MAX-PLANCK-GESELLSCHAFT

**Max Planck Institute Magdeburg**
**Preprints**

Peter Benner, Sergey Dolgov, Akwum Onwunta and Martin Stoll

# Low-rank solvers for unsteady Stokes-Brinkman optimal control problem with random data

MAX–PLANCK–INSTITUT
FÜR DYNAMIK KOMPLEXER
TECHNISCHER SYSTEME
MAGDEBURG

**MPIMD/15-10** **July 23, 2015**

## Abstract

We consider the numerical simulation of an optimal control problem constrained by the unsteady Stokes-Brinkman equation involving random data. More precisely, we treat the state, the control, the target (or the desired state), as well as the viscosity, as analytic functions depending on uncertain parameters. This allows for a simultaneous generalized polynomial chaos approximation of these random functions in the stochastic Galerkin finite element method discretization of the model. The discrete problem yields a prohibitively high dimensional saddle point system with Kronecker product structure. We develop a new alternating iterative tensor method for an efficient reduction of this system by the low-rank Tensor Train representation. Besides, we propose and analyze a robust Schur complement-based preconditioner for the solution of the saddle-point system. The performance of our approach is illustrated with extensive numerical experiments based on two- and three-dimensional examples. The developed Tensor Train scheme reduces the solution storage by two orders of magnitude.

**Impressum:**

**Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg**

**Publisher:**
Max Planck Institute for
Dynamics of Complex Technical Systems

**Address:**
Max Planck Institute for
Dynamics of Complex Technical Systems
Sandtorstr. 1
39106 Magdeburg

www.mpi-magdeburg.mpg.de/preprints

# 1 Introduction

The Brinkman model is a parameter-dependent combination of the Darcy and the Stokes models. It provides a unified approach to model flows of viscous fluids in a cavity and a porous media. As pointed out in [51], in practical applications, the location and number of the Darcy-Stokes interfaces might not be known a priori. Hence, the unified equations represent an advantage over the domain decomposition methods coupling the Darcy and the Stokes equations [10, 2]. The Brinkman model is typically applied in oil reservoir modeling [43], computational fuel cell dynamics [32, 56] or biomedical engineering [48].

The study of finite element-based solvers for the Brinkman model has, on the one hand, attracted much attention recently [43, 51, 52, 56]. It is a quite challenging task, essentially due to the high variability in the coefficients of the model, which may take very high or very small values. This feature adversely affects not only the preconditioning of the resulting linear system [51], but also the construction of stable finite element discretizations [36, 56]. On the other hand, the numerical simulation of optimization problems constrained by unsteady Brinkman equations has not yet received adequate attention. Generally speaking, optimization problems constrained by unsteady partial differential equations (PDEs) are a lot more computationally challenging because one needs 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 [41, 49]. Yet, more challenging than the aforementioned are the optimal control problems constrained by unsteady PDEs involving (countably many) parametric or uncertain inputs. This class of problems arises because the input parameters of the model, such as the viscosity or initial condition may be affected by uncertainty due, for example, to measurement errors, limited data or intrinsic variability in physical phenomenon being modeled. Hence, a convenient way to characterize the uncertainty in the problem consists in incorporating the uncertain parameters as random variables or space- and/or time-varying random fields. A major goal of this work is to specifically study the preconditioning of a linear system resulting from the discretization of the optimal control problem constrained by the unsteady Stokes-Brinkman flow involving random data.

In order to numerically simulate the Brinkman optimal control problem with stochastic inputs (SOCP), we assume that the state, the control and the target are analytic functions depending on some uncertain parameters. This allows for a simultaneous generalized polynomial chaos (PCE) approximation of these random functions [16, 17, 35, 44, 57] in the stochastic Galerkin finite element method (SGFEM) discretization of the model. However, these problems often lead to prohibitively high dimensional linear systems with Kronecker product structure.

To reduce the computational complexity, we impose the Kronecker product structure on the solution as well. More precisely, we seek an approximate solution in a low-rank tensor product representation, namely, the Tensor Train decomposition [38], also known as the Matrix Product States [26]. The tensor decomposition concept is similar to low-rank model reduction techniques, for example, the Proper Orthogonal Decomposition (POD) [31]. However, the POD solves the full problem in order to derive a reduced model. For really large-scale systems this is not feasible. Tensor methods aim to construct directly the reduced solution without a priori information. One of the most powerful tensor-based algorithms that can effectively accomplish this task is the alternating iterative method [21, 47, 54]. However, existing alternating solvers for linear systems require a positive definite matrix. Another novel contributions of this paper are the extension and adaptation of these algorithms to the saddle-points optimality system. We refer to [19, 18] for a more detailed overview of tensor methods.

This paper is structured as follows. In Section 2, we present the deterministic Stokes-Brinkman model. Section 3 introduces the Stokes-Brinkman optimal control problem with uncertain inputs and gives an overview of the SGFEM. Besides, it establishes the Kronecker-product structure of the discrete problem. Section 4 presents and analyzes our preconditioners for the corresponding saddle-point linear systems. In Section 5, we introduce the Tensor Train decomposition and alternating tensor algorithms, adjust them to the particular structure of the inverse problem and the Stokes-Brinkman model and discuss some implementation issues. Section 6 contains numerical results obtained for two- and three-dimensional examples using our approach. Finally, Section 7 gives a conclusion and outlines future research goals.

## 2 Deterministic Brinkman model

Let $\mathcal{D} \subset \mathbb{R}^d$ with $d \in \{1, 2, 3\}$, be a bounded open set with Lipschitz continuous simply connected boundary $\partial\mathcal{D}$. Herein, the spatial domain $\mathcal{D}$ consists of two parts, namely, a porous medium $\mathcal{D}_p$ and a viscous flow medium $\mathcal{D}_s$. That is, $\mathcal{D} = \mathcal{D}_p \cup \mathcal{D}_s$. Moreover, denote by $\mathcal{Q}$ the space-time cylinder $\mathcal{D}\times[0, T]$ and $\mathcal{T} = (0, T]$. The generalized unsteady Brinkman problem reads

$$\begin{cases} \dfrac{\partial v(\mathbf{x}, t)}{\partial t} - \nu\Delta v(\mathbf{x}, t) + K_0(\mathbf{x})v(\mathbf{x}, t) + \nabla p(\mathbf{x}, t) = u(\mathbf{x}, t), \text{ in } \mathcal{Q}, \\ \qquad\qquad\qquad\qquad -\nabla \cdot v(\mathbf{x}, t) = 0, \text{ on } \mathcal{Q}, \\ \qquad\qquad\qquad\qquad v(\mathbf{x}, t) = h(\mathbf{x}, t), \text{ on } \partial\mathcal{D} \times \mathcal{T}, \\ \qquad\qquad\qquad\qquad v(\mathbf{x}, 0) = v_0, \text{ in } \mathcal{D}, \end{cases} \tag{1}$$

where $v$ and $p$ are, respectively, the fluid velocity and the fluid pressure, and $h$ is the boundary condition. The parameter $\nu$ represents the fluid viscosity. Moreover, $K_0$ is the *inverse permeability tensor* of the medium. We assume here that $K_0 \in L^2(\mathcal{D}) \cap L^\infty(\mathcal{D})$ and that the source term $u \in L^2(\mathcal{D})$. The challenge of this problem is that the coefficient $K_0$ takes two extreme values: it is very small in the viscous flow medium $\mathcal{D}_s$ so that the PDE behaves like the unsteady Stokes flow, and very big in the porous medium $\mathcal{D}_p$ in which case the PDE behaves like the unsteady Darcy equations.

In this paper, we denote by $H^k(\mathcal{D})$ the Sobolev space of functions on $\mathcal{D}$ whose derivatives up to order $k$ are square-integrable. $H^k_0(\mathcal{D})$ denotes the closure in $H^k(\mathcal{D})$ of the set of finitely differentiable functions with compact support in $\mathcal{D}$. For some space $\mathcal{X}$ of functions on $\mathcal{D}$, let $L^2(0, T; \mathcal{X}) = L^2(0, T)\otimes\mathcal{X}$. The variational formulation of the Brinkman model (1) can thus be written in following form: find $v \in L^2(0, T; H^1_0(\mathcal{D}))$,

$p \in L^2(0,T;L^2(\mathcal{D}))$ and $\partial_t v \in L^2(0,T;H^{-1}(\mathcal{D}))$, such that $v|_{t=0} = v_0$ and *a.e* on $[0,T]$

$$\begin{cases} (\partial_t v(t), w) + \mathcal{B}(v(t), w) - \mathcal{C}(p(t), \text{div } w) = & (u, w),\ \forall w \in L^2(0,T;H_0^1(\mathcal{D})) \\ \mathcal{C}(\text{div } v, q) = & 0,\ \ \forall q \in L^2(0,T;L^2(\mathcal{D})), \end{cases}$$

where

$$\mathcal{B}(v(t), w) = (\nu \nabla v(t), \nabla w) + (K_0(\mathbf{x})v, w),$$
$$\mathcal{C}(p(t), \text{div } w) = (p(t), \nabla \cdot w),$$

and $(\cdot,\cdot)$ represents the $L^2$ inner product of a pair of functions on $\mathcal{D}$.

For a mixed finite element discretization of the Brinkman problem [36, 48, 51, 56] in the primal variables $v$ and $p$, let $V_h \subset L^2(0,T;H_0^1(\mathcal{D}))$ and $W_h \subset L^2(0,T;L^2(\mathcal{D}))$ be finite element spaces with stable elements (i.e. elements that satisfy the *inf-sup* condition, e.g. *mini elements* as discussed in [48]) such that $V_h = \text{span}\{\phi_1, \ldots, \phi_{J_v}\}$ and $W_h = \text{span}\{\varphi_1, \ldots, \varphi_{J_p}\}$. Performing a Galerkin projection on $V_h$ and $W_h$ and using implicit Euler for the temporal discretization, while taking into account the boundary conditions, leads to the following

$$\begin{cases} \dfrac{Mv_i - Mv_{i-1}}{\tau} + (\nu K + M_k)v_i + B^T p_i = Mu_i + g_i, \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad Bv_i = 0, \end{cases} \tag{2}$$

where $B = \left[-\int_{\mathcal{D}} \varphi_k \nabla \cdot \phi_{k'}\right]$ is the discrete divergence operator, $K = \left[\int_{\mathcal{D}} \nabla\phi_k : \nabla\phi_{k'}\right]$ is a matrix representing the vector Laplacian operator, $M = \left[\int_{\mathcal{D}} \phi_k \phi_{k'}\right]$ is the mass matrix and $M_k = \left[\int_{\mathcal{D}} K_0 \phi_k \phi_{k'}\right]$ is the matrix associated with the term which involves the inverse permeability coefficient $K_0(\mathbf{x})$, and $\tau$ is the size of the time step.

**Remark 1.** *In the special case where $M_k = 0$ in (2), we get the unsteady Stokes problem.*

# 3 Brinkman optimal control problem with random data

Suppose now that, even though the fluid viscosity $\nu$ is time-independent and spatially constant but that its value is not known precisely. Instead of guessing a value, we can model $\nu$ as a random variable defined on the complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$. This could be interpreted as a scenario where the volume of fluid moving into a channel is uncertain due to measurement error in $\nu$ or probably some other factors [45]. Here, $\Omega$ is a sample space of events whereas, $\mathcal{F}$ denotes a $\sigma$-algebra on $\Omega$ and is endowed with an appropriate probability measure $\mathbb{P}$. The corresponding Brinkman velocity and pressure are consequently also random and the numerical solution of the associated SOCP is far more challenging. More precisely, the SOCP which we will solve in the rest of this paper consists in minimizing the cost functional of tracking-type

$$\mathcal{J} = \frac{1}{2}||v - \bar{v}||^2_{L^2(0,T;\mathcal{D})\otimes L^2(\Omega)} + \frac{\alpha}{2}||\text{std}(v)||^2_{L^2(0,T;\mathcal{D})} + \frac{\beta}{2}||u||^2_{L^2(0,T;\mathcal{D})\otimes L^2(\Omega)} \tag{3}$$

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

$$\begin{cases} \dfrac{\partial v(\mathbf{x},t,\omega)}{\partial t} - \nu(\omega)\Delta v(\mathbf{x},t,\omega) + K_0(\mathbf{x})v(\mathbf{x},t,\omega) + \nabla p(\mathbf{x},t,\omega) = u(\mathbf{x},t,\omega), \text{ in } \mathcal{Q}\times\Omega, \\ \hfill -\nabla\cdot v(\mathbf{x},t,\omega) = 0, \text{ on } \mathcal{Q}\times\Omega, \\ \hfill v(\mathbf{x},t,\omega) = h, \text{ on } \partial\mathcal{D}\times\mathcal{T}\times\Omega, \\ \hfill v(\mathbf{x},0,\omega) = v_0, \text{ in } \mathcal{D}\times\Omega, \end{cases}$$

where $v, \bar{v}, p : \mathcal{D}\times\mathcal{T}\times\Omega \to \mathbb{R}$ are random fields [6] representing the state (velocity), the target (or the desired state) and the pressure. The forcing term on the right hand side $u : \mathcal{D}\times\mathcal{T}\times\Omega \to \mathbb{R}$ denotes a random control function. Moreover, the positive constant $\beta$ represents the parameter for the penalization of the norm of the control $u$, whereas $\alpha$ penalizes the standard deviation $\mathrm{std}(v)$ of the state $v$. Here, we have used the notation $L^2(\Omega) := L^2(\Omega, \mathcal{F}, \mathbb{P})$.

The viscosity $\nu$ in the state equations is modeled as a uniformly distributed random variable of the form

$$\nu(\omega) = \nu_0 + \nu_1\xi(\omega), \tag{4}$$

with $\nu_0, \nu_1 \in \mathbb{R}^+$ and $\xi \sim \mathcal{U}(-1,1)$. Furthermore, we assume that the control and the target satisfy

$$u, \bar{v} \in L^2(\mathcal{D}) \otimes L^2(\mathcal{T}) \otimes L^2(\Omega), \tag{5}$$

and that, for some $\nu_{\min}, \nu_{\max} \in \mathbb{R}^+$ satisfying $0 < \nu_{\min} < \nu_{\max} < +\infty$, we have

$$\mathbb{P}\left(\omega \in \Omega : \nu(\xi(\omega)) \in [\nu_{\min}, \nu_{\max}]\right) = 1. \tag{6}$$

## 3.1 A fully discrete problem

Two standard methods are used to discretize the optimal control problem introduced above - we can either discretize the model first and then optimize the discrete system (DTO method), or alternatively optimize first before discretizing the resulting optimality system (OTD method). 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 [33] for an overview). In what follows, we will adopt the DTO strategy because, for the optimal control problem considered in this paper, it leads to a symmetric saddle point linear system which fits in nicely with our preconditioning strategies.

Since our optimal control problem contains random coefficients, the stochastic discretization could be effected using either a projection-based method (e.g. stochastic Galerkin method in [44]) or a sampling method (e.g. stochastic collocation method in [46]). Due to its high convergence rate, the former is our preferred method in this paper. In order to use this method, we first assume that the pressure $p$, the state $v$, the control $u$ and the target $\bar{v}$ are analytic functions depending on the uncertain parameters. This allows for a simultaneous generalized polynomial chaos (PCE) approximation of these random functions [16, 44, 6]. Of course, $\bar{v}$ can equally be modeled deterministically. Together with the finite element method, the PCE yields an SGFEM for discretizing both the spatial and stochastic domains. More precisely, $p, u, v$, and $\bar{v}$ admit the following respective representations

$$\begin{aligned} p(\mathbf{x},t,\omega) &= \sum_{k=1}^{J_p}\sum_{j=1}^{P} p_{kj}(t)\varphi_k(\mathbf{x})\psi_j(\xi(\omega)), \\ u(\mathbf{x},t,\omega) &= \sum_{k=1}^{J_v}\sum_{j=1}^{P} u_{kj}(t)\phi_k(\mathbf{x})\psi_j(\xi(\omega)), \\ v(\mathbf{x},t,\omega) &= \sum_{k=1}^{J_v}\sum_{j=1}^{P} v_{kj}(t)\phi_k(\mathbf{x})\psi_j(\xi(\omega)), \\ \bar{v}(\mathbf{x},t,\omega) &= \sum_{k=1}^{J_v}\sum_{j=1}^{P} \bar{v}_{kj}(t)\phi_k(\mathbf{x})\psi_j(\xi(\omega)), \end{aligned} \tag{7}$$

where $\{\psi_j\}_{j=1}^{P}$ are univariate orthogonal polynomials of order $P-1$ satisfying

$$\langle\psi_1(\xi)\rangle = 1, \;\; \langle\psi_j(\xi)\rangle = 0, \; j>1, \;\; \langle\psi_j(\xi)\psi_k(\xi)\rangle = \left\langle\psi_j^2(\xi)\right\rangle\delta_{jk}, \tag{8}$$

with

$$\langle\psi_j(\xi)\rangle = \int_{\omega\in\Omega}\psi_j(\xi(\omega))\,d\mathbb{P}(\omega) = \int_{\xi\in\Gamma}\psi_j(\xi)\rho(\xi)\,d\xi, \tag{9}$$

where $\rho$ is the density of the random variable $\xi$ and $\Gamma$ is the support of $\rho$.

