FUNCTIONAL SYSTEM DYNAMICS

Norbert E. Ligterink

University Twente, The Netherlands
E-mail address: n.e.ligterink@utwente.nl

Key words and phrases. Modelling, Control, System Theory, PDE, Dynamics

ABSTRACT. Functional system dynamics is the analysis, modelling, and simulation of continuous systems usually described by partial differential equations. From the infinite degrees of freedom of such systems only a finite number of relevant variables have to be chosen for a practical model description. The proper input and output of the system are an important part of the relevant variables.

Proefschrift Universiteit Twente

graduation committee:
J. van Amerongen, A. J. van der Schaft, P. C. Breedveld, A. de Boer,
B. M. Maschke, J. J. W. van der Vegt, G. van Schothorst

Universiteit Twente

FUNCTIONAL SYSTEM DYNAMICS

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Twente,
op gezag van de rector magnificus,
prof.dr. W. H. M. Zijm,
volgens besluit van het College voor Promoties
in het openbaar te verdedigen
op vrijdag 31 augustus 2007 om 13.15 uur

door

Norbert Emiel Ligterink

geboren op 26 augustus 1968
te Enschede

Dit proefschrift is goedgekeurd door:

promotoren: Prof.dr. A. J. van der Schaft
                      Prof.dr.ir. J. van Amerongen

assistent-promotor: Dr.ir. P. C. Breedveld

Contents

Foreword

... untutored mind
His soul proud Science never thought to stray,
Far as the solar walk, or milky way;
Yet simple nature to his hope has given,
Behind the cloud-topped hill a humbler heaven.
[ALEXANDER POPE, AN ESSAY ON MAN]

In many respects this thesis is different from others. It does not treat a single topic in the traditional sense. It does not give a concise mathematical theory from a single perspective. This thesis is a collection of issues concerning the modelling of infinite-dimensional, continuous, or distributed, systems. The writer is someone who will first make the back-of-the-envelope calculation, before deciding on more complex tools such as finite element methods (FEM). He hopes the thesis will bridge some gaps between schools, which are based on tools rather than systems and problems, of thought concerning modelling.

To deceive yourself, and me with you, you speak too much.
[GEORGES BERNANOS, SOUS LE SOLEIL DE SATAN]

In some cases, for example, where well-known material is covered for completeness, the style from this work is very concise. Facts essential to the story are repeated three times in Hollywood movies, and every audience is triggered by different words and images. So the writer decided to use multiple views and different approaches to the same problems and methods, in the hope that many angles are covered. However, rather than repeating known and classical arguments, the thesis is written in the hope to supply a new insight in familiar and less familiar topics.

Eventually the understanding of systems is the main goal of most system engineers. Designing an optimal system for a particular purpose is not a process of trying to hit it on the head. Something has to be learned, even from complete failures. Insight in a system is not possible without having a language to think and talk about a system. It is the breadth of this language the thesis hopes to bring across. The ordering of topics and chapters is such that the physical examples such as elasticity and fluid dynamics appear and return throughout the thesis. Moreover, also the mathematical tools, such as finite elements and modal analysis, appear in several chapters. The language is that of variables, invariants and physical principles. In the theory of system engineering they have be introduced in different manners by different people. If this thesis follows any of these approaches more closely, it is the theory of port-based, or bond graph, modelling, introduced by Paynter, [Paynter, 1961] for discrete, or lumped, systems in the fifties.

> The less a science has advanced, the more its terminology to rest
> on a uncritical assumption of mutual understanding.
> [W.V. Quine, Truth by convention]

The principal concept is energy and energy conservation throughout the system. Each component can be described by variables or states. Changing a variable will result in a change in energy. The product of the enforced change and the system, or component, response is the power. Hence pair of variables: the action and the reaction, or the effort and the flow, are the power, or port, variables, common to all physical systems. A network representation expressing the possible interactions in terms of efforts and flows is called a bond graph.

From the other side energy also has played a crucial role in the underlying mathematics of distributed systems. Distributed systems are described by partial differential equations. In order to prove uniqueness and existence of a solution to a particular problem definition, called the initial boundary value problem, mathematicians have resorted often to defining an associated energy function. The energy function is positive and it has a unique minimum: the zero-energy solution, or the the null solution. The existence of the minimum is the basis of the existence of the solution, and that the difference of two solutions has zero energy proves the uniqueness of the solution.

In this thesis the opposite direction is followed. Rather than searching for an energy function for a problem, we state the problem in terms of its energy function or Hamiltonian, and seek admissible interactions, or boundary conditions, for this problem. The approach is similar to that of stability analysis using Lyapunov theory. Finding a Lyapunov function for a problem expressed in terms of equations of motion is difficult. However, for systems given by Hamiltonians, both the Lyapunov function and the equations of motion follow trivially, yielding essentially stable dynamics.

> I can't apologize for the banality of my dreams
> [Alice Munro, Bardon Bus]

If this thesis is to convey at least one message, this message is that physical principles should be retained in the numerical analysis of open, or connected, systems described by partial differential equations. As a consequence the boundary conditions are not the mathematical prerequisites for the analysis, but determine the essence of the relation of a system with its surroundings, and thereby the function of the system.

CHAPTER 1

Introduction

I was in a printing house in Hell and saw the method in which
knowledge is transmitted from generation to generation.
[WILLIAM BLAKE, MARRIAGE OF HEAVEN AND HELL, 1790]

Functional system dynamics is a convenient name for the study of continuous, distributed, or infinite-dimensional, systems from the perspective of effective low-order approximations in terms of a small number of modes, which are chosen appropriately and may depend on external parameters and conditions. Back-of-the-envelope calculations have been a tool of experts to estimate unwanted effects such as vibrations and deformations. However, in the cases these effects could not be ignored, a large effort is required to get a quantitative handle. The resulting analysis blocks fast development, and strains the relations between design, development, and testing.

1.1. The sad history of operational calculus

In the first half of the twentieth century there has been an ambitious project called operational calculus. [Heaviside, 1950, Churchill, 1944, van der Pol and Bremmer, 1959, Wiener, 1949] It aimed to simplify and automate the calculations of engineering mathematics, such that with limited understanding of the underlying theory, engineers could solve dynamical equations from tables and simple procedures, called calculus grammar. These procedures should be as simple as the rules used to differentiate functions. Not all derivatives are known. But with a basic set, and elementary rules a high-school student is able to apply differentiation rules to most functions at hand. The same set of procedures should have made it possible to solve any dynamical problem from an electric circuit, or a mechanical diagram, without much effort.

This program of operational calculus has died a silent death, unless one takes the implementation of mathematics in economics, called operations research as its successor and one takes it serious. Wiener considered Fourier-Laplace theory the successor of operational calculus, [Wiener, 1949] however, its aim is less ambitious. There are three main reasons for the failure of the program. First of all, the goal of the engineer was not to produce a specific numerical answer for a specific model in a specific case, but to investigate a system, determine its critical parts, and optimize it. This investigation is part of the design process. By decoupling the problem formulation from the problem solving, by lifting it and removing all design parameters, which are fixed in the numerical analysis, the feedback into the design process became impossible.

Secondly, the lifted program, without references to actual systems was too abstract and hard to understand for people who did not already apply the same methods in context to explicit cases. In lifting the problem all the irrelevant features for solving it are removed, such that a general but abstract problem remains, which can be treated with the generic methods of operational calculus. The abstraction made it harder to apply and check the method; engineers like most scientists use intuition and experience to guide them, which was not possible in the lifted approach.

Thirdly, computer modelling became the competition of operational calculus. Engineers discovered direct numerical implementation as a reasonably fast, trouble free, and accurate enough method to find results for their problems. It requires little mathematics to turn a partial differential equation into a finite element code. The ease of approximate numerical solutions plague all mathematicians, who require some strictness and bounds. Engineers seem to communicate in terms of toolboxes associated with software like Matlab. The inconsistencies and approximations required in such code become part of the engineering folklore. The situation could have been very different if computers were developed twenty years later. We might have had direct implementations of operational calculus, which now slowly arises through algebraic programming, implemented in computer programs like Reduce, Matlab, Maple, and Mathematica.

System theory may suffer from the same problems as operational calculus does. The engineer who has the experience and intuition on a particular field is not likely to resort to abstract methods to find answers. On the other hand his tools may be limited to a set of basic and approximate methods, which might fail for higher complexity or criticality of the system.

In order not to detach from the practice of the engineer one should formulate a system in the typical physical parameters which lie at the basis of such systems, and retain these parameters and variables throughout the formulation of problem, the method of solving, and the results. The physical principles underlying the system should be stressed by the method in order to pinpoint the critical elements of the design.

1.2. Complexity

Even more important than the increasing complexity of the methods available for analysis is the increasing complexity of the engineering systems. In the old days machines had only few moving parts, but with the industrial revolution machines, even simple ones like sewing machines and typewriters, have hundreds of moving parts. The complexity has increased significantly. However, before the electrical revolution the parts of a machine or system acted coherently, driven by a rotating single power source. With the revolution of the axis all the phases or states of the machine were determined. In the second half of the nineteenth century hydraulic systems in ships were the first systems where lags, or delays, played an important role. [Bennett, 1979] Effects of a single retardation were already known for governors of steam engines.[Maxwell, 1868] The dynamics of these systems was still limited. The single steam engine drove the single axis that powered the factory or ship. Electrical power changed all that. The parts, separately powered, have their own characteristics and timing.

Electrical engineers made things easy for themselves. The components they used were designed to have a single characteristic, commonly known as resistive,

capacitive, and inductive. In other words, they created a one-to-one mapping between ideal concept and physical component. It made them capable of designing complex electrical networks with hundreds, thousands of elements. Furthermore, these networks could again be separated into parts, each with its own function, such as bridges, delays, amplifiers, and switches. With electrical engineering the modular approach to engineering design was born. No longer a single engineer needed to be able to build and understand the machine as a whole. The components were designed and optimized independently and put together at a later stage. In the same way mathematicians use functions and algorithms as building blocks for more complex theory, and computer programmers write functions, subroutines, and libraries.

