SIAM J. Matrix Anal. Appl.
Vol. 37, No. 2, pp. 491–518

BLOCK-DIAGONAL PRECONDITIONING FOR OPTIMAL CONTROL PROBLEMS CONSTRAINED BY PDEs WITH UNCERTAIN INPUTS[*]

PETER BENNER[†], AKWUM ONWUNTA[‡], AND MARTIN STOLL[†]

Abstract. The goal of this paper is the efficient numerical simulation of optimization problems governed by either steady-state or unsteady partial differential equations involving random coefficients. This class of problems often leads to prohibitively high dimensional saddle-point systems with tensor product structure, especially when discretized with the stochastic Galerkin finite element method. Here, we derive and analyze robust Schur complement–based block-diagonal preconditioners for solving the resulting stochastic optimality systems with all-at-once low-rank iterative solvers. Moreover, we illustrate the effectiveness of our solvers with numerical experiments.

Key words. stochastic Galerkin system, iterative methods, PDE-constrained optimization, saddle-point system, low-rank solution, preconditioning, Schur complement

AMS subject classifications. 35R60, 60H15, 60H35, 65N22, 65F10, 65F50

DOI. 10.1137/15M1018502

1. Introduction.

Optimization problems constrained by deterministic steady-state partial differential equations (PDEs) are computationally challenging. This is even more so if the constraints are deterministic unsteady PDEs since one would then need to solve a system of PDEs coupled globally in time and space, and time-stepping methods quickly reach their limitations due to the enormous demand for storage [25]. Yet, more challenging than the aforementioned are problems constrained by unsteady PDEs involving (countably many) parametric or uncertain inputs. This class of problems often leads to prohibitively high dimensional linear systems with Kronecker product structure, especially when discretized with the stochastic Galerkin finite element method (SGFEM). Moreover, a typical model for an optimal control problem with stochastic inputs (SOCP) will usually be used for the quantification of the statistics of the system response—a task that could in turn result in additional enormous computational expense.

Stochastic finite element–based solvers for a large range of PDEs with random data have been studied extensively [1, 3, 12, 21, 24]. However, optimization problems constrained by PDEs with random inputs have, in our opinion, not yet received adequate attention. Hence, this study is aimed at pushing the research frontier with respect to the numerical simulation of the latter class of stochastic problems (that is, SOCPs) toward larger and more challenging problems. Some of the papers on SOCPs include [12, 13, 24]. While [12] studies the existence and the uniqueness of solutions to control problems constrained by elliptic PDEs with random inputs, the emphasis in

[*] Received by the editors April 27, 2015; accepted for publication (in revised form) by D. B. Szyld February 3, 2016; published electronically April 7, 2016.
http://www.siam.org/journals/simax/37-2/M101850.html

[†] Computational Methods in Systems and Control Theory, Max Planck Institute for Dynamics of Complex Technical Systems, 39106 Magdeburg, Germany (benner@mpi-magdeburg.mpg.de, stollm@mpi-magdeburg.mpg.de).

[‡] Corresponding author. Computational Methods in Systems and Control Theory, Max Planck Institute for Dynamics of Complex Technical Systems, 39106 Magdeburg, Germany (onwunta@mpi-magdeburg.mpg.de). The research of this author was supported by a research grant from the International Max Planck Research School (IMPRS) for Advanced Methods in Process and System Engineering (Magdeburg).

[13] is on solvers based on stochastic collocation methods (SCMs) for optimal control problems with random coefficients. Rosseel and Wells in [24] apply a one-shot method with both SGFEM and SCM approaches to an optimal control problem constrained by stochastic elliptic PDEs. One of their findings is that SGFEM generally exhibits superior performance compared to the SCM, in the sense that, unlike SGFEM, the nonintrusivity property of the SCM is lost when moments of the state variable appear in the cost functional, or when the control function is a deterministic function.

The fast convergence and other nice properties exhibited by SGFEM notwithstanding, the resulting large tensor-product algebraic systems associated with this intrusive approach unfortunately limit its attractiveness. Thus, for it to compete favorably with the sampling-based approaches, there is the need to develop efficient solvers for the resulting large linear systems. This is indeed the motivation for this work. More precisely, we apply an all-at-once approach, together with SGFEM, to two prototypical models, namely, optimization problems constrained by (a) stationary diffusion equations and (b) unsteady diffusion equations, and in each of the two cases, both the constraint equations and the objective functional have uncertain inputs. As these problems pose increased computational complexity due to enormous memory requirements, we here focus specifically on efficient low-rank preconditioned iterative solvers for the resulting linear systems representing the Karush–Kuhn–Tucker (KKT) conditions. In particular, inspired by a state-of-the-art preconditioning strategy employed in the deterministic framework [20, 25], we derive and analyze robust Schur complement–based block-diagonal preconditioners which we use in conjunction with low-rank solvers for the efficient solution of the optimality systems.

For the numerical simulation of the SOCPs considered in this work, we assume that the state, the control, and the target (or the desired state) are analytic functions depending on the uncertain parameters. However, we note here that, as pointed out in [24], problems in which the control is modeled as an unknown stochastic function constitute inverse problems, and they are different from those with deterministic controls. In the former, the stochastic properties of the control are unknown but will be computed. So, in most cases (as we assume in this work), the mean of the computed stochastic control could be considered as optimal. Depending on the application, the mean may not, in general, be the sought optimal control, though. In addition, computing the uncertainty in the system response might require additional computational challenges.

This paper is structured as follows. In section 2, we present our problem statement and give an overview of the SGFEM on which we shall rely throughout. Section 3 discusses efficient solution of our first model problem, namely, an optimization problem governed by a steady-state diffusion equation with uncertain inputs. As an extension of the concepts discussed in section 3, we proceed to section 4 to introduce and analyze our preconditioning strategy for the unsteady analogue of the steady-state model. Furthermore, we here briefly review the tensor-train (TT) toolbox, a software package which we shall use, in conjunction with MINRES, to solve our unsteady problems. Finally, section 5 presents some numerical experiments to demonstrate the performance of our solvers.

2. Problem statement.

In this paper, we study the numerical simulation of optimal control problems constrained by PDEs with uncertain coefficients. More precisely, we formulate our model problems as

$$\min_{y,u} \mathcal{J}(y,u) \ \text{ subject to } \ c(y,u)=0, \tag{1}$$

where the constraint equation $c(y,u)=0$ represents a PDE with an uncertain coefficient to be specified in what follows, and

$$\mathcal{J}(y,u) := \frac{1}{2}||y-\bar{y}||^2_{L^2(\mathcal{D})\otimes L^2(\Omega)} + \frac{\alpha}{2}||\mathrm{std}(y)||^2_{L^2(\mathcal{D})} + \frac{\beta}{2}||u||^2_{L^2(\mathcal{D})\otimes L^2(\Omega)} \tag{2}$$

is a cost functional of tracking type. The functions y, u, and $\bar{y}$ are, in general, real-valued random fields representing, respectively, the state, the control, and the prescribed target system response. We note here that $\bar{y}$ and u could also be modeled deterministically. The positive constant β in (2) represents the parameter for the penalization of the action of the control u, whereas α penalizes the standard deviation $\mathrm{std}(y)$ of the state y. The objective functional $\mathcal{J}(y,u)$ is a deterministic quantity with uncertain terms. In what follows, we shall focus on distributed control problems, although we believe that our discussion generalizes to boundary control problems.

Next, we recall that by a random field $z : \mathcal{D}\times\Omega \to \mathbb{R}$, we mean that $z(\mathbf{x},\cdot)$ is a random variable defined on the complete probability space $(\Omega,\mathcal{F},\mathbb{P})$ for each $\mathbf{x}\in\mathcal{D}$. Here, Ω is the set of outcomes, $\mathcal{F}\subset 2^\Omega$ is the σ-algebra of events, and $\mathbb{P}:\mathcal{F}\to[0,1]$ is an appropriate probability measure. Here, we assume that z is in the tensor-product Hilbert space $L^2(\mathcal{D})\otimes L^2(\Omega)$ which is endowed with the norm

$$||v||_{L^2(\mathcal{D})\otimes L^2(\Omega)} := \left(\int_\Omega ||v(\cdot,\omega)||^2_{L^2(\mathcal{D})}\, d\mathbb{P}(\omega)\right)^{\frac{1}{2}} < \infty,$$

where $L^2(\Omega) := L^2(\Omega,\mathcal{F},\mathbb{P})$. For any random variable g defined on $(\Omega,\mathcal{F},\mathbb{P})$, the standard deviation $\mathrm{std}(g)$ and the mean $\mathbb{E}(g)$ of g are given, respectively, by

$$\mathrm{std}(g) = \left[\int_\Omega (g-\mathbb{E}(g))^2\, d\mathbb{P}(\omega)\right]^{\frac{1}{2}} \text{ and } \mathbb{E}(g) = \int_\Omega g\, d\mathbb{P}(\omega) < \infty. \tag{3}$$

2.1. Representation of random inputs.

Suppose we are given a random field $z : \mathcal{D}\times\Omega\to\mathbb{R}$ with known continuous covariance function $C_z(\mathbf{x},\mathbf{y})$. Then one way to represent z with a finite number of random variables is through a truncated Karhunen–Lòeve expansion (KLE):

$$z_N(\mathbf{x},\omega) = \mathbb{E}[z](\mathbf{x}) + \sigma_z \sum_{i=1}^N \sqrt{\lambda_i}\varphi_i(\mathbf{x})\xi_i(\omega), \tag{4}$$

where σ_z is the standard deviation of z, the random variables $\{\xi_i\}_{i=1}^N$ are centered, normalized, and mutually uncorrelated with [1],

$$\xi_i(\omega) = \frac{1}{\sigma_z\sqrt{\lambda_i}} \int_\mathcal{D} (z(\mathbf{x},\omega) - \mathbb{E}[z](\mathbf{x}))\varphi_i(\mathbf{x})\, d\mathbf{x} \ \ \forall\lambda_i > 0,$$

and $\{\lambda_i,\varphi_i\}$ is the set of eigenvalues and eigenfunctions corresponding to $C_z(\mathbf{x},\mathbf{y})$; that is,

$$\int_\mathcal{D} C_z(\mathbf{x},\mathbf{y})\varphi_i(\mathbf{y})\, d\mathbf{y} = \lambda_i\varphi_i(\mathbf{x}).$$

The eigenfunctions $\{\varphi_i\}$ form a complete orthogonal basis in $L^2(\mathcal{D})$. The eigenvalues $\{\lambda_i\}$ form a sequence of nonnegative real numbers decreasing to zero and

$$\sum_{i=1}^\infty \lambda_i = \int_\mathcal{D} \mathbb{V}\mathrm{ar}[z](\mathbf{x})\, d\mathbf{x}.$$

Moreover, by Mercer's theorem [23, p. 245], we have

$$\sup_{\mathbf{x}\in\mathcal{D}} \mathbb{E}\left[(z-z_N)^2\right] = \sup_{\mathbf{x}\in\mathcal{D}} \sum_{i>N}^{\infty} \lambda_i \varphi_i^2(\mathbf{x}) \to 0 \text{ as } N\to\infty.$$

In what follows, we will employ the so-called *finite noise assumption,* which states that a random field $z(\mathbf{x},\omega)$ can be approximated with a prescribed finite number of random variables $\xi := \{\xi_1, \xi_2, \ldots, \xi_N\}$, where $N\in\mathbb{N}$ and $\xi_i(\omega) : \Omega \to \Gamma_i \subseteq \mathbb{R}$; this is, for instance, the case when we use a joint N-term KLE to approximate the random coefficient

$$a \equiv a(\mathbf{x},\omega) = a(\mathbf{x},\xi(\omega)) = a(\mathbf{x},\xi_1(\omega),\xi_2(\omega),\ldots,\xi_N(\omega)) \tag{5}$$

in the stochastic PDE $c(y,u)=0$. We also make the simplifying assumption that each random variable is independent and characterized by a probability density function $\rho_i : \Gamma_i \to [0,1]$. The random vector ξ has a bounded joint probability density function $\rho : \Gamma \to \mathbb{R}^+$, where $\Gamma := \prod_{i=1}^N \Gamma_i \subset \mathbb{R}^N$ and $\rho = \prod_{i=1}^N \rho_i(\xi_i)$. In particular, given the parametric representation (5) of $a(\mathbf{x},\omega)$, the Doob–Dynkin lemma (cf. [1]) guarantees that y, the solution corresponding to the stochastic PDE $c(y,u)=0$, admits exactly the same parametrization; that is, $y(\mathbf{x},\omega) = y(\mathbf{x},\xi_1(\omega),\xi_2(\omega),\ldots,\xi_N(\omega))$. Here, N has to be large enough so that the approximation error is sufficiently small.

