Applications of Mathematics

Michal Béreš

A comparison of approaches for the construction of reduced basis for stochastic Galerkin matrix equations

Applications of Mathematics, Vol. 65 (2020), No. 2, 191–225

Persistent URL: http://dml.cz/dmlcz/148109

A COMPARISON OF APPROACHES FOR THE CONSTRUCTION OF REDUCED BASIS FOR STOCHASTIC GALERKIN MATRIX EQUATIONS

Michal Béreš, Ostrava

Received October 3, 2019. Published online March 12, 2020.

Abstract. We examine different approaches to an efficient solution of the stochastic Galerkin (SG) matrix equations coming from the Darcy flow problem with different, uncertain coefficients in apriori known subdomains. The solution of the SG system of equations is usually a very challenging task. A relatively new approach to the solution of the SG matrix equations is the reduced basis (RB) solver, which looks for a low-rank representation of the solution. The construction of the RB is usually done iteratively and consists of multiple solutions of systems of equations. We examine multiple approaches and their modifications to the construction of the RB, namely the reduced rational Krylov subspace method and Monte Carlo sampling approach. We also aim at speeding up the process using the deflated conjugate gradients (DCG). We test and compare these methods on a set of problems with a varying random behavior of the material on subdomains as well as different geometries of subdomains.

Keywords: stochastic Galerkin method; reduced basis method; deflated conjugate gradients method; Darcy flow problem

MSC 2020: 65C05, 86-08, 60-08

1. Introduction and motivation

When modeling real-life problems, we often want to obtain the result not only for one combination of input parameters but for a whole set of acceptable parameters. This can be caused by the inherent uncertainty in the parameters and the need for a solution with a characterization of its dependence on uncertain parameters.

This work was supported by The Ministry of Education, Youth and Sports from the National Programme of Sustainability (NPU II) project IT4Innovations excellence in science—LQ1602. This works was also supported by TACR, project no. TK02010118.

Some examples of such problems are the stochastic homogenization, the uncertainty quantification or the construction of a response surface.

Until recently, due to the lack of computation power, the standard mathematical modeling approach was constrained to a limited number of model evaluations. Therefore, it usually used only one fixed combination of parameters (e.g. averages of single inputs). But nowadays, the fast development of computational resources allows computations of challenging problems. One of these computationally extensive problems is the solution of partial differential equations (PDEs) with uncertainties in parameters. Usual parameters affected by uncertainties are the boundary or initial conditions, coefficients (material fields) and forcing terms. Natural examples of such problems can be found for example in geosciences, where the parameters of geomaterials (aquifers) are usually highly variable. A particular example is e.g. the saturated groundwater flow. Here the permeability is unknown and there are only limited and expensive options on how to measure it. But if we use previous measurements of similar aquifers, we can build a reasonable statistical description of the permeability field (e.g. using a random field to describe porosity as in [20], [24], [15]) and approximate the porosity-permeability relationship as in [33]. In this paper, we will focus solely on the stationary Darcy flow (saturated groundwater flow).

Currently, there exists a broad palette of methods for solving such problems with uncertainties. These methods can be divided into two basic groups: intrusive and non-intrusive. Non-intrusive methods work primarily with single evaluations of deterministic models (treated as "black-boxes") using mostly an existing software. Non-intrusive methods are easy to use, but cannot utilize specific properties of the problems. This can often lead to extensive computations. Intrusive methods do not rely only on the "black-box" solvers but focus on the development of new approaches taking into account specific properties of these problems.

As the main representatives of non-intrusive approaches, we can mention the family of Monte Carlo (MC) methods [28], [10], which includes the Multi-level Monte Carlo method [4], [16] and the response surface methods which use various methods of the surrogate construction: space-filling curves [50], stochastic collocation [1], [35] or radial basis functions [11], [43].

The backbone of most of the intrusive methods is the stochastic Galerkin method (SGM), outlined in Section 3. It discretizes both the physical domain (using finite elements (FE)) and the parameter space (using e.g. polynomials). A thorough description of the method can be found in [30], [56], [2]. The Galerkin approach, in comparison to non-intrusive methods, leads to more accurate predictions and allows for a better control of numerical simulations through a posteriori error estimates [8], [17], [47] and estimation using adaptive refining [9], [26], [55], [46], [31]. The main difficulty of the use of SGM is the solution of the arising system of equations, which is very large.

The standard approach uses an iterative Krylov method with a suitable preconditioner, e.g. the block diagonal preconditioner [44], Kronecker preconditioner [53] hierarchical Schur preconditioner [52]. An improvement of the standard approach is the application of low-rank compression in each iteration, see e.g. [32], [5], [19], [29]. Theoretical results about the efficiency of the low-rank approximations can be found in [5]. A different approach is the Generalized Spectral Decomposition, where the problem is treated as an extended eigenvalue problem and the solution is built using the power-type method [36], [37], [41], [38], [40], [39]. There are also approaches specializing on specific problems with high dimensions which approximate the solution using the tensor train or the hierarchical Tucker format [27], [3], [18].

A relatively new approach to the solution of the tensor structured stochastic Galerkin systems of equations is the reduced basis (RB) approach, see Section 4. The RB solver for the elliptic problem that constructs the RB using the rational Krylov approximations can be found in [45]. Construction of the RB using MC sampling (greedy algorithm) was presented in [54]. An alternative construction of the RB using sampling via sparse grids is presented in [34], where it is applied to the mixed formulation of the elliptic problem. A comparison of the RB and the collocation approaches was presented in [13]. A thorough work on RB including the comparison between greedy and proper orthogonal decomposition approaches to the RB construction was presented in [14]. An alternative error estimator for the acceleration of the RB construction was presented in [12].

The main aim of this contribution is the development and comparison of different approaches to the construction of the RB. We are extending the results presented in [6]. Sections 2 and 3 provide the problem setting and a minimal necessary introduction to the SGM. Section 4 presents the reduced basis method and proposes several strategies for its efficient construction. Section 5 contains extensive numerical experiments showing the properties of the methods and their behavior for different problem settings. We summarize the obtained results in Section 6.

2. Problem setting

Let us consider the steady Darcy flow with uncertainties in input data. We assume that these uncertainties are represented by a finite dimensional random vector $\boldsymbol{Z}$. In the case of uncertainties given by a random process or a random field (infinite dimensional), we would need to use a truncation of the Karhunen-Loéve decomposition of such random process/field (see e.g. [30]). Therefore, all of our input data and the solution are viewed as functions of both the physical domain Ω and a random vector $\boldsymbol{Z}$: $u_D(x;\boldsymbol{Z})$, $u_N(x;\boldsymbol{Z})$, $f(x;\boldsymbol{Z})$, $k(x;\boldsymbol{Z})$, $u(x;\boldsymbol{Z})$: $\Omega \times \mathbb{R}^M \to \mathbb{R}$. The

formulation of our problem is

$$\begin{cases} -\mathrm{div}_x(k(x;\boldsymbol{Z})\cdot\nabla_x u(x;\boldsymbol{Z})) = f(x;\boldsymbol{Z}) & \forall\, x\in\Omega, \\ u(x;\boldsymbol{Z}) = u_D(x;\boldsymbol{Z}) & \forall\, x\in\Gamma_D, \\ \dfrac{\partial u(x;\boldsymbol{Z})}{\partial n(x)} = u_N(x;\boldsymbol{Z}) & \forall\, x\in\Gamma_N. \end{cases}$$

The uncertainty affecting only the right-hand side and boundary conditions $u_D(x;\boldsymbol{Z})$, $u_N(x;\boldsymbol{Z}), f(x;\boldsymbol{Z})$ (e.g. part of the random vector $\boldsymbol{Z}$) can be easily stripped away from the problems due to the superposition principle. We can use projection into orthogonal polynomials which will yield decoupled problems. Therefore, we focus only on the uncertainty contained in the permeability field $k(x;\boldsymbol{Z})$. Without loss of generality, we will restrict to the problem

$$\begin{cases} -\mathrm{div}_x(k(x;\boldsymbol{Z})\cdot\nabla_x u(x;\boldsymbol{Z})) = 0 & \forall\, x\in\Omega, \\ u(x;\boldsymbol{Z}) = u_D(x) & \forall\, x\in\Gamma_D, \\ \dfrac{\partial u(x;\boldsymbol{Z})}{\partial n(x)} = 0 & \forall\, x\in\Gamma_N, \end{cases} \tag{2.1}$$

where $\Omega = \langle 0,1\rangle^2$, $\Gamma_D = \{0,1\}\times\langle 0,1\rangle$, $\Gamma_N = \partial\Omega\setminus\Gamma_D$ and $u_D(x) = x_1$ (pressure 0 on the left-hand side of the rectangle and pressure 1 on the right-hand side of the rectangle). To complete the formulation, we need to specify $k(x;\boldsymbol{Z})$ together with some probability distribution of the vector $\boldsymbol{Z} = (Z_1,\ldots,Z_M)$. Here we assume that Ω is decomposed into subdomains Ω_i and the permeability field $k(x;\boldsymbol{Z})$ is piecewise constant. The constant value of $k(x;\boldsymbol{Z})$ on each subdomain Ω_i is controlled by Z_i. The splitting into subdomains is a natural outcome of the computed tomography or can be provided by geological investigation of an aquifer. Thus, we assume sharp knowledge of areas with different materials, but uncertainty in the knowledge of permeabilities. Additionally, we assume the log-normal distribution of the permeability on each subdomain, which is natural for the permeability of aquifers, see [24]. The random material field then takes the form

$$k(x;\boldsymbol{Z}) = \sum_{i=1}^{M} \chi_{\Omega_i}(x)\exp(\sigma_i Z_i + \mu_i),$$

where $\chi_{\Omega_i}(x)$ is a characteristic function of the subdomain Ω_i, $\exp(\sigma_i Z_i + \mu_i)$ describes the distribution of the permeability on the subdomain Ω_i, and μ_i, σ_i are the mean value and the standard deviation (std) of the underlying normal distribution. This qualitatively corresponds to the distribution of the porosity, see [33]. Then the components of $\boldsymbol{Z}$ are independent standard normal random variables.

2.1. Model problems. We want to investigate the behavior of RB construction techniques, depending on both the mean value/variance of the random variable and the geometry of the subdomain partitioning. We consider three different schemes for the partitioning of the subdomains (see Figure 2.1), which create subdomains of approximately the same sizes.

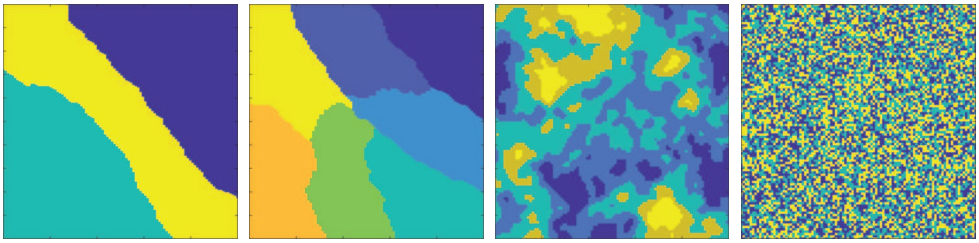


Figure 2.1. Examples of domain partition into subdomains. From the left: type 1 (3 subdomains), type 1 (7 subdomains), type 2 (5 subdomains), type 3 (3 subdomains).

Type 1: creates 100×100 grid and partitions cells to achieve the approximation of the lowest cut (number of edges between cells assigned to different domains) using the METIS package, see [25].

Type 2: generates a discretized replication of a Gaussian random field on 100×100 grid (see e.g. [30]) and divides the domain into subdomains according to the range of values.

Type 3: creates grid 100×100 and randomly assigns cells to the subdomains.

3. Stochastic Galerkin method

The SGM denotes the Galerkin method applied to the PDE with uncertain parameters in the form of functions of a random vector (here $k(x; \boldsymbol{Z})$). For a detailed introduction see [30], [7]. The SGM assumes a discretization of both the physical space (functions on the domain) and the stochastic/parametric space (functions of random variables).

We start with the variational formulation of the problem (2.1)

$$\begin{cases} \text{Find } u = u_h + u_D, & u_h \in V, \\ a(u, v) = 0, & \forall\, v \in V, \end{cases} \tag{3.1}$$

where

$$a(u, v) = \int_\Omega \left(\int_{\mathbb{R}^M} k(x; \boldsymbol{Z}) \nabla_x u(x; \boldsymbol{Z}) \nabla_x v(x; \boldsymbol{Z}) \, \mathrm{d}F\boldsymbol{Z} \right) \mathrm{d}x$$

and V is the tensor product of the Sobolev space $H_0^1(\Omega)$ and $L^2_{\mathrm{d}F\boldsymbol{Z}}(\mathbb{R}^M)$, which is the space of square integrable functions on $\mathbb{R}^M$ with respect to the distribution of $\boldsymbol{Z}$.

