The Stochastic Galerkin Method for Darcy Flow Problem with Log-Normal Random Field Coefficients

Michal BERES\textsuperscript{1,2,3}, Simona DOMESOVA\textsuperscript{1,2,3}

\textsuperscript{1}Department of Applied Mathematics, Faculty of Electrical Engineering and Computer Science, VSB–Technical University of Ostrava, 17. listopadu 15/2172, 708 33 Ostrava, Czech Republic
\textsuperscript{2}IT4Innovations National Supercomputing Center, VSB–Technical University of Ostrava, 17. listopadu 15/2172, 708 33 Ostrava, Czech Republic
\textsuperscript{3}Institute of Geonics of the Czech Academy of Sciences, Studentska 1768, 708 00 Ostrava, Czech Republic

michal.beres@vsb.cz, simona.domesova@vsb.cz

DOI: 10.15598/aeee.v15i2.2280

Abstract. This article presents a study of the Stochastic Galerkin Method (SGM) applied to the Darcy flow problem with a log-normally distributed random material field given by a mean value and an autocovariance function. We divide the solution of the problem into two parts. The first one is the decomposition of a random field into a sum of products of a random vector and a function of spatial coordinates; this can be achieved using the Karhunen-Loeve expansion. The second part is the solution of the problem using SGM. SGM is a simple extension of the Galerkin method in which the random variables represent additional problem dimensions. For the discretization of the problem, we use a finite element basis for spatial variables and a polynomial chaos discretization for random variables. The results of SGM can be utilised for the analysis of the problem, such as the examination of the average flow, or as a tool for the Bayesian approach to inverse problems.

Keywords

Darcy flow, Gaussian random field, Karhunen-Loeve decomposition, polynomial chaos, Stochastic Galerkin method.

1. Introduction and Problem Setting

The standard mathematical modelling approach usually works with deterministic parameters of solved problems, such as boundary conditions, material constants, etc. These simplifications neglect the inherent uncertainty in these parameters, which can lead to unreliable solutions. The fast development of computational resources allows computations of massive problems, e.g. solving the partial differential equations with uncertainty in parameters. Popular methods for calculation of these challenges are for example the Multilevel Monte Carlo method, stochastic collocation method (see [16]) or the stochastic Galerkin method which is presented in this article. The comparison between SGM and Monte Carlo method can be found in [17]. We focus on a model problem with a very high uncertainty in input parameters, for which we demonstrate the application of SGM.

Consider a 2D Darcy flow problem on the unit square according to Fig. 1. We assume that the pressure...
equals one on \( I_{D_1} \), zero on \( I_{D_2} \), and the flow equals zero on the rest of the boundary.

\[
\begin{align*}
\forall \omega \in \Omega : \\
-\text{div} \left( k (x; \omega) \nabla u (x; \omega) \right) &= 0, \forall x \in D, \\
u (x; \omega) &= 1, \forall x \in I_{D_1}, \\
u (x; \omega) &= 0, \forall x \in I_{D_2}, \\
\frac{\partial u (x; \omega)}{\partial n (x)} &= 0, \forall x \in \Gamma_N,
\end{align*}
\]

where \( \omega \) represents an event from the sample space \( \Omega \). In general the functions \( k (x; \omega) \) and \( u (x; \omega) : D \times \Omega \to \mathbb{R} \) are understood as random fields. That means that the function \( K_0 = k (x_0; \omega) : \Omega \to \mathbb{R} \) is a random variable for every fixed \( x_0 \in D \) and the function \( k (x; \omega) : D \to \mathbb{R} \) is a function on the spatial domain for every fixed \( \omega_0 \in \Omega \). A random field is a type of a random process, see for example [1]: chapter 5.

The problem Eq. [3] is very complex because the random field \( k \) represents an infinite number of random variables. To solve this kind of problems, we need to reduce the number of random variables to some finite, ideally small, number. We can achieve this using the Karhunen-Loeve (KL) decomposition, which will be discussed in the following section. For a more detailed insight into SGM and the KL decomposition, see [4].

We choose a log-normally distributed random material. This type of random field appears to be a reasonable choice for the description of permeability properties of natural materials like sandstone or gravel, see [6]. The log-normal random field is defined as an exponential of the Gaussian (normal) Random Field (GRF):

\[
k (x; \omega) = \exp (g (x; \omega)).
\]

The GRF is fully described by its mean value \( \mu (x) = E (g (x; \omega)) \) and its autocovariance function \( c (x, y) = \text{cov} (g (x; \omega), g (y; \omega)) \). In this paper, we consider the GRF specified by:

\[
\mu (x) = 0, c (x, y) = \sigma^2 \exp \left( \frac{\| x - y \|}{\lambda} \right),
\]

where \( \sigma \) is the scale parameter and \( \lambda \) is the correlation length parameter. The scale parameter affects the amplitude (equivalent samples with different values of \( \sigma \) differ only in scale) and the correlation length parameter changes the correlation/distance ratio between two points.

### 2. The Karhunen-Loeve Decomposition

In this section, we will discuss the approximation of the random field by a function of a random vector and a physical variable. This can be achieved using the truncated Karhunen-Loeve (KL) decomposition. The existence of the KL decomposition is stated in the following theorem, see [1]: theorem 7.52.

**Theorem 1.** Let \( D \subset \mathbb{R}^d \). Consider a random field \( \{ k (x; \omega) : x \in D \} \) and suppose that \( k \in L^2 (\Omega, L^2 (D)) \).

Then

\[
k (x; \omega) = \mu (x) + \sum_{j=1}^{\infty} \sqrt{\lambda_j} \psi_j (x) \xi_j (\omega),
\]

where the sum converges in \( L^2 (\Omega, L^2 (D)) \),

\[
\xi_j (\omega) := \frac{1}{\sqrt{\lambda_j}} \int_D (k (x; \omega) - \mu (x)) \psi_j (x) \, dx,
\]

and \( \{ \lambda_j, \psi_j \} \) denotes the eigenvalues and the eigenfunctions of the autocovariance operator \( C : L^2 (D) \to L^2 (D) \).