This vision inspired the analysis in other engineering disciplines. Mechanical parts of machines and chemical processes were decomposed in logical units. These logical units required connectors similar to the wires in electrical circuits. Different views led to different types of connectors. In the most general view an abstract signal was transmitted from one component to another. The term transmission implies a direction of transmission. In the behavioral approach to systems theory this direction remains absent, [Polderman and Willems, 1998] the system is only interconnected. Although an elegant formulation and an appropriate starting point for modelling a system, [Paynter, 1961] it misses the power of physical principles and heuristic insight. The structure of input and output; of cause and effect, and the conservation of a quality, such as energy and mass, yield natural restrictions within the framework of behavioral models, which are essential for consistency.

The general terms of input and output are reserved for such generic signals. However, in practice more details are known about the system and the signals. In the case of electrical networks, the Kirchhoff laws restrict the signals of currents and voltages such that charge conservation and energy conservation are implicit in the formulation of the system.

Physicists and chemists also have their own conservation laws: conservation of mass, charge, energy, momentum, and so on. Not all mathematical models of these systems satisfy all the conservation laws. In some cases it is extremely difficult, or even impossible to satisfy all conservation laws that apply, in an approximate model of the system. However, the restriction conservation laws impose on a model of a system are important for the quality of the model. If a conservation law is not built into the model of a system, it will be a good test for the validity of a model.

1.3. Port-based modelling

> When you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely in your thoughts advanced to the state of Science, whatever the matter may be.
> [W. Thomson, Lord Kelvin]

In order to connect components their relation should be clear. Newton wrote about action and reaction, where one component has the initiative, and the other component responses. The relation of cause and effect. However, the cause and effect can also be seen at more equal footing: if two ships are pulling, which is the

cause, and which is the effect? There is a single force and a single motion of the point on the line connecting the two ships. In the case of port-based modelling, the connection is the port defined by force and velocity, and which is the cause and which the effect are not determined. In the mental landscape of humans, we want to distinguish between cause and effect. Even if components seem to be entangled in an inseparable manner, they have been disentangled by picking a starting point and following the thread from one to the other and back. The interdependence of elements often gives rise to philosophical discussions, observational studies, and statistical studies, which only seem to confirm and re-confirm this interdependence. A pragmatic scientist or an engineer will choose, so to say, the x-axis and y-axis and will try to disentangle the intrinsic properties of each element and their relations, without deciding on the actor and the reactor among each two. Analysis is precisely that; to decompose into elements.

However, for simulation one requires to decide on cause and effect, or causality. In the case of the two ships tied together, we have to say, in order to resolve the impasse, that for the first ship it yields a force given a certain velocity, while for the other ship the force is given and the velocity is the result. In that way the two can be combined and the result match to yield a consistent dynamical situation with a force and a velocity of the whole system of two ships. For the model of the two ships it is not required to determine which is the actor: the force, and which will response with a velocity. However, for the simulation, and any numerical result it is relevant. Two pulling ships is in this case also a cumbersome combination. The common view is to see both as sources of force. The common causality is to have sources of force as precisely that: a given source causing a velocity or velocity change. Similarly, in electric networks it is common to have a potential source. The network as a whole responses by a certain current.

A connection, or port, consists of a pair of variables. The product of the two has the dimension of power, the rate of change of the energy. Such a general physical definition makes is possible to connect all kinds of different components together and still have a consistent model of the whole. This type of modelling is called bond-graph modelling, or port-based modelling. For discrete or lumped systems, such as networks or rigid-body dynamics it is a successful method. For continuous systems, described by partial differential equations, the approach is under development. There are no fundamental problems to generalize the approach. However, the properties and phenomena of the dynamics of continuous systems are much more involved than the dynamics of lumped components.

Part of the problem is to simplify, reduce, or restrict the continuous system to such an extent it can be used in the multicomponent analysis for simulation and control. So far the modelling of systems described by partial differential equations are research projects in their own right. The advent of multiphysics FEM packages show that these tasks can be automated. On the other hand a lot software deals with static solutions while dynamics requires time-dependent analysis and simulation. Again energy can play an important role in this case. The stability of solutions of partial differential equations are often analyzed using energy methods. Therefore methods of simulation and control of continuous systems based on energy methods have two advantages. First the simulations, based on energy principles, for a energy preserving model are inherently stable in the sense that given the energy the states are restricted to a subspace of the full state space. Secondly, the energy

methods yield automatically the boundary conditions which can be used to add the component model of the continuous system as part to a bond-graph model.

The natural boundary conditions that arise in energy-based methods are directly related to particular cases of ports. Many standard boundary conditions correspond to the absence of power flow through the boundary. These systems are isolated. More general boundary conditions have to be defined for modelling all the possible interactions, connections, or ports feasible. The definition of the change of energy in the system is augmented with a flow of energy through the boundary. In some respects such an approach resembles the systems of conservation laws, with one fundamental difference. The conserved energy is not a degree of freedom, or state, of the system. It can be expressed in terms of pairs of variables, like the charge density and the potential, or the volume and the pressure. In nature these pairs of variables are similar to the pairs of port variables. They are the bulk and boundary pairs of the same concept. The linkage between the two is the geometry.

The domain and its boundary are not only aspects of the topology of an object. Differential geometry links the two as well via Green identities and their generalizations by Cartan, Hodge, and Whitney. [Hodge, 1941, Ivey and Landsberg, 2000, Whitney, 1957] Another view onto the linkage of domain and boundary, between conserved quantities and conserved currents is offered by functional analysis. The energy is a norm on spaces of functions, i.e., the state variables of continuous systems. Combinations of operators on the domain have residual boundary operators. The Green identities yield relations between norms on the space and the boundary; the weak equivalent of the change of energy on a domain and its flow through the boundary.

The functional analysis [Yosida, 1980] formulation has one great advantage over the differential geometric formulation. Functional analysis allows us to restrict the function spaces, such that the results carry over from the continuous theory to any finite order discrete approximation we wish to use for numerical evaluation. In fact many of the results of functional analysis are based on the convergence of a set of approximations to the true solution in the continuous space. The typical convergence is called weak convergence as no distinction between some solutions, differing for example only on a finite set of positions, can be made. However, the results of differential geometry are not spoilt. The approximations in function space are made with the knowledge of the differential operators on elemental domains in the back of our mind. The construction of the boundary operator in the approximate space will be much simpler with the infinite continuous examples at hand.

1.4. Stages of existence

The numerical model of a system as a whole is the final stage before the numerical simulation and analysis. Parameters need to acquire a value, components need to be connected, and global boundary conditions need to be set to specify the dynamical situation in which the system is to be considered. Once this final stage is reached, the numerical model is finalized and few of the decisions in the process can be reverted easily, in order to study the effects of these decisions in the numerical results.

The power of simply analytical, often lumped, models is the easy study of parametric dependence, such as the material choice, and design decisions, such as

the manner in which components are interconnected. Furthermore, these interconnections are usually much more easily understood than the boundary conditions, internally and with the environment and the control in a complex distributed or FEM model. Many of the choices made to arrive at a numerical model are inherent in the approach and the tradition.

The models of the systems investigated in this thesis go through many stages of existence. From a system first the relevant qualities are selected, such as conductance in an electrical setting, or mass density in a mechanical setting. The related quantities such as current and displacement are the relevant state variables. A dynamical model of these continuous state variables are the partial differential equations of a system. In many cases the boundary conditions, which determine the interaction of the system with its surroundings, are used to select appropriate state variables. In this thesis we try to move away from this practice, to give a model of a component a status independent of its use, i.e., a particular interaction.

From the continuous, or distributed, model a, as large as possible, discrete, or FEM, model is usually the first stage in numerical analysis. However, in many cases the system is considered isolated for such models, and characteristics, such as vibrational modes and frequencies, are the outcome of such analysis. We want to postpone the specific choice of boundary conditions to a later stage, however, some choices have to be made. In this thesis several options are given. Model reduction techniques can focus on the recovery of lumped states of a system. For example, the lumped mass is a property of a system which can be used in many different settings, such as the total gravitational force it causes, the force it requires to set it in motion, or the force under impact. The next stage would be to determine the elasticity of the fixture, and express it in terms of a spring. The consistent decomposition of the system in terms of lumped properties rather than dynamic responses allows one to use the resulting model in several dynamical settings. Hence, keeping a modular design in mind focuses on the qualities of the system.

1.5. Boundary impedance

In the mathematical theory of partial differential equations, boundary conditions are set to generate a well-defined, closed problem. The system, described by the standard theory of partial differential equations, has a very limited, preset interaction with the environment, through the boundary conditions. Many choices of boundary conditions, such as Dirichlet and Neumann boundary conditions isolate the system completely from its environment, such that an energy function exists in conjunction with the partial differential equations, which is instrumental to determine existence and uniqueness, the well-posedness of the problem, of the solution to the partial differential equation.

Another type of boundary conditions is the transmission boundary condition. A constant flow goes in at one end and out at the other. Although the matter is in motion the velocity is stationary. This is one of the so-called quasi-stationary states, considered in this thesis. However, in these cases the well-posedness is conditional; it depends on the values at the boundary, not just the type of boundary condition. The system may be driven by the flow, yielding oscillatory effects or even turbulence.

In the most general case, under investigation here, we do not speak of boundary conditions, but of boundary impedance. The boundary impedance is the response of the surroundings on the behavior of the system at the common boundary. The term impedance comes from the theory of transmission lines, but is now widely used in scattering theory, electromagnetic and acoustic, to describe boundary conditions for boundaries with nontrivial response, such as absorbing boundary conditions.

The onset of turbulence for a fluid flow in a tube depends on the choice of boundary conditions. Experiments show, although the mean flow is the same, that a constant pressure at the boundary yields different results than a constant flow. The latter is difficult to achieve as most pumps to drive the flow have some response to the variations in pressure at for example a driving piston. The pump must be greatly overdimensioned to arrive at the constant-flow boundary condition. Even though the experiment should be isolated from the surroundings, the engineering solutions are not always able to do so; an ideal flow pump does not exists.

