Numerical Approximation of Operator Riccati Equations for Distributed Control of SPDE

Master Thesis

Leander Philipp Schroer
8th August 2016

Master Thesis
conducted under supervision of Dr. Adam Andersson and Prof. Dr. Etienne Emmrich
at the Institute for Mathematics of TU Berlin

Acknowledgements

A year has passed since my first encounter with the operator Riccati equation. Now, having almost completed my master thesis about it, I look back almost bemused about my first attempts to make sense of it. "Well, that wasn't that hard, was it?" is my usual reaction to the completion of a proof, regardless of the effort I had to put in to it. I believe this is a sensation I share with many mathematicians. This time however I realize how much I learned; About control theory, operator equations, PDE theory, computing and more, many things of which I had little to no knowledge about before. For this I very much want to thank Adam Andersson, who has been my supervisor for this project. He always found time and unswerving patience to discuss every question I had. Thank you for assigning me to this exciting topic. I always enjoyed working under your supervision, especially in the last months, which have been intensive but very fruitful.

I thank Etienne Emmrich and Raphael Kruse, without whom this project would not have been possible. I always felt welcome and communication was always pleasantly uncomplicated in the research group pf differential equations, this I never took for granted. I also want to point out the role Raphael Kruse's lectures played, when I decided to seek the opportunity to write my thesis in this field.

For having great laughs, heated discussions and overall merry company, I thank my friends, old and new. Especially Stefan Fortmeier, who was so kind as to help me with my English.

Finally I want to thank my family, who have been the unshakable foundation of my life and my safe harbour in windy times. There is no analogy which comes close to describing how much you mean to me. Therefore I want to keep it simple. I thank my brothers Adrian Schroer and Benjamin Trendelkamp-Schroer for countless hours, well spent on blatantly unproductive activities. I thank my parents Bernhard Schroer and Susanne Schroer for their integral support in my pursuit for a degree in mathematics and enabling me to engage in all the other adventures that would not have been possible without them.

Contents

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und eigenhändig sowie ohne unerlaubte fremde Hilfe und ausschließlich unter Verwendung der aufgeführten Quellen und Hilfsmittel angefertigt habe.

Die selbständige und eigenständige Anfertigung versichert an Eides statt:

Berlin, den 8. August 2016

_____

Introduction

Control theory is a sub-field of mathematical optimization. It concerns itself with the influence of a so called *control function* on a dynamic system. Therefore control theory has many applications in the natural sciences and engineering. A classical and rather old example from mechanics is the inverted pendulum problem. Examples of an inverted pendulum problem are the famous Segway or the first stage of the recent SpaceX Falcon 9 rocket. This problem belongs to a special class of optimal control problems, which are called linear quadratic (LQ) control problems. These problems are characterized by a controlled dynamical system, which is a linear differential equation, and the aim of minimizing a related quadratic cost functional which penalizes the control and the dynamics.

In this thesis we are interested in the following LQ-problem. We consider the stochastic heat equation with linear multiplicative noise and Dirichlet boundary condition. Let $H := L^2([0,1])$ be the space of square integrable functions on $[0,1]$ and $\Delta = \frac{\partial^2}{\partial x^2}$ be the Laplace operator. The stochastic heat equation is given by the following stochastic partial differential equation.

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}t}X(t,\xi) = \Delta X(t,\xi) + X(t,\xi)\cdot\dot{W}(t,\xi) \quad \forall t \in (0,T],\ \xi \in (0,1) \\ &X(t,0) = X(t,1) = 0 \quad \forall t \in (0,T], \\ &X(0) = x_0 \in H, \end{aligned} \tag{0.1}$$

where $x_0 \in H$ is the initial value and $\dot{W}$ represents space-time white noise on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$. We introduce a control $u : [0,T]\times\Omega \to H$ to this system. The controlled stochastic heat equation reads:

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}t}X^u(t,\xi) = \Delta X^u(t) + u(t) + X^u(t,\xi)\cdot\dot{W}(t,\xi), \quad \forall t \in (0,T],\ \xi \in (0,1) \\ &X^u(t,0) = X^u(t,1) = 0 \quad \forall t \in [0,T], \\ &X^u(0) = x_0 \in H. \end{aligned} \tag{0.2}$$

For this type of problem, when the control enters the interior of the domain, we speak of a distributed control problem. Another type of control is the boundary control, where the control influences the boundary conditions. The goal is to find a predictable square integrable optimal control $u \in L^2([0,T]\times\Omega; H)$, which minimizes the quadratic cost functional:

$$J(u) = \mathbb{E}\left[\int_0^T (\|X^u(t)\|^2 + \|u(t)\|^2)\mathrm{d}t + \|X^u(T)\|^2\right]. \tag{0.3}$$

The solution to this distributed linear quadratic control problem is well known. Let $\mathcal{L}(H)$ be the space of bounded linear operators on H and $P : [0,T] \to \mathcal{L}(H)$ an operator valued function. The optimal control is given by

$$\bar{u}(t) := -P(T-t)X^{\bar{u}}(t), \quad t \in [0,T], \tag{0.4}$$

where P solves a *variational operator Riccati equation*. Let $H_0^1 \subset H$ for the Sobolev space of square-integrable functions on $[0,1]$ with existing first order weak derivative and zero boundary conditions. We use the Notation $a(x,y) := \langle \nabla x, \nabla y \rangle_H = -\langle \Delta x, y \rangle_H$ for $x, y \in H_0^1$. Find $(P(t))_{t\in(0,T]} \subset \mathcal{L}(H)$ with $P(0) = \mathrm{id}_H$, such that

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle P(t)x, y\rangle_H + a\left(P(t)x, y\right) + a\left(P(t)y, x\right) + \langle P(t)x, P(t)y\rangle_H \\ = \langle P(t)Cx, Cy\rangle_{\mathcal{L}_2(H)} + \langle x, y\rangle_H, \quad \forall x, y \in H_0^1,\ t \in (0,T]. \tag{0.5}$$

Here $\mathcal{L}_2(H)$ denotes the Hilbert space of all Hilbert-Schmidt operators and C is the multiplication operator $(C(\phi)\psi)(x) = \phi(x)\cdot\psi(x),\ \phi, \psi \in H$.

The goal of this thesis is to study two semi-explicit numerical schemes, which approximate this function P and subsequently the optimal control function $\bar{u}$. This is vital to engineering applications. We embed this example into a more general framework to study the order of convergence of both schemes. The theoretical foundation of this thesis is based on a paper by Andersson, Djehiche and Larsson [1], which established the existence and uniqueness theory for a more general operator Riccati equation.

Let $k \in (0,1]$ be the temporal step size and $h \in (0,1]$ be the spatial step size. Let H^2 denote the Sobolev space of square-integrable functions on $[0,1]$ with existing weak derivatives up to second order. Let $V_h^1 \subset H_0^1$ be the finite element space of continuous piecewise linear functions, and $V_h^2 \subset H^2 \cap H_0^1$ the finite element space of continuous piecewise quadratic functions, both with respect to a triangulation with maximal mesh size h. Let $\Pi_h^i, i \in \{1,2\}$ denote the orthogonal projection of H onto V_h^i. The first scheme reads: Find a sequence $(P_{h,k}^j)_{j=0}^{N_k} \subset L(V_h^1)$ with $P_{h,k}^0 = \mathrm{id}_{V_h^1}$ such that

$$\begin{aligned}&\left\langle P_{h,k}^j\phi, \psi\right\rangle + ka\big(P_{h,k}^j\phi, \psi\big) + ka\big(P_{h,k}^j\psi, \phi\big) + k\big\langle P_{h,k}^{j-1}\phi, P_{h,k}^{j-1}\psi\big\rangle \\ &= \left\langle P_{h,k}^{j-1}\phi, \psi\right\rangle + k\big\langle P_{h,k}^{j-1}\Pi_h C\phi, C\psi\big\rangle_{\mathcal{L}_2(\mathcal{H};H)} + k\big\langle\phi, \psi\big\rangle, \quad \forall\phi, \psi \in V_h^1,\ j \in \{1,\dots,N_k\}.\end{aligned} \tag{0.6}$$

The second scheme reads: Find a sequence $(P_{h,k}^j)_{j=0}^{N_k} \subset L(V_h^1)$ with $P_{h,k}^0 = \mathrm{id}_{V_h^2}$ such that

$$\begin{aligned}&\left\langle P_{h,k}^j\phi, \psi\right\rangle + ka\big(P_{h,k}^j\phi, \psi\big) + ka\big(P_{h,k}^j\psi, \phi\big) + k\big\langle P_{h,k}^{j-1}\phi, P_{h,k}^{j-1}\psi\big\rangle + k^2\langle P_{h,k}^j\Pi_h\Delta\phi, \Delta\psi\rangle \\ &\quad= \left\langle P_{h,k}^{j-1}\phi, \psi\right\rangle + k\big\langle P_{h,k}^{j-1}\Pi_h C\phi, C\psi\big\rangle_{\mathcal{L}_2(\mathcal{H};H)} + k\big\langle Q\phi, Q\psi\big\rangle, \quad \forall\phi, \psi \in V_h^2,\ j \in \{1,\dots,N_k\}.\end{aligned} \tag{0.7}$$

We show that the discrete solutions to both schemes exist in a suitable sense, and are positive and bounded in every time-step. In order to show this, we assume a rather strict grid condition for the first scheme and mild assumptions for the second scheme. We further show that both schemes have an a-priori order of convergence of some $\gamma \in (0, 1/2)$ and a singularity in $t = 0$ of some order $\delta \in (\frac{1}{2}, 1)$. In both cases we rely on a coupling of the temporal and spatial step size. We choose such h, k, such that there exist constants $M, \nu > 0$ such that $k \leq Mh^{2+\nu}$.

Up to the best of our knowledge there has been work by Benner and Mena on the deterministic problem and its numerical approximation. An account on numerical solutions for the Riccati equation arising in deterministic control settings can be found in [20]. Further results can be found in [7], in which efficient algorithms to approximate the large scale Riccati equations are studied. Large scale Riccati equations in return approximate the integral Riccati equation as found by Gibson, see [14]. The approach by Gibson was further discussed and refined by Banks and Kunich in [5]. Aside from the recent work by Levajković, Mena and Tuffaha, [18], [19], we

are unaware of any publications that deal with the stochastic control case and the approximation of the arising Riccati equation but one. In [10] a study of the boundary control problem with multi-dimensional noise can be found, therein Backward Differentiation Formulas and splitting methods are discussed. We thus assume that it is new that we derive fully discrete variation of constants formulas for approximate Riccati equations driven by infinite dimensional noise. Up to the best of our knowledge it is also the first study of a fully discrete numerical schemes of Riccati equations in the semigroup framework with rigorous a-priori convergence rates. The second scheme we introduce is also new, its implementation is more elementary and improves computational time in comparison to the first scheme with piecewise quadratic functions. Our results are partly the base for a future paper on the finite element approximation of operator Lyapunov equations by Andersson, Lang, Pettersson and Schroer [3].

This thesis is presented in three chapters. In the first chapter we introduce the prerequisites for the error analysis. These mostly consist of basic results from the theory on strongly continuous semigroups and the formal introduction of the considered Riccati operator equation framework. A rigorous definition of a weak solution to the operator Riccati equation is given with respect to the variational operator Riccati equation (1.18). At the end of the first chapter we show that the example with the controlled stochastic heat equation complies with the assumptions and parameters of the first chapter, see Lemma 1.11. In the second chapter we derive useful reformulations of both schemes, show the existence and uniqueness theory for both discrete solutions and prove their convergence. For the existence see Lemmata 2.1 and 2.2. For the existence of the discrete solution to the first scheme we use the relation to the Sylvester equations. The key tool for the a-priori bounds and the a-priori order of convergence is Gronwall's Lemma. For the a-priori order of convergence see Theorems 2.4 and 2.6. In the third and final chapter we show some numerical results for the above example of the stochastic heat equation with multiplicative noise. Python codes are found in the appendix.

1 Preliminaries

The purpose of this chapter is to introduce notation and well known results about strongly continuous semigroups from the theory of partial differential equations and their numerical analysis. We use a notation, which is mainly in the style of [1]. Most of the results stated in the first section can be found in [17]. In the second section we introduce the operator Riccati equation and the corresponding linear quadratic control problem. At the end of this chapter we show that the stochastic heat equation from the introduction fits the previously specified framework and assumptions.

1.1 Semigroup Approximation

Let $(H, \|\cdot\|, \langle\cdot,\cdot\rangle), \mathcal{H}$ be a separable Hilbert space. Let $\mathcal{L}(U;V)$ be the space of bounded linear operators from Banach space U to Banach space V. A strongly continuous semigroup on H is a map $S : \mathbb{R}_+ \to \mathcal{L}(H)$, such that

(i) $S(0) = \mathrm{id}_H$,

(ii) $\forall t, s \geq 0 :\ S(t+s) = S(t)S(s)$,

(iii) $\forall x_0 \in H :\ \lim_{t \downarrow 0} \|S(t)x_0 - x_0\|_H = 0$.

Let $A : H \supset \mathcal{D}(A) \to H$ be a linear not necessarily bounded operator. We say A infinitesimally generates the semigroup S, if

$$Ax = \lim_{t \downarrow 0} \frac{1}{t} \left(S(t) - 1)\right) x, \quad \forall x \in \mathcal{D}(A). \tag{1.1}$$

Assumption 1.1. *The linear operator $A : H \supset \mathcal{D}(A) \to H$ is densely defined, self-adjoint and positive definite with compact inverse.*

If S is infinitesimally generated by A and A satisfies Assumption 1.1, then S satisfies the following smoothing property. There exists a constant M such that

$$\|A^\rho S(t)x\|_H \leq M t^{-\rho} \|x\|_H \quad \forall t \in \mathbb{R}_+,\ x \in H,\ \rho \in [0,1]. \tag{1.2}$$

We note that $A = -\Delta$ satisfies Assumption 1.1. Under Assumption 1.1 there exists a spectral representation of A. Every eigenvalue λ_n to eigen-function ϕ_n is positive for $n \in \mathbb{N}$. Hence $e^{-At}x = \sum_{n\in\mathbb{N}} e^{-\lambda_n t}\langle x, \phi_n\rangle \phi_n$ is well-defined and $(S(t))_{t\geq 0} := e^{-At}$ defines a strongly continuous semigroup. It holds true that $\frac{\mathrm{d}}{\mathrm{d}t}S(t)x = -AS(t)x$ for $t \geq 0$, $x \in \mathcal{D}(A)$. Therefore $-A$ infinitesimally generates $(S(t))_{t\geq 0}$. We note that $(S(t)u_0)_{t\geq 0}$ solves the following cauchy problem. Find $u : [0,T] \to H,\ T \in \mathbb{R}_+$, such that

$$\frac{\mathrm{d}}{\mathrm{d}t}u(t) + Au(t) = 0, \quad u(0) = u_0 \in H. \tag{1.3}$$

We consider the problem of finding a discrete approximation of $(S(t))_{t\geq 0}$. For this purpose we use finite elements for the spatial discretization. Define spaces $(H_r), r \in \mathbb{R}$ of fractional powers

of A. For $r \geq 0$ define $H_r := \mathcal{D}(A^r)$ subject to the norm $\|\cdot\|_{H_r} := \|A^r \cdot\|_H$. For $r \leq 0$ let H_r be the closure of H under the $\|\cdot\|_{H_r} = \|A^r \cdot\|$-norm. We define families of finite element spaces based on quasi-uniform grid families. A grid family $\Gamma = (\Gamma_h)_{h\in(0,1]}$ is called quasi-uniform, if there exists a number $M \in (0,1)$ such that every cell $\omega \in \Gamma_H$ contains a circle of radius $\rho_\omega \geq M\mathrm{diam}(\omega)$. Let $(W_h)_{h\in(0,1)}$ be a sequence of families of functions, if there exists a grid family Γ such that every $f \in W_h$ is uniquely defined by it's value on each vertex of Γ_h, we say $(W_h)_{h\in(0,1)}$ is based on that particular grid.

For $i \in \{1,2\}$ let $(V_h^i)_{h\in(0,1]} \subset H_{i/2}$ be a sequence of finite dimensional sub-spaces of $H_{1/2}$ and H_1, respectively. Define R_h to be the orthogonal projection of $H_{1/2}$ onto V_h with respect to the $H_{1/2}$-inner product $\langle A^{1/2}\cdot, A^{1/2}\cdot\rangle = \langle\cdot,\cdot\rangle_{H_{1/2}}$. In addition we assume that $(V_h^i)_{h\in(0,1]}$ complies with the following approximation property 1.2.

Assumption 1.2. *There exists a constant B such that*

$$\|R_h x - x\| \leq Bh^s \|x\|_{H_{\frac{s}{2}}}, \quad \forall x \in H_{\frac{s}{2}},\ s \in \{1,2\},\ h \in (0,1].$$

Note that both the standard finite element method and the spectral Galerkin method comply with Assumption 1.2, see [17, Examples 3.6, 3.7]. For V_h^2 it is possible to formulate a higher order approximation property, however it is never used in this thesis. Let $A_h^i \in \mathcal{L}(V_h^i)$ such that for $x_h \in V_h^i$, $A_h^i x_h$ is the unique element in V_h^i which satisfies

$$\langle Ax_h, y_h\rangle = \langle A_h^i x_h, y_h\rangle \quad \forall y_h \in V_h^i,\ i \in \{1,2\}. \tag{1.4}$$

This operator serves as a discrete version of A. We write $A_h = A_h^i$, since it is self-evident which one we mean in every given context. Note that A_h is both positive and self-adjoint. By $\Pi_h : H \to V_h$ we denote the orthogonal projection with respect to the H-inner product. We further assume the below inverse inequality.

Assumption 1.3. *There exists a positive constant M, such that*

$$\|A_h \Pi_h\| \leq Mh^{-2}, \quad h \in (0,1].$$

Finally we assume

Assumption 1.4. *There exists constant M, such that*

$$\|A_h^\alpha \Pi_h\| \leq M\|A^\alpha x\|; \quad x \in H_{\alpha/2}, \alpha \in [-1/2, 1/2].$$

Both Assumption 1.3 and Assumption 1.4 hold true, if V_h^i is based on a quasi uniform grid family, see [21, Equation 1.12] and [4, Equation 2.12]. For $k \in (0,1]$ let $t_j := jk, j \in \{0,\dots,N_k\} =: \mathcal{N}_k^0$ with $N_k k \leq T < (N_k+1)k$, then $(\mathcal{T}_k)_{k\in(0,1]} = (\{t_i\}_{i\in\mathcal{N}_k^0})_{k\in(0,1]}$ is a sequence of discretizations of the time interval $[0,T]$. We further define $\mathcal{N}_k := \mathcal{N}_k^0 \setminus \{0\}$. For $k,h \in (0,1]$ we refer to $(1+kA_h)^{-j}$ as the discrete semigroup. We introduce the following notation.

$$\bar{S}(t) := (1+kA_h)^{-j}, \quad \text{if } t \in [t_{j-1}, t_j),\ j \in \mathcal{N}_k. \tag{1.5}$$

Then we have the following estimates.

Lemma 1.5. *Let A and V_h satisfy Assumptions 1.1 and 1.2 respectively. Then following estimates hold true.*

(i) For all $0 \leq \nu \leq \mu \leq 1$ there exists a constant M such that

$$\left\|\left(\bar{S}(t)\Pi_h - S(t)\right)x\right\| \leq M(h^{2\mu} + k^\mu)t^{-(\mu-\nu)}\|x\|_{H_\nu}, \quad \forall x \in H_\nu,\ t > 0,\ h,k \in (0,1].$$

(ii) For all $0 \leq \rho \leq 1/2$ there exists a constant M such that

$$\left\| \left(\bar{S}(t)\Pi_h - S(t) \right) x \right\| \leq M t^{-\rho} \|x\|_{H_{-\rho}}, \quad \forall x \in H_{-\rho},\ t > 0,\ h, k \in (0, 1].$$

(iii) For all $0 \leq \rho \leq 1$ and $-\theta \leq \rho \leq \min(1/2, 1-\theta)$ there exists a constant M such that

$$\left\| \left(\bar{S}(t)\Pi_h - S(t) \right) x \right\| \leq M (h^{2\theta} + k^{\theta}) t^{-(\theta+\rho)} \|x\|_{H_{-\rho}}, \quad \forall x \in H_{-\rho},\ t > 0,\ h, k \in (0, 1].$$

For proofs see [17, Lemma 3.12] and [2, Lemma 5.1]. Thus $\bar{S}(t)$ converges towards $S(t)$ and is a discrete version of the semigroup. Note that, for $X \in \mathcal{L}(H)$, we denote $\|X\|_{\mathcal{L}(H)}$ as $\|X\|$. Further note that $\|S(t)\|$, $\|(1 + kA_h)^{-j}\Pi_h\|$, $\|\Pi_h\|$ are smaller than or equal to one for all $h, k \in (0, 1]$. Like the semigroup the discrete semigroup satisfies a smoothing property. There exists a constant M independent of k, h, j such that

$$\|A_h^{\rho}(1 + kA_h)^{-j} x_h\| \leq M t_j^{-\rho} \|x_h\|_H, \quad \forall j \in \mathbb{N}_1,\ x_h \in V_h,\ \rho \in [0, 1] \tag{1.6}$$

see [17, Equation 3.38]. This can be generalized further for A.

$$\|A^{\rho}(1 + kA_h)^{-j} x_h\| \leq M t_j^{-\rho} \|x_h\|_H, \quad \forall j \in \mathbb{N}_1,\ x_h \in V_h,\ \rho \in [0, 1/2] \tag{1.7}$$

see [17, Equation 3.42] for the case of $\rho = 1/2$, the case of $\rho = 0$ is obvious and the intermediate cases follow by an interpolation argument, see [21, Theorem 3.5]. We state two further Theorems, which are substantial to the second chapter.