Theorem 1: The truncated Karhunen-Loeve decomposition is a spectral decomposition of the operator \( C \), which we can obtain by solving the eigenvalue problem

\[
\int_D c (x, y) \psi_i (y) \, dy = \lambda_i \psi_i (x), \forall i \in \mathbb{N}.
\]

This is the Fredholm integral equation, which can be solved by the Galerkin method.

We can obtain the weak formulation of the Eq. [7] by multiplying it by a test function \( v \in L^2 (D) \) and integrating over the domain \( D \).

\[
\begin{align*}
\text{Find } \psi_i (x) &\in L^2 (D), \lambda_i \in \mathbb{R}^+, \\
\int_D \int_D c (x, y) \psi_i (y) \, dy \, dx &= \lambda_i \int_D \psi_i (x) \, dx, \forall \psi_i (x) \in L^2 (D).
\end{align*}
\]
The next step is the discretization of the weak formulation. First consider a basis \( \{ \phi_1(x), \ldots, \phi_n(x) \} = V_n \subset L^2(D) \), so the Galerkin formulation takes the form of:

\[
\begin{aligned}
\text{Find } \psi_i(x) &= \sum_{j=1}^{n} \psi_{ij} \phi_j(x), \quad \lambda_i \in \mathbb{R}^+ : \\
\int_D \phi_j(x) \int_D c(x, y) \psi_i(y) \, dy \, dx &= \lambda_i \int_D \phi_j(x) \psi_i(x) \, dx, \quad \forall \phi_j(x),
\end{aligned}
\]  

(9)

where \( \overline{\psi}_i \) is a representation of the eigenvector \( \psi_i \) in the \( V_n \) basis. The solution of Eq. (9) can be rephrased into a generalized eigenvalue problem:

\[
A \overline{\psi}_i = \lambda_i^p W \overline{\psi}_i,
\]  

(10)

where

\[
A_{ij} = \int_D \int_D c(x, y) \phi_i(y) \phi_j(x) \, dy \, dx,
\]  

(11)

\[
W_{ij} = \int_D \phi_i(x) \phi_j(x) \, dx.
\]  

(12)

2. The Choice of the Basis

We choose a discretization \( \{ \phi_1(x), \ldots, \phi_n(x) \} = V_n \) of the space \( L^2(D) \) \( (D = (a, b) \times (c, d)) \) as a tensor product of 1D bases of the spaces \( L^2((a, b)) \) and \( L^2((c, d)) \). Therefore basis functions are in the form:

\[
\phi_i(x) = \varphi_i^1(x_1) \varphi_i^2(x_2),
\]  

(13)

where \( \varphi_i^1(x_1) \in V_1^n := \langle \varphi_1^1(x_1), \ldots, \varphi_m^1(x_1) \rangle \) and \( \varphi_i^2(x_2) \in V_2^n := \langle \varphi_1^2(x_2), \ldots, \varphi_m^2(x_2) \rangle \) are 1D functions. The dimension of the space \( V_n = V_1^n \otimes V_2^n \) then equals \( m_1 \cdot m_2 \).

For the calculation of the problem Eq. (10), it is useful to consider the spaces \( V_1^n \) and \( V_2^n \) with specific properties. These properties will be demonstrated on 1D basis denoted by \( \{ \varphi_1(x), \ldots, \varphi_m(x) \} \).

1) The Orthogonality of \( \{ \varphi_1(x), \ldots, \varphi_m(x) \} \)

The complexity of the problem Eq. (10) decreases if we transform the generalized eigenvalue problem into a standard eigenvalue problem \( (W = I) \). We can achieve this by assuring all of the 1D bases \( \{ \varphi_1(x), \ldots, \varphi_m(x) \} \) to be orthonormal:

\[
\forall i, j : \int_a^b \varphi_i(x) \varphi_j(x) \, dx = \delta_{i,j}.
\]  

(14)

2) The (Anti)-Symmetry of \( (\varphi_1(x), \ldots, \varphi_m(x)) \)

The matrix \( A \) will generally be dense, but some unique properties of the chosen autocovariance function can be utilised to obtain partial sparsity of the matrix. First, we define the function

\[
p(x, y) = \sigma^2 \exp \left( -\frac{x^2 + y^2}{\lambda} \right),
\]  

(15)

which is related to the autocovariance function by

\[
p(|x_1 - y_1|, |x_2 - y_2|) = c((x_1, x_2), (y_1, y_2)).
\]  

(16)

Next, consider the translation of the integral over the domain \( D = (a, b) \times (c, d) \) into the integral over the domain \( (-\alpha, \alpha) \times (-\beta, \beta) \), where \( \alpha = \frac{b-a}{2} \) and \( \beta = \frac{d-c}{2} \) (translate the center of \( D \) into \( (0, 0) \)). Note, that this will not affect the function \( p(|x_1 - y_1|, |x_2 - y_2|) \).

For simplicity we denote the basis functions \( \varphi_i(x) = \varphi_i^1(x_1) \varphi_i^2(x_2) \) as before the translation of the integral. After these modifications, the formula for the elements of the matrix \( A \) is

\[
A_{i,j} = \int_{-\beta}^{\beta} \int_{-\beta}^{\beta} f(x_2, y_2) \varphi_i^2(x_2) \varphi_j^2(y_2) \, dy_2 \, dx_2,
\]  

(17)

where

\[
f(x_2, y_2) = \int_{-\alpha}^{\alpha} \int_{-\alpha}^{\alpha} p(|x_1 - y_1|, |x_2 - y_2|) \cdot \varphi_i^1(x_1) \varphi_j^1(y_1) \, dx_1 \, dy_1.
\]  

(18)

If \( \varphi_i^1(x_1) \) is symmetric and \( \varphi_j^1(y_1) \) is anti-symmetric (or vice versa), the function \( f(x_2, y_2) \) is constant zero and \( A_{i,j} = 0 \). The same holds for the functions \( \varphi_i^2(x_2) \) and \( \varphi_j^2(y_2) \).