Well-posedness of the problem (3.1) depends on the properties of $k(x;\boldsymbol{Z})$. If $k(x;\boldsymbol{Z}) \geqslant k_{\min} > 0$ for all x, $\boldsymbol{Z}$, we can prove the existence and uniqueness of the solution using the Lax-Milgram (LM) lemma. In the case of log-normal random variables, we do not have such strong properties of k. Here we have only $\mathbb{P}(k(x;\boldsymbol{Z}) > 0) = 1$ for all x and cannot use the LM lemma directly. The proof of the well-posedness of the problem (3.1) with log-normal random variables can be found in [21].

For the Galerkin discretization, we use a finite dimensional space with a basis constructed as the tensor product of the basis $\langle \varphi_1(x), \dots, \varphi_{N_d}(x) \rangle$ of standard linear elements on Ω and the basis $\langle \psi_1(\boldsymbol{Z}), \dots, \psi_{N_s}(\boldsymbol{Z}) \rangle$ of orthonormal polynomials with respect to the distribution of $\boldsymbol{Z}$. Due to the independence of random variables Z_i, the polynomial basis consists of the products of one dimensional Hermite polynomials. The basis functions which correspond to degrees of freedom on the Dirichlet part of the boundary $\partial\Omega$ are not considered as a part of the aforementioned FE basis.

Due to the separable nature of the material field, the values of the bilinear form on elements of the tensor product basis can be written as

$$a(\varphi_i\psi_j, \varphi_k\psi_l) = \sum_{m=1}^{M} \int_{\mathbb{R}^M} \psi_j\psi_l \exp(a_m Z_m + b_m)\,\mathrm{d}F\boldsymbol{Z} \int_{\Omega_m} \nabla\varphi_i \nabla\varphi_k \,\mathrm{d}x.$$

This leads to a large system ($N_s \times N_d$) of linear equations in the form of

$$A \cdot \overline{u}_h = \overline{b}, \tag{3.2}$$

where

$$A = \sum_{m=1}^{M} G_m \otimes K_m \quad \text{and} \quad b = \sum_{m=1}^{M} g_m \otimes f_m,$$

$$(G_m)_{j,l} := \int_{\mathbb{R}^M} \psi_j\psi_l \exp(a_m Z_m + b_m)\,\mathrm{d}F\boldsymbol{Z}, \quad (K_m)_{i,k} := \int_{\Omega_m} \nabla\varphi_i\nabla\varphi_k \,\mathrm{d}x,$$

$$(g_m)_j := \int_{\mathbb{R}^M} \psi_j \exp(a_m Z_m + b_m)\,\mathrm{d}F\boldsymbol{Z}, \quad (f_m)_i := \int_{\Omega_m} \nabla\varphi_i\nabla u_D \,\mathrm{d}x.$$

Matrices and vectors G_m, g_m correspond to the discretization of the parametric space (orthogonal polynomials). Matrices and vectors K_m, f_m correspond to the discretization of the physical space (FE).

3.1. Assembling of SGM systems. Assembling of the SGM system consists of two contributions of different origins. The matrix K_m and the vector f_m come from the finite element discretization and they can be assembled by the standard FE assembling software. The more difficult part is the assembly of G_m and g_m.

We assume normalized Hermite polynomials $\psi_i(\boldsymbol{Z})$. Note that each $\psi_i(\boldsymbol{Z}) = \prod_{j=1}^{M} \psi_{i,j}(Z_j)$ is a product of one dimensional normalized Hermite polynomials. The elements of the matrix G_m can be calculated as

$$\begin{aligned}(G_m)_{k,l} &= \int_{\mathbb{R}^M} \psi_k(\boldsymbol{Z})\psi_l(\boldsymbol{Z}) \exp(a_m Z_m + b_m)\,\mathrm{d}F\boldsymbol{Z} \\ &= \int_{\mathbb{R}} \psi_{k,m}(Z_m)\psi_{l,m}(Z_m) \exp(a_m Z_m + b_m)\,\mathrm{d}FZ_m \\ &\quad \times \prod_{i\neq m} \int_{\mathbb{R}} \psi_{k,i}(Z_i)\psi_{l,i}(Z_i)\,\mathrm{d}FZ_i,\end{aligned}$$

where $\prod_{i\neq m} \int_{\mathbb{R}} \psi_{k,i}(Z_i)\psi_{l,i}(Z_i)\,\mathrm{d}FZ_i$ is either 0 if some of the polynomials $\psi_{k,i}$ do not equal $\psi_{l,i}$, or 1 otherwise. The rest of the formula can be evaluated according to the following equalities

$$\begin{aligned}\int_{\mathbb{R}} f(Z)\exp(aZ+b)\,\mathrm{d}FZ &= \int_{\mathbb{R}} f(Z)\exp(aZ+b)\frac{1}{\sqrt{2\pi}}\exp\Big(-\frac{Z^2}{2}\Big)\,\mathrm{d}Z \\ &= \exp\Big(\frac{a^2}{2}+b\Big)\int_{\mathbb{R}} f(Z)\frac{1}{\sqrt{2\pi}}\exp\Big(-\frac{(Z-a)^2}{2}\Big)\,\mathrm{d}Z \\ &= \exp\Big(\frac{a^2}{2}+b\Big)\int_{\mathbb{R}} f(Z)\,\mathrm{d}F\widetilde{Z},\end{aligned}$$

where $\widetilde{Z} \sim \mathcal{N}(a,1)$. In the case of building matrices G_m, the function $f(Z)$ is a product of two polynomials (not orthogonal in measure $\mathrm{d}F\widetilde{Z}$). The integral $\int_{\mathbb{R}} f(Z)\,\mathrm{d}F\widetilde{Z}$ can be precisely computed using the Gauss-Hermite quadrature rule. Note that $\int_{\mathbb{R}} f(Z)\,\mathrm{d}F\widetilde{Z}$ can be also evaluated analytically (if f is polynomial), but the analytic formula can be numerically unstable due to operations between small and large values. The use of the Gaussian quadrature rule is also applicable to other types of random variables. It converges very fast because the integrand is usually an analytic function.

4. Reduced basis method

The work presented here is the continuation of [6], which extends the work in [45], [34]. For theoretical background see [45], here we deal with details only briefly.

We start with the equality

$$(A\otimes B)u = \mathrm{vec}(BUA^{\top}),\ U = \mathrm{mat}(u), \tag{4.1}$$

where $\mathrm{mat}(\cdot)$ denotes the operation of reshaping a vector into a matrix of appropriate size and $\mathrm{vec}(\cdot)$ denotes the operation of unfolding the matrix into a vector. Therefore, the tensor structure of the matrix $\mathbb{A}$ allows the system (3.2) to be viewed as the matrix equation

$$\sum_{m=1}^{M} K_m X G_m^\top = \sum_{m=1}^{M} f_m g_m^\top, \tag{4.2}$$

where $X = \mathrm{mat}(\overline{u}_h)$ denotes a matrix of size $N_d \times N_s$ arising from reshaping the original vector of unknowns $\overline{u}_h$ from (3.2), N_d is the number of degrees of freedom of the FE basis and N_s is the number of degrees of freedom of the polynomial basis. This formulation is algebraically equivalent to the system of equations (3.2), where the matrix and the right-hand side are sums of Kronecker products of matrices.

The core of the RB method is the assumption that there exists a good low-rank approximation $\widetilde{X}$ of the solution X of the equation (4.2):

$$X \approx \widetilde{X} = WY, \quad W \in \mathbb{R}^{N_d \times N_k}, \quad Y \in \mathbb{R}^{N_k \times N_s}. \tag{4.3}$$

The RB method then seeks one of the factors W, Y. In the context of (4.2), the factor W reduces the physical (spatial) dimension and the factor Y reduces the parametric dimension. Here, we aim at constructing the factor W. This is because the FE basis has usually much more degrees of freedom and its approximation using the RB allows us to incorporate existing FE software. In further text, we will address W as the RB of the problem. We describe the specifics of the RB construction using two different approaches in Subsections 4.3 and 4.4, we also consider a combination of these approaches in Subsection 4.5.

Our strategy is to construct the RB W iteratively. The RB iterations provide a sequence $W_0, W_1, \ldots$ In kth iteration ($k \geqslant 1$), we construct the RB W_k by extending the existing basis W_{k-1},

$$W_k = [w_{1,1}, \ldots, w_{1,n_1}, \ldots, w_{k,1}, \ldots, w_{k,n_k}] = [W_{k-1}, V_k] \in \mathbb{R}^{N_d \times N_k}, \quad N_k = \sum_{i=1}^{k} n_i,$$

where n_i denotes the number of new vectors added to the RB in the ith iteration and V_k is a matrix of n_k vectors added in the kth iteration.

We construct the RB in such a way that W_k is orthonormal ($W_k^\top W_k = I$). The enhancement of the RB in the kth iteration V_k is built in the following way:

- ▷ suggest new vectors $\widetilde{V}_k$ which expand the RB; this will be done by one of the methods described in Subsections 4.3, 4.4, and 4.5,

- ▷ orthogonalize $\widetilde{V}_k$ with respect to W_{k-1}: $\overline{V}_k = \widetilde{V}_k - W_{k-1}(W_{k-1}^\top \widetilde{V}_k)$,
- ▷ finally, we obtain V_k by orthonormalizing and compressing $\overline{V}_k$ using e.g. the singular value decomposition (SVD); in further text, we use drop tolerance for the compression step 10^{-12} (the threshold for singular values), this will filter vectors too close to the linear span of W_{k-1}.

After its construction, W_k serves for computation of a low-rank approximation $X_k = W_k Y_k$ of the solution X, and Y_k is a reduced solution, which can be obtained using the orthogonality between the residual of X_k and the RB W_k. The residual of the low-rank approximation X_k is

$$R_k := \sum_{m=1}^{M} K_m (W_k Y_k) G_m^\top - \sum_{m=1}^{M} f_m g_m^\top. \tag{4.4}$$

By applying the Galerkin orthogonality condition $W_k^\top R_k = 0$, we obtain the system

$$\sum_{m=1}^{M} W_k^\top K_m W_k Y_k G_m^\top = \sum_{m=1}^{M} W_k^\top f_m g_m^\top.$$

It can be easily seen that Y_k is the solution of the matrix equation that has the same structure as the original system, i.e.

$$\sum_{m=1}^{M} \overline{K}_{m,k} Y_k G_m^\top = \sum_{m=1}^{M} \overline{f}_{m,k} g_m^\top, \tag{4.5}$$

where $\overline{K}_{m,k} = W_k^\top K_m W_k$, $\overline{f}_{m,k} = W_k^\top f_m$. The dimension of the new system (4.5) is greatly reduced in comparison to the original system (4.2) and can be more easily solved by Krylov subspace iterative methods. We will elaborate further in Subsection 4.1.

The RB method can be now viewed as an iterative method with the kth iteration consisting of the following steps:

- ▷ construction of W_k,
- ▷ assembly and solution of the reduced system (4.5),
- ▷ construction of the approximation of the solution $X_k = W_k Y_k$ and evaluation of the approximation error.

We can evaluate the approximation error using the relative residual of X_k. This allows us to compare the results/convergence of the RB method with e.g. Krylov subspace methods applied to the original system. But the computation of the residual is very computationally/memory expensive because X_k is a dense matrix of large size. We briefly discuss some possibilities of error estimation of X_k in Subsection 4.2.

4.1. Solution of the reduced system. If we view the system (4.5) in the "system of linear equations" format, the matrix of the system is symmetric and positive definite (SPD). Therefore, we use the conjugate gradients method for its solution. For our numerical experiments in Section 5, we choose the mean value preconditioner

$$P = I \otimes W^\top K_0 W,$$

where K_0 denotes the stiffness matrix corresponding to the mean value of $k(x; \boldsymbol{Z})$, see [44]. Note that the application of the preconditioner can be viewed as the solution of the system with a block diagonal SPD matrix (see (4.1)).

