Efficient numerical approximation of a non-regular Fokker–Planck equation associated with first-passage time distributions

Boehm, U.; Cox, S.; Gantner, G.; Stevenson, R.

DOI
10.1007/s10543-022-00914-2

Publication date
2022

Document Version
Final published version

Published in
Bit : numerical mathematics

BIT Numerical Mathematics (2022) 62:1355–1382
https://doi.org/10.1007/s10543-022-00914-2

BIT

Efficient numerical approximation of a non-regular Fokker–Planck equation associated with first-passage time distributions

Udo Boehm[1] · Sonja Cox[2] · Gregor Gantner[3] · Rob Stevenson[2]

Received: 18 March 2021 / Accepted: 28 February 2022 / Published online: 4 April 2022

Abstract
In neuroscience, the distribution of a decision time is modelled by means of a one-dimensional Fokker–Planck equation with time-dependent boundaries and space-time-dependent drift. Efficient approximation of the solution to this equation is required, e.g., for model evaluation and parameter fitting. However, the prescribed boundary conditions lead to a strong singularity and thus to slow convergence of numerical approximations. In this article we demonstrate that the solution can be related to the solution of a parabolic PDE on a rectangular space-time domain with homogeneous initial and boundary conditions by transformation and subtraction of a known function. We verify that the solution of the new PDE is indeed more regular than the solution of the original PDE and proceed to discretize the new PDE using a space-time minimal residual method. We also demonstrate that the solution depends analytically on the

Communicated by Axel Målqvist.

GG was supported by the Austrian Science Fund (FWF) under grant J4379-N.

✉ Gregor Gantner
gregor.gantner@asc.tuwien.ac.at

Udo Boehm
u.bohm@uva.nl

Sonja Cox
s.g.cox@uva.nl

Rob Stevenson
r.p.stevenson@uva.nl

1 Department of Psychology, University of Amsterdam, PO Box 15906, 1001, NK Amsterdam, The Netherlands

2 Korteweg–de Vries (KdV) Institute for Mathematics, University of Amsterdam, PO Box 94248, 1090, GE Amsterdam, The Netherlands

3 Institute of Analysis and Scientific Computing, TU Wien, Wiedner Hauptstraße 8-10, 1040 Vienna, Austria

parameters determining the boundaries as well as the drift. This justifies the use of a sparse tensor product interpolation method to approximate the PDE solution for various parameter ranges. The predicted convergence rates of the minimal residual method and that of the interpolation method are supported by numerical simulations.

Keywords Fokker–Planck equation · Time-dependent spatial domain · Space-time variational formulation · Parameter-dependent PDE · Sparse tensor product interpolation

Mathematics Subject Classification 30B40 · 35A15 · 35B65 · 35K08 · 60H30 · 65D05 · 65M12

1 Introduction

In 1978 Ratcliff [24] introduced a model for binary decision processes based on diffusion processes. This model turned out to agree well with experimental data; Gold and Shadlen [17] provides a neurophysiological explanation for its success. Indeed, the solution $(X_t)_{t\geq 0}$ of a one-dimensional stochastic differential equation is assumed to describe the difference in activity of two competing neuron populations. At time $t = 0$, the value $X_0 = x_0 \in \mathbb{R}$ represents the resting-state activity of the neuron populations. A decision is triggered when $(X_t)_{t\geq 0}$ first reaches one of two (possibly time-dependent) critical values α or β, each reflecting an outcome of the decision process.

In a typical decision experiment, scientists can only measure the decision time and outcome. Parameter fitting thus requires access to the decision time distributions, which are rarely known explicitely. Ad hoc numerical simulations are costly whence efficient simulation methods are much sought-after [15, 18].

In this article we extend and improve a simulation method introduced in [30], which is based on the Fokker–Planck equation associated to the decision time. In particular, this article may be viewed as the theoretical counterpart of our publication [3], which is aimed at the neuroscientific community.

Linking the first hitting time of a stochastic differential equation to a Fokker–Planck equation is a well-known approach that has also been applied in e.g. astrophysics [7] and cell biology [20]; for an overview see [1]. In particular, although we only consider examples arising from neuroscience, the simulation method we introduce is also relevant for other applications.

To explain the Fokker–Planck based approach consider the following stochastic differential equation:

$$dX_t^y = \mu(t, X_t^y)\,dt + \sigma\,dW_t \quad t \in [0,\infty),\ X_0^y = y. \tag{1.1}$$

Here $(W_t)_{t\in[0,\infty)}$ is a Brownian motion, $\sigma \in (0,\infty)$ is the diffusion parameter, $\mu \in C([0,\infty)\times\mathbb{R})$ is the (time- and state-dependent) drift and $y \in \mathbb{R}$ is the initial value. Let $\alpha, \beta \in C^1([0,\infty))$ satisfy $\alpha \leq \beta$, and for all $y \in [\alpha(0), \beta(0)]$ define the stopping

times $\hat{\alpha}_y, \hat{\beta}_y$ by

$$\begin{aligned} \hat{\alpha}_y &:= \inf\{t \in [0,\infty)\colon X_t^y \le \alpha(t)\}, \\ \hat{\beta}_y &:= \inf\{t \in [0,\infty)\colon X_t^y \ge \beta(t)\}. \end{aligned} \tag{1.2}$$

The quantities of interest in neurophysiological decision models are the *first hitting time probabilities*: $\mathbb{P}[\hat{\alpha}_y \le \min(\tau, \hat{\beta}_y)]$, where $\tau \in (0,\infty)$ and $y \in [\alpha(0), \beta(0)]$. These probabilities can be linked to the solution of a parabolic PDE. Indeed, assume $\alpha < \beta$ on $[0,\tau]$ for some $\tau \in (0,\infty)$, set $Q := \{(t,x) \in (0,\tau)\times\mathbb{R}\colon \alpha(\tau - t) < x < \beta(\tau - t))\}$, and consider the following PDE:

$$\begin{cases} \partial_t F(t,x) = \frac{\sigma^2}{2}\partial_x^2 F(t,x) + \mu(\tau - t, x)\partial_x F(t,x) & (t,x) \in Q_\tau, \\ F(t,\alpha(\tau - t)) = 1, \quad F(t,\beta(\tau - t)) = 0 & t \in (0,\tau), \\ F(0,x) = 0 & x \in (\alpha(\tau), \beta(\tau)). \end{cases} \tag{1.3}$$

Under some additional regularity assumptions on α, β, and μ it can be shown that a solution to (1.3) exists and satisfies

$$\mathbb{P}[\hat{\alpha}_y \le \min(\tau, \hat{\beta}_y)] = F(\tau, y), \quad \alpha(0) \le y \le \beta(0). \tag{1.4}$$

(see [30, Appendix A] for the case that α and β are constant and μ does not depend on time or [23, Chapter 7] for general Fokker–Planck equations, also known in this setting as a backward Kolmogorov equation).

In [30], a Crank–Nicolson method is used to approximate solutions to (1.3) in the case that α, β, and μ are constant. One advantage of this setting is that one only needs to solve a single PDE of type (1.3) in order to obtain the first hitting time probabilities $\mathbb{P}[\hat{\alpha}_y \le \min(t, \hat{\beta}_y)]$ for all $t \in [0,\tau]$, $y \in [\alpha(0), \beta(0)]$. However, due to the fact that F is discontinuous at $(t,x) = (0, \alpha(\tau))$, no proof of convergence of the Crank–Nicolson for decreasing step-sizes seems available. At best, reduced rates are to be expected. Moreover, various authors have argued that time-dependent boundaries α and β and space-time-dependent drift μ provide a more realistic model for decision processes, for an overview see [18, 25].

In this article we *extend* [30] to include diffusion models with time-dependent boundaries and non-constant drift. We *improve* the efficiency of the numerical simulation by not approximating the solution F to (1.3) directly, instead, we approximate the solution to a parabolic PDE on a rectangular domain with homogeneous initial and boundary conditions constructed such that its difference with F (transformed to the same rectangular domain) is a function for which a rapidly converging series expansion is known.

More specifically, in Section 2 we demonstrate that if α, β are once continuously differentiable, then (1.3) can be transformed into a parabolic PDE on a rectangular domain with a space-time-dependent drift. Next, in Section 3 we demonstrate that by subtracting a known, discontinuous function, we obtain a parabolic PDE with homogeneous boundary conditions, see (3.1) below. We analyze the regularity of the solution e to this equation and verify that it is indeed smoother than F, see Corollary 3.1 and Theorem 3.1.

In Section 4 we apply a minimal residual method [2, 28, 29] to approximate the solution e to (3.1). This method is known to give quasi-best approximations from the selected trial space in the norm on a natural solution space being the intersection of two Bochner spaces. Taking as trial space the space of continuous piecewise bilinears with respect to a uniform partition of the space-time cylinder into rectangles with mesh width h, in Theorem 4.1 the optimal error bound of order h is shown for the solution e to (3.1).

In Section 5 we consider the situation that μ, α, and β can be parametrized analytically and verify that in this case the corresponding solution e to (3.1) (transformed onto the unit square) depends analytically on these parameters as well as on the final time τ, see Theorem 5.1. This justifies the use of a sparse tensor-product interpolation [22] to determine the solution e to (3.1) efficiently for multiple end-time and parameter values. Finally, in Section 6 we provide numerical simulations for three different decision models taken from the neurophysiological literature.

In our parallel publication [3] mentioned above, we provide further numerical experiments and code. There, we apply the Crank–Nicolson method (without giving any error analysis) to approximate the solution e to (3.1). In the examples we consider it appears that the Crank–Nicolson method leads to similar convergence as the minimal residual method. Although we only provide a rigorous error analysis for the minimal residual method, Crank–Nicolson may be preferred in practice as it is easier to implement. We refer to [3] for further details.

1.1 Notation

In this work, by $C \lesssim D$ we mean that C can be bounded by a multiple of D, independently of parameters which C and D may depend on. Obviously, $C \gtrsim D$ is defined as $D \lesssim C$, and $C \eqsim D$ as $C \lesssim D$ and $C \gtrsim D$.

For normed linear spaces E and F, by $\mathcal{L}(E, F)$ we denote the normed linear space of bounded linear mappings $E \to F$, and by $\mathcal{L}_{\mathrm{iso}}(E, F)$ its subset of boundedly invertible linear mappings $E \to F$.

2 Transforming the Fokker–Planck equation to a rectangular space-time domain

In this section we demonstrate that (1.3) can be transformed into a PDE on a rectangular space-time domain, see (2.3) below. The PDE in (2.3) below forms the starting point for the remainder of this article, which is why we use tildes in (2.1) below to distinguish the variables and coefficients of the non-transformed equation from those in (2.3). Indeed, let $\widetilde{T} \in (0, \infty]$, assume $a, b \in C^1([0, \widetilde{T}))$ satisfy $a(\tilde{t}) < b(\tilde{t})$ for all $\tilde{t} \in [0, \widetilde{T})$, set $\widetilde{Q} := \{(\tilde{t}, \tilde{x}) \in (0, \widetilde{T}) \times \mathbb{R} \colon a(\tilde{t}) < \tilde{x} < b(\tilde{t})\}$, let $\tilde{v} \in L_\infty(\widetilde{Q})$, and consider the

following parabolic initial- and boundary value problem:

$$\begin{cases} \partial_{\tilde{t}}\tilde{u}(\tilde{t},\tilde{x}) = \partial_{\tilde{x}}^2\tilde{u}(\tilde{t},\tilde{x}) + \tilde{v}(\tilde{t},\tilde{x})\partial_{\tilde{x}}\tilde{u}(\tilde{t},\tilde{x}) & (\tilde{t},\tilde{x}) \in \widetilde{Q}, \\ \tilde{u}(\tilde{t},a(\tilde{t})) = 1, \quad \tilde{u}(\tilde{t},b(\tilde{t})) = 0 & \tilde{t} \in (0,\widetilde{T}), \\ \tilde{u}(0,\tilde{x}) = 0 & \tilde{x} \in (a(0),b(0)). \end{cases} \tag{2.1}$$

Note that this is (1.3) with $\tilde{u}(\tilde{t},\tilde{x}) = F(\frac{2\tilde{t}}{\sigma^2},\tilde{x})$, $\widetilde{T} = \frac{\sigma^2\tau}{2}$, $a(\tilde{t}) = \alpha(\frac{2}{\sigma^2}(\widetilde{T}-\tilde{t}))$, $b(\tilde{t}) = \beta(\frac{2}{\sigma^2}(\widetilde{T}-\tilde{t}))$, $\tilde{v}(\tilde{t},\tilde{x}) = \frac{2}{\sigma^2}\mu(\frac{2}{\sigma^2}(\widetilde{T}-\tilde{t}),\tilde{x})$.

