ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Adaptive Galerkin Methods for Infinite Dimensional Parabolic Equations

Master Thesis

M. Verlinden

September 7, 2012

Advisor: Prof. Dr. C. Schwab.
SAM, ETH Zürich

Abstract

The following report follows very closely the guidelines provided in the paper of Schwab and Sülli [17] to solve parabolic problems on function spaces over a separable Hilbert space. Instead of a space time variational formulation, we consider here a Backward Euler scheme for the time discretization. The convergence of the Backward Euler scheme in a Hilbert space setting is proved. We present a class of high dimensional and infinite dimensional Fokker-Planck equations; for which a spectral Galerkin method is chosen in order to find a numerical approximation. The well-posedness of elliptic and parabolic problems in $L^2(H, \mu)$ spaces is discussed, and we show that the Wiener-Hermite polynomial chaos provides an appropriate basis for the discretization of variational operators. We show that the corresponding discrete operator equations in $\ell^2(\mathbb{N})$ can be approximated by a sequence of sparse problems that converge quasioptimally, in the sense of the best N-term rates possible for the exact solution.

Contents

Chapter 1

Introduction

This project presents numerical methods available to approximate the solutions of a class of high dimensional parabolic problems, possibly infinite dimensional. In particular we study the infinite dimensional Fokker-Planck equation (FP) for the Kolmogorov forward equation on a Hilbert space. The solutions to these kind of equations are useful in many different contexts (see, e.g., [2, 6] and the references therein). Citing [12, p. X] "parabolic equations on Hilbert spaces appear in mathematical physics to model systems with infinitely many degrees of freedom. Typical examples are provided by spin configurations in statistical mechanics and by crystals in solid state theory. Infinite-dimensional parabolic equations provide an analytic description of infinite-dimensional diffusion processes in such branches of applied mathematics as population biology, fluid dynamics, and mathematical finance". The numerical solutions to these equations have however received less attention and are generally done by path simulation of the corresponding stochastic partial differential equation. In this project we study instead the solution suggested in Schwab and Sülli [17]. Their approach offers a new, deterministic adaptive spectral Galerkin approach to the construction of finite-dimensional numerical approximations to the deterministic forward equation in infinite-dimensional spaces, which exhibit certain optimality properties. The equations are considered in a space-time variational formulation in Gelfand-triples of Sobolev spaces over a separable Hilbert space H with respect to a Gaussian measure μ. In this project however, we only consider a variational formulation in the space dimension, the Backward Euler method is used for integration in time. This reduces the parabolic problem to an elliptic problem at each time step. The solution is discretized via the choice of an appropriate Riesz basis, and approximated using the algorithms presented in [10] for adaptive Galerkin approximations of elliptic operator equations on bounded domains in $\mathbb{R}^d$. Most notably due to the lack of a suitable extension of Lebesgue measure to infinite dimensions, the study of infinite-dimensional Fokker–Planck equations is done on a separable Hilbert space H, equipped with a Gaussian measure μ. In this context we are in analogy with the finite dimensional study on weighted L^2 spaces of operators of the form $\partial_x(M\partial_x\cdot)$, where M is a density function. In the Gaussian case, the multivariate Hermite polynomials provide the Riesz basis for a *spectral Galerkin approximation*. We verify that for a particular class of second order operators, the associated discretized variational operator, viewed as a bi-infinite matrix, can be approximated by a sequence of sparse matrices in the sense of the operator norm. This verification ensures optimality of the adaptive procedure.

The project is structured as follows: Chapter 2 first presents the necessary theory of Gaussian measures on separable Hilbert spaces, together with the introduction of $L^2(H, \mu)$ spaces where μ is a Gaussian measure and H is infinite dimensional. Following [12, 15], we present the extension of the Laplacian and the associated heat-semigroup to the infinite dimensional setting, and see how to derive the variational formulation of second order operators. The rest of Chapter 2 is dedicated to the study of abstract elliptic problems and abstract parabolic problems. We define the solutions to such problems and precise the necessary conditions we shall assume for their existence. In Chapter 3 we first consider Fokker–Planck equations that arise in bead-spring chain models for d-dimensional polymeric flow $d \in \{2, 3\}$, with chains consisting of $K + 1$ beads whose kinematics are statistically described by a configuration vector $q \in \mathbb{R}^{Kd}$, $K \geqslant 1$. The probability density function $\psi = \psi(q, t)$ that is sought as the solution of the associated Fokker–Planck equation is therefore a function of Kd spatial variables with $K \geqslant 1$ and the time variable t. The aim is to embed this finite-dimensional problem of potentially very high dimension into an infinite-dimensional problem. Hence, Chapter 3 follows with the introduction of the infinite dimensional Fokker-Planck equation for which we verify well-posedness. In Chapter 4 we describe the first level of discretization in the time dimension. The Backward Euler method is chosen and we give the full proof of its consistency and stability in a Hilbert space setting. We then continue with the space discretization in Chapter 5 the adaptive procedures to approximate the solutions of abstract elliptic equations on separable Hilbert spaces. We prove the compressibility of the Backward Euler operator associated to the equations in Chapter 3. Finally in Chapter 6 we show some numerical results obtained for a canonical example of the Fokker-Planck equation from Chapter 3, in finite dimensions.

Chapter 2

Preliminaries

2.1 Gaussian measures on Hilbert spaces

2.1.1 Trace class operators

We let H be a separable Hilbert space over $\mathbb{R}$ with norm $|\cdot|$ associated with the inner-product $\langle\cdot,\cdot\rangle$. The space of all bounded linear operators on H will be denoted $L(H)$ and equipped with the operator norm $\|\cdot\|_{L(H)}$. We further let $L^+(H)$ denote the sub-space of symmetric non-negative operators on H, and for an operator $T \in L(H)$, we will denote by T^* its adjoint. We will also let $\mathcal{B}(H)$ denote the borel sigma algebra on H, associated to $|\cdot|$.

Definition 2.1 *An operator $T \in L(H)$ is said to be of trace class if there exist sequences $(a_n)_n$ and $(b_n)_n$ in H such that for all $f \in H$,*

$$Tf = \sum_{n=1}^{\infty} \langle f, a_n \rangle b_n, \quad \text{and} \quad \sum_{n=1}^{\infty} |a_n||b_n| < \infty. \tag{2.1}$$

We will denote by $L_1(H)$ the space of all linear operators of trace class. It is a Banach space with norm

$$\|T\|_{L_1(H)} = \inf \left\{ \sum_{n=1}^{\infty} |a_n||b_n| \ : \ Tf = \sum_{n=1}^{\infty} \langle f, a_n \rangle b_n \ \ \forall f \in H \right\} \tag{2.2}$$

$L_1^+(H) := L_1(H) \cap L^+(H)$. For an operator $R \in L_1(H)$, its trace, $\mathrm{Tr}(R)$ is given by

$$\mathrm{Tr}\ R = \sum_{n=1}^{\infty} \langle Re_n, e_n \rangle, \tag{2.3}$$

where $(e_n)_n \subset H$ is any complete orthonormal basis of H.

Definition 2.2 *An operator $T \in L(H)$ is Hilbert-Schmidt if there exists an orthonormal basis $(e_n)_n \subset H$ such that*

$$\sum_{n=1}^{\infty} |Te_n|^2 < \infty. \tag{2.4}$$

We will let $L_2(H)$ denote the space of all Hilbert-Schmidt operators on H. It is a Hilbert space for the scalar product

$$\langle T,S\rangle_{L_2(H)} = \sum_{i=1}^{\infty}\langle Te_n, Se_n\rangle, \tag{2.5}$$

where $(e_n)_n \subset H$ is any complete orthonormal system of H. The following well known spectral theorem also detailed later in Theorem 2.23, for bilinear forms, gives a useful characterization of the operators we shall use [12, p. 6].

Theorem 2.3 *Assume that S is a compact self-adjoint operator in $L(H)$. Then there exists a sequence $(\lambda_k)_k \subset \mathbb{R}$ and a complete orthonormal system $(e_k)_k \subset H$ such that $Se_k = \lambda_k e_k$, $k \in \mathbb{N}$. Moreover, $S \in L_1(H)$ if and only if $\sum_{k=1}^{\infty} |\lambda_k| < \infty$, in which case*

$$\|S\|_{L_1(H)} = \sum_{k=1}^{\infty} |\lambda_k|, \quad \textit{and} \quad \textit{Tr}\ S = \sum_{k=1}^{\infty} \lambda_k \ . \tag{2.6}$$

2.1.2 Gaussian measures on Hilbert spaces

Following [12], we first define Gaussian measures on $\mathbb{R}$. For $a \in \mathbb{R}$ and $\lambda > 0$, the Gaussian measure on $\mathbb{R}$ with mean a and variance λ is defined by

$$N_{a,\lambda}(\mathrm{d}x) := \frac{1}{\sqrt{2\pi\lambda}} \exp\left(\frac{-(x-a)^2}{2\lambda}\right) \mathrm{d}x \ . \tag{2.7}$$

For $a = 0$ we shall simply write $N_\lambda := N_{0,\lambda}$ for short.

Proposition 2.4 *For $a \in \mathbb{R}$ and $\lambda > 0$ we have,*

$$\begin{aligned} &\int_{\mathbb{R}} x N_{a,\lambda}(\mathrm{d}x) = a, \\ &\int_{\mathbb{R}} (x-a)^2 N_{a,\lambda}(\mathrm{d}x) = \lambda, \\ &\hat{N}_{a,\lambda}(h) = \int_{\mathbb{R}} e^{ixh} N_{a,\lambda}(\mathrm{d}x) = e^{iah - \frac{1}{2}\lambda h^2}, \quad h \in \mathbb{R}. \end{aligned}$$

We call a the mean, λ the variance and $\hat{N}_{a,\lambda}$ the Fourier transform (or characteristic function) of $N_{a,\lambda}$.

We now proceed to define the Gaussian measure $N_{a,Q}$ for an $a \in H$ and a $Q \in L^+(H)$, for H a finite dimensional Hilbert space. Let H be of dimension d, $Q \in L^+(H)$ and let $(e_1,\ldots,e_d)$ be an orthonormal basis of H such that $Qe_k = \lambda_k e_k$, $k = 1\ldots d$, for some $\lambda_1,\ldots,\lambda_d \in \mathbb{R}_+$. We set

$$x_k := \langle x, e_k\rangle, \quad x \in H,\ k = 1,\ldots,d, \tag{2.8}$$

and we identify H with $\mathbb{R}^d$ through the isomorphism

$$\gamma : H \to \mathbb{R}^d, \quad x \mapsto \gamma(x) = (x_1, x_2, \ldots, x_d). \tag{2.9}$$

The probability measure $N_{a,Q}$ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ is given by

$$N_{a,Q} = \bigotimes_{k=1}^{d} N_{a_k,\lambda_k}. \tag{2.10}$$

We have the following properties.

Proposition 2.5 *Let H be a finite dimensional Hilbert space, $a \in H$, $Q \in L^+(H)$ and $\mu = N_{a,Q}$ then,*

$$\int_H x N_{a,Q}(\mathrm{d}x) = a, \tag{2.11}$$

$$\int_H \langle x - a, y\rangle\langle x - a, z\rangle N_{a,Q}(\mathrm{d}x) = \langle Qy, z\rangle, \quad y, z \in H. \tag{2.12}$$

Moreover, the Fourier transform of the measure $N_{a,Q}$ is given by

$$\hat{N}_{a,Q}(h) := \int_H e^{i\langle h,x\rangle} N_{a,Q}(\mathrm{d}x) = e^{i\langle a,h\rangle - \frac{1}{2}\langle Qh,h\rangle}, \quad h \in H. \tag{2.13}$$

Finally, if the determinant of Q is positive, $N_{a,Q}$ is absolutely continuous with respect to Lebesgue measure in $\mathbb{R}^d$ and we have

$$N_{a,Q}(\mathrm{d}x) = \frac{1}{\sqrt{(2\pi)^d \det(Q)}} e^{-\frac{1}{2}\langle Q^{-1}(x-a),(x-a)\rangle} \mathrm{d}x. \tag{2.14}$$

We now let μ be a probability measure on $(H, \mathcal{B}(H))$ where H is any separable Hilbert space, possibly infinite dimensional. We assume that

$$\int_H |x| \mu(\mathrm{d}x) < \infty, \tag{2.15}$$

Then for any, $h \in H$, the linear functional $F : H \to \mathbb{R}$ defined as,

$$F(h) = \int_H \langle x, h\rangle \mu(\mathrm{d}x), \quad h \in H \tag{2.16}$$

is continuous since

$$|F(h)| \leqslant \int_H |x| \mu(\mathrm{d}x) |h|, \quad h \in H. \tag{2.17}$$

By the Riesz representation theorem there exists a unique $m \in H$ such that

$$F(h) = \langle m, h\rangle, \quad h \in H. \tag{2.18}$$

m is called the mean and we shall write,

$$m = \int_H x \mu(\mathrm{d}x). \tag{2.19}$$

We now assume that

$$\int_H |x|^2 \mu(\mathrm{d}x) < \infty. \tag{2.20}$$

Consequently, the bilinear form $G : H \times H \to \mathbb{R}$ defined as

$$G(h,k) = \int_H \langle x - m, h\rangle\langle x - m, k\rangle\mu(\mathrm{d}x), \quad h,l \in H \tag{2.21}$$

is also continuous, and by the Riesz isomorphism there exists a unique linear bounded operator $Q \in L(H)$ such that

$$G(h,k) = \langle Qh, k\rangle, \quad h,k \in H. \tag{2.22}$$

Q is called the covariance of μ.

Proposition 2.6 *Let μ be a probability measure on $(H, \mathcal{B}(H))$ with mean m and covariance Q. Then $Q \in L_1^+(H)$, i.e. Q is symmetric, positive and of trace class.*

For $a \in H$ and $Q \in L_1^+(H)$, a *Gaussian measure* $N_{a,Q}$ on $(H, \mathcal{B}(H))$ is a measure μ of mean a, covariance operator Q and Fourier transform

$$\hat{N}_{a,Q}(h) = \exp\left(i\langle a, h\rangle - \frac{1}{2}\langle Qh, h\rangle\right), \quad h \in H. \tag{2.23}$$

The Gaussian measure $N_{a,Q}$ is said non-degenerate if $\ker(Q) = \{0\}$. Since $Q \in L_1^+(H)$ there exists an orthonormal sequence $(e_k)_k \subset H$ and a sequence of non-negative numbers $(\lambda_k)_k$ with $Qe_k = \lambda_k e_k$, $k = 1, 2, \ldots$. For $x \in H$ we set $x_k = \langle x, e_k\rangle$, $k \in \mathbb{N}$. We now let $\mathbb{R}^\infty$ denote the space of all sequences $(x_n)_n \subset \mathbb{R}$ equipped with the metric

$$\mathfrak{d}(x,y) := \sum_{k=1}^{\infty} 2^{-k} \frac{|x_k - y_k|}{1 + |x_k - y_k|} \tag{2.24}$$

We will also let $\ell^2(\mathbb{N})$ denote the space of all sequences $x = (x_n)_n \in \mathbb{R}^\infty$ such that

$$\|x\|_{\ell^2(\mathbb{N})} := \left(\sum_{k=1}^{\infty} x_k^2\right)^{1/2} < \infty\ . \tag{2.25}$$

$\ell^2(\mathbb{N})$ is a Hilbert space with inner product $\langle x, y\rangle_{\ell^2(\mathbb{N})} := \sum_{k=1}^{\infty} x_k y_k$. In the next theorem, we identify H with $\ell^2(\mathbb{N})$ through the natural isomorphism $\gamma : H \to \ell^2$,

$$x \in H \mapsto (x_1, x_2, \ldots) \in \ell^2. \tag{2.26}$$

It is known ([12, p.10]) that for any Gaussian measure on H, the set $\ell^2(\mathbb{N}) \subset \mathbb{R}^\infty$ has measure 1. In order to define the Gaussian measure on $\ell^2(\mathbb{N})$ as an infinite product of Gaussian measures on $\mathbb{R}$, we introduce the projection operators, $p_J : \mathbb{R}^\infty \to \mathbb{R}^{|J|}$, $\quad J \subset \mathbb{N}$;

$$x = (x_n)_{n=1}^{\infty} \mapsto p_J(x) = (x_{j_1}, \ldots, x_{j_{|J|}}), \quad j_1, \ldots, j_{|J|} \in J. \tag{2.27}$$

For any $J \in \mathcal{F}(\mathbb{N})$, the set of all finite subsets of $\mathbb{N}$, the product σ-algebra and the product measure

$$\bigotimes_{j \in J} \Sigma_j, \quad \bigotimes_{j \in J} N_{a_j, \lambda_j}, \tag{2.28}$$

are understood in the usual way, with Σ_i the Borel σ-algebra on $\mathbb{R}$, $i \in \mathbb{N}$. The infinite product of the σ-algebras $\{\Sigma_i,\ i \in \mathbb{N}\}$, is defined as the smallest σ-algebra with respect to which the projections p_J are measurable, i.e,

$$\Sigma_0 := \bigotimes_{k=1}^{\infty} \Sigma_k = \sigma(p_J,\ J \in \mathcal{F}(N)). \tag{2.29}$$

There exists a unique measure μ on $(\mathbb{R}^\infty, \Sigma_0)$ such that

$$\mu \circ p_J^{-1} = \bigotimes_{j \in J} N_{a_j, \lambda_j}, \tag{2.30}$$

it shall be denoted $\bigotimes_{k=1}^{\infty} N_{a_k,\lambda_k}$.

Theorem 2.7 ([12, p. 9]) *Suppose that $a \in H$ and $Q \in L_1^+(H)$. Then, there exists a unique probability measure μ on $(H, \mathcal{B}(H))$ such that*

$$\int_H e^{i\langle h,x\rangle}\mu(\mathrm{d}x) = e^{i\langle a,h\rangle - \frac{1}{2}\langle Qh,h\rangle} \tag{2.31}$$

Moreover, μ is the restriction to H (identified with the Hilbert space $\ell^2(\mathbb{N})$) of the product measure

$$\bigotimes_{k=1}^{\infty} \mu_k = \bigotimes_{k=1}^{\infty} N_{a_k,\lambda_k}, \tag{2.32}$$

defined on $(\mathbb{R}^\infty, \mathcal{B}(\mathbb{R}^\infty))$.

We refer to $\mu := N_{a,Q}$ as the *Gaussian measure* associated to the *mean a* and the *covariance operator Q*. Theorem 2.7 implies that a random variable X with values in H is Gaussian if, and only if, for any $h \in H$ the real-valued random variable $\langle h, X\rangle$ is Gaussian.

2.1.3 L^2 and Sobolev spaces

We let $\mathcal{H} := L^2(H, \mu)$ denote the Hilbert space of equivalence classes of functions from H into $\mathbb{R}$ with inner-product

$$\langle u, v\rangle_{\mathcal{H}} := \int_H u(x)v(x)\mu(\mathrm{d}x), \quad u, v \in \mathcal{H}, \tag{2.33}$$

and norm

$$\|u\|_{\mathcal{H}} := \langle u, u\rangle_{\mathcal{H}}^{1/2} < \infty\ . \tag{2.34}$$

From now on, $\mu = N_Q := N_{0,Q}$ for some operator $Q \in L_1^+(H)$ with $\mathrm{Ker}(Q) = \{0\}$. We shall also suppose that there exists a complete orthonormal system $(e_k)_k$ in H and a sequence $(\lambda_k)_k$ of positive real numbers, the eigenvalues of Q (repeated according to their multiplicity and enumerated in decreasing order), such that $Qe_k = \lambda_k e_k$. The subspace $Q^{1/2}(H)$ is called the *reproducing kernel* of the measure N_Q. It is a dense subspace of H since we assumed $\mathrm{Ker}(Q) = \{0\}$. In fact, if $x_0 \in H$ is such that $\langle Q^{1/2}h, x_0\rangle = 0$ for all $h \in H$, then $Q^{1/2}x_0 = 0$, and therefore $Qx_0 = 0$, which implies that $x_0 = 0$. We now introduce the isomorphism

$$\begin{aligned} W : H &\to \mathcal{H}, \\ f &\mapsto W_f, \quad W_f(x) = \langle Q^{-1/2}f, x\rangle, \quad x \in H. \end{aligned}$$

One can check that we have

$$\int_H \langle W_f(x), W_g(x)\rangle \mu(\mathrm{d}x) = \langle f, g\rangle, \tag{2.35}$$

so that W is an isometry and can be uniquely extended to all $f \in H$. For every $f \in H$ we see that W_f is an $N_{|f|^2}$ real valued Gaussian random variable on $\mathbb{R}$.

We now define the Hermite polynomials on $\mathcal{H} = L^2(H, \mu)$. Let us consider to this end the set Γ of all mappings $\gamma : n \in \mathbb{N} \to \gamma_n \in \{0\} \cup \mathbb{N}$, such that $|\gamma| := \sum_{k=1}^{\infty} \gamma_k < \infty$. Clearly $\gamma \in \Gamma$ if, and only if, $\gamma_n = 0$ for all, except possibly finitely many, $n \in \mathbb{N}$. For any $\gamma \in \Gamma$ we define the Hermite polynomial

$$H_\gamma(x) = \prod_{k=1}^{\infty} H_{\gamma_k}\left(W_{e_k}(x)\right), \quad x \in H, \tag{2.36}$$

where the functions on the right hand side are defined by

$$H_n(\xi) = \frac{(-1)^n}{\sqrt{n!}} e^{\frac{\xi^2}{2}} \frac{\mathrm{d}^n}{\mathrm{d}\xi^n}\left(e^{-\xi^2/2}\right), \quad \xi \in \mathbb{R}, \quad n \in \{0\} \cup \mathbb{N}. \tag{2.37}$$

H_n is the classical Hermite polynomial of degree n with the first few terms given by,

$$H_0(\xi) \equiv 1, \quad H_1(\xi) = \xi, \quad H_2(\xi) = \frac{1}{\sqrt{2}}\left(\xi^2 - 1\right), \quad H_3(\xi) = \frac{1}{\sqrt{6}}\left(\xi^3 - 6\right) \dots \tag{2.38}$$

For the rest of this project, we shall use the convention $H_{-1} \equiv 0$. It is well known that the Hermite polynomials form an orthonormal basis of $L^2(\mathbb{R}, N_1)$. We also have the following relationships.

Proposition 2.8 *For $n \in \mathbb{N}$ and all $\xi \in \mathbb{R}$ we have*

$$\begin{aligned} \xi H_n(\xi) &= \sqrt{n+1} H_{n+1}(\xi) + \sqrt{n} H_{n-1}(\xi), && (2.39) \\ D_\xi H_n(\xi) &= \sqrt{n} H_{n-1}(\xi), && (2.40) \\ D_\xi^2 H_n(\xi) - \xi D_\xi H_n(\xi) &= -\, n H_n(\xi). && (2.41) \end{aligned}$$

The numerical methods presented in Chapter 5 make extensive use of the following theorem.

Theorem 2.9 *The system $(H_\gamma)_{\gamma \in \Gamma}$ is orthonormal and complete on $L^2(H, \mu)$.*

A proof may be found in [12, p.191]

2.1.4 The Sobolev space $\mathbf{W}^{1,2}(H,\mu)$

We will denote by $E(H)$ the linear space spanned by all exponential functions, that is all functions $\varphi : x \in H \mapsto \varphi(x) \in \mathbb{R}$ of the form

$$\varphi(x) = e^{\langle h,x \rangle}, \quad h \in H. \tag{2.42}$$

Proposition 2.10 *For any $h \in H$, the exponential function E_h, defined as*

$$E_h(x) = e^{\langle h,x \rangle}, \quad x \in H \tag{2.43}$$

belongs to $L^p(H,\mu)$, $p \geqslant 1$, and

$$\int_H e^{\langle h,x \rangle}\mu(dx) = e^{\frac{1}{2}\langle Qh,h \rangle}, \tag{2.44}$$

Moreover the subspace $E(H)$ is dense in $\mathcal{H}$.

For any $k \in \mathbb{N}$ we consider the partial derivative in the direction e_k, defined as

$$D_k\varphi(x) = \lim_{\epsilon \to 0} \frac{\varphi(x+\epsilon e_k) - \varphi(x)}{\epsilon}, \quad x \in H, \quad \varphi \in E(H). \tag{2.45}$$

When $\varphi \in E(H)$ with $\varphi(x) = e^{\langle f,x \rangle}$, $f \in H$, we have

$$D_k\varphi(x) = f_k e^{\langle f,x \rangle}, \quad \text{where} \quad f_k = \langle f, e_k \rangle \tag{2.46}$$

The following proposition is central to this project.

Proposition 2.11

$$D_k H_\gamma = \sqrt{\frac{\gamma_k}{\lambda_k}} H_{\gamma_k - 1}(W_{e_k}) H_\gamma^{(k)}, \tag{2.47}$$

with $H_\gamma^{(k)} = \prod_{j \neq k} H_{\gamma_j}(W_{e_j})$ and the convention $H_{-1}(W_{e_k}) = 0$. Moreover, the family

$$\left\{ H_{\gamma_k-1}(W_{e_k})H_\gamma^{(k)},\ \gamma \in \Gamma,\ \gamma_k > 0 \right\}, \tag{2.48}$$

is orthonormal in $\mathcal{H}$.

The verification of (2.47) can be done using the identity (2.40) and the fact that $W_{e_k}(x) = \lambda_k^{-1/2} x_k$. We now let Λ_0 denote the linear span of $\{H_\gamma \otimes e_k :\ \gamma \in \Gamma,\ k \in \mathbb{N}\}$, and D the linear operator

$$D : E(H) \subset \mathcal{H} := L^2(H,\mu) \to L(H,\mu;H),$$
$$\varphi \mapsto D\varphi \quad \text{with } D\varphi(x) := \sum_{k=1}^{\infty} D_k\varphi(x)e_k.$$

Thanks to Proposition 9.2.2 in Da Prato and Zabczyk [12], D_k is closable in $\mathcal{H}$ for all $k \in \mathbb{N}$. If φ belongs to the domain of the closure of D_k, which we shall still denote by D_k, we shall say that $D_k\varphi$ belongs to $\mathcal{H}$. Analogously, by Proposition 9.2.4 in [12], D is a closable linear operator. If φ belongs to the domain of the closure of D, which we shall still denote by D, we shall say that $D\varphi$ belongs to $L^2(H,\mu;H)$.

