hp-DG-QTT solution of high-dimensional degenerate diffusion equations

Eidgenössische Technische Hochschule Zürich

Ecole polytechnique fédérale de Zurich
Politecnico federale di Zurigo
Swiss Federal Institute of Technology Zurich

hp-DG-QTT solution of high-dimensional degenerate diffusion equations

V. Kazeev, O. Reichmann and Ch. Schwab

Research Report No. 2012-11
May 2012

Seminar für Angewandte Mathematik
Eidgenössische Technische Hochschule
CH-8092 Zürich
Switzerland

hp-DG-QTT solution of high-dimensional degenerate diffusion equations

Vladimir Kazeev* Oleg Reichmann* Christoph Schwab*

May 3, 2012

Abstract

We consider the discretization of degenerate, time-inhomogeneous Fokker-Planck equations for diffusion problems in high-dimensional domains. Well-posedness of the problem in time-weighted Bochner spaces is established. Analytic regularity of the time-dependence of the solution in countably normed, weighted Sobolev spaces is established. Time discretization by the *hp*-discontinuous Galerkin method is shown to converge exponentially. The resulting elliptic spatial problems are discretized with the use of the tensor-product "hat" finite elements constructed on uniform or patch-wise uniform (Shishkin) meshes and are solved in the *Quantized Tensor Train* representation. For numerical experiments we consider compatible and incompatible initial data in up to 40 and 18 dimensions respectively on a workstation.

Keywords: Fokker-Planck equation, degenerate diffusion, Gevrey regularity, *hp*-discontinuous Galerkin, time stepping, low-rank representation, Tensor Train (TT), Quantized Tensor Train (QTT).
Mathematics Subject Classification (2000): 15A69, 35K05, 35K15, 35K65, 65M12, 65M60, 65F99.

1 Introduction

We consider the efficient numerical solution of a class of degenerate, time-inhomogeneous diffusion problems on high-dimensional state spaces with tensorized geometry. While our primary motivation are Fokker-Planck equations which arise from certain time-inhomogeneous stochastic processes, among others, the present methods can be applied also in other contexts. The present approach is based on a *variational hp time-semidiscretization* of the evolution problem in time-weighted Bochner-Lebesgue spaces. These discretization are of *hp*- resp. of variable order, variable step size, and reduce the evolutionary Fokker-Planck equation to a *sequence of independent, stationary, elliptic problems*.

As we show here, the solution is, as a function of the time-variable, analytic resp. Gevrey-regular with respect to t. This implies the *exponential convergence* of the *hp*-DG time stepping on geometrically refined time-partitions and for linearly increasing polynomial orders. We then consider the efficient solution of the resulting sequence of high-dimensional spatial problems. To this end, we propose here a tensor product finite difference discretization of the spatial, elliptic systems and solve the resulting family of linear systems by the so-called *Quantized Tensor Train* (QTT) formatted matrix algebra. As we show in numerical experiments, the intrinsically nonlinear numerical approximations which are generated by the QTT discretizations are, as a rule, adapted in resolution with respect to the spatial dimension as well as in the "quantized" indexing which is employed in the QTT matrix format for all linear algebra operations. The compression realized by the *hp*-DG time stepping combined with the QTT representation is found to be superior to that of other sparse tensor discretizations of parabolic evolution problems.

*Seminar für Angewandte Mathematik, ETH Zürich. Rämistrasse 101, 8092 Zrich, Switzerland. {vladimir.kazeev,oleg.reichmann,christoph.schwab}@sam.math.ethz.ch.

2 Problem Formulation

2.1 Preliminaries

We consider evolution equations with coefficients which may be strongly degenerate functions of time. Therefore, solutions of these equations will *not*, in general, be continuous and bounded with respect to time up to $t = 0$. In order to introduce a suitable notion of solution, we shall require the following classes of time-weighted function spaces in order to state space time variational formulations of the evolution problems and, finally, to establish their well-posedness:

$$\begin{aligned}
\mathcal{X} &:= H^1_{t^{-\gamma/2}}(J;X^*) \cap L^2_{t^{\gamma/2}}(J;X) \cong \left(H^1_{t^{-\gamma/2}}(J) \otimes X^*\right) \cap \left(L^2_{t^{\gamma/2}}(J) \otimes X\right), && (2.1)\\
\mathcal{Y} &:= L^2_{t^{\gamma/2}}(J;X) \cong L^2_{t^{\gamma/2}}(J) \otimes X, && (2.2)\\
\mathcal{X}_{(0} &:= \{w \in \mathcal{X} : w(0,\cdot) = 0 \text{ in } X^*\}, && (2.3)\\
\mathcal{X}_{0)} &:= \{w \in \mathcal{X} : w(T,\cdot) = 0 \text{ in } X^*\}. && (2.4)
\end{aligned}$$

Here $X := H^1_0(D)$, $X^* = H^{-1}(D)$, $L^2_{t^{\gamma/2}}(J) = \overline{C^\infty(J)}^{\|\cdot\|_{L^2_{t^{\gamma/2}}(J)}}$ and $H^1_{t^{\gamma/2}}(J) = \overline{C^\infty(J)}^{\|\cdot\|_{H^1_{t^{\gamma/2}}(J)}}$, $J = (0,T)$. The weighted norms appearing in the above definitions are given by

$$\|u\|^2_{L^2_{t^{\gamma/2}}(J)} := \int_J u^2 t^\gamma \,\mathrm{d}t, \qquad \|u\|^2_{H^1_{t^{\gamma/2}}(J)} := \int_J u^2 t^\gamma \,\mathrm{d}t + \int_J \dot{u}^2 t^\gamma \,\mathrm{d}t\,.$$

To describe the smoothness of the initial data we define intermediate spaces between $H = L^2(D)$ and X by the real method of function space interpolation: specifically,

$$H_\theta = (H,X)_{\theta,2}, \quad 0 < \theta < 1\,,$$

where we define $H_0 := H$ and $H_1 := X$.

We next present the class of evolution problems to be considered. Since the problems are possibly degenerate in the time parameter t, we use the weighted spaces (2.1)-(2.4) and recapitulate the variational formulation and present some results on their well-posedness from [1].

2.2 γ-homogeneous diffusion

We consider an additive Markov process $X(t)$ (in the sense of Sato [2]) with characteristic triplet $(t^\gamma b, t^\gamma \mathcal{K}, 0)$ for $b \in \mathbb{R}^d, \gamma \in (-1,1)$ and positive-definite $\mathcal{K} \in \mathbb{R}^{d\times d}$ and focus on the Fokker-Planck equations of possibly degenerate diffusions $X(t)$. In particular, time-inhomogeneous processes $X(t)$ are considered for which the infinitesimal generator $\mathcal{A}(t)$ of $X(t)$ takes the (space-time separable) form

$$\mathcal{A}(t) = t^\gamma \left(b^\top \nabla + \mathrm{tr}[\mathcal{K}D^2]\right)\,, \tag{2.5}$$

where we use the following notation:

$$D^2 = \begin{pmatrix} \partial_{x_1}\partial_{x_1} & \cdots & \partial_{x_1}\partial_{x_d} \\ \vdots & \ddots & \vdots \\ \partial_{x_d}\partial_{x_1} & \cdots & \partial_{x_d}\partial_{x_d} \end{pmatrix},$$

so that $\mathrm{tr}[\mathcal{K}D^2] = \nabla^\top \mathcal{K} \nabla$ and the Laplacian $\triangle\varphi$ of φ can be written as $\triangle\varphi = \sum_{i=1}^d \partial_{x_i}\partial_{x_i}\varphi = \mathrm{tr}[D^2]\varphi$. The corresponding backward Fokker-Planck equation with initial data $u_0(x)$ reads: find $u(t,x)$ such that

$$\begin{aligned}
\partial_t u + \mathcal{A}(t)u - cu &= 0 \quad \text{in} \quad [0,T)\times\mathbb{R}^d,\\
u(T) &= u_0 \quad \text{in} \quad \mathbb{R}^d,
\end{aligned} \tag{2.6}$$

for some sufficiently smooth final condition u_0 and some constant $c > 0$. The drift and killing terms in (2.6) can be removed via a change of variable. Besides, we can approximate the Cauchy problem (2.6) by an initial boundary value problem on a bounded spatial domain. We refer to [3, 1] for a priori estimates of

the localization error. The main difficulty in the numerical solution (and in the formulation) of the initial boundary value problem resides in the possible weak degeneracy of the diffusion coefficient in time, due to the γ-homogeneity. We therefore consider the following model problem: find $u \in \mathcal{X}$ such that for all $v \in \mathcal{Y}$

$$\begin{aligned} \int_0^T \langle \partial_t u(t), v(t) \rangle + t^\gamma \langle \mathcal{B}\, u(t), v(t) \rangle \, \mathrm{d}t &= \int_0^T \langle f(t), v(t) \rangle \, \mathrm{d}t, \\ u(0) &= u_0 \end{aligned} \tag{2.7}$$

holds, where the spatial differential operator $\mathcal{B}$ is defined by

$$\mathcal{B} = -\sum_{p,q=1}^{d} \frac{\partial}{\partial x_p} \kappa^{pq}(x) \frac{\partial}{\partial x_q},$$

for $\gamma \in (-1, 1)$, a bounded Lipschitz domain $D \subset \mathbb{R}^d$, a finite time interval $J = (0, T)$ and smooth functions $\kappa^{pq}(x)$, $1 \le p, q \le d$, such that for some $\overline{\kappa}, \underline{\kappa} > 0$ it holds that $\underline{\kappa}\,|\xi|^2 \le \xi^\top \mathcal{K}(x)\, \xi \le \overline{\kappa}\,|\xi|^2$ for all $\xi \in \mathbb{R}^d$, $x \in D$ and $\mathcal{K} = \mathcal{K}^\top$ on D. The bilinear form $\mathbf{b}$ associated with the spatial differential operator $\mathcal{B}$ reads

$$\mathbf{b}(u, v) : X \times X \to \mathbb{R}, \qquad \mathbf{b}(u, v) = \langle \mathcal{B}\, u, v \rangle, \quad \forall u, v \in X, \tag{2.8}$$

where $X := H_0^1(D)$ and where $\langle \cdot, \cdot \rangle$ denotes the $H^{-1}(D) \times H_0^1(D)$ duality pairing.

Theorem 2.1 ([1])**.** *For every $f \in \mathcal{Y}^*$ and for homogeneous initial data $u_0 = 0$ in X^*, the evolution problem* (2.7) *admits a unique solution $u \in \mathcal{X}_{(0}$ and there holds the a priori error estimate*

$$\|u\|_{\mathcal{X}} \le \sqrt{2}\, \|f\|_{\mathcal{Y}^*} \ .$$

Remark 2.2. *Similar equations to* (2.7) *arise in the context of option pricing under fractional Brownian motion (FBM) models, we refer to [4] for an introduction to FBM and to [1] for details on the Kolmogorov equations for such processes.*

2.3 Gevrey time-inhomogeneity

The well-posedness result as well as the ensuing error analysis of *hp* time-discretizations will remain valid verbatim for slightly more general classes of time-dependent coefficients: instead of (2.5) we may consider

$$\mathcal{A}(t)\varphi(x) = -a(t)\, \mathcal{B}\, \varphi(x) = a(t) b^\top \nabla \varphi(x) + a(t)\, \mathrm{tr}[\mathcal{K} D^2 \varphi(x)] \ . \tag{2.9}$$

Here, b and $\mathcal{K}$ are as above (in particular, they are independent of t) and the coefficient $a(t)$ is *γ-pseudohomogeneous* in the following sense: it is independent of the spatial variable x and there exist $0 < c_1 \le c_2 < \infty$, possibly depending on T, such that

$$c_1 \le a(t)/t^\gamma \le c_2 \quad \forall t \in (0, T] \ . \tag{2.10}$$

Moreover, we may assume that $a(t)$ *is Gevrey-δ regular with respect to time*, i. e. $a(t) \in C^\infty(0, T]$ and

$$\exists c > 0,\, \delta \ge 1 :\ \forall m \in \mathbb{N}\ \forall t \in (0, T] : \qquad \left| \frac{d^m}{dt^m} a(t) \right| \le c^{m+1} (m!)^\delta t^{\gamma - m} \ . \tag{2.11}$$

Note that $\delta = 1$ implies analyticity of $a(t)$ with respect to t at every point $t \in (0, T]$. We have the following existence result.

Theorem 2.3. *Under assumption* (2.10)*, for every $f \in \mathcal{Y}^*$, $u_0 = 0$, the degenerate initial boundary value problem* (2.7) *admits a unique solution $u \in \mathcal{X}_{(0}$ and there holds the a priori error estimate*

$$\|u\|_{\mathcal{X}} \le C\, \|f\|_{\mathcal{Y}^*} \, .$$

Here, the constant $C > 0$ depends only on c_1 and c_2 in (2.10)*.*

3 DG semidiscretization in time

3.1 Model problem

We now address the time discretization of the abstract diffusion problem (2.7), i.e., find a sufficiently smooth $u(t,x)$ such that

$$\begin{aligned} \partial_t u + a(t)\,\mathcal{B}\,u &= g \text{ on } J \times D, \\ u(0) &= u_0, \end{aligned} \tag{3.1}$$

where $\mathcal{B}$ is the second order, elliptic self-adjoint operator with $\mathcal{B} : X \to X^*$ in (2.7), where $\gamma \in (-1,1)$. We have the following properties of $\mathbf{b}(u,v) = \langle \mathcal{B}\,u, v\rangle$: there exist $\alpha, \beta > 0$ such that for all $u, v \in X$ holds

$$|\mathbf{b}(u,v)| \leq \alpha \|u\|_X \|v\|_X \tag{3.2}$$

$$\mathbf{b}(u,u) \geq \beta \|u\|_X^2 \ . \tag{3.3}$$

Remark 3.1. *Note that the consideration of a localized equation can be justified in the case of a pure diffusion equation by the exponential decay of the truncation error with increasing computational domain. For the case of an FBM market model localization estimates have been derived in [1]. Throughout this section we use a generic positive constant C taking different values in different places, it is independent of the polynomial degree p, time t, the derivative order l and the interval length k.*

3.2 Formulation of the DG time semidiscretization

Definition 3.2. *Let $I = (0,1)$. For a function $u \in L^2(I;X)$ which is continuous at $t = 1$ we define $\Pi^p u \in \mathcal{P}^p(I,X)$, $r \geq 1$, via the conditions*

$$\int_I (\Pi^p u - u, q)_H \,\mathrm{d}t = 0, \quad \forall q \in \mathcal{P}^{p-1}(I;X) \tag{3.4}$$

and

$$\Pi^p u(+1) = u(+1) \in X. \tag{3.5}$$

For $p = 0$ we use only (3.5) to define Π^p, $H = L^2(D)$.

We consider the following DG-formulation:

Definition 3.3. *Let $\mathcal{M} = \{I_m\}_{m=1}^M$, $M \in \mathbb{N}$ be a partition of $J = (0,T)$, $\underline{p} \in \mathbb{N}_0^M$, then the DGFEM for (2.7) reads as follows: find $U \in \mathcal{V}^{\underline{p}}(\mathcal{M};X) := \{u : J \to X : u|_{I_m} \in \mathcal{P}^{p_m}(I_m,X), 1 \leq m \leq M\}$ such that*

$$B_{DG}(U,V) = F_{DG}(V), \quad \textit{where} \tag{3.6}$$

$$\begin{aligned} B_{DG}(U,V) &= \sum_{m=1}^M \int_{I_m} (U',V)_H \mathrm{d}t + \sum_{m=1}^M \int_{I_m} a(t)\mathbf{b}(U,V)\mathrm{d}t \\ &+ \sum_{m=2}^M ([U]_{m-1}, V_{m-1}^+)_H + (U_0^+, V_0^+)_H \end{aligned} \tag{3.7}$$

$$F_{DG}(V) = (u_0, V_0^+)_H + \sum_{m=1}^M \int_{I_m} (g(t),V)_{X^* \times X} \mathrm{d}t \tag{3.8}$$

for all $V \in \mathcal{V}^{\underline{p}}(\mathcal{M};X)$

3.3 Quasi-optimality

We now prove that, under the regularity assumption (2.11), for appropriate choices of the time step sizes $|I_m|$ and of the time discretization orders $p_m \geq 0$, that the time discretization error converges exponentially, under in a sense minimal regularity of the initial data u_0. Throughout this subsection, we shall assume that the coupled system of spatial, elliptic problems in each time step in the Definition 3.3 are solved exactly.