Now, set $T := \int_0^{\tilde{T}} |b(\tilde{s})-a(\tilde{s})|^{-2}\,d\tilde{s}$ (where possibly $T = \infty$) and define $\theta\colon [0,T) \to [0,\tilde{T})$ by $\theta(t) = \sup\left\{\tilde{r} \in [0,\widetilde{T})\colon \int_0^{\tilde{r}} |b(\tilde{s})-a(\tilde{s})|^{-2}\,d\tilde{s} \le t\right\}$, then θ is a bijection and $\theta^{-1}(\tilde{t}) = \int_0^{\tilde{t}} |b(\tilde{s})-a(\tilde{s})|^{-2}\,d\tilde{s}$. In particular, from $t = \theta^{-1}(\theta(t))$ we obtain that θ satisfies the following ODE

$$\theta'(t) = \big(b(\theta(t)) - a(\theta(t))\big)^2, \quad \theta(0) = 0. \tag{2.2}$$

With

$$\Omega := (0,1),$$

and $\xi\colon [0,\widetilde{T}) \times \overline{\Omega} \to \mathbb{R}$ defined by

$$\xi(\tilde{t},x) := (1-x)a(\tilde{t}) + xb(\tilde{t}),$$

we have that

$$[0,T) \times \Omega \to \widetilde{Q}\colon (t,x) \mapsto (\theta(t),\xi(\theta(t),x))$$

is a bijection with inverse

$$(\tilde{t},\tilde{x}) \mapsto \Big(\theta^{-1}(\tilde{t}), \frac{\tilde{x}-a(\tilde{t})}{b(\tilde{t})-a(\tilde{t})}\Big).$$

Defining $u, v\colon [0,T) \times \overline{\Omega} \to \mathbb{R}$ by

$$\begin{aligned} u(t,x) &:= \tilde{u}(\theta(t),\xi(\theta(t),x)), \\ v(t,x) &:= (b(\theta(t)) - a(\theta(t)))\big[\tilde{v}(\theta(t),\xi(\theta(t),x)) + (1-x)a'(\theta(t))) + xb'(\theta(t))\big], \end{aligned}$$

we have $u(t,0) = 1$, $u(t,1) = 0$ $(t \in (0,T))$, and $u(0,x) = 0$ $(x \in \Omega)$. Moreover, for $(t,x) \in (0,T) \times \Omega$, one has

$$\begin{aligned} \partial_x u(t,x) &= (b(\theta(t)) - a(\theta(t)))\partial_{\tilde{x}}\tilde{u}(\theta(t),\xi(\theta(t),x)), \\ \partial_x^2 u(t,x) &= (b(\theta(t)) - a(\theta(t)))^2\partial_{\tilde{x}}^2\tilde{u}(\theta(t),\xi(\theta(t),x)), \end{aligned}$$

and

$$\begin{aligned}\partial_t u(t,x) &= \theta'(t)\big\{\partial_{\tilde t}\tilde u(\theta(t),\xi(\theta(t),x)) + \partial_{\tilde t}\xi(\theta(t),x)\partial_{\tilde x}\tilde u(\theta(t),\xi(\theta(t),x))\big\}\\ &= (b(\theta(t))-a(\theta(t)))^2\Big\{\partial_{\tilde t}\tilde u(\theta(t),\xi(\theta(t),x))\\ &\quad + \big[(1-x)a'(\theta(t)) + xb'(\theta(t))\big]\partial_{\tilde x}\tilde u(\theta(t),\xi(\theta(t),x))\Big\}\\ &= (b(\theta(t))-a(\theta(t)))^2\Big\{\partial_{\tilde x}^2\tilde u(\theta(t),\xi(\theta(t),x))\\ &\quad + \big[\tilde v(\theta(t),\xi(\theta(t),x)) + (1-x)a'(\theta(t)) + xb'(\theta(t))\big]\partial_{\tilde x}\tilde u(\theta(t),\xi(\theta(t),x))\Big\}\\ &= \partial_x^2 u(t,x) + v(t,x)\partial_x u(t,x).\end{aligned}$$

In other words, with

$$I := (0,T),$$

(2.1) is equivalent to finding $u = u(v)$ that solves

$$\begin{cases}\partial_t u(t,x) = \partial_x^2 u(t,x) + v(t,x)\partial_x u(t,x) & (t,x)\in I\times\Omega,\\ u(t,0)=1,\quad u(t,1)=0 & t\in I,\\ u(0,x)=0 & x\in\Omega.\end{cases}\tag{2.3}$$

To be able to numerically solve (2.3), we assume from now on that $T<\infty$.

Example 2.1 Bowman, Kording, and Gottfried [4] suggested collapsing boundaries, i.e., in (1.3) they take $\alpha(t) := \frac{\beta_0 t}{2T_0}$ and $\beta(t) := \beta_0(1-\frac{t}{2T_0})$ for some fixed parameters $\beta_0, T_0 \in (0,\infty)$. Translating this to the setting of (2.1), this leads to $a(\tilde t) := \frac{\beta_0(\widetilde T-\tilde t)}{\sigma^2 T_0}$ and $b(\tilde t) := \beta_0(1-\frac{\widetilde T-\tilde t}{\sigma^2 T_0})$ (note that it only makes sense to consider $\widetilde T \in (0,\frac{\sigma^2 T_0}{2})$ in this setting). Note that it is easier to first determine $\theta^{-1}(\tilde t) = \int_0^{\tilde t} |b(\tilde s)-a(\tilde s)|^{-2}\,d\tilde s$ and then determine $T=\theta^{-1}(\widetilde T)$ and $\theta = (\theta^{-1})^{-1}$. Indeed, $\theta^{-1}(\tilde t) = \frac{\sigma^4 T_0^2\tilde t}{\beta_0^2(\sigma^2T_0-2\widetilde T)(\sigma^2T_0-2\widetilde T+2\tilde t)}$ and thus $T = \frac{\sigma^2T_0\widetilde T}{\beta_0^2(\sigma^2T_0-2\widetilde T)}$ and

$$\theta(t) = \frac{\beta_0^2(\sigma^2T_0-2\widetilde T)^2 t}{\sigma^4T_0^2 - 2\beta_0^2(\sigma^2T_0-2\widetilde T)t},\quad t\in[0,T).$$

By observing that $(1-x)a'(\theta(t)) + xb'(\theta(t)) = \frac{(2x-1)\beta_0}{\sigma^2T_0}$, and $b(\theta(t))-a(\theta(t)) = \frac{b_0(1-2\sigma^2T_0\widetilde T)}{\sigma^4T_0^2-2\beta_0^2(\sigma^2T_0-2\widetilde T)t}$, one obtains v in terms of $\tilde v$.

3 Regularity of the Fokker–Planck equation

Let $u(v)$ denote the solution to (2.3) for some given drift function v. Due to the discontinuity between boundary and initial data, it is clear that $u(v)$ is discontinuous at the corner $(t, x) = (0, 0)$. This reduces the rate of convergence of standard numerical methods and makes it difficult to provide a theoretical bound on the convergence rate. However, for *constant drift* v, a rapidly converging series expansion of $u(v)$ is known ([16]), which allows to efficiently approximate $u(v)$ within any given positive tolerance. Knowing this, our approach to approximate $u(v)$ for *variable* $v \in C(\overline{I \times \Omega})$ is to *approximate the difference*

$$e = e(v) = u(v) - u(v_0), \text{ where } v_0 := v(0, 0).$$

This function $e(v)$ solves

$$\begin{cases} \partial_t e(t,x) = \partial_x^2 e(t,x) + v(t,x)\partial_x e(t,x) + (v(t,x) - v_0)\partial_x u(v_0) & (t,x) \in I \times \Omega, \\ e(t,0) = 0, \quad e(t,1) = 0 & t \in I, \\ e(0,x) = 0 & x \in \Omega, \end{cases} \tag{3.1}$$

which we solve approximately with a numerical method. To derive a priori bounds for the approximation error, we analyze the smoothness of $e(v)$, see Section 3.3. In particular, under additional smoothness conditions on v, and using that $(v - v_0)(0, 0) = 0$, we show that

$$e(v) \textit{ is more smooth than } u(v_0), \textit{ and thus than } u(v),$$

which shows the benefit of applying the numerical method to (3.1) instead of directly to (2.3).

It turns out that for any v the smoothness of $u(v)$ is determined by that of the solution u_H of the heat equation on $(0, \infty) \times \mathbb{R}$ that is 0 at $t = 0$ and 1 at $x = 0$. Its smoothness is the topic of the next subsection.

3.1 The heat kernel

The function

$$H(t, x) := \frac{1}{2\sqrt{\pi t}} e^{-\frac{x^2}{4t}}$$

is the heat kernel. It satisfies

$$\begin{aligned} \partial_t H(t,x) &= \partial_x^2 H(t,x) && (t,x) \in (0,\infty) \times \mathbb{R}, \\ \lim_{t \downarrow 0} \int_{\mathbb{R}} H(t,x)\phi(x)\,dx &= \phi(0) && \text{for all } \phi \in \mathcal{D}(\mathbb{R}), \end{aligned}$$

the latter being the space of *test functions*.

Following [10, Ex. 2.14] and [14], for $(t, x) \in (0, \infty) \times \mathbb{R}$ we define

$$u_H(t, x) := 2 \int_x^\infty H(t, y)\, dy = \frac{2}{\sqrt{\pi}} \int_{\frac{x}{2\sqrt{t}}}^\infty e^{-s^2}\, ds = \operatorname{Erfc}(\tfrac{x}{2\sqrt{t}}).$$

Knowing that $\int_0^\infty \frac{1}{\sqrt{\pi t}} e^{-\frac{y^2}{4t}}\, dy = 1$, and $\lim_{t \downarrow 0} \int_x^\infty \frac{1}{\sqrt{\pi t}} e^{-\frac{y^2}{4t}}\, dy = 0$ for $x > 0$, we have

$$\begin{cases} 2\partial_t u_H(t, x) = \partial_x^2 u_H(t, x) & (t, x) \in (0, \infty) \times \mathbb{R}, \\ u_H(t, 0) = 1 & t > 0, \\ u_H(0, x) := \lim_{t \downarrow 0} u_H(t, x) = 0 & x > 0. \end{cases}$$

The following lemma turns out to be handy to analyze the smoothness of u_H restricted to $I \times \Omega$.

Lemma 3.1 *For $p > 0$, $\alpha, \beta \in \mathbb{R}$, it holds that $\int_0^T \int_0^1 |t^\alpha x^\beta e^{-\frac{x^2}{4t}}|^p\, dx\, dt < \infty$ if and only if $p\beta > -1$ and $p(2\alpha + \beta) > -3$.*

Proof The mapping

$$\Phi \colon \left\{(\lambda, x) \in (0, \infty) \times (0, 1) \colon x < 2\sqrt{\lambda T}\right\} \to (0, T) \times (0, 1) \colon (\lambda, x) \mapsto \left(\frac{x^2}{4\lambda}, x\right)$$

is a diffeomorphism, and $|D\Phi(\lambda, x)| = \frac{x^2}{4\lambda^2}$. One obtains

$$\begin{aligned} \int_0^T \int_0^1 |t^\alpha x^\beta e^{-\frac{x^2}{4t}}|^p\, dx\, dt &= \int_0^\infty \int_0^{\min(1, 2\sqrt{\lambda T})} \left(\frac{x^2}{4\lambda}\right)^{\alpha p} x^{\beta p} e^{-p\lambda} \frac{x^2}{4\lambda^2}\, dx\, d\lambda \\ &= 4^{-\alpha p - 1} \int_0^\infty \lambda^{-\alpha p - 2} e^{-p\lambda} \int_0^{\min(1, 2\sqrt{\lambda T})} x^{2\alpha p + \beta p + 2}\, dx\, d\lambda. \end{aligned}$$

The integral over x is finite if and only if $p(2\alpha + \beta) > -3$, and if so, the expression is equal to

$$\frac{4^{-\alpha p - 1}}{2\alpha p + \beta p + 3} \left[(2\sqrt{T})^{2\alpha p + \beta p + 3} \int_0^{\frac{1}{4T}} \lambda^{(\beta p - 1)/2} e^{-p\lambda} d\lambda + \int_{\frac{1}{4T}}^\infty \lambda^{-\alpha p - 2} e^{-p\lambda}\, d\lambda \right]$$

with the first integral being finite if and only if $p\beta > -1$. □

Following [31], we analyze the regularity of the solutions $u(v)$ and $e(v)$ of the parabolic problems (2.3) and (3.1), respectively, in (intersections of) Bochner spaces. In particular, the space $L_2(I; H^1(\Omega)) \cap H^1(I; H^{-1}(\Omega))$ plays an important role in this and following sections. For the precise definition of this space and some properties we refer to [31, Chapter 25]. With $H^1_{0,\{0\}}(I)$ denoting the closure in $H^1(I)$ of the functions in $C^\infty(I) \cap H^1(I)$ that vanish at 0, we have the following result concerning the smoothness of u_H restricted to $I \times \Omega$.