In spirit of [5, 49], we apply to the cost functional the trapezoidal rule for temporal discretization, and the *mini* finite elements [48], together with Legendre polynomial chaos in the SGFEM for spatial and stochastic discretizations [44], to get the following

$$\mathcal{J}(\mathbf{y},\mathbf{u}) := \frac{\tau}{2}(\mathbf{y}-\bar{\mathbf{y}})^T\boldsymbol{M}_a(\mathbf{y}-\bar{\mathbf{y}}) + \frac{\tau\alpha}{2}\mathbf{y}^T\boldsymbol{M}_b\mathbf{y} + \frac{\tau\beta}{2}\mathbf{u}^T\boldsymbol{M}_2\mathbf{u}, \tag{10}$$

where $\mathbf{y}^\top = \left[\mathbf{v}_1^\top, \mathbf{p}_1^\top, \ldots, \mathbf{v}_{n_t}^\top, \mathbf{p}_{n_t}^\top\right] \in \mathbb{R}^{JPn_t}$, $J := J_v + J_p$, and $\mathbf{u}^\top = \left[\mathbf{u}_1^\top, \ldots, \mathbf{u}_{n_t}^\top\right]$ denote the long vectors of all time snapshots of the state and control, respectively,

$$\begin{cases} \boldsymbol{M}_a = \operatorname{blkdiag}\left(\frac{1}{2}\mathcal{M}, 0, \mathcal{M}, 0, \ldots, \mathcal{M}, 0, \frac{1}{2}\mathcal{M}, 0\right), & \mathcal{M} := G_0 \otimes M, \\ \boldsymbol{M}_b = \operatorname{blkdiag}\left(\frac{1}{2}\mathcal{M}_t, 0, \mathcal{M}_t, 0, \ldots, \mathcal{M}_t, 0, \frac{1}{2}\mathcal{M}_t, 0\right), & \mathcal{M}_t := H_0 \otimes M, \\ \boldsymbol{M}_2 = \operatorname{blkdiag}\left(\frac{1}{2}\mathcal{M}, \mathcal{M}, \ldots, \mathcal{M}, \frac{1}{2}\mathcal{M}\right), \end{cases} \tag{11}$$

with $M$ the finite element mass matrix, and

$$\begin{cases} G_0 = \operatorname{diag}\left(\left\langle\psi_1^2(\xi)\right\rangle, \left\langle\psi_2^2(\xi)\right\rangle, \ldots, \left\langle\psi_P^2(\xi)\right\rangle\right), \\ H_0 = \operatorname{diag}\left(0, \left\langle\psi_2^2(\xi)\right\rangle, \ldots, \left\langle\psi_P^2(\xi)\right\rangle\right), \end{cases} \tag{12}$$

where the Kronecker product $\otimes$ is meant in the usual sense, $A \otimes B = [A_{ij}B]$.

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

$$\boldsymbol{K}\mathbf{y} - \boldsymbol{N}\mathbf{u} = \mathbf{g}, \tag{13}$$

where

$$\boldsymbol{K} = \begin{bmatrix} \bar{\mathcal{L}} & & & \\ -\bar{\mathcal{M}} & \bar{\mathcal{L}} & & \\ & \ddots & \ddots & \\ & & -\bar{\mathcal{M}} & \bar{\mathcal{L}} \end{bmatrix}, \ \boldsymbol{N} = \begin{bmatrix} \mathcal{N} & & & \\ & \mathcal{N} & & \\ & & \ddots & \\ & & & \mathcal{N} \end{bmatrix}, \ \mathbf{g} = \begin{bmatrix} \bar{\mathcal{M}}\mathbf{y}_0 + \mathbf{g}_1^0 \\ \mathbf{g}_2^0 \\ \vdots \\ \mathbf{g}_{n_t}^0 \end{bmatrix},$$

with

$$\mathcal{N} = G_0 \otimes N, \quad N = \begin{bmatrix} M \\ 0 \end{bmatrix}, \quad \bar{\mathcal{M}} = G_0 \otimes \tau^{-1}\bar{M}, \quad \bar{M} = \begin{bmatrix} M & 0 \\ 0 & 0 \end{bmatrix}, \tag{14}$$

and, in the notation of [45],

$$\bar{\mathcal{L}} = \begin{bmatrix} \mathcal{A} & \mathcal{B}^T \\ \mathcal{B} & 0 \end{bmatrix} \tag{15}$$

represents an instance of the time-dependent Brinkman problem with

$$\mathcal{A} = G_0 \otimes A + G_1 \otimes \nu_1 K, \quad A = \tau^{-1}M + \nu_0 K + M_k, \qquad \mathcal{B} = G_0 \otimes B, \tag{16}$$

and $G_1(j,j') = \langle \xi\psi_j(\xi)\psi_{j'}(\xi)\rangle$. Note that since we are using Legendre polynomials for SGFEM discretization, $G_0$ is a diagonal matrix whereas $G_1$ is a tridiagonal matrix with zeros on the main diagonal (see e.g., [44, 45]). This implies that the matrix $\mathcal{B}$ in (16) is block-diagonal. Furthermore, since the matrices $K$, $M$ and $M_k$ are positive definite, we know that $\mathcal{A}$ in (16) is sparse and positive definite. However, $\bar{\mathcal{L}}$ is an indefinite block sparse matrix with sparse blocks.

Later it will be convenient to work with the Kronecker product representations of the system matrices. To this end, we introduce the identity matrix $I_{n_t} \in \mathbb{R}^{n_t \times n_t}$, as well as the matrix

$$C = \begin{bmatrix} 0 & & & \\ -1 & 0 & & \\ & \ddots & \ddots & \\ & & -1 & 0 \end{bmatrix}, \tag{17}$$

and observe then that

$$\boldsymbol{K} = I_{n_t} \otimes G_0 \otimes \begin{bmatrix} A & B^\top \\ B & 0 \end{bmatrix} + I_{n_t} \otimes G_1 \otimes \begin{bmatrix} \nu_1 K & 0 \\ 0 & 0 \end{bmatrix} + C \otimes G_0 \otimes \begin{bmatrix} \tau^{-1}M & 0 \\ 0 & 0 \end{bmatrix}, \tag{18}$$

and

$$\boldsymbol{N} = I_{n_t} \otimes G_0 \otimes N. \tag{19}$$

The structure of the right-hand side is problem-dependent. However, in our experiments we will use $\mathbf{y}_0 = 0$ and a static deterministic $\mathbf{g}^0$ coming from Dirichlet boundary conditions, such that $\mathbf{g} = \mathbf{g}^0 = \mathbf{e} \otimes \mathbf{e}_1 \otimes \begin{bmatrix} \mathbf{g}_v^0 \\ \mathbf{g}_p^0 \end{bmatrix}$, where $\mathbf{e}$ is the vector of all ones, and $\mathbf{e}_1$ is the first unit vector.

Now, note from (10) and (13) that the discrete Lagrangian functional of the SOCP is given by

$$\mathfrak{L} := \frac{\tau}{2}(\mathbf{y} - \bar{\mathbf{y}})^T \boldsymbol{M}_a (\mathbf{y} - \bar{\mathbf{y}}) + \frac{\tau\alpha}{2} \mathbf{y}^T \boldsymbol{M}_b \mathbf{y} + \frac{\tau\beta}{2} \mathbf{u}^T \boldsymbol{M}_2 \mathbf{u} + \boldsymbol{\lambda}^T(-\boldsymbol{K}\mathbf{y} + \boldsymbol{N}\mathbf{u} + \mathbf{g}),$$

where $\boldsymbol{\lambda}$ is the Lagrange multiplier. Hence, applying the first order conditions to $\mathfrak{L}$ yields the Karush-Kuhn-Tucker (KKT) system

$$\underbrace{\begin{bmatrix} \tau \boldsymbol{M}_1 & 0 & -\boldsymbol{K}^T \\ 0 & \beta\tau \boldsymbol{M}_2 & \boldsymbol{N}^T \\ -\boldsymbol{K} & \boldsymbol{N} & 0 \end{bmatrix}}_{:=\mathfrak{A}} \begin{bmatrix} \mathbf{y} \\ \mathbf{u} \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{0} \\ \mathbf{g} \end{bmatrix}, \tag{20}$$

where $\mathbf{b}_1 = \tau \boldsymbol{M}_a \bar{\mathbf{y}}$, and

$$\boldsymbol{M}_1 = \boldsymbol{M}_a + \alpha \boldsymbol{M}_b = D \otimes G_\alpha \otimes \bar{M}, \qquad \boldsymbol{M}_2 = D \otimes \mathcal{M} = D \otimes G_0 \otimes M, \tag{21}$$

$$D = \operatorname{diag}\left(\frac{1}{2}, 1 \ldots, 1, \frac{1}{2}\right) \in \mathbb{R}^{n_t \times n_t}, \qquad G_\alpha := G_0 + \alpha H_0. \tag{22}$$

We note here that if the desired state is also static and deterministic, then one gets $\bar{\mathbf{y}} = \mathbf{e} \otimes \mathbf{e}_1 \otimes \begin{bmatrix} \bar{\mathbf{v}} \\ 0 \end{bmatrix}$.

## 4 Preconditioning

The KKT coefficient matrix $\mathfrak{A}$ in (20) is usually ill-conditioned and thus requires a suitable preconditioner to solve (20) efficiently. A block-diagonal preconditioner, discussed in the framework of deterministic unsteady Stokes control problem [50], is written in the form $\boldsymbol{P}_1 = \operatorname{blockdiag}(\tilde{\boldsymbol{M}}_1, \beta \boldsymbol{M}_2, \tilde{\boldsymbol{S}}_1)$, where $\tilde{\boldsymbol{S}}_1 = \frac{1}{\tau}\left(\boldsymbol{K} + \boldsymbol{M}_s\right) \tilde{\boldsymbol{M}}_1^{-1} \left(\boldsymbol{K}^T + \boldsymbol{M}_s\right)^T$ is the approximate Schur complement, and $\tilde{\boldsymbol{M}}_1$ is some perturbation to $\boldsymbol{M}_1$, since the latter is rank-deficient. Here, the matrix $\boldsymbol{M}_s$ is determined via a 'matching' argument. In particular, [50] suggest the following augmentation,

$$\tilde{\boldsymbol{M}}_1 = \begin{bmatrix} D \otimes G_\alpha \otimes M & \\ & D \otimes G_\alpha \otimes \left(\|M\|_2^2 \tau\beta\right) I \end{bmatrix},$$

where $I$ is the identity of the size of the pressure grid. However, this approach is tricky. For example, it may be quite suitable for preconditioning of MINRES, which works with the $\boldsymbol{P}_1^{-1}$-scalar product, but perform poorly in the Flexible GMRES, if we are to apply $\boldsymbol{P}_1^{-1}$ approximately. Besides, it is not obvious how to generalize it to the case when $\boldsymbol{M}_1$ is *numerically* rank-deficient, i.e. its eigenvalues form a gradually decaying sequence instead of two distinct clusters. This will occur in the low-rank tensor methods; consequently, instead of $\boldsymbol{M}_1$, we will work with its Galerkin projection in the sequel. More specifically, we proceed next to Section 4.1 to propose another preconditioner which circumvents this deficiency and yields faster convergence even with the original sparse $\boldsymbol{M}_1$.

## 4.1 A block-triangular preconditioner

Our point of departure is to replace the KKT coefficient matrix $\mathfrak{A}$ in (20) by $\tilde{\mathfrak{A}}$ given by

$$\tilde{\mathfrak{A}} := \mathfrak{A}\boldsymbol{\rho} = \begin{bmatrix} -\boldsymbol{K}^T & 0 & \tau\boldsymbol{M}_1 \\ \boldsymbol{N}^T & \beta\tau\boldsymbol{M}_2 & 0 \\ 0 & \boldsymbol{N} & -\boldsymbol{K} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Phi} & \boldsymbol{\Upsilon} \\ \boldsymbol{\Psi} & -\boldsymbol{K} \end{bmatrix},$$

where

$$\boldsymbol{\rho} = \begin{bmatrix} 0 & 0 & \boldsymbol{I} \\ 0 & \boldsymbol{I} & 0 \\ \boldsymbol{I} & 0 & 0 \end{bmatrix}, \quad \boldsymbol{\Phi} = \begin{bmatrix} -\boldsymbol{K}^T & 0 \\ \boldsymbol{N}^T & \beta\tau\boldsymbol{M}_2 \end{bmatrix}, \quad \boldsymbol{\Upsilon} = \begin{bmatrix} \tau\boldsymbol{M}_1 \\ 0 \end{bmatrix}, \quad \boldsymbol{\Psi} = \begin{bmatrix} 0 \\ \boldsymbol{N} \end{bmatrix}^T.$$

Note that the matrix $\boldsymbol{\rho}$ swaps the first and third columns of $\mathfrak{A}$ in the product $\mathfrak{A}\boldsymbol{\rho}$; it swaps the first and third rows of $\mathfrak{A}$ in the product $\boldsymbol{\rho}\mathfrak{A}$. Next, observe also that we can factorize the matrix $\tilde{\mathfrak{A}}$ as follows

$$\begin{bmatrix} \boldsymbol{\Phi} & \boldsymbol{\Upsilon} \\ \boldsymbol{\Psi} & -\boldsymbol{K} \end{bmatrix} = \begin{bmatrix} \boldsymbol{I} & 0 \\ \boldsymbol{\Psi}\boldsymbol{\Phi}^{-1} & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Phi} & \boldsymbol{\Upsilon} \\ 0 & -\boldsymbol{S}_2 \end{bmatrix},$$

where

$$\boldsymbol{\Phi}^{-1} = \begin{bmatrix} -\boldsymbol{K}^{-T} & 0 \\ \frac{1}{\tau\beta}\boldsymbol{M}_2^{-1}\boldsymbol{N}^T\boldsymbol{K}^{-T} & \frac{1}{\tau\beta}\boldsymbol{M}_2^{-1} \end{bmatrix}, \tag{23}$$

and $\boldsymbol{S}_2 = \boldsymbol{K} + \boldsymbol{\Psi}\boldsymbol{\Phi}^{-1}\boldsymbol{\Upsilon} = \boldsymbol{K} + \frac{1}{\beta}\boldsymbol{N}\boldsymbol{M}_2^{-1}\boldsymbol{N}^T\boldsymbol{K}^{-T}\boldsymbol{M}_1$. But then, from (14), (19) and (21), we obtain

$$\boldsymbol{N}\boldsymbol{M}_2^{-1}\boldsymbol{N}^T = D^{-1} \otimes G_0 \otimes \bar{M} = D^{-1} \otimes \begin{bmatrix} \tau\mathcal{M} & 0 \\ 0 & 0 \end{bmatrix} =: \boldsymbol{M}_{-1}. \tag{24}$$

Therefore,

$$\boldsymbol{S}_2 = \boldsymbol{K} + \boldsymbol{\Psi}\boldsymbol{\Phi}^{-1}\boldsymbol{\Upsilon} = \boldsymbol{K} + \frac{1}{\beta}\boldsymbol{M}_{-1}\boldsymbol{K}^{-T}\boldsymbol{M}_1. \tag{25}$$

We propose to right-precondition $\tilde{\mathfrak{A}}$ with the matrix

$$\boldsymbol{P}_D = \begin{bmatrix} \boldsymbol{\Phi} & \boldsymbol{\Upsilon} \\ 0 & -\boldsymbol{S}_2 \end{bmatrix}. \tag{26}$$

This implies that

$$\tilde{\mathfrak{A}}\boldsymbol{P}_D^{-1} = \mathfrak{A}\boldsymbol{\rho}\boldsymbol{P}_D^{-1} = \mathfrak{A}\boldsymbol{P}_2^{-1} = \begin{bmatrix} \boldsymbol{I} & 0 \\ \boldsymbol{\Psi}\boldsymbol{\Phi}^{-1} & \boldsymbol{I} \end{bmatrix}, \tag{27}$$

where the right preconditioner $\boldsymbol{P}_2$ for the original KKT matrix $\mathfrak{A}$ satisfies

$$\boldsymbol{P}_2^{-1} = \boldsymbol{\rho}\boldsymbol{P}_D^{-1} = \begin{bmatrix} 0 & 0 & -\boldsymbol{S}_2^{-1} \\ \frac{1}{\beta\tau}\boldsymbol{M}_2^{-1}\boldsymbol{N}^T\boldsymbol{K}^{-T} & \frac{1}{\beta\tau}\boldsymbol{M}_2^{-1} & \frac{1}{\beta}\boldsymbol{M}_2^{-1}\boldsymbol{N}^T\boldsymbol{K}^{-T}\boldsymbol{M}_1\boldsymbol{S}_2^{-1} \\ -\boldsymbol{K}^{-T} & 0 & -\boldsymbol{K}^{-T}\tau\boldsymbol{M}_1\boldsymbol{S}_2^{-1} \end{bmatrix}. \tag{28}$$

It can be noticed that (27) immediately implies $(\mathfrak{A}\boldsymbol{P}_2^{-1} - I)^2 = 0$; hence, such Krylov solvers as the generalized minimal residual (GMRES) method will converge in two iterations if $\boldsymbol{P}_2^{-1}$ is applied exactly, see e.g. [14, Section 8.1].

