Journal of Computational and Applied Mathematics 450 (2024) 115959

Contents lists available at ScienceDirect

Journal of Computational and Applied Mathematics

journal homepage: www.elsevier.com/locate/cam

Efficient low rank approximations for parabolic control problems with unknown heat source

Doghonay Arjmand [a,*], Maksat Ashyraliyev [b]

[a] Department of Information Technology, Division of Scientific Computing, Uppsala University, Sweden
[b] Division of Mathematics and Physics, Mälardalen University, 72123, Västerås, Box 883, Sweden

ARTICLE INFO

MSC:
0000
1111

Keywords:
Control problems
Low rank approximations
Parabolic PDEs
Arnoldi algorithm
Heat equation
Inverse problems

ABSTRACT

An inverse problem of finding an unknown heat source for a class of linear parabolic equations is considered. Such problems can typically be converted to a direct problem with non-local conditions in time instead of an initial value problem. Standard ways of solving these non-local problems include direct temporal and spatial discretization as well as the shooting method, which may be computationally expensive in higher dimensions. In the present article, we present approaches based on low-rank approximation via Arnoldi algorithm to bypass the computational limitations of the mentioned classical methods. Regardless of the dimension of the problem, we prove that the Arnoldi approach can be effectively used to turn the inverse problem into a simple initial value problem at the cost of only computing one-dimensional matrix functions while still retaining the same accuracy as the classical approaches. Numerical results in dimensions $d = 1, 2, 3$ are provided to validate the theoretical findings and to demonstrate the efficiency of the method for growing dimensions.

1. Introduction

The theory of inverse problems for differential equations has been extensively developed to tackle various problems in applied sciences and engineering. Typical applications turning this field into a popular area of research include geological exploration, medical diagnostics, and predictive material science. In a direct problem, typically a differential equation is given and its solution is determined uniquely from imposed initial and/or boundary conditions. In an inverse problem, on the contrary, a differential equation itself contains unknown data/parameters. For example, a differential equation may involve unknown coefficients or source terms, which need to be determined from a set of observations or measurements of the solution. In the study of inverse problems, the lack of knowledge about the data in the model problem is then compensated by introducing some additional conditions into the problem, which may eventually result in a well-posed mathematical model. For the general theory of inverse problems for differential equations we refer the interested reader to [1–3] and the references therein.

Motivated by real applications, inverse problems for diffusion equations can be categorized into two: (a) determining the coefficient of the equation, see e.g., [4–6] (b) determining the heat source. Latter is closely related to the goal of the present article. In particular, we are interested in the inverse problem of finding the pair of solutions $\{p, u\}$ such that

$$\begin{cases} \partial_t u(t,x) - \nabla \cdot (a(x)\nabla u(t,x)) = f(t,x) + p(x), \ \ (t,x) \in (0,T) \times K, \\ u(t,x) = 0, \quad (t,x) \in [0,T] \times \partial K, \\ u(0,x) = 0, \quad u(T,x) = \varphi(x), \quad x \in K, \end{cases} \tag{1}$$

* Corresponding author.
E-mail address: doghonay.arjmand@it.uu.se (D. Arjmand).

https://doi.org/10.1016/j.cam.2024.115959
Received 18 August 2023; Received in revised form 26 February 2024
Available online 1 May 2024

where $K := (0,1)^d$ is an open unit cube with a boundary ∂K and $a(x) \in \mathbb{R}^{d\times d}, \forall x \in K$ is a symmetric matrix function satisfying

$$0 < a_{\min}|\zeta|^2 \le \zeta^T a(x)\zeta \le a_{\max}|\zeta|^2, \quad \forall x \in K \text{ and } \zeta \in \mathbb{R}^d, \tag{2}$$

and f and φ are sufficiently smooth known functions. Problem (1) is highly relevant from an application point of view. Consider, for example, a domain K, for which there already exists a known (and practically difficult to remove) background heat source f. Suppose that our objective is to keep the temperature distribution at a specific time T in a desired manner. The problem is then to find the control data p accounting for yet another heat source, to be fine-tuned over K, such that the final temperature requirements are met. The inverse problem of simultaneous determination of temperature and a spacewise dependent source from given final data is related, for instance, to heat conduction problems with an unknown time-independent heat source. From an application point of view, it may help in controlling the heat distribution of solid structures (e.g., airplane wings) and other industrial devices in a non-local and non-dynamic fashion. In most of current physical applications, however, the unknown parameter may depend on time, and in this case, for the well-posedness of the problem the last condition in (1) must be changed, for instance replaced by $u(t,x^*) = \varphi(t),\ t \in [0,T]$, where $x^* \in K$. Although the principal ideas from this paper may be used also for this case, in general, the resulting inverse problems require a different mathematical machinery from the one we discuss in this paper and the numerical treatment of this case is left as a future work.

The problem of determining the heat source in a parabolic equation has been a topic of research for the past two decades and is still a popular area of research due to the need to improve the efficiency of numerical methods linked to the proposed strategies. A class of algorithms rely on obtaining a transformed equation, where the unknown heat source is no longer present, and the resulting system is then solved with a direct numerical simulation in time and space, see e.g., [7–10]. This idea typically leads to either a large linear system of equations to solve due to the urge to treat the additional temporal dimension similar to spatial dimensions rather than using a standard time-stepping method or alternatively exploiting iterative solution methods, which may be computationally expensive. More standard ways of tackling this problem is based on re-formulation of the original problem as a minimization problem, [11–15], which eventually needs an efficient iterative solution method to find the minimizer. There are other strategies which are limited as they assume a specific structure for the parabolic operator (e.g., constant coefficients), [16], or they have limitations due to dimension [17].

We may rewrite the above problem in abstract form as

$$\begin{cases} \dfrac{d}{dt}u(t) + Au(t) = f(t) + p, \quad t \in (0,T), \\ u(0) = 0, \quad u(T) = \varphi, \end{cases} \tag{3}$$

where $u(t)$, $t \in [0,T]$ and $f(t)$, $t \in (0,T)$ are vector valued functions in a suitable Hilbert space H. Similarly, $p, \varphi \in H$ and $A : H \to H$ is the corresponding abstract operator, where the boundary conditions are incorporated in A. In general, problem (3) is ill-posed. However, for sufficiently smooth data its well-posedness has been shown in the literature. The unique solvability of inverse source identification problem (3) in an arbitrary Banach space E with linear unbounded operator A was established in [18]. The stability estimates for the solution of problem (3) were obtained in [19]. Related to the theoretical foundations of the present work is also the well-posedness of an inverse parabolic problem with an unknown source term [20,21]; see also [22,23] for regularity conditions for the controllability of final time over-determination parabolic equations with a time and space dependent right hand side, and with superlinear nonlinearities.

The starting point that we follow is similar to [10] which relies on eliminating the unknown p in (3) by introducing a new abstract function v as

$$v(t) = u(t) - A^{-1}p, \quad t \in [0,T].$$

Indeed, it follows that $\frac{d}{dt}v(t) = \frac{d}{dt}u(t)$, $Av(t) = Au(t) - p$, $v(T) = \varphi - A^{-1}p$, and $v(0) = -A^{-1}p$. Therefore, the problem (3) is equivalent to the following problem:

$$\begin{cases} \dfrac{d}{dt}v(t) + Av(t) = f(t), \quad t \in (0,T), \\ v(0) = v(T) - \varphi. \end{cases} \tag{4}$$

Note that having the solution $v(t)$, $t \in [0,T]$ of problem (4) allows us to obtain the solutions of initial problem (3) directly as follows:

$$p = -Av(0), \quad u(t) = v(t) - v(0),\ t \in (0,T).$$

The problem (4) is direct since it does not involve an unknown source term p. However, it has a non-local nature due to the dependency of $v(0)$ on $v(T)$.

Remark 1. Apart from the well-posedness of the problem (3), the stability and uniqueness of the non-local problem (4) can be established using [24,25], where the authors consider a more general nonlocal condition that also applies to the setting of (4).