We can now replace the probability space $(\Omega,\mathcal{F},\mathbb{P})$ with $(\Omega,\mathbb{B}(\Gamma),\rho(\xi)d\xi)$, where $\mathbb{B}(\Gamma)$ denotes the Borel σ-algebra on Γ and $\rho(\xi)d\xi$ is the distribution measure of the vector ξ. In addition, denoting the space of square-integrable random variables with respect to the density ρ by $L^2_\rho(\Gamma)$, we introduce the space $L^2(\mathcal{D})\otimes L^2_\rho(\Gamma)$, equipped with the norm

$$||v||_{L^2(\mathcal{D})\otimes L^2_\rho(\Gamma)} := \left(\int_\Gamma ||v(\cdot,\xi)||^2_{L^2(\mathcal{D})}\rho(\xi)\, d\xi\right)^{\frac{1}{2}} < \infty. \tag{6}$$

Similarly, using (3), we have

$$\operatorname{std}(g) = \left[\int_\Gamma (g(\xi)-\mathbb{E}(g(\xi)))^2\rho(\xi)\, d\xi\right]^{\frac{1}{2}} \text{ and } \langle g\rangle = \int_\Gamma g(\xi)\rho(\xi)\, d\xi < \infty. \tag{7}$$

Furthermore, our cost functional $\mathcal{J}(y,u)$ now reads

$$\mathcal{J}(y,u) := \frac{1}{2}||y-\bar{y}||^2_{L^2(\mathcal{D})\otimes L^2_\rho(\Gamma)} + \frac{\alpha}{2}||\operatorname{std}(y)||^2_{L^2(\mathcal{D})} + \frac{\beta}{2}||u||^2_{L^2(\mathcal{D})\otimes L^2_\rho(\Gamma)}. \tag{8}$$

In this contribution, we shall rely on the SGFEM for the spatial and stochastic discretizations (see, e.g., [1, 12, 21, 24]), and our exposition here, in particular, follows closely the framework in [24]. In a nutshell, we recall that the SGFEM is a spectral approach in which one seeks y and u in a finite-dimensional subspace of the Hilbert space $H^1_0(\mathcal{D})\otimes L^2_\rho(\Gamma)$, consisting of tensor products of deterministic functions defined on the spatial domain and stochastic functions defined on the probability space. More precisely, suppose first that $V_h \subset H^1_0(\mathcal{D})$ is a space of standard Lagrangian finite element functions on a partition $\mathcal{T}$ into triangles (or rectangles) of the domain $\mathcal{D}$ defined by

$$V_h := \{v_h \in H^1_0(\mathcal{D}) : v_h \in P_k(\Xi) \;\; \forall \Xi \in \mathcal{T}\},$$

where $\Xi \in \mathcal{T}$ is a cell and P_k is the space of Lagrangian polynomials of degree k. In particular, let $V_h = \operatorname{span}\{\phi_j(\mathbf{x}),\ j = 1, \ldots, J\}$. Moreover, for approximation in the stochastic space, define the set $\mathcal{I}$ by

$$\mathcal{I} := \{i = (i_1, \ldots, i_N) \in \mathbb{N}^N : |i| \le n\},$$

and let $Y_n \subset L^2_\rho(\Gamma)$ be such that $Y_n := \operatorname{span}\{\psi_i(\xi) : i \in \mathcal{I}\}$. Here, $\{\psi_i(\xi)\}$ are N-variate orthogonal polynomials of degree at most n, whereas $\mathcal{I}$ is a set of all multi-indices of length N satisfying $|i| \le n$. It can then be shown that

$$P := \dim(Y_n) = \dim(\mathcal{I}) = 1 + \sum_{k=1}^{n} \frac{1}{k!} \prod_{j=0}^{k-1} (N + j) = \frac{(N+n)!}{N!n!}. \tag{9}$$

Hence, it turns out that there exists a bijection $\mu : \{1, \ldots, P\} \to \mathcal{I}$ that assigns a unique integer i to each multi-index $\mu(i) \in \mathcal{I}$.

To illustrate here how the space Y_n is constructed [21], consider the case of uniform random variables with $N = 2$ and $n = 3$. Then Y_n is a set of two-dimensional Legendre polynomials (products of a univariate Legendre polynomial in ξ_1 and a univariate Legendre polynomial in ξ_2) of degree less than or equal to three. Each of the basis functions is associated with a multi-index $\nu = (\nu_1, \nu_2)$, where the components represent the degrees of the polynomials in ξ_1 and ξ_2. Since the total degree of the polynomial is three, we have the possibilities ν = (0,0), (1,0), (2,0), (3,0), (0,1), (1,1), (2,1), (0,2), (1,2), and $(0, 3)$. Since the univariate Legendre polynomials of degrees $0, 1, 2, 3$ are $L_0(x) = 1,\ L_1(x) = x,\ L_2(x) = \frac{1}{2}(3x^2 - 1),$ and $L_3(x) = \frac{1}{2}(5x^3 - 3x)$, we have

$$\begin{aligned} Y_n &= \operatorname{span}\{\psi_i(\xi)\}_{i=0}^{9} \\ &= \Big\{1, \xi_1,\ \frac{1}{2}(3\xi_1^2 - 1),\ \frac{1}{2}(5\xi_1^3 - 3\xi_1),\ \xi_2,\ \xi_1\xi_2,\ \frac{1}{2}(3\xi_1^2 - 1)\xi_2,\ \frac{1}{2}(3\xi_2^2 - 1), \\ &\qquad \frac{1}{2}\xi_1(3\xi_2^2 - 1),\ \frac{1}{2}(5\xi_2^3 - 3\xi_2)\Big\}. \end{aligned}$$

So, spectral SGFEM essentially entails performing a Galerkin projection onto $W_{hn} := \mathcal{X}_h \otimes \mathcal{Y}_n \subset H_0^1(\mathcal{D}) \otimes L^2_\rho(\Gamma)$ using basis functions r_{hn} of the form

$$r_{hn} = \sum_{j=1}^{J} \sum_{k \in \mathcal{I}} r_{jk} \phi_j(\mathbf{x}) \psi_k(\xi) = \sum_{j=1}^{J} \sum_{k=0}^{P-1} r_{jk} \phi_j(\mathbf{x}) \psi_k(\xi), \tag{10}$$

where r_{ij} is a degree of freedom. Note, in particular, that

$$\mathbb{E}(r_{hn}) = \sum_{j=1}^{J} \sum_{k=0}^{P-1} r_{jk} \phi_j(\mathbf{x}) \left\langle \psi_k(\xi) \right\rangle = \sum_{j=1}^{J} r_{j0} \phi_j(\mathbf{x}), \tag{11}$$

where

$$\left\langle \psi_0(\xi) \right\rangle = 1,\ \ \left\langle \psi_j(\xi) \right\rangle = 0,\ j > 0,\ \ \left\langle \psi_j(\xi)\psi_k(\xi) \right\rangle = \left\langle \psi_j^2(\xi) \right\rangle \delta_{jk}. \tag{12}$$

We now proceed to section 3 to present our first SOCP whose constraint is a stationary diffusion equation. The idea is to use this model to motivate our discussion on the solvers for a time-dependent model problem in section 4.

3. A control problem with stationary diffusion equation.

Our first SOCP consists now in minimizing the cost functional $\mathcal{J}(y(\mathbf{x},\omega),u(\mathbf{x},\omega))$ defined in (2) such that, $\mathbb{P}$-almost surely, the following linear elliptic diffusion equation holds:

$$\left\{\begin{aligned} -\nabla\cdot(a(\mathbf{x},\omega)\nabla y(\mathbf{x},\omega)) &= u(\mathbf{x},\omega) \text{ in } \mathcal{D}\times\Omega,\\ y(\mathbf{x},\omega) &= 0 \text{ on } \partial\mathcal{D}\times\Omega,\end{aligned}\right. \tag{13}$$

where $a:\mathcal{D}\times\Omega\to\mathbb{R}$ is a random coefficient field and the forcing term on the right-hand side $u:\mathcal{D}\times\Omega\to\mathbb{R}$ denotes a random control function. Furthermore, we assume that

$$u\in L^2(\mathcal{D})\otimes L^2(\Omega) \text{ a.e.,} \tag{14}$$

and that there exist positive constants $a_{\min}$ and $a_{\max}$ such that

$$\mathbb{P}\left(\omega\in\Omega : a(\mathbf{x},\omega)\in[a_{\min},a_{\max}] \ \ \forall\mathbf{x}\in\mathcal{D}\right)=1. \tag{15}$$

For the weak formulation of the forward problem (13), we seek $y\in H_0^1(\mathcal{D})\otimes L^2(\Omega)$ such that, $\mathbb{P}$-almost surely,

$$\mathcal{B}(y,v)=\ell(u,v) \ \ \forall v\in H_0^1(\mathcal{D})\otimes L^2(\Omega), \tag{16}$$

where the bilinear form $\mathcal{B}(\cdot,\cdot)$ is given by

$$\mathcal{B}(y,v)=\int_\Omega\int_\mathcal{D} a(\mathbf{x},\omega)\nabla y(\mathbf{x},\omega)\cdot\nabla v(\mathbf{x},\omega)\,d\mathbf{x}d\mathbb{P}(\omega),\ v,y\in H_0^1(\mathcal{D})\otimes L^2(\Omega), \tag{17}$$

and

$$\ell(u,v)=\int_\Omega\int_\mathcal{D} u(\mathbf{x},\omega)v(\mathbf{x},\omega)\,d\mathbf{x}d\mathbb{P}(\omega),\ \ v,u\in H_0^1(\mathcal{D})\otimes L^2(\Omega). \tag{18}$$

The following existence and uniqueness result of the solution y to (13) follows from the Lax–Milgram lemma; see, e.g., [12].

THEOREM 1. *Under the assumptions* (14) *and* (15)*, there exists a unique solution* $y\in H_0^1(\mathcal{D})\otimes L^2(\Omega)$ *such that,* $\mathbb{P}$*-almost surely,* (16) *holds.*

Recasting the above SOCP given by (2) and (13) into a saddle-point formulation, Chen and Quarteroni in [5] prove the existence and uniqueness of its solution. More precisely, the following result holds.

THEOREM 2 (see [5, Theorem 3.5]). *Let* (14) *and* (15) *be satisfied and let* $\alpha=0$ *in* (8)*. Then there exists a unique optimal solution* (y,u,f) *to the SOCP* (8) *and* (13) *satisfying the stochastic optimality conditions*

$$\begin{aligned} \mathcal{B}(y,v)=\ell(u,v),&\quad v\in H_0^1(\mathcal{D})\otimes L^2(\Omega),\\ \ell(\beta u-f,w)=0,&\quad w\in L^2(\mathcal{D})\otimes L^2(\Omega),\\ \mathcal{B}'(y,r)+\ell(y,r)=\ell(\bar{y},r),&\quad r\in H_0^1(\mathcal{D})\otimes L^2(\Omega),\end{aligned}$$

where f *is the adjoint variable or Lagrangian parameter associated with the optimal solution and* $\mathcal{B}'$ *is the adjoint bilinear form of* $\mathcal{B}$ *as defined in* (17)*; that is,* $\mathcal{B}'(y,r)=\mathcal{B}(r,y)$.

We note here that the cost functional considered in [5, 12] does not include $||\mathrm{std}(y)||^2_{L^2(\mathcal{D})}$. But then, their results extend to the more general form of $\mathcal{J}(y,u)$ discussed in this paper due to the Frechét differentiability of $||\mathrm{std}(y)||^2_{L^2(\mathcal{D})}$; see, for example, [24].

As our major concern in this paper is to study efficient solvers resulting from the discretization of our model problems, we proceed next to recall the two common approaches in the literature to solve these optimization problems [26]. The first method is the so-called *optimize-then-discretize* (OTD) approach. Here, one essentially considers the infinite-dimensional problem, writes down the first order conditions, and then discretizes the first order conditions. An alternative strategy, namely, the *discretize-then-optimize* (DTO) approach, involves discretizing the problem first and then building a discrete Lagrangian functional with the corresponding first order conditions. The commutativity of DTO and OTD methods when applied to optimal control problems constrained by PDEs has been a subject of debate in recent times (see [15] for an overview). In what follows, we will adopt the DTO strategy because, for the SOCPs considered in this paper, it leads to a symmetric saddle-point linear system which fits in nicely with our preconditioning strategy.

To discretize the SOCP given by (2) and (13) using the SGFEM, consider first the constraint (13). Given a basis for $W_{hn} := V_h \otimes Y_n \subset H_0^1(\mathcal{D}) \otimes L^2_\rho(\Gamma)$ and a truncated KLE representation $a_N(\mathbf{x},\xi)$ (cf. (4)) of the random field a satisfying (15), we now seek a finite-dimensional $y_{hn}, u_{hn} \in W_{hn}$ satisfying

$$\int_\Gamma \int_{\mathcal{D}} a_N(\mathbf{x},\xi) \nabla y_{hn} \cdot \nabla v \; \rho(\xi) d\mathbf{x} d\xi = \int_\Gamma \int_{\mathcal{D}} u_{hn} v \; \rho(\xi) d\mathbf{x} d\xi \tag{19}$$

$\forall v \in W_{hn}$. Expanding y_{hn}, u_{hn}, and the test functions in the chosen basis in (19), we see that

$$y_{hn} = \sum_{k=0}^{P-1} \sum_{j=1}^{J} y_{jk} \phi_j(\mathbf{x}) \psi_k(\xi) = \sum_{k=0}^{P-1} y_k \psi_k(\xi)$$

and

$$u_{hn} = \sum_{k=0}^{P-1} \sum_{j=1}^{J} u_{jk} \phi_j(\mathbf{x}) \psi_k(\xi) = \sum_{k=0}^{P-1} u_k \psi_k(\xi)$$

yield the following linear system of dimension $JP \times JP$:

$$\mathcal{K}\mathbf{y} = \mathcal{M}\mathbf{u} \tag{20}$$

with block structure, where the blocks $\mathcal{K}_{p,q}$ of the stochastic Galerkin matrix $\mathcal{K}$ are linear combinations of $N+1$ weighted stiffness matrices of dimension J, with each of them having the same sparsity pattern equivalent to that of the corresponding deterministic problem. More specifically, for $p,q = 0,\ldots,P-1$, we have

$$\mathcal{K}_{p,q} = \langle \psi_p(\xi)\psi_q(\xi) \rangle K_0 + \sum_{i=1}^{N} \langle \xi_i \psi_p(\xi)\psi_q(\xi) \rangle K_i \tag{21}$$

and

$$\mathcal{M}_{p,q} = \langle \psi_p(\xi)\psi_q(\xi) \rangle M,$$

where the mass matrix $M \in \mathbb{R}^{J\times J}$ and the stiffness matrices $K_i \in \mathbb{R}^{J\times J}$, $i = 0,1,\ldots,N$, are given, respectively, by

$$M(j,k) = \int_{\mathcal{D}} \phi_j(\mathbf{x})\phi_k(\mathbf{x})\, d\mathbf{x}, \tag{22}$$

$$K_0(j,k) = \int_{\mathcal{D}} \mathbb{E}[a](\mathbf{x})\nabla\phi_j(\mathbf{x})\nabla\phi_k(\mathbf{x})\, d\mathbf{x}, \tag{23}$$

$$K_i(j,k) = \sigma_a\sqrt{\lambda_i}\int_{\mathcal{D}} \varphi_i(\mathbf{x})\nabla\phi_j(\mathbf{x})\nabla\phi_k(\mathbf{x})\, d\mathbf{x},\ i>0, \tag{24}$$

where $\mathbb{E}[a] > 0$ due to (15), so that K_0 is symmetric and positive definite. The block K_0 captures the mean information in the model and appears on the diagonal blocks of $\mathcal{K}$, whereas the other blocks $K_i,\ i = 1,\ldots,N$, represent the fluctuations in the model. In Kronecker product notation, one obtains

$$\mathcal{K} := G_0 \otimes K_0 + \sum_{i=1}^{N} G_i \otimes K_i, \qquad \mathcal{M} := G_0 \otimes M, \tag{25}$$

where

$$\begin{cases} G_0 = \operatorname{diag}\left(\left\langle\psi_0^2\right\rangle, \left\langle\psi_1^2\right\rangle, \ldots, \left\langle\psi_{P-1}^2\right\rangle\right), \\ G_i(j,k) = \left\langle \xi_i\psi_j\psi_k\right\rangle, \quad i = 1,\ldots,N, \end{cases} \tag{26}$$

due to the orthogonality of the stochastic basis functions with respect to the probability measure of the distribution of the chosen random variables (cf. (12)). Moreover, $\mathcal{K}$ is highly sparse as many of the sums in (21) are zero.

