for some **potential** $\Phi : U \times Y \rightarrow \mathbb{R}$. Note that:

- ▶ for fixed $u \in U$, $\Phi(u; \cdot) : Y \rightarrow \mathbb{R}$ is measurable and

$$\int_Y C(y) \exp(-\Phi(u; y)) Q_0(dy) = 1$$

- ▶ for a given instance $y \in Y$ fixed, $-\Phi(\cdot; y)$ is called the **log-likelihood**.

Definition 22.12 (Outer measure)

The outer measure on a set \mathcal{X} is a function $\mu : \mathcal{P}(\mathcal{X}) \rightarrow [0, \infty]$ such that

- ▶ $\mu(\emptyset) = 0$ and
- ▶ if $A \subseteq \bigcup_{j=1}^{\infty} A_j$ then $\mu(A) \leq \sum_{j=1}^{\infty} \mu(A_j)$ for arbitrary subsets A, A_1, A_2, \dots of \mathcal{X} ,

where $\mathcal{P}(\mathcal{X})$ is the power set of \mathcal{X} .

Definition 22.13 (μ measurability)

A subset A of a set \mathcal{X} is μ measurable if and only if

$$\mu(B) = \mu(B \cap A) + \mu(B \cap A^c)$$

for every subset B of \mathcal{X} (see also Definition 2.1).

Define the product measure ν_0 on $U \times Y$ via

$$\nu_0(\mathrm{d}u \otimes \mathrm{d}y) := \mu_0(\mathrm{d}u)Q_0(\mathrm{d}y),$$

that is the (product) under which u and y are independent.

$$\begin{aligned}\int f \circ \Phi \nu_0(\mathrm{d}u \otimes \mathrm{d}y) &= \int f \circ \Phi \mathrm{d}\mu_0 \underbrace{\frac{\mathrm{d}Q_0}{\mathrm{d}Q_u}}_{= C(y)^{-1} e^{\Phi(u;y)}} \mathrm{d}Q_u \\ &= C(y)^{-1} e^{\Phi(u;y)}\end{aligned}$$

Suppose that Φ is (jointly) ν_0 measurable. Then the random variable $(u, y) \in U \times Y$ is distributed according to the measure

$$\nu(\mathrm{d}u \otimes \mathrm{d}y) := \mu_0(\mathrm{d}u)Q_u(\mathrm{d}y).$$

Moreover, $\nu \ll \nu_0$ and

$$\frac{\mathrm{d}\nu}{\mathrm{d}\nu_0}(u, y) = c(y) \exp(-\Phi(u; y)).$$

We then get the infinite-dimensional analogue of Bayes' rule:

Theorem 22.14 (generalized Bayes' rule)

Suppose that $\Phi : U \times Y \rightarrow \mathbb{R}$ is ν_0 measurable and that

$$Z(y) := \int_U \exp(-\Phi(u; y)) \mu_0(\mathrm{d}u) > 0 \quad Q_0 - \text{a.s.}$$

Then the conditional distribution of $u|y$ exists under ν and is denoted by μ^y . Furthermore, $\mu^y \ll \mu_0$ and, for y ν -a.s.

$$\frac{\mathrm{d}\mu^y}{\mathrm{d}\mu_0}(u) = \frac{1}{Z(y)} \exp(-\Phi(u; y)).$$

The rigorous proof requires some technical results from advanced measure theory and is skipped here; see Stuart, Dashti & Stuart, and Dudley for details.

In order to implement Bayes' rule, one has to ensure **four key ingredients**:

- i define suitable (prior) measures μ_0 and Q_0 , whose independent product forms ν_0
- ii determine potential Φ such that

$$Q_u(dy) \sim \exp(-\Phi(u; y)) Q_0(dy)$$

- iii show that Φ is ν_0 measurable
 - iv show that the normalization constant $Z \equiv Z(y)$ is positive almost surely with respect to $y \sim Q_0$
- ~ this needs to be done on a case-by-case basis (in principle).

Remark 22.11

The stated version of Bayes' rule only asserts that the posterior is absolutely continuous with respect to the prior. Equivalence will occur when $\Phi(\cdot; y)$ is finite on U .