The main aim of this article is to develop numerical methods, based on low rank approximations, which bypasses the limitations of the existing methods with respect to dimension and structural properties of the operator, see also [26–28] for fully discrete

approaches available in the literature. To put everything into a framework, assume a spatial discretization of the problem (4) in the form

$$\begin{cases} \dfrac{d}{dt}v_h(t) + A_h v_h(t) = f_h(t), \quad 0 < t < T, \\ v_h(0) = v_h(T) - \varphi_h, \end{cases} \tag{5}$$

where $h = 1/N$ and v_h is a vector of approximations at spatial grid points

$$K_h = (hn_1, hn_2, \dots, hn_d), \quad 0 \le n_j \le N, \quad j = 1, 2, \dots, d,$$

and A_h is the second-order centered difference matrix operator[1] approximating the elliptic operator $-\nabla \cdot (a\nabla)$, incorporating also the homogeneous Dirichlet boundary conditions. By an additional time-stepping method, we can turn the continuous time solution $v_h(t)$ into a fully discrete one. Nevertheless, since problem (5) has a non-local condition in time, standard time stepping methods can not be exploited directly. Two natural ways of resolving the problem of non-locality in time is by (1) a direct discretization in time and space, which results in a linear system of extremely large size which is computationally expensive to solve, especially for two or three dimensional problems; [10], (2) developing a convergent shooting algorithm, see e.g., [29,30] for recent variants of it, to turn the non-local problem (5) into a purely initial-value problem, which is computationally and implementation-wise preferable in comparison to the former. In this article, we develop low rank approximations based on the Arnoldi algorithm to efficiently transform the problem (5) into an initial-value problem, and accurately approximate the unknown pair $\{p, u\}$, in a much more efficient way than the fully discrete approach and the shooting method.

This article is structured as follows. In Section 3 we discuss two standard methods: a full discretization of the problem (5), and a more efficient shooting algorithm, which is also proved to be convergent. Section 4 contains the main contribution of this paper, where a low rank approximation algorithm based on Arnoldi is introduced, and the foundations of the ideas are solidified by a convergence analysis for the low rank approximations of the emerging operators in the algorithm. We conclude the paper by providing numerical experiments in Section 5, to validate our theoretical findings.

2. Preliminaries

To improve readability and to synchronize the mathematical notation throughout the paper, (1) we present the common notations that will be subsequently used in the sequel, (2) we present a few intermediate theoretical results in advance (to be used in the analysis later).

- Let $B \in \mathbb{R}^{N^d \times N^d}$ be a matrix. We denote the set of eigenvalues of B by $\lambda(B)$. The spectral radius of the matrix B is denoted by ρ_B.
- Throughout the paper, the letter N represents the number of degrees of freedom in a spatial direction, while M represents the degrees of freedom coming from a temporal discretization.
- In order to avoid the confusion with the final time T in Eq. (1), we denote the transpose of a matrix B by B^*, even when the matrix B is real-valued.
- Superscripts are used to indicate the power of a matrix B^k, or the derivative of a function $f^{(k)}(t) := \frac{d^k f(t)}{dt^k}$, or as a means to represent the iteration number for a time-stepping method; see e.g. v_h^m in Eq. (6). The distinction between power of a quantity and the temporal iteration number is expected to be clear from the context.
- Properties of the matrix A_h:

  P1. The matrix $A_h \in \mathbb{R}^{N^d \times N^d}$ is symmetric and positive definite. The eigenvalues $\lambda_j \in \lambda(A_h)$ are real and positive and ordered as

  $$0 < \lambda_1 \le \lambda_2 \le \cdots \le \lambda_{N^d},$$

  and the corresponding eigenfunctions ϕ_j form an orthonormal basis for the finite dimensional space $l^2(K_h)$ equipped with the inner-product

  $$\langle f_h, g_h \rangle = h^d \sum_{j=1}^{N^d} f_{h,j} g_{h,j},$$

  and the norm $\|f_h\| := \sqrt{\langle f_h, f_h \rangle}$.

  P2. For sufficiently small h, we have $\lambda_1 \ge a_{\min} \frac{\pi^2}{d}$, up to an $O(h^2)$ discretization error.

  P3. The operator A_h generates the contraction semigroup $e^{-TA_h} : l^2(K_h) \to l^2(K_h)$ which satisfies

  $$\left\| e^{-TA_h} \right\| \le e^{-\lambda_1 T}.$$

[1] It is assumed that A_h is symmetric and positive definite.

Proof of P2. To see property 2, let us consider the continuous operator $A = -\nabla \cdot (a\nabla)$, which can be seen as the limit (as $h \to 0$) of the discrete operator A_h, and let $\{\lambda_1, \phi_1\}$ be the principal eigenvalue and eigen-function of A. Then it follows that

$$\int_K a(x)\nabla\phi_1(x) \cdot \nabla v(x)\, dx = \lambda_1 \int_K \phi_1(x)v(x)\, dx, \quad \forall v \in H_0^1(K).$$

Therefore, plugging $v = \phi_1$, we obtain

$$\lambda_1 = \frac{\int_K a(x)\nabla\phi_1(x) \cdot \nabla\phi_1(x)dx}{\|\phi_1\|^2_{L^2(K)}} \geq a_{\min} \frac{\|\nabla\phi_1\|^2_{L^2(K)}}{\|\phi_1\|^2_{L^2(K)}} \geq a_{\min} C_p(K)^{-2},$$

where the last inequality follows from the Poincare's inequality, and $C_p(K)$ is the Poincare's constant for the domain K, which is bounded by $C_p(K) \leq \frac{diam(K)}{\pi} = \frac{\sqrt{d}}{\pi}$, see [31]. Using this inequality, and the fact that $A_h \to A$, we obtain property 2 for sufficiently small h.

Proof of P3. To see the last property, assume $g = \sum_{j=1}^{N^d} g_j\phi_j$, where $g_j = \langle g, \phi_j\rangle$, and

$$\begin{aligned}\left\|e^{-TA_h}\right\|^2 &:= \sup_{\|g\|=1} \left\|e^{-TA_h}g\right\|^2 = \sup_{\|g\|=1} \|\sum_{j=1}^{N^d} e^{-T\lambda_j} g_j\phi_j\|^2 \\ &= \sup_{\|g\|=1} h^d \sum_{j,\ell=1}^{N^d} e^{-T\lambda_j} e^{-T\lambda_\ell} g_j g_\ell \langle\phi_j, \phi_\ell\rangle.\end{aligned}$$

Since the eigenfunctions are orthonormal, it follows that $\langle\phi_j, \phi_\ell\rangle = \delta_{i,j}$, where $\delta_{i,j}$ is the kronecker delta. Finally, using the discrete Parseval's equality, we obtain

$$\begin{aligned}\|e^{-TA_h}\|^2 &= \sup_{\|g\|=1} h^d \sum_{j=1}^{N^d} e^{-2T\lambda_j} g_j^2 \leq \sup_{\|g\|=1} e^{-2T\lambda_1} h^d \sum_{j=1}^{N^d} g_j^2 \\ &= \sup_{\|g\|=1} e^{-2T\lambda_1} \|g\|^2 = e^{-2T\lambda_1}.\end{aligned}$$

The result follows by taking the square roots of both sides.

3. Standard numerical methods for solving (5)

In this section, we present two strategies to tackle the problem of non-locality in time. First we present a direct approach, where the problem (5) is directly discretized in time, and we discuss potential numerical challenges with this approach. Second, we will present the shooting method, which is a more standard way of iteratively solving (5). We conclude this section by a proof of convergence for the shooting method.

3.1. A direct approach

Let $\tau = T/M$ and $t_m = m\tau$, $0 \leq m \leq M$. A temporal discretization of the problem (5), say using the Crank–Nicolson scheme, yields

$$\begin{cases} \dfrac{v_h^{m+1} - v_h^m}{\tau} + \dfrac{A_h v_h^{m+1} + A_h v_h^m}{2} = f_h(t_{m+1/2}), \ m = 0, 1, \dots, M-1, \\ v_h^0 = v_h^M - \varphi_h, \end{cases} \tag{6}$$

where v_h^k denotes the numerical approximations of $v_h(t_k)$. Since v_h^0 is not available, (6) cannot be solved recursively. In fact, it is a system of $M(N-1)^d$ linear equations with $M(N-1)^d$ unknowns. Therefore, one needs to invert the square matrix of size $M(N-1)^d$ which requires a huge amount of computation and memory space; in particular, for multi-dimensional problems with sufficiently small step sizes τ and h.