Lemma 1.6 (Gronwall's Lemma)**.** *Let $T > 0$ and $k \in (0, 1)$. Define $t_j := jk$ and $j \in \mathcal{N}_k$, such that $N_k \leq T < (N_k + 1)$. Let $A, B \geq 0$ and let $(\phi_j)_{j=1,\dots,N_k}$ be a non-negative sequence. If there exist $\alpha, \beta > 0$ such that*

$$\phi_n \leq A t_n^{-1+\alpha} + Bk \sum_{i=1}^{n-1} t_{n-i}^{-1+\beta} \phi_i, \quad \forall n \in \mathcal{N}_k, \tag{1.8}$$

then there exists a constant $C = C(A, B, T, \beta, \alpha)$ such that

$$\phi_n \leq C t_n^{-1+\alpha} \quad , \forall n \in \mathcal{N}_k. \tag{1.9}$$

Proof. The proof is based on applying (1.8) to itself $(n-1)$ times. A first application to itself yields

$$\phi_n \leq A t_n^{-1+\alpha} + Bk \sum_{i=1}^{n-1} t_{n-i}^{-1+\beta} A t_i^{-1+\alpha} + Bk \sum_{i=1}^{n-1} t_{n-i}^{-1+\beta} Bk \sum_{j=1}^{i-1} t_{i-j}^{-1+\beta} \phi_j \quad \forall n \in \mathcal{N}_k. \tag{1.10}$$

The following estimate holds true.

$$k \sum_{i=1}^{j-1} t_{j-i}^{-1+\beta} t_i^{-1+\alpha} \leq \int_0^{t_j} (t_j - s)^{-1+\beta} s^{-1+\alpha} \mathrm{d}s = t_j^{-1+\alpha+\beta} \mathrm{B}(\alpha, \beta) \leq T^{\beta} t_j^{-1+\alpha} \mathrm{B}(\alpha, \beta), \tag{1.11}$$

where $\mathrm{B}(\cdot, \cdot)$ denotes the beta function, which is well-defined for positive parameters. In terms of the Gamma function the Beta function is defined as

$$\mathrm{B}(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x + y)}, \quad x, y > 0. \tag{1.12}$$

We apply this to the above inequality and get

$$\phi_n \leq \left(A + BT^{\beta}\mathrm{B}(\alpha,\beta)\right) t_n^{-1+\alpha} + Bk\sum_{i=1}^{n-1} t_{n-i}^{-1+\beta} Bk \sum_{j=1}^{i-1} t_{i-j}^{-1+\beta}\phi_j, \quad \forall n \in \mathcal{N}_k. \tag{1.13}$$

We repeat this $(n-1)$-times. For some constant $C = C(A, B, T, \alpha, \beta)$ we get

$$\phi_n \leq Ct_n^{-1+\alpha} + Bk\sum_{i=1}^{n-1} t_{n-i}^{-1+\beta} Bk \sum_{j=1}^{i-1} t_{i-j}^{-1+\beta}\phi_j \ldots At_1^{-1+\alpha}, \quad \forall n \in \mathcal{N}_k. \tag{1.14}$$

After a final application of (1.11) we have a constant $C = C(A, B, T, \alpha, \beta)$ such that

$$\phi_n \leq Ct_n^{-1+\alpha}, \quad \forall n \in \mathcal{N}_k. \tag{1.15}$$

This completes the proof. □

This proof was taken from [11]. For the proof of the following theorem see [8].

Theorem 1.7 (Silvester-Rosenblum)**.** *If A and B are linear operators with disjoint spectra, then the Sylvester equation*

$$AX - XB = Y,$$

has a unique solution for every linear operator Y.

1.2 Operator Riccati Equations

In this section we introduce the abstract operator Riccati equation arising from a distributed control problem of a linear stochastic evolution equation. Let $(H, \|\cdot\|, \langle\cdot,\cdot\rangle)$ be a separable Hilbert space. Let $A : H \supset \mathcal{D}(A) \to H$ be a linear bounded operator satisfying Assumption 1.1. Let $\Sigma(H) \subset \mathcal{L}(H)$ be the space of self-adjoint linear bounded operators on H and let $\Sigma^+(H) \subset \Sigma(H)$ be the subset of non-negative linear operators. Let $\mathcal{L}_2(U;V)$ be the space of Hilbert-Schmidt operators from Hilbert space U to Hilbert space V. A Hilbert-Schmidt operator is a bounded operator $O \in \mathcal{L}(U;V)$ with finite Hilbert-Schmidt norm $\|\cdot\|_{\mathcal{L}_2(U;V)}$. This norm is defined by

$$\|O\|_{\mathcal{L}_2(U;V)} := \mathrm{Tr}(O^*O) := \sum_{i\in I}\langle Oe_i, Oe_i\rangle_V, \tag{1.16}$$

where $\{e_i : i \in I\}$ is any fixed orthonormal basis of U. The Hilbert-Schmidt norm is independent of choice of the basis $\{e_i : i \in I\}$. $\mathcal{L}_2(U;V)$, equipped with $\langle\cdot,\cdot\rangle_{\mathcal{L}_2(U;V)}$, is a Hilbert space. It has the inner product

$$\langle O_1, O_2\rangle_{\mathcal{L}_2(U;V)} := \mathrm{Tr}(O_1^*O_2) = \sum_{i\in I}\langle Oe_i, Oe_i\rangle_V, \quad O_1, O_2 \in \mathcal{L}_2(U;V). \tag{1.17}$$

Assumption 1.8. *Let $G, Q \in \mathcal{L}(H)$ and $C \in \mathcal{L}(H; \mathcal{L}_2(\mathcal{H}; H_{-\beta}))$ with regularity parameter $\beta \in [0, 1/2)$.*

We define the following scalar products on $H_{1/2}$ and H_1 respectively. A weak or variational solution to the operator Riccati equation (1.18) is a bounded and strongly continuous function $P : [0, T] \to \Sigma^+(H)$, such that

(i) $P(t) \in \mathcal{L}(H_{-\beta}; H_\beta), \quad \forall t \in (0,T],$

(ii) $((0,T] \ni t \mapsto \langle P(t)x, y\rangle) \in \mathcal{C}^1((0,T]; H), \quad \forall x, y \in H_{1/2},$

and $(P(t))_{t\in[0,T]}$ with $P(0) = G^*G$ satisfies the variational Riccati equation:

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle P(t)x, y\rangle + \langle P(t)x, Ay\rangle + \langle P(t)y, Ax\rangle + \langle P(t)x, P(t)y\rangle \\ = \langle P(t)Cx, Cy\rangle_{\mathcal{L}_2(\mathcal{H};H)} + \langle Qx, Qy\rangle, \quad t \in (0,T],\ \forall x, y \in H_{1/2}. \quad (1.18)$$

A mild solution to the Riccati eqation (1.18) is a bounded and strongly continuous function $P : [0,T] \to \Sigma^+(H)$, such that

(i) $P(t) \in \mathcal{L}(H_{-\beta}; H_\beta), \quad \forall t \in (0,T],$

(ii) $\sup_{t\in(0,T]} t^{2\beta}\|P(t)\|_{\mathcal{L}(H_{-\beta};H_\beta)} < \infty,$

and $(P(t))_{t\in[0,T]}$ satisfies the variation of constants integral equation for the Riccati equation:

$$P(t)x = S(t)G^*GS(t)x \\ + \int_0^t S(t-s)\left(Q^*Q + C^*P(s)C - P(s)^2\right)S(t-s)x\ \mathrm{d}s, \quad t \in [0,T],\ \forall x \in H. \quad (1.19)$$

The above integral is in the sense of a H-valued Bochner integral. An interpretation in terms of a $\mathcal{L}(H)$-valued integral is only formal since in general the natural sigma-algebra on $\mathcal{L}(H)$ is too fine.

Theorem 1.9. *There exists a unique mild solution $P : [0,T] \to \Sigma^+(H)$ to* (1.18)*. Furthermore P is a mild solution if and only if P is a weak solution to the Riccati equation* (1.18)*.*

For proof see [1, Theorems 2.2–2.3]. We emphasize that $\|P(t)\|$ means

$$\sup_{\|x\|_H=1}\left\|S(t)G^*GS(t)x + \int_0^t S(t-s)\left(Q^*Q + C^*P(s)C - P(s)^2\right)S(t-s)x\mathrm{d}s\right\|,$$

and does not mean the norm of a $\mathcal{L}(H)$-valued integral.

Theorem 1.10. *Let P be the solution to* (1.19)*. For all $\theta \in [0, \frac{1}{2})$ there exists a constant M, such that*

$$\|P(t)\|_{\mathcal{L}(H_{-\theta};H_\theta)} \le Mt^{-2\theta}, \quad t \in (0,T].$$

For all θ, ξ such that $2\theta + \xi < 1$ there exists a constant M, such that

$$\|P(t_2) - P(t_1)\|_{\mathcal{L}(H_{-\theta};H_\theta)} \le M\left(t_1 \wedge t_2\right)^{-2\theta-\xi}|t_2 - t_1|^\xi, \quad t_1, t_2 \in (0,T].$$

For the proof see [1, Theorem 2.4]. This theorem is essential to the analysis of the schemes. If $(P(t))_{t\in[0,T]}$ is a weak solution to the operator Riccati equation, we have that $y \mapsto \frac{\mathrm{d}}{\mathrm{d}t}\langle P(t)x, y\rangle$ is linear and $|\frac{\mathrm{d}}{\mathrm{d}t}\langle P(t)x, y\rangle| \le M\|x\|_{H_\theta}\|y\|_{H_\theta}$ for some constant M. By Riesz representation theorem there exists $\dot{P} \in \mathcal{L}(H)$, such that $\frac{\mathrm{d}}{\mathrm{d}t}\langle P(t)x, y\rangle = \langle \dot{P}(t)x, y\rangle$ for all $x, y \in H_\theta$. With this definition the following notation of the operator Riccati equation makes sense.

$$\dot{P}(t) + AP(t) + P(t)A + P(t)^2 = Q^*Q + C^*P(t)C, \quad t \in [0,T];\ P_0 = G^*G. \quad (1.20)$$

This illustrates the relation of this operator Riccati equation to the matrix Riccati equation. We use the weak formulation (1.18) to implement finite element methods, whereas the mild solution is substantial for later error analysis.

Let us verify that the example (0.5) arising from the control of a stochastic heat equation with multiplicative noise (0.2) and Dirichlet boundary conditions fits this framework. We choose $A := -\Delta$, $G := Q := \mathrm{id}_H$ and show that the multiplication operator is an element of $\mathcal{L}(H; \mathcal{L}_2(H; H_{-\beta})$ for some $\beta \in (1/4, 1/2)$, where $H := L^2([0,1])$.

Lemma 1.11. *Let $H := L^2([0,1])$ and let C be the multiplication operator, which is defined by $(C(v)u)(x) = u(x)v(x)$. Then C satisfies $C \in \mathcal{L}(H; \mathcal{L}_2(H; H_{-\beta}))$ for $\beta > 1/4$.*

Proof. Let $(e_k)_{k\in\mathbb{N}}$ be the orthonormal basis of $H = L^2([0,1])$, defined by $e_k(x) = 2^{1/2}\sin(k\pi x)$. This choice is also an eigenbasis for $A := -\Delta$. Because $(e_k)_{k\in\mathbb{N}}$ is a orthonormal basis, we have by Parsevals identity for any fixed $u \in H$ that

$$\|C(u)\|_{\mathcal{L}_2(H;H_{-\beta})} = \|A^{-\beta}C(u)\|_{\mathcal{L}_2(H)} = \sum_{k\in\mathbb{N}} \|A^{-\beta}C(u)e_k\|_H^2 = \sum_{k,\ell\in\mathbb{N}} \langle e_\ell, A^{-\beta}C(u)e_k\rangle^2$$

Since $(e_k)_{k\in\mathbb{N}}$ is also an eigenbasis for $A = -\Delta$ and A is selfadjoint, we have

$$\|C(u)\|_{\mathcal{L}_2(H;H_{-\beta})} = \sum_{k,\ell\in\mathbb{N}} \langle A^{-\beta}e_\ell, C(u)e_k\rangle^2 = \sum_{k,\ell\in\mathbb{N}} \|A^{-\beta}e_\ell\|_H^2 \langle e_\ell, C(u)e_k\rangle^2.$$

Further, because $C(u)$ is self-adjoint we get

$$\|C(u)\|_{\mathcal{L}_2(H;H_{-\beta})} = \sum_{\ell\in\mathbb{N}} \|A^{-\beta}e_\ell\|_H^2 \sum_{k\in\mathbb{N}} \langle e_k, C(u)e_\ell\rangle^2 = \sum_{\ell\in\mathbb{N}} \|A^{-\beta}e_\ell\|_H^2 \|C(u)e_\ell\|_H^2.$$

We use the specific choice of $e_k = 2^{1/2}\sin(k\pi x)$ and Hölder's inequality and estimate

$$\|C(u)e_k\|_H^2 = 2\int_0^1 u^2(x)\sin^2(k\pi x)\mathrm{d}x \le 2\int_0^1 u^2(x)\mathrm{d}x.$$

Since $(e_k)_{k\in\mathbb{N}} \subset H$ is orthonormal and $Ae_k = (\pi k)^2 e_k$, we use the definition of fractional powers to determine $A^{-\beta}$ and have

$$\|A^{-\beta}e_\ell\|_H = \Big\| \sum_{k\in\mathbb{N}} (\pi k)^{-2\beta} \langle e_\ell, e_k\rangle e_k \Big\| = (\pi k)^{-2\beta}.$$

We use this and have

$$\sum_{\ell\in\mathbb{N}} \|A^{-\beta}e_\ell\|_H^2 \|C(u)e_\ell\|_H^2 \le 2\sum_{\ell\in\mathbb{N}} \|A^{-\beta}e_\ell\|_H^2 \|u\|_H = 2\|A^{-\beta}\|_{\mathcal{L}_2(H)}^2 \|u\|_H.$$

We therefore have

$$\|C\|_{\mathcal{L}(H;\mathcal{L}_2(H;H_{-\beta}))} \le 2\|A^{-\beta}\|_{\mathcal{L}_2(H)}^2 = 2\sum_{k\in\mathbb{N}} \|A^{-\beta}e_k\|_H^2 = 2\pi^{-4\beta}\sum_{k\in\mathbb{N}} k^{-4\beta} = 2\pi^{-4\beta}\zeta(4\beta),$$

where $\zeta(s)$ denotes the Riemann zeta function, which is finite for $s > 1$. This completes the proof. □

This proof was taken from [16, § 5.2.1]. It shows that the intitial example fits the theoretical framework.

1.3 The generalized LQ-Problem

For the sake of completeness we briefly introduce the LQ-problem, which belongs to the abstract operator Riccati equation. Let $(H, \|\cdot\|, \langle\cdot,\cdot\rangle)$ be a separable Hilbert space and $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$ be a filtered probability space. Let $\mathcal{H} \subset H$ be a orthonormal basis of H. We define a cylindrical id_H-Wiener process $W : H \to L^2([0,T]\times\Omega;\mathbb{R})$ defined on $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$ to be the strong operator limit

$$W := \sum_{\eta\in\mathcal{H}} \xi^\eta \otimes \eta, \tag{1.21}$$

where $(\xi^\eta)^{\eta\in H}_{t\in[0,T]}$ denotes a sequence of independent standard Wiener processes defined on $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$. Let A, B, C, G, Q be linear operators, such that Assumptions 1.1 and 1.8 are satisfied. We consider the following dynamics

$$\mathrm{d}X^u(t) + AX^u(t)\mathrm{d}t = u(t)\mathrm{d}t + CX^u(t)\mathrm{d}W(t),\ t > 0;\ X^u_0 = x_0 \in H. \tag{1.22}$$

Consider the following optimization problem: Find a control function $u : [0,T)\times\Omega \to H$, which minimizes the cost functional

$$J(u) = \mathbb{E}\left[\int_0^T (\|QX^u(t)\|^2 + \|u(t)\|^2)\mathrm{d}t + \|GX^u(T)\|^2\right]. \tag{1.23}$$

We call a control function u admissible, if it is a predictable square-integrable function $u \in L^2([0,T]\times\Omega;H)$. Using Itô's formula and an approximation argument, which can be found in [1], one can show that

$$J(u) = \langle P_T x_0, x_0\rangle + \mathbb{E}\left[\int_0^T \|u_t + P(T-t)X^u(t)\|^2\mathrm{d}t\right], \tag{1.24}$$

where $P : [0,T] \to \mathcal{L}(H)$ is a weak solution to the operator Riccati equation (1.18). The stochastic process

$$\bar{u}(t) := -P(T-t)X^{\bar{u}}(t),\ t \in [0,T],$$

thus solves the optimization problem, if the controlled process $X^{\bar{u}}$ exists. Because this control function is a linear transformation of the dynamics, we say the optimal control is given in feedback form. The existence of $X^{\bar{u}}$ can be shown with a fixed point argument, see [1]. In literature the infinite dimensional deterministic problem was first studied in semigroup formulation in [9]. The stochastic case was later introduced in [15]. For the boundary control problem see [12] and [13].

2 A-priori Error Analysis

This chapter contains two sections. The first section introduces two schemes which we analyze in the second section. The first section also establishes the existence and uniqueness to both schemes. In the second section we provide proofs for a-priori order of convergence for both schemes. Throughout this chapter let $(H, \|\cdot\|, \langle\cdot,\cdot\rangle)$ be a separable Hilbert space, $A : H \supset \mathcal{D}(A) \to H$ a linear Operator such that Assumption 1.1 holds true. We write $a(x,y) := \langle A^{1/2}x, A^{1/2}y\rangle = \langle Ax, y\rangle,\ x, y \in H$. For $i \in \{1,2\}$ let $(V_h^i)_{h\in(0,1]}$ be a sequence of finite dimensional sub-spaces of $H_{i/2}$ satisfying Assumptions 1.2-1.4. Further let C, G, Q be operators such that Assumption 1.8 holds true. Throughout this chapter we denote by M not necessarily coinciding positive constants, which are independent of the parameters k, h, t_j , unless otherwise specified.

