Differential Equations with Constraints

Verdier, Olivier

2009

Differential Equations with Constraints

Olivier Verdier

Lund University

Faculty of Engineering
Centre for Mathematical Sciences
Mathematics

Mathematics
Centre for Mathematical Sciences
Lund University
Box 118
SE-221 00 Lund
Sweden

http://www.maths.lth.se/

Doctoral Theses in Mathematical Sciences 2009:4
ISSN 1404-0034

ISBN 978-91-628-7812-2
LUTFMA-1037-2009

Printed in Sweden by MediaTryck, Lund 2009

Organization	Document name
Centre for Mathematical Sciences Lund Institute of Technology Mathematics Box 118 SE-221 00 LUND	DOCTORATE THESIS IN MATHEMATICAL SCIENCES
	Date of issue April 2009
	Document Number LUTFMA-1037-2009
Author(s) Olivier Verdier	Supervisor Magnus Fontes
	Sponsoring organisation Lund University

Title and subtitle

Differential Equations with Constraints

Abstract

We study various differential equations subject to constraints. In the first part we study a partial differential equation, Burgers equation, subject to time-periodicity constraint. The forcing term is time-periodic and may be highly irregular. We prove an existence and uniqueness result which in a sense is optimal since we show that the operator corresponding to the Burgers equation is a diffeomorphism from a functional space to its dual.

In the second part we study general ordinary differential equations subject to general constraints. We first describe precisely what the index is. Subsequently we investigate the particular case of linear ordinary differential equation and derive a new normal form. We show that it is characterized by defect indices and we show the relation with the Kronecker normal form.

Key words

Classifiction system and/or index terms (if any)

Supplementary bibliography information

ISSN and key title	ISBN
1404-0034	978-91-628-7812-2

Language	Number of pages	Recipient's notes
English	127	

Security classification

Preface

This thesis is about constrained differential equations. It comprises two independent parts.

In the first part, we investigate Burgers equation with a periodic and highly irregular forcing term. We prove that notwithstanding the irregularity, there exists only one periodic solution. Besides, our result is optimal in the sense that the function spaces of the solution and that of the forcing term are dual to each others. The work of this first part is a joint work with M. Fontes.

The second part is devoted to the geometric study of ordinary differential equations with constraints. We will first consider ordinary differential equations in the most general terms, and give a precise definition of their index. We will also show how similarly stated differential equations with constraints may be related by pull-back. Later on, we will dive into the special case of linear differential equations, and exhibit a new normal form, which turns out to provide a new proof for the well-known Kronecker canonical form.

Aknowledgments

My coming to Sweden had nothing to do with mathematics. I was not prepared to start a Ph.D. thesis here. However, after my arrival in Malmö I felt a compelling urge to come by the university with the foolish hope that I might be of any use there.

I first met *Gustaf Söderlind*, at the time head of the department of numerical analysis. Although he couldn't find any funding for research projects, he generously offered me to make myself at home and to take one of the available desks of "room 139". That is how I gradually became one of the "hang arounds" at the department of numerical analysis. He also sent me further to the director of studies.

The director of studies was at that time *Claus Führer*, who offered me to teach some Matlab courses, in want of some more substantial financial support. Despite the colossal fiasco of my teaching Matlab in the Chemistry department[(a)], we stayed friends and he suggested that I should apply to some of the vacant Ph.D. positions. I first dismissed such an idea, and did not even apply to the first opened position. Claus then suggested me to talk to that professor who had recently come back from France, and had a new interesting course in mathematics.

This is how I met *Magnus Fontes*. I first started to take his course, and soon realised that I had a fairly good command of the material, having taken a similar course before in the rather highly paced "classes préparatoires" in France. Magnus then generously offered me to act as a teaching assistent in this new course. I was thrilled and made use of all the liberty that Magnus gave me. Working with Magnus turned out to be a treat. A new Ph.D. position opened, with Magnus as an advisor, and this time, I applied[(b)].

The reader would have understood by now that without the kindness and generosity of Gustaf, Claus and Magnus, I wouldn't be the proud author of this work.

As my supervisor, Magnus deserves a special mention. Along all these years, he never failed to be supporting and encouraging. Despite a heavy burden of

[a] It was no easy task, but it is only fair to say that the French education system had not made this any easier for me either

[b] ... and got refused, but the position after that one went along well.

administrative duties he never denied me the time for some mathematical question or a friendly chat. Moreover, he patiently coped with my compulsive habit of doing the exact opposite of what I am told to do, which is no small feat.

Besides my regular collaboration with Magnus, I stayed in contact with Gustaf and Claus during the whole period of my position. I came to teach with all three[c], all exceptional teachers, albeit with vastly different teaching styles. In all three cases, "Mathematical structures", "Numerical methods for differential equations", and "Scientific computing in Python", it was an extraordinary experience.

The last course was slightly different from the two others though. Despite being given the chance to contribute with many changes in the first two cases, the third course was an opportunity for me to practically create a course of my own. This also meant the reinforcement of a friendship with Claus[d] that has never receded since.

This collaboration with Claus came also, by chance, to influence my research. After having worked on the Burgers equation for two years, I had felt the urge to pit my limited skills against other problems. Rather by accident, I took up the study of the difficult problem of observability, which turned out to be too algebraically flavoured for my taste. My perspectives of research were rather grim until I suggested to Claus that we wrote an article together[e]. It then struck me that the observability problem and the theory of differential algebraic equations had in fact very similar structures, the latter being much more amenable than the former. This is what led to what has now become the second part of this thesis.

Along the way, many people have been helpful, encouraging and friendly. James Hakim and his coworkers have always been open for my odd requirements regarding the computer equipments. Anki has a never failing efficiency that most administrations would not even dream of being possible. The librarian Kerstin Brandt, and now Mikael Abrahamsson, have never complained at the high turnover of my book borrowing[f] and always kindly (and swiftly) complied at my article requests.

Many others made my stay a pleasant one, by discussions, scientific or not, and sometimes gatherings around a beer. To name but a few: Anders Sjöström, Klas Modin, Klas Josephson, Thomas Vallier, Charlotte Soneson, and I am forgetting many others.

I am also indebted to Claus and Magnus who proof-read this manuscript.

[c] as it happened, I practically only taught with Gustaf, Claus and Magnus during my time at the department.

[d] sponsored by the restaurant "Gladimat"

[e] which never happened...

[f] before the new library was built, Kerstin would even half-jokingly praise my hoarding of books in my office for it left more room in the library...

* * *

Finally, I would not be the person I am today without my family. My father gave me a taste for science that never left me, and I believe that I inherited only a modest fraction of his impressive intellectual abilities. Because life is not all science and knowledge, the role of my mother, who gave all her three sons love and self-confidence, has been and still is, invaluable. Lastly, my brothers David and Julien are my alter egos, without whom I dare not imagine what life would be like.

My dearest love goes to Linda, for having shared her life all these years with an absent-minded, late working, wannabe mathematician.

Olivier Verdier
Lund, Sweden
May 2009

Contents

Part I

Time-Periodic Constraints

1

PRELIMINARIES

1.1 Introduction

The study of the Burgers equation has a long history starting with the seminal papers by Burgers [Bur50], Cole [Col51] and Hopf [Hop50] where the Cole-Hopf transformation was introduced. The Cole-Hopf transformation transforms the homogeneous Burgers equation into the heat equation.

More recently there have been several articles dealing with the forced Burgers equation:

$$u_t - \nu u_{xx} + u u_x = f \tag{1.1}$$

The vast majority treats the initial value problem in time with homogeneous Dirichlet or periodic space boundary conditions (see for instance [KL89]).

Only recently the question of the time-periodic forced Burgers equation has been tackled. In most cases [JKM99; E99] the authors are chiefly interested in the inviscid limit (the limit when the viscosity ν tends to zero). The forcing term is usually chosen to take a particular form, for example a sum of products of white noises in time and smooth functions in space [E99; Sin91]. In [FS05] the space domain is the half line and the Dirichlet boundary conditions are time periodic and analytic.

The closest related work to ours is that of Jauslin, Kreiss and Moser [JKM99] in which the authors show existence and uniqueness of a space and time periodic solution of the Burgers equation for a space and time periodic forcing term which is smooth.

In this thesis we generalise these results and prove that the Burgers operator coming from the Burgers equation is in fact a diffeomorphism between appropriate time periodic anisotropic Sobolev spaces.

More precisely our main result (Theorem 1) shows that given a time periodic forcing term in $\mathrm{H}^{\left(-\frac{1}{2},-1\right)}$ we have existence and uniqueness of a time periodic solution in $\mathrm{H}_0^{\left(\frac{1}{2},1\right)}$. Furthermore we are able to prove smooth dependence on the forcing term.

To prove this result we will use a method similar to that of [FS04] which makes extensive use of anisotropic Sobolev spaces. We will also use a modification of the Cole-Hopf transformation to prove uniqueness of the solution.

We prove those results with homogeneous Dirichlet spatial boundary conditions but the results and the proof can be extended to inhomogeneous boundary conditions as well as spatial periodic boundary conditions.

The results of this part were published in [FV07].

1.2 Fractional Calculus

In this section we recall some well known facts and fix some general notations.

Fourier Analysis

We denote the *one dimensional torus* $\mathbb{T}$ by:

$$\mathbb{T} = \mathbb{R}/\mathbb{Z}$$

Let H denote a *complex Hilbert space*, then the space of smooth *Hilbert space-valued periodic test functions* $\mathcal{D}(\mathbb{T}, H)$ is denoted by:

$$\mathcal{D}(\mathbb{T}, H) = \mathcal{C}^\infty(\mathbb{T}, H)$$

endowed with the usual topology of test functions. Any test function $\varphi \in \mathcal{D}(\mathbb{T}, H)$ can be developed in a Fourier series:

$$\varphi = \sum_{n\in\mathbb{Z}} \varphi_n \cdot \mathrm{e}^{\mathrm{i}2\pi nt}$$

where $\varphi_n \in H$ is defined by:

$$\varphi_n = \int_{\mathbb{T}} \varphi(t) \cdot \mathrm{e}^{-\mathrm{i}2\pi nt} \mathrm{d}t$$

We denote by $\mathcal{D}'(\mathbb{T}, H^*)$ the space of continuous linear functionals on $\mathcal{D}(\mathbb{T}, H)$. It is naturally isomorphic to the Hilbert-space valued 1-periodic distributions on $\mathbb{R}$. For any periodic distribution $u \in \mathcal{D}'(\mathbb{T}, H^*)$ we then have:

$$u = \sum_{n\in\mathbb{Z}} u_n \cdot \mathrm{e}^{\mathrm{i}2\pi nt}$$

where $u_n \in H^*$ is defined by

$$\forall \psi \in H, \quad \langle u_n, \psi \rangle_{H^*,H} := \left\langle u, \mathrm{e}^{-\mathrm{i}2\pi nt} \cdot \psi \right\rangle_{\mathcal{D}',\mathcal{D}}$$

Fractional Derivatives

For any positive real number s we may define the ***fractional derivative of order*** s D^s defined on $\mathcal{D}'(\mathbb{T}, H^*)$ as follows:

$$\mathrm{D}^s u = \sum_{k\in\mathbb{Z}} (2\pi \mathrm{i} k)^s u_k \mathrm{e}^{\mathrm{i}2\pi kt} = \sum_{k\in\mathbb{Z}} |2\pi \mathrm{i} k|^s \, \mathrm{e}^{\mathrm{i}\,\mathrm{sgn}(k) s\frac{\pi}{2}} u_k \mathrm{e}^{\mathrm{i}2\pi kt}$$

where we have used the principal branch of the logarithm. The sign function sgn is defined as follows:

$$\mathrm{sgn}(k) := \begin{cases} \frac{k}{|k|} & \text{if } k \neq 0 \\ 0 & \text{if } k = 0 \end{cases}$$

For $s = 0$ we define $\mathrm{D}^0 = \mathbb{I}$. D^1 coincides with the usual differentiation operator on $\mathcal{D}'(\mathbb{T}, H^*)$. The familiar composition property also holds: $\mathrm{D}^s \circ \mathrm{D}^t = \mathrm{D}^{s+t}$ for any $t, s \geq 0$.

The *adjoint operator* D^s_* of D^s is defined by using the conjugate of the multiplier of D^s:

$$\mathrm{D}^s_* u = \sum_{k\in\mathbb{Z}} |2\pi \mathrm{i} k|^s \, \mathrm{e}^{-\mathrm{i}\,\mathrm{sgn}(k) s\frac{\pi}{2}} u_k \mathrm{e}^{\mathrm{i}2\pi kt}$$

D^s and D^s_* are adjoints in the sense that for any $u \in \mathcal{D}'(\mathbb{T}, H^*)$ and $\varphi \in \mathcal{D}(\mathbb{T}, H)$:

$$\langle \mathrm{D}^s u, \varphi \rangle = \langle u, \mathrm{D}^s_* \varphi \rangle$$

and similarly:

$$\langle \mathrm{D}^s_* u, \varphi \rangle = \langle u, \mathrm{D}^s \varphi \rangle$$

Hilbert Transform

The *Hilbert transform* $\mathcal{H}$ is defined using the multiplier $-\mathrm{i}\,\mathrm{sgn}\, k$. For $u \in \mathcal{D}'(\mathbb{T}, H^*)$ let

$$\mathcal{H} u = \sum_{k\in\mathbb{Z}} -\mathrm{i}\,\mathrm{sgn}\, k \, u_k \, \mathrm{e}^{\mathrm{i}2\pi kt}$$

Simple computations then give the following properties for $\mathrm{D}^{\frac{1}{2}}_*$:

$$\mathrm{D}^{\frac{1}{2}}_* = \mathrm{D}^{\frac{1}{2}} \circ \mathcal{H} = \mathcal{H} \circ \mathrm{D}^{\frac{1}{2}}$$

Notice that if H is a function space then $\mathcal{H}$ maps real functions to real functions. The following properties will be useful in the sequel:

$$\begin{gathered} \forall u \in \mathrm{H}^{(\frac{1}{2})}(\mathbb{T}, H) \quad \left(\mathrm{D}^{\frac{1}{2}} u, \mathrm{D}^{\frac{1}{2}}_* \mathcal{H} u\right)_{\mathrm{L}^2(\mathbb{T},H)} = -\left\|\mathrm{D}^{\frac{1}{2}} u\right\|^2_{\mathrm{L}^2(\mathbb{T},H)} \\ \forall u \in \mathrm{L}^2(\mathbb{T}\times\mathcal{I}) \quad \Re\Big((u, \mathcal{H}(u))_{\mathrm{L}^2(\mathbb{T}\times\mathcal{I})}\Big) = 0 \end{gathered} \tag{1.2}$$

where $\Re$ denotes the real part of the expression.

1.3 Sobolev Spaces

Fractional Sobolev Spaces

We define the fractional Sobolev spaces $\mathrm{H}^{(s)}(\mathbb{T}, H)$ in the following manner, for any $s \in \mathbb{R}$:

$$\mathrm{H}^{(s)}(\mathbb{T}, H) = \left\{ u \in \mathcal{D}'(\mathbb{T}, H^*); \quad \sum_{k \in \mathbb{Z}} \left|1 + k^2\right|^s \|u_k\|_H^2 < \infty \right\}$$

Of course $\mathrm{H}^{(0)}(\mathbb{T}, H) = \mathrm{L}^2(\mathbb{T}, H)$. When $s \geq 0$ then for an $u \in \mathrm{L}^2(\mathbb{T}, H)$: $u \in \mathrm{H}^{(s)}(\mathbb{T}, H) \iff \mathrm{D}^s u \in \mathrm{L}^2(\mathbb{T}, H)$. Moreover $\mathrm{H}^{(s)}(\mathbb{T}, H)$ is then a Hilbert space with the following scalar product:

$$(u, v) := (u, v)_{\mathrm{L}^2(\mathbb{T}, H)} + (\mathrm{D}^s u, \mathrm{D}^s v)_{\mathrm{L}^2(\mathbb{T}, H)}$$

The following classical result holds: $\left(\mathrm{H}^{(s)}(\mathbb{T}, H)\right)^* = \mathrm{H}^{(-s)}(\mathbb{T}, H^*)$.

Anisotropic Fractional Sobolev Spaces

Let $\mathcal{I}$ be an interval in $\mathbb{R}$ and $s \geq 0$. Let $\mathrm{H}^{(s)}(\mathcal{I})$ denote the usual fractional Sobolev space of real-valued s-times differentiable functions on $\mathcal{I}$. $\mathrm{H}_0^{(s)}(\mathcal{I})$ is the closure of $\mathcal{D}(\mathcal{I})$ in $\mathrm{H}^{(s)}(\mathcal{I})$. In that case we have $\left(\mathrm{H}_0^{(s)}(\mathcal{I})\right)^* = \mathrm{H}^{(-s)}(\mathcal{I})$. We will also use the following notations: for α, β nonnegative real numbers we define $\mathrm{H}^{(\alpha)(\beta)}(\mathbb{T} \times \mathcal{I})$:

$$\mathrm{H}^{(\alpha)(\beta)}(\mathbb{T} \times \mathcal{I}) := \mathrm{H}^{(\alpha)}(\mathbb{T}, \mathrm{H}^{(\beta)}(\mathcal{I}))$$

and $\mathrm{H}^{(\alpha,\beta)}(\mathbb{T} \times \mathcal{I})$:

$$\mathrm{H}^{(\alpha,\beta)}(\mathbb{T} \times \mathcal{I}) = \mathrm{H}^{(\alpha)(0)}(\mathbb{T} \times \mathcal{I}) \cap \mathrm{H}^{(0)(\beta)}(\mathbb{T} \times \mathcal{I})$$

We also introduce $\mathrm{H}_0^{(\alpha,\beta)}(\mathbb{T} \times \mathcal{I})$ as the closure of $\mathcal{D}(\mathbb{T} \times \mathcal{I})$ in $\mathrm{H}^{(\alpha,\beta)}(\mathbb{T} \times \mathcal{I})$. It is clear that $\mathrm{H}_0^{(\alpha,\beta)}(\mathbb{T} \times \mathcal{I}) = \mathrm{H}^{(\alpha)(0)}(\mathbb{T} \times \mathcal{I}) \cap \mathrm{L}^2(\mathbb{T}, \mathrm{H}_0^{(\beta)}(\mathcal{I}))$. Duals of such spaces are denoted as $\mathrm{H}^{(-\alpha,-\beta)}(\mathbb{T} \times \mathcal{I})$:

$$\begin{aligned} \mathrm{H}^{(-\alpha,-\beta)}(\mathbb{T} \times \mathcal{I}) := \left(\mathrm{H}_0^{(\alpha,\beta)}(\mathbb{T} \times \mathcal{I})\right)^* &= \mathrm{H}^{(-\alpha)}(\mathbb{T}, \mathrm{L}^2(\mathcal{I})) + \mathrm{L}^2(\mathbb{T}, \mathrm{H}^{(-\beta)}(\mathcal{I})) \\ &= \mathrm{H}^{(-\alpha)(0)}(\mathbb{T} \times \mathcal{I}) + \mathrm{H}^{(0)(-\beta)}(\mathbb{T} \times \mathcal{I}) \end{aligned}$$

1.4 Interpolation and regularity

If $s_k(\xi)$ is the Fourier transform $s_k(\xi) = \hat{u}(k,\xi)$ of a distribution u defined on $\mathbb{T}\times\mathbb{R}$, we have the following Hölder inequality for any $\theta \in [0,1]$:

$$\int_{\mathbb{R}} \sum_{k\in\mathbb{Z}} |k|^{2\alpha(1-\theta)} |\xi|^{2\beta\theta} |s_k(\xi)|^2 \,\mathrm{d}\xi \leq \left(\int_{\mathbb{R}} \sum_{k\in\mathbb{Z}} |k|^{2\alpha} |s_k(\xi)|^2 \,\mathrm{d}\xi\right)^{1-\theta} \left(\int_{\mathbb{R}} \sum_{k\in\mathbb{Z}} |\xi|^{2\beta} |s_k(\xi)|^2 \,\mathrm{d}\xi\right)^{\theta}$$

From this Hölder inequality we deduce

$$\mathrm{H}^{(\alpha,\beta)}(\mathbb{T}\times\mathbb{R}) \hookrightarrow \mathrm{H}^{((1-\theta)\alpha)}(\mathbb{T}, \mathrm{H}^{(\theta\beta)}(\mathbb{R}))$$

So using an extension operator from $\mathrm{H}^{(\theta\beta)}(\mathcal{I})$ to $\mathrm{H}^{(\theta\beta)}(\mathbb{R})$ one can prove the corresponding inclusion:

$$\mathrm{H}^{(\alpha,\beta)}(\mathbb{T}\times\mathcal{I}) \hookrightarrow \mathrm{H}^{((1-\theta)\alpha)(\theta\beta)}(\mathbb{T}\times\mathcal{I}) \tag{1.3}$$

For $\alpha = 1/2$ and $\beta = 1$ and $\theta = \frac{1}{3}$ we obtain (see Figure 1.1):

$$\mathrm{H}_0^{\left(\frac{1}{2},1\right)}(\mathbb{T}\times\mathcal{I}) \subset \mathrm{H}^{\left(\frac{1}{2},1\right)}(\mathbb{T}\times\mathcal{I}) \subset \mathrm{H}^{(1/3)(1/3)}(\mathbb{T}\times\mathcal{I})$$

Then the vectorial Sobolev inequalities yield:

$$\mathrm{H}_0^{\left(\frac{1}{2},1\right)}(\mathbb{T}\times\mathcal{I}) \subset \mathrm{H}^{(1/3)(1/3)}(\mathbb{T}\times\mathcal{I}) \hookrightarrow \mathrm{L}^4(\mathbb{T}, \mathrm{H}^{\left(\frac{1}{3}\right)}(\mathcal{I})) \hookrightarrow \mathrm{L}^4(\mathbb{T}, \mathrm{L}^4(\mathcal{I})) = \mathrm{L}^4(\mathbb{T}\times\mathcal{I}) \tag{1.4}$$

Here the injection $\mathrm{H}^{(1/3)(1/3)}(\mathbb{T}\times\mathcal{I}) \hookrightarrow \mathrm{L}^4(\mathbb{T}, \mathrm{H}^{(1/3)})$ is compact and thus the injection $\mathrm{H}_0^{\left(\frac{1}{2},1\right)}(\mathbb{T}\times\mathcal{I}) \hookrightarrow \mathrm{L}^4(\mathbb{T}\times\mathcal{I})$ is *compact*.

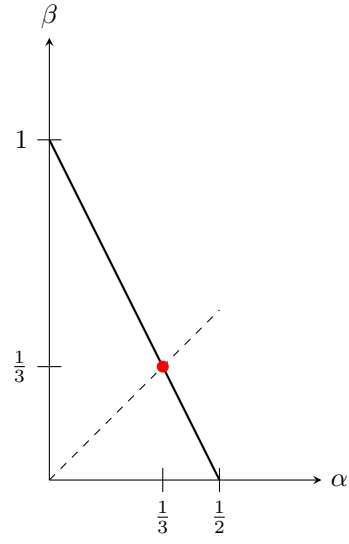


Figure 1.1: $\mathrm{H}_0^{(\frac{1}{2},1)}$ is included in $\mathrm{H}^{(\frac{1}{3})(\frac{1}{3})}$ which is included in L^6 by the usual Sobolev inclusion theorem. In particular, $\mathrm{H}_0^{(\frac{1}{2},1)}$ is included in L^4, so $u \in \mathrm{H}_0^{(\frac{1}{2},1)} \implies u^2 \in \mathrm{L}^2$. As a result the non-linear term of the Burgers equation may be written as $-(u^2, v_x)$ for a test function $v \in \mathrm{H}_0^{(\frac{1}{2},1)}$ since $v \in \mathrm{H}_0^{(\frac{1}{2},1)} \implies v_x \in \mathrm{L}^2$ by definition.

2

Burgers Equation: Estimates

> Dans ce qui suit, j'ai tâché, conformément à l'appel de N. Bourbaki, de substituer toujours les calculs aveugles aux idées lucides d'Euler.
>
> V. Arnol'd, [Arn66], 1966

2.1 Preliminaries

Scalings

For a period $T > 0$, a length $L > 0$, a non zero constant viscosity ν and a *time-periodic forcing term* g, the Burgers equation is formally defined on $\mathbb{R}/T\mathbb{Z} \times (0, L)$ by:

$$u_t + uu_x - \nu u_{xx} = g$$

For $\bar{t} \in \mathbb{T}$, $\bar{x} \in (0, 1)$ we define:

$$\bar{u}(\bar{t}, \bar{x}) := \frac{T}{L} u(\bar{t}T, \bar{x}L)$$

Then $\bar{u}$ is solution of

$$\bar{u}_t + \bar{u}\bar{u}_x - \mu \bar{u}_{xx} = f$$

where f and μ are defined by:

$$f(\bar{t}, \bar{x}) = \frac{T^2}{L} g(\bar{t}T, \bar{x}L)$$
$$\mu = \frac{\nu T}{L^2}$$

$\frac{1}{\mu}$ is often called the *Reynolds number*. The scalings allow us to restrict the study of the Burgers equation to the *normalised domain* $\mathbb{R}/\mathbb{Z} \times (0,1)$.

Notations

In the sequel we will use the following notation:

$$\begin{aligned}
\mathcal{I} &:= (0,1) \\
Q &:= \mathbb{T} \times \mathcal{I} \\
(u,v) &:= \int_Q u \cdot v \mathrm{d}t\mathrm{d}x \\
|u| &:= \sqrt{(u,u)} \\
\mathrm{H}_0^{\left(\frac{1}{2},1\right)} &:= \mathrm{H}_0^{\left(\frac{1}{2},1\right)}(Q) \\
\mathrm{H}^{\left(-\frac{1}{2},-1\right)} &:= (\mathrm{H}_0^{\left(\frac{1}{2},1\right)})^* \\
\mathrm{L}^p &:= \mathrm{L}^p(Q)
\end{aligned}$$

and for $u \in \mathrm{H}_0^{\left(\frac{1}{2},1\right)}$:

$$\begin{aligned}
\|u\| &:= \|u\|_{\mathrm{H}_0^{\left(\frac{1}{2},1\right)}} \\
\widetilde{u} &:= \mathcal{H}\, u \\
u_{\sqrt{t}} &:= \mathrm{D}^{\frac{1}{2}}\, u \in \mathrm{L}^2 \\
u_{\sqrt{t}*} &:= \mathrm{D}_*^{\frac{1}{2}}\, u \in \mathrm{L}^2 \\
u_x &:= \frac{\partial u}{\partial x} \in \mathrm{L}^2 \\
\text{For } f \in \mathrm{H}^{\left(-\frac{1}{2},-1\right)} \quad \langle f,u\rangle &:= \langle f,u\rangle_{\mathrm{H}^{\left(-\frac{1}{2},-1\right)},\mathrm{H}_0^{\left(\frac{1}{2},1\right)}}
\end{aligned}$$

2.2 Properties of the Burgers Operator

Functional Setting

By possibly changing the direction of time we may always assume that μ is a positive real number. We split the Burgers equation in a linear and a non-linear part by means of the two following operators:

Definition 1. We define $\mathcal{L}$ from $\mathrm{H}_0^{\left(\frac{1}{2},1\right)}$ to $\mathrm{H}^{\left(-\frac{1}{2},-1\right)}$ as:

$$\forall v \in \mathrm{H}_0^{\left(\frac{1}{2},1\right)} \quad \langle \mathcal{L}\, u, v\rangle := \left(u_{\sqrt{t}}, v_{\sqrt{t}*}\right) + \mu\,(u_x, v_x)$$

Definition 2. The (nonlinear) operator S from $\mathrm{H}_0^{\left(\frac{1}{2},1\right)}$ to $\mathrm{H}^{\left(-\frac{1}{2},-1\right)}$ is defined as:

$$\forall v \in \mathrm{H}_0^{\left(\frac{1}{2},1\right)} \quad \langle S(u), v\rangle := -\frac{1}{2}\left(u^2, v_x\right)$$

This operator is well-defined since $\mathrm{H}_0^{\left(\frac{1}{2},1\right)} \subset \mathrm{L}^4$ (cf. (1.4)).

Definition 3. The ***Burgers operator*** T defined from $\mathrm{H}_0^{\left(\frac{1}{2},1\right)}$ to $\mathrm{H}^{\left(-\frac{1}{2},-1\right)}$ is defined by:

$$\mathrm{T} = \mathcal{L} + S$$

Given $f \in \mathrm{H}^{\left(-\frac{1}{2},-1\right)}$ the Burgers equation becomes:

$$\mathrm{T}(u) = f \tag{2.1}$$

Main Result

Here is the main result of this part:

Theorem 1. The (nonlinear) Burgers operator T is a diffeomorphism from $\mathrm{H}_0^{\left(\frac{1}{2},1\right)}$ to $\mathrm{H}^{\left(-\frac{1}{2},-1\right)}$, i.e. it is a smooth bijection with smooth inverse.

The main ingredients in the proof of this result are an a priori estimate and the Cole-Hopf transformation. After giving the a priori estimate we will prove existence and then uniqueness. Before that we will make some initial observations.

Some elementary properties

If we denote by $\mathrm{T}'(m)$ the derivative of the operator T at $m \in \mathrm{H}_0^{\left(\frac{1}{2},1\right)}$ then the following holds for any u, v in $\mathrm{H}_0^{\left(\frac{1}{2},1\right)}$:

$$\mathrm{T}(u) - \mathrm{T}(v) = \mathrm{T}'\left(\frac{u+v}{2}\right) \cdot (u - v) \tag{2.2}$$

so T is injective iff $\mathrm{T}'(m)$ is injective for any $m \in \mathrm{H}_0^{\left(\frac{1}{2},1\right)}$.