Boundary impedance models these non-ideal boundary conditions. For fluid the pressure and the flow are related, for electrodynamics, the electric and magnetic fields are related at the boundary, for elastic bodies the pressure and the displacement are related at the boundary. All of these boundary conditions can be formulated as the work performed by, or on, the surroundings. Boundary impedance may guarantee global energy conservation, or power continuity.

1.6. Transmission versus vibration

> If such energy transfer can be slowed down or spread over a wider area, then the intensity of the force applied to the tissues per unit time is less. This is part of the function of a seat belt in a vehicle crash, where the stretching of the belt fabric extends the time of energy exchange and the considerable area of the belt is a preferable alternative to transferring all the kinetic energy via a few square centimeters of forehead against the windscreen.
> [BERNARD KNIGHT, FORENSIC PATHOLOGY]

A continuous piece of material is an abstract component in system without form or function. Its purpose should dictate the analysis which leads to a particular form and model, useful in the greater scheme of the system. Static descriptions are easily made, but the dynamics of a component, i.e., the time-dependent properties within the system as a whole, is conceptually more challenging. Three common forms spring to mind. First, simple functions, described by ideal component such as masses, springs, and dampers, in the mechanical domain. Secondly, the signal lines with delays and distortions. Thirdly, the vibrations of an isolated component.

The function of a component is generally not to vibrate or oscillate. The component is added to a larger system, and its function is to transmit or transform, in the most general sense. Parts of an engine of a car are there to turn the expansion of the exploded fuel mixture into forward motion of the car. They are all chains in network, transmitting the motion or force. To be more precise the network transmits motion under force or force producing motion. Hence a component must have fixtures, or ports, to receive and send the mechanical energy along.

Many types of analysis of components are based on vibrational motion of the isolated component. In this manner little information is gained on, for example, the

rigidity of a component, useful for the transmission of a motion under force. Even more, most forms of analysis do not have a language for the general transmission of energy, such as the bond-graph language for lumped port-based modelling. The typical boundary conditions for fixtures are invented to isolate a component from its surroundings.

For continuous systems, described by partial differential equations, boundary conditions are a necessary part to arrive at a well-defined mathematical problem. The typical boundary conditions are those for isolated problems and will lead, quite naturally, to vibrational problems in the dynamical regime, or static deformations in the stationary regime. Interpreting the static vibrational results and combining them with the design criterions for operation is a complex task, which may not necessarily say much about the actual function, or transmission, the component is designed to perform. Typically one would set as criterion that the lowest vibrational frequency must be higher than the highest operational frequency of the system. Whether a vibrational mode is excited by the motion of the system is often not determined. Furthermore, the actual required motion or operation of the system has often very little to do with the analysis of the component on its own. The mathematical boundary conditions do not allow for such global motion. For example, the body forces due to rotation in a mechanical component are often not taken into account in the vibration analysis, except for known problems related to the rotational motion, such as in turbines. [Biezeno and Grammel, 1953]

In this thesis the question of transmission through boundary conditions and the corresponding boundary variables is addressed. For the same component a number of problems in terms of different boundary conditions exist, depending on the particular setting. Take for example a ball. You can sit on it, which yields a stationary situation with a position boundary condition, at the contact point with the floor, free boundary conditions on the side, and a force boundary condition on the top. The mass of the ball does not play a role. The same ball may be thrown. In that case the mass will determine the acceleration given the force boundary condition. The acceleration will lead to body forces in the ball. These two situations correspond to two different problems in the standard mathematical language. In this thesis we will bring these two problems together, since we are talking about one and the same ball. The throwing of a ball, which falls and bounces, should be seen as a single operation and not three.

However, unlike a single ball, with its simple purpose of bouncing, components in a machine or system are connected together. The transmission means components are part of a network. The connections can be distributed or lumped interconnections or connections to actuators or sensors. The larger network should be kept in mind, even when dealing with some peculiarities of a single part. The component function is its reason for existence, in most cases. The relation to this function is often formulated in terms of the vibrational properties, as described above, or in terms of a transfer function, between input and output. Transfer functions describe the transfer of abstract signals, rather than mechanical, electrical, or fluidic transmission; the function of the component. In this thesis the focus is very much on quantifying the physical transmission. In particular the transmission of energy or power is used, which is central to most system modelling and stretches beyond particular physical domains, such as the electrical, mechanical, or thermodynamical. Eventually, it should get a place within the existing framework

FIGURE 1.1. Most local variations in the boundary conditions are not "transmitted" through the component, and in a global, network view, despite the distributed nature, average or lumped variables can be selected.

of lumped energy-based bond-graph network modelling. [Paynter, 1961, Karnopp et al., 2000]

Rather than starting from the network, and filling in the components, as parts, we start from the continuous components, since the actual function of a component, unlike in the lumped approaches, is not a priori clear. Whether a piece of structure acts more like a spring or a mass, or a combination of the two like a resonator, depends not only on the structure, but also on the operational setting. We try to approximate the component by the best simple lumped model, in terms of a set of modes. These modes are not the result the vibrational analysis, but in terms of the transmission analysis.

System theory is the theory of cutting and pasting. It is the global view, without loosing contact with the relevant details important for the function of the system as a whole. The other details of each of de modules or components are of less interest. Of interest are the qualities of the component which affect the other components in the systems, and the nature and possibilities of this interaction. These qualities of a component are very much related to a whole collection of simple dynamical states one can bring this component in, due to simple, typically stationary input. Hence the modular quality of a component is given by collecting the core states, or modes, of all the possible interactions it can undergo. The limitations of the fixtures, and thereby the stationary input, limits the network or global view and the number of qualities of a component one cares to distinguish. A beam attached only at both ends would have a core, or network, model of four degrees of freedom. With another fixture in the middle, one or two degrees of freedom extra, depending on the type of fixture, are required to describe the effects of interaction at one fixture on another.

A global, or network view of a system does, however, limit the complexity of the signals from one component to another. Fortunately, nature seems to have the same limitations in most cases. For example, the precise details of the heat flux between one component and another are irrelevant as the local temperature and flow differences are smoothed out, and the total amount of heat flow mainly determines state, or temperature, of the component. See Figure 1.1. The same principle in elasticity is called Saint-Venant principle. [Timoshenko, 1953, Truesdell and Noll, 1965] It states that the precise details of the distributed force at the boundary are irrelevant away from the boundary, only the force and the torsion are transmitted. The force and the torque are associated with the boundary conditions arising from the elastic energy in the system. These variables are therefore natural candidates of network or graph variables in physical network models, the distributed analogues of the bond-graph variables.
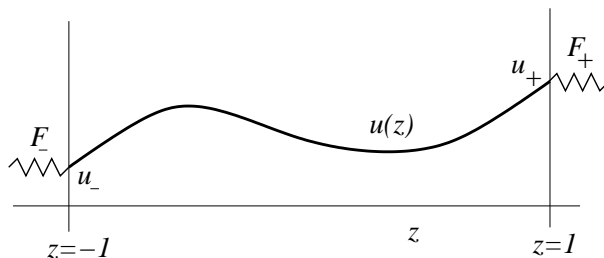
FIGURE 1.2. A string with its boundary conditions. Either the force or the position is given at the boundary, the other variable can be seen as string response.

1.7. Preliminary example

In order to set the stage for the approach advocated here, we examine a simple, well-known system of a string. [Courant and Hilbert, 1961, Fetter and Walecka, 1980, Sakamoto, 1982] The dynamics of the string is the wave equation:

$$(1.1) \qquad \rho\partial_t^2 u - \kappa\partial_z^2 u = 0 \ ,$$

where the string coordinate is $z \in [-1, 1]$, and the displacement of the string is $u(z, t)$. The relevant material properties are the mass density ρ and the string stiffness κ. The time-derivative is $\partial_t = \frac{\partial}{\partial t}$ and the spatial derivative is $\partial_z = \frac{\partial}{\partial z}$. The wave equation, Eq. 1.1, may equivalently be formulated as a set of first-order equations with two variables:

$$(1.2) \qquad \begin{pmatrix} \dot{u} \\ \dot{p} \end{pmatrix} = \begin{pmatrix} \frac{1}{\rho}\partial_z p \\ \kappa\partial_z u \end{pmatrix} \ ,$$

where the momentum $p = \rho\partial_z\dot{u}$ is an independent variable. An independent variable, or a state, requires its own initial condition; the value a certain time, to uniquely characterize or predict the systems behavior. Many other choices of the second variable are possible. We will show that they lead to different boundary conditions, and, thereby, to a variety of operational settings for the same string.

The energy in the string consists of kinetic and elastic energy:

$$(1.3) \qquad E = \frac{1}{2}\int_{-1}^{1} dz \left(\rho(\partial_t u)^2 + \kappa(\partial_z u)^2\right) \geq 0 \ .$$

The integrand alone is the energy density; the energy per length. The mass density ρ and the stiffness κ are the relevant material constant of the string. The wave velocity v is expressed in these material constants: $v = \sqrt{\kappa/\rho}$.

Boundary conditions are introduced to complete the problem. However, for an engineer the boundary conditions are the most essential part. Without boundary conditions the string is just an isolated object, possibly of interest of scientists, but not of engineers.

The boundary conditions come in a large variety. Each of the ends might fixed, or a force may be applied to it. Furthermore, position and force may vary in time. Eventually, the engineer cares little about the dynamics of the string. He cares about the displacement positions $u(\pm 1) = u_\pm$ and forces $F_\pm$ at both ends $z = \pm 1$. See Figure 1.2.

Preferably, an instantaneous approximation is sought, for example, the force F_- at the left end and the position u_+ at the right end as functions of the other force and position:

(1.4) $$u_-(F_-(t), u_+(t)) \quad \text{and} \quad F_+(F_-(t), u_+(t)) \; .$$

Instead of an instantaneous solution, the appropriate result would be the stationary solutions for static input. Only in that case the internal dynamics do not have to play a role. This is an important result, which might appear trivial, but is not always easy to accept.