2.1 Setting

First, recall the initial problem of finding a strongly continuous $P : (0,T] \to \Sigma(H) \cap \mathcal{L}(H_{-\beta}; H_\beta)$, which is a weak solution to (1.18), i.e. satisfies

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\langle P(t)x, y\rangle + a\left(P(t)x, y\right) + a\left(P(t)y, x\right) + \langle P(t)x, P(t)y\rangle \\ = \langle P(t)Cx, Cy\rangle_{\mathcal{L}_2(H)} + \langle Qx, Qy\rangle\,, \quad t \in (0,T],\ \forall x, y \in H_{1/2}.\end{aligned}$$

Let $h, k \in (0,1]$ and $N_k \in \mathbb{N}$, such that $kN_k \le T < k(N_k+1)$. The following equation defines a fully discrete semi-implicit Euler scheme. Find a sequence $(P_{h,k}^j)_{j=0}^{N_k} \subset \mathcal{L}(V_h)$ with $P_{h,k}^0 = \Pi_h G^* G \Pi_h$ such that

$$\begin{aligned}&\left\langle P_{h,k}^j x, y\right\rangle + ka\big(\langle P_{h,k}^j x, y\big) + ka\big(P_{h,k}^j y, x\big) + k\big\langle P_{h,k}^{j-1}x, P_{h,k}^{j-1}y\big\rangle \\ &\quad = \left\langle P_{h,k}^{j-1}x, y\right\rangle + k\big\langle P_{h,k}^{j-1}\Pi_h Cx, Cy\big\rangle_{\mathcal{L}_2(\mathcal{H};H)} + k\big\langle Qx, Qy\big\rangle, \quad \forall x, y \in V_h,\ j \in \mathcal{N}_k. \quad (2.1)\end{aligned}$$

We postpone stating the second scheme to later in this section, see equation (2.7). We remind ourselves of the definition of Π_h and A_h, see (1.4). The equation (2.1) is thus equivalent to

$$P_{h,k}^j + kA_hP_{h,k}^j + kP_{h,k}^jA_h + k\big(P_{h,k}^{j-1}\big)^2 = P_{h,k}^{j-1} + k\Pi_hQ^*Q + k\Pi_hC^*P_{h,k}^{j-1}\Pi_hC. \quad (2.2)$$

To prevent too many indices and long expressions, we use the notation

$$\begin{aligned}&S_{h,k}^j := (1+kA_h)^{-j} && j \in \mathcal{N}_k^0. \\ &\bar{S}(t) := (1+kA_h)^{-j} && \text{if } t \in [t_{j-1}, t_j),\ j \in \mathcal{N}_k, \\ &\bar{P}(t) := P_{h,k}^j && \text{if } t \in (t_j, t_{j+1}],\ j \in \mathcal{N}_k,\ \bar{P}(0) = G^*G.\end{aligned} \quad (2.3)$$

Our next goal is to derive a discrete variation of constants formula for the first scheme. We add $k^2 A_h P_{h,k}^j A_h$ to both sides of (2.2) and arrive at the following equation.

$$\begin{aligned} P_{h,k}^j + kA_h P_{h,k}^j + kP_{h,k}^j A_h + k^2 A_h P_{h,k}^j A_h \\ = P_{h,k}^{j-1} - k\big(P_{h,k}^{j-1}\big)^2 + kQ^*Q + k\Pi_h C^* P_{h,k}^{j-1} \Pi_h C + k^2 A_h P_{h,k}^j A_h, \quad j \in \mathcal{N}_k. \end{aligned} \tag{2.4}$$

We note that

$$P_{h,k}^j + kA_h P_{h,k}^j + kP_{h,k}^j A_h + k^2 A_h P_{h,k}^j A_h = (1+kA_h) P_{h,k}^j (1+kA_h).$$

Multiply (2.4) by $S_{h,k}^1 = (1+kA_h)^{-1}$ from the left and from the right. This gives

$$P_{h,k}^j = S_{h,k}^1 \left(P_{h,k}^{j-1} - k\big(P_{h,k}^{j-1}\big)^2 + kQ^*Q + k\Pi_h C^* P_{h,k}^{j-1} \Pi_h C + k^2 A_h P_{h,k}^j A_h \right) S_{h,k}^1.$$

Iterating this equation finally yields the discrete variation of constants formula.

$$\begin{aligned} P_{h,k}^j = S_{h,k}^j P_{h,k}^0 S_{h,k}^j + k \sum_{i=0}^{j-1} S_{h,k}^{j-i} \left(\Pi_h Q^*Q - k\big(P_{h,k}^i\big)^2 + \Pi_h C^* P_{h,k}^i \Pi_h C \right) S_{h,k}^{j-i} \\ + k^2 \sum_{i=0}^{j-1} S_{h,k}^{j-i} \left(kA_h P_{h,k}^{i+1} A_h \right) S_{h,k}^{j-i}, \quad j \in \mathcal{N}_k^0. \end{aligned} \tag{2.5}$$

We rewrite the discrete variation of constants formula. Using (2.3) we have

$$\begin{aligned} P_{h,k}^j = S_{h,k}^j \bar{P}(0) S_{h,k}^j + \int_0^{t_j} \bar{S}(t_j - s) \left(\Pi_h Q^*Q - \bar{P}(s)^2 + \Pi_h C^* \bar{P}(s) \Pi_h C \right) \bar{S}(t_j - s) \mathrm{d}s \\ + k^2 \sum_{i=0}^{j-1} S_{h,k}^{j-i} \left(A_h P_{h,k}^{i+1} A_h \right) S_{h,k}^{j-i}, \quad j \in \mathcal{N}_k^0. \end{aligned} \tag{2.6}$$

This formula allows for a nice comparison with the mild solution of the operator Riccati equation (1.19).

Lemma 2.1. *For all step-sizes $k, h \in (0,1)$ there exists a unique $(P_{h,k}^j)_{j=0}^{N_k} \subset \Sigma(V_h^1)$ with $P_{h,k}^0 = \Pi_h G^* G \Pi_h$, which solves (2.1) for every $j \in \mathcal{N}_k$.*

Proof. We have shown that (2.1) is satisfied, if and only if (2.4) is satisfied. X solves equation (2.4) for $P_{h,k}^j$ if and only if X solves

$$X + kA_h X + kXA_h = P_{h,k}^{j-1} + kQ^*Q + kC^* P_{h,k}^{j-1} C - k(P_{h,k}^{j-1})^2.$$

Denote the right hand side by Y. We write $\tilde{A} := (1 + kA_h)$ and $\tilde{B} := -kA_h$. We first note that $Y, \tilde{A}, \tilde{B} \in \mathcal{L}(V_h^1)$. The operator $\tilde{A}$ is positive and $\tilde{B}$ is negative, i.e. in particular $\tilde{A}, \tilde{B}$ have disjoint spectra. By Theorem 1.7 there exists a unique solution to the Silvester equation $\tilde{A}X + X\tilde{B} = Y$. By induction for every $j \in \mathcal{N}_k$ there exists a unique $P_{h,k}^j$, which solves (2.4). We have that the right hand side of the above display is self-adjoint, if $P_{h,k}^{j-1}$ is self-adjoint. Studying $X - X^*$ yields

$$\begin{aligned} 0 &= (X - X^*) + (kA_h X + kXA_h - k(A_h X)^* - k(XA_h)^*) + Y^* - Y \\ &= (X - X^*) + (kA_h X + kXA_h - kX^* A_h - kA_h X^*) \\ &= (X - X^*) + k(X - X^*)A_h + kA_h(X - X^*) \\ &= (1 + kA_h)(X - X^*) + (X - X^*)kA_h. \end{aligned}$$

We can choose a basis of V_h^1, which contains the eigen-vectors of A_h. For any x in that particular basis there exists $\lambda \in \mathbb{R}$ such that $A_h x = \lambda x$. We apply $\langle \cdot x, x\rangle$ to the above equation and arrive at $0 = (1+2k\lambda)\langle (X - X^*)x, x\rangle$, because A_h is self-adjoint. Subsequently $X - X^* = 0$ and X is self-adjoint. By induction we have that $P_{h,k}^j$ is self-adjoint for every $j \in \mathcal{N}_k$. □

Next, we introduce the second scheme. The idea of this scheme is to include $k^2\langle P_{h,k}^j \Pi_h Ax, Ay\rangle$ in the scheme, instead of adding $k^2 A_h P_{h,k}^j A_h$ to both sides of (2.4) . For this purpose we need $V_h^2 \subset \mathcal{D}(A)$. An example of such a space V_h^2 is the subspace of piecewise quadratic finite elements. Under this assumption find $(P_{h,k}^j)_{j=0}^{N_k} \subset \mathcal{L}(V_h^2) \subset \mathcal{L}(\mathcal{D}(A))$ with $P_{h,k}^0 = \Pi_h G^* G \Pi_h$ such that for $j \in \mathcal{N}_k$

$$\begin{aligned}\big\langle P_{h,k}^j x, y\big\rangle + ka\big(P_{h,k}^j x, y\big) + ka\big(P_{h,k}^j y, x\big) + k\big\langle P_{h,k}^{j-1}x, P_{h,k}^{j-1}y\big\rangle + k^2\langle P_{h,k}^j \Pi_h Ax, Ay\rangle \\ = \big\langle P_{h,k}^{j-1}x, y\big\rangle + k\big\langle P_{h,k}^{j-1}\Pi_h Cx, Cy\big\rangle_{\mathcal{L}_2(\mathcal{H};H)} + k\big\langle Qx, Qy\big\rangle, \quad \forall x, y \in V_h^2,\ j \in \mathcal{N}_k. \quad (2.7)\end{aligned}$$

Similar to the above procedure we pass over to the abstract operator equation

$$\begin{aligned}P_{h,k}^j + kA_h P_{h,k}^j + kP_{h,k}^j A_h + k^2 A_h P_{h,k}^j A_h \\ = P_{h,k}^{j-1} - k\left(P_{h,k}^{j-1}\right)^2 + k\Pi_h Q^* Q + k\Pi_h C^* P_{h,k}^{j-1}\Pi_h C, \quad j \in \mathcal{N}_k. \quad (2.8)\end{aligned}$$

In this case we can write down an explicit expression for the discrete variation of constants formula

$$P_{h,k}^j \quad = \quad S_{h,k}^j P_{h,k}^0 S_{h,k}^j \ + \ k\sum_{i=0}^{j-1} S_{h,k}^{j-i}\left(-\left(P_{h,k}^i\right)^2 + \Pi_h Q^* Q + \Pi_h C^* P_{h,k}^i \Pi_h C\right) S_{h,k}^{j-i}. \quad (2.9)$$

This coincides with the following version of this variation of constants formula

$$\begin{aligned}P_{h,k}^j = S_{h,k}^j \bar{P}(0) S_{h,k}^j \\ + \int_0^{t_j} \bar{S}(t_j - s)\left(-\bar{P}(s)^2 + \Pi_h Q^* Q + \Pi_h C^* \bar{P}(s)\Pi_h C\right)\bar{S}(t_j - s)\mathrm{d}s. \quad (2.10)\end{aligned}$$

Lemma 2.2. *Let $V_h^2 \subset \mathcal{D}(A)$. For all step-sizes $k, h \in (0,1)$ there exists a unique $(P_{h,k}^j)_{j=0}^{N_k} \subset \Sigma(V_h^2)$, with $P_{h,k}^0 = \Pi_h G^* G \Pi_h$, which solves (2.7) for every $j \in \mathcal{N}_k$.*

Proof. We have seen that (2.7) is satisfied, if and only if (2.8) is satisfied. However equation (2.9) uniquely defines the solution to (2.8) via recursion and thus the solution exists. This recursion defines $P_{h,k}^j$ as a sum of self-adjoint operators and thus $P_{h,k}^j$ is self-adjoint, too. □

2.2 Analysis of the First Scheme

The aim of this section is to establish convergence of the proposed schemes. The idea is that for sufficiently small step-sizes the discrete solution is positive. For any smaller step-sizes the discrete solution is also bounded, which we show in Lemma 2.3. With the help of this estimate for the norm of the discrete solution we can show that the discrete solution converges towards the mild solution. This convergence is in the sense of the difference between the discrete and the mild solution in the $\mathcal{L}(H_{-\beta}; H_\beta)$-norm, see Theorem 2.4. We begin with the analysis of the first scheme with the proof of the a-priori bound.

Lemma 2.3. *Let $(P_{h,k}^i)_{i=0}^{N_k}$ be the solution to (3.2). Let $B > 0$ and $\nu > 1/2$. There exist $\kappa, \eta \in (0,1]$, such that*

$$\|A_h^\theta P_{k,h}^j A_h^\theta \Pi_h\| \leq M t_j^{-2\theta},\ \forall j \in \mathcal{N}_k,\ \forall \theta \in [0,1/2),\ \forall (h,k) \in (0,\eta) \times (0,\kappa) \text{ with } k \leq B h^{2+\nu},$$

for some constant M, which is independent of step-sizes $k, h \in (0,1]$. Further $P_{h,k}^j$ is positive for every $j \in \mathcal{N}_k^0$.

Proof. We begin the proof of this a-priori with an estimate under the positivity assumption $\langle P_{h,k}^i x, x\rangle \geq 0, \forall x \in V_h^1, i \in \mathcal{N}_k$ and for $\theta \geq \beta$. In the second part of this proof we consider the cases of $\theta < \beta$. In the third part we finally prove that, under the given assumptions, $P_{h,k}^j$ is indeed positive for all $j \in \mathcal{N}_k$. We denote by B_1 the constant from Assumption 1.3 and by B_2 the constant such that $k \leq B_2 h^{2+\nu}$.

Part 1. If $P_{h,k}^j$ is positive, its positive square-root $(P_k^j)^{\frac{1}{2}}$ is well-defined. Therefore,

$$\left\|\left(P_{h,k}^j\right)^{\frac{1}{2}} A_h^\theta \Pi_h x\right\|^2 = \left\langle (P_{h,k}^j)^{\frac{1}{2}} A_h^\theta \Pi_h x, (P_{h,k}^j)^{\frac{1}{2}} A_h^\theta \Pi_h x\right\rangle = \left\langle P_k^j A_h^\theta \Pi_h x, A_h^\theta \Pi_h x\right\rangle. \tag{2.11}$$

For any $x \in H$ by insertion of scheme (2.5) into (2.11) we have

$$\begin{aligned}
\left\|\left(P_{h,k}^j\right)^{\frac{1}{2}} A_h^\theta \Pi_h x\right\|^2 = &\left\langle P(0) S_{h,k}^{j-i} A_h^\theta \Pi_h x, S_{h,k}^{j-i} A_h^\theta \Pi_h x\right\rangle \\
&- k \sum_{i=0}^{j-1} \left\langle P_{h,k}^i S_{h,k}^{j-i} A_h^\theta \Pi_h x, P_{h,k}^i S_{h,k}^{j-i} A_h^\theta \Pi_h x\right\rangle \\
&+ k \sum_{i=0}^{j-1} \left\langle P_{h,k}^i \Pi_h C S_{h,k}^{j-i} A_h^\theta \Pi_h x, \Pi_h C S_{h,k}^{j-i} A_h^\theta \Pi_h x\right\rangle_{\mathcal{L}_2(H)} \\
&+ k \sum_{i=0}^{j-1} \left\langle Q S_{h,k}^{j-i} A_h^\theta \Pi_h x, Q S_{h,k}^{j-i} A_h^\theta \Pi_h x\right\rangle \\
&+ k^2 \sum_{i=0}^{j-1} \left\langle P_{h,k}^{i+1} A_h S_{h,k}^{j-i} A_h^\theta \Pi_h x, A_h S_{h,k}^{j-i} A_h^\theta \Pi_h x\right\rangle.
\end{aligned}$$

The second term is negative and can be omitted in the next estimate. We write $P_{k,h}^j$ in terms of its square-roots, this yields the following expression in norms.

$$\begin{aligned}
\left\|\left(P_k^j\right)^{\frac{1}{2}} A_h^\theta x \Pi_h\right\|_H^2 \leq &\left\|G S_{h,k}^j A_h^\theta \Pi_h x\right\|_H^2 + k \sum_{i=0}^{j-1} \left\|Q S_{h,k}^{j-i} A_h^\theta \Pi_h x\right\|_H^2 \\
&+ k \sum_{i=0}^{j-1} \left\|\left(P_{h,k}^i\right)^{\frac{1}{2}} \Pi_h C S_{h,k}^{j-i} A_h^\theta \Pi_h x\right\|_{\mathcal{L}_2(H)}^2 \\
&+ k^2 \sum_{i=0}^{j-1} \left\|\left(P_{h,k}^{i+1}\right)^{\frac{1}{2}} A_h S_{h,k}^{j-i} A_h^\theta \Pi_h x\right\|_H^2.
\end{aligned}$$

Taking the supremum over all $x \in H, \|x\| \leq 1$ on the right hand side and then on the left hand side yields the following inequality in terms of the operator norm.

$$\begin{aligned}\left\|\left(P_k^j\right)^{\frac{1}{2}} A_h^\theta \Pi_h\right\|^2 \leq \left\|GS_{h,k}^j A_h^\theta \Pi_h\right\|^2 &+ k\sum_{i=0}^{j-1}\left\|QS_{h,k}^{j-i}A_h^\theta\Pi_h\right\|^2 \\ &+ k\sum_{i=0}^{j-1}\left\|(P_{h,k}^i)^{\frac{1}{2}}\Pi_h C S_{h,k}^{j-i}A_h^\theta\Pi_h\right\|^2_{\mathcal{L}(H;\mathcal{L}_2(H))} \\ &+ k^2\sum_{i=0}^{j-1}\left\|\left(P_{h,k}^{i+1}\right)^{\frac{1}{2}} A_h S_{h,k}^{j-i}A_h^\theta\Pi_h\right\|^2,\end{aligned}$$

and further

$$\begin{aligned}\left\|\left(P_k^j\right)^{\frac{1}{2}} A_h^\theta \Pi_h x\right\|^2 \leq \|G\|^2\left\|S_{h,k}^j A_h^\theta\Pi_h\right\|^2 &+ k\sum_{i=0}^{j-1}\|Q\|^2\left\|S_{h,k}^{j-i}A_h^\theta\Pi_h\right\|^2 \\ &+ k\sum_{i=0}^{j-1}\left\|\left(P_{h,k}^i\right)^{\frac{1}{2}} A_h^\theta\right\|^2\left\|A_h^{-\theta}CS_{h,k}^{j-i}A_h^\theta\Pi_h\right\|^2_{\mathcal{L}(H;\mathcal{L}_2(H))} \\ &+ k^2\sum_{i=0}^{j-1}\left\|\left(P_{h,k}^{i+1}\right)^{\frac{1}{2}} A_h^\theta\Pi_h\right\|^2\left\|A_h^{1-\theta}S_{h,k}^{j-i}A_h^\theta\Pi_h\right\|^2.\end{aligned}$$

We know that $(1+kA_h)^{-n}$ and A_h^θ commute for $n \in \mathbb{N}$ and any $\theta \in \mathbb{R}$. We further want to bring the above estimate into an explicit form. We split the sum and apply $k \leq B_2 h^{2+\nu}$ to the $(j-1)$-th summand.

$$\begin{aligned}&k^2\left\|\left(P_{h,k}^j\right)^{1/2} A_h^\theta\Pi_h\right\|^2\|A_h S_{h,k}^1\Pi_h\|^2 \\ &\quad\leq B_2 h^{4+2\nu}\left\|\left(P_{h,k}^j\right)^{1/2} A_h^\theta\Pi_h\right\|^2\|A_h\Pi_h\|^2 \leq B_1B_2h^{2\nu}\left\|\left(P_{h,k}^j\right)^{1/2} A_h^\theta\Pi_h\right\|^2.\end{aligned}$$

The last step follows by Assumption 1.3. We apply this to the previous inequality and change the index of the last sum $\sum_{i=0}^{j-2} \rightsquigarrow \sum_{i=1}^{j-1}$, we get

$$\begin{aligned}\left(1-B_1B_2h^{2\nu}\right)\left\|\left(P_k^j\right)^{\frac{1}{2}} A_h^\theta\Pi_h\right\|^2 \leq \|G\|^2\left\|A_h^\theta S_{h,k}^j\Pi_h\right\|^2 &+ k\sum_{i=0}^{j-1}\|Q\|^2\left\|A_h^\theta S_{h,k}^{j-i}\Pi_h\right\|^2 \\ &+ k\sum_{i=0}^{j-1}\left\|(P_{h,k}^i)^{\frac{1}{2}} A_h^\theta\right\|^2\left\|A_h^{-\theta}\Pi_h C A_h^\theta S_{h,k}^{j-i}\Pi_h\right\|^2 \\ &+ k^2\sum_{i=1}^{j-1}\left\|(P_{h,k}^i)^{\frac{1}{2}} A_h^\theta\Pi_h\right\|^2\left\|A_h S_{h,k}^{j-i-1}\Pi_h\right\|^2.\end{aligned}$$

We can divide by $(1 - B_1 B_2 h^{2\nu})$, because $B_1 B_2 h^{2\nu} < 1$. This yields the maximal spatial step-size η from Lemma 2.3. We further apply smoothing property (1.6) to this inequality. The following inequality holds true.

$$\begin{aligned}\left\| \left(P_k^j\right)^{\frac{1}{2}} A_h^\theta \Pi_h \right\|^2 \leq & M t_j^{-2\theta} + Mk \sum_{i=0}^{j-1} t_{j-i}^{-2\theta} + Mk \sum_{i=0}^{j-1} \left\| \left(P_{h,k}^i\right)^{\frac{1}{2}} A_h^\theta \right\|^2 \left\| A_h^{-\theta} \Pi_h C \right\|^2 t_{j-i}^{-2\theta} \\ & + Mk^2 \sum_{i=1}^{j-1} \left\| \left(P_{h,k}^i\right)^{\frac{1}{2}} A_h^\theta \Pi_h \right\|^2 \left\| A_h S_{h,k}^{j-i-1} \Pi_h \right\|^2 .\end{aligned}$$

Note that this is only possible if $\theta < 1/2$, as otherwise the smoothing property can not be applied. We interpret the second summand as a Riemann sum and bound it by an integral.

$$Mk \sum_{i=0}^{j-1} t_{j-i}^{-2\theta} \leq M \int_0^{t_j} (t_j - s)^{-2\theta} \mathrm{d}s = M t_j^{1-2\theta} \leq M.$$

In the third summand we insert $A^\beta A^{-\beta}$. Since $\theta \geq \beta$ we get

$$\left\| A_h^{-\theta} \Pi_h A^\beta A^{-\beta} C \right\|^2 \leq \left\| A_h^{\beta-\theta} \right\|^2 \left\| A_h^{-\beta} \Pi_h A^\beta \right\|^2 \left\| A^{-\beta} C \right\|^2 \leq M.$$

In order to use Gronwall's Lemma we need to preserve a factor of k in the last summand. We interject $A_h^{-\frac{1}{2}+\frac{\nu}{4}} A_h^{\frac{1}{2}-\frac{\nu}{4}}$ and use $k < Mh^{2+\nu}$, to get

$$Mk \left\| A_h S_{h,k}^{j-i} \Pi_h \right\|^2 \leq Mh^{2+\nu} \left\| A_h^{\frac{1}{2}+\frac{\nu}{4}} \Pi_h \right\|^2 \left\| A_h^{\frac{1}{2}-\frac{\nu}{4}} S_{h,k}^{j-i} \Pi_h \right\|^2 \leq Mh^{2-2+\nu-\nu} t_{j-i}^{\frac{\nu}{2}-1} \leq M t_{j-i}^{\frac{\nu}{2}-1}.$$

As long as $\nu > 0$, this singularity is not too strong for the application of Gronwall's lemma. For $\theta \geq \beta$ we now have

$$\left\| \left(P_{h,k}^j\right)^{\frac{1}{2}} A_h^\theta \Pi_h \right\|^2 \leq M t_j^{-2\theta} + M + Mk \sum_{i=0}^{j-1} t_{j-i}^{-2\theta \wedge (\frac{\nu}{2}-1)} \left\| \left(P_{h,k}^i\right)^{\frac{1}{2}} A_h^\theta \Pi_h \right\|^2, \quad j \in \mathcal{N}_k$$

The application of Gronwall's Lemma 1.6 yields

$$\left\| \left(P_{h,k}^j\right)^{\frac{1}{2}} A_h^\theta \Pi_h \right\|^2 \leq M t_j^{-2\theta}, \quad j \in \mathcal{N}_k.$$

The constant is independent of step-sizes. It follows that

$$\left\| A_h^\theta P_{h,k} A_h^\theta \right\| \leq \left\| \left(P_{h,k}^j\right)^{1/2} A_h^\theta \right\|^2 \leq M t_j^{-2\theta}, \quad j \in \mathcal{N}_k.$$

This completes the first part of this proof.