One can solve the nonlocal scheme (6) by using the locality of boundary conditions in one of the spatial variables. In fact, (6) can be written in the matrix form as follows:

$$\begin{cases} CV_{n+1} + BV_n + CV_{n-1} = \phi_n, \quad 1 \leq n \leq N-1, \\ V_0 = V_N = 0, \end{cases} \tag{7}$$

where B and C are constant square matrices of size $M(N-1)^{d-1}$, V_n and ϕ_n are column vectors with $M(N-1)^{d-1}$ entries. The solution of the matrix Eq. (7) can be found by using the modified Gauss elimination method [32] as follows:

$$\begin{cases} V_n = \alpha_{n+1}V_{n+1} + \beta_{n+1}, \quad n = N-1, \dots, 2, 1, \\ V_{N-1} = 0, \end{cases}$$

where α_n are square matrices and β_n are column vectors, calculated as

$$\alpha_{n+1} = -\left(B + C\alpha_n\right)^{-1} C,$$

$$\beta_{n+1} = \left(B + C\alpha_n\right)^{-1}\left(\phi_n - C\beta_n\right), \quad n = 1, 2, \ldots, N-1.$$

Here, α_1 is a zero matrix and β_1 is a zero vector.

The elimination method described above requires the inversion of a square matrix of size $M(N-1)^{d-1}$ at every iteration, which significantly limits its application to one and two dimensional problems only. Due to this computational limitation, we argue that the use of fully discrete scheme should be avoided in practice, and that more efficient strategies for approximating (5) are needed. Finally, we note that other boundary conditions such as the Neumann conditions can also be treated with a slight adjustment but with no major conceptual change in the algorithm.

3.2. Shooting method

An alternative way to solve the non-local problem (5) is by means of the shooting method, which is easy to implement and is favorable from a computational cost point of view. Let us denote by $v_h(t;\alpha)$ the solution of the initial value problem

$$\begin{cases} \dfrac{d}{dt}v_h(t;\alpha) + A_h v_h(t;\alpha) = f_h(t), \quad 0 < t < T, \\ v_h(0;\alpha) = \alpha. \end{cases} \tag{8}$$

For $v_h(t;\alpha)$ to be a solution of (5), the initial vector α must satisfy

$$\alpha = v_h(T;\alpha) - \varphi_h. \tag{9}$$

The corresponding fixed-point iterations can be constructed as

$$\alpha_{k+1} = v_h^M(\alpha_k) - \varphi_h, \quad k = 0, 1, 2, \ldots \tag{10}$$

where $v_h^M(\alpha_k)$ is the solution to the Crank–Nicolson approximation (6) of the problem (8) with initial data α_k, which reads as

$$\begin{aligned} v_h^{m+1}(\alpha_k) &= G(A_h;\tau)v_h^m(\alpha_k) + \tau\left(I + \frac{\tau A_h}{2}\right)^{-1} f_h(t_{m+1/2}), \\ v_h^0(\alpha_k) &= \alpha_k, \quad m = 0, 1, 2, \ldots, M-1, \quad \tau M = T, \end{aligned} \tag{11}$$

where $G(A_h;\tau) := \left(I + \frac{\tau A_h}{2}\right)^{-1}\left(I - \frac{\tau A_h}{2}\right)$.

Remark 2. Throughout the manuscript, we assume a Crank–Nicolson discretization in time. The efficiency of the time stepping can be further improved by employing either explicit stabilized integrators with large stability regions, such as [33] or symplectic Runge Kutta methods such as [34]. Nevertheless, note that employing different time-stepping methods will require completely different mathematical analysis, and the theories in this paper is valid only for the Crank–Nicolson scheme.

Theorem 1. *Let $\{\alpha_k\}$ be the sequence generated by fixed-point iterations (10), and let $v_h(0)$ satisfy $v_h(0) = v_h(T; v_h(0)) - \varphi_h$, where $v_h(T; v_h(0))$ is the solution at time T of problem (8) with initial data $v_h(0)$. Then it follows that*

$$\|\alpha_{k+1} - v_h(0)\| \le \|G\|^M \|\alpha_k - v_h(0)\| + C\tau^2\left(1 + \|G\|^M\right)\left(1 - \|G\|^M\right)^{-1},$$

where $\|G\| = \left|\frac{1-\tau\lambda_1/2}{1+\tau\lambda_1/2}\right| < 1$, λ_1 is the principal eigenvalue of the matrix A_h, and C is a constant independent of k but may depend on f_h and $v_h(0)$.

Proof. We re-write (11) in the form

$$\begin{aligned} v_h^{m+1}(\alpha_k) &= G^{m+1}\alpha_k + \tau\sum_{j=0}^{m} G^{m-j}F_h(t_{j+1/2}), \\ F_h(t_{j+1/2}) &:= \left(I + \frac{\tau A_h}{2}\right)^{-1} f_h(t_{j+1/2}), \quad m = 0, 1, 2, \ldots, M-1. \end{aligned}$$

Then the fixed-point iteration (10) can be written as

$$\alpha_{k+1} = G^M\alpha_k + \tau\sum_{j=0}^{M-1} G^{M-1-j}F_h(t_{j+1/2}) - \varphi_h, \quad k = 0, 1, 2, \ldots$$

Moreover, let α_∞ be the limiting value of α_k satisfying[2]

$$\alpha_\infty = G^M\alpha_\infty + \tau\sum_{j=0}^{M-1} G^{M-1-j}F_h(t_{j+1/2}) - \varphi_h.$$

[2] Indeed the fixed-point iteration is convergent since $\|G\| < 1$, and therefore the limiting value α_∞ exists.

Clearly

$$\|\alpha_{k+1}-\alpha_\infty\| \le \|G\|^M \|\alpha_k-\alpha_\infty\|. \tag{12}$$

On the other hand,

$$\begin{aligned}
\|\alpha_\infty - v_h(0)\|_2 &= \left\| v_h^M(\alpha_\infty) - \varphi_h - v_h(0)\right\| \\
&= \left\| v_h^M(\alpha_\infty) - \varphi_h - \left(v_h(T;v_h(0)) - \varphi_h\right)\right\|_2 \\
&= \left\| v_h^M(\alpha_\infty) - v_h^M(v_h(0)) + v_h^M(v_h(0)) - v_h(T;v_h(0))\right\|_2 \\
&\le \left\| v_h^M(\alpha_\infty) - v_h^M(v_h(0))\right\|_2 + \left\| v_h(T;v_h(0)) - v_h^M(v_h(0))\right\| \\
&\le \|G\|^M \|\alpha_\infty - v_h(0)\| + C\tau^2.
\end{aligned}$$

Therefore,

$$\|\alpha_\infty - v_h(0)\| \le C\tau^2\left(1-\|G\|^M\right)^{-1}.$$

Now using (12) together with the last estimate, we see that

$$\begin{aligned}
\|\alpha_{k+1} - v_h(0)\| &\le \|\alpha_{k+1}-\alpha_\infty\| + \|\alpha_\infty - v_h(0)\| \\
&\le \|G\|^M \|\alpha_k-\alpha_\infty\| + \|\alpha_\infty - v_h(0)\| \\
&\le \|G\|^M \|\alpha_k - v_h(0)\| + \left(1+\|G\|^M\right)\|\alpha_\infty - v_h(0)\| \\
&\le \|G\|^M \|\alpha_k - v_h(0)\| + C\tau^2\left(1-\|G\|^M\right)^{-1}\left(1+\|G\|^M\right).
\end{aligned}$$

Together with the fact that $\|G\| = \max_j \left|\frac{1-\tau\lambda_j/2}{1+\tau\lambda_j/2}\right| = \left|\frac{1-\tau\lambda_1/2}{1+\tau\lambda_1/2}\right| < 1$, we conclude the proof. □