We also suggest some techniques which aim at improving the efficiency of the solution of the reduced system (4.5):

- ▷ The matrices $\overline{K}_{m,k}$ and the right-hand sides $\overline{f}_{m,k}$ from the reduced system (4.5) can be gradually extended with iterations of W_k. The structure $W_k = [W_{k-1}, V_k]$ implies that

$$\overline{K}_{m,k} = W_k^\top K_m W_k = \begin{bmatrix} \overline{K}_{m,k-1} & W_{k-1}^\top K_m V_k \\ V_k^\top K_m W_{k-1} & V_k^\top K_m V_k \end{bmatrix},$$
$$\overline{f}_{m,k} = W_k^\top f_m = [\overline{f}_{m,k-1}, V_k^\top f_m].$$

- ▷ As we compute the solution of the reduced system (4.5) in every RB iteration, we can use the solution from the $(k-1)$st iteration filled by zeros $[Y_{k-1}^\top; 0]^\top$ as an initial guess for the kth iteration.
- ▷ We compute the reduced solution Y_k using the conjugate gradient method and obtain its approximation $\widetilde{Y}_k$ with the residual r_k. Therefore, we obtain the approximation $\widetilde{X}_k = W_k \widetilde{Y}_k$ with the residual $\widetilde{R}_k$. Note that the residual R_k of $X_k = W_k Y_k$ is fully determined by W_k. For the obtained residual $\widetilde{R}_k$, we have (using $\widetilde{W}_k = I \otimes W_k$, which represents the reduced basis W_k as the reduce basis for the linear system (3.2))

$$\begin{aligned} (4.6) \qquad \|\widetilde{R}_k\| &= \|A\widetilde{W}_k \mathrm{vec}(\widetilde{Y}_k) - \overline{b}\| = \|A\widetilde{W}_k(\mathrm{vec}(Y_k) + (\widetilde{W}_k^\top A\widetilde{W}_k)^{-1}\mathrm{vec}(r_k)) - \overline{b}\| \\ &\leqslant \|R_k\| + \|A\widetilde{W}_k(\widetilde{W}_k^\top A\widetilde{W}_k)^{-1}\|\|r_k\| = \|R_k\| + c_k\|r_k\|, \end{aligned}$$

  where c_k is an appropriate constant reflecting the properties of the matrix of the system (3.2) and W_k. The residual $\|R_k\|$ will be gradually smaller in each consecutive RB iteration. Therefore, we can adaptively adjust the precision of the iterative solution of the reduced system (4.5), so the contribution $c_k\|r_k\|$ to $\|\widetilde{R}_k\|$ is low enough. We do not need to know the value of c_k explicitly. It is sufficient to start the RB method with low accuracy for the reduced system and gradually increase it in the following iterations. If the decrease of $\|\widetilde{R}_k\|$ is almost the same

as the decrease of $\|r_k\|$, the precision of $\widetilde{Y}_k$ is likely insufficient and the residual error $\|\widetilde{R}_k\|$ is driven by $c_k\|r_k\|$. Otherwise, we found a reasonable precision for the current iteration and we can adjust it according to the decrease of $\|\widetilde{R}_k\|$ in the following iterations. For the numerical experiments see Subsection 5.7.

4.2. Stopping criterion. As mentioned above, the calculation of the residual (4.4) is memory and computationally very expensive. This is due to large dimensions of the dense matrix of the solution X_k. One possibility of dealing with memory complexity is to exploit specific properties of the tensor structure of the system matrix, compute the residual by parts and avoid the explicit construction of X_k. The squared norm of the residual takes the form

$$\|R_k\|^2 = \left\| \sum_{m=1}^{M} K_m(W_k Y_k) G_m^\top - \sum_{m=1}^{M} f_m g_m^\top \right\|^2.$$

First, we multiply factors of X_k by the corresponding matrices $\widetilde{K}_{m,k} := K_m W_k$, $\widetilde{G}_{m,k} := Y_k G_m^\top$. Matrix $\widetilde{G}_{m,k}$ is not needed as a whole, but it is sufficient to construct it by columns $[\widetilde{G}_{m,k}]_{\cdot,i}$. This will reduce the memory needed for the computation from $N_d \cdot N_s$ entries of X_k to $M \cdot N_k(N_d+1)$ entries for $\widetilde{K}_{m,k}, [\widetilde{G}_{m,k}]_{\cdot,i}$. Matrices $\widetilde{K}_{m,k}$ can be also constructed gradually as $\widetilde{K}_{m,k} = [\widetilde{K}_{m,k-1}, V_k K_m]$. Note that this is suitable only if $M \cdot N_k \ll N_s$, otherwise we do not save much memory. Then we can construct the squared norm by columns of the residual

$$\|R_k\|^2 = \sum_{i=1}^{N_s} \left\| \sum_{m=1}^{M} \widetilde{K}_{m,k} [\widetilde{G}_{m,k}]_{\cdot,i} - f_m [g_m^\top]_{\cdot,i} \right\|^2.$$

This will also lower the computational complexity because we do not need to multiply large sparse matrices K_m, G_m by the large dense matrix X_k.

Another option is to control the step size $\|X_k - X_{k-1}\|$ instead of the residual. As we do not want to construct solutions explicitly, we can estimate the step size using the reduced solutions Y_k (recall that $W_k = [W_{k-1}, V_k]$ and it is orthonormal)

$$\|X_k - X_{k-1}\|_2 = \|W_k Y_k - W_{k-1} Y_{k-1}\|_2 = \left\| W_k Y_k - W_k \begin{bmatrix} Y_{k-1} \\ 0 \end{bmatrix} \right\|_2$$
$$= \sqrt{\operatorname{trace}\left(\left(Y_k - \begin{bmatrix} Y_{k-1} \\ 0 \end{bmatrix} \right)^\top W_k^\top W_k \left(Y_k - \begin{bmatrix} Y_{k-1} \\ 0 \end{bmatrix} \right)^\top \right)} = \left\| Y_k - \begin{bmatrix} Y_{k-1} \\ 0 \end{bmatrix} \right\|_2.$$

This is very cheap, but it does not grant a lower bound for the error. Numerical experiments concerning the behavior of the step size, residual and error of the reduced solutions are presented in Subsection 5.6.

4.3. Monte Carlo approach. The Monte Carlo sampling is the simplest approach to the RB construction. The core of this approach is to sample realizations of $k(x; \boldsymbol{Z})$ and construct the RB from the corresponding solutions of the problem (2.1). We will call this approach the crude MC approach.

In the crude MC approach, each sample $k(x; \boldsymbol{Z})$ will lead to a solution of one deterministic system. This can be inefficient when the samples do not carry additional information to the RB. This can be caused e.g. by the close distance between samples or generally similar solutions for the samples. The crude MC approach can be modified to lower the possibility of such samples. This can be done in two main ways:

- ▷ Avoiding already generated samples on the level of the sampler. This approach is computationally inexpensive but does not directly translate to samples with a good contribution to the RB. We will denote this technique as "avoiding sampling".
- ▷ Proposing multiple samples $k(x; Z_i)$ which leads to systems $A_i u_i = b_i$. Computing the reduced solution $\widetilde{u}_i$ in the current RB W_k

$$W_k^\top A_i W_k \widetilde{u}_i = W_k^\top b_i$$

and evaluating the residual error of the reduced solution $\widetilde{u}_i$

$$R_i = A_i W_k \widetilde{u}_i - b_i.$$

Then we pick the sample with the highest residual error as the sample for the expansion of the RB. This approach will ensure that the new sample carries additional information, but it is computationally expensive.

The first modification, which should avoid the already generated samples, can be carried out in various ways. The simplest one would be the latin hypercube sampling, see [28]. The major disadvantage of the latin hypercube sampling is the fact that we can only generate N samples at one time and cannot add samples later. This is quite prohibitive for our intended usage. We propose a different approach to the sampling of $k(x; Z_i)$, which is done by altering the probability density function (pdf) of the sampler. This can be done e.g. by lowering the pdf around the already generated samples. Samples of $k(x; \boldsymbol{Z})$ are given solely by the value of $\boldsymbol{Z}$. Therefore, we will work with samples and the pdf of $\boldsymbol{Z}$. Assuming $f(\boldsymbol{Z})$ the pdf of $\boldsymbol{Z}$ and n the already generated samples X_i, the altered pdf function can take the form

$$\widetilde{f}_n(\boldsymbol{Z}) \propto f(\boldsymbol{Z}) \cdot \min_{i=1,\dots,n} w_i(\boldsymbol{Z}), \quad w_i(\boldsymbol{Z}) := 1 - \exp\Big(-\frac{1}{2}(\boldsymbol{Z} - X_i)^\top \Sigma^{-1} (\boldsymbol{Z} - X_i)\Big),$$

where $\propto$ denotes proportionality and Σ is a SPD matrix. The functions $w_i(\boldsymbol{Z})$ are zero at X_i and increase up to 1 with increasing distance from X_i. A reasonable choice

of Σ is the covariance matrix of $\boldsymbol{Z}$. Samples from $\widetilde{f}_n(\boldsymbol{Z})$ can be generated using the Metropolis-Hastings (MH) method, see [48]. Note that this way of sampling does not provide samples from the distribution given by $f(\boldsymbol{Z})$.

We can also compare the MC approach with some deterministic approach of sampling the points (replications of $\boldsymbol{Z}$). Therefore, we need a way of reasonably filling the space with points which allows iterative refinements. A possible way how to do this is to use integration points of nested sparse grids. A description of the nested Smolyak sparse grids with the Gaussian weight can be found in [42]. For our purposes, we will use the implementation of nested sparse grids from [23]. This approach of sampling points to construct the RB was used in [34].

Numerical experiments concerning the MC approach and the impact of its two modification are presented in Subsection 5.1.1.

4.4. Reduced rational Krylov subspace approach. The reduced rational Krylov subspace (RRKS) approach to the construction of the RB was presented in [45]. For more information about the rational Krylov subspace (RKS) approximation of matrix functions see [22].

In our application, we will operate only with SPD matrices. Therefore, the following definitions and relations are constrained only to these matrices although some of them can be extended to more general types of matrices.

Definition 4.1. Let A be a real SPD matrix with eigenvalues $\lambda_1, \ldots, \lambda_n$, $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_{k-1}]$, $\alpha_i \in \mathbb{R} \setminus \{\lambda_1, \ldots, \lambda_n\}$, and let v be a real nonzero vector. Then

$$\mathcal{K}_k(A, v, \boldsymbol{\alpha}) := \operatorname{span}\left\{ v, (A - \alpha_1 I)^{-1} v, \ldots, \prod_{j=1}^{k-1} (A - \alpha_j I)^{-1} v \right\}$$

is the RKS of order k associated with the matrix A, the vector v and parameters $\boldsymbol{\alpha}$.

Definition 4.1 is taken from [51]. Note that Definition 4.1 is specific to our problem and there are also other definitions of the RKS.

Here, we use the RRKS to create a subspace with a given set of SPD matrices $\{A_m\}_{m=1}^M$ and nonzero vectors $\{v_m\}_{m=1}^M$. We assume one common parameter α for all matrices, additionally we assume this parameter to be positive. We define partial subspace

$$\mathbb{K}(\alpha, v) := \operatorname{span}\{(A_m + \alpha I)^{-1} v\}_{m=1}^M.$$

The RB is built iteratively. We start with $W_0 = \operatorname{orth}(\{v_m\}_{m=1}^M)$, where $\operatorname{orth}(V)$ denotes the orthonormalization of the columns of V. Then we enhance W_{i-1} by $\mathbb{K}(\alpha_i, w_i)$ in each RB iteration. The vector w_i, which generates the expansion to the RB, is chosen as the ith column of W_{i-1} and α_i is the parameter of the RKS.

Other approach to the choice of w_i is discussed in Subsection 4.4.2. The choice of parameters α_i is generally hard (with respect to the approximation quality of W), we will focus on a single parameter value $\alpha_1 = \alpha_2 = \ldots = \alpha_{k-1}$ as it was a viable choice presented in [45]. The choice of this single parameter is numerically tested in Subsection 5.1.2.

In our case the matrices $\{K_m\}_{m=1,\ldots,M}$ are symmetric but only positive semidefinite. In the general case, these matrices can by indefinite if a decomposition of the permeability field

$$k(x; \boldsymbol{Z}) = \sum_{i=1}^{M} k_i(x) g_i(\boldsymbol{Z})$$

leads to factors $k_i(x)$ with negative values. To effectively use the RKS, we need to transform the problem. The matrices corresponding to the physical discretization need to be SPD and ideally with spectra in some small interval. We will proceed as in [45]. Recall the permeability field taking the form

$$k(x; \boldsymbol{Z}) = \sum_{i=1}^{M} \chi_{\Omega_i}(x) g_i(\boldsymbol{Z}).$$

We want the matrices K_i to contain as much information about $k(x; \boldsymbol{Z})$ as possible. Therefore, we assume that the matrices K_i correspond to the permeability fields $\chi_{\Omega_i}(x)\mathbb{E}(g_i(\boldsymbol{Z}))$, i.e., K_i contains additional information about the mean value of the permeability. Note that $\mathbb{E}(g_i(\boldsymbol{Z}))$ is a positive constant. Then we define the mean value matrix K_0 as the sum of K_i. The matrix K_0 is SPD. This is easy to see because in our case

$$\bigcup_{i=1}^{M} \Omega_i = \Omega \quad \text{and} \quad \mathbb{E}(g_i(\boldsymbol{Z})) > k_0 > 0 \quad \forall\, i = 1, \ldots, M.$$

Next, assume a factorization $K_0 = LL^\top$, where L can be e.g. the Cholesky factor of K_0.

