On Numerics for Deterministic and Stochastic Evolution Problems

MATTEO MOLTENI

CHALMERS
UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF GOTHENBURG

Division of Mathematics
Department of Mathematical Sciences
CHALMERS UNIVERSITY OF TECHNOLOGY
AND
UNIVERSITY OF GOTHENBURG
Gothenburg, Sweden 2016

On Numerics for Deterministic
and Stochastic Evolution Problems
Matteo Molteni
ISBN 978-91-7597-340-1

Doktorsavhandlingar vid Chalmers tekniska högskola
Ny serie nr 4021
ISSN 0346-718X

Department of Mathematical Sciences
Chalmers University of Technology and
University of Gothenburg
412 96 Gothenburg
Sweden
Phone: +46 (0)31-772 10 00

Printed in Gothenburg, Sweden 2016

On Numerics for Deterministic and Stochastic Evolution Problems

Matteo Molteni
Department of Mathematical Sciences
Chalmers University of Technology
and
University of Gothenburg

Abstract

This thesis is focused on the application and use of two different tools for the numerical solution of parabolic partial differential equations. The first part of the thesis investigates the possibility of using a weak space-time formulation in order to derive analytical results and to construct numerical schemes for linear parabolic equations, with the linear heat equation as reference problem. In the first paper the stochastic heat equation is considered and an alternative formulation is presented, which besides being consistent with known formulations of the same problem, naturally simplifies the construction of Petrov–Galerkin discretizations. The second paper is focused on certain features of an alternative discretization of the deterministic linear heat equation. It is shown that for the proposed discretization, a certain component of the solution, neglected in other works on the same topic, converges in space and time with a rate which is twice the rate of the "main" component of the solution. The third paper has a natural collocation in between the previous two, since it focuses on the quasi-optimality theory for parabolic problems with random coefficients. A spatial semidiscretization and a full discretization are considered, and results of quasi-optimality with explicit constants are derived, both path-wise and in an L^p-sense.

The second part of this thesis investigates the possible use of the Discrete Variational Derivative Method (DVDM) to construct numerical schemes that retain certain conservation properties. In the fourth paper this method is applied to construct energy-preserving numerical schemes to solve the geodesic Euler–Poincaré equation on the group of diffeomorphisms, also known as EPDiff. Three different schemes are presented, for which conservation properties, reversibility, convergence and computational cost are investigated both theoretically and empirically. The quality of the schemes is finally tested with a series of well-established benchmark problems.

Keywords: inf-sup theory, stochastic linear heat equation, Petrov–Galerkin discretization, random coefficients, DVDM, EPDiff

List of papers

A Stig Larsson and **Matteo Molteni**,
A weak space-time formulation for the linear stochastic heat equation,
Int. J. Appl. Comput. Math., 2016, electronic,
doi: 10.1007/s40819-016-0134-2.

B Stig Larsson and **Matteo Molteni**,
Numerical solution of parabolic problems based on a weak space-time formulation,
Preprint, arXiv:1603.03210.

C Stig Larsson, Christian Mollet, and **Matteo Molteni**,
Quasi-optimality of Petrov-Galerkin discretizations of parabolic problems with random coefficients,
Preprint, arXiv:1604.06611.

D Stig Larsson, Takayasu Matsuo, Klas Modin, and **Matteo Molteni**,
Discrete Variational Derivative Methods for the EPDiff equation,
Preprint, arXiv:1604.06224.

Paper not included in this thesis

E Marco Donatelli, **Matteo Molteni**, Vincenzo Pennati, and Stefano Serra-Capizzano,
Multigrid methods for cubic spline solution of two point (and 2D) boundary value problems,
Appl. Numer. Math., 104, 2016, pp. 15–29,
doi: 10.1016/j.apnum.2014.04.004.

In all the papers I have made major contributions to the writing and to the development of ideas, proofs, and numerical simulations.

Acknowledgements

I would like to express my sincere gratitude to my supervisor *Stig Larsson* for the continuous support of my Ph.D. studies, for his patience, motivation, and immense knowledge. His guidance has been of invaluable help during the whole process of writing this thesis.

I have deep gratitude for the many people who have helped me build this dissertation over the last years. In the most direct sense, a huge thanks to *Klas Modin*, for making me part of his research project and for his great help while working together, to *Takayasu Matsuo*, my amazing host and mentor in Tokyo, and to *Christian Mollet*, with whom I had the pleasure of working while he was visiting Chalmers.

I would also like to express my gratitude to *Daisuke Furihata* and *Yuto Miyatake* for the helpful discussions while working on the fourth paper of this thesis, and to *Marco Donatelli*, *Vincenzo Pennati* and *Stefano Serra-Capizzano*, for the nice work done together and for the support before and during my doctoral studies.

Further, I would like to thank *Adam Andersson* and *Fredrik Lindgren*, for the interesting seminars, lectures and discussions we have had during our time at Chalmers, and in general all the current and previous members in the CAM group.

On a more personal note, I would like to thank my fellow Ph.D. students for providing a sense of community and camaraderie; so many have over the years become not only good colleagues, but also great friends.

A very special thanks to my girlfriend, for always backing me up, standing by my side, and for reminding me that there is a whole world outside of my Ph.D.

Last, but not least, I want to thank my wonderful family for the neverending love and constant support. Without them nothing of this would have been possible.

Matteo Molteni
Gothenburg, April 2016

Contents

1

Weak space-time formulation of parabolic evolution problems

1.1 Introduction

The first part of this thesis deals with the weak space-time formulation of parabolic problems, with particular focus on the advantages that this approach offers both in terms of low-regularity of the solution and relatively easier error analysis based on the quasi-optimality theory. The main tool upon which we rely is the inf-sup theory and the Banach–Nečas–Babuška theorem. From a historical point of view, this approach was first used in connection to mixed formulations of elliptic problems rather than for parabolic problems. In 1972 Babuška and Aziz introduced an approximation theory for saddle point formulations of linear partial differential equations (see [4]), which constituted a powerful extension of the approximation theory based on Lax-Milgram and Céa's lemma, for positive-definite operators. The novelty of their approach was the possibility of using different spaces for trial and test functions. Natural applications for this were in the first place those problems which admitted a saddle-point or a mixed formulation. However, in 1989, Babuška and Janik (see [5] and [6]) proposed an application of the inf-sup theory in connection with numerics for parabolic problems, suggesting the possibility of a space-time formulation in which different test and trial space had to be used. In particular, up to that time, the finite element method was typically used in space only, reducing the problem to a system of ordinary differential equations to be then solved exactly or by means of the finite difference method. In [5] the finite element method was instead used both in space and in time for the first time, although a single element in space was considered (the authors referred to it as the *p-version*). In [6] we can instead find a real space-time finite element method, with a complete discussion of what the authors call *the h-p-version in time*. In particular, the error is measured with respect to the norm:

$$\int_0^T \|u(\cdot,t)\|^2_{H^1(\Lambda)}\,\mathrm{d}t, \tag{1.1}$$

which we will better specify later in this work, and which became the standard measure of error in the works dealing with space-time formulations of parabolic problems.

Since then, several authors have further investigated the use of space-time formulations of parabolic problems and the possible applications in numerics. We briefly summarize some of the most relevant features which inspired us to use this method:

- It naturally leads to the development of a quasi-optimal error analysis.
- It offers the possibility of constructing solutions under relatively weak assumptions of regularity.
- It allows us to keep track of every constant appearing in the bounds for the norm of the solution and in the error estimates.

1.2 The inf-sup theory

The main tool that we use in order to prove existence and uniqueness of the solution to our equation, once stated in its variational form, is the following theorem, here stated in its abstract form (see [4, 15]).

Theorem 1 (Banach–Nečas–Babuška (BNB)). *Let V and W be Banach spaces, V reflexive. Given a bounded bilinear form $\mathscr{B}\colon W \times V \to \mathbb{R}$,*

$$C_B := \sup_{0\neq w\in W} \sup_{0\neq v\in V} \frac{\mathscr{B}(w,v)}{\|w\|_W \|v\|_V} < \infty, \tag{BDD}$$

the associated linear operator $B\colon W \to V^*$, defined as

$${}_{V^*}\langle Bw, v\rangle_V := \mathscr{B}(w,v), \quad \forall w \in W, \forall v \in V, \tag{1.2}$$

is boundedly invertible if and only if the following two conditions are satisfied:

$$c_B := \inf_{0\neq w\in W} \sup_{0\neq v\in V} \frac{\mathscr{B}(w,v)}{\|w\|_W \|v\|_V} > 0, \tag{BNB1}$$

$$\forall v \in V, \quad \sup_{0\neq w\in W} \mathscr{B}(w,v) > 0. \tag{BNB2}$$

The constant c_B is called the *inf-sup constant*, while the constant C_B is called the *boundedness constant*. Whenever (BNB1)–(BNB2) hold and W is reflexive, we have the further identity

$$c_B^{-1} = \|B^{-1}\|_{\mathcal{L}(V^*,W)} = \|(B^*)^{-1}\|_{\mathcal{L}(W^*,V)}, \tag{1.3}$$

which leads to the following condition, equivalent to (BNB1)–(BNB2):

$$\inf_{0\neq w\in W} \sup_{0\neq v\in V} \frac{\mathscr{B}(w,v)}{\|w\|_W \|v\|_V} = \inf_{0\neq v\in V} \sup_{0\neq w\in W} \frac{\mathscr{B}(w,v)}{\|v\|_V \|w\|_W} > 0. \tag{1.4}$$

This allows to swap the spaces where the infimum and the supremum are taken, leading to the next corollary.

Corollary 2. *The variational problem*

$$w \in W : \mathscr{B}(w, v) = F(v), \quad \forall v \in V, \quad F \in V^*, \tag{1.5}$$

i.e., $Bw = F \in V^*$*, and its adjoint*

$$v \in V : \mathscr{B}(w, v) = G(w), \quad \forall w \in W, \quad G \in W^*, \tag{1.6}$$

i.e., $B^*v = G \in W^*$*, are well-posed whenever* (BDD)*,* (BNB1) *and* (BNB2) *hold. In particular, the well-posedness of the former is equivalent to the well-posedness of the latter and the norms of the solutions are bounded respectively as follows:*

$$\begin{aligned} \|w\|_W &\leq \|B^{-1}F\|_W \leq \frac{1}{c_B}\|F\|_{V^*}, \\ \|v\|_V &\leq \|(B^*)^{-1}G\|_V \leq \frac{1}{c_B}\|G\|_{W^*}. \end{aligned} \tag{1.7}$$

The two conditions expressed in (BNB1) and (BNB2) can be interpreted in terms of surjectivity and injectivity of the operator B (see [15, Remark 2.7]):

- (BNB1) ensures the surjectivity of $B^* \in \mathcal{L}(V, W^*)$, since it shows that $\mathrm{Ker}(B) = \{0\}$ and that $\mathrm{Im}(B)$ is closed.
- (BNB2) ensures that the operator $B^* \in \mathcal{L}(V, W^*)$ is injective, since it rules out the possibility of having $0 \neq v \in \mathrm{Ker}(B^*)$.

The next example helps to clarify that one condition alone does not ensure both existence and uniqueness:

Example 3 (Solvability of $\nabla \cdot w = F$ in $\mathbb{R}^3$)**.** *The bilinear form is in this case given by* $\mathscr{B}(w, v) = (w, \nabla v)$*, and we have that* $(w, \nabla v) = 0$ *does not necessarily imply that* $v = 0$ *because* w *can be a* curl *of some vector field. So, given a solution to* $\nabla \cdot w = F$ *we could add any* curl *to the solution and still get a solution. However the condition* (BNB2) *rules this out.*

It is quite important to make a comparison between this result and the well-known *Lax-Milgram theorem*.

Theorem 4 (Lax-Milgram (LM))**.** *Given a Hilbert space* V*, a bounded bilinear form* $a(\cdot, \cdot)$*, coercive on* V*, and a functional* $F \in V^*$*, there exists a unique solution to the problem*

$$a(u, v) = F(v), \quad \forall v \in V. \tag{1.8}$$

Although they both ensure the invertibility of an operator, the (BNB) theorem differs from the Lax-Milgram theorem in several ways:

BNB	**LM**
Equivalence	Implication
Banach spaces	Hilbert spaces
Test space $\neq$ Trial space	Test space $=$ Trial space

If we restrict ourselves to a finite-dimensional case, these differences become even more evident, since the two theorems respectively claim that:

LM: A positive-definite matrix is invertible.

BNB: A matrix is invertible if and only if it is indefinite.

An important difference is that the validity of the (BNB) theorem on a pair of spaces (W, V) does not imply its validity on two arbitrary subspaces $W_h \subset W$, $V_h \subset V$, the contrary to what happens with the (LM) theorem. One has thus to check again the validity of the three conditions expressed in Theorem 1.

Theorem 5. *Let $V_h \subset V$ and $W_h \subset W$ be subspaces of W and V introduced in Theorem 1. Given a bilinear form $\mathscr{B}\colon W_h \times V_h \to \mathbb{R}$,*

$$C_B^h := \sup_{0\neq w\in W_h} \sup_{0\neq v\in V_h} \frac{\mathscr{B}(w,v)}{\|w\|_W \|v\|_V} < \infty, \tag{BDDh}$$

the associated linear operator $B\colon W_h \to V_h^*$, defined as

$${}_{V^*}\langle Bw, v\rangle_V := \mathscr{B}(w,v), \quad \forall w \in W_h, \forall v \in V_h, \tag{1.9}$$

is boundedly invertible if and only if the following two conditions are satisfied:

$$c_B^h := \inf_{0\neq w\in W_h} \sup_{0\neq v\in V_h} \frac{\mathscr{B}(w,v)}{\|w\|_W \|v\|_V} > 0, \tag{BNB1h}$$

$$\forall v \in V_h, \quad \sup_{0\neq w\in W_h} \mathscr{B}(w,v) > 0. \tag{BNB2h}$$

The constant c_B^h is now called the *discrete inf-sup constant*, while the constant C_B^h is called the *discrete boundedness constant*. For the boundedness constant it is clear that $C_B^h \leq C_B$, so that boundedness in the continuous case implies boundedness in the discrete case. But this is not true for the inf-sup condition. In particular, c_B^h might depend on the chosen discretization and not

be uniform with respect to that. We can also notice that while checking the validity of conditions (BNB1h) and (BNB2h), if the spaces V_h and W_h are finite dimensional, a necessary condition is that the dimensions of the spaces are consistent, that is $\dim(V_h) = \dim(W_h)$. In this particular case, the two conditions (BNB1h) and (BNB2h) are equivalent, and we only need to check one of them.

1.3 Miscellaneous mathematical tools

Throughout the remaining of this chapter, and in the appended papers, we denote by $(H, \langle\cdot,\cdot\rangle_H)$ a separable Hilbert space.