Corollary 3.1 *$u_H \in L_2(I;H^1(\Omega)) \cap H^1_{0,\{0\}}(I;H^{-1}(\Omega))$, but $u_H \notin H^1_{0,\{0\}}(I;L_2(\Omega))$ and $u_H \notin L_2(I;H^2(\Omega))$. Furthermore, $t\partial_t\partial_x u_H, x\partial_x^2 u_H, t\partial_x^2 u_H \in L_2(I\times\Omega)$, and $x\partial_t\partial_x u_H \in L_2(I;H^{-1}(\Omega))$.*

Proof By applications of Lemma 3.1, we infer that $\partial_x u_H = -2H \in L_2(I\times\Omega)$, and that $\partial_t u_H(t,x) = \frac{1}{2\sqrt{\pi}} x t^{-\frac{3}{2}} e^{-\frac{x^2}{4t}} \notin L_2(I\times\Omega)$. This yields $u_H \in L_2(I;H^1(\Omega))$ and $u_H \notin H^1_{0,\{0\}}(I;L_2(\Omega))$.

If $\partial_x F = f$, then $f \in L_2(I;H^{-1}(\Omega))$ if and only if $F \in L_2(I\times\Omega)$. We have $\int_{-\infty}^x \partial_t u_H(t,y)\,dy = -\frac{t^{-\frac{1}{2}}}{\sqrt{\pi}} e^{-\frac{x^2}{4t}} \in L_2(I\times\Omega)$, so indeed $u_H \in H^1_{0,\{0\}}(I;H^{-1}(\Omega))$.

It holds

$$\partial_x^2 u_H(t,x) = -2\partial_x H(t,x) = \frac{1}{2\sqrt{\pi}} x t^{-\frac{3}{2}} e^{-\frac{x^2}{4t}} \notin L_2(I\times\Omega),$$

or $u_H \notin L_2(I;H^2(\Omega))$, but $x\partial_x^2 u_H, t\partial_x^2 u_H \in L_2(I\times\Omega)$.

We have $\partial_x\partial_t u_H = (t^{-\frac{3}{2}} - \frac{1}{2}x^2 t^{-\frac{5}{2}})\frac{e^{-\frac{x^2}{4t}}}{2\sqrt{\pi}}$, so $t\partial_x\partial_t u_H \in L_2(I\times\Omega)$. Proving that $x\partial_x\partial_t u_H \in L_2(I;H^{-1}(\Omega))$ amounts to proving $x t^{-\frac{3}{2}} e^{-\frac{x^2}{4t}}$, $t^{-\frac{5}{2}} x^3 e^{-\frac{x^2}{4t}} \in L_2(I;H^{-1}(\Omega))$, i.e., proving that $t^{-\frac{3}{2}}\int_{-\infty}^x y e^{-\frac{y^2}{4t}}dy$, $t^{-\frac{5}{2}}\int_{-\infty}^x y^3 e^{-\frac{y^2}{4t}}dy \in L_2(I\times\Omega)$. The first function equals $-2t^{-\frac{1}{2}}e^{-\frac{x^2}{4t}}$, which is in $L_2(I\times\Omega)$, and the second function equals $-8t^{-\frac{1}{2}}e^{-\frac{x^2}{4t}} - 2t^{-\frac{3}{2}}x^2 e^{-\frac{x^2}{4t}}$, which is also in $L_2(I\times\Omega)$. □

Finally in this subsection, notice that from $\partial_t u_H(t,x) = \frac{1}{2\sqrt{\pi}} x t^{-\frac{3}{2}} e^{-\frac{x^2}{4t}}$, it follows that for any $x > 0$ and $k \in \mathbb{N}_0$,

$$\lim_{t\downarrow 0} \partial_t^k u_H(t,x) = 0. \tag{3.2}$$

3.2 Regularity of the parabolic problem with homogeneous initial and boundary conditions

Knowing that $e(v)$ is the solution of the parabolic problem (3.1) that has homogeneous initial and boundary conditions, we study the regularity of such a problem.

Given functions $v \in L_\infty(I\times\Omega)$ and $f \in L_2(I;H^{-1}(\Omega))$, let w solve

$$\begin{cases} \partial_t w(t,x) = \partial_x^2 w(t,x) + v(t,x)\partial_x w(t,x) + f(t,x) & (t,x)\in I\times\Omega, \\ w(t,0) = 0, \quad w(t,1) = 0 & t \in I, \\ w(0,x) = 0 & x \in \Omega, \end{cases} \tag{3.3}$$

where the spatial differential operators at the right-hand side should be interpreted in a weak sense, i.e., $((\partial_x^2 + v\partial_x)\eta)(\zeta) := \int_D -\partial_x\eta\partial_x\zeta + v\partial_x\eta\,\zeta\,dx$. It is well-known

that

$$L(v) := w \mapsto f \in \mathcal{L}_{\text{iso}}(L_2(I; H_0^1(\Omega)) \cap H^1_{0,\{0\}}(I; H^{-1}(\Omega)), L_2(I; H^{-1}(\Omega))) \tag{3.4}$$

(see, e.g., [31, Thm. 26.1]). Under additional smoothness conditions on the right-hand side f beyond being in $L_2(I; H^{-1}(\Omega))$, additional smoothness of the solution w can be demonstrated:

Proposition 3.1 *a) If $v \in W^1_\infty(I \times \Omega)$, then*

$$L(v)^{-1} \in \mathcal{L}\Big(L_2(I; H^1(\Omega)) \cap H^1(I; H^{-1}(\Omega)), \\ H^1_{0,\{0\}}(I; H_0^1(\Omega)) \cap H^2(I; H^{-1}(\Omega)) \cap L_2(I; H^3(\Omega))\Big).$$

b) If $v \in L_\infty(I \times \Omega)$, then

$$L(v)^{-1} \in \mathcal{L}\Big(L_2(I \times \Omega), L_2(I; H^2(\Omega)) \cap H^1_{0,\{0\}}(I; L_2(\Omega))\Big),$$

Proof a) If $f \in L_2(I; H^1(\Omega)) \cap H^1(I; H^{-1}(\Omega))$, then also $f \in H^1(I; H^{-1}(\Omega))$, and $f(0,\cdot) \in L_2(\Omega)$ with $\|f(0,\cdot)\|_{L_2(\Omega)} \lesssim \|f\|_{L_2(I;H^1(\Omega))} + \|f\|_{H^1(I;H^{-1}(\Omega))}$ (see, e.g., [31, Thm. 25.5]). As shown in [31, Thm. 27.2 and its proof], from the last two properties of f, and $v \in W^1_\infty(I; L_\infty(\Omega))$, one has $w = L(v)^{-1} f \in H^1_{0,\{0\}}(I; H_0^1(\Omega)) \cap H^2(I; H^{-1}(\Omega))$ with

$$\|w\|_{H^1_{0,\{0\}}(I;H_0^1(\Omega))\cap H^2(I;H^{-1}(\Omega))} \lesssim \|f\|_{H^1(I;H^{-1}(\Omega))} + \|f(0,\cdot)\|_{L_2(\Omega)}.$$

To show the spatial regularity, i.e., $w \in L_2(I; H^3(\Omega))$, given a constant λ, we define $w_\lambda(t,\cdot) = w(t,\cdot)e^{-\lambda t}$, $f_\lambda(t,\cdot) = f(t,\cdot)e^{-\lambda t}$. One infers that

$$(-\partial_x^2 - v\partial_x + \lambda) w_\lambda = \underbrace{f_\lambda - \partial_t w_\lambda}_{g_\lambda :=} \quad \text{on } I\times\Omega, \quad w_\lambda(\cdot,0) = 0 = w_\lambda(\cdot,1) \quad \text{on } I, \tag{3.5}$$

where, as before, the spatial differential operators should be interpreted in a weak sense. Using that

$$\left| \int_I \int_D v(\partial_x w_\lambda)\, w_\lambda \, dx\, dt \right| \le \|v\|_{L_\infty(I\times\Omega)} \|\partial_x w_\lambda\|_{L_2(I\times\Omega)} \|w_\lambda\|_{L_2(I\times\Omega)}$$

and Young's inequality, one infers that for $\lambda > \frac{1}{4}\|v\|^2_{L_\infty(I\times\Omega)}$ the bilinear form defined by the left-hand side of (3.5) is bounded and *coercive* on $L_2(I; H_0^1(\Omega)) \times L_2(I; H_0^1(\Omega))$. Thus for $\lambda > \frac{1}{4}\|v\|^2_{L_\infty(I\times\Omega)}$ we have

$$A(v,\lambda) := w_\lambda \mapsto g_\lambda \in \mathcal{L}_{\text{iso}}(L_2(I; H_0^1(\Omega)), L_2(I; H^{-1}(\Omega))).$$

Realizing that $\|\cdot\|^2_{H^{k+2}(\Omega)} = \|\frac{\mathrm{d}^k}{\mathrm{d}x^k}\frac{\mathrm{d}^2}{\mathrm{d}x^2}\cdot\|^2_{L_2(\Omega)} + \|\cdot\|^2_{H^{k+1}(\Omega)}$, an induction and tensor product argument shows $A(0,0)^{-1} \in \mathcal{L}(L_2(I;H^k(\Omega)), L_2(I;H^{k+2}(\Omega)))$ for any $k \in \mathbb{N}_0$. Writing

$$A(v,\lambda)^{-1} - A(0,0)^{-1} = A(0,0)^{-1}(v\partial_x - \lambda\mathrm{Id})A(v,\lambda)^{-1},$$

and using that $v\partial_x \in \mathcal{L}(L_2(I;H^1(\Omega)), L_2(I;L_2(\Omega)))$ by $v \in L_\infty(I\times\Omega)$, one verifies that $A(v,\lambda)^{-1} \in \mathcal{L}(L_2(I\times\Omega), L_2(I;H^2(\Omega)))$. Repeating the argument, now using that $v\partial_x \in \mathcal{L}(L_2(I;H^2(\Omega)), L_2(I;H^1(\Omega)))$ by $v \in L_\infty(I;W^1_\infty(\Omega))$, one has $A(v,\lambda)^{-1} \in \mathcal{L}(L_2(I;H^1(\Omega)), L_2(I;H^3(\Omega)))$. Knowing that $f_\lambda - \partial_t w_\lambda \in L_2(I;H^1(\Omega))$ with $\|f_\lambda - \partial_t w_\lambda\|_{L_2(I;H^1(\Omega))} \lesssim \|f\|_{L_2(I;H^1(\Omega))} + \|f\|_{H^1(I;H^{-1}(\Omega))}$, one infers that w_λ and thus $w \in L_2(I;H^3(\Omega))$, and moreover $\|w\|_{L_2(I;H^3(\Omega))} \lesssim \|f\|_{L_2(I;H^1(\Omega))} + \|f\|_{H^1(I;H^{-1}(\Omega))}$.

b) Similar to Part a), it suffices to show that

$$L(v,\lambda)^{-1} := f_\lambda \mapsto w_\lambda \in \mathcal{L}\Big(L_2(I\times\Omega), L_2(I;H^2(\Omega)) \cap H^1_{0,\{0\}}(I;L_2(\Omega))\Big).$$

Knowing that $L(v,\lambda)^{-1} \in \mathcal{L}\big(L_2(I;H^{-1}(\Omega)), L_2(I;H^1_0(\Omega)) \cap H^1_{0,\{0\}}(I;H^{-1}(\Omega))\big)$, and $L(v,\lambda) - L(0,0) = -v\partial_x + \lambda\mathrm{Id} \in \mathcal{L}\big(L_2(I;H^1_0(\Omega)), L_2(I\times\Omega)\big)$, the proof is completed by $L(v,\lambda)^{-1} - L(0,0)^{-1} = L(0,0)^{-1}(L(0,0) - L(v,\lambda))L(v,\lambda)^{-1}$ and the *maximal regularity* result

$$L(0,0)^{-1} \in \mathcal{L}\Big(L_2(I\times\Omega), L_2(I;H^2(\Omega)) \cap H^1_{0,\{0\}}(I;L_2(\Omega))\Big)$$

from, e.g., [11, 12]. □

3.3 The regularity of $e(v) = u(v) - u(v_0)$

Recall that u_H denotes the solution of the heat equation studied in Section 3.1, that $u(v)$ denotes the solution to (3.1) for given $v \in C(\overline{I\times\Omega})$, and $v_0 := v(0,0)$. Since $e(v)$ solves (3.1), i.e., $e(v)$ is the solution w of (3.3) for forcing function f given by

$$\begin{aligned}&(v - v_0)\partial_x u(v_0)\\ &\quad = (v - v_0)\partial_x(u(v_0) - u(0)) + (v - v_0)\partial_x(u(0) - u_H) + (v - v_0)\partial_x u_H,\end{aligned} \tag{3.6}$$

in view of the regularity results proven in Proposition 3.1, we establish smoothness of $e(v)$ by demonstrating smoothness of each of the three terms at the right-hand side of (3.6).