We start our analysis of the time discretization error with some identities satisfied by the DG solution U, then verify Galerkin orthogonality of it and, finally, construct an interpolant for which we establish exponential convergence. The following identity holds due to [5, Lemma 1.8] for $B_{DG}(\cdot,\cdot)$ in (3.7).

Lemma 3.4. *Let B_{DG} be as in Definition 3.3, then for all $V, W \in \mathcal{V}^{\underline{p}}(\mathcal{M}; X)$*

$$\begin{aligned} B_{DG}(V,W) &= \sum_{m=1}^{M} \int_{I_m} (-V, W')_H + a(t)\mathbf{b}(V,W)\mathrm{d}t \\ &- \sum_{m=1}^{M} (V_m^-, [W]_m)_H + (V_M^-, W_M^-)_H, \\ B_{DG}(V-W, V-W) &= \sum_{m=1}^{M} \int_{I_m} a(t)\mathbf{b}(V-W, V-W)\mathrm{d}t + \frac{1}{2}\left\|(V-W)_0^+\right\|_H^2 \\ &+ \frac{1}{2}\sum_{m=1}^{M-1} \left\|[V-W]_m\right\|_H^2 + \frac{1}{2}\sum_{m=1}^{M-1} \left\|(V-W)_M^-\right\|_H^2 . \end{aligned}$$

Theorem 3.5. *Problem* (3.6) *has a unique solution $U \in \mathcal{V}^{\underline{p}}(\mathcal{M}; X)$. If u is the solution of* (2.7)*, then we have the Galerkin orthogonality*

$$B_{DG}(u-U, V) = 0 \text{ for all } V \in \mathcal{V}^{\underline{p}}(\mathcal{M}; X).$$

Proof. The proof follows as in [5, Proposition 1.7], where the case $\gamma = 0$ was treated. □

This implies

Theorem 3.6. *Let u be the exact solution of* (2.7) *and $U \in \mathcal{V}^{\underline{p}}(\mathcal{M}; X)$ the semidiscrete solution of* (3.6) *in $\mathcal{V}^{\underline{p}}(\mathcal{M}, X)$. Assume moreover that $u \in C([\varepsilon, T], X)$, for arbitrary $\varepsilon > 0$. Let $\mathcal{I}u \in \mathcal{V}^{\underline{p}}(\mathcal{M}, X)$ denote the interpolant of u which is defined on each time interval I_m as $\mathcal{I}u|_{I_m} = \Pi_{I_m}^{p_m}(u|_{I_m})$. Then there holds for some $C > 0$ independent of α, β, c_1, c_2*

$$\|u - U\|_{L^2_{t^{\gamma/2}}(I;X)} \leq C\left(1 + \frac{\alpha c_1}{\beta c_2}\right) \|u - \mathcal{I}u\|_{L^2_{t^{\gamma/2}}(I;X)} . \tag{3.9}$$

Proof. Using Lemma 3.4, the coercivity of $\mathbf{b}(\cdot,\cdot)$, the lower bound (2.10) and the Galerkin orthogonality we obtain

$$\begin{aligned} \beta \int_J t^\gamma \|U - \mathcal{I}u\|_X^2 \,\mathrm{d}t &\leq \frac{1}{c_1^2} B_{DG}(U - \mathcal{I}u, U - \mathcal{I}u) \\ &\leq \frac{1}{c_1^2} \left|B_{DG}(u - \mathcal{I}u, U - \mathcal{I}u)\right| . \end{aligned} \tag{3.10}$$

This implies

$$\int_J t^\gamma \|U - \mathcal{I}u\|_X^2 \,\mathrm{d}t \leq \frac{\alpha^2}{\beta^2 c_1^2} \int_J t^\gamma \|u - \mathcal{I}u\|_X^2 \,\mathrm{d}t \tag{3.11}$$

and therefore the claim follows using triangle inequality. □

Therefore it suffices to estimate the projection error to conclude the a priori error analysis. From [5, Corollary 1.20], we recall the following approximation result

Lemma 3.7. *Let $I=(a,b)$, $k=b-a$, $p\in\mathbb{N}_0$ and $u\in H^{s_0+1}$ for some $s_0\in\mathbb{N}_0$. Then*

$$\|u-\Pi_I^p u\|^2_{L^2(I,X)}\leq \frac{C}{\max\{1,p\}^2}\frac{\Gamma(p+1-s)}{\Gamma(p+1+s)}\left(\frac{k}{2}\right)^{2(s+1)}\|u\|^2_{H^{s+1}(I,X)},$$

for any $0\leq s\leq \min\{p,s_0\}$, s real and $C>0$.

3.4 Time-regularity

For data $u_0\in H$ and $g\in L^2_{t^{-\gamma/2}}(J;H)$ the solution of (2.7) can be represented as follows:

$$u(t)=\sum_{i=1}^{\infty}u_{\lambda_i}(t)(u_0,\varphi_i)_H\varphi_i+\sum_{i=1}^{\infty}\left(\int_0^t u_{\lambda_i}(t)(g(s),\varphi_i)_H ds\right)\varphi_i,$$

where $u_{\lambda_i}(t)$ is given by

$$u_{\lambda_i}(t)=e^{-\lambda_i\int_0^t a(s)ds}$$

and $\{\varphi_i\}_{i\in\mathbb{N}}$ denotes the family of eigenfunctions of $\mathcal{B}$, with $\varphi_i\in X$ for $i\in\mathbb{N}$. We assume that the real eigenvalues $\{\lambda_i\}_{i\in\mathbb{N}}$ are enumerated in non-decreasing order repeated according to multiplicity and that the sequence of eigenfunctions $\{\varphi_i\}_{i\in\mathbb{N}}$ forms an orthonormal basis of H.

Theorem 3.8. *Let the operator $\Theta_\gamma(t)$ for $u\in H$, $t\geq 0$ be given by*

$$\Theta_\gamma(t)u=\sum_{i=1}^{\infty}u_{\lambda_i}(t)(u,\varphi_i)_H\varphi_i.$$

Then for $a(t)=t^\gamma$, $-1<\gamma<1$, the following estimates hold for $\min(T,1)>t>0$ and $C,c>0$, independent of $l\geq 1$, and for all $l\in\mathbb{N}$

$$\left\|\Theta_\gamma^{(l)}(t)\right\|^2_{\mathcal{L}(H,X)} \leq Cc^{2l+1}(2l+2)!t^{-(2l+1)-\gamma(l+1)},\ \text{for } \gamma>0, \tag{3.12}$$

$$\left\|\Theta_\gamma^{(l)}(t)\right\|^2_{\mathcal{L}(H,X)} \leq Cc^{2l+1}(2l+2)!t^{-(2l+1)-\gamma},\ \text{for } \gamma\leq 0, \tag{3.13}$$

$$\left\|\Theta_\gamma^{(l)}(t)\right\|^2_{\mathcal{L}(X,X)} \leq Cc^{2l+1}(2l+2)!t^{-(2l)-\gamma l},\ \text{for } \gamma>0, \tag{3.14}$$

$$\left\|\Theta_\gamma^{(l)}(t)\right\|^2_{\mathcal{L}(X,X)} \leq Cc^{2l+1}(2l+2)!t^{-(2l)},\ \text{for } \gamma\leq 0. \tag{3.15}$$

For $l=0$ we obtain that there exists a constant $C>0$ such that, for $0<t\leq 1$,

$$\|\Theta_\gamma(t)\|^2_{\mathcal{L}(H,X)}\leq Ct^{-\gamma-1},\qquad \|\Theta_\gamma(t)\|^2_{\mathcal{L}(X,X)}\leq C\,.$$

Proof. For any $l\in\mathbb{N}$ and for $0<t\leq 1$, we have for $\gamma=0$ due to [5, Proposition 2.1] for $\Theta^{(l)}:=\Theta_0^{(l)}$

$$\left\|\Theta^{(l)}(t)\right\|^2_{\mathcal{L}(H,X)} \leq Cc_0^{2l+1}(2l+2)!t^{-(2l+1)}\qquad \|\Theta(t)\|^2_{\mathcal{L}(H,X)}\leq Ct^{-1},$$

for some constants $C,c_0>0$. We have the following relation between $\Theta(t)$ and $\Theta_\gamma(t)$

$$\Theta_\gamma(t)=\Theta(\omega(t)),$$

where $\omega(t) : \mathbb{R}_+ \to \mathbb{R}_+$, $\omega(t) = t^{\gamma+1}/(\gamma+1)$. We use the formula of Faa di Bruno, see [6]: with certain constants $c^k_{\alpha_1,\dots,\alpha_k}$, $\alpha_j \in \mathbb{N}_0$, $j,k \in \{1,\dots,l\}$ and estimate, with other constants $c_1, c_2, C_1, C_2 > 0$

$$\begin{aligned}
\left\|\Theta^{(l)}_\gamma(t)\right\|^2_{\mathcal{L}(H,X)} &= \left\|\sum_{k=1}^{l} \sum_{\substack{\alpha_1,\dots,\alpha_k \\ \sum \alpha_i = l}} c^k_{\alpha_1,\dots,\alpha_k} (\Theta^k(\omega(t))) \prod_{i=1}^{k} (\partial^{\alpha_i}\omega(t))\right\|^2_{\mathcal{L}(H,X)} \\
&\le \sum_{k=1}^{l} \sum_{\substack{\alpha_1,\dots,\alpha_k \\ \sum \alpha_i = l}} c^k_{\alpha_1,\dots,\alpha_k} \left(C(c_0(\gamma+1))^{2k+1}(2k+2)! t^{-(2k+1)(\gamma+1)}\right) \prod_{i=1}^{k} C_1 t^{\gamma-\alpha_i+1}\alpha_i! \\
&\le C_2 c_1^{2l+1} \left(\sum_{k=1}^{l} t^{-(\gamma+1)(2k+1)+k\gamma-l+k}\right) l! \sum_{k=1}^{l} \sum_{\substack{\alpha_1,\dots,\alpha_k \\ \sum \alpha_i = l}} c^k_{\alpha_1,\dots,\alpha_k} k! \prod_{i=1}^{k} \alpha_i
\end{aligned}$$

In order to estimate the double sum we use the identity shown in [7, Lemma 1.1.1]

$$\sum_{k=1}^{l} \sum_{\substack{\alpha_1,\dots,\alpha_k \\ \sum \alpha_i = l}} c^k_{\alpha_1,\dots,\alpha_k} k! \prod_{i=1}^{k} \alpha_i = 2^{l-1} l!$$

Therefore we have using the Stirling formula

$$\begin{aligned}
\left\|\Theta^{(l)}_\gamma(t)\right\|^2_{\mathcal{L}(H,X)} &\le C_2 c_2^{2l+1}(2l+2)! t^{-(2l+1)-\gamma(l+1)} \text{ for } \gamma > 0 \\
\left\|\Theta^{(l)}_\gamma(t)\right\|^2_{\mathcal{L}(H,X)} &\le C_2 c_2^{2l+1}(2l+2)! t^{-(2l+1)-\gamma} \text{ for } \gamma \le 0
\end{aligned}$$

The proof for (3.14)-(3.15) follows similarly using

$$\left\|\Theta^{(l)}(t)\right\|^2_{\mathcal{L}(X,X)} \le C c_0^{2l}(2l+2)! t^{-(2l+1)}$$

for some constants $C, c_0 > 0$, see [5, Proposition 2.1]. □

Remark 3.9. *Note that the estimates in the previous theorem coincide with the results of [5, Section 2.1],[8, 9] for $\gamma = 0$. In the following we restrict the analysis to the case of $-1 < \gamma \le 0$, the case of $1 > \gamma > 0$ will be treated elsewhere.*

We split the solution u of (2.7) into its homogeneous and inhomogeneous part, i.e., $u = u_1 + u_2$, where

$$\begin{aligned}
u_1' + t^\gamma \mathcal{B} u_1 &= 0, \quad u_1(0) = u_0, &(3.16)\\
u_2' + t^\gamma \mathcal{B} u_2 &= g, \quad u_2(0) = 0. &(3.17)
\end{aligned}$$

The behavior of both terms will be studied separately. The function $u_1(t)$ can be represented as $u_1(t) = \Theta_\gamma(t)u_0$

Theorem 3.10. *Let $u_0 \in H_\theta$ for $0 \le \theta \le 1$. Let u_1 be the solution of* (3.16)*. Then there holds for $l \ge 1$ and $C, c > 0$*

$$\begin{aligned}
\left\|u_1^{(l)}\right\|^2_X &\le C c^{2l+1} t^{-(2l+1+\gamma)+\theta(1+\gamma)}(2l+2)! \left\|u_0\right\|^2_{H_\theta} \quad \textit{and} \\
\left\|u_1\right\|^2_X &\le C t^{(-\gamma-1)(1-\theta)} \left\|u_0\right\|^2_{H_\theta}.
\end{aligned}$$

Proof. The proof follows from Theorem 3.8 and from $u_1(t) = \Theta_\gamma(t)u_0$. □

The solution u_2 of (3.17) can be represented by the "variation of constants" formula, see [10, Section 4.2, Definition 2.3].

$$u_2(t) = \int_0^t \Theta_\gamma(t-s)g(s)ds, \quad 0 \le t \le T\,.$$

We assume that the right hand side g in (3.17) is an analytic function of t; specifically, we assume g to satisfy with some constant $c > 0$, independent of l and t

$$\left\|g^{(l)}(t)\right\|_H \le l!c^{l+1}, \quad t \in [0,T], l \in \mathbb{N}_0. \tag{3.18}$$

Lemma 3.11. *Under Assumption* (3.18)*, we get for any* $t > 0$

i) $u_2(t) = \int_0^t \Theta_\gamma(s)g(t-s)ds$ *in* H*.*

ii)

$$u^{(l)}(t) = \sum_{i=0}^{l-1} \Theta_\gamma^{(i)}(t)g^{(l-i-1)}(0) + \int_0^t \Theta_\gamma(s)g^{(l)}(t-s)ds$$

for $l \ge 1$ *in* H*.*

Proof. The asserted identities are verified by direct evaluation. □

Lemma 3.12. *Assume* (3.18) *and let* u_2 *solve* (3.17)*. Then there exist constants* $C, c > 0$ *such that for* $\min(T,1) > t > 0$ *and for all* $l \ge 0$

$$\left\|u_2^{(l)}(t)\right\|_X \le Cc^l l!\left(t^{1/2-\gamma/2} + t^{-\frac{1}{2}-\frac{\gamma}{2}}\sum_{i=0}^{l-1} t^{-i}\right).$$

Proof.

$$\begin{aligned}\left\|u_2^{(l)}(t)\right\|_X &\le \sum_{i=0}^{l-1}\left\|\Theta_\gamma^{(i)}(t)\right\|_{\mathcal{L}(H,X)}\left\|g^{(l-1-i)}(0)\right\|_H \\ &\quad + \int_0^t \|\Theta_\gamma(s)\|_{\mathcal{L}(H,X)}\left\|g^{(l)}(t-s)\right\|_H ds := S_1 + S_2.\end{aligned}$$

We first bound S_1. We conclude from Theorem 3.10 and Stirling's formula that there exists $C, c_1, c_2 > 0$ such that for all $l \le 0$ and $\min(T,1) > t > 0$

$$\left\|\Theta_\gamma^{(l)}(t)\right\|_{\mathcal{L}(H,X)} \le Cc_1^{l-1}l!t^{(-l-1/2-\gamma/2)}. \tag{3.19}$$

Using (3.18) and (3.19) we conclude

$$\begin{aligned}S_1 &\le Cc_2^{l-1}\sum_{i=0}^{l-1} i!(l-i-1)!t^{-i-1/2-\gamma/2} \\ &\le Cc_2^{l-1}(l-1)!\sum_{i=0}^{l-1}\binom{l-1}{i}^{-1} t^{-i-1/2-\gamma/2} \le Cc_2^{l-1}(l-1)!\sum_{i=0}^{l-1} t^{-i-1/2-\gamma/2}.\end{aligned}$$

The bound on S_2 follows similarly:

$$S_2 \le Cc^l l!\int_0^t s^{-1/2-\gamma/2} = Cc^l l! t^{1/2-\gamma/2}, \quad 0 < t < \min(T,1),\ l \in \mathbb{N}_0.$$

□

Theorem 3.13. *Assume* (3.18) *and let* u_2 *solve* (3.17). *Then there exist constants* $C, c > 0$ *such that for* $\min(T,1) > t > 0$ *and for all* $l \in \mathbb{N}_0$

$$\left\| u_2^{(l)}(t) \right\|_X^2 \leq C c^{2l} (2l+1)! t^{-2l+1-\gamma}.$$

Proof. From Lemma 3.12 we have for any $l \geq 0$

$$\left\| u_2^{(l)}(t) \right\|_X \leq C c_1^l l! (l+1) t^{-l+1/2-\gamma/2},$$

for some $c_1 > 0$ independent of l. The claim follows using the properties of the Gamma function, i.e.,

$$\Gamma(c+1)^2 \leq \Gamma(c+1)\Gamma(c+3/2) = C_1 \Gamma(2c+2) 2^{-2(c+1)}, \quad \text{for } c \in \mathbb{R} \text{ and } C_1 > 0.$$

□

Theorem 3.14. *Let* u_0 *be in* H_θ, $I = (0,k)$ *and let* g *satisfy* (3.18), *then*

$$\int_0^k \|u(t) - u(k)\|_X^2 \, t^\gamma \mathrm{d}t \leq C k^{(\gamma+1)\theta}$$

for some $C > 0$.