We notice that by the inclusion (1.4): $S(u) = u \cdot u_x \in \mathrm{L}^{4/3} \hookrightarrow \mathrm{H}^{\left(-\frac{1}{2},-1\right)}$ and the last inclusion is the adjoint of the inclusion (1.4) and is thus compact as well. Since $u \to u^2$ is continuous from L^4 to L^2 we deduce that S is a non-linear compact operator (that is to say it is continuous and sends bounded sets of $\mathrm{H}_0^{\left(\frac{1}{2},1\right)}$ to relatively compact sets of $\mathrm{H}^{\left(-\frac{1}{2},-1\right)}$). Now as a general fact if S is compact and differentiable then $S'(m)$ is a compact linear operator at any point $m \in \mathrm{H}_0^{\left(\frac{1}{2},1\right)}$. We collect these elementary observations in the following Lemma:

Lemma 2.2.1. The nonlinear operator S is *compact* and for any $m \in \mathrm{H}_0^{\left(\frac{1}{2},1\right)}$ the derivative $S'(m)$ at the point m is a linear compact operator.

2.3 An a priori Estimate

We have the following a priori estimate of the solution set:

Theorem 2. Let $f \in \mathrm{H}^{\left(-\frac{1}{2},-1\right)}$. The set

$$\bigcup_{\lambda\in[0,1]} (\mathcal{L}+\lambda S)^{-1}(\{f\})$$

is bounded in $\mathrm{H}_0^{\left(\frac{1}{2},1\right)}$.

We get the following immediate corollary for the case $\lambda = 1$ since $\mathrm{T} = \mathcal{L}+S$:

Corollary 2.3.1. Let $f \in \mathrm{H}^{\left(-\frac{1}{2},-1\right)}$. The set $\mathrm{T}^{-1}(\{f\})$ is bounded in $\mathrm{H}_0^{\left(\frac{1}{2},1\right)}$.

To prove Theorem 2 we will use the same techniques as in [FS04]. In particular, the following Lemma:

Lemma 2.3.1. Given $f \in \mathrm{H}^{\left(-\frac{1}{2},-1\right)}$ and $\varepsilon > 0$ there exists g and h in $\mathrm{L}^2(Q)$ such that for any $\varphi \in \mathrm{H}_0^{\left(\frac{1}{2},1\right)}$ we have:

$$|g| \leq \varepsilon \tag{2.3}$$

$$\langle f, \varphi\rangle = (g, \varphi_{\sqrt{t}*}) - (h, \varphi_x) \qquad \forall \varphi \in \mathrm{H}_0^{\left(\frac{1}{2},1\right)} \tag{2.4}$$

In other words we have

$$f = g_{\sqrt{t}} + h_x$$

in the distribution sense, and g can be taken as small as we want in $\mathrm{L}^2(Q)$.

Proof of Lemma 2.3.1. This follows directly from the fact that $\mathrm{H}^{(0,-1)}$ is a dense subspace of $\mathrm{H}^{\left(-\frac{1}{2},-1\right)}$. Indeed given an $\varepsilon > 0$ there is a $\varphi \in \mathrm{H}^{(0,-1)}$ such that $\|f-\varphi\|_{\mathrm{H}^{\left(-\frac{1}{2},-1\right)}} \leq \varepsilon$. By the Hahn-Banach theorem there exist functions g, h_1 and h_2 in L^2 such that $f - \varphi = g_{\sqrt{t}} + h_{1x}$, $\varphi = h_{2x}$ and $|g| \leq \|f-\varphi\| \leq \varepsilon$. We take $h = h_1 + h_2$ and the Lemma is proved. □

To prove Theorem 2 we will also need the following Gagliardo-Nirenberg type inequality, for which we give an elementary proof for the convenience of the reader:

Lemma 2.3.2. There exists a constant $\mathcal{C} \in \mathbb{R}$ such that for any $u \in \mathrm{H}_0^{\left(\frac{1}{2},1\right)}(Q)$:

$$\int_Q |u(t,x)|^4 \,\mathrm{d}t\mathrm{d}x \leq \mathcal{C}^2 \left(\int_Q |u|^2 \,\mathrm{d}t\mathrm{d}x + \int_Q \left|u_{\sqrt{t}}\right|^2 \mathrm{d}t\mathrm{d}x\right) \cdot \left(\int_Q |u_x|^2 \,\mathrm{d}t\mathrm{d}x\right)$$

which implies that:

$$\left|u^2\right| \leq \mathcal{C}\,\|u\|\,|u_x| \tag{2.5}$$

Proof.

1. Using the standard Sovolev embedding:

$$\mathrm{H}^{\left(\frac{1}{2}\right)}(\mathbb{R}^2) \subset \mathrm{L}^4(\mathbb{R}^2)$$

we get by a scaling argument:

$$\forall v \in \mathrm{H}^{\left(\frac{1}{2}\right)}(\mathbb{R}^2)$$
$$\int_{\mathbb{R}^2} |v(t,x)|^4 \,\mathrm{d}t\mathrm{d}x \leq C \left(\int_{\mathbb{R}^2} \left|v_{\sqrt{t}}(t,x)\right|^2 \mathrm{d}t\mathrm{d}x \right) \cdot \left(\int_{\mathbb{R}^2} \left|v_{\sqrt{x}}(t,x)\right|^2 \mathrm{d}t\mathrm{d}x \right) \quad (2.6)$$

2. We use the partial Fourier transform in x:

$$\hat{v}(t,\xi) = \int_{\mathbb{R}} v(t,x) \mathrm{e}^{-\mathrm{i}2\pi x\xi} \mathrm{d}x$$

$$v_{\sqrt{x}}(t,x) = \int_{\mathbb{R}} \hat{v}(t,\xi) \sqrt{\mathrm{i}2\pi\xi} \mathrm{e}^{\mathrm{i}2\pi\xi} \mathrm{d}\xi$$

By Plancherel and Cauchy-Schwarz:

$$\begin{aligned} \left|v_{\sqrt{x}}\right|^2 &= \int_{\mathbb{R}^2} |\hat{v}(t,\xi)| \cdot |\hat{v}(t,\xi)| \, |\mathrm{i}2\pi\xi| \,\mathrm{d}\xi\mathrm{d}t \\ &\leq \sqrt{\int_{\mathbb{R}^2} |\hat{v}|^2} \cdot \sqrt{\int_{\mathbb{R}^2} |\hat{v}(t,\xi)|^2 (2\pi\xi)^2 \mathrm{d}\xi\mathrm{d}t} \\ &= \sqrt{\int_{\mathbb{R}^2} |v|^2 \cdot \int_{\mathbb{R}^2} |v_x|^2} \end{aligned}$$

3. From the last inequality together with (2.6), by extending functions by zero outside $\mathbb{R} \times \mathcal{I}$ we get for any $v \in \mathrm{H}_0^{\left(\frac{1}{2},1\right)}(\mathbb{R} \times \mathcal{I})$:

$$\int_{\mathbb{R}\times\mathcal{I}} |v|^4 \leq \int_{\mathbb{R}\times\mathcal{I}} \left|v_{\sqrt{t}}\right|^2 \sqrt{\int_{\mathbb{R}\times\mathcal{I}} |v|^2 \int_{\mathbb{R}\times\mathcal{I}} |v_x|^2}$$

The Poincaré inequality on $\mathbb{R} \times \mathcal{I}$:

$$\int_{\mathbb{R}\times\mathcal{I}} |v|^2 \leq \frac{1}{\pi} \int_{\mathbb{R}\times\mathcal{I}} |v_x|^2$$

then gives

$$\int_{\mathbb{R}\times\mathcal{I}} |v|^4 \leq \frac{C}{\pi} \int_{\mathbb{R}\times\mathcal{I}} \left|v_{\sqrt{t}}\right|^2 \int_{\mathbb{R}\times\mathcal{I}} |v_x|^2 \quad (2.7)$$

4. Finally, given $u \in \mathrm{H}_0^{\left(\frac{1}{2},1\right)}(Q)$ we define $\tilde{u}$ on $\mathbb{R} \times \mathcal{I}$ as the only 1-periodic function in t which is equal to u on $(0,1) \times \mathcal{I}$. Take φ in the Schwartz space $\mathcal{S}(\mathbb{R})$ such that $\mathrm{supp}(\hat{\varphi}) \subset (-\frac{1}{2}, \frac{1}{2})$, and $\varphi(0) = 1$. Moreover, given $0 < \delta < 1$, by means of scalings we may always choose φ such that $\varphi([-\frac{1}{2}, \frac{1}{2}]) \subset [1-\delta, 1+\delta]$.

a) $\tilde{u} \in \mathcal{S}'(\mathbb{R}, \mathrm{L}^2(\mathcal{I}))$ so $\mathcal{F}\,\tilde{u} = \sum_{k\in\mathbb{Z}} u_k \delta(\tau - k)$ where $u_k \in \mathrm{L}^2(\mathcal{I})$ are the Fourier coefficients of u. By the convolution formula we get the following Fourier expansion for $\varphi\tilde{u}$:

$$\mathcal{F}\,(\varphi\tilde{u}) = \sum_{k\in\mathbb{Z}} u_k \hat{\varphi}(\tau - k)$$

b) Thus

$$\begin{aligned}\int_{\mathbb{R}\times\mathcal{I}} \left|(\varphi\tilde{u})_{\sqrt{t}}\right|^2 &= \int_{\mathbb{R}} \left\|(\varphi\tilde{u})_{\sqrt{t}}\right\|^2_{\mathrm{L}^2(\mathcal{I})} \\ &= \int_{\mathbb{R}} |\tau| \left\| \sum_k u_k \hat{\varphi}(\tau - k) \right\|^2_{\mathrm{L}^2(\mathcal{I})} \mathrm{d}\tau \\ &= \sum_k \|u_k\|^2_{\mathrm{L}^2(\mathcal{I})} \int_{k-\frac{1}{2}}^{k+\frac{1}{2}} |\tau| \left|\hat{\varphi}(\tau - k)\right|^2 \mathrm{d}\tau\end{aligned}$$

Now the term on the right hand side can be estimated as follows:

$$\begin{aligned}\int_{k-\frac{1}{2}}^{k+\frac{1}{2}} |\tau| \left|\hat{\varphi}(\tau - k)\right|^2 \mathrm{d}\tau &= \int_{k-\frac{1}{2}}^{k+\frac{1}{2}} \underbrace{(|\tau| - |k|)}_{\leq 1} \left|\hat{\varphi}(\tau - k)\right|^2 + |k| \left|\hat{\varphi}(\tau - k)\right|^2 \mathrm{d}\tau \\ &\leq (1 + |k|) \int_{\mathbb{R}} \left|\hat{\varphi}\right|^2\end{aligned}$$

so we get:

$$\int_{\mathbb{R}\times\mathcal{I}} \left|(\varphi\tilde{u})_{\sqrt{t}}\right|^2 \leq \int_{\mathbb{R}} \left|\hat{\varphi}\right|^2 \cdot \left(\int_{\mathbb{T}\times\mathcal{I}} |u|^2 + \int_{\mathbb{T}\times\mathcal{I}} \left|u_{\sqrt{t}}\right|^2 \right) \tag{2.8}$$

c) Furthermore

$$\int_{\mathbb{R}\times\mathcal{I}} |\varphi\tilde{u}|^4 \geq (1 - \delta)^4 \int_{\mathbb{T}\times\mathcal{I}} |u|^4 \tag{2.9}$$

d) Since $\varphi \in \mathcal{S}(\mathbb{R})$ there exists an $A > 0$ such that $\forall t \in \mathbb{R} \quad |\varphi(t)| \leq A/(1 + t)$. Thus finally:

$$\begin{aligned}\int_{\mathbb{R}\times\mathcal{I}} |(\varphi\tilde{u})_x|^2 &= \sum_{k=0}^{\infty} \int_{[k,k+1]\cup[-k-1,-k]} |\varphi(t)|^2 \|\tilde{u}_x(t)\|^2_{\mathrm{L}^2(\mathcal{I})} \\ &\leq \sum_{k=0}^{\infty} 2 \left(\frac{A}{k+1}\right)^2 \int_Q |u_x|^2 \\ &= C \int_Q |u_x|^2\end{aligned} \tag{2.10}$$

5. By using (2.7) with $v = \varphi\tilde{u}$ and combining (2.8), (2.9) and (2.10) we get the desired inequality.

6. By using the Poincaré inequality once more one gets (2.5), which concludes the proof of Lemma 2.3.2.

□

We are now ready to proceed to the proof of the a priori estimate.

Proof of Theorem 2. By definition $\mathcal{L}u + \lambda S(u) = f$ means:

$$\forall v \in \mathrm{H}_0^{\left(\frac{1}{2},1\right)} \quad \left(u_{\sqrt{t}}, v_{\sqrt{t}*}\right) + \mu\left(u_x, v_x\right) - \frac{1}{2}\lambda\left(u^2, v_x\right) = \langle f, v\rangle \tag{2.11}$$

1. We notice that for smooth u:

$$\begin{aligned}\left(u^2, u_x\right) &= \int_Q u^2 u_x \\ &= \frac{1}{3}\int_Q (u^3)_x \\ &= 0\end{aligned}$$

and then by density and continuity this holds for all $u \in \mathrm{H}_0^{\left(\frac{1}{2},1\right)}$.

2. With $v = u$ in (2.11) we get:

$$\underbrace{\left(u_{\sqrt{t}}, u_{\sqrt{t}*}\right)}_{=0} + \mu\left(u_x, u_x\right) + \frac{1}{2}\lambda\underbrace{\left(u^2, u_x\right)}_{=0} = \langle f, u\rangle$$

which gives:

$$|u_x|^2 = \frac{\langle f, u\rangle}{\mu} \tag{2.12}$$

$$\leq \frac{\|f\|\,\|u\|}{\mu} \tag{2.13}$$

3. With $v = \widetilde{u}$ in (2.11) we get:

$$\left(u_{\sqrt{t}}, \widetilde{u}_{\sqrt{t}*}\right) + \mu\underbrace{\left(u_x, \widetilde{u}_x\right)}_{=0} + \frac{1}{2}\lambda\left(u^2, \widetilde{u}_x\right) = \langle f, \widetilde{u}\rangle$$

Using the identity (1.2), the fact that $\|\widetilde{u}\| = \|u\|$ and that $\lambda \leq 1$ we get:

$$\left|u_{\sqrt{t}}\right|^2 \leq \frac{1}{2}\left|\left(u^2, \widetilde{u}_x\right)\right| + \|f\|\,\|u\| \tag{2.14}$$

4. We estimate $\left|\left(u^2, \widetilde{u}_x\right)\right|$ using the Lemma 2.3.2:

$$\begin{aligned}\left|\left(u^2, \widetilde{u}_x\right)\right| &\leq \left|u^2\right|\,|u_x| \\ &\leq \mathcal{C}\,\|u\|\,|u_x|^2\end{aligned} \tag{2.15}$$

5. We use the Lemma 2.3.1 to write $f = g_{\sqrt{t}} + h_x$ together with (2.12) we have:

$$\begin{aligned}\mu |u_x|^2 &= \langle f, u\rangle \\ &= \left(g, u_{\sqrt{t}*}\right) - (h, u_x) \\ &\leq |g| \left|u_{\sqrt{t}}\right| + |h| |u_x| \\ &\leq |g| \left|u_{\sqrt{t}}\right| + |h| \sqrt{\frac{\|f\| \|u\|}{\mu}} \end{aligned} \tag{2.16}$$

6. Using the estimate (2.16) inside (2.15) and the fact that $\left|u_{\sqrt{t}}\right| \leq \|u\|$ we get:

$$\frac{1}{2} \left|\left(u^2, \widetilde{u}_x\right)\right| \leq R_0 \left[|g| \|u\|^2 + |h| \sqrt{\frac{\|f\|}{\mu}} \|u\|^{\frac{3}{2}}\right]$$

Where R_0 is defined as:

$$R_0 = \frac{\mathcal{C}}{2} \cdot \frac{1}{\mu}$$

So if we choose g small enough such that

$$R_0 |g| \leq \frac{1}{2}$$

then using (2.14) we get:

$$\left|u_{\sqrt{t}}\right|^2 \leq \|f\| \|u\| + \frac{1}{2} \|u\|^2 + R_0 |h| \sqrt{\frac{\|f\|}{\mu}} \|u\|^{\frac{3}{2}} \tag{2.17}$$

So with the notations:

$$a = 2\left(1 + \frac{1}{\mu}\right) \|f\|$$

and

$$b = R_0 |h| \sqrt{\frac{\|f\|}{\mu}}$$

from (2.13) and (2.17) we get

$$\|u\|^2 \leq a \|u\| + 2b \|u\|^{\frac{3}{2}}$$

A straightforward computation leads to the bound

$$\|u\| \leq (b + \sqrt{a + b^2})^2$$

Since that estimate does not depend on λ the theorem is proved.

□

2.4 Existence of solutions

Existence and Uniqueness in the Linear Case

Theorem 3. $\mathcal{L}$ is a continuous bijection.

Proof.

1. Let us define the following operator on $\mathrm{H}_0^{\left(\frac{1}{2},1\right)}$:
$$\mathrm{P}(u) = \frac{u - \widetilde{u}}{\sqrt{2}}$$
P is an isomorphism on $\mathrm{H}_0^{\left(\frac{1}{2},1\right)}$ since the corresponding Fourier multiplier has either module one or $1/\sqrt{2}$.
2. Now
$$\begin{aligned}\langle \mathrm{P}^* \mathcal{L}\, u, u\rangle &= \langle \mathcal{L}\, u, \mathrm{P}\, u\rangle \\ &= \frac{1}{\sqrt{2}}\left(\underbrace{\left(u_{\sqrt{t}}, u_{\sqrt{t}*}\right)}_{=0} - \left(u_{\sqrt{t}}, \widetilde{u}_{\sqrt{t}*}\right) + \mu\left(u_x, u_x\right) - \mu\underbrace{\left(u_x, \widetilde{u_x}\right)}_{=0}\right) \\ &= \frac{1}{\sqrt{2}}\left(\left|u_{\sqrt{t}}\right|^2 + \mu\left|u_x\right|^2\right) \\ &\geq \frac{\min\{1,\mu\}}{\sqrt{2}}\left\|u\right\|^2\end{aligned}$$
3. $\mathrm{P}^* \circ \mathcal{L}$ is therefore a coercive and continuous linear operator from $\mathrm{H}_0^{\left(\frac{1}{2},1\right)}$ to $\mathrm{H}^{\left(-\frac{1}{2},-1\right)}$. By the Lax-Milgram theorem (cf. for example [Bre93]) it is invertible. Since P is an isomorphism, so is P^*. We conclude that $\mathcal{L}$ is an isomorphism.

□

Existence for the General Case

Theorem 4. T is surjective.

Proof.

1. Because of Theorem 3, the equation $\mathrm{T}(u) = f$ can be rewritten:
$$\left[\mathbb{I} + K\right](u) = \mathcal{L}^{-1} f$$
where
$$K = \mathcal{L}^{-1} \circ S$$
so we only have to show that the application $\mathbb{I} + K$ is surjective. But since $\mathcal{L}^{-1}$ is continuous and by Lemma 2.2.1 S is compact, K is a compact map.
2. We choose an open ball U of $\mathrm{H}_0^{\left(\frac{1}{2},1\right)}$ that contains $\bigcup_{\lambda\in[0,1]}(\mathcal{L}+\lambda S)^{-1}(f)$. Theorem 2 ascertains that for all $\lambda \in [0,1]$, $\mathcal{L}^{-1} f \notin (\mathbb{I} + \lambda K)(\partial U)$. So the Leray-Schauder degree of $\mathbb{I} + K$ on U is equal to the one of $\mathbb{I}$ which is one:
$$D(\mathbb{I} + K, U, \mathcal{L}^{-1} f) = D(\mathbb{I}, U, \mathcal{L}^{-1} f) = 1$$
As a result, $\mathbb{I} + K$ is surjective and the theorem is proved. (For the Leray-Schauder degree theory, see for instance [Dei85]).

□

3

Cole-Hopf Transformation

3.1 Introduction

The aim of this chapter is to prove the following Theorem:

Theorem 5. T is injective.

The proof is quite long and involved and is split into Proposition 3.2.1, Proposition 3.3.1 and Proposition 3.4.1.

The first observation we make is that if two functions $u, v \in \mathrm{H}_0^{\left(\frac{1}{2},1\right)}$ satisfy $\mathrm{T}(u) = \mathrm{T}(v)$ then $w := u - v$ satisfies

$$\mathrm{T}(w) = -(vw)_x \tag{3.1}$$

Thus to prove Theorem 5 it suffices to prove that given any fixed $v \in \mathrm{H}_0^{\left(\frac{1}{2},1\right)}$ the equation (3.1) has only the trivial solution $w = 0$ in $\mathrm{H}_0^{\left(\frac{1}{2},1\right)}$.

This will be done using the Cole-Hopf transformation.

3.2 Cole-Hopf Transformation in Sobolev Spaces

In order to define it we shall need the following anisotropic Sobolev space with Neumann boundary conditions, denoted $\mathrm{H}_N^{(1,2)}$

$$\mathrm{H}_N^{(1,2)} = \left\{ u \in \mathrm{H}^{(1,2)}(Q);\ u_x(t,0) = u_x(t,1) = 0 \quad \forall t \in \mathbb{T} \right\}$$

Notice that by (1.3) (see also Figure 3.1) we have:

$$\mathrm{H}^{(1,2)}(Q) \subset \mathrm{H}^{\left(\frac{2}{3}\right)\left(\frac{2}{3}\right)}(Q) \subset \mathcal{C}^0\big(\overline{\mathbb{T}}, \mathcal{C}^0(\overline{\mathcal{I}})\big) = \mathcal{C}^0(\overline{Q}) \tag{3.2}$$

We may therefore define the following set $\mathrm{H}_{N+}^{(1,2)}$:

$$\mathrm{H}_{N+}^{(1,2)} = \left\{ u \in \mathrm{H}_N^{(1,2)};\ u > 0 \text{ on } \overline{Q} \right\}$$

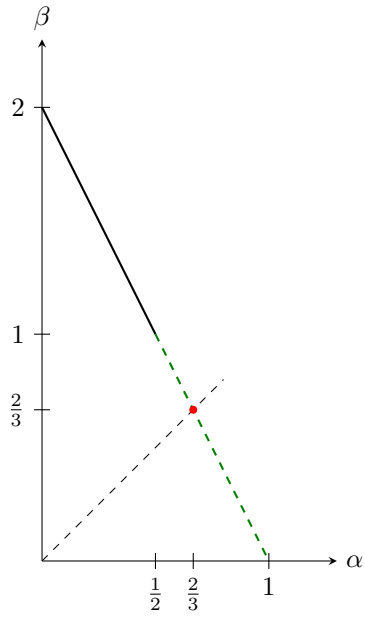


Figure 3.1: The first step of the Cole-Hopf Transformation is an integration in x. This function U obtained thus ends up in $\mathrm{H}^{(0)(1)} \cap \mathrm{H}^{\left(\frac{1}{2}\right)(1)}$, which delimits the plain line on the graph above. But it follows from $\mathrm{T}\, u \in \mathrm{H}^{(0)(-1)}$ that u is actually also in $\mathrm{H}^{(1)(-1)}$ so U ends up in $\mathrm{H}^{(1)(2)}$ and we have an inclusion in $\mathrm{H}^{\left(\frac{2}{3}\right)\left(\frac{2}{3}\right)}$ which is embedded in continuous Hölder functions.

We will also need the quotient sets $\mathrm{H}_N^{(1,2)}/\mathbb{R}$ and $\mathrm{H}_{N+}^{(1,2)}/\mathbb{R}_+$ where the latter is the quotient with respect to the action of the multiplicative group $(\mathbb{R}_+, \times)$ given by the scalar multiplication (i.e. $\varphi \sim \psi \iff \exists \eta > 0$ s.t. $\psi = \eta\varphi$).

We now define the following three solution sets S_1, S_2 and S_3, all depending on a fixed function $v \in \mathrm{H}_0^{\left(\frac{1}{2},1\right)}$.

Definition 4. We say that $w \in S_1$ if $w \in \mathrm{H}_0^{\left(\frac{1}{2},1\right)}$ and

$$\mathrm{T}(w) = -(vw)_x$$

Definition 5. We say that $([W], K) \in S_2$ if $[W] \in \mathrm{H}_N^{(1,2)}/\mathbb{R}$, $K \in \mathbb{R}$ and

$$W_t - \mu W_{xx} + \frac{1}{2}(W_x)^2 = -vW_x + K$$

Definition 6. We say that $([\varphi], K) \in S_3$ if $[\varphi] \in \mathrm{H}^{(1,2)}_{N+}/\mathbb{R}_+$, $K \in \mathbb{R}$ and

$$\varphi_t - \mu\varphi_{xx} + v\varphi_x + K\varphi = 0 \tag{3.3}$$

Notice that the definitions above are consistent since the equations used do not depend on the chosen representative. By the remark above, Theorem 5 will be proved if we can show that the cardinality of S_1 is one. We will do this by first proving that the cardinalities of S_1 and S_3 are the same (Proposition 3.2.1) and then finally by proving that $\mathrm{card}(S_3) = 1$ (Proposition 3.4.1).

We first prove an auxiliary Lemma that will be used to construct a bijection between S_2 and S_3:

Lemma 3.2.1. The exponential function is a bijection from $\mathrm{H}^{(1,2)}_N$ to $\mathrm{H}^{(1,2)}_{N+}$. The natural logarithm is its inverse. These functions can be naturally extended to bijections between $\mathrm{H}^{(1,2)}_N/\mathbb{R}$ and $\mathrm{H}^{(1,2)}_{N+}/\mathbb{R}_+$.

Proof.

1. Take $W \in \mathrm{H}^{(1,2)}_N$. By (3.2) we have $W \in \mathcal{C}^0(\overline{Q})$ and thus $\exp(W) \in \mathcal{C}^0(\overline{Q})$. The Lemma thus follows by simple computations and Sobolev injections. Indeed since $W_x \in \mathrm{H}_0^{\left(\frac{1}{2},1\right)}$ and $\mathrm{H}_0^{\left(\frac{1}{2},1\right)} \subset \mathrm{L}^4$ (by (1.4)), by considering $(\exp W)_{xx}$ we obtain:

$$(\exp W)_{xx} = \Big(\underbrace{W_{xx}}_{\in \mathrm{L}^2} + \underbrace{(W_x)^2}_{\in \mathrm{L}^4}\Big) \underbrace{\exp W}_{\in \mathcal{C}^0(\overline{Q})} \quad \in \mathrm{L}^2$$

   and since $(\exp W)_t \in \mathrm{L}^2$ we get $\exp(W) \in \mathrm{H}^{(1,2)}_{N+}$.

2. The proof goes along the same lines for the logarithm function.

3. The exponential and logarithm functions preserve the group actions used to define the quotient sets $\mathrm{H}^{(1,2)}_N/\mathbb{R}$ and $\mathrm{H}^{(1,2)}_{N+}/\mathbb{R}_+$ and can thus be extended to bijections between those sets.

□

We are now ready to prove our first Proposition:

Proposition 3.2.1. The cardinalities of the solution sets S_1, S_2 and S_3 defined above are the same.

Proof.

1. We shall explicitely construct two transformations, one from S_1 to S_2: $\phi_{21} : S_1 \to S_2$ and the other from S_2 to S_1: $\phi_{12} : S_2 \to S_1$ that are inverse to each other.
   For $w \in S_1$ we have $w_t = (\mu w_x - \frac{1}{2}w^2 - vw)_x$ so $w_t \in \mathrm{H}^{(0,-1)}$ and thus $\overline{W} := \int_0^x w(t,y)\mathrm{d}y \in \mathrm{H}^{(1,2)}$. Moreover since $w \in S_1$ we get

$$\mathrm{D_x}\Big(\overline{W}_t - \mu\overline{W}_{xx} + \frac{1}{2}(\overline{W}_x)^2 + v\overline{W}_x\Big) = 0$$

and thus

$$\overline{W}_t - \mu \overline{W}_{xx} + \frac{1}{2}(\overline{W}_x)^2 + v\overline{W}_x = g$$

for some $g \in \mathrm{L}^2(\mathbb{T})$.

Now define $h \in \mathrm{H}^{(1)}(\mathbb{T})$ by $g(t) = h'(t) + \int_{\mathbb{T}} g$ and then define $W := [\overline{W} - h]$ i.e. the projection of $\overline{W} - h \in \mathrm{H}_N^{(1,2)}$ onto $\mathrm{H}_N^{(1,2)}/\mathbb{R}$. An elementary computation shows that $(W, \int_{\mathbb{T}} g) \in S_2$. We put $\phi_{21}(w) = (W, \int_{\mathbb{T}} g)$.

2. On the other hand, given $([W], K) \in S_2$ with $W \in \mathrm{H}_N^{(1,2)}$, a straightforward computation shows that $W_x \in S_1$. Moreover W_x obviously does not depend on the chosen representative W. We put $\phi_{12}([W], K) = W_x$.

   Now $\phi_{12} \circ \phi_{21} = \mathbb{I}$ since for any $h \in \mathrm{L}^2(\mathbb{T})$:

$$\mathrm{D_x}\left(\int_0^x w(t,y)\mathrm{d}y - h(t)\right) = w$$

   Furthermore given $([W], K) \in S_2$ then $\phi_{21} \circ \phi_{12}([W], K) = ([U], \tilde{K}) \in S_2$ for some $U \in \mathrm{H}_N^{(1,2)}$ and $\tilde{K} \in \mathbb{R}$. One checks that $[U] = [W - h(t)]$ for an $h \in \mathrm{H}^{(1)}(\mathbb{T})$, which implies that $h'(t) = \tilde{K} - K$ and thus since h is periodic $K = \tilde{K}$ and h is constant, and thus $[W] = [W - h]$. As a result $\phi_{21} \circ \phi_{12} = \mathbb{I}$.