In general the result will be an integral over the history of the input $F_-(t)$ and $u_+(t)$:

(1.5) $$u_-(t) = \int_{-\infty}^{t} U_-(F_-(t'), u_+(t'))dt' \; ,$$

and likewise for the force F_+.

Furthermore, we can speak easily of a force F, but it does not appear in the equation of motion, the wave equation. Behind the existence of a force as a boundary condition lies the theory of mechanics and energy. As it turns out, the force is for the wave equation related to the derivative of the displacement at the boundary.

Force is not an arbitrary input to the system. Systems described by partial differential equations may behave oddly. However, the energy in the system determines the bounds on this behavior. In the case that the energy is zero, the system is at rest. The product of force and velocity is the power, or rate of change of the energy, supplied to the system. In general, if no force is applied to the system the energy remains the same. The state moves along lines with equal energy. A particular elegant way to describe the motion or dynamics is to rewrite the energy in such a way that it forms the basis of equations of motion, the Hamilton equations. The energy in this form is called the Hamiltonian. In the case of a Hamilton equations of motion, in terms of the energy functional, or Hamiltonian H, the applied force is additive to the change of momentum p:

(1.6)
$$\begin{pmatrix} \dot{q} \\ \dot{p} \end{pmatrix} = \begin{pmatrix} \delta_p H \\ -\delta_q H \end{pmatrix} + \begin{pmatrix} 0 \\ F_{\text{applied}} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} \delta_q H \\ \delta_p H \end{pmatrix} + \begin{pmatrix} 0 \\ F_{\text{applied}} \end{pmatrix} \; ,$$

which is the wave equation, Eq. 1.2, augmented with an input. The notation $\delta_p H$ and $\delta_q H$ represent the variational derivative of the Hamiltonian with respect to momentum p and position q.

The change of energy, due the additive force, follows from the time-derivative of the Hamiltonian:

(1.7) $$\dot{H} = \delta_q H \dot{q} + \delta_p H \dot{p} = \delta_q H \delta_p H + \delta_p H(-\delta_q H + F_{\text{applied}}) = \dot{q} F_{\text{applied}} \; .$$

The product of the generalized velocity $\dot{q} = \delta_p H$ with the applied force yields the power, or change of energy, of the system.

The set of first order equations, with two initial conditions $u(t = 0, z)$ and $p(t = 0, z)$, Eq. 1.2, can be derived from the wave equation by substituting $p = \partial_z \dot{u}$. However, in the Hamiltonian formulation, Eq. 1.6, the variables u and p do not naturally arise. In the port-Hamiltonian formulation [Van der Schaft and Maschke, 2002] the elastic-energy variable $\partial_z u$ is used instead of the configuration variable u. Likewise the momentum variable p changes. The choice of variables vary with

the approach and principles, but the pair of generalized coordinate and generalized momentum belong together. The explicit form of the port-Hamiltonian H' of the wave equation is:

$$(1.8) \qquad\qquad H' = \frac{1}{2\rho}p^2 + \frac{\kappa}{2}q^2 \quad,$$

where $q = \partial_z u$ and $p = \rho\dot{u}$. The equations of motion are therefore also not the standard equations, Eq. 1.6. The symplectic matrix is replaced by the differential operator:

$$(1.9) \qquad\qquad \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \rightarrow \begin{pmatrix} 0 & \partial_z \\ \partial_z & 0 \end{pmatrix} \quad,$$

yielding the same equations of motion, Eq. 1.2, for a different pair of variables. Furthermore, spatially constant configuration solutions $u_0(t)$ will not contribute to the port-Hamiltonian, expect for the constant of motion, p_0, the momentum of the string as a whole. In the port-Hamiltonian formulation these states can be excluded, by the appropriate choice of the function space for the momentum $p(z)$.

Often, we make no distinction between the Hamiltonian as the energy density of a system, and the Hamiltonian as the total energy, or the integral over the energy density. Their dimensions are different, but the difference is easily recognized and understood.

The inclusion of a differential operator ∂_z in the relation between velocity $\dot{u}(z)$ and momentum $p(z)$ is not the canonical choice, but related to the wave equation, Eq. 1.2. This, in turn, is related to the port-Hamiltonian form, discussed in Chapter 7.

The coordinates $q(z)$ and momenta $p(z)$ are a collection of infinitely many variables. For each string position z another value for $q(z)$ and $p(z)$ may occur. Such systems are also called continuous or distributed. From the infinitely many variables any subset may be selected to approximate the full system. However, care must be taken. The canonical pair at a point z, i.e., $q(z)$ and $p(z)$, seem to suggest that the same modes for $q(z)$ and $p(z)$ will lead to an appropriate discrete model. This is not the case. For example, for Eq. 1.2, if $q(z)$ is expanded in terms of $\cos k_i z$, the corresponding expansion for the momentum $p(z)$ is in terms of $\sin k_i z$.

The discretization of the local product $p(z)p(z)$ may no longer be local, depending on the approximation of the displacement. The momentum $p(z)$ is related to the velocity $\dot{u}$ in Eq. 1.1. This wave equation followed from the variational principles associated with the Lagrangian, similar to the Hamiltonian for the two first-order equations. [Goldstein, 1980] Hence in order to recover the proper momentum discretization we are forced to return to configuration-velocity space $(u, \dot{u})$ since the mode shapes of $\dot{u}(z)$ must be a subset of the mode shapes of $u(z)$.

In order to make a coordinate transformation and reduction, to go from functions $u(z)$ and $\dot{u}(z)$ to a set of modes, where the dynamics is given by the coefficients of the modes, the Lagrangian will serve for consistency. The Lagrangian is the invariant density which allows for contact coordinate transformations of u and $\dot{u}$ simultaneously from a function of z to mode shapes $u = q_i\phi_i(z)$, where q_i is the coefficient:

$$(1.10) \qquad\qquad L = \frac{\rho}{2}\dot{u}^2 - \frac{\kappa}{2}(\partial_z u)^2 \quad.$$

The displacement u plays two roles: it is the dynamical variable associated with the energy in the system, and it is the configuration and determines the possible, or allowed, velocities. The evolution of the system should be consistent which the configuration the system can attain.

Taking for example a set of modes $\phi_i(z) = z^i$:

$$(1.11) \qquad u(z) = q_0 + q_1 z + q_2 z^2 + q_3 z^3 \quad,$$

the velocities are directly given in terms of the time derivatives of the coefficients q_i:

$$(1.12) \qquad \dot{u}(z) = \dot{q}_0 + \dot{q}_1 z + \dot{q}_2 z^2 + \dot{q}_3 z^3 \quad,$$

The same modal subspace is spanned for u and $\dot{u}$. The canonical momenta $\mathbf{p} = (p_0, p_1, p_2, p_3)$ corresponding to the modes $\mathbf{q} = (q_0, q_1, q_2, q_3)$ are the variational derivatives:

$$(1.13) \qquad p_i = \delta_{\dot{q}_i} \int_{-1}^{1} dz\, L = [\mathbf{M}\dot{\mathbf{q}}]_i \quad,$$

where the mass matrix $\mathbf{M}$ is given by the integrals over products of modes as they occur in the Lagrangian:

$$(1.14) \qquad q_i M_{ij} \dot{q}_j = \dot{q}_i \left(\int_{-1}^{1} dz\, \rho \phi_i(z) \phi_j(z) \right) \dot{q}_j \quad.$$

The mass matrix has only constant values for a kinetic density $\frac{\rho}{2}\dot{u}^2$:

$$(1.15) \qquad \mathbf{M} = \rho \begin{pmatrix} 2 & 0 & 1 & 0 \\ 0 & 1 & 0 & \frac{2}{5} \\ 1 & 0 & \frac{2}{5} & 0 \\ 0 & \frac{2}{5} & 0 & \frac{2}{7} \end{pmatrix} \quad.$$

Hence, only for a particular choice of subspace, or modes, the Hamiltonian is properly defined. The momentum vector $\mathbf{p}$, paired with the position vector $\mathbf{q}$, is simply $\mathbf{p} = \mathbf{M}\dot{\mathbf{q}}$, with the given mass matrix. The local definition of momenta, $p(z)$ formally assigned, would arise from Fourier-Laplace theory [Hörmander, 1990] with quasi-periodic modes.

The Hamiltonian is for a quadratic Lagrangian itself quadratic:

$$(1.16) \qquad H = \frac{1}{2}\mathbf{p}^T \mathbf{M}^{-1}\mathbf{p} + \frac{1}{2}\mathbf{q}^T \mathbf{K}\mathbf{q} \quad,$$

where the stiffness matrix is recovered along the same lines as the mass matrix by integrating the modes $\phi_i = z^i$:

$$(1.17) \qquad \mathbf{K} = \kappa \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 3 \\ 0 & 0 & 4 & 0 \\ 0 & 3 & 0 & \frac{18}{5} \end{pmatrix} \quad.$$

The stiffness matrix has a null-space, corresponding to the constant displacement. If, instead of the polynomial expansion, we would have used a trigonometric expansion, the displacement would be associated with $\cos k_i z$ while the momenta would be associated with $k_i^2 \cos k_i z$, which vanishes for $k_i = 0$.

The modes are no longer local. For example, all of the modes $\mathbf{q}$ together determine the position of the ends $z = \pm 1$:

$$(1.18) \qquad u(\pm 1) = (1, \pm 1, 1, \pm 1) \cdot \mathbf{q} = \chi_{\pm}^T \mathbf{q} \quad.$$

For a given end position, for example $u(-1) = 0$, the Hamiltonian needs to be projected perpendicular to the subspace spanned by χ_-. The positions should satisfy $\chi_-^T \mathbf{q} = 0$, the momenta $\chi_-^T \mathbf{M}^{-1} \mathbf{p} = 0$. The second condition determines the rate of change of the constraint $\chi_-^T \dot{\mathbf{q}} = 0$. If the second condition is given, the constraint $\chi_-^T \mathbf{q} = 0$ only needs to be specified for the initial time $t = 0$.