1.3.1 Gelfand triple

Given a linear subspace $V \subset H$, densely embedded in H via the embedding map $J\colon V \hookrightarrow H$, there exists a canonical embedding $J^*\colon H^* \hookrightarrow V^*$, given by ${}_{V^*}\langle J^*\phi, v\rangle_V = \langle \phi, v\rangle_H$, $\forall v \in V, \phi \in H^*$. Provided that V is reflexive, the second embedding is also dense. By the identification of H with its dual, $H \equiv H^*$, one can thus obtain the *Gelfand triple:*

$$V \overset{J}{\hookrightarrow} H \equiv H^* \overset{J^*}{\hookrightarrow} V^*. \tag{1.10}$$

When possible, scalar product and dual pairing coincide:

$$\langle u, v\rangle_H = {}_{V^*}\langle u, v\rangle_V, \quad \forall u \in H, v \in V. \tag{1.11}$$

In general V is only required to be a Banach space but it might happen to be a Hilbert space itself, endowed with its own inner product; if this is the case, we cannot simultaneously identify $H \equiv H^*$ and $V \equiv V^*$, although an isometry from V onto V^* exists. This isometry is indeed no longer the identity, but rather another operator (for an instructive example see [7, Chapt. 5]). A typical example of Gelfand triple to which we often refer is given by

$$H_0^1(\Lambda) \hookrightarrow L^2(\Lambda) \hookrightarrow H^{-1}(\Lambda), \tag{1.12}$$

for some bounded domain $\Lambda \subset \mathbb{R}^d$.

1.3.2 About the operator A

The operator appearing in the problem of interest will often be a linear bounded coercive self-adjoint operator $A : V \to V^*$, associated to a bilinear form a given by $a(u, v) = {}_{V^*}\langle A\, u, v\rangle_V$. Generalizations of this operator, such as A non self-adjoint, time-dependent or dependent on a stochastic parameter ω will be also considered, but analysed case by case in the papers, and not presented here. We typically assume that the following conditions hold for some positive numbers $A_{\min}, A_{\max}$:

$$\begin{aligned} {}_{V^*}\langle Au, v\rangle_V &\le A_{\max}\|u\|_V\|v\|_V, \qquad && u, v \in V, \\ {}_{V^*}\langle Av, v\rangle_V &\ge A_{\min}\|v\|_V^2, && v \in V. \end{aligned} \tag{1.13}$$

The operator A is thus a bijection from V to V^*, and has a bounded inverse A^{-1}, which satisfies similar bounds:

$$\begin{aligned} {}_{V}\langle A^{-1}u, v\rangle_{V^*} &\le A_{\min}^{-1}\|u\|_{V^*}\|v\|_{V^*}, \qquad && u, v \in V^*, \\ {}_{V}\langle A^{-1}v, v\rangle_{V^*} &\ge A_{\max}^{-1}\|v\|_{V^*}^2, && v \in V^*. \end{aligned} \tag{1.14}$$

A typical example of an operator A satisfying the assumptions in (1.13) and (1.14) is given by

$$A := -\operatorname{Div}(\mathbb{A}\nabla\cdot), \tag{1.15}$$

defined on the spaces in (1.12), and with a matrix-valued function $\mathbb{A}$ such that:

$$A_{\min}|\zeta|^2 \le \zeta \cdot \mathbb{A}(\xi)\zeta \le A_{\max}|\zeta|^2, \qquad \xi \in \Lambda,\ \zeta \in \mathbb{R}^d. \tag{1.16}$$

1.3.3 Fractional powers

In order to measure the spatial regularity of the functions used in the appended papers, we make use of fractional powers of the operator A. To do so, we need to change the framework, moving from the Gelfand triple setting to something else. If the operator A is possibly unbounded, self-adjoint, defined on a domain $\mathscr{D}(A) \subset H$, with values in H, then it admits eigenpairs $(\lambda_n, e_n)_{n\in\mathbb{N}}$:

$$Ae_n = \lambda_n e_n, \tag{1.17}$$

where the $\{e_n\}_{n\in\mathbb{N}}$ constitute an orthonormal basis, and $\{\lambda_n\}_{n\in\mathbb{N}}$ is a sequence of positive numbers that tends to ∞. The analytic semigroup generated by $-A$

is denoted by

$$S_t := \mathrm{e}^{-tA}, \tag{1.18}$$

and defined as the strong operator limit

$$S_t := \sum_{n\in\mathbb{N}} \mathrm{e}^{-\lambda_n t}(e_n \otimes e_n). \tag{1.19}$$

The family $(S_t)_{t\geq 0}$ thus defined, fulfils the following properties:

$$\begin{aligned} &S_t \circ S_s = S_{t+s}, \qquad s,t \geq 0,\\ &S_0 := I_H,\\ &t \mapsto S_t \text{ is strongly continuous.} \end{aligned} \tag{1.20}$$

For an operator A as above, it is possible to define the square root, denoted by $A^{\frac{1}{2}}$, and in general any fractional powers. Given $\beta \in \mathbb{R}$, we define

$$A^{\frac{\beta}{2}} := \sum_{n\in\mathbb{N}} \lambda_n^{\frac{\beta}{2}}(e_n \otimes e_n), \tag{1.21}$$

with domain

$$\begin{aligned} &\mathscr{D}(A^{\frac{\beta}{2}}) = \Big\{v \in H : \sum_{n\in\mathbb{N}} \lambda_n^{\beta} |\langle v, e_n\rangle_H^2| < \infty\Big\}, \quad \beta > 0,\\ &\mathscr{D}(A^{\frac{\beta}{2}}) = H, \qquad\qquad\qquad\qquad\qquad\quad \beta \leq 0. \end{aligned} \tag{1.22}$$

When $\beta > 0$, we denote by $\dot{H}^\beta$ the set of all elements of H for which the first line in (1.22) holds. $\dot{H}^\beta$ is a Hilbert space, with inner product and norm defined by:

$$\begin{aligned} \langle u, v\rangle_{\dot{H}^\beta} &:= \sum_{n\in\mathbb{N}} \lambda_n^\beta \langle u, e_n\rangle_H \langle v, e_n\rangle_H,\\ \|u\|^2_{\dot{H}^\beta} &:= \sum_{n\in\mathbb{N}} \lambda_n^\beta |\langle u, e_n\rangle_H|^2. \end{aligned} \tag{1.23}$$

When $\beta \leq 0$, the spaces $\dot{H}^\beta$ are defined as the completion of H with respect to the norm defined above. Good references for this topic are given by [20] and [23]. Although there might seem to be a gap between the Gelfand triple formulation and the fractional powers defined above, it is actually possible to establish a precise connection and equivalence between the two. We report the explanation found in [11, Appendix 1] about how it is possible to switch from

the Gelfand triple framework to the fractional powers framework. For the other way round we refer instead to [24, Appendix F, Remark F.0.6], and refrain from presenting the details here.

We start by taking a Gelfand triple

$$V \overset{J}{\hookrightarrow} H \overset{\Phi}{\cong} H^* \overset{J^*}{\hookrightarrow} V^*, \tag{1.24}$$

where J and J^* are dense embeddings and Φ is the Riesz isomorphism. We want to modify the operator A introduced above, so that it becomes an unbounded operator $\tilde{A}$ from H into H. We define

$$\mathscr{D}(A) \subset H = \{v \in V : Av \in J^*\Phi(H)\}, \tag{1.25}$$

and a new operator $\tilde{A}$ as

$$\begin{aligned} \tilde{A} &: \mathscr{D}(\tilde{A}) \subset H \to H, \\ \mathscr{D}(\tilde{A}) &:= J(\mathscr{D}(A)), \\ \tilde{A} &:= \Phi^{-1}(J^*)^{-1}AJ^{-1}. \end{aligned} \tag{1.26}$$

In this way $\tilde{A}$ is an unbounded densely defined linear operator; in particular it is positive definite because of the coercivity of the bilinear form. If the bilinear form $a(\cdot,\cdot)$ associated to A is symmetric and J is a compact embedding, then $\tilde{A}$ is self-adjoint, boundedly invertible, with compact inverse $\tilde{A}^{-1} := JA^{-1}J^*\Phi$, and this implies that we can use the spectral theorem in order to define the semigroup and fractional powers of $\tilde{A}$. Alternatively, we can argue that such an operator is the generator of a strongly continuous semigroup of contractions and such a semigroup is holomorphic, as shown in [22, Theorem 1.52], and it is hence possible to define fractional powers of $\tilde{A}$. In order to simplify the notation, we finally omit the embeddings and denote $\tilde{A}$ by A.

1.3.4 Bochner–Lebesgue spaces

In order to present the abstract formulation of the heat equation, we will need the following *Bochner–Lebesgue spaces*:

$$\mathcal{Y} = L^2((0,T);V), \tag{1.27}$$

$$\mathcal{X} = L^2((0,T);V) \cap H^1((0,T);V^*), \tag{1.28}$$

which are Hilbert spaces respectively normed by

$$\|y\|_{\mathcal{Y}}^2 = \|y\|_{L^2((0,T);V)}^2 = \int_0^T \|y(t)\|_V^2 \,\mathrm{d}t, \tag{1.29}$$

and

$$\begin{aligned} \|x\|_{\mathcal{X}}^2 &= \|x(0)\|_H^2 + \|x\|_{L^2((0,T);V)}^2 + \|\dot{x}\|_{L^2((0,T);V^*)}^2 \\ &= \|x(0)\|_H^2 + \int_0^T \left(\|x(t)\|_V^2 + \|\dot{x}(t)\|_{V^*}^2 \right) \mathrm{d}t. \end{aligned} \tag{1.30}$$

In some books, as for example in [13], the space $\mathcal{X}$ is sometimes also denoted by $H^1((0,T);V,V^*)$ or $W^1((0,T);V,V^*)$.

The *trace theorem* for Bochner–Lebesgue spaces ([13, Theorem 1, Chapter XVIII.1]), says that $\mathcal{X}$ is densely embedded in $\mathscr{C}([0,T];H)$, so that $x(0)$ and $x(T)$ are defined in H. Whenever $x, y \in \mathcal{X}$ integration by parts is possible:

$$\begin{aligned} \int_0^T \left({}_{V^*}\langle \dot{x}(t), y(t) \rangle_V + {}_V\langle x(t), \dot{y}(t) \rangle_{V^*} \right) \mathrm{d}t \\ = \langle x(T), y(T) \rangle_H - \langle x(0), y(0) \rangle_H. \end{aligned} \tag{1.31}$$

The embedding constant M_{e}, defined as

$$M_{\mathrm{e}} := \sup_{0 \neq x \in \mathcal{X}} \frac{\|x(t)\|_{\mathscr{C}([0,T];H)}}{\|x\|_{\mathcal{X}}} < \infty, \tag{1.32}$$

is uniform in the choice of V. With our choice of norm on $\mathcal{X}$ we compute that $M_{\mathrm{e}} = 1$ because according to (1.31) we have that for any $r \in [0,T]$:

$$\begin{aligned} \|x(r)\|_H^2 &= \|x(0)\|_H^2 + \int_0^r {}_V\langle \dot{x}(t), x(t) \rangle_{V^*} \,\mathrm{d}t \\ &\leq \|x(0)\|_H^2 + \int_0^T |{}_{V^*}\langle \dot{x}(t), x(t) \rangle_V| \,\mathrm{d}t \\ &\leq \|x(0)\|_H^2 + \|x\|_{L^2((0,T);V)}^2 + \|\dot{x}\|_{L^2((0,T);V^*)}^2. \end{aligned} \tag{1.33}$$

Finally, we introduce the product space $L^2((0,T);V) \times H$, endowed with the product norm, for which we use the shorthand notation $\mathcal{Y}_H$, and the space $\mathcal{X}_{0,\{T\}}$, defined as the subspace of $\mathcal{X}$ of all the x's such that $x(T) = 0$.

1.3.5 Nuclear and Hilbert–Schmidt operators

We will need more tools from functional analysis in order to properly introduce stochastic evolution problems. Given a pair of separable Hilbert spaces

$(H, \langle\cdot,\cdot\rangle_H)$ and $(U, \langle\cdot,\cdot\rangle_U)$, we denote by $\mathcal{L}(U,H)$ the Banach space of bounded linear operators from U to H, with the simplified notation $\mathcal{L}(H)$ whenever $U = H$. For an operator $Q \in \mathcal{L}(H)$, we say that it is self-adjoint positive semi-definite (resp. positive definite) if $Q^* = Q$ and if $Q \geq 0$, i.e., $\langle Qv, v\rangle_H \geq 0$, $\forall v \in H$ (resp. if $Q > 0$, i.e., $\langle Qv, v\rangle_H > 0$, $\forall v \in H$, $v \neq 0$).

We denote by $\mathcal{L}_1(U,H)$ the space of *nuclear operators* from U to H, defined as the space of elements in $\mathcal{L}(U,H)$ for which there exists two sequences $\{a_j\}_{j\in\mathbb{N}} \subset H$, $\{b_j\}_{j\in\mathbb{N}} \subset U$ such that $\sum_{j\in\mathbb{N}} \|a_j\|_H \|b_j\|_U < \infty$ and such that $Tf = \sum_{j\in\mathbb{N}} \langle f, b_j\rangle_U a_j$, for every $f \in U$. Nuclear operators are sometimes called *trace-class* operators and form a Banach space normed by

$$\|T\|_{\mathcal{L}_1(U,H)} := \inf\Big\{ \sum_{j\in\mathbb{N}} \|a_j\|_H \|b_j\|_U < \infty : \\ Tf = \sum_{j\in\mathbb{N}} \langle f, b_j\rangle_U a_j,\ \forall f \in U \Big\}. \tag{1.34}$$

For operators $Q \in \mathcal{L}_1(H) = \mathcal{L}_1(H,H)$ the *trace* is well defined as

$$\mathrm{Tr}(Q) := \sum_{k\in\mathbb{N}} \langle Qe_k, e_k\rangle_H, \tag{1.35}$$

where $\{e_k\}_{k\in\mathbb{N}}$ is any orthonormal basis for H.

We say that an operator $Q \in \mathcal{L}(U,H)$ is a *Hilbert–Schmidt* operator if for an (hence for any) orthonormal basis $\{e_k\}_{k\in\mathbb{N}}$ of U, it holds that

$$\sum_{k\in\mathbb{N}} \|Qe_k\|_H^2 < \infty. \tag{1.36}$$

We denote by $\mathcal{L}_2(U,H)$ the space of Hilbert–Schmidt operators, endowed with the structure of Hilbert space induced by the inner product

$$\langle T, Q\rangle_{\mathcal{L}_2(U,H)} := \sum_{k\in\mathbb{N}} \langle Te_k, Qe_k\rangle_H. \tag{1.37}$$

It holds that $Q \in \mathcal{L}_2(U,H)$ if and only if $Q^* \in \mathcal{L}_2(H,U)$, with $\|Q\|_{\mathcal{L}_2(U,H)} = \|Q^*\|_{\mathcal{L}_2(H,U)}$ and that $Q \in \mathcal{L}_2(U,H)$ if and only if $QQ^* \in \mathcal{L}_1(H)$, with $\mathrm{Tr}(QQ^*) = \|Q\|^2_{\mathcal{L}_2(U,H)}$. We denote $\mathcal{L}_2(U,U)$ by $\mathcal{L}_2(U)$.