We will now consider the linear space $\mathcal{V} = W^{1,2}(H,\mu) \subset L^2(H,\mu)$, of functions $\varphi \in L^2(H,\mu)$ such that $D\varphi \in L^2(H,\mu;H)$. It is a Hilbert space with inner product

$$\langle u, v\rangle_{\mathcal{V}} = \langle u, v\rangle_{\mathcal{H}} + \int_H \langle Du(x), Dv(x)\rangle \mu(\mathrm{d}x), \tag{2.49}$$

and associated norm $\|u\|_{\mathcal{V}} = (\langle u, u\rangle_{\mathcal{V}})^{1/2}$.

Theorem 2.12 ([12, p. 199]) *A function $\varphi \in \mathcal{H}$ belongs to $\mathcal{V}$ if, and only if,*

$$\sum_{\gamma\in\Gamma} \langle \gamma, \lambda^{-1}\rangle_* |\varphi_\gamma|^2 < \infty, \tag{2.50}$$

where

$$\varphi_\gamma := \langle \varphi, H_\gamma\rangle, \text{ and } \langle \gamma, \lambda^{-1}\rangle_* := \begin{cases} \sum_{k=0}^{\infty} \gamma_k \lambda_k^{-1}, & \text{if } \gamma \neq 0, \\ 1 & \text{if } \gamma = 0, \end{cases} \tag{2.51}$$

and $(\lambda_k)_k$ is the sequence of (positive) eigenvalues (repeated according to their multiplicity) of the covariance operator $Q \in L_1^+(H)$, $Ker(Q) = \{0\}$. Moreover, if (2.50) holds, then

$$\|\varphi\|_{\mathcal{V}}^2 = \|\varphi\|_{\mathcal{H}}^2 + \sum_{\gamma\in\Gamma} \langle \gamma, \lambda^{-1}\rangle_* |\varphi_\gamma|^2. \tag{2.52}$$

Identifying $\mathcal{H}$ with its own dual $\mathcal{H}^*$, we obtain

$$\varphi \in \mathcal{V}^* \iff \sum_{\gamma\in\Gamma} \langle \gamma, \lambda^{-1}\rangle_*^{-1} |\varphi_\gamma|^2 < \infty. \tag{2.53}$$

Furthermore, the embedding of $\mathcal{V}$ into $\mathcal{H}$ is compact.

Proof The proof of (2.50) and (2.52) can be found in Da Prato and Zabczyk [12, p.200]. Using Proposition 2.11 we can further notice that

$$\Psi = (\psi_\gamma)_{\gamma\in\Gamma} := \left(\frac{H_\gamma}{(1 + \langle \gamma, \lambda^{-1}\rangle_*)^{1/2}} \right)_{\gamma\in\Gamma} \tag{2.54}$$

is an orthonormal basis of $\mathcal{V}$. In deed,

$$\begin{aligned}\langle \psi_\gamma, \psi_\nu \rangle_{\mathcal{V}} =& \langle \psi_\gamma, \psi_\nu \rangle_{\mathcal{H}} + \int_H \langle D\psi_\gamma(x), D\psi_\nu(x)\rangle \mu(\mathrm{d}x) \\ =& (1 + \langle \gamma, \lambda^{-1}\rangle_*)^{-1/2} (1 + \langle \nu, \lambda^{-1}\rangle)^{-1/2} \left(\delta_{\gamma,\nu} + \sum_{k=1}^{\infty} \int_H D_k H_\gamma(x) D_k H_\nu(x) \mu(\mathrm{d}x) \right) \\ =& (1 + \langle \gamma, \lambda^{-1}\rangle_*)^{-1} \left(\delta_{\gamma,\nu} + \sum_{k=1}^{\infty} \left(\frac{\gamma_k}{\lambda_k}\right)^{1/2} \left(\frac{\nu_k}{\lambda_k}\right)^{1/2} \delta_{\gamma,\nu} \right) = \delta_{\gamma,\nu}.\end{aligned}$$

Since $\left(|\varphi_\gamma|^2 (1 + \langle \gamma, \lambda^{-1}\rangle_*)^{-1/2} \right)_{\gamma\in\Gamma}$ converges if and only if $\left(|\varphi_\gamma|^2 \langle \gamma, \lambda^{-1}\rangle_*^{-1/2} \right)_{\gamma\in\Gamma}$ converges, by taking $\varphi_n \to \varphi$, $(\varphi_n)_n \subset \mathcal{H}$ we find,

$$\begin{aligned}\infty > \|\varphi\|_{\mathcal{V}*}^2 &\iff {}_{\mathcal{V}*}\langle \varphi, x\rangle_{\mathcal{V}}^2 < \infty, \ \forall x \in \mathcal{V}, \ \|x\|_{\mathcal{V}} \leqslant 1, & (2.55) \\ &\iff \lim_{n\to\infty} \langle \varphi_n, x\rangle_{\mathcal{H}}^2 < \infty, \ \forall x \in \mathcal{V}, \ \|x\|_{\mathcal{V}} \leqslant 1. & (2.56)\end{aligned}$$

By expanding φ_n into the polynomial chaos basis $\left(H_\gamma\right)_{\gamma\in\Gamma}$, and x into the orthonormal basis Ψ, we find

$$\begin{aligned}(2.56) \iff \lim_{n\to\infty} \sum_{\gamma\in\Gamma} \langle \gamma, \lambda^{-1}\rangle_*^{-1} |\varphi_\gamma^n|^2 < \infty,\\ \iff \sum_{\gamma\in\Gamma} \langle \gamma, \lambda^{-1}\rangle_*^{-1} |\varphi_\gamma|^2 < \infty. \qquad \square\end{aligned}$$

2.1.5 Variational formulation of second order operators

We are now interested in formulating a generalization of the Laplacian to infinite dimensional spaces. A straightforward generalization is impossible, mostly due to the lack of a suitable extension of Lebesgue measure to infinite dimensions. Following Chapter 3 in Da Prato and Zabczyk ([12]), we briefly look at strong solutions to the heat equation on infinite dimensional Hilbert spaces, we then present the $L^2(\mu)$ analysis following Ann Piech ([15]).

It is instructive to consider the problem

$$\partial_t u(t,x) = \frac{1}{2}\mathrm{Tr}(QD^2u(t,x)), \quad t>0,\ x\in H \tag{2.57}$$

$$u(0,x) = u_0 \in B, \tag{2.58}$$

where B is an appropriate Banach space to be determined, and D^2 denotes the second order Fréchet derivative at 0 of the function $g : H \to \mathbb{R}$ defined by $g(h) = u(t, x+h)$, $t > 0$, $x \in H$. Under suitable assumptions on B and Q described below, the solution to this problem is given by

$$u(t,x) = \int_H u_0(x+y) N_{tQ}(\mathrm{d}y). \tag{2.59}$$

In the case when $\dim(H) < \infty$ the solution is well understood, as the Radon-Nikodyn derivative of the measure $N_{tQ}(\mathrm{d}y)$ is simply given by the multivariate Gaussian density of mean 0 and covariance matrix tQ. When $\dim(H) = \infty$, we consider a sequence of finite rank operators Q_n converging strongly to Q in $L^+(H)$, and let $(u_n)_n$ be the sequence of solutions to the problems

$$\partial_t u_n(t,x) = \frac{1}{2}\mathrm{Tr}(Q_nD^2u_n(t,x)), \quad t>0,\ x\in H \tag{2.60}$$

$$u_n(0,x) = u_0. \tag{2.61}$$

When this sequence of solutions is convergent, we may take its limit as the solution to (2.57). We now see the conditions we must impose on Q to have this possibility. We let $C_b(H)$ denote the space of bounded continuous functions on H, taking values in $\mathbb{R}$.

Proposition 2.13 *Assume that $u_0 \in C_b(H)$ and $\lim_{|y|\to\infty} u_0(y) = 0$. If Tr$(Q) = \infty$ and $(Q_n)_n$ is a sequence of finite rank symmetric positive operators converging strongly to Q, then $\lim_{n\to\infty} u_n(t,x) = 0$, for all $t > 0$ and $x \in H$.*

Proposition 2.13 indicates that if $\mathrm{Tr}(Q) = \infty$ then, for a majority of initial functions u_0, the equation (2.57) does not have a continuous solution on $[0, +\infty) \times H$. This is why we will assume that $Q \in L^+(H)$ is of trace class.

Remark 2.14 *One can show that the heat equation (2.57) is the Kolmogorov equation corresponding to the simplest Ito equation*

$$\mathrm{d}X(t) = \mathrm{d}W(t), \quad X(0) = x, \quad , t > 0, \tag{2.62}$$

on a Hilbert space H, where W is a Wiener process on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ taking values in H, with covariance operator Q.

From [12], we can find a unique strong solution to (2.57), in the sense that $D_t u$ and $D^2 u$ exist, are continuous and bounded, and satisfy (2.57), for $t > 0$ and $x \in H$; provided that $u_0 \in UC_b^2(H)$, the space of functions having uniformly continuous and bounded derivatives of second order. In this case, the solution is given by (2.59). We shall denote by $(P_t)_{t>0}$ the *heat-semigroup* of operators defined by

$$P_t\varphi(x) = \int_H \varphi(x+y) N_{tQ}(\mathrm{d}y). \tag{2.63}$$

The infinitesimal generator of the heat semi-group is given by $\Delta_Q : D(\Delta_Q) \to \mathcal{H}$, which action can be interpreted as ([12])

$$\Delta_Q \varphi(x) := \sum_{k=1}^{\infty} \lambda_k D_k^2 \varphi(x), \quad x \in H. \tag{2.64}$$

It is known that when H is finite dimensional and $\varphi \in C_b(H)$, the function $u(t,x) = P_t\varphi(x)$ is of class C^∞ in t and x when $t > 0$. Moreover when $\dim(H) < \infty$, the semi-group $(P_t)_{t>0}$ is strongly continuous on $C_b(H)$. These results are not true in infinite dimensions ([12]). We refer to [12, Chap. 3] for a further analysis on spaces of continuous functions, and proceed to the variational formulation.

The above analysis has pointed out some of the major problems for extending the Laplacian to infinite dimensions. We shall now show that the operator (2.64) fails even to be symmetric.

Lemma 2.15 *Let φ, $\psi \in E(H)$. Then the following identity holds:*

$$\int_H \psi(x) D_k \varphi(x) \mu(\mathrm{d}x) + \int_H \varphi(x) D_k \psi(x) \mu(\mathrm{d}x) = \frac{1}{\lambda_k} \int_H x_k \varphi(x) \psi(x) \mu(\mathrm{d}x). \tag{2.65}$$

Proof Since $E(H)$ is dense in $W^{1,2}(H,\mu)$ it is enough to prove (2.65) for

$$\varphi(x) = e^{\langle f,x \rangle}, \ \psi(x) = e^{\langle g,x \rangle}, \ x \in H, \tag{2.66}$$

where $f, g \in H$. In this case we have

$$\int_H \psi(x) D_k \varphi(x) \mu(\mathrm{d}x) = \int_H f_k e^{\langle f+g,x \rangle} \mu(\mathrm{d}x) = \quad f_k e^{\frac{1}{2}\langle Q(f+g), f+g \rangle}, \tag{2.67}$$

$$\int_H \varphi(x) D_k \psi(x) \mu(\mathrm{d}x) = \int_H g_k e^{\langle f+g,x \rangle} \mu(\mathrm{d}x) = \quad g_k e^{\frac{1}{2}\langle Q(f+g), f+g \rangle}, \tag{2.68}$$

$$\int_H \varphi(x) x_k \psi(x) \mu(\mathrm{d}x) = \int_H x_k e^{\langle f+g,x \rangle} \mu(\mathrm{d}x), \tag{2.69}$$

$$=\frac{\mathrm{d}}{\mathrm{d}t}\int_H e^{\langle f+g+te_k,x\rangle}\mu(\mathrm{d}x)\restriction_{t=0}=\frac{\mathrm{d}}{\mathrm{d}t}e^{\frac{1}{2}\langle Q(f+g+te_k),f+g+te_k\rangle}\restriction_{t=0} \tag{2.70}$$

$$=\lambda_k(f_k+g_k)e^{\frac{1}{2}\langle Q(f+g),f+g\rangle}. \tag{2.71}$$

Since D is a closable operator for all $k\in\mathbb{N}$, we can easily extend Lemma 2.15 to $\varphi,\psi\in\mathcal{V}$. Using this extension and summing up over k leads to the formula, for all $G\in\operatorname{span}\{H_\gamma\otimes e_k,\ \gamma\in\Gamma,k\in\mathbb{N}\}$ and $\varphi\in\mathcal{V}$:

$$\int_H\langle D\varphi(x),G(x)\rangle\mu(\mathrm{d}x)+\int_H\varphi(x)\mathrm{div}(G(x))\mu(\mathrm{d}x)=\int_H\langle x,Q^{-1}G(x)\rangle\mu(\mathrm{d}x). \tag{2.72}$$

We shall often use the previous formula for $G=Du(x)$. Moreover, we have the following estimate.

Proposition 2.16 *Let $k\in\mathbb{N}$ and $D_k\varphi\in L^2(H,\mu)$. Then $x_k\varphi\in L^2(H,\mu)$ and the following estimate holds:*

$$\int_H x_k^2\varphi^2(x)\mu(\mathrm{d}x)\leqslant 2\lambda_k\int_H\varphi^2(x)\mu(\mathrm{d}x)+4\lambda_k^2\int_H(D_k\varphi(x))^2\mu(\mathrm{d}x). \tag{2.73}$$

Summing up over k in (2.73) leads to the following estimate.

Proposition 2.17 *Let $\zeta\in\mathcal{V}$. Then the function*

$$H\to\mathbb{R},\ x\mapsto|x|\zeta(x), \tag{2.74}$$

belongs to $\mathcal{H}$, and

$$\int_H|x|^2\zeta^2(x)\mu(dx)\leqslant 2Tr(Q)\int_H\zeta^2(x)\mu(\mathrm{d}x)+4\left\|Q\right\|^2\int_H|D\zeta(x)|^2\mu(\mathrm{d}x). \tag{2.75}$$

We now consider iterating the operators D_k, $k\in\mathbb{N}$.

Lemma 2.18 *Let $h,k\in\mathbb{N}$, then the linear operator D_hD_k , defined in $E(H)$, is closable.*

When φ belongs to the domain of the closure of D_hD_k , which we shall say that $D_hD_k\varphi$ is an element of $L^2(H,\mu)$. We now define $W^{2,2}(\mathcal{H},\mu)$ as the space of all functions $\varphi\in L^2(H,\mu)$ such that $D_hD_k\varphi\in L^2(H,\mu)$ for all $h,k\in\mathbb{N}$ and

$$\sum_{h,k\in\mathbb{N}}\|D_hD_k\varphi\|_{\mathcal{H}}^2<\infty. \tag{2.76}$$

Then $W^{2,2}(\mathcal{H},\mu)$ is a Hilbert space with the inner product

$$\langle u,v\rangle_{W^{2,2}(\mathcal{H},\mu)}=\langle u,v\rangle_{\mathcal{V}}+\sum_{h,k=1}^{\infty}\int_H\langle D_hD_ku(x),D_hD_kv(x)\rangle\mu(\mathrm{d}x). \tag{2.77}$$

If $\varphi\in W^{2,2}(H,\mu)$ we can define a Hilbert-Schmidt operator $D^2\varphi(x)$ on H for almost any $x\in H$ by setting

$$\langle D^2\varphi(x)\alpha,\beta\rangle=\sum_{h,k=1}^{\infty}D_hD_k\varphi(x)\alpha_h\beta_k,\quad\alpha,\beta\in H. \tag{2.78}$$

Unlike the embedding of $\mathcal{V}$ into $\mathcal{H}$, when H is infinite dimensional, the embedding of $W^{2,2}(\mathcal{H},\mu)$ into $\mathcal{V}$ is not compact. We can see this by setting $\varphi^{(n)}(x) = x_n, n \in \mathbb{N}$, which gives

$$\left\|\varphi^{(n)}\right\|^2_{L^2(H,\mu)} = \lambda_n \to 0, \quad \text{as } n \to \infty, \tag{2.79}$$

and

$$\left\|\varphi^{(n)}\right\|^2_{W^{1,2}(\mathcal{H},\mu)} = \left\|\varphi^{(n)}\right\|^2_{W^{2,2}(\mathcal{H},\mu)} = 1 + \lambda_n. \tag{2.80}$$

Therefore $\varphi^{(n)}$ converges to 0 in $L^2(H,\mu)$ but has no subsequence converging to 0 in $W^{1,2}(\mathcal{H},\mu)$. For any $\gamma \in \Gamma$, $k \in \mathbb{N}$ we also have $D_k^2 H_\gamma \in \mathcal{H}$ and

$$D_k^2 H_\gamma = \begin{cases} \frac{\sqrt{\gamma_k(\gamma_k-1)}}{\lambda_k} H_{\gamma_k-2}(W_{e_k})H_\gamma^{(k)}, \text{ if } \gamma_k \neq 0, \\ 0 \; otherwise, \end{cases} \tag{2.81}$$

and since $\left\{H_{\gamma_k-2}(W_{e_k})H_\gamma^{(k)},\ \gamma \in \Gamma,\ \gamma_k > 0\right\}$ is an orthonormal set,

$$\int_H |D_k^2\varphi(x)|^2\mu(\mathrm{d}x) = \sum_{\gamma\in\Gamma} \frac{\gamma_k(\gamma_k-1)}{\lambda_k^2}|\varphi_\gamma|^2. \tag{2.82}$$

Proceeding similarly, we also find for $h \neq k$,

$$\int_H |D_h D_k\varphi(x)|^2\mu(\mathrm{d}x) = \sum_{\gamma\in\Gamma} \frac{\gamma_h\gamma_k}{\lambda_k^2}. \tag{2.83}$$

This gives us a characterization of elements in $W^{2,2}(\mathcal{H},\mu)$ like Theorem 2.12.

Theorem 2.19 ([12, p. 203]) *A function $\varphi \in \mathcal{H}$ belongs to $W^{2,2}(\mathcal{H},\mu)$ if and only if*

$$\sum_{\gamma\in\Gamma}\langle\gamma,\lambda^{-1}\rangle_*^2|\varphi_\gamma|^2 - \sum_{\gamma\in\Gamma}\langle\gamma,\lambda^{-2}\rangle^2|\varphi_\gamma|^2 < \infty. \tag{2.84}$$

Using Lemma 2.15, we can derive the variational form of the generator of the heat semi-group Δ_Q in (2.64). Using the identification $\mathcal{H} \simeq \mathcal{H}^*$, for all $u \in \mathcal{D}(\Delta_Q) \cap \mathcal{V}$, $v \in \mathcal{V}$,

$$_{\mathcal{V}^*}\langle -\Delta_Q u, v\rangle_{\mathcal{V}} = -\sum_{k=1}^{\infty}\lambda_k\int_H v(x)D_k^2u(x)\mu(\mathrm{d}x) \tag{2.85}$$

$$= \sum_{k=1}^{\infty}\lambda_k\int_H D_ku(x)D_kv(x)\mu(\mathrm{d}x) - \sum_{k=1}^{\infty}\int_H x_kv(x)D_ku(x)\mu(\mathrm{d}x) \tag{2.86}$$

$$= \sum_{k=1}^{\infty}\lambda_k\int_H D_ku(x)D_kv(x)\mu(\mathrm{d}x) - \int_H\langle x, Du(x)\rangle v(x)\mu(\mathrm{d}x). \tag{2.87}$$

Since this operator is not symmetric, it is suggested in [15] to consider instead the operator L with $Lu(x) := \mathrm{Tr}(QD^2u(x)) - \langle x, Du(x)\rangle$. We are interested in the closure of $-L$; the operator $N = -\bar{L}$, also known as the *number operator* from quantum field theory. We have for all $u \in \mathcal{D}(\Delta_Q) \cap \mathcal{V}$, $v \in \mathcal{V}$,

$$_{\mathcal{V}*}\langle -Lu, v\rangle_{\mathcal{V}} = \sum_{k=1}^{\infty} \lambda_k \int_H D_k u(x) D_k v(x) \mu(\mathrm{d}x). \tag{2.88}$$

In view of Remark 2.14, we can see the heat semi-group $(P_t)_{t>0}$ as a Markov transition semi-group with transition probabilities $N_{x,tQ}(\mathrm{d}y)$, i.e

$$P_t f(x) = \int_H f(y) N_{x,tQ}(\mathrm{d}y), \quad x \in H. \tag{2.89}$$

It is in this sense that the *heat semi-group* is interpreted in [15], where it is shown that the semi-goup $(O_t)_{t>0}$ generated by L is defined by

$$O_t f(x) = \int_H f(y) N_{e^{-tx},(1-e^{-2t})Q}(\mathrm{d}y). \tag{2.90}$$

By verifying the claim on every element of the polynomial chaos basis (2.36), we have the following proposition.

Proposition 2.20 ([15, Proposition 1]) *$(O_t)_{t>0}$ forms a strongly continuous contraction semi-group on $L^2(H,\mu)$.*

It can be seen that $(O_t)_{t>0}$ is the generator of the Ito process

$$X(t) = e^{-t}x + W(1 - e^{-2t}), \tag{2.91}$$

where W is a Wiener process on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ taking values in H, with covariance operator Q.

The principal drawback of defining L this way is that $\mathrm{Tr}(\Delta_Q f(x))$ must exist separately from $\langle x, Du(x)\rangle$, i.e that $(\Delta_Q f(x))$ must be of trace class. Thanks to the following approach ([15]), we shall be able to only require the weaker assumption that $(\Delta_Q f(x))$ is Hilbert-Schmidt.

Definition 2.21 *Assume that $f \in L^2(H,\mu)$, $|Df(x)|$ exists for a.e. x and is in $L^2(\mu)$ and $\|D^2 f(x)\|_{H\text{-}S}$ exists for a.e. x and is in $L^2(\mu)$. Let P_n, be the orthogonal projection of H onto $\{e_1, , ..., e_n\}$. It is shown in [15, Proposition 4] show that*

$$\left\{ Tr(P_n D^2 f(x) - \langle x, P_n Df(x)\rangle) \right\}_{n\in\mathbb{N}} \tag{2.92}$$

is a Cauchy sequence in $L^2(\mu)$. Lf is defined as the limit of this sequence.

This definition makes use of the following useful lemma.

Lemma 2.22 *If f is a C^2 mapping from $\mathbb{R}^n$ to $\mathbb{R}^n$ with $|f(x)|$ and $|Df(x)|_{H\text{-}S}$ in $L^2(\mu_1)$ with $\mu_1 \sim N(0, I_n)$, then we have*

$$\int_{\mathbb{R}^n} (Tr(Df(x)) - x\cdot f(x))^2 \mu_1(\mathrm{d}x) \leqslant \int_{\mathbb{R}^n} \left(|f(x)|^2 + |D(f(x))|^2_{H\text{-}S}\right) \mu_1(\mathrm{d}x). \tag{2.93}$$

The analysis in [15, Prop. 11, 13] further shows that for all $x \in Q^{1/2}(H)$, and $f \in \mathcal{H}$, we have the generalized derivatives

$$\langle D(O_t f(x)), h\rangle = -(e^t(1-e^{-2t}))^{-1} \int_H f(y)\langle e^{-tx-y}, h\rangle N_{e^{-tx},(1-e^{-2t})Q}(\mathrm{d}y), \tag{2.94}$$

and for $h, k \in H$

$$\langle D^2 O_t f(x)h, k\rangle = (e^{2t}-1)^{-1} \int_H f(y) \tag{2.95}$$

$$\times \left((1-e^{-2t})^{-1}\langle e^{-t}x-y, h\rangle\langle e^{-t}x-y, k\rangle - \langle h, k\rangle\right) N_{e^{-tx},(1-e^{-2t})Q}(\mathrm{d}y), \tag{2.96}$$

with $|DO_t f(\cdot)|$ and $\left\|D^2 O_t f(\cdot)\right\|_{\text{H-S}}$ in $\mathcal{H}$. Any f in $\mathcal{D}(L)$ is also in the domain of N, and $-Nf = Lf$.