Similarly, applying SGFEM to the cost function (8), taking into account (11), (12), and the expression $\mathbb{V}\mathrm{ar}(y) = [\mathrm{std}(y)]^2 = \mathbb{E}(y^2) - [\mathbb{E}(y)]^2$, leads to

$$\frac{1}{2}(\mathbf{y}-\bar{\mathbf{y}})^T\mathcal{M}(\mathbf{y}-\bar{\mathbf{y}}) + \frac{\alpha}{2}\mathbf{y}^T\mathcal{M}_t\mathbf{y} + \frac{\beta}{2}\mathbf{u}^T\mathcal{M}\mathbf{u}, \tag{27}$$

where

$$\mathcal{M}_t := H_0 \otimes M, \quad H_0 := \operatorname{diag}\left(0, \left\langle\psi_1^2\right\rangle, \ldots, \left\langle\psi_{P-1}^2\right\rangle\right). \tag{28}$$

Our discrete SOCP now is to minimize (27) subject to (20). The Lagrangian functional $\mathcal{L}$ of this optimization problem is given by

$$\mathcal{L}(\mathbf{y},\mathbf{u},\mathbf{f}) := \frac{1}{2}(\mathbf{y}-\bar{\mathbf{y}})^T\mathcal{M}(\mathbf{y}-\bar{\mathbf{y}}) + \frac{\alpha}{2}\mathbf{y}^T\mathcal{M}_t\mathbf{y} + \frac{\beta}{2}\mathbf{u}^T\mathcal{M}\mathbf{u} + \mathbf{f}^T(-\mathcal{K}\mathbf{y} + \mathcal{M}\mathbf{u} - \mathbf{d}),$$

where $\mathbf{f}$ denotes the Lagrangian multiplier or adjoint associated with the constraint and the vector $\mathbf{d} := \operatorname{diag}(G_0)\otimes\tilde{\mathbf{d}}$, where $\tilde{\mathbf{d}}$ represents, in general, contributions from boundary conditions with respect to the spatial discretization. Now, applying the first order conditions to the Lagrangian yields the optimality system

$$\underbrace{\begin{bmatrix} \mathcal{M}_\alpha & 0 & -\mathcal{K}^T \\ 0 & \beta\mathcal{M} & \mathcal{M}^T \\ -\mathcal{K} & \mathcal{M} & 0 \end{bmatrix}}_{:=\mathcal{A}} \begin{bmatrix} \mathbf{y} \\ \mathbf{u} \\ \mathbf{f} \end{bmatrix} = \begin{bmatrix} \mathcal{M}\bar{\mathbf{y}} \\ \mathbf{0} \\ \mathbf{d} \end{bmatrix}, \tag{29}$$

where

$$\mathcal{M}_\alpha = \mathcal{M} + \alpha\mathcal{M}_t = (G_0 \otimes M) + \alpha(H_0 \otimes M) = G_\alpha \otimes M, \tag{30}$$

with $G_\alpha := G_0 + \alpha H_0$, so that

$$G_\alpha(j,k) = \begin{cases} \langle \psi_0^2 \rangle & \text{if} \quad j = k = 0, \\ (1+\alpha)\langle \psi_j^2 \rangle & \text{if} \quad j = k = 1, 2, \ldots, P-1, \\ 0 & \text{otherwise.} \end{cases} \tag{31}$$

We note from (25), (30), and (31) that if $\alpha = 0$, then $G_\alpha = G_0$ and, hence, $\mathcal{M}_\alpha = \mathcal{M}$. Moreover, we assume that the parameter N in the KLE of the random input a is chosen such that $\mathcal{K}$ stays symmetric and positive definite [21]. The system (29) is usually of huge dimension. As a result, the use of direct solvers for the system is out of the question. In what follows, we consider efficient iterative solvers instead. First, however, we discuss our preconditioning strategies.

3.1. Preconditioning the optimality system. Now, observe that (29) is an indefinite *saddle-point* system [8, Chapter 5]:

$$\mathcal{A} = \begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix}, \tag{32}$$

where

$$A = \begin{bmatrix} \mathcal{M}_\alpha & 0 \\ 0 & \beta\mathcal{M} \end{bmatrix}, \quad B = [-\mathcal{K} \;\; \mathcal{M}], \tag{33}$$

where A is symmetric and positive definite and B has full row rank. An appropriate Krylov subspace solver for the linear system is the MINRES algorithm (originally proposed by Paige and Saunders in [19]) with a suitable preconditioner.

Throughout this paper, we will focus mainly on block-diagonal preconditioners. More specifically, to solve (29), we precondition the MINRES algorithm with

$$\mathcal{P} := \begin{bmatrix} A & 0 \\ 0 & S \end{bmatrix} = \begin{bmatrix} \mathcal{M}_\alpha & 0 & 0 \\ 0 & \beta\mathcal{M} & 0 \\ 0 & 0 & S \end{bmatrix}, \tag{34}$$

where

$$S = BA^{-1}B^T = \mathcal{K}\mathcal{M}_\alpha^{-1}\mathcal{K} + \frac{1}{\beta}\mathcal{M} \tag{35}$$

is the (negative) *Schur complement.* We note here that (34) is only an ideal preconditioner for our saddle-point system (29) in the sense that it is not cheap to solve the system with it. In practice, one often has to approximate its three diagonal blocks in order to use $\mathcal{P}$ with MINRES. An effective approach to approximate blocks $(1,1)$ and $(2,2)$ is the application of Chebyshev semi-iteration to the mass matrices in each of the two blocks [27]. Approximating the Schur complement S, that is, block $(3,3)$, poses more difficulty, however. One possibility [22] is to approximate S by dropping the term $\frac{1}{\beta}\mathcal{M}$ to obtain

$$S_0 := \mathcal{K}\mathcal{M}_\alpha^{-1}\mathcal{K}^T. \tag{36}$$

An alternative and more robust approach, which we adopt here and in the rest of this paper, was proposed in [20] (see also [8, Chapter 5]) in the context of deterministic optimal control problems. Here, S is approximated by a matrix S_1 of the form

$$S_1 = (\mathcal{K} + \mathcal{M}_u)\, \mathcal{M}_\alpha^{-1}\, (\mathcal{K} + \mathcal{M}_u)^T\,, \tag{37}$$

where $\mathcal{M}_u$ is determined by "matching" the terms in the expressions for S_1 and S as given, respectively, in (37) and (35). More precisely, we ignore the cross terms (that is, $\mathcal{K}\mathcal{M}_\alpha^{-1}\mathcal{M}_u + \mathcal{M}_u\mathcal{M}_\alpha^{-1}\mathcal{K}$) in the expansion of S_1, to get

$$\mathcal{M}_u\mathcal{M}_\alpha^{-1}\mathcal{M}_u = \frac{1}{\beta}\mathcal{M} = \frac{1}{\beta}\mathcal{M}\mathcal{M}^{-1}\mathcal{M}. \tag{38}$$

Now, observe from (26), (30), and (31) that we have $\mathcal{M}_\alpha = G_\alpha \otimes M$. Moreover, note that ideally in (8), we have $\alpha \geq 0$. So, to derive an approximation to S_1, we consider first of all the case $\alpha = 0$. In this case, it is easy to see that (38) holds if we set

$$\mathcal{M}_u = \frac{1}{\sqrt{\beta}}\mathcal{M}, \tag{39}$$

since $\mathcal{M}_\alpha = \mathcal{M}$. If $\alpha > 0$, then we apply the following trick. We proceed first by replacing in (31) the $(0,0)$ entry in the diagonal matrix G_α with $(1+\alpha)\left\langle \psi_0^2 \right\rangle$, so that we can then obtain

$$\mathcal{M}_\alpha = G_\alpha \otimes M \approx (1+\alpha) G_0 \otimes M = (1+\alpha)\mathcal{M}.$$

It turns out then that (38) holds if and only if

$$\mathcal{M}_u = \sqrt{\frac{1+\alpha}{\beta}}\mathcal{M},$$

with which we recover (39) for $\alpha = 0$. Hence, we have

$$S_1 = \underbrace{\left(\mathcal{K} + \sqrt{\frac{1+\alpha}{\beta}}\mathcal{M}\right)}_{:=\mathcal{Z}} \mathcal{M}_\alpha^{-1} \left(\mathcal{K} + \sqrt{\frac{1+\alpha}{\beta}}\mathcal{M}\right)^T. \tag{40}$$

We point out here that the expression for $\mathcal{M}_u$ implies that the ignored cross terms are $\mathcal{O}(\beta^{-1/2})$ instead of $\mathcal{O}(\beta^{-1})$ in (36).

The effectiveness of the iterative solver used to solve our KKT system depends to a large extent on how well the approximation S_1 represents the exact Schur complement. To measure this, we need to consider the eigenvalues of the preconditioned Schur complement $S_1^{-1}S$. In what follows, we proceed to establish the spectrum $\lambda(S_1^{-1}S)$ of $S_1^{-1}S$ by examining the Raleigh quotient

$$R(x) := \frac{x^T S x}{x^T S_1 x}$$

for any nonzero vector x of appropriate dimension. We shall rely on the following result on positive definite matrices.

PROPOSITION 3 (see [16, Theorem 2]). *Let $X = AB + BA$, where A and B are positive definite, Hermitian square matrices. Then X is positive definite if*

$$\kappa(B) < \left(\frac{\sqrt{\kappa(A)}+1}{\sqrt{\kappa(A)}-1}\right)^2,$$

where $\kappa(Y)$ represents the spectral condition number of the matrix Y.

We can now prove the main result of this section.

Theorem 4. *Let $\alpha \in [0, +\infty)$. Then the eigenvalues of $S_1^{-1}S$ satisfy*

$$\lambda(S_1^{-1}S) \subset \left[\frac{1}{2(1+\alpha)}, 1\right) \quad \forall \alpha < \left(\frac{\sqrt{\kappa(\mathcal{C})}+1}{\sqrt{\kappa(\mathcal{C})}-1}\right)^2 - 1, \tag{41}$$

where $\mathcal{C} = \mathcal{M}^{-1/2}\mathcal{K}\mathcal{M}^{-1/2}$.

Proof. Suppose that $\alpha \in [0, +\infty)$. Define the diagonal matrices Υ and $\mathcal{E}_\alpha$ by

$$\Upsilon = \mathrm{diag}(0, I_{P-1}) \ \text{ and } \ \mathcal{E}_\alpha = (I_P + \alpha\Upsilon) \otimes I_J, \tag{42}$$

where I_n denotes an identity matrix of dimension $n \in \mathbb{N}$. Clearly,

$$I_{JP} \preceq \mathcal{E}_\alpha \preceq (1+\alpha)I_{JP} \ \text{ and } \ I_{JP} \succeq \mathcal{E}_\alpha^{-1} \succeq (1+\alpha)^{-1}I_{JP}, \tag{43}$$

where, for arbitrary square matrices X and Y, we write $X \succeq Y$ if $X - Y \geq 0$, and vice versa. Moreover, from (26), (28), (30), (42), and using the identity $(A \otimes B)(X \otimes Y) = AX \otimes BY$, we obtain

$$\begin{aligned}
\mathcal{M}_\alpha &= (G_0 + \alpha H_0) \otimes M \\
&= (G_0 I_P + \alpha G_0 \Upsilon) \otimes (M I_J) \\
&= (G_0 \otimes M)(I_P \otimes I_J) + (G_0 \otimes M)(\alpha\Upsilon \otimes I_J) \\
&= (G_0 \otimes M)\left[(I_P \otimes I_J) + (\alpha\Upsilon \otimes I_J)\right] \\
&= \mathcal{M}\left[(I_P + \alpha\Upsilon) \otimes I_J\right] \\
&= \mathcal{M}\mathcal{E}_\alpha = \mathcal{E}_\alpha\mathcal{M},
\end{aligned} \tag{44}$$

since both G_0 and $I_P + \alpha\Upsilon$ are diagonal matrices and therefore commute with each other. Now, recall from (40) that the approximation S_1 to the Schur complement S is given by

$$S_1 = \mathcal{K}\mathcal{M}_\alpha^{-1}\mathcal{K} + \frac{1+\alpha}{\beta}\mathcal{M}\mathcal{M}_\alpha^{-1}\mathcal{M} + \sqrt{\frac{1+\alpha}{\beta}}\left[\mathcal{K}\mathcal{M}_\alpha^{-1}\mathcal{M} + \mathcal{M}\mathcal{M}_\alpha^{-1}\mathcal{K}\right], \tag{45}$$

and that the preconditioned Schur complement $S_1^{-1}S$ is similar to the matrix

$$\mathcal{M}^{1/2}S_1^{-1}S\mathcal{M}^{-1/2} = (\mathcal{M}^{-1/2}S_1\mathcal{M}^{-1/2})^{-1}(\mathcal{M}^{-1/2}S\mathcal{M}^{-1/2}). \tag{46}$$