Fig. 2: Nonzero entries of matrix \( A \) (discretization 4 \times 4).
The nonzero pattern of the matrix $A$ for the standard ordering according to the polynomial degree (see Fig. 2(a)) can be reordered into a block diagonal structure (see Fig. 2(b)). The permutation can be done by reordering the basis functions according to their parity in each dimension. The block diagonal structure leads to the solution of four smaller eigenvalue problems instead of one bigger eigenvalue problem. Note that in 3D we get eight diagonal blocks.

The next part of the Galerkin method is the particular choice of the discrete basis (orthonormal, symmetric/anti-symmetric). The most suitable bases for the Galerkin discretization Eq. (9) are e.g. piecewise constant functions, polynomials, or trigonometric functions. For the solution of similar problem using the wavelet basis see [18].

3) The Piecewise Constant Basis

First consider an interval $(a,b)$ divided into an even number (say $2n$) of subintervals $t_i$ of the same size $|t_i| = \frac{b-a}{2n}$. Then the symmetric basis functions are

$$\forall i = 1,...,n : \varphi^S_i (x) = \begin{cases} \alpha, & x \in t_i, t_{2n-i+1}, \\ 0, & \text{otherwise}, \end{cases}$$

and the anti-symmetric basis functions are

$$\forall i = 1,...,n : \varphi^A_i (x) = \begin{cases} \alpha, & x \in t_i, \\ -\alpha, & x \in t_{2n-i+1}, \\ 0, & \text{otherwise}, \end{cases}$$

where $\alpha = \sqrt{\frac{b-a}{2}}$. This construction also ensures the orthonormality of the basis.

4) The Polynomial Basis

In the case of polynomial basis, the choice is simple; the Legendre polynomials are both orthogonal and alternately symmetric or anti-symmetric. For general use, the Legendre polynomials need to be shifted into the desired interval and normalised. An explicit formula for the standard Legendre polynomials on $(-1,1)$ is

$$P_n(x) = 2^n \sum_{k=0}^{n} (-1)^k \binom{n}{k} \binom{n+k-1}{k} \cdot \left( \frac{n+k-1}{n} \right).$$

The normalised Legendre polynomials on $(a,b)$ are defined as

$$\varphi_n (x) = \sqrt{\frac{2n+1}{b-a}} P_n \left( \frac{2}{b-a} x - \frac{a+b}{b-a} \right).$$

5) The Trigonometric Basis

As the trigonometric basis, we choose the solution of the same problem in 1D (on (0,1)). The 1D basis consists of eigenvectors of the operator

$$(Cf) (x) := \int_0^1 \sigma^2 \exp \left( \frac{\sqrt{\sigma^2 (x-y)^2}}{\lambda} \right) f(y) \, dy,$$

with the same parameters $\sigma$ and $\lambda$. This problem has an analytic solution, see [5]. Explicit formulas for the basis functions are

$$\varphi'_n (x) = a_n \cos (w_n x) + b_n \sin (w_n x),$$

where

$$a_n = \lambda w_n b_n, b_n = \left( \frac{\lambda^2 w^2_n + 1}{2} + \lambda \right)^{-\frac{1}{2}}$$

and $w_n$ is the $n$-th smallest solution of

$$\lambda^2 w^2 - 1 \sin(w) = 2 \lambda w \cos(w).$$

Then we shift the functions into the interval $(a,b)$:

$$\varphi_n (x) = \sqrt{\frac{1}{b-a}} \varphi'_n \left( \frac{1}{b-a} x - \frac{a}{b-a} \right).$$

2.2. Numerical Results

In this part, we compute the approximation of the KL decomposition and compare the bases mentioned above. To evaluate the integrals in the calculation of $A$ (see Eq. (11)), we use the Gauss-Legendre quadrature with 100 points per dimension. Note, that this is a computationally expensive task, because we need to evaluate the integrand at $10^4$ points for each nonzero entry of the matrix $A$.

First, we illustrate the fast decay of the eigenvalues (see Fig. 3) and the approximation of the first and thirteenth eigenfunction (see Fig. 4). The results are obtained using the trigonometric basis of 400 basis functions (20 in each dimension). Figure 3 shows that for a good approximation of the random field we need only the first few eigenpairs, i.e., the components of
that the convergence of the Galerkin method depends on the precision of the numerical integration, which is lower for fast oscillating functions like polynomials and trigonometric functions. On the other hand, the piecewise constant basis needs a lower number of integration points for a larger basis.

3. The Stochastic Galerkin Method

This section focuses on a solution of the model problem Eq. (1) with a random material properties in the form of the truncated KL decomposition of a random field. The fast decay of the eigenvalues (see Fig. 3) ensures a low Approximation Error (AE) of the truncated KL decomposition. We consider a log-normally distributed random material properties. Therefore our random material properties take the form

\[ k(x; Z) = \exp \left( \sum_{i=1}^{N} k_i(x) Z_i \right), \]

where \( Z = (Z_1, \ldots, Z_N) \) denotes a random vector of independent standard normal variables, and \( k_i(x) \) are known functions.

The formulation of the model problem with the material field \( k(x; Z) \) changes to the following:

\[
\forall Z = (Z_1, \ldots, Z_N), Z_i \sim N(0; 1): \\
\begin{aligned}
-\text{div} (k(x; Z) \nabla u(x; Z)) &= 0, \quad \forall x \in D, \\
u(x; Z) &= 1, \quad \forall x \in \Gamma_{D_1}, \\
u(x; Z) &= 0, \quad \forall x \in \Gamma_{D_2}, \\
\frac{\partial u(x; Z)}{\partial n(x)} &= 0, \quad \forall x \in \Gamma_N.
\end{aligned}
\]

We can understand the formulation Eq. (30) as a parametric differential equation with additional properties of the parameters defined by their distribution. The solution of this problem is the function \( u(x; Z) : D \times \mathbb{R}^N \rightarrow \mathbb{R} \), which is an approximation of the original solution \( u(x; \omega) : D \times \Omega \rightarrow \mathbb{R} \).