2.2 Abstract elliptic problems

Let $\mathcal{H}$ and $\mathcal{V}$ be two separable Hilbert spaces over $\mathbb{R}$, such that $\mathcal{V} \subset \mathcal{H}$ with dense and continuous injection. We will further assume that the canonical embedding of $\mathcal{V}$ into $\mathcal{H}$ is dense and compact. This will be denoted as $\mathcal{V} \hookrightarrow \mathcal{H}$. We will identify $\mathcal{H}$ with its dual space $\mathcal{H}^*$, so that the inner product on $\mathcal{H}$ denoted by $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ extends, by continuity, to the duality pairing ${}_{\mathcal{V}^*}\langle\cdot,\cdot\rangle_{\mathcal{V}}$ on $\mathcal{V}^* \times \mathcal{V}$. In this setting we have the Gelfand triple,

$$\mathcal{V} \hookrightarrow \mathcal{H} \cong \mathcal{H}^* \hookrightarrow \mathcal{V}^*. \tag{2.97}$$

Let $A \in L(\mathcal{V}, \mathcal{V}^*)$, and $f \in \mathcal{V}^*$. We are interested in solving the following *abstract elliptic equation* on $\mathcal{V}$,

$$Au = f, \quad u \in \mathcal{V}. \tag{2.98}$$

Defining the bilinear form $\mathfrak{a} : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$, $\mathfrak{a}(u,v) :={}_{\mathcal{V}^*} \langle Au, v\rangle_{\mathcal{V}}$ the equation reads as

$$\mathfrak{a}(u,v) = f(v) \quad \forall v \in \mathcal{V}. \tag{2.99}$$

We shall assume that A is selfadjoint, i.e., $A = A^*$ (which implies that the bilinear form $\mathfrak{a}(\cdot,\cdot)$ is symmetric on $V \times V$ and coercive on V, i.e., there exists a real number $\gamma_0 > 0$ such that $\forall u \in V : \mathfrak{a}(u,u) \geqslant \gamma_0 \|u\|_{\mathcal{V}}^2$. Our assumption $A \in L(\mathcal{V}, \mathcal{V}^*)$ implies the existence of a positive real number $\gamma_1 = \|A\|_{L(\mathcal{V},\mathcal{V}^*)} \geqslant \gamma_0$ such that $\forall u, v \in V$: $|\mathfrak{a}(u,v)| \leqslant \gamma_1 \|u\|_{\mathcal{V}} \|v\|_{\mathcal{V}}$; i.e., the bilinear form a is bounded. Under these assumptions, the *energy norm* $\|\cdot\|_{\mathfrak{a}}$ defined by $\|v\|_{\mathfrak{a}} = (\mathfrak{a}(v,v))^{1/2}$ is equivalent to the $\mathcal{V}$ norm and we have

$$\gamma_0 \|v\|_{\mathcal{V}}^2 \leqslant \|v\|_{\mathfrak{a}}^2 \leqslant \gamma_1 \|v\|_{\mathcal{V}}^2. \tag{2.100}$$

We recall the following version of the Hilbert–Schmidt theorem

Theorem 2.23 *Suppose that $\mathcal{H}$ and $\mathcal{V}$ are separable Hilbert spaces, with $\mathcal{V}$ densely and compactly embedded in $\mathcal{H}$. Let $\mathfrak{a} : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ be a nonzero, symmetric, coercive and bounded bilinear form. Then, there exists a sequence of real numbers $(\lambda_n)_n$ and a* *sequence of unit $\mathcal{H}$-norm elements $(\varphi_n)_n$ in $\mathcal{V}$, which solve the following eigenvalue problem : find $\lambda \in \mathbb{R}$ and $\varphi \in \mathcal{V}\setminus\{0\}$ such that*

$$\mathfrak{a}(\varphi, v) = \lambda \langle \varphi, v \rangle_{\mathcal{H}} \quad \forall v \in \mathcal{V}. \tag{2.101}$$

The real numbers $(\lambda_n)_n$, $n \in \mathbb{N}$, which can be assumed to be in increasing order with respect to the index $n \in \mathbb{N}$, are positive, bounded away from 0, *and* $\lim_{n\to\infty} \lambda_n = \infty$. *In addition, the φ_n , $n \in \mathbb{N}$, form an orthonormal system in $\mathcal{H}$ whose closed span in $\mathcal{H}$ is equal to $\mathcal{H}$, and the scaled elements $\varphi_n/\sqrt{\lambda_n}$, $n \in \mathbb{N}$, form an orthonormal system with respect to the inner product defined by the bilinear form $\mathfrak{a}$, whose closed span with respect to the norm $\|\cdot\|_{\mathfrak{a}}$ induced by $a(\cdot,\cdot)$ is equal to $\mathcal{V}$. Furthermore,*

$$h = \sum_{n=1}^{\infty} \langle h, \varphi_n \rangle_{\mathcal{H}} \varphi_n, \quad \textit{and} \quad \|h\|_{\mathcal{H}}^2 = \sum_{n=1}^{\infty} \langle h, \varphi_n \rangle_{\mathcal{H}}^2 \quad \forall h \in \mathcal{H}, \tag{2.102}$$

and

$$v = \sum_{n=1}^{\infty} \mathfrak{a}\left(v, \frac{\varphi_n}{\sqrt{\lambda_n}}\right) \frac{\varphi_n}{\sqrt{\lambda_n}} \quad \textit{and} \quad \|v\|_{\mathfrak{a}}^2 = \sum_{n=1}^{\infty} \mathfrak{a}\left(v, \frac{\varphi_n}{\sqrt{\lambda_n}}\right)^2 \quad \forall v \in \mathcal{V}, \tag{2.103}$$

and in addition for $h \in \mathcal{H}$,

$$h \in \mathcal{V} \Leftrightarrow \sum_{n=1}^{\infty} \lambda_n \langle h, \varphi_n \rangle_{\mathcal{H}}^2 < \infty. \tag{2.104}$$

Thanks to Theorem 2.23, we know that there exists a sequence $\sigma = (\lambda_n)_n \subset \mathbb{R}_+$ of eigenvalues of A, with accumulation point at ∞, and a sequence of $\mathcal{H}$-orthonormal elements $(\varphi_\lambda)_{\lambda\in\sigma} \subset \mathcal{H}$ of associated eigenfunctions, i.e

$$A\varphi_\lambda = \lambda\varphi_\lambda, \quad \lambda \in \sigma, \text{ and } \quad \langle \varphi_\lambda, \varphi_{\lambda'} \rangle_{\mathcal{H}} = \delta_{\lambda,\lambda'}, \ \lambda, \lambda' \in \sigma. \tag{2.105}$$

In particular, therefore, the system $\{\varphi_\lambda : \lambda \in \sigma\}$ forms a normalized Riesz basis of H. The next two lemmas whose proofs are elementary show that renormalized versions of $\{\varphi_\lambda : \lambda \in \sigma\}$ constitute Riesz bases in $\mathcal{V}$ and $\mathcal{V}^*$ as well.

Lemma 2.24 *The following two-sided bound holds for each $v \in V$:*

$$\gamma_0 \|v\|_{\mathcal{V}}^2 \leqslant \sum_{\lambda\in\sigma} \lambda |v_\lambda|^2 \leqslant \gamma_1 \|v\|_{\mathcal{V}}^2. \tag{2.106}$$

For $f \in \mathcal{V}^*$, we have that

$$f = \sum_{\lambda\in\sigma} f_\lambda \varphi_\lambda, \quad f_\lambda :=_{\mathcal{V}^*} \langle f, \varphi_\lambda \rangle_{\mathcal{V}}, \ \lambda \in \sigma. \tag{2.107}$$

Lemma 2.25 *The following two-sided bound holds for each $f \in V^*$:*

$$\frac{1}{\gamma_1} \|f\|_{\mathcal{V}^*}^2 \leqslant \sum_{\lambda\in\sigma} \frac{1}{\lambda} |f_\lambda|^2 \leqslant \frac{1}{\gamma_0} \|f\|_{\mathcal{V}^*}^2 \tag{2.108}$$

Since we can write the solution of (2.98) as

$$u = A^{-1}f = \sum_{\lambda\in\sigma} u_\lambda \varphi_\lambda \quad \text{with} \quad u_\lambda = \lambda^{-1} f_\lambda, \quad \lambda \in \sigma, \tag{2.109}$$

we have the identity,

$$\sum_{\lambda\in\sigma} \lambda |u_\lambda|^2 = \sum_{\lambda\in\sigma} \lambda^{-1} |f_\lambda|. \tag{2.110}$$

By lemma 2.24, the left hand side of this identity belongs to the interval $[\|u\|_{\mathcal{V}}^2 \gamma_0, \gamma_1 \|u\|_{\mathcal{V}}^2]$, while the right hand side belongs to $[\|f\|_{\mathcal{V}*}^2 \gamma_1^{-1}, \gamma_0^{-1} \|f\|_{\mathcal{V}*}^2]$. From this we deduce that

$$\gamma_0 \|u\|_{\mathcal{V}}^2 \leqslant \|f\|_{\mathcal{V}*}^2 \leqslant \gamma_1 \|u\|_{\mathcal{V}}^2, \tag{2.111}$$

and that A^{-1} is a (bi-Lipschitz) quasi-isometric isomorphism between $\mathcal{V}^*$ and $\mathcal{V}$, when these spaces are equipped with the norms $\|\cdot\|_{\mathcal{V}*}$ and $\|\cdot\|_{\mathcal{V}}$ respectively.

Example 2.26 *Within the framework of Section 2.1, we consider an operator A on $W^{2,2}(\mathcal{H},\mu)$ defined as follows,*

$$Au(x) = u(x) - Tr(D^2u(x)) + \langle x, Q^{-1}Du(x)\rangle. \tag{2.112}$$

Using the integration by parts formula (2.65), we have the bilinear form

$$\mathfrak{a}(u,v) := \langle -Au, v\rangle_{\mathcal{H}} = \int_H u(x)v(x)\mu(\mathrm{d}x) + \int_H \langle Du(x), Dv(x)\rangle \mu(\mathrm{d}x), \quad u,v \in \mathcal{V}. \tag{2.113}$$

This is nothing other the scalar product on $\mathcal{V}$, so that $\gamma_0 = \gamma_1 = 1$.

2.3 Abstract parabolic problems

We now introduce the Bochner spaces $L^2([0,T];\mathcal{V})$, $L^2([0,T];\mathcal{H})$ and $L^2([0,T];\mathcal{V}^*)$ and we define

$$H^1([0,T];\mathcal{V}) := \left\{ u \in L^2([0,T];\mathcal{V}) \ : \ u' \in L^2([0,T];\mathcal{V}) \right\}, \tag{2.114}$$

where u' will signify $\mathrm{d}u/\mathrm{d}t$ or $\partial u/\partial t$ depending on the context. We are interested in the following *abstract evolution problem* in $[0,T]$,

$$u'(t) + Au(t) = f(t) \quad \text{with} \quad u(0) \in \mathcal{H}, \tag{2.115}$$

where $f \in L^2([0,T];\mathcal{H})$, and $A \in L(\mathcal{V},\mathcal{V}^*)$ with $\|A\|_{L(\mathcal{V},\mathcal{V}^*)} = \gamma_1 > 0$, satisfies $A = A^*$ and

$$\exists \gamma_0 > 0 \quad \forall v \in \mathcal{V} \ : \ \mathfrak{a}(v,v) := {}_{\mathcal{V}*}\langle Av, v\rangle_{\mathcal{V}} \geqslant \gamma_0 \|v\|_{\mathcal{V}}^2. \tag{2.116}$$

The bilinear form $\mathfrak{a}$ associated to A is hence continuous and coercive. The following theorem from [11] gives the existence of a solution to the evolution problem (2.115) in a more general setting.

Theorem 2.27 *Given $f \in L^2([0,T];\mathcal{V}^*)$ and a bilinear form $a(t;u,v) : \mathcal{V}\times\mathcal{V} \to \mathbb{R}$ with the following properties*

1. *For every $u, v \in \mathcal{V}$ the function $t \mapsto a(t; u, v)$ is measurable,*

2. *$|a(t; u, v)| \leqslant M \|u\|_{\mathcal{V}} \|v\|_{\mathcal{V}}$ for a.e. $t \in [0, T]$, $u, v \in \mathcal{V}$,*

3. *$a(t; v, v) \geqslant \alpha \|v\|_{\mathcal{V}}^2 - C \|v\|_{\mathcal{H}}^2$ for a.e. $t \in [0, T]$, $v \in \mathcal{V}$,*

for some non-negative constants α, C, M. For all $u_0 \in \mathcal{H}$, there exists a unique solution $u \in L^2([0, T], \mathcal{V}) \cap H^1([0, T], \mathcal{V}^*)$ satisfying

$$_{\mathcal{V}^*}\langle u'(t), v \rangle_{\mathcal{V}} + a(t; u, v) = f(t), \quad \text{a.e. } t \in [0, T], \forall v \in \mathcal{V}, \tag{2.117}$$

and

$$u(0) = u_0. \tag{2.118}$$

We define the solution space of (2.115) as $\mathcal{X} := L^2([0, T], \mathcal{V}) \cap H^1([0, T], \mathcal{V}^*)$ equipped with the norm

$$\|u\|_{\mathcal{X}} := \left(\|u\|_{L^2([0,T];\mathcal{V})}^2 + \|u'\|_{L^2([0,T];\mathcal{V}^*)}^2 \right)^{1/2}. \tag{2.119}$$

Under the assumptions on $\mathcal{V}$, $\mathcal{H}$, and $\mathcal{V}^*$ we have the continuous embedding $\mathcal{X} \hookrightarrow C([0, T]; \mathcal{H})$ (in the sense that any $v \in \mathcal{X}$ is equal almost everywhere to a function that is uniformly continuous as a mapping from the nonempty closed interval $[0, T]$ of the real line into $\mathcal{H}$). Therefore, for $u \in \mathcal{X}$ and $0 \leqslant t \leqslant T$, the values $u(t)$ are well-defined in $\mathcal{H}$ and there exists a constant $C = C(T) > 0$ such that

$$\forall u \in \mathcal{X} \quad \forall t \in [0, T] \; : \quad \|u(t)\|_{\mathcal{H}} \leqslant C \|u\|_{\mathcal{X}}. \tag{2.120}$$

For the numerical schemes discussed in chapter 4, we require a precise representation of the solution of (2.115). To this end we exploit that A is coercive as a bilinear form on $\mathcal{V} \times \mathcal{V}$, we hence see that $\|(\lambda I + A)v\|_{\mathcal{H}} \geqslant \lambda \|v\|_{\mathcal{V}}$ for all $v \in \mathcal{V}$ and by invertibility $(\lambda I + A)$ is surjective. By the Lumer-Phillips Theorem [14, p.14], $(-A)$ generates a strongly continuous semi-group of contractions $\{S_t\}_{t \in [0,T]}$ on $\mathcal{H}$, solution to

$$\lim_{\epsilon \to 0} \left\| \frac{S_{t+\epsilon}x - S_t x}{\epsilon} + Ax \right\|_{\mathcal{H}} = 0 \quad \forall x \in \mathcal{V}, \quad t > 0. \tag{2.121}$$

We shall write $\mathcal{D}(A) := \mathcal{V}$ to refer to the domain of definition of A is this context. For every $u \in \mathcal{D}(A)$, we have

$$AS_t u = S_t A u. \tag{2.122}$$

The semi-group $\{S(t)\}_{t \in [0,T]}$ associated to $-A$ is in fact $\{e^{-tA}\}_{t \in [0,T]}$. We shall say that a function $f \in L^2(0, T; \mathcal{H})$ is strongly differentiable if there exists a function $f' \in L^2([0, T]; \mathcal{H})$ such that

$$\lim_{\epsilon \to 0} \left\| \frac{f(t+\epsilon) - f(t)}{\epsilon} \right\|_{\mathcal{H}} = 0. \tag{2.123}$$

It is strongly continuously differentiable if in addition f' is continuous in t with respect to the $\mathcal{H}$-norm.

Theorem 2.28 ([1, p. 203]) *Given $u_0 \in \mathcal{V}$ and $f \in L^2(0,T;\mathcal{H})$, such that f is strongly continuously differentiable in $(0,T)$, with continuous derivative in $[0,T]$, there exists a unique solution $u \in \mathcal{X}$ to the evolution problem*

$$u'(t) + Au(t) = f(t) \quad 0 < t < T, \tag{2.124}$$

moreover it has the representation

$$u(t) = e^{-tA}u(0) + \int_0^t e^{-(t-s)A} f(s)ds, \tag{2.125}$$

and satisfies $\|u(t) - u_0\|_{\mathcal{H}} \to 0$ as $t \to 0$.

By strong continuity of the semi-group $\{e^{-tA}\}_{t\in[0,T]}$, the integral in (2.125) is interpreted as a Riemann integral in the topology $\|\cdot\|_{\mathcal{H}}$. In chapter 4 this allows to apply bounded linear operators to such integrals and have the integral and the operator commute; because the integral is seen as a limit of Riemann sums.

Remark 2.29 *Let $A \in L(\mathcal{V},\mathcal{V}^*)$ have the same properties as above, and let $\lambda \in \mathbb{R}$. Then the problem*

$$u'(t) + Au(t) + \lambda u = f(t), \quad t \in [0,T], \tag{2.126}$$

$$u(0) = u_0, \tag{2.127}$$

reduces to the problem (2.115) by setting $v := e^{\lambda t}u(t)$. In deed we can check that v satisfies

$$v'(t) + Av(t) = f(t), \quad t \in [0,T], \tag{2.128}$$

$$v(0) = u_0. \tag{2.129}$$

Hence the evolution problem is analogous and given an operator of the form $A + \lambda I \in L(\mathcal{V},\mathcal{V}^*)$, we reduce ourselves to the study of the simpler evolution problem.

Chapter 3

The Fokker-Planck equation

In this section we present a typical parabolic problem, the Fokker-Planck equation. In each case the solution lies in a weighted Hilbert space allowing for representations in Hermite polynomials expansions from section 2.1.2.

3.1 The finite dimensional problem

For $k = 1, 2, \dots K$ and $d \in \{1, 2, 3\}$, let D_k be either a bounded ball, centered at the origin, of radius $\sqrt{b_k}$ in $\mathbb{R}^d$, either $\mathbb{R}^d$. We can unify the two scenarios by identifying $\mathbb{R}^d$ with an open ball of radius $+\infty$, and taking $b_k \in (0, \infty]$, with either k finite for all $k = 1, 2, \dots, K$, either k infinite for all $k = 1, 2, \dots, K$. We define $D := D_1 \times D_2 \times \dots D_K$. On the interval $[0, \frac{b_k}{2})$ we consider the function $U_k \in C^1[0, \frac{b_k}{2})$, referred to as a potential, such that $U_k(0) = 0$, U_k is strictly monotonic increasing and $\lim_{s \to b_k^-/2} U_k = +\infty$, $k = 1, \dots, K$. We then associate with U_k the *partial Maxwellian*, defined by

$$M_k(q_k) := \frac{1}{\mathcal{Z}_k} \exp\left(-U_k\left(\frac{1}{2}|q_k|^2\right)\right), \quad q_k \in D_k,$$

where

$$\mathcal{Z}_k = \int_{D_k} \exp\left(-U_k\left(\frac{1}{2}|p_k|^2\right)\right) dp_k,$$

for $k = 1, \dots, K$, and we define the (full) Maxwellian

$$M(q) := M_1(q_1) \cdots M_K(q_K), \quad q = (q_1^\top, \dots, q_K^\top)^\top \in D \subseteq \mathbb{R}^{Kd}.$$

Clearly, $M(q) > 0$ on D, $\int_D M(q) dq = 1$ and $\lim_{|q_k| \to b_k} M_k(q_k) = 0$, $k = 1, \dots, K$. When the domains D_k are bounded balls, we shall suppose that there exist positive constants C_{k_1} and C_{k_2}, and real numbers $\alpha_k > 1$, $k = 1, \dots, K$, such that

$$0 < C_{k_1} \leqslant \exp\left(-U_k\left(\frac{1}{2}|q_k|^2\right)\right) / (\mathrm{dist}(q_k, \partial D_k))^{\alpha_k} \leqslant C_{k_2} < \infty, \quad k = 1, \dots, K.$$

Alternatively, when $D_k = \mathbb{R}^d$ for all $k = 1, \dots, K$, we shall assume that $U_k(s) = s$, $k = 1, \dots, K$.

We are interested in solving the Fokker-Planck equation,

$$\partial_t \psi - \sum_{i,j=1}^{K} A_{ij} \nabla_{q_i} \cdot \left(M \nabla_{q_j} \left(\frac{\psi}{M} \right) \right) = f, \quad (q,t) \in D \times [0,T], \tag{3.1}$$

subject to the initial condition

$$\psi(q,0) = \psi_0(q), \quad q \in D, \tag{3.2}$$

where $A \in \mathbb{R}^{K \times K}$ is a symmetric positive definite matrix with minimal eigenvalue $\gamma_0 > 0$ and maximal eigenvalue $\gamma_1 > 0$, $f \in C(0,T;L^2(D))$ and $\psi_0 \in L^1(D)$ is a nonnegative function such that $\int_D \psi_0 dq = 1$. As ψ_0 is a probability density function, and this property needs to be propagated during the course of the evolution over the time interval $[0,T]$, the boundary condition on $\partial D \times [0,T]$ needs to be chosen so that $\psi(\cdot,t)$ remains a nonnegative function for all $t \in [0,T]$, and $\int_D \psi(q,t) dq = 1$ for all $t \in [0,T]$. This can be achieved, formally at least, by demanding that

$$\sum_{i=1}^{K} A_{ij} \nabla_{q_i} \cdot \left(M \nabla_{q_j} \left(\frac{\psi}{M} \right) \right) \cdot \frac{q_j}{|q_j|} \to 0, \quad \text{as} \quad |q_j|^2 \to b_j, \quad j = 1, \dots K, \tag{3.3}$$

where either $b_j \in (0,\infty)$ for all $j = 1,\dots,K$ when D_j is a bounded ball of radius b_j; or $b_j := +\infty$ for all $j = 1,\dots,K$ when $D_j = \mathbb{R}^d$. By writing,

$$\hat{\psi} := \frac{\psi}{M}, \quad \hat{\psi}_0 := \frac{\psi_0}{M},$$

the initial-value problem (3.1), (3.2) can be restated as follows:

$$M \partial_t \hat{\psi} - \sum_{i,j=1}^{K} A_{ij} \nabla_{q_i} \cdot \left(M \nabla_{q_j} \hat{\psi} \right) = Mf, \quad (q,t) \in D \times [0,T], \tag{3.4}$$

subject to the initial condition

$$\hat{\psi}(q,0) = \hat{\psi}_0(q), \quad q \in D, \tag{3.5}$$

together with the (formal) boundary condition

$$\sum_{i=1}^{K} A_{ij} \nabla_{q_i} \cdot \left(M \nabla_{q_j} \hat{\psi} \right) \cdot \frac{q_j}{|q_j|} \to 0, \quad \text{as} \quad |q_j|^2 \to b_j, \quad j = 1, \dots K. \tag{3.6}$$

We consider the Maxwellian-weighted L^2 space

$$L^2_M(D) = \{ \hat{\varphi} \in L^2_{\text{loc}}(D) | \sqrt{M} \hat{\varphi} \in L^2(D) \}$$

equipped with the inner product $(\cdot,\cdot)_{L^2_M(D)}$ and norm $||\cdot||_{L^2_M(D)}$, defined, respectively, by

$$(\hat{\psi}, \hat{\varphi})_{L^2_M(D)} := \int_D M(q) \hat{\psi}(q) \hat{\varphi}(q) dq, \quad ||\hat{\varphi}||_{L^2_M(D)} := (\hat{\varphi}, \hat{\varphi})^{\frac{1}{2}}_{L^2_M(D)},$$

and the associated Maxwellian-weighted H^1 space

$$H^1_M(D) := \{\hat{\varphi} \in L^2_M(D) | \nabla_{q_k}\hat{\varphi} \in L^2_M(D), k = 1, 2, \dots K\}$$

equipped with the inner product $(\cdot,\cdot)_{H^1_M(D)}$ and norm $||\cdot||_{H^1_M(D)}$ defined, respectively, by

$$(\hat{\psi},\hat{\varphi})_{H^1_M(D)} := (\hat{\psi},\hat{\varphi})_{L^2_M(D)} + \sum_{k=1}^{K}(\nabla q_k\hat{\psi}, \nabla q_k\hat{\varphi})_{[L^2_M(D)]^d}, \quad ||\hat{\varphi}||_{H^1_M(D)} := (\hat{\varphi},\hat{\varphi})^{\frac{1}{2}}_{H^1_M(D)}.$$

Adopting the notations introduced in the previous sections, we take $\mathcal{H} := L^2_M(D), \mathcal{V} := H^1_M(D)$, and consider the linear differential operator

$$A\hat{\varphi} = -\sum_{i,j=1}^{K} A_{ij}\nabla_{q_j}\cdot(M\nabla_{q_i}\hat{\varphi}), \quad \hat{\varphi} \in \mathcal{V} \tag{3.7}$$

that maps $\mathcal{V}$ into its dual space $\mathcal{V}^*$.

As described in [17], under the assumptions on M the embedding $H^1_M(D) \hookrightarrow L^2_M(D)$ is dense and compact. Moreover, if we assume that f is strongly continuously differentiable and we strengthen our original assumption $\psi_0 \in L^1(D)$ by demanding that $\hat{\psi}_0 \in L^2_M(D)$ $(||\psi_0||_{L^1(D)} \leqslant ||\hat{\psi}_0||_{L^2_M(D)} = ||\hat{\psi}_0||_{\mathcal{H}}$ for all $\psi_0 \in \mathcal{H} = L^2_M(D))$, and since A is continuous and coercive, we know from Theorem 2.27 that there exists a unique solution $\hat{\psi}$ to the variational formulation in space of the problem : given $\hat{\psi}_0 \in \mathcal{H}$, find $\hat{\psi} \in \mathcal{X} := L^2(0,T;\mathcal{V}) \cap H^1(0,T;\mathcal{V}^*)$ such that for a.e. $t \in [0,T]$,

$$(\frac{\mathrm{d}}{\mathrm{d}t}\hat{\psi}(t), v)_{L^2_M(D)} +_{\mathcal{V}^*} \langle\hat{\psi}(t), v\rangle_{\mathcal{V}} = \langle f, v\rangle_{\mathcal{H}}, \quad \forall v \in \mathcal{V}, \tag{3.8}$$

$$\hat{\psi}(q,0) = \hat{\psi}_0(q), \quad q \in D. \tag{3.9}$$

If we further assume that $\psi_0 \in \mathcal{V}$, we can use theorem 2.28 and see the problem as an evolution problem on $\mathcal{X} := L^2(0,T;\mathcal{V}) \cap H^1(0,T;\mathcal{V}^*)$;

$$\hat{\psi}'(t) + A\hat{\psi}(t) = f(t), \quad \hat{\psi}(0) = \psi_0, \tag{3.10}$$

with solution $\hat{\psi}(t) = e^{-tA}\hat{\psi}_0 + \int_0^t e^{-(t-s)}\hat{\psi}(s)f(s)\mathrm{d}s$.