2.1 Finite data case

For the special case of finite-dimensional data, i.e., $Y = \mathbb{R}^m$, $m \in \mathbb{N}$, the assumptions made above can be verified so that we restate Bayes' rule as:

Corollary 22.15

Suppose $G : U \rightarrow \mathbb{R}^m$ is continuous and that Q_0 has an absolutely continuous Lebesgue density on \mathbb{R}^m . If $u \sim \mu_0$, then $u|y \sim \mu^y$ with $\mu^y \ll \mu_0$ and

$$\frac{d\mu^y}{d\mu_0}(u) \sim \exp(-\Phi(u; y)).$$

Exercise 22.4

Explain why the assumptions in Corollary 22.15 imply the general measurability hypothesis in Bayes' rule.

Let $\rho : \mathbb{R}^m \rightarrow \mathbb{R}$ be the Lebesgue density of $\eta \sim Q_0$, that is,

$$\frac{dQ_0}{d\lambda}(y) = \rho(y).$$

From the observation model

$$y = G(u) + \eta \Leftrightarrow y - G(u) = \eta$$

we see that the measure Q_u (i.e., the translate of Q_0) has the Lebesgue density

$$\frac{dQ_u}{d\lambda}(y) = \rho(y - G(u))$$

$$Q_u(x) = \mathbb{P}(y \leq x | u) = \mathbb{P}(\eta \leq x - G(u) | u)$$

Consequently, the **potential** $\Phi : U \times \mathbb{R}^m \rightarrow \mathbb{R}$ is given by

$$\frac{\rho(y - G(u))}{\rho(y)} \sim \exp(-\Phi(u; y)),$$

provided that $\text{supp}(\rho) = \mathbb{R}^m$.

Example: If the noise is $\eta \sim \mathcal{N}(0, Q) \equiv Q_0$ for some symmetric positive-definite matrix $Q \in \mathbb{R}^{m \times m}$, then

$$\Phi(u; y) = \frac{1}{2} \|y - G(u)\|_{Q^{-1}}^2$$

and the posterior distribution is given by

$$\mu^y(\mathrm{d}u) \sim \exp\left(-\frac{1}{2} \|y - G(u)\|_{Q^{-1}}^2\right) \mu_0(\mathrm{d}u).$$

Remark 22.12

In the special case that U is also finite-dimensional, e.g., $U = \mathbb{R}^n$, and if the prior has a Lebesgue density $\mathrm{d}\mu_0 = \pi_0 \mathrm{d}\lambda$, then we recover the usual (finite-dim.) Bayes' rule. Indeed, the posterior μ^y then also has a Lebesgue density $\mathrm{d}\mu^y = \pi^y \mathrm{d}\lambda$, given by

$$\pi^y(u) = \frac{\rho(y - G(u)) \pi_0(u)}{\int_{\mathbb{R}^n} \rho(y - G(u)) \pi_0(u) \lambda(\mathrm{d}u)},$$

where $\rho = \mathrm{d}Q_0 / \mathrm{d}\lambda$ as before.

Observe that μ^y is usually not Gaussian, even if both μ_0 and Q_0 are Gaussian measures.

Exercise 22.5

Consider a Gaussian probability measure μ_0 on \mathbb{R}^n . Show that the posterior

$$\mu^y(\mathrm{d}u) \sim e^{-\Phi(u;y)} \mu_0(\mathrm{d}u)$$

is also a Gaussian measure on \mathbb{R}^n , if $\Phi(u; y)$ is quadratic in u .

Extend this result from $U = \mathbb{R}^n$ to U being a separable Banach space.

3. Well-posedness and Approximation

Now that we have established a rigorous infinite-dimensional analogue of Bayes' rule, we can discuss the well-posedness of the Bayesian inversion on Banach spaces. For simplicity, we will restrict our attention to **Gaussian prior measures**.

We will take the following as our standard assumptions on the potential Φ (i.e., the negative data log-likelihood). In essence, we wish to restrict our attention to potentials Φ that:

- ▶ are **Lipschitz** in both arguments;
- ▶ are **bounded** on bounded sets;
- ▶ do not decay to $-\infty$ at infinity “too fast”.