It is easier to apply an external force to the system. The force applied to one end, however, is distributed over all the modes.

We see that the collective modes give rise to some subtle results and effects in the finite approximations of the infinite theory. Another approach would be to retain as much of the local theory as possible. It is not the momentum $p(z)$ which is nonlocal, but the potential energy $\partial_z q(z)$ which couples the spatially separate dynamics $q(z)$ and $q(z')$ for $z \neq z'$. By introducing the moment variable $q'(z) = \partial_z q(z)$ the Hamiltonian $H(q, p)$ transforms to a port-Hamiltonian $H'(q', p)$, and the equations of motion, rather than the port-Hamiltonian itself, is the source of the nonlocal coupling of dynamics at distinct points z and z'.

1.8. Outline

This thesis addresses the question how continuous components can be analyzed and described such that the resulting model can be used in a larger system for system analysis and simulation. The question focusses very much on the treatment of a single component and its means to interact with its surroundings. The interaction takes the form of the power variables of bond-graph modelling representing the action-reaction pair which product is the energy flow in and out of the system.

The global question induces several more detailed questions. They fall in a number of groups:

- Models of a system: variables, properties, transformation of variables, reduction of variables, i.e., model order.
- Open, or modular, models: a model of a component, which can be inserted in all possible orientations and situations in a larger system.
- Qualities of infinite-dimensional, or continuous, models: energy, continuity relations, conservation laws, dimensionless constants.
- Dynamics of the model: time-behavior, input-output, stability, conserved flow.
- Mathematical tools to describe continuous systems: partial differential operators, differential geometry, variational principles, norms.
- Consistency of the model: constructing well-defined open mathematical problems, retaining qualities of the system in the model.
- Numerical evaluation: approximation of continuous operators, time-integration as simulation, numerical stability, constraints implementation.
- Mismatch between variables for energy and for the state, or configuration: null-spaces of operators, integrability, natural input variables.

These questions are addressed throughout the thesis. A number of preliminary chapters treat some of the known material in this light. The later chapters are the result of separate investigations.

For each chapter the purpose is described shortly.

- Chapter 1 **Introduction**
  Different concepts, such as complexity, model, port-based modelling, boundary impedance, and transmission, related to the work are discussed in a historical light. The chapter finishes with an illustrative example, this outline, and a dutch summary.

- Chapter 2 **Mathematical models of dynamical systems**
  The mathematical description of mainly isolated systems with increasing complexity, in terms of the number of degrees of freedom. Some generic methods associated with these models. The chapter begins with classical mechanics, and ends with general geometric, variational, and algebraic tools used in the description and analysis of dynamical models.

- Chapter 3 **Evolution**
  Finite-dimensional linear models for simulation, simulation through time integration, integration methods. The state-space, or **A,B,C,D**-model to be used throughout the thesis for finite-dimensional linear models, is defined, some of its properties are discussed, and the time-integration methods, for simulation, are analyzed and improved.

- Chapter 4 **Differential geometry**
  The spatial domain on which the states of the system are defined plays an important role in the choice and transformation of the state variables. Elements of differential geometry relevant for this choice and the possible transformations are discussed in this chapter: invariant, or coordinate-free, methods, integrability conditions, and conserved flow.

- Chapter 5 **Partial differential equations**
  A system described by a partial differential equations as input-output model, how to define energy, formulating it as an initial boundary value problem, where the initial conditions are the states, the boundary conditions the input. Results from the literature.

- Chapter 6 **Models of infinite-dimensional systems**
  Physical systems described in continuous variables, their formulation and intrinsic qualities. Elasticity, fluid dynamics, electromagnetism are discussed. Some models based on these theories illustrate modelling approaches. Dimensionless constants are discussed as an important tool for analyzing such systems.

- Chapter 7 **Port-based modelling of infinite systems**
  What are the equivalents of the lumped qualities, typical for bond-graph modelling, in continuous systems. Can they be isolated and assigned to boundary values? The port-Hamiltonian is constructed from variational principles with differential operators. The distributed equivalents of lumped properties of components are constructed as quasi-stationary states of the core model. A generalized view, which brings together a variety of physical boundary conditions, is introduced.

- Chapter 8 **Linear finite element methods**
  Short introduction into FEM, stencils, degrees of freedom, and their qualities and reduction. Special attention is paid to the initial-boundary value problem, dynamics, and the retention of conservation laws during FEM approximation of the continuous model.

- Chapter 9 **Analytic input**
  Model order reduction of large state-space models based on the time-dependence of the input, which is typical slow, and of low dimension, compared to the full frequency spectrum of the isolated evolution. The reduction falls into two main parts: First, the construction of modes, based on the input. Krylov method is used for this part. Secondly, the construction of a dynamical model based on these modes through projection. Different types of projections, time-independent and time-dependent, are discussed.

- Chapter 10 **Real and virtual dynamics**
  The connection between iterative numerical methods for systems of high dimension and the simulation of such systems. The iterative methods are expressed as dissipative systems, the time-step and the damping are the relevant parameters. The extension to nonlinear problems. Several lumped nonlinear models are used to test and illustrate the implementation of constraints.

- Chapter 11 **Nonlinear FEM**
  The invariant description of finite elements in terms of a consistent set of variables, for energy-conserving dynamics, decomposed in boundary port variables and energy-based state variables. The geometric invariant, constraint-free, finite-order approximation of an elastic curve is developed. The counting of degrees of freedom, for potential energy given by differential operators such as the curl operator, on square and triangular elements are discussed. The consistency conditions, arising from both the boundary energy and the differential operators are discussed. The examples of a triangular plate element and electrodynamics on the simplex are worked out in some detail.

- Chapter 12 **Force lines and velocity fields**
  A geometric view of the force balance in an extended object. Some aspects of a global view of continuous components. The balance laws implemented globally, to be adapted locally through material properties and dynamical principles. Given a harmonic approximation of the force balance between input and output, the material and dynamical aspects can be addressed locally, with simpler methods.

- Chapter 13 **Conclusions**
  A short summary and some conclusions concerning the conceptual development proposed in this thesis.

- Chapter 14 **Literature**
  Background reading.

Chapter 7 forms the conceptual core of this thesis. It is the juxtaposition between the earlier chapters which sketch the existing background within the framework of the problems pointed out in Chapter 7. The later chapters are more detailed answers to the questions posed. Each of these chapters contains stand-alone results, covering the different aspects, such as input-based dynamics, finite-element modelling, and global modelling.

1.9. Nederlandse samenvatting

Dit proefschrift "Functionele Systeem Dynamica" behandelt de vraag hoe continue componenten van een systeem goed beschreven kunnen worden met wiskundig modellen voor analyse en simulatie van het systeem als geheel. Het systeem als geheel wordt nauwelijks in beschouwing genomen, maar vooral de wisselwerking van een enkele component met zijn omgeving wordt onderzocht. Het paar variabelen, de kracht en de beweging, en het equivalent daarvan in andere fysische domeinen, zijn de poort-variabelen van deze wisselwerking. In de simulatie één van de variabelen is input, of actie, terwijl de andere de output, of de reactie, representeert. Het product is de overgedragen energie, ofwel het vermogen. Voor simpele modellen, met een beperkt aantal variabelen, dit modelleren is bekend als bondgraafmodelleren.

Voor continue componenten, die typisch beschreven worden door partiële differentiaalvergelijkingen, zijn een aantal aspecten van een dergelijke methode van modelleren onderzocht in dit proefschrift:

- De mogelijke variaties in het model, keuzes van variabelen, transformaties van het model, de relatie tussen de referentie-ruimte en de toestands-ruimte, het reduceren van het aantal variabelen.
- Een open model, de mogelijkheden van wisselwerking met de omgeving, de verschillende manieren en oriëntaties dat een model in een groter systeem ingepast kan worden.
- De eigenschappen en fysische wetten impliciet in de continue modellen, energie en andere behouden grootheden, dimensieloze constanten.
- De tijdsafhankelijkheid of dynamica van het model, tijdschalen, input-output, stabiliteit, en de stroom van een behouden grootheid.
- De wiskundige technieken geassocieerd met continue systemen, differentiaal operatoren, differentiaal geometrie, variationele principes en de bijbehorende normen en functionalen.
- Het modelleren zelf, het behoud van de fysische principes in simpele modellen.
- De numerieke evaluatie, benaderingen, tijd-integratie als simulatie, numerieke stabiliteit, en de implementatie van voorwaarden.
- Het verschil in configuratie toestanden en de variabelen relevant voor de energie in het systeem, de nulruimte van differentiaal operatoren, en de natuurlijke keuze van input-variabelen.

De resultaten leiden tot de volgende conclusies:

- De differentiaal operator in de moment- of energie-variabele maakt van de Hamiltoniaan van het gesloten systeem de poort-Hamiltoniaan van het open systeem.
- Het gebruik van basisfuncties, of modes, met exacte differentiatie en integratie in de eindige modelruimte, behouden de fysische principes van de continue modellen.
- De fysische randwaarden, waarmee de energie-flux door de rand gedefinieerd wordt, worden uitgedrukt in termen van de potentiaal, of "effort", en niet de toestanden.
- In een goede systeembeschrijving van een continue component wordt er een scheiding gemaakt tussen het kernmodel in termen van de mogelijke

randwaarden, en het model voor de interne dynamica, waarvoor begin-
waarden nodig zijn.