R e m a r k 4.2. Let K_i be the finite element matrix corresponding to the problem (2.1) and the permeability field $k_i(x) = \chi_{\Omega_i}(x)\mathbb{E}(g_i(\boldsymbol{Z}))$. Then K_i are symmetric positive semidefinite and their sum

$$K_0 = \sum_{i=1}^{M} K_i$$

is SPD with possible factorization $K_0 = LL^\top$. Then the spectrum of $L^{-1}K_iL^{-\top}$ is contained in $\langle 0, 1\rangle$.

P r o o f. $L^{-1}K_iL^{-\top}$ is symmetric. The largest eigenvalue of $L^{-1}K_iL^{-\top}$ can be expressed as (L is regular)

$$\lambda_{\max} = \sup_{x\colon \|x\|\neq 0} \frac{x^\top L^{-1}K_iL^{-\top}x}{x^\top x} = \sup_{y\colon \|y\|\neq 0} \frac{y^\top K_i y}{y^\top K_0 y} = \sup_{y\colon \|y\|\neq 0} \frac{y^\top K_i y}{\sum_{j=1}^M y^\top K_j y}.$$

Therefore, $\lambda_{\max} \leqslant 1$. Similarly, because K_i is positively semi-definite, $\lambda_{\min} \geqslant 0$. □

We can transform the system (4.2) to

$$\sum_{m=1}^M L^{-1}K_mL^{-\top}L^\top X G_m^\top = \sum_{m=1}^M L^{-1}f_m g_m^\top.$$

According to Remark 4.2, the spectrum of each $L^{-1}K_mL^{-\top}$ is in $\langle 0, 1\rangle$. After this transformation, we will seek the solution $\widetilde{X} = L^\top X$ instead of X. We will obtain SPD matrices by shifting the spectra by some positive value. This leads to the formulation

$$\sum_{m=1}^M (L^{-1}K_mL^{-\top} + \beta I)\widetilde{X}G_m^\top - \beta\sum_{m=1}^M \widetilde{X}G_m^\top = \sum_{m=1}^M L^{-1}f_m g_m^\top.$$

Now we have obtained the set of SPD matrices $A_m = (L^{-1}K_mL^{-\top} + \beta I)$, all with spectra in $\langle \beta, 1+\beta\rangle$.

Finally, we can implement RRKS with matrices $\{L^{-1}(K_m + \beta K_0)L^{-\top}\}_{m=1}^M$ and vectors $\{L^{-1}f_m\}_{m=1}^M$ to approximate the transformed solution $\widetilde{X}$. The shift β can be merged with RKS parameters α_i.

4.4.1. Construction of RB without the factor L. The usefulness of the aforementioned approach is limited by the need to build the factor L. We propose an alternative construction of the RB, which leads to the same approximation space without the factor L. First, let us write down how the vectors of W (approximating $\widetilde{X} = L^\top X$) look like. The first $i = 1, \ldots, M$ iterations of the RB will enhance the basis by the span of

$$\{L^\top(K_m + \alpha K_0)^{-1}\underbrace{LL^{-1}}_{I}f_i\}_{m=1}^M.$$

Other $i = M + k + M\cdot(l-1)$ ($k = 1, \ldots, M$, $l = 1, \ldots, M$) iterations will enhance W_{i-1} by the span of

$$\{L^\top(K_m + \alpha K_0)^{-1}\underbrace{LL^\top}_{K_0}(K_k + \alpha K_0)^{-1}\underbrace{LL^{-1}}_{I}f_l\}_{m=1}^M.$$

The process will continue with the same pattern. Note that the factor L remains only in front of the expression, the others are eliminated or form the matrix K_0. Therefore, the RB W can be expressed as $W = L^\top \widetilde{W}$. The RB $\widetilde{W}$ can be built in the same way as W, but with matrices $\{K_0^{-1}K_m\}_{m=1,\dots,M}$, starting vectors $\{K_0^{-1}f_m\}_{m=1,\dots,M}$ and RRKS parameter α. The first $i = 1, \dots, M$ iterations expand $\widetilde{W}_{i-1}$ by the span of

$$\{(K_m + \alpha K_0)^{-1} \underbrace{K_0 K_0^{-1}}_{I} f_i\}_{m=1}^{M}$$

and next $i = M + k + M \cdot (l-1)$ $(k = 1, \dots, M,\ l = 1, \dots, M)$ iterations expand $\widetilde{W}_{i-1}$ by the span of

$$\{(K_m + \alpha K_0)^{-1} K_0 (K_k + \alpha K_0)^{-1} \underbrace{K_0 K_0^{-1}}_{I} f_l\}_{m=1}^{M}.$$

Now, if the basis $W = L^\top \widetilde{W}$ approximates $\widetilde{X} = L^\top X$, we can use the basis $\widetilde{W}$ directly to approximate X. We can view the system as the system of linear equations with the corresponding RB $\overline{W} = I \otimes W$. Then, Remark 4.3 provides that the obtained approximation of X is the same whether using W or $\widetilde{W}$.

R e m a r k 4.3. Let A, B be regular matrices and $\widetilde{W}$ a RB designed to approximate the solution of the system

$$Ax = b,$$

i.e. $x \approx \widetilde{W}y$, $\widetilde{W}^\top A \widetilde{W} y = \widetilde{W}^\top b$. Further, let $W = \mathrm{orth}(B^\top \widetilde{W})$ be the RB for the transformed system

$$AB^{-\top}\widetilde{x} = b,\ x = B^{-\top}\widetilde{x},$$

or equivalently

$$B^{-1}AB^{-\top}\widetilde{x} = B^{-1}b,$$

i.e. $\widetilde{x} \approx W\widetilde{y}$, $W^\top B^{-1} A B^{-\top} W \widetilde{y} = W^\top B^{-1} b$. Then the reduced approximation of the solution is the same for both the original and the transformed system

$$x \approx \widetilde{W}y = B^{-\top} W \widetilde{y} \approx B^{-\top}\widetilde{x}.$$

P r o o f. W can be expressed in the form $W = B^\top \widetilde{W} G$, where G is a regular matrix, which arises from the orthonormalization of $B^\top \widetilde{W}$ (e.g. using the Gram-Schmidt process). The reduced approximation of the system $Ax = b$ with the RB $\widetilde{W}$ is $x \approx \widetilde{W}y$, where

$$y = (\widetilde{W}^\top A \widetilde{W})^{-1} \widetilde{W}^\top b.$$

The reduced approximation of the transformed system $B^{-1}AB^{-\top}\widetilde{x} = B^{-1}b$ with RB $W = B^{\top}\widetilde{W}G$ is $\widetilde{x} \approx B^{\top}\widetilde{W}G\widetilde{y}$, where

$$\widetilde{y} = G^{-1}(\widetilde{W}^{\top}A\widetilde{W})^{-1}\widetilde{W}^{\top}b \to \widetilde{y} = G^{-1}y.$$

Therefore, $\widetilde{x} \approx W\widetilde{y} = B^{\top}\widetilde{W}y$. □

4.4.2. Approaches to expansion vector choices. In the previous text, we assumed that the enhancement of the RB is based on the ith vector of the basis. This is not very optimal because we use vectors regardless of their quality. We can try to predict the quality of expansions given by vectors by a weighting. Then we use vector with the highest weight, which was not used yet. We will consider two different weighting schemes:

(1) Default, i.e. ordering of the vectors—the ith iteration uses the ith vector (as in [45]).
(2) Using the norms of the corresponding rows of the reduced solution matrix Y_i. When we build the approximation of the solution $X_k = W_kY_k$, the values in rows of Y_k correspond to the contribution of each vector of W_k to the reduced approximation. We can use the norms of these vectors of values as an indicator of how much is the vector significant in the reduced solution.

Numerical tests of these approaches can be found in Subsection 5.1.2.

4.5. Combination of MC and RRKS approach. Subsections 4.3 and 4.4 described two different approaches to the construction of the RB. Each of these approaches may be better for different types of problems. This can be a motivation for combining them to create a more robust approach. The combination of the MC and the RRKS approaches will be done on the level of RB iterations. This means that in some RB iterations we will expand the RB using the MC approach and in others using the RRKS approach. We propose two different schemes of choosing which approach will be performed in ith RB iteration:

▷ Alternate iterations—we perform the RRKS approach in odd iterations and the MC approach in even iterations.
▷ Weighted iterations—we try to estimate which approach is better in the current state of the RB by looking at the step size $\|X_i - X_{i-1}\|$. We start with alternate iterations for the first $2n$ iterations. Then we choose the approach using the sum of the step size in the last n RB iterations using the MC approach and the last n RB iterations using the RRKS approach. The approach with a larger sum of past step sizes will perform the current RB iteration.

4.6. Acceleration of reduced basis construction using deflated conjugate gradient method. During the construction of RB, we need to solve multiple similar FE systems. In the case of RRKS approach, these systems consist of M different matrices $(K_m + \alpha K_0)$ and various right-hand sides. In the case of the MC approach, these systems are the FE systems arising from random samples of $\boldsymbol{Z}$. Our aim is to speed up the solution of the subsequent systems using the information from the solution of the previous systems, more accurately we will use the available RB as the deflation space.

Deflated conjugate gradients (DCG) method is an extension of the standard conjugate gradient (CG or PCG if using a preconditioner) method, see [49]. The DCG method takes an additional parameter in the form of the deflation basis W. The deflation basis W should be able to describe the sought solution reasonably well. Therefore, the DCG method looks for the solution outside of the deflation basis W. In our application, we use the current RB as a deflation basis W.

The core of the DCG is to project the residual (or the preconditioned residual) using the projector $P = I - W(W^\top AW)^{-1}W^\top A$ during the CG routine. The most significant additional cost of this step is the solution of coarse systems with matrix $Q_i = W_i^\top AW_i$ (in ith iteration of the RB solver). In our application, the size of the matrix Q_i (corresponding to the size of the RB) is reasonably small and we can use e.g. explicit inversion. Additionally, if we have some fixed set of matrices (RRKS approach) and use the explicit inversion, we can exploit adaptive updates of deflation basis $W_i = [W_{i-1}, V_i]$ and calculate Q_i^{-1} using the Schur complement with known Q_{i-1}^{-1}. The deflation basis in the ith iteration of the RB solver is

$$W_i = [W_{i-1}\, V_i], Q_i = \begin{bmatrix} W_{i-1}^\top AW_{i-1} & W_{i-1}^\top AV_i \\ V_i^\top AW_{i-1} & V_i^\top AV_i \end{bmatrix},$$

where $W_{i-1}^\top AW_{i-1} = Q_{i-1}$. We reuse Q_{i-1}^{-1} from the previous iteration. Denoting $B_i = V_i^\top AW_{i-1}$ and $C_i = V_i^\top AV_i$, we get the following update for the calculation of the inversion:

$$Q_i^{-1} = \begin{bmatrix} Q_{i-1} & B_i^\top \\ B_i & C_i \end{bmatrix}^{-1} = \begin{bmatrix} Q_{i-1}^{-1} + Q_{i-1}^{-1}B_i^\top S_i^{-1}B_iQ_{i-1}^{-1} & -Q_{i-1}^{-1}B_i^\top S_i^{-1} \\ -S_i^{-1}B_iQ_{i-1}^{-1} & S_i^{-1} \end{bmatrix},$$

where $S_i = C_i - B_iQ_{i-1}^{-1}B_i^\top$ and has the dimension equal to the number of columns of V_i (can be one). Note that the solution with Q_i has to be performed with a sufficient precision otherwise we can encounter stagnation or break down of the DCG.

Numerical experiments considering the number of iterations saved using DCG compared to PCG can be found in Subsection 5.5.

5. Numerical experiments

The purpose of this section is to provide extensive numerical experiments of all aspects of the solution of the stochastic Galerkin equations by the introduced RB methods. We will assume the following notation and the settings of the methods and problems:

- ▷ according to Subsection 2.1, we assume three types of geometries denoted as $P1$, $P2$, $P3$,
- ▷ the number of subdomains, and therefore the number of random variables, is denoted by M,
- ▷ the behavior of the material on each subdomain is given by a random variable Z_i with mean value μ_i and std σ_i, the mean values μ_i are considered constant (except in Subsection 5.3.1) and given by Table 1,
- ▷ different FE grid sizes correspond to a regular discretization $100i \times 100i$ (denoted as "grid lvl i"); if not specifically stated otherwise, "grid lvl 1" is used in further tests,
- ▷ multiple sizes of complete polynomial bases with a given maximal degree; in further tests, we use polynomials up to degree 4, if not specifically stated otherwise,
- ▷ parameters of the methods are chosen according to the results of experiments in Subsection 5.1.