Lemma 3.2 *It holds that*

$$u(0) - u_H \in H^1_{0,\{0\}}(I;H^1_0(\Omega)) \cap H^2(I;H^{-1}(\Omega)) \cap L_2(I;H^3(\Omega)).$$

Proof The function $w(t,x) := u(0)(t,x) - (u_H(t,x) - xu_H(t,1))$ satisfies the homogeneous initial and boundary conditions from (3.3), and $\partial_t w(t,x) = \partial_x^2 w(t,x) + x\partial_t u_H(t,1)$. By (3.2) we have $(t,x) \mapsto x\partial_t u_H(t,1) \in L_2(I;H^1(\Omega)) \cap H^1(I;H^{-1}(\Omega))$, so that Proposition 3.1a) for $v = 0$ and $f(t,x) = x\partial_t u_H(t,1)$ shows that

$$w \in H^1_{0,\{0\}}(I;H^1_0(\Omega)) \cap H^2(I;H^{-1}(\Omega)) \cap L_2(I;H^3(\Omega)).$$

Because, again by (3.2), $(t,x) \mapsto xu_H(t,1)$ is in the same space, the proof is completed. □

Lemma 3.3 *For* any $v_0 \in \mathbb{R}$, $u(v_0) - u(0) \in L_2(I;H^2(\Omega)) \cap H^1_{0,\{0\}}(I;L_2(\Omega))$.

Proof The function $w := u(v_0) - u(0)$ satisfies the homogeneous initial- and boundary conditions from (3.3), and $\partial_t w(t,x) = \partial_x^2 w(t,x) + v_0\partial_x w - v_0\partial_x u(0)$. From $\partial_x u(0) \in L_2(I\times\Omega)$ by Corollary 3.1 and Lemma 3.2, an application of Proposition 3.1b) for $v = v_0$ and $f = -v_0\partial_x u(0)$ completes the proof. □

Lemma 3.4 *If* $v \in W^1_\infty(I\times\Omega) \cap L_\infty(I;W^2_\infty(\Omega))$, *then*

$$(v - v_0)\partial_x u_H \in L_2(I;H^1(\Omega)) \cap H^1(I;H^{-1}(\Omega)).$$

Proof Abbreviate $g := (v - v_0)\partial_x u_H$. Throughout the proof, we use the estimates for u_H proven in Corollary 3.1.

We start with proving $\partial_t g = (\partial_t v)\partial_x u_H + (v - v_0)\partial_t\partial_x u_H \in L_2(I;H^{-1}(\Omega))$. Using $v \in W^1_\infty(I;L_\infty(\Omega))$ and $\partial_x u_H \in L_2(I\times\Omega)$, the first term is even in $L_2(I\times\Omega)$. Writing the second term as

$$(v(t,x) - v_0)\partial_t\partial_x u_H(t,x) = \tfrac{v(t,x)-v(0,x)}{t} t\partial_t\partial_x u_H(t,x) + \tfrac{v(0,x)-v_0}{x} x\partial_t\partial_x u_H(t,x),$$

from $t\partial_t\partial_x u_H \in L_2(I\times\Omega)$ and $\frac{v(t,x)-v(0,x)}{t} \in L_\infty(I\times\Omega)$ by $v \in W^1_\infty(I;L_\infty(\Omega))$, we have $\frac{v(t,x)-v(0,x)}{t} t\partial_t\partial_x u_H(t,x) \in L_2(I\times\Omega)$. Similarly, from $x\partial_t\partial_x u_H(t,x) \in L_2(I;H^{-1}(\Omega))$ and $\frac{v(0,x)-v_0}{x} \in L_\infty(I;W^1_\infty(\Omega))$ by $v \in L_\infty(I;W^2_\infty(\Omega))$, we have $\frac{v(0,x)-v_0}{x} x\partial_t\partial_x u_H(t,x) \in L_2(I;H^{-1}(\Omega))$, so that $\partial_t g \in L_2(I;H^{-1}(\Omega))$.

It remains to show that $g \in L_2(I;H^1(\Omega))$. It is clear that $(v - v_0)\partial_x u_H \in L_2(I\times\Omega)$ and $(\partial_x v)\partial_x u_H \in L_2(I\times\Omega)$ by $v \in L_\infty(I;W^1_\infty(\Omega))$. Writing

$$(v(t,x) - v_0)\partial_x^2 u_H(t,x) = \tfrac{v(t,x)-v(0,x)}{t} t\partial_x^2 u_H(t,x) + \tfrac{v(0,x)-v_0}{x} x\partial_x^2 u_H(t,x),$$

from $\frac{v(t,x)-v(0,x)}{t}$, $\frac{v(0,x)-v_0}{x} \in L_\infty(I\times\Omega)$ by $v \in W^1_\infty(I\times\Omega)$, and both $t\partial_x^2 u_H(t,x)$ and $x\partial_x^2 u_H(t,x) \in L_2(I\times\Omega)$, we obtain $g \in L_2(I;H^1(\Omega))$, and the proof is completed. □

By combining the results of the preceding three propositions with the regularity result proven in Proposition 3.1 we obtain the following.

Theorem 3.1 *If* $v \in W^1_\infty(I \times \Omega) \cap L_\infty(I; W^2_\infty(\Omega))$, *then*

$$e(v) \in H^1_{0,\{0\}}(I; H^1_0(\Omega)) \cap H^2(I; H^{-1}(\Omega)) \cap L_2(I; H^3(\Omega)).$$

Proof We obtain $(v - v_0)\partial_x(u(v_0) - u_H) \in L_2(I; H^1(\Omega)) \cap H^1(I; H^{-1}(\Omega))$ from Lemma 3.2 and 3.3, whereas Lemma 3.4 implies that $(v-v_0)\partial_x u_H \in L_2(I; H^1(\Omega)) \cap H^1(I; H^{-1}(\Omega))$. We conclude that

$$(v - v_0)\partial_x u(v_0) \in L_2(I; H^1(\Omega)) \cap H^1(I; H^{-1}(\Omega)),$$

so that an application of Proposition 3.1a) completes the proof. □

Notice that as a consequence of Corollary 3.1, Lemma 3.2 and 3.3, $u(v_0) \notin H^1_{0,\{0\}}(I; L_2(\Omega)) \cup L_2(I; H^2(\Omega))$. Comparing Corollary 3.1 with Theorem 3.1, we conclude that

$$e(v) = u(v) - u(v_0) \textit{ is indeed smoother than } u(v_0), \textit{ and thus than } u(v),$$

confirming the claim we made at the beginning of Section 3.

4 Minimal residual method

For solving (3.3) (specifically for the forcing function f as in (3.6), i.e., for solving $e(v)$), we write it in variational form, i.e., we multiply it by test functions $z : I \times \Omega \to \mathbb{R}$ from a suitable collection, integrate it over $I \times \Omega$, and apply integration by parts with respect to x. We thus arrive at

$$\begin{aligned}&(Bw)(z) :=\\ &\int_I \int_D \partial_t w(t,x)z(t,x) + \partial_x w(t,x)\partial_x z(t,x) - v(t,x)\partial_x w(t,x)z(t,x)\,dx\,dt\\ &= \int_I \int_D f(t,x)z(t,x)\,dx\,dt =: f(z)\end{aligned}$$

for all those test functions. With

$$X := L_2(I; H^1_0(\Omega)) \cap H^1(I; H^{-1}(\Omega)), \quad Y := L_2(I; H^1_0(\Omega)),$$

it is known that $(B, \gamma_0) \in \mathcal{L}_{\text{iso}}(X, Y' \times L_2(\Omega))$, where $\gamma_0 := w \mapsto w(0, \cdot)$ denotes the initial trace operator, see, e.g., [31, Chapter IV] or [27].

Already because $X \neq Y \times L_2(\Omega)$, the well-posed system $(B, \gamma_0)w = (f, 0)$ cannot be discretized by simple Galerkin discretizations. Given a family $(X_h)_{h\in\Delta}$ of finite dimensional subspaces of X, as discrete approximations to w one may consider the minimizers $\operatorname{argmin}_{\bar{w}\in X_h} \|B\bar{w} - f\|^2_{Y'} + \|\gamma_0\bar{w}\|^2_{L_2(\Omega)}$. Since the dual norm $\|\cdot\|_{Y'}$ cannot be evaluated, this approach is not immediately feasible either. Therefore, for

$(Y_h)_{h\in\Delta}$ being a second family of finite dimensional subspaces, now of Y, for $h \in \Delta$ as a discrete approximation from X_h we consider

$$w_h := \operatorname{argmin}_{\bar{w}\in X_h} \|B\bar{w} - f\|^2_{Y_h'} + \|\gamma_0\bar{w}\|^2_{L_2(\Omega)}. \tag{4.1}$$

This minimal residual approach has been studied for general parabolic PDEs in, e.g., [2, 28, 29], where Ω can be a d-dimensional spatial domain for arbitrary $d \geq 1$.

For parabolic differential operators with a possibly asymmetric spatial part, in our setting caused by a non-zero drift function v, in [28, Thm. 3.1] it has been shown that if $X_h \subset Y_h$ and

$$\varrho := \inf_{h\in\Delta} \inf_{0\neq\bar{w}\in X_h} \frac{\|\partial_t\bar{w}\|_{Y_h'}}{\|\partial_t\bar{w}\|_{Y'}} > 0, \tag{4.2}$$

then

$$\|w - w_h\|_X \lesssim \min_{\bar{w}\in X_h} \|w - \bar{w}\|_X, \tag{4.3}$$

where the implied constant in (4.3) depends only on ϱ and an upper bound for $\|v\|_{L_\infty(I\times\Omega)}$, i.e., w_h is a *quasi-best approximation* from X_h with respect to the norm on X.

Remark 4.1 This quasi-optimality result has been demonstrated under the condition that the spatial part of the parabolic differential operator is *coercive* on $H^1_0(\Omega)\times H^1_0(\Omega)$ for a.e. $t \in I$, i.e.,

$$\int_D \eta'\eta' - v(t,\cdot)\eta'\eta\, dx \gtrsim \|\eta\|^2_{H^1(\Omega)} \quad (\eta \in H^1_0(\Omega)),$$

which holds true when $\partial_x v \leq 0$ or $\|v\|_{L_\infty(I\times\Omega)} \sup_{0\neq\eta\in H^1_0(\Omega)} \frac{\|\eta\|_{L_2(\Omega)}}{\|\eta'\|_{L_2(\Omega)}} < 1$, but which might be violated otherwise.

Although this coercivity condition might not be necessary, it can always be enforced by considering $w_\lambda(t,\cdot) := w(t,\cdot)e^{-\lambda t}$, $f_\lambda(t,\cdot) := f(t,\cdot)e^{-\lambda t}$ instead of w and f with λ sufficiently large, see also the proof of Proposition 3.1. By approximating w_λ by the minimal residual method, and by multiplying the obtained approximation by $e^{\lambda t}$, an approximation for w is obtained. Since qualitatively the transformations with $e^{\pm\lambda t}$ do not affect the smoothness of solution or right-hand side, for convenience in the following we pretend that coercivity holds true for (3.3).

