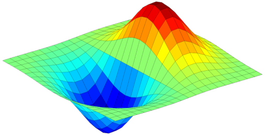
Solution of PDEs with uncertainties in parameters by the stochastic Galerkin method with geotechnical applications



SNA'23 January 23 - 27, 2023

Michal Béreš[1,2]

January 24, 2023

Institute of Geonics, Czech Academy of Sciences

Department of Applied Mathematics, FEECS, VŠB-TUO

prof. Radim Blaheta

千年古

肯德

Outline

Motivation

Deterministic case

Random variables/vectors/fields

Uncertainties in parameters

Stochastic Galerkin method

Assembling the SGM matrix

Solving the system

Use of SGM solution

Example - TSX experiment

Motivation

Uncertainty Quantification and SGM

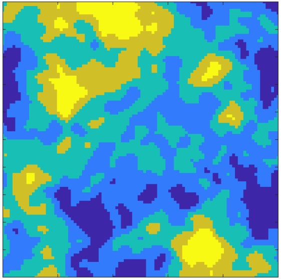
- models of geotechnical systems often have inherent uncertainty in the input data
    - unknown material parameters, imprecise measurements
    - e.g. rock properties for specific model
- probabilistic solutions to PDEs with random/uncertain inputs are useful in various applications
    - e.g. uncertainty quantification, sensitivity analysis, and design optimization
- can be solved using various methods such
    - Monte Carlo sampling, collocation methods
    - stochastic Galerkin method
- problems with uncertainties are much more computationally expensive
    - we need to exploit the structures of the problem
    - stochastic Galerkin method is more efficient for specific types of problem

Model uncertainty used here

- hydraulic conductivity field with unknown material properties

$$k\left(x,\boldsymbol{Z}\right)=\sum_{m=1}^{M_k} 1_{\mathcal{D}_m}\left(x\right)\exp\left(\sigma_m Z_m+\mu_m\right)$$

  i.e. we know where certain type of materials are, but do not know their exact properties



Deterministic case

Stationary Darcy flow

$$\begin{cases} -\operatorname{div}\left(k\left(x\right)\nabla u\left(x\right)\right) = f\left(x\right) & x \in \mathcal{D} \\ u\left(x\right) = u_0\left(x\right) & x \in \Gamma_D \\ -k\left(x\right)\frac{\partial u(x)}{n(x)} = g\left(x\right) & x \in \Gamma_N \end{cases}, \tag{1}$$

- $\mathcal{D} \subset \mathbb{R}^d$ $(d = 1, 2, 3)$ is a Lipschitz domain,
- $k\left(x\right)$ is a permeability field,
- $f\left(x\right)$ is a volume source,
- $u_0\left(x\right)$ are prescribed pressures on the Dirichlet boundary Γ_D,
- $g\left(x\right)$ represents sources on the Neumann boundary Γ_N.

Weak form

Find $u_H \in H^1_{0,\Gamma_D}(\mathcal{D})$, $\left(u = u_0 + u_H \in H^1(\mathcal{D})\right)$:

$$\begin{aligned}
a(u_H, v) &= b(v), \ \forall v \in H^1_{0,\Gamma_D}(\mathcal{D}), \\
a(u_H, v) &= \int_{\mathcal{D}} k(x)\nabla u_H(x) \cdot \nabla v(x)\, dx, \\
b(v) &= \int_{\mathcal{D}} f(x)\, v(x)\, dx - \int_{\Gamma_N} g(x)\, v(x)\, dx - \int_{\mathcal{D}} k(x)\nabla u_0(x) \cdot \nabla v(x)\, dx
\end{aligned}$$

- $k \in L^\infty(\mathcal{D})$, $0 < k_{\min} \leq k(x) \leq k_{\max} < \infty \ \forall x \in \mathcal{D}$,
- $f \in L^2(\mathcal{D})$,
- $u_0 \in H^1(\mathcal{D})$,
- $g \in L^2(\Gamma_N)$ (or $g \in H^{-1/2}(\Gamma_N)$).

Well-posedness

Proof of well-posedness via the Lax-Milgram theorem. We need:

- $H^1_{0,\Gamma_D}(\mathcal{D})$ to be a Hilbert space
- $a(u,v)$ to be continuous and elliptical bilinear operator

$$\exists C > 0 \; \forall u, v \in H^1_{0,\Gamma_D}(\mathcal{D}) : |a(u,v)| \leq C \|u\| \|v\|$$

$$\exists c > 0 \; \forall u \in H^1_{0,\Gamma_D}(\mathcal{D}) : a(u,u) \geq c \|u\|^2$$

- $b(v) \in H^{-1}_{0,\Gamma_D}(\mathcal{D})$

For the current problem all of these requirements are fulfilled and the problem is well-posed.

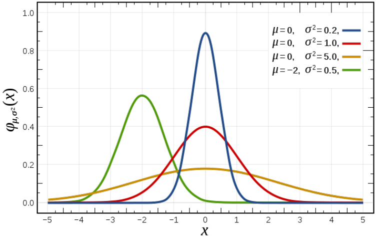
Random variables/vectors/fields

Random variables

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space

- continuous random variable Z is a map from sample space $\Omega \rightarrow \mathbb{R}$
- can be described by the probability density $f : \mathbb{R} \rightarrow \mathbb{R}_0^+$, $\int\limits_{\mathbb{R}} f(x)\,\mathrm{d}x = 1$
- distribution of a random variable Z defines a probability measure

$$\int\limits_{\Omega} g(Z(\omega))\, d\mathbb{P}(\omega) = \int\limits_{\mathbb{R}} g(Z)\,\mathrm{d}FZ = \int\limits_{\mathbb{R}} g(z)\, f(z)\,\mathrm{d}z = \mathbb{E}(g(Z))$$



Random vectors

- random vector is a vector of M (let assume continuous) random variables $\boldsymbol{Z} = (Z_1, \ldots, Z_M)$
- can be described by a joint probability density $f_{\boldsymbol{Z}} : \mathbb{R}^M \to \mathbb{R}_0^+$, $\int\limits_{\mathbb{R}^M} f_{\boldsymbol{Z}}(x) \, \mathrm{d}x = 1$
- random vector of **independent** random variables has a joint probability density in form

$$f_{\boldsymbol{Z}}(\boldsymbol{z}) = \prod_{i=1}^{M} f_{Z_i}(z_i)$$

Random fields

- real-valued random field $\{X(t) : t \in \mathcal{T}\}$ $X(t) : \Omega \to \mathbb{R}$,
    - indexed set of real valued random variables
- can be viewed as a function on both $\mathcal{T}, \Omega$: $X : \mathcal{T} \times \Omega \to \mathbb{R}$
- can be viewed as an H-valued random variable $X : \Omega \to \mathbb{R}^{\mathcal{T}}$
    - $\mathbb{R}^{\mathcal{T}}$ denotes a set of functions $\mathcal{T} \to \mathbb{R}$
    - important question is the regularity of random field = what properties does $\mathbb{R}^{\mathcal{T}}$ have
        - for some, can be answered by inspecting the properties of its covariance function
        - e.g. $L^2(\mathcal{T}), C(\mathcal{T}), \ldots$

Spaces on Ω

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space

- most often we need “just” square integrable functions/random variables
    - square integrable real valued random variable X fulfills

$$\int\limits_{\Omega} X(\omega)^2 \, d\mathbb{P}(\omega) < \infty$$

        - space of all real valued square integrable random variables $L^2(\Omega)$ (sometimes we stretch this notation to the space of random vectors)
    - square of integrable functions of a random vector $\boldsymbol{Z}$ creates the space

$$L^2_{\mathrm{d}F\boldsymbol{Z}}\left(\mathbb{R}^M\right) := \left\{ f : \mathbb{R}^M \to \mathbb{R} : \int\limits_{\Omega} f(\boldsymbol{Z}(\omega))^2 \, d\mathbb{P}(\omega) = \int\limits_{\mathbb{R}^M} f(\boldsymbol{Z})^2 \, dF\boldsymbol{Z} < \infty \right\}$$