Finally, for $Q \in \mathcal{L}(U)$, with $Q \geq 0$, by denoting with $Q^{\frac{1}{2}} \in \mathcal{L}(U)$ its unique square positive root, we define the *Cameron-Martin* space $U_0 := Q^{\frac{1}{2}}(U)$, the Hilbert space endowed with the inner product

$$\langle u, v\rangle_{U_0} := \langle Q^{-\frac{1}{2}}u, Q^{-\frac{1}{2}}v\rangle_H, \tag{1.38}$$

where $Q^{-\frac{1}{2}}$ indicates the pseudo-inverse of $Q^{\frac{1}{2}}$, i.e.,

$$Q^{-\frac{1}{2}} := \left(Q^{\frac{1}{2}} \Big|_{\mathrm{Ker}(Q^{\frac{1}{2}})^{\perp}} \right)^{-1}. \tag{1.39}$$

It holds that $Q^{\frac{1}{2}}$ is an isometric isomorphism between $(\mathrm{Ker}(Q^{\frac{1}{2}})^{\perp}, \langle \cdot, \cdot \rangle_U)$ and $(U_0, \langle \cdot, \cdot \rangle_{U_0})$, making the latter a Hilbert space as well. The notation $\mathcal{L}_2^0$ will sometimes be used to denote the space of Hilbert–Schmidt operators $\mathcal{L}_2(U_0, H)$.

1.4 The abstract parabolic problem

In this section we introduce the prototype problem we investigate in the first part of the thesis. We assume that $V \hookrightarrow H \hookrightarrow V^*$ are as in § 1.3.1, and that a and A are as in § 1.3.2.

1.4.1 Variational space-time formulations

Although some generalizations of this problem will be considered in the appended papers (the stochastic version of this problem in Paper A and the version with random coefficients in Paper C), we present a main overview of the known results of solvability and of "how to handle" the left-hand side for the prototype problem defined in its strong form as

$$\begin{aligned} &\dot{u}(t) + Au(t) = f(t), \quad t \in (0, T], \\ &u(0) = u_0. \end{aligned} \tag{1.40}$$

The *first* space-time variational formulation of Problem (1.40) is:

$$u \in \mathcal{X} : \mathscr{B}(u, y) = \mathscr{F}(y), \quad \forall y \in \mathcal{Y}_H, \tag{1.41}$$

where $y = (y_1, y_2)$ and where the following bilinear form and linear functional have been used:

$$\begin{aligned} &\mathscr{B} \colon \mathcal{X} \times \mathcal{Y}_H \to \mathbb{R}, \\ &\mathscr{B}(x, y) := \int_0^T \Big({}_{V^*}\langle \dot{x}(t), y_1(t) \rangle_V + a(x(t), y_1(t)) \Big) \, \mathrm{d}t + \langle x(0), y_2 \rangle_H, \\ &\mathscr{F} \colon \mathcal{Y}_H \to \mathbb{R}, \\ &\mathscr{F}(y) := \int_0^T {}_{V^*}\langle f(t), y_1(t) \rangle_V \, \mathrm{d}t + \langle u_0, y_2 \rangle_H. \end{aligned} \tag{1.42}$$

By means of a formal integration by parts we can derive the *second* space-time variational formulation

$$u \in \mathcal{Y}_H : \mathscr{B}^*(u, x) = \mathscr{F}(x), \quad \forall x \in \mathcal{X}, \tag{1.43}$$

where the bilinear form $\mathscr{B}^*(\cdot,\cdot)$ and the load functional $\mathscr{F}$ are now given by

$$\begin{aligned}
&\mathscr{B}^* : \mathcal{Y}_H \times \mathcal{X} \to \mathbb{R}, \\
&\mathscr{B}^*(y, x) := \int_0^T \Big({}_V\langle y_1(t), -\dot{x}(t)\rangle_{V^*} + a(y_1(t), x(t)) \Big) \,\mathrm{d}t + \langle y_2, x(T)\rangle_H, \\
&\mathscr{F} : \mathcal{X} \to \mathbb{R}, \\
&\mathscr{F}(x) := \int_0^T {}_{V^*}\langle f(t), x(t)\rangle_V \,\mathrm{d}t + \langle u_0, x(0)\rangle_H.
\end{aligned} \tag{1.44}$$

These two formulations are also often referred to as *primal* and *dual*, and the second is also called *weak* space-time formulation.

Whenever the solution u of the second problem is regular enough, that is, when $u_1 \in \mathcal{X}$, the two formulations are equivalent; in particular the two components of the solution are strictly related according to $u_2 = u_1(T)$. This is however not true in general, since the second component of the solution, u_2, must be in general understood as a continuous H-valued version of u_1, evaluated at time $t = T$. This will be of crucial importance in the stochastic case presented in Paper A, where the solution will not have the extra regularity required.

Traditionally, the weak-space time formulation is stated in terms of different spaces, $\mathcal{Y}$ and $\mathcal{X}_{0,\{T\}}$, where the latter is defined as in § 1.3.4 as:

$$\mathcal{X}_{0,\{T\}} := \{x \in \mathcal{X} : x(T) = 0\}. \tag{1.45}$$

The difference between the weak space-time formulation with spaces $(\mathcal{Y}_H, \mathcal{X})$ and the one with spaces $(\mathcal{Y}, \mathcal{X}_{0,\{T\}})$ is broadly discussed in Paper A and in Paper B. We can notice how the two formulations are in some sense one the adjoint of the other, as in (1.5) and (1.6), so that the proof of invertibility of the operator associated to the bilinear form $\mathscr{B}(\cdot,\cdot)$ is essentially the same in both cases.

1.4.2 Solvability

In the framework described above, Theorem 1 reduces to the following concrete statement:

Theorem 6. *If the bilinear form $\mathscr{B}^*\colon \mathcal{Y}_H \times \mathcal{X} \to \mathbb{R}$ is bounded,*

$$C_B := \sup_{0\neq y\in\mathcal{Y}_H}\sup_{0\neq x\in\mathcal{X}} \frac{\mathscr{B}^*(y,x)}{\|y\|_{\mathcal{Y}_H}\|x\|_{\mathcal{X}}} < \infty, \tag{BDD}$$

then the associated operator B is boundedly invertible if and only if the following two conditions are satisfied:

$$c_B := \inf_{0\neq y\in\mathcal{Y}_H}\sup_{0\neq x\in\mathcal{X}} \frac{\mathscr{B}^*(y,x)}{\|y\|_{\mathcal{Y}_H}\|x\|_{\mathcal{X}}} > 0, \tag{BNB1}$$

$$\forall x \in\mathcal{X}, \quad \sup_{0\neq y\in\mathcal{Y}_H} \mathscr{B}^*(y,x) > 0. \tag{BNB2}$$

The proof of this theorem is based on the observation in (1.4), so that Theorem 6 is proved with swapped spaces. In the appended papers we present and analyse in great details the validity of the first two conditions and we therefore refrain from presenting their proofs. Conditions (BDD) and (BNB1) contain in particular quantitative information of relevance for bounding the norm of the solution and for introducing the quasi-optimality theory presented in the Section 1.5. The condition expressed in (BNB2) is instead only qualitative, and is never explicitly proved in any of the manuscripts which compose this thesis. For the sake of completeness we present its proof in this subsection, by following two different references.

Proof of (BNB2) *according to [25].* To prove (BNB2), we start by considering a basis $\{\phi_i\}_{i=1}^{\infty}$ for the space V (subspace of a separable Hilbert space) and we define the family of finite dimensional subspaces $V_n := \text{span}\{\phi_i\}_{i=1}^{n}$. The key idea of this proof is to construct for any $\tilde{y} \in \mathcal{Y}_H$ an element $z \in \mathcal{X}$ as the limit of a solution to a finite dimensional problem, such that

$$\mathscr{B}^*(y,z) = \int_0^T a(y,\tilde{y}_1) + \langle y, \tilde{y}_2\rangle_H,\ \forall y = (y_1,y_2) \in \mathcal{Y}_H. \tag{1.46}$$

Using then the coercivity of the bilinear form $a(\cdot,\cdot)$ will prove the claim.

Given any $\tilde{y} = (\tilde{y}_1, \tilde{y}_2) \in \mathcal{Y}_H$, consider the finite dimensional problem of seeking $z_n(t) = \sum_{i=1}^{(n)} \mathbf{z}_i^{(n)}(t)\phi_i$ such that

$$\begin{aligned}
&\langle \xi_n, -\dot{z}_n(t)\rangle_H + a(\xi_n, z_n(t)) = a(\xi_n, \tilde{y}_1(t)), \quad \forall \xi_n \in V_n,\ \forall t \in [0,T),\\
&z_n(T) = \sum_{i=1}^{n} \tilde{\mathbf{y}}_{2,i}\phi_i,
\end{aligned} \tag{1.47}$$

where $\sum_{i=1}^{n} \tilde{\mathbf{y}}_{2,i}\phi_i \to \tilde{y}_2$ in H for $n \to \infty$. Such a problem admits a unique solution $z_n \in \mathscr{C}([0,T];V_n)$, whose derivative $\dot{z}_n$ belongs to $L^2((0,T);V_n)$. The first claim is that the sequence $(z_n)_{n\in\mathbb{N}}$ is bounded in $\mathcal{Y}$. In fact, by integrating in time (1.47) and choosing $\xi_n = z_n$, one can get

$$\int_0^T \langle z_n, -\dot{z}_n\rangle_H \,\mathrm{d}t + \int_0^T a(z_n, z_n)\,\mathrm{d}t = \int_0^T a(z_n, \tilde{y}_1)\,\mathrm{d}t, \tag{1.48}$$

i.e.,

$$\|z_n(0)\|_H^2 + 2\int_0^T a(z_n, z_n)\,\mathrm{d}t = 2\int_0^T a(z_n, \tilde{y}_1)\,\mathrm{d}t + \|z_n(T)\|_H^2. \tag{1.49}$$

Thus, using the coercivity of $a(\cdot,\cdot)$, for any $\epsilon > 0$ one gets:

$$\begin{aligned} 2A_{\min}\int_0^T \|z_n\|_V^2 &\le 2\int_0^T a(z_n, z_n)\,\mathrm{d}t \\ &\le \|z_n(0)\|_H^2 + 2\int_0^T a(z_n, z_n)\,\mathrm{d}t \\ &= 2\int_0^T a(z_n, \tilde{y}_1)\,\mathrm{d}t + \|z_n(T)\|_H^2 \\ &\le 2A_{\max}\int_0^T \sqrt{\epsilon}\|z_n\|_V \frac{1}{\sqrt{\epsilon}}\|\tilde{y}_1\|_V\,\mathrm{d}t + 2\|\tilde{y}_2\|_H^2, \end{aligned} \tag{1.50}$$

where in the last step, without loss of generality, it has been assumed that $\|z_n(T)\|_H \le \sqrt{2}\|\tilde{y}_2\|_H$. Moreover, by means of the elementary inequality $2ab \le a^2 + b^2$, one can obtain

$$\begin{aligned} &2A_{\min}\int_0^T \|z_n\|_V^2 \\ &\qquad \le A_{\max}\Big(\int_0^T \epsilon\|z_n\|_V^2\,\mathrm{d}t + \int_0^T \frac{1}{\epsilon}\|\tilde{y}_1\|_V^2\,\mathrm{d}t\Big) + 2\|\tilde{y}_2\|_H^2. \end{aligned} \tag{1.51}$$

Dividing now by $2A_{\min}$ and choosing $\epsilon = \frac{A_{\min}}{A_{\max}}$ the inequality becomes

$$\int_0^T \|z_n\|_V^2 \le \frac{1}{2}\int_0^T \|z_n\|_V^2\,\mathrm{d}t + \frac{A_{\max}^2}{2A_{\min}^2}\int_0^T \|\tilde{y}_1\|_V^2\,\mathrm{d}t + \frac{1}{A_{\min}}\|\tilde{y}_2\|_H^2, \tag{1.52}$$

leading to

$$\int_0^T \|z_n\|_V^2\,\mathrm{d}t \le \frac{A_{\max}^2}{A_{\min}^2}\int_0^T \|\tilde{y}_1\|_V^2\,\mathrm{d}t + \frac{2}{A_{\min}}\|\tilde{y}_2\|_H^2, \tag{1.53}$$

i.e., to

$$\|z_n\|_{\mathcal{Y}}^2 \leq \frac{A_{\max}^2}{A_{\min}^2}\|\tilde{y}_1\|_{\mathcal{Y}}^2 + \frac{2}{A_{\min}}\|\tilde{y}_2\|_H^2, \quad \forall n \in \mathbb{N}, \tag{1.54}$$

that proves the claim. Moreover for any $1 \leq i \leq n$ and for any $\theta \in \mathscr{C}([0,T];\mathbb{R})$ such that $\theta(0) = 0$, if z_n is the solution to the finite dimensional problem (1.47), it follows that

$$\int_0^T \langle \phi_i, -\dot{z}_n(t)\rangle_H \theta(t)\,\mathrm{d}t = \int_0^T a(\phi_i, \tilde{y}_1(t) - z_n(t))\theta(t)\,\mathrm{d}t, \tag{1.55}$$

thus, using integration by parts,

$$\begin{aligned}\int_0^T &\langle \phi_i, z_n(t)\rangle_H \dot{\theta}(t)\,\mathrm{d}t \\ &= \langle \phi_i, z_n(T)\rangle_H \theta(T) + \int_0^T a(\phi_i, \tilde{y}_1(t) - z_n(t))\theta(t)\,\mathrm{d}t.\end{aligned} \tag{1.56}$$

Since $(z_n)_{n\in\mathbb{N}}$ is a bounded sequence in $\mathcal{Y}$, there exists a subsequence $(z_{n_k})_{k\in\mathbb{N}}$ weakly convergent to an element $z \in \mathcal{Y}$. Using such a subsequence and taking the limit in the equation above, it follows that for any $n \in \mathbb{N}$

$$\begin{aligned}\int_0^T &\langle \phi_i, z(t)\rangle_H \dot{\theta}(t)\,\mathrm{d}t \\ &= \langle \phi_i, z(T)\rangle_H \theta(T) + \int_0^T a(\phi_i, \tilde{y}_1(t) - z(t))\theta(t)\,\mathrm{d}t \\ &= \langle \phi_i, \tilde{y}_2\rangle_H \theta(T) + \int_0^T a(\phi_i, \tilde{y}_1(t) - z(t))\theta(t)\,\mathrm{d}t.\end{aligned} \tag{1.57}$$

In particular this last equation holds for any $\theta \in \mathscr{D}((0,T);\mathbb{R})$, where $\mathscr{D}$ denotes the classical space of test functions.