3. We will again construct transformations between the two sets S_2 and S_3 which are inverse to each other: $\phi_{32} : S_2 \to S_3$ and $\phi_{23} : S_3 \to S_2$. Given an element $([W], K) \in S_2$ and one representative W one defines $\varphi = \mathrm{e}^{-\frac{W}{2\mu}}$. Then

$$\begin{cases} \varphi_t = -\dfrac{W_t}{2\mu}\varphi \\ \varphi_x = -\dfrac{W_x}{2\mu}\varphi \\ \varphi_{xx} = \left[-\dfrac{W_{xx}}{2\mu} + \left(\dfrac{W_x}{2\mu}\right)^2\right]\varphi \end{cases}$$

   So

$$\begin{aligned} \varphi_t - \mu\varphi_{xx} + v\varphi_x &= -\frac{1}{2\mu}\left[W_t - \mu W_{xx} + \frac{1}{2}(W_x)^2 + vW_x\right]\varphi \\ &= -\frac{K}{2\mu}\varphi \end{aligned}$$

   Using that $\varphi = \exp(-W/2\mu)$ and Lemma 3.2.1 one gets $\varphi \in \mathrm{H}_{N+}^{(1,2)}$. We denote $[\varphi]$ the projection of $\varphi \in \mathrm{H}_{N+}^{(1,2)}$ onto $\mathrm{H}_{N+}^{(1,2)}/\mathbb{R}_+$. So $([\varphi], \frac{K}{2\mu})$ is in S_3. Since $[\varphi]$ does not depend on the chosen representative W, the function ϕ_{32} which maps $([W], K)$ to $([\varphi], K/2\mu)$ is well defined.

4. We define in the same way ϕ_{23} by: $\phi_{23}(([\varphi], K)) = ([\log(\varphi)], 2\mu K)$. It is easy to see that ϕ_{23} is well defined, that it maps S_3 to S_2, and that ϕ_{23} and ϕ_{32} are inverse to each other.

□

We sum up the last result in the following corollary.

Corollary 3.2.1. The Cole-Hopf transformation Φ defined by $\Phi := \phi_{32} \circ \phi_{21}$ is a bijection from S_1 to S_3.

3.3 Uniqueness of the Ground Eigenvalue

We will now set out to prove that $S_3 = \{([1], 0)\}$. The first step consists in proving that if $(\varphi, K) \in S_3$ then $K = 0$. We will need a preliminary Lemma which proves the positivity of the evolution operator associated with the equation $\psi_t - \mu\psi_{xx} + v\psi_x = 0$.

Lemma 3.3.1. Given $v \in \mathrm{H}_0^{\left(\frac{1}{2},1\right)}$, for any $\psi \in \mathrm{H}_N^{(1,2)}((0,1) \times \mathcal{I})$ such that

$$\psi_t - \mu\psi_{xx} + v\psi_x = 0 \tag{3.4}$$

the traces $\psi(0) := (x \mapsto \psi(0,x))$ and $\psi(1) := (x \mapsto \psi(1,x))$ are well defined in $\mathcal{C}^0(\overline{\mathcal{I}})$ and the following holds:

$$\psi(0) \geq 0 \implies \psi(1) \geq 0$$

Proof.

1. The traces are well defined by the same argument as in (3.2):

$$\mathrm{H}^{(1,2)}((0,1) \times \mathcal{I}) \subset \mathrm{H}^{\left(\frac{2}{3}\right)\left(\frac{2}{3}\right)}((0,1) \times \mathcal{I}) \subset \mathcal{C}^0(\overline{(0,1) \times \mathcal{I}})$$

2. (3.4) implies that

$$\int_{\mathcal{I}} \psi_t \chi \mathrm{d}x + \mu \int_{\mathcal{I}} \psi_x \chi_x \mathrm{d}x + \int_{\mathcal{I}} v\psi_x \chi \mathrm{d}x = 0$$

for any test function $\chi \in \mathrm{H}^{(1,2)}((0,1) \times \mathcal{I})$. Using $\chi = \psi^- := \max(-\psi(t,x), 0)$ we get:

$$\int_{\mathcal{I}} \chi_t \chi \mathrm{d}x + \mu \int_{\mathcal{I}} \chi_x^2 \mathrm{d}x + \int_{\mathcal{I}} v\chi_x \chi \mathrm{d}x = 0 \tag{3.5}$$

3. We define

$$g(t) = \int_{\mathcal{I}} \chi(t,x)^2 \mathrm{d}x$$
$$h(t) = \int_{\mathcal{I}} \chi_x(t,x)^2 \mathrm{d}x$$

So we can rewrite (3.5) as:

$$\frac{1}{2} g'(t) + \mu h(t) - \frac{1}{2} \int_{\mathcal{I}} v_x \chi^2 \mathrm{d}x = 0 \tag{3.6}$$

4. We estimate the third term of the last equation:

$$\int_{\mathcal{I}} v_x \chi^2 = \int_{\mathcal{I}} v_x \left(\chi^2 - g(t)\right) \mathrm{d}x + \int_{\mathcal{I}} v_x g(t) \mathrm{d}x \tag{3.7}$$

We estimate the two last terms in the following way:

a)

$$\left| \int_{\mathcal{I}} v_x (\chi^2 - g(t)) \mathrm{d}x \right| \leq h_2(t) \sqrt{\int_{\mathcal{I}} (\chi^2 - g(t))^2}$$

where:

$$h_2(t) = \sqrt{\int_{\mathcal{I}} |v_x(t,x)|^2 \, \mathrm{d}x}$$

b) Notice that since for all $t \in [0,1]$ $\chi(t,\cdot) \in \mathrm{H}^{(1)}(\mathcal{I})$ we have:

$$\begin{aligned} \chi^2(t,x) - \chi^2(t,x_0) &= \int_{x_0}^{x} (\chi^2)_x(t,y) \mathrm{d}y \\ &= 2 \int \chi(t,y) \chi_x(t,y) \mathrm{d}y \\ &\leq 2 \sqrt{\int_{\mathcal{I}} \chi^2} \sqrt{\int_{\mathcal{I}} (\chi_x)^2} \end{aligned}$$

By integrating first with respect to x_0, squaring and then integrating with respect to x we get:

$$\int_{\mathcal{I}} (\chi^2(t,x) - g(t))^2 \mathrm{d}x \leq 4gh$$

so from (3.7) we obtain:

$$\int_{\mathcal{I}} v_x \chi^2 \leq h_2(t) \cdot 2 \cdot \sqrt{g \cdot h} + h_1(t) \cdot g(t) \tag{3.8}$$

where

$$h_1(t) = \int_{\mathcal{I}} |v_x(t,x)| \, \mathrm{d}x$$

c) By Young's inequality:

$$2h_2 \sqrt{gh} \leq \frac{g \cdot h_2^2}{2\mu} + 2\mu h$$

d) (3.8) now becomes:

$$\frac{1}{2} \int_{\mathcal{I}} v_x \chi^2 \leq \frac{g \cdot h_2^2}{4\mu} + \mu h(t) + \frac{1}{2} h_1(t) \cdot g(t)$$

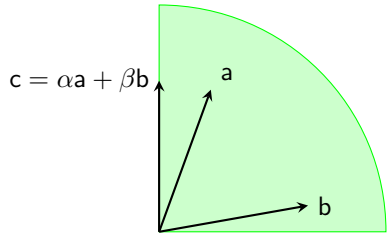


Figure 3.2: Illustration of the proof of Proposition 3.3.1. Assume that a and b are two eigenvectors of an operator A with respective eigenvalues λ and μ. Assume further that A preserves the cone represented in green on the picture. One can choose a particular linear combination $\mathsf{c} = \alpha\mathsf{a} + \beta\mathsf{b}$ such that c lies on the "border" of the cone. It is clear that the image $\mathrm{A}\,\mathsf{c}$ is in the cone iff $\mu \leq \lambda$. By repeating this argument on the other side of the cone one concludes that $\lambda = \mu$.

5. Combining the last estimate with (3.6) yields:
$$g'(t) \leq h_3(t) \cdot g(t)$$
where
$$h_3(t) = \left(\frac{h_2^2(t)}{2\mu} + h_1(t) \right)$$
6. h is integrable and $g(0) = 0$ so $g(t) = 0$ for any $t > 0$.

□

Proposition 3.3.1. If $([\varphi], K) \in S_3$ then $K = 0$.

Proof. From any representative φ we define $\psi \in \mathrm{H}^{(1,2)}((0,1)\times\mathcal{I})$ by $\psi(t,x) = \mathrm{e}^{-Kt}\varphi(t,x)$. A simple computation shows that $\psi_t - \mu\psi_{xx} + v\psi_x = 0$ on $(0,1)\times\mathcal{I}$ so we may use Lemma 3.3.1 to get that the trace $\psi(0) \in \mathcal{C}^0(\overline{\mathcal{I}})$ is well defined. We denote its minimum and maximum values in the following way:
$$\gamma_+ = \max_{x\in\mathcal{I}} \psi(0,x)$$
$$\gamma_- = \min_{x\in\mathcal{I}} \psi(0,x)$$
so
$$\psi_1 := \gamma_+ - \psi \geq 0$$
$$\psi_2 := \psi - \gamma_- \geq 0$$
Now ψ_1 and ψ_2 both qualify for Lemma 3.3.1 and $\psi_1(0) \geq 0$ and $\psi_2(0) \geq 0$. Moreover by construction we have $\psi(1) = \mathrm{e}^{-K}\psi(0)$ so applying Lemma 3.3.1 yields:
$$\psi_1(1) = \gamma_+ - \mathrm{e}^{-K}\psi(0) \geq 0$$
$$\psi_2(1) = \mathrm{e}^{-K}\psi(0) - \gamma_- \geq 0$$
so we get $\mathrm{e}^{-K} \leq 1$ and $\mathrm{e}^{-K} \geq 1$ and hence $K = 0$. □

3.4 Non-degeneracy of the Ground State

In the last section we proved that the "eigenvalue" K must be zero. We now prove that degeneracy for the corresponding eigenspace is impossible, i.e. we prove that it is one dimensional. Degeneracy implies indeed that the eigenspace would meet the boundary of the cone of positive functions. We will show that this cannot occur because S_2 is bounded in $\mathcal{C}^0(\overline{Q})/\mathbb{R}$. This fact, which follows from our a priori estimate of Corollary 2.3.1, is proved in the next Lemma.

Notice that by the previous Propositions S_2 is naturally embedded in $\mathrm{H}_N^{(1,2)}/\mathbb{R}$ (instead of $(\mathrm{H}_N^{(1,2)}/\mathbb{R}) \times \mathbb{R}$) which is itself embedded in $\mathcal{C}^0(\overline{Q})$. With this identification we have the following Lemma:

Lemma 3.4.1. *S_2 is bounded in $\mathcal{C}^0(\overline{Q})/\mathbb{R}$.*

Proof.

1. We first show that $\mathcal{L}$ (defined by $\mathcal{L}\,u = u_t - \mu u_{xx}$ for $u \in \mathrm{H}_N^{(1,2)}/\mathbb{R}$) is an isometry from $\mathrm{H}_N^{(1,2)}/\mathbb{R}$ to $L := \left\{ g \in \mathrm{L}^2(Q);\ \int_Q g = 0 \right\}$. Take $f \in L$. We define $\tilde{f}$ on $\mathbb{T} \times (-1, 1)$ by symmetrisation:
$$\begin{cases} \tilde{f}(t,x) = f(t,x) & \text{if } x \geq 0 \\ \tilde{f}(t,x) = f(t,-x) & \text{if } x \leq 0 \end{cases}$$
We can now regard $\tilde{f}$ as an element of $\mathrm{L}^2(\mathbb{T} \times (\mathbb{R}/2\mathbb{Z}))$. Notice that $\int_{\mathbb{T}\times(\mathbb{R}/2\mathbb{Z})} \tilde{f} = 0$. By Fourier analysis there is a unique $\tilde{u} \in \mathrm{H}^{(1,2)}(\mathbb{T} \times (\mathbb{R}/2\mathbb{Z}))/\mathbb{R}$ solution of
$$\tilde{u}_t - \mu \tilde{u}_{xx} = \tilde{f}$$
Now $\tilde{u}_x \in \mathrm{H}^{\left(\frac{1}{2},1\right)}(\mathbb{T} \times (\mathbb{R}/2\mathbb{Z}))$ so u_x has a trace on $\mathbb{T} \times \{0\}$ and $\mathbb{T} \times \{1\}$. By symmetry it must be zero in $\mathrm{L}^2(\mathbb{T})$.
2. For $(W, 0) \in S_2$ we have:
$$\mathcal{L}\,W = -\frac{1}{2}(W_x)^2 - vW_x$$
so by the previous result:
$$W = -\frac{1}{2}\,\mathcal{L}^{-1}\big(W_x(2v + W_x)\big)$$
3. Since $W \in S_2 \implies W_x \in S_1$ (cf. Proposition 3.2.1), and $W_x \in S_1 \iff \mathrm{T}(v + W_x) = \mathrm{T}(v)$, by the a priori estimate (Corollary 2.3.1) there exists $C > 0$ such that $\forall W \in S_2$, $\|W_x\| \leq C$. Using $\mathrm{H}_0^{\left(\frac{1}{2},1\right)} \subset \mathrm{L}^4$ we obtain that $W_x(2v + W_x)$ is bounded in L^2.
4. Combining the three preceding steps we conclude that S_2 is bounded in $\mathrm{H}_N^{(1,2)}/\mathbb{R}$. But $\mathrm{H}_N^{(1,2)}/\mathbb{R} \subset \mathcal{C}^0(\overline{Q})/\mathbb{R}$ so we get the result.

□

Proposition 3.4.1.
$$S_3 = \big\{([1], 0)\big\}$$

Proof. It is obvious that $([1], 0) \in S_3$. We proved (in Proposition 3.3.1) that $(\varphi, K) \in S_3$ implies $K = 0$. So we take $([\varphi], 0) \in S_3$ and we will now show that $[\varphi] = [1]$.

1. Given one representative φ of $[\varphi]$, let us choose a point $x_0 \in \overline{Q}$ where φ takes its minimum γ:
$$\gamma = \min_{\overline{Q}} \varphi = \varphi(x_0)$$
and let us define for $n \in \mathcal{N}$ the function ψ_n on $\overline{Q}$ by
$$\psi_n := \varphi - \gamma + \frac{1}{n}$$
By construction we have $\psi_n(x_0) \xrightarrow[n\to\infty]{} 0$ and thus $\log \psi_n(x_0) \xrightarrow[n\to\infty]{} \infty$. It is also clear that for any $n \in \mathcal{N}$ we have $\psi_n \in S_3$ so by Lemma 3.2.1, $[\log \psi_n] \in S_2$.
2. If we now assume that $[\varphi] \neq [1]$ then there exists x_1 such that $\varphi(x_1) \neq \gamma$ so the sequence $\{\log(\psi_n(x_1))\}_{n\in\mathcal{N}}$ is bounded. As a result $\|\log(\psi_n)\|_{\mathcal{C}^0/\mathbb{R}} \xrightarrow[n\to\infty]{} \infty$. That is a contradiction to the fact that S_2 is bounded in $\mathcal{C}^0(\overline{Q})/\mathbb{R}$. We may therefore conclude that $[\varphi] = [1]$ and the Proposition is proved.

□

3.5 Smoothness

At this point we have all the ingredients to prove our main result, the Theorem 1:

Proof of Theorem 1.

1. $\mathrm{T}'(m)$ is injective for all $m \in \mathrm{H}_0^{\left(\frac{1}{2},1\right)}$ because of the observation (2.2) and Theorem 5.
2. $\mathrm{T}'(m) = \mathcal{L} + S'(m)$ but $S'(m)$ is compact (Lemma 2.2.1) so the Fredholm index $\mathrm{ind}\big(\mathrm{T}'(m)\big) = \mathrm{ind}(\mathcal{L}) = 0$ since $\mathcal{L}$ is an isomorphism (cf. Theorem 3). Thus $\mathrm{T}'(m)$ is surjective. Since $\mathrm{T}'(m)$ is continuous, linear and bijective it is a homeomorphism.
3. We can use the inverse mapping theorem in Banach spaces (see [Hör03]) to assert that T is locally a diffeomorphism.
4. T is a surjection (Theorem 4) so it is a global diffeomorphism.

□

3.6 Extensions

Our method can be adapted to cover the case of non homogeneous Dirichlet boundary conditions as well as the case of periodic spatial boundary conditions (prescribing the momentum $\int u(t,x)\mathrm{d}x$ at $t = 0$ for the solution).

Bibliography

[Arn66] Arnold, V. *Sur la géométrie différentielle des groupes de Lie de dimension infinie et ses applications à l'hydrodynamique des fluides parfaits.* Ann. Inst. Fourier (Grenoble), 16(fasc. 1):319–361, 1966. Available from: http://www.numdam.org/item?id=AIF_1966__16_1_319_0.
(Cited p. 21)

[Bre93] Brezis, H. *Analyse fonctionnelle.* Sciences sup. Dunod, Paris, 1993. Théorie et applications. isbn:978-2-100-49336-4.
(Cited p. 29)

[Bur50] Burgers, J. M. *Correlation problems in a one-dimensional model of turbulence. I.* Nederl. Akad. Wetensch., Proc., 53:247–260, 1950.
(Cited p. 15)

[Col51] Cole, J. D. *On a quasi-linear parabolic equation occurring in aerodynamics.* Quart. Appl. Math., 9:225–236, 1951.
(Cited p. 15)

[Dei85] Deimling, K. *Nonlinear functional analysis.* Springer-Verlag, Berlin, 1985. isbn:978-3-540-13928-7.
(Cited p. 29)

[E99] E, W. *Aubry-Mather theory and periodic solutions of the forced Burgers equation.* Comm. Pure Appl. Math., 52(7):811–828, 1999. doi:10.1002/(SICI)1097-0312(199907)52:7<811::AID-CPA2>3.0.CO;2-D.
(Cited p. 15)

[FS04] Fontes, M., and Saksman, E. *Optimal results for the two dimensional Navier-Stokes equations with lower regularity on the data.* In *Actes des Journées Mathématiques à la Mémoire de Jean Leray*, volume 9 of *Sémin. Congr.*, pages 143–154. Soc. Math. France, Paris, 2004. Available from: http://www.emis.de/journals/SC/2004/9/html/smf_sem-cong_9_143-154.html, isbn:978-2-85629-160-3.
(Cited p. 16, 24)

[FS05] Fokas, A. S., and Stuart, J. T. *The time periodic solution of the Burgers equation on the half-line and an application to steady streaming.* J. Nonlinear Math. Phys., 12(suppl. 1):302–314, 2005. Available from: http://www.damtp.cam.ac.uk/user/tf227/Time-periodic.ps, doi:10.2991/jnmp.2005.12.s1.24.
(Cited p. 15)

[FV07] Fontes, M., and Verdier, O. *Time-periodic solutions of the burgers equation.* J. math. fluid mech., 2007. doi:10.1007/s00021-007-0260-z.
(Cited p. 16)

[Hop50] Hopf, E. *The partial differential equation $u_t + uu_x = \mu u_{xx}$.* Comm. Pure Appl. Math., 3:201–230, 1950. doi:10.1002/cpa.3160030302.
(Cited p. 15)

[Hör03] Hörmander, L. *The analysis of linear partial differential operators. I.* Classics in Mathematics. Springer-Verlag, Berlin, 2003. Distribution theory and Fourier analysis, Reprint of the second (1990) edition [Springer, Berlin; MR1065993 (91m:35001a)]. isbn:978-3-540-00662-6.
(Cited p. 39)

[JKM99] Jauslin, H. R., Kreiss, H. O., and Moser, J. *On the forced Burgers equation with periodic boundary conditions.* In *Differential equations: La Pietra 1996 (Florence)*, volume 65 of *Proc. Sympos. Pure Math.*, pages 133–153. Amer. Math. Soc., Providence, RI, 1999. isbn:978-0-8218-0610-4.
(Cited p. 15)

[KL89] Kreiss, H.-O., and Lorenz, J. *Initial-boundary value problems and the Navier-Stokes equations*, volume 136 of *Pure and Applied Mathematics.* Academic Press Inc., Boston, MA, 1989. isbn:978-0-12-426125-9.
(Cited p. 15)

[Sin91] Sinaĭ, Y. G. *Two results concerning asymptotic behavior of solutions of the Burgers equation with force.* J. Statist. Phys., 64(1-2):1–12, 1991. doi:10.1007/BF01057866.
(Cited p. 15)

Part II

Geometric Constraints

4

Geometric Index

In der ganzen modernen Mathematik ist die Theorie der Differentialgleichungen die wichtigste Disziplin

Lie, Leipziger Berichte, 1895

4.1 Introduction

The usual picture of a differential equation is that of a flow on a manifold. This is only one of many aspects of differential equations. There are in particular three levels at which to look at differential equations.

Vector fields

This is the most familiar way of looking at ordinary differential equations. One has an equation of the form:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = f(x)$$

where f is a *vector field* of a given manifold.

The vector field doesn't depend on time here but this is not a restriction since we may always add the trivial equation $\frac{\mathrm{d}t}{\mathrm{d}t} = 1$ and the new variable $\overline{x} := (t, x)$.

A not-so-innocent remark is that this implicitly models immobility, since immobility corresponds to the particular vector field zero. Technically speaking, from the space-time perspective, we have introduced a *splitting* of time and space. It makes sense to say that "position doesn't change" (or that "time doesn't change"). This splitting between time and space will be relevant in some of our applications but not in the development of the general theory.

Jet spaces

The most general way to describe an ordinary differential equation is by a number of equations:

$$F^j\Big(t, x^1, \dots, x^n, \frac{\mathrm{d}x^1}{\mathrm{d}t}, \dots, \frac{\mathrm{d}x^n}{\mathrm{d}t}\Big) = 0 \qquad 1 \le j \le m$$

which fixes a relation between the time t (the independent variable), x^i, the dependent variables, along with their derivatives $\frac{\mathrm{d}x^i}{\mathrm{d}t}$.

From now on we will denote the sequence x^i by x:

$$x \equiv (x^1, \dots, x^n)$$

Lie ([Lie85; Lie77; Sto00; Val45]) had already imagined a geometric theory, where these equations would define a surface (a manifold) as follows:

$$\mathcal{M} := \big\{(t, x, p) :\ F^j(t, x, p) = 0 \qquad 1 \le j \le m\big\}$$

in the space comprising $2n+1$ variables t, x^i, and p^i. Now the variables p^i are just extra variables, not yet related to x^i.

That *dynamics*[a] is provided by hyperplanes[b] at every point, which indicates the relations between p and x. The hyperplane corresponding to the coordinate i at the point (t_0, x_0, p_0) would take the form:

$$P^i(t_0, x_0, p_0) := \big\{(t, x, p) :\ (x^i - x^i_0) - p^i_0(t - t_0) = 0\big\} \tag{4.1}$$

At each point $(t_0, x_0, p_0) \in \mathcal{M}$ the *intersection* of the planes $P^i(t_0, x_0, p_0)$ models the dynamics.

Let us denote this direction by P:

$$P(t_0, x_0, p_0) := \bigcap_i P^i(t_0, x_0, p_0)$$

The intersection of P and the tangent plane of $\mathcal{M}$ at (t_0, x_0, p_0) produces a direction, tangent to the manifold $\mathcal{M}$ at this point. We use the word "direction" in a loose sense here, to mean any subspace at this point (not necessary a line).

Now a graph of a solution to the differential equation is obtained by integrating those directions on the surface $\mathcal{M}$. If those directions are not lines, there will possibly be more than one solution going through a given point.

What we presented so far is a general point of view of jet spaces that allows to handle any differential equation, partial or ordinary. It was already considered by Lie. See also [Olv93; Vin84] for similar points of view.

So, loosely speaking, the *dynamics* is given by fields of hyperplanes P^i, while the *constraints* are given by the functions F^j.

---

[a] we call "dynamics" the relation between the variables p^i and x^i

[b] called "contact planes"

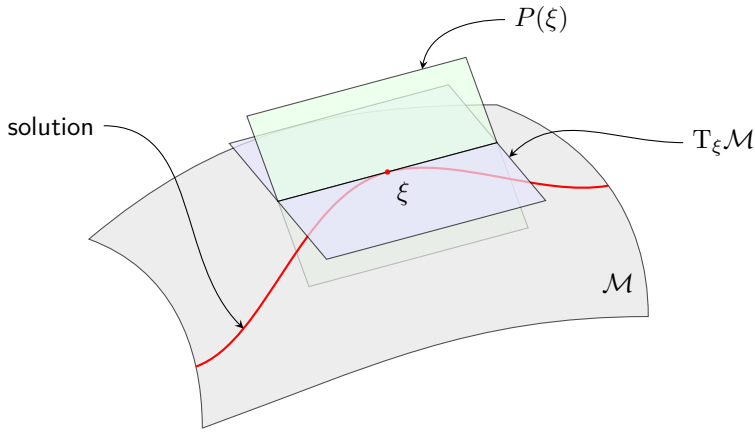


Figure 4.1: An illustration of a differential equation on a manifold.

Planes and differential forms

Let us emphasise that the planes $P^i(t_0, x_0, p_0)$ are only defined *infinitesimally close* to the point (t_0, x_0, p_0) on $\mathcal{M}$. More precisely, they are only defined on the tangent space at $\xi_0 = (t_0, x_0, p_0)$.

One convenient way to describe those hyperplanes is via differential forms. For instance the plane (4.1) is represented by the differential form

$$\theta^i(t_0, x_0, p_0) := \mathrm{d}x^i - p_0^i \,\mathrm{d}t$$

In general a differential form ω is expressed in coordinates (ξ^i), at the point ξ_0, as:

$$\omega(\xi_0) = \sum_i \omega_i(\xi_0)\,\mathrm{d}\xi^i$$

A differential form is nothing but a point dependent linear form. At every point its kernel defines an hyperplane. Differential forms are thus convenient to describe hyperplanes varying smoothly with respect to the position.

For the differential form above, the corresponding hyperplane would have the equation:

$$P(\xi_0) := \Big\{\xi : \sum_i \omega_i(\xi_0)(\xi^i - \xi_0^i) = 0\Big\}$$

Note again that those planes are only defined locally around the point ξ_0.

The previous construction, with variables (t, x, p) and contact planes (4.1), is referred to as the "jet-space" formulation. A submanifold of a jet-space naturally describes a differential equation.

Quasilinear equations

Our point of view will be slightly more general than the jet-space one. We will now consider any possible planes (not only the contact ones), and will thus no longer need the "third coordinates" p.

Let us consider the equation:

$$\sum_i e_i^j(t,x)\frac{\mathrm{d}x^i}{\mathrm{d}t} + a^j(t,x) = 0 \qquad 1 \leq j \leq m \tag{4.2}$$

This type of equation is called *quasilinear* because the highest order derivative appears linearly in the equation.

An equivalent way of formulating this equation is to directly describe the *dynamics* by the following differential forms:

$$\sum_i e_i^j(t,x)\mathrm{d}x^i + a^j(t,x)\mathrm{d}t \qquad 1 \leq j \leq m \tag{4.3}$$

As we see there, we have now lost any idea of a splitting between time and space. So, given a manifold that models space-time, *any* differential form is acceptable to describe the dynamics.

Now the graph of the solution of that equation will be a curve tangent to the following hyperplanes (cf. Figure 4.1):

$$P^j(t_0,x_0) := \left\{(t,x) :\ e^j(t,x)(x-x_0) + a^j(t,x)(t-t_0) = 0\right\} \qquad 1 \leq j \leq m$$

As before, we denote the intersection of those hyperplanes by:

$$P(t_0,x_0) := \bigcap_j P^j(t_0,x_0)$$

Vertical Solutions

Notice that going from (4.2) to (4.3) we have, formally, "multiplied by $\mathrm{d}t$". Doing this allows us to forget the distinction between the dependent and independent variables (time and space) but that comes at a cost. Intuitively, now, "$\mathrm{d}t$ is allowed to be zero".

Let us choose for instance $e(t,x) = 0$ and $a(t,x) = x$. The dynamics is thus modelled by the differential form:

$$x\,\mathrm{d}t \tag{4.4}$$

corresponding to the trivial differential equation

$$x = 0 \tag{4.5}$$

In that case, the curve $\mathcal{C}$ parametrized by:

$$s \mapsto (t_0, x_0 + s)$$

is a curve tangent to the plane field P at every point, and going through (t_0, x_0). We conclude that there is a solution to the problem described by (4.4).

However, the only relevant solution to (4.5) is obviously only $x = 0$. Clearly, the curve $\{x = 0\}$ is also a solution to (4.4).

We see the advantage of the formulation with differential forms: instead of restricting the set of possible initial conditions, we allow any initial condition, but we must be prepared to obtained unwanted, "vertical", solutions, i.e. solutions where "time stands still".

The next step is to define what those unwanted solutions exactly are.

Motivation for the Reduction Procedure

The simplest model of a differential equation is now given by a manifold $\mathcal{M}$ and some differential forms defined on it. It has thus become necessary to describe those unwanted, "vertical" directions, in order to eliminate the corresponding unwanted solutions. We therefore assume that we are given vertical directions at every point. In other words, at any given point we are given a hyperplane describing the "unwanted direction". For instance such a hyperplane could be given by:

$$T(t_0, x_0) := \big\{(t, x) : \ x - x_0 = 0\big\}$$

Two remarks are in order here:

First, this amounts to introduce a notion of *simultaneity*. This is not the same as having a complete splitting of time and space. In fact, this is not even equivalent to considering space-time as a bundle over the time line, if the field of hyperplanes T is not integrable.

Second, as a physicist would notice, not all physical theories allow to model simultaneity. In particular, the theory of relativity specifies other "forbidden directions" at every point (the ones with negative Lorentz length). Even though we will assume that the forbidden directions span a hyperplane at every point, the whole definition of the reduction procedure, described later in this Chapter, will go unchanged if one is given a *set* of directions at every point instead.