4. Arnoldi approximation

Arnoldi approximation, [35–37], relies on Krylov subspace methods and is typically used to approximate matrix functions $F(B)$ or $F(B)b$, where $B \in \mathbb{C}^{N^d\times N^d}$ is a large matrix, and $b \in \mathbb{C}^{N^d}$ is a vector. The idea is to bypass the computation of large matrix functions and do the matrix computations on a lower dimensional subspace, which may result in tremendous computational gain. Consider, for example, the matrix exponential $F(B) = e^{-B}$, where $B \in \mathbb{R}^{N^d\times N^d}$ is a positive definite and symmetric matrix, which is also central to the goal of the present article. The starting point is a unitary transformation of the matrix B in the form $H = Q^*BQ$, where $Q \in \mathbb{R}^{N^d\times k}$ with $Q^*Q = I$, and $H \in \mathbb{R}^{k\times k}$, where $k \ll N^d$. We can then approximate $F(B)b$ for $b \in \mathbb{R}^{N^d}$ as follows

$$F(B)b \approx QF(H)Q^*b.$$

We summarize the properties of the matrix H, whenever the original matrix B is symmetric and positive definite. We can make the following immediate observation: Assume that the matrix Q has rank k. If the matrix B is positive definite and symmetric, the matrix H is also positive definite and symmetric. To prove the symmetric property, we write

$$H^* = Q^*B^*Q = Q^*BQ = H.$$

To prove positiveness, assume $y \in \mathbb{R}^k \neq \mathbf{0}$. Then

$$y^*Hy = y^*Q^*BQy > 0.$$

The computational efficiency here originates from the fact that now the matrix exponential is computed for a matrix of much lower rank. The following theorem provides an error estimate for the difference $F(B)b - QF(\mathcal{H})Q^*b$.

Theorem 2. *[Hochbruck,Lubich [37]] Let $B \in \mathbb{C}^{N^d\times N^d}$ be a Hermitian positive semi-definite matrix with eigenvalues in $[0,\rho]$. Moreover, let $H = Q^\star BQ$ be a unitary transformation of B via an Arnoldi procedure with $H \in \mathbb{R}^{k\times k}$ and $Q \in \mathbb{R}^{N^d\times k}$. Then the following estimate holds*

$$\|e^{-B}b - Qe^{-H}Q^\star b\|_2 \le \|b\|_2 \begin{cases} 10e^{-\frac{4k^2}{5\rho}}, & \sqrt{\rho} \le k \le \rho/2, \\ \dfrac{40}{\rho}e^{-\frac{\rho}{4}}\left(\dfrac{e\rho}{4k}\right)^k, & k \ge \rho/2. \end{cases} \tag{13}$$

Theorem 2 results in a particular computational advantage when applied to the second order difference operator A_h. The spectral radius ρ_{A_h} of the matrix A_h scales as h^{-2}. This will then imply that in order to obtain an exponential accuracy of the form e^{-C} (for some positive constant C), it suffices to choose $k = \sqrt{\frac{5\rho_{A_h}}{4}(C+\ln 10)} = O\left(h^{-1}\right)$. In other words, regardless of the dimension of the problem, we can compute $e^{-A_h}b$, where $b \in \mathbb{R}^{N^d}$ at the cost of computing the matrix exponential of a one dimensional problem.

In what follows, we will present two different ways of exploiting the Arnoldi method to gain computational efficiency in comparison to the direct approach from Section 3.1, and the shooting method from Section 3.2. Both approaches rely on relating the

solution $v(T)$ of the parabolic PDE (4) at time T to the initial value $v(0)$ and then approximating the emerging operators using a low rank Arnoldi approximation. This will then result in significantly lower computational cost while retaining the same accuracy of the direct discretization and the shooting method. For both approaches, we also include separate convergence analysis, which shows that the emerging matrix functions can be approximated accurately at the cost of matrix function computations of a one-dimensional problem.

Remark 3. Note that matrix exponentials of the form e^{-B} can also be computed by a standard eigenfunction expansion. This is, however, computationally very expensive and must be avoided in computations, see [38] for a full error analysis.

4.1. A hybrid shooting-Arnoldi approximation

Step 1. The starting point is to write the solution $v_h(T)$ of (5) in terms of the initial value and the right hand side

$$v_h(T) = e^{-TA_h} v_h(0) + \int_0^T e^{-(T-s)A_h} f_h(s) ds. \tag{14}$$

Step 2. Approximate the operators in (14) using the Arnoldi algorithm, and use the fixed point iteration to find $v_h(0)$ such that $v_h(0) = v_h(T) - \varphi_h$ is met, up to a desired tolerance. In other words, solve for

$$\alpha_n = Qe^{-TH} Q^* \alpha_{n-1} + \int_0^T Qe^{-(T-s)H} Q^* f_h(s) ds - \varphi_h, \tag{15}$$

for sufficiently large n such that $\alpha_n \approx v_h(0)$.

Step 3. Solve (5) as an initial value problem, where the initial data comes from Step 2.

4.1.1. Analysis

Here, we aim at establishing an error bound for the difference between α_n, defined by (15), and the true value of the initial data. The main result is stated in Lemma 2.

Lemma 1. *Suppose $A_h \in \mathbb{R}^{N^d \times N^d}$ is the second order difference matrix in (5). Let $v_h(0)$ be the vector satisfying the relations (14) and $v_h(0) = v_h(T) - \varphi_h$ exactly, and $\{\tilde{\alpha}_n\}$ be the sequence generated by the fixed point iterations*

$$\tilde{\alpha}_n = e^{-TA_h} \tilde{\alpha}_{n-1} + \int_0^T e^{-(T-s)A_h} f_h(s) ds - \varphi_h, \quad n = 1, 2, \dots. \tag{16}$$

Then it follows that

$$\|\tilde{\alpha}_n - v_h(0)\| \le e^{-Tn\lambda_1} \|\tilde{\alpha}_0 - v_h(0)\|, \quad n = 1, 2, \dots,$$

where $\lambda_1 > a_{\min} \frac{\pi^2}{d}$ is the smallest eigenvalue of the operator A_h.

Proof. Using the relation $v_h(0) = v_h(T) - \varphi_h$ together with (14) we see that

$$v_h(0) = e^{-TA_h} v_h(0) + \int_0^T e^{-(T-s)A_h} f_h(s) ds - \varphi_h.$$

Using the last equation and (16), we immediately see that

$$\begin{aligned} \|\tilde{\alpha}_n - v_h(0)\| &= \|e^{-TA_h} \left(\tilde{\alpha}_{n-1} - v_h(0)\right)\| \\ &\le \|e^{-TA_h}\| \|\tilde{\alpha}_{n-1} - v_h(0)\| \le e^{-T\lambda_1} \|\tilde{\alpha}_{n-1} - v_h(0)\|, \end{aligned}$$

where property 3 in Section 2 was used for establishing the last inequality. □

Lemma 2. *Let $A_h \in \mathbb{R}^{N^d \times N^d}$ be the same matrix as in Lemma 1 and $0 < \varepsilon < T$. Suppose $Q \in \mathbb{R}^{N^d \times k}$ and $H \in \mathbb{R}^{k \times k}$ with*

$$\sqrt{\frac{5T\rho_{A_h}}{4} \ln \frac{10}{1 - e^{-T\lambda_1}}} < k < \frac{\varepsilon \rho_{A_h}}{2} \tag{17}$$

are the corresponding matrices coming from an Arnoldi procedure applied to the matrix A_h. Moreover, let $v_h(0)$ be the vector satisfying the relations (14) and $v_h(0) = v_h(T) - \varphi_h$ exactly, and α_n be the solution of the fixed point iteration (15), with $\alpha_0 = \tilde{\alpha}_0$, where $\tilde{\alpha}_0$ is the initial guess for the iteration (16). Then it follows that

$$\begin{aligned} \|\alpha_n - v_h(0)\| \le\ & C^n \|\alpha_0 - v_h(0)\| \\ &+ 10e^{-\frac{4k^2}{5T\rho_{A_h}}} \left(\|v_h(0)\| + T \sup_{0 \le s \le T} \|f_h(s)\| \right) \frac{1 - C^n}{1 - C} \\ &+ \varepsilon \left(1 + \|Q\| \|Q^*\|\right) \sup_{0 \le s \le T} \|f_h(s)\| \frac{1 - C^n}{1 - C}, \end{aligned} \tag{18}$$

where $C = e^{-T\lambda_1} + 10e^{-\frac{4k^2}{5T\rho_{A_h}}} < 1$.