The seeming complicated structure of (28) notwithstanding, matrix-vector product with $\boldsymbol{P}_2^{-1}$ can be implemented fairly easily. For instance, suppose now that we want to solve $\mathbf{x} = \boldsymbol{P}_2^{-1}\mathbf{y}$, where $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3]^T$, $\mathbf{y} = [\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3]^T$. Then, it can easily be shown that an efficient way to implement the matrix-vector product is

$$\begin{cases} \mathbf{x}_1 = -\boldsymbol{S}_2^{-1}\mathbf{y}_3 \\ \mathbf{x}_3 = -\boldsymbol{K}^{-T}(\mathbf{y}_1 - \tau\boldsymbol{M}_1\mathbf{x}_1) \\ \mathbf{x}_2 = \tau^{-1}\beta^{-1}\boldsymbol{M}_2^{-1}(\mathbf{y}_2 - \boldsymbol{N}^T\mathbf{x}_3). \end{cases} \tag{29}$$

Next, following a state-of-the-art preconditioning strategy in [42], we approximate the Schur complement $\boldsymbol{S}_2$ in (25) with a matrix of the form

$$\begin{aligned} \tilde{\boldsymbol{S}}_2 &= (\boldsymbol{K} + \boldsymbol{M}_l)\,\boldsymbol{K}^{-T}\left(\boldsymbol{K}^T + \boldsymbol{M}_r\right). \\ &= \boldsymbol{K} + \boldsymbol{M}_l\boldsymbol{K}^{-T}\boldsymbol{M}_r + \boldsymbol{M}_l + \boldsymbol{K}\boldsymbol{K}^{-T}\boldsymbol{M}_r, \end{aligned} \tag{30}$$

where $\boldsymbol{M}_l$ and $\boldsymbol{M}_r$ are determined using also a 'matching' argument between the exact Schur complement $\boldsymbol{S}_2$ and the approximation $\tilde{\boldsymbol{S}}_2$. More precisely, we ignore the last two terms in (30) and match the first and second terms with those in (25) to get $\boldsymbol{M}_r = \beta^{-1/2}\boldsymbol{M}_1$, and $\boldsymbol{M}_l = \beta^{-1/2}\boldsymbol{M}_{-1}$, where $\boldsymbol{M}_1$ and $\boldsymbol{M}_{-1}$ are as defined, respectively, in (21) and (24). Hence, we have

$$\tilde{\boldsymbol{S}}_2 = \left(\boldsymbol{K} + \frac{1}{\sqrt{\beta}}\boldsymbol{M}_{-1}\right)\boldsymbol{K}^{-T}\left(\boldsymbol{K}^T + \frac{1}{\sqrt{\beta}}\boldsymbol{M}_1\right). \tag{31}$$

For matrix-vector products, the factors $\left(\boldsymbol{K} + \frac{1}{\sqrt{\beta}}\boldsymbol{M}_{-1}\right)$ and $\left(\boldsymbol{K}^T + \frac{1}{\sqrt{\beta}}\boldsymbol{M}_1\right)$ can be kept as sums of four Kronecker products, with the first three coming from $\boldsymbol{K}$ in (18), and the fourth corresponding to $\boldsymbol{M}_{-1}$ in (24) and $\boldsymbol{M}_1$ in (21), respectively. However, our ultimate goal is to apply $\tilde{\boldsymbol{S}}_2^{-1}$, where it appears that solving a linear system with exact factors is difficult. As a result, we instead approximate them by one Kronecker-product term: we approximate $\boldsymbol{K}$ by the first term from (18), whereas we set $\boldsymbol{M}_1 \approx I_{n_t} \otimes (1+\alpha)G_0 \otimes \bar{M}$ and $\boldsymbol{M}_{-1} \approx I_{n_t} \otimes G_0 \otimes \bar{M}$; therefore,

$$\left(\boldsymbol{K} + \frac{1}{\sqrt{\beta}}\boldsymbol{M}_{\mathtt{i}}\right) \approx I_{n_t} \otimes G_0 \otimes \begin{bmatrix} A + \eta_{\mathtt{i}} M & B^\top \\ B & 0 \end{bmatrix}, \tag{32}$$

where $\mathtt{i} \in \{-1, 1\}$, and $\eta_{-1} = 1/\sqrt{\beta}$, $\eta_1 = (1+\alpha)/\sqrt{\beta}$. Inside alternating tensor methods (cf. Section 5.5), the matrix $I_{n_t} \otimes G_0$ will be further reduced, but the concept of the one-term preconditioner remains the same.

### 4.2 Preconditioning of the forward Stokes-Brinkman problem

In linear systems of the form (32), $I_{n_t}$ and $G_0$ can be inverted straightforwardly, while the spatial matrix may require a special treatment. To this end, we can use either the GMRES or the inexact Uzawa algorithm (see e.g. [50]), together with the block-triangular preconditioner

$$P_s = \begin{bmatrix} \tilde{A} & 0 \\ B & -S_0 \end{bmatrix}, \tag{33}$$

where $S_0 = B\tilde{A}^{-1}B^\top$ is the Schur complement and $\tilde{A} = \nu_0 K + M_k + (\tau^{-1} + \eta)M$ with $\eta = \frac{1}{\sqrt{\beta}}$ or $\eta = \frac{1+\alpha}{\sqrt{\beta}}$. So, we need $P_s^{-1}$, that is,

$$P_s^{-1} = \begin{bmatrix} \tilde{A}^{-1} & 0 \\ S_0^{-1}B\tilde{A}^{-1} & -S_0^{-1} \end{bmatrix}. \tag{34}$$

In what follows, we derive the approximation to the blocks of $P_s^{-1}$. First, to approximate $\tilde{A}$, we can use algebraic multigrid methods, since $\tilde{A}$ is symmetric and positive definite. Next, we need an approximation to the Schur complement $S_0$. As was pointed out in [14], the pressure mass matrix is a very effective approximation for $S_0$ in the case of stationary Stokes equations. However, as we are considering unsteady Stokes-Brinkman constraint, this does not apply since $\tilde{A}$ has an entirely different structure. Thus, following [50], we proceed to derive the so-called Cahouet-Chabard approximation to $S_0$ using a technique for the steady Navier-Stokes equation, which is based on the least squares commutator (see Chapter 8 of [14]) defined by

$$\mathbb{E} := (\mathbb{L})\nabla - \nabla(\mathbb{L}_p),$$

where $\mathbb{L} = (\tau^{-1} + \eta)I + \Delta + K_0$ and $\mathbb{L}_p = (\tau^{-1} + \eta)I_p + \Delta_p + K_{0_p}$ is defined similarly but on the pressure space. As was noted in [50], these operators are only used for the purpose of deriving matrix preconditioners and no function spaces or boundary conditions are defined here. Assuming the least squares commutator is small, we obtain the following discretization of the differential operators

$$\mathbb{E}_h = (M^{-1}\tilde{A})M^{-1}B^T - M^{-1}B^T(M_p^{-1}\tilde{A}_p) \approx 0, \tag{35}$$

where $\tilde{A}, B$ and $M$ are as defined previously, and

$$\tilde{A}_p \quad = \quad \nu_0 K_p + M_{k_p} + (\tau^{-1} + \eta)M_p. \tag{36}$$

Next, we pre-multiply the expression (35) by $B\tilde{A}^{-1}M$ and post-multiply it by $\tilde{A}_p^{-1}M_p$ to obtain

$$BM^{-1}B^T\tilde{A}_p^{-1}M_p - B\tilde{A}^{-1}B^T \approx 0. \tag{37}$$

Now, since $BM^{-1}B^T$ is spectrally equivalent to the Laplacian defined on the pressure space[1], $K_p$, and $B\tilde{A}^{-1}B^T$ is the sought $S_0$, we obtain

$$S_0 \approx K_p \tilde{A}_p^{-1} M_p. \tag{38}$$

Hence, from (36) and (38), we have

$$S_0^{-1} \approx M_p^{-1} \left(\nu_0 K_p + M_{k_p} + (\tau^{-1} + \eta) M_p\right) K_p^{-1}. \tag{39}$$

The inverse of the pressure Laplacian $K_p^{-1}$ is approximated using algebraic multigrid methods, whereas the use of the Chebyshev semi-iteration will suffice for $M_p^{-1}$. We note here that, as pointed out in Chapter 5 of [14], the pressure Laplacian represents a Neumann problem because the pressure basis functions form a partition of unity. Indeed, this property is independent of the boundary conditions attached to the flow problem. To solve the problem of indefiniteness of $K_p$ we just pin a boundary node in $K_p$ (see, e.g., [8]). Afterwards, we use the AMG package provided by [9].

## 4.3 Spectral analysis

The effectiveness of the iterative solver for our KKT linear system (20) depends to a large extent on how well the exact Schur complement is represented by its approximation. To measure this, we need to consider the eigenvalues of the preconditioned Schur complement $\boldsymbol{S}_2^{-1}\tilde{\boldsymbol{S}}_2$. We are, however, unable to give a general estimate. Instead, we restrict our analysis to the regularization parameters.

**Theorem 1.** *If the system matrix $\boldsymbol{K}$ in (18) and its velocity block are invertible, then*

$$\begin{aligned} \operatorname{cond}(\boldsymbol{S}_2^{-1}\tilde{\boldsymbol{S}}_2) &\leq (1 + C_1\beta^{1/2}) \quad \textit{if} \quad \beta \ll 1, \\ \operatorname{cond}(\boldsymbol{S}_2^{-1}\tilde{\boldsymbol{S}}_2) &\leq (1 + C_2\beta^{-1/2}) \quad \textit{if} \quad \beta \gg 1. \end{aligned}$$

*Proof.* Recall first that if

$$\boldsymbol{K}^T = \begin{bmatrix} \boldsymbol{A}^T & \boldsymbol{B}^T \\ \boldsymbol{B} & 0 \end{bmatrix},$$

where $\boldsymbol{B} = I_{n_t} \otimes G_0 \otimes B$,

$$\boldsymbol{A} = I_{n_t} \otimes G_0 \otimes (\nu_0 K + M_k + \tau^{-1} M) + I_{n_t} \otimes G_1 \otimes \nu_1 K + C \otimes G_0 \otimes \tau^{-1} M,$$

and that both $\boldsymbol{K}^T$ and $\boldsymbol{A}$ are non-singular, then

$$\boldsymbol{K}^{-T} = \begin{bmatrix} \boldsymbol{A}^{-T} - \boldsymbol{A}^{-T}\boldsymbol{B}^T\boldsymbol{S}^{-1}\boldsymbol{B}\boldsymbol{A}^{-T} & \boldsymbol{A}^{-T}\boldsymbol{B}^T\boldsymbol{S}^{-1} \\ \boldsymbol{S}^{-1}\boldsymbol{B}\boldsymbol{A}^{-T} & -\boldsymbol{S}^{-1} \end{bmatrix}, \tag{40}$$

[1] We may argue that $BM^{-1}B^T \approx K_p$ as follows [14]. At the continuous level, it is clear that $-\nabla \cdot \nabla = -\nabla^2$. Since in the finite element space, $K_p$ corresponds to the operator $-\nabla^2$, $B$ represents a discretization of the negative of the divergence operator, $B^T$ corresponds to the gradient operator and $M$ relates to the identity operator, we see that the approximation of $K_p$ by $BM^{-1}B^T$ is a natural one.

and

$$\boldsymbol{K}\boldsymbol{K}^{-T} = \begin{bmatrix} \boldsymbol{A}\boldsymbol{A}^{-T}(I-\boldsymbol{P}_K)+\boldsymbol{P}_K & (\boldsymbol{A}\boldsymbol{A}^{-T}-I)\boldsymbol{B}^T\boldsymbol{S}^{-1} \\ 0 & I \end{bmatrix},$$

where $\boldsymbol{S} = \boldsymbol{B}\boldsymbol{A}^{-T}\boldsymbol{B}^T$, $\boldsymbol{P}_K = \boldsymbol{B}^\top\boldsymbol{S}^{-1}\boldsymbol{B}\boldsymbol{A}^{-T}$, and $I$ is an identity of suitable sizes, see e.g. [7]. Notice that $\boldsymbol{P}_K = \boldsymbol{P}_K^2$; that is, the matrix $\boldsymbol{P}_K$ is a projector. From (21) we have that

$$\beta^{-1}\boldsymbol{M}_{-1}\boldsymbol{K}^{-T}\boldsymbol{M}_1 = \begin{bmatrix} \boldsymbol{M}_\star & 0 \\ 0 & 0 \end{bmatrix}, \tag{41}$$

where

$$\boldsymbol{M}_\star = \beta^{-1}\mathtt{M}_{-1}\boldsymbol{K}_{11}\mathtt{M}_1, \tag{42}$$

$\mathtt{M}_{-1} = D^{-1}\otimes G_0 \otimes M$ and $\mathtt{M}_1 = D\otimes G_\alpha \otimes M$ are the velocity submatrices of $\boldsymbol{M}_{-1}$ and $\boldsymbol{M}_1$, as given by in (24) and (21) respectively, and $\boldsymbol{K}_{11} = \boldsymbol{A}^{-T}(I-\boldsymbol{P}_K)$ denotes the (1,1) block of $\boldsymbol{K}^{-T}$. Thus, using (42), (41) and (25), we get

$$\boldsymbol{S}_2 = \boldsymbol{K} + \beta^{-1}\boldsymbol{M}_{-1}\boldsymbol{K}^{-T}\boldsymbol{M}_1 = \begin{bmatrix} \boldsymbol{A}_\star & \boldsymbol{B}^T \\ \boldsymbol{B} & 0 \end{bmatrix}, \tag{43}$$

where $\boldsymbol{A}_\star = \boldsymbol{A} + \boldsymbol{M}_\star$. Next, observe from (30) that

$$\tilde{\boldsymbol{S}}_2 - \boldsymbol{S}_2 = \beta^{-1/2}(\boldsymbol{M}_{-1} + \boldsymbol{K}\boldsymbol{K}^{-T}\boldsymbol{M}_1) = \begin{bmatrix} \boldsymbol{U} & 0 \\ 0 & 0 \end{bmatrix}, \tag{44}$$

where $\boldsymbol{U} = \beta^{-1/2}\left(\mathtt{M}_{-1} + \left(\boldsymbol{A}\boldsymbol{A}^{-T}(I-\boldsymbol{P}_K)+\boldsymbol{P}_K\right)\mathtt{M}_1\right)$. Hence, using (40), (43) and (44), we have