M \ i	1	2	3	4	5	6	7
1	-6						
2	-15	-6					
3	-15	-10.5	-6				
4	-15	-12	-9	-6			
5	-15	-12.75	-10.5	-8.25	-6		
6	-15	-13.2	-11.4	-9.6	-7.8	-6	
7	-15	-13.5	-12	-10.5	-9	-7.5	-6

Table 1. Values of μ_i for different number of variables M.

Apart from the testing of the settings and modifications of selected approaches to the construction of RB, we will test the impact of different geometries of a subdomain, mean values μ_i and stds σ_i. We can view the analytic solution of problem (2.1) in the generalized spectral decomposition form

$$u(x;\boldsymbol{Z}) = \sum_{i=1}^{\infty} \lambda_i \varphi_i(x) \psi_i(\boldsymbol{Z}),$$

where $\varphi_i(x)$ and $\psi_i(\boldsymbol{Z})$ are orthonormal bases of the corresponding spaces. Then the rate of the decay of the spectra λ_i determines the efficiency of the low-rank approximation. Our assumption is that different settings lead to a different rate of decay of the spectra of the solution operator. This would mean that for some problems a low-rank approximation does not bring a sufficiently low error. We expect that these types of behavior will decrease the rate of decay of the spectra:

- ▷ an increasing number of subdomains (and variables),
- ▷ more complex geometries; we assume that $P3$ is the most complex geometry, in comparison with $P2$ and then $P1$,
- ▷ higher contrast of materials on subdomains (higher differences in mean values μ_i),
- ▷ higher stds σ_i (larger state space to cover).

5.1. Impact of solvers parameters choice. Approaches presented in Section 4 have several parameters. Here we test the impact of these parameters on the convergence of the RB method and choose some setting for further usage.

5.1.1. Monte Carlo approach. As mentioned in Subsection 4.3, we assume two possible modifications of the crude MC approach. The first is to propose multiple samples and then to choose the one with the greatest residual of the reduced solution in the current RB. The second is an alternative sampling (avoiding sampling) of proposal samples using an altered pdf with lowered probability around already used samples. We can also mix these modifications together. Sampling from the altered pdf will be done using the MH method with 5000 steps and the proposal pdf equal to the original pdf of $\boldsymbol{Z}$. As the MC approach includes a significant amount of randomness, its convergence will differ in each replication (for the same problem). Therefore, we use 100 sample solutions and then evaluate their results. We will be interested in both the mean of the convergence and the variance of the convergence. When comparing two convergence curves, we are more interested in their ratio than their difference. Therefore, we will use log-transformed average denoted as log-average, which is equivalent to the geometric mean (in this section $N = 100$)

$$\overline{X}_g = \exp\left(\frac{1}{N}\sum_{i=1}^{N}\log(X_i)\right) = \left(\prod_{i=1}^{N} X_i\right)^{1/N}, \tag{5.1}$$

and log transformed std denoted as log-std which is equivalent to the geometric std

$$s_g = \exp\sqrt{\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{X_i}{\overline{X}_g}\right)^2}. \tag{5.2}$$

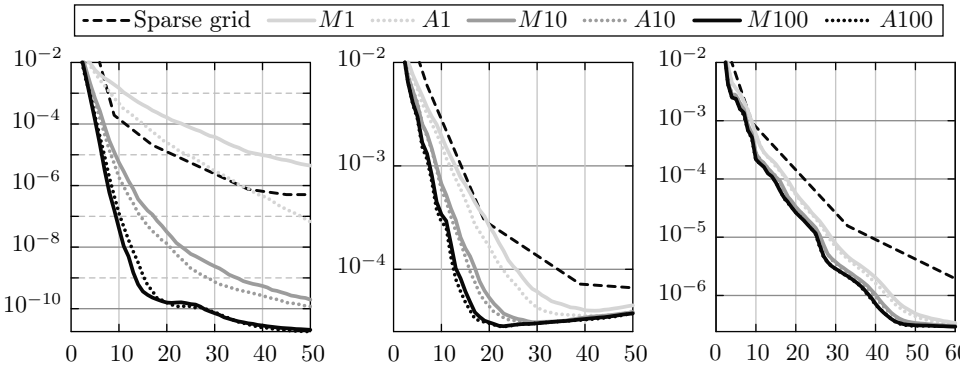
For a more convenient view of the results, we can use Chebyshev's inequality, e.g. interval $\langle \overline{X}_g / s_g^3, \overline{X}_g \cdot s_g^3 \rangle$ contains 81% values.

We perform tests on geometry $P2$, FE grid lvl 1, complete polynomials up to degree 4 and use three different settings ($M = 2$, $\sigma_i = 2$), ($M = 3$, $\sigma_i = 1$), ($M = 4$, $\sigma_i = 0.2$). The mean values of Z_i are chosen according to Table 1. We test combinations of 1, 10, 100 proposal samples and standard/avoiding sampling (non-avoiding variant with one proposal sample is the crude MC approach). Here, one RB iteration is understood as adding one vector to the RB, therefore one RB iteration corresponds to one solution of a deterministic FE system. The results of log-averages can be found in Figure 5.1a and the results of log-std can be found in Figure 5.1b. The results show that avoiding sampling as well as multiple proposal samples have a significant impact on the convergence speed. The impact is higher for the problems with a lower number of variables and high std σ_i of Z_i. The benefits of modifications are not only in the faster convergence but also in the more consistent results. From the results of log-std, we can see that multiple proposal samples and avoiding sampling lowers the variance of single algorithm runs. This is again more significant for a lower number of variables and high σ_i.

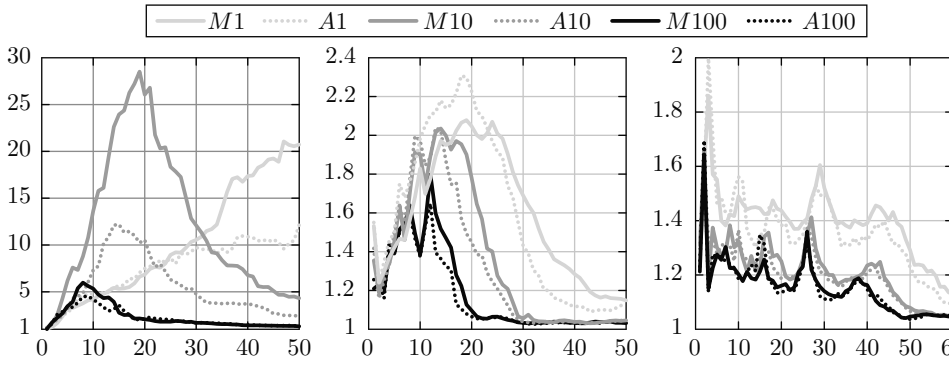
The results also show the sudden stagnation of convergence when the MC approach reaches a certain precision (relative residual). This is characteristic for all results obtained with the MC approach and will be further discussed in Subsection 5.4.2.

Figure 5.1a also presents the comparison with deterministically chosen points using the sparse grid points. The usage of sparse grid points leads to a worse convergence than the MC approach with modifications, but it brings no additional computational costs (e.g. in the computation of reduced solutions). It can be preferred in some cases.

5.1.2. Reduced rational Krylov subspace approach and combined approach. The first numerical experiment concerns the choice of parameter α of the RRKS approach (and also the combined approach). We perform tests using FE grid lvl 1, complete polynomials up to degree 4, $M = 5$ and $\sigma_i = 0.2$. We test α in the range $1/10$ to 10 and for comparison use all three types of geometries. The combined approach uses an alternate choice of approaches, even RB iterations use the RRKS approach and odd RB iterations use the MC approach using 10 proposal samples and avoiding sampling. One RB iteration of RRKS expands the RB with M vectors. Therefore, one RB iteration of the MC approach also expands the RB with M vectors (M inner steps of the MC approach) to balance the combined approach. The results of numerical experiments concerning the impact of α on the relative residual in the fixed RB iteration can be found in Figure 5.2. The optimal value of α differs slightly between geometry types. We can see that values $\alpha < 1$ yield bad results in



(a) log-average (see (5.1)) compared to the relative residual of sparse grid sample selection. Horizontal axis: RB iteration, vertical axis: mean of RB relative residual.



(b) log-std (see (5.2)). Horizontal axis: RB iteration, vertical axis: std of RB relative residual.

Figure 5.1. Results of the relative residual of 100 sample runs of MC approach with different settings (crude—M/Avoided—A; 1, 10, 100 proposal samples) on geometry type $P2$. Both figures from the left: ($M = 2$, $\sigma_i = 2$), ($M = 3$, $\sigma_i = 1$), ($M = 4$, $\sigma_i = 0.2$).

all tested settings and that the results get slightly worse with high values of α, but any choice of $\alpha > 1$ leads to satisfactory results. We will use $\alpha = 2$ in all further tests. We can link these results to the fact that the spectra of preconditioned matrices $L^{-1}K_m L^{-\top}$ are in $\langle 0, 1\rangle$. Therefore, very high values of α would override the information in matrices and very low values of α would lead to badly scaled matrices (high condition number).

Another possible parameter is the scheme for the choice of the new expansion vector in the RRKS approach (see Subsection 4.4.2). We test the weighted and non-weighted approach for all three geometry types and for $M = 3, 5, 7$, the results can be seen in Figure 5.3. The results show that the weighted approach performs slightly

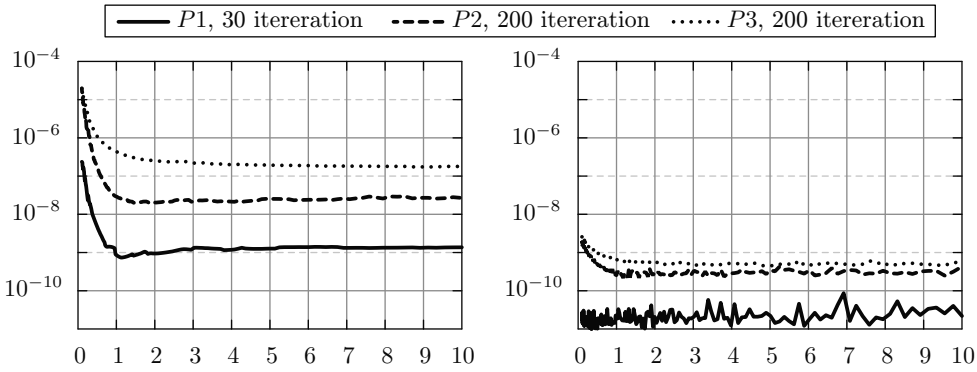


Figure 5.2. Dependence of relative residual in fixed RB iteration on the choice of RRKS parameter α (multiple geometry types ($P1, P2, P3$), $M = 5$, $\sigma_i = 0.2$). Left: RRKS approach, Right: combined approach. Horizontal axis: α, vertical axis: relative residual of RB.

better. The gain in convergence speed is not significant, but it is consistent. We will use the weighted approach in all further tests.

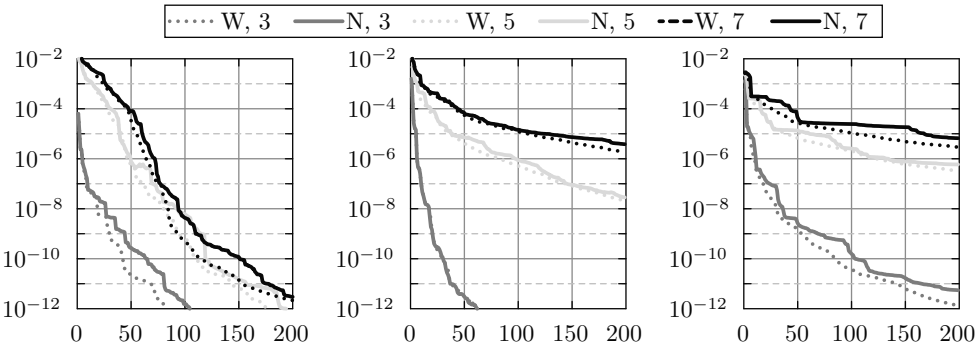


Figure 5.3. Convergence of weighted (W) and non-weighted (N) RRKS approach ($M \in \{3, 5, 7\}$, $\sigma = 0.2$). Geometry types from left: $P1$, $P2$, $P3$. Horizontal axis: RB iteration, vertical axis: RB relative residual.

The last numerical experiment concerns the choice of approach selection in the combined approach (see Subsection 4.5). We test the alternate option and the weighted option. The weighted option uses the step sizes computed during the RB iterations to decide which approach performs better in the current state of the RB computation. The weighted approach can be used with a different size of the buffer (how many of last iterations affect the decision). The results of testing can be found in Figure 5.4. We show results only for geometry types $P2$ and $P3$, there were no differences between approaches in geometry type $P1$ due to very fast convergence (under 10 RB iterations). The results show an improved convergence when using the

weighted approach. This is more significant in the regions where it is clear which approach performs better (start—MC, end—RRKS). The gain of the convergence speed can be up to 2×, when one approach stagnates. The size of the buffer decreases the speed of switching between approaches but brings more stability to the process. We will use the weighted approach with buffer size 4 in all further experiments.