Proof. Let $\tilde{\alpha}_n$ and $v_h(0)$ be defined in the same way as in the proof of Lemma 1. We start by splitting the error $\alpha_n - v_h(0)$ into two parts as follows:

$$\|\alpha_n - v_h(0)\| \le \|\alpha_n - \tilde{\alpha}_n\| + \|\tilde{\alpha}_n - v_h(0)\|. \tag{19}$$

An upper bound for the second term in the right hand side was already established in Lemma 1. To bound the first term in the right hand side, we write

$$\begin{aligned}
\tilde{\alpha}_n - \alpha_n &= e^{-TA_h}\tilde{\alpha}_{n-1} + \int_0^T e^{-(T-s)A_h} f_h(s)ds - \varphi_h \\
&- \left(Qe^{-TH}Q^*\alpha_{n-1} + \int_0^T Qe^{-(T-s)H}Q^* f_h(s) - \varphi_h \right) \\
&= e^{-TA_h}\left(\tilde{\alpha}_{n-1} - \alpha_{n-1}\right) - \left(e^{-TA_h} - Qe^{-TH}Q^*\right)\left(\tilde{\alpha}_{n-1} - \alpha_{n-1}\right) \\
&+ \left(e^{-TA_h} - Qe^{-TH}Q^*\right)\left(\tilde{\alpha}_{n-1} - v_h(0)\right) \\
&+ \left(e^{-TA_h} - Qe^{-TH}Q^*\right) v_h(0) \\
&+ \int_0^T \left(e^{-(T-s)A_h} - Qe^{-(T-s)H}Q^*\right) f_h(s)ds.
\end{aligned}$$

Then,

$$\begin{aligned}
\|\tilde{\alpha}_n - \alpha_n\| &\le \left(\|e^{-TA_h}\| + \|e^{-TA_h} - Qe^{-TH}Q^*\|\right)\|\tilde{\alpha}_{n-1} - \alpha_{n-1}\| \\
&+ \|e^{-TA_h} - Qe^{-TH}Q^*\|\|\tilde{\alpha}_{n-1} - v_h(0)\| \\
&+ \|e^{-TA_h} - Qe^{-TH}Q^*\|\|v_h(0)\| \\
&+ \int_0^{T-\varepsilon} \|e^{-(T-s)A_h} - Qe^{-(T-s)H}Q^*\|\|f_h(s)\|ds \\
&+ \int_{T-\varepsilon}^T \|e^{-(T-s)A_h} - Qe^{-(T-s)H}Q^*\|\|f_h(s)\|ds.
\end{aligned}$$

Now, using the fact that $\|e^{-TA_h}\| \le e^{-T\lambda_1}$ (see property 3 in Section 2), as well as $\|e^{-(T-s)A_h} - Qe^{-(T-s)H}Q^*\| \le 10e^{-\frac{4k^2}{5(T-s)\rho_{A_h}}}, 0 \le s \le T-\varepsilon$ (see Theorem 2), we obtain

$$\begin{aligned}
\|\tilde{\alpha}_n - \alpha_n\| &\le \left(e^{-T\lambda_1} + 10e^{-\frac{4k^2}{5T\rho_{A_h}}}\right)\|\tilde{\alpha}_{n-1} - \alpha_{n-1}\| \\
&+ 10e^{-\frac{4k^2}{5T\rho_{A_h}}}\|\tilde{\alpha}_{n-1} - v_h(0)\| + 10e^{-\frac{4k^2}{5T\rho_{A_h}}}\|v_h(0)\| \\
&+ 10e^{-\frac{4k^2}{5T\rho_{A_h}}}(T-\varepsilon)\sup_{0\le s\le T-\varepsilon}\|f_h(s)\| \\
&+ \varepsilon\left(1 + \|Q\|\|Q^*\|\right)\sup_{T-\varepsilon\le s\le T}\|f_h(s)\|.
\end{aligned}$$

We note that the assumption (17) on k implies that the amplification factor $C = e^{-T\lambda_1} + 10e^{-\frac{4k^2}{5T\rho_{A_h}}} < 1$. Moreover, using Lemma 1 for the term $\|\tilde{\alpha}_{n-1} - v_h(0)\|$, we get

$$\begin{aligned}
\|\tilde{\alpha}_n - \alpha_n\| &\le C\|\tilde{\alpha}_{n-1} - \alpha_{n-1}\| + 10e^{-\frac{4k^2}{5T\rho_{A_h}}}e^{-T(n-1)\lambda_1}\|\tilde{\alpha}_0 - v_h(0)\| \\
&+ 10e^{-\frac{4k^2}{5T\rho_{A_h}}}\left(\|v_h(0)\| + T\sup_{0\le s\le T}\|f_h(s)\|\right) \\
&+ \varepsilon\left(1 + \|Q\|\|Q^*\|\right)\sup_{0\le s\le T}\|f_h(s)\|.
\end{aligned}$$

Then, by a trivial induction we have

$$\begin{aligned}
\|\tilde{\alpha}_n - \alpha_n\| &\le C^n\|\tilde{\alpha}_0 - \alpha_0\| + 10e^{-\frac{4k^2}{5T\rho_{A_h}}}\|\tilde{\alpha}_0 - v_h(0)\|\sum_{\ell=0}^{n-1} C^\ell e^{-T(n-1-\ell)\lambda_1} \\
&+ 10e^{-\frac{4k^2}{5T\rho_{A_h}}}\left(\|v_h(0)\| + T\sup_{0\le s\le T}\|f_h(s)\|\right)\sum_{\ell=0}^{n-1} C^\ell \\
&+ \varepsilon\left(1 + \|Q\|\|Q^*\|\right)\sup_{0\le s\le T}\|f_h(s)\|\sum_{\ell=0}^{n-1} C^\ell.
\end{aligned}$$

The final estimate follows then from (19) under the assumption $\tilde{\alpha}_0 = \alpha_0$.

Remark 4. Note that the estimate (18) in Lemma 2 is sub-optimal due to the presence of an additional $O(\varepsilon)$ term in the upper bound. Indeed, in numerical simulations we only observe the exponentially decaying part of the error, see the numerical results in Section 5.1.

4.2. An approach based on a pure Arnoldi approximation

Step 1. Unlike the previous approach, instead of using iterations to solve for $v_h(0)$, one can also formulate a direct equation by exploiting (14), together with $v_h(0) = v_h(T) - \varphi_h$ to represent the initial data $v_h(0)$ as follows

$$v_h(0) = (I - e^{-TA_h})^{-1} \left(-\varphi_h + \int_0^T e^{-(T-s)A_h} f_h(s) ds \right). \tag{20}$$

Step 2. Approximate the matrix operators $(I - e^{-TA_h})^{-1}$, and $e^{-A_h(T-s)}$ using the Arnoldi algorithm, and solve for $v_h(0)$.
Step 3. Solve (5) as an initial value problem, where the initial data comes from Step 2.

4.2.1. Analysis

In this approach, in addition to approximations of exponential functions, we also need to approximate the operator $(I-e^{-TA_h})^{-1}$. In what follows, we aim at establishing error bounds for an Arnoldi approximation for this operator. The main result is Theorem 3. However, we will need the following lemma prior to proving Theorem 3.