Proof.

$$\begin{aligned} \|u(t) - u(k)\|^2_{L^2_{t^{\gamma/2}}(I;X)} &\leq C \|u_1(t)\|^2_{L^2_{t^{\gamma/2}}(I;X)} + C \|u_1(k)\|^2_{L^2_{t^{\gamma/2}}(I;X)} \\ &\quad + C \|u_2(t)\|^2_{L^2_{t^{\gamma/2}}(I;X)} + C \|u_2(k)\|^2_{L^2_{t^{\gamma/2}}(I;X)} \\ &:= T_1 + T_2 + T_3 + T_4 \, . \end{aligned}$$

We first bound T_1 using Theorem 3.10: $T_1 \leq t^{\gamma+1} \|u_0\|_X^2$ and $T_1 \leq C \|u_0\|_H^2$. Therefore, by interpolation, we have $T_1 \leq C k^{(\gamma+1)\theta} \|u_0\|^2_{H_\theta}$, $0 < \theta < 1$. We can also bound T_2 with Theorem 3.10, which gives $T_2 \leq C k^{(1+\gamma)\theta} \|u_0\|^2_{H_\theta}$. T_3 and T_4 can be bounded by Theorem 3.13: $\max\{T_3, T_4\} \leq C k^2$. □

Theorem 3.15. *Let* $u_0 \in H_\theta$ *for* $0 < \theta \leq 1$ *and let* $g \in L^2(J,H)$ *satisfy* (3.18), *then there exist constants* $C, c > 0$ *such that* u *satisfies for* $0 < a \leq b < \min(1,T)$ *and for every* $l \geq 0$, $\min(T,1) > t > 0$

$$\left\| u^{(l)}(t) \right\|_X^2 \leq C c^{2l} (2l)! t^{-(2l+1)+\theta+\gamma(\theta-1)}, \tag{3.20}$$

$$\int_a^b \left\| u^{(l)}(t) \right\|_X^2 t^\gamma \mathrm{d}t \leq C c^{2l} (2l)! a^{-2l+\theta+\theta\gamma}. \tag{3.21}$$

Proof. We split the solution u into u_1 and u_2. Estimate (3.20) follows directly from Theorem 3.10 and 3.13. Estimate (3.21) can be obtained from (3.20):

$$\int_a^b \left\| u^{(l)}(t) \right\|_X^2 t^\gamma \mathrm{d}t \leq C c^{2l} (2l)! \int_a^\infty t^{-(2l+1)+\theta} \mathrm{d}t \leq C c^{2l} (2l)! a^{-2l+\theta} \, .$$

□

Lemma 3.16. *Let* $u_0 \in H_\theta$ *for* $0 \leq \theta \leq 1$ *and let* $g \in L^2(J,H)$ *satisfy* (3.18), *then there exist* $C, c > 0$ *such that* u *satisfies for* $0 < a \leq b < \min(1,T)$ *and every* $s \geq 0$

$$\|u(t)\|^2_{H^s_{t^{\gamma/2}}((a,b),X)} \leq C c^{2s} \Gamma(2s+3) a^{-2s+\theta(1+\gamma)} \, .$$

Proof. The result follows by interpolation of the statement in Theorem 3.15. □

3.5 *hp*-DG timestepping

For exponential convergence, we require geometric time partitions and so called linear degree vectors.

Definition 3.17. *A geometric time partition $\mathcal{M}_{M,\sigma} = \{I_m\}_{i=1}^M$ with grading factor $\sigma \in (0,1)$ and M time steps I_m, $m = 1, 2, ..., M$ is given by the nodes*

$$t_0 = 0, \quad t_m = T\sigma^{M-m}, \quad 1 \le m \le M.$$

The time steps $k_m = t_m - t_{m-1}$ satisfy

$$k_m = \lambda t_{m-1}, \quad \lambda = \frac{1-\sigma}{\sigma},$$

for $2 \le m \le M$.

Definition 3.18. *A polynomial degree vector $\underline{p} = \{p_m\}_{m=1}^M$ is called* linear with slope $\nu \ge 0$ *on the geometric partition $\mathcal{M}_{M,\sigma}$ of the time interval $(0,T)$ if $r_1 = 0$ and $p_m = \lfloor \nu m \rfloor$ for $2 \le m \le M$.*

Lemma 3.19. *Fix an interval $I_m \in \mathcal{M}_{M,\sigma}$, for $2 \le m \le M$ and set $s_m = \alpha_m p_m$ with $\alpha_m \in (0,1)$. Then for every $\gamma \in (-1,0)$, $\theta \in (0,1]$ there exist constants $C, c > 0$ such that for all M, $m = 2, \dots, M$*

$$\left\| u - \Pi_{I_m}^{p_m} u \right\|^2_{L^2_{t^{\gamma/2}}(I_m,X)} \le C\sigma^{(M-m+1)\theta(1+\gamma)+\gamma} \left((\mu c)^{2\alpha_m} \frac{(1-\alpha_m)^{1-\alpha_m}}{(1+\alpha_m)^{1+\alpha_m}} \right)^{p_m},$$

where $\mu = \max\{1, \lambda\}$ and $\lambda = \frac{1-\sigma}{\sigma}$. The constants $C, c > 0$ only depend on $u_0 \in H_\theta$ and γ, $g \in L^2_{t^{-\gamma/2}}(I;H)$ satisfying (3.18).

Proof. We omit for simplicity the dependence of I, p, α, k and s on m in the following and set $t = t_{m-1}$.

$$\begin{aligned}
& \left\| u - \Pi_l^p u \right\|^2_{L^2_{t^{\gamma/2}}(I;X)} \le \max\left(a^\gamma, b^\gamma\right) \left\| u - \Pi_l^p u \right\|^2_{L^2(I;X)} \\
\le\ & Ca^\gamma \frac{\Gamma(p+1-s)}{p^2\Gamma(p+1+s)} \left(\frac{k}{2}\right)^{2(s+1)} \|u\|^2_{H^{s+1}(I;X)} \\
\le\ & C\left(\frac{a}{b}\right)^\gamma \frac{\Gamma(p+1-s)}{p^2\Gamma(p+1+s)} \left(\frac{k}{2}\right)^{2(s+1)} \|u\|^2_{H^{s+1}_{t^{\gamma/2}}(I;X)} \\
\le\ & C\left(\frac{a}{b}\right)^\gamma \frac{\Gamma(p+1-s)}{p^2\Gamma(p+1+s)} \left(\frac{\mu c}{2}\right)^{2s+2} \Gamma(2s+5) t^{2(s+1)} t^{-2s-2+\theta(1+\gamma)} \\
\le\ & C\left(\frac{\mu c}{2}\right)^{2s} \sigma^\gamma \frac{\Gamma(p+1-s)}{p^2\Gamma(p+1+s)} \Gamma(2s+1) t^{\theta(1+\gamma)}.
\end{aligned}$$

Using the Stirling formula we obtain

$$\frac{\Gamma(p+1-s)}{\Gamma(p+1+s)}\Gamma(2s+1) \le p^{1/2} 2^{2s} \left(\frac{(1-\alpha)^{1-\alpha}}{(1+\alpha)^{1+\alpha}} \right)^p.$$

□

Theorem 3.20. *Consider the parabolic problem* (2.7) *on the time interval $J = (0,1)$ with initial data $u_0 \in H_\theta$ for some $\theta \in (0,1]$ with $\gamma \in (-1,0]$ and with right hand side g satisfying* (3.18). *The weak formulation is discretized in time using the DGFEM as given in Definition 3.3 on a geometric partition $\mathcal{M}_{M,\sigma}$ of the time interval J. Then there exists $\nu_0 > 0$ such that for all linear polynomial degree vectors $\underline{p} = \{p_m\}_{m=1}^M$ with slope $\nu \ge \nu_0$ the semidiscrete DGFEM solution $U \in \mathcal{V}^{\underline{p}}(\mathcal{M}_{M,\sigma}, X)$ satisfies the error estimate*

$$\|u - U\|_{L^2_{t^{\gamma/2}}(J;X)} \le C \exp\left(-bN^{\frac{1}{2}}\right),$$

for some constants $b, C > 0$.

Proof. We present the details only in the analytic case, i.e. $\delta = 1$ in (2.11). A scaling argument allows to consider wlog. the case $T = 1$. Let

$$\nu > \max\left\{1, \frac{\theta \ln(\sigma)}{\ln(f_{\min})}\right\}, \tag{3.22}$$

where $f_{\min}$ will be defined below. Set $r_1 = 0$ and $p_m = \lfloor \nu m \rfloor \geq 1$ for $2 \leq m \leq M$. As before $s_m = \alpha_m p_m$, for $\alpha_m \in (0,1)$ to be selected. We start from (3.9) and use Theorem 3.14 to estimate the error on the first interval I_1 near the origin and Lemma 3.19 to estimate the error on $I_2, \ldots, I_M$. This yields

$$\begin{aligned}
\|u - U\|^2_{L^2_{t^{\gamma/2}}(J;X)} &\leq C\sigma^{(M-1)\theta(1+\gamma)} + C\sum_{m=2}^{M} \sigma^{(M-m+1)\theta(1+\gamma)+\gamma} f_{\mu,c}(\alpha_m)^{p_m} \\
&\leq C\sigma^{(M-1)\theta(1+\gamma)}\left(1 + \sigma^{\gamma}\sum_{m=2}^{M} \sigma^{(2-m)\theta(1+\theta)} \left(f_{\mu,c}(\alpha_m)\right)^{p_m}\right),
\end{aligned}$$

where $f_{\mu,c}(\alpha) = (\mu c)^{2\alpha}\left(\frac{(1-\alpha)^{1-\alpha}}{(1+\alpha)^{1+\alpha}}\right)$. The function $f_{\mu,c}$ satisfies

$$0 < \inf_{0<\alpha<1} f_{\mu,c}(\alpha) = f_{\mu,c}(\alpha_{\min}) < 1 \text{ with } \alpha_{\min} = \frac{1}{\sqrt{1+\mu^2 d^2}}.$$

Set $f_{\min} = f_{\mu,c}(\alpha_{\min})$ and select $\alpha_m = \alpha_{\min}$ for $2 \leq m \leq M$. Hence for every $\gamma \in (-1, 0]$, $\sigma \in (0,1)$ there exists a constant $C > 0$ such that for all $M \in \mathbb{N}$

$$\|u - U\|^2_{L^2_{t^{\gamma/2}}(J;X)} \leq C\sigma^{(M-1)\theta(1+\gamma)}\left(1 + \sigma^{\gamma}\sum_{m=2}^{M} \sigma^{(2-m)\theta(1+\gamma)} f_{\min}^{p_m}\right). \tag{3.23}$$

Since

$$\sigma^{(2-m)\theta(1+\gamma)} f_{\min}^{p_m} \leq C\sigma^{2\theta(1+\gamma)}\left(\frac{f_{\min}^{\nu}}{\sigma^{\theta(1+\gamma)}}\right)^m$$

and $f_{\min}^{\nu} < \sigma^{\theta}$ by (3.22), we conclude that the sum in (3.23) can be bounded by

$$\sum_{m=2}^{M} \sigma^{(2-m)\theta(1+\gamma)} f_{\min}^{p_m} \leq C\sigma^{2\theta(1+\gamma)} \sum_{m=2}^{M} q^m$$

with $q = f_{\min}^{\nu}/\sigma^{\theta(1+\gamma)} < 1$. Therefore $\sum_{m=2}^{\infty} q^m < \infty$ holds and we conclude

$$\|u - U\|^2_{L^2_{t^{\gamma/2}}(J;X)} \leq C_\sigma \sigma^{(M-1)\theta(1+\gamma)} \quad \text{for} \quad -1 < \gamma \leq 0.$$

Taking into account $N = \operatorname{nrdof}(\mathcal{V}^{\underline{p}}(\mathcal{M}_{M,\sigma}, X)) \leq \mathcal{O}(M^2)$, as $M \to \infty$ with $\mathcal{O}$ dependent on ν, concludes the proof. □

Remark 3.21. *The estimates in the case $\delta > 1$ are analogous with slightly modified coefficients, resulting in the rate* $\exp\left(-bN^{1/(1+\delta)}\right)$.

3.6 Time-semidiscrete problem

The discontinuous Galerkin semidiscretization (in time), which is given in Definition 3.3, reduces the initial parabolic problem (2.7) to the successive solution of M systems of linear elliptic equations, where the m-th system consists of $p_m + 1$ coupled equations posed on the corresponding time interval I_m of length $k_m = t_m - t_{m-1}$ for $m = 1, \ldots, M$. We shall now derive these systems.

For $1 \le m \le M$ let us consider the space $\mathcal{Q}_m = \mathcal{P}_{p_m}(I_m, X)$ of polynomials of degree $p_m \ge 0$ with coefficients in X, equipped with the norm of $L^2(t_{m-1}, t_m) \otimes X$. At the current time step we seek $\widehat{U}_m = \widehat{U}\Big|_{I_m} \in \mathcal{Q}_m$ which satisfies

$$\Big\langle \widehat{U}_m(t_{m-1}), \widehat{W}(t_{m-1}) \Big\rangle + \int_{t_{m-1}}^{t_m} \Big[\Big\langle \widehat{U}_m'(t), \widehat{W}(t) \Big\rangle + a(t)\, \mathbf{b}\left(\widehat{U}_m(t), \widehat{W}(t)\right) \Big] \,\mathrm{d}t$$
$$= \Big\langle \widehat{U}_{m-1}(t_{m-1}), \widehat{W}(t_{m-1}) \Big\rangle + \int_{t_{m-1}}^{t_m} \Big\langle g(t), \widehat{W}(t) \Big\rangle_{X^* \times X} \,\mathrm{d}t \quad \text{for all} \quad \widehat{W} \in \mathcal{Q}_m, \quad (3.24)$$

where $\widehat{U}_0(t_0)$ stands for the initial value u_0.