It therefore follows from (35), (40), (44), (45), and (46) that

$$\begin{aligned}
S_1^{-1}S &\sim \left(\mathcal{C}\mathcal{E}_\alpha^{-1}\mathcal{C} + \frac{1+\alpha}{\beta}\mathcal{E}_\alpha^{-1} + \sqrt{\frac{1+\alpha}{\beta}}\left(\mathcal{C}\mathcal{E}_\alpha^{-1} + \mathcal{E}_\alpha^{-1}\mathcal{C}\right)\right)^{-1}\left(\mathcal{C}\mathcal{E}_\alpha^{-1}\mathcal{C} + \beta^{-1}I_{JP}\right) \\
&= \left(\beta\mathcal{C}\mathcal{E}_\alpha^{-1}\mathcal{C} + (1+\alpha)\mathcal{E}_\alpha^{-1} + \sqrt{\beta(1+\alpha)}\left(\mathcal{C}\mathcal{E}_\alpha^{-1} + \mathcal{E}_\alpha^{-1}\mathcal{C}\right)\right)^{-1}\left(\beta\mathcal{C}\mathcal{E}_\alpha^{-1}\mathcal{C} + I_{JP}\right),
\end{aligned}$$

where $\sim$ implies similarity transformation and $\mathcal{C} := \mathcal{M}^{-1/2}\mathcal{K}\mathcal{M}^{-1/2}$. Now, observe that the matrix $\mathcal{C}$ is symmetric and positive definite so that $\lambda(\mathcal{C}) \subset (0, +\infty)$. Consider now the Raleigh quotient

$$R(x) := \frac{x^T\left[\beta\mathcal{C}\mathcal{E}_\alpha^{-1}\mathcal{C} + I_{JP}\right]x}{x^T\left[\beta\mathcal{C}\mathcal{E}_\alpha^{-1}\mathcal{C} + (1+\alpha)\mathcal{E}_\alpha^{-1} + \sqrt{\beta(1+\alpha)}\left(\mathcal{C}\mathcal{E}_\alpha^{-1} + \mathcal{E}_\alpha^{-1}\mathcal{C}\right)\right]x}.$$

But then $\kappa(\mathcal{E}_\alpha^{-1}) = 1 + \alpha$, and hence, by Proposition 3, we have that

$$x^T(\mathcal{C}\mathcal{E}_\alpha^{-1} + \mathcal{E}_\alpha^{-1}\mathcal{C})x > 0 \ \text{ for } \ \alpha + 1 < \left(\frac{\sqrt{\kappa(\mathcal{C})}+1}{\sqrt{\kappa(\mathcal{C})}-1}\right)^2.$$

This, in turn, implies that the denominator of $R(x)$ is also strictly positive. Hence,

$$R(x) \leq \frac{x^T\left[\beta\mathcal{C}\mathcal{E}_\alpha^{-1}\mathcal{C} + (1+\alpha)\mathcal{E}_\alpha^{-1}\right]x}{x^T\left[\beta\mathcal{C}\mathcal{E}_\alpha^{-1}\mathcal{C} + (1+\alpha)\mathcal{E}_\alpha^{-1} + \sqrt{\beta(1+\alpha)}\left(\mathcal{C}\mathcal{E}_\alpha^{-1} + \mathcal{E}_\alpha^{-1}\mathcal{C}\right)\right]x} < 1,$$

from which we deduce that $\lambda_{\max} := \max R(x) < 1$.

Now, observe that $x^T\mathcal{C}\mathcal{E}_\alpha^{-1}x = x^T\mathcal{E}_\alpha^{-1}\mathcal{C}x$. Moreover, for any two vectors z_1, z_2 of appropriate dimensions, Cauchy–Schwarz inequality implies $\langle z_1^T z_2\rangle^2 \leq (z_1^T z_1)(z_2^T z_2)$. Thus, setting $z_1^T = x^T\mathcal{C}\mathcal{E}_\alpha^{-1/2}$ and $z_2 = \mathcal{E}_\alpha^{-1/2}x$, we obtain

$$\left(x^T\mathcal{C}\mathcal{E}_\alpha^{-1}x\right)^2 \leq \left(x^T\mathcal{C}\mathcal{E}_\alpha^{-1}\mathcal{C}x\right)\left(x^T\mathcal{E}_\alpha^{-1}x\right). \tag{47}$$

Hence, using (47), together with the fact that $(a+b)^2 \leq 2(a^2+b^2)$, $a, b \in \mathbb{R}$, yields

$$\begin{aligned}
R(x) &= \frac{x^T\left[\beta\mathcal{C}\mathcal{E}_\alpha^{-1}\mathcal{C} + I_{JP}\right]x}{x^T\left[\beta\mathcal{C}\mathcal{E}_\alpha^{-1}\mathcal{C} + (1+\alpha)\mathcal{E}_\alpha^{-1} + \sqrt{\beta(1+\alpha)}\left(\mathcal{C}\mathcal{E}_\alpha^{-1} + \mathcal{E}_\alpha^{-1}\mathcal{C}\right)\right]x} \\
&\geq \frac{x^T\beta\mathcal{C}\mathcal{E}_\alpha^{-1}\mathcal{C}x + x^T I_{JP}x}{\left[\beta^{1/2}(x^T\mathcal{C}\mathcal{E}_\alpha^{-1}\mathcal{C}x)^{1/2} + (1+\alpha)^{1/2}(x^T\mathcal{E}_\alpha^{-1}x)^{1/2}\right]^2} \\
&\geq \frac{x^T\beta\mathcal{C}\mathcal{E}_\alpha^{-1}\mathcal{C}x + x^T I_{JP}x}{2\left[\beta x^T\mathcal{C}\mathcal{E}_\alpha^{-1}\mathcal{C}x + (1+\alpha)x^T\mathcal{E}_\alpha^{-1}x\right]} \\
&\geq \frac{x^T\beta\mathcal{C}\mathcal{E}_\alpha^{-1}\mathcal{C}x + x^T\mathcal{E}_\alpha^{-1}x}{2\left[\beta x^T\mathcal{C}\mathcal{E}_\alpha^{-1}\mathcal{C}x + (1+\alpha)x^T\mathcal{E}_\alpha^{-1}x\right]} \\
&\geq \frac{x^T\mathcal{E}_\alpha^{-1}x}{2(1+\alpha)x^T\mathcal{E}_\alpha^{-1}x} = \frac{1}{2(1+\alpha)},
\end{aligned} \tag{48}$$

which shows that $\lambda_{\min} := \min R(x) \geq \frac{1}{2(1+\alpha)}$. □

Note that, in the context of a deterministic optimal control problem, Pearson and Wathen in [20, Theorem 4] have independently obtained, specifically for $\alpha = 0$, a similar result to that of Theorem 4; see also [8, Lemma 5.2]. We, however, point out herein that, in addition to the generalization of the said result, ours yields a sharper bound than the one obtained by these authors. Moreover, save the parameter α, the result of Theorem 4 is independent of the discretization parameters in the system.

The following result is an immediate consequence of Theorem 4.

THEOREM 5. *Let $\mathcal{A}$ be the KKT matrix given by* (32) *and define $\mathcal{P}_0$ by*

$$\mathcal{P}_0 = \begin{bmatrix} A & 0 \\ 0 & S_1 \end{bmatrix},$$

where A and S_1 are given, respectively, by (33) *and* (40). *Moreover, assume that $\alpha < \left(\frac{\sqrt{\kappa(\mathcal{C})}+1}{\sqrt{\kappa(\mathcal{C})}-1}\right)^2 - 1$, where $\mathcal{C}$ is as defined in Theorem* 4. *Then the eigenvalues of the matrix $\mathcal{P}_0^{-1}\mathcal{A}$ satisfy*

$$\lambda(\mathcal{P}_0^{-1}\mathcal{A}) = \{1\} \cup \mathcal{I}^- \cup \mathcal{I}^+, \tag{49}$$

where

$$\mathcal{I}^- = \left(\frac{1}{2}(1-\sqrt{5}), \frac{1}{2}\left(1-\sqrt{1+\frac{2}{1+\alpha}}\right)\right], \quad \mathcal{I}^+ = \left[\frac{1}{2}\left(1+\sqrt{1+\frac{2}{1+\alpha}}\right), \frac{1}{2}(1+\sqrt{5})\right).$$

Proof. First, we note that $\mathcal{P}_0^{-1}\mathcal{A}$ shares the same eigenvalues with the symmetric matrix given by

$$\mathcal{P}_0^{-1/2}\mathcal{A}\mathcal{P}_0^{-1/2} = \left[\begin{array}{cc} I & A^{-1/2}B^T S_1^{-1/2} \\ S_1^{-1/2}BA^{-1/2} & 0 \end{array}\right].$$

Now, using [9, Lemma 2.1], we know that the eigenvalues of $\mathcal{P}_0^{-1/2}\mathcal{A}\mathcal{P}_0^{-1/2}$ are either 1 or have the form $\frac{1}{2}\left(1 \pm \sqrt{1+4s^2}\right)$, where s is a singular value of $X := S_1^{-1/2}BA^{-1/2}$; in other words, s^2 is an eigenvalue of XX^T. Since $S_1^{-1}S$ is similar to XX^T, the result (49) follows immediately from Theorem 4. □

The robustness of S_1 notwithstanding, we cannot implement it as it is, since this would be equivalent to solving the forward problem twice per iteration due to the presence of $\mathcal{Z} := \mathcal{K} + \sqrt{\frac{1+\alpha}{\beta}}\mathcal{M}$ and its transpose in (40). Hence, we need to derive an appropriate approximation for $\mathcal{Z}$. To this end, observe first, from (25), that since

$$\mathcal{Z} = \mathcal{K} + \sqrt{\frac{1+\alpha}{\beta}}\mathcal{M} = \sum_{i=0}^{N} G_i \otimes \tilde{K}_i, \tag{50}$$

with $\tilde{K}_0 := K_0 + \sqrt{\frac{1+\alpha}{\beta}}M$, $\tilde{K}_i = K_i$, $i = 1, \ldots, N$, one could approximate $\mathcal{Z}$ using, for example, the block-diagonal *mean-based preconditioner* [3, 21]:

$$\mathcal{Z}_0 := G_0 \otimes \tilde{K}_0. \tag{51}$$

For a practical algorithm, S_1 could then be implemented using multigrid techniques for $\tilde{K}_0$ in $\mathcal{Z}_0$. Note that (51) is best suited for systems for which the variance of the random input a is small relative to its mean. Its performance, unfortunately, deteriorates with increasing σ_a, since in this case we see from (23), (24), and (25), that the off-diagonal blocks of the global stochastic Galerkin matrix $\sum_{i=1}^{N} G_i \otimes K_i$ become more significant and are not represented in the preconditioner [21].

As an alternative, in our numerical experiments we therefore additionally consider solves with $\mathcal{Z}$ (that is, $\mathcal{Z}\mathbf{x} = \mathbf{b}$) via the inexact Uzawa method as given by Algorithm 1. In our experience, the latter approach proved more efficient (with just a few iterations) than the former, especially as we increased the variance of a.

Algorithm 1 Inexact Uzawa method for $\mathcal{Z}\mathbf{x} = \mathbf{b}$.

```
Select $\mathbf{x}_0$
Set $\widehat{\mathcal{P}} := \mathcal{Z}_0$
for $k = 0, 1, 2, \ldots$ do
    $\mathbf{r}_k = \mathbf{b} - \mathcal{Z}\mathbf{x}_k$
    $\mathbf{x}_{k+1} = \mathbf{x}_k + \widehat{\mathcal{P}}^{-1}\mathbf{r}_k$
end for
```

In a nutshell, we outline below the dominant operations in the application of our proposed block-diagonal preconditioner $\mathcal{P}$ in (34).

- **(1,1)**: 1 Chebyshev semi-iteration for the mass matrix M.
- **(2,2)**: 1 Chebyshev semi-iteration for the mass matrix M.
- **(3,3)**: 2 multigrid (or inexact Uzawa) operations: 1 for $\mathcal{Z}_0$ (resp., $\mathcal{Z}$) and 1 for its transpose.
- **Total**: 2 Chebyshev semi-iterations and 2 multigrid (or inexact Uzawa) operations.

Having been equipped with a suitable preconditioner, we proceed to the next section to discuss our Krylov subspace solver.