Spaces on Ω

- Random field $X : \Omega \to V$ (V is normed space) is square integrable (second order) random field if

$$\int\limits_{\Omega} \|X(\omega)\|_V^2 \, d\mathbb{P}(\omega) < \infty$$

  - with corresponding space $L^2(\Omega, V)$

- second order random field $X : \Omega \to L^2(\mathcal{T})$, $X \in L^2\left(\Omega, L^2(\mathcal{T})\right) \approx L^2(\Omega) \otimes L^2(\mathcal{T})$
- Karhunen-Loève decomposition

$$X(\omega, x) = \mu_X(x) + \sum_{j=1}^{\infty} \sqrt{\lambda_j} \psi_j(x) \xi_j(\omega)$$

  - as $\sqrt{\lambda_j}$ decreases, we can truncate the sum (many standard random fields have error estimates for truncation)

Uncertainties in parameters

Darcy flow with uncertainties (infinite)

$$\begin{cases} -\mathrm{div}_x\left(k\left(x,\omega\right)\nabla_x u\left(x,\omega\right)\right) = f\left(x,\omega\right), & \forall\omega\in\Omega, \forall x\in\mathcal{D} \\ u\left(x,\omega\right) = u_0\left(x,\omega\right), & \forall\omega\in\Omega, \forall x\in\Gamma_D \\ -k\left(x,\omega\right)\frac{\partial u(x,\omega)}{\partial n(x)} = g\left(x,\omega\right), & \forall\omega\in\Omega, \forall x\in\Gamma_N \end{cases},$$

$k\left(x,\omega\right)$, $f\left(x,\omega\right)$, $u_0\left(x,\omega\right)$, $g\left(x,\omega\right)$, $u\left(x,\omega\right) : \mathcal{D}\times\Omega\to\mathbb{R}$ are understood as random fields

- **there is no uncertainty in geometry!** (very different types of problem, out of scope for this talk)
- we obtain a deterministic system for each $\omega\in\Omega$
    - necessary requirements is that each of these systems is well-posed
    - $u\left(x,\omega\right)$ can be then viewed as mapping of the sample set to deterministic solutions $u:\Omega\to H^1\left(\mathcal{D}\right)$, i.e. the path-wise solution

Darcy flow with uncertainties (discretized)

- An infinite dimensional case cannot be solved directly, we need to replace random fields with functions of random variables.
- e.g. via the Karhunen-Loève decomposition or the projection into orthogonal polynomials

We obtain similar problem, but "parametric dimension" is now finite = there are M parameters forming a random vector

$$\begin{cases} -\mathrm{div}_x \left(k\left(x, \boldsymbol{Z}\left(\omega\right)\right) \nabla_x u\left(x, \boldsymbol{Z}\left(\omega\right)\right)\right) = f\left(x, \boldsymbol{Z}\left(\omega\right)\right), & \forall x \in \mathcal{D}, \boldsymbol{Z} \in \mathbb{R}^M \\ u\left(x, \boldsymbol{Z}\left(\omega\right)\right) = u_0\left(x, \boldsymbol{Z}\left(\omega\right)\right), & \forall x \in \Gamma_D, \boldsymbol{Z} \in \mathbb{R}^M \\ -k\left(x, \boldsymbol{Z}\left(\omega\right)\right) \frac{\partial u(x,\boldsymbol{Z})}{\partial n(x)} = g\left(x, \boldsymbol{Z}\left(\omega\right)\right), & \forall x \in \Gamma_N, \boldsymbol{Z} \in \mathbb{R}^M \end{cases},$$

$k\left(x, \boldsymbol{Z}\right)$, $f\left(x, \boldsymbol{Z}\right)$, $u_0\left(x, \boldsymbol{Z}\right)$, $g\left(x, \boldsymbol{Z}\right)$, $u\left(x, \boldsymbol{Z}\right) : \mathcal{D} \times \mathbb{R}^M \to \mathbb{R}$ are understood as functions of random vector $\boldsymbol{Z}\left(\omega\right) \in L^2\left(\Omega, \mathbb{R}^M\right)$ (this includes its probability distribution)

Variational formulation

Find $u_H \in L^2\left(\Omega, H^1_{0,\Gamma_D}(\mathcal{D})\right)$, $\left(u = u_0 + u_H \in L^2\left(\Omega, H^1(\mathcal{D})\right)\right)$:

$$\begin{aligned} a(u_H, v) &= b(v), \ \forall v \in L^2\left(\Omega, H^1_{0,\Gamma_D}(\mathcal{D})\right), \\ a(u_H, v) &= \int\limits_{\mathbb{R}^M} \int\limits_{\mathcal{D}} k(x, \boldsymbol{Z}) \nabla_x u_H(x, \boldsymbol{Z}) \cdot \nabla_x v(x, \boldsymbol{Z}) \, dx \, dF\boldsymbol{Z}, \\ b(v) &= \int\limits_{\mathbb{R}^M} \int\limits_{\mathcal{D}} f(x, \boldsymbol{Z}) v(x, \boldsymbol{Z}) \, dx \, dF\boldsymbol{Z} - \int\limits_{\mathbb{R}^M} \int\limits_{\Gamma_N} g(x, \boldsymbol{Z}) v(x, \boldsymbol{Z}) \, dx \, dF\boldsymbol{Z} \\ &\quad - \int\limits_{\mathbb{R}^M} \int\limits_{\mathcal{D}} k(x, \boldsymbol{Z}) \nabla_x u_0(x, \boldsymbol{Z}) \cdot \nabla_x v(x, \boldsymbol{Z}) \, dx \, dF\boldsymbol{Z} \end{aligned}$$

- $k \in L^2\left(\Omega, L^\infty(\mathcal{D})\right)$, $0 < k_{\min} \le k(x, \boldsymbol{Z}(\omega)) \le k_{\max} < \infty \ \forall x \in \mathcal{D}, \forall \omega \in \Omega$,
- $f \in L^2\left(\Omega, L^2(\mathcal{D})\right)$, $u_0 \in L^2\left(\Omega, H^1(\mathcal{D})\right)$, $g \in L^2\left(\Omega, L^2(\Gamma_N)\right)$

Well-posedness

Similarly as in the deterministic case, we can show that

- $L^2\left(\Omega, H^1_{0,\Gamma_D}(\mathcal{D})\right)$ is a Hilbert space
- $a(u, v)$ is continuous and elliptical bilinear operator

$$\exists C > 0\ \forall u, v \in L^2\left(\Omega, H^1_{0,\Gamma_D}(\mathcal{D})\right) : |a(u, v)| \leq C \|u\| \|v\|$$

$$\exists c > 0\ \forall u \in L^2\left(\Omega, H^1_{0,\Gamma_D}(\mathcal{D})\right) : a(u, u) \geq c \|u\|^2$$

- $b(v) \in L^2\left(\Omega, H^1_{0,\Gamma_D}(\mathcal{D})\right)^*$

Types of uncertainties

Input uncertainties $k\left(x,\boldsymbol{Z}\right),f\left(x,\boldsymbol{Z}\right),u_0\left(x,\boldsymbol{Z}\right),g\left(x,\boldsymbol{Z}\right)$ can be divided into two groups:

- uncertainties whose affect $a\left(\cdot,\cdot\right)$
  - $k\left(x,\boldsymbol{Z}\right)$
- uncertainties whose only affect $b\left(\cdot\right)$
  - $f\left(x,\boldsymbol{Z}\right),u_0\left(x,\boldsymbol{Z}\right),g\left(x,\boldsymbol{Z}\right)$

Uncertainties in $b(\cdot)$

If $k(x, \boldsymbol{Z}) = k(x)$ (without uncertainty) and other uncertain inputs are separable

$$f(x, \boldsymbol{Z}) = \sum_{i=1}^{M_f} f_i^D(x)\, f_i^S(\boldsymbol{Z}),\; u_0(x, \boldsymbol{Z}) = \sum_{i=1}^{M_u} u_{0,i}^D(x)\, u_{0,i}^S(\boldsymbol{Z}),\; g(x, \boldsymbol{Z}) = \sum_{i=1}^{M_g} g_i^D(x)\, g_i^S(\boldsymbol{Z})$$

The linearity of the problem yields

$$u(x, \boldsymbol{Z}) = \sum_{i=1}^{M_f} u_i^f(x)\, f_j^S(\boldsymbol{Z}) + \sum_{i=1}^{M_u} u_i^u(x)\, u_{0,i}^S(\boldsymbol{Z}) + \sum_{i=1}^{M_g} u_i^g(x)\, g_i^S(\boldsymbol{Z})$$

Where $u_i^f(x), u_i^u(x), u_i^g(x)$ are the solutions of deterministic problems.