3.2 The Fokker-Planck equation in countably many dimensions

For $k \in \mathbb{N}$, let us denote by D_k the set $\mathbb{R}^d$ equipped with the Gaussian measure

$$\mu_k(\mathrm{d}q_k) := N_{a_k,\Sigma_k} = \frac{1}{(2\pi)^{d/2}\det(\Sigma_k)^{1/2}}\exp\left(-\frac{1}{2}(q_k - a_k)^\top\Sigma_k^{-1}(q_k - a_k)\right)\mathrm{d}q_k \tag{3.11}$$

with mean $a_k \in \mathbb{R}^d$ and positive definite covariance matrix $\Sigma_k \in \mathbb{R}^{d\times d}$. We shall assume henceforth that $a_k = 0$ for all $k \in \mathbb{N}$, and that the covariance operator Q, represented by the bi-infinite block- diagonal matrix

$$\Sigma = \mathrm{diag}(\Sigma_1, \Sigma_2, \dots) \tag{3.12}$$

with $d \times d$ diagonal blocks Σ_k, $k = 1, 2, \ldots$, is trace class. We define

$$D := \bigtimes_{k=1}^{\infty} D_k, \tag{3.13}$$

so that

$$q = (q_1^\top, q_2^\top, \ldots)^\top \in D, \quad q_k \in D_k, \quad k = 1, 2, \ldots \tag{3.14}$$

We equip the domain D with the product measure

$$\mu = \bigotimes_{k=1}^{\infty} \mu_k = \bigotimes_{k=1}^{\infty} N_{0,\Sigma_k}. \tag{3.15}$$

Let $\mathbf{M} = \{M_{ij}\}_{i,j=1}^{\infty} \in \mathbb{R}^{\infty \times \infty}$ be a symmetric infinite matrix, i.e., $M_{ij} = M_{ji}$ for all $i, j \in \mathbb{N}$. Suppose further that there exists a real number $\gamma_0 > 0$ such that

$$\sum_{i,j=0}^{\infty} M_{i,j} \xi_i \xi_j \geqslant \gamma_0 \left\|\xi\right\|_{\ell^2}^2, \quad \forall \xi = (\xi_i)_{i=0}^{\infty} \in \ell^2(\mathbb{N}), \tag{3.16}$$

a real number $\gamma_1 > 0$ such that

$$|\sum_{i,j=0}^{\infty} M_{i,j} \xi_i \eta_j| \leqslant \gamma_1 \left\|\xi\right\|_{\ell^2} \left\|\eta\right\|_{\ell^2}, \quad \forall \xi = (\xi_i)_{i=0}^{\infty}, \eta = (\eta_i)_{i=0}^{\infty} \in \ell^2(\mathbb{N}). \tag{3.17}$$

Using the abstract framework in section 2.3 we select $\mathcal{H} = L^2(D, \mu)$, $\mathcal{V} = W^{1,2}(D, \mu)$ and define

$$\mathcal{X} := L^2(0, T; \mathcal{V}) \cap H^1(0, T; \mathcal{V}^*). \tag{3.18}$$

With these spaces, given $f \in L^2(0, T; \mathcal{H})$ strongly continuously differentiable, and $\hat{\psi}_0 \in \mathcal{X}$ we formulate the *infinite dimensional Fokker-Planck equation* as the problem of finding $\hat{\psi} \in \mathcal{X}$ such that

$$_{\mathcal{V}^*}\langle \frac{\mathrm{d}}{\mathrm{d}t} \hat{\psi}(t), v \rangle_{\mathcal{V}} + \mathfrak{a}(\hat{\psi}(t), v) = \langle f(t), v \rangle_{\mathcal{H}}, \quad \text{a.e. } t \in [0, T] \quad \forall v \in \mathcal{V}, \tag{3.19}$$

$$\hat{\psi}(q, 0) = \hat{\psi}_0(q), \quad q \in D, \tag{3.20}$$

where

$$\mathfrak{a}(u, v) := \sum_{i,j=1}^{\infty} M_{i,j} \left\langle \nabla_{q_i} u, \nabla_{q_j} v \right\rangle_{[L^2(D,\mu)]^d}. \tag{3.21}$$

Assuming for instance that $d = 1$ for all $k \in \mathbb{N}$, we shall heuristically derive the strong formulation of this problem. We see in this case that the spectral decomposition of the operator Q is given. In this case, the variables q coincide with the decomposition on the complete μ-orthonormal system $(e_k)_k \subset H$ with sequence $(\lambda_n)_n \subset \mathbb{R}$ such that $Qe_k = \lambda_k e_k$, $k = 1, \ldots, n$, from Chapter

2.1. Assuming first that $M \in L_1^+(H)$, we may use formula (2.78) to have

$$\operatorname{Tr}(MD^2u(x)) = \sum_{k=1}^{\infty} \langle MD^2u(x)e_k, e_k\rangle \tag{3.22}$$

$$= \sum_{k=1}^{\infty} \langle D^2u(x)e_k, Me_k\rangle \tag{3.23}$$

$$= \sum_{k=1}^{\infty}\sum_{l=1}^{\infty} D_kD_lu(x)\langle Me_k, e_l\rangle \tag{3.24}$$

$$= \sum_{k=1}^{\infty}\sum_{l=1}^{\infty} M_{kl}D_kD_lu(x) \tag{3.25}$$

Using formula (2.65) and summing up over l and k shows that

$$-\int_H \operatorname{Tr}(MD^2u(x))v(x)\mu(\mathrm{d}x) = \sum_{k,l=1}^{\infty} \int_H M_{kl}\langle D_ku(x), D_lv(x)\rangle\mu(\mathrm{d}x) \tag{3.26}$$

$$-\int_H \langle x, v(x)Q^{-1}MDu(x)\rangle\mu(\mathrm{d}x). \tag{3.27}$$

Hence we infer that this operator corresponds to the second order operator L with $Lu(x) = -\operatorname{Tr}(MD^2u(x)) + \langle x, Q^{-1}MDu(x)\rangle$.

By chapter 2.1.2 we know that $\mathcal{V} \hookrightarrow \mathcal{H} \cong \mathcal{H}^* \hookrightarrow \mathcal{V}^*$, and we use again the abstract framework from section 2.3 to see this as an evolution problem in $\mathcal{X}$;

$$\hat{\psi}'(t) + A\hat{\psi}(t) = f(t), \quad \hat{\psi}(0) = \psi_0 \in \mathcal{V}, \tag{3.28}$$

where A is the linear operator from $\mathcal{V}$ into $\mathcal{V}^*$ induced by $\mathfrak{a}$, and the solution has a representation with the time propagator e^{-tA}.

Example 3.1 *We may take A tridiagonal depending on a bounded sequence $\epsilon = (\epsilon_i)_{i=1}^{\infty}$ with $\|\epsilon\|_{\ell^\infty(\mathbb{N})} < 1/2$. The matrix*

$$\mathbf{A}[\epsilon] = \operatorname{tridiag}\{(\epsilon_i, 1, \epsilon_i),\ i = 1, 2, \ldots\} \tag{3.29}$$

satisfies (3.16) with $\gamma_0 = 1 - 2\|\epsilon\|_{\ell^\infty}$ and (3.17) with $\gamma_1 = 1 + 2\|\epsilon\|_{\ell^\infty}$.

Chapter 4

Semi discretization in time

4.1 Semi-group approximations

We consider the general setting of abstract evolution problems presented in chapter 2.3, we let $\mathcal{H}$ and $\mathcal{V}$ be separable Hilbert spaces as in the triple (2.97). Moreover we assume that $A \in \mathcal{L}(\mathcal{V}, \mathcal{V}^*)$ is a linear, selfadjoint, positive definite operator, and $f \in L^2(0,T;\mathcal{H})$ is strongly continuously differentiable. We are interested in solving the initial value problem,

$$u'(t) + Au(t) = f(t) \quad \text{a.e } t \in [0,T], \quad \text{with} \quad u(0) = u_0 \in \mathcal{V}. \tag{4.1}$$

We use the representation given by Theorem 2.28 to write the solution as

$$u(t) = e^{-tA}u(0) + \int_0^t e^{-(t-s)A} f(s) \mathrm{d}s. \tag{4.2}$$

We first assume $f \equiv 0$. In order to approximate the solution (4.2) on the interval $[0,T]$, we introduce a time discretization. To this end we partition the interval $[0,T]$ into M equal time intervals $[t_i, t_{i+1}]$, $i = 0, \dots, M-1$, with $|t_{i+1} - t_i| = h$. We approximate the solution $u(t)$ on the nodes $\{t_i\}_{i=1,\dots,M}$, by a sequence of elements $\{U^n\}_{n=0,\dots,M} \subset \mathcal{V}$ with the Backward Euler scheme,

$$(I + hA)U^{n+1} = U^n, \quad U^0 = u_0. \tag{4.3}$$

We now introduce the Backward Euler operator,

$$B_h := (I + hA)^{-1}, \tag{4.4}$$

where $h > 0$ is small. As we shall see with Proposition 4.2 we have $\|B_h\|_{\mathcal{L}(\mathcal{H})} \leqslant 1$.

Proposition 4.1 *Let A be as above, then*

$$\frac{1}{h}(I - B_h)Av = AB_h v = B_h Av \quad \forall v \in \mathcal{V}, \quad \textit{and} \quad \lim_{h \to 0} B_h v = v \quad \forall v \in \mathcal{H}. \tag{4.5}$$

Proof The left identity can be verified easily and also follows from (2.122) and Proposition 4.2. To prove the other statement, we first assume $v \in \mathcal{D}(A)$. Then

$$\|v - B_h v\|_{\mathcal{H}} = h \left\| \frac{1}{h} B_h A v \right\|_{\mathcal{H}} \leqslant h \|Av\|_{\mathcal{H}}. \tag{4.6}$$

Taking $h \to 0$ gives the desired conclusion. □

The following proposition gives a useful representation of B_h. We shall use the notation S_t for the semi-group generated by $-A$.

Proposition 4.2 *With A as above, for every $u_0 \in \mathcal{H}$,*

$$B_h u_0 = \int_0^\infty \frac{e^{-t/h}}{h} S_t u_0 \mathrm{d}t. \tag{4.7}$$

Proof We first notice that since $\|S_t\|_{\mathcal{L}(\mathcal{H})} \leqslant 1\ \forall t \geqslant 0$, the integral is well defined. By setting

$$\tilde{B}_h u_0 := \int_0^\infty \frac{e^{-t/h}}{h} S_t u_0 \mathrm{d}t, \quad \forall u_0 \in \mathcal{H}, \tag{4.8}$$

we will show that $(I + hA)\tilde{B}_h u_0 = u_0$, for all $u_0 \in \mathcal{H}$. In deed, for any $\epsilon > 0$, using (2.122) one has

$$\frac{S_\epsilon \tilde{B}_h u_0 - \tilde{B}_h u_0}{\epsilon} = \frac{1}{\epsilon} \int_0^\infty \frac{e^{-t/h}}{h} (S_{t+\epsilon} u_0 - S_t u_0) \mathrm{d}t \tag{4.9}$$

$$= \frac{1}{h\epsilon} \int_0^\infty (e^{-(t-\epsilon)/h} - e^{-t/h}) S_t u_0 \mathrm{d}t - \frac{1}{h\epsilon} \int_0^\epsilon e^{-(t-\epsilon)/h} S_t u_0 \mathrm{d}t \tag{4.10}$$

$$= \frac{e^{\epsilon/h} - 1}{h\epsilon} \int_0^\infty e^{-t/h} S_t u_0 \mathrm{d}t - \frac{e^{\epsilon/h}}{h\epsilon} \int_0^\epsilon e^{-t/h} S_t u_0 \mathrm{d}t. \tag{4.11}$$

By taking the limit as $\epsilon \to 0^+$ one has,

$$\lim_{\epsilon \to 0^+} \frac{S_\epsilon \tilde{B}_h u_0 - \tilde{B}_h u_0}{\epsilon} = \frac{1}{h} \tilde{B}_h u_0 - \frac{1}{h} u_0, \tag{4.12}$$

which shows that $\tilde{B}_h u_0 \in \mathcal{D}(-A)$ and $A\tilde{B}_h u_0 = \frac{1}{h} u_0 - \frac{1}{h} \tilde{B}_h u_0$. From this we can retrieve $(I + hA)\tilde{B}_h u_0 = u_0$, for all $u_0 \in \mathcal{H}$. From the injectivity of $(I + hA)$ we see that in deed $\tilde{B}_h u_0 = (I + hA)^{-1} u_0$, $\forall u_0 \in \mathcal{H}$. □

4.2 Convergence analysis

By Proposition 4.2, we first see that for every $u_0 \in \mathcal{H}$,

$$\frac{\|B_h u_0\|_{\mathcal{H}}}{\|u_0\|_{\mathcal{H}}} \leqslant \frac{1}{\|u_0\|_{\mathcal{H}}} \int_0^\infty \frac{e^{-t/h}}{h} \|S_t\|_{\mathcal{L}(\mathcal{H})} \|u_0\|_{\mathcal{H}} \mathrm{d}t \leqslant 1, \tag{4.13}$$

since $\|S_t\|_{\mathcal{L}(\mathcal{H})} \leqslant 1$ for $t \geqslant 0$. This gives us the stability estimates

$$\|B_h\|_{\mathcal{L}(\mathcal{H})} \leqslant 1, \quad \text{and} \quad \|u(t)\|_{\mathcal{H}} \leqslant \|u_0\|_{\mathcal{H}}, \ \forall t \geqslant 1. \tag{4.14}$$

Also by Proposition 4.2, we can see the backward-Euler operator as a weighted average of the semi-group. More precisely, given a random variable X on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with exponential distribution of parameter $1/h$ the proposition allows us to write

$$B_h u_0 = \mathbb{E}(S_X u_0), \quad \forall u_0 \in \mathcal{H}. \tag{4.15}$$

In deed, the density w of X is $w(t) = h^{-1}e^{-t/h}\mathbb{1}_{[0,\infty[}(t)$,where $\mathbb{1}_{[0,\infty[}(t)$ is the indicator function of the positive real line. Furthermore, the density of a sum of n independent exponential random variables $X_1, \ldots, X_n$ of parameter $1/h$ is the n-convolution $w_n := w * w * \cdots * w$. By induction using (4.7) we can see that

$$B^n u_0 = \int_0^\infty w_n(t) S_t u_0 \mathrm{d}t \quad \forall u_0 \in \mathcal{H}, \tag{4.16}$$

which allows to write $B^n u_0 = \mathbb{E}(S_{\frac{1}{n}(X_1+\cdots+X_n)} u_0)$ for some independent and identically distributed exponential random variables $X_1, \ldots, X_n$ of parameter $1/h$. In our discretization setting, we take $h = T/M$ and approximate $u(T)$ by

$$U^M = B^M u_0 = \mathbb{E}(S_{\frac{1}{M}(X_1+\cdots+X_M)} u_0). \tag{4.17}$$

By the Law of Large Numbers, the quantity $\frac{1}{n}(X_1 + \cdots + X_n)$ converges almost surely to T, and since the semi-group is strongly continuous $S_{\frac{1}{M}(X_1+\cdots+X_M)} v$ converges almost surely to $S_T v$ in $\mathcal{H}$ $\forall v \in \mathcal{H}$. Finally since $\|S_t\|_{\mathcal{L}(\mathcal{H}} \leqslant 1$ for $t \geqslant 0$, by dominated convergence $B^M u_0 = \mathbb{E}(S_{\frac{1}{M}(X_1+\cdots+X_M)} u_0) \to S_T u_0$. Meaning that the Backward Euler approximations are consistent. We shall now proceed to the general case $f \neq 0$ and find the rate of convergence. The scheme is now given by

$$\frac{U^{n+1} - U^n}{h} + AU^{n+1} = f(t_{n+1}) \quad n = 1 \ldots M, \quad U^0 = u_0. \tag{4.18}$$

We shall make use of the following theorem ([4]).

Theorem 4.3 *Assume $u_0 \in \mathcal{D}(A^k)$ for some integer $k \geqslant 2$. Then the solution in Theorem 2.28 given by (4.2) satisfies*

$$u \in C^{k-j}([0,T], \mathcal{D}(A^j)) \quad \forall j = 0, \ldots, k. \tag{4.19}$$

Theorem 4.3 allows us to find easily the rate of convergence for a particular class of initial solutions.

Theorem 4.4 *For a given $u_0 \in \mathcal{D}(A^2)$ and $f \in L^2(0,T;\mathcal{H})$, such that f is strongly continuously differentiable in $(0,T)$, with continuous derivative in $[0,T]$, let $\{U_h^n\}_{n=1,\ldots,M}$ be the backward Euler approximations of $u(t_n)$, $n = 1, \ldots, M$. Then*

$$\left\|u(t_n) - U_h^n\right\|_{\mathcal{H}} = \mathcal{O}(h), \text{ as } h \to 0. \tag{4.20}$$

Proof Assuming that we have the exact solution at t_n, using the definition the approximation error for $u(t_{n+1})$ is given by

$$\hat{\epsilon}_n = u(t_{n+1}) - u(t_n) - h\left(f(t_{n+1}) - Au(t_{n+1})\right) \quad n = 1 \ldots M. \tag{4.21}$$

Since $\|B_h\|_{\mathcal{L}(\mathcal{H})} \leqslant 1$, we have the estimate

$$\left\|u(t_n) - U_h^n\right\|_{\mathcal{H}} \leqslant \sum_{k=1}^{M} \|\hat{\epsilon}_n\|_{\mathcal{H}}. \tag{4.22}$$

Consequently we are reduced to estimating $\|\hat{e}_n\|_{\mathcal{H}}$ for $n = 1 \dots M$. We can do it in the following way.

$$\|\hat{e}_n\|_{\mathcal{H}} = \|u(t_{n+1}) - u(t_n) - h\left(f(t_{n+1}) - Au(t_{n+1})\right)\|_{\mathcal{H}} \tag{4.23}$$

$$= \left\| \int_{t_n}^{t_{n+1}} u'(t)\mathrm{d}t - hu'(t_{n+1}) \right\|_{\mathcal{H}} \tag{4.24}$$

$$= \left\| h \sum_{i=1}^{m_n} \beta_{n,i} \left(u'(\tau_{n,i}) - u'(t_{n+1})\right) + w_\epsilon \right\|_{\mathcal{H}} \tag{4.25}$$

$$\leqslant h \sum_{i=1}^{m_n} \beta_{n,i} \left\|u'(\tau_{n,i}) - u'(t_{n+1})\right\|_{\mathcal{H}} + \epsilon. \tag{4.26}$$

Since by Theorem 4.3 $u'(t)$ is Riemann integrable, for $n = 1, \dots, M$ the nodes $\tau_{n,i}$ and the weights $\beta_{n,i}$ $i = 1, \dots, m_n$ exist (with $\sum_i \beta_{n,i} = 1$) as to approximate the integral of $u'(t)$ up to an error w_ϵ, with $\|w_\epsilon\|_{\mathcal{H}} < \epsilon$. By Theorem 4.3 $u'(t)$ is of bounded variation, so taking $\tau_{n,i}$ and $\beta_{n,i}$ $i = 1, \dots, m_n$ independent of n (as is possible) and summing over n then exchanging the summation over i and n gives,

$$\sum_{k=1}^{n} \|\hat{e}_n\|_{\mathcal{H}} \leqslant hV_0^T(u') + \tilde{\epsilon} = \mathcal{O}(h), \tag{4.27}$$

where $V_0^T(u')$ is the variation of u' on $[0, T]$. □

We now turn to the case $u_0 \in \mathcal{V} = \mathcal{D}(A)$. We need the following lemma ([4]).

Lemma 4.5 *Let $u_0 \in \mathcal{D}(A)$. Then $\forall \epsilon > 0$ $\exists \bar{u}_0 \in \mathcal{D}(A^2)$ such that $\|u_0 - \bar{u}_0\|_{\mathcal{H}} < \epsilon$ and $\|Au_0 - A\bar{u}_0\|_{\mathcal{H}} < \epsilon$. Which shows that $\mathcal{D}(A^2)$ is dense in $\mathcal{D}(A)$ (for the graph norm).*

Proof We set $\bar{u}_0 = B_h u_0$ for some $h > 0$ to be fixed later. We have

$$\bar{u}_0 \in \mathcal{D}(A), \quad \text{and} \quad \bar{u}_0 + hA\bar{u}_0 = \bar{u}_0. \tag{4.28}$$

Thus $A\bar{u}_0 \in \mathcal{D}(A)$, so that $\bar{u}_0 \in \mathcal{D}(A^2)$. On the other hand, by Proposition 4.1

$$\lim_{h\to0} \|B_h u_0 - u_0\|_{\mathcal{H}} = 0, \quad \lim_{h\to0} \|B_h Au_0 - Au_0\|_{\mathcal{H}} = 0, \quad \text{and} \quad B_h Au_0 = AB_h u_0. \tag{4.29}$$

The conclusion follows by taking $h > 0$ small enough. □

By Lemma 4.5, given $u_0 \in \mathcal{D}(A)$ we can construct a sequence (u_{0n}) in $\mathcal{D}(A^2)$ such that $u_{0n} \to u_0$ and $Au_{0n} \to Au_0$. By Theorem 4.4 we know that the solution u_n to

$$u_n'(t) + Au_n(t) = f(t) \quad \text{on } [0, T], \quad u_n(0) = u_{n0}, \tag{4.30}$$

is approximated at rate h by the backward Euler Scheme. On the other hand, the stability estimate gives

$$\|u_n(t) - u_m(t)\|_{\mathcal{H}} \leqslant \|u_{0n} - u_{0m}\|_{\mathcal{H}} \underset{m,n\to\infty}{\to} 0 \tag{4.31}$$

$$\left\|u_n'(t) - u_m'(t)\right\|_{\mathcal{H}} \leqslant \|Au_{0n} - Au_{0m}\|_{\mathcal{H}} \underset{m,n\to\infty}{\to} 0. \tag{4.32}$$

Therefore

$$u_n(t) \to u(t), \quad \text{uniformly on } [0,T], \tag{4.33}$$

$$u_n'(t) \to u'(t), \quad \text{uniformly on } [0,T], \tag{4.34}$$

with $u \in C^1([0,T],\mathcal{V})$. Passing to the limit in (4.30), using the fact that A is a closed operator, we see that $u(t) \in \mathcal{D}(A)$ and u is solution to (4.1). Since the functions of bounded variation form a Banach space, by (4.33) and (4.34) we have that u' is of bounded variation, moreover since $\mathcal{D}(A) = \mathcal{V} \subset\subset \mathcal{H}$ and B_h is continuous on $\mathcal{H}$, up to a subsequence

$$\left\| B_h^n u_0 - u(t_n) \right\|_{\mathcal{H}} \leqslant \underbrace{\left\| B_h^n u_0 - B_h^n u_{m0} \right\|_{\mathcal{H}}}_{\substack{\to 0 \\ m\to\infty}} + \mathcal{O}(h) + \underbrace{\left\| u_m(t_n) - u(t_n) \right\|_{\mathcal{H}}}_{\substack{\to 0 \\ m\to\infty}}. \tag{4.35}$$

Showing that the backward Euler approximations are $\mathcal{O}(h)$ also if $u_0 \in \mathcal{V}$.

Chapter 5

Adaptive Galerkin discretization in space

5.1 Discrete operator representations

5.1.1 Riesz bases

Definition 5.1 *A sequence of elements $\Phi^{\Xi} := \{\varphi_\nu : \nu \in \Xi\}$ in a Hilbert space $\mathcal{H}$ is called a Riesz basis for $\mathcal{H}$ if its associated synthesis operator,*

$$T_\Phi : \ell^2(\Xi) \to \mathcal{H} \;:\quad c = (c_\nu)_{\nu\in\Xi} \in \ell^2(\Xi) \mapsto \sum_{\nu\in\Xi} c_\nu \varphi_\nu \tag{5.1}$$

is boundedly invertible.

By identifying $\ell^2(\Xi)$ with its dual, the adjoint of T_Φ, known as the analysis operator, is

$$T_\Phi^* : \mathcal{H}^* \to \ell^2(\Xi) \;:\quad g \mapsto [g(\varphi_\nu)]_{\nu\in\Xi}. \tag{5.2}$$

The two values,

$$b_\Phi := \left\| T_\Phi^{-1} \right\|_{\mathcal{H}\to\ell^2(\Xi)} \quad \text{and} \quad B_\Phi := \|T_\Phi\|_{\ell^2(\Xi)\to\mathcal{H}} \tag{5.3}$$

are called the *Riesz bounds* of Φ. For all $f \in \mathcal{H}^*$,

$$b_\Phi \|f\|_{\mathcal{H}^*} \leqslant \left(\sum_{\nu\in\Xi} |f(\varphi_\nu)|^2 \right)^{1/2} \leqslant B_\Phi \|f\|_{\mathcal{H}^*}. \tag{5.4}$$

The Riesz basis Φ is called a *Parseval frame* if $b_\Phi = B_\Phi$. The *Riesz operator* is the self adjoint linear map,

$$S_\Phi := T_\Phi T_\Phi^* : \mathcal{H}^* \to \mathcal{H}, \quad f \mapsto \sum_{\nu\in\Xi} f(\varphi_\nu)\varphi_\nu. \tag{5.5}$$

The sequence $\Phi^* := S_\Phi^{-1}\Phi$ is a Riesz basis of $\mathcal{H}^*$, called the *canonical dual basis*. Its synthesis operator is $T_{\Phi^*} = S_\Phi^{-1} T_\Phi$. Since S_Φ^{-1} is self-adjoint, the Riesz operator of Φ^* is given by

$$S_{\Phi^*} = T_{\Phi^*} T_{\Phi^*}^* = S_\Phi^{-1} T_\Phi T_\Phi^* S_\Phi^{-1} = S_\Phi^{-1} S_\Phi S_\Phi^{-1} = S_\Phi^{-1}. \tag{5.6}$$

Moreover since

$$T_\Phi T_{\Phi^*}^* = T_\Phi T_\Phi^* S_\Phi^{-1} = \mathrm{id}_{\mathcal{H}} \quad \text{and} \quad T_{\Phi^*} T_\Phi^* = S_\Phi^{-1} T_\Phi T_\Phi^* = \mathrm{id}_{\mathcal{H}^*}, \tag{5.7}$$

writing the elements of Φ^* as $\varphi_\nu^* := S_\Phi^{-1}\varphi_\nu$, we have

$$w = \sum_{\nu\in\Xi} \varphi_\nu^*(w)\varphi_\nu, \ \ \forall w \in \mathcal{H} \quad \text{and} \quad f = \sum_{\nu\in\Xi} f(\varphi_\nu)\varphi_\nu^*. \tag{5.8}$$

For a Riesz basis Y indexed by Ξ, We shall denote by [Y] the infinite vector with entries indexed by Ξ. This will be convenient for further matrix notation hereafter.