Now we are able to define a set where there are at least some non forbidden directions. This set is simply defined by the points where T and P are *transverse*:

$$\overline{\mathcal{M}} := \big\{\xi \in \mathcal{M} : \ P(\xi) \not\subset T(\xi)\big\}$$

In the event of $\overline{\mathcal{M}}$ being a *submanifold*, to make the point clearer, we denote that submanifold by:

$$\mathcal{M}' := \overline{\mathcal{M}}$$

The essential observation now is that we obtain a *new quasilinear equation* by *restricting* the differential forms on the submanifold $\mathcal{M}'$. To restrict a differential form on a submanifold simply means that its domain is now restricted to the tangent spaces of the new manifold $\mathcal{M}'$.

A simple example

Let us work out a simple example. Let $\mathcal{M} = \mathbb{R}^3$ with coordinates labelled as (t, x, y). The differential forms defining the dynamics are:

$$\begin{cases} \mathrm{d}y - x\,\mathrm{d}t \\ y\,\mathrm{d}t \end{cases} \tag{4.6}$$

Let us assume that the simultaneity directions are given by:

$$T(t_0, x_0, y_0) = \big\{(t, x, y) :\ t = t_0\big\}$$

It is easy to see that the intersection P of P_1 and P_2 is "vertical" as soon as $y_0 \neq 0$:

$$P(t_0, x_0, y_0) = \big\{(t, x, y) :\ t = t_0\big\} \subset T(t_0, x_0, y_0) \quad \text{if } y_0 \neq 0$$
$$P(t_0, x_0, 0) = \big\{(t, x, y) :\ y - x_0(t - t_0) = 0\big\}$$

So we see here that the set where we have hope to find non-vertical solutions is simply:

$$\overline{\mathcal{M}} := \big\{(t, x, y) :\ y = 0\big\}$$

In this simple case, this set turns out to be a submanifold of $\mathcal{M}$. Note that this will also be the case in most practical examples in this report.

According to the last section, all we have to do now is to compute the restriction of the differential forms (4.6) to the submanifold $\mathcal{M}' := \overline{\mathcal{M}}$. The tangent space of $\mathcal{M}'$ is easy enough to compute, because in the set of coordinates we are working with, it is a linear subspace. So in those coordinates, a tangent vector to $\mathcal{M}'$ has coordinates $(\alpha, \beta, 0)$. As a consequence, the differential form $\mathrm{d}y$ is zero, when restricted to $\mathcal{M}'$.

Obviously, the form $y\,\mathrm{d}t$ is also zero when restricted to $\mathcal{M}'$, so the new, "reduced", equation is now formulated as:

$$x\,\mathrm{d}t$$

A similar computation would lead to a new reduced set $\overline{\mathcal{M}'} = \{x = 0\}$, and it is again a submanifold. On this manifold, all the original differential forms are now zero when restricted on it. Any curve on $\mathcal{M}''$ is thus a solution. Since $\mathcal{M}''$ is one dimensional, it is the only solution to the problem.

4.2 Codistributions

This section is devoted to the definition of codistributions. Codistributions are a means to describe fields of subspaces of the tangent space. In simple cases, as a field of lines, or planes, the dimension does not change. In those cases one may describe those subspace fields by using spans of vector fields, or spans of covector fields.

However, when the dimension is allowed to change, it is crucial to choose either spans of vector fields (called "distribution" in the realm of differential geometry), or spans of covector fields, which are the ones that prove useful in the study of differential equations with constraints.

The reader willing to skip this section may content himself with the following definition of codistributions:

> A codistribution is a (point dependent) intersection of a fixed number of smoothly varying hyperplanes.

The reason why they occur naturally in this form, is that differential equations are usually provided by a list of smooth functions defined on a jet-space, being equal to zero. Each such function will give rise to a hyperplane by differentiation, the intersection of which giving the directions to integrate in order to construct the solutions.

Definition

Definition 4.1. A codistribution Ω is a mapping which for every point $\xi \in \mathcal{M}$ associates a subspace of $\mathrm{T}^*_\xi\mathcal{M}$. This mapping must also fulfill the following property:

For any $\xi \in \mathcal{M}$, there exists a neighbourhood N of ξ and a (possibly infinite) family of *smooth* covector fields $\{\varphi^a\}_{a\in A}$ on N such that:

$$\forall \zeta \in N \quad \Omega_\zeta = \mathrm{span}_{a\in A}\{\varphi^a_\zeta\} \tag{4.7}$$

Remark 4.2.1.

> As opposed to some other definitions of codistributions (e.g. [AMR88]), *we do not require the dimension of Ω_ξ to be constant with respect to ξ.* This will turn out to be important for the formulation of differential equations.

Operations on Codistributions

We also define the *sum* of two codistributions as the point-wise sum of the subspaces in the cotangent space, i.e.:

$$(\Omega_1 + \Omega_2)_\xi := (\Omega_1)_\xi + (\Omega_2)_\xi \qquad \forall \xi \in \mathcal{M}$$

The reader will be easily convinced that the result of this operation is a smooth codistribution, since it is generated by the sum of the generating covector fields of Ω_1 and Ω_2.

Similarly we will denote the *point-wise intersection* by:

$$(\Omega_1 \cap \Omega_2)_\xi := (\Omega_1)_\xi \cap (\Omega_2)_\xi$$

Note that this codistribution *need not be smooth*. For instance on $\mathbb{R}^2$ with coordinates (x, y), the intersection of $\Omega_1 = \text{span}\{\mathrm{d}y\}$ and $\Omega_2 = \text{span}\{x\,\mathrm{d}x + \mathrm{d}y\}$ is zero everywhere except on the line described by the equation $x = 0$. Such a codistribution cannot be smooth.

Pull-backs

Let us now recall the definition of a pull-back.

Definition 4.2. Let a smooth function φ maps a manifold $\mathcal{N}$ to a manifold $\mathcal{M}$

$$\varphi : \mathcal{N} \to \mathcal{M}$$

Recall that the derivative, or Jacobian, of φ is denoted by φ_*.

The ***pull-back*** of a one-form (or covector) $\theta \in \mathrm{T}^*\mathcal{M}$ is the one-form $\varphi^*(\theta)$ defined by:

$$\langle \varphi^*(\theta), X \rangle := \langle \theta, \varphi_*(X) \rangle \qquad \forall X \in \mathrm{T}\mathcal{N}$$

It is clear that we may then define pull-backs of codistributions by pulling-back the subspaces of the cotangent space at each point. Notice that *the pull-back of a smooth codistribution is always smooth.*

4.3 Implicit Differential Equation

General Definitions

Definition 4.3. An ***Implicit Differential Equation*** (***IDE***) is defined as the triple

$$(\mathcal{M}, \Omega, \tau)$$

where

- $\mathcal{M}$ is a smooth manifold
- Ω is a smooth co-distribution on $\mathcal{M}$
- τ is a smooth co-distribution on $\mathcal{M}$

Remark 4.3.1.

Conceptually we will regard this triplet as follows:

- $\mathcal{M}$ models *space-time*
- the co-distribution Ω models the *dynamics*
- the co-distribution τ models *simultaneity*

In all our examples the codistribution τ will be spanned by the differential form $\mathrm{d}t$ where t is the first variable in our coordinate system. It is only a *practical convention that we will follow in all our examples.*

In other words, in all our examples the coordinates will be labeled $(t, x^1, x^2, \ldots, x^d)$ we will assume the following:

Notation 4.3.1 (**Blanket hypothesis**)**.** Unless stated otherwise, the simultaneity codistribution τ will be defined by:

$$\tau = \mathrm{span}(\mathrm{d}t) \tag{4.8}$$

Example 4.3.1. Of course our framework applies to explicitly defined ordinary differential equations as well. Suppose the equation

$$\frac{\mathrm{d}x^i}{\mathrm{d}t} = f^i(t, x) \qquad 1 \leq i \leq d \tag{4.9}$$

has to be solved in $\mathbb{R}^d$. In our framework this can be written as the IDE:

$$\Omega = \mathrm{span}\left\{\mathrm{d}x^i - f^i(t, x)\,\mathrm{d}t;\ 1 \leq i \leq d\right\} \tag{4.10}$$

on the space

$$\mathcal{M} = \mathbb{R} \times \mathbb{R}^d = \mathbb{R}^{d+1}$$

and where the simultaneity codistribution is as usual:

$$\tau = \mathrm{span}(\mathrm{d}t)$$

where t denotes the first coordinate in $\mathbb{R}^{d+1}$.

Notation 4.3.2. We will use a special notation for IDEs throughout this report. The notation will be as follows: each time we will provide the number of space variables plus time, so that the total number of variables is $d + 1$.

Given one-forms θ_1, θ_2,..., θ_n defined on $\mathbb{R}^{d+1}$, and real-valued functions f_1, f_2,..., f_m defined from $\mathbb{R} \times \mathbb{R}^d$, we will write the IDE at hand as:

$$\begin{cases} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_n \\ f_1(t, x) = 0 \\ f_2(t, x) = 0 \\ \vdots \\ f_m(t, x) = 0 \end{cases}$$

By this we mean that:

$$\mathcal{M} := \{t, x \in \mathbb{R} \times \mathbb{R}^d : \ f_i(t,x) = 0 \quad \forall i \leq m\}$$

In all our examples, $\mathcal{M}$ will be a smooth submanifold of $\mathbb{R}^{d+1}$, and the Jacobian of the function f with coordinates $(f_1, f_2, \ldots, f_m)$ will have full rank.

$$\Omega := \text{span}_{i \leq n}(\theta_i)|_{\mathcal{M}}$$

(notice that this is the *restriction* of a certain codistribution on $\mathcal{M}$)

$$\tau := \text{span}(\mathrm{d}t)|_{\mathcal{M}}$$

Definition 4.4. A (generalized) ***solution*** to an IDE $(\mathcal{M}, \Omega, \tau)$ is a one dimensional manifold $\mathcal{C}$ and a smooth immersion:

$$\varphi : \ \mathcal{C} \to \mathcal{M}$$

such that the pull-back of Ω by φ is zero. In other words if:

$$\varphi^* \Omega = 0$$

Remark 4.3.2.

Note that the restriction (pull-back) of a codistribution is always well defined and smooth.

Remark 4.3.3.

Those solutions are "generalized" because the definition does not make use of the simultaneity co-distribution τ. In particular generalized solutions may be curves in space-time along which time is constant! Note that this concept was already considered by Lie [Sto00, §2.2]!

Example 4.3.2. A solution of the IDE set up in Example 4.3.1 is a curve in $\mathbb{R}^{n+1}$ representing the *graphs* of solutions in the ordinary sense. In other words, if the function x is a solution of (4.9), the curves parametrized by:

$$s \mapsto (s, x^1(s), \ldots, x^d(s))$$

are orthogonal to the codistribution (4.10).

In a similar manner, curves orthogonal to (4.10) must take the form above.

Practical Computations

In practice a submanifold is often given by the *locus* of a given function G, i.e. by the set:

$$\mathcal{N} := \{\xi \in \mathcal{M} : \ G(\xi) = 0\}$$

We shortly explain how to compute the restriction of a given codistribution Ω on $\mathcal{N}$ in this case.

Obviously for any tangent vector X to this submanifold $\mathcal{N}$ we have

$$\langle \mathrm{d}G, X\rangle = 0$$

Now if we denote the injection of $\mathcal{N}$ in $\mathcal{M}$ (which is also an *immersion* since $\mathcal{N}$ is a submanifold):

$$\mathrm{i} : \ \mathcal{N} \to \mathcal{M}$$

and in a coordinate system (which we denote by $t, x_1, \ldots, x_d$ if $\mathcal{M}$ had dimension $d+1$) this simply means that one obtains the following relations:

$$\frac{\partial G}{\partial t}\, \mathrm{i}^*(\mathrm{d}t) + \sum_{j=1}^{d} \frac{\partial G}{\partial x^j}\, \mathrm{i}^*(\mathrm{d}x^j) = 0$$

4.4 Jet spaces

In this section we show how to generate plenty of IDEs from differential equations, possibly with constraints.

Definition 4.5. Given a manifold $\mathcal{X}$ and an open interval $\mathcal{I} \subset \mathbb{R}$ we first trivially define the ***zero-th order jet space*** J_0 by

$$\mathrm{J}_0(\mathcal{I} \times \mathcal{X}) := \mathcal{I} \times \mathcal{X}$$

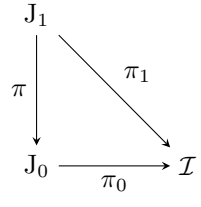
and the ***first order jet space*** J_1 by

$$\mathrm{J}_1(\mathcal{I} \times \mathcal{X}) := \mathcal{I} \times \mathrm{T}\mathcal{X}$$

Notice that if $\mathcal{Y} \subset \mathcal{X}$ is a submanifold of $\mathcal{X}$ one has the natural embedding:

$$\mathrm{J}_1(\mathcal{I} \times \mathcal{Y}) \hookrightarrow \mathrm{J}_1(\mathcal{I} \times \mathcal{X})$$

Note that the first jet space admits the following canonical projections:



Remark 4.4.1.

In what follows we will often denote coordinates of $\mathrm{J}_1(\mathcal{I},\mathcal{X})$ by $(t,x,\dot{x})$. The reader must be aware that Newton's notation $\dot{x}$ will **never** be a derivative. **All** the derivatives will be denoted either with Leibniz's symbol $'$, or explicitly with the usual notation:

$$f'(t) = \frac{\mathrm{d}f}{\mathrm{d}t}(t)$$

The jet space $\mathrm{J}_1(\mathcal{I}\times\mathcal{X})$ also admits a codistribution generated by the ***contact form*** which in coordinates are defined as:

$$\theta^i := \mathrm{d}x^i - \dot{x}^i\,\mathrm{d}t \qquad 1 \le i \le n \tag{4.11}$$

Remark 4.4.2.

One may define the contact distribution without resorting to the coordinates by considering the *lifting operation* from $\mathcal{I}\times\mathcal{X}$ to $\mathrm{J}(\mathcal{I}\times\mathcal{X})$. The lifting is defined as follows: for any section γ of the bundle $\mathcal{I}\times\mathcal{X}$ (a function from $\mathcal{I}$ to $\mathcal{X}$) we define its *lift* as being the curve

$$t \mapsto \left(t, \gamma(t), \frac{\mathrm{d}\gamma(t)}{\mathrm{d}t}\right) \in \mathrm{J}(\mathcal{X})$$

Taking the tangent of each such curve we thus obtain a vector of $\mathrm{TJ}_1(\mathcal{I}\times\mathcal{X})$. The linear span of all those vectors is sometimes called the *Cartan distribution* on the jet space $\mathrm{J}_1(\mathcal{I}\times\mathcal{X})$ ([Vin84]). The codistribution orthogonal to it is the ***contact codistribution***.

Definition 4.6. The jet space $\mathrm{J}_1(\mathcal{I}\times\mathcal{X})$ will now be implicitly equipped with an ***IDE structure*** given by the *contact codistribution* spanned by the contact forms (4.11) and the trivial simultaneity codistribution τ as in (4.8).

4.5 Pull-backs of IDEs

Introduction

We will often have to deal with differential equations stated in very similar ways. For instance what is the difference with the two following systems:

$$\begin{cases} \mathrm{d}x - p\,\mathrm{d}t \\ p = x \end{cases}$$

and the system simply described by the form:

$$\mathrm{d}x - x\,\mathrm{d}t$$

Clearly, those two systems are different, the second one sporting one less variable than the first. However, it is also clear that they are related, the first one being a reformulation of the second one with an auxiliary variable p.

The aim of this section is to give a precise account of how different IDE may be related. The key notion is that of *pull-back*. For example, the second system above is just the pull-back (i.e. "restriction") of the first system on the submanifold of equations $\{p = x\}$.

Definitions

Definition 4.7. Given a smooth map φ:

$$\mathcal{M}_1 \xrightarrow{\varphi} \mathcal{M}_0$$

and two IDEs

$$(\mathcal{M}_1, \Omega_1, \tau_1) \quad \text{and} \quad (\mathcal{M}_0, \Omega_0, \tau_0)$$

we define the ***pull-back*** of $(\mathcal{M}_0, \Omega_0, \tau_0)$ on $(\mathcal{M}_1, \Omega_1, \tau_1)$ by:

$$\varphi^*\big((\mathcal{M}_0, \Omega_0, \tau_0)\big) := (\mathcal{M}_1, \Omega_1 + \varphi^*\Omega_0, \tau_1 + \varphi^*\tau_0)$$

One way to see the pull-back operation is that it restricts another IDE $(\mathcal{M}_0, \Omega_0, \tau_0)$ using the mapping φ, and *adds* the structure obtained to the current IDE $(\mathcal{M}_1, \Omega_1, \tau_1)$. This will be especially useful when the IDE on which one pulls-back is the full jet-space (Definition 4.6), i.e. for instance $(\mathbb{R}^{2d+1}, \Omega, \tau)$, where Ω is the contact codistribution made of the contact forms $\mathrm{d}x^i - p^i\,\mathrm{d}t$, in the coordinates (t, x^i, p^i).

But the most frequent use will be to pull-back on submanifolds. To do this and still be able to use the definition above, we need to give any manifold an IDE structure. This is done via the notion of a trivial IDE:

Definition 4.8. A manifold $\mathcal{M}$ may always be equipped with a ***trivial IDE*** structure by choosing $\Omega = 0$ and $\tau = 0$.

In the sequel we will identify manifolds with trivial IDEs.

$$\mathcal{M} \equiv (\mathcal{M}, 0, 0) \tag{4.12}$$

Remark 4.5.1.

> Identifying manifolds with trivial IDEs allows us to define an IDE for *any submanifold* of a given IDE. Specifically, if
>
> $$\varphi : \mathcal{M}_1 \to \mathcal{M}_0$$

then the pull-back of the IDE $(\mathcal{M}_0, \Omega_0, \tau_0)$ on $\mathcal{M}_1$ is defined as the pull-back of $(\mathcal{M}_0, \Omega_0, \tau 0)$ on $(\mathcal{M}_1, 0, 0)$ (according to convention (4.12)) using Definition 4.7, i.e.:

$$\varphi^*(\mathcal{M}_0, \Omega_0, \tau_0) = (\mathcal{M}_1, \varphi^*\Omega_0, \varphi^*\tau_0)$$

Remark 4.5.2.

In particular, any submanifold of $\mathrm{J}_1(\mathcal{I} \times \mathcal{X})$ has now a natural structure of implicit differential equation following the convention of Remark 4.5.1.

Remark 4.5.3.

In the sequel this will mostly be used when φ is an *immersion*, its image thus defining a sub-manifold $\mathcal{N}$. In that case the pulled back IDE is simply given by the *restriction* of the codistributions Ω and τ on the submanifold $\mathcal{N}$.

4.6 Degeneracy

Degeneracy occurs at the points where the codimension of Ω is greater than one. The meaning of degeneracy is that the codistribution does no longer ensure uniqueness of solutions. We will see later however that degeneracy of the codistribution at a given point does not necessarily imply that several solutions pass through this point (see e.g. the Example 4.8.9).

Let us denote the set containing those points by:

$$\mathcal{N} := \left\{ x \in \mathcal{M} : \ \dim \Omega^{\perp} > 1 \right\}$$

Proposition 4.6.1. *$\mathcal{N}$ is a closed set.*

Proof. This is due to the fact that Ω is smooth, so the codimension of Ω may not locally increase. □

Definition 4.9. When $\mathcal{N}$ is a submanifold of $\mathcal{M}$ we will call it the ***degenerate manifold***.

Example 4.6.1. The equation of conservation of energy of an harmonic oscillator in one dimension is:

$$x^2 + \dot{x}^2 = 1$$

It is interesting to consider this equation as an IDE on the jet space $\mathrm{J}(\mathbb{R} \times \mathbb{R})$. We may parametrize this cylinder with $(t, \cos(\theta), \sin(\theta))$. The pull-back of the contact form $\mathrm{d}x - \dot{x}\,\mathrm{d}t$ is

$$-\sin(\theta)(\mathrm{d}\theta + \mathrm{d}t)$$

The solutions passing through the point $\big(t_0, \cos(\theta_0), \sin(\theta_0)\big)$ are thus the curves $(t, x(t), \frac{\mathrm{d}x}{\mathrm{d}t}(t))$ parametrized as:

$$x(t) = \cos(t - t_0 + \theta_0)$$

when $\sin(\theta) = 0$ there is another solution, parametrised by: $(t_0 + s, \pm 1, 0)$

Note that the points where $x = \pm 1$ are important also for numerical reasons. Let us investigate the solvability of a simple midpoint scheme:

$$\frac{(x_1 - x_0)^2}{h^2} + \frac{(x_1 + x_0)^2}{4} = 1$$

Using the notation $b := h/2$ and $\alpha = (1 - b^2)/(1 + b^2)$ we obtain:

$$(x_1 - x_0\alpha)^2 = \frac{4b^2}{(1 + b^2)^2}(1 - x_0^2 - b^2)$$

For this numerical equation to be solvable in x_1 we need the right-hand side to be positive, which will not be the case if $1 - x_0^2$ is too small. Note also that for such a scheme to work one has to keep track of the *sign* of $\dot{x}$.

The following three examples are taken from [Arn88, § 1.3.C]. See also [BG92, § 5.39, § 7.25]. For the Clairaut equation see also [HNW93, § I.2].

Example 4.6.2. The *Clairaut equation* is defined as

$$x = t\dot{x} - f(\dot{x})$$

The restriction of Ω is

$$\dot{x}\,\mathrm{d}t + t\,\mathrm{d}\dot{x} - f'(\dot{x})\,\mathrm{d}\dot{x} - \dot{x}\,\mathrm{d}t = (t - f'(\dot{x}))\,\mathrm{d}\dot{x}$$

We may define the *degenerate manifold*

$$\mathcal{N} := \{(t, x, \dot{x}) \in \mathcal{M} : \ t = f'(\dot{x})\}$$

The manifold $\mathcal{N}$ *is itself a (degenerate) solution*. Notice that when f is strictly convex, this regular solution is the *lift of the Legendre transformation of* f, defined as

$$g(t) := \sup_t (t\dot{x} - f(\dot{x}))$$

The other solutions, passing through points $(t_0, x_0, \dot{x}_0)$ outside $\mathcal{N}$ are given by:

$$s \mapsto (t_0 + s, (t_0 + s)\dot{x}_0 - f(\dot{x}_0), \dot{x}_0)$$

So in this case as in the earlier cylinder example the solutions are unique except on the singular manifold where there are exactly two of them.

Example 4.6.3.

$$\dot{x}^2 = x$$

This is an example similar to the two previous ones. The solution is unique except on the degenerate manifold

$$\mathcal{N} := \{x = 0 : \ \dot{x} = 0\}$$

$\mathcal{N}$ itself is a solution. The other solutions passing through $(t_0, \dot{x}_0^2, \dot{x}_0)$ are parametrized as:

$$s \mapsto (t_0 + 2s, (\dot{x}_0 + s)^2, \dot{x}_0 + s)$$

Notice that the degenerate solutions[c] are in fact *envelopes* of the regular solutions.

Example 4.6.4.

$$\dot{x}^2 = t$$

This is an example of an equation with "regular singular" points ([Arn88, § 1.4.E]). From our point of view there is no particular pathology: the solution through any point $(\dot{x}_0^2, x_0, \dot{x}_0)$ exists and is unique; it is parametrised by:

$$s \mapsto \left((\dot{x}_0 + s)^2, \frac{2}{3}\left((\dot{x}_0 + s)^3 - \dot{x}_0^3\right) + x_0, \dot{x}_0 + s\right)$$

However, the π_1-projections are *not* differentiable at $t = 0$ since their graph here is a *cusp*.

---

[c] also called "singular solutions" in [BG92]

4.7 The Reduction Procedure and the Index

Vertical Solutions

One of the possible pathologies that may occur is that the solutions be *vertical* with respect to the simultaneity co-distribution τ. Vertical directions are basically the directions where time does not change, in other words the directions which are orthogonal to τ.

Horizontal forms are forms which vanish on vertical directions. Those forms are "bad" in the sense that they force solutions to be vertical. Those vertical solution cannot be graphs of smooth time-depending functions.

Definition 4.10. We call a form θ ***horizontal*** if it belongs to the co-distribution τ:

$$\theta \in \tau$$

Notice that the set of horizontal forms of Ω is simply:

$$\Omega \cap \tau$$

We are able to state an elementary result on pull-backs of IDEs and horizontal forms:

Proposition 4.7.1. *Assume that the IDE $(\mathcal{M}, \Omega, \tau)$ is the pullback of $(\mathcal{M}_0, \Omega_0, \tau_0)$ by φ on $(\mathcal{M}_1, \Omega_1, \tau_1)$:*

$$\big(\mathcal{M}, \Omega_1 + \varphi^*(\Omega_0), \tau_1 + \varphi^*(\tau_0)\big) = \varphi^*(\mathcal{M}_0, \Omega_0, \tau_0)$$

Then

$$\varphi^*(\Omega \cap \tau) \subset \Omega_1 \cap \tau_1$$

Proof. The proof is a direct consequence of the general (elementary) fact that for two co-distribution Ω_0 and τ_0 one has:

$$\varphi^*(\Omega_0 \cap \tau_0) \subset \varphi^*(\Omega_0) \cap \varphi^*(\tau_0)$$

so we obtain the result since $\varphi^*(\Omega_0) \subset \Omega = \varphi^*(\Omega_0) + \Omega_1$, and $\varphi^*(\tau_0) \subset \tau = \varphi^*(\tau_0) + \tau_1$. □

Note that the inclusion in the theorem is not strict in general. See Example 4.8.2 for more details.

Reduction Procedure

One is rarely interested in vertical solutions because they do not correspond to any graph of any time-depending function. This is why one is led to the reduction procedure which eventually allows to define local solvability for *non-vertical* solutions.

Definition 4.11. We define the ***reduced set*** as the set $\overline{\mathcal{M}}$ defined by:

$$\overline{\mathcal{M}} := \left\{ \xi \in \mathcal{M} : \ (\Omega \cap \tau)_\xi = 0 \right\}$$

Remark 4.7.1.

> That reduced set obviously depends on the codistributions Ω and τ. However, in order to reduce the notational overhead, and if no confusion is possible, we will not mention this dependence in the notation.

Definition 4.12. Assume that $\mathcal{M}'$ is a submanifold of $\mathcal{M}$ such that:

$$\mathcal{M}' \subset \overline{\mathcal{M}}$$

The pull-back of Ω on $\mathcal{M}'$ is denoted by Ω'. In other words:

$$\mathcal{M}' \overset{\mathrm{i}}{\hookrightarrow} \mathcal{M} \qquad \Omega' = \mathrm{i}^*\Omega \tag{4.13}$$

By pulling back the simultaneity co-distribution τ one obtains a new IDE as pointed out in Remark 4.5.1 and Remark 4.5.3.

So from any choice of any submanifold contained in $\overline{\mathcal{M}}$ we obtain a new IDE. We may thus proceed further and define further reduction sets and continue until the process stops. This is the purpose of the *reduction chains*, which will help us to define the *index* of an IDE at every point.

4.8 Definition of the Index

Reduction Chains

Definition 4.13. For a point $\xi \in \mathcal{M}$ and a neighbourhood N of ξ, a ***reduction chain*** is a finite sequence of submanifolds, $\mathcal{M}^{(k)}$ for $0 \leq k \leq n$ of $\mathcal{M} \cap N$ such that:

1. $$\xi \in \mathcal{M}^{(k)} \qquad 0 \leq k \leq n$$

2. $$\overline{\mathcal{M}^{(n)}} = \mathcal{M}^{(n)} \subset \overline{\mathcal{M}^{(n-1)}} \subsetneq \mathcal{M}^{(n-1)} \subset \cdots \subset \mathcal{M}' \subset \overline{\mathcal{M}} \subsetneq \mathcal{M}$$

The *length* of the chain is the integer n.

The maximum of the lengths of all such possible chains in the neighbourhood N is the integer which will denoted by:

$$\operatorname{ind}_{\mathcal{M}}^{N} \xi$$

We are now able to define the index of a given point:

Definition 4.14. The ***index*** of a point $\xi \in \mathcal{M}$ is the maximum, over all neighbourhoods N of ξ, of the maximum over all the lengths of reduction chains in N at ξ, i.e. the index at ξ, $\operatorname{ind}_{\mathcal{M}} \xi$ is defined by:

$$\operatorname{ind}_{\mathcal{M}} \xi := \max_{N} \operatorname{ind}_{\mathcal{M}}^{N} \xi$$

Remark 4.8.1.

> The same remark as Remark 4.7.1 is in order. The index depends obviously on the codistributions Ω and τ. Again, we will not emphasise that dependency in the notation.

Example 4.8.1. For example, if $\overline{\mathcal{M}}$ is a neighbourhood of $\xi \in \mathcal{M}$ then

$$\operatorname{ind} \xi = 0$$

We may now also define the global index of an IDE:

Definition 4.15. The ***global index*** of an IDE is the maximum of all indices at all points:

$$\operatorname{ind}(\mathcal{M}, \Omega, \tau) := \max_{\xi \in \mathcal{M}} \operatorname{ind} \xi$$

Another important concept is that of a regular point, a point where all the reduced sets are submanifolds:

Definition 4.16. A point $\xi \in \mathcal{M}$ is said to be ***regular*** if there exists a neighbourhood N of ξ such that the successive reduced sets are in fact submanifolds, i.e. ξ has the following reduction chain:

$$\overline{\mathcal{M}^{(n)}} = \mathcal{M}^{(n)} = \overline{\mathcal{M}^{n-1}} \subset \mathcal{M}^{(n-1)} \subset \cdots \subset \mathcal{M}' = \overline{\mathcal{M}} \subset \mathcal{M}$$

We will call that chain the ***regular reduction chain*** at ξ.