Lemma 3 (*See Chapter 11, [39]*). *If a function f has a power series representation*

$$f(z) = \sum_{k=0}^{\infty} c_k z^k$$

on an open disk containing the eigenvalues $\lambda(B)$ of a matrix $B \in \mathbb{C}^{N^d \times N^d}$, then

$$f(B) = \sum_{k=0}^{\infty} c_k B^k,$$

and,

$$\| f(B) - \sum_{j=0}^{k-1} c_j B^j \| \leq \frac{N^d}{k!} \max_{0 \leq s \leq 1} \| B^k f^{(k)}(Bs) \|. \tag{21}$$

Now suppose $b(z) = (1 - e^{-z})^{-1}$. We are interested in deriving an error estimate for

$$\| b(TA_h)g - Qb(TH)Q^* g \|.$$

Theorem 3. *Let $A_h \in \mathbb{R}^{N^d \times N^d}$ be as in Lemma 2, and $Q \in \mathbb{R}^{N^d \times k}$ and $H \in \mathbb{R}^{k \times k}$ with $\mu_1 = \min \lambda(H)$, be the matrices coming from the Arnoldi procedure. Then the following estimate holds:*

$$\begin{aligned} \| \, b(TA_h)g \, - \, Qb(TH)Q^* g \, \| \leq & \left(N^d e^{-T\lambda_1 m} \frac{1}{(1 - e^{-T\lambda_1})^{m+1}} \right. \\ & \left. + \, Cke^{-T\mu_1 m} \frac{1}{(1 - e^{-T\mu_1})^{m+1}} + 10(m-1) e^{-\frac{4k^{2-\beta}}{5\rho}} \right) \| g \|, \end{aligned}$$

where $m = \lfloor k^\beta \rfloor$, $0 < \beta < 1$, and the constant $C = \|Q\| \|Q^*\|$.

Proof. Let $f(z) = (1 - z)^{-1}$. Then for $|z| < 1$, we can write

$$f(z) = \sum_{j=0}^{\infty} z^j = f_k(z) + \sum_{j=k}^{\infty} z^j, \quad |z| < 1,$$

where $f_k(z) = \sum_{j=0}^{k-1} z^j$. Moreover, let $B = e^{-TA_h}$, then clearly $\lambda(B) \in [e^{-T\rho_{A_h}}, e^{-T\lambda_1}]$, and by Lemma 3 it follows that

$$b(TA_h) = f(B) = f_k(B) + \sum_{j=k}^{\infty} B^j,$$

where $f_k(B)$ is given by

$$f_k(B) = \sum_{j=0}^{k-1} B^j.$$

Note that the spectral radius of B is strictly less than 1. Now, assume $0 < \beta < 1$ and $m = \lfloor k^\beta \rfloor$, and consider the following decomposition of the overall error:

$$\| f(B)g - Qf(e^{-TH})Q^* g \| \leq \| f(B)g - f_m(B)g \| + \| f_m(B)g - Qf(e^{-TH})Q^* g \|.$$

To bound the first term, we use inequality (21) in Lemma 3

$$\begin{aligned}\|f(B)g - f_m(B)g\| &\le \|f(B) - f_m(B)\|\|g\| \\ &\le \frac{N^d}{m!}\max_{0\le s\le 1}\|B^m f^{(m)}(Bs)\|\|g\| \\ &\le \frac{N^d}{m!}\|B^m\|\max_{0\le s\le 1}\|f^{(m)}(Bs)\|\|g\| \\ &\le \frac{N^d}{m!}e^{-T\lambda_1 m}\max_{0\le s\le 1}\|\frac{m!}{(I-Bs)^{m+1}}\|\|g\| \\ &\le N^d e^{-T\lambda_1 m}\frac{1}{(1-e^{-T\lambda_1})^{m+1}}\|g\|.\end{aligned}$$

For the second term, we consider the decomposition

$$\begin{aligned}\left\|f_m(B)g - Qf\left(e^{-TH}\right)Q^*g\right\| &\le \left\|f_m(B)g - Qf_m\left(e^{-TH}\right)Q^*g\right\| \\ &+ \left\|Q\left(f\left(e^{-TH}\right) - f_m\left(e^{-TH}\right)\right)Q^*g\right\|.\end{aligned}$$

The first term in the right hand side can be bounded as follows

$$\begin{aligned}\left\|f_m(B)g - Qf_m\left(e^{-TH}\right)Q^*g\right\| &\le \sum_{\ell=1}^{m-1} 10\|g\|e^{-\frac{4k^2}{5\ell\rho}} \\ &\le 10(m-1)\|g\|e^{-\frac{4k^2}{5(m-1)\rho}} \\ &\le 10(m-1)\|g\|e^{-\frac{4k^2-\beta}{5\rho}}.\end{aligned}$$

The second term, on the other hand, is bounded in a similar way as $\|f(B)g - f_m(B)g\|$ with an additional constant accounting for the l^2 norms of Q and Q^*, i.e.,

$$\left\|Q\left(f\left(e^{-TH}\right) - f_m\left(e^{-TH}\right)\right)Q^*g\right\| \le Cke^{-T\mu_1 m}\frac{1}{\left(1-e^{-T\mu_1}\right)^m}\|g\|.$$

Remark 5. Choosing $0 < \beta < 1$ sufficiently small and $k = O(N)$, Theorem 3 results in exponentially decaying error bounds in terms of N, while the cost would be comparable to computing the matrix functions of sizes $k \times k$.

4.3. Numerical integration of matrix exponentials

In the analysis provided in Sections 4.1 and 4.2, it was assumed that the integral $\int_0^T e^{-(T-t)A_h} f_h(t)\ dt$ and it's approximation $\int_0^T Qe^{-(T-t)H}Q^* f_h(t)\ ds$ are given exactly. In practice these integrals need to be approximated by a quadrature rule. Here, we draw the attention of the reader to a subtlety in relation with the quadrature rule. For simplicity, we consider only the integral $\int_0^T e^{-(T-s)A_h} f_h(s)\,ds$. A quadrature rule based on the standard midpoint rule yields

$$\int_0^T e^{-(T-t)A_h} f_h(t)dt = \tau\sum_{k=1}^{M} e^{-(T-t_{k-1/2})A_h} f_h(t_{k-1/2}) + O(\tau^2).$$

The error bound for such an approximation includes the second derivative of the integrand $e^{-(T-t)A_h} f_h(t)$, which would scale as $\|A^2\| \approx (1/h)^4 \approx N^4$. Therefore, a direct application of the midpoint rule will be very costly since very small time steps will be needed to achieve reasonable error tolerances. The problematic part with the error bound for the standard midpoint rule is that the exponential term in the integrand is also approximated at the midpoint. We will now, instead, derive a variant of the midpoint rule which bypasses the approximation of the exponential part of the integrand. This would then result in a much smaller pre-factor in the error bound as we will see in a nutshell. To achieve this, we start by

$$\begin{aligned}\int_0^T e^{-(T-t)A_h} f_h(t)dt &= \sum_{k=1}^{M}\int_{t_{k-1}}^{t_k} e^{-(T-t)A_h} f_h(t)dt \\ &\approx \sum_{k=1}^{M}\int_{t_{k-1}}^{t_k} e^{-(T-t)A_h}\ dt f_h(t_{k-1/2})\end{aligned}$$

Moreover,

$$\int_{t_{k-1}}^{t_k} e^{-(T-t)A_h}dt = e^{-TA_h}\int_{t_{k-1}}^{t_k} e^{tA_h}dt = e^{-TA_h}\left(e^{t_k A_h} - e^{t_{k-1}A_h}\right)A_h^{-1}.$$

Therefore,

$$\int_0^T e^{-(T-t)A_h} f_h(t)dt = \sum_{k=1}^{M}\left(e^{-(T-t_k)A_h} - e^{-(T-t_{k-1})A_h}\right)A_h^{-1} f_h(t_{k-1/2}) + O(\tau).$$

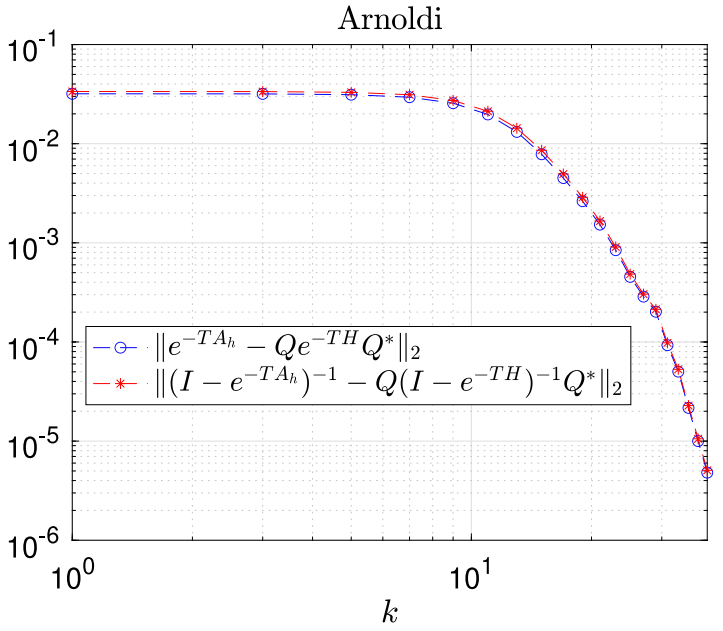


Fig. 1. Approximation error for the matrix functions computations via Arnoldi.