Part 2. In the second part we treat the cases of $\theta \leq \beta$. We follow the procedure of the first part of this proof until we make the following estimate.

$$\sum_{i=0}^{j-1} \left\| \left(P_{h,k}^i\right)^{\frac{1}{2}} \Pi_h C S_{h,k}^{j-i} A_h^\theta \Pi_h \right\|^2 \leq \sum_{i=0}^{j-1} \left\| \left(P_{h,k}^i\right)^{\frac{1}{2}} A_h^\theta \right\|^2 \left\| A_h^{-\theta} C S_{h,k}^{j-i} A_h^\theta \Pi_h \right\|^2 .$$

Instead we use the following inequality.

$$\sum_{i=0}^{j-1} \left\| \left(P_{h,k}^i\right)^{\frac{1}{2}} \Pi_h C S_{h,k}^{j-i} A_h^\theta \Pi_h \right\|^2 \leq \sum_{i=0}^{j-1} \left\| \left(P_{h,k}^i\right)^{\frac{1}{2}} A_h^\beta \right\|^2 \left\| A_h^{-\beta} C S_{h,k}^{j-i} A_h^\theta \Pi_h \right\|^2.$$

We continue with the procedure from the first part, with respect to the above change, until we get

$$\begin{aligned}\left\| \left(P_{h,k}^j\right)^{\frac{1}{2}} A_h^\theta \Pi_h \right\|^2 \leq & M t_j^{-2\theta} + M + Mk \sum_{i=0}^{j-1} t_{j-i}^{-2\theta} \left\| \left(P_{h,k}^i\right)^{\frac{1}{2}} A_h^\beta \Pi_h \right\|^2 \\ & + Mk \sum_{i=0}^{j-1} t_{j-i}^{\frac{\nu}{2}-1} \left\| \left(P_{h,k}^i\right)^{\frac{1}{2}} A_h^\theta \Pi_h \right\|^2.\end{aligned}$$

We insert the bound $\left\| \left(P_{h,k}^j\right)^{1/2} A_h^\beta \right\|^2 \leq M t_j^{-2\beta}$ from the first part of this proof, this yields

$$\left\| \left(P_{h,k}^j\right)^{\frac{1}{2}} A_h^\theta \Pi_h \right\|^2 \leq M t_j^{-2\theta} + M + Mk \sum_{i=0}^{j-1} t_{j-i}^{-2\theta} t_i^{-2\beta} + Mk \sum_{i=0}^{j-1} t_{j-i}^{\frac{\nu}{2}-1} \left\| \left(P_{h,k}^i\right)^{\frac{1}{2}} A_h^\theta \Pi_h \right\|^2.$$

Similar to the proof of Lemma 1.6, we us the Beta function (1.12) and get

$$\left\| \left(P_{h,k}^j\right)^{\frac{1}{2}} A_h^\theta \Pi_h \right\|^2 \leq M t_j^{-2\theta} + M + M\mathrm{B}(1-2\theta, 1-2\beta) + Mk \sum_{i=0}^{j-1} t_{j-i}^{\frac{\nu}{2}-1} \left\| \left(P_{h,k}^i\right)^{\frac{1}{2}} A_h^\theta \Pi_h \right\|^2.$$

The application of Gronwall's Lemma 1.6 finally yields

$$\left\| A_h^\theta P_{h,k} A_h^\theta \right\| \leq \left\| \left(P_{h,k}^j\right)^{1/2} A_h^\theta \right\|^2 \leq M t_j^{-2\theta}, \quad j \in \mathcal{N}_k,\ \theta \in [0, \beta).$$

In conclusion this estimate is satisfied for every $\theta \geq 0$ under the assumption of positivity of the discrete solution.

Part 3. In the third part of the proof we use an inductive argument to show that $P_{h,k}^j$ is positive for every $j \in \mathcal{N}_k^0$. From the second part of the proof we get a constant $D(T) < \infty$, such that $\|P_{h,k}^i\| \leq D(T)$ for all $i \in \mathcal{N}_k^0$ under the assumption of positivity. In particular it holds true that if $P_{h,k}^i$ is positive for every $i \leq j$, then there exists a constant M such that

$$\|P_{h,k}^j\| \leq M \leq D(T).$$

We show that if $kD(T) \leq 1$ and $P_{h,k}^i$ is positive for all $i < j$ then $P_{h,k}^j$ is also positive. Note that the base case $j = 0$ is true, since $\Pi_h G^* G \Pi_h$ is positive. Since $P_{h,k}^{j-1}$ and $P_{h,k}^j$ solve equation (2.4) by definition, it holds that

$$P_{h,k}^j + kA_h P_{h,k}^j + kP_{h,k}^j A_h = P_{h,k}^{j-1} + kQ^*Q + kC^* P_{h,k}^{j-1} C - k(P_{h,k}^{j-1})^2. \tag{2.12}$$

First we show that under the assumption of Lemma 2.3 the operator $R := P_{h,k}^{j-1} - k(P_{h,k}^{j-1})^2$ is positive. We have that R is a self-adjoint linear operator on a finite dimensional subspace $V_h^1 \subset H$. Therefore R is diagonalizable. Hence R is positive, if it has only positive eigenvalues. Let (x, λ) be any eigenpair of $P_{h,k}^{j-1}$, then $Rx = \lambda(1 - k\lambda)x$. Since the operator norm coincides

with the largest eigenvalue we have that $\lambda \leq D(T)$ and subsequently $(1 - k\lambda) \geq 0$ since $kD(T) \leq 1$. Therefore the right hand side of the above equation (2.12) is positive. Subsequently we have

$$0 \leq \langle P^j_{h,k}x, x\rangle + k\langle A_h P^j_{h,k}x, x\rangle + k\langle P^j_{h,k}A_h x, x\rangle = \langle (P^j_{h,k} + 2kAP^j_{h,k})x, x\rangle \quad \forall x \in V^1_h.$$

Now assume $P^j_{h,k}$ is not positive and hence $P^j_{h,k}$ has a negative eigenvalue $\bar{\lambda}$ to $\bar{x} \in V^1_h$, then

$$0 \leq \langle (P^j_{h,k} + 2kA_h P^j_{h,k})\bar{x}, \bar{x}\rangle = \bar{\lambda}\langle (1 + 2kA_h)\bar{x}, \bar{x}\rangle < 0.$$

In conclusion $P^j_{h,k}$ is positive. Therefore the induction step holds true. It holds true that $P^j_{h,k}$ is positive as long as preceding $P^i_{h,k}$ are positive and therefore satisfies the a-priori bound independent of step-sizes. This completes the proof of Lemma 2.3.

□

Theorem 2.4. *Let $(P(t))_{t\in[0,T]} \subset \mathcal{L}(H)$ be the mild solution to (1.18) and $(P^j_{h,k})_{0\leq j\leq N_k}$ the solution to (2.6). Let $B, \nu > 0$ be positive constants. For every $\epsilon \in (0, 1 - 2\beta)$ there exists a constant M such that*

$$\|P^j_{h,k}\Pi_h - P(t_j)\|_{\mathcal{L}(H_{-\beta};H_\beta)} \leq M t_j^{-\eta}\left(h^{2-4\beta-2\epsilon} + k^{1-2\beta-\epsilon}\right), \quad \forall j \in \mathcal{N}_k,\ \forall h \in (0,1],\ k \in (0, Bh^{2+\nu}],$$

where $\eta := (2\beta) \vee (1 - 2\epsilon)$. The constant M is independent of step-size k and h.

Proof. We compare the discrete solution with the mild solution (1.19) in the $\mathcal{L}(H_{-\beta}; H_\beta)$-norm. The difference can be written

$$\begin{aligned}
\left(P^j_{h,k}\Pi_h - P(t_j)\right)x &= \left(S^j_{h,k}\bar{P}(0)\Pi_h S^j_{h,k}\Pi_h - S(t_j)P(0)S(t_j)\right)x \\
&+ \int_0^{t_j}\left(\bar{S}(t_j - s)\Pi_h Q^*Q\bar{S}(t_j - s)\Pi_h - S(t_j - s)Q^*QS(t_j - s)\right)x\mathrm{d}s \\
&+ \int_0^{t_j}\left(\bar{S}(t_j - s)\Pi_h C^*\bar{P}(s)C\bar{S}(t_j - s)\Pi_h - S(t_j - s)C^*P(s)CS(t_j - s)\right)x\mathrm{d}s \\
&+ \int_0^{t_j}\left(\bar{S}(t_j - s)\bar{P}(s)^2\bar{S}(t_j - s)\Pi_h - S(t_j - s)P(s)^2S(t_j - s)\right)x\mathrm{d}s \\
&+ \sum_{i=0}^{j-1} S^{j-i}_{h,k}\left(k^2 A_h P^{i+1}_{h,k}A_h\right)S^{j-i}_{h,k}\Pi_h x.
\end{aligned}$$

The discretization error $\|P^j_{h,k}\Pi_h - P(t_j)\|_{\mathcal{L}(H_{-\beta};H_\beta)}$ decomposes via triangle inequality into the

following increments.

$$
\begin{aligned}
I_1 &:= \left\| S_{h,k}^j \bar{P}(0)\Pi_h S_{h,k}^j \Pi_h - S(t_j)P(0)S(t_j) \right\|_{\mathcal{L}(H_{-\beta};H_\beta)}, \\
I_2 &:= \left\| \int_0^{t_j} \bar{S}(t_j-s)\Pi_h Q^* Q \bar{S}(t_j-s)\Pi_h - S(t_j-s)Q^*QS(t_j-s)\mathrm{d}s \right\|_{\mathcal{L}(H_{-\beta};H_\beta)}, \\
I_3 &:= \left\| \int_0^{t_j} \bar{S}(t_j-s)\Pi_h C^* \bar{P}(s) C \bar{S}(t_j-s)\Pi_h - S(t_j-s)C^*P(s)CS(t_j-s)\mathrm{d}s \right\|_{\mathcal{L}(H_{-\beta};H_\beta)}, \\
I_4 &:= \left\| \int_0^{t_j} \bar{S}(t_j-s)\bar{P}(s)^2 \bar{S}(t_j-s)\Pi_h - S(t_j-s)P(s)^2S(t_j-s)\mathrm{d}s \right\|_{\mathcal{L}(H_{-\beta};H_\beta)}, \\
I_5 &:= \left\| \sum_{i=0}^{j-1} S_{h,k}^{j-i} \left(k^2 A_h P_{h,k}^{i+1} A_h\right) S_{h,k}^{j-i} \right\|_{\mathcal{L}(H_{-\beta};H_\beta)}.
\end{aligned}
$$

For the proof of this theorem we introduce the error operator.

$$E(t) := \bar{S}(t)\Pi_h - S(t), \quad t \in [0,T].$$

The aim of this proof is to apply Gronwall's Lemma 1.6. We prove suitable bounds for the above summands to make this application possible. The estimates for the semigroup discretization from Lemma 1.5 are the keys to this endeavor. We divide this proof into several parts which all deal with one of the above summands.

Part 1. Note once more that $P_{h,k}^j$ is self-adjoint. It follows that

$$
\begin{aligned}
I_1 =& \left\| S_{h,k}^j \Pi_h P(0) S_{h,k}^j \Pi_h - S(t_j)P(0)S(t_j) \right\|_{\mathcal{L}(H_{-\beta};H_\beta)} \\
=& \left\| A^\beta \bar{S}(t_j)\Pi_h \bar{P}(0)\bar{S}(t_j)\Pi_h A^\beta - A^\beta S(t_j)P(0)S(t_j)A^\beta \right\| \\
\leq& \left\| A^\beta \left(\bar{S}(t_j)\Pi_h - S(t_j)\right) P(0)\bar{S}(t_j)\Pi_h A^\beta \right\| + \left\| A^\beta S(t_j)P(0)\left(\bar{S}(t_j)\Pi_h - S(t_j)\right)A^\beta \right\| \\
\leq& M t_j^{-\beta} \left(\left\| A^\beta E(t_j)P(0) \right\| + \left\| P(0)E(t_j)A^\beta \right\| \right) \\
\leq& 2M t_j^{-\beta} \left\| G \right\|^2 \left\| \left(\bar{S}(t_j)\Pi_h - S(t_j)\right) A^\beta \right\|.
\end{aligned}
$$

By Lemma 1.5 (iii), applied with $\rho = \beta$ and $\theta = 1 - 2\beta - \epsilon$, it holds for $\epsilon \in (0, 1-2\beta)$ that

$$\left\| A^\beta E(t_j) \right\| \leq M t_j^{-1+2\beta+\epsilon} \left(h^{2(1-2\beta-\epsilon)} + k^{1-2\beta-\epsilon} \right). \tag{2.13}$$

Using 2.13 we arrive at

$$I_1 \leq M t_j^{-1+\epsilon} \left(h^{2(1-2\beta-\epsilon)} + k^{1-2\beta-\epsilon} \right). \tag{2.14}$$

Part 2. For I_2 we have by triangle inequality

$$
\begin{aligned}
I_2 =& \left\| \int_0^{t_j} \bar{S}(t_j-s)\Pi_h Q^* Q \bar{S}(t_j-s)\Pi_h - S(t_j-s)Q^*QS(t_j-s)\mathrm{d}s \right\|_{\mathcal{L}(H_{-\beta};H_\beta)} \\
\leq& \left\| A^\beta \int_0^{t_j} E(t_j-s)Q^*Q\bar{S}(t_j-s)\Pi_h A^\beta \mathrm{d}s \right\| + \left\| A^\beta \int_0^{t_j} \bar{S}(t_j-s)Q^*QE(t_j-s)A^\beta \mathrm{d}s \right\|.
\end{aligned}
$$

It further holds by 2.13 that

$$\begin{aligned} I_2 \leq & M \int_0^{t_j} \left\| A^\beta E(t_j - s) \right\| \|Q^* Q\| \left\| \bar{S}(t_j - s) \Pi_h A^\beta \right\| \mathrm{d}s \\ \leq & M \int_0^{t_j} (t_j - s)^{-\beta} \left\| A^\beta E(t_j - s) \right\| \mathrm{d}s \\ \leq & M \int_{t_{j-1}}^{t_j} (t_j - s)^{-1+\epsilon} \left(h^{2(1-2\beta-\epsilon)} + k^{1-2\beta-\epsilon} \right) \mathrm{d}s. \end{aligned}$$

In conclusion for $\epsilon \in (0, 1-2\beta)$ we have

$$I_2 \leq \frac{M t_j^\epsilon}{\epsilon} \left(h^{2(1-2\beta-\epsilon)} + k^{1-2\beta-\epsilon} \right). \tag{2.15}$$

Part 3. We want to use the Hölder continuity for I_3, since it does not hold in $t = 0$, we decompose I_3 as follows

$$\begin{aligned} I_3 = & \left\| \int_0^{t_j} \bar{S}(t_j - s) \Pi_h C^* \bar{P}(s) \Pi_h C \bar{S}(t_j - s) \Pi_h - S(t_j - s) C^* P(s) C S(t_j - s) \mathrm{d}s \right\|_{\mathcal{L}(H_{-\beta}; H_\beta)} \\ \leq & \sum_{i=1}^{j-1} \int_{t_i}^{j-1} \left\| A^\beta \bar{S}(t_j - s) \Pi_h C^* (\bar{P}(s) \Pi_h - P(t_i)) C \bar{S}(t_j - s) \Pi_h A^\beta \right\| \mathrm{d}s \\ & + \sum_{i=1}^{j-1} \int_{t_i}^{t_{i+1}} \left\| A^\beta \bar{S}(t_j - s) \Pi_h C^* (P(t_i) - P(s)) C \bar{S}(t_j - s) \Pi_h A^\beta \right\| \mathrm{d}s \\ & + \sum_{i=1}^{j-1} \int_{t_i}^{t_{i+1}} \left\| A^\beta \left(\bar{S}(t_j - s) \Pi_h C^* P(s) C \bar{S}(t_j - s) \Pi_h - S(t_j - s) C^* P(s) C S(t_j - s) \right) A^\beta \right\| \mathrm{d}s \\ & + \int_0^k \left\| A^\beta \left(\bar{S}(t_j - s) \Pi_h C^* \bar{P}(s) \Pi_h C \bar{S}(t_j - s) \Pi_h - S(t_j - s) C^* P(s) C S(t_j - s) \right) A^\beta \right\| \mathrm{d}s \\ =: & I_{3,1} + I_{3,2} + I_{3,3} + I_{3,4}. \end{aligned}$$

For the first term we get the estimate

$$\begin{aligned} I_{3,1} \leq & \sum_{i=1}^{j-1} \int_{t_i}^{i+1} \left\| A^\beta \bar{S}(t_j - s) \Pi_h \right\|^2 \|C^* A^{-\beta}\|^2 \|A^\beta \bar{P}(s) \Pi_h - P(t_i)\| \mathrm{d}s \\ \leq & M \sum_{i=1}^{j-1} \int_{t_i}^{i+1} \|A^\beta (1 + k A_h)^{-j+i}\|^2 \|\bar{P}(t_i) \Pi_h - P(t_i))\|_{\mathcal{L}(H_{-\beta}; H_\beta)} \mathrm{d}s \\ \leq & M k \sum_{i=0}^{j-1} t_{j-i}^{-2\beta} \|P_{h,k}^i \Pi_h - P(t_i))\|_{\mathcal{L}(H_{-\beta}; H_\beta)}. \end{aligned}$$

The last inequality holds by the smoothing property (1.7). We use the local Hölder-continuity

from Theorem 1.10 to bound $I_{3,2}$. First, we bound $I_{3,2}$ as follows.

$$\begin{aligned} I_{3,2} \leq & \sum_{i=1}^{j-1} \int_{t_i}^{i+1} \|P(t_i) - P(s)\|_{\mathcal{L}(H_{-\beta};H_\beta)} \left\|A^{-\beta} C \bar{S}(t_j - s) \Pi_h A^\beta\right\|^2_{\mathcal{L}(H;\mathcal{L}_2(H;H))} \mathrm{d}s \\ \leq & \sum_{i=1}^{j-1} \int_{t_i}^{i+1} \|P(t_i) - P(s)\|_{\mathcal{L}(H_{-\beta};H_\beta)} \|C\|^2_{\mathcal{L}(H;\mathcal{L}_2(H;H_{-\beta}))} \|A^\beta \bar{S}(t_j - s) \Pi_h\|^2 \mathrm{d}s. \end{aligned}$$

By definition of $\bar{S}$, it holds that $\bar{S}(t_j - s) = (1 + kA_h)^{-(j-i)}$ for $s \in (t_i, t_{i+1})$. An application of the Hölder-continuity from Theorem 1.10 with $\xi = 1 - 2\beta - \epsilon$ and the estimate (1.7) yields

$$\begin{aligned} I_{3,2} \leq & M \sum_{i=1}^{j-1} \int_{t_i}^{t_{i+1}} \left\|A^\beta (1 + kA_h)^{-(j-i)}\right\|^2 \|P(t_i) - P(s)\|_{\mathcal{L}(H_{-\beta};H_\beta)} \mathrm{d}s \\ \leq & M \sum_{i=1}^{j-1} \int_{t_i}^{t_{i+1}} (t_j - t_i)^{-2\beta} t_i^{-1+\epsilon} (s - t_i)^{1-2\beta-\epsilon} \mathrm{d}s \\ \leq & M \sum_{i=1}^{j-1} (t_j - t_i)^{-2\beta} t_i^{-1+\epsilon} k^{2-2\beta-\epsilon}. \end{aligned}$$