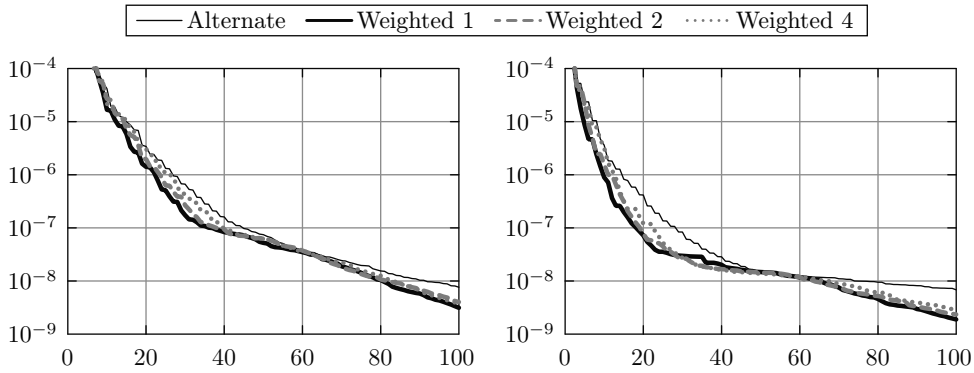


Figure 5.4. The convergence of MC/RRKS selection techniques (Alternate, Weighted with buffer size 1, 2, 4) for the combined approach ($M = 5$, $\sigma = 0.2$). Geometry types from left: $P2$, $P3$. Horizontal axis: RB iteration, vertical axis: RB relative residual.

5.2. Impact of problem geometry. Next experiments concern the impact of different geometries of the subdomains. First, we test the impact of adding more subdomains (which increases the number of random variables). We use the geometry type $P2$ because it allows the scaling of the number of subdomains while keeping the arrangement of the subdomain fixed (see Subsection 2.1). This should eliminate at least some of the behavior caused by different subdomain arrangement. The results can be found in Figure 5.5. Note that the x axis describes the number of solutions (M times RB iterations) and it is in the logarithmic scale. The results show that

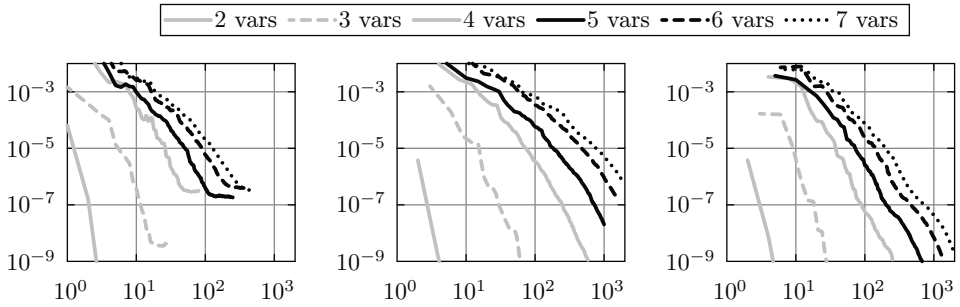


Figure 5.5. Convergence of different approaches ($P2$, $M = 2, 3, 4, 5, 6, 7$, $\sigma_i = 0.2$). Approaches from left: MC, RRKS, combined. Horizontal axis: no. of solutions, vertical axis: RB relative residual.

all the methods deal with an increasing number of variables in the same way. The increase in the number of variables leads to a slower convergence rate. We can also observe the stagnation of the MC approach, this will be further discussed in Subsection 5.4.2.

The second experiment tests the impact of the different geometry arrangement. For this, we use the geometry type $P2$ and generate different Gaussian random field samples on which the geometry is based. In this way, we get different geometries of similar properties. The results are in Figure 5.6. The log-average exhibits the same behavior as the previous tests: the MC approach converges fast and then stagnates, the RRKS approach keeps a steady but slower convergence rate, and the combined approach mixes the good properties of these two behaviors together. The log-std shows that there are fairly significant differences between the geometry samples. The MC and the combined approach have a higher log-std than the RRKS approach, this is probably due to its random nature. An interesting fact is that the MC approach has a high log-std after log-average converges. This is due to different precisions in which the stagnation of convergence occurs.

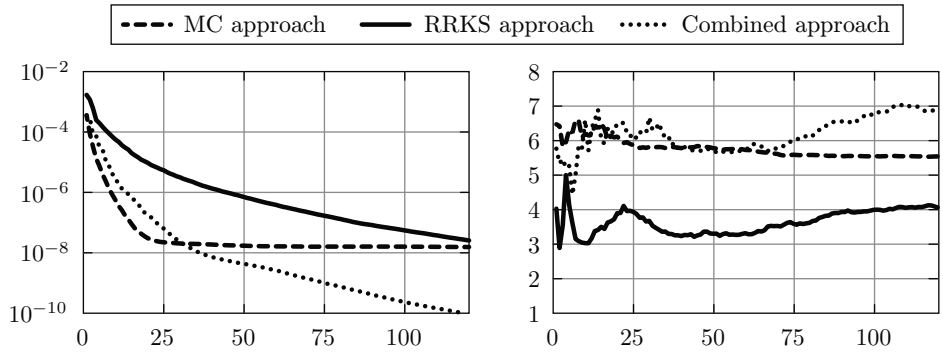
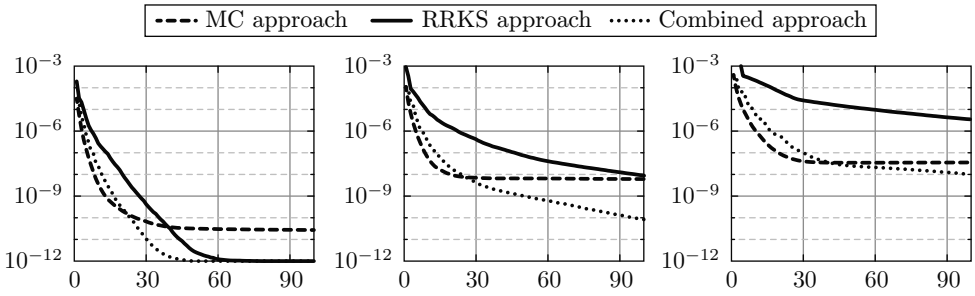


Figure 5.6. Convergence of 20 random samples of geometry type $P2$ ($M = 5$, $\sigma_i = 0.2$). Left: log-average (see (5.1)), Right: log-std (see (5.2)). Horizontal axis: RB iteration, vertical axis: log mean of RB realative residual.

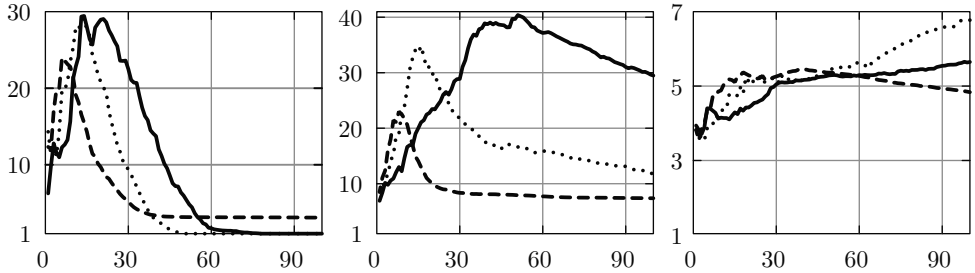
5.3. Impact of random variables mean and variance. Aside from the geometry of subdomains, a fairly large impact on the convergence can be expected from the mean values and the stds of Z_i.

5.3.1. Mean values of random variables. The impact of the mean value will be tested using the solutions of problems with randomly generated mean values $\mu_i \sim \mathcal{N}(0, 3)$. We performed 30 samples with all geometry types with $M = 5$, $\sigma_i = 0.2$. The results are in Figure 5.7. We can again see that the MC approach stagnates at some point (will be further discussed in Subsection 5.4.2). The stagnation of

convergence of the RRKS in $P1$ is caused by the stopping condition in 10^{-12}. Overall the convergence of the log-average is in agreement with the previous tests. The log-std values show that the difference in the convergence due to differing mean values is significant. Compared to the difference caused by geometry samples, it has about 4 to 5 times higher log-std. An exception to this is the geometry type $P3$, where the log-std is 6 times lower than in $P1$ and $P2$. This is caused probably by the fact that the geometry type $P3$ represents a problem where the solution does not have efficient low-rank approximation regardless of the mean values on subdomains.



(a) log-average (see (5.1)). Horizontal axis: RB iteration, vertical axis: RB relative residual.



(b) log-std (see (5.2)). Horizontal axis: RB iteration, vertical axis: log-std of RB relative residual.

Figure 5.7. The relative residual of problems with random mean value on subdomains $\mu_i \sim \mathcal{N}(0, 3)$ (30 samples, $M = 5$, $\sigma_i = 0.2$). On both sub-figures from left: $P1$, $P2$, $P3$.

5.3.2. Variance of random variables. The results of the tests for stds $\sigma_i = 0.1, 0.2, 1, 2$ are shown in Figure 5.8. We show the results on the geometry type $P2$, but the behavior of the convergence curves is the same for other geometry types too. We can observe the predicted behavior, higher values of std lead to a more complex problems and therefore slower convergence rate. The decrease in the convergence rate is similar for all types of the problems and the approaches. Again, we see the stagnation of the MC approach (see Subsection 5.4.2).

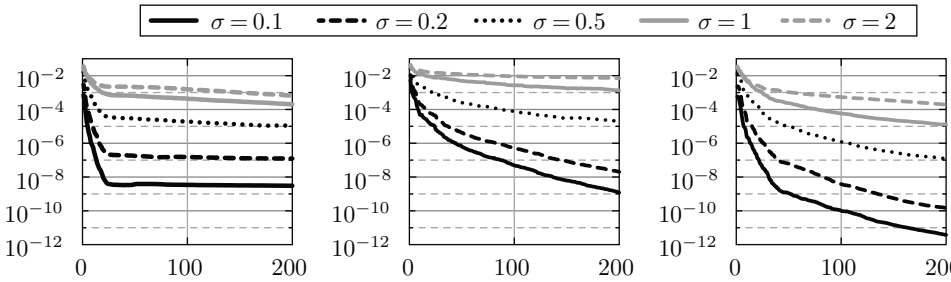


Figure 5.8. The comparison of convergence for different values of σ_i ($P2$, $M = 5$). Approaches from left: MC, RRKS, combined. Horizontal axis: RB iteration, vertical axis: RB relative residual.

5.4. Impact of discretization size. In this section, we focus on the examination of the impact of different discretizations on the resulting approximation error. In comparison to the other tests, here we use the RRKS approach with non-weighted expansion vector selection. We will use the geometry type $P2$ and 5 subdomains.

5.4.1. Finite element grid sizes. We test three different FE grid levels $(1, 2, 4)$. The results are in Figure 5.9. We can observe that the convergence does differ only by a small factor. The relative residual using finer grids is smaller. This is caused by the usage of the relative residual as the convergence criterion, because finer discretizations lead to a higher norm of the right-hand side vector. Doubling the discretization level will increase the norm of the right-hand side vector approximately by the factor of $\sqrt{2}$. This is in agreement with the measurements, where single convergence curves are scaled by the multiple of $\sqrt{2}$. Therefore, there is no effect of different FE discretizations on the convergence of the RB.

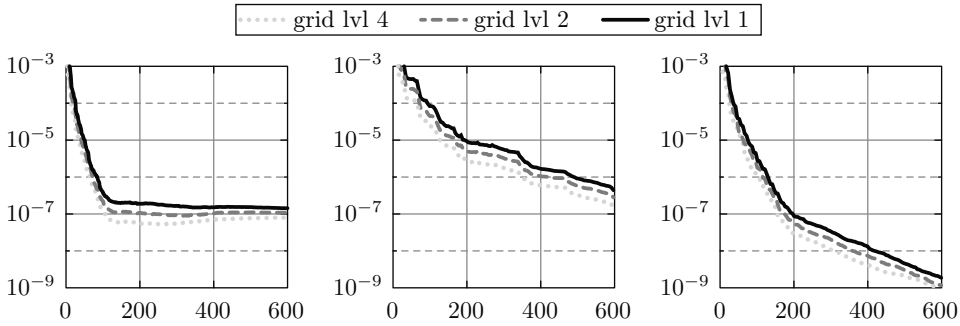


Figure 5.9. The convergence of RB with different finite element discretizations ($P2$, $M = 5$, $\sigma_i = 0.2$). Approaches from left: MC, RRKS, combined. Horizontal axis: no. of solutions, vertical axis: RB relative residual.