- De randwaarden en randcondities zijn niet uniek maar afhankelijk van het
  type wisselwerking met de omgeving.
- Voor het construeren van een dynamisch eindig element moeten de rand-
  waarden en beginwaarden passen bij de bewegingsvergelijkingen. De nul-
  ruimte van de bewegingsvergelijkingen speelt daarbij een belangrijke rol.
  Het verband tussen de energie-variabelen en de configuratie-variabelen is
  belangrijk voor consistentie.
- De quasi-stationaire toestand voor de stationaire input is de belangrijkste
  toestand voor het beschrijven van dynamica. Deze toestand is de eerste in
  een rij van toestanden, of modes, die de consequentie zijn van polynoom-
  achtige tijdsafhankelijke input.
- Het aantal toestanden, of modes, van een FEM model overtreft het aan-
  tal relevant of bruikbaar voor simulatie en regeling. De selectie van de
  nuttige modes kan gedaan worden op grond van het inputsignaal en de
  systeemrespons.
- De globale visie of systeemvisie staat op gespannen voet met de differ-
  entiaalmethoden voor continue systemen. De gesloten stroomlijnen van
  behouden groetheden en balanswetten vormen een mogelijke brug.

In het algemeen kan gesteld worden de modellen in termen van partiële differen-
tiaalvergelijkingen voor continue systemen leiden tot een naar binnen gerichte blik,
waarin de grote variëteit aan intrinsieke eigenschappen van de continue systemen
centraal staat. De randwaarden zijn daarin bijzaak. De systeemvisie is een naar
buiten gerichte blik, waarbij de randwaarden en de respons juist de hoofdzaak zijn.
De beperkte systeemeigenschappen van de component volgen uit de wisselwerking
met de omgeving.

CHAPTER 2

Mathematical models of dynamical systems

> We perceive motion and time together. It sometimes happens
> that we perceive the flow of time though we may not sense a
> particular sensible motion.
> [THOMAS AQUINAS, ON ARISTOTLE'S PHYSICS]

Science is about quantifying and predicting. In order to test our understanding
of a system, we need to be able to predict its behavior, before it happens. Only if
we can predict a systems behavior in a wide variety of circumstances, we can say
we truly understand this system. In order to be able to predict appropriately, we
need to be able to describe; to separate different types of behaviors and states of a
system. Part of predicting is therefore also specifying the variables that describe the
system and its time-dependent behavior fully and uniquely. Some variables do not
vary, these are either irrelevant, or parameters or intrinsic material constants, or
constants of motion, in our language. The variables which evolve with time are the
initial conditions, or state variables, or kinematical variables, in, respective, mathe-
matical, system-theoretic, and physical language. Typically, the state variables are
the minimal, or independent, set of variables needed for the unique characterization
of the behavior of the system. Behavior is the generic name of state as function of
time. The greater part of this thesis is about separating and combining the input
from outside and the intrinsic behavior of an isolated system. In this chapter we
deal mainly with isolated systems, commonly referred to as dynamical systems.

Dynamical systems are the dressed down versions of the general models, where
only the essential variables for the description and evolution of a system are re-
tained. Dynamical systems is the study of the time-dependence of systems. In or-
der to have a well-defined evolution of a system, there is a minimal set of variables
required, these are the kinematical variables. The restricted study of kinematical
systems is, in this thesis, the unique characterization of a system at a certain time.

2.1. Introduction

In this chapter we treat the mathematical models from the perspective of
Hamiltonian dynamics and energy conservation. These models are classified by
the number of degrees of freedom. With the growing number new features arise in
dynamical systems. For a more complete introduction into Hamiltonian dynamics
the reader is referred to the classical literature, [Whittaker, 1917, Carathéodory,
1935] physics literature, [Landau and Lifschitz, 1976, Goldstein, 1980, Fetter and
Walecka, 1980], or the mathematical literature. [Gantmacher, 1970, Hirsch and
Smale, 1974, Arnold, 1989]

Energy conservation is a central principle in physical dynamical systems. Con-
servation means the energy remains equal with time. The notion of energy has

been around since classical times, Huygens was the first person to equate two forms
of energy: kinetic and gravitational, and thereby quantifying the concept. [Huy-
gens, 1673, Theobald, 1966] But its central role, given to it by Hamilton, since the
eighteenth century, reversed its function. Rather than energy, E, being a quality
of a system, it became a dynamical principle. If energy is conserved, the state of
an isolated system must retain the same energy, which restricts the dynamics, i.e.,
the change of the state in the time. The resulting equations of motion are known
as the Hamilton, sometimes Hamilton-Jacobi, equations of motion. The energy E
in an appropriate choice of dynamical, or state variables $\mathbf{x}$ of the system is the
Hamiltonian: $E = H(\mathbf{x})$.

The energy of a system has later been related to the translational invariance
in the time. For physical systems the results of experiments will typically not
depend on whether an experiment is done today or yesterday. This invariance is a
symmetry of the system and the resulting constant of motion is the energy. Noether
[Noether, 1918] and Weyl [Weyl, 1952] have been instrumental in the introduction
of this connection between time and energy. In that respect the Hamiltonian H was
identified as a single component of a space-time tensor, [Frankel, 2004] 4×4 matrix
that transforms with the transformation of space-time coordinates, $T^{00} = H$. The
0-component indicates the time direction, while the other $1, 2, 3$-components are the
spatial directions x, y, z. The full conserved energy-momentum tensor $\{T^{\mu\nu}\}^3_{\mu,\nu=0}$
expresses all of the space-time symmetries; translational, rotational, and boost.
[Felsager, 1981] These aspects are further discussed in Chapters 5 and 7.

2.2. One-dimensional dynamical system

The most generic dynamical system is the one-dimensional harmonic oscilla-
tor. In many cases, dynamical systems are reduced or approximated by harmonic
oscillators, or independent copies of multiple harmonic oscillators, in a linear ap-
proximation. In the section we will discuss the harmonic oscillator in some detail
and introduce most of the concepts we will use later for finite-dimensional and
infinite-dimensional systems. The harmonic oscillator can be described by a second
order differential equation:

$$(2.1) \qquad m\ddot{x}(t) + kx(t) = 0 \ ,$$

where the position x is a function of the time t. The time derivatives are denoted
by the dots $\dot{x} = \partial_t x = \frac{\partial x}{\partial t}, \ddot{x} = \partial_t \partial_t x = \partial_t^2 x$, etc. Two parameters play a role:
the mass m and the stiffness k. There exists a continuous set of solutions to this
problem characterized by two parameters: the amplitude x_0 and the angle ϕ:

$$(2.2) \qquad x(t) = x_0 \sin\left(\sqrt{\frac{k}{m}}t + \phi\right) = x_0 \cos\phi \sin\sqrt{\frac{k}{m}}t + x_0 \sin\phi \cos\sqrt{\frac{k}{m}}t \ .$$

The constant $\sqrt{\frac{k}{m}}$ is the radial frequency usually denoted by ω. In order to define
uniquely the evolution of the system $x(t)$ for $t > 0$, we require the knowledge of two
variables at $t = 0$, the position $x(0)$ and the velocity $\dot{x}(0)$, these are the kinemat-
ical variables. The system, at all times, is uniquely described by the kinematical
variables at a given time. The point $x(0) = 0$ and $\dot{x}(0) = 0$, is however unique.
This point is a critical point, in the sense that the state remains the constant over
time. For no other point $(x, \dot{x})$ this is the case.
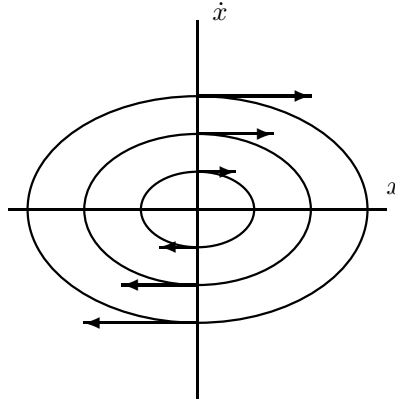
FIGURE 2.1. Some trajectories of the harmonic oscillator in configuration-velocity space, the vectors indicate the velocities. The closed orbits in the form of ellipses are the contours of equal energy.

The critical point is also singular point in the polar coordinates (x_0, ϕ). The angle variable is irrelevant at $x_0 = 0$, as the system will remain stuck at $x(t) = 0$, hence it is characterized by its zero amplitude $x_0 = 0$. If we draw the trajectories of the kinematical variables as a function of time we find they are ellipses around the critical point. See Figure 2.1. The closed orbits of the harmonic oscillator are very interesting. It tells us that the time evolution is described by the angle $\phi(t)$ only; the amplitude x_0 remains constant. The amplitude is associated with the energy E in the system. Since two variables are needed to describe a system, we might as well pick two variables which brings the system to the general canonical form for arbitrary vectors $\mathbf{x}$, which coefficients are the kinematical or state variables:

$$(2.3) \qquad \dot{\mathbf{x}} = A\mathbf{x} \ ,$$

which means using the momentum $p = m\dot{x}$ instead of the velocity $\dot{x}$, since the leading order term in the differential equation $m\ddot{x}$ must be on the left hand side of the canonical evolution equation. The matrix $\mathbf{A}$ is called the evolution operator.