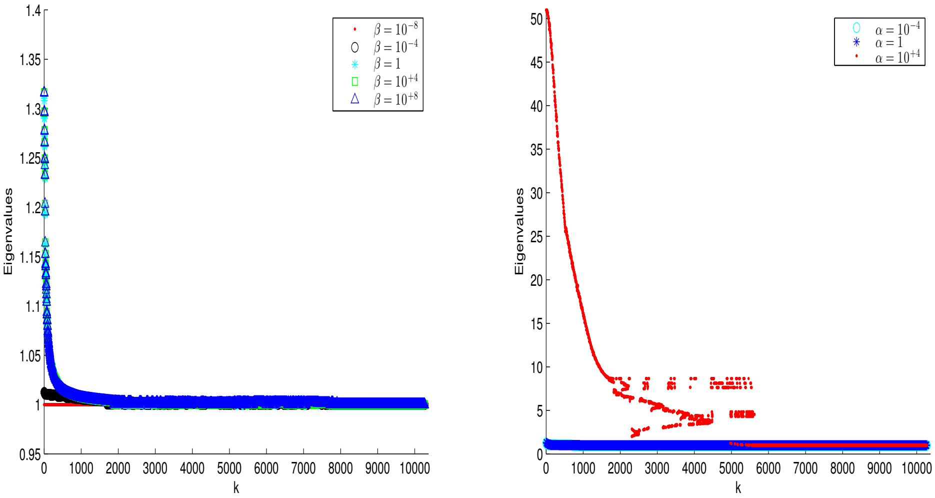
$$\begin{aligned} \boldsymbol{S}_2^{-1}\tilde{\boldsymbol{S}}_2 &= \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} + \begin{bmatrix} \boldsymbol{A}_\star & \boldsymbol{B}^T \\ \boldsymbol{B} & 0 \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{U} & 0 \\ 0 & 0 \end{bmatrix} \\ &= \begin{bmatrix} I + \boldsymbol{A}_\star^{-1}(I-\boldsymbol{P}_\star)\boldsymbol{U} & 0 \\ \boldsymbol{S}_\star^{-1}\boldsymbol{B}\boldsymbol{A}_\star^{-1}\boldsymbol{U} & I \end{bmatrix}, \end{aligned} \tag{45}$$

where $\boldsymbol{S}_\star = \boldsymbol{B}\boldsymbol{A}_\star^{-1}\boldsymbol{B}^T$ and $\boldsymbol{P}_\star = \boldsymbol{B}^T\boldsymbol{S}_\star^{-1}\boldsymbol{B}\boldsymbol{A}_\star^{-1}$ is another projector. Thus, the eigenvalues of $\boldsymbol{S}_2^{-1}\tilde{\boldsymbol{S}}_2$ are contained in the set $\{1\}\cup\sigma\left(I+\boldsymbol{A}_\star^{-1}(I-\boldsymbol{P}_\star)\boldsymbol{U}\right)$, where $\sigma(X)$ represents the spectrum of a square matrix $X$.

Now, if $\beta$ is small, then the norm of $\boldsymbol{M}_\star$ is large[2], and hence $\boldsymbol{A}_\star \approx \boldsymbol{M}_\star$. In particular, we have $\|\boldsymbol{A}_\star^{-1}\| \le C_1\beta$. Similarly, the norm of the projector $\boldsymbol{P}_\star$ is asymptotically $\beta$-independent. Finally, $\|\boldsymbol{U}\| \le C_2\beta^{-1/2}$, and $\|\boldsymbol{A}_\star^{-1}(I-\boldsymbol{P}_\star)\boldsymbol{U}\| \le C_3\beta^{1/2}$. That is, $\lambda(\boldsymbol{S}_2^{-1}\tilde{\boldsymbol{S}}_2) \in [1-C_3\beta^{1/2}, 1+C_3\beta^{1/2}] \to \{1\}$ when $\beta \to 0$.

[2] This means, from (42), that $\|\boldsymbol{M}_\star\| = \|\beta^{-1}\mathtt{M}_{-1}\boldsymbol{K}_{11}\mathtt{M}_1\| \gg \|\boldsymbol{A}\|$, which in turn implies that $\beta \ll \|\mathtt{M}_{-1}\boldsymbol{K}_{11}\mathtt{M}_1\|/\|\boldsymbol{A}\|$, where the bound depends on $J$, $n_t$ and other model and discretization parameters.

Figure 1: Eigenvalue distribution of the matrix $I+\boldsymbol{A}_\star^{-1}(I-\boldsymbol{P}_\star)\boldsymbol{U}$ using the parameters $\nu_1 = 0.1, J = 642,\ P = 4,\ n_t = 4$. Left: $\alpha = 1$ and $\beta$ is varied. Right: $\beta = 1$ and $\alpha$ is varied.



On the other hand, when $\beta$ is large, the norm of $\boldsymbol{M}_\star$ is small, and $\boldsymbol{A}_\star \approx \boldsymbol{A}$, a matrix independent of $\beta$. The only multiplication with $\beta$ comes from $\boldsymbol{U}$; therefore, $\|\boldsymbol{A}_\star^{-1}(I - \boldsymbol{P}_\star)\boldsymbol{U}\| \le C_4\beta^{-1/2} \to 0$ when $\beta \to \infty$. Again, the matrix $\boldsymbol{S}_2^{-1}\tilde{\boldsymbol{S}}_2$ becomes well conditioned in $\beta$ in the limit.

□

For intermediate $\beta$, we expect that $\tilde{\boldsymbol{S}}_2$ is still a good approximation to $\boldsymbol{S}_2$, and do observe that in practice. For small matrices we have illustrated the distribution of the eigenvalues of $I + \boldsymbol{A}_\star^{-1}(I - \boldsymbol{P}_\star)\boldsymbol{U}$ explicitly in Figure 1. As we can see from the left figure, as $\beta$ is varied, the eigenvalues are mostly clustered between 1 and 1.4, regardless of the value of $\beta$. On the other hand, Figure 1 (right) shows that, keeping $\beta = 1$, the eigenvalues of $I + \boldsymbol{A}_\star^{-1}(I - \boldsymbol{P}_\star)\boldsymbol{U}$ are clustered around 1 if $0 \le \alpha \le 1$, but drastically increase for $\alpha > 1$. This observation confirms the deterioration in the performance of our solver as $\alpha$ increases in Section 6.5 below. The scenario $\alpha \gg 1$ is not of much practical interest anyway, as this would imply a very low value of the variance, in which case we lose the point of uncertainty quantification in the problem.

# 5 Tensor Train solver

To develop an efficient tensor-based iterative solver for our problem, we separate variables $\mathbf{x}$, $\omega$ and $t$, but not the inner components of $\mathbf{x}$. In what follows, we shall rely specifically on the *Tensor Train* (TT) decomposition introduced in [38] to solve our linear systems. For our purposes, we proceed to first give a simplified presentation of the TT decomposition for three independent variables. A detailed discussion on TT decomposition can be found in recent surveys and books [19, 18, 25].

## 5.1 Tensor Train decomposition

The first operation we need for high-dimensional data is reshaping. To this end, suppose $\mathbf{y}$ is the solution of (20). Its elements can be naturally enumerated by three indices $i, j, k$, corresponding to the discretized variables $t$, $\omega$ and $\mathbf{x}$, respectively. Introducing a *multi-index*

$$\overline{ijk} = (i-1)PJ + (j-1)J + k,$$

we can denote $\mathbf{y} = \left[\mathbf{y}(\overline{ijk})\right]_{i,j,k=1}^{n_t,P,J}$, and consider $\mathbf{y}$ as a three-dimensional *tensor* with elements $\mathbf{y}(i,j,k)$. The *Tensor Train* (or simply TT) decomposition aims to approximate $\mathbf{y}$ as follows,

$$\mathbf{y}(i,j,k) \approx \sum_{s_1,s_2=1}^{r_1,r_2} \mathbf{y}^{(1)}_{s_1}(i)\mathbf{y}^{(2)}_{s_1,s_2}(j)\mathbf{y}^{(3)}_{s_2}(k) \quad \Leftrightarrow \quad \mathbf{y} \approx \sum_{s_1,s_2=1}^{r_1,r_2} \mathbf{y}^{(1)}_{s_1} \otimes \mathbf{y}^{(2)}_{s_1,s_2} \otimes \mathbf{y}^{(3)}_{s_2}. \quad (46)$$

The summation indices $r_1, r_2$ are called *TT ranks*, the factors $\mathbf{y}^{(m)}$, $m = 1,2,3$ are called *TT blocks* and have the sizes $\mathbf{y}^{(1)} \in \mathbb{R}^{n_t \times r_1}$, $\mathbf{y}^{(2)} \in \mathbb{R}^{r_1 \times P \times r_2}$ and $\mathbf{y}^{(3)} \in \mathbb{R}^{r_2 \times J}$. Notice that we can fix some of the indices, e.g. $\mathbf{y}^{(2)}(j) \in \mathbb{R}^{r_1 \times r_2}$ is a matrix *slice*, $\mathbf{y}^{(2)}_{s_1,s_2} \in \mathbb{R}^P$ is a vector, and $\mathbf{y}^{(2)}_{s_1,s_2}(j)$ is a scalar. The total number of elements in all factors is $n_t r_1 + r_1 P r_2 + r_2 J = \mathcal{O}(Jr + Pr^2)$, where $r \geq r_1, r_2$, since in our case $J \sim n_t \gg P$. Therefore, if $r \ll J$, the amount of memory consumed by the TT format is much less than $JPn_t$, needed for the full vector $\mathbf{y}$.

Particular values of $r_1, r_2$ depend on the accuracy we enforce in Eq. (46). Although it is difficult in general to estimate the TT ranks theoretically, there is a reliable numerical TT-SVD procedure, which computes a quasi-optimal TT decomposition, using a sequence of singular value decompositions (SVD) [38].

The complexity of the TT-SVD is $\mathcal{O}(J^2Pn_t)$ when we compress a full tensor. However, in the course of computations we mostly need to re-compress a tensor, given already in the TT format, but with (overly) larger ranks. For example, given a matrix as a sum of Kronecker products, $\boldsymbol{A} = \sum_{q=1}^{R} A_q \otimes B_q \otimes C_q$ and a vector $\mathbf{y}$ in the format (46), the matrix-vector product can be written as follows [47, 38],

$$\mathbf{g} = \boldsymbol{A}\mathbf{y} = \sum_{s_1,s_2=1}^{r_1,r_2} \sum_{q_1,q_2=1}^{R,R} \left(A_{q_1}\mathbf{y}^{(1)}_{s_1}\right) \otimes \left(\delta_{q_1,q_2} B_{q_1}\mathbf{y}^{(2)}_{s_1,s_2}\right) \otimes \left(C_{q_2}\mathbf{y}^{(3)}_{s_2}\right), \quad (47)$$

where $\delta_{q_1,q_2} = 1$ if $q_1 = q_2$ and zero otherwise. Similarly, a linear combination $\mathbf{y} + \mathbf{g}$ of vectors can be recast to their TT blocks $\mathbf{y}^{(m)}, \mathbf{g}^{(m)}$. Each bracket in the right-hand side of (47) is a larger TT block, the new rank indices are $s_1' = \overline{s_1 q_1}$, $s_2' = \overline{s_2 q_2}$, and hence the TT ranks are $Rr_1, Rr_2$. However, $\mathbf{g}$ might be approximated accurately enough with much smaller ranks. When applied to the TT format (47) instead of the full tensor, the TT-SVD requires $\mathcal{O}(JR^2r^2 + PR^3r^3)$ operations. These properties allow to adopt classical iterative methods such as MINRES or GMRES in an *inexact* fashion, keeping all Krylov vectors in the TT format and performing the TT-SVD re-compression (or TT truncation) [30, 4, 1, 11].

## 5.2 Alternating iterative methods

Notwithstanding the TT truncation, the Krylov vectors may still develop rather large TT ranks – much larger than the ranks of the exact solution, in particular. Unless a very good preconditioner is available, such that the method converges in about 10 iterations, the TT-GMRES approach may become too expensive. For problems of some special forms, such as the Lyapunov equations, one can employ ADI [53] or tensor product Krylov methods [29]. For more general problems we have to employ more general alternating methods [21, 47].

The main idea behind the alternating tensor methods is to reduce the problem to the elements of a particular TT block and iterate over different TT blocks until convergence is achieved. In the mathematical community, the concept started with the Alternating Least Squares (ALS) method used to minimize the misfit of a tensor by a low-rank tensor model, see the surveys [27, 18]. This was later extended to the solution of linear systems [21, 40]. In quantum physics, a powerful realization of the alternation idea is the Density Matrix Renormalization Group (DMRG) algorithm [54], which is mainly used for eigenvalue problems, but also for linear systems [24]. Later on, the ALS/DMRG methods were combined with the classical gradient descent iteration: besides the ALS iteration, the TT blocks are explicitly augmented by the partial TT format of the residual surrogate. The DMRG algorithm with a single center site [55] uses the surrogate of the Krylov vector, and the Alternating Minimal Energy (AMEn) method [13] uses the actual residual, which was later adopted for eigenvalue problems as well [23, 28]. Details of these algorithms can be found in the corresponding papers. Here we give a brief idea of the AMEn algorithm, adapted to 3-dimensional tensors, and then extended for saddle-point systems.

Let us consider a linear system $\boldsymbol{A}\mathbf{y} = \mathbf{g}$, where $\mathbf{y}$ is sought in the TT format (46), and some initial guess for $\mathbf{y}$ is given. The ALS method reduces this system to the elements of a chosen TT block $\mathbf{y}^{(m)}$ in the course of iteration $m = 1, 2, 3$. Using the multi-indices, we may stretch the TT block to a vector, for example, as follows: $\mathbf{y}^{(2)}(\overline{s_1 j s_2}) = \mathbf{y}^{(2)}_{s_1,s_2}(j)$. Notice that the TT format is linear w.r.t. each particular TT block, i.e. we can write

$$\mathbf{y} = \boldsymbol{Y}_1 \mathbf{y}^{(1)} = \boldsymbol{Y}_2 \mathbf{y}^{(2)} = \boldsymbol{Y}_3 \mathbf{y}^{(3)},$$

where the *frame* matrices $\boldsymbol{Y}_m$, $m = 1, 2, 3$, are constructed as follows:

$$\begin{aligned}
\boldsymbol{Y}_1 &= I_{n_t} \otimes \sum_{s_2=1}^{r_2} \left(\mathbf{y}^{(2)}_{s_2}\right)^\top \otimes \left(\mathbf{y}^{(3)}_{s_2}\right)^\top \in \mathbb{R}^{n_t PJ \times n_t r_1}, \\
\boldsymbol{Y}_2 &= \mathbf{y}^{(1)} \otimes I_P \otimes \left(\mathbf{y}^{(3)}\right)^\top \in \mathbb{R}^{n_t PJ \times r_1 P r_2}, \\
\boldsymbol{Y}_3 &= \sum_{s_1=1}^{r_1} \mathbf{y}^{(1)}_{s_1} \otimes \mathbf{y}^{(2)}_{s_1} \otimes I_J \in \mathbb{R}^{n_t PJ \times r_2 J}.
\end{aligned} \tag{48}$$

Recall from (46) that $\mathbf{y}^{(2)}_{s_2} \in \mathbb{R}^{r_1 \times P}$ and $\mathbf{y}^{(2)}_{s_1} \in \mathbb{R}^{P \times r_2}$. In other words, each frame matrix $\boldsymbol{Y}_m$ constitutes the TT format (46) with the block $\mathbf{y}^{(m)}$ replaced by the identity matrix of the corresponding size. These frame matrices are used to project the linear system. The ALS method updates the TT format by solving the following Galerkin systems

$$\left(\boldsymbol{Y}_1^\top \boldsymbol{A} \boldsymbol{Y}_1\right) \mathbf{y}^{(1)} = \boldsymbol{Y}_1^\top \mathbf{g}, \tag{49}$$

$$\left(\boldsymbol{Y}_2^\top \boldsymbol{A} \boldsymbol{Y}_2\right) \mathbf{y}^{(2)} = \boldsymbol{Y}_2^\top \mathbf{g}, \tag{50}$$

$$\left(\boldsymbol{Y}_3^\top \boldsymbol{A} \boldsymbol{Y}_3\right) \mathbf{y}^{(3)} = \boldsymbol{Y}_3^\top \mathbf{g}, \tag{51}$$

and so on from the first step. Using the QR decompositions of the properly reshaped TT blocks, it is easy to make the frame matrices orthogonal, and therefore preserve the stability of the Galerkin systems, if $\boldsymbol{A}$ is positive definite. For example, it is enough to make $\mathbf{y}^{(1)}$ column-orthogonal and $\mathbf{y}^{(3)}$ row-orthogonal to make the whole $\boldsymbol{Y}_2$ (column-)orthogonal. Since this step is never a bottleneck, we always assume that the frame matrices are orthogonal, before solving (49)–(51).