3.2. Computing low-rank approximation of the solution to the stationary problem. As we have already pointed out in section 3.1, the MINRES algorithm is an optimal solver for the system (29). Hence, we will use it, together with (34), to solve (29). In particular, our approach is based on the low-rank version of MINRES presented in [25]. In this section, we give a brief overview of this low-rank iterative solver. Now, observe first that using the identity

$$\text{vec}(WXV) = (V^T \otimes W)\text{vec}(X), \tag{52}$$

where $\text{vec}(X) = (x_1, \ldots, x_m)^T \in \mathbb{R}^{mn\times 1}$ is a column vector obtained by stacking the columns of the matrix $X = [x_1, \ldots, x_m] \in \mathbb{R}^{n\times m}$ on top of each other, the linear system (29) can be rewritten as $\mathcal{A}\mathcal{X} = \mathcal{R}$, where

$$\mathcal{A} = \begin{bmatrix} G_\alpha \otimes M & 0 & -\sum_{i=0}^{N} G_i \otimes K_i \\ 0 & \beta(G_0 \otimes M) & G_0 \otimes M \\ -\sum_{i=0}^{N} G_i \otimes K_i & G_0 \otimes M & 0 \end{bmatrix},$$

$$\mathcal{X} = \begin{bmatrix} \text{vec}(Y) \\ \text{vec}(U) \\ \text{vec}(F) \end{bmatrix}, \quad \mathcal{R} = \begin{bmatrix} \text{vec}(R_1) \\ 0 \\ \text{vec}(R_3) \end{bmatrix},$$

and

$$Y = [y_0, \ldots, y_{P-1}], \quad U = [u_0, \ldots, u_{P-1}], \quad F = [f_0, \ldots, f_{P-1}],$$
$$R_1 = \text{vec}^{-1}((G_0 \otimes M)\bar{\mathbf{y}}), \quad R_3 = \text{vec}^{-1}(\mathbf{d}).$$

Hence, (52) implies that

$$\mathcal{A}\mathcal{X} = \text{vec}\left(\begin{bmatrix} MYG_\alpha^T - \sum_{i=0}^{N} K_i F G_i^T \\ \beta M U G_0^T + M F G_0^T \\ -\sum_{i=0}^{N} K_i Y G_i^T + M U G_0^T \end{bmatrix}\right) = \text{vec}\left(\begin{bmatrix} R_1 \\ 0 \\ R_3 \end{bmatrix}\right). \tag{53}$$

Our approach is essentially based on the assumption that both the solution matrix $\mathcal{X}$ and the right-hand side matrix $\mathcal{R}$ admit low-rank representations; that is,

$$\begin{cases} Y = W_Y V_Y^T, & \text{with } W_Y \in \mathbb{R}^{J\times k_1}, \ V_Y \in \mathbb{R}^{P\times k_1}, \\ U = W_U V_U^T, & \text{with } W_U \in \mathbb{R}^{J\times k_2}, \ V_U \in \mathbb{R}^{P\times k_2}, \\ F = W_F V_F^T, & \text{with } W_F \in \mathbb{R}^{J\times k_3}, \ V_F \in \mathbb{R}^{P\times k_3}, \end{cases} \tag{54}$$

where $k_{1,2,3}$ are small relative to P. Substituting (54) in (53) and ignoring the vec operator, we then obtain

$$\begin{bmatrix} MW_Y V_Y^T G_\alpha^T - \sum\limits_{i=0}^{N} K_i W_F V_F^T G_i^T \\ \beta M W_U V_U^T G_0^T + M W_F V_F^T G_0^T \\ -\sum\limits_{i=0}^{N} K_i W_Y V_Y^T G_i^T + M W_U V_U^T G_0^T \end{bmatrix} = \begin{bmatrix} R_{11} R_{12}^T \\ 0 \\ R_{31} R_{32}^T \end{bmatrix}, \tag{55}$$

where $R_{11}R_{12}^T$ and $R_{31}R_{32}^T$ are the low-rank representations of R_1 and R_3, respectively.

The attractiveness of this approach lies therefore in the fact that one can rewrite the three block rows on the left-hand side in (55), respectively, as

$$\begin{cases} \text{(first block row)} \begin{bmatrix} MW_Y & -\sum\limits_{i=0}^{N} K_i W_F \end{bmatrix} \begin{bmatrix} V_Y^T G_\alpha^T \\ V_F^T G_i^T \end{bmatrix}, \\[2ex] \text{(second block row)} \begin{bmatrix} \beta M W_U & M W_F \end{bmatrix} \begin{bmatrix} V_U^T G_0^T \\ V_F^T G_0^T \end{bmatrix}, \\[2ex] \text{(third block row)} \begin{bmatrix} -\sum\limits_{i=0}^{N} K_i W_Y & M W_U \end{bmatrix} \begin{bmatrix} V_Y^T G_i^T \\ V_U^T G_0^T \end{bmatrix}, \end{cases} \tag{56}$$

so that the low-rank nature of the factors guarantees fewer multiplications with the submatrices while maintaining smaller storage requirements. More precisely, keeping in mind that

$$x = \operatorname{vec}\left(\begin{bmatrix} X_{11} X_{12}^T \\ X_{21} X_{22}^T \\ X_{31} X_{32}^T \end{bmatrix} \right)$$

corresponds to the associated vector x from a vector-based version of MINRES, matrix-vector multiplication in our low-rank MINRES is given by Algorithm 2. Note that an important feature of low-rank MINRES is that the iterates of the solution matrices Y, U, and F in the algorithm are truncated by a truncation operator $\mathcal{T}_\epsilon$ with a prescribed tolerance ϵ. This is accomplished via QR decomposition as in [14] or truncated singular value decomposition (SVD) as in [3, 25].

Algorithm 2 Matrix-vector multiplication in low-rank MINRES.

```
Input: $W_{11}, W_{12}, W_{21}, W_{22}, W_{31}, W_{32}$
Output: $X_{11}, X_{12}, X_{21}, X_{22}, X_{31}, X_{32}$
$X_{11} = [MW_{11} \quad -K_0 W_{31} \quad \cdots \quad -K_N W_{31}]$
$X_{12} = [G_\alpha W_{12} \quad G_0 W_{32} \quad \cdots \quad G_N W_{32}]$
$X_{21} = [\beta M W_{21} \quad M W_{31}]$
$X_{22} = [G_0 W_{22} \quad G_0 W_{32}]$
$X_{31} = [-K_0 W_{11} \quad \cdots \quad -K_N W_{11} \quad M W_{21}]$
$X_{32} = [\ G_0 W_{12} \quad \cdots \quad G_N W_{12} \quad G_0 W_{22}]$
```

The truncation operation is necessary because the new computed factors could have increased ranks compared to the original factors in (56). Hence, a truncation of all the factors after the matrix-vector products is used to construct new factors; for instance,

$$[\tilde{X}_{11}, \tilde{X}_{12}] := \mathcal{T}_\epsilon\left([X_{11}, X_{12}]\right) = \mathcal{T}_\epsilon\left(\left[\begin{array}{cc} MW_{11} & -\sum_{i=0}^{N} K_i W_{31} \end{array}\right]\left[\begin{array}{c} W_{12}^T G_\alpha^T \\ W_{32}^T G_i^T \end{array}\right]\right).$$

Moreover, in order to ensure that the inner products within the iterative low-rank solver are computed efficiently, we use the fact that

$$\langle x, y\rangle = \mathrm{vec}\,(X)^T\,\mathrm{vec}\,(Y) = \mathrm{trace}\left(X^T Y\right)$$

to deduce that

$$\mathrm{trace}\left(\underbrace{\left(X_1 X_2^T\right)^T}_{\text{Large}}\underbrace{\left(Y_1 Y_2^T\right)}_{\text{Large}}\right) = \mathrm{trace}\left(\underbrace{Y_2^T X_2}_{\text{Small}}\underbrace{X_1^T Y_1}_{\text{Small}}\right),$$

where $X = X_1 X_2^T$ and $Y = Y_1 Y_2^T$, which allows us to compute the trace of small matrices rather than of those from the full model.

For more details on implementation issues, we refer the interested reader to [3, 25]. In section 5, we use numerical experiments to illustrate the performance of low-rank MINRES, together with the preconditioners discussed in section 3.1.

Next, we proceed to section 4 to present a time-dependent analogue of the model problem considered so far.

4. A stochastic parabolic optimal control problem.

In an attempt to extend our discussion on the above model problem to a time-dependent case, we henceforth replace $L^2(\mathcal{D})$ in (2) by the space

$$L^2(0,T;\mathcal{D}) = \left\{f \in L^2(\mathcal{D}) : \int_0^T [f(t)]^2\, dt < \infty\right\}$$

and then consider a parabolic SOCP now given by

$$\begin{aligned}\mathcal{J}(t,y,u) = {} & \frac{1}{2}||y-\bar{y}||^2_{L^2(0,T;\mathcal{D})\otimes L^2(\Omega)} + \frac{\alpha}{2}||\mathrm{std}(y)||^2_{L^2(0,T;\mathcal{D})} \\ & + \frac{\beta}{2}||u||^2_{L^2(0,T;\mathcal{D})\otimes L^2(\Omega)} \end{aligned} \tag{57}$$

subject, $\mathbb{P}$-almost surely, to

$$\left\{\begin{aligned} \frac{\partial y(t,\mathbf{x},\omega)}{\partial t} - \nabla\cdot(a(\mathbf{x},\omega)\nabla y(t,\mathbf{x},\omega)) = u(t,\mathbf{x},\omega) & \text{ in } (0,T]\times\mathcal{D}\times\Omega, \\ y(t,\mathbf{x},\omega) = 0 & \text{ on } (0,T]\times\partial\mathcal{D}\times\Omega, \\ y(0,\mathbf{x},\omega) = y_0 & \text{ in } \mathcal{D}\times\Omega, \end{aligned}\right. \tag{58}$$

where the random control function satisfies

$$u \in L^2(0,T;\mathcal{D})\otimes L^2(\Omega) \text{ a.e.,}$$

and, as before, $a(\mathbf{x},\omega)$ is assumed to be uniformly positive in $\mathcal{D}\times\Omega$. We note here that the time-dependence of this problem introduces an additional degree of freedom which makes the system matrix here (a lot) larger than the system matrix in the steady-state case.

We use the trapezoidal rule for temporal discretization (as was done for deterministic problems in, e.g., [25]) and SGFEM in the spatial and the stochastic domains to get the discrete objective function

$$\mathcal{J}(t,y,u) = \frac{\tau}{2}(\mathbf{y}-\bar{\mathbf{y}})^T\mathcal{M}_a(\mathbf{y}-\bar{\mathbf{y}}) + \frac{\tau\alpha}{2}\mathbf{y}^T\mathcal{M}_b\mathbf{y} + \frac{\tau\beta}{2}\mathbf{u}^T\mathcal{M}_2\mathbf{u}, \tag{59}$$

where τ represents the time step size, and

$$\begin{cases} \mathcal{M}_a = \text{blkdiag}\left(\frac{1}{2}\mathcal{M}, \mathcal{M}, \dots, \mathcal{M}, \frac{1}{2}\mathcal{M}\right), \\ \\ \mathcal{M}_b = \text{blkdiag}\left(\frac{1}{2}\mathcal{M}_t, \mathcal{M}_t, \dots, \mathcal{M}_t, \frac{1}{2}\mathcal{M}_t\right), \end{cases} \tag{60}$$

with $\mathcal{M}$ and $\mathcal{M}_t$ as defined in (25) and (28), respectively. Note that $\mathcal{M}_2 = \mathcal{M}_a$. Here, denoting the number of time steps by n_t, we also note that

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_{n_t} \end{bmatrix}, \quad \bar{\mathbf{y}} = \begin{bmatrix} \bar{\mathbf{y}}_1 \\ \vdots \\ \bar{\mathbf{y}}_{n_t} \end{bmatrix}, \text{ and } \mathbf{u} = \begin{bmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_{n_t} \end{bmatrix},$$

with $\mathbf{y}_i, \bar{\mathbf{y}}_i, \mathbf{u}_i \in \mathbb{R}^{JP\times 1}$, $i = 1, \dots, n_t$.

For an all-at-once discretization of the state equation (58), we use the implicit Euler method together with SGFEM to get

$$\mathcal{K}_t\mathbf{y} - \tau\mathcal{N}\mathbf{u} = \mathbf{d},$$

where

$$\mathcal{K}_t = \begin{bmatrix} \mathcal{L} & & & \\ -\mathcal{M} & \mathcal{L} & & \\ & \ddots & \ddots & \\ & & -\mathcal{M} & \mathcal{L} \end{bmatrix}, \ \mathcal{N} = \begin{bmatrix} \mathcal{M} & & & \\ & \mathcal{M} & & \\ & & \ddots & \\ & & & \mathcal{M} \end{bmatrix}, \ \mathbf{d} = \begin{bmatrix} \mathcal{M}\mathbf{y}_0 \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

where $\mathcal{L} := G_0 \otimes (M + \tau K_0) + \tau \sum_{i=1}^{N} G_i \otimes K_i$. Observe that the matrix $\mathcal{K}_t$ in this case is not symmetric, unlike the matrix $\mathcal{K}$ in the stationary case.

Applying first order conditions to the Lagrangian functional for this constrained optimization problem yields

$$\begin{bmatrix} \tau\mathcal{M}_1 & 0 & -\mathcal{K}_t^T \\ 0 & \beta\tau\mathcal{M}_2 & \tau\mathcal{N}^T \\ -\mathcal{K}_t & \tau\mathcal{N} & 0 \end{bmatrix} \begin{bmatrix} \mathbf{y} \\ \mathbf{u} \\ \mathbf{f} \end{bmatrix} = \begin{bmatrix} \tau\mathcal{M}_a\bar{\mathbf{y}} \\ \mathbf{0} \\ \mathbf{d} \end{bmatrix}, \tag{61}$$

where, from (60), (30), and (25),

$$\begin{aligned} \mathcal{M}_1 &= \mathcal{M}_a + \alpha\mathcal{M}_b \\ &= (D \otimes \mathcal{M}) + \alpha(D \otimes \mathcal{M}_t) \\ &= D \otimes (\mathcal{M} + \alpha\mathcal{M}_t) \\ &= D \otimes G_\alpha \otimes M = D \otimes \mathcal{M}_\alpha, \end{aligned} \tag{62}$$

with G_α as defined in (31), and

$$D = \operatorname{diag}\left(\frac{1}{2}, 1 \ldots, 1, \frac{1}{2}\right) \in \mathbb{R}^{n_t \times n_t}. \tag{63}$$

We note here that

$$\mathcal{K}_t = (I_{n_t} \otimes \mathcal{L}) + (C \otimes \mathcal{M}) = I_{n_t} \otimes \left[\sum_{i=0}^{N} G_i \otimes \widehat{K}_i\right] + (C \otimes G_0 \otimes M), \tag{64}$$

where $\widehat{K}_0 = M + \tau K_0$, $\widehat{K}_i = \tau K_i$, $i = 1, \ldots, N$. The matrix $C \in \mathbb{R}^{n_t \times n_t}$ comes from the implicit Euler discretization and is given by

$$C = \begin{bmatrix} 0 & & & \\ -1 & 0 & & \\ & \ddots & \ddots & \\ & & -1 & 0 \end{bmatrix},$$

and I_{n_t} is an identity matrix of dimension n_t. The use of other temporal discretizations is, of course, possible. The Crank–Nicolson scheme, for instance, can be written in a similar way. Moreover,

$$\mathcal{N} = I_{n_t} \otimes G_0 \otimes M, \quad \mathcal{M}_2 = D \otimes G_0 \otimes M. \tag{65}$$

Hence, each of the block matrices $\mathcal{K}_t, \mathcal{N}, \mathcal{M}_1$, and $\mathcal{M}_2$ belongs to $\mathbb{R}^{JPn_t \times JPn_t}$, since $G_i \in \mathbb{R}^{P \times P}$, $i = 0, \ldots, P-1$, and $M, K_i \in \mathbb{R}^{J \times J}$, $i = 0, \ldots, N$. So, the overall coefficient matrix in (61) is of dimension $3JPn_t \times 3JPn_t$.