Let $\{\phi_j\}_{j=0}^{p_m}$ be a basis of the polynomial space $\mathcal{P}_{p_m}(-1,1)$. Then the time shape functions on the time interval I_m are $\phi_j \circ F_m^{-1}$, $0 \le j \le p_m$, where the affine mapping $F_m \colon (-1,1) \to I_m$ is given by

$$t = F_m(\tau) = \frac{1}{2}(t_{m-1} + t_m) + \frac{1}{2} k_m \tau, \qquad k_m = t_m - t_{m-1}, \qquad \tau \in (-1,1).$$

By rewriting (3.24) in terms of the temporal basis functions for $\widehat{U}_m(t) = \sum_{j=0}^{p_m} \widehat{U}_{m,j} \cdot (\phi_j \circ F_m^{-1})(t)$, where $\widehat{U}_{m,j} \in X$ for $0 \le j \le p_m$, we obtain

$$\sum_{j=0}^{p_m} \left(C_{mij} \cdot \Big\langle \widehat{U}_{m,j}, W \Big\rangle + \frac{k_m}{2}\, G_{mij} \cdot \mathbf{b}\left(\widehat{U}_{m,j}, W\right) \right) = \widehat{g}_{m,i}(W) + \frac{k_m}{2}\, \tilde{g}_{m,i}(W), \quad (3.25)$$

which holds true for all $W \in X$ and $0 \le i \le p_m$. The matrices and right-hand side vectors in (3.25) are given by

$$C_{mij} = \int_{-1}^{1} \phi_j'(\tau)\, \phi_i(\tau)\,\mathrm{d}\tau + \phi_j(-1)\, \phi_i(-1), \quad G_{mij} = \int_{-1}^{1} (a \circ F_m)(\tau)\, \phi_j(\tau)\, \phi_i(\tau) \mathrm{d}\tau.$$
$$\widehat{g}_{m,i}(W) = \phi_i(-1) \Big\langle \widehat{U}_{m-1}(t_{m-1}), W \Big\rangle, \; \tilde{g}_{m,i}(W) = \int_{-1}^{1} \langle (g \circ F_m)(\tau), W \rangle_{X^* \times X}\, \phi_i(\tau)\,\mathrm{d}\tau,$$

Equation (3.25) is a system of $p_m + 1$ coupled linear second order elliptic equations in the weak form, which needs to be solved at the m-th time step. We may rewrite it as

$$\mathbf{C}_m\, \widehat{\mathbf{u}}_m + \frac{k_m}{2}\, \mathbf{G}_m\, \widehat{\mathbf{v}}_m = \mathbf{g}_m \quad \text{for all} \quad W \in X, \quad (3.26)$$

where, for given $W \in X$ and for $0 \le i, j \le p_m$,

$$(\widehat{\mathbf{u}}_m)_{\,j} = \Big\langle \widehat{U}_{m,j}, W \Big\rangle, \quad (\widehat{\mathbf{v}}_m)_{\,j} = \mathbf{b}\left(\widehat{U}_{m,j}, W\right), \quad (\mathbf{g}_m)_{\,i} = \widehat{g}_{m,i}(W) + \frac{k_m}{2}\, \tilde{g}_{m,i}(W).$$

Remark 3.22. *Let us use the temporal shape functions $\phi_i = (i+1/2)^{1/2} L_i$ with L_i denoting the i-th Legendre polynomial on $(-1,1)$, normalized so that $L_i(1) = 1$.*

Then the elements of the matrix $\mathbf{C}_m$ in (3.26) *are*

$$C_{mij} = \varsigma_{ij} \sqrt{i + \tfrac{1}{2}} \sqrt{j + \tfrac{1}{2}}, \quad i,j = 0, \dots, p_m, \quad \textit{where} \quad \varsigma_{ij} = \begin{cases} (-1)^{i+j} & \textit{if } j > i, \\ 1 & \textit{otherwise.} \end{cases}$$

Also, if $a(t) \equiv 1$, then $\mathbf{G}_m = \mathbb{I}_{p_m+1}$ in (3.26).

From now on we use the normalized Legendre polynomials $\phi_i = (i+1/2)^{1/2} L_i$, $i \ge 0$, as temporal shape functions.

Let us assume that the right-hand side is time-space separable of rank R, i .e. it holds in $L^2_{t^{-\gamma/2}}(J;X)$ that

$$g(t) = \sum_{\alpha=1}^{R} g^{[\alpha]}(t) \cdot f^{[\alpha]}, \quad t \in J, \tag{3.27}$$

for some $g^{[\alpha]} \in L^2_{t^{-\gamma/2}}(J)$ and for $f^{[\alpha]} \in X^*$, $1 \le \alpha \le R$. Then the right-hand side of (3.26) has the following structure:

$$\mathbf{g}_m = \mathbf{g}_m^{[0]} \cdot f_m^{[0]}(W) + \frac{k_m}{2} \sum_{\alpha=1}^{R} \mathbf{g}_m^{[\alpha]} \cdot f^{[\alpha]}(W), \tag{3.28}$$

where

$$\left(\mathbf{g}_m^{[0]}\right)_i = \phi_i(-1), \quad \left(\mathbf{g}_m^{[\alpha]}\right)_i = \int_{-1}^{1} (g^{[\alpha]} \circ F_m)(\tau)\phi_i(\tau)\,\mathrm{d}\tau \tag{3.29}$$

for $0 \le i \le p_m$ and

$$f_m^{[0]}(W) = \left\langle \widehat{U}_{m-1}, W \right\rangle, \quad f^{[\alpha]}(W) = \left\langle f^{[\alpha]}, W \right\rangle_{X^* \times X}.$$

4 Tensor-structured spatial discretization

4.1 Derivation of the linear system for the tensor-product FEM

In this section we describe the spatial discretization of the linear system (3.26) with the use of tensor-product finite elements constructed in a tensor-product domain $D \subset \mathbb{R}^d$. For the sake of simplicity we assume $D = (0,1)^d$ and $X = H_0^1(D)$, although this is not essential for the following.

Assume that we use some finite elements $\psi_{i_k}^{(k)}$, $i_k = 1, \dots, n_k$ in the k-th dimension, $1 \le k \le d$, and define finite elements in D as $\psi_{i_1,\dots,i_d} = \psi_{i_1}^{(1)} \otimes \dots \otimes \psi_{i_d}^{(d)}$. Then we seek the discrete solution to (3.26) in the form

$$\widehat{U}_{m,j} = \sum_{j_1=1}^{n_1} \dots \sum_{j_d=1}^{n_d} \mathbf{u}_{m,j,j_1,\dots,j_d} \psi_{j_1,\dots,j_d}, \quad j = 0, \dots, p_m$$

and, by testing the variational problem with $W = \psi_{i_1,\dots,i_d}$, $1 \le i_k \le n_k$ for $k = 1, \dots, d$, we obtain the linear system

$$\left(\mathbf{C}_m \otimes \mathbf{M} + \frac{k_m}{2}\,\mathbf{G}_m \otimes \mathbf{S}\right) \mathbf{u}_m = \mathbf{g}_m \tag{4.1}$$

for the m-th time step, where the right-hand side, under the assumption of (3.27), has the following structure:

$$\mathbf{g}_m = \mathbf{g}_m^{[0]} \otimes \mathbf{f}_m^{[0]} + \frac{k_m}{2} \sum_{\alpha=1}^{R} \mathbf{g}_m^{[\alpha]} \otimes \mathbf{f}^{[\alpha]} \tag{4.2}$$

with $\mathbf{g}_m^{[0]}$ and $\mathbf{g}_m^{[\alpha]}$ defined in (3.29), $\mathbf{f}_{m\,i_1,\dots,i_d}^{[0]} = \left\langle \widehat{U}_{m-1}, \psi_{i_1,\dots,i_d} \right\rangle$ and $\mathbf{f}^{[\alpha]}{}_{i_1,\dots,i_d} = \left\langle f^{[\alpha]}, \psi_{i_1,\dots,i_d} \right\rangle$ for $\alpha = 1, \dots, R$. The mass matrix $\mathbf{M}$ is

$$\mathbf{M} = \mathbf{M}_1 \otimes \dots \otimes \mathbf{M}_d, \tag{4.3}$$

where

$$(\mathbf{M}_k)_{i_k j_k} = \int_0^1 \psi_{i_k}^{(k)}(x_k) \cdot \psi_{j_k}^{(k)}(x_k)\,\mathrm{d}x_k, \quad 1 \le i_k, j_k \le n_k, \tag{4.4}$$

for $1 \le k \le d$. The stiffness matrix $\mathbf{S}$ reads

$$\mathbf{S}_{\substack{i_1,\dots,i_d \\ j_1,\dots,j_d}} = \mathbf{b}\left(\psi_{i_1,\dots,i_d}, \psi_{j_1,\dots,j_d}\right) = \sum_{p,q=1}^{d} \int_{[0,1]^d} \kappa^{pq}(x_1,\dots,x_d) \cdot \left(\frac{\partial}{\partial x_p}\psi_{i_1}^{(1)}(x_1) \cdot \dots \cdot \psi_{i_d}^{(d)}(x_d)\right)\left(\frac{\partial}{\partial x_q}\psi_{j_1}^{(1)}(x_1) \cdot \dots \cdot \psi_{j_d}^{(d)}(x_d)\right)\mathrm{d}x_1 \dots \mathrm{d}x_d \tag{4.5}$$

for $1 \le i_k, j_k \le n_k$.

The system (4.1) is of order $(p_m + 1) \times n_1 \times \ldots \times n_d$ and, in the usual elementwise representation, bears the "curse of dimensionality" [11] and, even in case d is moderate, becomes numerically intractable quickly as we increase n_k, $k = 1, \ldots, d$, which are the numbers of degrees of freedom in each dimension. To avoid this, we employ the low-parametric tensor representation of the vectors and matrices involved, based on the approximate separation of variables.

4.2 The TT and QTT representations

To represent vectors and matrices involved in the solution of the fully discrete system (4.1), we exploit the *Tensor Train* (TT) format [12, 13, 14, 15] by Oseledets and Tyrtyshnikov. For a d-dimensional $n_1 \times \ldots \times n_d$-vector $\mathbf{u}$ it reads

$$\mathbf{u}_{j_1,\ldots,j_d} = \sum_{\alpha_1=1}^{r_1} \ldots \sum_{\alpha_{d-1}=1}^{r_{d-1}} U_1(j_1, \alpha_1) \cdot U_2(\alpha_1, j_2, \alpha_2) \cdot \ldots \cdot U_{d-1}(\alpha_{d-2}, j_{d-1}, \alpha_{d-1}) \cdot U_d(\alpha_{d-1}, j_d), \tag{4.6}$$

where $1 \le j_k \le n_k$, $1 \le k \le d$. The two- and three-way arrays U_k are referred to as *cores* of the decomposition (4.6); the summation indices and summation limits on the right-hand side are called, respectively, *rank indices* and *TT ranks* of the decomposition. For every $k = 1, \ldots, d-1$ the decomposition (4.6) implies a rank-r_k representation of the unfolding matrix $\mathbf{U}^{(k)}$ defined as follows:

$$\mathbf{U}^{(k)}{}_{i_1 \ldots i_k;\, i_{k+1} \ldots i_d} = \mathbf{u}_{i_1 \ldots i_d}.$$

This relation between the TT and matrix ranks renders the TT format well-defined and allows the robust quasi-optimal rank truncation; see, for example, [16, Theorem 2.1] or [15]. The TT format is also applied to a d-level $(m_1 \times \ldots \times m_d) \times (n_1 \times \ldots \times n_d)$-matrix $\mathbf{A}$ after its vectorization and a permutation of its indices, which results in a representation of the form

$$\mathbf{A}_{\substack{i_1,\ldots,i_d \\ j_1,\ldots,j_d}} = \sum_{\alpha_1=1}^{r_1} \ldots \sum_{\alpha_{d-1}=1}^{r_{d-1}} A_1(i_1, j_1, \alpha_1) \cdot A_2(\alpha_1, i_2, j_2, \alpha_2) \cdot \ldots \cdot A_{d-1}(\alpha_{d-2}, i_{d-1}, j_{d-1}, \alpha_{d-1}) \cdot A_d(\alpha_{d-1}, i_d, j_d), \tag{4.7}$$

involving three- and four-dimensional arrays A_k as TT cores.

The way the low-rank representation of a vector (matrix) in the TT format is related to the low-rank representation of its unfolding matrices allows efficient and robust computations in this format, which can be based on standard matrix algorithms (e. g. SVD coupled with QR).

The complexity of a TT representation is formally linear w. r. t. d, but also depends drastically on the TT ranks of the decompositions in question. For instance, storage costs and the complexity of basic TT arithmetics operations applied to the representation (4.6) (such as the dot product, multi-dimensional contraction, matrix-vector multiplication, rank reduction and orthogonalization of a decomposition, etc.) is $d \cdot n \cdot \mathtt{poly}(r_1, \ldots, r_{d-1})$ in case $n_1 = \ldots = n_d = n$. In general, the TT ranks $r_1, \ldots, r_{d-1}$ in (4.6) may depend on d and n. However, in various applications the TT ranks of vectors and matrices prove to be feasibly low and the TT format turns out to be an advantageous framework for the efficient solution of high-dimensional problems.

With the aim of further reduction of the complexity, the TT format can be applied to a "quantized" vector (matrix), which leads to the *Quantized Tensor Train* (QTT) format [17, 18, 19]. The idea of quantization consists in introducing l_k "virtual" dimensions (levels) corresponding to the k-th original "physical" dimension [20], provided that the corresponding mode size n_k can be factorized as $n_k = \prod_{\kappa_k=1}^{k} n_{k\kappa_k} = n_{k1} \cdot n_{k2} \cdot \ldots \cdot n_{kl_k}$ in terms of integral factors $n_{k\kappa_k} \ge 2$, $\kappa_k = 1, \ldots, l_k$, for $1 \le k \le d$. This corresponds to reshaping the k-th mode of size n_k into l_k modes of sizes $n_{k1}, \ldots, n_{kl_k}$.

A TT decomposition of a vector (matrix) under such a transformation is referred to as a *QTT decomposition* of the vector (matrix). The TT ranks of this decomposition are called *QTT ranks*. In this sense (4.6) and (4.7),

with d being replaced with l, also present QTT representations of ranks $r_1, \ldots, r_{l-1}$ of a one-dimensional vector $\widehat{\mathbf{u}}$ and of a one-dimensional matrix $\widehat{\mathbf{A}}$ with entries $\widehat{\mathbf{u}}_{\overline{i_1,\ldots,i_l}} = \mathbf{u}_{i_1,\ldots,i_l}$ and $\widehat{\mathbf{A}}_{\overline{i_1,\ldots,i_l},\overline{j_1,\ldots,j_l}} = \mathbf{A}_{i_1,\ldots,i_l,j_1,\ldots,j_l}$.

The QTT format represents more structure in the data by splitting all the "virtual" dimensions introduced. It involves more rank numbers which can be higher than for the TT decomposition (with no quantization). Typically one tends to introduce as fine (i. e. with small $n_{k\kappa_k}$) quantization as possible and wind up with as many virtual modes as possible. This corresponds to seeking as much low-rank QTT structure in the data as possible. The low-rank QTT structure corresponding to the "finer" quantization allows to further reduce the complexity.

As an example of the finest possible quantization one may consider the representation of every "physical" scalar index $i = \overline{i_1, \ldots, i_l} \equiv 1 + \sum_{k=1}^{l} 2^{l-k} (i_k - 1)$ varying from 1 to 2^l in terms of "virtual" indices $i_1, \ldots, i_l$ taking values 1 and 2. This binary encoding reshapes a one-dimensional 2^l-component vector into a l-dimensional $2 \times \ldots \times 2$-tensor; and a d-dimensional $2^{l_1} \times \ldots \times 2^{l_d}$-tensor into an $l_1 + \ldots + l_d$-dimensional $2 \times \ldots \times 2$-tensor.

Since a TT decomposition of a d-dimensional tensor has $d-1$ ranks, it is convenient to introduce aggregate characteristics, such as the *effective rank* of a TT decomposition: for an $n_1 \times \ldots \times n_d$-tensor given in a TT decomposition of ranks $r_1, \ldots, r_{d-1}$ we define it as the positive root r of the equation

$$n_1 r_1 + \sum_{k=2}^{d-1} r_{k-1} n_k r_k + r_{d-1} n_d = n_1 r + \sum_{k=2}^{d-1} r\, n_k r + r\, n_d \tag{4.8}$$

which equates the memory needed to store the given decomposition (left-hand side) and a decomposition in the same format, i. e. of an $n_1 \times \ldots \times n_d$-tensor, but with equal $d-1$ ranks $r, \ldots, r$ (right-hand side). In this sense, "effective rank" is understood with respect to memory. However, the notion of effective rank allows to evaluate exactly the complexity of some TT-structured operations, such as the matrix-vector multiplication and Hadamard product, and also estimates the complexity of other operations such, e. g. the TT rank truncation. The concept of the effective rank applies in the same way to QTT decompositions. A similar definition of effective rank was used in [18, Section 3.4].

Let us note that instead of the TT representation and its "quantized" counterpart QTT one may also use the Hierarchical Tensor format [21, 22, 23] along with the "tensorization" [24]. Surveys on tensor representations can be found in [25, 26, 27].

4.3 The TT structure of the fully discrete linear system

The matrix of the fully discrete system (4.1) is time-space separable of rank 2:

$$\mathbf{A}_m = \mathbf{C}_m \otimes \mathbf{M} + \frac{k_m}{2}\, \mathbf{G}_m \otimes \mathbf{S}, \tag{4.9}$$

where the temporal factors $\mathbf{C}_m$ and $\mathbf{G}_m$ are of moderate order, namely $p_m + 1$, and the spatial mass and stiffness matrices $\mathbf{M}$ and $\mathbf{S}$ suffer from the curse of dimensionality. The mass matrix $\mathbf{M}$ arises in the rank-1 separable form (4.3), while the structure of the stiffness matrix (4.5) depends on that of the diffusion coefficient $\mathcal{K}$.