As in [28, 29], we equip Y_h in (4.1) with the energy norm

$$\|z\|^2_Y := (A_s z)(z) \quad (z \in Y_h),$$

where

$$(A_s z)(\bar{z}) := \int_I \int_D \partial_x z(t,x)\partial_x\bar{z}(t,x) - \frac{v(t,x)}{2}(\partial_x z(t,x)\bar{z}(t,x) + z(t,x)\partial_x\bar{z}(t,x))\, dx\, dt$$

denotes the symmetric part of the spatial differential operator. Equipping Y_h and X_h with bases $\Phi^h = \{\phi_i^h\}$ and $\Psi^h = \{\psi_j^h\}$, respectively, and denoting by $\boldsymbol{w}^h$ the representation of the minimizer w_h with respect to Ψ_h, $\boldsymbol{w}^h$ is found as the second component of the solution of

$$\begin{bmatrix} \boldsymbol{A}_s^h & \boldsymbol{B}^h \\ \boldsymbol{B}^{h\top} & \boldsymbol{C}^h \end{bmatrix} \begin{bmatrix} \boldsymbol{\mu}^h \\ \boldsymbol{w}^h \end{bmatrix} = \begin{bmatrix} \boldsymbol{f}^h \\ 0 \end{bmatrix}, \tag{4.4}$$

where $(\boldsymbol{A}_s^h)_{ij} := (A_s\phi_j^h)(\phi_i^h)$, $\boldsymbol{B}_{ij}^h := (B\psi_j^h)(\phi_i^h)$, $\boldsymbol{C}_{ij}^h := \int_D \psi_j^h(0,x)\psi_i^h(0,x)\,dx$, and $\boldsymbol{f}_i^h := f(\phi_i^h)$. The operator A_s can be replaced by any other spectrally equivalent operator on Y_h without compromising the quasi-optimality result (4.3). We refer to [28, 29] for details.

Let P_1 be the set of polynomials of degree one. Taking for $n := 1/h \in \mathbb{N}$,

$$\begin{aligned} V_{x,h} &:= \big\{\eta \in H_0^1(\Omega)\colon \eta|_{((i-1)h,ih)} \in P_1 \text{ for } i = 1,\dots,n\big\},\\ V_{t,h} &:= \big\{\zeta \in H^1(I)\colon \zeta|_{((i-1)hT,ihT)} \in P_1 \text{ for } i = 1,\dots,n\big\},\\ X_h &:= V_{t,h} \otimes V_{x,h}, \end{aligned} \tag{4.5}$$

it is known, cf. [29, Sect. 4], that condition (4.2) is satisfied for

$$Y_h := \big\{\zeta \in L_2(I)\colon \zeta|_{((i-1)hT,ihT)} \in P_1 \text{ for } i = 1,\dots,n\big\} \otimes V_{x,h}, \tag{4.6}$$

where obviously also $X_h \subset Y_h$.

Applying this approach for $f = (v - v_0)\partial_x u(v_0)$, in view of (4.3) the error of the obtained approximation for $e(v)$ with respect to the X-norm can be bounded by the error of the best approximation from X_h. To bound the latter error we recall from Theorem 3.1 that for $v \in W_\infty^1(I \times \Omega) \cap L_\infty(I; W_\infty^2(\Omega))$, it holds that

$$e(v) \in \big(H^1_{0,\{0\}}(I) \otimes H_0^1(\Omega)\big) \cap \big(H^2(I) \otimes H^{-1}(\Omega)\big) \cap \big(L_2(I) \otimes H^3(\Omega)\big).$$

With $Q_{x,h}$, $Q_{t,h}$ denoting the $L_2(\Omega)$- or $L_2(I)$-orthogonal projectors onto $V_{x,h}$ or $V_{t,h}$, respectively, $Q_{t,h} \otimes Q_{x,h}$ is a projector onto X_h. Writing

$$\mathrm{Id} - Q_{t,h} \otimes Q_{x,h} = (\mathrm{Id} - Q_{t,h}) \otimes Q_{x,h} + \mathrm{Id} \otimes (\mathrm{Id} - Q_{x,h}),$$

and using that

$$\begin{aligned} &\|\mathrm{Id} - Q_{x,h}\|_{\mathcal{L}(H_0^1(\Omega)\cap H^2(\Omega), H_0^1(\Omega))} \lesssim h, \quad \|Q_{x,h}\|_{\mathcal{L}(H_0^1(\Omega), H_0^1(\Omega))} \lesssim 1,\\ &\|\mathrm{Id} - Q_{t,h}\|_{\mathcal{L}(H^1(I), L_2(I))} \lesssim h, \quad \|\mathrm{Id}\|_{\mathcal{L}(L_2(I), L_2(I))} = 1 \end{aligned}$$

by standard interpolation estimates and uniform H^1-boundedness of these L_2-orthogonal projectors, see e.g. [5, §3], one infers that

$$\|\mathrm{Id} - Q_{t,h} \otimes Q_{x,h}\|_{\mathcal{L}((L_2(I)\otimes(H_0^1(\Omega)\cap H^2(\Omega)))\cap(H^1(I)\otimes H_0^1(\Omega)), L_2(I)\otimes H_0^1(\Omega))} \lesssim h.$$

Similarly using that

$$\begin{aligned}
&\|\mathrm{Id} - Q_{x,h}\|_{\mathcal{L}(L_2(\Omega),H^{-1}(\Omega))} = \|\mathrm{Id} - Q_{x,h}\|_{\mathcal{L}(H_0^1(\Omega),L_2(\Omega))} \lesssim h,\\
&\|Q_{x,h}\|_{\mathcal{L}(H^{-1}(\Omega),H^{-1}(\Omega))} = \|Q_{x,h}\|_{\mathcal{L}(H_0^1(\Omega),H_0^1(\Omega))} \lesssim 1,\\
&\|\mathrm{Id} - Q_{t,h}\|_{\mathcal{L}(H^2(I),H^1(I))} \lesssim h, \quad \|\mathrm{Id}\|_{\mathcal{L}(H^1(I),H^1(I))} = 1,
\end{aligned}$$

one infers that

$$\|\mathrm{Id} - Q_{t,h} \otimes Q_{x,h}\|_{\mathcal{L}((H^1(I)\otimes L_2(\Omega))\cap(H^2(I)\otimes H^{-1}(\Omega)),H^1(I)\otimes H^{-1}(\Omega))} \lesssim h.$$

Our findings are summarized in the following theorem.

Theorem 4.1 *For $v \in W^1_\infty(I \times \Omega) \cap L_\infty(I; W^2_\infty(\Omega))$ and X_h, Y_h as defined in (4.5) and (4.6), the numerical approximation $e_h = e_h(v) \in X_h$ to $e = e(v)$ obtained by the application of the minimal residual method to (3.1)[1] satisfies*

$$\|e - e_h\|_X \lesssim h.$$

Notice that for this space X_h of continuous piecewise bilinears, this linear decay of the error $\|e - e_h\|_X$ as function of h is generally the best that can be expected. In view of the order of the space X_h, one may hope that $\|e - e_h\|_{L_2(I\times\Omega)}$ is $\mathcal{O}(h^2)$, but on the basis of the smoothness demonstrated for e, even for $\inf_{\bar{e}\in X_h} \|e - \bar{e}\|_{L_2(I\times\Omega)}$ this cannot be shown.

5 Interpolation for parametrized drift, boundaries, and final time

In this section we consider the case that v and T in (2.3) depend on a number of parameters $(\rho_1, \ldots, \rho_N) \in [-1, 1]^N$, and that one is interested in the solution $u(v)$ to (2.3) for multiple values of these parameters. As explained in Section 3, in order to find $u(v)$ it suffices to obtain the solution $e(v)$ to (3.1). Instead of simply solving $e(v)$ for each of the desired parameter values, under the provision that v and T depend smoothly on the parameters, one may attempt to *interpolate* $e(v)$ from its a priori computed approximations for a carefully selected set of parameters in $[-1, 1]^N$.

In order to be able to do so, first of all we need to get rid of the parameter dependence of the domain $I \times \Omega = (0, T) \times (0, 1)$. With $\hat{I} := (0, 1)$, the function $\hat{u}$ on $\hat{I} \times \Omega$ defined by $\hat{u}(t, x) := u(tT, x)$ solves

$$\begin{cases}
\partial_t \hat{u}(t,x) = T[\partial_x^2 \hat{u}(t,x) + \hat{v}(t,x)\partial_x \hat{u}(t,x)] & (t,x) \in \hat{I} \times D,\\
\hat{u}(t,0) = 1, \quad \hat{u}(t,1) = 0 & t \in \hat{I},\\
\hat{u}(0,x) = 0 & x \in D,
\end{cases} \tag{5.1}$$

[1] If necessary taking into account the transformations discussed in Remark 4.1.

where analogously $\hat{v}(t,x) := v(tT,x)$. Denoting this $\hat{u}$ as $\hat{u}(\hat{v},T)$, the difference

$$\hat{e} = \hat{e}(\hat{v},T) := \hat{u}(\hat{v},T) - \hat{u}(v_0,T)\colon (t,x) \mapsto e(tT,x)$$

solves

$$\begin{cases} \partial_t \hat{e}(t,x) = T[\partial_x^2 \hat{e}(t,x) + \hat{v}(t,x)\partial_x \hat{e}(t,x)] \\ \qquad\qquad + T(\hat{v}(t,x) - v_0)\partial_x \hat{u}(v_0,T) & (t,x) \in \hat{I} \times \Omega, \\ \hat{e}(t,0) = 0, \quad \hat{e}(t,1) = 0 & t \in \hat{I}, \\ \hat{e}(0,x) = 0 & x \in \Omega. \end{cases} \tag{5.2}$$

By simply replacing $I = (0,T)$ by $\hat{I} = (0,1)$ and in particular X as well as Y by

$$\hat{X} := L_2(\hat{I}; H_0^1(\Omega)) \cap H^1(\hat{I}; H^{-1}(\Omega)), \quad \hat{Y} := L_2(\hat{I}; H_0^1(\Omega)),$$

in a number of places, it is clear that the results that we obtained about the smoothness of e and its numerical approximation e_h by the minimal residual method apply equally well to $\hat{e}$ and its minimal residual approximation that we denote as $\hat{e}_h$.

Since the domain of $\hat{e}$ is independent of parameters, we can apply the idea of interpolation. One option is to perform a 'full' tensor product interpolation. In this case, the number of interpolation points required for a fixed polynomial degree, i.e., the number of values of the parameters for which a numerical approximation for $\hat{e} \in \hat{X}$ has to be computed, grows exponentially with the number N of parameters. As this is undesirable, we instead apply a sparse tensor product interpolation. More specifically, we choose the Smolyak construction, based on Clenshaw–Curtis abscissae in each parameter direction, see [22]: For $i \in \mathbb{N}$ let I_{i+1} denote the univariate interpolation operator with abscissae $\cos j2^{-i}\pi$, $j = 0, \dots, 2^i$, onto the space of polynomials of degree 2^i, let I_1 be the interpolation operator with abscissa 0 and let $I_0 := 0$. Then, for an integer $q \geq N$, we apply the sparse interpolator

$$\mathcal{I}_q := \sum_{\{\mathbf{i} \in \mathbb{N}_0^N \colon \sum_{n=1}^N i_n \leq q\}} \bigotimes_{n=1}^N (I_{i_n} - I_{i_n - 1}).$$

It is known that the resulting interpolation error in $C([-1,1]^N; \hat{X})$ (for arbitrary Banach space $\hat{X}$), equipped with $\|\cdot\|_{L_\infty([-1,1]^N;\hat{X})}$, decays subexponentially in the number of interpolation points when $\hat{e}$ as function of each of the parameters ρ_n has an extension to a differentiable mapping on a neighbourhood Σ of $[-1,1]$ in $\mathbb{C}$. For details about this statement we refer to [22, Thm. 3.11]. [22] also mentions that the result requires relatively large values of q. Thus, the authors additionally prove algebraic convergence under the same assumptions but for arbitrary q [22, Thm. 3.10].

Instead of $\hat{e}$, we interpolate a numerical approximation $\hat{e}_h$, specifically the one obtained by the minimal residual method described in Section 4. For the additional

error we have

$$\|\mathcal{I}_q(\hat{e}-\hat{e}_h)\|_{L_\infty([-1,1]^N;\hat{X})} \le \|\mathcal{I}_q\|_{\mathcal{L}(C([-1,1]^N),C([-1,1]^N))}\|\hat{e}-\hat{e}_h\|_{L_\infty([-1,1]^N;\hat{X})}.$$

In [8, Sect. 5.3] it has been shown that the factor $\|\mathcal{I}_q\|_{\mathcal{L}(C([-1,1]^N),C([-1,1]^N))}$, known as the Lebesgue constant, is bounded by $(\#\{\mathbf{i}\in\mathbb{N}_0^N:\ \sum_{n=1}^N i_n \le q\})^2$, which is only of polylogarithmic order as function of the number of interpolation points.