Uncertainties in $b(\cdot)$

$$i = 1, \ldots M_f, u_i^f : \begin{cases} -\mathrm{div}\left(k(x)\nabla u_i^f(x)\right) = f_i^D(x) & x \in \mathcal{D} \\ u(x) = 0 & x \in \Gamma_D \\ -k(x)\frac{\partial u(x)}{n(x)} = 0 & x \in \Gamma_N \end{cases},$$

$$i = 1, \ldots M_u, u_i^u : \begin{cases} -\mathrm{div}\left(k(x)\nabla u_i^u(x)\right) = 0 & x \in \mathcal{D} \\ u(x) = u_{0,i}^D(x) & x \in \Gamma_D \\ -k(x)\frac{\partial u(x)}{n(x)} = 0 & x \in \Gamma_N \end{cases},$$

$$i = 1, \ldots M_g, u_i^g : \begin{cases} -\mathrm{div}\left(k(x)\nabla u_i^g(x)\right) = 0 & x \in \mathcal{D} \\ u(x) = 0 & x \in \Gamma_D \\ -k(x)\frac{\partial u(x)}{n(x)} = g_i^D(x) & x \in \Gamma_N \end{cases},$$

Uncertainties in $b(\cdot)$

- This works in the same way if $k(x, \boldsymbol{Z})$ is not deterministic
    - sub-problems are also stochastic, but only with stochastic $k(x, \boldsymbol{Z})$
- separable representation of $f(x, \boldsymbol{Z}), u_0(x, \boldsymbol{Z}), g(x, \boldsymbol{Z})$ can be done with the projection into orthogonal polynomials on Ω
    - due to well-posedness, good separable approximations yields good approximations of original u
- In usual cases, $f(x, \boldsymbol{Z}), u_0(x, \boldsymbol{Z}), g(x, \boldsymbol{Z})$ depends on different random variables than $k(x, \boldsymbol{Z})$
    - stochastic sub-problems with $k(x, \boldsymbol{Z})$ have lower dimension
    - if not, this approach will not bring much benefit
- for simplicity we proceed with deterministic $f(x), u_0(x), g(x)$
    - extension to stochastic $f(x, \boldsymbol{Z}), u_0(x, \boldsymbol{Z}), g(x, \boldsymbol{Z})$ will not change the complexity of the problem

Vanishing material field

- the condition $0 < k_{\text{min}} \leq k\left(x, \boldsymbol{Z}\left(\omega\right)\right) \leq k_{\text{max}} < \infty$ may be too strong
- if we are interested in a path-wise solution $\left(\forall\omega \in \Omega\right)$, the deterministic problem for each $\omega \in \Omega$ is well posed if

$$0 < k_{\text{min}}\left(\omega\right) \leq k\left(x, \boldsymbol{Z}\left(\omega\right)\right) \leq k_{\text{max}}\left(\omega\right) < \infty$$

- but the condition for the whole problem throughout Ω can be broken:

$$\inf_{\omega\in\Omega} k_{\text{min}}\left(\omega\right) = 0, \sup_{\omega\in\Omega} k_{\text{max}}\left(\omega\right) = \infty$$

- this is the case for a log-normal random variable $\exp\left(Z\right)$, $Z \sim \mathcal{N}\left(\mu, \sigma\right)$

Well-posedness of vanishing material field

- well posedness can be achieved using weighted spaces

$$L^2_{\varrho}\left(\Omega, V\right) := L^2\left(\left(\Omega, \mathcal{F}, \varrho d\mathbb{P}\right), V\right) = \left\{ f : \Omega \to V \text{ measurable } : \mathbb{E}\left(\|f\|_V^2 \, \varrho\right) \right\}$$

- for $k_{\text{min}}\left(\omega\right)$ from previous slide, we define a space

$$U_{k_{\text{min}}^{-1}} := L^2_{k_{\text{min}}^{-1}}\left(\Omega, H^1_{0,\Gamma_D}\left(\mathcal{D}\right)\right)$$

- if $b \in U_k^*$ (e.g. $f \in L^2_{k_{\text{min}}^{-1}}\left(\Omega, L^2\left(\mathcal{D}\right)\right)$), the problem can be shown to be well-posed in U_k
    - but still, the problem is not as nice as if $k\left(x, \omega\right)$ was uniformly bounded throughout Ω
    - Galerkin approximation inside U_k may be a problem

A. Mugler, H.-J. Starkloff: On the convergence of the stochastic Galerkin method for random elliptic partial differential equations, 2013 ESAIM

Weighted formulation for vanishing material field

- For our simple case, the problem can be reformulated into

$$-\mathrm{div}_x\left(\frac{k(x,\omega)}{k_{\min}(\omega)}\nabla_x u(x,\omega)\right)=\frac{f(x,\omega)}{k_{\min}(\omega)},$$

  where $1\leq\frac{k(x,\omega)}{k_{\min}(\omega)}\ \forall\omega\in\Omega$ and $\frac{f(x,\omega)}{k_{\min}(\omega)}$ need to be from $L^2\left(\Omega,H^1_{0,\Gamma_D}(\mathcal{D})\right)$

- the reformulated problem is well-posed in $L^2_{kk_{\min}^{-1}}\left(\Omega,H^1_{0,\Gamma_D}(\mathcal{D})\right)$
- $L^2_{kk_{\min}^{-1}}\left(\Omega,H^1_{0,\Gamma_D}(\mathcal{D})\right)$ is continuously embedded in $L^2\left(\Omega,H^1_0(\mathcal{D})\right)$

Examples of broken convergence

$$\begin{cases} -\operatorname{div}\left(\exp\left(Z\right)\nabla u\left(x,Z\right)\right) = \exp\left(Z\right)\left|Z-1\right| & x \in \mathcal{D},\ Z \sim \mathcal{N}\left(0,1\right) \\ u\left(x,Z\right) = 0 & x \in \partial\mathcal{D},\ Z \sim \mathcal{N}\left(0,1\right) \end{cases}$$

$$\begin{cases} -\left(\exp\left(-\frac{Z^2 x}{10}\right) u'\left(x,Z\right)\right)' = 1 & x \in \mathcal{D} = \left(0,1\right),\ Z \sim \mathcal{N}\left(0,1\right) \\ u\left(0,Z\right) = 0 & Z \sim \mathcal{N}\left(0,1\right) \\ -\exp\left(-\frac{Z^2 x}{10}\right) u'\left(1,Z\right) = -\exp\left(-6Z^2\right) & Z \sim \mathcal{N}\left(0,1\right) \end{cases}$$