Once more we use the trick of bounding the sum with the help of the Beta function defined in (1.12). We have

$$\begin{aligned} I_{3,2} \leq & M k^{1-2\beta-\epsilon} \int_0^{t_j} (t_j - s)^{-2\beta} s^{-1+\epsilon} \mathrm{d}s \\ \leq & M k^{1-2\beta-\epsilon} t_j^{-2\beta+\epsilon} \int_0^1 (1 - s)^{-2\beta} s^{-1+\epsilon} \mathrm{d}s \\ \leq & M k^{1-2\beta-\epsilon} t_j^{-2\beta} \mathrm{B}(\epsilon, 1 - 2\beta). \end{aligned}$$

Thus $I_{3,2} \leq M k^{1-2\beta-\epsilon} t^{-2\beta}$. We further have

$$\begin{aligned} I_{3,3} \leq & \sum_{i=1}^{j-1} \int_{t_i}^{t_{i+1}} \left\|A^\beta E(t_j - s) C^* P(s) C \bar{S}(t_j - s) \Pi_h A^\beta\right\| \mathrm{d}s \\ & + \sum_{i=1}^{j-1} \int_{t_i}^{t_{i+1}} \left\|A^\beta S(t_j - s) C^* P(s) C E(t_j - s) A^\beta\right\| \mathrm{d}s \\ \leq & M \sum_{i=1}^{j-1} \int_{t_i}^{t_{i+1}} (t_j - s)^{-\beta} \left\|A^\beta E(t_j - s) C^* P(s) C\right\| \mathrm{d}s. \end{aligned}$$

Since, by Theorem 1.10 it holds true that $\|A^\beta P(s)A^\beta\| \leq Ms^{-2\beta}$ we get

$$\begin{aligned}
I_{3,3} \leq & M\left(h^{2(1-2\beta-\epsilon)} + k^{1-2\beta-\epsilon}\right)\sum_{i=1}^{j-1}\int_{t_i}^{t_{i+1}}(t_j-s)^{-1+\epsilon}\|C^*P(s)C\|\,\mathrm{d}s \\
\leq & M\|C\|^2_{\mathcal{L}(H;\mathcal{L}_2(H;H_{-\beta}))}\left(h^{2(1-2\beta-\epsilon)} + k^{1-2\beta-\epsilon}\right)\sum_{i=0}^{j-1}\int_{t_i}^{t_{i+1}}(t_j-s)^{-1+\epsilon}\left\|A^\beta P(s)A^\beta\right\|\mathrm{d}s \\
\leq & M\left(h^{2(1-2\beta-\epsilon)} + k^{1-2\beta-\epsilon}\right)\sum_{i=1}^{j-1}\int_{t_i}^{t_{i+1}}(t_j-s)^{-1+\epsilon}s^{-2\beta}\mathrm{d}s \\
\leq & M\left(h^{2(1-2\beta-\epsilon)} + k^{1-2\beta-\epsilon}\right)t_j^{1-2\beta-\epsilon}\mathrm{B}(1-2\beta,\epsilon)s.
\end{aligned}$$

Finally we deal with $I_{3,4}$.

$$\begin{aligned}
I_{3,4} \leq & \int_0^k \left\|A^\beta\bar{S}(t_j-s)\Pi_h C^*\bar{P}(s)\Pi_h C\bar{S}(t_j-s)\Pi_h A^\beta\right\|\mathrm{d}s \\
& + \int_0^k \left\|A^\beta\bar{S}(t_j-s)\Pi_h C^*P(s)C\bar{S}(t_j-s)\Pi_h A^\beta\right\|\mathrm{d}s \\
& + \int_0^k \left\|A^\beta E(t_j-s)C^*P(s)C\bar{S}(t_j-s)\Pi_h A^\beta\right\|\mathrm{d}s \\
& + \int_0^k \left\|A^\beta\bar{S}(t_j-s)\Pi_h C^*P(s)CE(t_j-s)A^\beta\right\|\mathrm{d}s \\
& + \int_0^k \left\|A^\beta E(t_j-s)C^*P(s)CE(t_j-s)\Pi_h A^\beta\right\|\mathrm{d}s \\
= & I_{3,4,1} + I_{3,4,2} + I_{3,4,3} + I_{3,4,4} + I_{3,4,5}.
\end{aligned}$$

We use $k \leq Mh^{2+\nu}$ for $I_{3,4,1}$ to get

$$\begin{aligned}
I_{3,4,1} \leq & \int_0^k \left\|A^\beta S^j_{h,k}\Pi_h\right\|^2\|C\|^2_{\mathcal{L}(H;\mathcal{L}_2(H;H_{-\beta}))}\left\|A_h^{-\beta}\Pi_h A^\beta\right\|^2\left\|A_h^\beta\right\|^2\|\bar{P}(s)\|\mathrm{d}s \\
\leq & \int_0^k Mt_j^{-2\beta}h^{-4\beta}\mathrm{d}s = kMt_j^{-2\beta}h^{-4\beta} \\
\leq & Mt_j^{-2\beta}h^{2(1-2\beta)+\nu}.
\end{aligned}$$

We proceed with $I_{3,4,2}$. We use the spatial regularity from Theorem 1.10 in order to get

$$\begin{aligned}
I_{3,4,2} \leq & \int_0^k \left\|S^j_{h,k}\Pi_h A^\beta\right\|^2\|C\|^2_{\mathcal{L}(H;\mathcal{L}_2(H;H_{-\beta}))}\left\|A^\beta P(s)A^\beta\right\|\mathrm{d}s \\
\leq & \int_0^k Mt_j^{-2\beta}s^{-2\beta}\mathrm{d}s \\
\leq & Mt_j^{-2\beta}k^{1-2\beta}.
\end{aligned}$$

We notice that $I_{3,4,3} = I_{3,4,4}$ and make the following estimate

$$\begin{aligned}
I_{3,3,3} + I_{3,4,4} \leq & 2\int_0^k \left\|A^\beta E(t_j - s)\right\| \|C\|^2_{\mathcal{L}(H;\mathcal{L}_2(H;H_{-\beta}))} \left\|A^\beta P(s) A^\beta\right\| \left\|S^j_{h,k}\Pi_h A^\beta\right\| \mathrm{d}s \\
\leq & \int_0^k M\left(h^{2(1-2\beta-\epsilon)} + k^{1-2\beta-\epsilon}\right)(t_j - s)^{-1+\beta+\epsilon} s^{-2\beta} t_j^{-\beta} \mathrm{d}s \\
\leq & M\left(h^{2(1-2\beta-\epsilon)} + k^{1-2\beta-\epsilon}\right) t_j^{-\beta} \int_0^k (t_j - s)^{-1+\beta+\epsilon} s^{-2\beta} \mathrm{d}s \\
\leq & M\left(h^{2(1-2\beta-\epsilon)} + k^{1-2\beta-\epsilon}\right) t_j^{-2\beta+\epsilon} \mathrm{B}(1-2\beta, \beta+\epsilon).
\end{aligned}$$

Finally we have

$$\begin{aligned}
I_{3,4,5} \leq & \int_0^k \left\|A^\beta E(t_j - s)\right\|^2 \|C\|^2_{\mathcal{L}(H;\mathcal{L}_2(H;H_{-\beta}))} \left\|A^\beta P(s) A^\beta\right\| \mathrm{d}s \\
\leq & \int_0^k M\left(\left(h^{1-2\beta-\epsilon} + k^{\frac{1}{2}-\beta-\frac{\epsilon}{2}}\right)(t_j - s)^{-\frac{1}{2}+\frac{\epsilon}{2}}\right)^2 s^{-2\beta} \mathrm{d}s \\
\leq & M\left(h^{2(1-2\beta-\epsilon)} + k^{1-2\beta-\epsilon}\right) \int_0^k (t_j - s)^{-1+\epsilon} s^{-2\beta} \mathrm{d}s \\
\leq & M\left(h^{2(1-2\beta-\epsilon)} + k^{1-2\beta-\epsilon}\right) t_j^{-2\beta+\epsilon} \mathrm{B}(\epsilon, 1-2\beta).
\end{aligned}$$

In summary for every $\epsilon \in (0, 1-2\beta)$, there exists a constant M such that I_3 is bounded by the following expression.

$$I_3 \leq M\left(\left(h^{2(1-2\beta-\epsilon)} + k^{1-2\beta-\epsilon}\right) t_j^{-2\beta} + k\sum_{i=0}^{j-1} t_{j-i}^{-2\beta} \|P^i_{h,k}\Pi_h - P(t_i)\|_{\mathcal{L}(H_{-\beta};H_\beta)}\right). \tag{2.16}$$

Part 4. Regarding I_4, split the term once again into three parts.

$$\begin{aligned}
I_4 \leq & \sum_{i=0}^{j-1} \int_{t_i}^{t_{i+1}} \left\|\bar{S}(t_j - s)\bar{P}(s)^2 \bar{S}(t_j - s)\Pi_h - S(t_j - s)P(s)^2 S(t_j - s)\right\|_{\mathcal{L}(H_{-\beta};H_\beta)} \mathrm{d}s \\
\leq & \sum_{i=0}^{j-1} \int_{t_i}^{t_{i+1}} \left\|\bar{S}(t_j - s)\left(\bar{P}(s)^2 - \Pi_h P(t_i)^2\right)\bar{S}(t_j - s)\Pi_h\right\|_{\mathcal{L}(H_{-\beta};H_\beta)} \mathrm{d}s \\
& + \sum_{i=0}^{j-1} \int_{t_i}^{t_{i+1}} \left\|\bar{S}(t_j - s)\Pi_h\left(P(t_i)^2 - P(s)^2\right)\bar{S}(t_j - s)\Pi_h\right\|_{\mathcal{L}(H_{-\beta};H_\beta)} \mathrm{d}s \\
& + \sum_{i=0}^{j-1} \int_{t_i}^{t_{i+1}} \left\|\bar{S}(t_j - s)\Pi_h P(s)^2 \bar{S}(t_j - s)\Pi_h - S(t_j - s)P(s)^2 S(t_j - s)\right\|_{\mathcal{L}(H_{-\beta};H_\beta)} \mathrm{d}s \\
= & : I_{4,1} + I_{4,2} + I_{4,3}.
\end{aligned}$$

Once more by Theorem 1.10 and with the same arguments as for $I_{3,2}$ it holds true that

$$\begin{aligned}
I_{4,1} \leq & \sum_{i=0}^{j-1} \int_{t_i}^{t_{i+1}} \left\| A^\beta (1+kA_h)^{-j-i} \right\|^2 \left\| \bar{P}(s)^2 \Pi_h - P(t_i)^2 \right\| \mathrm{d}s \\
\leq & M \sum_{i=0}^{j-1} \int_{t_i}^{t_{i+1}} t_{j-i}^{-2\beta} \left(\|\bar{P}(s)\Pi_h\| + \|P(t_i)\| \right) \|\bar{P}(s)\Pi_h - P(t_i)\| \mathrm{d}s \\
\leq & M \left(\sup_{t_i \leq t_j} \|\bar{P}(t_i)\Pi_h\| + \sup_{t_i \leq t_j} \|P(t_i)\| \right) k \sum_{i=0}^{j-1} t_{j-i}^{-2\beta} \|\bar{P}(t_i)\Pi_h - P(t_i)\| \\
\leq & Mk \sum_{i=0}^{j-1} t_{j-i}^{-2\beta} \|\bar{P}(t_i)\Pi_h - P(t_i)\|.
\end{aligned}$$

Moreover

$$\begin{aligned}
I_{4,2} = & \sum_{i=0}^{j-1} \int_{t_i}^{t_{i+1}} \left\| \bar{S}(t_j - s)\Pi_h \left(P(t_i)^2 \Pi_h - P(s)^2 \right) \bar{S}(t_j - s)\Pi_h \right\|_{\mathcal{L}(H_{-\beta}; H_\beta)} \mathrm{d}s \\
\leq & \sum_{i=0}^{j-1} \int_{t_i}^{t_{i+1}} \|A^\beta \bar{S}(t_j - s)\Pi_h\|^2 \left\| P(t_i)^2 - P(s)^2 \right\| \mathrm{d}s \\
\leq & \sum_{i=0}^{j-1} \int_{t_i}^{t_{i+1}} \|A^\beta \bar{S}(t_j - s)\Pi_h\|^2 \left(\|(P(t_i) - P(s))P(t_i)\| + \|P(s)(P(s) - P(t_i)\| \right) \mathrm{d}s \\
\leq & \, 2 \sup_{s \in [0, t_j]} \|P(s)\| \sum_{i=0}^{j-1} \int_{t_i}^{t_{i+1}} \|A^\beta \bar{S}(t_j - s)\|^2 \|P(t_i) - P(s)\| \mathrm{d}s \\
\leq & M t_j^{-1+\epsilon} k^{1-2\beta-\epsilon}.
\end{aligned}$$

For the next term we use the same arguments as for I_2. We get

$$\begin{aligned}
I_{4,3} \leq & \sum_{i=0}^{j-1} \int_{t_i}^{t_{i+1}} \left\| E(t_j - s)P(s)^2 \bar{S}(t_j - s)\Pi_h \right\|_{\mathcal{L}(H_{-\beta}; H_\beta)} \mathrm{d}s \\
& + \sum_{i=0}^{j-1} \int_{t_i}^{t_{i+1}} \left\| S(t_j - s)P(s)^2 E(t_j - s) \right\|_{\mathcal{L}(H_{-\beta}; H_\beta)} \mathrm{d}s \\
\leq & M \int_0^{t_j} (t_j - s)^{-\beta} \left\| A^\beta E(t_j - s)P(s)^2 \right\| \mathrm{d}s \\
\leq & \, M \sup_{s \in [0, t_j]} \|P(s)\|^2 \left(h^{2(1-2\beta-\epsilon)} + k^{1-2\beta-\epsilon} \right) \int_0^{t_j} (t_j - s)^{-1+\epsilon} \mathrm{d}s \\
\leq & \, M \left(h^{2(1-2\beta-\epsilon)} + k^{1-2\beta-\epsilon} \right).
\end{aligned}$$

After summation I_4 satisfies

$$I_4 \leq M \left(h^{2-4\beta} + k^{1-2\beta-\epsilon} + k \sum_{i=0}^{j-1} t_{j-i}^{-2\beta} \|\bar{P}(t_i)\Pi_h - P(t_i)\| \right), \tag{2.17}$$

with $\epsilon \in (0, 1-2\beta)$ as above.

Part 5. In the statement we assumed that $k \leq Mh^{2+\nu}$, we have

$$\begin{aligned} I_5 =& k^2 \left\| \sum_{i=0}^{j-1} S_{h,k}^{j-i} \left(A_h P_{h,k}^{i+1} A_h \right) S_{h,k}^{j-i} \Pi_h \right\|_{\mathcal{L}(H_{-\beta};H_\beta)} \\ \leq& k^2 \sum_{i=0}^{j-1} \left\| A^\beta S_{h,k}^{j-i} A_h^{\frac{1}{2}-\frac{\epsilon}{4}-\beta} \Pi_h \right\|^2 \|A_h^{\beta+\frac{\epsilon}{2}}\|^2 \left\| A_h^{\frac{1}{2}-\frac{\epsilon}{4}} P_{h,k}^{i+1} A_h^{\frac{1}{2}-\frac{\epsilon}{4}} \Pi_h \right\| \\ \leq& Mh^{2+\nu-4\beta-2\epsilon} k \sum_{i=0}^{j-1} t_{j-i}^{-1+\epsilon} t_{i+1}^{-1+\epsilon}. \end{aligned}$$

This follows by (1.7) and Lemma 2.3. We use the Beta function (1.12) and write

$$I_5 \leq Mh^{2(1-2\beta-\epsilon)+\nu} k \sum_{i=0}^{j-1} t_{j-i}^{-1+\epsilon} t_i^{-1+\epsilon} \leq Mh^{2(1-2\beta-\epsilon)+\nu} t_j^{-1+2\epsilon} \mathrm{B}(\epsilon,\epsilon). \tag{2.18}$$

In summary we established that there exists a constant M, independent of the step-sizes h and k, such that for $\epsilon \in (0, 1-2\beta)$ and $\eta = (2\beta) \vee (1-2\epsilon) < 1$,

$$I \leq M(\epsilon) \left(\left(h^{2-4\beta-\epsilon} + k^{1-2\beta-\epsilon} \right) t_j^{-\eta} + k \sum_{i=0}^{j-1} t_{j-i}^{-2\beta} \|\bar{P}(t_i)\Pi_h - P(t_i)\|_{\mathcal{L}(H_{-\beta};H_\beta)} \right). \tag{2.19}$$

This estimate arises from summation of $I_1, \ldots, I_5$, i.e. the inequalities (2.14), (2.15), (2.16), (2.17), (2.18). Note that one has to pass over to the stronger norm in (2.17) to get above inequality. By Gronwall's Lemma 1.6 it holds true that

$$I \leq M t_j^{-\eta} \left(h^{2(1-2\beta-\epsilon)} + k^{1-2\beta-\epsilon} \right). \tag{2.20}$$

This completes the proof. □

2.3 Analysis of the Second Scheme

In the first section of this chapter we introduced a second scheme, which is defined by the equation (2.8) or rather (2.7). In this section let $\bar{P}, P_{h,k}$ denote the discrete solution to (2.7).

Lemma 2.5. *Let $(P_{h,k}^i)_{i=0}^{N_k} \subset V_h^2$ be the solution to* (2.7)*. Then*

$$\|A_h^\theta P_{h,k}^j A_h^\theta \Pi_h\| \leq M t^{-2\theta}, \quad \forall j \in \mathcal{N}_k,\ \forall \theta \in [0, 1/2),\ \forall k, h \in (0,1] \times (0,1],$$

for some constant M, which is independent of step-sizes $k, h \in (0,1]$. Further $P_{h,k}^i$ is positive for all $i \in \mathcal{N}_k^0$.

Proof. The proof of this Lemma is analogue to the proof of Lemma 2.3. Instead of what we get

for the first scheme, we get under positivity assumption

$$\begin{aligned}\left\|\left(P_{h,k}^{j}\right)^{\frac{1}{2}}A_h^{\theta}\Pi_h x\right\|^2 &= \left\langle P(0)S_{h,k}^{j-i}A_h^{\theta}\Pi_h x, S_{h,k}^{j-i}A_h^{\theta}\Pi_h x\right\rangle \\ &\quad - k\sum_{i=0}^{j-1}\left\langle P_{h,k}^{i}S_{h,k}^{j-i}A_h^{\theta}\Pi_h x, P_{h,k}^{i}S_{h,k}^{j-i}A_h^{\theta}\Pi_h x\right\rangle \\ &\quad + k\sum_{i=0}^{j-1}\left\langle P_{h,k}^{i}\Pi_h C S_{h,k}^{j-i}A_h^{\theta}\Pi_h x, \Pi_h C S_{h,k}^{j-i}A_h^{\theta}\Pi_h x\right\rangle_{\mathcal{L}_2(H)} \\ &\quad + k\sum_{i=0}^{j-1}\left\langle QS_{h,k}^{j-i}A_h^{\theta}\Pi_h x, QS_{h,k}^{j-i}A_h^{\theta}\Pi_h x\right\rangle.\end{aligned}$$

We proceed by omitting the negatively signed sum. After some transformation we apply Gronwall's Lemma 1.6. Because the implicit $k^2A_hP_k^jA_h$-term is omitted in this scheme, it is not necessary to assume $k \leq Mh^{2+\nu}$ nor is an upper bound for h needed. The proof for positivity is just slightly different. We again show that $R := P_{h,k}^{j-1} - k(P_{h,k}^{j-1})^2$ is positive for any small enough k. We then get the positivity by the recursion

$$P_{h,k}^{j} = S_{h,k}^{1}\left(P_{h,k}^{j-1} - k\big(P_{h,k}^{j-1}\big)^2 + k\Pi_h Q^*Q + k\Pi_h C^*P_{h,k}^{j-1}\Pi_h C\right)S_{h,k}^{1}, \quad j \in \mathcal{N}_k.$$

This completes the proof. □

Theorem 2.6. *Let $P \in \mathcal{L}(H)$ be the solution to (1.20) and $(P_{h,k}^{j})_{0\leq j\leq N_k}$ be the discrete solution to (2.7). Let $k \leq Mh^{2+\nu}$ for some positive constants M, ν. For every $\epsilon \in (0, 1-2\beta)$ there exists a constant M such that*

$$\|P_{h,k}^{j}\Pi_h - P(t_j)\|_{\mathcal{L}(H_{-\beta};H_\beta)} \leq Mt_j^{-2\beta}\left(h^{2(1-2\beta-\epsilon)} + k^{1-2\beta-\epsilon}\right).$$

The constant M is independent of step-size k and h.