It then follows that by interpreting $\dot{z} \in \mathscr{D}^*((0,T);V) \hookrightarrow \mathscr{D}^*((0,T);V^*)$,

$$\langle \phi_i, -\dot{z}(\theta)\rangle_H = \langle \phi_i, \int_0^T A^*(\tilde{y}_1(t) - z(t))\theta(t)\,\mathrm{d}t\rangle_H, \tag{1.58}$$

which reads

$$-\dot{z} = A(\tilde{y}_1 - z) \text{ in } \mathscr{D}^*((0,T);V^*). \tag{1.59}$$

Since $\tilde{y}_1 - z \in \mathcal{Y}$ and $A\colon \mathcal{Y} \to L^2((0,T);V^*)$ is bounded, it follows that $\dot{z} \in L^2((0,T);V^*)$, i.e., $z \in \mathcal{X}$.

By replacing it in (1.57), after another integration by parts, one can thus obtain $z(T) = \tilde{y}_2$ and, by density of $\mathscr{D}((0,T);\mathbb{R}) \otimes V$ in $\mathcal{Y}_H$, it is possible to finally obtain that for any $y \in \mathcal{Y}_H$

$$\mathscr{B}^*(y,z) = \int_0^T a(y_1,\tilde{y}_1)\,\mathrm{d}t + \langle y_2, \tilde{y}_2\rangle_H. \tag{1.60}$$

If we then choose $y = \tilde{y}$, we have that for any $y \in \mathcal{Y}_H$

$$\begin{aligned}\sup_{x\in\mathcal{X}} \mathscr{B}^*(y,x) &\geq \mathscr{B}^*(y,z)\\ &= \int_0^T a(y_1,y_1)\,\mathrm{d}t + \langle y_2, y_2\rangle_H \geq \min\{1, A_{\min}\}\|y\|^2_{\mathcal{Y}_H},\end{aligned} \tag{1.61}$$

and (BNB2) is hence proved. □

Proof of (BNB2) *according to [27].* The second way to prove (BNB2) relies on proving that if there exists a $y \in \mathcal{Y}_H$ such that

$$\mathscr{B}^*(y,x) = 0, \quad \forall x \in \mathcal{X}, \tag{1.62}$$

then it must hold that $y = 0$.

To this end we observe that for all $x \in \mathscr{C}^\infty((0,T);V)$ the following inequality holds:

$$\begin{aligned}\int_0^T {}_V\langle y_1(t), -\dot{x}(t)\rangle_{V^*}\,\mathrm{d}t &= \int_0^T -{}_V\langle y_1(t), Ax(t)\rangle_{V^*}\,\mathrm{d}t\\ &\leq A_{\max}\|x\|_{L^2((0,T);V)}\|y\|_{L^2((0,T);V)}\end{aligned} \tag{1.63}$$

Since $\mathscr{C}^\infty((0,T);V)$ is dense in $L^2((0,T);V)$, we can conclude from (1.63) that y_1 has a weak derivative in $L^2((0,T);V)^* \simeq L^2((0,T);V^*)$, that is $y_1 \in \mathcal{X}$. If we integrate by parts in (1.62), we can see that y_1 is the solution to

$$\begin{aligned}&\langle x(T), y_2(T) - y_1(T)\rangle_H + \langle x(0), y_1(0)\rangle_H\\ &\qquad + \int_0^T \left({}_V\langle \dot{y}_1(t) + Ay_1(t), x(t)\rangle_{V^*}\right)\mathrm{d}t = 0.\end{aligned} \tag{1.64}$$

By using suitable test functions $x \in \mathcal{X}$, we can derive that:

$$\begin{aligned}\dot{y}_1 + Ay_1 &= 0,\\ y_1(T) - y_2(T) &= 0,\\ y_1(0) &= 0.\end{aligned} \tag{1.65}$$

By using these facts and by choosing $x = y_1$ in (1.62), we can finally conclude that $y_1 = 0$ and $y_2 = 0$. □

Provided that the three conditions (BDD), (BNB1) and (BNB2) are satisfied, the first and the second space-time formulation of (1.40) admit a unique solution whenever the load functional $\mathscr{F}$ defined in (1.42) and (1.44) belongs to the proper dual space. It is not difficult to see that this is the case whenever $u_0 \in H$ and $f \in L^2((0,T);V^*)$. Problem (1.41) admits therefore a unique solution, which satisfies the following bound:

$$\|u\|_{H^1((0,T);V,V^*)} \leq \frac{1}{c_B}\Big(\|u_0\|_H^2 + \|f\|^2_{L^2((0,T);V^*)}\Big)^{\frac{1}{2}}. \tag{1.66}$$

Similarly, Problem (1.43) admits also a unique solution, which satisfies the following bound:

$$\|u\|_{L^2((0,T);V)} \leq \frac{1}{c_B}\Big(\|u_0\|_H^2 + \|f\|^2_{L^2((0,T);V^*)}\Big)^{\frac{1}{2}}. \tag{1.67}$$

It is important to notice at this point that the results of invertibility of B^* and B hold independently from the choice of the right-hand side. This means that for whatever choice of functional $\mathscr{F}$, such that it belongs to the dual space of $\mathcal{X}$ or of $\mathcal{Y}_H$, the problem is uniquely solvable. This property is of particular relevance for the weak space-time formulation, because a broader class of functional than the one presented in (1.44) can actually be proven to belong to $\mathcal{X}^*$. This class includes, amongst others, stochastic integrals and nowhere differentiable functions.

1.5 Quasi-optimality

The main advantage of using the inf-sup theory described in Section 1.2 is the possibility of deriving results of quasi-optimality in a natural way. The importance of a quasi-optimality result is that it states the equivalence between the error of the method we investigate and the *best possible error* that would be committed by approximating the solution with a function living in the same subspace where the discrete solution lives. In particular, from the quasi-optimality result, we can deduce error bounds of optimal order, and this is an implication, rather than an equivalence.

If we denote by $w_h \in W_h$ the discrete solution to (1.5) on the couple of spaces (W_h, V_h), subspaces of (W, V), the quasi-optimality constant is defined

as the smallest constant q for which the following inequality holds:

$$\|w - w_h\|_W \leq q \inf_{v \in W_h} \|w - v\|_W. \tag{1.68}$$

The first result of quasi-optimality can be traced back to Céa, under the assumption of having the same Hilbert space W as both test and trial space, and by assuming that the bilinear form defining problem (1.5) is symmetric and coercive, with coercivity constant α_B. Under these assumptions, q could be bounded from above as

$$q \leq \sqrt{\frac{C_B}{\alpha_B}}. \tag{1.69}$$

The great contribution of Babuška, in [3], was to get rid of these restrictive assumptions, proving in the setting described in Section 1.2 that an upper bound for q is given by

$$q \leq 1 + \frac{C_B}{c_B^h}. \tag{1.70}$$

This estimate was finally sharpened by Xu and Zikatanov in [31], where by exploiting the properties of idempotent operators onto Hilbert spaces, the authors achieved a sharper upper bound for q, given by

$$q \leq \frac{C_B}{c_B^h}. \tag{1.71}$$

A major contribution in the investigation of the quasi-optimality theory for Petrov-Galerkin discretizations of evolution problems based on a space time formulation can be found in [27]. The author investigates a particular choice of discretization that leads to the backward Euler time stepping. The theory of quasi-optimality is first analysed in an abstract way, proving a counterpart for the estimate in (1.71) for the case of non-conforming discretizations, to then achieve concrete quasi-optimal error estimates for a spatial semidiscretization, for a temporal semidiscretization, and for a fully discrete scheme based on a temporal evolution that resembles the backward Euler time stepping. The author shows in particular how the best quasi-optimality constant is equal to the norm of the Ritz projection from the space where the continuous solution lives, W, to the space where the discrete solution lives, W_h. In the case of spatial semidiscrete schemes, it is shown that the boundedness of the $L^2(H^1)$-projection is a

sufficient and necessary condition for the stability of the method and for the quasi-optimality of the error estimates. These estimates constitute the starting point and main inspiration for Paper B and Paper C:

- In Paper B the results of quasi-optimality are used to derive schemes that are superconvergent at the temporal nodes, based on a temporal discretization with piecewise polynomial of arbitrary degree q.
- In Paper C we extend the results of quasi-optimality to evolution problems with random coefficients, in the spirit of what is done for elliptic problems in [9] and [28]. The possibility of keeping track of all the constants appearing in the error estimates, and of how they depend on each other, allows us to treat numerics for equations with stochastically unbounded and non-uniformly coercive coefficients. We thus generalize some recent results for these problems, where having uniformly bounded and coercive coefficients was a necessary assumption (see, for example, [16]).

1.6 Probabilistic tools

In this section we try to recap the main concepts and tools needed in order to introduce stochastic partial differential equations, and we establish the notation and the preliminaries needed to facilitate the reading of Paper A and C.

1.6.1 The probabilistic environment

We assume that the Hilbert space $(H, \langle \cdot, \cdot \rangle_H)$ is endowed with its Borel sigma-algebra, denoted by $\mathcal{B}(H)$, and with a probability measure μ. A random variable is any measurable function $X \colon (H, \mathcal{B}(H)) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, and its law is given by $\mu \circ X^{-1}$. We say that a probability measure μ on $(H, \mathcal{B}(H))$ is Gaussian if for every $v \in H$, v^* has a Gaussian law as a real-valued random variable on $(H, \mathcal{B}(H), \mu)$, where v^* is defined as $v^*(u) := \langle v, u \rangle_H$.

We recall that a finite measure μ on $(H, \mathcal{B}(H))$ is Gaussian if and only if its Fourier transform satisfies the following

$$\hat{\mu}(u) = \mathrm{e}^{\mathrm{i}\langle m, u \rangle_H - \frac{1}{2}\langle Qu, u \rangle_H}, \tag{1.72}$$

for some $m \in H$, $Q \in \mathcal{L}(H)$, with $Q \geq 0$ and $\mathrm{Tr}(Q) < \infty$. This is often denoted by $\mu = N(m, Q)$, where m and Q are respectively called *mean* and *co-*

variance operator. A Gaussian measure is uniquely characterized by these two quantities. The definition may be naturally extended to any H-valued random variable X on a probability space $(\Omega, \Sigma, \mathbb{P})$, that from now on will be assumed to be complete, by saying that X is a Gaussian random variable if it is a measurable map such that $\mathbb{P} \circ X^{-1} = N(m, Q)$. It holds in particular that for any $u, v \in H$:

$$\begin{aligned} &\mathbb{E}\Big[\langle X, u\rangle_H\Big] = \langle m, u\rangle_H, \\ &\mathbb{E}\Big[\langle X - m, u\rangle_H \langle X - m, v\rangle_H\Big] = \langle Qu, v\rangle_H, \\ &\mathbb{E}\Big[\|X - m\|_H^2\Big] = \mathrm{Tr}(Q). \end{aligned} \tag{1.73}$$

An equivalent characterization of Gaussian random variables can be given in terms of the eigenpairs of their covariance operator.

Given $m \in H$ and $Q \in \mathcal{L}(H)$, with $Q \geq 0$ and $\mathrm{Tr}(Q) < \infty$, we say that $X \colon (\Omega, \Sigma, \mathbb{P}) \to (H, \mathcal{B}(H))$ is Gaussian, with $X = N(m, Q)$, if and only if

$$X = m + \sum_{k \in \mathbb{N}} \sqrt{\lambda_k} \beta_k e_k, \tag{1.74}$$

where the (λ_k, e_k)'s are the eigenpairs of Q, the β_k's are independent real-valued random variables, with $\beta_k = N(0, 1)$ if $\lambda_k > 0$ or with $\beta_k = 0$ otherwise. The series converges in $L^2(\Omega, \Sigma, \mathbb{P}; H)$.

Given $[0, T] \subset \mathbb{R}$, an H-valued stochastic process $\{X(t)\}_{t \in [0,T]}$ is a set of H-valued random variables $X(t)$ on $(\Omega, \Sigma, \mathbb{P})$. Given two stochastic processes $\{X(t)\}_{t \in [0,T]}$ and $\{Y(t)\}_{t \in [0,T]}$, we say that they are *modifications* (or versions) of each other if $\mathbb{P}(\{X(t) \neq Y(t)\}) = 0$ for all $t \in [0, T]$ and that they are *indistinguishable* if $\mathbb{P}(\cup_{t \in [0,T]} \{X(t) \neq Y(t)\}) = 0$.

1.6.2 Wiener processes and martingales

An H-valued process $\{W(t)\}_{t \in [0,T]}$ with almost surely continuous paths, such that $W(0) = 0$ and such that it has independent, Gaussian distributed increments, i.e., $\mathbb{P} \circ (W(t) - W(s))^{-1} = N(0, Q(t - s))$, $t > s$, is called a *nuclear Q-Wiener process*. As in the case of H-valued Gaussian distributed random variables, also Q-Wiener processes have an equivalent characterization in terms of the eigenpairs of the covariance operator. Given $m \in H$ and $Q \in \mathcal{L}(H)$, with $Q \geq 0$ and $\mathrm{Tr}(Q) < \infty$, we say that $\{W(t)\}_{t \in [0,T]}$ is an H-valued Q-Wiener

process, with $X = N(m, Q)$, if and only if

$$W(t) = m + \sum_{k\in\mathbb{N}} \sqrt{\lambda_k}\beta_k(t)e_k, \tag{1.75}$$

where the (λ_k, e_k)'s are the eigenpairs of Q, the β_k's are independent real-valued standard Brownian motions on $(\Omega, \Sigma, \mathbb{P})$ if $\lambda_k > 0$, $\beta_k = 0$ otherwise. Here the series converges in $L^2(\Omega, \Sigma, \mathbb{P}; \mathscr{C}([0, T]; H))$.

A filtration $\{\Sigma_t\}_{t\in[0,T]}$ is called *normal* if Σ_0 contains all the null-sets of Σ and if $\Sigma_t = \Sigma_{t^+} := \cap_{s>t}\Sigma_s$ for every $t \in [0, T]$. If not otherwise specified we will always assume that $\Sigma_t,\ t \in [0, T]$, is a normal filtration. A process $\{X(t)\}_{t\in[0,T]}$ is said to be *adapted* to $\{\Sigma_t\}_{t\in[0,T]}$ if $X(t)$ is Σ_t-measurable for any $t \in [0, T]$. It is said to be *predictable* if, considered as a mapping from $\Omega \times [0, T]$, it is measurable with respect to the sigma algebra generated by the left-continuous processes. It is said to be *progressively measurable* if for any time $t \in [0, T]$ the map $(s, \omega) \mapsto X(s, \omega)$ is $\mathcal{B}([0, t]) \otimes \Sigma_t$-measurable.

We say that $\{W(t)\}_{t\in[0,T]}$ is a Q-Wiener process with respect to the filtration $\{\Sigma_t\}_{t\in[0,T]}$ if $\{W(t)\}_{t\in[0,T]}$ is adapted to $\{\Sigma_t\}_{t\in[0,T]}$ and if the random variable $(W(t)-W(s))$ is independent of Σ_s for every $s \in [0, t]$. It holds in particular that if $\{W(t)\}_{t\in[0,T]}$ is an H-valued Q-Wiener process on $(\Omega, \Sigma, \mathbb{P})$, then it is possible to construct a normal filtration with respect to which $\{W(t)\}_{t\in[0,T]}$ is an H-valued Q-Wiener process.