5.1.2 Best N-term approximations and approximation classes

We are concerned with the general problem of approximating an element $f \in \mathcal{H}$, up to a desired accuracy, with as few memory usage as possible. To this end, we assume given a Riesz basis $\Psi = (\psi_\lambda)_{\lambda\in\Lambda}$ of the Hilbert space $\mathcal{H}$ and introduce the nonlinear N-term approximation manifold

$$\Sigma_N(\Psi) := \left\{ \mathbf{c}^\top \Psi \ : \ \#\{\lambda \in \Lambda \ : \ c_\lambda \neq 0\} \leqslant N \right\} \subset \mathcal{H},$$

and the N-term approximation error

$$\sigma_N(f,\Psi) := \inf_{g\in\Sigma_N} \|f-g\|_{\mathcal{H}}, \tag{5.9}$$

for an element $f \in \mathcal{H}$. We further define the nonlinear approximation space

$$\mathcal{A}^s(\Psi,\mathcal{H}) := \left\{ f \in \mathcal{H} \ : \ \sigma_N(\mathcal{H},\Psi) \lesssim N^{-s} \right\}, \tag{5.10}$$

with norm

$$\|f\|_{\mathcal{A}^s(\Psi,\mathcal{H})} := \inf\left\{ C > 0 \ : \ \sigma_N(f,\Psi) \leqslant CN^{-s} \right\}. \tag{5.11}$$

For $f \in \mathcal{H}$ we let $P_N(f)$ be the element of Σ_N that minimizes $\|f - f_N\|_{\mathcal{H}}$ over $f_N \in \Sigma_N$. The space $\mathcal{A}^s(\Psi,\mathcal{H})$ consists of the elements f of $\mathcal{H}$ whose best N-term approximations converge with rate s to f. It is in general not possible to find best N-term approximations, especially when the vector to be approximated is defined implicitly through a matrix equation. Nevertheless we present hereafter methods to approximate such solutions u, which whenever $u \in \mathcal{A}^s(\Psi,\mathcal{H})$, converge to the solution u with rate s. Moreover, the complexity of these methods is linear in N; the cardinality of the set of "active" coefficients necessary to represent the finitely supported approximation. That is, for $f \in \mathcal{A}^s(\Psi,\mathcal{H})$

$$\|f - P_N(f)\|_{\mathcal{H}} \leqslant \|f\|_{\mathcal{A}^s} N^{-s}, \quad \forall N \in \mathbb{N}_0. \tag{5.12}$$

When the space Hilbert space $\mathcal{H}$ will be ℓ^2, we shall omit the precision $\mathcal{H}$ in every of the above notations.

5.1.3 Compressibility

We now consider the problem of discretizing operators $A \in \mathcal{L}(\mathcal{H}, \mathcal{H}^*)$. Assuming that we have Riesz bases $\Psi = (\psi_\nu)_{\nu\in\Xi}$ and $\tilde{\Psi} = (\tilde{\psi}_\nu)_{\nu\in\Xi}$ of $\mathcal{H}$ and $\mathcal{H}^*$ respectively. We further assume that $\tilde{\Psi}$ is the canonical dual base of Ψ. We consider the action of A on an element f of $\mathcal{H}$ at the matrix level.

Proposition 5.2 *For a given $f \in \mathcal{H}^*$, an element $u \in \mathcal{H}$ satisfies*

$$Au = f, \tag{5.13}$$

if and only if $\mathbf{u} = T_\Psi^{-1} u$ satisfies

$$\mathbf{A}\mathbf{u} = \mathbf{f}, \tag{5.14}$$

where $\mathbf{A} = (\langle A\psi_\mu, \psi_\nu \rangle)_{\mu,\nu\in\Xi} \in \mathcal{L}(\ell^2(\Xi))$, and $f = T_{\tilde{\Psi}}\mathbf{f}$.

Consequently, we hereafter present methods to efficiently compute abstract matrix vector multiplications of the form

$$\mathbf{c}^\top \in \ell^2(\Xi) \mapsto \mathbf{A}\mathbf{c}. \tag{5.15}$$

Since the index sets of Ψ and $\tilde{\Psi}$ are the same we shall denote $\ell^2(\Xi)$ by ℓ^2.

Example 5.3 *With the framework from Chapter 2, we consider the elliptic equation in $\mathcal{V}$*

$$Au = f, \tag{5.16}$$

where A is given by

$$Au(x) = -Tr(D^2 u(x)) + \langle x, Q^{-1} Du(x) \rangle, \tag{5.17}$$

with associated bilinear form

$$\mathfrak{a}(u,v) = \int_H \langle Du(x), Dv(x) \rangle \mu(\mathrm{d}x), \quad u, v \in \mathcal{V}. \tag{5.18}$$

Choosing the polynomial chaos basis $\mathrm{Y} := (H_\gamma)_{\gamma\in\Gamma}$ from Theorem 2.9 for $\mathcal{H}$ and the rescaled orthonormal base $\Psi := (H_\gamma/(1 + \langle \gamma, \lambda^{-1} \rangle_*)^{1/2})_{\gamma\in\Gamma}$ from Theorem 2.12 for $\mathcal{V}$ and applying the above methodology gives the matrix $\mathbf{A}$ associated to this problem,

$$\mathbf{A} = \mathfrak{a}([\Psi]\,([\Psi]). \tag{5.19}$$

Denoting by $\mathbf{D}^\lambda$ the matrix $\operatorname{diag}\left\{ \left(1 + \langle \gamma, \lambda^{-1} \rangle_*\right)^{1/2}, \gamma \in \Gamma \right\}$, we have $\mathbf{D}^\lambda = \|[\mathrm{Y}]\|_{\mathcal{V}}$ and $(\mathbf{D}^\lambda)^{-1} = \|[\mathrm{Y}]\|_{\mathcal{V}^*}$, and we may write

$$[\mathrm{Y}]_{\mathcal{V}} := [\Psi] = (\mathbf{D}^\lambda)^{-1}\,[\mathrm{Y}]\,. \tag{5.20}$$

Using the bilinear form associated to A given by (5.18), we may write the entries $(\mathbf{A}_{\gamma,\nu})_{\gamma,\nu\in\Gamma}$ of the matrix with

$$(\mathbf{A})_{\gamma,\nu} = \mathfrak{a}(\psi_\gamma, \psi_\nu) = \sum_{k\in\mathbb{N}} (\mathbf{D}^\lambda_\gamma)^{-1} \left\langle D_k H_\gamma, D_k H_\nu \right\rangle_{\mathcal{H}} (\mathbf{D}^\lambda_\nu)^{-1}. \tag{5.21}$$

Clearly, if $\gamma = 0$, then $(A)_{\gamma,\nu} = \delta_{\gamma,\nu}$ for all $\nu \in \Gamma$; similarly, if $\nu = 0$, then $(A)_{\gamma,\nu} = \delta_{\gamma,\nu}$ for all $\gamma \in \Gamma$. We shall therefore focus our attention on the nontrivial case when $\gamma, \nu \in \Gamma \backslash \{0\}$; for any such γ, ν, we have by Proposition 2.11 that

$$(\mathbf{A})_{\gamma,\nu} = \sum_{k\in\mathbb{N},\gamma_k\geqslant 1,\nu_k\geqslant 1} (\mathbf{D}_\gamma^\lambda)^{-1}\sqrt{\frac{\gamma_k\nu_k}{\lambda_k^2}}\left\langle H_{\gamma_k-1}(W_{e_k})H_\gamma^{(k)}, H_{\nu_k-1}(W_{e_k})H_\nu^{(k)}\right\rangle_{\mathcal{H}}(\mathbf{D}_\nu^\lambda)^{-1} \tag{5.22}$$

$$=\delta_{\gamma,\nu}(\mathbf{D}_\gamma^\lambda)^{-1}(\mathbf{D}_\nu^\lambda)^{-1}\sum_{k\in\mathbb{N},\gamma_k\geqslant 1}\sqrt{\frac{\gamma_k^2}{\lambda_k^2}} \tag{5.23}$$

$$=\frac{\delta_{\gamma,\nu}\langle\gamma,\lambda^{-1}\rangle_*}{1+\langle\gamma,\lambda^{-1}\rangle_*}. \tag{5.24}$$

On the other hand, if we consider the operator A from Example 2.26; $Au(x) = u(x) - Tr(D^2u(x)) + \langle x, Q^{-1}Du(x)\rangle$, the bilinear form $\mathfrak{a}$ associated to A is no other than $\langle\cdot,\cdot\rangle_{\mathcal{V}}$. In this case the associated matrix is the identity since $\Psi := (H_\gamma/(1+\langle\gamma,\lambda^{-1}\rangle_*)^{1/2})_{\gamma\in\Gamma}$ is an orthonormal basis of $\mathcal{V}$.

Definition 5.4 *An operator $\mathbf{A} \in \mathcal{L}(\ell^2)$ is said n-sparse if each column contains at most n non-zero entries. It is s^*-compressible for $s^* \in (0,\infty]$ if there exists a sequence $(\mathbf{A}_j)_{j\in\mathbb{N}}$ such that $\mathbf{A}_j$ is n_j-sparse, with $(n_j)_{j\in\mathbb{N}} \in \mathbb{N}^{\mathbb{N}}$ satisfying*

$$c_{\mathbf{A}} := \sup_{j\in\mathbb{N}}\frac{n_{j+1}}{n_j} < \infty \tag{5.25}$$

and for every $s \in (0, s^*)$,

$$d_{\mathbf{A},s} := \sup_{j\in\mathbb{N}} n_j^s \left\|\mathbf{A} - \mathbf{A}_j\right\|_{\ell^2\to\ell^2} < \infty. \tag{5.26}$$

The operator $\mathbf{A}$ is strictly s^*-compressible if, in addition,

$$\sup_{s\in(0,s^*)} d_{\mathbf{A},s} < \infty. \tag{5.27}$$

We will use the approximation errors $e_{\mathbf{A},j} := \left\|\mathbf{A} - \mathbf{A}_j\right\|_{\ell^2\to\ell^2}$. The definition implies that these approximation errors satisfy $e_{\mathbf{A},j} \leqslant d_{\mathbf{A},s}n_j^{-s}$.

Definition 5.5 *An operator $\mathbf{A} \in \mathcal{L}(\ell^2)$ is s^*-computable for $s^* \in (0,\infty]$ if it is s^*-compressible with approximating sequence $(A_j)_{j\in\mathbb{N}}$ as in definition 5.4 such that there exists a routine*

$$\texttt{Build}_A[j,k] \mapsto \left[(l_i)_{i=1}^{n_j}, (a_i)_{i=1}^{n_j}\right], \tag{5.28}$$

with the k-th column of $\mathbf{A}_j$ equal to

$$\sum_{i=1}^{n_j} a_i \epsilon_{l_i}, \tag{5.29}$$

where ϵ_{l_i} is the Kronecker sequence that is 1 at l_i and 0 elsewhere, and there is a constant b_A such that the number of arithmetic operations and storage locations used by a call of $\texttt{Build}_A[j,k]$ is less than $b_A n_j$ for any $j \in \mathbb{N}$ and $k \in \mathbb{N}$.

Example 5.6 *We now reconsider Example 3.1, with*

$$\mathbf{M}[\epsilon] = \text{tridiag}\left\{(\epsilon_i, 1, \epsilon_i),\ i = 1, 2, \ldots\right\}, \tag{5.30}$$

for a bounded sequence $\epsilon = (\epsilon_i)_{i=1}^{\infty}$ with $\|\epsilon\|_{\ell^\infty(\mathbb{N})} < 1/2$. We follow [17], however for implementation considerations, we present here a construction of the bi-infinite matrix associated to the operator (3.21) 'column'-wise. The bilinear form $\mathfrak{a}$ reduces to

$$\begin{aligned} \mathfrak{a}(u, v) &= \sum_{i,j=1}^{\infty} M_{ij} \left\langle D_{q_i} u, D_{q_j} v \right\rangle_{L^2(D,\mu)} && (5.31)\\ &= \sum_{i=1}^{\infty} \left\langle D_{q_i} u, D_{q_i} v \right\rangle_{L^2(D,\mu)} && (5.32)\\ &+ \sum_{j=2}^{\infty} \epsilon_{j-1} \left\langle D_{q_{j-1}} u, D_{q_j} v \right\rangle_{L^2(D,\mu)} && (5.33)\\ &+ \sum_{j=1}^{\infty} \epsilon_j \left\langle D_{q_{j+1}} u, D_{q_j} v \right\rangle_{L^2(D,\mu)}, && (5.34) \end{aligned}$$

with the diagonal term having already been discussed. By superposition we may now confine ourselves to investigating the computability of the matrix $\mathbf{A}$ when the matrix $\mathbf{M}$ has entries $M_{ij} = \epsilon_i \delta_{i,j-1}$, $i, j = 1, 2, \ldots$, and when $\mathbf{M}$ has entries $M_{ij} = \epsilon_i \delta_{i,j+1}$, $i, j = 1, 2, \ldots$. The matrices $\mathbf{A}$ corresponding to these two cases will be denoted below by $\mathbf{A}^{(-)}$ and $\mathbf{A}^{(+)}$, and we denote their respective entries by $\mathbf{A}^{(-)}_{\gamma\nu}$ and $\mathbf{A}^{(+)}_{\gamma\nu}$, with $\gamma, \nu \in \Gamma$. For the sake of excluding trivial situations we shall assume henceforth that $\epsilon_i \neq 0, i = 1, 2, \ldots$. Instead of the orthonormal basis $(H_\gamma/(1 + \langle \gamma, \lambda^{-1} \rangle_*)^{1/2})_{\gamma\in\Gamma}$ of $\mathcal{V}$, we shall now take the Riesz basis $(H_\gamma/(\langle \gamma, \lambda^{-1} \rangle_*)^{1/2})_{\gamma\in\Gamma}$; as we shall see this basis has the advantage of scaling to 1 the diagonal entries of $\mathbf{A}$. Given $\gamma, \nu \in \Gamma\backslash\ \{0\}$, writing, as before,

$$\mathbf{D}^\lambda := \text{diag}\left\{ \left(\langle \gamma, \lambda^{-1} \rangle_* \right)^{1/2}, \gamma \in \Gamma \right\}, \tag{5.35}$$

we calculate that

$$\mathbf{A}^{(\pm)}_{\gamma\nu} = \sum_{\substack{i,j \in \mathbb{N}, \\ \gamma_i \geqslant 1, \nu_j \geqslant 1}} \epsilon_i \delta_{i\pm1,j} (\mathbf{D}^\lambda_\gamma)^{-1} \sqrt{\frac{\gamma_i \nu_j}{\lambda_i \lambda_j}} \left\langle H_{\gamma_i - 1} H^{(i)}_\gamma, H_{\nu_j - 1} H^{(j)}_\nu \right\rangle_{\mathcal{H}} (\mathbf{D}^\lambda_\nu)^{-1} \tag{5.36}$$

$$= \sum_{\substack{\binom{j \geqslant 2}{j \geqslant 1}, \\ \gamma_{j\mp1} \geqslant 1, \nu_j \geqslant 1}} \epsilon_{\substack{j-1 \\ j}} (\mathbf{D}^\lambda_\gamma)^{-1} \sqrt{\frac{\gamma_{j\mp1} \nu_j}{\lambda_{j\mp1} \lambda_j}} \left\langle H_{\gamma_{j\mp1} - 1} H^{(j\mp1)}_\gamma, H_{\nu_j - 1} H^{(j)}_\nu \right\rangle_{\mathcal{H}} (\mathbf{D}^\lambda_\nu)^{-1}, \tag{5.37}$$

with the notational convention $\nu_0 := 0$ and $\lambda_0 := \lambda_1 (> 0)$. Thus, for example, if $\gamma = 1_1 := (1, 0, 0, \ldots)$, then $\mathbf{A}^{(+)}_{\gamma,\nu} = 0$ for all $\nu \in \Gamma\backslash\{0\}$; $\mathbf{A}^{(-)}_{\gamma,\nu} = \epsilon_1$ for $\nu =$

$(0,\ 1,\ 0,\ 0,\ldots)$, and $\mathbf{A}^{(-)}_{\gamma\nu} = 0$ for all other $\nu \in \Gamma\backslash\{0\}$. Also, if $\gamma = 0$, then $\mathbf{A}^{(\pm)}_{0\nu} = \delta_{0\nu}$ for all $\nu \in \Gamma$; and, analogously, if $\nu = 0$, then $\mathbf{A}^{(\pm)}_{0\nu} = \delta_{\gamma 0}$ for all $\gamma \in \Gamma$.

It therefore remains to compute $\mathbf{A}^{(\pm)}_{\gamma\nu}$ for $\gamma, \nu \in \Gamma\backslash\{0, 1_1\}$. Hence, by defining for a fixed integer $j \geqslant 1$ and for $\gamma, \nu \in \Gamma\backslash\{0\}$ such that $\gamma_{j\mp 1} \geqslant 1$ and $\nu_j \geqslant 1$ the expression

$$\mathfrak{h}^{(\pm),(j)}_{\gamma\nu} := \left\langle H_{\gamma_{j\mp 1}-1} H^{(j\mp 1)}_{\gamma}, H_{\nu_j - 1} H^{(j)}_{\nu} \right\rangle_{\mathcal{H}}, \tag{5.38}$$

we deduce that

$$\mathfrak{h}^{(\pm),(j)}_{\gamma\nu} = \begin{cases} 1 & \textit{if } \gamma_j = \nu_j - 1 \wedge \gamma_{j\mp 1} = \nu_{j\mp 1} + 1 \wedge \nu_l = \gamma_l,\ l \in \{j, j\mp 1\}^c \cap \mathbb{N}, \\ 0 & \textit{otherwise.} \end{cases} \tag{5.39}$$

For $\gamma \in \Gamma$ let us define the support of γ by supp$(\gamma) := \{i \in \mathbb{N} : \gamma_i \neq 0\}$. Then, $\mathbf{A}_{\gamma\nu} = 0$ if supp$(\gamma)\cap$ supp$(\gamma) = \varnothing$. Therefore $\mathbf{A}$ is sparse. Now for each $\gamma, \nu \in \Gamma$ there is at most one $j \in$ supp(ν) such that $\mathfrak{h}^{(\pm),(j)} \neq 0$, when it exists, let it be denoted by $j_\pm$, in which case we have

$$\mathbf{A}^{(\pm)}_{\gamma\nu} = \epsilon_{\substack{j_\pm - 1 \\ j_\pm}} \sqrt{\frac{\gamma_{j_\pm \mp 1}\nu_{j_\pm}}{\lambda_{j_\pm \mp 1}\lambda_{j_\pm}}} \langle \gamma, \lambda^{-1}\rangle_*^{-1/2} \langle \nu, \lambda^{-1}\rangle_*^{-1/2}, \tag{5.40}$$

In any case, we deduce that in each 'column' with index $\nu \in \Gamma\backslash\{0, 1_1\}$ the matrix $\mathbf{A}^{(+)}$ contains nonzero off-diagonal entries only in 'rows' ν; for which there exists $j_+ \in$ supp(ν) such that $\mathfrak{h}^{(\pm),(j_+)} \neq 0$, and analogously for $\mathbf{A}^{(-)}$, showing that $\mathbf{A}$ is still very sparse.

We now refer to Definition 5.5 and verify condition (5.28): i.e., we wish to show that for $\bar{s} > 0$

$$c_{\mathbf{A},\bar{s}} := \sup_{N \in \mathbb{N}} N \left\| \mathbf{A} - \mathbf{A}^{[N]} \right\|^{1/\bar{s}}_{\ell^2(\Gamma)\to\ell^2(\Gamma)} < \infty \tag{5.41}$$

where $(\mathbf{A}^{[N]})_{N=1}^{\infty}$ is a sequence of infinite matrices, which we shall define below. We shall make use of Stechkin's lemma.

Lemma 5.7 (Stechkin) *Let $0 < p \leqslant q \leqslant \infty$ and assume that $\alpha = (\alpha_\gamma)_{\gamma\in\Gamma} \in \ell^p(\Gamma)$. For $N \geqslant 1$, let $\Gamma_N \subset \Gamma$ denote the set of indices corresponding to the N largest values of $|\alpha_\gamma|$. Then,*

$$\left(\sum_{\gamma \notin \Gamma_N} |\alpha_\gamma|^q \right)^{1/q} \leqslant N^{-r} \|\alpha\|_{\ell^p(\Gamma)}, \quad \textit{with } r := \frac{1}{p} - \frac{1}{q} \geqslant 0. \tag{5.42}$$

We define $\mathbf{A}^{[N]}$ 'column'-wise for $N \in \mathbb{N}$ as follows: if $N = 1$, we select $\mathbf{A}^{[N]}$ to be the diagonal part of $\mathbf{A}$. If $N > 1$, we define $\mathbf{A}^{[N]}$ to contain, in the off-diagonal of the 'column' associated with index $\nu \in \Gamma$, at most N nonzero elements of $\mathbf{A}_{\gamma\nu}$ where $\gamma = \gamma(\nu, i)$ with the index i such that $i \in \{j : \nu_j \neq 0 \wedge (\nu_j - 1 \neq 0 \vee \nu_j + 1 \neq 0)\} \cap \left\{j : \epsilon^{[N]}_j \neq 0\right\}$. Here, for a given sequence $\epsilon \in \ell^2(\mathbb{N})$ in the definition (5.30) of the infinite, tridiagonal matrix $\mathbf{M}$ appearing in the bilinear form $\mathfrak{a}(\cdot,\cdot)$, we denote by

$\epsilon^{[N]}$ its best $(N-1)$-term approximation in $\ell^2(\mathbb{N})$. Then, for all $N \in \mathbb{N}$ and for every $0 < p \leqslant 2$,

$$\left\|\epsilon - \epsilon^{[N]}\right\|_{\ell^2(\mathbb{N})} \leqslant N^{-(1/p-1/2)} \left\|\epsilon\right\|_{\ell^p(\mathbb{N})}. \tag{5.43}$$

We note that by definition of $\epsilon^{[N]}$, $\mathbf{A}^N$ is defined completely analogously 'row'-wise for every $\gamma \in \Gamma$, 'row' γ has at most N nonzero elements of $\mathbf{A}_{\gamma\nu}$ where $\nu = \nu(\gamma, i)$ with the index i satisfying a symmetric set of conditions. As described in [17] we have

$$\left\|\mathbf{A} - \mathbf{A}^{[N]}\right\|_{\ell^2 \to \ell^2} \leqslant \sup_{\gamma\in\Gamma} \sum_{\nu\in\Gamma} \left|\mathbf{A}_{\gamma\nu} - \mathbf{A}_{\gamma\nu}^{[N]}\right|^2. \tag{5.44}$$

It therefore follows from the definition of the entries $\mathbf{A}_{\gamma\nu}$ of $\mathbf{A}$ and from the above calculations that

$$\forall N \in \mathbb{N} : \left\|\mathbf{A} - \mathbf{A}^{[N]}\right\|_{\ell^2(\Gamma)\to\ell^2(\Gamma)} \leqslant \left\|\epsilon - \epsilon^{[N-1]}\right\|_{\ell^2(\mathbb{N})} \leqslant 2^{1/p-1/2} N^{-(1/p-1/2)} \left\|\epsilon\right\|_{\ell^p(\mathbb{N})}, \tag{5.45}$$

from which we deduce, with $C_p := 2^{(1/p-1/2)/\bar{s}}$ that

$$c_{\mathbf{A},\bar{s}} = \sup_{N\in\mathbb{N}} N \left\|\mathbf{A} - \mathbf{A}^{[N]}\right\|_{\ell^2(\Gamma)\to\ell^2(\Gamma)}^{1/\bar{s}} \leqslant C_p \sup_{N\in\mathbb{N}} \left(N^{1-(1/p-1/2)/\bar{s}}\right) \left\|\epsilon\right\|_{\ell^p(\mathbb{N})}^{1/\bar{s}} < \infty \tag{5.46}$$

provided that $\epsilon \in \ell^p(\mathbb{N})$ with $0 < p < 2$ and $\bar{s} = \bar{s}(p)$ is chosen as

$$0 < \bar{s} := 1/p - 1/2. \tag{5.47}$$

Referring to the definition of s^*-computability (cf. Definition 5.5), we infer that $\mathbf{A}$ is s^*-computable with any $0 < s \leqslant s^*(p)$ if the sequence in Example 3.1 belongs to $\ell^p(\mathbb{N})$ with some $0 < p < 2$, resp. with $s^* = 1/p - 1/2$ (this encompasses the previous case, if $p = 0$ is understood to indicate that ϵ is the zero sequence).

The routine `Build`$_A[\nu, k, \epsilon]$ presented hereafter gives the step by step procedure to construct a column of $\mathbf{A}_k$, when the sequence ϵ is decreasing, so that it's best k approximation is it's first k terms. For every $\gamma \in \Gamma$ let $(\gamma^{(i,j)}, a, b)$ denote the multi-index γ whose value on i and j is replaced by a and b respectively.

Routine 5.1 `Build`$_A[\nu, k, \epsilon] \mapsto \mathbf{v}$

```
$\mathbf{v} = \mathbf{1}_\nu$
for $j$ in $\mathrm{supp}(\nu)$ do
    if $j < k$ then
        if $\nu_{j-1} \geqslant 1$ then
            $\gamma_j^+ = (\nu^{(j,j-1)}, \nu_j + 1, \nu_{j-1} - 1)$
            $\mathbf{v}_{\gamma_j^+} = \epsilon_{j-1}\sqrt{\frac{\gamma_{j-1}\nu_j}{\lambda_{j-1}\lambda_j}}\langle\gamma_j^+, \lambda^{-1}\rangle_*^{-1/2}\langle\nu, \lambda^{-1}\rangle_*^{-1/2}$
        if $\nu_{j+1} \geqslant 1$ then
            $\gamma_j^- = (\nu^{(j,j+1)}, \nu_j + 1, \nu_{j+1} - 1)$
            $\mathbf{v}_{\gamma_j^-} = \epsilon_j\sqrt{\frac{\gamma_{j+1}\nu_j}{\lambda_{j+1}\lambda_j}}\langle\gamma_j^-, \lambda^{-1}\rangle_*^{-1/2}\langle\nu, \lambda^{-1}\rangle_*^{-1/2}$
```

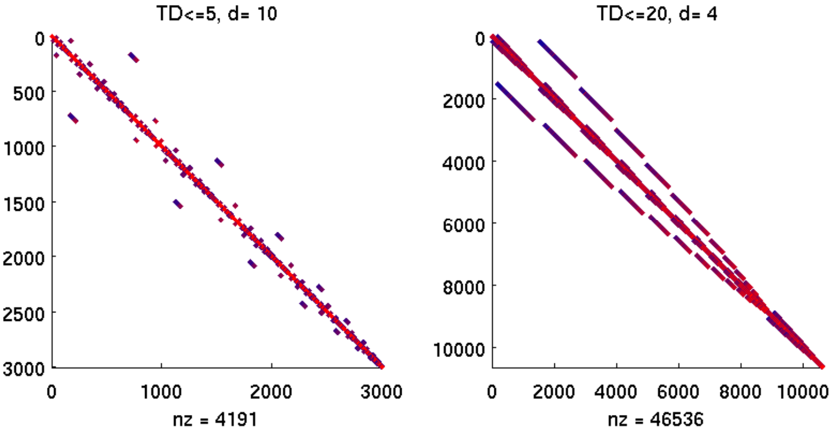

Figure 5.1: Visualization of finite parts of the matrix $\mathbf{A}$. The left figure shows a construction of the columns of $\mathbf{A}$ on a set of multi-indices $E \subset \Gamma$ with total degree (TD) less than or equal to 5 in the first 10 dimensions (d). On the right is the same construction for a set of multi-indices with total degree less than 21 in 4 dimensions. A color ranging from red to blue is associated to each entry, depending on its ℓ^∞ weight, with red chosen for the entries of maximum weight, and blue those of minimum. In both cases it is chosen $\epsilon_i = i^{-1.1}$ and Q diagonal with $\lambda_k = k^{-2}$.