Stochastic Galerkin method

Discretization spaces

- we seek the solution u (u_H) in space $L^2\left(\Omega, H^1_{0,\Gamma_D}(\mathcal{D})\right)$ which is isometrically isomorphic with $H^1_{0,\Gamma_D}(\mathcal{D}) \otimes L^2(\Omega)$
  - this means that u can be represented as

$$u(x,\omega) = \sum_{i=1}^{\infty} u_i^D(x)\, u_i^S(\omega), \;\; u_i^D(x) \in H^1_{0,\Gamma_D}(\mathcal{D}), u_i^S(\omega) \in L^2(\Omega)$$

- we use the tensor structure of the solution/test space to construct the finite-dimensional solution/test space

$$V_{h,K} := V_h \otimes V_K, \;\; V_h \subset H^1_{0,\Gamma_D}(\mathcal{D}), V_K \subset L^2(\Omega)$$

$$V_h := \{\varphi_1(x), \ldots, \varphi_{N_D}(x)\}, \;\; V_K := \{\psi_1(\omega), \ldots, \psi_{N_S}(\omega)\}$$

  - the dimension of $V_{h,K}$ is $N_D N_S$ with the basis

$$\xi_{i,j}(x,\omega) = \varphi_i(x)\, \psi_j(\omega), \;\; \forall i = 1, \ldots, N_D, j = 1, \ldots, N_S$$

Discretization of $H^1_{0,\Gamma_D}(\mathcal{D})$

- Discretization of $H^1_{0,\Gamma_D}(\mathcal{D})$ is usually done using finite elements.
  - the same choice as for the deterministic counterpart of the problem
- as the basis is not adaptive with respect to Ω, its good to consider what possible grids will be needed throughout different possible realisations of ω

Discretization of $L^2(\Omega)$

- we use the transition from $u(x,\omega)$ into $u(x,\boldsymbol{Z}(\omega))$
    - $\boldsymbol{Z}$ is a random vector
- $\psi_i(\omega) = \psi_i(\boldsymbol{Z})$ (a shift from abstract function on the stochastic space, into the functions of a real valued vector)
- there is (in all standard cases) no benefit from picking local basis functions
- the best choice are the polynomials
    - complete polynomials = multivariate polynomials with bounded total degree

$$V_K = \operatorname{span}\left\{\psi(\boldsymbol{Z}) = \prod_{i=1}^{M} Z_i^{\alpha_i} : \sum_{i=1}^{M} \alpha_i \leq K\right\}$$

    - tensor product polynomials = multivariate polynomials with uniformly bounded degree

$$V_K = \operatorname{span}\left\{\psi(\boldsymbol{Z}) = \prod_{i=1}^{M} Z_i^{\alpha_i} : \alpha_i \leq K\,\forall i\right\}$$

Polynomials of random vector

- although the space V_K itself is given now, it is very important to pick the right basis
- we need to consider the following issues:
    - numerical stability, e.g. Z^{20} will range from very low to very high values
    - potential sparsity of resulting system
- both of these issues can be (at least partially) solved by using the orthogonal polynomials with respect to the distribution of $\boldsymbol{Z}$
    - for a general $\boldsymbol{Z}$, we need to construct the polynomials for the whole $\boldsymbol{Z}$ (e.g. Gram-Schmidt - very difficult to avoid numerical stability issues)
    - for $\boldsymbol{Z}$ consisting of independent random variables Z_i, it can be easily constructed as a product of orthogonal polynomials on each variable Z_i

Orthogonal polynomials of independent random vector

- sometimes called "polynomial chaos"
- many "standard" random variables have well-known orthogonal polynomials with understood properties and methods of construction
  - e.g. Hermite polynomials, Laguerre polynomials, Jacobi polynomials, ...
  - organized into Askey scheme
  - recurrence relation is very useful for stable evaluation of the polynomials

$$P_n(x) = (A_n x + B_n) P_{n-1}(x) + C_n P_{n-2}(x)$$

- orthogonal polynomials on $\boldsymbol{Z}$ product of 1d polynomials of Z_i

$$\psi_i(\boldsymbol{Z}) = \prod_{k=1}^{M} \psi_{i_k}(Z_k),$$

where i denotes the multi-index of size M

Assembling the SGM matrix

Separable form of input data

- we could now assemble the matrix (and right hand side) but it would be semi-dense $N_D N_S \times N_D N_S$ matrix (usually $N_D > 10^6$, $N_S > 10^3$) which would probably not fit into memory
  - $N_D N_S = 10^9$ and 0.0001% fill it would take ≈ 1.5 TB (terabyte) in the sparse format (CRS)
- we need to assemble the matrix in a compressed form
- the way to achieve this is to have all input data in separable form (same as before for elimination of uncertainty in $b(\cdot)$)
  - for simplicity we still assume only the permeability field with uncertainties

$$k(x, \boldsymbol{Z}) = \sum_{m=1}^{M_k} k_m^D(x)\, k_m^S(\boldsymbol{Z})$$

Bilinear form on tensor product space

Recall the bilinear form of the problem

$$a\left(u, v\right) = \int\limits_{\mathbb{R}^M} \int\limits_{\mathcal{D}} k\left(x, \boldsymbol{Z}\right) \nabla_x u\left(x, \boldsymbol{Z}\right) \cdot \nabla_x v\left(x, \boldsymbol{Z}\right) dx\, dF\boldsymbol{Z}$$

- the solution u is in the form

$$u\left(x, \boldsymbol{Z}\right) = \sum_{i=1}^{N_D} \sum_{j=1}^{N_S} \overline{u}_{i,j} \varphi_i\left(x\right) \psi_j\left(\boldsymbol{Z}\right)$$

Combined with the separability of $k\left(x, \boldsymbol{Z}\right)$, we obtain for $v = \varphi_\ell\left(x\right) \psi_n\left(\boldsymbol{Z}\right)$:

$$a\left(u, v\right) = \sum_{m=1}^{M_k} \sum_{i=1}^{N_D} \sum_{j=1}^{N_S} \overline{u}_{i,j} \int\limits_{\mathbb{R}^M} \int\limits_{\mathcal{D}} k_m^D\left(x\right) k_m^S\left(\boldsymbol{Z}\right) \nabla_x \varphi_i\left(x\right) \cdot \nabla_x \varphi_\ell\left(x\right) \psi_j\left(\boldsymbol{Z}\right) \psi_n\left(\boldsymbol{Z}\right) dx\, dF\boldsymbol{Z}$$

Bilinear form on tensor product space

The integrals can be separated now

$$a\left(u,v\right)=\sum_{m=1}^{M_k}\sum_{i=1}^{N_D}\sum_{j=1}^{N_S}\overline{u}_{i,j}\int\limits_{\mathcal{D}}k_m^D\left(x\right)\nabla_x\varphi_i\left(x\right)\cdot\nabla_x\varphi_\ell\left(x\right)dx\int\limits_{\mathbb{R}^M}k_m^S\left(\boldsymbol{Z}\right)\psi_j\left(\boldsymbol{Z}\right)\psi_n\left(\boldsymbol{Z}\right)dF\boldsymbol{Z}$$

and the matrix of the system can be represented as (assuming indexing $ij\times mn$)

$$\mathbb{A}=\sum_{m=1}^{M_k}G_m\otimes K_m,$$

$$\left(K_m\right)_{im}=\int\limits_{\mathcal{D}}k_m^D\left(x\right)\nabla\varphi_i\left(x\right)\cdot\nabla\varphi_\ell\left(x\right)dx,$$

$$\left(G_m\right)_{jn}=\int\limits_{\mathbb{R}^M}k_m^S\left(\boldsymbol{Z}\right)\psi_j\left(\boldsymbol{Z}\right)\psi_n\left(\boldsymbol{Z}\right)dF\boldsymbol{Z}.$$

Kronecker product

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{m1}\mathbf{B} & \cdots & a_{mn}\mathbf{B} \end{bmatrix}$$