Proof. Proceed exactly like in the proof of Theorem 2.4 until you estimate I_5, which we simply skip because it did not arise in the difference $P_{h,k}^{j} - P(t_j)$. In this case the difference is

$$\begin{aligned}\left(P_{h,k}^{j}\Pi_h - P(t_j)\right)x &= \left(S_{h,k}^{j}\bar{P}(0)\Pi_h S_{h,k}^{j}\Pi_h - S(t_j)P(0)S(t_j)\right)x \\ &\quad + \int_0^{t_j}\left(\bar{S}(t_j-s)\bar{P}(s)^2\bar{S}(t_j-s)\Pi_h - S(t_j-s)P(s)^2S(t_j-s)\right)x\mathrm{d}s \\ &\quad + \int_0^{t_j}\left(\bar{S}(t_j-s)\Pi_h C^*\bar{P}(s)C\bar{S}(t_j-s)\Pi_h - S(t_j-s)C^*P(s)CS(t_j-s)\right)x\mathrm{d}s \\ &\quad + \int_0^{t_j}\left(\bar{S}(t_j-s)\Pi_h Q^*Q\bar{S}(t_j-s)\Pi_h - S(t_j-s)Q^*QS(t_j-s)\right)x\mathrm{d}s.\end{aligned}$$

We note that in this lemma the prerequisite $k \leq Mh^{2+\nu}$ is in fact needed again in order to get the estimate for $I_{3,4,1}$, where we made the estimate

$$\begin{aligned}I_{3,4,1} &\leq \int_0^k \left\|A^{\beta}S_{h,k}^{j}\Pi_h\right\|^2\|C\|^2_{\mathcal{L}(H;\mathcal{L}_2(H;H_{-\beta}))}\left\|A_h^{-\beta}\Pi_h A^{\beta}\right\|^2\left\|A_h^{\beta}\right\|^2\|\bar{P}(s)\|\mathrm{d}s \\ &\leq \int_0^k Mt_j^{-2\beta}h^{-4\beta}\mathrm{d}s = Mt_j^{-2\beta} \\ &\leq Mt_j^{-2\beta}h^{2(1-2\beta)+\nu}.\end{aligned}$$

□

3 Numerical Experiment

3.1 Discrete equations for the first scheme

In this section we derive the discrete equations for the first scheme. We recall the example from the introduction. Let $H := L^2([0,1])$ and $A = -\Delta$, where Δ denotes the Laplacian. Then $\mathcal{D}(A) = H_0^1 \cap H^2$, where H^2 denotes the Sobolev space of square-integrable functions on $[0,1]$ with existing weak derivatives up to second order. Further let $G, Q := \mathrm{id}_H$ and C be given by the multiplication operator, i.e. the mapping $C : (\phi, \psi)(x) \mapsto \phi(x) \cdot \psi(x)$. Let $V_h \subset L^2([0,1])$ be the the finite element space of continuous piecewise linear functions. Let $(\phi_i)_{i=1}^{N_h} \subset V_h$ be the set of basis functions $\phi_i(x) = (1 - |ih - x|/h)_+$, $x \in [0,1]$, where $h(N_h + 1) = 1$. For every $\phi, \psi \in H$ we define the tensor product $(\phi \otimes \psi)\xi = \phi\langle\psi, \xi\rangle,\ \xi \in H$. The operator $P_{h,k}^\tau : V_h \to V_h$ can be written as

$$P_{h,k}^\tau = \sum_{i,j=1}^{N_h} p_{i,j}^\tau \, (\phi_i \otimes \phi_j)\,. \tag{3.1}$$

Note that a natural one to one map $f : \mathcal{L}(V_h) \ni X \mapsto Y \in \mathbb{R}^{N_h \times N_h}$ is defined by $(f(X))_{j,i} = \langle X\phi_i, \phi_j\rangle$. We recall the definition of the first scheme (2.1). We state it with respect to the finite element basis. Find a sequence $(P_{h,k}^u)_{u=0}^{N_k} \subset \mathcal{L}(V_h)$ such that for all $\ell, m \in \mathcal{N}_h^0, \tau \in \mathcal{N}_k$ it holds true that $P_{h,k}^0 = \langle G^*G\phi_m, \phi_\ell\rangle$ and

$$\begin{aligned}\langle P_{h,k}^\tau \phi_m, \phi_\ell\rangle - \langle P_h^{\tau\,-1}\phi_m, \phi_\ell\rangle + k\langle P_{h,k}^\tau \phi_m, A_h\phi_\ell\rangle + k\langle P_{h,k}^\tau \phi_\ell, A_h\phi_m\rangle \\ = k\langle Q\phi_m, Q\phi_\ell\rangle - k\langle P_{h,k}^{\tau\,-1}\phi_m, P_{h,k}^{\tau\,-1}\phi_\ell\rangle + k\langle P_{h,k}^{\tau\,-1}\Pi_h C\phi_m, C\phi_\ell\rangle_{\mathcal{L}_2(H;H)}.\end{aligned} \tag{3.2}$$

It is handy to write the discrete operators as matrices. We use the notation.

$$\begin{aligned}&\bar{P}_t := \left(p_{i,j}^\tau\right)_{j,i=1}^{N_h}, && \bar{Q} := \left(\langle Q\phi_i, Q\phi_j\rangle\right)_{j,i=1}^{N_h}, \\ &\bar{M} := \left(\langle\phi_i, \phi_j\rangle\right)_{j,i=1}^{N_h} && \bar{C}_k := \left(\langle C(\phi_m)\phi_i, \phi_j\rangle\right)_{j,i=1}^{N_h} \quad m \in \{1, \dots, N_h\}, \\ &\bar{A} := \left(\langle A\phi_i, \phi_j\rangle\right)_{j,i=1}^{N_h}.\end{aligned} \tag{3.3}$$

The matrix $\bar{M}$ is called the mass-matrix and $\bar{A}$ the stiffness matrix. All of the above matrices are self-adjoint. We use this notation and have

$$\begin{aligned}\left(\langle P_{h,k}^\tau \phi_m, \phi_\ell\rangle\right)_{\ell,m=1}^{N_h} &= \left(\sum_{i,j=1}^{N_h} p_{i,j}^\tau \langle(\phi_i \otimes \phi_j^*)(\phi_m), \phi_\ell\rangle\right)_{\ell,m=1}^{N_h} \\ &= \left(\sum_{i,j=1}^{N_h} p_{i,j}^\tau \langle\phi_j, \phi_m\rangle\langle\phi_i, \phi_\ell\rangle\right)_{\ell,m=1}^{N_h} = \bar{M}\bar{P}_t\bar{M}.\end{aligned} \tag{3.4}$$

We proceed similarly with the other terms of (3.2). For the second term we have

$$\left(\langle P_{h,k}^{\tau}\phi_m, A\phi_\ell\rangle\right)_{\ell,m=1}^{N_h} = \left(\sum_{i,j=1}^{N_h} p_{i,j}^{\tau}\langle\phi_j,\phi_m\rangle\langle\phi_i, A\phi_\ell\rangle\right)_{\ell,m=1}^{N_h} = \bar{A}\bar{P}_t\bar{M}. \tag{3.5}$$

The last equation holds true, because A is self-adjoint. Therefore the third term of (3.2) yields the matrix $\bar{M}\bar{P}_t\bar{A}$. We further have

$$\left(\langle P_{h,k}^{t}\phi_m, P_{h,k}^{t}\phi_\ell\rangle\right)_{\ell,m=1}^{N_h} = \left(\sum_{i,j=1}^{N_h} p_{i,j}^{\tau}\langle\phi_j,\phi_m\rangle\langle\phi_i, P_{h,k}^{\tau}\phi_\ell\rangle\right)_{\ell,m=1}^{N_h} = \bar{M}\bar{P}_t\bar{M}\bar{P}_t\bar{M}. \tag{3.6}$$

For any fixed orthnormal basis $(e_n)_{n\in I}$ of V_h we have

$$\begin{aligned}
\langle P_{h,k}^{t}\Pi_h C\phi_m, C\phi_\ell\rangle_{\mathcal{L}_2(H;H)} &= \sum_{n\in I}\left\langle P_{h,k}^{t}\Pi_h\left(C\phi_m\right)e_n, \left(C\phi_\ell\right)e_n\right\rangle \\
&= \sum_{n\in I}\sum_{i,j=1}^{N_h} p_{i,j}\left\langle\phi_j, (C\phi_m)e_n\right\rangle\left\langle\phi_i, (C\phi_\ell)e_n\right\rangle \\
&= \sum_{n\in I}\sum_{i,j=1}^{N_h} p_{i,j}\left\langle(C\phi_m)\phi_j, e_n\right\rangle\left\langle(C\phi_\ell)\phi_i, e_n\right\rangle \\
&= \sum_{i,j=1}^{N_h} p_{i,j}\left\langle(C\phi_m)\phi_j, \sum_{n\in I}\langle(C\phi_\ell)\phi_i, e_n\rangle\, e_n\right\rangle \\
&= \sum_{i,j=1}^{N_h} p_{i,j}\left\langle(C\phi_m)\phi_j, (C\phi_\ell)\phi_i\right\rangle.
\end{aligned}$$

Now, we use the notation and the above equalities in order to rewrite the whole equation (3.2). We receive

$$\begin{aligned}
&\bar{M}\bar{P}_t\bar{M} + k\bar{A}\bar{P}_t\bar{M} + k\bar{M}\bar{P}_t\bar{A} \\
&\quad = \bar{M}\bar{P}_{\tau-1}\bar{M} + k\bar{Q} - k\bar{M}\bar{P}_{\tau-1}\bar{M}\bar{P}_{\tau-1}\bar{M} + k\left(\sum_{i,j=1}^{N_h} p_{i,j}\left\langle(C\phi_m)\phi_j, (C\phi_\ell)\phi_i\right\rangle\right)_{\ell,m=1}^{N_h}.
\end{aligned} \tag{3.7}$$

We substitute $Y_t := \bar{M}\bar{P}_t\bar{M}$, this yields

$$\begin{aligned}
&\left(\frac{1}{2} + k\bar{A}\bar{M}^{-1}\right)Y_t + Y_t\left(\frac{1}{2} + k\bar{M}^{-1}\bar{A}\right) \\
&\qquad = Y_{\tau-1} + k\bar{Q} - kY_{\tau-1}\bar{M}^{-1}Y_{\tau-1} + k\left(\sum_{i,j=1}^{N_h} p_{i,j}\left\langle(C\phi_m)\phi_j, (C\phi_\ell)\phi_i\right\rangle\right)_{\ell,m=1}^{N_h}.
\end{aligned} \tag{3.8}$$

This is a Silvester equation. We use the Bartels-Stewart algorithm to solve this equation. For the Sylvester equation $R_1X - XR_2 = R_3$ this algorithm comprises using QR algorithms to

obtain R_1, R_2 in Schur form, and solving the resulting problem by back substitution. For more details see [6]. We define $R_1 := 1/2 + k\bar{A}\bar{M}^{-1}, R_2 = 1/2 + k\bar{M}^{-1}\bar{A}$ and R_3 to be the right side of (3.8). Let U_1, U_2 be unitary matrices such that we have the following Schur decomposition.

$$R_1' = U_1 R_1 U_1^*, \quad R_2' = U_2 R_2 U_2^*$$

Since U_1, U_2 are unitary we have that $U_1^* U_1 = U_2^* U_2 = \mathrm{id}$. We define $Y_t' := U_1 Y_t U_2^*$ and $R_3' := U_1 R_3 U_2^*$ and we get an equation, which is equivalent to (3.8).

$$\begin{aligned} & R_1' Y_t' + Y_t' R_2' = R_3' \\ \Leftrightarrow \quad & U_1 R_1 U_1^* U_1 Y_t U_2^* + U_1 Y_t U_2^* U_2 R_2 U_2^* = U_1 R_3 U_2^* \\ \Leftrightarrow \quad & U_1 R_1 Y_t U_2^* + U_1 Y_t R_2 U_2^* = U_1 R_3 U_2^* \\ \Leftrightarrow \quad & R_1 Y_t + Y_t R_2 = R_3 \end{aligned}$$

Note that in the case that both R_1 and R_2 are self-adjoint, and therefore normal, the respective Schur forms are diagonal matrices. This simplifies solving the above equation. Since we do not have that in general we use the *scipy.linalg.solve_sylvester* routine for our Python program. In order to further save computation time we can calculate $\left(\sum_{i,j=1}^{N_h} p_{i,j} \langle (C\phi_m)\phi_j, (C\phi_\ell)\phi_i \rangle\right)_{\ell,m=1}^{N_h}$ beforehand. We have

$$\begin{aligned} \sum_{i,j=1}^{N_h} p_{i,j} \langle (C\phi_m)\phi_j, (C\phi_\ell)\phi_i \rangle &= \sum_{i,j=1}^{N_h} p_{i,j} \int_0^1 \phi_i(x)\phi_j(x)\phi_\ell(x)\phi_m(x)\mathrm{d}x \\ &= \begin{cases} \frac{2h}{5} p_{\ell,\ell} + \frac{h}{30}(p_{\ell+1,\ell+1} + p_{\ell-1,\ell-1}) + \frac{2h}{20}(p_{\ell,\ell+1} + p_{\ell,\ell-1}), & \text{if } \ell = m \\ \frac{h}{20}(p_{\ell,\ell} + p_{\ell+1,\ell+1}) + \frac{2h}{30} p_{\ell,\ell+1}, & \text{if } m = \ell + 1 \\ \frac{h}{20}(p_{m,m} + p_{m+1,m+1}) + \frac{2h}{30} p_{m,m+1}, & \text{if } m = \ell - 1 \\ 0, & \text{if } |\ell - m| > 1 \end{cases} \end{aligned} \tag{3.9}$$

For the remaining non trivial cases we need $\bar{P}_{\tau-1} = \bar{M}^{-1} Y_{\tau-1} \bar{M}^{-1}$. We use the Schur decomposition of $\bar{M}$ to solve $\bar{M}\bar{P}_{\tau-1}\bar{M} = Y_{\tau-1}$. For an unitary matrix U, we have $U\bar{M}U^* U\bar{P}_{\tau-1}U^* U\bar{M}U^* = UY_{\tau-1}U^*$. Since the Schur form of the self-adjoint matrix $\bar{M}$ is diagonal, the problem of solving the resulting equation is benevolent.

3.2 Discrete equations for the second scheme

In this section we derive derive the discrete equations for the second scheme. Let $H := L^2([0,1])$ and $A = -\Delta$, where Δ denotes the Laplacian, $\mathcal{D}(A) = H_0^1 \cap H^2$. Let $Q, G := \mathrm{id}_H$ and C be given by the multiplication operator, i.e. the mapping $C : (\phi, \psi)(x) \mapsto \phi(x) \cdot \psi(x)$. Let $V_h \subset L^2([0,1])$ be the finite element space of continuous piecewise quadratic functions, then $V_h \subset \mathcal{D}(\Delta)$. Let $(\phi_i)_{i=1}^{N_h} \subset V_h$ be the set of basis functions, where $N_h := 2/h - 1$. The finite element functions are given for odd $i \in \mathcal{N}_h$ by

$$\phi_i(x) = \begin{cases} 1 - \frac{4}{h^2}(x - i\frac{h}{2})^2, & \text{if } x \in [(i-1)\frac{h}{2}, (i+1)\frac{h}{2}] \\ 0, & \text{otherwise;} \end{cases}$$

and for even $i \in \mathcal{N}_h$ by

$$\phi_i(x) = \begin{cases} 1 + \frac{3}{h}(x - i\frac{h}{2}) + \frac{2}{h^2}(x - i\frac{h}{2})^2, & \text{if } x \in [(\frac{i}{2} - 1)h, \frac{i}{2}h], \\ 1 - \frac{3}{h}(x - i\frac{h}{2}) + \frac{2}{h^2}(x - i\frac{h}{2})^2, & \text{if } x \in [\frac{i}{2}h, (\frac{i}{2} + 1)h], \\ 0, & \text{otherwise.} \end{cases}$$

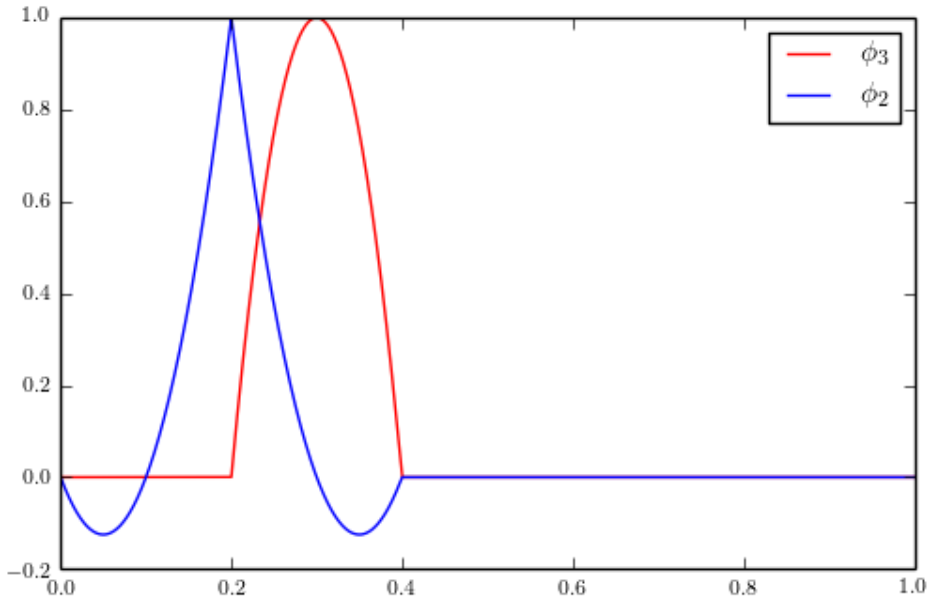

Figure 3.1: piecewise Quadratic Basis Elements for $N_h = 5$

We recall the definition of the second scheme (2.7). We state it with respect to the finite element basis. Find a sequence $(P^u_{h,k})_{u=0}^{N_k} \subset \mathcal{L}(V_h)$ such that for all $\ell, m \in \mathcal{N}_h^0, \tau \in \mathcal{N}_k$ it holds true that $P^0_{h,k} = \langle G^* G\phi_m, \phi_\ell\rangle$ and

$$\begin{aligned}\big\langle P^\tau_{h,k}\phi_\ell, \phi_m\big\rangle - \big\langle P^{\tau-1}_{h,k}\phi_\ell, \phi_m\big\rangle + k\big\langle P^\tau_{h,k}\phi_\ell, A_h\phi_m\big\rangle + k\big\langle P^\tau_{h,k}\phi_m, A_h\phi_\ell\big\rangle + k^2\langle P^\tau_{h,k}\Pi_h A_h\phi_\ell, A_h\phi_m\rangle \\ = k\big\langle Q\phi_\ell, Q\phi_m\big\rangle - k\big\langle P^{\tau-1}_{h,k}\phi_\ell, P^{\tau-1}_{h,k}\phi_m\big\rangle + k\big\langle P^{\tau-1}_{h,k}\Pi_h C\phi_\ell, C\phi_m\big\rangle_{\mathcal{L}_2(\mathcal{H};H)}. \quad (3.10)\end{aligned}$$

We use notation (3.3) and receive analogous to the previous section

$$\begin{aligned}&\bar{M}\bar{P}_\tau\bar{M} + k\bar{A}\bar{P}_\tau\bar{M} + k\bar{M}\bar{P}_\tau\bar{A} + k^2\bar{A}\bar{P}_\tau\bar{A} \\ &= \bar{M}\bar{P}_{\tau-1}\bar{M} + k\bar{Q} - k\bar{M}\bar{P}_{\tau-1}\bar{M}\bar{P}_{\tau-1}\bar{M} + k\left(\sum_{i,j=1}^{N_h} p_{i,j}\,\langle (C\phi_m)\phi_j, (C\phi_\ell)\phi_i\rangle\right)_{\ell,m=1}^{N_h}. \quad (3.11)\end{aligned}$$

Hence we have

$$\begin{aligned}&\left(\bar{M} + k\bar{A}\right)\bar{P}_\tau\left(\bar{M} + k\bar{A}\right) \\ &= \bar{M}\bar{P}_{\tau-1}\bar{M} + k\bar{Q} - k\bar{M}\bar{P}_{\tau-1}\bar{M}\bar{P}_{\tau-1}\bar{M} + k\left(\sum_{i,j=1}^{N_h} p_{i,j}\,\langle (C\phi_m)\phi_j, (C\phi_\ell)\phi_i\rangle\right)_{\ell,m=1}^{N_h}. \quad (3.12)\end{aligned}$$

Since $\bar{M} + k\bar{A}$ is self-adjoint, we can solve this efficiently using the Schur-decomposition of $\bar{M} + k\bar{A}$, which is a diagonal matrix. From this point the implementation is quite "straight forward". This scheme differs from the first scheme in the detail that we do not have to solve a Sylvester equation.