Concerning the factor $\|\hat{e}-\hat{e}_h\|_{L_\infty([-1,1]^N;\hat{X})}$, in our derivation of Theorem 4.1 we have seen that for each parameter value $(\rho_1,\dots,\rho_N)\in[-1,1]^N$ the expression $h^{-1}\|\hat{e}-\hat{e}_h\|_{\hat{X}}$ can be bounded by a constant multiple, only dependent on an upper bound for $\|\hat{v}\|_{L_\infty(\hat{I}\times\Omega)}$ and for the norm of $\hat{e}$ in $H^1_{0,\{0\}}(\hat{I};H^1_0(\Omega))\cap H^2(\hat{I};H^{-1}(\Omega))\cap L_2(\hat{I};H^2(\Omega))$. For uniformly bounded T and T^{-1}, and $\hat{v}$ that varies over a bounded set in $W^1_\infty(\hat{I}\times\Omega)\cap L_\infty(\hat{I};W^2_\infty(\Omega))$, inspection of the estimates from Sect. 3 reveals that the latter norm of $\hat{e}$ is uniformly bounded. So assuming that these conditions on T, T^{-1} and v hold true for $(\rho_1,\dots,\rho_N)\in[-1,1]^N$, we have that $\|\hat{e}-\hat{e}_h\|_{L_\infty([-1,1]^N;\hat{X})}\lesssim h$.

What remains is to establish the differentiability of the solution $\hat{e}$ as function of each of the parameters which is done in the following theorem.

Theorem 5.1 *For an open $[-1,1]\subset\Sigma\subset\mathbb{C}$, let $(\hat{v},T)\colon\Sigma\to C(\overline{\hat{I}};W^1_\infty(\Omega))\times(0,\infty)$ be differentiable. For $\rho\in\Sigma$ let $\hat{e}(\hat{v}(\rho),T(\rho))\in\hat{X}$ be the solution to (5.2). Then $\rho\mapsto\hat{e}=\hat{e}(\hat{v}(\rho),T(\rho))\colon\Sigma\to\hat{X}$ is differentiable.*

Proof The proof is based on the fact that $\hat{e}$ is the solution of a well-posed PDE with coefficients and a forcing term that are differentiable functions of ρ.

Analogously to (3.4), denoting by $L(\hat{v},T)$ the map $w\mapsto f$ defined by $\partial_t w = T(\partial_x^2+\hat{v}\partial_x)w+f$ on $\hat{I}\times\Omega$, $w(t,0)=0=w(t,1)$ $(t\in\hat{I})$, and $w(0,x)=0$ $(x\in\Omega)$, one has

$$\hat{e}(\hat{v}(\rho),T(\rho)) = L(\hat{v}(\rho),T(\rho))^{-1}T(\rho)(\hat{v}(\rho)-v_0(\rho))\partial_x\hat{u}(v_0(\rho),T(\rho)), \tag{5.3}$$

where $v_0(\rho):=\hat{v}(\rho)(0,0)$. Below we demonstrate that

$$\rho\mapsto T(\rho)(\hat{v}(\rho)-v_0(\rho))\colon\Sigma\to L_\infty(\hat{I};W^1_\infty(\Omega)) \text{ is differentiable,} \tag{5.4}$$

$$\rho\mapsto\hat{u}(v_0(\rho),T(\rho))\colon\Sigma\to L_2(\hat{I}\times\Omega) \text{ is differentiable,} \tag{5.5}$$

so that, from $\partial_x\in\mathcal{L}(L_2(\hat{I}\times\Omega),\hat{Y}')$ and $L_\infty(\hat{I};W^1_\infty(\Omega))$-functions being pointwise multipliers in $\mathcal{L}(\hat{Y}',\hat{Y}')$,

$$\rho\mapsto T(\rho)(\hat{v}(\rho)-v_0(\rho))\partial_x\hat{u}(v_0(\rho),T(\rho))\colon\Sigma\to\hat{Y}' \text{ is differentiable.} \tag{5.6}$$

We proceed below to show that

$$\rho\mapsto L(\hat{v}(\rho),T(\rho))^{-1}\colon\Sigma\to\mathcal{L}(\hat{Y}',\hat{X}) \text{ is differentiable.} \tag{5.7}$$

Together, (5.6) and (5.7) complete the proof.

From $\rho \mapsto \hat{v}(\rho)\colon \Sigma \to C(\overline{\hat{I}}; W^1_\infty(\Omega))$ being differentiable, it follows that $\rho \mapsto v_0(\rho)\colon \Sigma \to \mathbb{C}$ is differentiable, which together with $T\colon \Sigma \to (0,\infty)$ being differentiable shows (5.4).

To show (5.7), we fix some arbitrary $\rho_0 \in \Sigma$, abbreviate $L := L(\hat{v}(\rho), T(\rho))$ as well as $L_0 := L(\hat{v}(\rho_0), T(\rho_0))$ and write

$$L^{-1} = L_0^{-1} + L_0^{-1}[L_0 - L]L_0^{-1} + L^{-1}\{[L_0 - L]L_0^{-1}\}^2.$$

This decomposition and the fact that $L(\hat{v}(\rho), T(\rho))^{-1}$ is bounded in $\mathcal{L}(\hat{Y}', \hat{X})$ for ρ in a neighbourhood of ρ_0 ([27, Thm. 5.1]) imply that it suffices to show that for some $K(\rho_0) \in \mathcal{L}(\mathbb{C}, \mathcal{L}(\hat{X}, \hat{Y}'))$,

$$L(\hat{v}(\rho_0), T(\rho_0)) - L(\hat{v}(\rho), T(\rho)) = K(\rho_0)(\rho - \rho_0) + o(\rho - \rho_0) \text{ in } \mathcal{L}(\hat{X}, \hat{Y}'). \quad (5.8)$$

We have

$$\begin{aligned} &L(\hat{v}(\rho_0), T(\rho_0)) - L(\hat{v}(\rho), T(\rho)) \\ &= [T(\rho) - T(\rho_0)]\partial_x^2 + [(T(\rho) - T(\rho_0))\hat{v}(\rho) + T(\rho_0)(\hat{v}(\rho) - \hat{v}(\rho_0))]\partial_x. \end{aligned}$$

From $T(\rho) - T(\rho_0) = DT(\rho_0)(\rho - \rho_0) + o(\rho - \rho_0)$, $\hat{v}(\rho) - \hat{v}(\rho_0) = D\hat{v}(\rho_0)(\rho - \rho_0) + o(\rho - \rho_0)$ in $C(\overline{I}_1, W^1_\infty(\Omega)) \hookrightarrow L_\infty(\hat{I} \times \Omega)$, $\partial_x^2 \in \mathcal{L}(\hat{X}, \hat{Y}')$, $\partial_x \in \mathcal{L}(\hat{X}, L_2(\hat{I} \times \Omega))$, $L_\infty(\hat{I} \times \Omega)$-functions being pointwise multipliers in $\mathcal{L}(L_2(\hat{I} \times \Omega), L_2(\hat{I} \times \Omega))$, and $L_2(\hat{I} \times \Omega) \hookrightarrow \hat{Y}'$, one concludes (5.8), and so (5.7).

To show (5.5), i.e., differentiability of $\rho \mapsto \hat{u}(v_0(\rho), T(\rho))$, we repeat the argument that led to (5.3) to obtain

$$\begin{aligned} \hat{u}(v_0(\rho), T(\rho)) &= \hat{u}(0, T(\rho)) + \hat{u}(v_0(\rho), T(\rho)) - \hat{u}(0, T(\rho)) \\ &= \hat{u}(0, T(\rho)) + T(\rho)v_0(\rho)L(v_0(\rho), T(\rho))^{-1}\partial_x \hat{u}(0, T(\rho)), \end{aligned}$$

and show that

$$\rho \mapsto \hat{u}(0, T(\rho))\colon \Sigma \mapsto L_2(\hat{I} \times \Omega) \text{ is differentiable.} \quad (5.9)$$

Then $\rho \mapsto \partial_x \hat{u}(0, T(\rho))\colon \Sigma \mapsto \hat{Y}'$ is differentiable, and from both $\rho \mapsto T(\rho)v_0(\rho)\colon \Sigma \to \mathbb{C}$ and $\rho \mapsto L(v_0(\rho), T(\rho))^{-1}\colon \Sigma \to \mathcal{L}(\hat{Y}', \hat{X})$ being differentiable one infers (5.5).

To show (5.9), we apply our approach for the third time. Picking some $\bar{\rho} \in \Sigma$, we write

$$\hat{u}(0, T(\rho)) = \hat{u}(0, T(\bar{\rho})) + (T(\bar{\rho}) - T(\rho))L(0, T(\rho))^{-1}\partial_x^2 \hat{u}(0, T(\bar{\rho})).$$

Knowing that $\partial_x^2 \hat{u}(0, T(\bar{\rho})) \in \hat{Y}'$, and $\rho \mapsto L(0, T(\rho))^{-1}\colon \Sigma \to \mathcal{L}(\hat{Y}', \hat{X})$ and $\rho \mapsto T(\rho)\colon \Sigma \to \mathbb{C}$ are differentiable, the proof of (5.9) and thus of the theorem is completed. □

6 Numerical results

We consider three relevant examples of the form (1.3) (or its equivalent reformulation (2.1)) with $\sigma = 1$ from the literature. We transform the solution $\tilde{u}$ of (2.1), which might live on a time-dependent spatial domain, to u, which satisfies (2.3) on the domain $(0, T) \times (0, 1)$. In each example the resulting drift function v as well as the end time point T depend on an up to $N = 5$ dimensional parameter $\boldsymbol{\rho} \in [-1, 1]^N$.

As $u(v(\boldsymbol{\rho})(0, 0), T(\boldsymbol{\rho}))$ can be computed efficiently as a truncated series, it suffices to consider the difference

$$e(v(\boldsymbol{\rho}), T(\boldsymbol{\rho})) = u(v(\boldsymbol{\rho}), T(\boldsymbol{\rho})) - u(v(\boldsymbol{\rho})(0, 0), T(\boldsymbol{\rho})),$$

which satisfies equation (3.1) and is provably smoother than u (Theorem 3.1).

Thinking of a multi-query setting, instead of approximating this difference for each individual parameter value of interest we want to use (sparse) interpolation in the parameter domain $[-1, 1]^N$. To that end, defining $\hat{e}(t, x) := e(tT(\boldsymbol{\rho}), x)$, we get rid of the parameter-dependent domain $(0, T(\boldsymbol{\rho})) \times \Omega$ on which e lives. This function $\hat{e}(t, x)$ satisfies the parabolic problem equation (5.2) on the space-time domain $\hat{I} \times \Omega = (0, 1)^2$ with forcing term

$$\begin{aligned}\bar{w} \mapsto & \int_0^1 \int_D (\hat{v}(t, x) - v_0)\partial_x \hat{u}(v_0)(t, x)\bar{w}(t, x)\, dx\, dt \\ = & \int_0^1 \int_D \hat{u}(v_0)(t, x)\big(- \partial_x \hat{v}(t, x)\bar{w}(t, x) - (\hat{v}(t, x) - v_0)\partial_x \bar{w}(t, x)\big)\, dx\, dt\end{aligned}$$

for all $\bar{w} \in \hat{X} = L_2(\hat{I}; H_0^1(\Omega)) \cap H^1(\hat{I}; H^{-1}(\Omega))$, and $v_0 := v(\boldsymbol{\rho})(0, 0)$ and corresponding $\hat{u}(v_0)$ solving (5.1) with $\hat{v} = v_0$.

For all sparse interpolation points, by applying the minimal residual method from Section 4 we approximate $\hat{e}$ by the continuous piecewise affine function $\hat{e}_h$ on a uniform tensor mesh with mesh-size h, where $\hat{u}(v_0)$ inside the forcing term can be efficiently approximated at high accuracy as a truncated series.

Finally, for all parameter values $\boldsymbol{\rho}$ of interest, we apply the sparse tensor product interpolation analyzed in Section 5 giving rise to an overall error

$$\|\hat{e} - \mathcal{I}_q \hat{e}_h\|_{\hat{X}} \le \|\hat{e} - \hat{e}_h\|_{\hat{X}} + \|\hat{e}_h - \mathcal{I}_q \hat{e}_h\|_{\hat{X}} \approx \|\hat{e}_{h/2} - \hat{e}_h\|_{\hat{X}} + \|\hat{e}_h - \mathcal{I}_q \hat{e}_h\|_{\hat{X}}$$

with q the parameter that steers the accuracy of the sparse interpolation. For each of the considered three examples, we compute the latter two errors for different h and q and parameter test set

$$\boldsymbol{\rho} \in \{-1, -0.5, 0.5, 1\}^N. \tag{6.1}$$

By Theorem 4.1, we expect $\|\hat{e}_{h/2} - \hat{e}_h\|_{\hat{X}} = \mathcal{O}(h)$ for the first term. Section 5 suggests subexponential convergence of the second term $\|\hat{e}_h - \mathcal{I}_q \hat{e}_h\|_{\hat{X}}$ as function of the number of interpolation points (this was shown for $\|\hat{e} - \mathcal{I}_q \hat{e}\|_{\hat{X}}$). However, we already mentioned there that subexponential convergence is only observed for very high q and in practice one should rather expect algebraic convergence.