5.1.4 Adaptive application of s*-computable operators

We now let $\mathbf{A} \in \mathcal{L}(\ell^2)$ be an s^*-computable operator. With approximating sequence $(\mathbf{A}_k)_{k\in\mathbb{N}}$ satisfying

$$\|\mathbf{A} - \mathbf{A}_k\|_{\ell^2\to\ell^2} \leqslant \bar{e}_{A,k}. \tag{5.48}$$

We present here a method to efficiently apply the matrix $\mathbf{A}$ to sequences $\mathbf{v} \in \ell^2$. Using the algorithms described in Appendix A such as

$$\texttt{BucketSort}[\mathbf{v},\epsilon] \mapsto \left[(v_{[p]})_{p=1}^P, (\Xi_p)_{p=1}^P\right], \tag{5.49}$$

we partition the vector $\mathbf{v} \in \ell^2$ into $v_{[p]} := v\restriction_{\Xi_p}$, $p = 1,\dots,P$, where $v_{[1]}$ contains the largest elements of $\mathbf{v}$, $v_{[2]}$ the next largests, and so on. The integer P is minimal with

$$2^{-P/2}\,\|\mathbf{v}\|_{\ell^\infty}\sqrt{\#\mathrm{supp}\,\mathbf{v}} \leqslant \epsilon. \tag{5.50}$$

Moreover the number of operations and storage locations required by a call of `BucketSort` is bounded by

$$\#\mathrm{supp}\,\mathbf{v} + \max(1, \lceil \log(\|\mathbf{v}\|_{\ell^\infty})\sqrt{\#\mathrm{supp}\ \mathbf{v}}/\epsilon \rceil). \tag{5.51}$$

Routine 5.2 $\texttt{Apply}_A[\mathbf{v}, \epsilon] \mapsto z$

```
$(\mathbf{v}_{[p]})_{p=1}^P \leftarrow \texttt{BucketSort}[\mathbf{v}, \frac{\epsilon}{2\bar{e}_{A,0}}]$
compute the minimal $l \in \{1, \ldots, P\}$ s.t $\delta := \bar{e}_{A,0} \left\| \mathbf{v} - \sum_{p=1}^l v_{[p]} \right\|_{\ell^2} \leqslant \frac{\epsilon}{2}$
$\mathbf{k} = (k_p)_{p=1}^l \leftarrow (0)_{p=1}^l$
while $\zeta_{\mathbf{k}} > \epsilon - \delta$ do
    $\mathbf{k} \leftarrow \texttt{NextOpt}[\mathbf{k}]$ with objective $-\zeta_{\mathbf{k}} = -\sum_{p=1}^l \bar{e}_{\mathbf{A},k_p} \|\mathbf{v}_{[p]}\|_{\ell^2(\Xi_p)}$
  and cost $\sigma_{\mathbf{k}} = \sum_{p=1}^l n_{k_p}(\#\mathrm{supp}\,\mathbf{v}_{[p]})$
$z \leftarrow \sum_{p=1}^l \mathbf{A}_{k_p} v_{[p]}$
```

The algorithm $\texttt{BucketSort}[\mathbf{v}, \epsilon]$ consists of three subtasks. The first task is to partition $\mathbf{v}$ into vectors with decreasing ℓ^∞ weight. The smaller elements are neglected, and this truncation produces an error of at most $\delta \leqslant \epsilon/2$. Next, the greedy algorithm `NextOpt` also detailed in appendix A assigns a sparse operator A_{k_p} to each section $v_{[p]}$, $p = 1, \ldots, P$. This optimization ensures that the final step, consisting of the sum of each ordinary matrix vector multiplication of the sparse operators with their assigned section approximates the abstract matrix vector multiplication up to an error ϵ, as desired. The algorithm enjoys additional properties which are detailed in the following theorem. However, to ensure that the algorithm terminates, we must assume that $(\bar{e}_{A,k})_{k \in \mathbb{N}_0}$ is nonincreasing and converges to 0; $n_0 = 0$ with $(n_k)_{k \in \mathbb{N}_0}$ strictly increasing, and

$$\eta_k := \frac{\bar{e}_{A,k} - \bar{e}_{A,k+1}}{n_{k+1} - n_k} \quad \text{nonincreasing in } k. \tag{5.52}$$

Moreover we assume

$$\bar{r}_A := \sup_{k \in \mathbb{N}_0} \frac{\bar{e}_{A,k}}{\bar{e}_{A,k+1}} < \infty. \tag{5.53}$$

Theorem 5.8 *For any finitely supported* $\mathbf{v} \in \ell^2$ *and any* $\epsilon > 0$, *a call of* $\texttt{Apply}_A[\mathbf{v}, \epsilon]$ *terminates, its output is a finitely supported* $\mathbf{z} \in \ell^2$ *with*

$$\|\mathbf{A}\mathbf{v} - \mathbf{z}\|_{\ell^2} \leqslant \delta + \zeta_{\mathbf{k}} \leqslant \epsilon, \tag{5.54}$$

where $\mathbf{k} = (k_p)_{p=1}^P$ *is the vector constructed by the greedy algorithm in* $\texttt{Apply}_A[\mathbf{v}, \epsilon]$. *Furthermore, the number of arithmetic operations required by the final step of* $\texttt{Apply}_A[\mathbf{v}, \epsilon]$ *is bounded by*

$$\sum_{p=1}^P n_{k_p}(\#\mathrm{supp}\ v_{[p]}) \tag{5.55}$$

if the relevant entries of $\mathbf{A}_{k_p}$ *are precomputed. If* $\sup_{k\in\mathbb{N}} \bar{e}_{A,k} n_k^s < \infty \; \forall s \in (0, s^*)$*, then for any* $s \in (0, s^*)$*,*

$$\#\mathrm{supp}\, \mathbf{z} \leqslant \sigma_{\mathbf{k}} \lesssim \epsilon^{-1/s} \|\mathbf{v}\|_{\mathcal{A}^s}^{1/s}. \tag{5.56}$$

5.1.5 Time stepping as a bi-infinite matrix vector equation

In what follows, we consider the backward Euler scheme (4.18). For $n = 0, \ldots, M-1$ we would like to solve

$$(I + hA)U^{n+1} = I\, U^n + hf(t_{n+1}), \tag{5.57}$$

where I is the identity operator from $\mathcal{V}$ to $\mathcal{V}^*$ with respect to the triple (2.97). We assume that we have a Riesz basis $\Upsilon = (\psi_\lambda)_{\lambda\in\Gamma}$ of $\mathcal{H}$ which renormalized in $\mathcal{V}$ or $\mathcal{V}^*$ gives the Riesz bases

$$\Psi = \left(\frac{\psi_\lambda}{\|\psi_\lambda\|_{\mathcal{V}}}\right)_{\lambda\in\Gamma} \quad \text{and} \quad \tilde{\Psi} = \left(\frac{\psi_\lambda}{\|\psi_\lambda\|_{\mathcal{V}^*}}\right)_{\lambda\in\Gamma} \tag{5.58}$$

of $\mathcal{V}$ or $\mathcal{V}^*$ respectively. We have their associated Riesz operators T_Ψ and $T_{\tilde{\Psi}}$. If we let $B := I + hA \in \mathcal{L}(\mathcal{V}, \mathcal{V}^*)$, the problem resumes to finding $\mathbf{U}^n \in \ell^2(\Gamma)$ for $n = 0, \ldots, M-1$ such that

$$\mathbf{B}\mathbf{U}^{n+1} = \mathbf{g}^n, \tag{5.59}$$

where $\mathbf{B} := T^*_{\tilde{\Psi}} \circ B \circ T_\Psi$, and $\mathbf{g}^n = T_{\tilde{\Psi}}(I\, U^n + hf(t_{n+1}))$. Letting $\mathbf{M} \in \mathcal{L}(\ell^2)$ denote the matrix $\langle I\Psi, \Psi\rangle_{\mathcal{H}}$ and $\mathbf{A} \in \mathcal{L}(\ell^2)$ the matrix $\langle A\Psi, \Psi\rangle_{\mathcal{H}}$, we may represent the operator $\mathbf{B}$ as

$$(\mathbf{B})_{\lambda,\lambda'} = \frac{\delta_{\lambda,\lambda'}}{\|\psi_\lambda\|_{\mathcal{V}} \|\psi_{\lambda'}\|_{\mathcal{V}}} + \frac{h}{\|\psi_\lambda\|_{\mathcal{V}} \|\psi_{\lambda'}\|_{\mathcal{V}}} \langle A\psi_\lambda, \psi_{\lambda'}\rangle_{\mathcal{H}} \tag{5.60}$$

where $(\mathbf{B})_{\lambda,\lambda'} = \left(T^*_{\tilde{\Psi}} \circ B\psi_\lambda\right)_{\lambda'}$. We may notice that the two terms in the above sum represent the additive contribution of the operators I and hA taken separately. We have the Riesz constants,

$$\Lambda_\Psi := \|T_\Psi\|_{\ell^2(\Gamma)\to\mathcal{V}} = \sup_{\mathbf{c}\in\ell^2(\Gamma)} \frac{\|T_\Psi \mathbf{c}\|_{\mathcal{V}}}{\|\mathbf{c}\|_{\ell^2(\Gamma)}}, \tag{5.61}$$

and,

$$\Lambda_{\tilde{\Psi}} := \|T_{\tilde{\Psi}}\|_{\ell^2(\Gamma)\to\mathcal{V}^*} = \sup_{\mathbf{c}\in\ell^2(\Gamma)} \frac{\|T_{\tilde{\Psi}} \mathbf{c}\|_{\mathcal{V}^*}}{\|\mathbf{c}\|_{\ell^2(\Gamma)}}, \tag{5.62}$$

giving us upper bounds for $\|\mathbf{B}\|_{\ell^2(\Gamma)\to\ell^2(\Gamma)}$ and $\|\mathbf{B}^{-1}\|_{\ell^2(\Gamma)\to\ell^2(\Gamma)}$;

$$\|\mathbf{B}\|_{\ell^2(\Gamma)\to\ell^2(\Gamma)} \leqslant \|B\|_{\mathcal{L}(\mathcal{V},\mathcal{V}^*)} \Lambda_\Psi \Lambda_{\tilde{\Psi}}, \tag{5.63}$$

$$\left\|\mathbf{B}^{-1}\right\|_{\ell^2(\Gamma)\to\ell^2(\Gamma)} \leqslant \frac{\|B^{-1}\|_{\mathcal{L}(\mathcal{V}^*,\mathcal{V})}}{\Lambda_\Psi \Lambda_{\tilde{\Psi}}}. \tag{5.64}$$

Using (5.63) and (5.64), we may bound the condition number of $\mathbf{B}$,

$$k_{\mathbf{B}} := \|\mathbf{B}\|_{\ell^2(\Gamma)\to\ell^2(\Gamma)} \left\|\mathbf{B}^{-1}\right\|_{\ell^2(\Gamma)\to\ell^2(\Gamma)} \leqslant \|B\|_{\mathcal{L}(\mathcal{V},\mathcal{V}^*)} \left\|B^{-1}\right\|_{\mathcal{L}(\mathcal{V}^*,\mathcal{V})}. \tag{5.65}$$

Taking $h < 1/\|A\|_{\mathcal{L}(\mathcal{V},\mathcal{V}*)}$, we may use a Neumann series argument to infer that $B^{-1} = (I + hA)^{-1}$ is well defined and continuous. We also obtain an estimate for the condition number,

$$k_{\mathbf{B}} \leqslant 1, \tag{5.66}$$

which holds uniformly in h for $h \leqslant 1/\|A\|_{\mathcal{L}(\mathcal{V},\mathcal{V}*)}$.

5.2 Adaptive Galerkin methods

We are now interested in solving the discrete operator equations (5.59). We take a more general approach and consider solving in $\ell^2(\Gamma)$ the bi-infinite matrix vector equation

$$\mathbf{A}\mathbf{u} = \mathbf{f}, \tag{5.67}$$

for a given $\mathbf{A} \in \mathcal{L}(\ell^2(\Gamma), \ell^2(\Gamma'))$ and $f \in \ell^2(\Gamma')$. We further assume that $\mathbf{A}$ is symmetric and positive definite, which is the case in (5.59). We assume that the action of $\mathbf{A}$ can be approximated by a routine

$$\texttt{Apply}_{\mathbf{A}}[\mathbf{v}, \epsilon] \to \mathbf{z}, \quad \|\mathbf{A}\mathbf{v} - \mathbf{z}\|_{\ell^2(\Gamma')} \leqslant \epsilon, \tag{5.68}$$

for finitely supported vectors $\mathbf{v}$. Similarly, we require a routine that given $\epsilon > 0$ produces an approximation

$$\texttt{RHS}_{\mathbf{f}}[\epsilon] \to \mathbf{g}, \quad \|\mathbf{f} - \mathbf{g}\|_{\ell^2(\Gamma')} \leqslant \epsilon, \tag{5.69}$$

to approximate the right hand side up to arbitrary precision. Moreover we require,

$$\texttt{RHS}_s := \sup_{0<\epsilon<\|\mathbf{f}\|_{\ell^2}} [\text{\# operations required by the call } \texttt{RHS}_{\mathbf{f}}]^s < \epsilon. \tag{5.70}$$

Some efficient methods to compute these approximations where presented in section 5.1.4 and in [10]. These two methods are combined into $\texttt{Residual}_{\mathbf{A},\mathbf{f}}$ which enables to compute the error between a given approximation and the solution to (5.67).

Routine 5.3 $\texttt{Residual}_{A,f}[\epsilon, \mathbf{v}, \eta_0, \chi, \omega, \beta] \mapsto [\mathbf{r}, \eta, \zeta]$

```
Require: $\zeta \leftarrow \chi\eta_0$
  repeat
    $\mathbf{r} \leftarrow \texttt{RHS}_f[\beta\zeta] - \texttt{Apply}_A[\mathbf{v}, (1-\beta)\zeta]$
    $\eta \leftarrow \|\mathbf{r}\|_{\ell^2}$
  if $\zeta \leqslant \omega\eta$ or $\eta + \zeta \leqslant \epsilon$ then
    break
  $\zeta \leftarrow \omega\frac{1-\omega}{1+\omega}(\eta + \zeta)$
```

Let $\|\mathbf{A}\| \leqslant \hat{\alpha}$ and $\|\mathbf{A}^{-1}\| \leqslant \check{\alpha}$. Then $k_{\mathbf{A}} := \hat{\alpha}\check{\alpha}$ is an upper bound for the condition number $\|\mathbf{A}\| \|\mathbf{A}^{-1}\|$. Furthermore, we let $\|\mathbf{f}\| \leqslant \lambda$. We now present an adaptive solver for the bi-infinite matrix equation.

Routine 5.4 $\texttt{Solve}_{A,f}[\epsilon,\chi,\theta,\omega,\sigma,\beta] \mapsto [u_\epsilon,\bar{\epsilon}]$

```
$\Xi^{(0)} \leftarrow \varnothing$,   $\tilde{\mathbf{u}}^{(0)} \leftarrow \mathbf{0}$,   $\delta_0 \leftarrow \check{\alpha}^{1/2}\lambda$
for $k = 0,1,\dots$ do
  if $\delta_k \leqslant \epsilon$ then
    break
    $[\mathbf{r}_k,\eta_k,\zeta_k] \leftarrow \texttt{Residual}_{A,f}[\epsilon\check{\alpha}^{-1/2},\tilde{\mathbf{u}}^{(k)},\hat{\alpha}^{1/2}\delta_k,\chi,\omega,\beta]$
    $\bar{\delta}_k \leftarrow \check{\alpha}^{1/2}(\eta_k+\zeta_k)$
  if $\bar{\delta}_k \leqslant \epsilon$ then
    break
    $[\Xi^{(k+1)},\rho_k] \leftarrow \texttt{Refine}[\Xi^{(k)},\mathbf{r}_k,\sqrt{\eta_k^2-(\zeta_k+\theta(\eta_k+\zeta_k))^2}]$
    $\bar{\theta}_k \leftarrow \left(\sqrt{\eta_k^2-\rho_k^2}-\zeta_k\right)/(\eta_k+\zeta_k)$
    $[\tilde{\mathbf{u}}^{(k+1)},\tau_{k+1}] \leftarrow \texttt{Galerkin}_{A,f}[\Xi^{(k+1)},\tilde{\mathbf{u}}^{(k)},\sigma\min(\delta_k,\bar{\delta}_k)]$
    $\delta_{k+1} \leftarrow \tau_{k+1}+\sqrt{1-\bar{\theta}_k^2\kappa_A^{-1}}\min(\delta_k,\bar{\delta}_k)$
  $u_\epsilon \leftarrow \tilde{\mathbf{u}}^{(k)}$
  $\bar{\epsilon} \leftarrow \min(\delta_k,\bar{\delta}_k)$
```

The method $\texttt{Solve}_{A,f}$ uses approximate residuals computed with $\texttt{Residual}_{A,f}$ to adaptively select and iteratively solve a finite section of (5.67). For a finite index set $\Xi \subset \mathbb{N}$ and a finitely supported $\mathbf{r} \in \ell^2$ and $\epsilon > 0$, the routine

$$\texttt{Refine}[\Xi,\mathbf{r},\epsilon] \mapsto [\tilde{\Xi},\rho], \tag{5.71}$$

constructs a set $\tilde{\Xi} \supset \Xi$ such that $\rho := \|r - r\restriction_{\tilde{\Xi}}\|_{\ell^2} \leqslant \epsilon$, and $\#\tilde{\Xi}$ is minimal with this property, up to a constant factor $\hat{c}$. For $\hat{c} = 1$ this can be done by sorting $\mathbf{r}$ and adding to Ξ the indices for which $|\mathbf{r}_i|$ is largest. Using an appropriate sorting algorithm, this can be done at a computational cost of order $\#\mathrm{supp}\,\mathbf{r}$.

The function $\texttt{Galerkin}_{A,f}$ approximates the solution of (5.67) restricted to the index set Ξ. It is the Galerkin projection on this set, that is

$$\mathbf{u}_{\text{Galerkin}}(\Xi) = \operatorname*{argmin}_{\#\mathrm{supp}\,\mathbf{w}\subset\Xi} \langle \mathbf{A}(\mathbf{w}-\mathbf{u}),\mathbf{w}-\mathbf{u}\rangle_{\ell^2} = \operatorname*{argmin}_{\#\mathrm{supp}\,\mathbf{w}\subset\Xi} \|\mathbf{w}-\mathbf{u}\|_{\mathbf{A}}\,. \tag{5.72}$$

A convergence analysis and optimality properties of this algorithm is detailed in [5], [7] and [10]. We have the following theorem.

Theorem 5.9 *Let $\epsilon,\chi,\theta,\omega > 0$, $\omega+\theta+\omega\theta \leqslant 1$, $0 < \sigma < 1-\sqrt{1-\theta^2\kappa_A^{-1}}$, and $0 < \beta < 1$, then $\texttt{Solve}_{A,f}$ constructs a finitely supported $\mathbf{u}_\epsilon$ with*

$$\|\mathbf{u}-\mathbf{u}_\epsilon\|_A \leqslant \epsilon, \tag{5.73}$$

and for all $k \in \mathbb{N}_0$,

$$\kappa_A^{-1/2}\frac{1-\omega}{1+\omega}\bar{\delta}_k \leqslant \left\|\mathbf{u}-\tilde{\mathbf{u}}^{(k)}\right\|_A \leqslant \min(\delta_k,\bar{\delta}_k). \tag{5.74}$$

Moreover, if $\mathbf{u} \in \mathcal{A}^s$ for some $s > 0$ and,

$$\hat{\theta} := \frac{\theta(1+\omega)+2\omega}{1-\omega} < \kappa_A^{-1/2}, \tag{5.75}$$

then by iteration k,

$$\left\| \mathbf{u} - \tilde{\mathbf{u}}^{(k)} \right\|_{\ell^2} \leqslant 2^s \hat{c}^s \kappa_A \tau^{-1} \rho (1 - \rho^{1/s})^{-s} \tfrac{1+\omega}{1-\omega} \left\| \mathbf{u} \right\|_{\mathcal{A}^s} \left(\# \Xi^{(k)} \right)^{-s} \tag{5.76}$$

for $\rho = \sigma + \sqrt{1 - \theta^2 \kappa_A^{-1}}$ *and* $\tau = \sqrt{1 - \theta^2 \kappa_A}$.

5.3 Convergence of the fully discrete problem

We shall now approximate the solution U^n, $n = 0, 1, \dots, M$ of the backward Euler scheme (5.57), with the algorithms presented above. Let $\tilde{U}^n$ be the fully discrete and finitely supported approximation of U_h^n, given by

$$\tilde{U}_h^n = T_\Psi \mathbf{U}^n, \tag{5.77}$$

and similarly we define $\tilde{U}_h^{n,(k)} = T_\Psi \tilde{\mathbf{U}}^{n,(k)}$, the approximation of $\tilde{U}_h^n$ by (5.59) after the k-th iteration of the previous algorithm. We assume that $U_h^n \in \mathcal{A}^s$, $n = 1, \dots, M$. The following theorem shows how the computational error $\left\| u(t_n) - \tilde{U}_h^n \right\|_{\mathcal{H}}$ is of order $\min(1, s)$.

Theorem 5.10 *Let* $u_0 \in \mathcal{V}$ *and* $f \in L^2(0, T; \mathcal{H})$*, such that* f *is strongly continuously differentiable in* $(0, T)$*, with continuous derivative in* $[0, T]$*. We assume that for* $n = 1, \dots, M$*,* $f(t_n)$ *satisfies the computability conditions for the right hand side (5.69) and (5.70). We further assume that* A *is an* s^**-computable operator satisfying all the properties of Theorem 2.23. Then for* $\epsilon, \chi, \theta, \omega > 0$*,* $\omega + \theta + \omega\theta \leqslant 1$*,* $0 < \sigma < 1 - \sqrt{1 - \theta^2 \kappa_A^{-1}}$*, and* $0 < \beta < 1$*, the approximation method defined by (5.59) satisfies,*

$$\left\| u(t_n) - \tilde{U}_h^{n,(k)} \right\|_{\mathcal{H}} \leqslant h V_0^T(u') \tag{5.78}$$

$$+ 2^s \hat{c}^s \kappa_A \tau^{-1} \rho (1 - \rho^{1/s})^{-s} \tfrac{1+\omega}{1-\omega} \left\| \mathbf{u} \right\|_{\mathcal{A}^s} \left(\# \Xi^{(k)} \right)^{-s}. \tag{5.79}$$

Proof We simply use the triangular inequality,

$$\left\| u(t_n) - \tilde{U}_h^{n,(k)} \right\|_{\mathcal{H}} \leqslant \left\| u(t_n) - U_h^n \right\|_{\mathcal{H}} + \left\| U_h^n - \tilde{U}_h^{n,(k)} \right\|_{\mathcal{H}}. \tag{5.80}$$

Assuming that a step $n - 1$ we have approximated $u(t_n)$ with precision $\epsilon/2$, under the assumption on f we can use $\texttt{RHS}_f[\epsilon/2]$ to approximate $I\ U^n + hf(t_{n+1}))$ up to precision ϵ. Estimation (5.79) is hence deduced from Theorem 5.9 (5.76), while (5.78) has been estimated in Chapter 4. □

Chapter 6

Implementation

In this final part of the project, we present numerical solutions to problem (2.115), in the particular case where A is the generator of the heat semigroup. In this setting, we take Q diagonal and solve for a given $u_0 \in UC_b(H)$, the problem

$$u'(t) - \Delta_Q u(t) = f(t) \tag{6.1}$$

$$u(0) = u_0, \tag{6.2}$$

with f strongly continuously differentiable from $[0, T]$ to H. The exact solution to this problem is given by

$$u(t,x) = \int_H u_0(y) N_{x,tQ}(\mathrm{d}y) + \int_0^t \int_H f(s,y) N_{x,(t-s)Q}(\mathrm{d}y) ds. \tag{6.3}$$

6.1 Associated matrix equation

When $A = -2\Delta_Q$, we have seen in (2.85) that the associated bilinear form is given by

$$_{\mathcal{V}*}\langle Au, v\rangle_{\mathcal{V}} = \sum_{k=1}^{\infty} \lambda_k \int_H (D_k u(x)^2 \mu(\mathrm{d}x) - \int_H \langle x, v(x) Du(x)\rangle \mu(\mathrm{d}x). \tag{6.4}$$

Let $\mathfrak{d}$ denote the nonsymmetric bilinear form on $\mathcal{V} \times \mathcal{V}$ defined by

$$\mathfrak{d}(u,v) := -\int_H \langle x, v(x) Du(x)\rangle \mu(\mathrm{d}x). \tag{6.5}$$

Following [17, Section 10] we now show that $\mathfrak{d}$ is continuous.