The procedure above results in a first order method but the pre-factor in the upper bound is independent of A, and includes only the first derivative $\partial_t f_h(t)$. In our simulations below, we also use a Richardson extrapolation to achieve a globally second-order method in time, which proves to be much more efficient than a standard midpoint rule.

5. Numerical results

In this section, we provide numerical examples to validate the theoretical findings. In Section 5.1, we provide a numerical example to demonstrate the exponential convergence for the approximation of the matrix exponential e^{-TA_h} as well as $(I-e^{-TA_h})^{-1}$. In Sections Section 5.2 we provide numerical examples in one, two, and three dimensions to demonstrate the advantage of using the full Arnoldi approximation or the hybrid approach in comparison to the standard shooting method.

5.1. Matrix function approximations-numerical results

In this section, we provide numerical evidence for exponentially decaying errors originating from the Arnoldi approximation of the operators e^{-TA_h}, and $(I - e^{-TA_h})^{-1}$. The size of the matrix $A_h \in \mathbb{R}^{N^2\times N^2}$ is 1600×1600. Fig. 1 demonstrates the exponential decay of the error corroborating the results of Theorems 2 and 3. In particular, we observe accuracies down to 10^{-5} tolerances for reasonably low rank approximation; i.e., $k \approx 40$. Note that, in this simulation the value of the parameter T is chosen to be $T = 0.1$.

5.2. Approximation of u and p

In this section, we consider the inverse problem of determining the pair of solutions $\{p, u\}$ in dimensions $d = 1, 2, 3$:

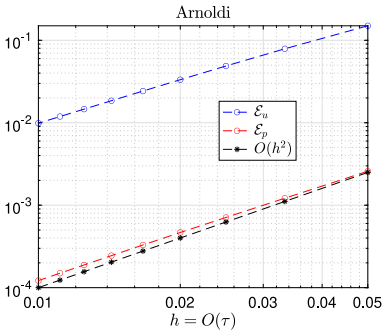
$$\begin{aligned} \partial_t u(t,x) \; - \; \Delta u(t,x) &= f(t,x) + p(x), \quad \text{in } K = (0,1)^d \times (0,T), \\ u(t,x) &= 0, \quad \text{on } \partial K \times (0,T), \\ u(0,x) &= 0, \quad \text{in } K, \\ u(T,x) &= \varphi(x), \quad \text{in } K. \end{aligned}$$

To study the convergence properties, we assume an exact solution u (only for the sake of comparison) of the form

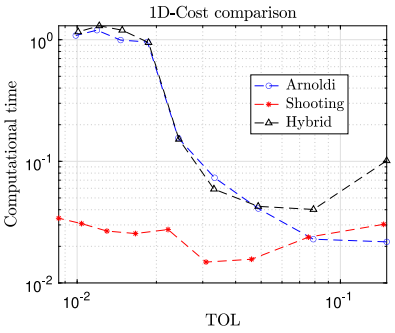
$$u(t,x) = \left(e^{-t} - 1\right) \prod_{j=1}^{d} \sin^2(2\pi x_j),$$

and

$$f(t,x) \; = \; -e^{-t} \left(\prod_{j=1}^{d} \sin^2(2\pi x_j) + 8\pi^2 \sum_{j=1}^{d} \cos(4\pi x_j) \prod_{k\neq j}^{d} \sin^2(2\pi x_k) \right),$$

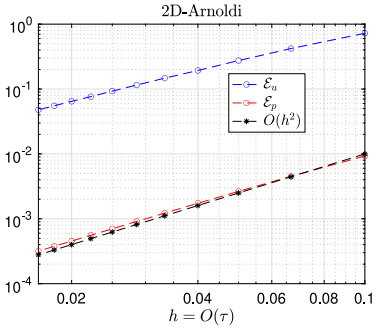


(a) Convergence of u and p in $d = 1$

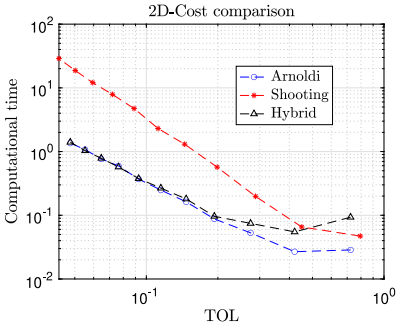


(b) Cost comparison in $d = 1$

Fig. 2. Numerical results in one dimension.



(a) Convergence of u and p in $d = 2$



(b) Cost comparison in $d = 2$

Fig. 3. Numerical results in two dimensions.

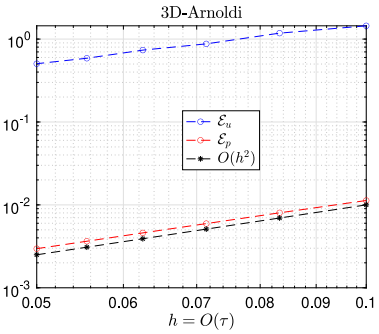
$$p(x) = 8\pi^2 \sum_{j=1}^{d} \cos(4\pi x_j) \prod_{k\neq j}^{d} \sin^2(2\pi x_k),$$
$$\varphi(x) = u(T, x).$$

In Figs. 2(a),3(a),4(a), the relative errors

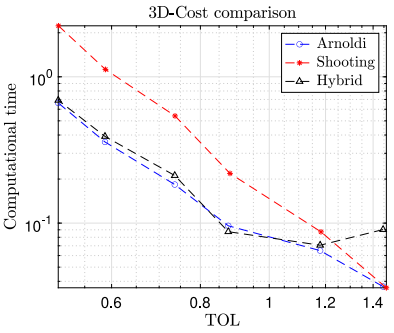
$$\mathcal{E}_u := \max_{k,j} \left| \frac{u(t_k, x_j) - u_{k,j}}{u(t_k, x_j)} \right|, \quad \mathcal{E}_p := \max_{j} \left| \frac{p(x_j) - p_j}{p(x_j)} \right|$$

are depicted for dimensions $d = 1, 2, 3$ respectively. Here $u_{k,j}$ and p_j are numerical approximations to $u(t_k, x_j)$ and $p(x_j)$ respectively. The error plots include only the error corresponding the full Arnoldi approach. This is intentional since the error plots for the hybrid approach as well as the shooting algorithm are almost identical, and all show second order convergence rates in time and space. Note that in all of the simulations the time-step τ is simultaneously refined (proportional to the spatial stepsize h). More precisely, we use $\tau = 0.1h$ for the one- and two-dimensional cases and $\tau = 0.05h$ for the three-dimensional case. Moreover, the final time T is set as $T = 0.1$, and the number of basis vectors in the Arnoldi algorithm is chosen as $k = N = h^{-1}$ in all of the simulations.