As can be seen from (64), for instance, the time-dependent problem leads to an additional Kronecker product. Indeed, although the low-rank solver presented in the stationary case reduces storage problems in large-scale simulations, the low-rank factors become infeasible in higher dimensions. Further data compression can, fortunately, be achieved with more advanced high-dimensional tensor product decompositions. Together with preconditioned MINRES, we henceforth solve the linear system discussed in this section using an elegant and robust tensor format called the *tensor train* (TT) format which was introduced in [17]. To this end, we proceed next to section 4.1 to give an overview of the TT-format.

4.1. Solving the optimality systems from the unsteady problem. First, we recall that a tensor $\mathbf{y} := \mathbf{y}(i_1, \ldots, i_d)$, $i_k = 1, \ldots, n_k$, is an $n_1 \times n_2 \times \cdots \times n_d$ multidimensional array, where the integers $n_1, n_2, \ldots, n_d$ are called the mode sizes and d is the order of $\mathbf{y}$. The tensor $\mathbf{y}$ admits a TT-decomposition or TT- format [17, 6] if it can be expressed as

$$\mathbf{y}(i_1, \ldots, i_d) = \mathbf{y}_1(i_1)\mathbf{y}_2(i_2) \cdots \mathbf{y}_d(i_d),$$

where $\mathbf{y}_k(i_k)$ is an $r_{k-1} \times r_k$ matrix for each fixed i_k, $1 \le i_k \le n_k$. Moreover, the numbers r_k are called the TT-ranks, whereas $\mathbf{y}_k(i_k)$ are the cores of the decomposition. More precisely, $\mathbf{y}_k(i_k)$ is a three-dimensional array, and it can essentially be treated as an $r_{k-1} \times n_k \times r_k$ array with elements $\mathbf{y}_k(\alpha_{k-1}, i_k, \alpha_k) = \mathbf{y}^{(k)}_{\alpha_{k-1},\alpha_k}(i_k)$. Here, the boundary conditions $r_0 = r_d = 1$ are imposed on the decomposition to make the matrix-by-matrix products a scalar. The decomposition can be expressed in index form as

$$\mathbf{y}(i_1,\dots,i_d)=\sum_{\alpha_1\dots\alpha_{d-1}=1}^{r_1\dots r_{d-1}}\mathbf{y}_1(\alpha_0,i_1,\alpha_1)\mathbf{y}_2(\alpha_1,i_2,\alpha_2)\cdots\mathbf{y}_{d-1}(\alpha_{d-1},i_d,\alpha_d),\tag{66}$$

where $\alpha_0=\alpha_d=1$. It turns out that TT-decomposition yields a low-rank format for tensors as it is derived by a repeated application of low-rank approximation [17]. To see this, let

$$\overline{i_2\cdots i_d}=i_2+(i_3-1)n_2+\cdots+(i_d-1)n_2n_3\cdots n_{d-1}.\tag{67}$$

Then, by *regrouping* of indices, one can rewrite $\mathbf{y}$ as a matrix $Y_1\in\mathbb{R}^{n_1\times n_2\cdots n_d}$ with $Y_1(i_1,\overline{i_2\cdots i_d})=\mathbf{y}(i_1,\dots,i_d)$. Thus, applying a low-rank SVD to the matrix Y_1 yields

$$Y_1\approx U_1\Sigma_1V_1^T,\quad U_1\in\mathbb{R}^{n_1\times r_1},\ \ V_1\in\mathbb{R}^{n_2\cdots n_d\times r_1}.$$

The first factor U_1 is of moderate dimension and can be stored as $\mathbf{y}^{(1)}_{\alpha_1}(i_1)=U_1(i_1,\alpha_1)$, where $\alpha_1=1,\dots,r_1$ and $i_1=1,\dots,n_1$. The remaining matrix $\Sigma_1V_1^T$ depends on α_1 and $i_2\cdots i_d$. Next, we regroup these indices as follows:

$$Y_2(\overline{\alpha_1 i_2},\overline{i_3\cdots i_d})=\Sigma_1(\alpha_1,\alpha_1)V_1^T(\alpha_1,\overline{i_2\cdots i_d}),$$

and we compute the next SVD:

$$Y_2\approx U_2\Sigma_2V_2^T,\quad U_2\in\mathbb{R}^{r_1n_2\times r_2},\ \ V_2\in\mathbb{R}^{n_3\cdots n_d\times r_2}.$$

Again, U_2 can be reshaped to a three-dimensional tensor $\mathbf{y}^{(2)}_{\alpha_1,\alpha_2}(i_2)=U_2(\overline{\alpha_1 i_2},\alpha_2)$ of moderate size, and the decomposition also applied to $\Sigma_2V_2^T$. Proceeding in this manner, one eventually obtains the TT-format:

$$\mathbf{y}(i_1,\dots,i_d)=\sum_{\alpha_1\dots\alpha_{d-1}=1}^{r_1\dots r_{d-1}}\mathbf{y}^{(1)}_{\alpha_1}(i_1)\mathbf{y}^{(2)}_{\alpha_1,\alpha_2}(i_2)\cdots\mathbf{y}^{(d-1)}_{\alpha_{d-2},\alpha_{d-1}}(i_{d-1})\mathbf{y}^{(d)}_{\alpha_{d-1}}(i_d),\tag{68}$$

with the total storage of at most dnr^2 memory cells, where $r_k\le r,\ n_k\le n$. In particular, if r is small, then this requirement is much smaller than the storage of the full array, n^d. A similar construction can be made for discretized operators in high dimensions. To this end, consider a matrix $A=A(\overline{i_1\cdots i_d},\overline{j_1\cdots j_d})\in\mathbb{R}^{(n_1\cdots n_d)\times(n_1\cdots n_d)}$. We decompose A as follows:

$$A(\overline{i_1\cdots i_d},\overline{j_1\cdots j_d})\approx\sum_{\beta_1\dots\beta_{d-1}=1}^{R_1\dots R_{d-1}}\mathbf{A}^{(1)}_{\beta_1}(i_1,j_1)\mathbf{A}^{(2)}_{\beta_1,\beta_2}(i_2,j_2)\cdots\mathbf{A}^{(d)}_{\beta_{d-1}}(i_d,j_d),\tag{69}$$

which is consistent with the Kronecker product $A=A^{(1)}\otimes A^{(2)}$ in the case $d=2$ and $R_1=1$, and allows a natural multiplication with (68) returning the result in the same format.

As pointed out in [6], the TT-format is stable in the sense that one can always find the best approximation of tensors computed via a sequence of QR and SVD decompositions of auxiliary matrices. The TT-decomposition algorithm is implemented in the TT-Toolbox [18] and comes with a number of basic linear algebra operations, such as addition, subtraction, and matrix-by-vector product. Unfortunately, these operations lead to prohibitive increase in the TT-ranks. Thus, one necessarily has to truncate (or round) the resulting tensor after implementing each of the operations. This enhances the efficiency of the method when used with any standard iterative method such as MINRES. We point out that although solving the KKT system in the TT-format (and, in general, with low-rank solvers) introduces further error in the simulation due to the low-rank truncations, the relative tolerance of the truncation operator can be so tightened that the error will become negligible. This is investigated in [3] for a low-rank conjugate gradient iterative solver; see also, e.g., [6, 7] for TT iterative solvers.

We remark here that there are, of course, other tensor formats, such as canonical, hierarchical, and Tucker formats, which could be used to represent tensors [11] and hence solve our linear systems. However, our choice of TT-format (or TT-Toolbox) is due to its relative elegance and convenience in implementation. The details of its implementation are found in [18]. A comprehensive overview of low-rank tensor decompositions can be found in [11] and the references therein. In our numerical experiments, we use preconditioned MINRES, together with the TT-Toolbox, to solve the linear system (61).

4.2. Preconditioning the optimality system. As in the case of the optimality system associated with the stationary model problem, we need a good preconditioner to solve (61). To this end, we will proceed as before and rewrite the saddle-point system (61) as

$$A = \begin{bmatrix} \tau\mathcal{M}_1 & 0 \\ 0 & \tau\beta\mathcal{M}_2 \end{bmatrix}, \quad B = [-\mathcal{K}_t \;\; \tau\mathcal{N}] \tag{70}$$

in the notation of (32). Again, we are interested in a block-diagonal preconditioner to approximate the solution to (61). More precisely, we seek a preconditioner of the form

$$\widehat{\mathcal{P}} = \begin{bmatrix} A_1 & & \\ & A_2 & \\ & & S_2 \end{bmatrix},$$

where the blocks $A_1 \approx \tau D \otimes G_\alpha \otimes M$ and $A_2 \approx \tau\beta D \otimes G_0 \otimes M$, and as we noted before, both approximations could be accomplished by applying a Chebyshev semi-iteration on the mass matrix M in the blocks. The matrices D, G_0, and G_α are easy to invert since they are diagonal matrices. Moreover, S_2 is an approximation to the (negative) Schur complement $S_t = BA^{-1}B^T$, that is,

$$S_t := \frac{1}{\tau}\mathcal{K}_t\mathcal{M}_1^{-1}\mathcal{K}_t^T + \frac{\tau}{\beta}\mathcal{N}\mathcal{M}_2^{-1}\mathcal{N}^T. \tag{71}$$

As in the time-independent case, we consider the following approximation of the Schur complement:

$$S_2 = \frac{1}{\tau}\left(\mathcal{K}_t + \widehat{\mathcal{M}}_u\right)\mathcal{M}_1^{-1}\left(\mathcal{K}_t + \widehat{\mathcal{M}}_u\right)^T, \tag{72}$$

where $\widehat{\mathcal{M}}_u$ is again determined via the "term-matching" procedure so that both the first and second terms in S_t and S_2 are matched, but the cross terms in S_2 are ignored; that is, we have

$$\widehat{\mathcal{M}}_u\mathcal{M}_1^{-1}\widehat{\mathcal{M}}_u = \frac{\tau^2}{\beta}\mathcal{N}\mathcal{M}_2^{-1}\mathcal{N}^T,$$

from which we deduce that $\widehat{\mathcal{M}}_u = \gamma\mathcal{N}$, with $\gamma = \tau\sqrt{\frac{1+\alpha}{\beta}}$, by using arguments similar to those used before, so that

$$S_2 = \frac{1}{\tau}\underbrace{\left(\mathcal{K}_t + \tau\sqrt{\frac{1+\alpha}{\beta}}\mathcal{N}\right)}_{:=\widehat{\mathcal{Z}}}\mathcal{M}_1^{-1}\left(\mathcal{K}_t + \tau\sqrt{\frac{1+\alpha}{\beta}}\mathcal{N}\right)^T. \tag{73}$$

Moreover, as in the stationary case, we have the following result regarding the eigenvalues of the preconditioned Schur complement $S_2^{-1}S_t$.

THEOREM 6. *Let $\alpha \in [0, +\infty)$. Then the eigenvalues of $S_2^{-1}S_t$ satisfy*

$$\lambda(S_2^{-1}S_t) \subset \left[\frac{1}{2(1+\alpha)}, 1\right) \quad \forall\alpha < \left(\frac{\sqrt{\kappa(\mathcal{K})}+1}{\sqrt{\kappa(\mathcal{K})}-1}\right)^2 - 1, \tag{74}$$

where $\mathcal{K} = \sum_{i=0}^N G_i \otimes K_i$.

Proof. Let $I_{n_t} := I$, and observe first from (64) that we can rewrite $\mathcal{K}_t$ as

$$\mathcal{K}_t = (I + C) \otimes (G_0 \otimes M) + I \otimes \tau\sum_{i=0}^{N}(G_i \otimes K_i) = J_0 \otimes \mathcal{M} + \tau I \otimes \mathcal{K}, \tag{75}$$

where

$$J_0 = I + C = \begin{bmatrix} 1 & & & \\ -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{bmatrix},$$

and $\mathcal{K}$, the coefficient matrix associated with the stationary forward problem, is positive definite. Now, using (42), (44), (62), (65), we see that

$$\begin{aligned} \mathcal{M}_1 &= D \otimes \mathcal{M}_\alpha \\ &= D \otimes \mathcal{M}\mathcal{E}_\alpha \\ &= (D \otimes \mathcal{M})(I \otimes \mathcal{E}_\alpha) \\ &= \mathcal{M}_2\mathcal{F}_\alpha = \mathcal{F}_\alpha\mathcal{M}_2, \end{aligned} \tag{76}$$

where $\mathcal{F}_\alpha = I \otimes \mathcal{E}_\alpha$. Next, define the matrix $\mathcal{X}$ by

$$\begin{aligned} \mathcal{X} &:= (D \otimes I)\mathcal{M}_2^{-1/2}\mathcal{K}_t\mathcal{M}_2^{-1/2} \\ &= D^{1/2}J_0D^{-1/2} \otimes I + \tau I \otimes \mathcal{M}^{-1/2}\mathcal{K}\mathcal{M}^{-1/2}. \end{aligned} \tag{77}$$