The next step is a weak formulation of Eq. (30), the details of which can also be found in [16]. A random field is understood as an operator from the space \( L^2(\Omega; L^2(D)) \), as mentioned in Sec. 2. In the case of the solution \( u(x; \omega) \) (also a random field), the homogeneous part \( u_H(x; \omega) \) is in the space \( H_0^1(D) \) for every fixed \( \omega \). Therefore the resulting random field is in the space \( V := L^2(\Omega; H_0^1(D)) + u_0(x; \omega) \), where \( u_0(x; \omega) \) is an arbitrary function from the space \( L^2(\Omega; H^1(D)) \) that satisfies the Dirichlet boundary conditions. When using the truncated KL decomposition form of the random material properties, the solution \( u(x; Z) \) is in

\[ k(x; Z) = \exp \left( \sum_{i=1}^{N} k_i(x) Z_i \right), \]

where \( Z = (Z_1, \ldots, Z_N) \) denotes a random vector of independent standard normal variables, and \( k_i(x) \) are known functions.

The formulation of the model problem with the material field \( k(x; Z) \) changes to the following:

\[
\forall Z = (Z_1, \ldots, Z_N), Z_i \sim N(0; 1): \\
\begin{aligned}
-\text{div} (k(x; Z) \nabla u(x; Z)) &= 0, \quad \forall x \in D, \\
u(x; Z) &= 1, \quad \forall x \in \Gamma_{D_1}, \\
u(x; Z) &= 0, \quad \forall x \in \Gamma_{D_2}, \\
\frac{\partial u(x; Z)}{\partial n(x)} &= 0, \quad \forall x \in \Gamma_N.
\end{aligned}
\]

We can understand the formulation Eq. (30) as a parametric differential equation with additional properties of the parameters defined by their distribution. The solution of this problem is the function \( u(x; Z) : D \times \mathbb{R}^N \rightarrow \mathbb{R} \), which is an approximation of the original solution \( u(x; \omega) : D \times \Omega \rightarrow \mathbb{R} \).

The next step is a weak formulation of Eq. (30), the details of which can also be found in [16]. A random field is understood as an operator from the space \( L^2(\Omega; L^2(D)) \), as mentioned in Sec. 2. In the case of the solution \( u(x; \omega) \) (also a random field), the homogeneous part \( u_H(x; \omega) \) is in the space \( H_0^1(D) \) for every fixed \( \omega \). Therefore the resulting random field is in the space \( V := L^2(\Omega; H_0^1(D)) + u_0(x; \omega) \), where \( u_0(x; \omega) \) is an arbitrary function from the space \( L^2(\Omega; H^1(D)) \) that satisfies the Dirichlet boundary conditions. When using the truncated KL decomposition form of the random material properties, the solution \( u(x; Z) \) is in
Let \( Z \) be a random variable with all \( 2n \) moments finite, i.e. \( \forall n \in \mathbb{N}: E(\mathbb{Z}^{2n}) < \infty \). The PC of such variable is a set of \( N \) polynomials \( \psi_i(Z) \), \( i \in 1, \ldots, N \) (\( i \) denotes the degree of the polynomial) satisfying
\[
E(\psi_i(Z) \psi_j(Z)) = \gamma_i \delta_{ij}, \quad \gamma_i = E(\psi_i(Z)^2).
\]
In the following text, we consider a normalised PC (\( \gamma_i = 1 \)). We can understand the PC of a random variable \( Z \) as an orthonormal basis of the space \( L^2_{\mathcal{F}_Z}(\mathbb{R}) \). Consequently, we can approximate every function \( f(Z) \in L^2_{\mathcal{F}_Z}(\mathbb{R}) \) by the PC of the random variable \( Z \):
\[
f(Z) \approx f_n(Z) = \sum_{i=0}^{n} \psi_i(Z) E(\psi_i(Z) f(Z)),
\]
where \( f_n(Z) \) is the projection of \( f(Z) \) into the space of polynomials up to the degree \( n \), and
\[
\lim_{n \to \infty} \|f_n(Z) - f(Z)\|_{L^2_{\mathcal{F}_Z}(\mathbb{R})} = 0,
\]
see \cite{2, 19} and \cite{20}.

In the case of a random vector \( Z \) of \( N \) independent random variables, the space \( L^2_{\mathcal{F}_Z}(\mathbb{R}^N) \) can be handled as a tensor product \( L^2_{\mathcal{F}_Z}(\mathbb{R}) \otimes \ldots \otimes L^2_{\mathcal{F}_Z}(\mathbb{R}) \). The PC of the random vector \( Z \) consists of products of the PC basis functions of the random variables \( Z \):
\[
\psi_i(Z) = \prod_{k=1}^{N} \psi_{i_k}(Z_k),
\]
where \( i \) denotes a multi-index of size \( N \).

The PC expansions are well known for most of the standard probability distributions, see the Wiener-Askey scheme in Tab. 1. According to this scheme, the polynomial chaos for a normal random variable consists of the Hermite polynomials (we will use the normalised Hermite polynomials).

**Tab. 1:** The Wiener-Askey scheme, see \cite{1}: Table 5.1.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Polynomial chaos</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaussian</td>
<td>Hermite polynomials</td>
</tr>
<tr>
<td>Uniform</td>
<td>Legendre polynomials</td>
</tr>
<tr>
<td>Gamma</td>
<td>Laguerre polynomials</td>
</tr>
<tr>
<td>Beta</td>
<td>Jacobi polynomials</td>
</tr>
<tr>
<td>Poisson</td>
<td>Charlier polynomials</td>
</tr>
<tr>
<td>Negative Binomial</td>
<td>Mieixner polynomials</td>
</tr>
<tr>
<td>Binomial</td>
<td>Krawtchouk polynomials</td>
</tr>
<tr>
<td>Hypogeometric</td>
<td>Hahn polynomials</td>
</tr>
</tbody>
</table>