For numerical experiments in the current paper we assume that $\mathcal{K} = \mathbb{I}_d$ in D, so that $\mathcal{B} = -\triangle$ and the stiffness matrix $\mathbf{S}$ has the form

$$\mathbf{S} = \mathbf{S}_1 \otimes \mathbf{M}_2 \ldots \otimes \mathbf{M}_d + \ldots + \mathbf{M}_1 \otimes \ldots \otimes \mathbf{M}_{d-1} \otimes \mathbf{S}_d, \tag{4.10}$$

which includes d terms. Such a matrix can be represented in the TT format in terms of one-dimensional matrices $\mathbf{M}_k$ and $\mathbf{S}_k$, $k = 1, \ldots, d$, with ranks $2, \ldots, 2$, see [28, Lemma 5.1].

However, the stiffness matrix (4.5) can be proved to admit low-rank TT representations in more general cases: in [29] we derive the low-rank TT structure of the stiffness matrix $\mathbf{S}$ under certain requirements on the form of $\mathcal{K}$, which are satisfied, e. g., in the multi-dimensional Black-Scholes model in real price variables, see [30, 31], and in the Heston model or the multiscale stochastic volatility model, see [32].

4.4 QTT structure of the fully discrete linear system

For the tensor-structured solution of the system (4.1) we use the following QTT format. In the k-th spatial dimension, where $1 \le k \le d$, we use $n_k = 2^{l_k}$ degrees of freedom. This allows us to introduce l_k virtual levels corresponding to the k-th real dimension, as we describe in Section 4.2. The solution "quantized" in space in such a way turns into a $(p_m+1) \times 2 \times \ldots \times 2$-tensor with $1 + l_1 + \ldots + l_d$ dimensions which we split in the TT format:

$$\mathbf{u}_{j,j_1,\ldots,j_d} = \sum_{\alpha=1}^{r} \sum_{\alpha_1=1}^{r_1} \ldots \sum_{\alpha_{d-1}=1}^{r_{d-1}} U(j,\alpha) \cdot U_1(\alpha, j_1, \alpha_1) \cdot U_2(\alpha_1, j_2, \alpha_2) \cdot \ldots \cdot U_{d-1}(\alpha_{d-2}, j_{d-1}, \alpha_{d-1}) \cdot U_d(\alpha_{d-1}, j_d), \tag{4.11}$$

where

$$U_k\left(\alpha_{k-1}, \overline{j_{k,1}, \ldots, j_{k,l_k}}, \alpha_k\right) = \sum_{\alpha_{k,1}=1}^{r_{k,1}} \ldots \sum_{\alpha_{k,d-1}=1}^{r_{k,d-1}} U_{k,1}(\alpha_{k-1}, j_{k,1}, \alpha_{k,1}) \cdot U_{k,2}(\alpha_{k,1}, j_{k,2}, \alpha_{k,2}) \cdot \ldots \cdot U_{k,d-1}(\alpha_{k,d-2}, j_{k,d-1}, \alpha_{k,d-1}) \cdot U_{k,d}(\alpha_{k,d-1}, j_{k,d}, \alpha_k) \tag{4.12}$$

for $1 \le k \le d$ and the index α_d is void. The QTT ranks in (4.11)–(4.12) are

$$\mathbf{r}, r_{1,1}, \ldots, r_{1,l_1}, \mathbf{r_1}, r_{2,1}, \ldots, r_{2,l_2}, \mathbf{r_2}, \ldots\ldots, \mathbf{r_{d-1}}, r_{d,1}, \ldots, r_{d,l_d},$$

where we emphasize in boldface the TT ranks, i. e. the ranks of the separation of "physical" dimensions. A decomposition of a matrix in the corresponding format is analogous to (4.11)–(4.12), each j-index being accompanied by an i-index, cf. (4.6) and (4.7).

Provided that matrices $\mathbf{M}$ (4.3) and $\mathbf{S}$ (4.5) can be represented with low ranks in the TT format in terms of coordinate-wise matrices $\mathbf{M}_k$ and $\mathbf{S}_k$, $k = 1, \ldots, d$, the representations of them in the QTT format (4.11)–(4.12) relies on the QTT structure of the coordinate-wise matrices. Let us consider a finite element basis in the k-th dimension chosen as the set of piecewise-linear "hat" functions $\psi_{i_k}^{(k)}$, $i_k = 1, \ldots, n_k$, constructed on an equidistant mesh with the nodes $x_{i_k}^{(k)} = \frac{i_k}{n_k+1}$, $i = 0, \ldots, n_k+1$, so that $\psi_{i_k}^{(k)}\left(x_{j_k}^{(k)}\right) = \boldsymbol{\delta}_{i_k j_k}$ for any $i_k = 1, \ldots, n_k$ and $j_k = 0, \ldots, n_k+1$. Then the one-dimensional mass and stiffness matrices involved in (4.3) and (4.10) are the following:

$$\mathbf{M}_k = \frac{h_k}{6} \begin{pmatrix} 4 & 1 & & \\ 1 & \ddots & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & 4 \end{pmatrix}, \quad \mathbf{S}_k = \frac{1}{h_k} \begin{pmatrix} 2 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{pmatrix},$$

where $h_k = \frac{1}{n_k+1}$. Both can be represented in the QTT format with ranks $3, \ldots, 3$ [28, Lemma 3.1]. As a consequence, the multi-dimensional mass and stiffness matrices (4.3) and (4.10) can be represented in the QTT format with ranks $3, \ldots, 3, \mathbf{2}, 6, \ldots, 6, \mathbf{2}, \ldots\ldots, \mathbf{2}, 6, \ldots, 6, \mathbf{2}, 6, \ldots, 6$, see [28, Lemma 5.2 and Remark 5.4]. Therefore we make the following conclusion.

Proposition 4.1. *Assume that $\mathcal{K} = \mathbb{I}_d$ and the finite elements described in the sections 4.1 and 4.4 are used for the spatial discretization. Then the matrix $\mathbf{A}_m$ (4.9) of the time step system (4.1) has a QTT decomposition of ranks*

$$\mathbf{2}, 3, \ldots, 3, \mathbf{2}, 6, \ldots, 6, \mathbf{2}, \ldots\ldots, \mathbf{2}, 6, \ldots, 6, \mathbf{2}, 6, \ldots, 6.$$

5 Numerical experiments

In our numerical experiments we consider the model problem (2.7) in the time-space cylinder $J \times D$, where $J = (0, T]$ and $D = (0,1)^d$ is a hypercube, with $\mathcal{B} = -\triangle$, $\gamma \in (-1, 1)$ and $f = 0$: find $u \in \mathcal{X}$ such that

$$\int_J \left[\langle \dot{u}(t), v(t) \rangle_{L^2(D)} + t^\gamma \langle \nabla u(t), \nabla v(t) \rangle_{L^2(D)} \right] \mathrm{d}t = 0 \quad \text{for all} \quad v \in \mathcal{Y}, \quad u(0) = u_0, \tag{5.1}$$

where and $\mathcal{X} = H^1_{t^{-\gamma/2}}\left(J; H^{-1}(D)\right) \cap L^2_{t^{\gamma/2}}\left(J; H^1_0(D)\right) \simeq H^1_{t^{-\gamma/2}}(J) \otimes H^{-1}(D) \cap L^2_{t^{\gamma/2}}(J) \otimes H^1_0(D) \mathcal{Y} = L^2_{t^{\gamma/2}}\left(J; H^1_0(D)\right) \simeq L^2_{t^{\gamma/2}}(J) \otimes H^1_0(D)$.

For $\alpha \in \mathbb{N}$ let $\sqrt{\lambda_\alpha} = \pi(2\alpha - 1)$ and $\psi_\alpha(x) = \sin\left(\sqrt{\lambda_\alpha}\, x\right)$, $x \in (0,1)$. Then the eigenpairs $\{\Lambda_\alpha, \Psi_\alpha\}_{\alpha \in \mathbb{N}^d}$ of $\mathcal{B}$ are the following: $\Psi_{\alpha_1,\ldots,\alpha_d} = \psi_{\alpha_1} \otimes \ldots \otimes \psi_{\alpha_d}$ and $\Lambda_{\alpha_1,\ldots,\alpha_d} = \lambda_{\alpha_1} + \ldots + \lambda_{\alpha_d}$. If we consider the expansion $u_0 = \sum_{\alpha \in \mathbb{N}^d} C_\alpha \Psi_\alpha$, where $C_\alpha = 2^d \langle u_0, \Psi_\alpha \rangle_{L^2(D)}$, then the function u defined as

$$u(t,\cdot) = \sum_{\alpha \in \mathbb{N}^d} \exp\left(-\Lambda_\alpha \frac{t^{\gamma+1}}{\gamma+1}\right) C_\alpha \Psi_\alpha \quad \text{for} \quad t \in J \tag{5.2}$$

solves (5.1). We consider both compatible and incompatible initial data u_0 below. Here, by compatible initial data we mean data which satisfies the compatibility conditions up to an arbitrary order, see [33, Chapter 4, Section 5, p. 319] for a definition of the compatibility conditions. For the sake of simplicity, we assume u_0 to be symmetric rank-1 separable, i. e. $u_0 = \left(u_0^{(1)}\right)^{\otimes d}$ for some univariate function $u_0^{(1)}$.

We implemented the *hp*-DG approach in time with the QTT-structured discretization in space in MATLAB using the *TT Toolbox* for basic QTT-structured operations and the solution of linear systems (the toolbox is publicly available at `http://spring.inm.ras.ru/osel` and `http://github.com/oseledets/TT-Toolbox`; we used the GitHub version of November 23, 2011). We run the resulting *hp*-DG-QTT solver in MATLAB 7.12.0.635 (R2011a) on a laptop with a 2.7 GHz dual-core processor and 4 GB RAM, and report the computation times in seconds.

We apply the DG time discretization introduced in Section 3 with the following parameters: the number of time intervals M, the grading factor σ of the geometric refinement of the temporal mesh, the slope ν of polynomial degrees. As basis functions for the polynomial spaces of order $p_m + 1$ on the m-th time interval we use Legendre polynomials, where $p_m = \lfloor \nu m \rfloor$ for $j = 1, \ldots, M$. The space discretization is defined by the number l of virtual levels (levels of quantization) in each of the d dimensions. We construct the FEM discretization on an equidistant mesh with $2^l \times \ldots \times 2^l$ nodes, as described in Section 4.

For the solution of linear systems in the QTT format we use the DMRG solver proposed in [34] and available as the function `dmrg_solve2` of the TT Toolbox. The method still lacks a rigorous analysis of its convergence properties, which is currently common for tensor-structured solvers. However, it proves to be highly efficient in many applications including our numerical examples presented below. The crucial parameters of the solver are the desired relative residual `res` of the linear system and the maximum number `swp` of its iterations ("DMRG sweeps"). The iterations continue until either their number reaches `swp` or the relative residual is less than or equal to `res`. In every particular run we use the same values of `res` and `swp` at every time interval to solve the linear system (4.1). As output we obtain the block solution $\mathbf{u}_m$, which is a $(p_m + 1) \times 2^l \times \ldots \times 2^l$-component tensor in the QTT format, see (4.11)–(4.12). We contract the block solution $\mathbf{u}_m$ along the first mode with the values of the Legendre polynomials at 1 to obtain the tensor $\mathbf{u}_m^0$ of interpolation coefficients of the numerical solution u_m^0 evaluated at the time point t_m. Then we truncate $\mathbf{u}_m^0$ in the QTT format with the relative accuracy $\delta = 10^{-15}$ to exclude obviously excessive QTT components. This gives us a tensor of coefficients $\mathbf{u}_m^\delta$ and the corresponding approximation u_m^δ to the solution. We compute

$$\mathtt{err}\left[u_m^\delta\right] = \frac{\|u_m^\delta - \widehat{u}_m\|_{L^2(D)}}{\|\widehat{u}_m\|_{L^2(D)}} = \frac{\|\mathbf{u}_m^\delta - \widehat{\mathbf{u}}_m\|_{\mathbf{M}}}{\|\widehat{\mathbf{u}}_m\|_{\mathbf{M}}}, \tag{5.3}$$

which is the relative L^2-error of the fully discrete numerical solution u_m^δ with respect to the discretized reference solution $\widehat{u}_m$ evaluated at $t = t_m$ (it quantifies the effects of the temporal and spatial discretizations, as well as the QTT-structured representation of the latter). By $\widehat{\mathbf{u}}_m$ we denote the tensor of interpolation coefficients of $\widehat{u}_m$; and by $\mathbf{M}$, the mass matrix from (4.3).

5.1 Compatible initial data in d dimensions

First, we consider $T = 0.005$ and the following initial data which is compatible with the boundary conditions:

$$u_0(x_1, \ldots, x_d) = \prod_{k=1}^{d} \sin \pi x_k \quad \text{for} \quad x_k \in (0,1), \quad 1 \le k \le d, \tag{5.4}$$

	$\gamma = -\frac{1}{2}$ $M = 30$		$\gamma = 0$ $M = 10$		$\gamma = \frac{1}{2}$ $M = 10$	
d	$\texttt{err}\left[u_M^\delta\right]$	time	$\texttt{err}\left[u_M^\delta\right]$	time	$\texttt{err}\left[u_M^\delta\right]$	time
5	$1.1 \cdot 10^{-8}$	12.2	$8.8 \cdot 10^{-10}$	3.9	$1.0 \cdot 10^{-11}$	4.1
10	$3.1 \cdot 10^{-8}$	24.2	$1.4 \cdot 10^{-9}$	7.5	$6.9 \cdot 10^{-11}$	7.5
20	$5.6 \cdot 10^{-8}$	47.4	$2.4 \cdot 10^{-9}$	15.2	$1.7 \cdot 10^{-10}$	14.6
30	$9.0 \cdot 10^{-8}$	71.8	$3.1 \cdot 10^{-9}$	23.1	$1.9 \cdot 10^{-10}$	21.6
40	$1.9 \cdot 10^{-7}$	96.4	$3.7 \cdot 10^{-9}$	31.6	$2.8 \cdot 10^{-10}$	29.3

Table 1: Compatible initial data in d dimensions: errors at $t = T$ and computation times.

so that $u_0 \in C_0^\infty(D)$. We consider the cases $\gamma = 0$, $\gamma = \frac{1}{2}$ and $\gamma = -\frac{1}{2}$. In the latter cases the diffusion operator $-t^\gamma \triangle$ degenerates at $t = 0$. However, the *hp*-DG time discretization allows to resolve such a degeneracy. We regard the case $\gamma = 0$ relatively simple and consider it as a reference for more challenging problems corresponding to $\gamma = \pm\frac{1}{2}$. We use the FEM discretization in space, constructed on equidistant meshes with $l = 14$ levels of quantization (see Section 4), which corresponds to $2^l = 4096$ degrees of freedom in each dimension. As for the time discretization, we use the temporal geometric mesh with M time intervals, which are graded geometrically by a subdivision ratio of $\sigma = 0.5$, and linearly increasing time-stepping orders, where the slope of polynomial degrees equal $\nu = 2$. We choose M so as to ensure the minimal achievable $\texttt{err}\left[u_M^\delta\right]$ which is the error (5.3) at $t = T$.

For the solution of the linear systems (4.1) we use $\texttt{res} = 10^{-8}$ and $\texttt{swp} = 10$. However, in this series of experiments the solution is very simple and has low QTT rank, so that the DMRG solver requires mere 2–3 iterations to solve the systems with relative residuals not greater than 10^{-8}.

The results are presented in Table 1. For time-singular problems with $\gamma = -\frac{1}{2}$ we use a finer time discretization ($M = 30$) in order to resolve the degeneracy of the operator at small times t, since we observe that the error thus introduced propagates in time till $t = T$.

In all three cases our results show a linear dependence of the error and computation time on the number of dimensions d.

5.2 Incompatible initial data

In this section we present numerical experiments for the model problem (5.1) with $T = 1$, $\gamma = 0$ and for the incompatible initial data

$$u_0(x_1, \ldots, x_d) = 1 \quad \text{for} \quad x_k \in (0, 1), \quad 1 \le k \le d, \tag{5.5}$$

therefore $u_0 \in H^{1/2-\varepsilon}$ for $\varepsilon > 0$. Such initial data arise in financial models; for example, in the context of binary contracts.

Although for $\gamma = 0$ the operator $-t^\gamma \triangle$ does not degenerate at $t = 0$, even in this case it is well-known that the solution can, in general, not be extended smoothly to $t = 0$. The incompatibility of the initial data gives rise to a boundary layer at small t, which fades at positive times due to parabolic smoothing. The Fourier coefficients in (5.2) are explicitly given by $C_\alpha = (4\pi)^{-d} \prod_{k=1}^d (2\alpha_k - 1)^{-1}$, and the residual of the series after the truncation at a certain number of terms can be estimated with the help of the exponential integral (in particular, we do so to obtain the reference solution). The coefficients before Ψ_α in (5.2) decay exponentially with respect to time. The high-frequency modes with large $|\alpha| = \sum_{k=1}^d \alpha_k$ which account for the boundary layer at small times $t > 0$, appearing due to incompatible initial data, get "squeezed" exponentially as time advances by the parabolic evolution operator. For example, at $t = 1$ the first term of (5.2) with $\alpha = (1, \ldots, 1)$ approximates the solution with the accuracy not worse than $d \cdot 10^{-7}$ in the Chebyshev norm.