However, the convergence of this algorithm is questionable. It is possible that the systems (49)–(51) remain the same within machine precision in two consecutive iterations, while the true residual of the initial linear system $\mathbf{g} - \boldsymbol{A}\mathbf{y}$ is large. The AMEn algorithm [13] was developed to circumvent this problem. In addition to the solution, we approximate the residual in the TT format,

$$\mathbf{g} - \boldsymbol{A}\mathbf{y} \approx \mathbf{z} = \sum_{\zeta_1, \zeta_2 = 1}^{\rho_1, \rho_2} \mathbf{z}^{(1)}_{\zeta_1} \otimes \mathbf{z}^{(2)}_{\zeta_1, \zeta_2} \otimes \mathbf{z}^{(3)}_{\zeta_2}. \tag{52}$$

A very low accuracy is often sufficient for the residual (in our experiments we use $\rho_1 = \rho_2 = 2$), so we can use the simple ALS method to approximate the residual. Along the lines of (48), we construct the orthogonal *residual frame* matrices $\boldsymbol{Z}_m$ from (52) and compute $\mathbf{z}^{(m)} = \boldsymbol{Z}_m^\top (\mathbf{g} - \boldsymbol{A}\mathbf{y})$ in a sequence $m = 1, 2, 3$, and so on. Since both $\boldsymbol{A}$ and $\mathbf{g}$ are given in the TT format, this computation is inexpensive. Moreover, it is enough to compute $\mathbf{z}^{(m)}$ only once after the $m$-th step of (49)–(51), i.e. perform one ALS iteration for $\mathbf{z}$ whenever the solution changes.

The crucial step now is the *enrichment* of the solution. Having solved (49), for example, we concatenate $\mathbf{y}^{(1)}$ and $\mathbf{z}^{(1)}$ as follows,

$$\mathbf{y}^{(1)}_{s_1'}(i) = \begin{cases} \mathbf{y}^{(1)}_{s_1'}(i), & s_1' = 1, \dots, r_1, \\ \mathbf{z}^{(1)}_{s_1'-r_1}(i), & s_1' = r_1+1, \dots, r_1+\rho_1, \end{cases}$$

and so on. The enrichment has a two-fold consequence. First, the residual can be well approximated in the basis of columns of the frame matrices, which prevents the Galerkin projection from a premature stagnation. Second, we can start from a low-rank initial guess and increase the TT ranks gradually, preventing them from becoming significantly larger than the ranks of the exact solution.

## 5.3 Block alternating iteration

The AMEn method performs well for positive definite matrices. However, if we try to solve the KKT system (20) with the saddle-point matrix, the method may fail. The Galerkin projections (49)–(51) obey the Poincaré Separation Theorem [22, Section 4.3], and since the spectrum has both positive and negative parts, some of the eigenvalues may interlace to zero. Consequently, the projected matrices become degenerate and the calculation stops.

To avoid this problem, we store the state $\mathbf{y}$, control $\mathbf{u}$ and adjoint $\boldsymbol{\lambda}$ vectors in the *shared*, or *block* TT format [12], and preserve the KKT structure in the reduced system. Suppose that $\mathbf{y}, \mathbf{u}, \boldsymbol{\lambda}$ have the same sizes (although this is not the case initially, we will make it true in the next subsection), and collect them into a long vector $\mathbf{w}^\top = [\mathbf{w}_1^\top, \mathbf{w}_2^\top, \mathbf{w}_3^\top] = [\mathbf{y}^\top, \mathbf{u}^\top, \boldsymbol{\lambda}^\top]$. The block TT format for $\mathbf{w}$ can now be written in either of three variants:

$$\mathbf{w}_l(i,j,k) = \sum_{s_1,s_2=1}^{r_1,r_2} \mathbf{w}^{(1)}_{s_1}(i,l)\mathbf{w}^{(2)}_{s_1,s_2}(j)\mathbf{w}^{(3)}_{s_2}(k), \tag{53}$$

$$\mathbf{w}_l(i,j,k) = \sum_{s_1,s_2=1}^{r_1,r_2} \mathbf{w}^{(1)}_{s_1}(i)\mathbf{w}^{(2)}_{s_1,s_2}(j,l)\mathbf{w}^{(3)}_{s_2}(k), \tag{54}$$

$$\mathbf{w}_l(i,j,k) = \sum_{s_1,s_2=1}^{r_1,r_2} \mathbf{w}^{(1)}_{s_1}(i)\mathbf{w}^{(2)}_{s_1,s_2}(j)\mathbf{w}^{(3)}_{s_2}(k,l). \tag{55}$$

The only difference between these three variants is in which TT block the index $l$ ($l = 1,2,3$) is placed, but we need these different representations in different AMEn steps, as explained below. It is easy to switch between the representations using SVD [12]. Given the variant (53), we reshape $\mathbf{w}^{(1)}$ to a matrix $W^{(1)} \in \mathbb{R}^{n_t \times 3r_1}$, compute the truncated SVD, namely, $W^{(1)} \approx U\Sigma V^\top$, where $U \in \mathbb{R}^{n_t \times r_1'}$, so the elements of $U$ can be enumerated by two indices, $U(i, s_1')$, $i = 1, \dots, n_t$, $s_1' = 1, \dots, r_1'$. Therefore, $\mathbf{w}^{(1)}$ in (54) or (55) can be replaced by $U$. Then the matrix $\Sigma V^\top$ is reshaped to a matrix $R \in \mathbb{R}^{3r_1' \times r_1}$, indexed as $R(\overline{ls_1'}, s_1)$, and multiplied with $\mathbf{w}^{(2)}$ as follows:

$$\hat{\mathbf{w}}^{(2)}_{s_1',s_2}(j,l) := \sum_{s_1=1}^{r_1} R(\overline{ls_1'}, s_1)\mathbf{w}^{(2)}_{s_1,s_2}(j).$$

We notice that the result $\hat{\mathbf{w}}^{(2)}$ can overwrite $\mathbf{w}^{(2)}$ in (54), since it has the same form. In the same way, we can convert (54) to (55), or the other way around. Generally, the TT ranks change after such transformations. However, in the numerical practice the ranks remain comparatively the same in different block representations.

The transition from one block variant to another is performed routinely in the AMEn iteration. Note that each of the variants (53)–(55) induces *only one* frame matrix $\boldsymbol{W}_m$ of the form (48), since the frame matrices do not depend on $l$:

$$\boldsymbol{W}_1 = I_{n_t} \otimes \sum_{s_2=1}^{r_2} \left(\mathbf{w}_{s_2}^{(2)}\right)^\top \otimes \left(\mathbf{w}_{s_2}^{(3)}\right)^\top,$$

$$\boldsymbol{W}_2 = \mathbf{w}^{(1)} \otimes I_P \otimes \left(\mathbf{w}^{(3)}\right)^\top,$$

$$\boldsymbol{W}_3 = \sum_{s_1=1}^{r_1} \mathbf{w}_{s_1}^{(1)} \otimes \mathbf{w}_{s_1}^{(2)} \otimes I_J.$$

Therefore, to assemble the first reduced system (49) we need the first block representation (53), for the second system (50) we need (54), and so on. However, each frame matrix has the column size $JPn_t$, which coincides with the sizes of *submatrices* of (20), not the whole KKT matrix. Besides, we need a system of equations w.r.t. the index $l$, carried in the TT block under consideration. Thus, a natural generalization of (49)–(51) is the following,

$$\begin{bmatrix} \boldsymbol{W}_m^\top \tau \boldsymbol{M}_1 \boldsymbol{W}_m & 0 & -\boldsymbol{W}_m^\top \boldsymbol{K}^\top \boldsymbol{W}_m \\ 0 & \boldsymbol{W}_m^\top \beta\tau \boldsymbol{M}_2 \boldsymbol{W}_m & \boldsymbol{W}_m^\top \boldsymbol{N}^\top \boldsymbol{W}_m \\ -\boldsymbol{W}_m^\top \boldsymbol{K} \boldsymbol{W}_m & \boldsymbol{W}_m^\top \boldsymbol{N} \boldsymbol{W}_m & 0 \end{bmatrix} \mathbf{w}^{(m)} = \begin{bmatrix} \boldsymbol{W}_m^\top \tau \boldsymbol{M}_a \bar{\mathbf{y}} \\ 0 \\ \boldsymbol{W}_m^\top \mathbf{g} \end{bmatrix}, \tag{56}$$

for $m = 1, 2, 3$. After this system is solved, we use the SVD procedure outlined above to switch to the next block TT representation, compute the residual and enrich the new $\mathbf{w}^{(m)}$ (which does not contain $l$ anymore). The residual is also kept in the block form, $\mathbf{z}^\top = [\mathbf{z}_1^\top, \mathbf{z}_2^\top, \mathbf{z}_3^\top]$, where $\mathbf{z}_l$ denotes the residual in the $l$-th row of the KKT system (20), and is approximated in the appropriate block TT representation. Its active block is computed as $\mathbf{z}^{(m)}(l) = \boldsymbol{Z}_m^\top(\mathbf{g}_l - \mathfrak{A}_{l,:}\mathbf{w})$, and then the index $l$ is replaced to the next TT block by the same SVD procedure.

Since each of the submatrices $\boldsymbol{M}_1$, $\boldsymbol{M}_2$ is symmetric and semidefinite, the same property is inherited by the corresponding blocks in (56). However, $\boldsymbol{K}$ is the Stokes-Brinkman matrix, which is indefinite. We could consider the 2 × 2 Stokes-Brinkman block structure and the 3 × 3 KKT structure on the same level, and solve the 5 × 5 block system. However, the second row of the Stokes-Brinkman matrix has a very particular meaning, which we can exploit to reduce the complexity in what follows.

### 5.4 Pressure elimination in the reduced model

The low-rank separation of space and time variables has been used for a while in the numerical simulation of the Navier-Stokes equation. The Proper Orthogonal Decomposition (POD) is a well-known approach to model reduction [31]. It reshapes the *velocity* component of the solution to a matrix $Y = [\mathbf{y}(\overline{ij}, k)]$, computes the truncated SVD $Y \approx U\Sigma V^\top$, and uses the columns of $V$ for the Galerkin reduction of the velocity operators. If we were solving the continuous equation, we would have a vector-valued function $V = V(x) \in \mathbb{R}^r$, where $r$ is the number of POD terms, and the reduced solution sought in the form $y(x,t) \approx V(x)a(t)$. Plugging this into the Stokes-Brinkman equation, and projecting the velocity equation onto $V$, we have

$$\begin{cases} \frac{da}{dt} - \nu\langle V^\top, \Delta V\rangle a + \langle V^\top, K_0 V\rangle a + \langle V^\top, \nabla p\rangle & = & \langle V^\top, u\rangle, \\ \nabla \cdot Va & = & 0. \end{cases}$$

Since $a(t)$ is not fixed a priori, from the second row we have $\nabla \cdot V(x) = 0$. However, then in the first row $\langle V^\top, \nabla p\rangle = -\langle \nabla \cdot V^\top, p\rangle = 0$; that is, the reduced model contains no pressure at all. In the discrete formulation, we have the system (15), and the pressure part $V^\top \mathcal{B}^\top \mathbf{p}$ is not exactly zero due to the boundary conditions. Nevertheless, it is often heuristically assumed that its magnitude is small [3]. If it is not the case, there are nonlinear corrections available [37]. They are important for the POD approach, since the last step there is the solution of the time-dependent reduced model. However, the alternating methods are different: we may stop the iteration at the spatial TT block and return the block TT format of the form (55), instead of (53) in the POD counterpart. Therefore, we perform the pressure exclusion trick (even if $V^\top \mathcal{B}^\top \mathbf{p}$ is not small) differently.

When we solve (56) for the spatial TT block ($m = 3$), we consider the $5 \times 5$ Stokes-KKT structure

$$\begin{bmatrix} \tau\hat{M}_1 & 0 & 0 & -\hat{A} & -\hat{B}^\top \\ 0 & 0 & 0 & -\hat{B} & 0 \\ 0 & 0 & \beta\tau\hat{M}_2 & \hat{N}^\top & 0 \\ -\hat{A} & -\hat{B}^\top & \hat{N} & 0 & 0 \\ -\hat{B} & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \hat{\mathbf{w}}^{(3)}(1) \\ \hat{\mathbf{w}}^{(3)}(2) \\ \hat{\mathbf{w}}^{(3)}(3) \\ \hat{\mathbf{w}}^{(3)}(4) \\ \hat{\mathbf{w}}^{(3)}(5) \end{bmatrix} = \begin{bmatrix} \hat{\mathbf{b}}_\mathbf{1} \\ 0 \\ 0 \\ \hat{\mathbf{g}}_\mathbf{v} \\ \hat{\mathbf{g}}_\mathbf{p} \end{bmatrix}, \tag{57}$$

where $\hat{M}_1 = \widehat{D}_\alpha \otimes M$, $\hat{M}_2 = \widehat{D}_0 \otimes M$, $\hat{N} = \widehat{I}_0 \otimes M$, $\hat{B} = \widehat{I}_0 \otimes B$,

$$\hat{A} = \widehat{I}_0 \otimes \left(\tau^{-1}M + \nu_0 K + M_k\right) + \widehat{I}_1 \otimes \nu_1 K + \widehat{C}_0 \otimes \tau^{-1} M,$$

the reduced matrices corresponding to the time $t$ and the event $\omega$ are computed as

$$\begin{aligned} &\widehat{I}_0 = \mathcal{W}_3^\top \left(I \otimes G_0\right) \mathcal{W}_3, \quad \widehat{I}_1 = \mathcal{W}_3^\top \left(I \otimes G_1\right) \mathcal{W}_3, \quad \widehat{C}_0 = \mathcal{W}_3^\top \left(C \otimes G_0\right) \mathcal{W}_3, \\ &\widehat{D}_0 = \mathcal{W}_3^\top \left(D \otimes G_0\right) \mathcal{W}_3, \quad \widehat{D}_\alpha = \mathcal{W}_3^\top \left(D \otimes G_\alpha\right) \mathcal{W}_3, \end{aligned} \tag{58}$$

whereas the right-hand side parts are

$$\hat{\mathbf{b}}_\mathbf{1} = \mathcal{W}_3^\top \left(D\mathbf{e} \otimes G_0 \mathbf{e}_1\right) \otimes \tau\bar{\mathbf{v}}, \qquad \begin{bmatrix} \hat{\mathbf{g}}_\mathbf{v} \\ \hat{\mathbf{g}}_\mathbf{p} \end{bmatrix} = \mathcal{W}_3^\top \left(\mathbf{e} \otimes \mathbf{e}_1\right) \otimes \begin{bmatrix} \mathbf{g}_v^0 \\ \mathbf{g}_p^0 \end{bmatrix},$$

and $\mathcal{W}_3 = \sum\limits_{s_1=1}^{r_1} \mathbf{w}_{s_1}^{(1)} \otimes \mathbf{w}_{s_1}^{(2)} \in \mathbb{R}^{n_t P \times r_2}$ is a chunk of the frame matrix $\boldsymbol{W}_3$. We had to introduce this chunk and the Kronecker structures above in order to explain the preconditioner in the next section. We see that the solution components $\hat{\mathbf{w}}^{(3)}(2)$ and $\hat{\mathbf{w}}^{(3)}(5)$ denote the state and adjoint pressures, respectively. The new TT block is assembled from the remaining components only, $\mathbf{w}^{(3)} = \left[\hat{\mathbf{w}}^{(3)}(1), \hat{\mathbf{w}}^{(3)}(3), \hat{\mathbf{w}}^{(3)}(4)\right]$. For the subsequent AMEn steps ($m = 2, 1$), we do not assume the pressure components to be small, but we assume that they will *not change* significantly. Therefore, their contributions to the velocity equations can be recast to the right-hand side. More precisely, we construct the TT formats

$$\delta\mathbf{b_1} = \sum_{s_1,s_2} \mathbf{w}^{(1)}_{s_1} \otimes G_0 \mathbf{w}^{(2)}_{s_1,s_2} \otimes B^\top \hat{\mathbf{w}}^{(3)}_{s_2}(5), \quad \delta\mathbf{g} = \sum_{s_1,s_2} \mathbf{w}^{(1)}_{s_1} \otimes G_0 \mathbf{w}^{(2)}_{s_1,s_2} \otimes B^\top \hat{\mathbf{w}}^{(3)}_{s_2}(2),$$

and correct the right-hand side of (20) as follows,

$$\begin{bmatrix} \mathbf{b_1} \\ 0 \\ \mathbf{g} \end{bmatrix} \quad \rightarrow \quad \begin{bmatrix} \mathbf{b_1} + \delta\mathbf{b_1} \\ 0 \\ \mathbf{g} + \delta\mathbf{g} \end{bmatrix}.$$