Definition 4.17. An IDE is said to be ***regular*** if it is regular at all points.

Proposition 4.8.1. *At a regular point $\xi \in \mathcal{M}$ the index is the length of the regular reduction chain.*

Proof. The proof follows from an induction on the length of the regular reduction chain and the fact that if $\mathcal{N}$ is a submanifold of $\mathcal{M}$ then the reduction chains on $\mathcal{N}$ are at most as long as the reduction chains on $\mathcal{M}$. □

Example 4.8.2. A trivial example of reduction is given by the following linear system:

$$\begin{cases} \mathrm{d}y - x\,\mathrm{d}t \\ \mathrm{d}z - y\,\mathrm{d}t \\ z = 0 \end{cases}$$

It is straightforward to see that the reduced set are the manifolds $\mathcal{M}' := \{y = 0\} \cap \mathcal{M}$ and $\mathcal{M}'' := \{x = 0\} \cap \mathcal{M}'$. In particular, this IDE is regular (all linear IDEs are, see Proposition 5.1.2).

According to the definition of the reduced manifolds we have

$$\mathcal{M}' \overset{\mathrm{i}}{\hookrightarrow} \mathcal{M}$$

and the surjection:

$$\Omega \overset{\mathrm{i}^*}{\to} \Omega'$$

Notice now that by definition of the reduction: $\mathrm{i}^*(\Omega \cap \tau) = 0$ although $\Omega' \cap \tau' \neq 0$, so the inclusion in Proposition 4.7.1 is in general strict.

Example 4.8.3. We take an example from [CG95, Example 3] to illustrate the case of a non-regular IDE.

$$\begin{cases} \dot{x} = 1 \\ \dot{y} = z \\ 0 = xz - y \end{cases}$$

Those equations define a manifold $\mathcal{M}$ in the jet space $\mathrm{J}(\mathbb{R} \times \mathbb{R}^3) \equiv \mathbb{R}^7$ and hence an IDE by pull-back (Remark 4.5.2). The reduced set $\overline{\mathcal{M}}$ is defined by

$$\overline{\mathcal{M}} := \{x\dot{z} = 0\} \cap \mathcal{M}$$

Let us define the two submanifolds

$$\mathcal{N}_0 := \left\{\dot{z} = 0\right\} \cap \mathcal{M} \quad \text{and} \quad \mathcal{N}_1 := \left\{x = 0\right\} \cap \mathcal{M}$$

So we have:

$$\overline{\mathcal{M}} = \mathcal{N}_0 \cup \mathcal{N}_1$$

There are only vertical solutions on $\mathcal{N}_1$, parametrized by

$$\mathbb{R} \ni s \mapsto (t_0, 0, 0, z_0, 1, z_0, \dot{z}_0 + s)$$

As a result, the reduction of $\mathcal{N}_1$ is $\overline{\mathcal{N}_1} = \emptyset$. We conclude that the index of any point of $\mathcal{N}_1$ with respect to $\mathcal{N}_1$ is zero.

Now, the restriction of Ω on $\mathcal{N}_0$ generates a one-dimensional distribution which yields the solution passing through the point $(t_0, x_0, x_0 z_0, z_0, 1, z_0, 0)$:

$$(t_0 + s, x_0 + s, z_0(x_0 + s), z_0, 1, z, 0)$$

As a result, $\mathcal{N}_0$ is totally reduced, i.e. $\overline{\mathcal{N}_0} = \mathcal{N}_0$. We conclude that the index of any point of $\mathcal{N}_0$ with respect to $\mathcal{N}_0$ is zero.

We conclude that the index of all points in $\overline{\mathcal{M}}$ is one, whereas the index outside $\overline{\mathcal{M}}$ is zero. According to our definition of the index, the index of this system is one.

Remark 4.8.2.

The Example 4.8.3 above may be defined from the following IDE in $\mathbb{R}^6$ with coordinates $(t, x, y, z, \dot{x}, \dot{y})$:

$$\begin{cases} \mathrm{d}x - \mathrm{d}t \\ \mathrm{d}y - z\,\mathrm{d}t \\ \dot{x} = 1 \\ \dot{y} = z \\ 0 = xz - y \end{cases} \tag{4.14}$$

Consider now the projection π from the IDE $\mathrm{J}(\mathcal{I} \times \mathbb{R}^3)$ to $\mathbb{R}^6$ defined by "forgetting" the last component:

$$\pi(t, x, y, z, \dot{x}, \dot{y}, \dot{z}) := (t, x, y, z, \dot{x}, \dot{y})$$

the Example 4.8.3 is now the pull-back of (4.14) by π on $\mathrm{J}(\mathcal{I} \times \mathbb{R}^3)$. See also Example 4.8.9 for another pull-back of (4.14).

Example 4.8.4. The reduced manifold may also be an open subset of $\mathcal{M}$ (and thus have the same dimension). Take for example $\mathcal{M} = \mathbb{R}^2$ with coordinates (t, x). The codistribution is generated by $2x\,\mathrm{d}x - \mathrm{d}t$ (this corresponds to the differential equation $2x\dot{x} = 1$). The reduced manifold is

$$\mathcal{M}' := \{x \neq 0\}$$

Indeed, the solutions passing through (t_0, x_0) are parametrised as $s \mapsto ((x_0 + s)^2 - x_0^2 + t_0, x_0 + s)$. Those curves are not graphs of smooth functions on $\mathcal{M}'$.

Proposition 4.8.2. *A solution on a reduced manifold $\mathcal{M}'$ is also a solution of the IDE defined on $\mathcal{M}$.*

Proof. A solution is a manifold $\mathcal{C} \overset{\mathrm{i}}{\hookrightarrow} \mathcal{M}'$ such that $\mathrm{i}^*\,\Omega' = 0$. But with $\mathcal{M}' \overset{\mathrm{j}}{\hookrightarrow} \mathcal{M}$ we have by definition (cf. (4.13)) $\Omega' = \mathrm{j}^*\Omega$ so $(\mathrm{j}\circ\mathrm{i})^*\Omega = \mathrm{i}^*(\mathrm{j}^*(\Omega)) = 0$. □

Example 4.8.5. A simple example from [RLW01, Example 7] may illustrate the efficiency of calculus via pull-backs.

$$\begin{cases} t\dot{x}^2 - 2x\dot{x} + 9t^2 = 0 \\ \mathrm{d}x - \dot{x}\,\mathrm{d}t \end{cases}$$

Computing the pull-back of the contact form $\mathrm{d}x - \dot{x}\,\mathrm{d}t$ on $\mathcal{M}$ yields:

$$(2t\dot{x} - 3x)\,\mathrm{d}\dot{x} - 2(\dot{x}^2 - 9t)\,\mathrm{d}t$$

which immediately allows to conclude that:

- when $2t\dot{x} - 3x \neq 0$ there exists a unique solution through any point
- when $2t\dot{x} - 3x = 0$ then the solutions are vertical unless $\dot{x}^2 - 9t = 0$ which is a one dimensional manifold on which the pulled back contact form is zero so that manifold is a (singular) solution.

Remark 4.8.3.

The index computed this way might not be the same as the classical indices. Indeed, the equation:

$$\begin{cases} \dfrac{\mathrm{d}x}{\mathrm{d}t} = f(x,y) \\ g(x,y) = 0 \end{cases}$$

where g_y is invertible is considered to have index 1 in most definitions of the index. In following the convention of the Notation 4.3.2, *this equation has index* 0!

Indeed by the implicit function theorem we may solve the variable y in function of x, in other words, locally, there exists a function h such that:

$$g(x,y) = 0 \iff y = h(x)$$

Now the equation above is simply:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = f(x, h(x))$$

so it seems appropriate to regard it as an explicit ordinary differential equation.

The index discrepancy is due to the confusion between the two following systems:

$$\begin{cases} \mathrm{d}x - f(x,y)\,\mathrm{d}t \\ g(x,y)\,\mathrm{d}t \end{cases} \quad \text{and} \quad \begin{cases} \mathrm{d}x - f(x,y)\,\mathrm{d}t \\ g(x,y) = 0 \end{cases}$$

The first system (the manifold $\mathcal{M}$ is the entire space) has index 1 whereas the second system (with $\mathcal{M} = \{g(x,y) = 0\}$) has index 0. Indeed the second system is the reduced version of the first!

Notice that the discretisation of those systems are entirely different. For example the Radau method ([HLR89, § 2]) for the second system is *nothing but the usual Runge-Kutta method* on a differential equation!

In fact, one often assumes that an index-1 system is in "Hessenberg form", meaning that the constraint equations are separated from the proper differential equations. This is another way of saying that the system is reduced to an index-0 form.

Index-0 may also lead to discretization problems, but milder ones. For instance it may lead to unexpected time-step restrictions ([HMT03]).

More serious problems crop up when the index is one.

Example 4.8.6. In the literature the following type of system is often mentioned:

$$\begin{cases} \mathrm{d}x - f(x,y)\,\mathrm{d}t \\ g(x) = 0 \end{cases}$$

where $g'f_y$ is assumed to be invertible.

In that case, this system has index one. Indeed it is clear that

$$\mathcal{M}' = \{x, y : \ g'(x)f(x,y) = 0\}$$

Now if $g'f_y$ is invertible then

$$\Omega' \cap \tau = 0$$

so $\mathcal{M}'' = \mathcal{M}'$ and this system has index one.

Example 4.8.7. An example of a system which index should be considered to be 0 is the *singular van der Pol equation*. In a reduced form (the system comes from a singular differential equation) the system, in $\mathbb{R}^3$ with coordinates (t, x, y):

$$\begin{cases} \mathrm{d}x - y\,\mathrm{d}t \\ (1-x^2)y - x = 0 \end{cases}$$

The pull-back of the one-form above is:

$$(1-x^2)\,\mathrm{d}y - (2xy+1)y\,\mathrm{d}t$$

The function x^2-1 is nonzero on the manifold $\mathcal{M}$ so the system is totally reduced (i.e. $\mathcal{M} = \mathcal{M}'$). In fact it is clear that this system is equivalent to the differential equation:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \frac{x}{1-x^2}$$

so it makes sense to regard it as an index 0 IDE.

Example 4.8.8. A classical example that illustrates the difficulties in the numerical discretisation is the following ([HLR89, § 2] and [GP84, §3]).

$$\begin{cases} \mathrm{d}x + \eta t\,\mathrm{d}y + (1+\eta)y\,\mathrm{d}t - g(t)\,\mathrm{d}t \\ x + \eta t y = f(t) \end{cases}$$

This problem may rewritten as

$$\begin{cases} x + \eta t y = f(t) \\ \mathrm{d}(x + \eta t y - f(t)) - (g(t) - y - f'(t))\,\mathrm{d}t \end{cases}$$

it has thus index one and the reduced manifold is $\mathcal{M}' = \{g(t) - f'(t) - y = 0\} \cap \mathcal{M}$.

Index and pull-backs

The purpose of reduction is to obtain another system with a lower index. We illustrate this with a couple of examples.

Example 4.8.9. It is instructive to revisit Example 4.8.3. Formulated as it was, that IDE had index one. Let us now look at a similar IDE given by the *pull-back* of (4.14) by the following injection mapping:

$$\varphi(t, x, y, z) = (t, x, y, z, 1, z)$$

We obtain:

$$\begin{cases} \mathrm{d}x - \mathrm{d}t \\ \mathrm{d}y - z\,\mathrm{d}t \\ 0 = xz - y \end{cases}$$

In that case *this system has index zero*! This is clear since, provided that $x \neq 0$ the system above may be written as:

$$\begin{cases} \dfrac{\mathrm{d}x}{\mathrm{d}t} = 1 \\ \dfrac{\mathrm{d}y}{\mathrm{d}t} = \dfrac{y}{x} \end{cases}$$

so it is an ordinary differential equation in explicit form.

The pull back of the codistribution on $\mathcal{M}$ contains the element $x\,\mathrm{d}z$, which means that there is a degenerate manifold $\mathcal{N} = \{x = 0\} \cap \mathcal{M}$. On the other hand there is no distinct reduced manifold, i.e. $\mathcal{M}' = \mathcal{M}$. Outside $\mathcal{N}$ one finds the unique standard solution

$$s \mapsto (t_0 + s, x_0 + s, z_0(x_0 + s), z_0)$$

on $\mathcal{N}$ there are other solutions, vertical ones, given by

$$s \mapsto (t_0, 0, 0, z_0 + s)$$

In other words this second point of view is very similar to the Example 4.6.1. A crucial difference though is that the degenerate manifold $\mathcal{N}$ is not a solution since the pullback of the codistribution on it has full rank.

Example 4.8.10. Notice that the formulation of a DAE on a jet space as a general IDE may *change the index*. An example of an IDE on a jet space is

$$\begin{cases} \dot{x} = y \\ \dot{y} = z \\ x = g(t) \end{cases} \qquad (\spadesuit)$$

The codistribution is, as usual, the contact codistribution. This IDE has index *3* since the reduced manifolds are $\mathcal{M}' = \{g'(t) = y\} \cap \mathcal{M}$, $\mathcal{M}'' = \{g''(t) = z\} \cap \mathcal{M}'$ and $\mathcal{M}''' = \{g'''(t) = \dot{z}\}$.

Consider now the seemingly equivalent IDE defined on $\mathbb{R}^4$:

$$\begin{cases} \mathrm{d}x - y\,\mathrm{d}t \\ \mathrm{d}y - z\,\mathrm{d}t \\ x = g(t) \end{cases} \qquad (\clubsuit)$$

The relation with (♠) is that (♠) is the pull-back of (♣) on the submanifold $\mathcal{M} = \{\dot{x} = y;\ \dot{y} = z;\ x = g(t)\}$ of the jet-space with three variables by the following projection:

$$\pi(t, x, y, z, y, z, \dot{z}) := (t, x, y, z, y, z)$$

The IDE (♣) now has only index *2* since the reduced manifolds are $\mathcal{M}' = \{g'(t) = y\} \cap \mathcal{M}$ and $\mathcal{M}'' = \{g''(t) = z\} \cap \mathcal{M}'$.

This reflects the fact that z is, in a sense, a *Lagrange multiplier* corresponding to the constraint. For this reason it does not really make sense to have a differential equation for this variable.

Solvability

An important aspect of IDEs is of course solvability that we only define for regular IDEs for simplicity:

Proposition 4.8.3. *A regular IDE will be said to be* ***solvable*** *if the corresponding totally reduced IDE has exactly one (non-vertical) solution passing through every point.*

Example 4.8.11. Here is an example where the degeneracy strictly increases during the reduction

$$\begin{cases} \mathrm{d}z + (x + y)\,\mathrm{d}t \\ \mathrm{d}x + \mathrm{d}y \\ z\,\mathrm{d}t \end{cases}$$

The reduced manifold is obviously $\mathcal{M}' = \{z = 0\}$. Similarly one finds $\mathcal{M}' = \{x + y = 0\} \cap \mathcal{M}$. However, on $\mathcal{M}''$, the form $\mathrm{d}x + \mathrm{d}y$ is pulled back to zero so this system is degenerate.

Since this system is linear we will see later that we can describe exactly how this system fails to be solvable using normal forms (cf. Example 5.6.2).

4.9 Reduction in Jet Spaces

Relation with the Projection Approach

In [Rei90, §4] the author defines the reduction process by *projecting* using the projection π_1. A detailed investigation of this approach is made in [RR94]. This

approach is named by some authors as the Rabier-Rheinboldt reduction [RLW01; MT98]. We adapt that approach to the jet-space setting (as in [RLW01]) and we will show that our definition is broader than theirs.

Example 4.9.1. For an example of the fact that the index defined in [CG95, Definition 6] (called *uniform index*) is not geometric consider the example in [CG95, Example 9]:

$$\begin{cases} \dot{y}y + x = 0 \\ y = 0 \end{cases}$$

Note that the *manifold* defined by those equation is exactly

$$\begin{cases} x = 0 \\ y = 0 \end{cases}$$

But those two formulations have different uniform indices. The index defined in [CG95], according to the authors, is 2 for the first system whereas it is 1 in the second case. In our approach the index is 1 independently of the equations defining the manifold.

Example 4.9.2. Consider the jet space $\mathrm{J}(\mathcal{I} \times \mathbb{R}^3)$ with coordinates:

$$(t, x, y, z, \dot{x}, \dot{y}, \dot{z})$$

and the submanifold $\mathcal{M}$ defined by the equations:

$$\begin{cases} x = g(\dot{z})^2 \\ y = g(\dot{z})^3 \\ z = 0 \end{cases}$$

where g is a smooth function from $\mathbb{R}$ to $\mathbb{R}$.

The restriction of the contact codistribution Ω on $\mathcal{M}$ is given by

$$\begin{aligned} 2g'(\dot{z})\,\mathrm{d}\dot{z} - \dot{x}\,\mathrm{d}t \\ (3g(\dot{z})\dot{x} - 2\dot{y})\,\mathrm{d}t \\ -\dot{z}\,\mathrm{d}t \end{aligned}$$

The reduced manifold is thus

$$\mathcal{M}' := \Big\{(t, x, y, z, \dot{x}, \dot{y}, \dot{z}) \in \mathcal{M} :\ \dot{z} = 0,\ 3g(\dot{z})\dot{x} = 2\dot{y}\Big\}.$$

which can be reformulated as

$$\begin{cases} x = g(0)^2 \\ y = g(0)^3 \\ z = 0 \\ 3g(0)\dot{x} = 2\dot{y} \\ \dot{z} = 0 \end{cases}$$

Outside $\mathcal{M}'$ there are by definition only vertical solutions; here, if $g'(\dot{z}_0) \neq 0$ there are many of them: any curve in the vertical plane:

$$\mathbb{R}^2 \ni (u, v) \mapsto \Big(t_0, g(\dot{z}_0)^2, g(\dot{z}_0)^3, 0, \dot{x}_0 + u, \dot{y}_0 + v, \dot{z}_0\Big)$$

is a vertical solution. More generally, vertical solutions are of the form

$$s \mapsto (t_0, g(\dot{z}_0)^2, g(\dot{z}_0)^3, 0, \dot{x}(s), \dot{y}(s), \dot{z}(s))$$

where the constraint on the function $s \mapsto \dot{z}(s)$ is that $g(\dot{z}(s)) = 0$ (indeed if $g'(\dot{z}_0) \neq 0$ then $\dot{z}'(s) = 0$).

There are vertical solution on $\mathcal{M}'$ too, parametrized by

$$s \mapsto \Big(t_0, g(0)^2, g(0)^3, 0, \dot{x}_0 + s, \frac{3}{2}g(0)(\dot{x}_0 + s), 0\Big)$$

$\mathcal{M}'$ is further reduced to

$$\mathcal{M}'' = \mathcal{M}' \cap \{\dot{x} = 0\}$$

Now $\mathcal{M}''$ has dimension one and the restriction of Ω on it is zero. The only solution is thus $\mathcal{M}''$ itself, i.e. the constant solutions $s \mapsto (t_0+s, g(0)^2, g(0)^3, 0, 0, 0, 0)$.

Remark 4.9.1.

Our approach is *more general* than that of Rabier and Rheinboldt. Indeed in the Example 4.9.2 if $g(0) = 0$ and $g'(0) \neq 0$ (take for instance $g(\dot{x}) = \dot{x}$) then the π_1 projection is a cusp with singular point along $\{x = y = 0\}$. But this set turns out to contain the (non-vertical) solutions. Since in the approach of Rabier and Rheinboldt such singularities are excluded they would conclude that this system has no solutions. In our framework this system has a solution and a well-defined index.

Example 4.9.3. We now present another example where the approach of Rabier and Rheinboldt fails. Consider $\mathbb{R}^6$ parametrized with the variables $(t, x, y, z, \dot{y}, \dot{z})$ and the following system:

$$\begin{cases} \mathrm{d}y - \dot{y}\,\mathrm{d}t \\ \mathrm{d}z - \dot{z}\,\mathrm{d}t \\ \cos(\dot{y}) = x \\ \cos(\dot{z}) = y \\ z = 1 \end{cases}$$

In that case the successive reduced manifolds are $\mathcal{M}' = \{\dot{z} = 0\} \cap \mathcal{M}$ and $\mathcal{M}'' = \{\dot{y} = 0\} \cap \mathcal{M}'$. The index is thus 2 and $\mathcal{M}''$ is the only solution. In other words the solutions may be trivially parametrised as $s \mapsto (t_0 + s, 1, 1, 1, 0, 0)$.

What makes this example interesting is that the π_1 projection of $\mathcal{M}$ on the π_1-base space is *not a subimmersion* precisely at the interesting points, for example at $\{\dot{z} = 0\}$. The algorithm of [RR94] will thus conclude that at points where the projection is a subimmersion there is no solution. Our approach allows to correctly compute the index and to find the correct solution.

Notice also that the approach of [CM95] (see also [HW96, § VII.2]) also fails in this case. Indeed one part of the assumption is that the projection of the prolonged manifold on the π_1-base of the jet space has constant rank which is not the case here.

4.10 Notes

One of the first attempts at a systematic study of the geometric properties of differential algebraic equations is to be found in [Rei90]. It was later reformulated by projections in jet spaces[(d)] in [RR94].

Differential algebraic equations are special cases in the Cartan-Kuranishi theory of solvability of differential equations (see [AVL91, § 5.6.1], or [CHS00] for an elementary introduction, and [BCG$^+$91; Kur57; Her65] for detailed accounts). Several connexions with DAEs are explained in [Sei99]. The connexion between the approach of Rabier and Rheinboldt and the Cartan-Kuranishi prolongation is detailed in [RLW01].

Note however that the latter approach is more algebraic in nature. It is a study of the properties of the equations themselves, rather than the properties of their locus.

An attempt was made at defining reduction in a purely geometric fashion in [MLRR99]. The authors failed however to notice the recursive structure of the reduction.

[d] To be precise, in tangent bundles, but the generalization to jet-spaces is straightforward.

The idea of considering a simultaneity codistribution is very closely related to the idea of *independency condition*, as it appears in [BCG$^+$91]. It was considered more generally as a codistribution in [Her65, p. 281] in the same form as ours.

Our definition of the index in such general terms is new, because the usual requirements are much harsher than ours. In particular, the Cartan-Kuranishi theory makes repeated use of the Cauchy-Kowalewska theorem which requires analytic regularity. Besides, the Pfaffian system studied is often assumed to have constant dimension ([Olv95; vNRM98]), which discards any possibility of reduction as we described it. This is not so surprising given that the focus of the Cartan-Kuranishi theory is shifted towards the integrability condition called *involutivity*, which is relevant only for partial differential equations.

The notion of pull-back of IDE is new, and it is a convenient tool to investigate the precise relation between different formulations of the same IDE. A related notion, but which does not apply for general IDEs, is that of absolute equivalence of Pfaffian systems in [vNRM98], related to *differential flatness* and *dynamic feedback equivalence*.

5

Linear Autonomous IDE

> [L'art] de démontrer les vérités déjà trouvées et de les éclaircir de telle sorte que la preuve en soit invincible est le seul que je veux donner ; et je n'ai pour cela qu'à expliquer la méthode que la géométrie y observe ;
>
> Pascal, Pensées, 1670

Introduction

We study in details the particular case of IDEs which may be written with linear time-independent matrices in some coordinates. We show how the geometric reduction naturally leads to a canonical form equivalent to that of Kronecker.

This chapter is independent of the previous one, with the notable exception of the first section.

5.1 Geometric Setting

In this section we consider the ambient manifold to be

$$\mathcal{M} = \mathcal{I} \times M$$

where M is a finite dimensional ***linear space*** and $\mathcal{I} \subset \mathbb{R}$ as usual.

We consider two linear operators E and A from M to an other finite dimensional linear space V (which we will refer to as the *codomain*).

$$\mathsf{A}, \mathsf{E} : M \to V$$

We proceed to define an IDE by giving a precise meaning to the equation:

$$\mathsf{E}\,\mathrm{d}\mathsf{x} + \mathsf{A}\mathsf{x}\,\mathrm{d}t$$

which will be the natural way to regard the ordinary differential equation

$$\mathsf{E}\frac{\mathrm{d}\mathsf{x}}{\mathrm{d}t} + \mathsf{A}\mathsf{x} = 0$$

as an IDE.

Vector-valued Differential Forms

Given a finite dimensional vector space V considered as a manifold, its tangent bundle TV is canonically "parallelized", in the sense that one can, in a canonical way, translate a tangent vector to a reference point, say the origin zero. It will be easy to define the differential of forms taking values in vector spaces (or, in fact, in any parallelized manifold).

More precisely, there exists a canonically defined parallelization mapping P_V which is a vector bundle morphism:

$$P_V : \mathrm{T}V \longrightarrow V$$

where V is identified with the vector bundle over a base of dimension zero. In other words, P is linear when acting on the tangent vectors of V.

The ***differential*** d of a mapping

$$f : \mathcal{M} \longrightarrow V$$

is defined by:

$$\mathrm{d}f := f^* P_V$$

In other words the function $\mathrm{d}f$ is defined on a tangent vector X of $\mathcal{M}$ at $\xi \in \mathcal{M}$ as:

$$\langle \mathrm{d}f, X \rangle := P_V(f_*(X))$$

Note that when the vector space V happens to be the space of real numbers $\mathbb{R}$, then this definition coincides with the usual definition of the differential of a scalar function[a].

Formal Definition of a Linear IDE

Let us define the projections of $\mathcal{I} \times M$ on $\mathcal{I}$, respectively M by t, respectively x. We may now define properly the *vector valued differential form* θ:

$$\theta = \mathsf{E}\,\mathrm{d}\mathsf{x} + \mathsf{A}\mathsf{x}\,\mathrm{d}t$$

[a] In fact, some authors ([Mal72]) define the differential of a scalar function in this way.

This form defines a codistribution in the following way. For each vector $\varphi \in V^*$, we obtain the (scalar) one-form:

$$\theta_\varphi := \varphi \circ \theta$$

and the codistribution Ω corresponding to the linear IDE at hand may now be written as:

$$\Omega := \operatorname{span}\{\theta_\varphi;\ \varphi \in V^*\}$$

In other words, if a tangent vector $X \in \mathrm{T}_\xi\mathcal{M}$ is split into a time part and a space part as

$$X = (s, y) \in \mathbb{R} \times M \equiv \mathrm{T}_\xi\mathcal{M}$$

where $s = P_{\mathbb{R}} t_*(X)$ and $y = P_M \mathsf{x}_*(X)$ then the requirement of it being orthogonal to the codistribution is simply:

$$\mathsf{E}y + \mathsf{A}\mathsf{x}(\xi)s$$

Remark 5.1.1.

> Notice that at each fixed point $\xi \in \mathcal{M}$ the form θ is a mapping from $\mathrm{T}_\xi\mathcal{M}$ to V, so at each point, its kernel defines a subspace. The codistribution Ω is its orthogonal:
>
> $$\Omega = (\ker\theta)^\perp$$

In the sequel we will use the functions t and x as *coordinates*, following the tradition of differential geometry, although with the vector-coordinate x.

Reduction

Proposition 5.1.1. *The reduced set is a manifold and may be written as*

$$\mathcal{M}' = \mathcal{I} \times M' \qquad (\diamondsuit)$$

where M' is the linear subspace

$$M' := \{\mathsf{x} \in M :\ \mathsf{A}\mathsf{x} \in \operatorname{Im}\mathsf{E}\} = \ker\big((\mathsf{E}M)^\perp \mathsf{A}\big)$$

Proof. According to the definition of the reduced set in Definition 4.11 a point $(t, \mathsf{x}) \in \mathcal{M}$ is *not* in $\overline{\mathcal{M}}$ iff:

$$\exists \mathsf{q} \in V^* \qquad \mathsf{q}\mathsf{E}\,\mathrm{d}\mathsf{x} + \mathsf{q}\mathsf{A}\mathsf{x}\,\mathrm{d}t = \mathrm{d}t$$

i.e.:

$$\exists \mathsf{q} \in V^* \quad \text{s.t.} \quad \begin{cases} \mathsf{q}\mathsf{E} = 0 \\ \mathsf{q}\mathsf{A}\mathsf{x} \neq 0 \end{cases}$$

which is equivalent to:

$$\exists \mathsf{q} \in (\mathsf{E}M)^{\perp} \quad \mathsf{q}\mathsf{A}\mathsf{x} \neq 0$$

Reformulating the last line we obtain:

$$(t, \mathsf{x}) \in \overline{\mathcal{M}} \quad \Longleftrightarrow \quad \forall \mathsf{q} \in (\mathsf{E}M)^{\perp} \quad \mathsf{q}\mathsf{A}\mathsf{x} = 0$$

□

Remark 5.1.2.

> One of the first occurrence of the definition of that subspace M' is to be found in [Rei92, §7], although with a different purpose than ours.

Recalling the Definition 4.17 of regularity of an IDE, the last Proposition has the following immediate consequence:

Proposition 5.1.2. *A linear IDE is regular.*

Proposition 5.1.1 paves the way to the description of reduced systems, which is the aim of the two following Proposition:

Proposition 5.1.3. *The linear system* (E, A) *is totally reduced (i.e.* $M' = M$*) iff*

$$\mathsf{A}M \subset \mathsf{E}M$$

Proposition 5.1.4. *An equation for the reduced system is given by*

$$\mathsf{E}|_{M'}\, \mathrm{d}\mathsf{x} + \mathsf{A}|_{M'}\mathsf{x}\, \mathrm{d}t$$

where $\mathsf{E}|_{M'}$ *and* $\mathsf{A}|_{M'}$ *are the restrictions of respectively* E *and* A *on* M'*.*

Proof. The result follows from the observation that if i is the natural inclusion of M' in M then:

$$\mathrm{i}^*(\mathsf{E}\,\mathrm{d}\mathsf{x}) = \mathsf{E}\,\mathrm{d}(\mathrm{i}\,\mathsf{x}) = \mathsf{E}\,\mathrm{i}\,\mathrm{d}\mathsf{x}$$

□

5.2 Determinacy Degree

Definitions

We define an integer which gives an indication of whether the system is under or over determined (or neither). Loosely speaking it is the number of equations minus the number of variables.