The ellipses in the configuration-velocity space $(x, \dot{x})$ can be mapped to ellipses in the phase space $(x, p) = (x, m\dot{x})$. The energy, yielding the families of trajectories, is given by:

$$(2.4) \qquad E = \frac{1}{2m}p^2 + \frac{k}{2}x^2 \ .$$

The motion is perpendicular to the lines with constant E. Hence it can be treated geometrically rather than as a differential equation of time. The two variables, the position $x_1 = x$ and the momentum $x_2 = p$ form a vector $\mathbf{x}$:

$$(2.5) \qquad \mathbf{x} = \left(\begin{array}{c} x_1 \\ x_2 \end{array} \right) = \left(\begin{array}{c} x \\ p \end{array} \right) \ .$$

The direction in which the system moves in time: $\dot{\mathbf{x}}$ is a vector tangent to the line $E(\mathbf{x}) = $ constant, which is most easily constructed as the vector perpendicular the
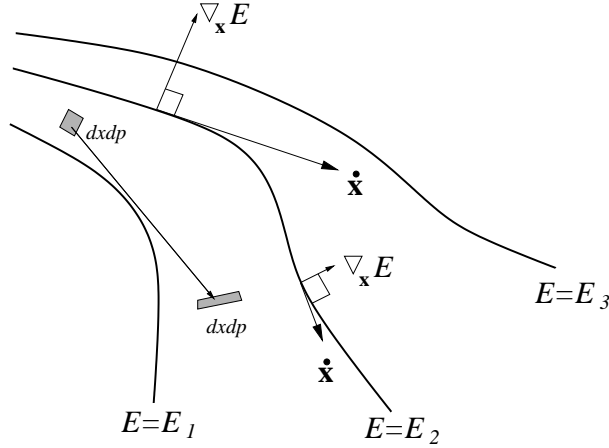
FIGURE 2.2. The gradient, in $\mathbf{x} = (x, p)$, of the energy is a vector perpendicular to the lines of constant energy. The time evolution is perpendicular to the gradient, and therefore along the lines of constant energy. If the lines are closer together, the gradient $\nabla_{\mathbf{x}} E$ is larger, and so is the evolution $\dot{\mathbf{x}}$. The area of the phase space $dxdp$ is preserved.

gradient of the energy:

$$(2.6) \qquad \dot{\mathbf{x}} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \nabla_{\mathbf{x}} E(\mathbf{x}) \ .$$

The choice of factors m^{-1} and k and the variables q and p for the energy E is now clear, as it yields the correct time dependence. Even if the trajectories are not simple ellipses but some complicated closed orbit in phase space, the time evolution of the system $\dot{\mathbf{x}}$ would still be given by the gradient of the energy, for the simple reason that if the energy is conserved; it does not vary in time:

$$(2.7) \qquad 0 = \dot{E} = \nabla_{\mathbf{x}} E(\mathbf{x})^T \dot{\mathbf{x}} \ ,$$

the vectors $\nabla_{\mathbf{x}} E(\mathbf{x})$ and $\dot{\mathbf{x}}$ should be perpendicular. It is an application of the chain rule of differentiation, [Rudin, 1966] and the equations of motion. For example, if the symplectic matrix $\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$, guaranteeing the orthogonality, would not be constant, but a function of $\mathbf{x}$, it would have led to a diagonal terms under differentiation with as consequence $\dot{E} \neq 0$.

An interesting consequence of the symplectic rotation, i.e., the matrix $\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ in front of the gradient, and the gradient itself is the conservation of an area of the phase space. See Figure 2.2. This is known as Liouville theorem. [Arnold, 1989, Thirring, 1997] The configuration-velocity space does not have this property. If we consider an area $\mathbf{x} \in \mathcal{S}$ spanned in phase space, the area remains constant under time evolution. Heisenberg's uncertainty principle $\Delta p \Delta x \geq \hbar$ can therefore be a fundamental consequence of quantum mechanics independent of coordinate and time. In simple terms, no quantum system can be squeezed into an area of the phase space smaller than $\hbar$, Planck's constant. Furthermore, the phase space of a single degree of freedom can be mapped to the problem of two-dimensional

incompressible flows, which is equivalent to conformal mappings of the complex plane. [Spiegel, 1964] See also Section 4.5. indexconformal

If the phase space consists of simple orbits around the critical point. The position and momentum variables can be mapped to action and angle variables. [Fetter and Walecka, 1980] The action indicates the orbit, or the amplitude, and the angle the position along the orbit. The time evolution of the angle variable ϕ is therefore given by a function of the amplitude. However, it would be more convenient to use a function of the energy as the second variable in the kinematical pair. Such a set of kinematical variables are called action-angle variables. The action $W(E)$ is given as the first kinematical variable, hence, in order to have a dynamical system with a conserved energy in the canonical form:

$$(2.8) \qquad \left(\begin{array}{c} \dot{W} \\ \dot{\phi} \end{array} \right) = \left(\begin{array}{cc} 0 & 1 \\ -1 & 0 \end{array} \right) \left(\begin{array}{c} \partial_W E(W) \\ \partial_\phi E(W) \end{array} \right) = \left(\begin{array}{c} 0 \\ \partial_W E(W) \end{array} \right) \quad ,$$

which would yield a linear change of the angle $\phi(t) = \partial_W E(W)t + \phi(0)$. In the case of a harmonic oscillator the action variable W is the energy: $W = E$. Consequently, for the harmonic oscillator, the frequency, the time it takes to traverse each orbit, is constant and independent of the energy or amplitude. In more general cases of dynamical systems this is no longer true: the frequency will depend on the amplitude. It will not always be simple to construct the action variable. The reason why the harmonic oscillator still is the cornerstone of dynamical systems is that around most critical points, for small amplitudes, the system can be approximated by a harmonic oscillator.

The mathematical pendulum is a point mass on a massless stick which can rotate in the plane under the influence of gravity. The differential equation governing its motion is:

$$(2.9) \qquad ml\ddot{x} + mlg\cos x = 0 \quad ,$$

where x is the angle, g the gravitational constant, and l the length of the stick. It can be derived easily from the energy considerations. The kinetic energy T is the same as before:

$$(2.10) \qquad T = \frac{p^2}{2m} = \frac{ml^2\dot{x}^2}{2} \quad .$$

The potential energy V is the gravitational energy with a constant force downward or vertical, such that the potential depends on the vertical position of the mass:

$$(2.11) \qquad V = mgl\cos x \quad .$$

Inserting the total energy $E = T + V$ into the canonical dynamical equation and using $p = ml\dot{x}$, the differential equation for the mathematical pendulum arises, again described by two kinematical variables, in this case x and p. A closer look at the phase space we see it is more complicated than the elliptic orbits of the harmonic oscillator. See Figure 2.3. The numerical evaluation of the mathematical pendulum, as a constrained system in two dimensions, is discussed in Section 10.10.

This is the first of many nonlinear systems we will encounter. In this case the equations of motions are nonlinear, because the potential V is not a quadratic function of the energy. Another nonlinearity arises if the mass $m(x)$ is a function of the position. In most cases we will consider the energy to be only quadratic in the momentum, yielding always a linear dependence on the momentum in the
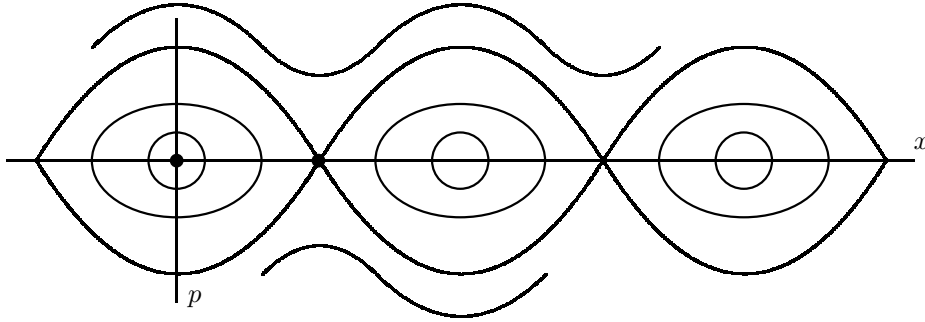
FIGURE 2.3. Some trajectories of the mathematical pendulum in phase space. The crossing lines are homoclinic orbits, reaching the saddle points only after infinite time. The angle variable leads to a 2π-periodic phase space.

equations of motion. While investigating dynamical systems in higher dimensions, more complex nonlinearities and higher order time-derivatives will arise. However, often these systems can be mapped to products of one-dimensional systems (x_i, p_i), at least in a small part of the multi-dimensional phase space. There are also many different ways to characterize the nonlinearity of a system. The phase space picture is generally a good one, for second-order systems, or independent copies thereof. However, care must be taken. Transformations of the system, e.g., $x \to x'(x)$ might change V and m beyond recognition, but leave the dynamics invariant.

For the global motion and distinguishing different types of behavior the critical points are essential. For the mathematical pendulum, the critical points lie on the $p = 0$ axis and are alternating stable critical points, with harmonic motion in its vicinity, and unstable critical points of the saddle type, stable in the p-direction and unstable in the x-direction. Such points can never be reached in finite time, as the system will slow down as it comes closer to the critical point. Such orbits are called homoclinic orbits and separate the phase space in regions with typically different global behavior, like periodic and non-periodic. See Figure 2.3.

So far, we have ignored damping or external forces, we have looked at the system as an isolated system. The energy was conserved for this reason. It allowed us to bring the equations of motion in the convenient Hamilton form. Energy is conserved in general, however, it might take other forms than we care to consider. For example, friction or damping will turn kinetic energy into heat or vibrations. Hence we will put the friction force, and external forces, on the right-hand side of the dynamical equation as input or output of the dynamical system $\mathbf{x} = (x, p)$ under consideration: $(p = m\dot{x})$

$$(2.12) \qquad \dot{p} + kx = m\ddot{x} + kx = -\frac{\gamma}{m}p + F = -\gamma\dot{x} + F \quad .$$

where F is the applied force and $-\gamma\dot{x}$ is the induced friction. In terms of the Hamilton equations of motion, this is an augmentation of the second, or bottom-row equation, Eq. 2.6.

The change of energy $\dot{E}$ is a function of the right-hand side. The definition of energy will not change, however, it is no longer conserved. The change of energy is

given by the equations of motion:

$$(2.13) \qquad \dot{E} = \nabla_{\mathbf{x}} E^T \dot{\mathbf{x}} = -\gamma \dot{x}^2 + \dot{x} F \quad .$$

The friction can only lower the energy of the system, since $-\gamma \dot{x}^2 < 0$. Friction is a dissipative force. The applied force F can both lower and increase the energy of the system. The product of velocity $\dot{x}$ and force F is called the power product, as it indicates the rate of change of energy put in or taken out of the system. It will play an essential role in constructing connections between separate systems and defining boundary variables. Traditionally, in order to define an isolated system in a general setting, either $\dot{x} = 0$ is set, or $F = 0$. The latter corresponds to free motion, the first to fixed position.