After that, we conduct AMEn steps $m = 2, 1, 2$ with the system of the form (56), where $\boldsymbol{K}$ contains now only the velocity equation, and hence is positive definite. When we come back to $m = 3$, we drop the right-hand side corrections and solve the full system (57). If we are to stop the iteration, we return the full solution, including $\hat{\mathbf{w}}^{(3)}(2)$ and $\hat{\mathbf{w}}^{(3)}(5)$. Due to the Galerkin projection, the accuracy depends only on how good the common TT blocks $\mathbf{w}^{(1)}$ and $\mathbf{w}^{(2)}$ represent all solution components. Although it is unclear whether it is allowed in general to 'freeze' some components, in our numerical experiments we observed that the solution is accurate enough; that is, the blocks $\mathbf{w}^{(1)}$ and $\mathbf{w}^{(2)}$ are computed accurately using only the velocity information.

### 5.5 Practical implementation

The preconditioner developed in Section 4.1 needs to be adjusted to the local problem (57). Although the reduced matrices (58) are small, they are dense, and it is impractical to compute the blocks of (57) explicitly. However, note that all of them are single Kronecker products except $\hat{A}$. Moreover, if the norms of $K$ and $M_k$ are sufficiently large, and $\nu_1$ is small, then the first term in $\hat{A}$ dominates. Therefore, we replace $\hat{A}$ by its first term $\widehat{I}_0 \otimes \left(\tau^{-1}M + \nu_0 K + M_k\right)$ during the preconditioning. This also allows to avoid the second level of preconditioning for the Stokes-Brinkman system (33). Since $\hat{B}$ contains $\widehat{I}_0$, we can assemble the Stokes-Brinkman matrix in the Kronecker form as well,

$$\hat{\mathcal{K}} = \widehat{I}_0 \otimes \begin{bmatrix} \tau^{-1}M + \nu_0 K + M_k & B^\top \\ B & 0 \end{bmatrix}.$$

In the computation of $\mathbf{x}_3$ in an analog of (29), we can solve linear systems with $\hat{\mathcal{K}}$ directly. For two-dimensional cases, this approach is faster than iterations with (33). In the same way we approximate the factors of the Schur complement (31), e.g.

$$\hat{\mathcal{K}}^\top + \hat{\mathcal{M}}_r \approx \widehat{I}_0 \otimes \begin{bmatrix} \left(\frac{1}{\tau} + \frac{1}{\sqrt{\beta}} \frac{\|\widehat{D}_\alpha\|}{\|\widehat{I}_0\|}\right) M + \nu_0 K + M_k & B^\top \\ B & 0 \end{bmatrix}, \tag{59}$$

where we approximated $\hat{\mathcal{M}}_r = \frac{1}{\sqrt{\beta}}\widehat{D}_\alpha \otimes M$ by $\widehat{I}_0 \frac{\|\widehat{D}_\alpha\|}{\|\widehat{I}_0\|\sqrt{\beta}} \otimes M$, and $\widehat{D}_\alpha$ and $\widehat{I}_0$ are defined in (58). For three dimensions (Section 6.10), the matrices become more dense, and we have to use iterative methods, preconditioning the velocity block by a multigrid cycle. Similar rank-1 approximation is performed for the TT blocks $\mathbf{w}^{(1)}$ and $\mathbf{w}^{(2)}$. Although they are smaller than the spatial block, they are still rather large to form and solve the systems (56) directly. The crucial point here, fortunately, is that the new preconditioner does not need to invert $\boldsymbol{M}_1$.

# 6 Numerical experiments

A systematic study of the proposed technique will be conducted on two- and three-dimensional examples. We first consider the Stokes(-Brinkman) flow constraints on $\mathcal{D} = [0,1]^2$ with the inflow boundary conditions

$$v_1|_{x_1=0} = x_2(1-x_2), \quad v_2|_{x_1=0} = 0, \qquad v|_{x_2=0} = v|_{x_2=1} = 0,$$

and 'do-nothing' boundary conditions at $x_1 = 1$. The velocity operators are discretized with the *mini* elements [48] and the pressure operators are discretized with the piecewise linear finite elements. The stiffness matrices are assembled in FEniCS 1.5.0 package [34]. For the Stokes-Brinkmann equation, the coefficient is chosen as follows:

$$K_0(\mathbf{x}) = \begin{cases} \bar{K}_0, & (x_1-0.5)^2 + (x_2-0.5)^2 \leq 0.15^2, \\ 0, & \text{otherwise.} \end{cases}$$

The right-hand side and the initial condition are zeros. The desired state is the deterministic stationary solution of the forward Stokes-Brinkman problem.

The model is characterized by 8 parameters: the spatial grid size $J$, the number of time steps $n_t$, the time interval $T$, regularization parameters $\alpha$ and $\beta$, variance $\nu_1$, a threshold for the tensor approximation and the AMEn algorithm $\varepsilon$, and the porosity coefficient $\bar{K}_0$. For the sake of brevity, we perform 8 experiments, fixing all parameters to their default values and varying only one of them. The default parameters are the following: one-dimensional spatial grid size $n = 64$ (so that $J = 29059$), time grid size $n_t = 2^{10}$, time interval $T = 1$, regularization parameters $\beta = 10^{-6}$ and $\alpha = 1$, variance parameter[3] $\nu_1 = 0.1$, approximation tolerance $\varepsilon = 10^{-6}$, and pure Stokes coefficient $\bar{K}_0 = 0$. The mean viscosity is always fixed at $\nu_0 = 1$, since the behavior of the model is the same if $\nu_0 \sim 1/T$, so we can investigate either of these parameters. The stochastic polynomial degree $P = 16$.

We investigate several kinds of discrepancies, such as the residual, the misfit w.r.t. the desired state, and so on. Therefore, it is convenient to introduce a unifying notation. All errors are measured in the Frobenius norm, i.e. given the reference $\mathbf{y}_\star$ and

[3] In applications involving highly heterogeneous media, such as subsurface diffusion, the variance of a random field may be several orders in magnitude. However, a highly viscous fluid is more or less homogeneous, and the 10% variance is realistic. This is the case in biomedical modeling, for example.

Table 1: 2D Stokes, comparison of spatial preconditioners

| | $\boldsymbol{P}_1$ | | $\boldsymbol{P}_2$ | |
|---|---|---|---|---|
| $\beta$ | Iterations | CPU time | Iterations | CPU time |
| $10^{-2}$ | 1264 | 6197 | 194 | 2015 |
| $10^{-4}$ | 738 | 3700 | 201 | 1968 |
| $10^{-6}$ | 196 | 759 | 108 | 700 |
| $10^{-8}$ | 163 | 465 | 72 | 322 |

trial $\mathbf{y}$ vectors, we compute

$$\mathcal{E}(\mathbf{y}, \mathbf{y}_\star) = \|\mathbf{y} - \mathbf{y}_\star\|_F / \|\mathbf{y}_\star\|_F. \tag{60}$$

By 'residual', we mean the maximal relative residual among the KKT system rows:

$$\text{residual} = \max\left(\mathcal{E}(\tau \boldsymbol{M}_1 \mathbf{y} - \boldsymbol{K}^\top \boldsymbol{\lambda}, \tau \boldsymbol{M}_a \bar{\mathbf{y}});\ \mathcal{E}(\tau\beta \boldsymbol{M}_2 \mathbf{u}, \mathbf{N}^\top \boldsymbol{\lambda});\ \mathcal{E}(-\boldsymbol{K}\mathbf{y} + \boldsymbol{N}\mathbf{u}, \mathbf{g})\right).$$

Since the KKT matrix is rather ill-conditioned, we also estimate the Frobenius-norm errors of the state and control components of the solution as follows. For each experiment, we solve the system with two thresholds, $\varepsilon$ and $0.1\varepsilon$. The solution components of the latter run, denoted as $\mathbf{y}_\star$ and $\mathbf{u}_\star$, are taken as the reference ones, and the relative errors are computed by (60).

The complexity indicators are the CPU time, memory consumption and the number of iterations. The CPU time is measured for a sequential MATLAB R2012b program, run under Linux at Intel Xeon X5650 CPU with 2.67GHz. The TT algorithms are implemented within the TT-Toolbox [39]. The memory consumption is reported as the memory compression ratio by the TT format. It is computed as the number of TT elements over the total number of degrees of freedom in the solution, i.e.

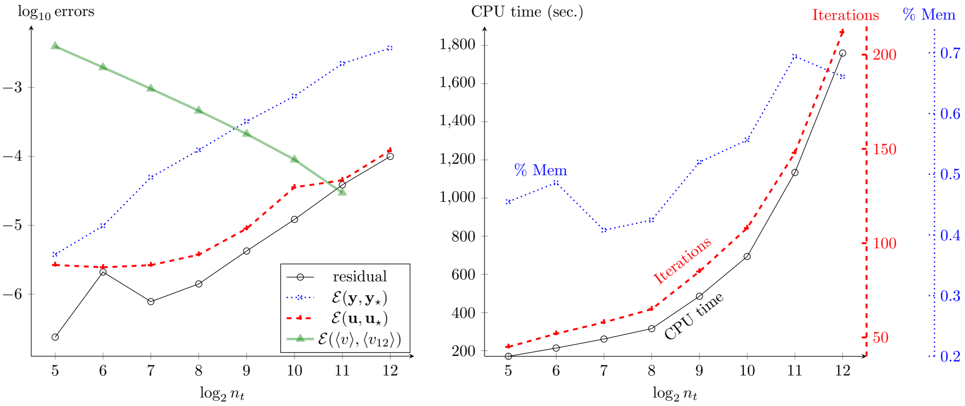
$$\%\ \text{Mem} = \frac{n_t r_1 + r_1 P r_2 + r_2 J}{J P n_t} \cdot 100.$$

By 'iterations', we mean the total number of FGMRES iterations, spent in solving the reduced systems (57) for the spatial TT block, in all AMEn steps. The FGMRES is used with the block-triangular preconditioner (29) for the KKT level only (the Stokes-like systems (59) are solved directly in two-dimensional examples).

## 6.1 Performance of the new block-triangular preconditioner

It is illustrative to compare the new preconditioner (29) with the established block-diagonal preconditioner $\boldsymbol{P}_1$ from [50], mentioned at the beginning of Section 4. We test $\boldsymbol{P}_1$ using the MINRES method, for the spatial TT block only. The comparison with $\boldsymbol{P}_2$ (29) is given in Table 1. We see that $\boldsymbol{P}_2$ provides faster convergence in terms of both iterations and time. Therefore, we use it in all the remaining experiments in this paper.

Figure 2: 2D Stokes, experiment with $n_t$. Left: Residual, errors w.r.t. the reference solutions, and the mean value error w.r.t. the time grid level. Right: CPU time, total number of iterations in spatial systems, memory compression ratio.



## 6.2 Experiment with $n_t$ (Figure 2)

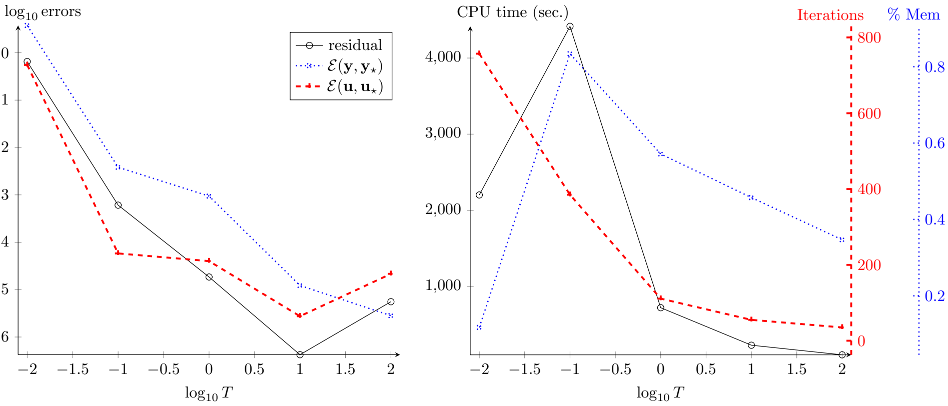
In the first test, we vary the number of time steps from $2^5$ to $2^{12}$. In addition to general errors, we report also the convergence of the mean value of the velocity with the time grid refinement. The mean value is computed over all variables:

$$\langle v\rangle = \frac{\mathcal{T}}{T}\sum_{k,k',i=1}^{J_v,J_v,n_t} M(k,k')D(i,i)y(i,1,k') \approx \int_{\mathcal{D}}\int_{\Omega}\frac{1}{T}\int_0^T v(\mathbf{x},\omega,t)dtd\mathbb{P}(\omega)d\mathbf{x}.$$

Note that $\mathbf{y}$ has the form $[\mathbf{v},\mathbf{p}]$ w.r.t. the index $k$, so that the summation $k,k' = 1,\dots,J_v$ extracts only the velocity. The reference value $\langle v_{12}\rangle$ is computed on the grid $n_t = 2^{12}$. The distance from $\langle v\rangle$ decays proportionally to $2^{-n_t}$, as expected for the Euler scheme.

The errors grow proportionally to the grid size, since the matrix becomes more ill-conditioned. However, the CPU times and the numbers of iterations grow only as a small power of $\log n_t$. The behavior of the CPU time is very close to the behavior of the iterations, while the TT ranks (and hence the memory) are almost stable w.r.t. $n_t$. This shows that the main reason for the increase in time is the deterioration of the quality of the preconditioner (since we use the rank-1 approximation (59)). A more robust (in terms of iterations) preconditioner should also involve the term related to the time derivative. However, each iteration might become more costly. Future research is needed to make the preconditioner suitable for extreme parameters.

Figure 3: 2D Stokes, experiment with $T$. Left: Residual and errors w.r.t. the reference solutions. Right: CPU time, total number of iterations in spatial systems, memory compression ratio.