These conditions are, more precisely, formulated as:

Assumptions on Φ :

Assumption 22.1 (Assumptions on Φ)

Suppose that the potential $\Phi : U \times Y \rightarrow \mathbb{R}$ satisfies:

(Decay) For every $\varepsilon > 0$ and $r > 0$ there is an $M = M(\varepsilon, r) \in \mathbb{R}$ such that for all $u \in U$ and all $y \in Y$ with $\|y\|_Y < r$:

$$\Phi(u; y) \geq M - \varepsilon \|u\|_U^2.$$

(Bdd.) For every $r > 0$ there is $K = K(r) > 0$ such that for all $u \in U$ and $y \in Y$ with $\max \{\|u\|_U, \|y\|_Y\} < r$: $\Phi(u; y) \leq K$.

(Lip. in u) For every $r > 0$ there exists an $L = L(r) > 0$ such that for all $u_1, u_2 \in U$ and all $y \in Y$ with $\|u_1\|_U, \|u_2\|_U, \|y\|_Y < r$: $|\Phi(u_1; y) - \Phi(u_2; y)| \leq L \|u_1 - u_2\|_U$.

(Lip. in y) For every $\varepsilon > 0$ and $r > 0$ there is $C = C(\varepsilon, r) > 0$ such that for all $u \in U$ and all $y_1, y_2 \in Y$ with $\|y_1\|_Y, \|y_2\|_Y < r$: $|\Phi(u; y_1) - \Phi(u; y_2)| \leq \exp(\varepsilon \|u\|_U^2 + C) \|y_1 - y_2\|_Y$.

For the special case that Y is finite dimensional, e.g., $Y = \mathbb{R}^m$, and $Q_0 = \mathcal{N}(0, Q)$ non-degenerate, we know that:

$$\Phi(u; y) = \frac{1}{2} \|y - G(u)\|_{Q^{-1}}^2.$$

It is then natural to directly impose conditions on the observation operator $G : U \rightarrow \mathbb{R}^m$.

Assumption 22.2 (Assumptions on G for finite data)

The observation operator $G : U \rightarrow \mathbb{R}^m$ satisfies:

(Decay-G) *For every $\varepsilon > 0$ there exist an $M = M(\varepsilon) > 0$ such that for all $u \in U$:*

$$\|G(u)\|_{Q^{-1}} \leq \exp(\varepsilon \|u\|_U^2 + M).$$

(Bdd.-G) *For every $r > 0$ there is $K = K(r) > 0$ such that for all $u_1, u_2 \in U$ with $\max\{\|u_1\|_U, \|u_2\|_U\} < r$:*

$$\|G(u_1) - G(u_2)\|_{Q^{-1}} \leq K \|u_1 - u_2\|_U.$$

Lemma 22.16

Assume that $G : U \rightarrow \mathbb{R}^m$ satisfies the finite data assumptions (Decay – G) and ($Bdd.$ – G). Then $\Phi : U \times \mathbb{R}^m \rightarrow \mathbb{R}$ given by $\Phi(u; y) := \frac{1}{2} \|y - G(u)\|_{Q^{-1}}^2$ satisfies the standard assumptions on Φ with $(Y, \|\cdot\|_Y) = (\mathbb{R}^m, \|\cdot\|_{Q^{-1}})$.

Exercise 22.6

Prove the Lemma.

Next, we show that, for Gaussian priors, the standard assumptions on Φ yield a well-defined posterior measure for each instance of the observed data.

Theorem 22.17 (well-defined measure)

Let $\Phi : U \times Y \rightarrow \mathbb{R}$ satisfy the standard assumptions (Decay), (Bdd.), and (Lip. in u). Furthermore, let μ_0 be a Gaussian measure on U . Then, for each $y \in Y$, the posterior measure μ^y given by

$$\frac{d\mu^y}{d\mu_0}(u) = \frac{\exp(-\Phi(u; y))}{Z(y)}, \quad Z(y) = \int_U \exp(-\Phi(u; y)) \mu_0(du),$$

is a well-defined measure on U .