Note then that $\mathcal{X}$ is similar to $J_0 \otimes I + \tau I \otimes \mathcal{M}^{-1}\mathcal{K} = (D \otimes I)\mathcal{M}_2^{-1}\mathcal{K}_t$. Moreover, since

$$S_2^{-1}S_t \sim (D \otimes I)^{-1}\mathcal{M}_2^{1/2}S_2^{-1}S_t\mathcal{M}_2^{-1/2}(D \otimes I), \tag{78}$$

we see, from (71), (73), (76), (77), and (78), that

$$\begin{aligned} &S_2^{-1}S_t \sim \left[(D \otimes I)\mathcal{M}_2^{-1/2}S_2\mathcal{M}_2^{-1/2}(D \otimes I)\right]^{-1}\left[(D \otimes I)\mathcal{M}_2^{-1/2}S_t\mathcal{M}_2^{-1/2}(D \otimes I)\right] \\ &= \left[\beta\mathcal{X}\mathcal{F}_\alpha^{-1}\mathcal{X}^T + \tau^2(1+\alpha)\mathcal{F}_\alpha^{-1} + \tau\sqrt{\beta(1+\alpha)}\left(\mathcal{X}\mathcal{F}_\alpha^{-1} + \mathcal{F}_\alpha^{-1}\mathcal{X}^T\right)\right]^{-1}\left(\beta\mathcal{X}\mathcal{F}_\alpha^{-1}\mathcal{X}^T + \tau^2 I\right). \end{aligned}$$

Now, consider the Raleigh quotient

$$R(x) := \frac{x^T \left[\beta \mathcal{X}\mathcal{F}_\alpha^{-1}\mathcal{X}^T + \tau^2 I\right] x}{x^T \left[\beta \mathcal{X}\mathcal{F}_\alpha^{-1}\mathcal{X}^T + \tau^2(1+\alpha)\mathcal{F}_\alpha^{-1} + \tau\sqrt{\beta(1+\alpha)}\left(\mathcal{X}\mathcal{F}_\alpha^{-1} + \mathcal{F}_\alpha^{-1}\mathcal{X}^T\right)\right] x}.$$

But then

$$\mathcal{X}\mathcal{F}_\alpha^{-1} + \mathcal{F}_\alpha^{-1}\mathcal{X}^T = D^{1/2}(J_0 + J_0^T)D^{-1/2} \otimes \mathcal{E}_\alpha^{-1} + \tau I \otimes \mathcal{M}^{-1/2}(\mathcal{K}\mathcal{E}_\alpha^{-1} + \mathcal{E}_\alpha^{-1}\mathcal{K})\mathcal{M}^{-1/2}.$$

Since the matrix $D^{1/2}(J_0 + J_0^T)D^{-1/2}$ is the sum of two positive definite matrices, it is therefore positive definite. In addition, by Proposition 3, one gets

$$\mathcal{K}\mathcal{E}_\alpha^{-1} + \mathcal{E}_\alpha^{-1}\mathcal{K} \succ 0 \quad \forall \alpha < \left(\frac{\sqrt{\kappa(\mathcal{K})}+1}{\sqrt{\kappa(\mathcal{K})}-1}\right)^2 - 1.$$

It follows that $\mathcal{X}\mathcal{F}_\alpha^{-1} + \mathcal{F}_\alpha^{-1}\mathcal{X}^T \succ 0$. Furthermore, it is easy to check that both $\mathcal{X}\mathcal{F}_\alpha^{-1}\mathcal{X}^T$ and $\mathcal{F}_\alpha^{-1}$ are also positive definite. Hence, using (43), we obtain

$$R(x) \le \frac{x^T \left[\beta \mathcal{X}\mathcal{F}_\alpha^{-1}\mathcal{X} + \tau^2(1+\alpha)\mathcal{F}_\alpha^{-1}\right] x}{x^T \left[\beta \mathcal{X}\mathcal{F}_\alpha^{-1}\mathcal{X}^T + \tau^2(1+\alpha)\mathcal{F}_\alpha^{-1} + \tau\sqrt{\beta(1+\alpha)}\left(\mathcal{X}\mathcal{F}_\alpha^{-1} + \mathcal{F}_\alpha^{-1}\mathcal{X}^T\right)\right] x} < 1,$$

from which we deduce that $\lambda_{\max} := \max R(x) < 1$.

The proof of $\lambda_{\min} := \min R(x) \ge \frac{1}{2(1+\alpha)}$ follows arguments similar to those used in the second part of the proof of Theorem 4, with $\mathcal{C}$ and $\mathcal{E}_\alpha$ replaced, respectively, by $\mathcal{X}$ and $\mathcal{F}_\alpha$. □

Remark 7. Note that, using arguments similar to those in Theorem 5, we can as well characterize the spectrum of the preconditioned KKT system in the unsteady case if we define $\mathcal{A}$ as the global coefficient matrix and $\mathcal{P}_0$ as

$$\mathcal{P}_0 = \begin{bmatrix} A & 0 \\ 0 & S_2 \end{bmatrix},$$

where A and S_2 are given by (70) and (73), respectively.

It turns out that if we specifically use Legendre polynomials and piecewise linear (or bilinear) approximation in the SGFEM discretization of the SOCPs considered herein, then the following result proved by Powell and Elman enables us to further bound the parameter α in Theorem 6 above.

PROPOSITION 8 (see [21, Lemma 3.7]). *Let the matrices G_k in (26) be defined using normalized Legendre polynomials in uniform random variables on a bounded symmetric interval $[-\nu, \nu]$, and suppose that piecewise linear (or bilinear) approximation is used for the spatial discretization, on quasi-uniform meshes. Let (λ_i, φ_i) be the eigenpairs associated with the N-term KLE of the random field a_N. Then $\kappa(\mathcal{K}) \le \Phi/\Psi$, where $\Phi = c_2\mathbb{E}(a) + \eta$ and $\Psi = c_1 h^2\mathbb{E}(a) - \eta$, with*

$$\eta = c_2\sigma_a C_{n+1}^{\max} \sum_{i=1}^{N} \sqrt{\lambda_i}||\varphi_i(\mathbf{x})||_\infty,$$

where $C_{n+1}^{\max}$ is the maximal root of the Legendre polynomial of degree $n+1$, σ_a is the standard deviation of the random field a, h is the spatial discretization parameter, and c_1 and c_2 are constants independent of h, N, and n.

We can now state the following result.

COROLLARY 9. *Let $\alpha \in [0, +\infty)$. Then the spectrum of $S_2^{-1}S_t$ satisfies*

$$\lambda(S_2^{-1}S_t) \subset \left[\frac{1}{2(1+\alpha)}, 1\right) \quad \forall \alpha < \tilde{\mu}^2 - 1, \tag{79}$$

where $\tilde{\mu} = \frac{1+p+2\sqrt{p}}{p-1}$, $p \neq 1$, and $p = \sqrt{\Phi/\Psi}$, with Φ and Ψ as defined in Proposition 8.

Proof. The proof is a direct consequence of Proposition 8 and Theorem 6. ∎

Next, as the approximation S_2 is impractical, we proceed next to derive its practical version. Now, observe from (64), (65), and (73) that

$$\begin{aligned}
\widehat{\mathcal{Z}} &:= \mathcal{K}_t + \gamma\mathcal{N} \\
&= [(I_{n_t} \otimes \mathcal{L}) + (C \otimes \mathcal{M})] + \gamma(I_{n_t} \otimes \mathcal{M}) \\
&= I_{n_t} \otimes \left[\left(G_0 \otimes (M + \tau K_0) + \tau \sum_{i=1}^{N} G_i \otimes K_i\right) + \gamma(G_0 \otimes M)\right] + (C \otimes \mathcal{M}) \\
&= I_{n_t} \otimes \left[G_0 \otimes ((1+\gamma)\, M + \tau K_0) + \tau \sum_{i=1}^{N} G_i \otimes K_i\right] + (C \otimes \mathcal{M}) \\
&= I_{n_t} \otimes \left[G_0 \otimes \mathcal{Y} + \tau \sum_{i=1}^{N} G_i \otimes K_i\right] + (C \otimes G_0 \otimes M), \qquad (80)
\end{aligned}$$

where $\mathcal{Y} = (1+\gamma)\, M + \tau K_0$. Hence, using arguments similar to those in section 3.1 we can now approximate $\widehat{\mathcal{Z}}$ using

$$\widehat{\mathcal{Z}}_0 := I_{n_t} \otimes G_0 \otimes \mathcal{Y}. \tag{81}$$

In practice, we thus approximate S_2 by applying a cheap multigrid process to $\mathcal{Y}$ in each of the diagonal blocks of $\widehat{\mathcal{Z}}_0$ and $\widehat{\mathcal{Z}}_0^T$. The expression (81) is admittedly not the best possible approximation to S_2 due essentially to the same reasons provided in the case of S_1 in section 3.1. In addition, the absence of the term $C \otimes G_0 \otimes M$ in $\widehat{\mathcal{Z}}_0$ would likely impact negatively on the performance of $\widehat{\mathcal{Z}}_0$. Again, solves with $\widehat{\mathcal{Z}}$ via the inexact Uzawa algorithm could substantially mitigate these shortcomings.

5. Numerical experiments.

In this section, we present some numerical results. The numerical experiments were performed on a Linux machine with 80 GB RAM using MATLAB 7.14 together with a MATLAB version of the AMG code HSL MI20 [4]. We implement each of the mean-based preconditioners $\mathcal{Z}_0$ and $\widehat{\mathcal{Z}}_0$ as given, respectively, by (51) and (81) using one V-cycle of AMG with symmetric Gauss–Seidel (SGS) smoothing to approximately invert $\tilde{K}_0$. We remark here that we apply the method as a black-box in each experiment, and the setup of the approximation to $\tilde{K}_0$ only needs to be performed once. Unless otherwise stated, in all the simulations, MINRES is terminated when the relative residual error, measured in the Euclidean norm, is reduced to $tol = 10^{-5}$. Note that tol should be chosen such that the truncation tolerance $\epsilon \leq tol$; otherwise, one would be essentially iterating on the "noise" from the low-rank truncations, as it were. In particular, we have chosen herein $\epsilon = 10^{-8}$.

For our simulations, the random input a is characterized by the covariance function

$$C_a(\mathbf{x}, \mathbf{y}) = \sigma_a^2 \exp\left(-\frac{|x_1 - y_1|}{\ell_1} - \frac{|x_2 - y_2|}{\ell_2}\right) \quad \forall (\mathbf{x}, \mathbf{y}) \in [-1, 1]^2,$$

Table 1

Simulation results showing the total number of iterations from low-rank preconditioned MINRES and the total CPU times (in seconds) using the mean-based preconditioner in (51) *with* $\alpha = 1$, $\beta \in \{10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$, $\sigma_a = 0.1$, *and selected spatial* (J) *and stochastic* (P) *degrees of freedom.*

LR-MINRES	# iter (t)	# iter (t)	# iter (t)	# iter (t)
J P	481	1985	8065	32513
$\beta = 10^{-2}$				
20	25 (32.8)	25 (115.4)	27 (250.5)	29 (736.6)
84	25 (119.7)	27 (380.4)	27 (582.2)	29 (1619.6)
210	25 (141.6)	27 (392.8)	27 (594.69)	29 (1673.9)
$\beta = 10^{-3}$				
20	21 (25.7)	21 (113.8)	25 (260.9)	25 (666.8)
84	21 (128.9)	23 (363.7)	25 (607.6)	25 (1438.1)
210	21 (145.6)	23 (385.5)	25 (600.8)	25 (1471.8)
$\beta = 10^{-4}$				
20	19 (8.2)	21 (17.4)	23 (67.4)	23 (618.3)
84	19 (18.8)	21 (42.5)	23 (229.7)	23 (1313.7)
210	19 (19.6)	21 (44.9)	23 (276.9)	23 (1450.0)
$\beta = 10^{-5}$				
20	17 (19.6)	17 (84.8)	21 (223.7)	21 (578.3)
84	17 (99.9)	19 (306.4)	21 (520.7)	21 (1217.2)
210	17 (115.4)	19 (313.63)	21 (515.6)	23 (1322.6)

where the correlation lengths $\ell_1 = \ell_2 = 1$ and the mean $\mathbb{E}[a] = 1$. The forward problem has been extensively studied in, for instance, [21]. The eigenpairs (λ_j, φ_j) of the KLE of the random field a are given explicitly in [10].

Next, we investigate the behavior of the solvers (low-rank MINRES and TT-MINRES) for different values of the stochastic discretization parameters J, P, σ_a, as well as α and β. Moreover, we choose $\xi = \{\xi_1, \ldots, \xi_N\}$ such that $\xi_j \sim \mathcal{U}[-1, 1]$, and $\{\psi_j\}$ are N-dimensional Legendre polynomials with support in $[-1, 1]^N$. The spatial discretization uses $\mathbf{Q}_1$ spectral elements. In the considered unsteady SOCP example (that is, in section 4), the resulting linear systems were solved for time $T = 1$. Moreover, our target (or desired state) in both models is the stochastic solution of the forward model with right-hand side 1 and zero Dirichlet boundary conditions.[1]

Tables 1, 3, 4, and 5 show the results from the low-rank preconditioned MINRES for the model constrained by a steady-state diffusion equation. In Table 2 we give the total dimensions of the KKT systems in (29) for various discretization parameters used to obtain the results in Tables 1. Herein, h is the spatial mesh size and the dimensions range between 28,000 and 20 million. The results in Tables 1, 4, and 5 were obtained with $\alpha = 1$, whereas those in Table 3 were computed with $\alpha = 0$. We have solved the linear systems using our proposed block-diagonal preconditioner, together with the approximation S_1 for the Schur complement S. To compare their practical performances, we use both the mean-based preconditioner in (51) (denoted henceforth by MBP) and the inexact Uzawa method in Algorithm 1 (with just 4 iterations, and denoted by IUM) for approximating $\mathcal{Z}$ in S_1.

We observe first that Table 1 confirms our theoretical prediction that with a relatively low variance (here $\sigma_a = 0.1$), our proposed block-diagonal preconditioner, when used together with MBP, is robust with respect to the discretization parameters.

[1] Note that this is not an "inverse crime" as the right-hand side of the forward model used is deterministic, unlike in the state equation.

Table 2
Dimension of global coefficient matrix $\mathcal{A}$ in (29); *here dim*($\mathcal{A}$) = $3JP$.