3.2. Assembling the Stiffness Matrix

As mentioned before, discretization of the space \( V_H \) can be considered as a tensor product of discretization of the space \( H^1_0(D) \) and discretization of the space \( L^2_{\mathcal{F}_Z}(\mathbb{R}^N) \). As discretization
Calculation of elements of the matrix
form of and the discretized particular solution will take the
hand side the aforementioned multi-index into the simple index).
We denote the discretized test function space as $V^h := V^h_D \otimes V^h_I \subset V^h$. Then, the discretized solution and the discretized particular solution will take the form of

$$u^h_H (x; Z) = \sum_{i=1}^{N_D} \sum_{j=1}^{N_P} (\overline{w}^h_{ij}) \varphi_i (x) \psi_j (Z), \quad (37)$$

and

$$u^0_H (x, Z) = \sum_{i=N_D+1}^{N_{DD}} \sum_{j=1}^{N_P} (\overline{w}_0)^i \varphi_i (x) \psi_1 (Z), \quad (38)$$

where $\psi_1 (Z)$ is the constant element of the PC basis. The Galerkin formulation of the problem:

$$\left\{ \begin{array}{l}
\text{Find } \pi_{FT} \in \mathbb{R}^{N_{DD} \times N_P}, \forall \varphi \in V^h_D, \forall \psi \in V^h_I : \\
a (u^h_H, \varphi \psi) = -a (u^0_H, \varphi \psi),
\end{array} \right. \quad (39)$$

is equivalent to the system of linear equations

$$A \pi_{FT} = \overline{b}, \quad (40)$$

where individual elements of $A$ and $\overline{b}$ are defined as:

$$(A)_{ij,kl} = \int_{\mathbb{R}^N} \int_D \exp \left( \sum_{m=1}^{N} k_m (x) Z_m \right) \cdot \nabla \varphi_i (x) \psi_j (Z) \nabla \varphi_k (x) \psi_l (Z) \, dx \, dFZ, \quad (41)$$

$$(\overline{b})_{ij} = \int_{\mathbb{R}^N} \int_D \exp \left( \sum_{m=1}^{N} k_m (x) Z_m \right) \nabla \varphi_i (x) \cdot \psi_j (Z) \sum_{k=N_D+1}^{N_{DD}} (\overline{w}_0)^i \varphi_k (x) \psi_1 (Z) \, dx \, dFZ. \quad (42)$$

Calculation of elements of the matrix $A$ and the right-hand side $\overline{b}$ seems complicated, but the integral over the stochastic domain $\mathbb{R}^N$ can be evaluated analytically. Using the commutativity of multiplication and Fubini’s theorem, we can transform the expression of $(A)_{ij,kl}$ and $(\overline{b})_{ij}$ to

$$(A)_{ij,kl} = \int_D \nabla \varphi_i (x) \nabla \varphi_k (x) k_{j,l} (x) \, dx, \quad (43)$$

$$(\overline{b})_{ij} = \int_D \nabla \varphi_i (x) \sum_{k=N_D+1}^{N_{DD}} (\overline{w}_0)^i \varphi_k (x) k_{j,l} (x) \, dx. \quad (44)$$

We obtain the values of $k_{i,j} (x)$ for some $x \in D$ by rewriting the expression using the definition of the PC basis (see Eq. (40)):

$$k_{i,j} (x) = \prod_{m=1}^{N} \int_{\mathbb{R}} \exp \left( k_m (x) Z_m \right) \cdot \psi_{im} (Z_m) \psi_{jm} (Z_m) \, dFZ_m, \quad (46)$$

and merging the part of the random field $\exp (k_m (x) Z_m)$ corresponding to the random variable $Z_m$ with its probability density function (currently denoted as $dFZ_m$). The elements of the product take the form of

$$\int_{\mathbb{R}} \psi_{im} (Z_m) \psi_{jm} (Z_m) \, dFZ_m \cdot \frac{1}{\sqrt{2\pi}} \exp \left( \frac{2k_m (x) Z_m - Z_m^2}{2} \right) \, dZ_m, \quad (47)$$

which can be expressed by the shifted random variables $\tilde{Z}_m \sim N (k_m (x), 1)$. The probability density of $\tilde{Z}_m$ is

$$dF\tilde{Z}_m := \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{(Z_m - k_m (x))^2}{2} \right) \, dZ_m. \quad (48)$$

Therefore $k_{i,j} (x)$ can be expressed as:

$$k_{i,j} (x) = \prod_{m=1}^{N} \exp \left( k_m (x) \right)^2 \cdot \int_{\mathbb{R}} \psi_{im} (Z_m) \psi_{jm} (Z_m) \, dF\tilde{Z}_m. \quad (49)$$

Calculation of the stochastic part of the integral now reduces to the evaluation of

$$\int_{\mathbb{R}} \psi_{im} (Z_m) \psi_{jm} (Z_m) \, dF\tilde{Z}_m, \forall m = 1, \ldots, N. \quad (50)$$

The functions $\psi_{im} (Z_m)$ are polynomials; therefore the expression Eq. (50) can be written as a linear combination of the moments of $\tilde{Z}_m$ (normal distribution with given mean), which are known.

Another approach to the construction of the matrix $A$ is using the PC expansion of the truncated KL decomposition of the random material properties

$$\sum_{i=1}^{N_{PC}} \tilde{k}_i (x) \psi_i (Z) \approx \exp \left( \sum_{i=1}^{N} k_i (x) Z_i \right). \quad (51)$$
This approach allows us to separate the stochastic and the physical part of the integral in the matrix $A$ calculation

$$(A)_{ij,kl} = \sum_{m=1}^{NPC} \int_{D} \vec{k}_m(x) \nabla \varphi_k(x) \nabla \varphi_i(x) \, dx \cdot \int_{\mathbb{R}^N} \psi_m(Z) \psi_j(Z) \psi_l(Z) \, dFZ. \quad (52)$$

The separation leads to the “tensor form” of the matrix $A$:

$$A = \sum_{m=1}^{NPC} G_m \otimes K_m, \quad (53)$$

where

$$(K_m)_{ij} = \int_{D} \vec{k}_m(x) \nabla \varphi_i(x) \nabla \varphi_j(x) \, dx, \quad (54)$$

$$(G_m)_{ij} = \int_{\mathbb{R}^N} \psi_m(Z) \psi_i(Z) \psi_j(Z) \, dFZ. \quad (55)$$

This approach allows compressed storage of the matrix $A$ in the form of pairs of the matrices $(K_m, G_m)$; the matrix $A$ does not need to be constructed. On the other hand, this approach suffers from an additional AE of the PC representation of the random material properties, see Eq. (51). The tensor form of the matrix can also be assembled by rephrasing the problem, see [10], but this approach leads to a non-symmetric system.