## 6.3 Experiment with $T$ (Figure 3)

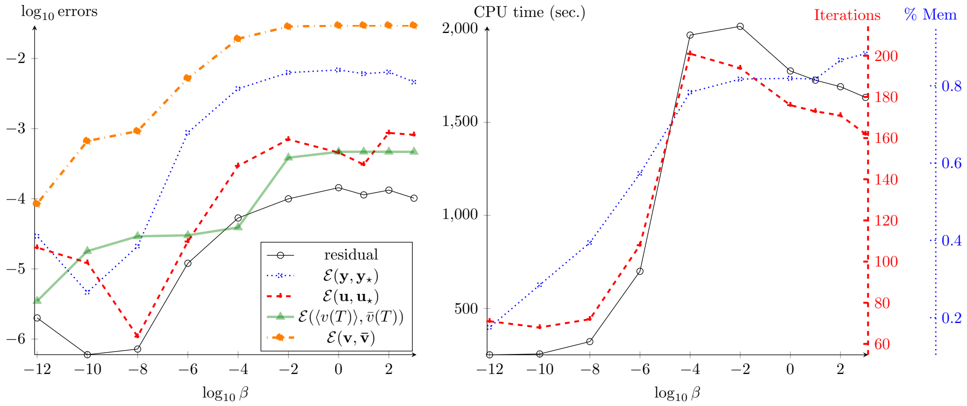
Since the initial condition is zero, while the desired state is not for any time step, the time interval influences the model significantly. The smaller is the interval, the larger the force (in our terminology, control) that must be exerted to drive the system to the desired state. This is true not only for the physical behavior, but also for the computational efforts required to solve the system. For $T = 0.01$, the matrix becomes too ill-conditioned, and 800 iterations are not enough to compute the spatial TT block accurately enough. For larger $T$, both the error and the complexity decrease.

## 6.4 Experiment with $\beta$ (Figure 4)

Although there are rigorous mathematical ways to estimate $\beta$ for a given problem, such as the L-curve analysis [20] or the discrepancy principle [15], we do not follow them here for a couple of reasons. First, the value of $\beta$ may be suggested by the physical considerations (i.e. the maximal force available). Second, we want to demonstrate robustness of our approach for as wide range as possible. Therefore, we vary $\beta$ from $10^{-12}$ to $10^3$.

We see that the errors are smaller for smaller $\beta$ and stabilize at some levels when $\beta$ increases. When $\beta$ is small, the model reconstructs the deterministic Stokes solution quite accurately, as can be seen from the discrepancy $\mathcal{E}(\mathbf{v}, \bar{\mathbf{v}})$. In addition, we report the deviation of the mean solution at the final time from the desired state. This quantity is much smaller and less dependent on $\beta$ than the global misfit: since the initial state is zero, the misfit in the first time steps will always be rather large, but in the latter steps the systems converges to the stationary solution. From the complexity figure, we see that the most difficult are the cases with intermediate $\beta$. The memory consumption increases with $\beta$, since the solution drives away from the rank-1 desired

Figure 4: 2D Stokes, experiment with $\beta$. Left: Residual and errors w.r.t. the reference solutions, and the distance to the desired state. Right: CPU time, total number of iterations in spatial systems, memory compression ratio.



state.

## 6.5 Experiment with $\alpha$ (Figure 5)

This parameter is supposed to penalize the standard deviation of the velocity. The (discrete) deviation is defined as follows,

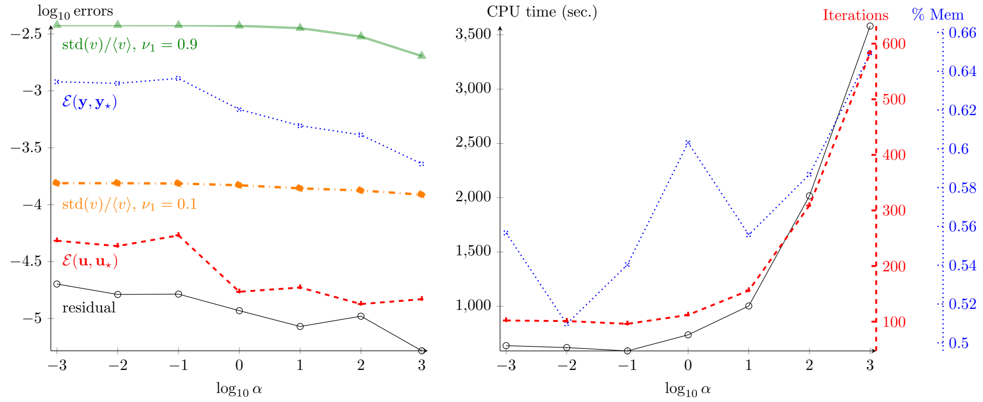
$$\operatorname{std}(v) = \sqrt{\frac{\tau}{T} \sum_{k,k',i=1}^{J_v,J_v,n_t} \sum_{j=2}^{P} M(k,k')G_0(j,j)D(i,i)y^2(i,j,k')}.$$

In Fig. 5, we report the relative deviations for two variance parameters, $\nu_1 = 0.1$ and $\nu_1 = 0.9$. We see that in both cases the deviation decreases only marginally with $\alpha$ varying from $10^{-3}$ to $10^2$. In particular, for $\nu_1 = 0.1$, it seems that the minimization of $\|\mathbf{v} - \bar{\mathbf{v}}\|$ with a deterministic $\bar{\mathbf{v}}$ delivers $\mathbf{v}$ with already a quasi-minimal variance as well. For larger $\nu_1$, the deviation decreases more significantly. We could expect this effect to develop further for $\alpha > 10^3$. However, the preconditioner deteriorates rapidly with larger $\alpha$. In particular, for $\alpha = 10^4$, the GMRES did not converge below the threshold $\varepsilon = 10^{-6}$ after 900 iterations. Further investigation is needed to develop reliable methods for damping the solution variance.

## 6.6 Experiment with $\nu_1$ (Figure 6)

The ratio of maximal and minimal viscosities due to the stochasticity is $\nu_{\max}/\nu_{\min} = (1+\nu_1)/(1-\nu_1)$. If $\nu_1 \ll 1$, it grows almost linearly, $\nu_{\max}/\nu_{\min} \approx 1 + 2\nu_1$. If $\nu_1$ is close to 1, the behavior becomes essentially nonlinear, e.g. for $\nu_1 = 0.9$ we have $\nu_{\max}/\nu_{\min} = 19$. The same can be seen in both error and complexity figures. The

Figure 5: 2D Stokes, experiment with $\alpha$. Left: Residual and errors w.r.t. the reference solutions, and the relative standard deviation. Right: CPU time, total number of iterations in spatial systems, memory compression ratio.



residuals and errors are almost stable for small $\nu_1$, and the standard deviation grows linearly, while for $\nu_1 > 0.5$, all quantities grow faster. In particular, the distance to the desired state becomes larger since the Stokes system becomes more stiff. All three complexity indicators grow rapidly as $\nu_1 \to 1$ as well.

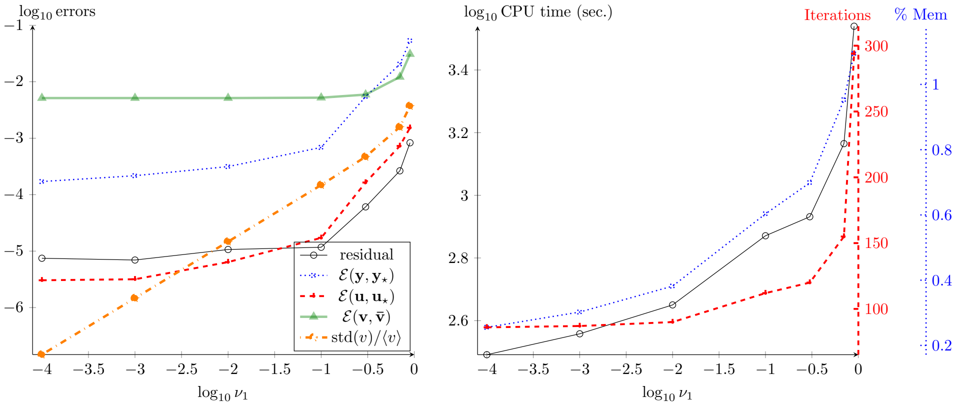
## 6.7 Experiment with the tensor approximation tolerance (Figure 7)

In experiments with positive definite matrices, it was observed that residuals and errors decay proportionally to $\varepsilon$. In this problem, this is only the case for $\varepsilon$ between $10^{-4}$ and $10^{-5}$. For smaller tolerances the residual and the control error are approximately proportional to $\varepsilon^{0.5}$, and the state error almost stagnates. This may be caused by the indefiniteness of the problem and the pressure exclusion trick. Unfortunately, we are unable to study their effects separately in the meantime, as the reduced systems (49), (50) and (51) become degenerate if we try to apply the AMEn to an indefinite system directly.

## 6.8 Experiment with $n$ (Figure 8)

The mesh generator in FEniCS is initialized with the number of mesh steps in one dimension $n$. The number of degrees of freedom for the pressure is $(n+1)^2$, since the pressure is discretized with linear elements, but together with the cubic mini elements for two components of the velocity, the total number of DoFs $J \approx 7n^2$. As in the time grid test, in addition to the residual and errors w.r.t. the reference solution, we investigate the error decay w.r.t. the grid refinement. The reference velocity for this test, $\langle v_8 \rangle$, is the mean value computed at the grid $n = 2^8$. The approximation error decays with the rate $n^{-1.4}$.

Figure 6: 2D Stokes, experiment with $\nu_1$. Left: Residual and errors w.r.t. the reference solutions, the relative standard deviation and the distance to the desired state. Right: CPU time, total number of iterations in spatial systems, memory compression ratio.



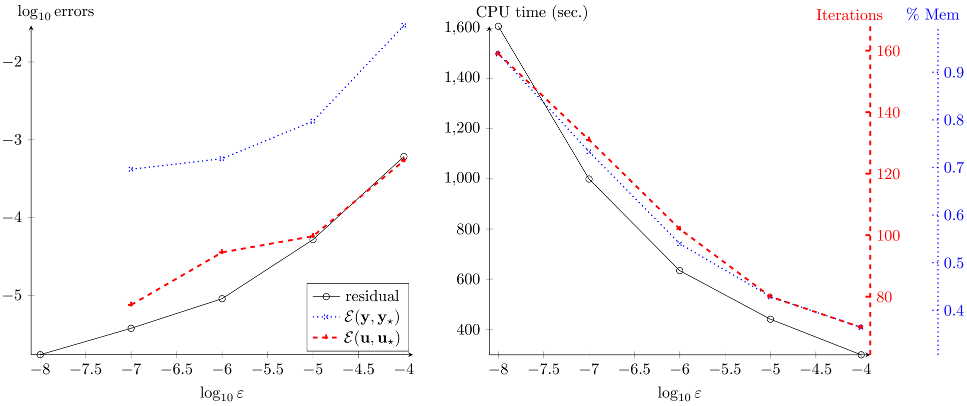
The most time-consuming stage in the scheme is the solution of the system for the spatial TT block. The sparsity of the spatial matrix allows its efficient factorization, such that the CPU time grows proportionally to $n^2$, i.e. linear w.r.t. the total number of spatial degrees of freedom. Interestingly, the number of iterations, TT ranks and the residual are smaller for larger $n$. This is due to the rank-1 approximation used for the factors of the preconditioner (31). For larger $n$, the norm of the discrete Laplace operator becomes larger, and the rank-1 term becomes a better approximation to the whole matrix.

## 6.9 Experiment with $\bar{K}_0$ (Figure 9)

Finally, we take $\bar{K}_0$ nonzero and investigate the Stokes-Brinkman model. For some reasons, with $n = 64$ and $\bar{K}_0 > 10^5$, the velocity matrix becomes indefinite. This might be due to the Gibbs phenomenon of the quadrature rule employed in FEniCS in computation of the stiffness matrix elements corresponding to the interface of $K_0(\mathbf{x})$. A detailed study would require interfering with the FEniCS source codes and this was not conducted. As a remedy, we perform this test with $n = 128$. This produces correct matrices up to $\bar{K}_0 = 10^6$.

We see that the scheme is quite robust in the considered range of the coefficient. The error estimates decay with increasing $\bar{K}_0$, since the system becomes closer to the Darcy model. The CPU time and the number of iterations show the chaotic behavior, but this fluctuation is only 10–20% compared to the average values.

Figure 7: 2D Stokes, experiment with $\varepsilon$. Left: Residual and errors w.r.t. the reference solutions. Right: CPU time, total number of iterations in spatial systems, memory compression ratio.



## 6.10 3D problem (Figure 10)

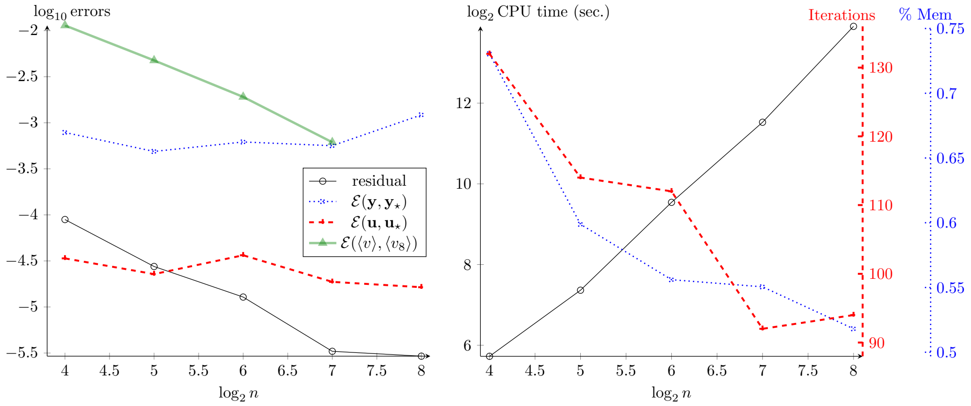
Finally, we demonstrate that our approach is suitable for larger 3D problems. We consider the three-dimensional Stokes-Brinkman problem on the domain $[0,1]\times[0,1]\times[0,5]$ as constraints, with the coefficient

$$K_0(\mathbf{x}) = \begin{cases} 10^4, & (x_1-0.5)^2+(x_2-0.5)^2+(x_3-2.5)^2 \le 0.1^2, \\ 0, & \text{otherwise,} \end{cases}$$

and the inflow boundary condition $v_1|_{x_1=0} = x_2(1-x_2)$ and zero conditions at other boundaries. The one-dimensional grid sizes are $16, 16, 32$ for $x_1, x_2, x_3$, respectively, which results in $J_v = 212355$ degrees of freedom for the velocity. Other parameters are the same as in the 2D tests except $\nu_1 = 0.01$ and $\varepsilon = 10^{-4}$.

Since the direct elimination is too expensive for such matrices, we used the commutator-based preconditioner (38) for the Schur complement in the Stokes matrices, and the velocity matrix was approximated by one V-cycle of the HSL MI20 algebraic multigrid [9]. The iterative method is two-level. First, we employed the block-triangular preconditioner for the KKT structure in the FGMRES method with unlimited number of iterations. Second, for all Stokes-like matrices in the preconditioning step, e.g. in (59), we used another FGMRES method with 50 iterations, preconditioned by (38) with the multigrid. That many inner iterations are needed because the commutator preconditioner deteriorates rapidly with the size of the porosity region. The KKT solver conducted in total 152 iterations, which took 148985 seconds of the CPU time. Nevertheless, the maximal TT rank of the solution is 8, so the TT format consumed only 0.2% of the memory required for the full solution. The final residual is $4.1\cdot10^{-4}$, and the misfit with the desired state $\mathcal{E}(\mathbf{v},\bar{\mathbf{v}}) = 2.8\cdot10^{-3}$. The mean and the standard deviation of the solution at the final time are shown in Fig. 10. We notice a clear perturbation around the region with nonzero Brinkman coefficient. In particular, the

Figure 8: 2D Stokes, experiment with $n$. Left: Residual and errors w.r.t. the reference solutions, and the mean value error w.r.t. the spatial grid level. Right: CPU time, total number of iterations in spatial systems, memory compression ratio.