Remark 5.1. *As we know that Ψ_α is of QTT ranks $2, \ldots, 2$ for every α (see, e. g. [35, Section 2.3]), we expect the "squeezing" property of the parabolic operator to ultimately reduce the QTT ranks of the numerical solution to 2 for any spatial dimension d.*

However, if one is interested in the solution to (5.1) with initial data (5.5) for small times $t > 0$, the contribution of high-frequency modes may be significant, so that the accurate resolution of the temporal boundary layer requires a careful numerical treatment. We elaborate on this issue in the univariate case.

Unlike the problem considered in Section 5.1, for small times $t > 0$ several modes contribute significantly to the solution of (5.2) with the initial data (5.5). Therefore we report the effective QTT rank $r\left[\mathbf{u}_m^\delta\right]$ of u_m^δ, which is defined in (4.8) and need not be an integer, at time points t_m, $1 \le m \le M$. The memory required to store the corresponding QTT decomposition equals $2\,(dl-2)\,r\left[\mathbf{u}_m^\delta\right]^2 + 4r\left[\mathbf{u}_m^\delta\right]$ which can be considered as the number of "effective degrees of freedom" through which the QTT decomposition represents the solution.

Remark 5.2. *When the Neumann boundary conditions are imposed in the problem* (5.1)*, the exact solution reads* (5.2) *with the summation over* $\alpha \in \mathbb{N}_0^d = (\mathbb{N} \cup \{0\})^{\times d}$*, where* $\sqrt{\lambda_0} = 0$ *and* $\psi_0(x) = 1$*,* $x \in (0,1)$*. Then the lowest-frequency mode* $\Psi_{0,\dots,0}$ *is constant, and the first essential modes* Ψ_α *with* $|\alpha| = 1$ *contribute the following* d *terms to the solution:*

$$\exp\left(-\lambda_1 \frac{t^{\gamma+1}}{\gamma+1}\right)\,\left(C_{1,0,\dots,0}\,\psi_1 \otimes 1 \otimes \dots \otimes 1 + \dots + C_{0,\dots,0,1}\,1 \otimes \dots \otimes 1 \otimes \psi_1\right). \tag{5.6}$$

According to [28, Lemma 5.1], this contribution can be represented in the TT format with ranks $\mathbf{2},\dots,\mathbf{2}$ *in terms of the one-dimensional factors* $\psi_0 = 1$ *and* ψ_1*, which themselves have QTT decompositions of ranks* $1,\dots,1$ *and* $2,\dots,2$ *respectively. Therefore, by [28, Lemma 5.2] we conclude that the contribution* (5.6) *of the modes* Ψ_α *with* $|\alpha| = 1$ *can be represented in the QTT format with ranks*

$$3,\dots,3,\mathbf{2},4,\dots,4,\mathbf{2},\dots,\dots,\mathbf{2},4,\dots,4,\mathbf{2},3,\dots,3$$

bounded by 4 *from above. Therefore, the effective rank of the numerical solution orthogonalized to the constant mode* $\Psi_{0,\dots,0}$ *at large times* $t > 0$ *should be expected to be slightly below* 4 *in the case of the Neumann boundary conditions.*

5.2.1 Univariate case

Boundary layer and the spatial discretization for small times $t > 0$ The first step of the *hp*-DG time discretization is similar to a step of the implicit Euler scheme, which for $d = 1$ leads to the following singularly perturbed problem.

$$-\varepsilon^2 v'' + v = f \quad \text{in} \quad G = (-1,1), \quad v(1) = \alpha^+, \quad v(-1) = \alpha^-. \tag{5.7}$$

Formally, for $\varepsilon > 0$ it is a second-order differential equation. But, if $\varepsilon = 0$, then (5.7) is of order zero. Hence, if the boundary data is not compatible, i. e. $\alpha^\pm \neq f(\pm 1)$, then the solution to (5.7) contains boundary layer terms of the form

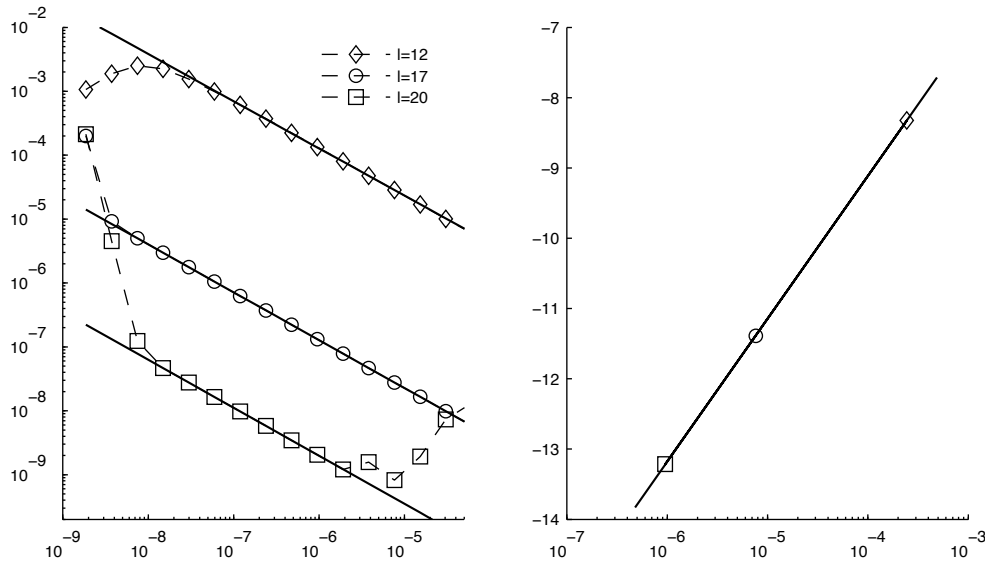
$$v^-(x) = e^{-(x+1)/\varepsilon}, \quad v^+(x) = e^{-(1-x)/\varepsilon}.$$

The following decomposition of the solution of (5.7) holds true, cf. [36].

Theorem 5.3. *Let* $n \in \mathbb{N}$ *be such that* $f \in H^{4n+2}(G)$*. Then* $v(x) = v_n(x) + A_n v^-(x) + B_n v^+(x)$ *for* $x \in G$*, where the constant* $C = C(n, \alpha^-, \alpha^+, f)$ *is independent of* ε *and* $\|v_n\|_{H^l(G)} + |A_n| + |B_n| \le C$ *for* $l = 0, 1, \dots, 2n+2$*.*

Theorem 5.3 suggests that the spatial finite element space is required to resolve accurately the boundary layer and the smooth part of the solution in order to achieve high accuracy for small times $t > 0$. The most straightforward way to ensure this within the discretization approach described in Section 4.4 is to use extremely fine meshes to construct the finite element spaces.

We discretize the problem in space using linear finite elements constructed on an equidistant mesh with step size $h_l = \left(2^l + 1\right)^{-1}$, as described in Section 4.4. The initial data satisfies $u_0 \in H_\theta$ for $\theta \in \left[0, \frac{1}{2}\right)$, and for the semidiscrete (discrete in space, continuous in time) equation one obtains theoretically the convergence $\mathcal{O}\left(h^2 t^{-1+\theta/2}\right)$ of the relative spatial L^2-error (see [37, Theorem 1] and [5, Proposition 2.2]). We use $l = 12$, $l = 17$ and $l = 20$ for the spatial discretization. The fully discrete problem involves also the *hp*-DG time-stepping, for which we choose $M = 30$ time intervals and the grading factor $\sigma = 0.5$. In order to ensure that the error of the time discretization is negligible compared to the error due to the space discretization and tensor-structured representation, we use the slope $\nu = 2$ of polynomial degrees of the temporal shape functions.



(a) $\texttt{err}\left[u_m^\delta\right]$ (5.3) vs. t_m for $m = 1, \ldots, 15$ and (b) A_l vs. h_l and the linear fit $A_l = B_l + \beta \log_{10} h_l$ linear fits $\log_{10} \texttt{err}\left[u_m^\delta\right] = A_l - \alpha_l \log_{10} t_m$ for $l = 12, 17, 20$

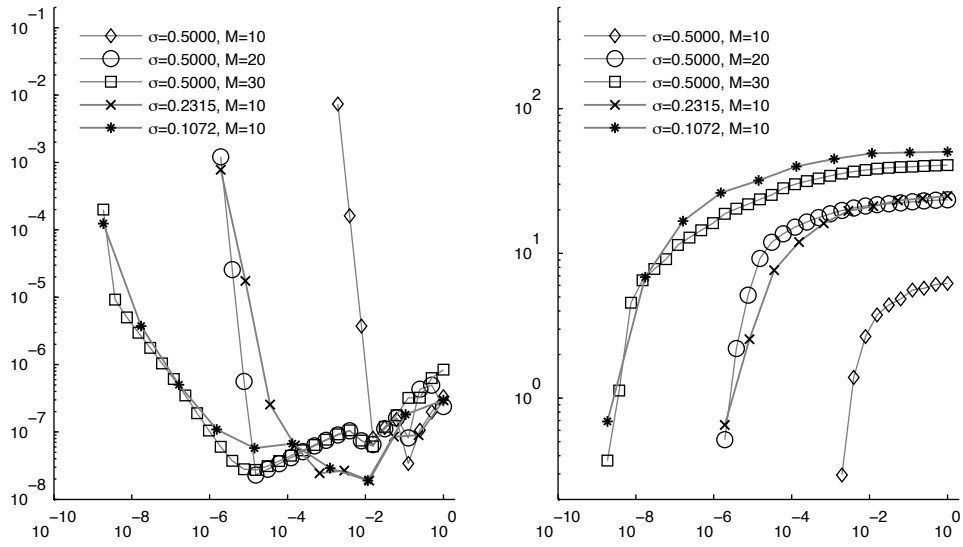
Figure 1: The refinement of the spatial equidistant mesh and the error of the fully discrete QTT-structured numerical solution at small times $t_m > 0$.

The first time step has the length $t_1 - t_0$, where $t_0 = 0$ and $t_1 = 2^{-29} \approx 1.86 \cdot 10^{-9}$. We set $\texttt{res} = 10^{-9}$ and $\texttt{swp} = 10$ in order to avoid that the error due to the tensor-structured representation dominates. The value of $\texttt{swp}$ proves not to limit the accuracy significantly and does not affect certain difficulties the DMRG solver encounters at the second half of the time steps ($m \geq 15$), which happens earlier for finer meshes.

The step sizes of the spatial meshes are $h_{12} \approx 2.44 \cdot 10^{-7}$, $h_{17} \approx 7.63 \cdot 10^{-7}$ and $h_{20} \approx 9.54 \cdot 10^{-7}$. In these three cases we observe the convergence rates $\alpha_{12} \approx -0.7382$, $\alpha_{17} \approx -0.7491$ and $\alpha_{20} \approx -0.7517$ with respect to t, see Figure 1(a). This is satisfactory, as the theoretical value is $-3/4$. Figure 1(b) plots the shift constants A_l of the linear fits $\log_{10} \texttt{err}\left[u_m^\delta\right] = A_l - \alpha_l \log_{10} t_m$ shown in Figure 1(a) vs. h_l. The data is fitted by $A_l = B_l + \beta \log_{10} h_l$ with $\beta \approx 2.0318$, while theoretical considerations suggest the rate of 2.

Figure 1(a) shows the characteristic features of the error, related to the spatial and temporal discretizations. For the coarsest spatial mesh ($l = 12$) the values of $\texttt{err}\left[u_m^\delta\right]$ at small t_m lie below the linear fit in the plot. This happens because the finite element space with $l = 12$ does not resolve the boundary layer at the first time steps, and the resolution of the error is too crude in the vicinity of the boundary; this effect disappears if we decrease the step size. However, for finer spatial meshes the values of $\texttt{err}\left[u_m^\delta\right]$ for small t_m lie above the corresponding linear fits. This appears due to the time discretization being too coarse at the first steps and may be easily avoided: in our experiments we observed that this effect vanishes when more temporal shape functions are used at first time steps; e. g., $p_m = 30 + m$ instead of $p_m = 2m$ for $m = 1, \ldots, 30$.

Refinement of the time discretization In this series of experiments we start with a time discretization with the slope of polynomial degrees $\nu = 2$, $M = 10$ time intervals and the grading factor $\sigma = 0.5$ (then $t_1 \approx 1.95 \cdot 10^{-3}$). We consider two refinement strategies. First, we keep $\nu = 2$ and $\sigma = 0.5$ and increase the number of time steps to $M = 20$ and $M = 30$, so that $t_1 \approx 1.91 \cdot 10^{-6}$ and $t_1 \approx 1.86 \cdot 10^{-9}$ respectively. Second, we reconsider $M = 10$, choose $\nu = 3$ and obtain approximately the same refined values of t_1 by decreasing the grading factor: we set $\sigma = 0.2315$ and $\sigma = 0.1072$, so that the length of the first time interval is $t_1 \approx 1.91 \cdot 10^{-6}$ and $t_1 \approx 1.87 \cdot 10^{-9}$ respectively. In all five experiments we use an equidistant mesh with $l = 17$ for the spatial discretization to ensure that the error is dominated by that of the temporal discretization. The results are presented in Figure 2(a) and Figure 2(b).



(a) $\texttt{err}\left[u_m^\delta\right]$ (5.3) vs. t_m (b) computation time vs. t_m

Figure 2: Comparison of DG-discretizations in time

Figure 2(a) demonstrates that by decreasing t_1 we improve the accuracy at small times $t > 0$. However, we see that the error is smoothened out at large times, which is typical for parabolic problems, and all 5 temporal discretizations yield approximately the same accuracy at $t = 1$. Figure 2(b) shows that they also require similar computation time.

For small times $t > 0$ the time discretization with the smallest t_1 is accurate enough to show the algebraic convergence of the error with respect to t, which we have already discussed above. Another observation is that both refinement strategies yield approximately the same accuracies.

Spatial discretization on Shishkin meshes From the previous paragraph we see that one has to use very fine spatial meshes in order to resolve the boundary layer in the solution at the first time steps. However, as the solution smoothens in time, it can be resolved on far coarser meshes at the last time steps. Moreover, finer spatial meshes may even yield worse accuracies at the last time steps: the iterations of the DMRG solver we exploit to solve linear systems (4.1) get stuck at relatively large residuals at the last time steps. We believe that this happens mostly due to the conditioning of the systems we solve, which relates to the conditioning of "local problems" that constitute a single DMRG iteration (see [34] for details). Thus, using the same equidistant mesh for the accurate solution of (5.2), (5.5) on all intervals may be inefficient. Instead, one may use alternative approaches to the spatial discretization.

An obvious possibility is to use equidistant meshes in space with different mesh sizes on different time intervals. This involves prolongation and restriction operations, which can be performed efficiently in the QTT format. For the case of mixed Dirichlet-Neumann boundary conditions, when the numbers of degrees of freedom are even for all the meshes involved, these operations were considered in [28, Section 4].

In the present paper we apply another approach, which exploits the spatial localization of the non-smooth features of the solution. We use *Shishkin meshes*, which are piecewise-equidistant meshes with a mesh width h in the interior of the domain and a mesh width $\tilde{h}$ used in a ρ-neighborhood of the boundary. For example, for a singularly perturbed elliptic problem (5.7) ρ is given as $\rho = \mathcal{O}\left(\varepsilon \ln \varepsilon^{-1}\right)$ and $\tilde{h} = \mathcal{O}\left(\varepsilon h \ln h^{-1}\right)$, see [38, Chapter I.2.4] for a detailed analysis including uniform (in ε) convergence results. Parabolic problems with boundary layers are considered in [38, Chapter II.3.4.2 and Chapter II.4.2.1]; for applications see, e. g. [39]. In our problem, at the first time step the *hp*-DG time-stepping with geometric mesh refinement corresponds to a singularly perturbed problem with $\varepsilon^2 = \sigma^{M-1}$.