Proof: Assumption (*Bdd.*) implies that $Z(y)$ is bounded below. Indeed:

$$\begin{aligned} Z(y) &= \int_U \exp(-\Phi(u; y)) \mu_0(\mathrm{d}u) \\ &\geq \int_{\{u: \|u\|_U \leq r\}} \exp(-\Phi(u; y)) \mu_0(\mathrm{d}u) \\ &\stackrel{(\textit{Bdd.})}{\geq} \int_{\{u: \|u\|_U \leq r\}} \exp(-K(r)) \mu_0(\mathrm{d}u) = e^{-K(r)} \mu_0(\{u : \|u\|_U \leq r\}) \\ &> 0 \end{aligned}$$

for $r > 0$, since μ_0 is a strictly positive measure on U . It follows from (*Lip. in u*) that $\Phi(\cdot; y)$ is μ_0 -measurable, so that μ^y is a well-defined measure. It remains to verify that μ^y is a probability measure.

By (Decay), for $\|y\|_Y \leq r$ and some $\varepsilon > 0$, we find

$$\begin{aligned} Z(y) &= \int_U \exp(-\Phi(u; y)) \mu_0(\mathrm{d}u) \\ &\leq \int_U \exp(\varepsilon \|u\|_U - M(\varepsilon, r)) \mu_0(\mathrm{d}u) \\ &\leq C \exp(-M(\varepsilon, r)) < \infty, \end{aligned}$$

since μ_0 is Gaussian, and we may select $\varepsilon > 0$ small enough, so that Fernique's Theorem applies (see Thm (Fernique, 1970)). It thus follows that μ^y can be normalized indeed. □

Next, we discuss the **continuous dependence** on the observed data. To do this, we need to assess the “distance” between two probability distributions. One way is to use the following distance.

Definition 22.18 (Hellinger distance)

Let μ and ν be σ -finite measures on a common measurable space $(\mathcal{X}, \mathcal{F})$. The **Hellinger distance** between μ and ν is

$$d_H(\mu, \nu) := \sqrt{\int_{\mathcal{X}} \frac{1}{2} \left| \sqrt{\frac{d\mu}{d\nu}} - 1 \right|^2 d\nu} = \sqrt{\int_{\mathcal{X}} \frac{1}{2} \left| \sqrt{\frac{d\mu}{d\rho}} - \sqrt{\frac{d\nu}{d\rho}} \right|^2 d\rho}$$

for any reference measure ρ on $(\mathcal{X}, \mathcal{F})$ with respect to which both μ and ν are absolutely continuous.

Remark 22.13

- a One can show that d_H is independent of the particular reference measure ρ (e.g., such as $\mu + \nu$); see [Sullivan, Lemma 5.8]
- b The Hellinger distance and the total variation metric induce the same topology on the space of probability measures:

$$d_H(\mu, \nu)^2 \leq d_{\text{TV}}(\mu, \nu) \leq 2 d_H(\mu, \nu).$$

The Hellinger distance offers the following useful property.

Lemma 22.19

Let $(V, \|\cdot\|)$ be a Banach space and suppose that $f : (\Omega, \mathcal{A}) \rightarrow V$ is such that $f \in L^2(\Omega, \mathcal{A}, \mu; V) \cap L^2(\Omega, \mathcal{A}, \nu; V)$. Then:

$$\|\mathbb{E}_\mu(f) - \mathbb{E}_\nu(f)\| \leq 2 \sqrt{\mathbb{E}_\mu(\|f\|^2) + \mathbb{E}_\nu(\|f\|^2)} d_H(\mu, \nu).$$

Exercise 22.7

Prove the Lemma.

Now we can state the result regarding the continuous dependence of the posterior with respect to the input perturbations.

Theorem 22.20 (continuous dependence on the data)

Let Φ satisfy standard Assumptions (Decay), (Bdd.), and (Lip. in y).

Suppose further that μ_0 is a Gaussian probability measure on U , and that
 $\hookrightarrow \mu_0(U) = 1$

$\mu^y \ll \mu_0$ with Radon–Nikodym derivative given by the generalized Bayes' rule for each $y \in Y$. Then there exists a constant $C > 0$, such that, for all $y, y' \in Y$:

$$d_H(\mu^y, \mu^{y'}) \leq C \|y - y'\|_Y.$$

Proof: From the well-posedness proof, we know that (Bdd.) gives a uniform lower bound on $Z(y)$. We also have the following Lipschitz bound for the difference of the normalization constants for y and y' :

$$\begin{aligned} |Z(y) - Z(y')| &\leq \int_U \left| e^{-\Phi(u;y)} - e^{-\Phi(u;y')} \right| \mu_0(du) \\ &\leq \int_U \max \left\{ e^{-\Phi(u;y)}, e^{-\Phi(u;y')} \right\} |\Phi(u; y) - \Phi(u; y')| \mu_0(du) \end{aligned}$$

by the mean-value theorem.