Assembly of sub-matrices G_m, K_m

- K_m are "standard" finite elements matrices (although material can be zero or even negative)
- G_m are possibly hard to assemble, generally can be dense and every entry represents integral over $\mathbb{R}^M$ with measure given by $\boldsymbol{Z}$
  - if $\boldsymbol{Z}$ is independent and $k_m^S(\boldsymbol{Z}) = \prod_{s=1}^{M} k_{m,s}^S(Z_s)$

$$\int\limits_{\mathbb{R}^M} k_m^S(\boldsymbol{Z})\,\psi_j(\boldsymbol{Z})\,\psi_n(\boldsymbol{Z})\,dF\boldsymbol{Z} = \prod_{s=1}^{M}\int\limits_{\mathbb{R}} k_{m,s}^S(Z_s)\,\psi_{j_s}(Z_s)\,\psi_{n_s}(Z_s)\,dFZ_s$$

  - in usual cases k^S are smooth (analytical) and we can estimate the integrals very efficiently via the Gaussian quadrature rule

Resulting system in the form of matrix equations

$$\mathbb{A}\overline{u} = \overline{b}$$

$$\mathbb{A} = \sum_{m=1}^{M_k} G_m \otimes K_m, \overline{b} = \sum_{m=1}^{M_b} g_m \otimes k_m$$

here we simplify the right hand side as sum over M_b terms as it will look differently based on exact problem (f, u_0, ...) at maximum it would be (considering all input data with uncertainties) $M_b = M_k M_u + M_f + M_g$

The system can be viewed as matrix equations, assuming reshaping $\overline{u}$ into $N_D \times N_S$ matrix $\boldsymbol{u}$

$$\sum_{m=1}^{M_k} K_m \boldsymbol{u} G_m^T = \sum_{m=1}^{M_b} k_m g_m^T$$

Solving the system

Preconditioning

Let K_0 be matrix of deterministic counterpart with $k_0\left(x\right) = \mathbb{E}\left(k\left(x,\omega\right)\right)$

- block-diagonal (mean field) preconditioner: $P = I \otimes K_0$
  - Powell, Elman: Block-Diagonal Preconditioning for Spectral Stochastic Finite-Element Systems. 2009
- kronecker proconditioner: $P = G \otimes K_0,\ G = \sum_{m=1}^{M_k} G_m \frac{\mathrm{trace}\left(K_m^T K_0\right)}{\mathrm{trace}\left(K_0^T K_0\right)}$
  - Ullmann: A Kronecker Product Preconditioner for Stochastic Galerkin Finite Element Discretizations. 2010
- hirearchical Schur preconditioner (specific matrices G_m)
  - Sousedík, Ghanem, Phipps: Hierarchical Schur Complement Preconditioner for the Stochastic Galerkin Finite Element Methods. 2014

Reduced basis method

Solution of the original system can be prohibitively difficult

$$\sum_{m=1}^{M_k} K_m \boldsymbol{u} G_m^T = \sum_{m=1}^{M_b} k_m g_m^T$$

- due to the size of bases V_h and V_K
- remedy can be the solution only on some subspace (reduced basis)
  - it make sense to create the reduced basis of V_h (it is the larger one and we have tools to create a meaningful subspace of it)
  - system with the reduced basis should fulfill all the conditions needed to be well-posed (e.g. discrete inf-sup condition)
  - for SPD problems, we can pick any linearly independent reduced basis W and obtain a valid system

$$\sum_{m=1}^{M_k} W^T K_m W \tilde{\boldsymbol{u}} G_m^T = \sum_{m=1}^{M_b} W^T k_m g_m^T$$

Reduced basis method

There are different methods for creating a reduced basis. For SPD systems it can be utilized

- Monte Carlo sampling → W is constructed from the solutions
- Reduced Rational Krylov subspace method, generating rational Krylov subspace from matrices K_m

In case of e.g. saddle point matrices

- we can use Monte Carlo sampling
- need to assure discrete inf-sup condition
    - can be done with enriching the reduced basis with supremizer functions with respect to original space V_h

Reduced basis method

```
$l = 0,\ W_0 = \emptyset,\ R_0 = \sum_{m=1}^{M_b} k_m g_m^T$
while $\left\| R_l \right\| / \left\| R_0 \right\| > \varepsilon$
    $l = l + 1$
    ① propose an enhancement of RB: $V_l$
    $W_l = \mathrm{orth}\left( \left[ W_{l-1}, V_l \right] \right)$
    ② find $\boldsymbol{y}_l$ as a solution of RB system
        $$\sum_{m=1}^{M_k} W_l^T K_m W_l \boldsymbol{y}_l G_m^T = \sum_{m=1}^{M_b} W_l^T k_m g_m^T$$
    ③ compute $\left\| R_l \right\|$
        $$\left\| R_l \right\| = \left\| \sum_{m=1}^{M_k} K_m W_l \boldsymbol{y}_l G_m^T - \sum_{m=1}^{M_b} k_m g_m^T \right\|$$
end
return $\boldsymbol{u} \approx \tilde{\boldsymbol{u}}_l = W_l \boldsymbol{y}_l$
```

Monte Carlo sampling for the construction of RB

1. draw N_{MC} samples $Z_1, \ldots, Z_{N_{MC}}$ of random vector $\boldsymbol{Z}$
2. for each Z_j assemble and solve the reduced system

$$W_l^T A_j W_l \widetilde{u}_j = W_l^T b_j$$

3. compute indicators (higher number = better sample)

$$f_{\boldsymbol{Z}}\left(\boldsymbol{Z}_j\right) \left\| A_j W_l \widetilde{u}_j - b_j \right\|^2$$

4. select P (for simplicity, we use $P = 1$) highest values of identificators and compute solutions at corresponding samples Z_j

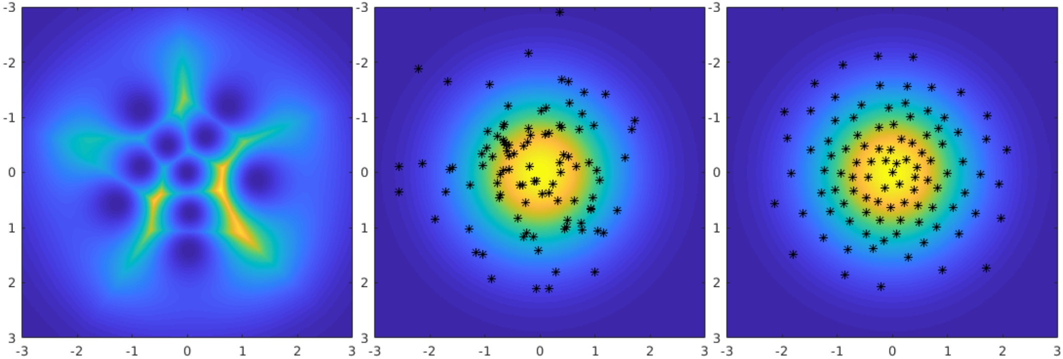
$$A_j u_j = b_j$$

5. use the collected solutions to expand the reduced basis W_l and check if the expanded reduced basis is good enough

Monte Carlo sampling - improved sampling

Computing reduced solutions and their residuals at samples $\boldsymbol{Z}^j$ is costly ⇒ avoid samples around those already contributing to RB

$$\tilde{f}_l\left(\boldsymbol{Z}\right) \propto f\left(\boldsymbol{Z}\right) \min_{i=1,\dots,l} w_i\left(\boldsymbol{Z}\right), w_i\left(\boldsymbol{Z}\right) = \left(1 - \exp\left(- \left\|\boldsymbol{Z} - X_i\right\|^2_{\Sigma^{-1}} /2\right)\right)^{\beta}$$