The matrix $A$ can have two different types of structure, which depends on the ordering of the basis functions:

- $(A)_{ij,kl}$ denotes ordering $(\varphi_1 \psi_1), (\varphi_2 \psi_1), \ldots, (\varphi_{N_D} \psi_1), (\varphi_1 \psi_2), \ldots$, or
- $(A)_{ji,kl}$ denotes ordering $(\varphi_1 \psi_1), (\varphi_1 \psi_2), \ldots, (\varphi_{N_P} \psi_1), (\varphi_2 \psi_1), \ldots$.

We can see the nonzero pattern of the matrix $A$ for both structures in Fig. 7. We can use this property for the construction of a preconditioner.

### 3.3. Numerical Results

This part reports on numerical results for the solved problem. For the numerical experiments, we choose a regular $50 \times 50$ grid for the physical domain. We solve the problem using the full matrix $A$ construction to avoid the AE of the PC expansion of the random material properties.

The main aims of this section are the evaluation of the mean and the variance/standard deviation of the resulting random field and the comparison of the SGM solution with the standard FEM solution.

The system of equations (structure $(A)_{ij,kl}$) was solved using the flexible GMRES method and a double block diagonal preconditioner

$$M = D_2^{-1} P^T D_1^{-1} P, \quad (56)$$

where $P$ is a permutation matrix for reordering a vector from the $(A)_{ij,kl}$ ordering to the $(A)_{ji,kl}$ ordering, $D_1$ is the block diagonal of the matrix $(A)_{ji,kl}$ and $D_2$ is the block diagonal of the matrix $(A)_{ij,kl}$.

There exist other types of preconditioners for the SGM systems, for example: block-diagonal preconditioner, see [12]; kronecker preconditioner, see [15]; or hierarchical Schur preconditioner, see [11]. Some insight into effectivity of preconditioners can be obtained using the strengthened Cauchy-Bunyakowski-Schwarz constant, see [13] and [14].

1) Mean Value and Standard Deviation of the Solution

The SGM solution of the given problem offers an easy way to obtain the mean and the variance of the solution. We can express the mean value of the solution

$$\mu_{u^h}(x) : D \to \mathbb{R} := \int_{\mathbb{R}^N} u^h(x, Z) \, dFZ, \quad (57)$$

as

$$\mu_{u^h}(x) = \int_{\mathbb{R}^N} \sum_{j=1}^{N_P} u_j(x, Z) \psi_j(Z) \, dFZ = u_1(x), \quad (58)$$

where

$$u_j(x) = \sum_{i=1}^{N_{DD}} (\Pi)_{ij} \varphi_i(x). \quad (59)$$

The variance of the solution

$$\sigma_{u^h}^2(x) : D \to \mathbb{R} := \int_{\mathbb{R}^N} (u^h(x, Z) - \mu_{u^h}(x))^2 \, dFZ, \quad (60)$$
can be expressed as

$$
\sigma_{u_h}^2 (x) = \int_{\mathbb{R}^N} \left( \sum_{j=2}^{N_p} u_j (x) \psi_j (\mathbf{Z}) \right)^2 dFZ
$$

$$
= \int_{\mathbb{R}^N} \sum_{j=2}^{N_p} \sum_{i=2}^{N_p} u_j (x) \psi_j (\mathbf{Z}) u_i (x) \psi_i (\mathbf{Z}) dFZ
$$

$$
= \sum_{j=2}^{N_p} u_j^2 (x). \quad (61)
$$

The mean value and the standard deviation of solution are shown in Fig. 8. The results are obtained using $N = 5$ elements of the truncated KL decomposition and the PC up to the degree 10. The results agree with the intuition - the mean value corresponds to the deterministic solution with the homogeneous material and the standard deviation is higher for the points farther from the Dirichlet boundary.

![Fig. 8: Characteristics of the solution.](image)

2) Comparing SGM and FEM Solutions

The main advantage of the SGM solution is that it gives us an approximation of $u(x, \mathbf{Z})$ for all possible values of $\mathbf{Z}$ (random material instances). Therefore we can use the SGM solution instead of the FEM calculation for a specific material sample. In the following text, $u(x, \mathbf{Z})$ denotes the FEM solution for a random material sample (using the truncated KL decomposition) and $u^h(x, \mathbf{Z})$ denotes the SGM solution. The AE of SGM will be measured using the $H^1(\Omega)$ norm and the $L^2(\Omega)$ norm of the difference between the FEM and the SGM solution; the precision of the SGM approximation over the whole domain of $\mathbf{Z}$ will be measured using

$$
\text{err}_{H^1}(u^h) := \int_{\mathbb{R}^N} \left\| u(x, \mathbf{Z}) - u^h(x, \mathbf{Z}) \right\|_{H^1} dFZ, \quad (62)
$$

and

$$
\text{err}_{L^2}(u^h) := \int_{\mathbb{R}^N} \left\| u(x, \mathbf{Z}) - u^h(x, \mathbf{Z}) \right\|_{L^2} dFZ. \quad (63)
$$

The integral in the calculation of $\text{err}_{H^1}(u^h)$ and $\text{err}_{L^2}(u^h)$ will be approximated by the Monte Carlo method with $10^9$ samples.

The first numerical test describes the dependence of the AE on the iterative solution precision and on the maximum polynomial degree of the PC basis. We construct the PC basis with the maximum polynomial degree $n$ as

$$
\left\langle \psi_a(\mathbf{Z}) := \prod_{i=1}^{N} \psi_{a_i}(Z_i) \right\rangle_{\sum a_i \leq n}. \quad (64)
$$

The results of the numerical tests are shown in Fig. 9. We can observe the following:

- The convergence in the $L^2$ norm is comparable with the convergence in the $H^1$ norm. In the rest of the text, we will use only the $H^1$ norm.
- The precision of the iterative solution affects the SGM AE only slightly, the difference between relative residual error $10^{-5}$ and $10^{-9}$ is negligible.
- The convergence is almost linear when increasing the maximum polynomial degree.