Definition 5.1. The ***determinacy degree*** Δ is defined to be

$$\Delta := \dim(\mathsf{A}M + \mathsf{E}M) - \dim M$$

This allows us to precisely define over and under-determinacy for a linear system:

Definition 5.2. A linear IDE (E, A) will be called:

- ***overdetermined*** if $\Delta > 0$
- ***well-determined*** if $\Delta = 0$
- ***underdetermined*** if $\Delta < 0$

We may now show that the determinacy degree decreases with the reduction. See also Proposition 5.4.4 for further details.

Proposition 5.2.1. *The determinacy degree* decreases *with the reduction:*

$$\Delta' \leq \Delta$$

where Δ' *is the determinacy degree of the reduced system defined by* $(\mathsf{E}|_{M'}, \mathsf{A}|_{M'})$*, i.e.:*

$$\Delta' = \dim(\mathsf{A}M' + \mathsf{E}M') - \dim M' \tag{5.1}$$

The proof will be a direct consequence of this Lemma:

Lemma 5.2.1.

$$\Delta' = \dim(\mathsf{A}M + \mathsf{E}M') - \dim M$$

Proof.

1. First notice that
$$\begin{aligned} \mathsf{A}M' &= \{\mathsf{A}x : \ \mathsf{A}x \in \mathsf{E}M\} \\ &= \mathsf{A}M \cap \mathsf{E}M \end{aligned}$$
2. It is clear that $\ker \mathsf{A} \subset M'$ so $\ker \mathsf{A} = \ker(\mathsf{A}|_{M'})$.
3. We thus obtain:
$$\begin{aligned} \dim M' &= \dim \ker(\mathsf{A}|_{M'}) + \dim \mathsf{A}M' \\ &= \dim \ker \mathsf{A} + \dim(\mathsf{A}M \cap \mathsf{E}M) \end{aligned} \tag{♣}$$
4.
$$\begin{aligned} \dim(\mathsf{A}M' + \mathsf{E}M') &= \dim \mathsf{A}M' + \dim \mathsf{E}M' - \dim(\mathsf{A}M' \cap \mathsf{E}M') \\ &= \dim(\mathsf{A}M \cap \mathsf{E}M) + \dim \mathsf{E}M' - \dim(\mathsf{A}M \cap \mathsf{E}M') \end{aligned} \tag{♠}$$

5. Using (♣) and (♠) and the definition of Δ' (5.1) we obtain:

$$\Delta' = \dim \mathsf{E}M' - \dim(\mathsf{A}M \cap \mathsf{E}M') - \dim \ker \mathsf{A}$$

6. Using

$$\dim M = \dim \ker \mathsf{A} + \dim \mathsf{A}M$$

and

$$\dim(\mathsf{A}M + \mathsf{E}M') = \dim \mathsf{A}M + \dim \mathsf{E}M' - \dim(\mathsf{A}M \cap \mathsf{E}M')$$

we obtain:

$$\Delta' = \dim(\mathsf{A}M + \mathsf{E}M') - \dim M$$

□

Proof of Proposition 5.2.1. Using Lemma 5.2.1 we obtain directly:

$$\begin{aligned}\Delta' &= \dim(\mathsf{A}M + \mathsf{E}M') - \dim M \\ &\leq \dim(\mathsf{A}M + \mathsf{E}M) - \dim M = \Delta\end{aligned}$$

□

The meaning of Proposition 5.2.1 is that an overdetermined linear IDE may become well-determined (or underdetermined) after the reduction procedure and that an underdetermined IDE will always stay underdetermined after the reduction.

Solvability

Let us make an obvious remark concerning reduced systems. Loosely speaking, a totally reduced system can be seen as the "control part" of a control system in descriptor form. There are then as many solutions as there are control inputs. In other words, the solution is unique if and only if there is no control part.

Proposition 5.2.2. *Assume that* (E, A) *is totally reduced. Then it is solvable iff* E *is injective.*

Proof.

1. We may change coordinates in such a way that E is represented by a block matrix with one identity matrix as:

$$\mathsf{E} = \begin{bmatrix} \mathbb{I} & 0 \\ 0 & 0 \end{bmatrix}$$

2. Since the system is totally reduced, $\mathsf{A}M \subset \mathsf{E}M$ and A is represented as:

$$\mathsf{A} = \begin{bmatrix} \mathsf{A}_1 & \mathsf{A}_2 \\ 0 & 0 \end{bmatrix}$$

3. Notice that
$$\mathsf{A}_2 = 0 \iff \ker \mathsf{E} = 0$$

4. The equations are in this coordinate system:
$$\frac{\mathrm{d}\xi}{\mathrm{d}t} + \mathsf{A}_1\xi + \mathsf{A}_2 u$$

5. Obviously if $\mathsf{A}_2 = 0$ then the system has a unique solution for any initial condition. If $\mathsf{A}_2 \neq 0$ then for any smooth function $t \mapsto u(t)$ sucht that
$$u(t_0) = u_0 \qquad (\spadesuit)$$
we obtain a unique solution of the system. In other words there are as many different solutions as there are smooth functions u satisfying the condition ($\spadesuit$).

□

Example 5.2.1. In Example 4.8.11 the determinacy degree strictly decreases during the second reduction. The matrices for that system are:
$$\mathsf{E} = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \qquad \text{and} \qquad \mathsf{A} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$
The reduction for this system is:
$$M' = \{z = 0\} \qquad \mathsf{E}' = \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix} \qquad \mathsf{A}' = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}$$
$$M'' = \{x + y = 0\} \cap M' \qquad \mathsf{E}'' = 0 \qquad \mathsf{A}'' = 0$$
So the sequence of determinacy degrees is easily determined to be:
$$(\Delta, \Delta', \Delta'') = (0, 0, -1)$$

Example 5.2.2. Here is an example of an apparently overdetermined system which turns out to be solvable:
$$\mathsf{E} = \begin{bmatrix} \boxed{1} & & & \\ & \boxed{0} & \boxed{\begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix}} & \\ & & \boxed{\begin{matrix} 0 & 0 \\ 0 & 0 \end{matrix}} & \end{bmatrix} \qquad \mathsf{A} = \begin{bmatrix} \boxed{1} & & & \\ & \boxed{1} & \boxed{\begin{matrix} 0 & 0 \\ 0 & 0 \end{matrix}} & \\ & & \boxed{\begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix}} & \end{bmatrix}$$

Let us call the variables u, x, y_1, y_2. The reduced subspace are $M' = \{y_1 = y_2 = 0\}$ and $M'' = \{x = 0\} \cap M'$.

Notice that although this system is solvable, it will *not* be solvable with a source term, unless some conditions on the source term are fulfilled.

For instance here if we add a source term with coordinates $\mathsf{f} = (f, g_1, g_2, h_1, h_2)$ the system

$$\mathsf{E}\frac{\mathrm{d}\mathsf{x}}{\mathrm{d}t} + \mathsf{A}\mathsf{x} = \mathsf{f}$$

will have no solution unless $g_2' = f_2$.

Proposition 5.2.3. *A totally reduced linear IDE system* (E, A) *is* solvable *iff it is well-determined, i.e. iff:*

$$\Delta = 0$$

Proof. Using Proposition 5.1.3 and Definition 5.1 we get:

$$\Delta = \dim(\mathsf{E}M) - \dim M$$

which implies that $\Delta = 0$ iff E is injective, so we conclude with Proposition 5.2.2. □

Remark 5.2.1.

Notice that it may very well happen that the totally reduced subspace is zero, as in this example:

$$\mathsf{E} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \qquad \mathsf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

This system has only solutions for the initial condition zero.

5.3 Nested Reductions

Reduced Codomain

It will prove useful to define the *reduced codomain* of the operators E and A:

Definition 5.3. Given a linear IDE (E, A) we define its ***reduced codomain*** V' as:

$$V' := \mathsf{E}M$$

One motivation of that definition is the following elementary observation:

Proposition 5.3.1.

$$\mathsf{E}M' \subset V' \qquad \mathsf{A}M' \subset V'$$

This suggest the definition of the *reduced operators* as follows:

Definition 5.4. Given a linear IDE (E, A) we define the ***reduced operators*** E' and A' as follows:

$$\mathsf{E}', \mathsf{A}' : \ M' \longrightarrow V'$$

such that:

$$\mathsf{E}' = \mathsf{E}|_{M'} \qquad \mathsf{A}' = \mathsf{A}|_{M'}$$

The only difference between E' and $\mathsf{E}|_{M'}$, or A' and $\mathsf{A}|_{M'}$ is the codomain of E' and A' which is now V'.

Iterated Reduction

We may iterate the process described in the last section on the new IDE $(\mathsf{E}', \mathsf{A}')$. This leads to a sequence of nested spaces $M^{(k)}$. Using Proposition 4.8.1, the smallest integer for which this sequence stalls is the *index* of the IDE:

Proposition 5.3.2. *The smallest integer n for which*

$$M^{(n+1)} = M^{(n)}$$

is the index of the IDE.

Remark 5.3.1.

> Notice that even if the system $(\mathsf{E}^{(n)}, \mathsf{A}^{(n)})$ is totally reduced, $\mathsf{E}^{(n)}$ may still fail to be surjective, which entails $V^{(n+1)} = \mathsf{E}M^{(n)} \subsetneq V^{(n)}$. This is summarized in the following description of the sequences of nested spaces:

$$\begin{aligned} M^{(n+1)} =& M^{(n)} \subset \ldots \subset M'' \subset M' \subset M \\ V^{(n+2)} = V^{(n+1)} \subset& V^{(n)} \subset \ldots \subset V'' \subset V' \subset V \end{aligned}$$

Notation 5.3.1. In order to avoid mentioning the index n of a given linear IDE, we will denote, for an IDE of index n: $\mathsf{E}^{(\infty)}$ $\mathsf{A}^{(\infty)}$ $M^{(\infty)}$ $V^{(\infty)}$

$$\begin{aligned} \mathsf{E}^{(\infty)} &:= \mathsf{E}^{(n+1)} \\ \mathsf{A}^{(\infty)} &:= \mathsf{A}^{(n+1)} \\ M^{(\infty)} &:= M^{(n+1)} \\ V^{(\infty)} &:= V^{(n+1)} \end{aligned}$$

Note that although we defined $M^{(\infty)} = M^{(n+1)}$ for consistency with the other definitions, we also have by definition of n:

$$M^{(\infty)} = M^{(n)}$$

Note also that we could simply have defined, say $\mathsf{E}^{(\infty)}$ by the limit of the sequence of operators $\mathsf{E}^{(k)}$, which explains the notation "∞".

Let us notice that by construction $\mathsf{E}^{(\infty)}$ is surjective:

Proposition 5.3.3. *The totally reduced operator $\mathsf{E}^{(\infty)}$ is surjective.*

5.4 Supplementary Spaces

Definitions

In this section we briefly study the general properties of supplementary spaces to M'.

Let us choose a supplementary space N from M' to M:

$$M = M' \oplus N \tag{5.2}$$

Please take good notice that this splitting is by no means unique! We will see later in Section 5.5 and Section 5.6 how to choose the supplementary space N in an advantageous way.

Notice now the following facts, which are direct consequences of the definition of M':

Proposition 5.4.1.

$$\mathsf{A}N \cap \mathsf{E}M = 0$$

$$\mathsf{E}M \oplus \mathsf{A}N = \mathsf{E}M + \mathsf{A}M$$

Recalling the definition of the reduced codomain $V' = \mathsf{E}M$ (Definition 5.3) we may *choose* a supplementary space Z such that the codomain V is decomposed as:

$$V = V' \oplus \mathsf{A}N \oplus Z \tag{5.3}$$

We use the previous observations to produce a simple matrix representation:

Proposition 5.4.2. *By choosing in $\mathsf{A}N$ the image of the basis in N by A we obtain the matrix form of Figure 5.1.*

Proof. The block decomposition is obtained by noticing that: $(\mathsf{E}M)^{\perp}\mathsf{E}M = 0$, $(\mathsf{A}N + \mathsf{E}M)^{\perp}\mathsf{A}N = 0$ and $(\mathsf{E}M)^{\perp}\mathsf{A}M' = 0$. □

Remark 5.4.1.

In a sense the space Z (the letter "Z" stands for *zero*) models the extra "non equations" of the system. It simply means that the space V is too big.

The reader might think that it would be a serious modelling mistake to choose too big a space V. However, this phenomenon of "non-equations" may occur at a higher level of the reduction, namely for one of the reduced systems $(\mathsf{E}^{(k)}, \mathsf{A}^{(k)})$, for which we cannot directly choose the corresponding codomain $V^{(k)}$.

Remark 5.4.2.

In the same spirit as Remark 5.3.1, it makes sense to define Z even if the system (E, A) is totally reduced, i.e. when $M' = M$. In that case, the only possible choice for the supplementary space of (5.2) is $N = 0$ and (5.3) becomes:

$$V = V' \oplus Z$$

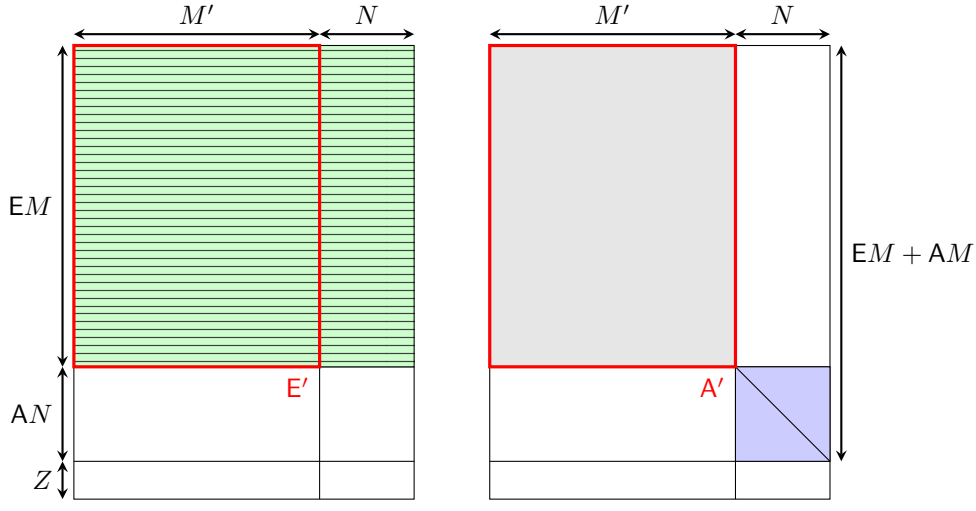


Figure 5.1: In the space decompositions described in Equation 5.2 and Equation 5.3 the operators E and A have those matrix representations. Blue squares with diagonals are identity matrices. White areas are zeroes. The grey area may be any matrix. The green striped rectangle matrix must have *full rank*. The red-framed submatrices are the reduced matrices E′ and A′ of the next step in the reduction. Notice that the result of Proposition 5.2.1 is apparent on this figure.

Iterated Supplementary Spaces

By repeating the procedure of (5.2) and (5.3), we may *choose* the following spaces:

$$N, N', \dots, N^{(k)}, \dots$$

$$Z, Z', \dots, Z^{(k)}, \dots$$

Assume that the index of the system defined by (E, A) is n, i.e., according to Proposition 5.3.2, that n is the first integer such that $M^{(n+1)} = M^{(n)}$

Then the only possible choice for $N^{(n)}$ is:

$$N^{(n)} = 0 \tag{5.4}$$

As for the sequence of spaces $Z^{(k)}$ it will generally stall at one step further, as explained in Remark 5.4.2:

$$Z^{(n+1)} = 0$$

The iterated decompositions (5.3) and (5.2) are written as:

$$M = M^{(n)} \oplus N^{(n-1)} \oplus \cdots \oplus N$$
$$V = V^{(n+1)} \oplus Z^{(n)} \oplus \mathsf{A}N^{(n-1)} \oplus Z^{(n-1)} \oplus \cdots \oplus \mathsf{A}N \oplus Z$$

For convenience we will also define the spaces $W^{(k)}$:

$$W^{(k)} := \mathsf{A}N^{(k)} \oplus Z^{(k)}$$

One may thus now also write:

$$V = V^{(n+1)} \oplus W^{(n)} \oplus W^{(n-1)} \oplus \cdots \oplus W$$

We can already prove the following relation between the dimensions of $Z^{(k)}$ and $\ker \mathsf{E}^{(\infty)}$:

Proposition 5.4.3. *The following identity hold:*

$$\dim M - \dim \ker \mathsf{E}^{(\infty)} = \dim V - \sum_k \dim Z^{(k)}$$

Proof. The identity follows from:

$$\dim M = \dim M^{(\infty)} + \sum_k \dim N^{(k)}$$

$$\dim V = \dim V^{(\infty)} + \sum_k \dim Z^{(k)} + \sum_k \dim \mathsf{A}N^{(k)}$$

$$\dim M^{(\infty)} = \dim \mathsf{E}M^{(\infty)} + \dim \ker \mathsf{E}^{(\infty)}$$

and $\dim \mathsf{A}N^{(k)} = \dim N^{(k)}$ for all k, plus the property that $V^{(\infty)} = \mathsf{E}M^{(\infty)}$. □

Remark 5.4.3.

Here is an example of a calculation of Z'. It is given by the *pure derivator* in control theory. Its homogeneous part (without the sources) is given by the matrices:

$$\mathsf{E} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \qquad \mathsf{A} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$$

In that example the decomposition is:

$$N = M$$

$$V = Z' \oplus \mathsf{A}N$$

and

$$Z' = \{y = z = 0\}$$

The source term to observe the derivator behaviour is

$$q(t) = \big(0, u(t), y(t)\big)$$

The existence of solutions then imposes the relation:

$$y(t) = u'(t)$$

Determinacy Degree and Defect Spaces

The sequence of spaces $\left\{Z^{(k)}\right\}_{k\geq 0}$ is related to the indeterminacy of the system. In fact the dimensions of those spaces exactly describe the determinacy drop observed in Proposition 5.2.1:

Proposition 5.4.4. *The determinacy degree drop* $\Delta - \Delta'$ *is given by:*

$$\Delta - \Delta' = \dim Z'$$

Proof.

1. Using Lemma 5.2.1 one has:
$$\Delta - \Delta' = \dim(\mathsf{A}M + \mathsf{E}M) - \dim(\mathsf{A}M + \mathsf{E}M')$$
2. Now by definition of Z':
$$\mathsf{E}M = \mathsf{E}M' \oplus \mathsf{A}N' \oplus Z'$$
in particular notice that
$$Z' \cap \mathsf{E}M' = 0 \qquad (\spadesuit)$$

3. As a result:
$$\operatorname{Im}\mathsf{A} + \operatorname{Im}\mathsf{E} = \operatorname{Im}\mathsf{A} + \mathsf{E}M' + Z'$$
4. Using the definition of Z' and (♠) we obtain:
$$\operatorname{Im}\mathsf{A} + \mathsf{E}M' + Z' = (\operatorname{Im}\mathsf{A} + \mathsf{E}M') \oplus Z'$$
which finishes the proof.

□

The following Proposition gives a restriction on the dimension drop of the hidden manifolds $M^{(k)}$. More precisely this dimension drop may only *decrease*. See an illustration on Figure 5.2.

Proposition 5.4.5.
$$\dim Z' + \dim N' \leq \dim N$$

Proof.

1. We have
$$M = M' \oplus N$$
so
$$\mathsf{E}M = \mathsf{E}M' + \mathsf{E}N$$
which implies
$$\dim \mathsf{E}M \leq \dim \mathsf{E}M' + \dim \mathsf{E}N$$
2. Now by definition of Z':
$$\dim \mathsf{E}M = \dim \mathsf{E}M' + \dim \mathsf{A}N' + \dim Z'$$
3. The facts that $\dim \mathsf{A}N' = \dim N'$ and that $\dim \mathsf{E}N \leq \dim N$ finish the proof.

□

5.5 Coupling

Coupling spaces

All the results up to now are valid independently of the choice of the supplementary space N. To improve the final normal form we will now restrict our choice of the supplementary space.

The strategy is to try to choose N in the same direction as the part of $\ker \mathsf{E}$ that remains out of M'. First we define what this space is by decomposing the kernel of E in the part that is included in M' and some supplementary space. This is achieved by choosing any supplementary space K such that:
$$\ker \mathsf{E} = (\ker \mathsf{E} \cap M') \oplus K$$

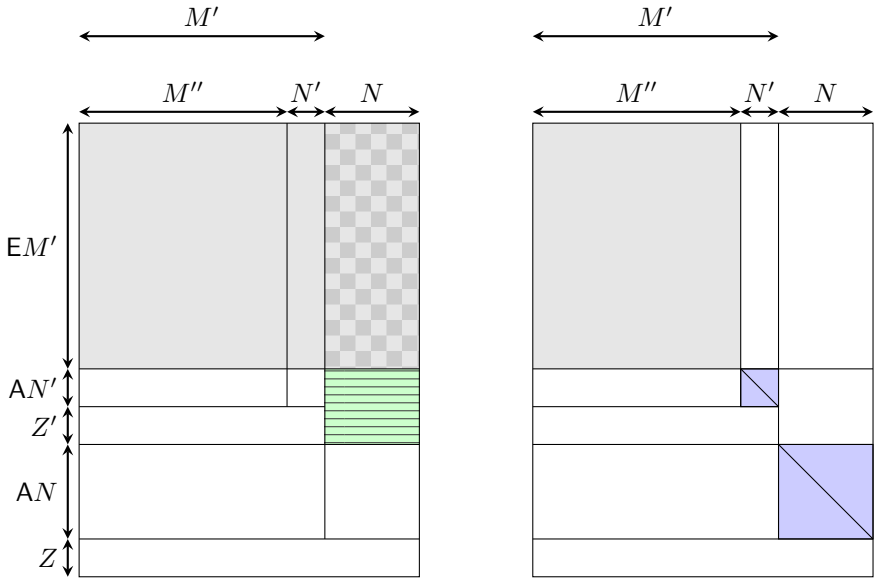


Figure 5.2: Illustration of Proposition 5.4.5. The picture represents the matrices E and A. The green striped rectangle matrix must have full rank. This implies that $\dim \mathsf{A}N' + \dim Z' \leq \dim N$. We will see in Section 5.6 that, by a careful choice of basis, the whole checkered area may be set to zero.

Then since, by construction, $K \cap M' = 0$ one may complete M' by choosing a supplementary space C such that:

$$M = M' \oplus C \oplus K$$

We now choose N as:

$$N := C \oplus K$$

Notice now that the space K roughly speaking corresponds to the variables that are *decoupled* from the rest of the system. They are sometimes called the *algebraic constraints*.

Example 5.5.1. Let us illustrate the previous remark by a trivial example. Consider the following simple system:

$$\begin{cases} \mathrm{d}x - x\,\mathrm{d}t \\ y\,\mathrm{d}t \end{cases}$$

The variable y is *decoupled* from the rest of the system.

Remark 5.5.1.

Let us notice the well known (obvious) identity:

$$\dim V = \dim(\operatorname{Im} \mathsf{E}) + \dim C + \dim K + \dim Z$$

since

$$\dim \mathsf{A}N = \dim N = \dim C + \dim K$$

Remark 5.5.2.

Some authors ([KM94]) define the "strangeness" as what turns out to be the following quantity:

$$s = \dim C$$

Roughly speaking it is the number of constraints that, when differentiated, will help reduce the system.

We call them the **number of coupled constraints** instead, given the lack of insight that the English word "strangeness" provides.

5.6 Complete Decomposition

In this section we will see how to choose the spaces $K^{(k)}$ and $C^{(k)}$ in such a way that the operators E and A are *simultaneously* decomposed in an advantageous way.

The main tool will be this elementary result from linear algebra:

Lemma 5.6.1. *Given a surjective operator* E *defined on a space* M_0 *to a space* V_1 *and a subspace* $M_1 \subset M_0$ *with a given decomposition:*

$$V_1 = \mathsf{E}M_1 \oplus W$$

then there exists subspaces $C \subset M_0$ *and* $K \subset M_0$ *such that*

$$M_0 = M_1 \oplus C \oplus K$$

and

$$\mathsf{E}C = W$$

$$\ker \mathsf{E} \cap C = 0$$

$$K \subset \ker \mathsf{E}$$

Proof.

1. There exists a right inverse F to E such that:

$$\mathsf{EF} = \mathbb{I}$$

   Note that F is necessarily injective.

2. Consider the space C defined by

$$C = \mathsf{F}W$$

   Note that

$$\dim C = \dim W$$

   since F is injective.

3. Now if we denote

$$K_1 := \ker \mathsf{E} \cap M_1$$

   we decompose:

$$K_0 := \ker \mathsf{E} = K_1 \oplus K$$

   It is clear that, by definition of K:

$$K \cap M_1 = 0$$

4. Since

$$\dim M_0 = \dim \mathsf{E}M_0 + \dim K_0$$

   and similarly:

$$\dim M_1 = \dim \mathsf{E}M_1 + \dim K_1$$

   we obtain:

$$\dim M_0 - \dim M_1 = \dim W + \dim K$$

   which in turn yields:

$$\dim M_0 = \dim M_1 + \dim C + \dim K$$

5. This implies that

$$M_0 = M_1 \oplus K \oplus C$$

   Besides, by construction we have

$$\mathsf{E}K = 0$$

   and

$$\mathsf{E}C = W$$

   and

$$\ker \mathsf{E} \cap C = 0$$

□

We are now ready to proceed to the main Theorem of this Chapter.

Theorem 5.1. *Consider two linear operators* E *and* A*, both acting from a finite dimensional vector space* M *to a finite dimensional vector space* V*. Assume that the corresponding IDE* (E, A) *has index* n*.*

There exists a decomposition (cf. Figure 5.3):

$$M = S \oplus \ker \mathsf{E}^{(\infty)} \oplus \bigoplus_{i=0}^{n-1} N^{(i)}$$

and

$$V = V^{(\infty)} \oplus \bigoplus_{i=0}^{n} W^{(i)}$$

such that

$$M^{(\infty)} = S \oplus \ker \mathsf{E} M^{(\infty)}$$

and further decompositions:

$$N^{(i)} = C^{(i)} \oplus K^{(i)}$$

and

$$W^{(i)} = Z^{(i)} \oplus \mathsf{A} N^{(i)}$$

with

$$N^{(n)} = 0$$

There exists a choice of basis in those spaces such that E *is represented as a block matrix which is*

- *identity on the block of coordinates* $(S, V^{(n)})$
- *identity on the "upper diagonal blocks" of coordinates* $(C^{(i)}, W^{(i+1)})$ *for* $0 \leq i \leq n-1$
- *zero on all other blocks*

As for the operator A*, it is represented as a block matrix*

- *a nonspecific matrix on the block* $(M^{(\infty)}, V^{(\infty)})$
- *identity on the "diagonal" blocks* $(N^{(i)}, \mathsf{A} N^{(i)})$ *for* $0 \leq i \leq n$
- *zero on all other blocks*

Proof.

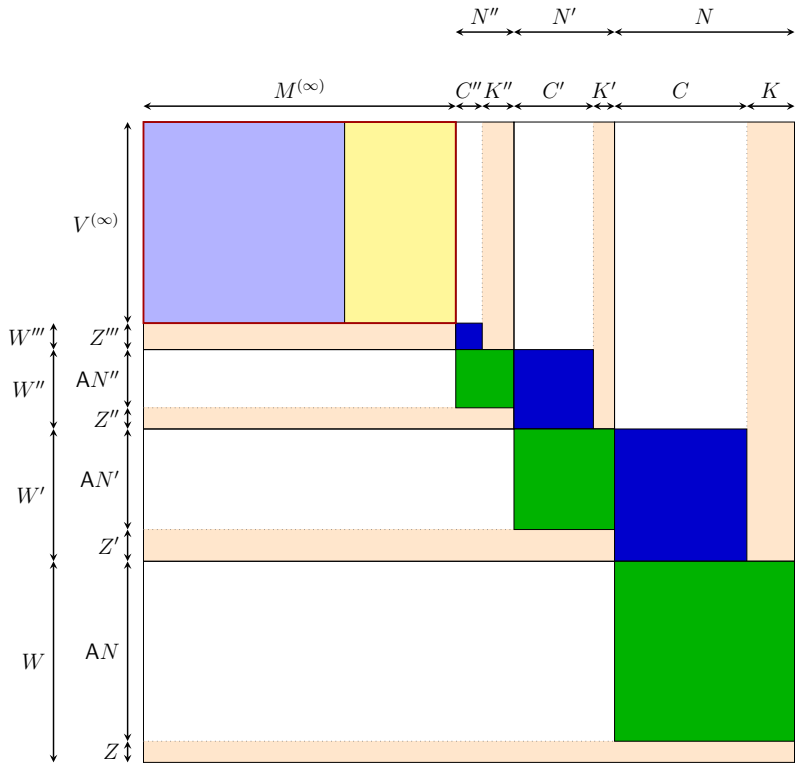


Figure 5.3: (See legend on p. 119). An illustration of the decomposition described in Theorem 5.1.

1. We will proceed by induction (cf. Figure 5.4). Let us assume that the statement holds for systems of index $n-1$. Given a system (E,A) of index n, the reduced system $(\mathsf{E}',\mathsf{A}')$ has index $n-1$. Recall the $(\mathsf{E}',\mathsf{A}')$ are operators defined from M' to V':

$$M' \longrightarrow V' := \mathsf{E}M$$

So we get a decomposition of the spaces M' and V' as described in the Theorem. For convenience we will shift the indices of all the spaces produced by the Theorem. For example, the space $W^{(k)}$ will be now denoted $W^{(k+1)}$, so we may write the decomposition as:

$$V' = \mathsf{E}M = \mathsf{E}M^{(n)} \oplus W^{(n)} \oplus \cdots \oplus W'' \oplus W'$$

   a) By definition of W', we have:

$$\mathsf{E}M = V' = V'' \oplus W' = \mathsf{E}M' \oplus W'$$

   W' is moreover equipped with a basis.