5.4.2. Polynomial chaos degrees. Here, we examine the behavior of the RB convergence when the discretization of the parameter space changes. We test the

discretizations consisting of complete polynomials with the maximal degree from 1 to 8. The results can be found in Figure 5.10. The tested RB construction approaches (MC and RRKS) build the RB independently of the discretization in the parametric space. Therefore, we expected that when we increase the degree of the polynomial the error of the current RB approximation will slightly increase. This happened in the RRKS approach and also in the MC approach before the stagnation occurred.

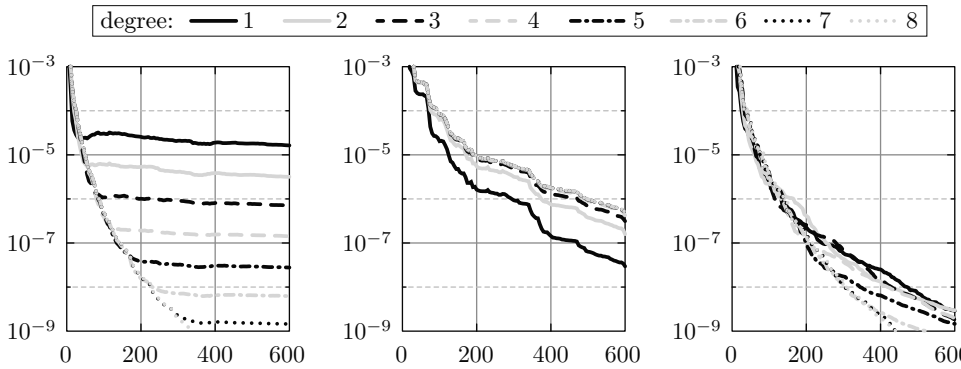


Figure 5.10. Convergence of RB with different polynomial chaos discretizations ($P2$, $M = 5$, $\sigma_i = 0.2$). Approaches from left: MC, RRKS, combined. Horizontal axis: number of solutions, vertical axis: RB relative residual.

The interesting part of our results is the behavior of the stagnation of the MC approach. It is clear that the stagnation occurs later when we use a finer discretization in the parametric space. This is probably caused by the approximation properties of the basis constructed by the MC approach. The MC approach constructs the RB from the samples of the solution $u(\cdot; \boldsymbol{Z})$. These samples can dramatically differ from the equivalent samples of the discretized SGM solution when the degree of polynomials is not sufficient to describe the solution properly. This discrepancy probably leads to a stagnation of convergence of the MC approach. Therefore, the stagnation of the MC approach is not a bad property of the approach, but a beneficial one indicating that the approximation of the parametric space is insufficient.

5.5. Acceleration of inner FE systems computation using the DCG. We compare the difference between a number of iterations when using the DCG with the current RB as the deflation basis and the PCG. We perform tests with three preconditioners:

- ▷ additive Schwarz preconditioner, set with 30 subdomains = equal column slices of our square domain of the size 1/20,
- ▷ incomplete Cholesky preconditioner with no filling allowed,
- ▷ diagonal preconditioner (Jacobi).

Graphical comparison of the iterations count can be seen in Figure 5.11. We can see that the usage of the RB as the deflation basis significantly decreases the number of iterations. The number of iterations when using PCG is almost constant during the RB iterations, but the number of DCG iterations gets gradually lower. Note that the descend of the number of DCG iterations is not granted. It is possible that the RB method encounters a proposal vector containing new information outside the deflation basis. In that case, the number of iterations will be higher than in the previous steps of RB, but never higher than in PCG. Cumulative results are shown in Table 2. We can see that the usage of the DCG can save up to 80 % of the CG iterations during the RB computation.

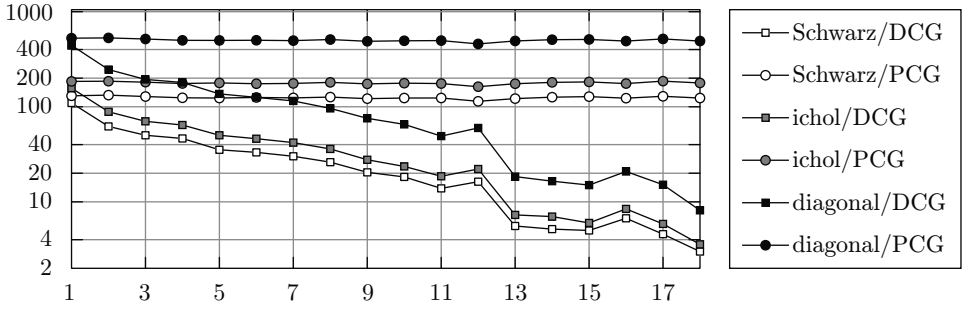


Figure 5.11. The dependence of the average number of the DCG/PCG iterations per inner system on RB iterations (until relative residual 10^{-9}), (MC approach, $P1$, $M = 5$, $\sigma_i = 0.2$). Horizontal axis: RB iteration, vertical axis: average DCG/PCG iteration.

	Ad. Schwarz p.	diagonal p.	ichol (nofill) p.
RRKS savings in %	72.32%	73.47%	73.33%
MC savings in %	82.58%	83.48%	83.06%

Table 2. Computational savings using DCG with the RB as a deflation basis.

5.6. Stopping criterion. In Subsection 4.2, we discussed different stopping criteria. Here we test the behavior of the relative residual and the step size in comparison to the relative error. The relative error is computed from the solution obtained using the PCG and the mean value preconditioner [44], note that this is fairly expensive and we can perform such computation only for small FE grids/polynomial bases. We perform the experiment using the combined approach and three different settings of the problem: ($P2$, $M = 3$, $\sigma_i = 1$, degree $= 4$), ($P2$, $M = 5$, $\sigma_i = 0.2$, degree $= 7$) and ($P3$, $M = 5$, $\sigma_i = 0.2$, degree $= 8$). The results can be found in Figure 5.12. We can see that both the relative residual and the step size have convergence very similar to the relative error. It exhibits a little noisy behavior, which can be corrected by

averaging with a buffer of last step sizes. Overall the step size seems to be a viable stopping criterion.

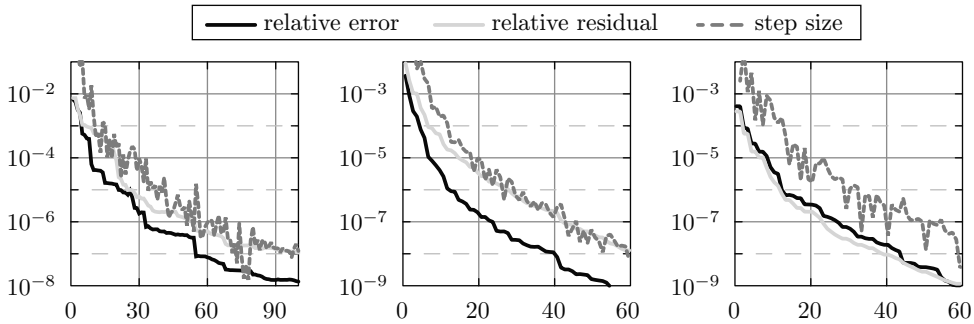


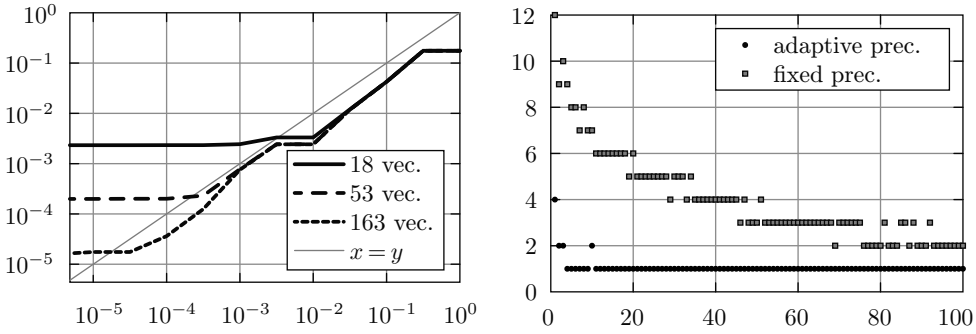
Figure 5.12. Comparison of different stopping criteria using the combined approach. From left: ($P2$, $M = 3$, $\sigma_i = 1$, degree $= 4$), ($P2$, $M = 5$, $\sigma_i = 0.2$, degree $= 7$), ($P3$, $M = 5$, $\sigma_i = 0.2$, degree $= 8$). Horizontal axis: RB iteration, vertical axis: stopping crit.

5.7. Solutions of the reduced system. In Subsection 4.1, we proposed the use of the adaptive precision when solving the reduced solution. Here we test the dependence between the residual of the reduced solution (CG precision) and the residual of the RB, which leads to an estimation of c_k from (4.6). The test is performed using the RRKS approach, $P2$, $M = 2$, $\sigma_i = 0.2$. Results can be seen on the left side of Figure 5.13. We can see that c_k is very close to 1 in all tested RB sizes. This is a very good behavior, because our method estimates c_k by lowering CG precision between RB iterations and compares it with the descend of the RB relative residual (or step size). Therefore, it will find a good approximation very fast. Next, we compare the benefits (in terms of the iterations saved) of adaptive precision. Results can be found on the right side of Figure 5.13. These results are very good, we need to perform only a few iterations in the first few RB iterations and then it is enough to perform only one iteration per RB iteration. This lowers the computational cost significantly without affecting the resulting accuracy.

6. Conclusions

The main contribution of this paper can be divided into two parts: modifications of the existing approaches to the construction of the RB and the examination of the behavior of these approaches for various problem settings.

The first of the examined approaches was the MC (Monte Carlo) approach. We proposed an alternative sampling procedure, which could be combined with the multiple sample proposal, see Subsection 4.3. Its benefits were more significant for



(a) Horizontal axis: CG precision, vertical axis: relative residual of RB.

(b) Horizontal axis: RB iteration, vertical axis: number of CG iteration.

Figure 5.13. Left: CG precision and the corresponding relative residual of selected reduced bases. Right: number of iterations of CG in each RB iteration.

problems with a lower number of variables, see Subsection 5.1.1. We observed that the MC approach always stagnated at some point. The precision where the MC approach stagnated was lower for more complex problems (high standard deviation, more subdomains).

Further, we considered the RRKS (reduced rational Krylov subspace) approach. In Subsection 4.4 we presented an alternative construction without the need for computing the Cholesky factorization of the FE matrix. We also improved the expansion vector selection scheme (see Subsection 4.4.2) and the results were presented in Subsection 5.1.2. The RRKS approach converged more slowly than the MC approach, but did not encounter stagnation.

We also combined the above approaches into a hybrid method adaptively switching between MC and RRKS approaches, see Subsections 4.5 and 5.1.2. The combined approach was able to combine the positive properties of both approaches. It converged fast at the start of RB computation due to switching to MC approach and preserved at least a lower convergence rate later due to switching to RRKS approach.

All of the mentioned approaches are built around solutions of FE deterministic counterparts of the problem. This consumes most of the computational time of the RB method. Therefore, we proposed to use the DCG (deflated conjugate gradients) with the RB as a basis of the deflation space for the solution of these systems (Subsection 4.6). This does not add any significant overhead to the CG computation and decreases the number of iterations dramatically (Subsection 5.5).

Apart from the construction of the RB, the RB method includes the solution of the reduced system. This is done many times during the RB iterations. In Subsection 4.1, we proposed the usage of an adaptive precision for these solutions. This significantly

reduced the number of iterations needed in each RB step, see Subsection 5.7. We also tested different stopping criteria, see Subsections 4.2 and 5.6.

For numerical experiments, we chose a set of problems which should reflect realistic problems i.e. subdomains with a log-normal distribution. We tested different geometries of subdomains as well as mean values and stds of controlling variables. The increasing std and the number of subdomains/variables affect the convergence of the RB method most significantly (Subsections 5.2 and 5.3.2). Less significant, but still notable, were differences in convergence caused by different geometries of subdomains and the contrast between the mean values on the subdomains (Subsections 5.2 and 5.3.1).

We also tested the impact of the used discretization. The convergence results for different FE grid sizes were almost identical and from our observations, the FE grid size does not affect the convergence (Subsection 5.4.1). Specific results were obtained from testing different discretizations of the parametric space (complete polynomials up to a given degree), see Subsection 5.4.2. The RRKS approach behaved as expected, the convergence of the problem with a higher polynomial degree had a slightly lower rate of convergence. The MC approach behaved in the same way before it reached stagnation. The interesting observation was that the precision, where the MC approach stagnates, increases with an increasing polynomial degree. This means that the stagnation observed in the MC approach is caused by the discrepancy between MC samples of solution (point values of the analytic solution) and the appropriate point values of the SG approximation of the solution. Therefore, stagnation indicates a need for the increase of the polynomial degree in the discretization of the parametric space.