![Fig. 9: Comparison of error for different iterative precision and polynomial degree.](image)

The next numerical test will compare two basic approaches to the construction of the PC base. The first construction is given by Eq. (64), where the sum of degrees of polynomials over all dimensions is bounded by $n$. The second construction is

$$
\left\langle \psi_a(\mathbf{Z}) := \prod_{i=1}^{N} \psi_{a_i}(Z_i) \right\rangle_{\forall i: a_i \leq n}. \quad (65)
$$

where the degree of a polynomial is bounded by $n$ in every dimension. The size of the PC basis is $\binom{n + N}{n}$ for the first construction and $n^N$ for the second construction.

The error $\text{err}_{H^1}(u^h)$ for both PC bases is shown in Fig. 10 and Fig. 11. Figure 10 shows the dependence of
First, we investigate the AE in some specific subsets of the stochastic dimension. We want to observe the AE when changing only one random variable and the AE when fixing one random variable to zero. Therefore we propose two error indicators. First, the AE measured only across the \( Z_i \) part of the stochastic dimension

\[
err_{H^1}(u^h) := \int_{\mathbb{R}^N} \| u(x, Z_i) - u^h(x, Z_i) \|_{H^1} \, dF Z, \tag{66}
\]

where \( Z_i = (0, \ldots, 0, Z_i, 0, \ldots, 0) \). Second, the AE measured as the complement to \( err_{H^1}(u^h) \)

\[
err_{H^1}^c(u^h) := err_{H^1}(u^h) - \int_{\mathbb{R}^N} \| u(x, Z^c_i) - u^h(x, Z^c_i) \|_{H^1} \, dF Z, \tag{67}
\]

where \( Z^c_i = (Z_1, \ldots, Z_{i-1}, 0, Z_{i+1}, \ldots, Z_N) \).

![Fig. 10: Error dependence on the maximal degree of the PC basis.](image1)

![Fig. 11: Error dependence on the size of the PC basis.](image2)

The results of the numerical test of the two proposed error indicators are shown in Fig. 12. The test was performed for \( N = 15 \) and \( n = 2, 3, 4 \) (bounded sum of degrees). Both \( err_{H^1}(u^h) \) and \( err_{H^1}^c(u^h) \) exhibit a similar behaviour, the relative impact on the AE is the greatest for the 3rd and 4th KLV. The results of \( err_{H^1}^c(u^h) \) in Fig. 12 may be more credible, because they include a dependence between single KLVs.

The behaviour observed in the previous paragraph can be supported by the sensitivity analysis. The sensitivity analysis, in the context of this paper, is understood as a comparison of \( err_{H^1}(u^h) \) for different bases.

The tested bases will be constructed as

\[
\psi^m := \psi^n \cup \psi^m, \tag{68}
\]

where

\[
\psi^n := \left\{ \psi_{\alpha}(Z) := \prod_{j=1}^{N} \phi_{\alpha_j}(Z_j) \right\}, \tag{69}
\]

\[
\psi^m := \left\{ \psi_{\alpha}(Z) := \prod_{j=1, j \neq i}^{N} \phi_{\alpha_j}(Z_j) \right\}. \tag{70}
\]

We performed numerical tests for \( \psi_{4,3} \) (denoted as “pol. degree -1”), \( \psi_{4,2} \) (denoted as “pol. degree -2”) and \( \psi_{4,1} \) (denoted as “pol. degree -3”). The results (see Fig. 13) reflect the increase of \( err_{H^1}(u^h) \) when removing the polynomials of the highest degree for a specific random variable. Note, that the sensitivity analysis calculation is extremely computationally expensive; therefore it cannot be used for the evaluation of the importance of a KLV.
The performed numerical tests are computationally expensive and require the calculation of the approximations of the solution. Therefore we cannot use them for the construction of the PC basis. Another way to estimate the importance of the KLV is the examination of the functions \( k_1(x) \) from the truncated KL decomposition form of the random material properties, see Eq. (29). The different importance of single KLVs is caused only by the disparity between the functions \( k_i(x) \).

The objective is to find a function \( f : L^2(D) \rightarrow \mathbb{R}^+ \) which equals to the \( \sqrt{\text{div}^2(k(x))} \). However, a single \( k(x) \) returns similar values as the relative impact indicators or sensitivity analysis for single KLVs. A straightforward approach to estimate the relative impact on the AE is to take the \( L^2(D) \) norm of \( k_i(x) \). However, a single \( k_i(x) \) has the same relative impact on the AE as the constant \( k_i(x) = k_i^c(x) + c_i \) is added to it, because

\[
k(x; Z) = c(Z) k^c(x; Z), \quad (71)
\]

where \( c(Z) = \exp\left(\sum_{i=1}^{N} c_i Z_i\right) \) and \( k^c(x; Z) = \exp\left(\sum_{i=1}^{N} k_i^c(x) Z_i\right) \), and

\[
-\text{div} \left(c(Z) k^c(x; Z) \nabla u(x; Z)\right) = 0
\]

\[
-\text{div} \left(k^c(x; Z) \nabla u(x; Z)\right) = 0. \quad (72)
\]

Therefore the following approximation is proposed:

\[
f(k_i(x)) := \min_{c \in \mathbb{R}} \|k_i(x) - c\|_{L^2(D)}, \quad (73)
\]

which equals to the

\[
f(k_i(x)) = \left\| k_i(x) - \int_D k_i(y) \, dy \right\|_{L^2(D)}. \quad (74)
\]

The values of the \( f(k_i(x)) \) for the KL decomposition elements 1, \ldots, 15 can be seen in Fig. 13. Compared to the previous relative impact indicators and sensitivity analysis, the function \( f(k_i(x)) \) has a similar behaviour.