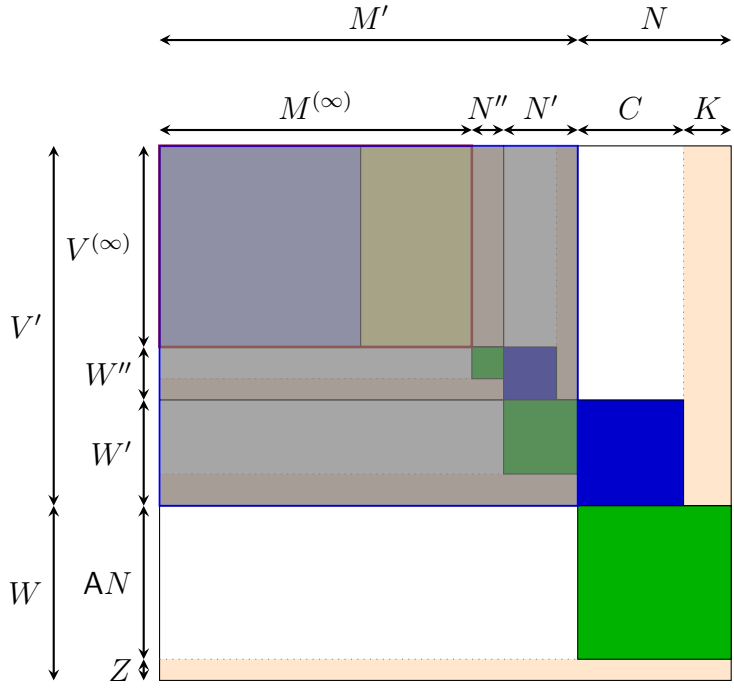


Figure 5.4: (See legend on p. 119). An illustration of the proof of Theorem 5.1 on an index three system. The grey shaded part pictures the previous step of the recursion. Starting with W', one constructs the spaces C and K using Lemma 5.6.1, and defines $N := C \oplus K$. One then constructs Z such that $V = V' \oplus \mathsf{A}N \oplus Z$. This in turn defines $W := \mathsf{A}N \oplus Z$.

b) Using Lemma 5.6.1 we obtain the existence of C and K such that

$$M = M' \oplus C \oplus K$$

with

$$\mathsf{E}C = W'$$

and

$$\mathsf{E}K = 0$$

and

$$\ker \mathsf{E} \cap C = 0$$

Note that, given a basis in W' we can choose a basis on C such that E is represented by the identity matrix when restricted from C to W'.

Let us now denote for brevity:

$$N = C \oplus K$$

We choose a basis of K so that we now have a basis for N.

c) Now we notice as we did earlier in Proposition 5.4.1

$$\mathsf{A}N \cap \mathsf{E}M = 0$$

so we define Z as in (5.3), i.e. such that:

$$V = \mathsf{E}M \oplus \mathsf{A}N \oplus Z$$

and we choose in $\mathsf{A}N$ the image of the basis of N by A, so that A is represented by the identity matrix when restricted from N to $\mathsf{A}N$.

d) Using $(V')^{\perp}\mathsf{E}M = 0$, $(W')^{\perp}\mathsf{E}C = 0$, $\mathsf{E}K = 0$ we obtain the desired block structure for E. Similarly, using $(V')^{\perp}\mathsf{A}M' = 0$, $(\mathsf{A}N)^{\perp}\mathsf{A}N = 0$ and $(V' + \mathsf{A}M)^{\perp}\mathsf{A}M = 0$, we obtain the desired block structure for A.

2. We have to check the first step of the recursion. When the index is zero we simply have $\mathsf{A}M \subset \mathsf{E}M$ and we just choose W such that:

$$V = \mathsf{E}M \oplus W$$

We choose an arbitrary basis in W.

Lastly we choose S as a supplementary space of $\ker \mathsf{E}$ in M:

$$M = S \oplus \ker \mathsf{E}$$

Clearly, E is injective on S. Given a basis on S we choose $\mathsf{E}S$ as a basis on $\mathsf{E}M$.

□

Let us note the following easy Corollaries to the Theorem:

Corollary 5.6.1. *It follows from Theorem 5.1 that:*

(i) $\dim C^{(i)} = \dim W^{(i+1)} \qquad \forall 0 \leq i \leq n-1$

(ii) $\dim W^{(i)} \leq \dim N^{(i-1)} \leq \dim W^{(i-1)}$ *or more precisely:*

$$\begin{aligned} \dim N^{(i-1)} &= \dim W^{(i)} + \dim K^{(i)} \\ \dim W^{(i)} &= \dim N^{(i)} + \dim Z^{(i)} \end{aligned} \tag{5.5}$$

Corollary 5.6.2. *The following identity hold:*

$$\dim W = \sum_k \dim Z^{(k)} + \sum_k \dim K^{(k)}$$

Proof. From:

$$\dim M = \dim \mathsf{E}M^{(\infty)} + \dim \ker \mathsf{E}^{(\infty)} + \sum_k \dim C^{(k)} + \sum_k \dim K^{(k)}$$

$$\dim V = \dim \mathsf{E}M^{(\infty)} + \sum_{k\geq 1} \dim W^{(k)} + \dim W$$

and the fact that $\dim W^{(k)} = \dim C^{(k-1)}$ for $k \geq 1$, we obtain:

$$\dim M - \dim \ker \mathsf{E}^{(\infty)} - \sum_k \dim K^{(k)} = \dim V - \dim W$$

Now we use Proposition 5.4.3 to obtain the result. □

We also make the crucial observation that Theorem 5.1 provides us with a splitting of M and V such that E and A are acting separately on those parts:

Corollary 5.6.3. *Given the decomposition provided by Theorem 5.1, and defining* $\overline{M}$ *and* $\overline{V}$ *as:*

$$\overline{M} := \bigoplus_k N^{(k)} \qquad \overline{V} := \bigoplus_k W^{(k)}$$

then, by construction:

$$M = M^{(\infty)} \oplus \overline{M} \qquad V = V^{(\infty)} \oplus \overline{V}$$

and the following holds:

(i)
$$\mathsf{E}M^{(\infty)} \subset V^{(\infty)} \qquad \mathsf{A}M^{(\infty)} \subset V^{(\infty)}$$

(ii)
$$\mathsf{E}\overline{M} \subset \overline{V} \qquad \mathsf{A}\overline{M} \subset \overline{V}$$

Proof.

1. Notice first that:
$$(V^{(\infty)})^{\perp}\mathsf{E}M^{(\infty)} = 0 \qquad (V^{(\infty)})^{\perp}\mathsf{A}M^{(\infty)} = 0$$

2.
$$\overline{V}^{\perp}\mathsf{E}N^{(k)} = 0 \qquad \overline{V}^{\perp}\mathsf{A}N^{(k)} = 0 \qquad \forall k$$

since we have:
$$\mathsf{E}N^{(k)} \subset W^{(k+1)}$$

and
$$\mathsf{A}N^{(k)} \subset W^{(k)}$$

□

Example 5.6.1. Let us compute the normal form of a linearised mechanical system. Such a system may be written in the following way, in coordinates denoted by (x, v, λ):

$$\mathsf{E} = \begin{bmatrix} \mathbb{I} & 0 & 0 \\ 0 & \mathbb{I} & 0 \\ 0 & 0 & 0 \end{bmatrix} \qquad \mathsf{A} = \begin{bmatrix} 0 & -\mathbb{I} & 0 \\ \mathsf{K} & \mathsf{D} & \mathsf{G}^{\mathsf{T}} \\ \mathsf{G} & 0 & 0 \end{bmatrix}$$

where we implicitly inverted the mass matrix.

Let us change coordinates in such a way that G is written as:

$$\mathsf{G} = \begin{bmatrix} 0 & \mathbb{I} \end{bmatrix}$$

where we assumed that G has full rank.

We now change coordinates to $(x_1, x_2, v_1, v_2, \lambda)$ so that

$$\ker \mathsf{G} = \{x_2 = 0\}$$

We may split the matrices K and D in block matrices. For example K is split into: $\mathsf{K}_{11}, \mathsf{K}_{12}, \mathsf{K}_{21}, \mathsf{K}_{22}$.

We now have following equations in the variables $(x_1, v_1, \lambda, v_2, x_2)$

$$\begin{cases} \mathrm{d}x_1 - v_1 \,\mathrm{d}t \\ \mathrm{d}v_1 + (\mathsf{K}_{11}x_1 + \mathsf{D}_{11}v_1 + \mathsf{D}_{12}v_2 + \mathsf{K}_{12}x_2)\,\mathrm{d}t \\ \mathrm{d}v_2 + (\mathsf{K}_{21}x_1 + \mathsf{K}_{22}x_2 + \mathsf{D}_{21}v_1 + \mathsf{D}_{22}v_2 + \lambda)\,\mathrm{d}t \\ \mathrm{d}x_2 - v_2 \,\mathrm{d}t \\ \quad - x_2 \,\mathrm{d}t \end{cases}$$

First we may eliminate the term containing $x_2 \,\mathrm{d}t$ by multiplying the last line by K_{12} or K_{22} and adding it to the corresponding lines (second and third).

Similarly, we multiply the next-to-last line by D_{12} and add it to the second line. The second line is now:

$$\mathrm{d}v_1 + \mathsf{D}_{12}\,\mathrm{d}x_2 + (\mathsf{K}_{11}x_1 + \mathsf{D}_{11}v_1)\,\mathrm{d}t$$

Now it is easy to obtain the normal form by choosing the new coordinates:

$$\bar{\lambda} := -(\mathsf{K}_{21}x_1 + \mathsf{D}_{21}v_1 + \mathsf{D}_{22}v_2 + \lambda)$$

and

$$\bar{v}_1 := v_1 + \mathsf{D}_{12}x_2$$

Clearly the change of coordinates defined by

$$(x_1, v_1, \lambda, v_2, x_2) \longmapsto (x_1, \bar{v}_1, \bar{\lambda}, v_2, x_2)$$

is invertible since it is given by a triangular matrix with ones on the diagonal.

Now in those coordinates $(x_1, \bar{v}_1, \bar{\lambda}, v_2, x_2)$ we obtain the matrix representations:

$$\mathsf{E} = \begin{bmatrix} \mathbb{I} & 0 & & & \\ 0 & \mathbb{I} & & & \\ & & 0 & \mathbb{I} & \\ & & & 0 & \mathbb{I} \\ & & & & 0 \end{bmatrix} \qquad \mathsf{A} = \begin{bmatrix} 0 & \mathbb{I} & & & \\ -\mathsf{K}_{11} & -\mathsf{D}_{11} & & & \\ & & \mathbb{I} & & \\ & & & \mathbb{I} & \\ & & & & \mathbb{I} \end{bmatrix}$$

which is the normal form for that system.

Example 5.6.2. As a simple example, the system of Example 4.8.11 has the normal form:

$$\mathsf{E} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \qquad \mathsf{A} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

In this normal form the non-solvability is directly apparent: the first variable does not appear anywhere in the equation.

5.7 Square Systems

Definition

We briefly recall basic results of linear IDEs expressed in our framework. We are able to give original proofs of those results.

In this section we will focus on "square matrices", i.e. operators for which the domain and codomain have the same dimension.

Definition 5.5. A ***square linear system*** is a linear IDE (E, A) for which the domain M and codomain V have the same dimension:

$$\dim M = \dim V$$

The regular pencil Theorem

In this section we give an original proof of the so-called "regular pencil theorem" (see e.g. [Gan59, § XII.7]), which relates solvability and the "regularity" of the pencil given by E and A, namely the matrix-valued polynomial $z\mathsf{E} + \mathsf{A}$.

It should be noticed that this subsection is independent from Section 5.6.

We start by relating the appearance of a overdetermination defect[b] (i.e. a non-zero $Z^{(k)}$ space) to the polynomial defined by:

$$F(z) := \det(z\mathsf{E} + \mathsf{A})$$

Lemma 5.7.1. *For a square system* (E, A)*, if* $Z \neq 0$ *then* $z\mathsf{E} + \mathsf{A}$ *is not surjective for any* $z \in \mathbb{C}$*.*

Proof. If $Z \neq 0$ then, by Proposition 5.4.1 and (5.3): $\mathsf{A}M + \mathsf{E}M \neq V$. In particular:

$$(z\mathsf{E} + \mathsf{A})M \subset \mathsf{A}M + \mathsf{E}M$$

so the operator $z\mathsf{E} + \mathsf{A}$ cannot be surjective for any $z \in \mathbb{C}$. □

In fact, Proposition 5.4.3 provides us with the exact relation between the dimension of the kernel of the totally reduced system and the dimensions of the spaces $Z^{(k)}$:

Proposition 5.7.1. *For a square system we have:*

$$\dim \ker \mathsf{E}^{(\infty)} = \sum_k \dim Z^{(k)}$$

Proof. It is a direct consequence of Proposition 5.4.3.

□

Lemma 5.7.2. *For a square system* (E, A)*, if* $Z = 0$ *then there exists a choice of basis in which:*

$$\det(z\mathsf{E} + \mathsf{A}) = \det(z\mathsf{E}' + \mathsf{A}') \quad \forall z \in \mathbb{C}$$

Proof. It is a direct consequence of the block decomposition of Proposition 5.4.2, illustrated on Figure 5.1. □

Of course we could have use the decomposition of Theorem 5.1 in the last proof.

We may now prove the Lemma relating the overdetermination defects and the determinant:

[b] See Section 5.8 for more details on defects.

Lemma 5.7.3. *Given a square system* (E, A)*, the following assertions are equivalent:*

(i)
$$\big(z \mapsto \det(z\mathsf{E} + \mathsf{A})\big) \not\equiv 0$$

(ii)
$$\dim Z^{(k)} = 0 \quad \forall k \geq 0$$

Proof. Let us proceed by induction on the index. Obviously if $\dim Z \neq 0$ then $\det(z\mathsf{E} + \mathsf{A}) \equiv 0$. On the other hand, if $\dim Z = 0$ we may use Lemma 5.7.2 and use the induction hypothesis on the reduced operators $(\mathsf{E}', \mathsf{A}')$ since $\dim M' = \dim V'$ by Corollary 5.6.1. Finally, if the system is totally reduced then $\dim Z = 0$ is equivalent to E being surjective, (and thus invertible since $\dim M = \dim V$), which implies $\det(z\mathsf{E} + \mathsf{A}) \neq 0$. □

We may now prove the following Proposition, relating solvability and the totally reduced system $(\mathsf{E}^{(\infty)}, \mathsf{A}^{(\infty)})$:

Proposition 5.7.2. *The square system* (E, A) *is solvable iff* $\mathsf{E}^{(\infty)}$ *is invertible.*

Proof. Any solution of the IDE defined by (E, A) must be a solution of $(\mathsf{E}^{(\infty)}, \mathsf{A}^{(\infty)})$. We then use Proposition 5.2.2. □

We are now in position to state an improved version of the regular pencil theorem, which relates solvability and the dimensions of the overdetermination defect spaces $Z^{(k)}$ (see Figure 5.5).

Theorem 5.2. *Given a square system* (E, A) *the following assertions are equivalent:*

(i) *the linear system* (E, A) *is solvable*

(ii)
$$\big(z \mapsto \det(z\mathsf{E} + \mathsf{A})\big) \not\equiv 0$$

(iii)
$$\dim Z^{(k)} = 0 \quad \forall k \geq 0$$

Proof. It is a direct consequence of Proposition 5.2.3, Lemma 5.7.1, Lemma 5.7.2 and the fact that for a square matrix M, $\det(\mathsf{M}) = 0$ is invariant by equivalence transformations (i.e. transformations of the type PMQ for P, Q invertible matrices.)

□

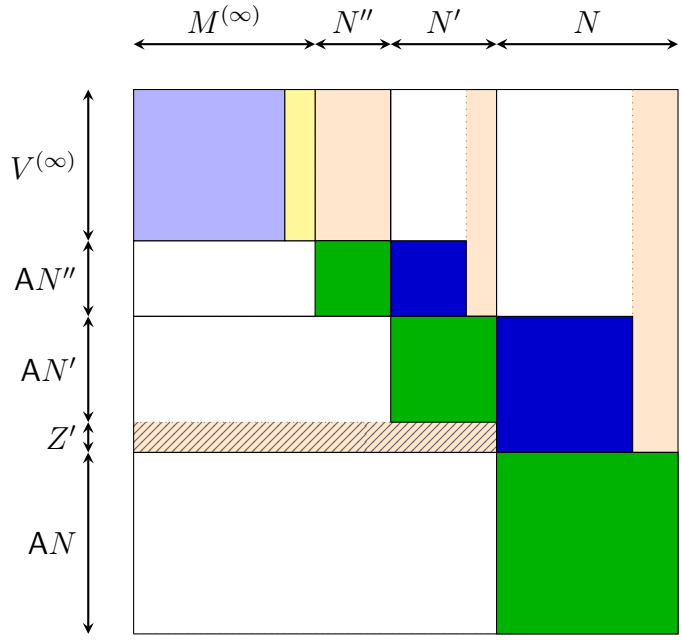


Figure 5.5: (See legend on p. 119). Illustration of Theorem 5.2. As long as an equation defect is present (here $Z' \neq 0$) and if the matrices are square (i.e. $\dim M = \dim V$) then the intrinsic dynamical system will be "squeezed" to a rectangle form, i.e. the kernel of the totally reduced operator $\mathsf{E}^{(\infty)} = \mathsf{E}''$ is not zero. On the figure, the result of Proposition 5.7.1 is clear. In that precise case, $\dim \ker \mathsf{E}^{(\infty)} = \dim Z'$.

5.8 Full Canonical Form

Jordan Canonical Form

The decomposition of Theorem 5.1 is not quite a canonical form yet, because it leaves parts of the representation of A undetermined. In one case though, one may proceed to reduce A to its Jordan canonical form by similarity transformations.

Proposition 5.8.1. *If the linear IDE* (E, A) *is totally reduced and if*

$$\ker \mathsf{E} \subset \ker \mathsf{A} \tag{$\diamondsuit$}$$

one can represent the system (E, A) *by the matrices:*

$$\mathsf{E} = \begin{bmatrix} \mathbb{I} & 0 \\ 0 & 0 \end{bmatrix} \qquad \mathsf{A} = \begin{bmatrix} \mathsf{J} & 0 \\ 0 & 0 \end{bmatrix} \tag{$\clubsuit$}$$

where J *is in Jordan canonical form.*

Proof.

1. One first represents E as in (♣), and using that the system is totally reduced we obtain:
$$A = \begin{bmatrix} \mathsf{A}_1 & \mathsf{A}_2 \\ 0 & 0 \end{bmatrix}$$
2. (♢) yields $\mathsf{A}_2 = 0$.
3. One reduces A_1 to a Jordan canonical form by similarity transformation:
$$\mathsf{A}_1 = \mathsf{P}_1 \mathsf{J} \mathsf{P}_1^{-1}$$
4. It is now clear that we can put (E, A) in the form (♣) using a similarity transformation with the matrix:
$$\mathsf{P} := \begin{bmatrix} \mathsf{P}_1 & 0 \\ 0 & \mathbb{I} \end{bmatrix}$$

□

Second sweep of the decomposition

In the specific case of E being injective, the decomposition of Theorem 5.1 is slightly simplified.

Proposition 5.8.2. *If E is injective, then in the decomposition of Theorem 5.1 we get:*
$$\begin{aligned} \ker \mathsf{E}^{(\infty)} &= 0 \\ K^{(k)} &= 0 \qquad \forall k \geq 0 \end{aligned}$$

Now we make the following observation, stemming from the fact that by definition $\mathsf{E}^{(\infty)}$ is surjective (Proposition 5.3.3).

Proposition 5.8.3. *Given a linear IDE (E, A), the totally reduced operator $\mathsf{E}^{(\infty)*}$ is injective.*

Now, using Corollary 5.6.3, we are in a position to run the decomposition of Theorem 5.1 for the IDE $(\mathsf{E}^{(\infty)*}, \mathsf{A}^{(\infty)*})$ and obtain a decomposition of M and V.

Theorem 5.3. *On top of the decomposition given by Theorem 5.1, the spaces $M^{(\infty)}$ and $V^{(\infty)}$ may be decomposed as (cf. Figure 5.6):*
$$M^{(\infty)} = V_*^{(\infty)} \bigoplus_k W_*^{(k)}$$
$$V^{(\infty)} = M_*^{(\infty)} \bigoplus_k N_*^{(k)}$$
with $W_*^{(k)} = Z_*^{(k)} \oplus C_*^{(k)}$ and such that E and A are zero, except on the following blocks:

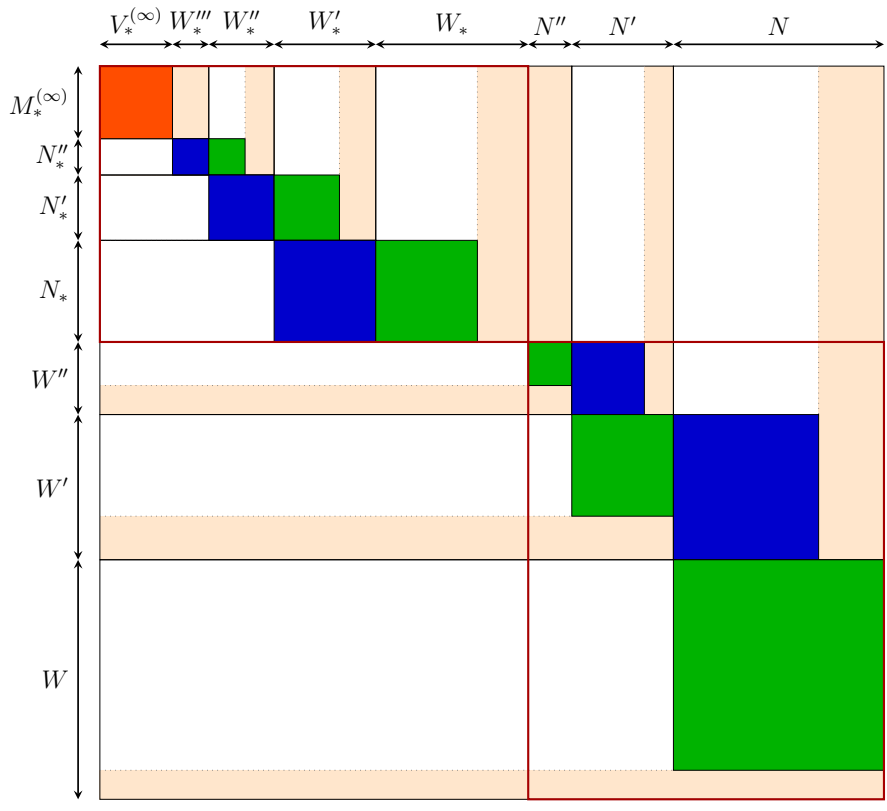


Figure 5.6: (See legend on p. 119). An illustration of the full decomposition. The first decomposition leads to M'' and the corresponding space $V''' = \mathsf{E}M''$, at which point the algorithm stalls. The second step consists in transposing the reduced operators $\mathsf{E}^{(\infty)}$ and $\mathsf{A}^{(\infty)}$, run algorithm, and transposing back again. The red area denotes the identity for E, and a non specific matrix for A. Notice that this block is completely separated from the rest, so one may now reduce the A to Jordan blocks by a similarity transformation.

(i) A *is the identity on the blocks* $(C_*^{(k)}, N_*^{(k)})$

(ii) E *is the identity on the blocks* $(W_*^{(k+1)}, N_*^{(k)})$

(iii) E *is the identity on the block* $(V_*^{(\infty)}, M_*^{(\infty)})$

(iv) A *is in Jordan form on the block* $(V_*^{(\infty)}, M_*^{(\infty)})$

Example 5.8.1. The following example is taken from [Sjö08, Example 2.4].

$$\mathsf{E} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \qquad \mathsf{A} = \begin{bmatrix} 2 & 0 & 0 & 1 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

The normal form for that system is:

$$\mathsf{E} = \left[\begin{array}{cccc|c} 1 & 0 & & & \\ & 1 & 0 & & \\ & & 1 & 0 & \\ \hline & & & & 0 \end{array}\right] \qquad \mathsf{A} = \left[\begin{array}{cccc|c} 0 & 1 & & & \\ & 0 & 1 & & \\ & & 0 & 1 & \\ \hline & & & & 1 \end{array}\right]$$

In that case, there is no dynamical part remaining.

5.9 Defect Indices

In order to facilitate the description of the various decompositions, we now decide on a way to write the various defects involved in the canonical decomposition (see Figure 5.7).

Definition 5.6. The ***constraint defects*** α are the dimensions of the space $K^{(k)}$:

$$\alpha_k(\mathsf{E}, \mathsf{A}) := \dim K^{(k-1)} \qquad k \geq 1$$

The ***overdetermination defects*** β^+ are the dimensions of the spaces $Z^{(k)}$:

$$\beta_k^+(\mathsf{E}, \mathsf{A}) := \dim Z^{(k-1)} \qquad k \geq 1$$

The ***underdetermination defects*** β^- are the overdetermination defects of the dual reduced system $(\mathsf{E}^{(\infty)*}, \mathsf{A}^{(\infty)*})$:

$$\beta_k^-(\mathsf{E}, \mathsf{A}) := \beta_k^+(\mathsf{E}^{(\infty)*}, \mathsf{A}^{(\infty)*})$$

Sometimes we will also denote the dimension of the remaining ***pure dynamical part***:

$$\delta := \dim \mathrm{Im}(\mathsf{E}^{(\infty)*})^{(\infty)}$$

It is important to notice that those indices are defined in an invariant manner, i.e. their definition does not depend on the chosen basis. In fact, it is possible to define directly in terms of invariant quantities as follows:

Proposition 5.9.1.

$$\alpha_k = \dim M^{(k-1)} - \dim M^{(k)} - (\dim V^{(k)} - \dim V^{(k+1)})$$
$$\beta_k^+ = \dim V^{(k-1)} - \dim(\mathsf{E}M^{(k-1)} + \mathsf{A}M^{(k-1)})$$

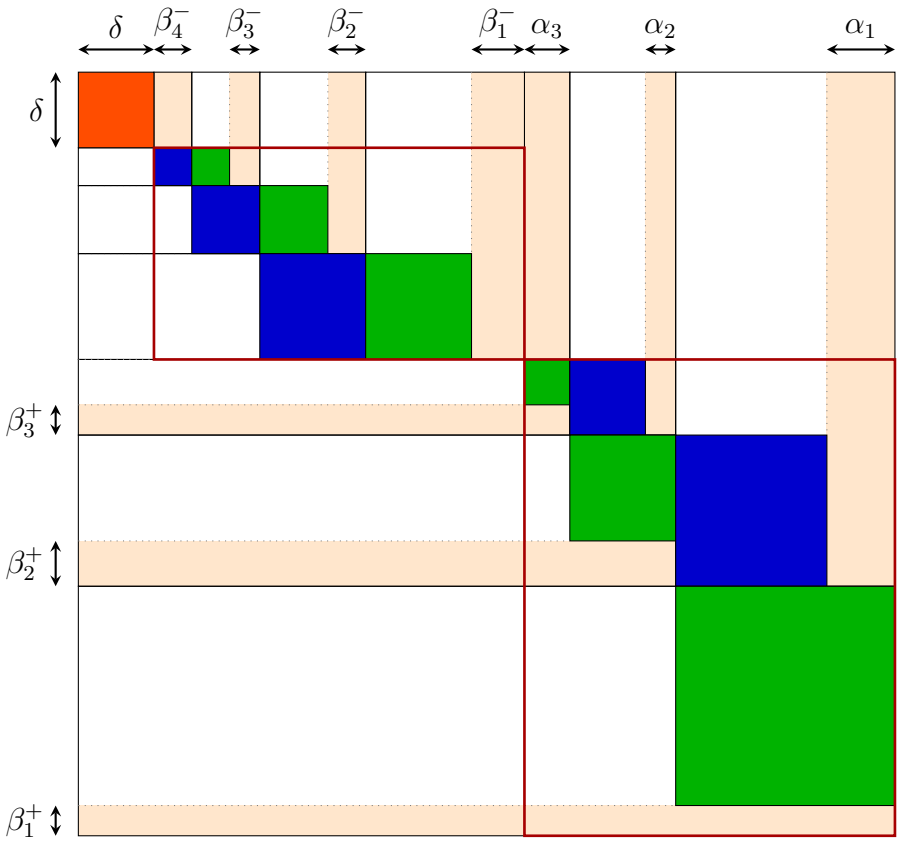


Figure 5.7: (See legend on p. 119). Illustration of Definition 5.6. The definition of the defects α, β^+ and β^-.