We set the width of the boundary zone ρ and mesh sizes h and $\tilde{h} < h$, and construct equidistant meshes

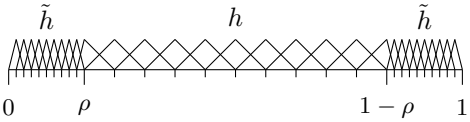


Figure 3: A Shishkin mesh in 1D. The meshwidths are h and $\tilde{h}$, the width of the boundary zone is ρ.

mesh	l	h	$\tilde{h}$	ρ	$h/\tilde{h}$
`Eq0`	12	$2.44 \cdot 10^{-4}$			
`Sh1`	15	$2.44 \cdot 10^{-4}$	$7.88 \cdot 10^{-6}$	0.05	31
`Eq1`	17	$7.63 \cdot 10^{-6}$			
`Sh2`	14	$2.44 \cdot 10^{-4}$	$1.39 \cdot 10^{-6}$	0.02	176
`Eq2`	20	$9.54 \cdot 10^{-7}$			

Table 2: Spatial discretizations on equidistant and Shishkin meshes

of step sizes $\tilde{h}$, h and $\tilde{h}$ in $[0, \rho]$, $[\rho, 1-\rho]$ and $[1-\rho, 1]$ respectively. The spatial finite elements are defined as described in Section 4 (see Figure 3). We choose the mesh parameters so that the number of degrees of freedom is 2^l for some $l \in \mathbb{N}$.

The mass and stiffness matrices can be represented in the QTT format with low ranks. Specifically, in the case of homogeneous Dirichlet boundary conditions the following can be shown: first, the one-dimensional stiffness and mass matrices $\mathbf{M}_k$ and $\mathbf{S}_k$ involved in (4.3) and (4.10) have QTT representations of ranks bounded by 6. Second, the d-dimensional stiffness and mass matrices $\mathbf{M}$ (4.3) and $\mathbf{S}$ (4.10) have QTT representations of ranks bounded by 16 (cf. the QTT rank estimates for equidistant meshes discussed in Sections 4.3 and 4.4). These rank estimates do not depend on d, l_k, $k = 1, \ldots, d$ and the mesh parameters ρ, h and $\tilde{h}$ (they also do not have to be the same in different dimensions). For the sake of brevity, we do not present the corresponding QTT decompositions in this paper. Also, the discretization of a function on a Shishkin mesh may have higher QTT ranks than on an equidistant mesh.

In this section we consider the problem (5.2), (5.5) in one dimension and apply the hp-DG time discretization with $M = 30$ time intervals, the grading factor $\sigma = 0.5$ and the slope $\nu = 2$ of polynomial degrees of temporal shape functions.

We compare five spatial discretizations (see Table 2). We start with an equidistant mesh with $l = 12$ and again refine it to $l = 17$ and $l = 20$. Also we consider two Shishkin meshes with such parameters that their mesh size in the boundary zone is approximately the same as for the refined equidistant meshes, and in the rest of the domain it is equal to that of the coarsest equidistant mesh. We use the DMRG solver with the parameters $\texttt{res} = 10^{-7}$ and $\texttt{swp} = 5$ (further iterations of the solver do not improve the accuracy significantly). For meshes `Eq1` and `Eq2` the relative residuals of linear systems rise up to approximately 10^{-6} ($m = 30$) and 10^{-3} ($m \geq 24$) respectively. For other meshes the residuals remain below 10^{-7}.

For this series of experiments we plot $r\left[\mathbf{u}_m^\delta\right]$, which is the effective QTT rank of $\mathbf{u}_m^\delta$, defined in (4.8), vs t_m. However, the error contained in the numerical solution u_m^δ may include high-rank QTT components of small norms (compared to the accuracy of the numerical solution). We drop them by truncating $\mathbf{u}_m^\delta$ to $\mathbf{u}_m^{\varepsilon_m}$. The truncation parameter ε_m depends on $\texttt{err}\left[u_m^\delta\right]$ and is chosen so that $\texttt{err}\left[u_m^{\varepsilon_m}\right] \leq 1.01 \cdot \texttt{err}\left[u_m^\delta\right]$, where

$$\texttt{err}\left[u_m^{\varepsilon_m}\right] = \frac{\|u_m^{\varepsilon_m} - \widehat{u}_m\|_{L^2(D)}}{\|\widehat{u}_m\|_{L^2(D)}} = \frac{\|\mathbf{u}_m^{\varepsilon_m} - \widehat{\mathbf{u}}_m\|_{\mathbf{M}}}{\|\widehat{\mathbf{u}}_m\|_{\mathbf{M}}} \tag{5.8}$$

similarly to (5.3). The effective QTT rank of the QTT representation of $\mathbf{u}_m^{\varepsilon_m}$, obtained in this way, is denoted by $r\left[\mathbf{u}_m^{\varepsilon_m}\right]$. It shows how much memory (or how many "effective degrees of freedom") one needs to represent the numerical solution in the QTT format with its actual accuracy.

Figure 4(a) shows the spatial error vs. time for all five spatial discretizations, all of which yield approximately the same accuracies after the last time steps. The evolution of the spatial error for the three equidistant meshes for small times $t > 0$ is similar to what we see in Figure 1(a). Also these experiments show that the finest spatial mesh leads to a substantial loss of accuracy at large times. Shishkin meshes `Sh1` and `Sh2` provide

the same accuracies as their equidistant counterparts `Eq1` and `Eq2` for very small times. The error rises when the boundary layer propagates outside the boundary zone of width ρ, and smoothens afterwards. The peak is less pronounced for `Sh1`, as its boundary zone is wider and the ratio $h/\tilde{h}$ is smaller. Also, the propagation of the boundary layer outside the boundary zone increases the QTT ranks of the numerical solution. This can be seen in Figure 4(c) for the QTT representations which include high-rank components of the error, and also in Figure 4(d), for their counterparts obtained by the accuracy-based QTT truncation. This rise of the QTT ranks makes the DMRG-based solution of linear systems more time-consuming, as can be seen in Figure 4(b).

We also observe in Figure 4(c) and Figure 4(d) that for large times $t > 0$ the effective QTT rank of the solution tends to 2 (for equidistant meshes) or a slightly larger constant (for Shishkin meshes) logarithmically fast. This agrees with the discussion of the "squeezing" property of the parabolic operator in Remark 5.1. The QTT ranks of a single term of the expansion (5.2) discretized on a Shishkin mesh are higher than for an equidistant mesh. On the other hand, the numbers of the "quantization levels" which we use for Shishkin meshes is less than for their equidistant counterparts. Overall, we find Shishkin meshes to require approximately the same computational effort as equidistant meshes, see Figure 4(b), and to allow resolving the boundary layer at small times without losing accuracy at large times.

5.2.2 Multivariate case

Next we consider the case of high dimensions. We use the equidistant mesh with $l = 8$ to construct tensor-product finite element subspaces in d spatial dimensions, where d varies from 1 to 18. As for the temporal discretization, the geometric partition of the time interval we use in one dimension appears to be unsuitable for high d. In particular, we observe that last time intervals become too large in order to allow the accurate time discretization with reasonable polynomial degrees. Therefore in this series of experiments we use the following combinative geometric-uniform partition of the time interval and the following polynomial degrees:

$$t_m = \begin{cases} (m-10)\cdot 0.05, & 12 \le m \le 30, \\ 2^{(m-11)}\cdot 0.05, & 1 \le m \le 11, \end{cases} \quad \text{and} \quad p_m = \begin{cases} 36, & 12 \le m \le 30, \\ 3m, & 1 \le m \le 11. \end{cases}$$

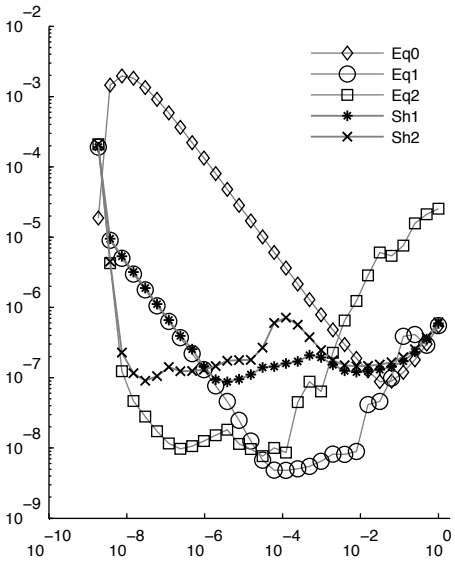
The first 11 time intervals are graded geometrically by the factor of $\sigma = 0.5$ and involve linearly increasing polynomial degrees with the slope $\nu = 3$. The remaining 19 intervals have the same length 0.05 and involve the constant polynomial degree 36. Such a partition is suitable for the entire range of d we consider in this experiment.

In Figure 5(a) and Figure 5(b) the spatial error and computation time exhibit the same behavior as we see in one dimension for the coarsest grid and corresponding times t in Figure 4(a) and Figure 4(b) respectively. Figure 5(c) shows that the numerical solution evolves logarithmically fast into a function of QTT ranks $2, \dots, 2$ independently of d, which gives an illustration to Remark 5.1. From Figure 5(d) we see that both the complexity and spatial error evaluated at $t = 1$ grow linearly with respect to d.

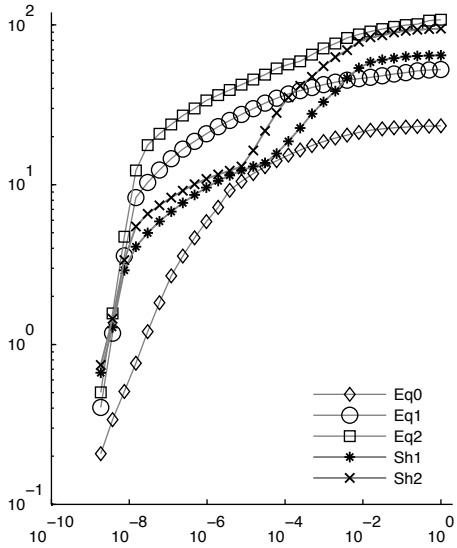
5.3 General remarks

The numerical experiments considered above represent two important cases in which the problem (5.1) requires a careful treatment at small times $t > 0$. First, if the operator is degenerate (namely, when $-1 < \gamma < 0$), the time discretization has to approximate the solution at small times with sufficient accuracy. Otherwise, even low-frequency modes of the solution in the Fourier expansion (5.2), get polluted and the accuracy gets lost irrevocably at the very first time step. Second, even in case $\gamma = 0$, i. e. in the absence of the time singularity, the incompatibility of the initial data with the boundary conditions is well-known to give rise to a boundary layer. However, the inconsistency due to this incompatibility decays exponentially as time advances. Therefore, high resolution is crucial for the accuracy at small times only. The *hp*-DG-timestepping and the QTT-Shishkin space discretization proposed here prove to be efficient in dealing with both these complications.

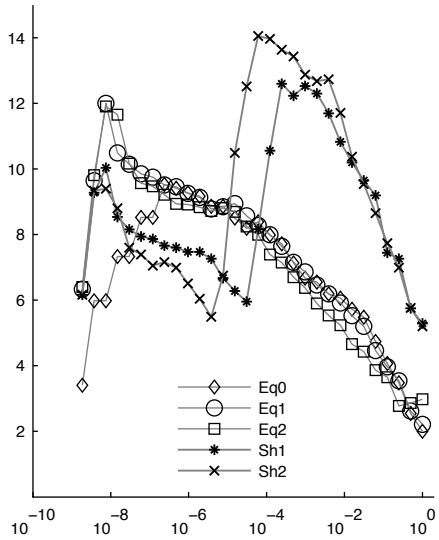
Our examples involve symmetric rank-1 separable solutions, which allows to obtain easily the reference solution and is therefore convenient for numerical experiments. On the other hand, this also implies the linear scaling of the algorithm with respect to d (because the TT ranks splitting the "physical" dimensions are $1, \dots, 1$ independently of d). However, the algorithm uses no a priori information about the structure of the solution and has to find it adaptively with the help of the QTT format. In the very same way the



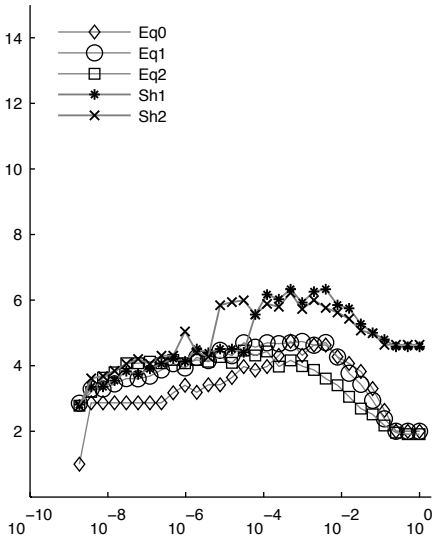
(a) $\mathtt{err}\,[u_m^{\varepsilon_m}]$ (5.8) vs. t_m



(b) computation time vs. t_m

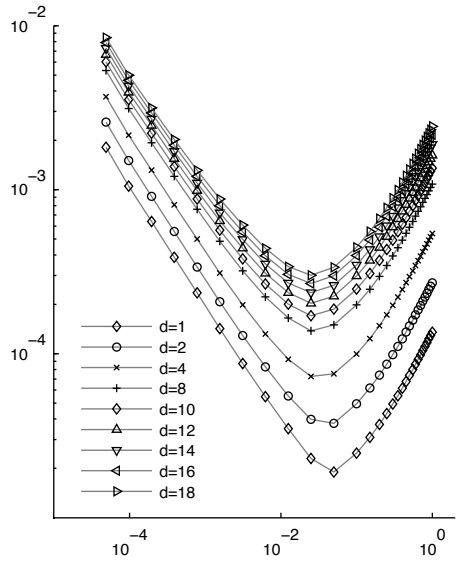


(c) Effective QTT rank $r\left[\mathbf{u}_m^{\delta}\right]$ vs. t_m

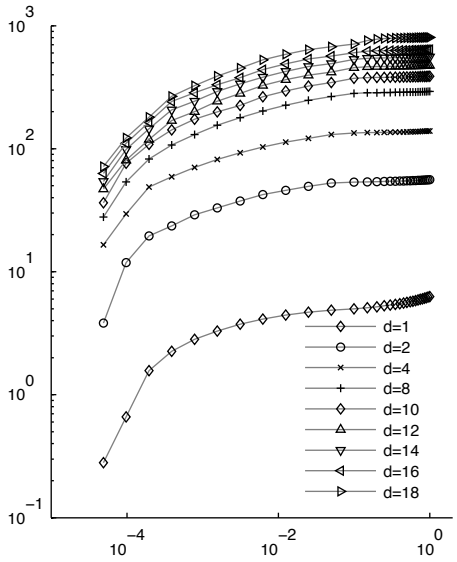


(d) Effective QTT rank $r\left[\mathbf{u}_m^{\varepsilon_m}\right]$ vs. t_m

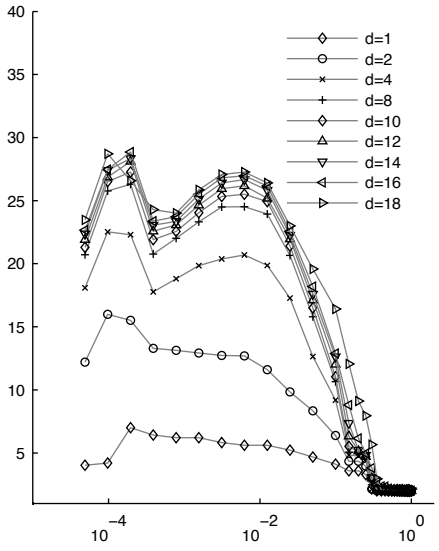
Figure 4: Spatial discretization on equidistant and Shishkin meshes



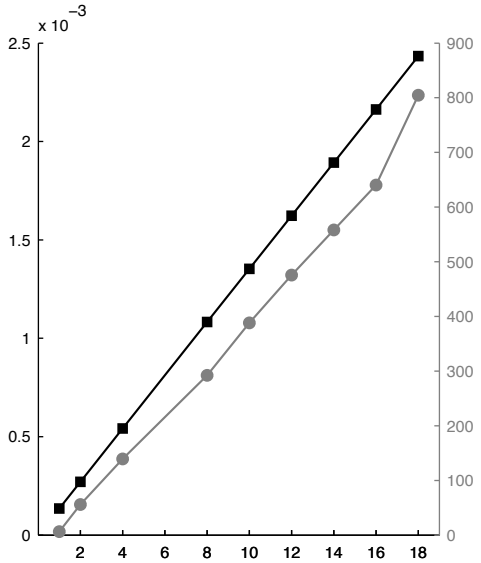
(a) $\mathtt{err}\left[u_m^{\varepsilon_m}\right]$ (5.8) vs. t_m



(b) computation time vs. t_m



(c) Effective QTT rank $r\left[\mathbf{u}_m^{\delta}\right]$ vs. t_m



(d) $\mathtt{err}\left[u_M^{\varepsilon_M}\right]$ (5.8) (black) and total computation time (gray) vs. d

Figure 5: Multivariate case with incompatible initial data

algorithm can be applied to more practical, *essentially* multi-dimensional problems, in which the solution can be assumed to have the TT ranks growing reasonably slow with respect to d.