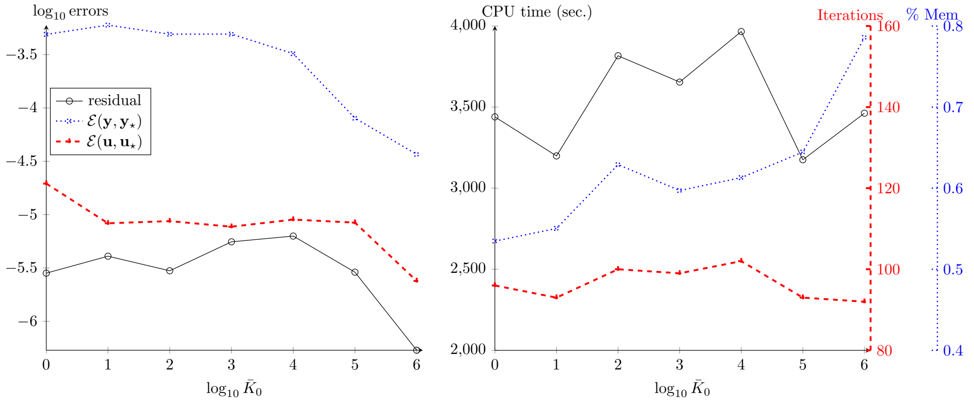
largest deviations are attained at the interface, while in the homogeneous region the velocity is almost deterministic. The deviation of the pressure grows proportionally to the mean magnitude (note that the mean pressure is mostly negative, while the deviation is not, hence the color map in the right middle figure was reversed). The control exhibits a clear interface around the Brinkman region. Another interesting feature is that the deviation of the control is larger than its mean.

## 7 Conclusions and outlook

We have considered a low-rank solution to an optimal control problem constrained by Stokes-Brinkman with uncertain inputs. The discretized solution can be naturally indexed by three independent parameters, coming from the spatial, stochastic and time variables. Each of these parameters can vary in a considerable range, hence the straightforward storage of the solution consumes a vast amount of memory. By employing tensor product decomposition methods, we have reduced it by two–three orders of magnitude. However, the optimal control problem yields a saddle-point linear system, which requires a special treatment. We have extended the alternating minimal energy algorithm such that it preserves the saddle-point structure and solves this system robustly. Moreover, we have proposed a new Schur complement-based preconditioner which is free from auxiliary perturbations and provides smaller condition numbers of the preconditioned matrix.

Several directions of future research are possible. A natural extension is to apply our techniques to the nonlinear Navier-Stokes model. The preconditioner still needs an improvement, especially for large stochastic variance parameter $\nu_1$, variance-penalizing regularization parameter $\alpha$ and many time steps. More complex models, such as those

Figure 9: 2D Stokes-Brinkmann, experiment with $\bar{K}_0$. Left: Residual and errors w.r.t. the reference solutions. Right: CPU time, total number of iterations in spatial systems, memory compression ratio.



with uncertain boundary conditions and random domain, are also a challenging topic for future investigation.

## References

[1] R. Andreev and C. Tobler, *Multilevel preconditioning and low rank tensor iteration for space-time simultaneous discretizations of parabolic PDEs*, Numerical Linear Algebra with Applications, 22 (2015), pp. 317–337.

[2] H. Antil, M. Heinkenschloss, and R.H.W. Hoppe, *Domain decomposition and balanced truncation model reduction for shape optimization of the Stokes system*, Optimization Methods and Software, 26 (2011), pp. 643–669.

[3] M. J. Balajewicz, E. H. Dowell, and B. R. Noack, *Low-dimensional modelling of high-Reynolds-number shear flows incorporating constraints from the Navier-Stokes equation*, Journal of Fluid Mechanics, 729 (2013), pp. 285–308.

[4] J. Ballani and L. Grasedyck, *A projection method to solve linear systems in tensor format*, Numerical Linear Algebra with Applications, 20 (2013), pp. 27–43.

[5] P. Benner, A. Onwunta, and M. Stoll, *Block-diagonal preconditioning for optimal control problems constrained by PDEs with uncertain inputs*, MPI Magdeburg Preprint 2015-05, (2015).

[6] ———, *Low-rank solution of unsteady diffusion equations with stochastic coefficients*, SIAM/ASA Journal on Uncertainty Quantification, To appear, (2015).

[7] M. Benzi, G. H. Golub, and J. Liesen, *Numerical solution of saddle point problems*, Acta Numerica, 14 (2005), pp. 1 – 137.

[8] P. Bochev and R. B. Lehoucq, *On the finite element solution of the pure Neumann problem*, SIAM Review, 47 (2005), pp. 50–66.

[9] J. Boyle, M. D. Mihajlovic, and J. A. Scott, *HSL MI20: An efficient AMG preconditioner*, Tech. Report RAL-TR-2007-021, Rutherford Appleton Laboratory (CCLRC), 2007.

[10] A. Caiazzo, V. John, and U Wilbrandt, *On classical iterative subdomain methods for the Stokes-Darcy problem*, Computational Geosciences, 18 (2014), pp. 711–728.

[11] S. V. Dolgov, *TT-GMRES: solution to a linear system in the structured tensor format*, Russian Journal of Numerical Analysis and Mathematical Modelling, 28 (2013), pp. 149–172.

[12] S. V. Dolgov, B. N. Khoromskij, I. V. Oseledets, and D. V. Savostyanov, *Computation of extreme eigenvalues in higher dimensions using block tensor train format*, Computer Physics Communications, 185 (2014), pp. 1207–1216.

[13] S. V. Dolgov and D. V. Savostyanov, *Alternating minimal energy methods for linear systems in higher dimensions*, SIAM Journal on Scientific Computing, 36 (2014), pp. A2248–A2271.

[14] H. C. Elman, D. J. Silvester, and A. J. Wathen, *Finite Elements and Fast Iterative Solvers with Applications in Incompressible Fluid Dynamics*, Numerical Mathematics and Scientific Computation, Oxford University Press, New York, 2005.

[15] W. H. Engl, *Discrepancy principles for Tikhonov regularization of ill-posed problems leading to optimal convergence rates*, Journal of Optimization Theory and Applications, 52 (1987), pp. 209–215.

[16] O. G. Ernst, A. Mugler, H.-J. Starkloff, and E. Ullmann, *On the convergence of generalized polynomial chaos expansions*, ESAIM: Mathematical Modelling and Numerical Analysis, 46 (2012), pp. 317–339.

[17] R. G. Ghanem and R. M. Kruger, *Numerical solution of spectral stochastic finite element systems*, Computer Methods in Applied Mechanics and Engineering, 129 (2005), pp. 289–303.

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

[19] W. Hackbusch, *Tensor Spaces And Numerical Tensor Calculus*, Springer–Verlag, Berlin, 2012.

[20] P. C. Hansen and D. P. O'Leary, *The use of the L-curve in the regularization of discrete ill-posed problems*, SIAM Journal on Scientific Computing, 14 (1993), pp. 1487–1503.

[21] S. Holtz, T. Rohwedder, and R. Schneider, *The alternating linear scheme for tensor optimization in the tensor train format*, SIAM Journal on Scientific Computing, 34 (2012), pp. A683–A713.

[22] Roger A. Horn and Charles R. Johnson, *Matrix analysis*, Cambridge University Press, Cambridge, 1990.

[23] C. Hubig, I. P. McCulloch, U. Schollwöck, and F. A. Wolf, *Strictly single-site DMRG algorithm with subspace expansion*, Physical Review B, 91 (2015), p. 155115.

[24] E. Jeckelmann, *Dynamical density matrix renormalization group method*, Physical Review B, 66 (2002), p. 045114.

[25] B. N. Khoromskij, *Tensor numerical methods for high-dimensional PDEs: Basic theory and initial applications*, arXiv preprint 1409.7970, 2014. to appear in ESAIM: Proceedings.

[26] A. Klümper, A. Schadschneider, and J. Zittartz, *Matrix product ground states for one-dimensional spin-1 quantum antiferromagnets*, Europhysics Letters, 24 (1993), pp. 293–297.

[27] T. G. Kolda and B. W. Bader, *Tensor decompositions and applications*, SIAM Review, 51 (2009), pp. 455–500.

[28] D. Kressner, M. Steinlechner, and A. Uschmajew, *Low-rank tensor methods with subspace correction for symmetric eigenvalue problems*, SIAM Journal on Scientific Computing, 36 (2014), pp. A2346–A2368.

[29] D. Kressner and C. Tobler, *Krylov subspace methods for linear systems with tensor product structure*, SIAM Journal on Matrix Analysis and Applications, 31 (2010), pp. 1688–1714.

[30] D. Kressner and C. Tobler, *Low-rank tensor Krylov subspace methods for parametrized linear systems*, SIAM Journal on Matrix Analysis and Applications, 32 (2011), pp. 273–290.

[31] K. Kunisch and S. Volkwein, *Galerkin POD methods for parabolic problems*, Numerische Mathematik, 90 (2001), pp. 117–148.

[32] J. Larminie and A. Dicks, *Fuel cell systems explained*, vol. 2nd edition, Wiley, 2013.

[33] D. Leykekhman, *Investigation of commutative properties of Discontinuous Galerkin methods in PDE constrained optimal control problems*, Journal of Scientific Computing, 53 (2012), pp. 483 – 511.

[34] A. Logg, K.-A. Mardal, and G. N. Wells (Eds.), *Automated Solution of Differential Equations by the Finite Element Method*, Springer, 2012.

[35] G. J. Lord, C. E. Powell, and T. Shardlow, *An introduction to computational stochastic PDEs*, Cambridge University Press, 2014.

[36] K. A. Mardal, X. C. Tai, and R. Winther., *A mixed formulation for the Brinkman problem*, SIAM Journal on Numerical Analysis, 40 (2002), pp. 1605 – 1631.

[37] B. R. Noack, P. Papas, and P. A. Monkewitz, *The need for a pressure-term representation in empirical Galerkin models of incompressible shear flows*, Journal of Fluid Mechanics, 523 (2005), pp. 339–365.

[38] I. V. Oseledets, *Tensor-train decomposition*, SIAM Journal on Scientific Computing, 33 (2011), pp. 2295–2317.

[39] 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.

[40] I. V. Oseledets and S. V. Dolgov, *Solution of linear systems and matrix inversion in the TT-format*, SIAM Journal on Scientific Computing, 34 (2012), pp. A2718–A2739.

[41] J. W. Pearson, M. Stoll, and A. J. Wathen, *Regularization-robust preconditioners for time-dependent PDE-constrained optimization problems*, SIAM Journal on Matrix Analysis and Applications, 33 (2012), pp. 1126–1152.

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

[43] P. Popov, Y. Efendiev, and G. Qin., *Multiscale modeling and simulations of flows in naturally fractured Karst reservoirs*, Communications in Compuational Physics, 6 (2009), pp. 162 – 184.

[44] C. E. Powell and H. C. Elman, *Block-diagonal preconditioning for spectral stochastic finite-element systems*, IMA Journal of Numerical Analysis, 29 (2009), pp. 350–375.

[45] C. E. Powell and D. J. Silvester, *Preconditioning steady-state Navier-Stokes equations with random data*, SIAM Journal on Scientific Computing, 34 (2012), pp. A2482 – A2506.

[46] E. Rosseel and G. N. Wells, *Optimal control with stochastic PDE constraints and uncertain controls*, Computer Methods in Applied Mechanics and Engineering, 213-216 (2012), pp. 152–167.

[47] U. Schollwöck, *The density–matrix renormalization group*, Reviews of Modern Physics, 77 (2005), pp. 259–315.

[48] J. Sogn, *Stabilized finite element methods for the Brinkman equation on fitted and fictitious domains*, vol. Master's Thesis, University of Oslo, 2014.

[49] M. Stoll and T. Breiten, *A low-rank in time approach to PDE-constrained optimization*, SIAM Journal on Scientific Computing, 37 (2015), pp. B1 – B29.

[50] M. Stoll and A. Wathen, *All-at-once solution of time-dependent Stokes control*, Journal of Computational Physics, 232 (2013), pp. 498–515.

[51] P. S. Vassilevski and U. Villa, *A block-diagonal algebraic multigrid preconditioner for the Brinkman problem*, SIAM Journal on Scientific Computing, 35 (2013), pp. S3 – S17.

[52] ——, *A mixed formulation for the Brinkman problem*, SIAM Journal on Numerical Analysis, 52 (2014), pp. 258 – 281.

[53] E. L. Wachspress, *The ADI Model Problem*, Springer, New York, 2013.

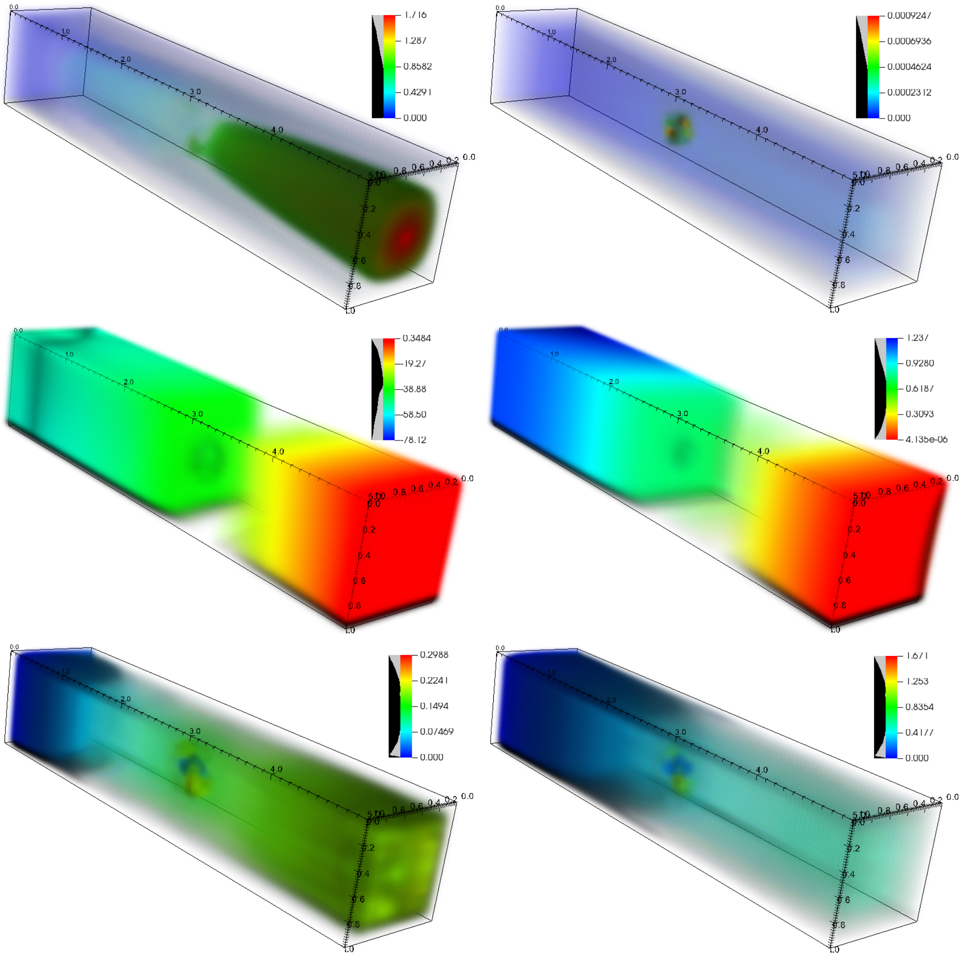
[54] S. R. White, *Density matrix algorithms for quantum renormalization groups*, Physical Review B, 48 (1993), pp. 10345–10356.

[55] ——, *Density matrix renormalization group algorithms with a single center site*, Physical Review B, 72 (2005), p. 180403.

[56] X. P. Xie, J. C. Xu, and G. R. Xue, *Uniformly stable finite element methods for Darcy-Stokes-Brinkman models*, Journal of Computational Mathematics, 26 (2008), pp. 437 – 455.

[57] D. Xiu and J. Shen, *Efficient stochastic Galerkin methods for random diffusion*, Journal of Computational Physics, 228 (2009), pp. 266–281.

Figure 10: 3D Stokes-Brinkman. Left: mean values at the last time step, right: standard deviations. Top: velocity, middle: pressure, bottom: control.



**Max Planck Institute Magdeburg Preprints**