$J(h)$ $P(N,n)$	$481\left(\frac{1}{2^4}\right)$	$1985\left(\frac{1}{2^5}\right)$	$8065\left(\frac{1}{2^6}\right)$	$32513\left(\frac{1}{2^7}\right)$
20 ($N=3, n=3$)	28,860	119,100	483,900	1,950,780
84 ($N=6, n=3$)	121,212	500,220	2,032,380	8,193,276
210 ($N=6, n=4$)	303,030	1,250,550	5,080,950	20,483,190

Table 3
Simulation results using the mean-based preconditioner in (51) *with* $\sigma_a = 0.1$, $\alpha = 0$, $\beta \in \{10^{-3}, 10^{-4}, 10^{-5}\}$, *and* $J = 1985$ $(h = \frac{1}{2^5})$.

LR-MINRES	# iter (t)	# iter (t)	# iter (t)
P	20	84	210
dim($\mathcal{A}$) = $3JP$	119,100	500,220	1,250,550
$\beta = 10^{-3}$	19 (96.4)	21 (336.0)	21 (347.93)
$\beta = 10^{-4}$	17 (86.3)	19 (302.6)	19 (305.64)
$\beta = 10^{-5}$	15 (77.4)	17 (273.6)	17 (283.24)

Furthermore, Table 5 shows that the preconditioner performs relatively better with IUM than it does with MBP, especially as the standard deviation σ_a increases from 1% to 40%. Indeed, the iterations are clearly indicative of benign dependence of IUM on σ_a, unlike MBP. In general, the iterations obtained with MBP took slightly less CPU time, though. So, these results generally suggest that IUM should be preferred to MBP when dealing with higher fluctuations in the random input a. We remark here, though, that for $\sigma_a > 0.5$, we can no longer guarantee the positive-definiteness of the matrix $\mathcal{K}$ corresponding to the forward problem [21].

Observe from Tables 1 and 5 that the timings for $P = 84$ and $P = 210$ are nearly constant but much higher than those for $P = 20$. Our extensive numerical experiments revealed that the timings, in general, have a strong dependence on the number of random variables N (all things being equal), which in turn determines P; see also [3]. Now, recall that for $P = 20$ the (stochastic) matrices $G_k \in \mathbb{R}^{P\times P}$ used in the simulations were obtained with only $N = 3$ random variables, whereas the other two cases were obtained with $N = 6$ random variables (cf. Table 2). So, we believe, in particular, that the timings which are roughly constant for simulations with $P = 84$ and $P = 210$ are due to the fact that both cases were computed with exactly the same value of N.

We have reported in Table 4 the values of the tracking term and the cost functional for $\alpha = 1$ and $\sigma_a = 0.1$. As expected, the tracking term gets smaller and smaller as the regularization parameter β decreases, and the cost functional also decreases accordingly converging, respectively, to 1.2×10^{-4} and 2.5×10^{-4}.

Next, we present in Table 6 our results for the unsteady diffusion-constrained model as discussed in section 4. Here, for $\alpha \in \{0, 1\}$ and different values of β, we present the outputs of our simulations showing the total CPU times and the total number of iterations from preconditioned TT-MINRES. Also, DoF $= J \cdot P \cdot n_t$ is the size of each of the 9 block matrices in KKT matrix $\mathcal{A}$; that is, $\mathcal{A}$ is of dimension 3DoF. In particular, we have done the computations with $J = 1985$ $(h = \frac{1}{2^5})$, $P = 56$ $(N = 5, n = 3)$, $\sigma_a = 0.1$, and different numbers of total time steps n_t.

As in the steady-state case, we see from Table 6 that TT-MINRES, when used together with our mean-based preconditioner as given by (81), is quite robust, but in general yields fewer iterations for $\alpha = 0$ than for $\alpha = 1$. We remark here that we used

Table 4

Tracking term and the cost functional in the steady-state model using the mean-based preconditioner $\mathcal{Z}_0$ in (51) *for different values of β and with $\alpha = 1,\ \sigma_a = 0.1,\ J = 1985\ (h = \frac{1}{2^5}),\ P = 84\ (N = 6, n = 3)$.*

β	10^{-2}	10^{-4}	10^{-6}	10^{-10}
$\lvert\lvert y-\bar{y}\rvert\rvert^2_{L^2(\mathcal{D})\otimes L^2_\rho(\Gamma)}$	5.1×10^{-3}	1.8×10^{-4}	1.2×10^{-4}	1.2×10^{-4}
$\mathcal{J}(y,u)$	1.4×10^{-2}	4.2×10^{-4}	2.5×10^{-4}	2.5×10^{-4}

Table 5

Simulation results comparing mean-based preconditioning (MBP) and the inexact Uzawa method (IUM) in approximating S_1 in low-rank preconditioned MINRES with $\alpha = 1,\ \beta = 10^{-4}$.

LR-MINRES	# iter (t)	# iter (t)	# iter (t)	# iter (t)
$J(h)$ P	$481\left(h=\frac{1}{2^4}\right)$	$1985\left(h=\frac{1}{2^5}\right)$	$8065\left(h=\frac{1}{2^6}\right)$	$32513\left(h=\frac{1}{2^7}\right)$
$\sigma_a = 0.01$ with MBP				
20	17 (7.4)	19 (16.7)	19 (53.4)	21 (544.8)
84	17 (17.0)	19 (39.0)	19 (190.0)	21 (1190.0)
210	17 (18.4)	19 (40.4)	19 (470.0)	21 (1230.2)
$\sigma_a = 0.1$ with MBP				
20	19 (8.2)	21 (17.4)	23 (67.4)	23 (618.3)
84	19 (18.8)	21 (42.5)	23 (229.7)	23 (1313.7)
210	19 (19.6)	21 (44.9)	23 (276.9)	23 (1450.0)
$\sigma_a = 0.4$ with MBP				
20	33 (13.8)	37 (28.0)	41 (115.3)	43 (1049.8)
84	35 (33.8)	41 (84.5)	45 (447.0)	47 (2610.4)
210	41 (41.9)	47 (98.4)	47 (782.3)	55 (3161.1)
$\sigma_a = 0.01$ with IUM				
20	15 (13.5)	15 (20.5)	17 (82.2)	19 (1142.4)
84	15 (31.8)	15 (57.6)	17 (332.5)	19 (2117.4)
210	15 (35.0)	15 (61.9)	17 (314.9)	19 (2777.2)
$\sigma_a = 0.1$ with IUM				
20	15 (14.2)	17 (23.6)	17 (100.4)	19 (560.0)
84	15 (32.4)	17 (66.5)	17 (350.6)	19 (2124.0)
210	15 (34.4)	17 (82.9)	17 (375.5)	19 (2463.9)
$\sigma_a = 0.4$ with IUM				
20	15 (13.6)	17 (27.1)	19 (109.2)	21 (1158.3)
84	15 (34.6)	17 (78.7)	19 (402.7)	19 (2577.4)
210	15 (34.5)	17 (80.7)	19 (414.5)	21 (2958.2)

Table 6

Simulation results using the mean-based preconditioner $\widehat{\mathcal{Z}}_0$ in (81) *with the model with time-dependent diffusion constraint for selected parameter values and degrees of freedom.*

TT-MINRES	# iter (t)	# iter (t)	# iter (t)
n_t	2^5	2^6	2^8
$\dim(\mathcal{A}) = 3JPn_t$	10,671, 360	21,342,720	85,370,880
$\alpha = 1,\ \mathrm{tol} = 10^{-3}$			
$\beta = 10^{-5}$	6 (285.5)	6 (300.0)	8 (372.2)
$\beta = 10^{-6}$	4 (77.6)	4 (130.9)	4 (126.7)
$\beta = 10^{-8}$	4 (56.7)	4 (59.4)	4 (64.9)
$\alpha = 0,\ \mathrm{tol} = 10^{-3}$			
$\beta = 10^{-5}$	4 (207.3)	6 (366.5)	6 (229.5)
$\beta = 10^{-6}$	4 (153.9)	4 (158.3)	4 (172.0)
$\beta = 10^{-8}$	2 (35.2)	2 (37.8)	2 (40.0)

a smaller tolerance $tol = 10^{-3}$ in the unsteady case because MATLAB took a lot more time due to the rapid growth of TT-ranks. Although not reported here, we also got robust two-digit TT-MINRES iterations when we used $tol = 10^{-5}$; these iterations were, as expected, even better with the inexact Uzawa method.

6. Conclusions and outlook. In this paper, we have derived and implemented block-diagonal Schur complement–based preconditioners for linear systems arising from the SGFEM discretization of SOCPs constrained by either stationary or unsteady PDEs with random inputs. Crucially, we presented detailed analyses of the spectra of the derived preconditioners. Our approach to the solution of the KKT linear systems entails a formulation that solves the systems at once (for all time steps in the unsteady case). This strategy leads to a large system that cannot be solved with direct solvers. However, combining our proposed preconditioners with appropriate low-rank iterative solvers has proven efficient in accomplishing the tasks. In particular, solves with the derived Schur complements via a few iterations of the inexact Uzawa method seem quite promising.

Although the TT-MINRES works quite well for the time-dependent problem considered in this paper, the rapid growth of the TT-ranks is not a trivial issue. In a related work [2], we are currently exploiting some capabilities of the TT-toolbox to minimize the rank growth and hence make the solver a lot more efficient.

Acknowledgments. The authors would like to thank Sergey Dolgov for introducing TT-MINRES to them; they are also grateful to Eldad Haber and Xin Liang for fruitful discussions in the course of writing this paper.

REFERENCES

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

[2] P. Benner, S. Dolgov, A. Onwunta, and M. Stoll, *Low-rank solvers for unsteady Stokes-Brinkman optimal control problem with random data*, Comput. Methods Appl. Mech. Engrg., 304 (2016), pp. 26–54.

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

[4] J. Boyle, M. D. Mihajlovic, and J. A. Scott, *HSL_MI20: An efficient AMG preconditioner for finite element problems in 3D*, Internat. J. Numer. Methods Engrg., 82 (2010), pp. 64–98.

[5] P. Chen and A. Quarteroni, *Weighted reduced basis method for stochastic optimal control problems with elliptic PDE constraint*, SIAM/ASA J. Uncertain. Quantif., 2 (2014), pp. 364–396.

[6] S. V. Dolgov, *TT-GMRES: On the solution to a linear system in the structured tensor format*, Russian J. Numer. Anal. Math. Modelling, 28 (2013), pp. 149–172.

[7] S. V. Dolgov, J. W. Pearson, D. V. Savostyanov, and M. Stoll, *Fast tensor product solvers for optimization problems with fractional differential equations as constraints*, Appl. Math. Comput., 273 (2016), pp. 604–623.

[8] H. Elman, D. Silvester, and A. Wathen, *Finite Elements and Fast Iterative Solvers*, 2nd ed., Oxford University Press, Oxford, 2014.

[9] B. Fischer, A. Ramage, D. J. Silvester, and A. J. Wathen, *Minimum residual methods for augmented systems*, BIT, 38 (1998), pp. 527–543.

[10] R. G. Ghanem and P. Spanos, *Stochastic Finite Elements: A Spectral Approach*, Springer-Verlag, New York, 1996.

[11] L. Grasedyck, D. Kressner, and C. Tobler, *A literature survey of low-rank tensor approximation techniques*, GAMM-Mitt., 36 (2013), pp. 53–78.

[12] L. S. Hou, J. Lee, and H. Manouzi, *Finite element approximations of stochastic optimal control problems constrained by stochastic elliptic PDEs*, J. Math. Anal. Appl., 384 (2011), pp. 87–103.

[13] D. P. Kouri, M. Heinkenschloss, D. Ridzal, and B. G. van Bloemen Waanders, *A trust-region algorithm with adaptive stochastic collocation for PDE optimization under uncertainty*, SIAM J. Sci. Comput., 35 (2013), pp. A1847–A1879.

[14] D. Kressner and C. Tobler, *Low-rank tensor Krylov subspace methods for parametrized linear systems*, SIAM J. Matrix Anal. Appl., 32 (2011), pp. 1288–1316.

[15] D. Leykekhman, *Investigation of commutative properties of discontinuous Galerkin methods in PDE-constrained optimal control problems*, J. Sci. Comput., 53 (2012), pp. 483–511.

[16] D. W. Nicholson, *Eigenvalue bounds for $AB+BA$ with A, B positive definite matrices*, Linear Algebra Appl., 24 (1979), pp. 173–183.

[17] I. V. Oseledets, *Tensor-train decomposition*, SIAM J. Sci. Comput., 33 (2011), pp. 2295–2317.

[18] I. V. Oseledets, S. Dolgov, V. Kazeev, D. Savostyanov, O. Lebedeva, P. Zhlobich, T. Mach, and L. Song, *TT-Toolbox*, https://github.com/oseledets/TT-Toolbox.

[19] C. C. Paige and M. A. Saunders, *Solution of sparse indefinite systems of linear equations*, SIAM J. Numer. Anal., 12 (1975), pp. 617–629.

[20] J. W. Pearson and A. J. Wathen, *A new approximation of the Schur complement in preconditioners for PDE-constrained optimization*, Numer. Linear Algebra Appl., 19 (2012), pp. 816–829.

[21] C. E. Powell and H. Elman, *Block-diagonal preconditioning for spectral stochastic finite-element systems*, IMA J. Numer. Anal., 29 (2009), pp. 350–375.

[22] T. Rees, H. S. Dollar, and A. J. Wathen, *Optimal solvers for PDE-constrained optimization*, SIAM J. Sci. Comput., 32 (2010), pp. 271–298.

[23] F. Riesz and B. Sz.-Nagy, *Functional Analysis*, Dover, New York, 1990.

[24] E. Rosseel and G. N. Wells, *Optimal control with stochastic PDE constraints and uncertain controls*, Comput. Methods Appl. Mech. Engrg., 213-216 (2012), pp. 152–167.

[25] M. Stoll and T. Breiten, *A low-rank in time approach to PDE-constrained optimization*, SIAM J. Sci. Comput., 37 (2015), pp. B1–B29.

[26] M. Stoll and A. Wathen, *All-at-once solution of time-dependent Stokes control*, J. Comput. Phys., 232 (2013), pp. 498–515.

[27] A. J. Wathen and T. Rees, *Chebyshev semi-iteration in preconditioning for problems including the mass matrix*, Electron. Trans. Numer. Anal., 34 (2008), pp. 125–135.