Given a Banach space V, we say that a V-valued random variable X,

$$X\colon (\Omega, \Sigma, \mathbb{P}) \to (V, \mathcal{B}(V)), \tag{1.76}$$

is Bochner integrable if it is measurable and if

$$\int_\Omega \|X(\omega)\|_V \,\mathrm{d}\mathbb{P}(\omega) < \infty. \tag{1.77}$$

A V-valued stochastic process $\{M(t)\}_{t\in[0,T]}$ on $(\Omega, \Sigma, \mathbb{P})$ is said to be a *martingale* with respect to a filtration $\{\Sigma_t\}_{t\in[0,T]}$ if:

$$\begin{aligned} &\mathbb{E}(\|M(t)\|_V) < \infty, \quad \forall t \in [0, T], \\ &\{M(t)\}_{t\in[0,T]} \text{ is adapted to } \{\Sigma_t\}_{t\in[0,T]}, \\ &\mathbb{E}(M(t)|\Sigma_s) = M(s), \quad \forall 0 \le s \le t. \end{aligned} \tag{1.78}$$

We denote by $\mathscr{M}^2_T(V)$ the space of V-valued $\{\Sigma_t\}_{t\in[0,T]}$-adapted martin-

gales with almost surely continuous paths, $M(t)$, such that

$$\sup_{t\in[0,T]} \int_\Omega \|M(t)\|_V^2 \, \mathrm{d}\mathbb{P} < \infty. \tag{1.79}$$

The space $\mathscr{M}_T^2(V)$ has a Banach space structure when endowed with the norm

$$\|M\|_{\mathscr{M}_T^2(V)} := \sup_{t\in[0,T]} \left(\mathbb{E}\Big[\|M(t)\|_V^2\Big]\right)^{\frac{1}{2}} = \left(\mathbb{E}\Big[\|M(T)\|_V^2\Big]\right)^{\frac{1}{2}}, \tag{1.80}$$

where the last equality follows from Doob's maximal inequality. In particular, any H-valued Q-Wiener process $\{W(t)\}_{t\geq 0}$, defined on $(\Omega, \Sigma, \mathbb{P})$, with respect to a normal filtration $\{\Sigma_t\}_{t\geq 0}$, belongs to $\mathscr{M}_T^2(H)$ for any positive T.

1.6.3 Stochastic integrals

The first references for the theory of stochastic integral can be traced back to Wiener, when the integrand is deterministic (see [30]), and to Ito (see [17]), when the integrand is stochastic. More actual references for a complete and comprehensive introduction to the topic are given by, for example, [12] or [19]. Given a pair of separable Hilbert spaces $(H, \langle\cdot,\cdot\rangle_H)$ and $(U, \langle\cdot,\cdot\rangle_U)$, we say that a $\mathcal{L}(U,H)$-valued process $\{\Phi(t)\}_{t\in[0,T]}$ is *elementary* if there exists a sequence $0 = t_0 \leq \ldots \leq t_N = T$ such that

$$\Phi(t) = \sum_{i=0}^{N-1} \Phi_i \, \chi_{(t_i, t_{i+1})}, \tag{1.81}$$

where each Φ_i is a strongly Σ_{t_i}-measurable $\mathcal{L}(U,H)$-valued random variable that only takes a finite number of values in $\mathcal{L}(U,H)$. The space of elementary processes is usually denoted by $\mathcal{E}$.

For an elementary process $\Phi \in \mathcal{E}$ its stochastic integral with respect to a U-valued Q-Wiener process is defined as:

$$\int_0^t \Phi(s)\, \mathrm{d}W(s) := \sum_{i=0}^{N-1} \Phi_i \Delta W_i(t), \tag{1.82}$$

where $\Delta W_i(t) := W(t_{i+1} \wedge t) - W(t_i \wedge t)$. For any $\Phi \in \mathcal{E}$, the stochastic integral is a continuous square-integrable $\Sigma_{t\in[0,T]}$-martingale, i.e.,

$$\left\{ \int_0^t \Phi \, \mathrm{d}W \right\}_{t\in[0,T]} \in \mathscr{M}_T^2(H). \tag{1.83}$$

In particular its expectation is 0 and the so called *Itô-isometry* holds:

$$\mathbb{E}\left[\left\| \int_0^T \Phi \, \mathrm{d}W \right\|^2\right] = \mathbb{E}\left[\int_0^T \|\Phi(s) Q^{\frac{1}{2}}\|^2_{\mathcal{L}_2(U,H)} \, \mathrm{d}s \right]. \tag{1.84}$$

For simplicity, the following notation is also used

$$\|\Phi\|_T^2 = \mathbb{E}\left[\int_0^T \|\Phi(s) Q^{\frac{1}{2}}\|^2_{\mathcal{L}_2(U,H)} \, \mathrm{d}s \right], \tag{1.85}$$

and $\|\cdot\|_T$ defines a norm on $\mathcal{E}$, once we re-define $\mathcal{E}$ to be the quotient space $\mathcal{E}/\mathcal{E}_0$, where

$$\mathcal{E}_0 := \left\{ \Phi \in \mathcal{E} \colon \Phi = 0 \text{ on } Q^{\frac{1}{2}}(U),\ \mathrm{d}t \otimes \mathbb{P} - \text{a.s.} \right\}. \tag{1.86}$$

The stochastic integral thus defines a continuous isometry between the space $(\mathcal{E}, \|\cdot\|_T)$ and the complete space $(\mathscr{M}_T^2, \|\cdot\|_{\mathscr{M}_T^2})$, that can hence be extended to $\bar{\mathcal{E}}$, abstract completion of $\mathcal{E}$, which will be denoted by $\mathscr{N}_W^2([0,T];H)$ and that can be explicitly characterized as follows:

$$\begin{aligned} \mathscr{N}_W^2([0,T];H) := \Big\{ & \Phi \colon [0,T] \times \Omega \to \mathcal{L}_2^0 \colon \\ & \Phi \text{ is predictable and } \|\Phi\|_T < \infty \Big\}. \end{aligned} \tag{1.87}$$

By a localization procedure it is possible to further extend the stochastic integral to an even broader space by dropping the assumption on the boundedness of the $\|\cdot\|_T$-norm and requiring only that

$$\mathbb{P}\left(\int_0^T \|\Phi\|^2_{\mathcal{L}_2^0} \, \mathrm{d}s < \infty \right) = 1. \tag{1.88}$$

Such a space is denoted by $\mathscr{N}_W([0,T];H)$.

Finally, it is possible to consider even the case when $\mathrm{Tr}(Q) = \infty$, as for example when Q is the identity operator. Indeed it is always possible to find a Hilbert space $(\tilde{U}, \langle\cdot,\cdot\rangle_{\tilde{U}})$ such that the embedding $J \colon U_0 \to \tilde{U}$ is Hilbert–Schmidt and define $\tilde{Q} \colon \tilde{U} \to \tilde{U}$ as $\tilde{Q} = JJ^*$, so that it is bounded, positive semi-definite and trace-class. The series

$$\tilde{W}(t) = \sum_{k \geq 1} \beta_k(t) \, Je_k, \quad t \in [0,T] \tag{1.89}$$

then converges in $\mathscr{M}_T^2(\tilde{U})$ and defines a $\tilde{Q}$-Wiener process on $\tilde{U}$, where in particular $\tilde{U}_0 = J(U_0)$ and $J \colon U_0 \to \tilde{U}_0$ is an isometric isomorphism. The process

$\{\tilde{W}(t)\}_{t\in[0,T]}$ is called *cylindrical Wiener process* and for processes $\Phi \in \mathscr{N}_W^2$, the stochastic integral with respect to a cylindrical Wiener process is defined as

$$\int_0^t \Phi(s)\,\mathrm{d}W(s) := \int_0^t \Phi(s)J^{-1}\,\mathrm{d}\tilde{W}(s). \tag{1.90}$$

An extension of the previous definitions is given by the *weak stochastic integral*, which is defined for any $\Phi \in \mathscr{N}_W([0,T];H)$ and for any continuous, Σ_t-adapted and H-valued process f as:

$$\int_0^T \langle f(t), \Phi(t)\,\mathrm{d}W(t)\rangle_H := \int_0^T \tilde{\Phi}_f(t)\,\mathrm{d}W(t), \tag{1.91}$$

where $\tilde{\Phi}_f(t)(u) := \langle f(t), \Phi(t)u\rangle_H$ for any $u \in U_0$. It holds in particular that $\tilde{\Phi}_f\colon \Omega\times[0,T] \to \mathcal{L}_2(U_0,\mathbb{R})$ is a $\mathcal{P}_T/\mathcal{B}(\mathcal{L}_2(U_0,\mathbb{R}))$-measurable process, whose norm satisfies

$$\|\tilde{\Phi}_f(t)\|_{\mathcal{L}_2(U_0,\mathbb{R})} = \|\Phi^*(t)f(t)\|_{U_0}, \tag{1.92}$$

and

$$\begin{aligned}\int_0^T \|\tilde{\Phi}_f(t)\|^2_{\mathcal{L}_2(U_0,\mathbb{R})}\,\mathrm{d}t \\ \leq \sup_{0\leq t\leq T}\|f(t)\|_H \int_0^T \|\Phi(t)\|^2_{\mathcal{L}_2^0}\,\mathrm{d}t < \infty,\ \mathbb{P}\text{-a.s.}.\end{aligned} \tag{1.93}$$

Here $\mathcal{P}_T$ denotes the predictable sigma-algebra on $\Omega\times[0,T]$ and the notation $\mathcal{P}_T/\mathcal{B}(\mathcal{L}_2(U_0,\mathbb{R}))$-measurable indicates that the process is measurable when its domain and co-domain are endowed respectively with $\mathcal{P}_T$ and $\mathcal{B}(\mathcal{L}_2(U_0,\mathbb{R}))$ as sigma-algebras.

1.7 Stochastic evolution equations

In order to facilitate the reading of Paper A, which deals with the stochastic version of (1.40), we devote this section to the formal introduction of stochastic evolution problems. A good reference about this topic, which has been an important source of inspiration for proving some of the results in Paper A, can be found in [10].

We assume throughout this whole section that the following objects are given:

- A progressively measurable map $A : [0,T]\times\Omega\times V \to V^*$, with associated bilinear form $a(\cdot,\cdot;\cdot,\cdot)$ that is bounded and weakly coercive. This is a generalization of the operator A defined in § 1.3.2.

- A map $f \in L^2(\Omega \times (0,T); V^*)$ that it is a predictable process with Bochner integrable trajectories on $[0,T]$.

- A Q-Wiener process $\{W(t)\}_{t\in[0,T]}$, where we assume that the covariance operator $Q \in \mathcal{L}(H)$ is trace class, or, equivalently, that $Q^{1/2} \in \mathcal{L}_2(H)$.

The problem of interest reads, in its abstract formulation:

$$\begin{aligned} &\mathrm{d}U(t) + A(t)U(t)\,\mathrm{d}t = f(t)\,\mathrm{d}t + \mathrm{d}W(t), \quad t \in (0,T], \\ &U(0) = U_0. \end{aligned} \tag{1.94}$$

Such a notation, with $\mathrm{d}W(\cdot)$, is purely symbolical and indeed refers to an underlying stochastic integral equation. In order to give it a meaning, we have to introduce a formal and well defined concept of solution. This is done in great detail in the first part of Paper A and here we only recall the main features.

We say that a continuous H-valued (Σ_t)-adapted process $\{U(t)\}_{t\in[0,T]}$ is a *variational solution*[1] to (1.94) if for its $\mathrm{d}t\otimes\mathbb{P}$ equivalence class $\hat{U}$ we have $\hat{U} \in L^2(\Omega\times(0,T), \mathrm{d}t\otimes\mathbb{P}; V)$ and $\mathbb{P}$-a.s.

$$U(t) = U(0) - \int_0^t A(s)\bar{U}(s)\,\mathrm{d}s + \int_0^t f(s)\,\mathrm{d}s + \int_0^t \mathrm{d}W(s), \tag{1.95}$$

for any $t \in [0,T]$, where $\bar{U}$ is any V-valued progressively measurable $\mathrm{d}t\otimes\mathbb{P}$ version of $\hat{U}$. The following *Itô formula* holds:

$$\begin{aligned} \mathbb{E}\Big[\|U(t)\|_H^2\Big] &= \mathbb{E}\Big[\|U_0\|_H^2\Big] \\ &\quad + \int_0^t \mathbb{E}\Big[2_{V^*}\langle A(s)\bar{U}(s), \bar{U}(s)\rangle_V + \|Q^{\frac{1}{2}}\|_{\mathcal{L}_2(H)}\Big]\,\mathrm{d}s, \end{aligned} \tag{1.96}$$

and, for any $t \in [0,T]$, we have that

$$\mathbb{E}\Big[\sup_{t\in[0,T]} \|U(t)\|_H^2\Big] < \infty. \tag{1.97}$$

[1] See [24, Chapt. 4].

If the operator A is now possibly unbounded, independent of ω and t, and defined on a certain domain $\mathscr{D}(A)$, $A : \mathscr{D}(A) \subset H \to H$, as in § 1.3.3, an H-valued, predictable stochastic process $\{U(t)\}_{t\in[0,T]}$ which is Bochner integrable $\mathbb{P}$-a.s. and satisfies

$$\begin{aligned}\langle U(t), v\rangle_H = \langle U(0), v\rangle_H - \int_0^t \langle U(s), A^* v\rangle_H \,\mathrm{d}s \\ + \int_0^t \langle f(s), v\rangle_H \,\mathrm{d}s + \int_0^t \langle \,\mathrm{d}W(s), v\rangle_H,\end{aligned} \tag{1.98}$$

$\mathbb{P}$-a.s., $\forall v \in \mathscr{D}(A^*)$, $t \in [0,T]$, is called a *weak solution*[2] to (1.94). If $-A$ is the generator of a strongly continuous semigroup $S(\cdot)$ in H and if

$$\int_0^T \|S(t)Q^{\frac{1}{2}}\|^2_{\mathcal{L}_2(H)} \,\mathrm{d}t < \infty, \tag{1.99}$$

then the unique weak solution coincides with the *mild solution*, whose expression is given for all $t \in [0,T]$ by:

$$U(t) = S(t)U_0 + \int_0^t S(t-s)f(s) \,\mathrm{d}s + \int_0^t S(t-s) \,\mathrm{d}W(s). \tag{1.100}$$

Assuming for simplicity that $f = 0$, it is known, see for example [32], that the mild solution, in the hypothesis that $U_0 \in L^2(\Omega; \dot{H}^\beta)$, where $\dot{H}^\beta := \mathscr{D}(A^{\frac{\beta}{2}})$, and that $\|A^{\frac{\beta-1}{2}}\|_{\mathcal{L}_2^0} < \infty$ for some $\beta \geq 0$, satisfies for any $t \in [0,T]$

$$\|U(t)\|_{L^2(\Omega;\dot{H}^\beta)} \leq C\Big(\|U_0\|_{L^2(\Omega;\dot{H}^\beta)} + \|A^{\frac{\beta-1}{2}}\|_{\mathcal{L}_2^0}\Big), \tag{1.101}$$

and, in particular, if Q is a trace-class operator,

$$\|U(t)\|_{L^2(\Omega;\dot{H}^1)} \leq C\Big(\|U_0\|_{L^2(\Omega;\dot{H}^1)} + [\mathrm{Tr}(Q)]^{\frac{1}{2}}\Big). \tag{1.102}$$

Several results about the numerical approximation of the mild solution with semidiscrete or fully discrete schemes are known in literature; however we will not mention them, considering that the main goal with the appended papers is not to deal with numerics for this type of problems. The reader can refer to the survey article [18] and references therein in order to get an idea about the state of the art of this topic.