By (Decay) and (Lip. in y) we thus find

$$\begin{aligned} |Z(y) - Z(y')| &\leq \int_U \exp(\varepsilon \|u\|_U^2 - M) \cdot \exp(\varepsilon \|u\|_U^2 + C) \|y - y'\|_Y \mu_0(du) \\ &\leq C \|y - y'\|_Y \end{aligned}$$

└ can be chosen suff. small ($r = \max(\|y\|, \|y'\|)$)

using Fernique's Thm (see Thm (Fernique, 1970)).

Using the Hellinger distance with μ_0 as reference measure, we obtain

$$\begin{aligned} d_H(\mu^y, \mu^{y'}) &= \int_U \left| \frac{1}{\sqrt{Z(y)}} e^{\frac{-\Phi(u;y)}{2}} - \frac{1}{\sqrt{Z(y')}} e^{\frac{-\Phi(u;y')}{2}} \right|^2 \mu_0(du) \\ &= \frac{1}{Z(y)} \int_U \left| e^{\frac{-\Phi(u;y)}{2}} - \sqrt{\frac{Z(y)}{Z(y')}} e^{\frac{-\Phi(u;y')}{2}} \right|^2 \mu_0(du) \\ &\leq I_1 + I_2, \end{aligned}$$

where

$$\begin{aligned} I_1 &:= \frac{1}{Z(y)} \int_U \left| e^{\frac{-\Phi(u;y)}{2}} - e^{\frac{-\Phi(u;y')}{2}} \right|^2 \mu_0(du) \\ I_2 &:= \left| \frac{1}{\sqrt{Z(y)}} - \frac{1}{\sqrt{Z(y')}} \right| \int_U e^{\frac{-\Phi(u;y')}{2}} \mu_0(du). \end{aligned}$$

Similar to bounding $|Z(y) - Z(y')|$ above, the mean-value theorem together with (Decay) and (Lip. in y), and Fernique's theorem yield:

$$\begin{aligned} I_1 &\leq \frac{1}{Z(y)} \int_U \max \left\{ \frac{1}{2} e^{\frac{-\Phi(u;y)}{2}}, \frac{1}{2} e^{\frac{-\Phi(u;y')}{2}} \right\}^2 |\Phi(u; y - \Phi(u; y'))|^2 \mu_0(\mathrm{d}u) \\ &\leq \frac{1}{4Z(y)} \int_U e^{2 \cdot \frac{\varepsilon}{2} \|u\|_U^2 - \frac{M}{2} \cdot 2} e^{2 \cdot \varepsilon \|u\|_U^2 + 2C} \|y - y'\|_Y^2 \\ &\leq C \|y - y'\|_Y^2. \end{aligned}$$

Furthermore, it follows from (Lip. in y) and Fernique's Thm., that

$$\int_U e^{\frac{-\Phi(u;y')}{2}} \mu_0(\mathrm{d}u) \leq C < \infty.$$

As $Z(y) > 0 \forall y \in Y$, the mean-value theorem implies:

$$\begin{aligned} \left| \frac{1}{\sqrt{Z(y)}} - \frac{1}{\sqrt{Z(y')}} \right|^2 &\leq C \max \left\{ \frac{1}{Z(y)^3}, \frac{1}{Z(y')^3} \right\} |Z(y) - Z(y')|^2 \\ &\leq C \|y - y'\|_Y^2, \end{aligned}$$

so that $I_2 \leq C \|y - y'\|_Y^2$, which completes the proof. □

Similarly, the next theorem asserts the well-posedness with respect to Gaussian priors when only an approximate observational (forw.) operator $G_h : U \rightarrow Y$ is available, instead of $G : U \rightarrow Y$. For example, G_h could be the FE-based approximation of G . For brevity, we suppress the dependence on the observed data $y \in Y$ below, which we assumed to be fixed.