3.3 Numerical results

We want to study an experimental error of this scheme. Therefore, we compare some numerical solutions $(P_{h,k}^{N_k})_{(h,k)\in I}$,$I \subset (0,1)^2$, to a numerical solution $P_{\hat{h},\hat{k}}^{N_{\hat{k}}}, (\bar{h},\bar{k}) \in (0,1)^2$ with considerably smaller step-size $\hat{h} \ll h, \hat{k} \ll k, (h,k) \in I$. Using the above notation we get $\bar{P}_{h,k}^{N_k} \in \mathbb{R}^{N_h \times N_h}$. Subsequently, solutions based on different step-sizes are not ad-hoc comparable. We write ϕ_i^h for the i-th Lagrangian basis function for the finite element space V_h. We then define a generalized mass matrix

$$M^{(h_i,h_j)} := \left(\langle \phi_m^{h_i}, \phi_\ell^{h_j} \rangle\right)_{l\in\mathcal{N}_{h_j}, m\in\mathcal{N}_{h_i}}. \tag{3.13}$$

We have for fixed $h_1 > h_2$ and by (3.4)

$$\left(\langle P_{h_1,k}^\tau \phi_m^{h_2}, \phi_\ell^{h_2}\rangle\right)_{\ell,m=1}^{N_{h_2}} = \left(\sum_{i,j=1}^{N_{h_1}} p_{i,j}^\tau \langle \phi_j^{h_1}, \phi_m^{h_2}\rangle \langle \phi_i^{h_1}, \phi_\ell^{h_2}\rangle\right)_{\ell,m=1}^{N_{h_2}} = M^{(h_2,h_1)} \bar{P}_{h_1,k}^\tau M^{(h_1,h_2)}. \tag{3.14}$$

We then have for $h_1 > h_2, k_1 > k_2$ that

$$\begin{aligned} \left\|P_{h_1,k_1}^\tau - P_{h_2,k_2}^\tau\right\|_{\mathcal{L}(V_{h_2})} &= \sup_{x\in V_{h_2}\setminus\{0\}} \frac{\left\|(P_{h_1,k_1}^\tau - P_{h_2,k_2}^\tau)x\right\|_{V_{h_2}}}{\|x\|_{V_{h_2}}} \\ &= \sup_{x\in V_{h_2}\setminus\{0\}} \frac{1}{\|x\|_{V_{h_2}}} \left(\left\langle P_{h_1,k_1}^\tau x, P_{h_1,k_1}^\tau x\right\rangle - 2\left\langle P_{h_1,k_1}^\tau x, P_{h_2,k_2}^\tau x\right\rangle + \left\langle P_{h_2,k_2}^\tau x, P_{h_2,k_2}^\tau x\right\rangle\right)^{\frac{1}{2}}. \end{aligned} \tag{3.15}$$

We calculate $\left\langle P_{h_1,k_1}^\tau x, P_{h_2,k_2}^\tau x\right\rangle$ with $x = \sum_{i=1}^{N_{h_2}} v_i \phi_i$.

$$\begin{aligned} \left\langle P_{h_1,k_1}^\tau x, P_{h_2,k_2}^\tau x\right\rangle &= \sum_{\ell,m=1}^{N_{h_2}} v_\ell v_m \left\langle P_{h_1,k_1}^\tau \phi_m^{h_2}, P_{h_2,k_2}^\tau \phi_\ell^{h_2}\right\rangle \\ &= \sum_{\ell,m=1}^{N_{h_2}} \sum_{i,j=1}^{N_{h_1}} v_\ell v_m p_{i,j}^{h_1,k_1} \left\langle \phi_j^{h_1}, \phi_m^{h_2}\right\rangle \left\langle \phi_i^{h_1}, P_{h_2,k_2}^\tau \phi_\ell^{h_2}\right\rangle \\ &= \sum_{\ell,m=1}^{N_{h_2}} \sum_{n,o=1}^{N_{h_2}} \sum_{i,j=1}^{N_{h_1}} v_\ell v_m p_{i,j}^{h_1,k_1} p_{n,o}^{h_2,k_2} \left\langle \phi_j^{h_1}, \phi_m^{h_2}\right\rangle \left\langle \phi_i^{h_1}, \phi_n^{h_2}\right\rangle \left\langle \phi_o^{h_2}, \phi_\ell^{h_2}\right\rangle \\ &= v^\tau M^{(h_2,h_2)} \bar{P}_{h_2,k_2}^\tau M^{(h_2,h_1)} \bar{P}_{h_1,k_1}^\tau M^{(h_1,h_2)} v. \end{aligned}$$

The other terms can be calculated analogously. We have for the error

$$\begin{aligned}\left\|P^\tau_{h_1,k_1}-P^\tau_{h_2,k_2}\right\|_{\mathcal{L}(V_{h_2})} = \sup_{v\in\mathbb{R}^{N_{h_2}}\setminus\{0\}} \frac{1}{|v^\tau M^{h_2,h_2}v|^{1/2}}\Big|v^\tau\Big(&M^{(h_2,h_2)}\bar{P}^\tau_{h_2,k_2}M^{(h_2,h_2)}\bar{P}^\tau_{h_2,k_2}M^{(h_2,h_2)}\\ &- M^{(h_2,h_1)}\bar{P}^\tau_{h_1,k_1}M^{(h_1,h_2)}\bar{P}^\tau_{h_2,k_2}M^{(h_2,h_2)}\\ &- M^{(h_2,h_2)}\bar{P}^\tau_{h_2,k_2}M^{(h_2,h_1)}\bar{P}^\tau_{h_1,k_1}M^{(h_1,h_2)}\\ &+ M^{(h_2,h_1)}\bar{P}^\tau_{h_1,k_1}M^{(h_1,h_1)}\bar{P}^\tau_{h_1,k_1}M^{(h_1,h_2)}\Big)v\Big|^{1/2}\\ =: \Big(\sup_{v\in\mathbb{R}^{N_{h_2}}\setminus\{0\}} \frac{1}{|v^\tau M^{(h_2,h_2)}v|}&\big|v^\tau D((h_2,k_2),(h_1,k_1))v\big|\Big)^{1/2}.\end{aligned}$$

We further break this down into

$$\begin{aligned}\left\|P^\tau_{h_1,k_1}-P^\tau_{h_2,k_2}\right\|_{\mathcal{L}(V_{h_2})} &= \Big(\sup_{v\in\mathbb{R}^{N_{h_2}}\setminus\{0\}} \frac{|v^T(M^{h_2,h_2})^{-1/2}D(h_2,h_1)(M^{h_2,h_2})^{-1/2}v|}{|v^Tv|}\Big)^{1/2}\\ &= \|(M^{h_2,h_2})^{-1/2}D(h_2,h_1)(M^{h_2,h_2})^{-1/2}\|_{\mathcal{L}(\mathbb{R}^{N_{h_2}})}\\ &=: e(P^\tau_{h_2,k_2},P^\tau_{h_1,k_1}).\end{aligned}$$

We depict this as the *Error* in the below experiments. For two successive values of $(h_1,k_1),(h_2,k_2)\in I$ we define the *experimental order of convergence* (EOC) to be

$$EOC(h_1,h_2) := \frac{\log\Big(e\big(P^{N_{k_2}}_{h_2,k_2},P^{N_{k_3}}_{h_3,k_3}\big)\Big)-\log\Big(e\big(P^{N_{k_1}}_{h_1,k_1},P^{N_{k_3}}_{h_3,k_3}\big)\Big)}{\log(N_{h_1})-\log(N_{h_2})}, \tag{3.16}$$

where h_3,k_3 denote benchmark step-sizes.

We conduct one experiment for each type of finite element basis functions. In each experiment we approximate the operator Riccati equation and display the errors produced by different degrees of freedom with respect to a fixed benchmark solution. We take the experimental setup from the introduction, we further specify the following: We set the final temporal point to be $T=0.2$, this merely reduces the number of temporal steps which we simulate. We fix the step-sizes $h=k$, disregarding the analytic conditions on h,k from Theorem 2.4. We also choose benchmark solutions with $N_h=1499$ *degrees of freedom* for both experiments because we do not have any exact solution at hand. We compare the degrees of freedom $N_h=\{9,19,39,79,159\}$. This means that the spatial step-size of each solution is half the step-size of the preceding solution. We use a 2,6 GHz machine with 7.6 GiB memory for our simulations. Given the above setting we need to compute $K=\lfloor(N_h+1)*0.2\rfloor$ time-steps for the piecewise linear functions case, whereas it is necessary to compute $K=\lfloor((N_h+1)/2)*0.2\rfloor$ time-steps for the piecewise quadratic functions case. We recall that for the same spatial step size h the dimension of V^1_h is about half that of V^2_h. In the first experiment we compute the errors for the first scheme with respect to a finite element space of piecewise linear functions, see Table 3.1 and Figure 3.3. In the second experiment we compute the errors for both schemes with respect to a finite element space of piecewise quadratic basis functions, see Table 3.2 and Figure 3.2. For the second experiment we create the benchmark solution with the second scheme.

	1st Scheme		
$N_h + 1$	Error	EOC	Time/Step (s)
10	1.10756e-01	-	3.62873e-04
20	5.65968e-02	0.96860	7.57933e-04
40	2.72603e-02	1.05391	2.56515e-03
80	1.28485e-02	1.08520	1.45240e-02
160	5.92741e-03	1.11613	1.20316e-01
320	2.57581e-03	1.20237	8.46205e-01

Table 3.1: Piecewise Linear Basis Functions, $N_h = 1499$-Benchmark Solution.

	1st Scheme			2nd Scheme		
$N_h + 1$	Error	EOC	Time/Step (s)	Error	EOC	Time/Step (s)
10	5.14806e-02	-	5.74112e-04	7.94467e-02	-	1.29223e-04
20	1.70083e-02	1.59779	1.40405e-03	3.84840e-02	1.04573	5.14984e-04
40	0.46385e-02	1.87451	4.37999e-03	1.72308e-02	1.15927	1.01995e-03
80	0.12650e-02	1.87453	1.98770e-02	0.75029e-02	1.19947	1.55687e-03
160	0.03781e-02	1.74227	2.02142e-01	0.31449e-02	1.25445	9.39894e-03
320	0.01286e-02	1.55634	1.93454e-00	0.12025e-02	1.38688	5.60880e-02

Table 3.2: Piecewise Quadratic Basis Functions, $N_h = 1499$-Benchmark Solution.

Our Analysis suggests that we can expect a rate of convergence of one in space and of one half in time. The first experiment backs an overall convergence rate of one for the first scheme with piecewise linear functions which is better than our initial expectation. Figure 3.3 shows a slope which resembles the Mh^1-line. Note that the computational time grows considerably for higher N_h. We remark that this increased convergence rate is unlikely the assymptotic rate, but is only visible in the refinement levels we consider. If this claim was wrong it would contradict well-known results from theory of weak convergence of SPDE. The second experiment suggests that the first scheme converges faster than the second one. We also observe that the computational time grows considerably faster for the first scheme. The difference in computational time can be explained by the fact that the second scheme is more elementary. Solving the Sylvester equation in the first scheme is heavier than the inversion in the second scheme. In the long run the computational time for the first scheme will grows so fast that the second scheme might be more practical even though memory is an issue in this case.

As we have chosen $h = k$ we expect, in fact, rate $1/2$ from the temporal convergence rate. We have tried in different ways to make the rate $1/2$ visible but did not manage to achieve it. For instance, we tried larger $T = 1$ or larger step sizes $k = ch$ for large h. It seems like the spatial error dominates the temporal error anyhow. We have no convincing explanation to this issue, but dare to state a weak conjecture that a finer benchmark solution might reveal the rate $1/2$. The *EOC* almost appears to grow for higher N_h and fixed benchmark solution, which led us to this guess. In [10] experimental orders of convergence of one and two were found for the Lie splitting and Strang splitting methods respectively for a multidimensional noise framework. Further see [7] for numerical experiments for large algebraic Riccati equations.

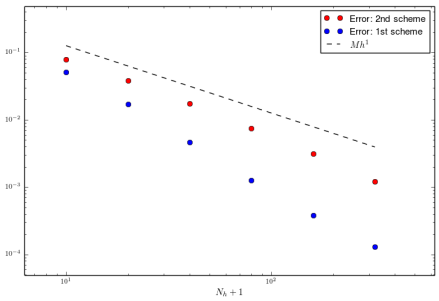

Figure 3.2: Piecewise Quadratic Basis Functions, Experimental Error

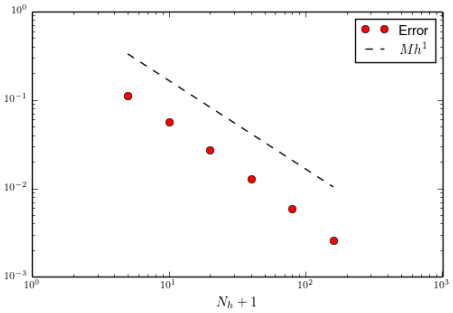

Figure 3.3: Piecewise Linear Basis Functions, Experimental Error

4 Conclusion

We conclude that the a-priori convergence rate shown in Theorem 2.4 and Theorem 2.6 was outmatched in our implementation of the example with the stochastic heat equation from the introduction. We do not attribute this to a better performance but rather to the experimental setting itself. The cause of this effect remains dubious to us. In this setting we notice that the second scheme is faster than the first one with piece-wise quadratic functions. The good news is that we make these observations without imposing the harsh conditions from the analysis in the second chapter. However in this analysis of the semi-implicit Euler scheme for the operator Riccati equation arising from distributed Control of SPDE in the semi-group framework we see the main contribution of this thesis. Making use of a discrete variations of constants formula we rigorously showed existence, uniqueness and convergence under rather strict conditions. To the best of our knowledge this is new for the given setting. We provided a blueprint of implementing this scheme. We see the introduction of the second scheme, which is new, as a nice addition to the main result. There are several interesting generalizations of the frame-work of this thesis. Especially the study of the boundary control problem would be a major improvement, since the distributed control framework seems rather unrealistic when looking for a "real life" application. As far as we can tell the difficulty in this is a more complicated quadratic term arising in the Riccati equation, which does not permit the here given argument for positivity. As mentioned before, the computational costs are quite high, in which we see a second area for improvement. We saw that even in the one dimensional case we get really large matrices. The given approach therefore demands a lot of memory when one wants to deal with a higher dimensional problem.

5 Appendix

5.1 Code

The below python code is based on the first scheme and the second scheme respectively. It was used to create Table 3.1, Table 3.2, Figure 3.2 and Figure 3.3. Note that the trick of inversion by Schur decomposition might yield bigger errors than using a routine such as *numpy.linalg.solve* or *numpy.linalg.inv*.

Listing 5.1: 1st Scheme

```
import numpy as np
import scipy.linalg as sc

#mass matrix
def Mf(N):
    h = 1.0/float(N)
    M = np.zeros((N-1,N-1))
    for i in range(N-1):
        M[i,i] = 2.0/3*h
    for i in range(N-2):
        M[i,i+1] = 1.0/6*h
        M[i+1,i] = M[i,i+1]
    return M

#stiffness matrix
a = 0.1
def Af(N):
    h = 1.0/float(N)
    A = np.zeros((N-1,N-1))
    for i in range(N-1):
        A[i,i] = 2.0/h
    for i in range(N-2):
        A[i,i+1] = -1.0/h
        A[i+1,i] = A[i,i+1]
    return a*A

#solver
def s(q,a,b,c,N):
    #spatial step-size
    h = 1.0/float(N)
    #temporal step-size
    k = h
    #number of temporal steps
```

```
    K = int(1.0/k)
    #mass matrix and stiffness matrix
    M = Mf(N)
    A = Af(N)
    #M**(-1)(.)M**(-1) with schur decomposition
    s1,s2 = sc.schur(M, output='real')
    def tran(E):
        X = np.dot(s2.transpose(),np.dot(E,s2))/(s1.diagonal()[:,
            np.newaxis]*s1.diagonal()[np.newaxis,:])
        return np.dot(s2,np.dot(X,s2.transpose()))
    #Silvester equation parameters
    I = np.identity(N-1)/2
    R2 = k*np.dot(tran(A),M)
    R2 = I+R2
    R1 = R2.transpose()
    #Define the solver for each temporal step
    def sylvester(q):
        return sc.solve_sylvester(R1, R2, q)
    #Compute solution
    P = M
    i = 1
    while i <= K:
        PT = tran(P)
        D1 = q*k*M+P-b*k*np.dot(P,np.dot(PT,M))
        D2 = np.zeros((N-1,N-1))
        D2[0,0] = h*(2*PT[0,0]/5+2*PT[0,1]/20+PT[1,1]/30)
        D2[N-2,N-2] = h*(2*PT[N-2,N-2]/5+PT[N-3,N-3]/30+2*PT[N-2,
            N-3]/20)
        for n in range(1,N-2):
            D2[n,n] = h*(PT[n,n]*(2/5)+2*PT[n,n+1]/20+PT[n+1,n
                +1]/30+PT[n-1,n-1]/30+2*PT[n,n-1]/20)
        for n in range(N-2):
            D2[n,n+1] = h*(PT[n,n]/20+2*PT[n,n+1]/30+PT[n+1,n
                +1]/20)
            D2[n+1,n] = D2[n,n+1]
        P = sylvester(D1+c*k*D2)
        i = i+1
    return P
```