[2] See [12].

1.8 Two motivating examples of stochastic evolution equations

Stochastic evolution equations in infinite dimensions are a natural generalization of stochastic ordinary equations and beside a natural mathematical interest, their theory has motivations coming also from other fields, such as physics, chemistry and biology. We present in this section two examples, taken from [12], of stochastic PDE's coming from biology and from physics.

Stochastic semilinear equations have been used in population genetics to model changes in the structure of population in both time and space. Given a population $p(t,\cdot)$ at a time $t \geq 0$, a way to model the mass distribution of p is given by the equation

$$\mathrm{d}p(t,\xi) = a\Delta p(t,\xi)\,\mathrm{d}t + b\sqrt{p_+(t,\xi)}\,\mathrm{d}W, \qquad \xi \in \mathbb{R}^d. \tag{1.103}$$

Here a and b are positive constants and W is a H-valued Wiener process with nuclear covariance operator Q. The space H is in this case given by $L^2(\mathbb{R}^d)$, the operator A is given by $a\Delta$, with domain $\mathscr{D}(A) := H^2(\mathbb{R}^d)$, and

$$(\Psi x)u(\xi) := b\sqrt{x_+(\xi)}u(\xi). \tag{1.104}$$

For more details about this example, the reader can refer to [14, Appendix I].

Another example is given by the stochastic diffusion-reaction equation. The equation reads, in its deterministic form:

$$\frac{\partial u}{\partial t}(t,\xi) = \sigma^2\frac{\partial^2 u}{\partial \xi^2}(t,\xi) + f(u(t,\xi)), \qquad t \geq 0,\ \xi \in [0,T]. \tag{1.105}$$

The many-particles nature of a real system, leads to having internal fluctuations, which can be modelled according to the following equation:

$$\frac{\partial u}{\partial t}(t,\xi) = \sigma^2\frac{\partial^2 u}{\partial \xi^2}(t,\xi) + f(u(t,\xi)) + \dot{W}(t,\xi), \quad t \geq 0,\ \xi \in [0,T], \tag{1.106}$$

with $\dot{W}$ being a temporal and spatial white noise. This equation is clearly of the same type as (1.94), and constitute a further motivation for investigating this sort of problems. For more details about this last example, the reader is referred to [2].

1.9 Summary of Paper A

In Paper A we deploy the idea of Section 1.4 in connection with the linear stochastic heat equation. The paper provides the first application of the inf-sup theory in order to prove existence and uniqueness for the solution to the linear stochastic heat equation, once the problem is formulated in a weak space-time form. This approach offers two advantages: results of existence and uniqueness are obtained in a relatively simple way and the problem is set up in a way that naturally allows Petrov-Galerkin discretization. This kind of approach has been widely used by other authors in the deterministic case (see [21, 25, 29]), and in a stochastic/random setting (see [16, 26]). Our work can be viewed as a tool to be used for future research on numerical aspects of the same equation, in the same way as the deterministic counterpart of this theory has been used in the past to construct and analyse numerical solutions of evolution equations. In particular, the comprehensive analysis of our concept of solution, the consistency with known concepts of solutions, the bounds derived for the norm of the solution, and the sufficient conditions to have further spatial regularity, are of crucial importance when deriving error bounds for the numerical solution of the same equation. Although the core of our work is based on a linear problem with additive noise, which is the setting in which the inf-sup theory naturally takes place, we also show how our findings extend to more general equations, possibly semilinear and with multiplicative noise.

1.10 Summary of Paper B

In Paper B we use the weak space-time formulation of the heat equation in order to investigate the numerical property of the schemes obtained by discretizing the problem on proper piecewise polynomial tensor subspaces. The novelty in the approach we propose is to exploit the presence of a pointwise component of the solution otherwise neglected in other works (see [21] or [29], for example). This component is the pointwise evaluation of the solution one would obtain by discretizing the problem stated in its first space-time variational formulation, with polynomials of one degree higher with respect to time (see Chapter 3 for more details). We prove that this component of the solution can be constructed on any arbitrary grid point, and has the remarkable property of giving superconvergence of the error, with order $2(q+1)$, with q being the polynomial degree of

the numerical solution, and where the error is measured with respect to the norm $\max \|\cdot\|_H$.

1.11 Summary of Paper C

In Paper C we investigate the theory of quasi-optimality for the heat equation with random coefficients. We assume that both the operator A and the function f appearing in the equation depend on a random parameter ω. The novelty of this paper is that the operator A is not assumed to be uniformly coercive and uniformly bounded with respect to ω. We prove the existence of p-moments of the solution in terms of the integrability of A and f, by exploiting the inf-sup theory. The main advantage of our approach is that every constant appearing in the estimates we provide is known explicitly, so that we can track down all of them in order to provide the sharpest possible estimate for the norm of the solution. Under the further assumption that the operator A satisfies a certain property of "not having its minimum and maximum too far apart as functions of ω", we prove a result of quasi-optimality for the error obtained by a Petrov-Galerkin semidiscretization and full-discretization of the problem similar to the one used in Paper B. For the semidiscretization, the quasi-optimality constant that we obtain is in than absolute constant that does not depend on ω, so that we have optimal rates of convergence in $L^p(\Omega;\cdot)$ under the same assumptions needed to ensure existence and uniqueness of the solution. In the fully discrete case we show instead that the quasi-optimality constant has an ω-dependence which apparently cannot be avoided, and that affects the error estimates.

References

[1] ANDREEV, R., *Stability of Space-Time Petrov-Galerkin Discretizations for Parabolic Evolution Equations*, PhD thesis, 2012.

[2] ARNOLD, L., *Mathematical models of chemical reactions*, Stoch. syst., Vol. 78, 1981, pp. 111–134.

[3] BABUŠKA, I., *Error bounds for finite element method*, Numer. Math., Vol. 16, 1971, pp. 322–333.

[4] BABUŠKA, I. AND AZIZ, A. K., *The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations* Proc. Sympos., Univ. Maryland, Baltimore, Md., 1972, pp. 1–359.

[5] BABUŠKA, I. AND JANIK, T., *The h-p version of the finite element method for parabolic equations. I. The p-version in time*, Numer. Methods Partial Differential Equations, Vol. 5, 1989, pp. 363–399.

[6] BABUŠKA, I. AND JANIK, T., *The h-p version of the finite element method for parabolic equations. II. The h-p version in time*, Numer. Methods Partial Differential Equations, Vol. 6, 1990, pp. 343–369.

[7] BREZIS, H., *Functional Analysis, Sobolev Spaces and Partial Differential Equations*, Springer, 2011.

[8] CÉA, J., *Approximation variationelle des problèmes aux limites*, Inst. Fourier, Vol. 14(2), 1964, pp. 345–444.

[9] CHARRIER, J., *Strong and weak error estimates for elliptic partial differential equations with random coefficients*, SIAM J. Numer. Anal., Vol. 50(1), 2012, pp. 216–246.

[10] CHOW, P. L., *Stochastic Partial Differential Equations*, Chapman & Hall/CRC, 2007.

[11] CIOICA, P. A., DAHLKE, S., DÖHRING, N., FRIEDRICH, U., KINZEL, S., LINDNER, F., RAASCH, T., RITTER, K., AND SCHILLING, R. L., *Convergence analysis of spatially adaptive Rothe methods*, Found. Comput. Math., Vol. 14(5), 2014, pp. 863–912.

[12] DA PRATO, G. AND ZABCZYK, J., *Stochastic Equations in Infinite Dimensions*, Cambridge University Press, 1992.

[13] DAUTRAY, R. AND LIONS, J. L., *Mathematical Analysis and Numerical Methods for Science and Technology. Vol. 5*, Springer-Verlag, 1992.

[14] DAWSON, D. A., *Stochastic evolution equations*, Math. Biosci., vol. 5(1), 1972, pp. 1–52.

[15] ERN, A. AND GUERMOND, J. L., *Theory and Practice of Finite Elements*, Springer-Verlag, 2004.

[16] GITTELSON, C. J., ANDREEV, R. AND SCHWAB, CH., *Optimality of adaptive Galerkin methods for random parabolic partial differential equations*, J. comput. appl. math, Vol. 263, 2014, pp. 189–201.

[17] ITO, K., *Stochastic integral*, Proc. Imp. Acad. Tokyo, Vol. 20, 1944, pp. 519–524.

[18] JENTZEN, A. AND KLOEDEN, P. E., *The numerical approximation of stochastic partial differential equations*, Milan J. Math., Vol. 77(1), 2009, pp. 205–244.

[19] KOVACS, M. AND LARSSON, S., *Introduction to Stochastic Partial Differential Equations,*, Lecture notes, 2008.

[20] LUNARDI, A., *Interpolation Theory*, Edizioni della Normale, Pisa, 2009.

[21] MOLLET, CH., *Stability of Petrov-Galerkin discretizations: application to the space-time weak formulation for parabolic evolution problems*, Comput. Methods Appl. Math., Vol. 14(2), 2014, pp. 231–255.

[22] OUHABAZ, E., *Analysis of Heat Equations on Domains*, London Mathematical Society Monographs Series, 2005.

[23] PAZY, A., *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer, 1983.

[24] PRÉVÔT, C., AND RÖCKNER, M., *A Concise Course on Stochastic Partial Differential Equations*, Springer, 2007.

[25] SCHWAB, CH., STEVENSON, R., *Space-time adaptive wavelet methods for parabolic evolution problems*, Math. Comp., Vol. 78, 2009, pp. 1293–1318.

[26] SCHWAB, CH. AND SÜLI, E., *Adaptive Galerkin approximation algorithms for Kolmogorov equations in infinite dimensions*, Stoch. Partial Differ. Equ. Anal. Comput., Vol 1(1), 2013, pp. 204–239.

[27] TANTARDINI, F., *Quasi-Optimality in the Backward Euler-Galerkin Method for Linear Parabolic Problems*, PhD Thesis, 2013.

[28] TECKENTRUP, A. L., *Multilevel Monte Carlo Methods and Uncertainty Quantification*, PhD Thesis, 2013.

[29] URBAN, K. AND PATERA, A. T., *An improved error bound for reduced basis approximation of linear parabolic problems*, Math. Comp., Vol 83, 2014, pp. 1599–1615.

[30] WIENER, N., *Generalised Harmonic Analysis*, Acta Math., Vol. 55, 1930, 117–258.

[31] XU, J. AND ZIKATANOV, L., *Some observations on Babuška and Brezzi theories*, Numer. Math., Vol. 94(1), 2003, pp. 195–202.

[32] YAN, Y., *Semidiscrete Galerkin approximation for a linear stochastic parabolic partial differential equation driven by an additive noise*, BIT., Vol. 44(4), 2004, pp. 829–847.

2

Discrete Variational Derivative Method (DVDM)

2.1 Introduction and motivation

The fourth paper included in this thesis deals with the construction of numerical schemes for solving a certain non-linear evolution problem so that its energy and momenta are preserved. This particular equation belongs to a wider target of equations for which the total energy, either remain constant (conservative PDE's), or monotonically decrease with time (dissipative PDE's). It is in general desirable while solving conservative PDE's to use a numerical scheme that retains the conservation property. This is both due to the fact that these schemes will in general be more stable from the numerical point of view, and to the fact that the properties preserved might have some practical meaning themselves, from an engineering or physical point of view. The field of structure preserving numerical integration algorithms, called *geometric numerical integration* started in the mid-eighties for symplectic integration of Hamiltonian ODE's; in later years there has also been some work on similar approaches for Hamiltonian PDE's. For ODE's, several unified approaches are well established, and cover not only the case of conservative and dissipative equations, but also equations with many other geometric structures, such as the symplectic method for Hamiltonian systems, the Lie group method for constrained mechanical systems, methods that preserve first-integrals, and methods for ODE's evolving on manifolds. A good text about structure-preserving methods for ODE's can be found in [3]. A comprehensive and unified approach to numerically deal with conservative and dissipative PDE's, used in Paper D to solve the EPDiff equation, is the Discrete Variational Derivative Method (DVDM), which we describe in the next section.

2.2 Discrete Variational Derivative Method

The DVDM is a general theory that easily allows the construction of conservative/dissipative schemes for real/complex-valued PDE's. In order to facilitate the reading, we present it only for a specific case:

- We restrict ourselves to the case of periodic boundary conditions.
- We restrict ourselves to the case of real-valued conservative PDE's.

This is not a necessary restriction and can easily be avoided. The DVDM is indeed usable in a more general framework, and covers the following:

- Non-periodic boundary conditions.
- System of equations.
- Dissipative equations.
- Complex-valued equations

For more details about this, as well as for further details on what is presented in this section, we refer to [2]. We summarize in Table 2.1 the main features of DVDM, making a comparison between how the energy is conserved in the continuous equation and how this can be done for its discretization. The main

Continuous	**Discrete**
Energy function: $G(u, u_x)$	Discrete energy function: $G_d(U^{(m)})$
Variational derivative: $\frac{\delta G}{\delta u}$	Discrete variational derivative: $\frac{\delta G_d}{\delta (U^{(m)}, U^{(m+1)})}$
Definition of the PDE: $\frac{\partial u}{\partial t} = H \frac{\delta G}{\delta u}$, H skew-symmetric.	Definition of a FD-scheme: $\frac{U_k^{(m+1)} - U_k^{(m)}}{\Delta t} = H_d \frac{\delta G_d}{\delta (U^{(m)}, U^{(m+1)})_k}$, H_d skew-symmetric discretization of H.
Consequence: Conservation property $\frac{\mathrm{d}u}{\mathrm{d}t} \int_0^L G(u, u_x)\, \mathrm{d}x = 0$	Consequence: Discrete conservation property $\sum_{k=0}^{\mathcal{K}-1} G_{d,k}(U^{(m+1)} - U^{(m)})\Delta x$

Table 2.1: *Continuous calculus versus discrete calculus*

difference between the discrete variational derivative method, and other structure preserving methods, is therefore what gets discretized first. With DVDM we discretize the energy and compute a discrete variation; the conservation property, together with the definition of the scheme, comes as a side product, while the conventional approach is to discretize the equations directly, and only thereafter investigate conservation properties (see figure 2.1).