Monte Carlo method for the RB construction

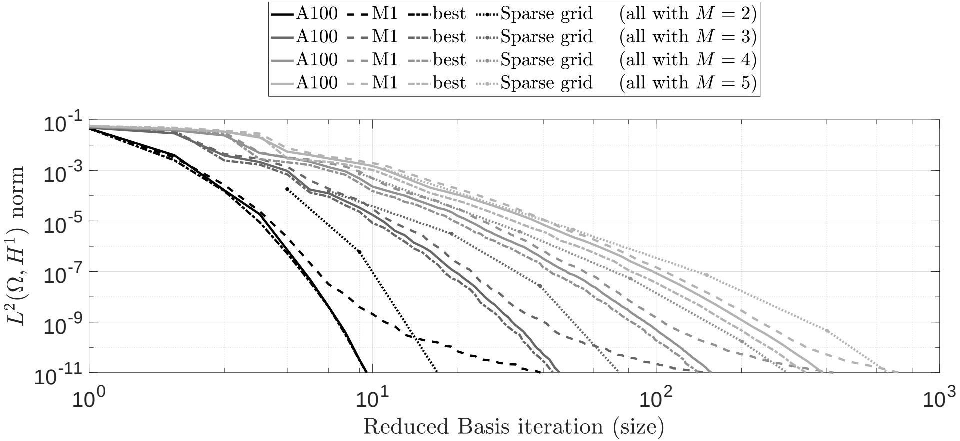


Figure 1: Comparison of convergence of greedy MC with the "best" scenario (basis obtained via SVD of the solution) and basis consisting of solutions in sparse grid points

MC error indicators and adaptive polynomial selection

- $L^2\left(\Omega, H^1\right)$ error ("true error"), error approximated via 1000 MC samples
- ε_1: estimation of $L^2\left(\Omega, H^1\right)$ error between the SGM solution and the path-wise solution obtained from all samples used for RB construction
- ε_2: estimation of $L^2\left(\Omega, H^1\right)$ error between RB reduced solution and path-wise solution obtained from P samples prepared to be added to the RB at the current iteration
- ε_3: smoothed ε_2 via moving geometric average with window of length 5variables Z_i, it can be easily constructed as product of orthogonal polynomials on each variable Z_i

$\varepsilon_2/\varepsilon_3$ does not require the reduced solution of the SGM system (and the polynomial basis) → we can build reduced basis independently of the discretization of the stochastic space

GMC error indicators and adaptive polynomial selection

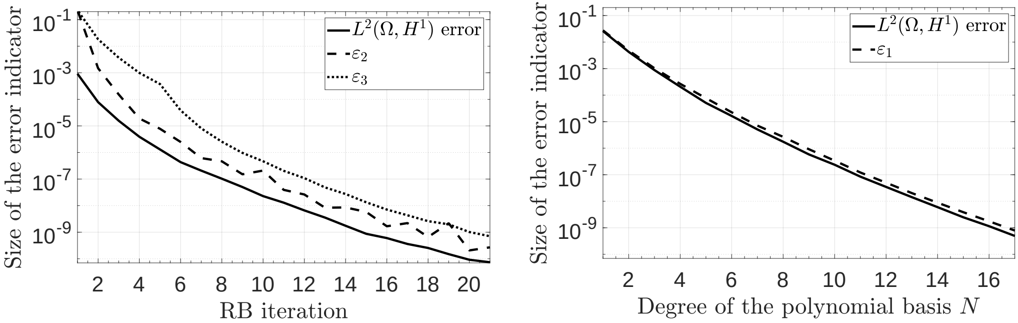


Figure 2: Two phase solution of SGM problem with adaptive polynomial degree selection. Left: phase 1 - construction of RB; Right: phase 2 - selection of maximum polynomial degree

Tensor train approximation

- can be used if we use tensor polynomials, the random vector $\boldsymbol{Z}$ consist of independent random variables, and the input data are separable (including $\boldsymbol{Z}$)
  - i.e. matrix and rhs of the SGM system can be expressed in the canonical form of the $M+1$ dimensional tensor (M is number of random variables)
- TT approximation is stable “low-rank” approximation of the higher dimensional tensor (counterpart of the SVD for matrices)
- TT-toolbox used for the computation, specifically the Alternating minimal energy method for the TT approximation of the solution of linear system
- implicitly preconditioned system was solved
  - using mean field preconditioner

Tensor train approximation

three problem settings: $S1$: $\sigma = 0.3, \mu = 0$;

$S2$: $\sigma = (0.1, 0.1, 0.1, 0.1, 0.3, 0.3, 0.1, 0.1, 0.1, 0.1)$, $\mu = (0, 0, 0, 0, -5, -5, 0, 0, 0, 0)$;
$S3$: $\sigma = (0.01, 0.01, 0.01, 0.01, 0.3, 0.3, 0.01, 0.01, 0.01, 0.01)$, $\mu = (0, 0, 0, 0, -10, -10, 0, 0, 0, 0)$

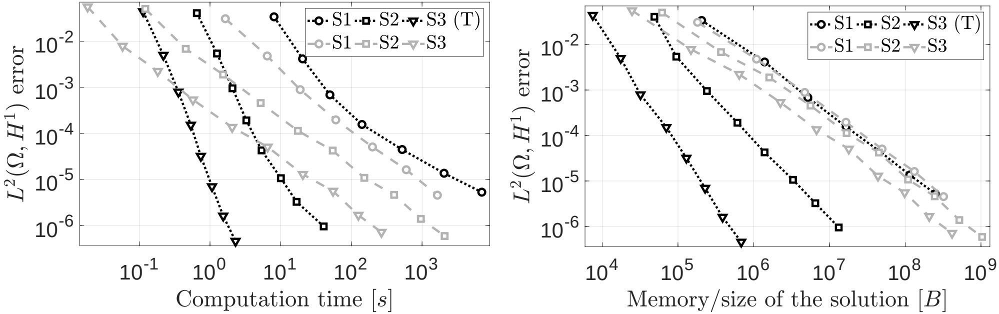


Figure 3: Comparison of TT approximation and CG solution using complete polynomial basis. Left: computational time; Right: memory size of the solution

Acceleration of systems solutions using deflation

- using conjugate gradients (CG) as the solved systems are SPD
- deflated CG (DCG) takes an additional parameter in the form of the deflation basis W
  - W should be able to describe the sought solution reasonably well
  - DCG looks for the solution only in the complement of W by projecting the residual (or the preconditioned residual) using the projector

    $$P = I - W\left(W^T A W\right)^{-1} W^T A$$

    during the CG routine
- good choice of the deflation basis W is the current RB

Acceleration of systems solutions using deflation

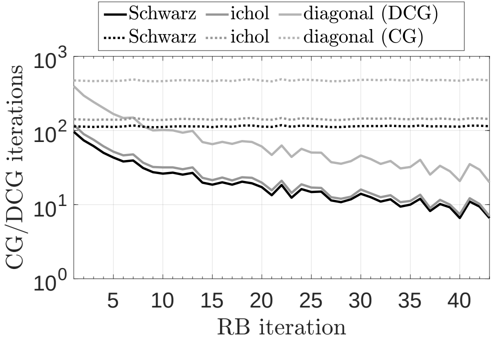


Figure 4: Comparison of mean number of CG iterations for the solution of deterministic counterparts

- over 80 % of iterations saved across the different approaches (RRKS, GMC), problem settings, and target precision

Use of SGM solution

Mean value and standard deviation of the solution

SGM solution can be easily used for the calculation of mean and variance (standard deviation) of resulting random field

- we assume solution in the form

$$u\left(x, \boldsymbol{Z}\right) = \sum_{i=1}^{N_D} \sum_{j=1}^{N_S} \overline{u}_{i,j} \varphi_i\left(x\right) \psi_j\left(\boldsymbol{Z}\right)$$

  and $\psi_1\left(\boldsymbol{Z}\right) = 1$
- mean is

$$\mathbb{E}\left(u\left(x, \boldsymbol{Z}\right)\right) = \sum_{i=1}^{N_D} \overline{u}_{i,1} \varphi_i\left(x\right)$$