Proposition 6.1 *Assume that the covariance operator Q of the Gaussian measure μ on H is of trace class. Then, $\mathfrak{d} : W^{1,2}(H,\mu) \times W^{1,2}(H,\mu) \to \mathbb{R}$ is continuous and*

$$|\mathfrak{d}(u,v)| \leqslant \left(2\mathrm{Tr}(Q) \int_H |v(x)|^2 \mu(\mathrm{d}x) + 4\,\|Q\|^2 \int_H |Dv(x)|^2 \mu(\mathrm{d}x)\right)^{1/2} \|Du\|_{L^2(H,\mu)}\,. \tag{6.6}$$

Proof We estimate

$$|\mathfrak{d}(u,v)| \leqslant \left(\sum_{k=1}^{\infty} \| |x_k| v\|^2_{L^2(H,\mu)}\right)^{1/2} \left(\sum_{k\geqslant 1} \|D_k u\|^2_{L^2(H,\mu)}\right)^{1/2} \tag{6.7}$$

$$= \left(\sum_{k=1}^{\infty} \| |x_k| v\|^2_{L^2(H,\mu)}\right)^{1/2} \|Du\|^2_{L^2(H,\mu)} . \tag{6.8}$$

Using [12, Prop. 9.2.10] and the assumption that $v \in W^{1,2}(H,\mu)$, we deduce that, for a Gaussian measure μ with trace-class covariance Q,

$$\int_H \|x\|^2 (v(x))^2 \mu(\mathrm{d}x) \leqslant 2\mathrm{Tr}(Q) \int_H |v(x)|^2 \mu(\mathrm{d}x) + 4\|Q\|^2 \int_H \|Dv(x)\|^2 \mu(\mathrm{d}x). \tag{6.9}$$

This then yields the desired inequality (6.6). □

The bound (6.6) implies a Garding inequality for the Fokker–Planck operator with drift. Let $\mathfrak{a} : \mathcal{V} \times \mathcal{V}$ be defined as

$$\mathfrak{a}(u,v) := \sum_{k=1}^{\infty} \lambda_k \int_H D_k u(x) D_k v(x) \mu(\mathrm{d}x) \tag{6.10}$$

Proposition 6.2 *Assume that the covariance operator Q of the Gaussian measure μ is trace-class. Then, the bilinear form*

$$\mathfrak{a}(\cdot,\cdot) + \mathfrak{d}(\cdot,\cdot) : \mathcal{V} \times \mathcal{V} \to \mathbb{R} \tag{6.11}$$

is continuous and satisfies the Garding inequality (i.e it satisfies 2 and 3 in Theorem 2.27) in the triple $\mathcal{V} \subset \mathcal{H} \simeq \mathcal{H}^* \subset \mathcal{V}^*$. In particular, the variational problem is well-posed.

Proof The continuity of the bilinear form $\mathfrak{a}(\cdot,\cdot) + \mathfrak{d}(\cdot,\cdot)$ is evident from the previous proposition and the continuity of $\mathfrak{a}(\cdot,\cdot)$. The Garding inequality follows from the coercivity of $\mathfrak{a}(\cdot,\cdot)$ on $\mathcal{V} \times \mathcal{V}$ and from the continuity estimate (6.6) using a Cauchy inequality. □

6.2 s^*-computability of $\mathbf{B}_h$

We take $\Psi = (\psi_\gamma)_{\gamma\in\Gamma} := \left(H_\gamma / \langle \gamma, \lambda^{-1}\rangle_*^{1/2}\right)_{\gamma\in\Gamma}$ as a Riesz basis of $\mathcal{V}$. Let $\mathbf{B}_h :=_{\mathcal{V}^*} \langle [\Psi], [\Psi]\rangle_{\mathcal{V}} + h\left(\mathfrak{a}([\Psi],[\Psi]) + \mathfrak{d}([\Psi],[\Psi])\right)$. From the previous section we can easily compute the contributions of ${}_{\mathcal{V}^*}\langle [\Psi], [\Psi]\rangle_{\mathcal{V}}$ and $h\mathfrak{a}([\Psi],[\Psi])$, since for all γ, $\nu \in \Gamma$, using Proposition 2.11,

$${}_{\mathcal{V}^*}\langle \psi_\gamma, \psi_\nu\rangle_{\mathcal{V}} = \frac{\delta_{\gamma,\nu}}{\langle \gamma, \lambda^{-1}\rangle_*}, \tag{6.12}$$

$$h\mathfrak{a}(\psi_\gamma, \psi_\nu) = \frac{h}{\langle \gamma, \lambda^{-1}\rangle_*^{1/2} \langle \nu, \lambda^{-1}\rangle_*^{1/2}} \sum_{k=1}^{\infty} \lambda_k \int_H D_k H_\gamma D_k H_\nu \mu(\mathrm{d}x) \tag{6.13}$$

$$= \frac{h\delta_{\gamma,\nu}}{\langle \gamma, \lambda^{-1}\rangle_*} \sum_{k\in\mathrm{supp}(\gamma)} \gamma_k . \tag{6.14}$$

$$\tag{6.15}$$

We are hence interested in computing $\mathfrak{d}(\psi_\gamma, \psi_\nu)$ for $\gamma, \nu \in \Gamma$. We compute for $k \geqslant 1$,

$$\int_H x_k D_k H_\gamma(x) H_\nu(x) \mu(\mathrm{d}x). \tag{6.16}$$

Since $D_k H_\gamma(x) = \sqrt{\frac{\gamma_k}{\lambda_k}} H_\gamma^{(k)}(x) H_{\gamma_k - 1}(W_{e_k}(x))$ with

$$H_\gamma^{(k)}(x) = \prod_{j \neq k} H_{\gamma_j}(W_{e_j}(x)), \tag{6.17}$$

and $x_k = \lambda_k^{1/2} W_{e_k}(x)$, we rewrite (6.16) as

$$\begin{aligned}
(6.16) =& \sqrt{\gamma_k} \int_H W_{e_k}(x) H_{\gamma_k - 1}(W_{e_k}(x)) H_\gamma^{(k)}(x) H_\nu(x) \mu(\mathrm{d}x), && (6.18)\\
=& \sqrt{\gamma_k} \int_H \left(\sqrt{\gamma_k} H_{\gamma_k}(W_{e_k}(x)) + \sqrt{\gamma_k - 1} H_{\gamma_k - 2}(W_{e_k}(x)) \right) H_\gamma^{(k)}(x) H_\nu(x) \mu(\mathrm{d}x), \\
& && (6.19)\\
=& \gamma_k \delta_{\gamma,\nu} + \sqrt{\gamma_k(\gamma_k - 1)} \delta_{\nu, \gamma - 2\mathbb{1}_k}. && (6.20)
\end{aligned}$$

where we have used that $\xi H_n(\xi) = \sqrt{n+1} H_{n+1}(\xi) + \sqrt{n} H_{n-1}(\xi)$ and $\mathbb{1}_k$ is the multi-index with value one at k and 0 elsewhere, is the function γ with the k-th component decreased by two units. Since there is at most one k such that for $\gamma, \nu \in \Gamma$, $\delta_{\nu, \gamma - 2\mathbb{1}_k} \neq 0$, summing over k the last expression gives

$$(\mathbf{B}_h)_{\gamma,\nu} = \begin{cases} \frac{1}{\langle \gamma, \lambda^{-1} \rangle_*} & \text{if } \gamma = \nu, \\ \frac{-h\sqrt{\gamma_k(\gamma_k - 1)}}{\langle \gamma, \lambda^{-1} \rangle_*^{1/2} \langle \nu, \lambda^{-1} \rangle_*^{1/2}} & \text{if } \nu_k = \gamma_k - 2 \text{ and } \nu_l = \gamma_l,\ l \neq k. \end{cases} \tag{6.21}$$

For $m \in \mathbb{N}$, the following procedure enables the approximation of a column ν of $\mathbf{B}_h$, having at most $m + 1$ non-zeros entries.

Routine 6.1 `Build`$_B[\nu, m, \lambda] \mapsto \mathbf{v}$

```
$\mathbf{v}_\nu = \frac{1}{\langle \nu, \lambda^{-1} \rangle_*}$
for $j = 1$ to $m$ do
    $\mathbf{v}_{\nu + 2\mathbb{1}_j} = \frac{-h\sqrt{(\nu_j + 2)(\nu_j + 1)}}{\langle \nu + 2\mathbb{1}_j, \lambda^{-1} \rangle_*^{1/2} \langle \nu, \lambda^{-1} \rangle_*^{1/2}}$
```

Let $\mathbf{B}_h^N$ denote the matrix obtained when approximating all the columns $\nu \in \Gamma$ of $\mathbf{B}_h$ according to `Build`$_B[\nu, N, \lambda]$. To show s^*-computability we must bound

$$C_{\mathbf{B},\bar{s}} := \sup_{N \in \mathbb{N}} N \left\| \mathbf{B}_h - \mathbf{B}_h^N \right\|_{\ell^2(\Gamma) \to \ell^2(\Gamma)}^{1/\bar{s}}. \tag{6.22}$$

We shall use Schur's Lemma (cf. [16], p.6, ¶2, and [3] p.449, Theorem B) stated hereafter.

Theorem 6.3 (Schur) *A matrix $A := (a_{ij}) \in L(\ell^2(\mathbb{N}))$ if and only if there exist positive numbers C_1 and C_2 and a positive sequence $u := (u_j)_j$ such that*

$$\sum_{j=1}^{\infty} a_{ij} u_j^{1/2} \leqslant C_1 u_i^{1/2}, \quad i = 1, 2, \ldots \tag{6.23}$$

and

$$\sum_{i=1}^{\infty} a_{ij} u_j^{1/2} \leqslant C_2 u_j^{1/2}, \quad j = 1, 2, \dots, \tag{6.24}$$

in which case we have $\|A\|_{\ell^2(\mathbb{N})} \leqslant C_1^{1/2} C_2^{1/2}$.

We observe that for all $\gamma \in \Gamma$, $(\mathbf{B}_h^N)_{\gamma\nu} \neq 0$ only for $\nu = \gamma$ or $\nu = \gamma - 2\mathbb{1}_k$, $k \leqslant N$, and $(\mathbf{B}_h^N)_{\gamma\nu} = (\mathbf{B}_h)_{\gamma\nu}$ otherwise. This shows that for any 'row' $\gamma \in \Gamma$ of the matrix $\mathbf{B}_h - \mathbf{B}_h^N$ there are column entries for every $k > N$ such that $\gamma_k \geqslant 2$. On the other hand, for every 'column' $\nu \in \Gamma$, there are row entries for every $k > N$. We now verify the hypothesis of Schur's lemma (Theorem 6.3) for the positive sequence $u_\gamma = 1$, $\gamma \in \Gamma$ meaning that all rows and columns of $(\mathbf{B}_h)_{\gamma\nu} - (\mathbf{B}_h^N)_{\gamma\nu}$ are in $\ell^1(\Gamma)$. For any 'row' $\gamma \in \Gamma$,

$$\begin{aligned}
\sum_{\nu\in\Gamma} \left| (\mathbf{B}_h)_{\gamma\nu} - (\mathbf{B}_h^N)_{\gamma\nu} \right| &= h \sum_{\substack{k>N \\ \gamma_k \geqslant 2}} \left| \frac{\sqrt{\gamma_k(\gamma_k - 1)}}{\langle \gamma, \lambda^{-1} \rangle_*^{1/2} \langle \gamma - 2\mathbb{1}_k, \lambda^{-1} \rangle_*^{1/2}} \right| && (6.25) \\
&\leqslant h \left(\sum_{\substack{k>N \\ \gamma_k \geqslant 2}} \frac{\lambda_k (\gamma_k/\lambda_k)}{\langle \gamma, \lambda^{-1} \rangle_*} \right)^{1/2} \left(\sum_{\substack{k>N \\ \gamma_k \geqslant 2}} \frac{\gamma_k - 1}{\langle \gamma, \lambda^{-1} \rangle_* - 2\lambda_k^{-1}} \right)^{1/2} && (6.26) \\
&\leqslant h \left(\sum_{\substack{k>N \\ \gamma_k \geqslant 2}} \lambda_k \right)^{1/2} \left(\sum_{\substack{k>N \\ \gamma_k \geqslant 2}} 2\lambda_k \right)^{1/2} && (6.27) \\
&\leqslant h\sqrt{2} N^{-r} \|\lambda\|_{\ell^p(\mathbb{N})}, && (6.28)
\end{aligned}$$

where the last line follows from Stechkin's lemma (Lemma 5.7) for $0 < p < 1$ and $r = 1/p - 1$. We have also used the fact that for any $\gamma \in \Gamma$, $\gamma_k/\lambda_k \leqslant \langle \gamma, \lambda^{-1} \rangle_*$, $k \in \operatorname{supp}(\gamma)$. Also, if $\gamma_k > 2$,

$$\frac{\gamma_k - 1}{\langle \gamma, \lambda^{-1} \rangle_* - 2\lambda_k^{-1}} \leqslant \frac{\gamma_k - 1}{\gamma_k/\lambda_k - 2\lambda_k^{-1}} = \lambda_k \frac{\gamma_k - 1}{\gamma_k - 2} \leqslant 2\lambda_k, \tag{6.29}$$

while if $\gamma_k = 2$ we have

$$\frac{\gamma_k - 1}{\langle \gamma, \lambda^{-1} \rangle_* - 2\lambda_k^{-1}} = \frac{1}{\langle \gamma^{(k)}, \lambda^{-1} \rangle_*} \leqslant \frac{\lambda_j}{\gamma_j}, \tag{6.30}$$

with $j > k$ the next index $j \in \operatorname{supp}(\gamma)$ such that there is no index $l \in \operatorname{supp}(\gamma)$ with $k < l < j$. If no such index exists it is possible to pick again $j = \max(\operatorname{supp}(\gamma))$ without changing the validity of (6.27). On the other hand, for each column $\nu \in \Gamma$ we also have

$$\begin{aligned}
\sum_{\gamma\in\Gamma}\left|(\mathbf{B}_h)_{\gamma\nu} - (\mathbf{B}_h^N)_{\gamma\nu}\right| =& h\sum_{k>N}\left|\frac{\sqrt{(\nu_k+1)(\nu_k+2)}}{\langle\nu,\lambda^{-1}\rangle_*^{1/2}\langle\nu+2\mathbb{1}_k,\lambda^{-1}\rangle_*^{1/2}}\right| && (6.31)\\
\leqslant& h\left(\sum_{k>N}\frac{\lambda_k(\nu_k+2)/\lambda_k}{\langle\nu+2\mathbb{1}_k,\lambda^{-1}\rangle_*}\right)^{1/2}\left(\sum_{k>N}\frac{\nu_k+1}{\langle\gamma,\lambda^{-1}\rangle_*}\right)^{1/2} && (6.32)\\
\leqslant& h\left(\sum_{k>N}\lambda_k\right)^{1/2}\left(\sum_{\substack{k>N\\ \gamma_k\geqslant 2}}2\lambda_k\right)^{1/2} && (6.33)\\
\leqslant& h\sqrt{2}N^{-r}\left\|\lambda\right\|_{\ell^p(\mathbb{N})}, && (6.34)
\end{aligned}$$

for $0 < p < 1$ and $r = 1/p - 1$. By the above estimations, we can use Schur's Lemma (Theorem 6.3) for $C_1 = C_2 = h\sqrt{2}N^{-r}\left\|\lambda\right\|_{\ell^p(\mathbb{N})}$ to find that

$$\left\|\mathbf{B}_h - \mathbf{B}_h^N\right\|_{\ell^2(\Gamma)\to\ell^2(\Gamma)} \leqslant h\sqrt{2}N^{-r}\left\|\lambda\right\|_{\ell^p(\mathbb{N})}. \qquad (6.35)$$

This shows that for $\lambda \in \ell^p(\mathbb{N})$, $\mathbf{B}_h$ is s^*-computable with $s^* = 1/p - 1$. Moreover the approximating sequence satisfies (5.52) and (5.53).

6.3 Right hand side and initial condition

For the sake of simplicity, we consider a uniform forcing term of the form

$$f(t,x) = g(t), \quad t \in [0,T],\ x \in H. \qquad (6.36)$$

Trivially, $f(t,x) = g(t)H_0(x)$ so that the corresponding vector in $\ell^2(\Gamma)$ is given by $(g(t),0,0,\ldots)$. Moreover, we shall also assume given the full scaled Polynomial Chaos expansion of the initial condition,

$$u_0 = \sum_{\gamma\in\Gamma} u_0^\gamma \psi_\gamma, \quad \text{with } u_0^\gamma = \psi_\gamma^*(u_0), \qquad (6.37)$$

where $(\psi_\gamma^*)_{\gamma\in\Gamma}$ is the canonical dual base corresponding to Ψ defined through the Riesz operator S_Ψ in (5.5). In fact, $\Psi^* = \left(\langle\gamma,\lambda^{-1}\rangle_* H_\gamma\right)_{\gamma\in\Gamma}$ and the action on an element $f \in \mathcal{V}$ is defined as $\psi_\gamma^*(f) = \int_H\langle\gamma,\lambda^{-1}\rangle_* H_\gamma(x)f(x)\mu(\mathrm{d}x)$. We require the availability of a routine $\texttt{RHS}[g,\epsilon] \to \mathbf{g}$ such that for $g = u_0$ and $g = f$,

$$\|\mathbf{g} - T_\Upsilon^* g\|_{\ell^2(\Gamma)} \leqslant \epsilon, \text{ and } \#\mathrm{supp}(\mathbf{g}) \lesssim \min\left\{N : \|T_\Upsilon^* g\|_{\ell^2(\Gamma)-P_N g}\,\ell^2(\Gamma) \leqslant \epsilon\right\}, \qquad (6.38)$$

with the number of arithmetic operations and storage locations used by the call $\texttt{RHS}[g,\epsilon]$ being bounded by some absolute multiple of $\#\mathrm{supp}(\mathbf{g}) + 1$.

Example 6.4 *Consider for $0 < a < T$ in $[a,T] \times \mathbb{R}^n$ the parabolic problem*

$$\partial_t u(x,t) - div\left(M(x)A\nabla\left(\frac{u(x,t)}{M(x)}\right)\right) = 0, \quad (x,t) \in \mathbb{R}^n \times [a,T] \tag{6.39}$$

$$u(x,a) = u_0(x,a) \quad x \in \mathbb{R}^n, \tag{6.40}$$

$$M(x) = \frac{1}{(2\pi)^{n/2}} e^{-\frac{\|x\|^2}{2}}, \quad x \in \mathbb{R}^n. \tag{6.41}$$

where A is a symmetric positive definite matrix on $\mathbb{R}^n$. The case $a = 0$ can also be determined but is not considered here. By applying the methodology in section 3.1, we view this as the problem of finding $\tilde{u} := u/M$ solving

$$M(x)\partial_t \tilde{u}(x,t) - div\left(M(x)A\nabla\tilde{u}\right) = 0, \quad (x,t) \in \mathbb{R}^n \times [a,T] \tag{6.42}$$

$$\tilde{u}(x,a) = u_0(x,a)/M(x) \quad x \in \mathbb{R}^n. \tag{6.43}$$

For the sake of simplicity, we shall first take $A = I_n$ and consider the associated variational problem on the Maxwellian weighted L^2 space $\mathcal{H} := L^2(\mathbb{R}^n, N_I)$. Using the theory in Section 2.1.5, in particular (2.65) (2.90) and (2.91), we shall consider the solution to this equation given by

$$\tilde{u}(x,t) = (1 - e^{-2t})^{-n/2} \exp\left(-\frac{\|x\|^2}{2(e^{2t}-1)}\right), \tag{6.44}$$

with $u_0(x,a) = \tilde{u}(x,a)$. One can check that the function (6.44) satisfies (6.42).

In order to test the numerical schemes discussed, we take $\Upsilon := (H_\gamma)_{\gamma\in\Gamma}$ (with $\Gamma = \mathbb{N}^n$) as a complete orthonormal system on $\mathcal{H}$, and look for the solution under the form

$$\tilde{u}(x,t) = \sum_{\gamma\in\Gamma} \left\langle \tilde{u}(t), H_\gamma \right\rangle_{\mathcal{H}} H_\gamma(x). \tag{6.45}$$

Fortunately, in this particular case it is possible to find the full Hermite expansion of the solution and to compare it with numerical results. In deed, for any $t \in [a,T]$

$$\left\langle \tilde{u}(t), H_\gamma \right\rangle_{\mathcal{H}} = \int_{\mathbb{R}^n} \tilde{u}(t,x) H_\gamma(x) M(x) \mathrm{d}x \tag{6.46}$$

$$=(1 - e^{-2t})^{-n/2} \prod_{k=1}^{n} \int_{\mathbb{R}} \exp\left(-\frac{x_k^2}{2(e^{2t}-1)}\right) H_{\gamma_k}(x_k) N_1(\mathrm{d}x_k). \tag{6.47}$$

We are hence interested in computing for $m \geqslant 0$ the integral

$$I_m := \int_{\mathbb{R}} e^{-x^2/2\theta(t)} H_m(x) e^{x^2/2}(x) \mathrm{d}x, \tag{6.48}$$

with $\theta(t) := e^{2t} - 1$. By symmetry, we see that (6.48) vanishes for odd m. For m even, we shall use Proposition 9.1.1 in Da Prato and Zabczyk ([12, p.188]) inferring that for all $m \in \mathbb{N}$

$$H_m(x) = \frac{(-1)^n}{\sqrt{m!}} e^{x^2/2} \frac{\mathrm{d}^m}{\mathrm{d}x^m}(e^{-x^2/2}). \tag{6.49}$$

Hence using integration by parts ([8, p.22-29]),

$$C_m := I_{2m} = \frac{1}{\sqrt{(2m)!}} \int_{\mathbb{R}} e^{-x^2/2\theta(t)} \frac{\mathrm{d}^{2m}}{\mathrm{d}x^{2m}} (e^{-x^2/2}) \mathrm{d}x \tag{6.50}$$

$$= \frac{1}{\sqrt{(2m)!}} \int_{\mathbb{R}} \frac{x}{\theta(t)} e^{-x^2/2\theta(t)} \frac{\mathrm{d}^{2m-1}}{\mathrm{d}x^{2m-1}} (e^{-x^2/2}) \mathrm{d}x \tag{6.51}$$

$$= \frac{-\sqrt{(2m-1)!}}{\theta(t)\sqrt{(2m!)}} \int_{\mathbb{R}} e^{-x^2/2\theta(t)} x H_{2m-1}(x) e^{x^2/2} \mathrm{d}x \tag{6.52}$$

$$= \frac{-1}{\theta(t)\sqrt{2m}} \int_{\mathbb{R}} e^{-x^2/2\theta(t)} \sqrt{2m} H_{2m}(x) e^{-x^2/2} \mathrm{d}x \tag{6.53}$$

$$+ \frac{-1}{\theta(t)\sqrt{2m}} \int_{\mathbb{R}} e^{-x^2/2\theta(t)} \sqrt{2m-1} H_{2m-2}(x) e^{-x^2/2} \mathrm{d}x \tag{6.54}$$

$$= - \frac{1}{\theta(t)} C_m - \frac{1}{\theta(t)} \sqrt{\frac{2m-1}{2m}} C_{m-1} \tag{6.55}$$

$$= (-1)^m (\theta(t)+1)^{-m} \sqrt{\frac{(2m-1)(2m-3)\cdots 1}{2m(2m-2)\cdots 1}} C_0 \tag{6.56}$$

$$= (-1)^m \sqrt{\frac{2\pi\theta(t)}{\theta(t)+1}} (\theta(t)+1)^{-m} \frac{\sqrt{(2m)!}}{2^m (m!)}. \tag{6.57}$$

with $C_0 = \sqrt{2\pi\theta(t)/(\theta(t)+1)}$. *Using Stirling's approximation for* $m \geqslant M$ *and a predefined* $M \in \mathbb{N}$ *we can assume that the last expression can be evaluated in constant time. In this case, the full development of the solution* $\tilde{u}(t,x)$ *of (6.39) is found on the set of* $\tilde{\Gamma} := \{\gamma \in \Gamma : \gamma_k = 2m_k,\ \ k \in supp(\gamma)\}$ *and is given by* $\tilde{u}(x,t) = \sum_{\gamma \in \tilde{\Gamma}} \langle \tilde{u}(t), H_\gamma \rangle_{\mathcal{H}} H_\gamma(x)$ *with*

$$\langle \tilde{u}(t), H_\gamma \rangle_{\mathcal{H}} = \prod_{k=1}^{n} (-1)^{m_k} e^{-2m_k t} \frac{\sqrt{(2m_k)!}}{2^{m_k}(m_k!)}. \tag{6.58}$$

We also note here that the above development of the solution converges in $L^2(\mathbb{R}^n, N_I(\mathrm{d}x))$ *and not with respect to Lebesgue measure. Hence, approximations are good locally around the origin essentially.*

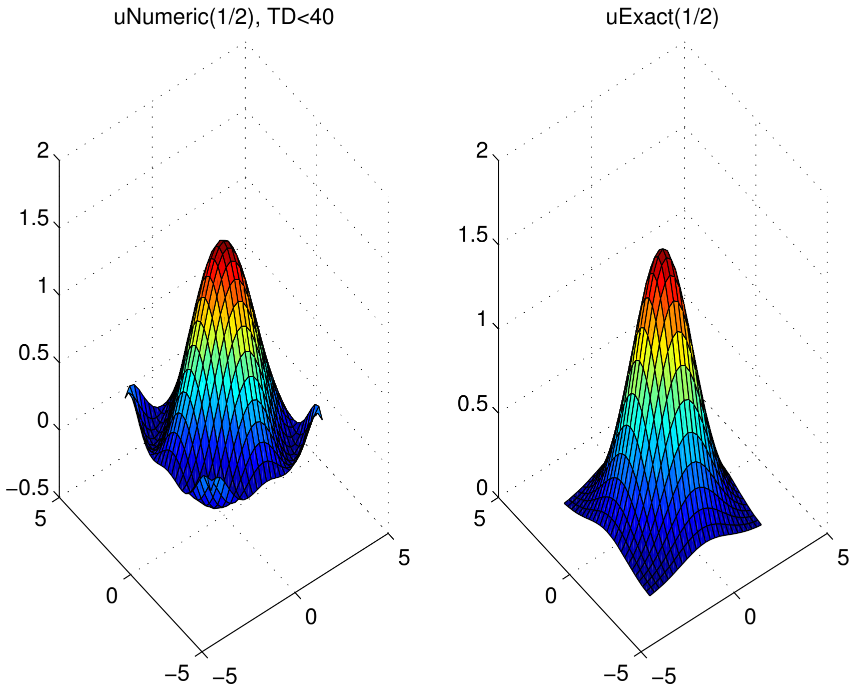

Figure 6.1: Visualization of the solution of (6.42) for $d = 2$, on the left is plotted the expansion of the solution on a set of bivariate Hermite polynomials with total degree at most 6. On the right is the exact solution at time $t = 0.5$. One notices a good approximation around the origin, which degrades after about two standard deviations away from 0. This is compensated after the change of variables in (6.39).