Notice that $\|\cdot\|_{\hat{X}}$ involves a negative order Sobolev norm. Thus, we compute an equivalent version of $\|\cdot\|_{\hat{X}}$ for functions in the discrete trial space $\bar{w} \in \hat{X}_h \subset \hat{X}$ (similarly for $\bar{w} \in \hat{X}_{h/2}$) (see [28, Proof of Thm. 3.1])

$$\|\bar{w}\|_{\hat{X}}^2 \eqsim (\bar{\boldsymbol{w}}^h)^\top (\boldsymbol{B}^h)^\top (\boldsymbol{A}^h)^{-1} \boldsymbol{B}^h \bar{\boldsymbol{w}}^h + (\bar{\boldsymbol{w}}^h)^\top \boldsymbol{C}^h \bar{\boldsymbol{w}}^h. \tag{6.2}$$

Here, $\bar{\boldsymbol{w}}^h$ is the coefficient vector of $\bar{w}$ in the standard nodal basis $\Psi^h = \{\psi_i^h\}$, $\boldsymbol{B}^h$ and $\boldsymbol{C}^h$ are defined as in (4.4) with the standard nodal basis $\Phi^h = \{\phi_i^h\}$, and $\boldsymbol{A}_{ij}^h := \int_I \int_\Omega \partial_x \phi_j^h(t,x) \partial_x \phi_i^h(t,x)\, dx\, dt$.

6.1 Time-dependent hyperbolic drift function

As in [9, 19], we consider

$$\mu(t,x) := \mu_0 + \mu_1 \frac{t}{t+t_0}$$

from Section 1 with parameters $\mu_0, \mu_1 \in \mathbb{R}$ and $t_0 > 0$. The left and right boundary are given as

$$\alpha(t) := 0 \quad \text{and} \quad \beta(t) := \beta_0$$

with parameter $\beta_0 > 0$. Following [9, 19], we particularly consider the following practical ranges: $\mu_0 \in [-1.97, -1.64]$, $\mu_1 \in [-2.31, -0.99]$, $t_0 \in [0.13, 0.40]$, $\beta_0 \in [1.38, 2.26]$, and $\tau \in [0.1, 2.5]$ for the end-time point. We have $N = 5$ different parameters on which $\tilde{v}$ and thus v depend. After rescaling, the parameter space hence has the form $[-1, 1]^5$.

In Figure 1, we plot the maximal error $\hat{e}_{h/2} - \hat{e}_h \approx \hat{e} - \hat{e}_h$ measured in the (equivalent) $\hat{X}$-norm (6.2) over the test set (6.1) for different values of h. Figure 2 depicts the maximal interpolation error $\hat{e}_h - \mathcal{I}_q \hat{e}_h$ over the test set (6.1) for different values of h and q.

6.2 Space-dependent linear drift function

As in [26], we consider

$$\mu(t,x) := \mu_0 + \mu_1(\beta_0 - x)$$

from Section 1 with parameters $\beta_0 > 0$ and $\mu_0, \mu_1 \in \mathbb{R}$. The left and right boundary are again given as

$$\alpha(t) := 0 \quad \text{and} \quad \beta(t) := \beta_0.$$

Motivated by [21, 26], we particularly consider the following practical ranges: $\mu_0 \in [-2, 2]$, $\mu_1 \in [-4, 4]$, and $\beta_0 \in [0.5, 2]$, and choose the end-time point as $\tau := 2.5$.

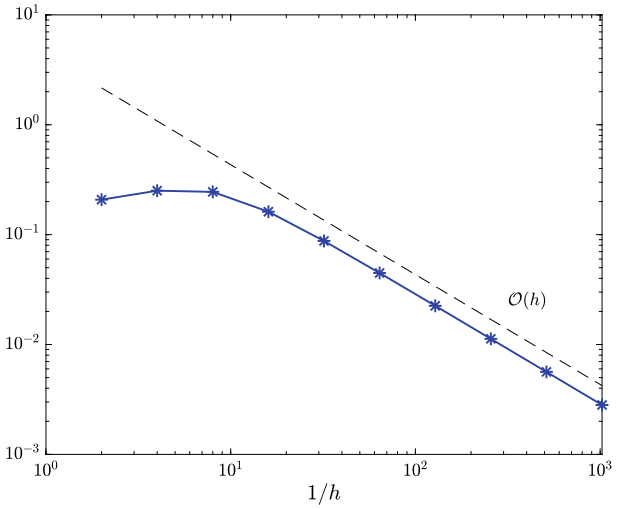

Fig. 1 Maximal error $\hat{e}_{h/2}(\hat{v}(\boldsymbol{\rho})) - \hat{e}_h(\hat{v}(\boldsymbol{\rho}))$ measured in (equivalent) $\hat{X}$-norm over all $\boldsymbol{\rho} \in \{-1, -0.5, 0, 0.5, 1\}^5$ for time-dependent hyperbolic drift function from Sect. 6.1

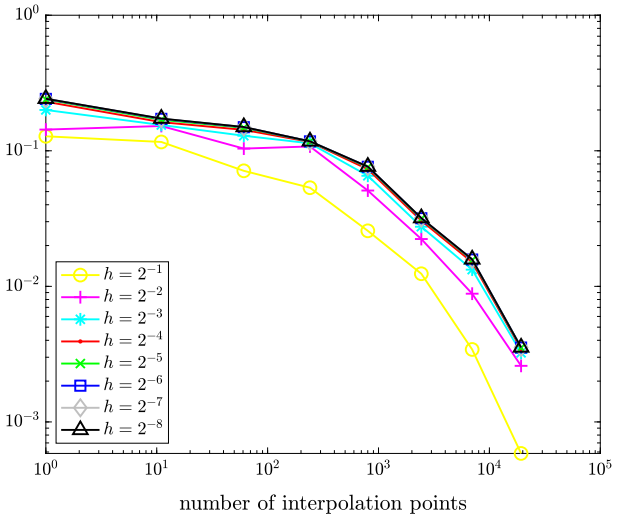

Fig. 2 Maximal interpolation error $\hat{e}_h(\hat{v}(\boldsymbol{\rho})) - (\mathcal{I}_q \hat{e}_h(\hat{v}(\cdot)))(\boldsymbol{\rho})$ for various choices of h measured in (equivalent) $\hat{X}$-norm over all $\boldsymbol{\rho} \in \{-1, -0.5, 0, 0.5, 1\}^5$ for time-dependent hyperbolic drift function from Sect. 6.1

We have $N = 3$ different parameters on which $\tilde{v}$ and thus v depend. After rescaling, the parameter space hence has the form $[-1, 1]^3$.

In Figure 3, we plot the maximal error $\hat{e}_{h/2} - \hat{e}_h \approx \hat{e} - \hat{e}_h$ measured in the (equivalent) $\hat{X}$-norm (6.2) over the test set (6.1). Figure 4 depicts the maximal interpolation error $\hat{e}_h - \mathcal{I}_q \hat{e}_h$ over the test set (6.1) for different values of h and q.

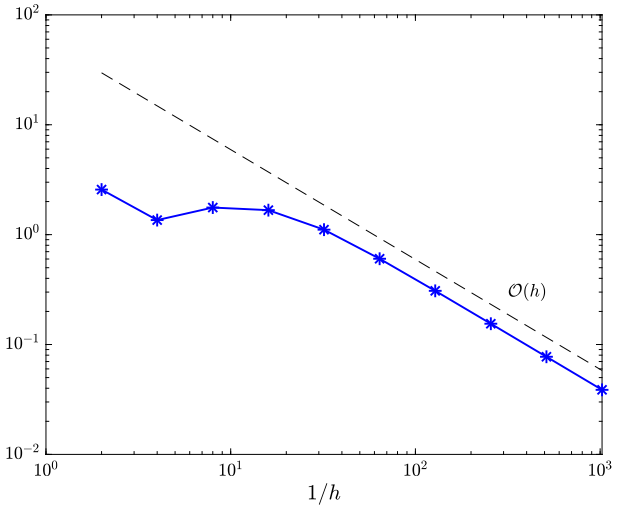

Fig. 3 Maximal error $\hat{e}_{h/2}(\hat{v}(\boldsymbol{\rho})) - \hat{e}_h(\hat{v}(\boldsymbol{\rho}))$ measured in (equivalent) $\hat{X}$-norm over all $\boldsymbol{\rho} \in \{-1, -0.5, 0, 0.5, 1\}^3$ for space-dependent linear drift function from Sect. 6.2

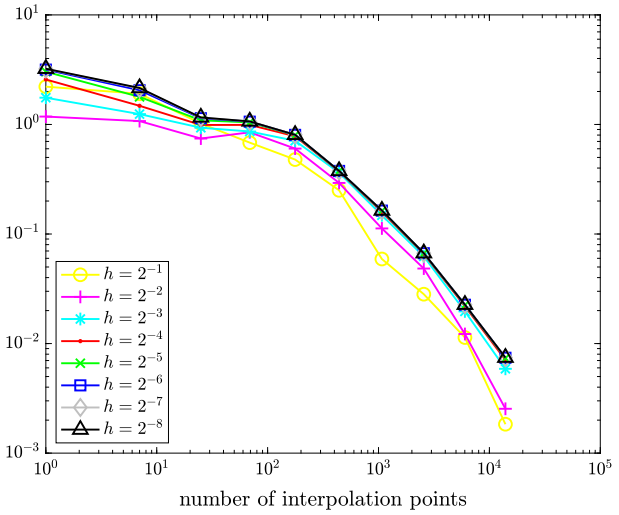

Fig. 4 Maximal interpolation error $\hat{e}_h(\hat{v}(\boldsymbol{\rho})) - (\mathcal{I}_q \hat{e}_h(\hat{v}(\cdot)))(\boldsymbol{\rho})$ with $h = 2^{-1}, \ldots, 2^{-8}$ measured in (equivalent) $\hat{X}$-norm over all $\boldsymbol{\rho} \in \{-1, -0.5, 0, 0.5, 1\}^3$ for space-dependent linear drift function from Sect. 6.2

6.3 Constant drift function and time-dependent linear spatial domain

We consider a constant drift function

$$\mu(t, x) := \mu_0$$

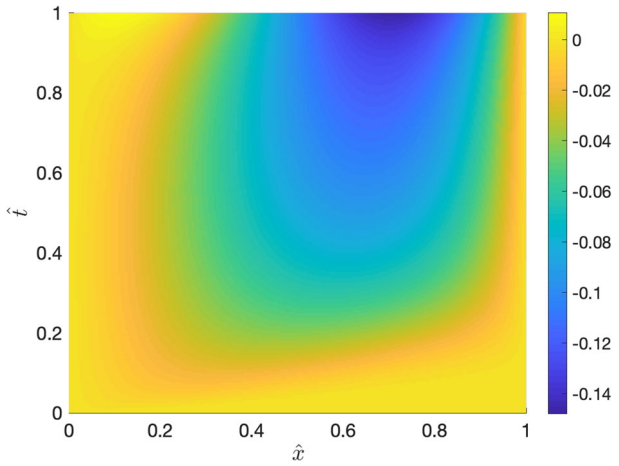

Fig. 5 An approximation of solution $\hat{e}$ to (5.2) using the minimal residual method

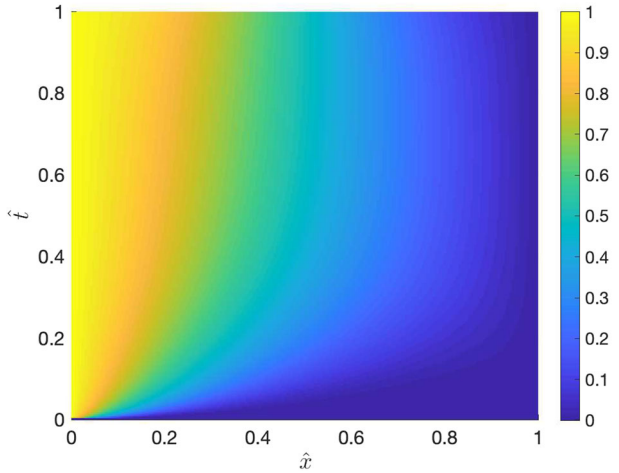

Fig. 6 An approximation of the solution $\hat{u}$ to (5.1) obtained by adding $\hat{u}(v_0, T)$ to $\hat{e}$

with parameter $\mu_0 \in \mathbb{R}$. As in [13] (see also Example 2.1), we choose the left and right boundary as

$$\alpha(t) := \beta_0 \frac{t}{2T_0} \quad \text{and} \quad \beta(t) := \beta_0 \Big(1 - \frac{t}{2T_0}\Big)$$

with parameters $\beta_0, T_0 > 0$. Recall from Example 2.1 that

$$\theta(t) = \frac{\beta_0^2 (T_0 - 2\widetilde{T})^2 t}{T_0^2 - 2\beta_0^2 (T_0 - 2\widetilde{T}) t}, \quad t \in [0, T)$$