Theorem 22.21 (well-posedness for approximate forw. operator)

Let μ and μ^h be the posterior probability measures arising from the potentials Φ and Φ^h , respectively. Suppose that μ and μ^h are absolutely continuous with respect to μ_0 for all $h > 0$. Moreover, suppose that Φ and Φ^h satisfy the standard assumptions (Decay) and (Bdd.) uniformly in $h > 0$. If, for all $\varepsilon > 0$, there exists $K = K(\varepsilon) > 0$ such that

$$|\Phi(u; y) - \Phi^h(u; y)| \leq K \exp(\varepsilon \|u\|_U^2) \psi(h),$$

with $\lim_{h \rightarrow 0} \psi(h) = 0$, then there is a constant $C > 0$, independent of h and

$$d_H(\mu, \mu^h) \leq C \psi(h).$$

Exercise 22.8

Prove Theorem 22.21.

Definition 22.22 (Wasserstein distance)

Consider a metric space (\mathcal{X}, d) and let

$$M_p(\mathcal{X}) := \left\{ \mu : \int_{\mathcal{X}} d(x, y)^p \mu(dx) < \infty \right\}, \quad p \geq 1.$$

The p -Wasserstein distance of the probability measures $\mu, \nu \in M_p(\mathcal{X})$ is defined as

$$W_p(\mu, \nu) := \inf_{\gamma \in \Gamma(\mu, \nu)} \left(\int_{\mathcal{X} \times \mathcal{X}} d(x, y)^p \gamma(dx, dy) \right)^{1/p},$$

where

$$\Gamma(\mu, \nu) := \left\{ \gamma : \int_{\mathcal{X}} \gamma(x, y) dy = \mu(x), \int_{\mathcal{X}} \gamma(x, y) dx = \nu(y) \right\}$$

is the set of all couplings between μ and ν .

Remark 22.14

Under the assumptions studied above, we have seen that the posterior μ^y will be well-posed, in the sense of being well-defined and being absolutely continuous with respect to the prior μ_0 for any observed data $y \in Y$. Hence, μ^y cannot associate positive weight with events that are null-sets under the prior.

However, in infinite-dimensional spaces it can be very difficult for probability measures to be absolutely continuous with respect to each other (see Theorem 22.9). Therefore, the choice of an infinite-dimensional prior μ_0 is a very strong modelling assumption. In fact, if done “wrong”, cannot be “corrected” even by large amounts of data. Hence, it is not reasonable to expect that Bayesian inference on function spaces should be well-posed with respect to “small” perturbations to the prior μ_0 , e.g., by a shift of mean that lies outside the Cameron–Martin space (see Definition 22.8), at least with respect to the Hellinger distance. In fact, recent results indicate that one has such a prior robustness with respect to another distance (Wasserstein distance) under suitable conditions; see [Ernst, Pichler, Sprungk. 2022]. Nonetheless, the general practical significance of the infinite-dimensional well-posedness result is immense, as it allows the design of finite-dimensional (i.e., discretized) Bayesian inverse problems with good stability properties with respect to the (discretization) dimension.

4. Algorithms

For a given observation $y \in Y$, we have seen that the posterior μ^y on U is given by

$$\frac{d\mu^y}{d\mu_0} = \frac{1}{Z(y)} \exp(-\Phi(\cdot; y)).$$

The Bayesian approach to inverse problems thus entails characterizing the poster distribution. For example, we may want to extract the MAP or the (conditional) mean and/or covariance of μ^y .

The question of how to effectively access the Bayesian posterior is still an active field of applied mathematical research. Indeed, there are various approaches. Here, we will briefly discuss (basic) **Markov Chain Monte Carlo (MCMC)** techniques.

For these sampling-based techniques, the precise form of the posterior as $d\mu^y = Z(y)^{-1} e^{-\Phi(\cdot; y)} d\mu_0$ is irrelevant. Instead, we will simply refer to that measure as the **target distribution** in the context of MCMC and denote it as μ .

MCMC methods are based on a **simple idea**:

- ▶ design a Markov chain $(u_k)_{k \geq 0}$ that is distributed according to the target μ .

This simple idea leads to a very broad class of algorithms with great potential for innovation.

Here, we will, for simplicity, focus on the particular class of methods known as **Metropolis-Hastings** (MH) methods.