Listing 5.2: 2nd Scheme

```
import numpy as np

#mass matrix
def Mf(N):
    h = 2.0/(N-1)
    M = np.zeros((N-2,N-2))
    for i in range(0,N-2,2):
        M[i,i] = 16
    for i in range(1,N-2,2):
```

```
        M[i,i] = 8
    for i in range(N-3):
        M[i,i+1] = 2
        M[i+1,i] = M[i,i+1]
    for i in range(1,N-4,2):
        M[i,i+2] = -1
        M[i+2,i] = M[i,i+2]
    return M*h/30

#stiffness matrix
def Af(a,N):
    h = 2.0/(N-1)
    print 1/h
    A = np.zeros((N-2,N-2))
    for i in range(0,N-2,2):
        A[i,i] = 16
    for i in range(1,N-2,2):
        A[i,i] = 14
    for i in range(N-3):
        A[i,i+1] = -8
        A[i+1,i] = A[i,i+1]
    for i in range(1,N-4,2):
        A[i,i+2] = 1
        A[i+2,i] = A[i,i+2]
    return a*A/(3*h)

#piece-wise quadratic basis elts
def phif(N,i,x):
    h = 2.0/(N-1)
    def f(i,x):
        if x >= (i-1)*h/2 and x <= (i+1)*h/2:
            return 1-(4/(h**2))*(x-i*h/2)**2
        else:
            return 0

    def g(i,x):
        if x >= (i/2 -1)*h and x <= h*i/2:
            return 1+(x-i*h/2)*3/h+(x-i*h/2)**2*2/h**2
        elif x >= h*i/2 and x <= (i/2+1)*h:
            return 1-(x-i*h/2)*3/h+(x-i*h/2)**2*2/h**2
        else:
            return 0
    #odd and even changed
    if i in range(0,N-2,2):
        return g(i,x)
    else:
        return f(i,x)

#integrals for tr(C^*P C)-term
```

```
def Cf(N):
    h = 2.0/(N-1)
    def phi(i,x):
        return phif(N,i,x)
    I = np.zeros((4,5))
    # 0->(2,2) 1->(2,3) 2->(2,4) 3->(3,3) 4->(4,4)
    I[0,0] = integrate.quad(lambda x: phi(2,x)*phi(2,x)*phi(2,x)*
       phi(2,x), (1-1)*h, (1+1)*h)[0]
    I[0,1] = integrate.quad(lambda x: phi(2,x)*phi(2,x)*phi(2,x)*
       phi(3,x), (1-1)*h, (1+1)*h)[0]
    I[0,2] = integrate.quad(lambda x: phi(2,x)*phi(2,x)*phi(2,x)*
       phi(4,x), (1-1)*h, (1+1)*h)[0]
    I[0,3] = integrate.quad(lambda x: phi(2,x)*phi(2,x)*phi(3,x)*
       phi(3,x), (1-1)*h, (1+1)*h)[0]
    I[0,4] = integrate.quad(lambda x: phi(2,x)*phi(2,x)*phi(4,x)*
       phi(4,x), (1-1)*h, (1+1)*h)[0]

    I[1,0] = integrate.quad(lambda x: phi(2,x)*phi(3,x)*phi(2,x)*
       phi(2,x), (1-1)*h, (1+1)*h)[0]
    I[1,1] = integrate.quad(lambda x: phi(2,x)*phi(3,x)*phi(2,x)*
       phi(3,x), (1-1)*h, (1+1)*h)[0]
    I[1,2] = integrate.quad(lambda x: phi(2,x)*phi(3,x)*phi(2,x)*
       phi(4,x), (1-1)*h, (1+1)*h)[0]
    I[1,3] = integrate.quad(lambda x: phi(2,x)*phi(3,x)*phi(3,x)*
       phi(3,x), (1-1)*h, (1+1)*h)[0]
    I[1,4] = integrate.quad(lambda x: phi(2,x)*phi(3,x)*phi(4,x)*
       phi(4,x), (1-1)*h, (1+1)*h)[0]

    I[2,0] = integrate.quad(lambda x: phi(2,x)*phi(4,x)*phi(2,x)*
       phi(2,x), (1-1)*h, (1+1)*h)[0]
    I[2,1] = integrate.quad(lambda x: phi(2,x)*phi(4,x)*phi(2,x)*
       phi(3,x), (1-1)*h, (1+1)*h)[0]
    I[2,2] = integrate.quad(lambda x: phi(2,x)*phi(4,x)*phi(2,x)*
       phi(4,x), (1-1)*h, (1+1)*h)[0]
    I[2,3] = integrate.quad(lambda x: phi(2,x)*phi(4,x)*phi(3,x)*
       phi(3,x), (1-1)*h, (1+1)*h)[0]
    I[2,4] = integrate.quad(lambda x: phi(2,x)*phi(4,x)*phi(4,x)*
       phi(4,x), (1-1)*h, (1+1)*h)[0]

    I[3,0] = integrate.quad(lambda x: phi(3,x)*phi(3,x)*phi(2,x)*
       phi(2,x), (1-1)*h, (1+1)*h)[0]
    I[3,1] = integrate.quad(lambda x: phi(3,x)*phi(3,x)*phi(2,x)*
       phi(3,x), (1-1)*h, (1+1)*h)[0]
    I[3,2] = integrate.quad(lambda x: phi(3,x)*phi(3,x)*phi(2,x)*
       phi(4,x), (1-1)*h, (1+1)*h)[0]
    I[3,3] = integrate.quad(lambda x: phi(3,x)*phi(3,x)*phi(3,x)*
       phi(3,x), (1-1)*h, (1+1)*h)[0]
    I[3,4] = integrate.quad(lambda x: phi(3,x)*phi(3,x)*phi(4,x)*
       phi(4,x), (1-1)*h, (1+1)*h)[0]
```

```
    return I

#solver
def s(N):
    #spatial step-size
    h = 2.0/(N-1)
    #temporal step-size
    k = h
    #number of temporal steps
    K = int(1.0/k)
    #overlapp/mass matrix
    M = Mf(N)
    #stiffness matrix
    A = Af(1,N)
    #integrals for tr(C^*P C)-term
    I = Cf(N)
    #S**(-1)(.)S**(-1) with schur decomposition
    s1,s2 = sc.schur(M, output='real')
    def tran(E):
        X = np.dot(s2.transpose(),np.dot(E,s2))/(s1.diagonal()[:,
            np.newaxis]*s1.diagonal()[np.newaxis,:])
        return np.dot(s2,np.dot(X,s2.transpose()))
    #(S+kA)**(-1)(.)(S+kA)**(-1) with schur decomposition
    s3,s4 = sc.schur(M+k*A, output='real')
    def sol(E):
        X = np.dot(s4.transpose(),np.dot(E,s4))/(s3.diagonal()[:,
            np.newaxis]*s3.diagonal()[np.newaxis,:])
        return np.dot(s4,np.dot(X,s4.transpose()))
    # Solve one step
    P = tran(M)
    j = 0
    while j <= K-1:
        D1 = k*M+np.dot(M,np.dot(P,M))-k*np.dot(M,np.dot(P,np.dot
            (M,np.dot(P,M))))
        #tr(C^*P C)-term
        D2 = k*np.zeros((N-2,N-2))
        for i in range(1,N-2,2):
            D2[i,i] = P[i,i]*(I[0,0]+4*I[0,1]+4*I[0,2]+2*I
                [0,3]+2*I[0,4])
            D2[i,i+1] = P[i,i+1]*(I[1,0]+2*I[1,1]+2*I[1,2]+I
                [1,3]+I[1,4])
            D2[i+1,i] = D2[i,i+1]
            D2[i-1,i] = P[i-1,i]*(I[1,0]+2*I[1,1]+2*I[1,2]+I
                [1,3]+I[1,4])
            D2[i,i-1] = D2[i-1,i]
        D2[1,1] = P[1,1]*(I[0,0]+2*I[0,1]+2*I[0,2]+1*I[0,3]+1*I
            [0,4])
        D2[N-4,N-4] =  P[N-4,N-4]*(I[0,0]+2*I[0,1]+2*I[0,2]+1*I
            [0,3]+1*I[0,4])
```

```
        for i in range(1,N-4,2):
            D2[i,i+2] = P[i,i+2]*(I[2,0]+2*I[2,1]+2*I[2,2]+I
                [2,3]+I[2,4])
            D2[i+2,i] = D2[i,i+2]
        for i in range(0,N-2,2):
            D2[i,i] = P[i,i]*(I[3,0]+2*I[3,1]+2*I[3,2]+I[3,3]+I
                [3,4])
        #solve for P(j+1)
        P = sol(D1+k*D2)
        j = j+1
    return np.dot(M,np.dot(P,M))
```

5.2 Zusammenfassung

Kontrolltheorie ist ein Gebiet der mahematischen Optimierung. Sie untersucht den Einfluss einer sogenannten *Kontrollfunktion* auf ein dynamisches System. Die Kontrolltheorie findet daher klassischer Weise viel Anwendung in den Natur- und Ingeneurswissenschaften. Ein klassisches Beispiel aus der Mechanik ist das inverse Pendel Problem. Das Segway oder die erste Stufe der SpaceX Falcon9 Rakete lassen sich beispielsweise als umgekehrtes Pendel verstehen. Dieses Problem zählt man zu den linear quadratischen Kontrollproblemen (LQ-Probleme). Die linearen quadratischen Kontrollprobleme werden durch ein kontrolliertes dynamisches System charakterisiert, welches durch eine lineare Differentialgleichung beschrieben werden kann. In Abhängigkeit von dieser Dynamik gilt es eine zeitsteitge quadratische Kostenfunktion abhängig von der Dynamik und der Kontrolle zu minimieren.

In dieser Arbeit interessieren wir uns für das folgende LQ-Problem. Wir betrachten die stochastische Wärmeleitungsgleichung mit multiplikativem Rauschen und Dirichlet Randbedingung. Sei $H := L^2([0,1])$ der Raum der integrablen Funktionen über $[0,1]$ und $\Delta = \frac{\partial^2}{\partial x^2}$ sei der Laplace-Operator. Die stochastische Wärmeleitungsgleichung ist geben durch die folgende stochastische partielle Differentialgleichung.

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}t}X(t,\xi) = \Delta X(t,\xi) + X(t,\xi)\cdot \dot{W}(t,\xi) \quad \forall t \in (0,T],\ \xi \in (0,1) \\ &X(t,0) = X(t,1) = 0 \quad \forall t \in (0,T], \\ &X(0) = x_0 \in H, \end{aligned} \tag{5.1}$$

wobei $x_0 \in H$ der Anfangswert ist und $\dot{W}$ weißes Rauschen in Ort und Zeit darstellt, definiert auf einem filtrierten Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$. Wir fügen diesem System eine Kontrolle (bzw. einen Regler) $u : [0,T]\times\Omega \to H$ hinzu. Wir erhalten die folgende kontrollierte stochastische Wäremleitungsgleichung:

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}t}X^u(t,\xi) = \Delta X^u(t) + u(t) + X^u(t,\xi)\cdot \dot{W}(t,\xi), \quad \forall t \in (0,T],\ \xi \in (0,1) \\ &X^u(t,0) = X^u(t,1) = 0 \quad \forall t \in [0,T], \\ &X^u(0) = x_0 \in H. \end{aligned} \tag{5.2}$$

In diesem Fall, da der Regler das Innere des Definitionsbereiches kontrolliert, sprechen wir von einem *verteiltem Kontrollproblem*. In Abgrenzung hirzu gibt es ferner das *Randkontrollproblem*, in diesem Fall wird die Dynamik nur auf dem Rand des Definitionsbereiches beeinflusst. Das Ziel ist es einen vorhersehbaren quadratintegrierbaren optimalen Regulator $u \in L^2([0,T]\times\Omega; H)$ zu finden, welcher die folgende quadratische Kostenfunktion minimiert:

$$J(u) = \mathbb{E}\left[\int_0^T (\|X^u(t)\|^2 + \|u(t)\|^2)\mathrm{d}t + \|X^u(T)\|^2\right] \tag{5.3}$$

Die Lösung dieses verteilten linearen quadratischen Kontrollproblems ist bereits bekannt. Sei $\mathcal{L}(H)$ der Raum der beschränkten linearen Operatoren auf H und $P : [0,T] \to \mathcal{L}(H)$ eine Operator-wertige Funktion. Die optimale Kontrolle ist gegeben durch

$$\bar{u}(t) := -P(T-t)X^{\bar{u}}(t), \quad t \in [0,T], \tag{5.4}$$

falls P die folgende *Riccati-Operator-Variationsgleichung* löst. Wir schreiben $H_0^1 \subset H$ für den Sobolev-Raum der quadratintegrierbaren Funktionen auf $[0,1]$ mit erster schwacher Ableitung

und Dirichlet Randbedingung. Wir gebrauchen die Notation $a(x,y) := \langle \nabla x, \nabla y\rangle_H = -\langle \Delta x, y\rangle_H$ für $x, y \in H_0^1$. Finde $(P(t))_{t\in(0,T]} \subset \mathcal{L}(H)$ mit $P(0) = \mathrm{id}_H$, sodass

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\langle P(t)x, y\rangle_H + a\left(P(t)x, y\right) + a\left(P(t)y, x\right) + \langle P(t)x, P(t)y\rangle_H \\ = \langle P(t)Cx, Cy\rangle_{\mathcal{L}_2(H)} + \langle x, y\rangle_H, \quad \forall x, y \in H_0^1, t \in (0,T]. \quad (5.5)\end{aligned}$$

In diesem Fall bezeichnet $\mathcal{L}_2(H)$ den Hilbertraum aller Hilbert-Schmidt-Operatoren und C den Multiplikationsoperator $(C(\phi)\psi)(x) = \phi(x)\cdot\psi(x),\ \phi, \psi \in H$.

Das Ziel dieser Arbeit ist es zwei semi-explizite numerische methoden zu studieren welche diese Funktion P und somit die optimale Kontrollfunktion $\bar{u}$ nähern. Dies ist unerlässlich für die praktische Anwendung. Wir binden dieses Beispiel in einen größeren theoretischen Rahmen ein, um die Konvergenz beider Methoden zu untersuchen. Das theoretische Fundament hierfür liefert ein Paper [1] von Andersson, Djehiche und Larsson, in welchem die Existenz- und Eindeutigkeitstheorie für allgemeinere Operator Riccati Gleichung entwickelt wird.

Sei $k \in (0,1]$ der Zeitschritt und $h \in (0,1]$ der Ortsschritt. Sei H^2 der Sobolevraum quadratintegrierbarer Funktionen mit zweiten schwachen Ableitungen. Sei V_h^1 der Finite-Elemente-Raum der stetigen stückweise linearen Funktionen und $V_h^2 \subset H^2 \cap H_0^1$ der Finite-Elemente-Raum der stückweise quadratischen Funktionen, jeweils auf einer Triagulation mit maximaler Gittergröße von h. Sei $\Pi_h^i, i \in \{1,2\}$ die orthogonale Projektion von H auf V_h^i. Die erste Methode ist definiert durch: Finde eine Folge $(P_{h,k}^j)_{j=0}^{N_k}$ mit $P_{h,k}^0 = \mathrm{id}_{V_h^1}$, sodass

$$\begin{aligned}&\left\langle P_{h,k}^j\phi, \psi\right\rangle + ka\big(P_{h,k}^j\phi, \psi\big) + ka\big(P_{h,k}^j\psi, \phi\big) + k\left\langle P_{h,k}^{j-1}\phi, P_{h,k}^{j-1}\psi\right\rangle \\ &\quad = \left\langle P_{h,k}^{j-1}\phi, \psi\right\rangle + k\left\langle P_{h,k}^{j-1}\Pi_h C\phi, C\psi\right\rangle_{\mathcal{L}_2(\mathcal{H};H)} + k\left\langle \phi, \psi\right\rangle, \quad \forall \phi, \psi \in V_h^1, j \in \{1,\dots,N_k\}. \quad (5.6)\end{aligned}$$

Die zweite Methode ist definiert durch: Finde eine Folge $(P_{h,k}^j)_{j=0}^{N_k}$ mit $P_{h,k}^0 = \mathrm{id}_{V_h^2}$, sodass

$$\begin{aligned}&\left\langle P_{h,k}^j\phi, \psi\right\rangle + ka\big(P_{h,k}^j\phi, \psi\big) + ka\big(P_{h,k}^j\psi, \phi\big) + k\left\langle P_{h,k}^{j-1}\phi, P_{h,k}^{j-1}\psi\right\rangle + k^2\langle P_{h,k}^j\Pi_h\Delta\phi, \Delta\psi\rangle \\ &\quad = \left\langle P_{h,k}^{j-1}\phi, \psi\right\rangle + k\left\langle P_{h,k}^{j-1}\Pi_h C\phi, C\psi\right\rangle_{\mathcal{L}_2(\mathcal{H};H)} + k\left\langle Q\phi, Q\psi\right\rangle, \quad \forall \phi, \psi \in V_h^2, j \in \{1,\dots,N_k\}. \\ &\qquad (5.7)\end{aligned}$$

Wir zeigen, dass die diskrete Lösung zu beiden Methoden in einer angemessenen Weise existiert und dass diese in jedem Zeitschritt positiv und beschränkt sind. Um dies zeigen zu können, fordern wir einige recht starke Annahmen für die erste Methode und etwas schwächere für die zweite. Weiter zeugen wir das beide Methoden eine a-priori Kovergenzordnung von einem $\gamma \in (0, 1/2)$ und eine Singularität in $t = 0$ von einer Ordnung $\delta \in (\frac{1}{2}, 1)$ haben. In beiden Fällen nutzen wir eine Kopplung des Zeit- und des Raumschritts. Wir wählen solche h, k, dass Konstanten $M, \nu > 0$ existieren, sodass $k \le Mh^{2+\nu}$.

Soweit es uns bekannt ist, gab es bereits Arbeiten von Benner und Mena zu dem deterministischen Problem und seiner numerischen Annäherung. Ergebnisse zu numerischen Lösungen für die Riccatigleichung für deterministische Kontrollprobleme können in [20] gefunden werden. Weitere Ergebnisse befinden sich in [7], worin effiziente Algorithmen für große Riccati Gleichungen untersucht werden. Große Riccatigleichungen nähern ihrerseits die integrale Riccatigleichung an, welche von Gibson gefunden wurde, siehe [14]. Der Ansatz von Gibson fand weitere Anwendung bei Banks und Kunich, siehe [5]. Abgesehen von der neuesten Arbeit von Levajković, Mena and Tuffaha, [18] und [19], sind uns keine weiteren Publikationen bekannt, welche den stochastische Fall und die Näherung der entsprechenden Riccatigleichung behandeln.

Daher gehen wir davon aus, dass es neu ist, dass wir die Variation der Konstanten zur Diskretisierung von Riccatigleichungen in Raum und Zeit benutzen. Soweit wir wissen ist dies außerdem die erste Arbeit in der Zeit- und Raum-diskerte numerische Methoden für die Riccatigleichung mit Hilfe von Halbgruppen und deren rigorose a-priori Kovergenzordnungen untersucht werden. Die zweite Methode ist neu und simpler zu implementieren. Wählt man für beide Methoden stückweise quadratische Funktionen zur Näherung, so verbessert die zweite Methode die Laufzeit im Vergleich zur ersten. Unsere Ergebnisse sind zum Teil Basis für ein zukünftiges Paper über finite Elemente-Näherung für Operator-Lyapunov-Gleichungen von Andersson, Lang, Pettersson und Schroer [3].

Diese Arbeit setzt sich aus drei Kapiteln zusammen. Im ersten Kapitel wiederholen wir die nötigen Vorraussetzungen und Werkzeuge für die spätere Fehleranalyse. Diese bestehen hauptsätzlich aus grundlegenden Ergebnissen aus der Theorie über stark stetige Halbgruppen und der Formulierung der relevanten Riccati-Operator-Gleichung. Eine rigorose Definition einer schwachen Lösung der Riccati-Operator-Gleichung geben wir in Bezug auf die Riccati-Operator-Variationsgleichung 1.18. Am Ende des ersten Kapitels zeigen wir, dass der gegebene theoretische Rahmen tatsächlich zu dem gewählten Szenario passt, siehe Lemma 1.11. Im zweiten Kapitel zeigen wir nützliche Neuformulierungen der beiden Methoden, die Existenz und Eindeutigkeit beider Methoden und beweisen deren Konvergenz. Für die Existenz siehe Lemmata 2.1 und 2.2. Für die Existenz der ersten Methode gebrauchen wir deren Verwandtschaft zur Sylvester Gleichung. Das Hauptwerkzeug im Beweis der Konvergenz ist Gronwalls Lemma, hierfür siehe Theorem 2.4 und Theorem 2.6. Im dritten und letzten Kapitel zeigen wir einige numerische Ergebnisse für das obige Beispiel der stochastischen Wärmeleitungsgleichung. Der entsprechende Python Code befindet sich im Appendix.

Bibliography

[1] Adam Andersson, Boualem Djehiche, and Stig Larsson. Riccati equations for the boundary control and filtering of stochastic reaction diffusion equations: Existence and uniquness. Work in Progress.

[2] Adam Andersson, Raphael Kruse, and Stig Larsson. Duality in refined Sobolev–Malliavin spaces and weak approximation of SPDE. *Stochastic Partial Differential Equations: Analysis and Computations*, 4:113–149, 2016.

[3] Adam Andersson, Annika Lang, Andreas Pettersson, and Leander Schroer. Finite element approximation of operator Lyapunov equations verses multi level Monte Carlo methods for the computation of quadratic functionals of SPDE. Work in Progress.

[4] Adam Andersson and Stig Larsson. Weak convergence for a spatial approximation of the nonlinear stochastic heat equation. *Mathematics of Computation*, 2016.

[5] HT Banks and Karl Kunisch. The linear regulator problem for parabolic systems. *SIAM Journal on Control and Optimization*, 22(5):684–698, 1984.

[6] Richard H. Bartels and GW Stewart. Solution of the matrix equation ax+ xb= c [f4]. *Communications of the ACM*, 15(9):820–826, 1972.

[7] Peter Benner, Jing-Rebecca Li, and Thilo Penzl. Numerical solution of large-scale Lyapunov equations, Riccati equations, and linear-quadratic optimal control problems. *Numerical Linear Algebra with Applications*, 15(9):755–777, 2008.

[8] Rajendra Bhatia and Peter Rosenthal. How and why to solve the operator equation ax-xb= y. *Bulletin of the London Mathematical Society*, 29(01):1–21, 1997.

[9] Ruth F Curtain and AJ Pritchard. The infinite-dimensional Riccati equation. *Journal of Mathematical Analysis and Applications*, 47(1):43–57, 1974.

[10] Tobias Damm, Hermann Mena, and Tony Stillfjord.

[11] Charles M Elliott and Stig Larsson. Error estimates with smooth and nonsmooth data for a finite element method for the Cahn-Hilliard equation. *Mathematics of Computation*, 58(198):603–630, 1992.

[12] Franco Flandoli. Riccati equation arising in a stochastic optimal control problem with boundary control. *Bollettino dell'Unione Matematica Italiana*, 6(1):377–393, 1982.

[13] Franco Flandoli. Direct solution of a Riccati equation arising in a stochastic control problem with control and observation on the boundary. *Applied Mathematics and Optimization*, 14(1):107–129, 1986.

[14] JS Gibson. The Riccati integral equations for optimal control problems on Hilbert spaces. *SIAM Journal on Control and Optimization*, 17(4):537–565, 1979.

[15] Akira Ichikawa. Dynamic programming approach to stochastic evolution equations. *SIAM Journal on Control and Optimization*, 17(1):152–174, 1979.

[16] Arnulf Jentzen. Numerical analysis of stochastic partial differential equations. 2014.

[17] Raphael Kruse. *Strong and Weak Approximation of Semilinear Stochastic Evolution Equations*, volume 2093. Springer Lecture Notes in Mathematics, 2014.

[18] Tijana Levajković, Hermann Mena, and Amjad Tuffaha. A numerical approximation framework for the stochastic linear quadratic regulator on Hilbert spaces. *Applied Mathematics & Optimization*, pages 1–25, 2016.

[19] Tijana Levajković, Hermann Mena, and Amjad Tuffaha. The stochastic linear quadratic optimal control problem in Hilbert spaces: A polynomial chaos approach. *Evolution Equations and Control Theory*, 5(1):105–134, 2016.

[20] H Mena. *Numerical Solution of Differential Riccati Equations Arising in Optimal Control Problems for Parabolic Partial Differential Equations*. PhD thesis, PhD thesis, Escuela Politécnica Nacional (Ecuador) partnership program with TU Berlin (Germany), 2007.

[21] Vidar Thomée. *Galerkin finite element methods for parabolic problems*, volume 1054. Springer, 1984.