Proof. The first assertion follows from (5.5), the second one from the defining property of Z being a supplementary subspace, in (5.3). □

The following elementary observation stems from the definition of the reduction:

Proposition 5.9.2. *The relation between the defect indices of a system* (E, A) *and the corresponding reduced system* $(\mathsf{E}', \mathsf{A}')$ *is as follows:*

$$\left.\begin{aligned} \alpha_k(\mathsf{E}', \mathsf{A}') &= \alpha_{k+1}(\mathsf{E}, \mathsf{A}) \\ \beta_k^+(\mathsf{E}', \mathsf{A}') &= \beta_{k+1}^+(\mathsf{E}, \mathsf{A}) \\ \beta_k^-(\mathsf{E}', \mathsf{A}') &= \beta_k^-(\mathsf{E}, \mathsf{A}) \end{aligned}\right\} \qquad \forall k \geq 1$$

Proof. It follows from the inductive definition of the reduced operators E' and A'. □

The dimensions of the spaces W and N are related to the defect indices as follows:

Lemma 5.9.1.

$$\dim N^{(k)} = \sum_{j \geq k+1} (\alpha_j + \beta^+_{j+1})$$
$$\dim W^{(k)} = \sum_{j \geq k+1} (\alpha_j + \beta^+_j)$$

Proof. The second identity is proved using Corollary 5.6.2 and Proposition 5.9.2. The first identity follows from the second identity and Corollary 5.6.1. □

We also note that the constraint defects determine the index of the IDE

Proposition 5.9.3. *The index n of a linear IDE (E, A) is given by:*

$$n = \max\big\{k : \ \alpha_k(\mathsf{E}, \mathsf{A}) + \beta^+_{k+1}(\mathsf{E}, \mathsf{A}) \neq 0\big\}$$

Proof. Using the defining property of $N^{(k)}$, the index n may be defined as

$$n = \min_k \dim N^{(k)} = 0$$

Using Lemma 5.9.1 we obtain directly:

$$\dim N^{(k)} = 0 \iff \alpha_j + \beta^+_{j+1} = 0 \quad \forall j \geq k+1$$

which proves the claim. □

In the case of overdetermination defects we obtain readily:

Corollary 5.9.1. *The index n of an IDE (E, A) without overdetermination defects (i.e. $\beta^+ = 0$) is the biggest index of non-zero constraint defects:*

$$n = \max\big\{k : \ \alpha_k(\mathsf{E}, \mathsf{A}) \neq 0\big\}$$

The choice of the name "defect" may seem overly negative, but those indices really measure how far an IDE is from a standard, explicit and solvable ordinary differential equation. This is the essence of the following proposition:

Proposition 5.9.4. *An IDE with no defect (i.e. all defect indices are zero) is a totally reduced, solvable IDE.*

Proof. Lemma 5.9.1 yields that the system is totally reduced ($\dim N = 0$) and that E is surjective ($\dim W = 0$). Now since $\beta^- = 0$, then by reasoning on the dual system which is $\mathsf{E}^* = \mathsf{E}^{(\infty)*}$ (since we just proved that (E, A) was totally reduced), we finally obtain that E^* is surjective, i.e. E is injective, so the system is solvable. □

Proposition 5.9.5. *The dimensions of M, V, the defects α, β^+ and β^- are related by the following formulae:*

$$\dim M = \delta + \sum_{k\geq 1} k\alpha_k + \sum_{k\geq 1} k\beta_k^- + \sum_{k\geq 2} k\beta_k^+$$
$$\dim V = \delta + \sum_{k\geq 1} k\alpha_k + \sum_{k\geq 1} k\beta_k^+ + \sum_{k\geq 2} k\beta_k^-$$

Proof. The proof is by induction on the index using Lemma 5.9.1, Corollary 5.6.2 and Theorem 5.3. □

Example 5.9.1. The system of Example 5.6.2 has non-zero defects:

$$\beta_3^+ = 1,\ \beta_1^- = 1$$

Example 5.9.2. Let us examine the defects of some simple two-by-two systems. In the right column we only mention the non zero defect indices. For convenience we also indicate the size δ of the purely dynamical part, although it is easily computed using Proposition 5.9.5.

$$\left(\begin{bmatrix}1&0\\0&1\end{bmatrix}, \begin{bmatrix}0&0\\0&0\end{bmatrix}\right) \quad \text{all defects are zero, } \delta = 2$$
$$\left(\begin{bmatrix}1&0\\0&0\end{bmatrix}, \begin{bmatrix}0&0\\0&0\end{bmatrix}\right) \quad \beta_1^+ = \beta_1^- = 1,\ \delta = 1$$
$$\left(\begin{bmatrix}0&1\\0&0\end{bmatrix}, \begin{bmatrix}0&0\\0&1\end{bmatrix}\right) \quad \beta_2^+ = \beta_1^- = 1,\ \delta = 0$$
$$\left(\begin{bmatrix}1&0\\0&0\end{bmatrix}, \begin{bmatrix}0&1\\0&0\end{bmatrix}\right) \quad \beta_1^+ = \beta_2^- = 1,\ \delta = 0$$
$$\left(\begin{bmatrix}1&0\\0&0\end{bmatrix}, \begin{bmatrix}0&0\\0&1\end{bmatrix}\right) \quad \alpha_1 = 1,\ \delta = 1$$
$$\left(\begin{bmatrix}0&0\\0&0\end{bmatrix}, \begin{bmatrix}0&1\\0&0\end{bmatrix}\right) \quad \alpha_1 = \beta_1^+ = \beta_1^- = 1,\ \delta = 0$$
$$\left(\begin{bmatrix}0&0\\0&0\end{bmatrix}, \begin{bmatrix}1&0\\0&1\end{bmatrix}\right) \quad \alpha_1 = 2,\ \delta = 0$$

Example 5.9.3. The system presented in Example 5.8.1 has defect indices:

$$\alpha = (1, 0, \ldots)$$
$$\beta^+ = (0, \ldots)$$
$$\beta^- = (0, 0, 1, 0, \ldots)$$

Example 5.9.4. The mechanical system of Example 5.6.1 has defect indices $\beta^+ = 0$, $\beta^- = 0$ and:

$$\alpha = (0, 0, n, 0, \ldots)$$

where n is the number of constraints, i.e. the rank of G. Notice that in general if G has rank deficiency m then the β defect indices are not zero anymore and one has:

$$\beta_1^+ = \beta_1^- = m$$

The system is clearly solvable iff $m = 0$, i.e. if G has full rank.

5.10 Kronecker Decomposition

The Kronecker canonical form makes use of special blocks, each of which having a variant for the matrices E and A.

First the rectangular "L-***blocks***" $\mathsf{L}_k^{\mathsf{E}}$ and $\mathsf{L}_k^{\mathsf{A}}$:

$$\mathsf{L}_k^{\mathsf{E}} := \begin{bmatrix} 1 & & & \\ 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \\ & & & 0 \end{bmatrix} \Bigg\} k \qquad \mathsf{L}_k^{\mathsf{A}} := \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ & \ddots & \ddots & \\ & & 1 & 0 \\ & & & 1 \end{bmatrix} \Bigg\} k$$

One also uses the ***nilpotent blocks*** $\mathsf{N}_k^{\mathsf{E}}$ and $\mathsf{N}_k^{\mathsf{A}}$ as follows:

$$\mathsf{N}_k^{\mathsf{E}} := \begin{bmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & 0 & 1 \\ & & & & 0 \end{bmatrix} \Bigg\} k \qquad \mathsf{N}_k^{\mathsf{A}} := \begin{bmatrix} 1 & 0 & & & \\ & 1 & 0 & & \\ & & \ddots & \ddots & \\ & & & 1 & 0 \\ & & & & 1 \end{bmatrix} \Bigg\} k$$

Theorem 5.4. *A decomposition with defects α, β^+, β^- produces a Kronecker decomposition with:*

- α_k *block of type* N_k
- β_k^+ *blocks of type* L_k
- β_k^- *blocks of type* L_k^T

Proof. The proof is a combinatorial rearrangement of the basis of M and V provided by Theorem 5.3, best seen on Figure 5.8. □

Conjugate Decomposition

We may now show the relation between the decomposition of Theorem 5.1 on a pair (E, A) and on the adjoints $(\mathsf{E}^*, \mathsf{A}^*)$.

Theorem 5.5. *The conjugate decomposition switches the defects β^+ and β^-, i.e. it produces the following defects:*

$$\begin{aligned}\alpha(\mathsf{E}^*, \mathsf{A}^*) &= \alpha(\mathsf{E}, \mathsf{A})\\ \beta^+(\mathsf{E}^*, \mathsf{A}^*) &= \beta^-(\mathsf{E}, \mathsf{A})\\ \beta^-(\mathsf{E}^*, \mathsf{A}^*) &= \beta^+(\mathsf{E}, \mathsf{A})\end{aligned}$$

Weierstraß decomposition

In the case of square, solvable linear systems, the Kronecker decomposition is called the *Weierstraß decomposition* ([Wei68; HW96]) and is as follows:

$$\mathsf{E} = \begin{bmatrix}\mathbb{I} & 0\\ 0 & \mathsf{N}\end{bmatrix} \quad \mathsf{A} = \begin{bmatrix}\mathsf{C} & 0\\ 0 & \mathbb{I}\end{bmatrix}$$

where C may be in Jordan normal form and N is a block diagonal matrix of blocks of type N_k^E.

Then the matrix block N consists of nilpotent blocks as follows:

$$\mathsf{N} = \operatorname{diag}\Big(\mathsf{N}_{k_1}^E(0), \mathsf{N}_{k_2}^E(0), \ldots, \mathsf{N}_{k_m}^E(0)\Big)$$

Proposition 5.10.1. *The Weierstraß decomposition produces a decomposition with*

$$\alpha_k \text{ blocks } \mathsf{N}_k^\mathsf{E} \qquad 1 \leq k \leq n$$

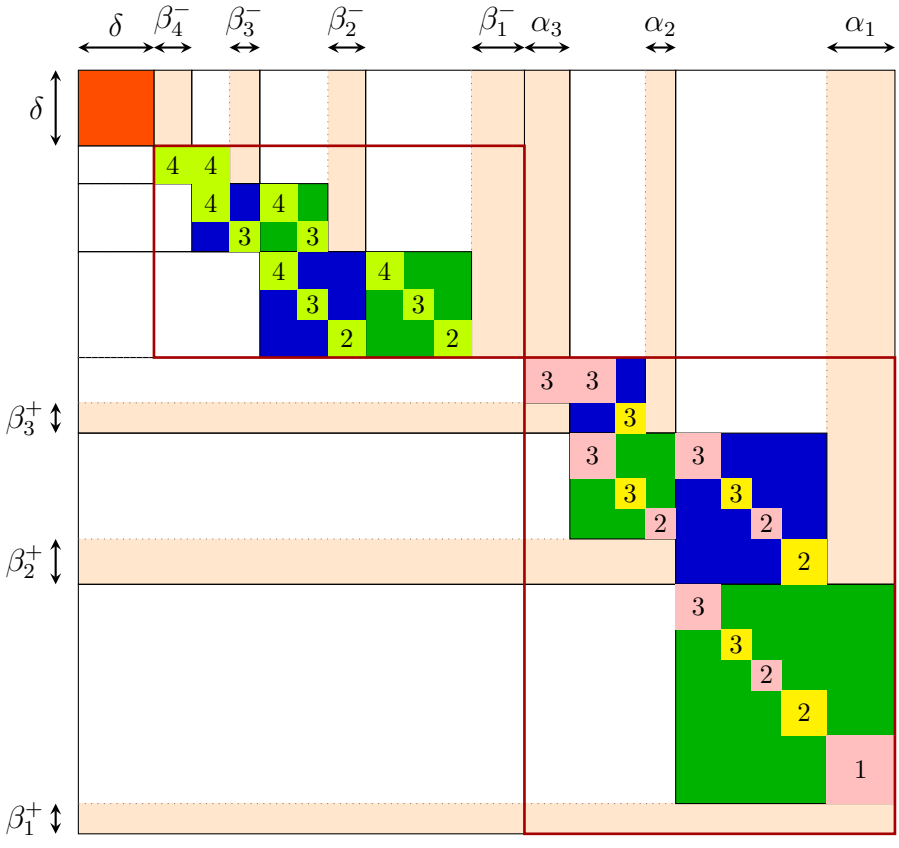


Figure 5.8: (See legend on p. 119). An illustration of Theorem 5.4. The difference of size of the squares is exactly given by the defects α, β^+ and β^-. The pink squares bearing the number j represent all the nilpotent blocks N_j; there are α_j such blocks. The yellow squares bearing the number j represent the L-blocks L_j. There are β_j^+ such blocks. The light green squares bearing the number j represent the L-blocks L_j^T. There are β_j^- such blocks.

Proof. It is just a special case of Theorem 5.4. □

5.11 Notes

The material presented in this Chapter is entirely new. The Kronecker decomposition Theorem is a well-know result in linear algebra ([Kro90], [Gan59, § XII.4], [GLR06, § A.7]), with applications in control theory and numerical analyis ([Kai79, § 6.3], [KM94], [ESF98, § 2.6.2], [HW96, VII.1]). Our proof has a geometric flavour that is lacking in the existing proofs. The same holds for our proof of the regular pencil theorem.

Some interesting references on variants of the Kronecker decomposition is to be found in [Joh05]. As far as we know, the decomposition presented in this Chapter has never been used for practical computation of the Kronecker canonical form.

BIBLIOGRAPHY

[AMR88] Abraham, R., Marsden, J. E., and Ratiu, T. *Manifolds, tensor analysis, and applications*, volume 75 of *Applied Mathematical Sciences*. Springer-Verlag, New York, second edition, 1988. isbn:978-0-387-96790-5.
(Cited p. 51)

[Arn88] Arnold, V. I. *Geometrical methods in the theory of ordinary differential equations*, volume 250 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, New York, second edition, 1988. isbn: 978-0-387-96649-6.
(Cited p. 59, 60)

[AVL91] Alekseevskij, D. V., Vinogradov, A. M., and Lychagin, V. V. *Basic ideas and concepts of differential geometry*. In *Geometry, I*, volume 28 of *Encyclopaedia Math. Sci.*, pages 1–264. Springer, Berlin, 1991. isbn: 978-3-540-51999-7.
(Cited p. 73)

[BCG+91] Bryant, R. L., Chern, S. S., Gardner, R. B., Goldschmidt, H. L., and Griffiths, P. A. *Exterior differential systems*, volume 18 of *Mathematical Sciences Research Institute Publications*. Springer-Verlag, New York, 1991. isbn:978-0-387-97411-8.
(Cited p. 73, 74)

[BG92] Bruce, J. W., and Giblin, P. J. *Curves and singularities*. Cambridge University Press, Cambridge, second edition, 1992. A geometrical introduction to singularity theory. isbn:978-0-521-41985-7.
(Cited p. 59, 60)

[CG95] Campbell, S. L., and Gear, C. W. *The index of general nonlinear DAEs*. Numer. Math., 72(2):173–196, 1995. doi:10.1007/s002110050165.
(Cited p. 64, 71)

[CHS00] Calmet, J., Hausdorf, M., and Seiler, W. *A constructive introduction to involution*. ISACA 2000, World Scientific, Singapore, 2000. Available from: http://iaks-www.ira.uka.de/calmet/papers/isaca2.ps.
(Cited p. 73)

[CM95] Campbell, S. L., and Moore, E. *Constraint preserving integrators for general nonlinear higher index DAEs*. Numer. Math., 69(4):383–399, 1995. doi:10.1007/s002110050099.
(Cited p. 73)

[ESF98] Eich-Soellner, E., and Führer, C. *Numerical methods in multibody dynamics*. European Consortium for Mathematics in Industry. B. G. Teubner, Stuttgart, 1998. Second Edition 2008. Available from: http://www.maths.lth.se/na/staff/claus/NMMD2/, isbn:978-91-631-2929-2.
(Cited p. 111)

[Gan59] Gantmacher, F. R. *The theory of matrices. Vols. 1, 2*. Translated by K. A. Hirsch. Chelsea Publishing Co., New York, 1959.
(Cited p. 99, 111)

[GLR06] Gohberg, I., Lancaster, P., and Rodman, L. *Invariant subspaces of matrices with applications*, volume 51 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2006. Reprint of the 1986 original. isbn:978-0-89871-608-5.
(Cited p. 111)

[GP84] Gear, C. W., and Petzold, L. R. *Ode methods for the solution of differential/algebraic systems*. SIAM Journal on Numerical Analysis, 21(4):716–728, 1984. doi:10.1137/0721048.
(Cited p. 68)

[Her65] Hermann, R. *E. Cartan's geometric theory of partial differential equations*. Advances in Math., 1(fasc. 3):265–317 (1965), 1965. doi:10.1016/0001-8708(65)90040-X.
(Cited p. 73, 74)

[HLR89] Hairer, E., Lubich, C., and Roche, M. *The numerical solution of differential-algebraic systems by Runge-Kutta methods*, volume 1409 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1989. isbn:978-0-387-51860-2.
(Cited p. 67, 68)

[HMT03] Higueras, I., März, R., and Tischendorf, C. *Stability preserving integration of index-1 DAEs*. Appl. Numer. Math., 45(2-3):175–200, 2003. doi:10.1016/S0168-9274(02)00215-5.
(Cited p. 67)

[HNW93] Hairer, E., Nørsett, S. P., and Wanner, G. *Solving ordinary differential equations. I*, volume 8 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, second edition, 1993. Nonstiff problems. isbn:978-3-540-56670-0.
(Cited p. 59)

[HW96] Hairer, E., and Wanner, G. *Solving ordinary differential equations. II*, volume 14 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, second edition, 1996. Stiff and differential-algebraic problems. isbn:978-3-540-60452-5.
(Cited p. 73, 109, 111)

[Joh05] Johansson, S. *Canonical forms and stratification of orbits and bundles of system pencils*. Technical report, Umeå University, Department of Computing Science, 2005. Available from: http://www.cs.umu.se/~stefanj/research.html [as of 2009-05].
(Cited p. 111)

[Kai79] Kailath, T. *Linear Systems*. Prentice-Hall Information and System Science Series. Prentice Hall, November 1979. isbn:978-0-13-536961-6.
(Cited p. 111)

[KM94] Kunkel, P., and Mehrmann, V. *Canonical forms for linear differential-algebraic equations with variable coefficients*. J. Comput. Appl. Math., 56(3):225–251, 1994. Available from: http://germain.math.tu-berlin.de/ePrints/papers/pdf/KunM94a.pdf [as of 2009-03], doi:10.1016/0377-0427(94)90080-9.
(Cited p. 90, 111)

[Kro90] Kronecker, L. *Algebraische Reduction der Schaaren bilinearer Formen*. Sitzungberichte der Königlich Preußischen Akademie der Wissenschaften zu Berlin, pages 1225–1237, 1890.
(Cited p. 111)

[Kur57] Kuranishi, M. *On E. Cartan's prolongation theorem of exterior differential systems*. Amer. J. Math., 79:1–47, 1957. doi:10.2307/2372381.
(Cited p. 73)

[Lie85] Lie, S. *Allgemeine Untersuchungen über Differentialgleichungen, die eine continuirliche, endliche Gruppe gestatten*. Math. Ann., 25(1):71–151, 1885.
(Cited p. 46)

[Lie77] Lie, S. *Geometrie der Berührungs transformationen*. Chelsea Publishing Co., New York, corrected edition, 1977. With editorial assistance by Georg Scheffers. isbn:978-0-8284-0291-0.
(Cited p. 46)

[Mal72] Malliavin, P. *Géométrie différentielle intrinseque.* Number 14 in Collection Enseignement des sciences. Hermann, 1972. isbn:978-2-7056-5696-6.

(Cited p. 76)

[MLRR99] Muñoz-Lecanda, C., M., and Roman-Roy, N. *Implicit quasilinear differential systems: a geometrical approach.* Electron.n. J. Diff. Eqns., 1999(10):1–33, 1999. Available from: http://ejde.math.txstate.edu/Volumes/1999/10/abstr.html [as of 2009-03].
(Cited p. 73)

[MT98] Macutan, Y. O., and Thomas, G. *Theory of formal integrability and daes: effective computations.* Numerical Algorithms, 19(1):147–157, 1998. doi:10.1023/A:1019162624913.
(Cited p. 71)

[Olv93] Olver, P. J. *Applications of Lie groups to differential equations*, volume 107 of *Graduate Texts in Mathematics.* Springer-Verlag, New York, second edition, 1993. isbn:978-0-387-94007-6.
(Cited p. 46)

[Olv95] Olver, P. J. *Equivalence, invariants, and symmetry.* Cambridge University Press, Cambridge, 1995. isbn:978-0-521-47811-3.
(Cited p. 74)

[Rei90] Reich, S. *Beitrag zur Theorie der Algebrodifferentialgleichungen.* PhD thesis, TU Dresden, 1990. gbook:KPbvHAAACAAJ.
(Cited p. 70, 73)

[Rei92] Reich, S. *Existence and uniqueness results for nonlinear differential-algebraic equations.* Seminar notes, Humboldt Universität zu Berlin, 1992. Available from: http://www.mathematik.hu-berlin.de/publ/SB-92-1/s_dae.html [as of 2009-04].
(Cited p. 78)

[RLW01] Reid, G. J., Lin, P., and Wittkopf, A. D. *Differential elimination-completion algorithms for DAE and PDAE.* Stud. Appl. Math., 106(1):1–45, 2001. Available from: http://www.math.nus.edu.sg/~matlinp/WWW/SiAM_ReidLinWittkopf_01.pdf [as of 2008-08], doi:10.1111/1467-9590.00159.
(Cited p. 66, 71, 73)

[RR94] Rabier, P. J., and Rheinboldt, W. C. *A geometric treatment of implicit differential-algebraic equations.* J. Differential Equations, 109(1):110–146, 1994. doi:10.1006/jdeq.1994.1046.
(Cited p. 70, 73)

[Sei99] Seiler, W. M. *Indices and solvability for general systems of differential equations.* In *Computer algebra in scientific computing—CASC'99 (Munich)*, pages 365–385. Springer, Berlin, 1999. Available from: http://www.mathematik.uni-kassel.de/~seiler/Papers/HTML/Index.html [as of 2009-03].
(Cited p. 73)

[Sjö08] Sjöberg, J. *Optimal Control and Model Reduction of Nonlinear DAE Models.* Linköping studies in science and technology. dissertations. no. 1166, Linköping University, April 2008. Available from: http://www.control.isy.liu.se/publications/doc?id=2063 [as of 2009-04], isbn:978-91-7393-964-5.
(Cited p. 104)

[Sto00] Stormark, O. *Lie's structural approach to PDE systems*, volume 80 of *Encyclopedia of Mathematics and its Applications.* Cambridge University Press, Cambridge, 2000. isbn:978-0-521-78088-9.
(Cited p. 46, 54)

[Val45] Valiron, G. *Équations Fonctionnelles. Applications.* Masson et Cie, Paris, 1945. isbn:978-2-87647-061-3.
(Cited p. 46)

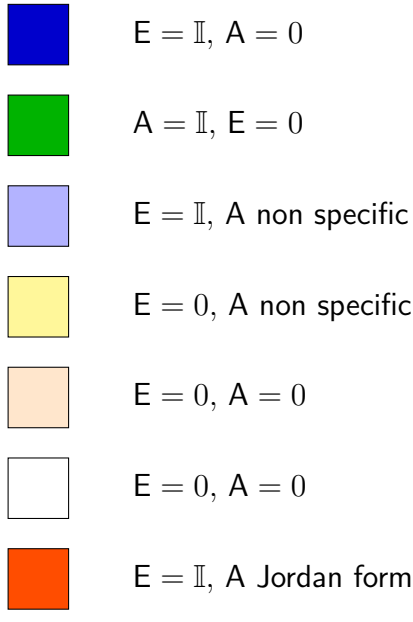
[Vin84] Vinogradov, A. M. *Local symmetries and conservation laws.* Acta Appl. Math., 2(1):21–78, 1984. Available from: http://diffiety.ac.ru/djvu/local_symmetries.djvu [as of 2008-08], doi:10.1007/BF01405491.
(Cited p. 46, 56)

[vNRM98] van Nieuwstadt, M., Rathinam, M., and Murray, R. M. *Differential flatness and absolute equivalence of nonlinear control systems.* SIAM J. Control Optim., 36(4):1225–1239, 1998. Available from: http://caltechcdstr.library.caltech.edu/6/ [as of 2009-03], doi:10.1137/S0363012995274027.
(Cited p. 74)

[Wei68] Weierstrass. *On the theory of bilinear and quadratic forms. (Zur Theorie der bilinearen und quadratischen Formen.).* J. Berl. Monatsber., 1868.
(Cited p. 109)

Legend

Legend of the various matrix figures.



List of Symbols

Part I

T Burgers operator, 23

$\mathrm{D}_*^{\frac{1}{2}}$ Dual operator of $\mathrm{D}^{\frac{1}{2}}$, 17

D^s Fractional derivative operator, 17

D_*^s Adjoint differentiation operator, 17

$u_{\sqrt{t}}$ $\mathrm{D}^{\frac{1}{2}} u$, 22

$\mathcal{D}(\mathbb{T}, H)$ Hilbert space-valued periodic test functions, 16

$\mathcal{D}'(\mathbb{T}, H^*)$ Space of periodic, Hilbert space valued distributions, 16

$u_{\sqrt{t}*}$ $\mathrm{D}_*^{\frac{1}{2}} u$, 22

$\langle f, u \rangle$ Duality bracket between $\mathrm{H}_0^{\left(\frac{1}{2},1\right)}$ and its dual $\mathrm{H}^{\left(-\frac{1}{2},-1\right)}$, 22

$\mathrm{H}_0^{\left(\frac{1}{2},1\right)}$ $\mathrm{H}_0^{\left(\frac{1}{2},1\right)}(Q)$, 22

$\widetilde{u}$ $\mathcal{H} u$, 22

H Complex Hilbert space, 16

$\mathcal{H}$ Hilbert transform, 17

$\mathrm{H}_{N+}^{(1,2)}$ Cone of strictly positive functions in $\mathrm{H}_N^{(1,2)}$, 31

$\mathrm{H}_{N+}^{(1,2)}/\mathbb{R}_+$ Quotient of $\mathrm{H}_{N+}^{(1,2)}$ by multiplication by positive numbers, 32

$\mathrm{H}^{\left(-\frac{1}{2},-1\right)}$ $(\mathrm{H}_0^{\left(\frac{1}{2},1\right)})^*$, 22

$\mathrm{H}_N^{(1,2)}$ Sobolev space with Neumann boundary conditions, 31

$\mathrm{H}_N^{(1,2)}/\mathbb{R}$ Quotient of $\mathrm{H}_N^{(1,2)}$ by addition of constants, 32

$\mathcal{I}$ Interval $(0,1)$, 22

L^p $\mathrm{L}^p(Q)$, 22

$\mathcal{L}$ Linear part of the Burgers equation, 22

μ Normalized viscosity, 21

$\|u\|$ $\|u\|_{\mathrm{H}_0^{(\frac{1}{2},1)}}$, 22

$|u|$ L^2 norm of u, 22

ν Viscosity, 21

(u,v) L^2 scalar product on Q, 22

Q Cylinder $\mathbb{T}\times\mathcal{I}$, 22

sgn Sign function, 17

$\mathrm{H}^{(-\alpha,-\beta)}(\mathbb{T}\times\mathcal{I})$ Dual of $\mathrm{H}_0^{(\alpha,\beta)}(\mathbb{T}\times\mathcal{I})$, 18

$\mathrm{H}^{(\alpha,\beta)}(\mathbb{T}\times\mathcal{I})$ Sobolev space with regularity α in time and β in space, 18

$\mathrm{H}^{(s)}(\mathbb{T},H)$ Fractional Sobolev spaces, 18

$\mathrm{H}^{(\alpha)(\beta)}(\mathbb{T}\times\mathcal{I})$ Sobolev space with regularity α in time, taking values in a Sobolev space with regularity β in space, 18

S_1 First stage of the Cole-Hopf transformation, 32

S_3 Final stage of the Cole-Hopf transformation, 32

S_2 Intermediate stage of the Cole-Hopf transformation, 32

$\mathbb{T}$ One dimensional torus, 16

S Nonlinear part of the Burgers equation, 23

Part II

A Operator for linear IDEs, 75

$\mathsf{A}^{(\infty)}$ Totally reduced operator A, 83

$\mathsf{A}|_{M'}$ Restriction of A on M', 78

d Differential, 76

E Operator for linear IDEs, 75

$\mathsf{E}^{(\infty)}$ Totally reduced operator E, 83

$\mathsf{E}|_{M'}$ Restriction of E on M', 78

V Vector space, 76

$\operatorname{ind}_{\mathcal{M}} \xi$ Index, 63

J_0 Zero-th order jet-space, 55

J_1 First order jet-space, 55

$\mathcal{M}$ Ambient manifold; space-time, 52

$\mathsf{L}^{\mathsf{A}}_k$ A part of the Kronecker L-blocks, 108

$\mathsf{L}^{\mathsf{E}}_k$ E part of the Kronecker L-blocks, 108

N Matrix of nilpotent blocks, 109

$\mathsf{N}^{\mathsf{A}}_k$ A part of the nilpotent blocks, 108

$\mathsf{N}^{\mathsf{E}}_k$ E part of the nilpotent blocks, 108

θ Vector-valued differential form for a linear IDE, 76

M Domain of the linear IDEs, 75

M' Reduced domain of linear IDEs, 77

$M^{(\infty)}$ Totally reduced domain M, 83

N Supplementary space from M' to M, 84

$\overline{\mathcal{M}}$ Reduced set, 62

$\overline{M}$ Complement of $M^{(\infty)}$, 96

$\overline{V}$ Complement of $V^{(\infty)}$, 96

V Codomain of linear IDEs, 75

V' Reduced codomain, 82

$V^{(\infty)}$ Totally reduced codomain V, 83

$\dot{x}$ Jet-space coordinate, 56

$W^{(k)}$ Supplementary space from $V^{(k+1)}$ to $V^{(k)}$, 86

Z Supplementary space from $V' \oplus \mathsf{A}N$ to V, 84

C Supplementary space of the coupled constraints, 89

K Supplementary space in $\ker \mathsf{E}$, 88

Ω Dynamics codistribution, 52

α Constraint defect, 104

β^+ Overdetermination defect, 104

β^- Underdetermination defect, 104

π_0 Jet-space projection, 55

τ Simultaneity codistribution (equal to $\mathrm{span}(\mathrm{d}t)$ if not otherwise prescribed), 52

φ_* Linearized map (Jacobian), 52

Δ Determinacy degree, 79

IDE Implicit Differential Equation, 52

Index