The next step is the construction of the PC basis according to the given weights. We proceed as follows:

- we normalize the weights from Eq. (74) so that the maximum is 0.5:

\[
w_i = f(k_i(x)) / \left(2 \max_{k \in 1, \ldots, N} f(k_i(x))\right),
\]

the vector \( w = (w_1, \ldots, w_N) \) is obtained,

- we calculate the weight of each polynomial of the PC basis from its multi-index \( a = (a_1, \ldots, a_N) \)

\[
w_a = \prod_{i=1}^{N} w_i^{a_i}, \quad (75)
\]

- the PC basis of \( N_P \) polynomials is constructed from the first \( N_P \) polynomials with the highest weight \( w_a \).

The numerical results of the AE dependence on the size of the PC basis for the weighted polynomial ordering compared to the previous construction approaches can be seen in Fig. 14. Note that in addition to the lower AE for the same size of the PC basis, the proposed construction also allows an arbitrary size of the basis.

\[
\begin{array}{c}
\text{Pol. degree} -3 \\
\text{Pol. degree} -2 \\
\text{Pol. degree} -1 \\
\text{Base error}
\end{array}
\]

\[
\begin{array}{c}
\text{H} \\
\text{error}
\end{array}
\]

4. Conclusion

The paper focused on the complete solution of the Darcy flow problem with a random material. We outlined a path from the differential equation with a random field as an input parameter to an approximation of the solution, which is also a random field. We dealt with the spectral decomposition of the operator given...
by the autocovariance function, and we used the out-
coming decomposition for assembling the SGM system of
equations.

In the section about the KL decomposition, we also
provided an enhanced approach to the construction of
the Galerkin solution, which leads to the block dia-
gonal structure of the outcoming system, and numerical
experiments concerning the choice of the basis for the
Galerkin solution.

In the section about the SGM solution, we performed
extensive numerical experiments regarding the conver-
genence in the stochastic dimension and the AE of the
different PC bases. Then we proposed the construc-
tion of the PC basis with a better AE.

Acknowledgment

This work was supported by The Ministry of Educa-
tion, Youth and Sports from the National Programme
of Sustainability (NPU II) project “IT4Innovations ex-
cellence in science - LQ1602”. The work is partially
supported by Grant of SGS No. SP2017/56, VSB–
Technical University of Ostrava, Czech Republic.

References

[1] LORD, G. J., C. E. POWELL and T. SHARD-
LOW. An introduction to computational stochas-
tic PDEs. New York: Cambridge University Press,

[2] LIGHT, W. A. and E. W. CHENEY. Approxima-
tion theory in tensor product spaces. New York:

[3] DA PRATO, G. and J. ZABCZYK. Second or-
der partial differential equations in Hilbert spaces.

[4] XIU, D. Numerical methods for stochastic com-
putations: a spectral method approach. New Jersey:
14212-8.

perturbation approach for flow in random porous
media via Karhunen–Loeve and polynomial ex-
pansions. Journal of Computational Physics. 2004,
vol. 194, iss. 2, pp. 773–794. ISSN 0021-9991.

[6] HOEKSEMA, R. J. and P. K. KITANIDIS. Anal-
ysis of the spatial structure of properties of se-
lected aquifers. Water Resources Research. 1985,
vol. 21, iss. 4, pp. 563–572. ISSN 0043-1397.
DOI: 10.1029/WR021i004p00563.

F. LI and I. M. DAVIS. A Fourier–Karhunen–
Loeve discretization scheme for stationary random
material properties in SFEM. International jour-
nal for numerical methods in engineering. 2008,
vol. 73, iss. 13, pp. 1942–1965. ISSN 0029-5981.
DOI: 10.1002/nme.2160.

[8] GITTELSON, C. J. Stochastic Galerkin
discretization of the log-normal isotropic
diffusion problem. Mathematical Models
and Methods in Applied Sciences. 2010,
vol. 20, iss. 2, pp. 237–263. ISSN 0218-2025.
DOI: 10.1142/S0218202510004210.

stochastic differential equations. SIAM
Journal on Scientific Computing. 2003,
vol. 24, iss. 2, pp. 619–644. ISSN 1064-8275.
DOI: 10.1137/S1064827501387826.

ERNST. Efficient Iterative Solvers for Stochastic
Galerkin Discretizations of Log-Transformed Ran-
dom Diffusion Problems. SIAM Journal on Scien-
tific Computing. 2012, vol. 34, iss. 2, pp. A639-
A682. ISSN 1064-8275. DOI: 10.1137/110836675.

PHIPPS. Hierarchical Schur Complement Pre-
conditioner for the Stochastic Galerkin Finite
Element Methods: Dedicated to Professor Ivo
Marek on the Occasion of His 80th Birthday.
Numerical Linear Algebra with Applications.
DOI: 10.1002/nla.1869.

[12] POWELL, C. E. and H. C. ELMAN. Block-
Diagonal Preconditioning for Spectral Stochastic
Finite-Element Systems. JMA Journal of Numeri-
ISSN 0272-4979. DOI: 10.1093/imanum/drn014.

[13] PULTAROVA, I. Adaptive Algorithm for Stochas-
tic Galerkin Method. Applications of Mathemat-
ics. 2015, vol. 60, iss. 5, pp. 551–571. ISSN 0862-

[14] PULTAROVA, I. Hierarchical Preconditioning for
the Stochastic Galerkin Method: Upper Bounds
on the Strengthened CBS Constants. Computers
and Mathematics with Applications. 2016,
vol. 71, iss. 4, pp. 949–964. ISSN 0898-1221.
DOI: 10.1016/j.camwa.2016.01.006.

[15] ULLMANN, E. A Kronecker Product Precon-
ditioner for Stochastic Galerkin Finite Eleme


About Authors

Michal BERES graduated from the Faculty of Electrical Engineering and Computer Science, VSB–Technical University of Ostrava in 2015. He is currently a Ph.D. student at the same university. His research interests include uncertainty in the numerical models.

Simona DOMESOVA graduated from VSB–Technical University of Ostrava in 2015, where she studied Computational Mathematics. Currently, she is a Ph.D. student at the same university. Her research interests include stochastic modeling.