In order to introduce the next example, we assume to have an inner product defined on $\mathbb{R}^{\mathcal{K}}$, $\mathcal{K} \in \mathbb{N}$, by:

$$\langle \mathbf{v}, \mathbf{w} \rangle := \sum_{k=0}^{\mathcal{K}-1} v_k w_k \Delta x, \qquad \mathbf{v}, \mathbf{w} \in \mathbb{R}^{\mathcal{K}}. \tag{2.1}$$

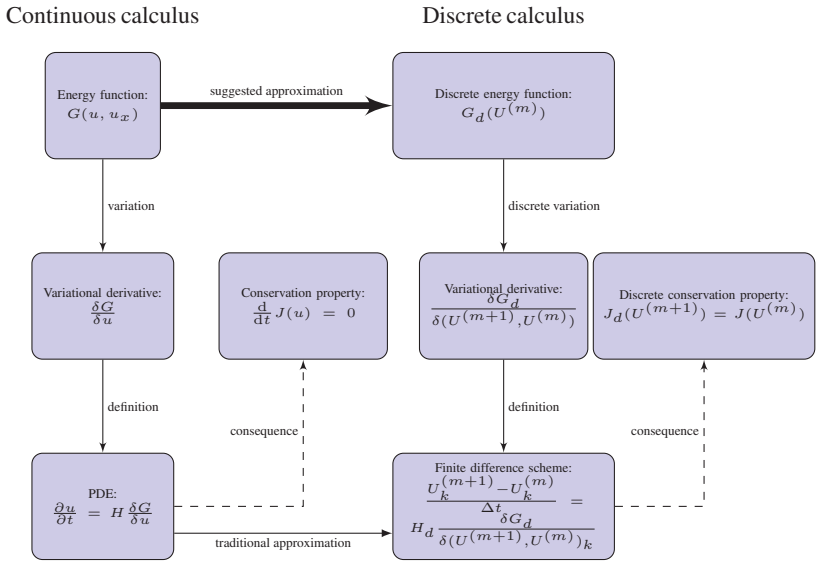

Figure 2.1: *Standard strategy versus DVDM ([2, Chapter 1])*

This induces a natural norm, given by:

$$\|\mathbf{w}\|^2 := \sum_{k=0}^{\mathcal{K}-1} w_k^2 \Delta x, \qquad \mathbf{w} \in \mathbb{R}^{\mathcal{K}}. \tag{2.2}$$

We use the following standard notation for centred finite differences:

$$\delta_k^{<1>} f_k := \frac{f_{k+1} - f_{k-1}}{2\Delta x}. \tag{2.3}$$

Example 7. We consider a one-dimensional domain $[0, L]$, with periodic boundary conditions, and we assume that that an energy G is given and it is defined as:

$$G(u, u_x) = \frac{u^2}{2}. \tag{2.4}$$

This gives us a *total energy*, defined as:

$$J(u) := \int_0^L G(u, u_x)\, \mathrm{d}x. \tag{2.5}$$

If we introduce a variation $\frac{\delta G}{\delta u}$, we obtain what is called *variational derivative*, which in this case is given by:

$$\frac{\delta G}{\delta u} = u. \tag{2.6}$$

The PDE corresponding to the energy G is then constructed having the conservation of the energy as starting point rather than arrival point:

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x} u. \tag{2.7}$$

In this way the conservation property is within the equation itself, and it is not something imposed in a second moment. It is easy to see that the variation of the total energy is zero, since

$$\frac{\mathrm{d}u}{\mathrm{d}t} J(u) = \frac{\mathrm{d}u}{\mathrm{d}t} \int_0^L G(u, u_x)\, \mathrm{d}x = \frac{\mathrm{d}u}{\mathrm{d}t} \int_0^L \frac{u^2}{2}\, \mathrm{d}x = 0. \tag{2.8}$$

The idea behind the DVDM is to mimic what happens in the continuous case. For the linear convection equation introduced above, we start by defining a discrete energy, given by

$$G_d(U^{(m)}) = \frac{[U^{(m)}]^2}{2}. \tag{2.9}$$

Here U denotes the discrete solution we are looking for, and the superscript (m) refers to a given time instant t_m. Having a discrete energy allows us to introduce a discrete variation to G_d, which is the counterpart to the variational derivative $\frac{\delta G}{\delta u}$ previously used. The term discrete variation has here a very precise meaning, and can be formally defined in mathematical terms, but we refrain from presenting the details here, and we limit ourselves to observe that it leads to the following expression:

$$\frac{\delta G_d}{\delta(U^{(m)}, U^{(m+1)})} = \frac{U^{(m+1)} + U^{(m)}}{2}. \tag{2.10}$$

The idea is now to construct a scheme starting from the discrete energy, in the same way in the continuous case the equation was defined from the energy, and not the other way round. In this particular case, we have:

$$\frac{U_k^{(m+1)} - U_k^{(m)}}{\Delta t} = \delta_k^{<1>} \frac{U_k^{(m+1)} + U_k^{(m)}}{2}, \tag{2.11}$$

where $\delta_k^{<1>}$ is the discrete counterpart to ∂_x. It is now easy to see that:

$$\sum_{k=0}^{\mathcal{K}-1} G_{d,k}(U^{(m+1)})\Delta x = \sum_{k=0}^{\mathcal{K}-1} G_{d,k}(U^{(m)})\Delta x. \tag{2.12}$$

Moreover, we have a further discrete conservation law holds, as a side product:

$$\sum_{k=0}^{\mathcal{K}-1} U^{(m)}\Delta x = \sum_{k=0}^{\mathcal{K}-1} U^{(m+1)}\Delta x. \tag{2.13}$$

This does not occur all the time and strongly depend on the kind of discretization chosen for the energy (see for example Scheme 3 in Paper D and the remark after Theorem 5).

We can see that we have two main ingredients that are needed in order to make the DVDM work:

- A discrete variation.
- A scheme based on the discrete variation.

We try to clarify both points, by further exploiting the example given by the linear convection equation. The discrete variation can be obtained by computing the quantity:

$$\frac{1}{\Delta t}\Big(\sum_{k=0}^{\mathcal{K}-1} G_{d,k}(U^{(m+1)}) - \sum_{k=0}^{\mathcal{K}-1} G_{d,k}(U^{(m)})\Big)\Delta x. \tag{2.14}$$

This fact is true in general, and although formulas for computing the discrete variational derivative exist, we found it easier and more instructive to compute the discrete variational derivative directly, case by case. For the discrete energy G_d introduced in Example 7, this gives:

$$\begin{aligned}
&\frac{1}{\Delta t}\Big(\sum_{k=0}^{\mathcal{K}-1} G_{d,k}(U^{(m+1)}) - \sum_{k=0}^{\mathcal{K}-1} G_{d,k}(U^{(m)})\Big)\Delta x \\
&\qquad = \frac{1}{\Delta t}\sum_{k=0}^{\mathcal{K}-1}\Big(\frac{(U_k^{(m+1)})^2}{2} - \frac{(U_k^{(m)})^2}{2}\Big)\Delta x \\
&\qquad = \sum_{k=0}^{\mathcal{K}-1}\Big(\underbrace{\frac{U_k^{(m+1)} - U_k^{(m)}}{\Delta t}}_{\text{Approximation of } \dot{u}}\Big)\Big(\underbrace{\frac{U_k^{(m+1)} + U_k^{(m)}}{2}}_{\text{Discrete variational derivative}}\Big)\Delta x.
\end{aligned} \tag{2.15}$$

The two quantities highlighted in the previous equations are the quantities upon which we construct our scheme:

$$\underbrace{\frac{U_k^{(m+1)} - U_k^{(m)}}{\Delta t}}_{\text{Approximation of } \dot{u}} = \overbrace{\delta_k^{<1>}}^{\text{Discretely skew-symmetric}} \underbrace{\frac{U_k^{(m+1)} + U_k^{(m)}}{2}}_{\text{Discrete variational derivative}}, \tag{2.16}$$

where it is crucial to have a discrete operator which is the skew-symmetric discretization of the continuous one, as it will be more clear in the next equation. We can now see how and why conservation of the energy happens, by using (2.16) in (2.15), which gives:

$$\begin{aligned} &\frac{1}{\Delta t}\Big(\sum_{k=0}^{\mathcal{K}-1} G_{d,k}(U^{(m+1)}) - G_{d,k}(U^{(m)})\Big)\Delta x \\ &\quad = \sum_{k=0}^{\mathcal{K}-1} \Big(\frac{U_k^{(m+1)} - U_k^{(m)}}{\Delta t}\Big)\Big(\frac{U_k^{(m+1)} + U_k^{(m)}}{2}\Big)\Delta x \\ &\quad = \sum_{k=0}^{\mathcal{K}-1} \Big(\delta_k^{<1>}\frac{U_k^{(m+1)} + U_k^{(m)}}{2}\Big)\frac{U_k^{(m+1)} + U_k^{(m)}}{2}\Delta x = 0, \end{aligned} \tag{2.17}$$

where the last equality holds because $\delta_k^{<1>}$ is skew-symmetric with respect to the discrete inner product introduced in (2.1). Even if at first glance what we have done might not appear so innovative, since at the end of the day we are doing nothing but rediscovering the Crank–Nicolson scheme, it is worth stressing some important features, which have been the main motivation for investigating this method:

- The method works "black box" in more complicated cases, such as nonlinear equations, where the derivation of an energy preserving scheme might not be so easy.
- The conservation property is not imposed after the derivation of the scheme, but it comes as a consequence of how the schemes are built.
- There is a general formula to compute all the discrete quantities involved in the scheme (see [2, Chap.3]), if one does not want to derive them case by case.

For all these reasons we decide to use the DVDM as the main tool to investigate the numerical solution of the EPDiff equation.

2.3 Summary of Paper D

In the paper we construct and investigate numerical methods for the EPDiff equation. The EPDiff equation can be thought as a multidimensional generalization of the Camassa–Holm equation for shallow water (see [4]), and is of particular importance in shape analysis, where it can be shown (see [10]) that the problem of finding the "best" continuous warp between medical images and shapes, is equivalent to solving the EPDiff equation. The equation has important features, in particular it is a Hamiltonian system with respect to a Lie-Poisson structure, which implies that it has conservation laws. For Poisson structure preserving discretizations of the EPDiff equation, the only known approaches are to use particle methods (see [1, 7]). The Compatible Differencing Algorithm (CDA, see [5, 6, 9]) is another approach that has been suggested (see [4]), but it is unclear to what extent such methods preserve structure. Instead of focusing on conservation of structure, CDA is based on the fact that the equations can be rewritten in a form that contains divergence, gradient and curl operators. In this paper we develop energy conserving geometric integrators for the EPDiff equation, in two spatial dimensions. Our schemes conserve the total energy and, in some cases, also total momentum. They are based on the DVDM approach described in Section 2.2 and on a generalization of the DVDM schemes for the Camassa–Holm equation suggested in [8]. The methods are tested with a series of benchmark problems of singular wave fronts interactions, first proposed in [4], and later also used in [1].

References

[1] CHERTOCK, A., TOIT, P. D., AND MARSDEN, J. E., *Integration of the EPDiff equation by particle methods,* ESAIM: Math. Model. Numer. Anal., Vol. 46(3), 2012, pp. 515–534.

[2] FURIHATA, D., MATSUO, T., *Discrete Variational Derivative Method: a Structure-Preserving Numerical Method for Partial Differential Equations,* Chapman and Hall/CRC, 2011.

[3] HAIRER, E., LUBICH, C. AND WANNER, G., *Geometric Numerical Integration,* Springer–Verlag, 2002.

[4] HOLM, D. D. AND STALEY, M. F., *Interaction dynamics of singular wave fronts,* Preprint, arXiv:1301.1460, 2013.

[5] HYMAN, J. M. AND SHASHKOV, M., *The adjoint operators for the natural discretizations of the divergence, gradient and curl on logically rectangular grids,* IMACS J. Appl. Num. Math., Vol. 25, 1997, pp. 413–442

[6] HYMAN, J. M. AND SHASHKOV, M., *Natural discretizations for the divergence, gradient, and curl on logically rectangular grids,* Comput. Math. Appl., Vol. 33(4), 1997, pp. 81–104.

[7] MCLACHLAN, R. I. AND MARSLAND, S., *N-particle dynamics of the Euler equations for planar diffeomorphisms,* Dyn. Syst., Vol. 22(3), 2007, pp. 269–290.

[8] MIYATAKE, Y., MATSUO, T. AND FURIHATA, D., *Invariants-preserving integration of the modified Camassa–Holm equation,* Japan J. Indust. Appl. Math., Vol. 28(3), 2011, pp. 351–381.

[9] SHASHKOV, M., *Conservative finite-difference methods on general grids,* CRC Press, Boca Raton, FL, 1996.

[10] YOUNES, L., *Shapes and Diffeomorphisms,* Springer-Verlag Berlin Heidelberg, 2010.

3

Material not included in the papers

3.1 Excluded from Paper B: Connection between the discrete solutions to primal and dual space-time formulation

We collect in this section some results that for reason of space or completeness have not been included in Paper B, but that are of interest to better grasp the connection between the method we presented and the "standard method" originated by a discretization based on the first space-time formulation. We adopt the notation introduced in Paper B; the results presented below would virtually fit at the end of Section 3 in the paper, complementing the subsection *The roles of* U_1 *and* U_2. In order to further clarify the role of each component of the solution and associate them to a degree of freedom, we need to bridge the gap between primal and weak discrete formulation, in the same way we did for the continuous formulation in Paper B. We start by considering the original problem and we assume for simplicity that A does not depend on time:

$$\begin{aligned} &\dot{u} + Au = f, \\ &u(0) = \xi \end{aligned} \tag{3.1}$$

The primal space-time formulation leads to the following discretization:

$$\begin{aligned} &W \in \mathcal{X}^n_{h,k,q+1}, \qquad Y \in \mathcal{X}^n_{h,k,q}, \\ &\int_0^{t_n} {}_{V^*}\langle \dot{W}(s) + AW(s), Y(s)\rangle_V \,\mathrm{d}s = \int_0^{t_n} {}_{V^*}\langle f(s), Y(s)\rangle_V \,\mathrm{d}s, \\ &W(0) = \xi. \end{aligned} \tag{3.2}$$

The weak space-time formulation gives instead:

$$\begin{aligned} &U_1 \in \mathcal{Y}^n_{h,k,q},\ U_2 \in V_h \qquad X \in \mathcal{X}^n_{h,k,q+1}, \\ &\int_0^{t_n} {}_{V}\langle U_1(s), -\dot{X}(s) + A^* X(s)\rangle_{V^*} \,\mathrm{d}s + \langle U_2, X(t_n)\rangle_H \\ &\quad = \int_0^{t_n} {}_{V^*}\langle f(s), X(s)\rangle_V \,\mathrm{d}s + \langle \xi, X(0)\rangle_H, \end{aligned} \tag{3.3}$$

where, in particular, we can split the scheme as in Paper B, to get values $U_2^{(i)}$ at each time point t_i. Problem (3.2) and (3.3) have both a unique solution. The next theorem states that the discrete solution to the primal and to the dual formulation of (1.40) are in a certain sense the same, whenever $f = 0$.