Key ingredient of these methods is a probability measure on U , parametrized by $u \in U$, specifically, a **Markov transition kernel** $q(u, dv)$. The kernel is used to propose a move from the current state u_K , say, to a potentially new state distributed as $q(u_k, \cdot)$. The proposed state is then **accepted or rejected** based on a criterion that depends on the target μ . When constructed correctly, the resulting Markov chain has the desired property of preserving the target. Hence, the key aspect is selecting a “good” proposal q .

4.1 Construction and Examples

Before we discuss two simple possibilities for selecting the kernel q , let's make the idea outlined above concrete.

Given a kernel $q(\cdot, \cdot)$ and a target μ , we define a new product measure on $U \times U$ by

$$\underbrace{\nu(\mathrm{d}u \otimes \mathrm{d}v)}_{\equiv \nu(\mathrm{d}u, \mathrm{d}v)} = q(u, \mathrm{d}v)\mu(\mathrm{d}u).$$

We then define also another product measure by simply reversing the roles of current state u and proposed state v :

$$\nu^\top(\mathrm{d}v \otimes \mathrm{d}u) = q(v, \mathrm{d}u)\mu(\mathrm{d}v).$$

If q is such that $\nu^\top \ll \nu$, then we may introduce

$$\alpha(u, v) := \min \left(1, \frac{\mathrm{d}\nu^\top}{\mathrm{d}\nu}(u, v) \right).$$

Next, we define a random variable $\gamma \equiv \gamma(u, v)$, independent of the probability space underlying the transition kernel q , with the property that

$$\gamma(u, v) = \begin{cases} 1, & \text{with probability } \alpha(u, v); \\ 0, & \text{otherwise.} \end{cases}$$

Based on this on-off characteristic, we eventually define the Markov chain $(u_k)_{k \geq 0}$ recursively by

$$(MH - MCMC) \left\{ \begin{array}{ll} \cdot \text{given} & u_k; \\ \cdot \text{propose} & v_k \sim q(u_k, \cdot); \\ \cdot \text{set} & u_{k+1} = \gamma(u_k, v_k)v_k + (1 - \gamma(u_k, v_k))u_k. \end{array} \right.$$

Observe that this construction gives rise to a well-defined Markov chain on U , if we choose the randomness in the proposal v_k and the one in $\gamma(u_k, v_k)$ independent of each other, and independently of their values for different (iterations) k .

Before going into technical details, under certain conditions on $\alpha(\cdot, \cdot)$ (and hence q and μ), one can show that:

- ▶ the target μ is an invariant measure for the Markov chain $(u_k)_{k \geq 0}$: if $u_0 \sim \mu$, then $u_k \sim \mu \ \forall k \geq 0$
- ▶ the Markov chain is ergodic: for any $M \geq 0$

$$\frac{1}{N} \sum_{K=1}^N \phi(u_{K+M}) \xrightarrow{N \rightarrow \infty} \int_U \phi(u) \mu(du),$$

for μ_0 -a.a. $u_0 \in U$ and all $\phi \in C_b(U)$. That is, the empirical distribution induced by the Markov chain converges (weakly) to that of the target measure μ .

It remains to address how to actually construct an MH-MCMC. If $U = \mathbb{R}^n$ and the target μ has a well-defined, positive Lebesgue-density, then this is “easy” as any transition kernel $q(u, dv)$ will do, provided it too has a positive Lebesgue density. This is because it follows that ν^\top and ν are equivalent (see also below).

Example: A widely used (because simple) proposal kernel is simply that of a **random walk**; if $\mu_0 = \mathcal{N}(0, C)$ it is natural to propose new candidates via

$$v = u + \sqrt{2\delta}\xi, \quad \xi \sim \mathcal{N}(0, C),$$

for some $\delta > 0$. That is,

$$q(u, \cdot) = \mathcal{N}(u, 2\delta C).$$

For a target measure of the form

$$d\mu = \frac{1}{Z} \exp(-\Phi(\cdot; y)) d\mu_0,$$

it follows that

$$\alpha(u, v) = \min \{1, \exp(I(u) - I(v))\},$$

where $I(u) := \frac{1}{2} \|u\|_{C^{-1}}^2 + \Phi(u)$.

Consequently, if the proposed state v corresponds to a lower value of the **regularized least-squares functional** I (i.e., $I(u) - I(v) > 0$), then the proposed state is automatically accepted; otherwise it will be accepted with a probability depending on $I(u) - I(v)$.