Overall, the MC approach performed best in all of our tests in terms of the speed of convergence. It experienced stagnation but, as was mentioned above, this just meant that the degree of polynomial approximation was insufficient. In the rare case when the system is given with its polynomial discretization and we need to solve it with high precision, we can effectively use the combined approach. But this is not a realistic case, because the precision of the RB approximation of the SG solution is then unnecessarily higher than the precision of the SG approximation itself. We should also take into consideration that the MC iterations are more expensive than the RRKS iterations. This is due to multiple proposal samples and a need to assemble the FE matrix for each system. The matrix of each system can be assembled from subdomain matrices, but this is still not cheap as we need to initialize preconditioners. The ratio of the computational cost of MC and the RRKS iterations depends on many factors and mostly on the implementation. It would be a topic of a further research.

References

[1] *I. Babuška, F. Nobile, R. Tempone*: A stochastic collocation method for elliptic partial differential equations with random input data. SIAM J. Numer. Anal. *45* (2007), 1005–1034. zbl MR doi

[2] *I. Babuška, R. Tempone, G. E. Zouraris*: Galerkin finite element approximations of stochastic elliptic partial differential equations. SIAM J. Numer. Anal. *42* (2004), 800–825. zbl MR doi

[3] *J. Ballani, L. Grasedyck*: A projection method to solve linear systems in tensor format. Numer. Linear Algebra Appl. *20* (2013), 27–43. zbl MR doi

[4] *A. Barth, C. Schwab, N. Zollinger*: Multi-level Monte Carlo finite element method for elliptic PDEs with stochastic coefficients. Numer. Math. *119* (2011), 123–161. zbl MR doi

[5] *P. Benner, A. Onwunta, M. Stoll*: Low-rank solution of unsteady diffusion equations with stochastic coefficients. SIAM/ASA J. Uncertain. Quantif. *3* (2015), 622–649. zbl MR doi

[6] *M. Béreš*: An efficient reduced basis construction for stochastic Galerkin matrix equations using deflated conjugate gradients. AETA 2018-Recent Advances in Electrical Engineering and Related Sciences: Theory and Application (I. Zelinka et al., eds.). Lecture Notes in Electrical Engineering 554, Springer, Cham, 2019, pp. 175–184. doi

[7] *M. Béreš, S. Domesová*: The stochastic Galerkin method for Darcy flow problem with log-normal random field coefficients. Advances in Electrical and Electronic Engineering *15* (2017), 13 pages. doi

[8] *A. Bespalov, C. E. Powell, D. Silvester*: Energy norm a posteriori error estimation for parametric operator equations. SIAM J. Sci. Comput. *36* (2014), A339–A363. zbl MR doi

[9] *A. Bespalov, D. Silvester*: Efficient adaptive stochastic Galerkin methods for parametric operator equations. SIAM J. Sci. Comput. *38* (2016), A2118–A2140. zbl MR doi

[10] *R. E. Caflisch*: Monte Carlo and quasi-Monte Carlo methods. Acta Numerica *7* (1998), 1–49. zbl MR doi

[11] *C. S. Chen, Y. C. Hon, R. A. Schaback*: Scientific Computing with Radial Basis Functions. Department of Mathematics, University of Southern Mississippi, Hattiesburg, 2005.

[12] *Y. Chen, J. Jiang, A. Narayan*: A robust error estimator and a residual-free error indicator for reduced basis methods. Comput. Math. Appl. *77* (2019), 1963–1979. MR doi

[13] *P. Chen, A. Quarteroni, G. Rozza*: Comparison between reduced basis and stochastic collocation methods for elliptic problems. J. Sci. Comput. *59* (2014), 187–216. zbl MR doi

[14] *P. Chen, A. Quarteroni, G. Rozza*: Reduced basis methods for uncertainty quantification. SIAM/ASA J. Uncertain. Quantif. *5* (2017), 813–869. zbl MR doi

[15] *G. Christakos*: Random Field Models in Earth Sciences. Academic Press, San Diego, 1992. doi

[16] *K. A. Cliffe, M. B. Giles, R. Scheichl, A. L. Teckentrup*: Multilevel Monte Carlo methods and applications to elliptic PDEs with random coefficients. Comput. Vis. Sci. *14* (2011), 3–15. zbl MR doi

[17] *A. J. Crowder, C. E. Powell*: CBS constants & their role in error estimation for stochastic Galerkin finite element methods. J. Sci. Comput. *77* (2018), 1030–1054. zbl MR doi

[18] *S. Dolgov, B. N. Khoromskij, A. Litvinenko, H. G. Matthies*: Polynomial chaos expansion of random coefficients and the solution of stochastic partial differential equations in the tensor train format. SIAM/ASA J. Uncertain. Quantif. *3* (2015), 1109–1135. zbl MR doi

[19] *H. C. Elman, T. Su*: A low-rank multigrid method for the stochastic steady-state diffusion problem. SIAM J. Matrix Anal. Appl. *39* (2018), 492–509. zbl MR doi

[20] *R. A. Freeze*: A stochastic-conceptual analysis of one-dimensional groundwater flow in nonuniform homogeneous media. Water Resources Research *11* (1975), 725–741. doi

[21] *C. J. Gittelson*: Stochastic Galerkin discretization of the log-normal isotropic diffusion problem. Math. Models Methods Appl. Sci. *20* (2010), 237–263. zbl MR doi
[22] *S. Güttel*: Rational Krylov approximation of matrix functions: Numerical methods and optimal pole selection. GAMM-Mitt. *36* (2013), 8–31. zbl MR doi
[23] *F. Heiss, V. Winschel*: Likelihood approximation by numerical integration on sparse grids. J. Econom. *144* (2008), 62–80. zbl MR doi
[24] *R. J. Hoeksema, P. K. Kitanidis*: Analysis of the spatial structure of properties of selected aquifers. Water Resources Research *21* (1985), 563–572. doi
[25] *G. Karypis, V. Kumar*: A fast and high quality multilevel scheme for partitioning irregular graphs. SIAM J. Sci. Comput. *20* (1998), 359–392. zbl MR doi
[26] *A. Keese, H. G. Mathhies*: Adaptivity and sensitivity for stochastic problems. Computational Stochastic Mechanics 4 (P. D. Spanos et al., eds.). Millpress, Rotterdam, 2003, pp. 311–316.
[27] *B. N. Khoromskij, C. Schwab*: Tensor-structured Galerkin approximation of parametric and stochastic elliptic PDEs. SIAM J. Sci. Comput. *33* (2011), 364–385. zbl MR doi
[28] *D. P. Kroese, T. Taimre, Z. I. Botev*: Handbook for Monte Carlo Methods. Wiley Series in Probability and Statistics, Wiley, Hoboken, 2011. zbl doi
[29] *K. Lee, H. C. Elman*: A preconditioned low-rank projection method with a rank-reduction scheme for stochastic partial differential equations. SIAM J. Sci. Comput. *39* (2017), S828–S850. zbl MR doi
[30] *G. J. Lord, C. E. Powell, T. Shardlow*: An Introduction to Computational Stochastic PDEs. Cambridge Texts in Applied Mathematics, Cambridge University Press, Cambridge, 2014. zbl MR doi
[31] *L. Mathelin, O. Le Maître*: Dual-based a posteriori error estimate for stochastic finite element methods. Commun. Appl. Math. Comput. Sci. *2* (2007), 83–115. zbl MR doi
[32] *H. G. Matthies, E. Zander*: Solving stochastic systems with low-rank tensor compression. Linear Algebra Appl. *436* (2012), 3819–3838. zbl MR doi
[33] *P. H. Nelson*: Permeability-porosity relationships in sedimentary rocks. Log Analyst *35* (1994), 38–62.
[34] *C. J. Newsum, C. E. Powell*: Efficient reduced basis methods for saddle point problems with applications in groundwater flow. SIAM/ASA J. Uncertain. Quantif. *5* (2017), 1248–1278. zbl MR doi
[35] *F. Nobile, R. Tempone, C. G. Webster*: A sparse grid stochastic collocation method for partial differential equations with random input data. SIAM J. Numer. Anal. *46* (2008), 2309–2345. zbl MR doi
[36] *A. Nouy*: A generalized spectral decomposition technique to solve a class of linear stochastic partial differential equations. Comput. Methods Appl. Mech. Eng. *196* (2007), 4521–4537. zbl MR doi
[37] *A. Nouy*: Construction of generalized spectral bases for the approximate resolution of stochastic problems. Mecanique et Industries *8* (2007), 283–288. doi
[38] *A. Nouy*: Generalized spectral decomposition method for solving stochastic finite element equations: Invariant subspace problem and dedicated algorithms. Comput. Methods Appl. Mech. Eng. *197* (2008), 4718–4736. zbl MR doi
[39] *A. Nouy*: Recent developments in spectral stochastic methods for the numerical solution of stochastic partial differential equations. Arch. Comput. Methods Eng. *16* (2009), 251–285. zbl MR doi
[40] *A. Nouy*: Proper generalized decompositions and separated representations for the numerical solution of high dimensional stochastic problems. Arch. Comput. Methods Eng. *17* (2010), 403–434. zbl MR doi

[41] *A. Nouy, O. P. Le Maitre*: Generalized spectral decomposition for stochastic nonlinear problems. J. Comput. Phys. *228* (2009), 202–235. zbl MR doi
[42] *K. Petras*: Smolyak cubature of given polynomial degree with few nodes for increasing dimension. Numer. Math. *93* (2003), 729–753. zbl MR doi
[43] *M. J. D. Powell*: Radial basis function methods for interpolation to functions of many variables. HERCMA 2001. Proceedings of the 5th Hellenic-European Conference on Computer Mathematics and Its Applications (E. A. Lipitakis, ed.). LEA, Athens, 2002, pp. 2–24. zbl
[44] *C. E. Powell, H. C. Elman*: Block-diagonal preconditioning for spectral stochastic finite-element systems. IMA J. Numer. Anal. *29* (2009), 350–375. zbl MR doi
[45] *C. E. Powell, D. Silvester, V. Simoncini*: An efficient reduced basis solver for stochastic Galerkin matrix equations. SIAM J. Sci. Comput. *39* (2017), A141–A163. zbl MR doi
[46] *I. Pultarová*: Adaptive algorithm for stochastic Galerkin method. Appl. Math., Praha *60* (2015), 551–571. zbl MR doi
[47] *I. Pultarová*: Hierarchical preconditioning for the stochastic Galerkin method: Upper bounds to the strengthened CBS constants. Comput. Math. Appl. *71* (2016), 949–964. MR doi
[48] *C. P. Robert, G. Casella*: Monte Carlo Statistical Methods. Springer Texts in Statistics, Springer, New York, 2004. zbl MR doi
[49] *Y. Saad, M. Yeung, J. Erhel, F. Guyomarc'h*: A deflated version of the conjugate gradient algorithm. SIAM J. Sci. Comput. *21* (2000), 1909–1926. zbl MR doi
[50] *T. J. Santner, B. J. Williams, W. I. Notz*: The Design and Analysis of Computer Experiments. Springer Series in Statistics, Springer, New York, 2003. zbl MR doi
[51] *V. Simoncini*: Analysis of the rational Krylov subspace projection method for large-scale algebraic Riccati equations. SIAM J. Matrix Anal. Appl. *37* (2016), 1655–1674. zbl MR doi
[52] *B. Sousedík, R. G. Ghanem, E. T. Phipps*: Hierarchical Schur complement preconditioner for the stochastic Galerkin finite element methods. Numer. Linear Algebra Appl. *21* (2014), 136–151. zbl MR doi
[53] *E. Ullmann*: A Kronecker product preconditioner for stochastic Galerkin finite element discretizations. SIAM J. Sci. Comput. *32* (2010), 923–946. zbl MR doi
[54] *S. Ullmann, J. Lang*: Stochastic Galerkin reduced basis methods for parametrized linear elliptic PDEs. Available at `https://arxiv.org/abs/1812.08519` (2018), 20 pages.
[55] *X. Wan, G. E. Karniadakis*: An adaptive multi-element generalized polynomial chaos method for stochastic differential equations. J. Comput. Phys. *209* (2005), 617–642. zbl MR doi
[56] *D. Xiu*: Numerical Methods for Stochastic Computations. A Spectral Method Approach. Princeton University Press, Princeton, 2010. zbl MR doi

Author's address: *Michal Béreš*, Institute of Geonics of the CAS, Studentská 1768, 708 00 Ostrava-Poruba, Czech Republic; Department of Applied Mathematics, Faculty of Electrical Engineering and Computer Science, VŠB-Technical University of Ostrava, 17. listopadu 2172/15, 708 00 Ostrava-Poruba, Czech Republic, e-mail: `michal.beres@vsb.cz`.