Theorem 8. *If $W \in \mathcal{X}^n_{h,k,q+1}$ and $U_1 \in \mathcal{Y}^n_{h,k,q}$, $U_2 \in V_h$ are respectively solutions to* (3.2) *and* (3.3) *with $f = 0$, then:*

$$\begin{aligned} U_2^{(n)} &= W(t_n), \\ U_1 &= \Pi^{(q)} W. \end{aligned} \tag{3.4}$$

In particular, if the weak-space time solution is obtained with the splitting proposed in Paper B, it also holds that

$$U_2^{(i)} = W(t_i), \quad i = 1, \dots, n, \tag{3.5}$$

Proof. We consider the pair $\Big(\Pi^{(q)} W, W(t_n)\Big)$, for a given t_n, where $W \in \mathcal{X}^n_{h,k,q+1}$ is solution to the primal formulation:

$$\int_0^{t_n} {}_{V^*}\langle \dot{W}(s) + AW(s), Y(s) \rangle_V \, \mathrm{d}s = 0 \qquad \forall Y \in \mathcal{Y}^n_{h,k,q}. \tag{3.6}$$

This can be rewritten as

$$\int_0^{t_n} {}_{V^*}\langle \dot{W}(s) + AW(s), \Pi^{(q)} X(s) \rangle_V \, \mathrm{d}s = 0, \qquad \forall X \in \mathcal{X}^n_{h,k,q+1}. \tag{3.7}$$

We can integrate by part the first term, obtaining:

$$\begin{aligned} &\sum_{i=0}^{n-1} \int_{I_i} {}_{V^*}\langle \dot{W}(s), \Pi_i^{(q)} X(s) \rangle_V \, \mathrm{d}s = \sum_{i=0}^{n-1} \int_{I_i} {}_{V^*}\langle \dot{W}(s), X(s) \rangle_V \, \mathrm{d}s \\ &= \sum_{i=0}^{n-1} \Big(\int_{I_i} {}_{V}\langle W(s), \dot{X}(s) \rangle_{V^*} \, \mathrm{d}s + \\ &\qquad \langle W(t_i), X(t_i) \rangle_H - \langle W(t_{i-1}), X(t_{i-1}) \rangle_H \Big). \end{aligned} \tag{3.8}$$

The previous expression reduces to:

$$\begin{aligned} &= \sum_{i=0}^{n-1} \Big(\int_{I_i} {}_{V}\langle \Pi_i^{(q)} W(s), \dot{X}(s) \rangle_{V^*} \, \mathrm{d}s \Big) + \langle W(t_n), X(t_n) \rangle_H - \langle \xi, X(0) \rangle_H \\ &= \int_0^{t_n} {}_{V}\langle \Pi^{(q)} W(s), \dot{X}(s) \rangle_{V^*} \, \mathrm{d}s + \langle W(t_1), X(t_1) \rangle_H - \langle \xi, X(0) \rangle_H. \end{aligned} \tag{3.9}$$

Thus, for any $X \in \mathcal{X}^n_{h,k,q+1}$ we have:

$$\begin{aligned} &\int_0^{t_n} {}_{V}\langle \Pi^{(q)} W(s), -\dot{X}(s) + A^* X(s) \rangle_{V^*} \, \mathrm{d}s + \langle W(t_n), X(t_n) \rangle_H \\ &\qquad = \langle \xi, X(0) \rangle_H. \end{aligned} \tag{3.10}$$

Since the solution to such a problem is unique, the first part of the claim follows. In particular, since t_n is arbitrary, the second part of the claim also holds true. □

The previous result extends naturally even when $f \neq 0$; now the two solutions are no longer the same but differ up to a term which is proportional to the interpolation error of f.

Theorem 9. *Under the same assumptions of Theorem 8, but with $f \neq 0$, such that $f^{(\gamma)} \in L^2([0,t_n];V)$ for some $\gamma \in \mathbf{N}$, then*

$$\begin{aligned} \|U_1 - \Pi^{(q)} W\|_{L^2((0,t_n);V)} + \|U_2^{(n)} - W(t_n)\|_H \\ \leq C k^{\theta+1} \|f^{(\theta)}\|_{L^2((0,t_n);V)}, \end{aligned} \tag{3.11}$$

where $\theta := \min\{q+1, \gamma\}$. In particular, if the weak-space time solution is obtained with the splitting proposed in Paper B, it also holds that

$$\begin{aligned} \|U_1 - \Pi^{(q)} W\|_{L^2((0,t_n);V)} + \max_{i=1,\dots,N} \|U_2^{(i)} - W(t_i)\|_H \\ \leq C k^{\theta+1} \|f^{(\theta)}\|_{L^2((0,t_n);V)}. \end{aligned} \tag{3.12}$$

Proof. The crucial difference with the previous proof is that the non-zero right-hand side gives us:

$$\begin{aligned} \mathscr{B}_n^*((U_1 - \Pi^{(q)} W, U_2^{(n)} - W(t_n), X) \\ = \int_0^{t_n} {}_{V^*}\langle f, X\rangle_V \, \mathrm{d}s - \int_0^{t_n} {}_{V^*}\langle f, \Pi^{(q)} X\rangle_V \, \mathrm{d}s. \end{aligned} \tag{3.13}$$

Since the right-hand side is no longer zero, we cannot argue that the two solutions coincides by uniqueness; however, we can bound the norm of the solution in terms of the data:

$$\|U_1 - \Pi^{(q)} W\|^2_{L^2((0,t_n);V)} + \|U_2^{(n)} - W(t_n)\|_H^2 \leq C \|\widetilde{\mathscr{F}}\|^2_{(\mathcal{X}_k)^*}, \tag{3.14}$$

where

$$\widetilde{\mathscr{F}}_n := \int_0^{t_n} {}_{V^*}\langle (I - \Pi^{(q)}) f(s), X(s)\rangle_V \, \mathrm{d}s. \tag{3.15}$$

Here we can use either $|\cdot|_{\mathcal{X}_k}$ or $\|\cdot\|_{\mathcal{X}_k}$ to compute the dual norm. Since constants play no role in this analysis, we decide to use the latter. We have that:

$$\begin{aligned}\left|\widetilde{\mathscr{F}_n}(X)\right| &= \left|\int_0^{t_n} {}_{V^*}\langle (I-\Pi^{(q)})f(s), X(s)\rangle_V \,\mathrm{d}s\right| \\ &= \left|\int_0^{t_n} {}_{V^*}\langle (I-\Pi^{(q)})f(s), (I-\Pi^{(q)})X(s)\rangle_V \,\mathrm{d}s\right| \\ &\leq \Big(\int_0^{t_n} \|(I-\Pi^{(q)})f(s)\|_V^2 \,\mathrm{d}s\Big)^{\frac{1}{2}} \Big(\int_0^{t_n} \|(I-\Pi^{(q)})X(s)\|_{V^*}^2 \,\mathrm{d}s\Big)^{\frac{1}{2}} \\ &\leq C\Big(\sum_{i=0}^{N-1} k_i^{2\theta} \int_{I_i} \|f^{(\theta)}(s)\|_V^2 \,\mathrm{d}s\Big)^{\frac{1}{2}} \Big(\sum_{i=0}^{N-1} k_i^2 \int_{I_i} \|\dot{X}(s)\|_{V^*}^2 \,\mathrm{d}s\Big)^{\frac{1}{2}}.\end{aligned} \tag{3.16}$$

This gives us

$$\|\widetilde{\mathscr{F}_n}\|_{(\mathcal{X}_k,\|\cdot\|_{\mathcal{X}_k})^*} \leq Ck^{\theta+1}\|f^{(\theta)}\|_{L^2((0,t_n);V)}. \tag{3.17}$$

It follows that

$$\begin{aligned}\|U_1 - \Pi^{(q)}W\|_{L^2((0,t_n);V)} + \|U_2^{(n)} - W(t_n)\|_H \\ \leq Ck^{\theta+1}\|f^{(\theta)}\|_{L^2((0,t_n);V)},\end{aligned} \tag{3.18}$$

which proves the first part of the claim. Since t_n is arbitrary, the second part follows in the same way. □

3.2 Excluded from Paper D: A possible fourth scheme

We devote this section to the presentation of a possible fourth scheme to solve the EPDiff, which has been omitted in the final draft. We refer to the appended paper for the notation and the missing definitions, in order to avoid unnecessary repetitions. A logical choice for a possible fourth scheme would be based on the following definition of discrete energy, which combines the definitions used in scheme 2 and 3:

$$H_{k,j}^{(n+\frac{1}{2})} = \frac{M_{1;k,j}^{(n+\frac{1}{2})} U_{1;k,j}^{(n+\frac{1}{2})} + M_{2;k,j}^{(n+\frac{1}{2})} U_{2;k,j}^{(n+\frac{1}{2})}}{2}. \tag{3.19}$$

This can be written explicitly as:

$$\begin{aligned} & H_{k,j}^{(n+\frac{1}{2})} \\ & = \frac{M_{1;k,j}^{(n+1)} U_{1;k,j}^{(n+1)} + M_{1;k,j}^{(n)} U_{1;k,j}^{(n)} + M_{1;k,j}^{(n+1)} U_{1;k,j}^{(n)} + M_{1;k,j}^{(n)} U_{1;k,j}^{(n+1)}}{8} \\ & + \frac{M_{2;k,j}^{(n+1)} U_{2;k,j}^{(n+1)} + M_{2;k,j}^{(n)} U_{2;k,j}^{(n)} + M_{2;k,j}^{(n+1)} U_{2;k,j}^{(n)} + M_{2;k,j}^{(n)} U_{2;k,j}^{(n+1)}}{8}. \end{aligned} \tag{3.20}$$

In particular, if we denote respectively by ${}^2H_{k,j}^{(n+\frac{1}{2})}$ and ${}^3H_{k,j}^{(n+\frac{1}{2})}$ the discrete energies for scheme 2 and scheme 3, it holds that:

$$H_{k,j}^{(n+\frac{1}{2})} = \frac{1}{2}\Big({}^2H_{k,j}^{(n+\frac{1}{2})} + {}^3H_{k,j}^{(n+\frac{1}{2})}\Big). \tag{3.21}$$

It follows that

$$\begin{aligned} & \frac{1}{\Delta t} \sum_{j=0}^{\mathcal{J}-1} \sum_{k=0}^{\mathcal{K}-1} (H_{k,j}^{(n+\frac{1}{2})} - H_{k,j}^{(n-\frac{1}{2})}) \Delta x \Delta y \\ & = \frac{1}{2\Delta t} \sum_{j=0}^{\mathcal{J}-1} \sum_{k=0}^{\mathcal{K}-1} \Big[({}^2H_{k,j}^{(n+\frac{1}{2})} - {}^2H_{k,j}^{(n-\frac{1}{2})}) + ({}^3H_{k,j}^{(n+\frac{1}{2})} - {}^3H_{k,j}^{(n-\frac{1}{2})}) \Big] \Delta x \Delta y, \end{aligned} \tag{3.22}$$

which, in turn, reduces to

$$\begin{aligned} & = \sum_{j=0}^{\mathcal{J}-1} \sum_{k=0}^{\mathcal{K}-1} \Big(\frac{M_{1;k,j}^{(n+1)} - M_{1;k,j}^{(n-1)}}{2\Delta t} \frac{U_{1;k,j}^{(n+1)} + 2U_{1;k,j}^{(n)} + U_{1;k,j}^{(n-1)}}{4} \\ & \qquad + \frac{M_{2;k,j}^{(n+1)} - M_{2;k,j}^{(n-1)}}{2\Delta t} \frac{U_{2;k,j}^{(n+1)} + 2U_{2;k,j}^{(n)} + U_{2;k,j}^{(n-1)}}{4} \Big) \Delta x \Delta y. \end{aligned} \tag{3.23}$$

We have the following discrete variational derivative which approximates the continuous one by

$$\frac{\delta H}{\delta(\mathbf{M}^{(n+1)},\mathbf{M}^{(n)},\mathbf{M}^{(n-1)})}_{k,j} := \begin{bmatrix} \frac{U_{1;k,j}^{(n+1)}+2U_{1;k,j}^{(n)}+U_{1;k,j}^{(n-1)}}{4} \\ \frac{U_{2;k,j}^{(n+1)}+2U_{2;k,j}^{(n)}+U_{2;k,j}^{(n-1)}}{4} \end{bmatrix}. \tag{3.24}$$

The scheme becomes

$$\frac{\underline{M}_{k,j}^{(n+1)} - \underline{M}_{k,j}^{(n-1)}}{2\Delta t} = -\widetilde{\Gamma}_{\mathbf{m}}^{(n)} \frac{\delta H}{\delta(\mathbf{M}^{(n+1)},\mathbf{M}^{(n)},\mathbf{M}^{(n-1)})}_{k,j}, \tag{3.25}$$

where $\widetilde{\Gamma}_{\mathbf{m}}^{(n)}$ is as in Paper D. The following result of conservation holds:

Theorem 10. *Under the discrete periodic boundary conditions, the numerical solution produced by Scheme 4 conserves the following invariant, for each* $n = 1, 2, \ldots$*:*

$$\sum_{j=0}^{\mathcal{J}-1}\sum_{k=0}^{\mathcal{K}-1} H_{k,j}^{(n+\frac{1}{2})}\Delta x\Delta y = \sum_{j=0}^{\mathcal{J}-1}\sum_{k=0}^{\mathcal{K}-1} H_{k,j}^{(\frac{1}{2})}\Delta x\Delta y. \tag{3.26}$$

This scheme does not conserve momenta and is computationally more expensive than the two schemes upon which it is based, and has therefore been omitted in the analysis presented in Paper D. However, it is worth noticing that the energy preserved by this scheme is "the real numerical energy" defined at $n + \frac{1}{2}$, while Scheme 2 and Scheme 3 only preserve part of it, or, otherwise stated, an alternative definition of numerical energy ad $n + \frac{1}{2}$.

We append the numerical results obtained for Scheme 4, while testing it for the same benchmark problem used to test the other schemes. We present a comparison with Scheme 1, which makes more visible the advantages and the drawbacks of this scheme.

Table 3.1: *Conservation of the discrete energy*

	Total Variation	$\|\cdot\|_\infty$
Scheme 1	$1.8529 \cdot 10^{-8}$	$1.8529 \cdot 10^{-8}$
Scheme 4	$6.0348 \cdot 10^{-10}$	$1.9753 \cdot 10^{-11}$

Table 3.2: *Conservation of the linear momentum in the x-direction*

	Total Variation	$\|\cdot\|_\infty$
Scheme 1	$3.1130 \cdot 10^{-9}$	$3.1127 \cdot 10^{-9}$
Scheme 4	3.1008	0.0088

Table 3.3: *Conservation of the linear momentum in the y-direction*

	Total Variation	$\|\cdot\|_\infty$
Scheme 1	$2.6557 \cdot 10^{-16}$	$8.0264 \cdot 10^{-17}$
Scheme 4	$2.9119 \cdot 10^{-09}$	$1.7376 \cdot 10^{-10}$