6 Conclusion

We presented a scheme for the efficient numerical solution of singular and degenerate parabolic evolution equations in high-dimensional domains. The approach is based on an *hp* time-stepping procedure with variable approximation order and variable time step. This allows to exploit the parabolic smoothing and yields exponential convergence of the time-semidiscrete solution despite the degeneracy at time $t = 0$. The spatial discretization of the sequence of elliptic problems resulting from the time-stepping is based on the the so-called *Quantized Tensor Train* format coupled with very straightforward (full tensor product) finite elements.

The parabolic nature of the problem manifests itself also in a strong reduction of the effective degrees of freedom, needed to describe the solution. The inherently adaptive nature of the QTT compression was shown to identify and localize the "active" degrees of freedom very efficiently. With the use of the approach proposed, parabolic problems in dimension d up 40 are solved on a laptop in minutes.

We mention that the approach presented here is found to be competitive with wavelet-based sparse, compressive tensor product discretizations, such as the ones presented in [40, 41], which were to some extent tailored to the particular problem (heat equation) at hand, whereas the present approach is, to a large extent, based on the black-box TT-structured arithmetics and solver of linear systems included in the TT Toolbox (publicly available at `http://spring.inm.ras.ru/osel`).

Let us also comment briefly on the recent related work [42]. There, a non-degenerate time-homogeneous parabolic Fokker-Planck problem is considered and two approaches are proposed. The first approach ("time stepping") is based on the Crank-Nicolson scheme with an equidistant mesh in time. At every time step it leads to an elliptic problem in space, which, after the FDM discretization, is also solved in the QTT format with the use of the TT Toolbox. The second approach ("block solution") suggests the QTT-structured solution of the simultaneous time-space FDM discretization of the initial parabolic problem. However, the authors note that this method appears to be unsuitable for large time intervals throughout which the solution changes significantly, and apply it for a partitioned time interval with restarts.

Our method differs from the time-stepping approach of [42] in the following respects. First, as we mentioned, it handles the time inhomogeneity of the diffusion operator. Second, it exploits the localization of the singularity. For the time stepping we use the geometrically graded mesh and variable approximation order, which allows us to distribute the computational effort related to the time stepping according to the features of the solution and, thus, yields a substantial reduction in the number of spatial problems to be solved. As for the spatial discretization, we propose to use Shishkin meshes in order to resolve the boundary layer due to incompatible initial data.

We compare the performance of the *hp*-DG-QTT method proposed in the present paper to the approaches from [42], see [42, Table 2] and Figure 5(b) in the present paper. The class of parabolic problems under consideration is (5.1) with the incompatible initial data (5.5) in the case of $d = 2$ dimensions. For the spatial discretization $l = 8$ levels of quantization are used. About the same accuracy $\approx 2 \cdot 10^{-4}$ at $t = 1$ is obtained with 4096 and 30 time steps in 6232 and 56 seconds by the "time stepping" approach [42] and the *hp*-DG-QTT method respectively. The "block solution" [42] approach yields a similar accuracy for 2048 time steps in 69 seconds, which is comparable to our result.

Let us note that, for the ease of the presentation, in our numerical experiments we only considered spatially homogeneous isotropic diffusion. The major conclusions of this paper remain true, however, in the presence of the diffusion anisotropy and for non-analytic temporal coefficients a.

References

[1] *O. Reichmann.* Optimal space-time adaptive wavelet methods for degenerate parabolic PDEs (to appear in Numer. Math. (2012)): Tech. Rep. 03: SAM, ETH, 2011. `http://www.sam.math.ethz.ch/reports/2011/03`. 2, 3, 4

[2] *K.-I. Sato*. Lévy processes and infinitely divisible distributions. — Cambridge: Cambridge Studies in Advanced Mathematics, 1999. 2

[3] *N. Reich, Ch. Schwab, Ch. Winter*. On Kolmogorov Equations for Anisotropic Multivariate Lévy Processes // *Finance and Stochastics*. 2010. V. 14. P. 527–567. 2

[4] *F. Biagini, Y. Hu, B. Øksendal, T. Zhang*. Stochastic Calculus for Fractional Brownian Motion and Applications. — New York: Springer, 2008. 3

[5] *D. Schötzau*. *hp*-DGFEM for parabolic evolution problems: Ph.D. thesis / ETH Zürich. — 1999. 5, 6, 7, 19

[6] *S. Roman*. The formula of Faà di Bruno // *Amer. Math. Monthly*. 1980. V. 87. P. 805–809. 7

[7] *M. Costabel, M. Dauge, S. Nicaise*. Corner Singularities and Analytic Regularity for Linear Elliptic Systems. Part I: Smooth domains. `http://hal.archives-ouvertes.fr/hal-00453934`. 7

[8] *D. Schötzau, Ch. Schwab*. An *hp* a priori error analysis of the DG time-stepping method for initial value problems // *Calcolo*. 2000. V. 37, No. 4. P. 207–232. 7

[9] *D. Schötzau, Ch. Schwab*. *hp*-discontinuous Galerkin time-stepping for parabolic problems // *C.R.Acad. Sci. Paris*. 2001. V. 333. P. 1121-1126. 7

[10] *A. Pazy*. Semigroups of Linear Operators and Applications to Partial Differential Equations. — Berlin: Springer-Verlag, 1983. 8

[11] *R. Bellman*. Adaptive Control Processes: A Guided Tour. — Princeton, NJ: Princeton University Press, 1961. 14

[12] *I. Oseledets, E. Tyrtyshnikov*. Recursive decomposition of multidimensional tensors // *Doklady Mathematics*. 2009. V. 80. P. 460–462. 10.1134/S1064562409040036. `http://dx.doi.org/10.1134/S1064562409040036`. 14

[13] *I. Oseledets*. A new tensor decomposition // *Doklady Mathematics*. 2009. V. 80. P. 495–496. DOI: `10.1134/S1064562409040115`. `http://dx.doi.org/10.1134/S1064562409040115`. 14

[14] *I. V. Oseledets, E. E. Tyrtyshnikov*. Breaking the curse of dimensionality, or how to use SVD in many dimensions // *SIAM Journal on Scientific Computing*. 2009, October. V. 31, No. 5. P. 3744–3759. DOI: `10.1137/090748330`. `http://epubs.siam.org/sisc/resource/1/sjoce3/v31/i5/p3744_s1`. 14

[15] *I. V. Oseledets*. Tensor Train decomposition // *SIAM Journal on Scientific Computing*. 2011. V. 33, No. 5. P. 2295–2317. DOI: `10.1137/090752286`. `http://dx.doi.org/10.1137/090752286`. 14

[16] *I. V. Oseledets, E. E. Tyrtyshnikov*. TT-cross approximation for multidimensional arrays // *Linear Algebra and its Applications*. 2010, January. V. 432, No. 1. P. 70–88. DOI: `10.1016/j.laa.2009.07.024`. `http://www.sciencedirect.com/science/article/pii/S0024379509003747`. 14

[17] *I. Oseledets*. Approximation of matrices with logarithmic number of parameters // *Doklady Mathematics*. 2009. V. 80. P. 653–654. DOI: `10.1134/S1064562409050056`. `http://dx.doi.org/10.1134/S1064562409050056`. 14

[18] *B. Khoromskij*. $\mathcal{O}(d \log N)$-Quantics Approximation of N-d Tensors in High-Dimensional Numerical Modeling // *Constructive Approximation*. 2011. V. 34, No. 2. P. 257–280. DOI: `10.1007/s00365-011-9131-1`, 10.1007/s00365-011-9131-1. `http://www.springerlink.com/content/06n7q85q14528454/`. 14, 15

[19] *I. V. Oseledets*. Approximation of $2^d \times 2^d$ matrices using tensor decomposition // *SIAM Journal on Matrix Analysis and Applications*. 2010. V. 31, No. 4. P. 2130–2145. DOI: `10.1137/090757861`. `http://link.aip.org/link/?SML/31/2130/1`. 14

[20] *E. E. Tyrtyshnikov*. Tensor approximations of matrices generated by asymptotically smooth functions // *Sbornik: Mathematics*. 2003. V. 194, No. 5. P. 941–954. DOI: 10.1070/SM2003v194n06ABEH000747. http://iopscience.iop.org/1064-5616/194/6/A09. 14

[21] *W. Hackbusch, S. Kühn*. A New Scheme for the Tensor Representation // *Journal of Fourier Analysis and Applications*. 2009. V. 15. P. 706–722. DOI: 10.1007/s00041-009-9094-9, 10.1007/s00041-009-9094-9. http://www.springerlink.com/content/t3747nk47m368g44/. 15

[22] *L. Grasedyck*. Hierarchical Singular Value Decomposition of Tensors // *SIAM Journal on Matrix Analysis and Applications*. 2010. V. 31, No. 4. P. 2029–2054. DOI: 10.1137/090764189. http://link.aip.org/link/?SML/31/2029/1. 15

[23] *D. Kressner, Ch. Tobler*. Preconditioned low-rank methods for high-dimensional elliptic PDE eigenvalue problems: Research Report 48: Seminar for Applied Mathematics, ETHZ, 2011. http://www.sam.math.ethz.ch/reports/2011/48. 15

[24] *L. Grasedyck*. Polynomial Approximation in Hierarchical Tucker Format by Vector-Tensorization: Preprint 308: Institut für Geometrie und Praktische Mathematik, RWTH Aachen, 2010, April. http://www.igpm.rwth-aachen.de/Download/reports/pdf/IGPM308_k.pdf. 15

[25] *T. G. Kolda, B. W. Bader*. Tensor Decompositions and Applications // *SIAM Review*. 2009, September. V. 51, No. 3. P. 455–500. DOI: 10.1.1.153.2059. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.153.2059&rep=rep1&type=pdf. 15

[26] *B. N. Khoromskij*. Tensors-structured numerical methods in scientific computing: Survey on recent advances // *Chemometrics and Intelligent Laboratory Systems*. 2011. V. 110, No. 1. P. 1–19. DOI: 10.1016/j.chemolab.2011.09.001. http://www.sciencedirect.com/science/article/pii/S0169743911001808. 15

[27] *W. Hackbusch*. Tensor Spaces and Numerical Tensor Calculus. — Springer, 2012. — V. 42 of *Springer Series in Computational Mathematics*. http://www.springerlink.com/content/l62t86. 15

[28] *V. A. Kazeev, B. N. Khoromskij*. Low-rank explicit QTT representation of the Laplace operator and its inverse // *To appear in SIAM Journal on Matrix Analysis and Applications*. 2012. 15, 16, 19, 21

[29] *V. Kazeev, O. Reichmann, Ch. Schwab*. Low-rank tensor structure of linear diffusion operators in the TT and QTT formats // *In progress*. 15

[30] *N. Hilber, S. Kehtari, Ch. Schwab, Ch. Winter*. Wavelet finite element method for option pricing in highdimensional diffusion market models: Tech. rep.: SAM Report 01-2010, ETH Zurich, 2010. http://www.sam.math.ethz.ch/reports/2010/01. 15

[31] *Ch. Reisinger, G. Wittum*. Efficient hierarchical approximation of high-dimensional option pricing problems // *SIAM J. Sci. Comput.* 2007. V. 29, No. 1. 15

[32] *N. Hilber*. Stabilized Wavelet Method for Pricing in High Dimensional Stochastic Volatility Models: Ph.D. thesis / SAM, ETH Dissertation No. 18176. — 2009. http://e-collection.ethbib.ethz.ch/view/eth:41687. 15

[33] *O. Ladyzenskaja, V. Solonnikov, N. Ural'ceva*. Linear and quasi-linear equations of parabolic type. Translations of Math. Monog. — AMS, Philadelphia, 1968. 17

[34] *S. V. Dolgov, I. V. Oseledets*. Solution of linear systems and matrix inversion in the TT-format: Preprint 19: Max-Planck-Institut für Mathematik in den Naturwissenschaften, 2011. http://www.mis.mpg.de/publications/preprints/2011/prepr2011-19.html. 17, 21

[35] *I. V. Oseledets*. Constructive representation of functions in tensor formats: Preprint 4: Institute of Numerical Mathematics of RAS, 2010, August. http://pub.inm.ras.ru/pub/inmras2010-04.pdf. 18

[36] *Ch. Schwab, M. Suri.* The p and hp versions of the finite element method for problems with boundary layers // *Math. Comput.* 1996. V. 65, No. 216. P. 1403–1429. 19

[37] *A.-M. Matache, Ch. Schwab, T. Wihler.* Linear complexity solution of parabolic integro-differential equations // *Numer. Math.* 2006. V. 104, No. 1. P. 69–102. 19

[38] *H.-G. Roos, M. Stynes, L. Tobiska.* Robust Numerical Mothods for Singulary Perturbed Differential Equations. — Second edition. — New York: Springer, 2008. 21

[39] *S. Franz, R. B. Kellogg, M. Stynes.* Galerkin and streamline diffusion finite element methods on a Shishkin mesh for a convection-diffusion problem with corner singularities // *Mathematics of Computation.* 2011. V. 81, No. 278. P. 661–685. http://www.ams.org/journals/mcom/2012-81-278/S0025-5718-2011-02526-3/S0025-5718-2011-02526-3.pdf. 21

[40] *T. von Petersdorff, Ch. Schwab.* Numerical solution of parabolic equations in high dimensions // *M2AN Math. Model. Numer. Anal.* 2004. V. 38, No. 1. P. 93–127. DOI: 10.1051/m2an:2004005. http://dx.doi.org/10.1051/m2an:2004005. 26

[41] *T. J. Dijkema, Ch. Schwab, R. Stevenson.* An adaptive wavelet method for solving high-dimensional elliptic PDEs // *Constr. Approx.* 2009. V. 30, No. 3. P. 423–455. DOI: 10.1007/s00365-009-9064-0. http://dx.doi.org/10.1007/s00365-009-9064-0. 26

[42] *S. V. Dolgov, B. N. Khoromskij, I. V. Oseledets.* Fast solution of multi-dimensional parabolic problems in the TT/QTT-format with initial application to the Fokker-Planck equation: Preprint 80: Max-Planck-Institut für Mathematik in den Naturwissenschaften, 2011. http://www.mis.mpg.de/publications/preprints/2011/prepr2011-80.html. 26

Research Reports

No.	Authors/Title
12-11	*V. Kazeev, O. Reichmann and Ch. Schwab* *hp*-DG-QTT solution of high-dimensional degenerate diffusion equations
12-10	N.H. Risebro and F. Weber A note on front tracking for the Keyfitz-Kranzer system
12-09	*U. Koley and N.H. Risebro* Convergence of finite difference schemes for symmetric Keyfitz-Kranzer system
12-08	*S. Mishra, Ch. Schwab and J. Šukys* Monte Carlo and multi-level Monte Carlo finite volume methods for uncertainty quantification in nonlinear systems of balance laws
12-07	*A. Hiltebrand and S. Mishra* Entropy stable shock capturing space-time discontinuous Galerkin schemes for systems of conservation laws
12-06	*R. Hiptmair, A. Moiola and I. Perugia* Trefftz discontinuous Galerkin methods for acoustic scattering on locally refined meshes
12-05	*C. Winteler, R. Käppeli, A. Perego, A. Arcones, N. Vasset, N. Nishimura, M. Liebendörfer and F.-K. Thielemann* Magneto-rotationally driven Supernovae as the origin of early galaxy r-process elements?
12-04	*P. Grohs* Intrinsic localization of anisotropic frames
12-03	*P. Grohs* Geometric multiscale decompositions of dynamic low-rank matrices
12-02	*D. Kressner and C. Tobler* htucker - A Matlab toolbox for tensors in hierarchical Tucker format
12-01	*F.Y. Kuo, Ch. Schwab and I.H. Sloan* Quasi-Monte Carlo methods for high dimensional integration - the standard (weighted Hilbert space) setting and beyond
11-72	*P. Arbenz, A. Hiltebrand and D. Obrist* A parallel space-time finite difference solver for periodic solutions of the shallow-water equation
11-71	*M.H. Gutknecht* Spectral deflation in Krylov solvers: A theory of coordinate space based methods