To compare the efficiency of the methods, we also report, in Figs. 2(b), 3(b),4(b), the actual computational time (measured in seconds) to reach a desired error tolerance. For dimensions $d = 2, 3$, the results clearly demonstrate the advantage of using the hybrid and the full Arnoldi approaches over the classical shooting method, whereas the shooting method seems to be more efficient only in the one-dimensional setting. Note that the case of $d = 1$ is somewhat special. The Arnoldi algorithm requires computing a matrix exponential in one-dimension, which is more costly than solving a one-dimensional parabolic PDE using an implicit method, whose cost is dominated by inverting a matrix of size $N^d \times N^d$. This is due to the fact that the overall cost of the Arnoldi algorithm is $O(N^d k^2)$, where k is the number of basis functions in the Krylov-subspace, see [38] for details, whereas the inversion of a matrix of size $N^d \times N^d$, with bandwidth $p = N^{d-1}$ entails $O(N^d p^2) = O(N^{3d-2})$ operations. Therefore, the advantage of the Arnoldi algorithm over the shooting method can not be seen unless $d \geq 2$.



(a) Convergence of u and p in $d = 3$



(b) Cost comparison in $d = 3$

Fig. 4. Numerical results in three dimensions.

Data availability

No data was used for the research described in the article.

References

[1] A.I. Prilepko, D.G. Orlovsky, I.A. Vasin, Methods for solving inverse problems in mathematical physics, in: Chapman & Hall/CRC Pure and Applied Mathematics, CRC Press, 2000.
[2] V. Isakov, Inverse problems for partial differential equations, in: Applied Mathematical Sciences, vol. 127, Springer, 2006.
[3] S.I. Kabanikhin, Inverse and ill-posed problems: Theory and applications, in: Inverse and Ill-Posed Problems Series, vol. 55, De Gruyter, 2011.
[4] T. Kimura, T. Suzuki, A parabolic inverse problem arising in a mathematical model for chromatography, SIAM J. Appl. Math. 53 (6) (1993) 1747–1761.
[5] Q. Chen, J. Liu, Solving an inverse parabolic problem by optimization from final measurement data, J. Comput. Appl. Math. 193 (1) (2006) 183–203.
[6] L. Yang, J.-N. Yu, Z.-C. Deng, An inverse problem of identifying the coefficient of parabolic equation, Appl. Math. Model. 32 (10) (2008) 1984–1995.
[7] M. Dehghan, Determination of a control parameter in the two-dimensional diffusion equation, Appl. Numer. Math. 37 (4) (2001) 489–502.
[8] M. Dehghan, F. Shakeri, Method of lines solutions of the parabolic inverse problem with an overspecification at a point, Numer. Algorithms 50 (2009) 417–437.
[9] L. Yang, M. Dehghan, J.-N. Yu, G.-W. Luo, Inverse problem of time-dependent heat sources numerical reconstruction, Math. Comput. Simulation 81 (8) (2011) 1656–1672.
[10] A. Ashyralyev, A.S. Erdogan, O. Demirdag, On the determination of the right-hand side in a parabolic equation, Appl. Numer. Math. 62 (11) (2012) 1672–1683.
[11] M.K. e Oshagh, M. Abbaszadeh, E. Babolian, H. Pourbashash, An adaptive wavelet collocation method for the optimal heat source problem, Internat. J. Numer. Methods Heat Fluid Flow 32 (7) (2021) 2360–2382.
[12] Z.-C. Deng, L. Yang, Unicity of solution for a semi-infinite inverse heat source problem, AIMS Math. 7 (4) (2022) 7026–7039.
[13] U. Biccari, Y. Song, X. Yuan, E. Zuazua, A two-stage numerical approach for the sparse initial source identification of a diffusion-advection equation, 2022, http://dx.doi.org/10.48550/arXiv.2202.01589, arXiv.
[14] C.H. Huang, M.N. Ozisik, Inverse problem of determining the unknown strength of an internal plane heat source, J. Franklin Inst. 329 (4) (1992) 751–764.
[15] F.-B. Liu, A modified genetic algorithm for solving the inverse heat transfer problem of estimating plan heat source, Int. J. Heat Mass Transfer 51 (15) (2008) 3745–3752.
[16] J. Cheng, J. Liu, An inverse source problem for parabolic equations with local measurements, Appl. Math. Lett. 103 (2020) 106213.
[17] M.B. Pintarelli, Solving the heat source inverse problem with moments problems techniques, London J. Res. Sci.: Nat. Formal 20 (1) (2020).
[18] Y.S. Eidelman, An inverse problem for an evolution equation, Math. Notes 49 (5) (1991) 535–540.
[19] A. Ashyralyev, On the problem of determining the parameter of a parabolic equation, Ukrainian Math. J. 62 (9) (2011) 1397–1408.
[20] M. Choulli, M. Yamamoto, Generic well-posedness of a linear inverse parabolic problem with diffusion parameters, J. Inverse Ill-Posed Probl. 7 (3) (1999) 241–254.
[21] A. Hasanov, M. Slodička, An analysis of inverse source problems with final time measured output data for the heat conduction equation: A semigroup approach, Appl. Math. Lett. 26 (2) (2013) 207–214.
[22] Y. Cao, M. Gunzburger, J. Turner, The controllability of systems governed by parabolic differential equations, J. Math. Anal. Appl. 215 (1) (1997) 174–189.
[23] H. Amann, P. Quittner, Optimal control problems with final observation governed by explosive parabolic equations, SIAM J. Control Optim. 44 (4) (2005) 1215–1238.
[24] A. Ashyralyev, C. Ashyralyyev, On the stability of parabolic differential and difference equations with a time-nonlocal condition, Comput. Math. Math. Phys. 62 (2022) 962–973.
[25] V.N. Starovoitov, Unique solvability of a linear parabolic problem with nonlocal time data, Sib. Math. J. 62 (2021) 337–340.
[26] J. Martín-Vaquero, S. Sajavičius, The two-level finite difference schemes for the heat equation with nonlocal initial condition, Appl. Math. Comput. 342 (2019) 166–177.
[27] Y. Lin, Analytical and numerical solutions for a class of nonlocal nonlinear parabolic differential equations, SIAM J. Math. Anal. 25 (6) (1994) 1577–1594.
[28] A. Ashyralyev, C. Ashyralyyev, Numerical solution of time-nonlocal problem for parabolic equation, AIP Conf. Proc. 2483 (1) (2022) 060001, http://dx.doi.org/10.1063/5.0114878.
[29] M. Geiger, Adaptive multiple shooting for boundary value problems and constrained parabolic optimization problems, 2015, URL https://api.semanticscholar.org/CorpusID:126378561.
[30] Y.-W. Chen, A backward-forward Lie-group shooting method for nonhomogeneous multi-dimensional backward heat conduction problems under a long time span, Int. J. Heat Mass Transfer 133 (2019) 226–246.
[31] L.E. Payne, H.F. Weinberger, An optimal Poincaré inequality for convex domains, Arch. Ration. Mech. Anal. 5 (1960) 286–292.

[32] A.A. Samarskii, E.S. Nikolaev, Numerical Methods for Grid Equations: Direct Methods, Birkhäuser, 1988.
[33] I. Almuslimani, G. Vilmart, Explicit stabilized integrators for stiff optimal control problems, SIAM J. Sci. Comput. 43 (2) (2021) A721–A743.
[34] X. Liu, J. Frank, Symplectic Runge–Kutta discretization of a regularized forward–backward sweep iteration for optimal control problems, J. Comput. Appl. Math. 383 (2021) 113133.
[35] S. Güttel, Rational Krylov approximation of matrix functions: Numerical methods and optimal pole selection, GAMM-Mitt. 36 (1) (2013) 8–31.
[36] N.J. Higham, Functions of Matrices: Theory and Computation, Society for Industrial and Applied Mathematics, 2008.
[37] M. Hochbruck, C. Lubich, On Krylov subspace approximations to the matrix exponential operator, SIAM J. Numer. Anal. 34 (5) (1997) 1911–1925.
[38] A. Abdulle, D. Arjmand, E. Paganoni, An elliptic local problem with exponential decay of the resonance error for numerical homogenization, Multiscale Model. Simul. 21 (2) (2023) 513–541.
[39] G.H. Golub, C.F. Van Loan, Matrix computations, in: Johns Hopkins Studies in the Mathematical Sciences, Johns Hopkins University Press, 2013.