- variance is

$$\mathrm{Var}\left(u\left(x, \boldsymbol{Z}\right)\right) = \sum_{j=2}^{N_S} \left(\sum_{i=1}^{N_D} \overline{u}_{i,j} \varphi_i\left(x\right) \right)^2$$

Evaluating at samples - deterministic counterpart approximate

Once we have SG solution, we can easily create approximations of the deterministic counterparts

- we just need to evaluate the polynomials $\psi_j(\boldsymbol{Z})$
- recurrent formulas are very useful for the evaluation of $\psi_j(\boldsymbol{Z})$
- for the solution reshaped into matrix $\boldsymbol{u}$ and matrix of evaluated polynomials in samples $\boldsymbol{Z}^i$: $\Psi = \left[\overline{\psi}\left(\boldsymbol{Z}^1\right), \ldots\right]$, the approximations of deterministic counterparts are columns of $\boldsymbol{u}\Psi$
- all of the steps are straightforwardly vectorized or parallelized

Example - TSX experiment

Problem setting

Stationary Darcy flow, $\mathcal{D} = (0,100) \times (0,100) \setminus E$ (E is ellipse with center $[50,50]$ and height 2×1.75 and width 2×2.1875)

$$\begin{cases} -\mathrm{div}_x \left(k\left(x,\boldsymbol{Z}\right) \nabla_x u\left(x,\boldsymbol{Z}\right)\right) = 0, & \forall x \in \mathcal{D}, \boldsymbol{Z} \in \mathbb{R}^3 \\ u\left(x,\boldsymbol{Z}\right) = 3 \cdot 10^6, & \forall x \in \Gamma_1, \boldsymbol{Z} \in \mathbb{R}^3 \\ u\left(x,\boldsymbol{Z}\right) = 0, & \forall x \in \Gamma_2, \boldsymbol{Z} \in \mathbb{R}^3 \end{cases},$$

where

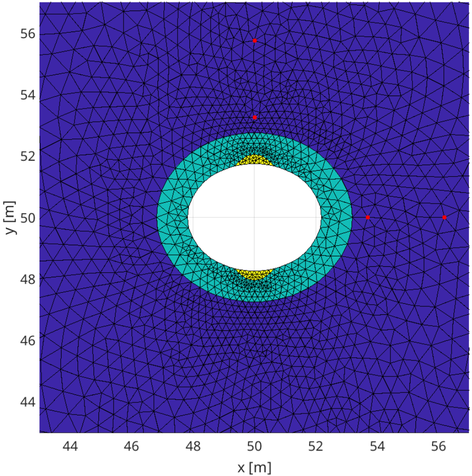
$$k\left(x,\boldsymbol{Z}\right) = \sum_{i=1}^{3} 1_{\Omega_i}\left(x\right) 10^{Z_i}$$

$Z_1 \sim \mathcal{N}\left(\mu = -16, \sigma = \frac{1}{3}\right)$, $Z_1 \sim \mathcal{N}\left(\mu = -18, \sigma = \frac{1}{3}\right)$, $Z_1 \sim \mathcal{N}\left(\mu = -21, \sigma = \frac{1}{3}\right)$

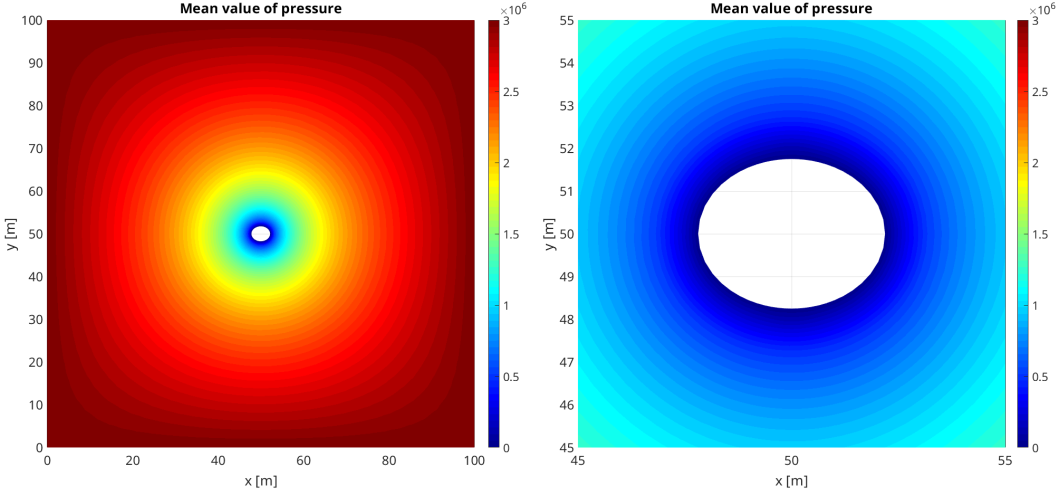
Γ_1 is outer boundary of the rectangle, Γ_2 is boundary of cut-off ellipse

Problem setting

Zoom around the cut-off with marked measuring points

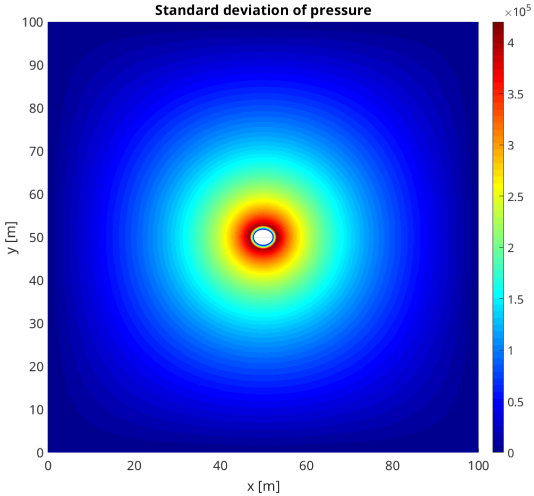


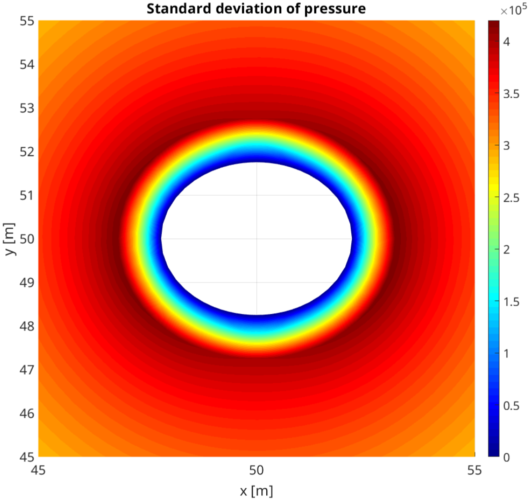
Mean value



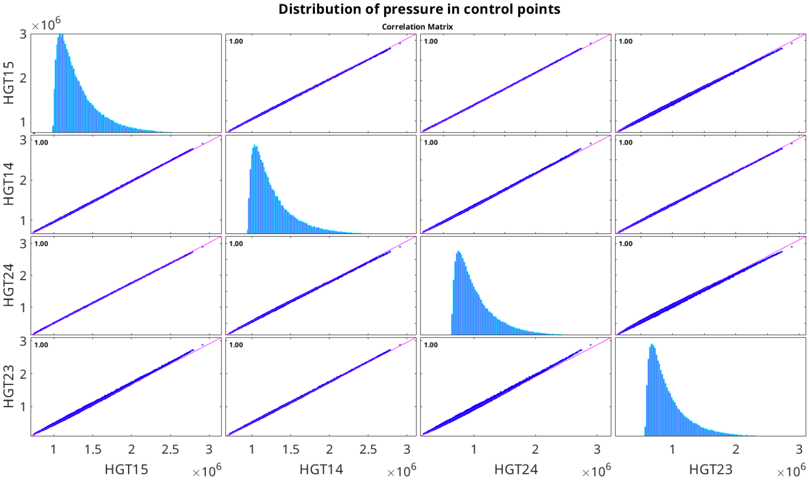
Mean value of pressure
100
90
80
70
60
50
40
30
20
10
0
y [m]
0
20
40
60
80
100
x [m]
×10^6
3
2.5
2
1.5
1
0.5
0
Mean value of pressure
55
54
53
52
51
50
49
48
47
46
45
y [m]
45
50
55
x [m]
×10^6
3
2.5
2
1.5
1
0.5
0