The parameter $\delta > 0$ controls the size of the proposed moves (recall that these are not informed):

- ▶ δ large \rightsquigarrow proposals unlikely to be accepted
 - \rightsquigarrow high correlation in Markov chain
- ▶ δ small \rightsquigarrow also high correlation due to dependence

The key aspect is to find a good compromise.

More advanced MH-MCMC algorithms improve upon the simple random walk by incorporating “more information”, such as $D\Phi$ (cf. MALA), into the proposal step as an attempt to facilitate moves to regions of higher probability.

In infinite-dimensions, things are not so straightforward anymore: a random walk will typically not provide the required condition that $\nu^\top \ll \nu$. For example, if $\mu_0 = \mathcal{N}(0, C)$ and U is infinite-dimensional, then the random walk proposal $q(u, \cdot) = \mathcal{N}(u, 2\delta C)$ will not satisfy the property (cf., (22.8)). However, a simple fix can help already.

Example: To ensure the desired absolute continuity of ν^\top with respect to ν , one can simply modify the proposal to:

$$\nu = (1 - 2\delta)^{\frac{1}{2}} u + \sqrt{2\delta} \xi, \quad \xi \sim \mathcal{N}(0, C),$$

where $0 < \delta \leq \frac{1}{2}$. That is,

$$q(u, \cdot) = \mathcal{N}\left((1 - 2\delta)^{\frac{1}{2}} u, 2\delta C\right).$$

This proposal will satisfy the desired condition for any $\delta > 0$. For a target measure of the form

$$d\mu = \frac{1}{Z} \exp(-\Phi(\cdot; y)) d\mu_0, \quad \mu_0 = \mathcal{N}(0, C),$$

the corresponding acceptance probability is

$$\alpha(u, v) = \min \{1, \exp(\Phi(u) - \Phi(v))\}.$$

Therefore, one automatically accepts the proposed state v , if $\Phi(v) < \Phi(u)$. In the context of Bayesian inversion, where the target is of said form, this proposal is also called preconditioned [Crank–Nicolson proposal](#).

To intuitively see why this modification allows to view it as an appropriate analogue of the random walk for the infinite-dimensional setting, observe that: if $u \sim \mathcal{N}(0, C)$ and $v = (1 - 2\delta)^{\frac{1}{2}} u + \sqrt{2\delta}\xi$ with $\xi \sim \mathcal{N}(0, C)$ ($u \perp \xi$), and $0 < \delta \leq 1/2$, then $v \sim \mathcal{N}(0, C)$. That is, **the proposal preserves the underlying reference** (i.e., prior) **measure μ_0** . This is not the case for the classic random walk: if $u \sim \mathcal{N}(0, C)$ and $v = u + \sqrt{2\delta}\xi$ with $\xi \sim \mathcal{N}(0, C)$ ($u \perp \xi$), then $v \sim \mathcal{N}(0, (1 + 2\delta)C)$. Note that $\delta = \frac{1}{2}$ gives the preconditioned Crank–Nicolson proposal

$$v = (1 - 2\delta)^{\frac{1}{2}} u + \sqrt{2\delta}\xi = \xi,$$

which is known as an **independence sampler**, where the proposal is drawn from the prior measure $\xi \sim \mu_0$, independently of the current state u of the chain.

Finally, we mention that there are also various improved proposals for the infinite-dimensional setting that incorporate “more information” about the target, e.g., by using $D\Phi$.

Remark 22.15

In computational practice, we always implement an MCMC sampling method in finite-dimensions, of course. The error incurred on the Bayesian posterior in doing so is quantified in the previous section. We emphasize that the value of deriving MH-MCMC methods in infinite-dimensions is that any MH-MCMC method in finite-dimensions that does not correspond to a well-defined limiting MH-MCMC method in the infinite-dimensional limit (i.e., a function space setting) will degenerate as the number of dimensions increases. The essential take home message therefore is: the function space viewpoint to MCMC is not just useful, but essential, and allows developing improved algorithms.

Further reading:

- ▶ Cotter et al. MCMC algorithms for functions: Modifying old algorithms to make them faster. *Stat. Science.* 2013, vol 28(3), pp. 424–446