Equation (6.4) gives us the spectral decomposition from Theorem 2.23; the Hermite polynomials being the eigenfunctions of this problem. In deed, one has

$$
\begin{aligned}
D_k\left(e^{-\|x\|^2/2} D_k H_{\gamma_k}(x_k)\right) =& \sqrt{\gamma_k} D_k\left(e^{-\|x\|^2/2} H_{\gamma_k - 1}\right) && (6.59)\\
=& \sqrt{\gamma_k}(-1)^{\gamma_k - 1}(\gamma_k!)^{-1/2} D_k\left(D_k^{\gamma_k - 1}\left(e^{-\|x\|^2/2}\right)\right) && (6.60)\\
=& -\gamma_k e^{-\|x\|^2/2} H_{\gamma_k}. && (6.61)
\end{aligned}
$$

Hence, for an initial condition under the form $u_0 = \sum_\gamma u_0^\gamma H_\gamma$ we know that the solution at time t is given by $u(t) = \sum_\gamma u_0^\gamma e^{-|\gamma| t} H_\gamma$. It is possible to compute the energy of the solution at time $t > 0$ using (6.44), but it is also possible to use the above decomposition to see that

$$
\|u(t)\|_{\mathcal{H}}^2 = 1 + \mathcal{O}(e^{-t}), \quad \text{as } t \to \infty. \tag{6.62}
$$

For higher dimensions, due to the impossibility of plotting the resulting function, we show in Figure 6.2 the L^2 norm of the numerical solutions.

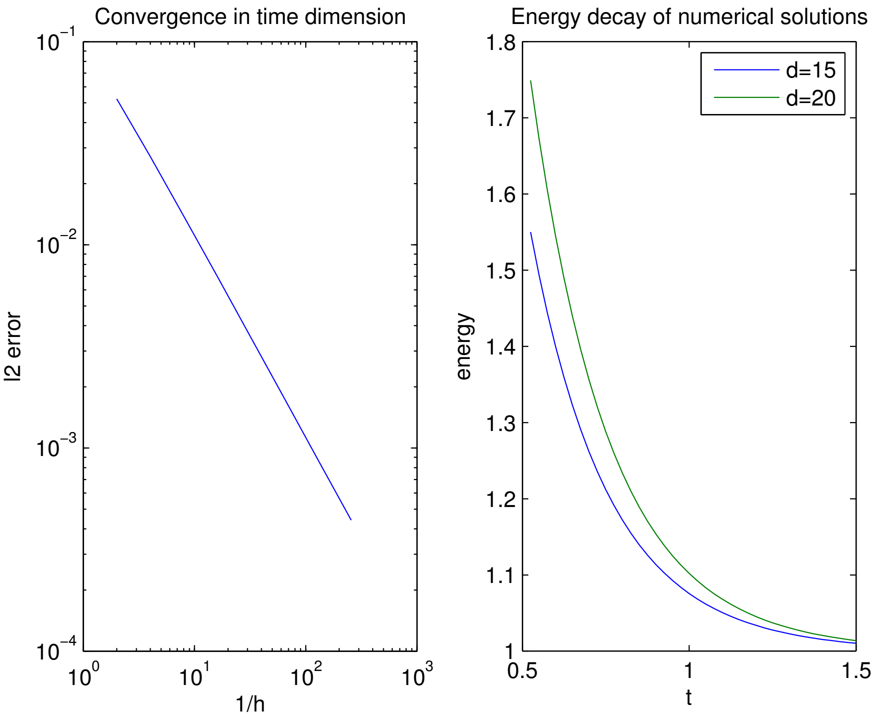

Figure 6.2: On the left we see the convergence of the Backward Euler approximations to the representation (6.58) as the step size decreases. The experiment was realized in dimension $d = 20$ and maximum polynomial degree 4. On the right is the plot of the energy decay of the solutions for high dimensional instances ($n = d$). The maximum polynomial degree is in this case also 4.

We have also considered numerical solutions to (6.42) for $A = A[\epsilon]$ *with*

$$\mathbf{A}[\epsilon] = \text{tridiag}\left\{(\epsilon_i, 1, \epsilon_i),\ i = 1, 2, \ldots, K\right\}. \tag{6.63}$$

In this particular case, we know from Example 5.6 that the associated bi-infinite matrix is symmetric. Moreover, for any support set of the approximation of the initial condition, we can make the approximate matrix square, and of full rank. This implies that the set of active coefficients is invariant during the numerical integration in time. The matrix is also very sparse, it is hence possible to consider an exact solver for relatively high dimensions, depending on the sparsity of the initial condition.

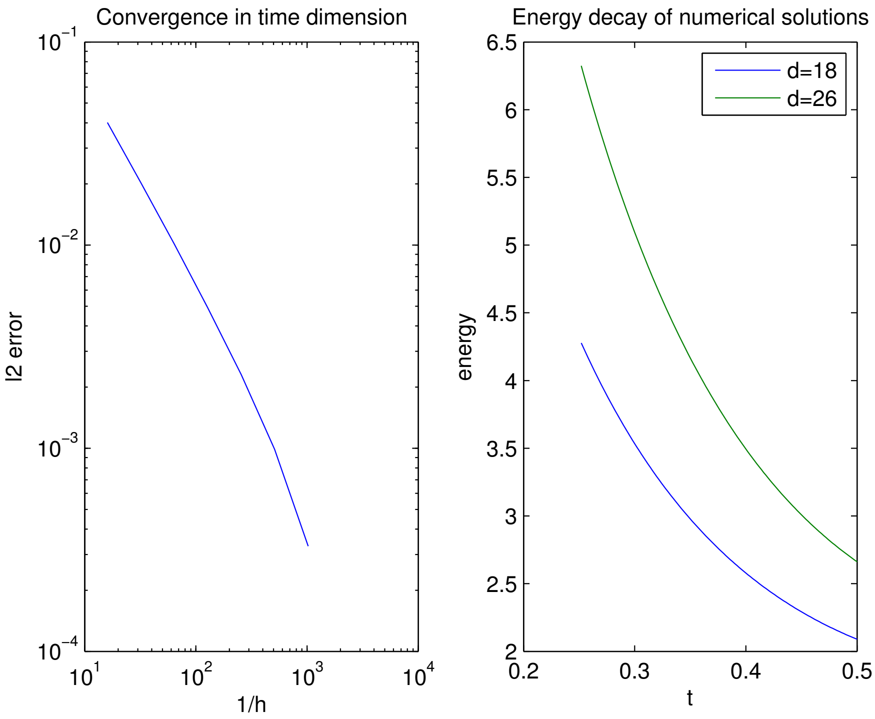

Figure 6.3: Test case for $K = 18, 26$ and $\epsilon_i = 2,\ i = 1, 2, \ldots, K$. The left plot shows linear convergence of the Backward Euler approximations. The right plot shows the L^2 norm of the solution decaying in time.

Bibliography

[1] A. V. Balakrishnan, *Applied functional analysis,* Springer 1976.

[2] V. Bogachev, G. Da Prato, and M. Rockner, *Existence and uniqueness of solutions for Fokker-Planck equations on Hilbert spaces*, J. Evol. Equ. 10 (2010), no. 3, 487–509.

[3] D. Borwein, X. Gao, *Matrix operators on ℓ^p to ℓ^q* , Canad. Math. Bull. **37** (1994), no. 4, 448-446.

[4] H. Brezis *Functional Analysis, Sobolev Spaces and Partial Differential Equations*, Universitext, 2010.

[5] A. Cohen, W. Dahmen, and R. DeVore, *Adaptive wavelet methods for elliptic operator equations – Convergence rates*, Math. Comp 70 (2001), 27–75.

[6] S. Cerrai, *Second Order PDE's in Finite and Infinite Dimension*, Lecture Notes in Mathematics, vol. 1762, Springer-Verlag, Berlin, 2001, A probabilistic approach.

[7] —, *Adaptive wavelet methods II - Beyond the elliptic case*, Found. Comput. Math. 2 (2002), no. 3, 203–245.

[8] E. Feldheim, *Quelques nouvelles relations pour les polynomes d'Hermite*, Journal London Math. Soc, 13 (1938).

[9] C. J. Gittelson, *Adaptive Wavelet Methods for Elliptic Partial Differential Equations with Random Operators*, Tech. Rep. 2011-37, Seminar for Applied,Mathematics, ETH Zurich, 2011.

[10] C. J. Gittelson, *Adaptive Galerkin Methods for Parametric and Stochastic Operator Equations*, PhD thesis, ETH Zurich, 2011. ETH Dissertation No. 19533.

[11] J.L. Lions, E. Magenes, *Non-homogeneous Boundary Value Problems and Applications* (3 volumes), Springer, 1972.

[12] G. Da Prato and J. Zabczyk, *Second Order Partial Differential Equations in Hilbert Spaces*, London Mathematical Society Lecture Note Series, vol. 293, Cambridge University Press, Cambridge, 2002. (2004e:47058)

[13] G. Da Prato, *An introduction to infinite-dimensional analysis*, Berlin : Springer, 2006 .

[14] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer Verlag, New York, 1983.

[15] A. Piechs, *The Ornstein-Uhlenbeck Semigroup in an Infinite Dimensional L^2 Setting*, J. Funct. Analy. (1975), 271-285.

[16] I. Schur, *Bemerkungen zur Theorie der Beschrankten Bilinearformen mit unendlich vielen Veranderlichen*, J.reine angew. Math. 140 (1911), 1-28.

[17] Ch. Schwab, E. Sülli *Adaptive Galerkin approximation algorithms for partial differential equations in infinite dimensions*, Tech. Rep. 2011-69, Seminar for Applied Mathematics, ETH Zurich, 2011.

[18] R. Spigler, M. Vianello (1995), *Convergence analysis of the semi-implicit euler method for abstract evolution equations*, Numerical Functional Analysis and Optimization, 16:5-6, 785-803

[19] T. Vidar *Galerkin finite element methods for parabolic problems* Springer, 2006.

Appendix A

Generalized knapsack problems

A.1 Problem setting

The adaptive solver relies on the possibility of iteratively finding appropriate sparse operators increasing the *value* of an approximate solution, in terms of residuals, at a corresponding computational *cost*. The problem of choosing the appropriate operators is hence seen as *generalized knapsack problem*. We specify this hereafter following [10]. Let $\mathcal{M} \in \mathbb{N}_0$, and for each $m \in \mathcal{M}$ let $(c_j^m)_{j\in\mathbb{N}_0}$ and $(\omega_j^m)_{j\in\mathbb{N}_0}$ be two increasing sequences, interpreted as costs and values. We associate to each $\mathbf{j} = (j_m)_{m\in\mathcal{M}} \in \mathbb{N}_0^{\mathcal{M}}$ a cost

$$c_{\mathbf{j}} = \sum_{m\in\mathcal{M}} c_{j_m}^m, \tag{A.1}$$

and a value

$$\omega_{\mathbf{j}} = \sum_{m\in\mathcal{M}} \omega_{j_m}^m. \tag{A.2}$$

We are interested in maximizing $\omega_{\mathbf{j}}$ under a constraint on $c_{\mathbf{j}}$, or equivalently minimizing $c_{\mathbf{j}}$ under a constraint on $\omega_{\mathbf{j}}$.

A.2 A sequence of optimal solutions

For each $m \in \mathcal{M}$ and all $j \in \mathbb{N}_0$, let

$$\Delta c_j^m : c_{j+1}^m - c_j^m \quad \text{and} \quad \Delta\omega_j^m := \omega_{j+1}^m - \omega_j^m. \tag{A.3}$$

We furthermore define quantities q_j^m as the quotient of these two increments;

$$q_j^m := \frac{\Delta\omega_j^m}{\Delta c_j^m}, \quad m \in \mathcal{M},\ j \in \mathbb{N}_0. \tag{A.4}$$

These values are interpreted as the value to cost ratio of passing to $j+1$ from j in the index $m \in \mathcal{M}$. We shall iteratively build a sequence $(\mathbf{j}^k)_{k\in\mathbb{N}_0}$ in $\mathbb{N}^{\mathcal{M}}$, such that each $\mathbf{j}^k$ is optimal under some assumptions.

Assumption A.1 *For all $m \in \mathcal{M}$,*

$$c_0^m = 0 \quad \text{and} \quad \Delta c_j^m > 0 \quad j \in \mathbb{N}_0, \tag{A.5}$$

i.e $(c_j^m)_{j \in \mathbb{N}_0}$ is strictly increasing. Also, $(\omega_0^m)_{m \in \mathcal{M}} \in \ell^1(\mathcal{M})$ and $(\omega_j^m)_{j \in \mathbb{N}_0}$ is nondecreasing for all $m \in \mathcal{M}$, i.e $\Delta\omega_j^m \geqslant 0$ for all $j \in \mathbb{N}_0$. Furthermore, for each $m \in \mathcal{M}$, the sequence $(q_j^m)_{j \in \mathbb{N}_0}$ is nonincreasing, i.e if $i \geqslant j$, then $q_i^m \leqslant q_j^m$. Finally, for any $\epsilon > 0$, there are only finitely many $m \in \mathcal{M}$ for which $q_0^m \geqslant \epsilon$.

The assumption that $(q_j^m)_{j \in \mathbb{N}_0}$ is nonincreasing is equivalent to

$$\frac{\omega_i^m}{\omega_j^m} \leqslant \frac{\Delta c_i^m}{\Delta c_j^m} \quad \text{if } i \geqslant j \tag{A.6}$$

if $\Delta\omega_j^m > 0$. In this sense, $(\omega_j^m)_{j \in \mathbb{N}_0}$ increases more slowly than $(c_j^m)_{j \in \mathbb{N}_0}$. Also, this assumption implies that if $\Delta\omega_j^m = 0$, then $\omega_i^m = \omega_j^m$ for all $i \geqslant j$. We define a total order on $\mathcal{M} \times \mathbb{N}_0$ by

$$(m, j) \succ (n, i) \text{ if } \begin{cases} q_j^m > q_i^m & \text{or} \\ q_j^m = q_i^n \quad \text{and} \quad m < n & \text{or} \\ q_j^m = q_i^n \quad \text{and} \quad m = n \quad \text{and} \quad j < i. \end{cases} \tag{A.7}$$

To any sequence $\mathbf{j} = (j_m)_{m \in \mathcal{M}}$ in $\mathbb{N}_0$ we associate the set

$$\{\{\mathbf{j}\}\} := \{(m, j) \in \mathcal{M} \times \mathbb{N}_0; j < j_m\}\,. \tag{A.8}$$

We now construct the sequence $(\mathbf{j}^k)_{k \in \mathbb{N}_0}$ in $\mathbb{N}^{\mathcal{M}}$. Let $\mathbf{j}^0 := \mathbf{0} \in \mathbb{N}^{\mathcal{M}}$, and for all $k \in \mathbb{N}_0$ we construct $\mathbf{j}^{k+1}$ from $\mathbf{j}^k$ as follows. Let $m_k \in \mathbb{N}_0$ maximize $q_{j_m^k}^m$. Existence of such maximum is guaranteed by Assumption A.1. If the maximum is not unique, we select the smallest m_k among all maxima. We then define $j_{m_k}^{k+1} : j_{m_k}^k + 1$, and set $j_m^{k+1} := j_m^k$ if $m_k \neq m$. For this sequence we abbreviate $c_k := c_{\mathbf{j}^k}$ and $\omega_k := \omega_{\mathbf{j}^k}$.

Lemma A.2 *For all $k \in \mathbb{N}_0$, $\{\{k\}\} := \left\{\left\{\mathbf{j}^k\right\}\right\}$ consists of the fist k terms of $\mathcal{M} \times \mathbb{N}_0$ with respect to the order $\succ$.*

Proof The case $k = 0$ is trivial. By induction, if the assumption holds for some $k \in \mathbb{N}_0$,

$$\{\{k+1\}\} = \{\{k\}\} \cup \left\{(m_k, j_{m_k}^k)\right\}, \tag{A.9}$$

with $(m_k, j_{m_k}^k)$ the $\succ$-minimal element of the set $\left\{(m, j_m^k); m \in \mathcal{M}\right\}$. For each $m \in \mathcal{M}$, Assumption A.1 implies $q_i^m \leqslant q_{j_{m_k}^k}^m$ for all $i \geqslant j_{m_k}^k + 1$. Therefore, $(m_k, j_{m_k}^k) \succ (m, i)$ for all $i \geqslant j_m^k + 1$, and consequently $(m_k, j_{m_k}^k)$ is the $\succ$-minimal element of $(\mathcal{M} \times \mathbb{N}_0) \setminus \{\{k\}\}$. □

Theorem A.3 *For all $k \in \mathbb{N}_0$, the sequence $\mathbf{j}^k$ maximizes $\omega_{\mathbf{j}}$ among all finitely supported sequences $\mathbf{j} = (j_m)_{m \in \mathcal{M}}$ in $\mathbb{N}_0$ with $c_{\mathbf{j}} \leqslant c_k$. Furthermore, if $c_| < c_k$ and there exist k pairs $(m, i) \in \mathcal{M} \times \mathbb{N}_0$ with $\Delta\omega_i^m > 0$, then $\omega_{\mathbf{j}} < \omega_k$.*

Proof Let $k \in \mathbb{N}$ and let $\mathbf{j} = (j_m)_{m\in\mathcal{M}}$ be a finitely supported sequence of in $\mathbb{N}_0$ with $c_{\mathbf{j}} \leqslant c_k$. By definition,

$$\omega_{\mathbf{j}} = \sum_{m\in\mathcal{M}} \omega_0^m + \sum_{m\in\mathcal{M}} \sum_{i=0}^{j_m-1} q_i^m \Delta c_i^m = \omega_{\mathbf{j}^0} + \sum_{(m,i)\in\{\{\mathbf{j}\}\}} q_i^m \Delta c_i^m. \tag{A.10}$$

Therefore, the assertion reduces to

$$\sum_{(m,i)\in\{\{\mathbf{j}\}\}\setminus\{\{k\}\}} q_i^m \Delta c_i^m \leqslant \sum_{(m,i)\in\{\{k\}\}\setminus\{\{\mathbf{j}\}\}} q_i^m \Delta c_i^m. \tag{A.11}$$

Note that by (A.1) and (A.3)

$$\sum_{(m,i)\in\{\{\mathbf{j}\}\}\setminus\{\{k\}\}} \Delta c_i^m = c_{\mathbf{j}} - c' \quad \text{for} \quad c' := \sum_{((m,i)\in\{\{\mathbf{j}\}\}\cap\{\{k\}\}} \Delta c_i^m. \tag{A.12}$$

By Lemma A.2 and (A.7), $q := q^{m_{k-1}}_{\mathbf{j}^{k-1}_{m_{k-1}}}$ satisfies $q \leqslant q_i^m$ for all $(m,i) \in \{\{k\}\}$, and $q_i^m \leqslant q$ for all $(m,i) \in (\mathcal{M} \times \mathbb{N}_0)\setminus\{\{k\}\}$. In particular, $q > 0$ if there exist k pairs $(m,i) \in \mathcal{M} \times \mathbb{N}_0$ with $q_i^m > 0$ since $\#\{\{k\}\} = k$. Consequently,

$$\begin{aligned} \sum_{(m,i)\in\{\{\mathbf{j}\}\}\setminus\{\{k\}\}} q_i^m \Delta c_i^m \leqslant & q \sum_{(m,i)\in\{\{\mathbf{j}\}\}\setminus\{\{k\}\}} \Delta c_i^m & \quad (A.13)\\ \leqslant & q(c_k - c') \leqslant \sum_{(m,i)\in\{\{k\}\}\setminus\{\{\mathbf{j}\}\}} q_i^m \Delta c_i^m, & \quad (A.14) \end{aligned}$$

and this inequality is strict if $q_i > 0$ and $c_k > c_{\mathbf{j}}$. □

A.3 Numerical construction

We now present greedy algorithms to construct the sequence $(\mathbf{j}^k)_{k\in\mathbb{N}_0}$. We shall assume that for each $m \in \mathcal{M}$, the sequences $(c_j^m)_{j\in\mathbb{N}_0}$ and $(c_j^m)_{j\in\mathbb{N}_0}$ are stored as linked lists. We shall first assume that $\mathcal{M}$ is finite with $\#\mathcal{M} := M$. We define a list $\mathcal{N}$ as the set of triples (m, j_m^k, q^m, j_m^k), sorted in ascending order, with respect to $\succ$. We assume to have a data structure enabling the removal of the minimal element of the list, and the insertion of new elements.

Routine A.1 `NextOpt`$[\mathbf{j}, \mathcal{N}] \to [\mathbf{j}, m, \mathcal{N}]$

```
$m \leftarrow$ PopMin($\mathcal{N}$)
$j_m \leftarrow j_m + 1$
$q \leftarrow (\omega^m_{j_m+1} - \omega^m_{j_m})/(c^m_{j_m+1} - c^m_{j_m})$
$\mathcal{N} \leftarrow$ Insert[$\mathcal{N}, (m, j_m, q)$]
```

Proposition A.4 *Let $\mathcal{N}_0$ be initialized as $\{(m, 0, q_0^m); m \in \mathcal{M}\}$ and $\mathbf{j}^0 := \mathbf{0} \in \mathbb{N}_0^{\mathcal{M}}$. Then the recursive application of*

$$\texttt{NextOpt}[\mathbf{j}^k, \mathcal{N}_k] \to [\mathbf{j}^{k+1}, m_k, \mathcal{N}_{k+1}] \tag{A.15}$$

constructs the sequence $(\mathbf{j}^k)_{k\in\mathbb{N}_0}$ as defined above. Initialization of the data structure $\mathcal{N}_0$ requires $\mathcal{O}(M\log M)$ operations and $\mathcal{O}(M)$ memory. One step of (A.15) requires $\mathcal{O}(M)$ operations if $\mathcal{N}$ is realized as a linked list, and $\mathcal{O}(\log M)$ operations if $\mathcal{N}$ is realized as a tree. The total number of operations required by the first k steps is $\mathcal{O}(kM)$ in the former case and $\mathcal{O}(k\log M)$ in the latter. In both cases, the total memory requirement for the first k steps is $\mathcal{O}(M+k)$.

Proof Recursive application of `NextOpt` as in (A.15) constructs the sequence $(\mathbf{j}^k)_{k\in\mathbb{N}_0}$ by Lemma A.2 and the definition of $\succ$. In the k-th step, the element m_k is removed from $\mathcal{N}$ and reinserted in a new position. Therefore, the size of $\mathcal{N}$ remains constant at M. The computational cost of (A.15) is dominated by the insert oepration on $\mathcal{N}$, which has the complexity stated above. □

We now turn to the case when $\mathcal{M}$ is countably infinite. By enumerating the elements of $\mathcal{M}$, it suffices suffices to consider the case $\mathcal{M} = \mathbb{N}$. We assume in this case that the sequence $(q_0^m)_{m\in\mathcal{M}}$ is nonincreasing.

As above, we use a list $\mathcal{N}$ of triples $(m, j_m^k, q_{\mathbf{j}_m^k}^m)$ to construct the sequence $(\mathbf{j}^k)_{k\in\mathbb{N}_0}$. However, $\mathcal{N}$ should only store triples for which m is candidate for the next value of m_k, i.e all m with $\mathbf{j}_m^k \neq 0$ and the smallest m with $\mathbf{j}_m^k = 0$. As in the finite case, $\mathcal{N}$ cn be realized as a linked list or a tree. The data structure should provide a function for removing the smallest element with respect to $\succ$, and for inserting a new element.

Routine A.2 `NextOptInf`$[\mathbf{j}, \mathcal{N}, M] \to [\mathbf{j}, m, \mathcal{N}, M]$

```
$m \leftarrow$ PopMin($\mathcal{N}$)
$j_m \leftarrow j_m + 1$
$q \leftarrow (\omega^m_{j_m+1} - \omega^m_{j_m})/(c^m_{j_m+1} - c^m_{j_m})$
$\mathcal{N} \leftarrow$ Insert[$\mathcal{N}, (m, j_m, q)$]
if $m = M$ then
    $M \leftarrow M + 1$
    $q \leftarrow (\omega^M_1 - \omega^M_0)/c^M_1$
    $\mathcal{N} \leftarrow$ Insert[$\mathcal{N}, (M, 1, q)$]
```

Proposition A.5 *Let $\mathcal{N}_0$ be initialized as $\{(1, 0, q^1)\}$, $M_0 := 1$ and $\mathbf{j}^0 := \mathbf{0} \in \mathbb{N}_0^{\mathcal{M}}$. Then the recursion*

$$\texttt{NextOptInf}[\mathbf{j}^k, \mathcal{N}_k, M_k] \to [\mathbf{j}^{k+1}, m_k, \mathcal{N}_{k+1}, M_{k+1}] \tag{A.16}$$

constructs the sequence $(\mathbf{j}^k)_{k\in\mathbb{N}_0}$ as defined above. For all $k \in \mathbb{N}_0$, the ordered set $\mathcal{N}_k$ contains exactly M_k elements, and $M_k \leqslant k$. The k-th step of (A.16) requires $\mathcal{O}(k)$ operations if $\mathcal{N}$ is realized as a linked list, and $\mathcal{O}(k\log k)$ if $\mathcal{N}$ is realized as a tree. The total number of operations required by the first k steps is $\mathcal{O}(k^2)$ in the former case and $\mathcal{O}(k\log k)$ in the latter. In both cases, the total memory requirement for the first k steps is $\mathcal{O}(k)$.

Proof It follows from the definitions that recursive application of `NextOptInf` as in (A.16) constructs the sequence $(\mathbf{j}^k)_{k\in\mathbb{N}_0}$. In the k-th step, the element m_k is removed from $\mathcal{N}$ and reinserted in a new position. If $m_k = M$, an additional element is inserted, and M is incremented. Therefore, the number of elements in $\mathcal{N}$ is M, and $M \leqslant k$. The computational cost of (A.16) is dominated by the insert operation on $\mathcal{N}$, which has the complexity stated above. □

Remark A.6 *As mentioned above, $(c_j^m)_{j \in \mathbb{N}_0}$ and $(\omega_j^m)_{j \in \mathbb{N}_0}$ are assumed to be stored in a linked list for each$m \in \mathcal{M}$. By removing the first element from the $\mathcal{M}_k$-th list in the k-th step of (A.15) or (A.16),* `NextOpt` *and* `NextOptInf` *only ever access the first two elements of one of these lists, which takes $\mathcal{O}(1)$ time. The memory locations of the lists can be stored in a hash table for efficient access.*

Remark A.7 *An appropriate way to store $(\mathbf{j}^k)_{k \in \mathbb{N}_0}$ is to collect $(m_k)_{k \in \mathbb{N}_0}$ in a linked list. Then $\mathbf{j}^k$ can be reconstructed by reading the first k elements of the list, which takes $\mathcal{O}(k)$ time independantly of the size of the list. Also, the total memory requirement is $\mathcal{O}(k)$ is the first k elements are stored.*