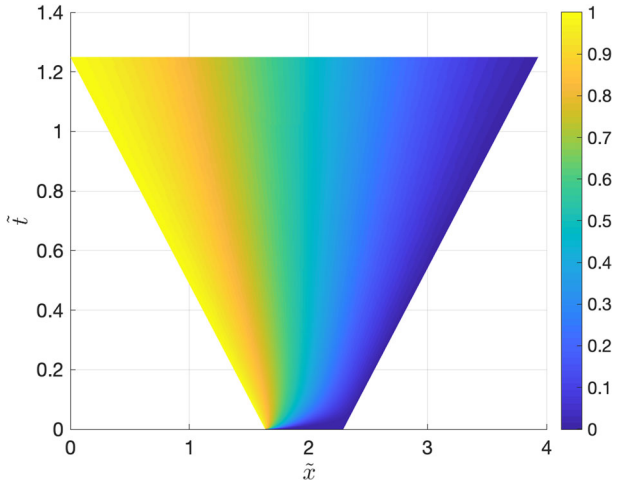

Fig. 7 An approximation of the solution $\tilde{u}$ to (2.1) obtained by transforming the approximation of $\hat{u}$ back to the original domain

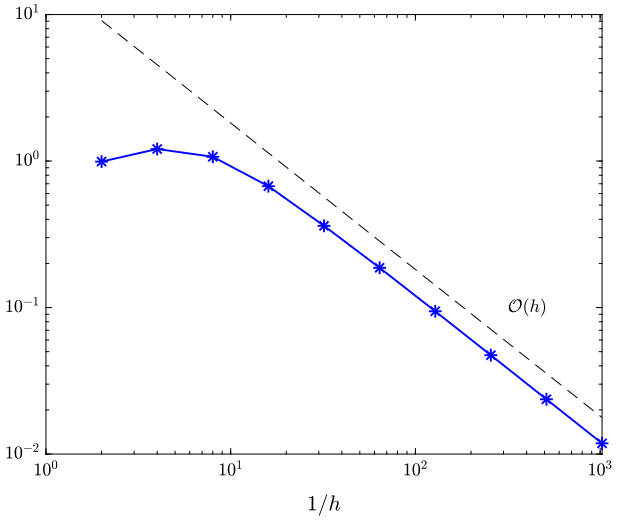

Fig. 8 Maximal error $\hat{e}_{h/2}(\hat{v}(\boldsymbol{\rho}), T(\boldsymbol{\rho})) - \hat{e}_h(\hat{v}(\boldsymbol{\rho}), T(\boldsymbol{\rho}))$ measured in (equivalent) $\hat{X}$-norm over all $\boldsymbol{\rho} \in \{-1, -0.5, 0, 0.5, 1\}^4$ for constant drift function with time-dependent linear spatial domain from Sect. 6.3

with $T = \theta^{-1}(\widetilde{T}) = \frac{T_0\widetilde{T}}{\beta_0^2(T_0-2\widetilde{T})}$. Following [13], we particularly consider the following practical ranges: $\mu_0 \in [-5.86, 0]$, $\beta_0 \in [0.56, 3.93]$, $T_0 \in [3, 20]$, and $\tau \in [0.1, 2.5]$ for the end-time point. We have $N = 4$ different parameters on which $\tilde{v}$ and thus v depend. After rescaling, the parameter space hence has the form $[-1, 1]^4$. Figures 5, 6, and 7 show approximations of the solution $\hat{e}$ to (5.2), the solution $\hat{u}$ to (5.1), and the solution $\tilde{u}$ to the original problem (2.1), with parameter values $\mu_0 = 0$, $\beta_0 = 3.93$, $T_0 = 3$, and $\tau = 2.5$. In Figure 8, we plot the maximal error $\hat{e}_{h/2} - \hat{e}_h \approx \hat{e} - \hat{e}_h$

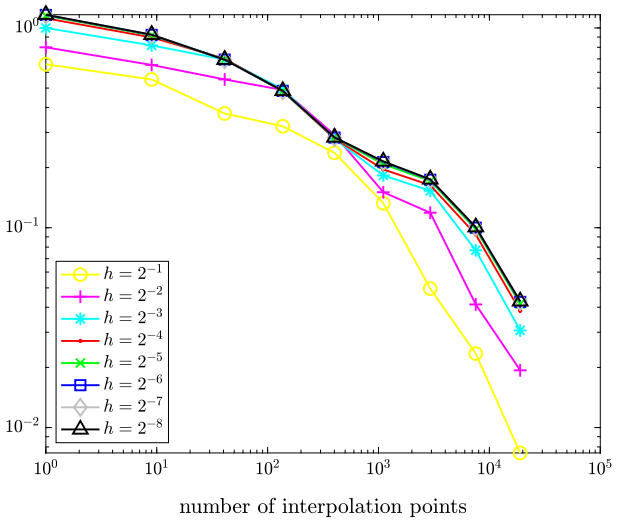

Fig. 9 Maximal interpolation error $\hat{e}_h(\hat{v}(\boldsymbol{\rho}), T(\boldsymbol{\rho})) - (\mathcal{I}_q \hat{e}_h(\hat{v}(\cdot), T(\cdot)))(\boldsymbol{\rho})$ for various choices of h measured in (equivalent) $\hat{X}$-norm over all $\boldsymbol{\rho} \in \{-1, -0.5, 0, 0.5, 1\}^4$ for constant drift function with time-dependent linear spatial domain from Sect. 6.3

measured in the (equivalent) $\hat{X}$-norm (6.2) over the test set (6.1). Figure 9 depicts the maximal interpolation error $\hat{e}_h - \mathcal{I}_q \hat{e}_h$ over the test set (6.1) for different values of h and q.

7 Conclusion

We have developed a numerical solution method for solving the Fokker–Planck equation on a one-dimensional spatial domain and with a discontinuity between initial and boundary data and time-dependent boundaries. We first transformed the equation to an equation on a rectangular time-space domain. We then demonstrated that the solution of a corresponding equation with a suitable constant drift function, whose solution is explicitly available as a fast converging series expansion, captures the main singularity present in the solution for a variable drift function. The equation for the difference of both these solutions, which is thus more regular than both terms, is solved with a minimal residual method. This method is known to give a quasi-best approximation from the selected trial space.

Finally, in order to efficiently solve Fokker–Planck equations that depend on multiple parameters, we demonstrate that the solution is a holomorphic function of these parameters. Consequently, a sparse tensor product interpolation method can be shown to converge at a subexponentional rate as function of the number of interpolation points. In one test example, this interpolation method works very satisfactory, but the results are less convincing in two other cases. We envisage that in those cases better

results can be obtained by an adaptive sparse interpolation method as the one proposed in [6].

Funding Open access funding provided by TU Wien (TUW).

References

1. Artime, O., Khalil, N., Toral, R., San Miguel, M.: First-passage distributions for the one-dimensional Fokker-Planck equation. Phys. Rev. E **98**(4), 042143 (2018)
2. Andreev, R.: Stability of sparse space-time finite element discretizations of linear parabolic evolution equations. IMA J. Numer. Anal. **33**(1), 242–260 (2013)
3. Boehm, U., Cox, S., Gantner, G., Stevenson, R.: Fast solutions for the first-passage distribution of diffusion models with space-time-dependent drift functions and time-dependent boundaries. J. Math. Psych. **105**, 102613 (2021)
4. Bowman, N.E., Kording, K.P., Gottfried, J.A.: Temporal integration of olfactory perceptual evidence in human orbitofrontal cortex. Neuron **75**, 916–927 (2012)
5. Bramble, J.H., Xu, J.: Some estimates for a weighted L^2 projection. Math. Comp. **56**, 463–476 (1991)
6. Chkifa, A., Cohen, A., Schwab, Ch.: High-dimensional adaptive sparse polynomial interpolation and applications to parametric PDEs. Found. Comput. Math. **14**(4), 601–633 (2014)
7. Chandrasekhar, Subrahmanyan: Dynamical friction. I. General considerations: The coefficient of dynamical friction. Astrophys. J. **97**, 255–262 (1943)
8. Chkifa, A.: Sparse polynomial methods in high dimension: Application to parametric PDE, Ph.D. thesis, Université Pierre et Marie Curie - Paris VI, (2014)
9. Churchland, A.K., Kiani, R., Shadlen, M.N.: Decision-making with multiple alternatives. Nat. Neurosci. **11**(6), 693–702 (2008)
10. Costabel, M.: Boundary integral operators for the heat equation. Integr. Equ. Op. Theor. **13**(4), 498–552 (1990)
11. Denk, R., Hieber, M., Prüss, J.: $\mathcal{R}$-boundedness, Fourier multipliers and problems of elliptic and parabolic type, Mem. Amer. Math. Soc. **166**(788), viii+114 (2003)
12. de Simon, L.: Un'applicazione della teoria degli integrali singolari allo studio delle equazioni differenziali lineari astratte del primo ordine. Rend. Sem. Mat. Univ. Padova **34**, 205–223 (1964)
13. Evans, N.J., Trueblood, J.S., Holmes, W.R.: A parameter recovery assessment of time-variant models of decision-making. Behav. Res. Meth. **52**, 193–206 (2020)
14. Flyer, N., Fornberg, B.: Accurate numerical resolution of transients in initial-boundary value problems for the heat equation. J. Comput. Phys. **184**(2), 526–539 (2003)
15. Fengler, A., Frank, M., Govindarajan, L., Chen, T.: Likelihood Approximation Networks (LANs) for fast inference of simulation models in cognitive neuroscience. eLife **10**, e65074 (2021)
16. Gondan, M., Blurton, S.P., Kesselmeier, M.: Even faster and even more accurate first-passage time densities and distributions for the Wiener diffusion model. J. Math. Psych. **60**, 20–22 (2014)
17. Gold, J.I., Shadlen, M.N.: Neural computations that underlie decisions about sensory stimuli. Trends Cognit. Sci. **5**(1), 10–16 (2001)
18. Hawkins, G.E., Forstmann, B.U., Wagenmakers, E.-J., Ratcliff, R., Brown, S.D.: Revisiting the evidence for collapsing boundaries and urgency signals in perceptual decision-making. J. Neurosci. **35**(6), 2476–2484 (2015)
19. Hanks, T., Kiani, R., Shadlen, M.N.: A neural mechanism of speed-accuracy tradeoff in macaque area LIP. eLife **3**, e02260 (2014)

20. Holcman, D., Schuss, Z.: Stochastic narrow escape in molecular and cellular biology, vol. 48. Springer, New York (2015)
21. Matzke, D., Wagenmakers, E.J.: Psychological interpretation of the ex-Gaussian and shifted Wald parameters: A diffusion model analysis. Psychon. Bull. Rev. **16**(5), 798–817 (2009)
22. Nobile, F., Tempone, R., Webster, C.G.: A sparse grid stochastic collocation method for partial differential equations with random input data. SIAM J. Numer. Anal. **46**(5), 2309–2345 (2008)
23. Øksendal, B.: Stoch. Differ. Equ., 5th edn. Springer, Berlin (1998)
24. Ratcliff, R.: A theory of memory retrieval. Psychol. Rev. **85**(2), 59–108 (1978)
25. Michael, N.: Shadlen and Roozbeh Kiani, Decision making as a window on cognition. Neuron **80**(3), 791–806 (2013)
26. Smith, P.L.: From Poisson shot noise to the integrated Ornstein-Uhlenbeck process: Neurally principled models of information accumulation in decision-making and response time. J. Math. Psych. **54**, 266–283 (2010)
27. Schwab, Ch., Stevenson, R.P.: A space-time adaptive wavelet method for parabolic evolution problems. Math. Comp. **78**, 1293–1318 (2009)
28. Stevenson, R.P., Westerdiep, J.: Minimal residual space-time discretizations of parabolic equations: Asymmetric spatial operators, . Comput. Math. Appl. **101**, 107–118 (2021)
29. Stevenson, R.P., Westerdiep, J.: Stability of Galerkin discretizations of a mixed space-time variational formulation of parabolic evolution equations. IMA J. Numer. Anal. **41**(1), 28–47 (2021)
30. Voss, A., Voss, J.: A fast numerical algorithm for the estimation of diffusion model parameters. J. Math. Psych. **52**(52), 1–9 (2008)
31. Wloka, J.: Partial differential equations. Cambridge University Press, Cambridge (1987)