Standard deviation

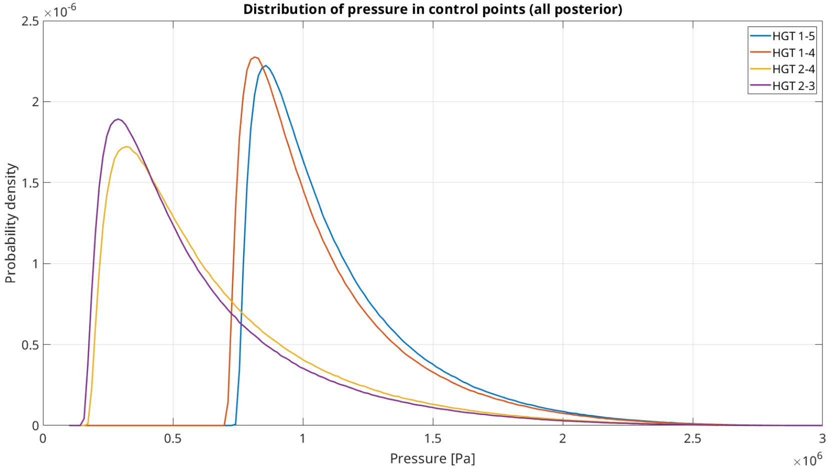




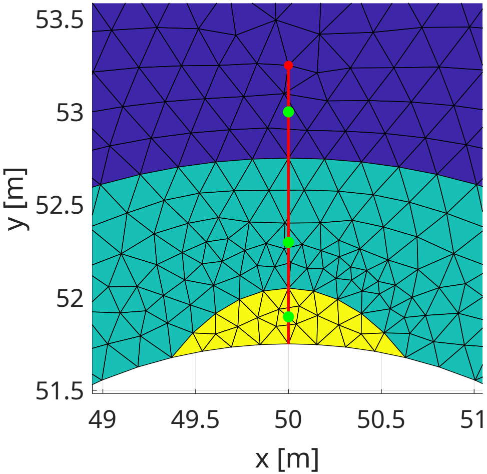
Distribution of pressure in measuring points



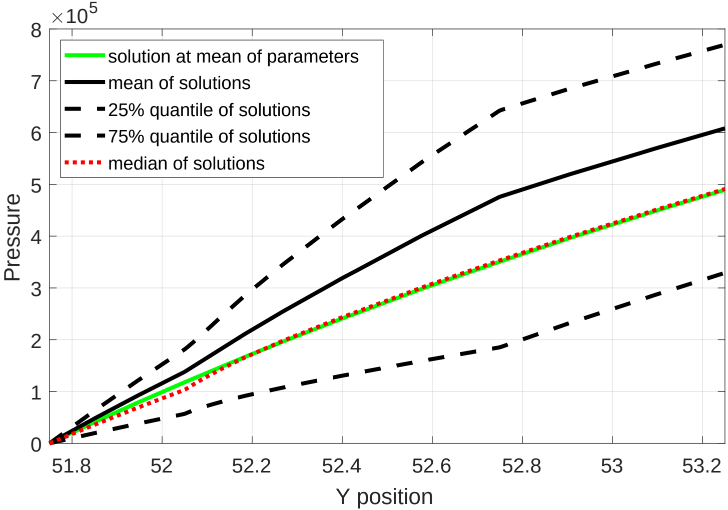
Distribution of pressure in measuring points



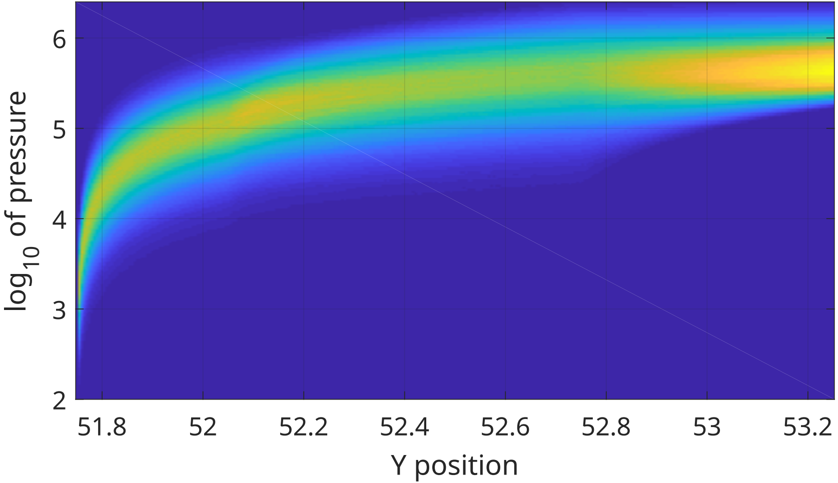
Geometry - cutoff and behaviour on a line



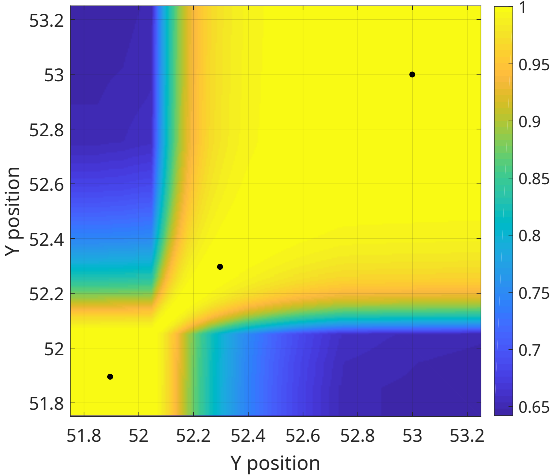
Behaviour on vertical line



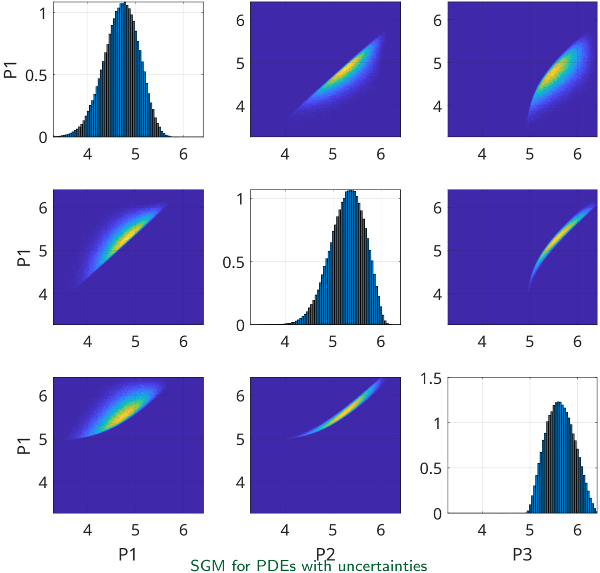
Behaviour on vertical line - distribution



Behaviour on vertical line - correlation



Distribution in selected points



The End

Thank you for your attention!