Treating in friction on the same footing as the energy-conserving dynamics of the harmonic oscillator can be a formidable task. [Soper, 1976] The Rayleigh and the Gossick terms [Gossick, 1967] allow for the inclusion of damping in the equations of motion. However, friction or damping is not the loss of energy. It is the conversion of energy, to the thermal or vibrational domain. [Breedveld, 1984] Friction represents a port, or connection, between domains, just as applied forces are. Damping and friction are input and output between components. These components are not spatially separated, but conceptually.

There is another form of input, which is the constrained motion, which means specifying $x(t)$ or $\dot{x}(t)$. For a particular initial condition, $x(0), \dot{x}(0)$ at time $t = 0$, specifying $x(t)$ or $\dot{x}(t)$ is equivalent, since $\dot{x}(t)$ can be integrated with the initial condition $x(0)$ to yield $x(t)$. They specify a fixed motion in phase space, hence a trivial dynamics. However, in higher dimensions, fixing one position $x_i(t)$, would yield time-dependent equations of motion. In practice it is important to keep track of the force $f_i(t)$ associated with the constrained coordinate $x_i(t)$, as the pair tells how the energy changes in the system:

$$(2.14) \qquad \dot{E} = \frac{\partial E}{\partial x_i} \dot{x}_i = f_i \dot{x}_i \quad .$$

At this point it makes no difference whether we constrain the motion $x_i(t)$, or whether we applied the correct force $f_i(t)$ which yields this particular motion. In practice it will make a difference, physical models of motors which yield the applied force have a particular response which is a function of input $u(t)$, $x_i(t)$ and $f_i(t)$: $\chi(x_i, f_i, u) = 0$.

Apart from identifying a dynamical system by its phase space, one could also identify the system by its amplitude x_0 and phase ϕ, as specified in Eq. 2.2. The response is given in terms of amplitude and phase of an oscillatory $x(t)$ to an external oscillatory force $F(t) = F \cos(\omega t)$, or usually in frequency space $F(\omega)$. The plot of amplitude and phase against the frequency of the applied force $F(t)$ is called a Bode plot. For a linear system this identification suffices, the function $H(\omega) = \frac{x_0(\omega)}{F(\omega)}$ is called the transfer function from applied force to amplitude. However, in the nonlinear case this will not do. If the phase is included, one generally has much more insight in the system. However, global knowledge, as depicted in the phase space, is often lost. To reach certain amplitudes for, for example, the mathematical pendulum with a limited force max $F < F_0$, the frequency ω has to decrease with time to drive the system close to resonance with the proper phase.

2.3. Finite-dimensional dynamical systems

From the one-dimensional dynamical systems there are many routes along which we can extend to higher-dimensional systems. The simplest approach would be to consider n copies of a one-dimensional system, with $2n$ variables in canonical pairs. The Hamiltonian for n copies of a one-dimensional system is the sum of n copies of the one-dimensional Hamiltonian:

$$(2.15) \qquad H(\mathbf{q}, \mathbf{p}) = \sum_{i=1}^{n} H_i(q_i, p_i) \quad .$$

The Hamilton equation of motion is identical to the one-dimensional case, even for an interacting system:

$$(2.16) \qquad \begin{pmatrix} \dot{\mathbf{q}} \\ \dot{\mathbf{p}} \end{pmatrix} = \begin{pmatrix} 0 & \mathbf{1} \\ -\mathbf{1} & 0 \end{pmatrix} \begin{pmatrix} \nabla_{\mathbf{q}} H(\mathbf{q}, \mathbf{p}) \\ \nabla_{\mathbf{p}} H(\mathbf{q}, \mathbf{p}) \end{pmatrix} \quad .$$

The mass and stiffness matrices are given for a value of the position vector $\mathbf{q}$ and the momentum vector $\mathbf{p}$:

$$(2.17) \qquad M_{ij}^{-1}(\mathbf{q}, \mathbf{p}) = \partial_{p_i} \partial_{p_j} H(\mathbf{q}, \mathbf{p}) \quad ,$$

and

$$(2.18) \qquad K_{ij}(\mathbf{q}, \mathbf{p}) = \partial_{q_i} \partial_{q_j} H(\mathbf{q}, \mathbf{p}) \quad .$$

In the mathematical literature, [Arnold, 1989] however, a finite-dimensional Hamiltonian is more general than the n copies with $2n$ canonical variables. It might have an odd number of variables $\mathbf{x} = (\mathbf{q}, \mathbf{p})$, where there is no longer a clear distinction between position and momentum variables.

Considering just geometrical arguments would be one approach to arrive at such general Hamiltonians. The link with rotation may clarify some of the peculiarities. In that case only the energy, and the symplectic structure are of interest. If we still want to satisfy energy conservation, the only requirement is that the evolution of the system $\dot{\mathbf{x}}$ is perpendicular lies in the $E =$ constant surfaces, which will be of $n-1$ dimensions, if the space of $\mathbf{x}$ is n dimensional. Hence, $\dot{\mathbf{x}}$ has to be perpendicular to the gradient of the energy. A vector perpendicular to the gradient of the energy is necessarily the generator of rotation. Given the evolution equation:

$$(2.19) \qquad \dot{\mathbf{x}} = J \nabla_{\mathbf{x}} E(\mathbf{x}) \quad .$$

The symplectic matrix J should satisfy:

$$(2.20) \qquad J^T = -J \quad .$$

In a simplified case, where the energy surfaces, $E =$ contant, are surfaces of a sphere the canonical energy is $E = \frac{1}{2} \mathbf{x}^T \mathbf{x}$. The symplectic matrix corresponds to a rotation $\mathbf{X}$ of the vector $\mathbf{x}$:

$$(2.21) \qquad \int e^{Jt} dt = \mathbf{X}(t) \quad \text{with} \quad \mathbf{X}^T(t) \mathbf{X}(t) = 1 = \mathbf{X}(t) \mathbf{X}(t)^T \quad .$$

Rotations leave the sphere invariant: a point $\mathbf{x}$ with energy E is mapped unto a point $\mathbf{x}'$ with the same energy. In dynamical systems, from a configuration space, each position x_i has an associated velocity $\dot{x}_i$, yielding a canonical pair (x_i, p_i). On the other hand, the symplectic matrix J can be odd-dimensional. For example, for

the rotation of a free rigid body, no potential appears and the symplectic matrix J is the three-dimensional generator of rotations:

$$(2.22) \qquad \begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{pmatrix} = J \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \equiv \begin{pmatrix} 0 & \omega_{12} & -\omega_{13} \\ -\omega_{12} & 0 & \omega_{23} \\ \omega_{13} & -\omega_{23} & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \quad ,$$

where ω_{ij} are the components of the rotation vector. The three-dimensional generator is build up of three basic generators J_{12}:

$$J(\omega) = \begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \omega_{12} + \begin{pmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \omega_{13} + \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{pmatrix} \omega_{23}$$

$$(2.23) \qquad\qquad = J_{12}\omega_{12} + J_{13}\omega_{13} + J_{23}\omega_{23} = J \cdot \omega \quad .$$

The rotation dynamics of the rigid body is a well-known example of an odd-dimensional dynamical systems. Since the rotation of a rigid body has no associated potential energy, only the three angular velocities x_i are of interest, the actual orientation angles not. Hence only three of six degrees of freedom are kept in the dynamics. There are many other examples of reduced dynamics, they are the result of constraint relations, conserved quantities such as momentum, or approximations. One could argue that such reduced dynamics are the fundamental dynamics. However, the connection between position and velocity, useful for consistency, is often severed. For the high-dimensional and infinite-dimensional systems considered later the reduction will play a role in the index of the operators at hand. In practice, reducing a high-dimensional system by a couple of orders is not worth the trouble. Only for consistency, it is important to single out irrelevant variables.

Since a symplectic approach to dynamics is more general than a local product of one-dimensional systems, we will generally stick to the geometrical, symplectic approach. However, Darboux theorem yields an important connection between the two approaches. Locally, that is: in a small region of phase space, a symplectic structure $\{J, E\}$ can be mapped, through the right coordinate transformation $\mathbf{x}'(\mathbf{x})$, to a set of canonical pairs of the canonical energies, with remainder coordinates mapped to the evolution null-space, i.e., fixed x_i':

$$(2.24) \qquad \dot{\mathbf{x}}' = \begin{pmatrix} 0 & -1 & & & & & \\ 1 & 0 & & & & & \\ & & \ddots & & & & \\ & & & 0 & -1 & & \\ & & & 1 & 0 & & \\ & & & & & 0 & \\ & & & & & & \ddots & \\ & & & & & & & 0 \end{pmatrix} \mathbf{x}' \quad .$$

Hence, from the general mathematical description, in terms of the symplectic matrix, it is again possible to arrive at a local coordinate system, in some small neighborhood of the phase space, where the system corresponds again to n copies of the one-dimensional oscillator.

For the case of the three-dimensional rotation of a rigid body this theorem has a simple geometrical interpretation. The surface of constant energy is a sphere.
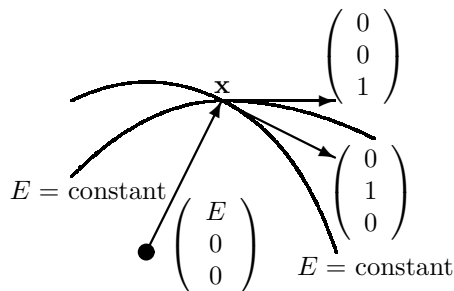
FIGURE 2.4. On the energy surface in three-dimensional space, the two directions perpendicular to the surface normal form a canonical pair.

At a particular point of the sphere, there are only two tangential directions in the energy surface. To first order, the third direction remains constant. For example, if $x_2 = x_3 = 0$, the generator J_{23} will change the system. For example, if we map the position on the sphere to $\mathbf{x} = (1, 0, 0)^T$, the two tangential directions, forming a canonical pair, undergoing a symplectic evolution, are $(0, 1, 0)^T$ and $(0, 0, 1)^T$. See Figure 2.4.

The energy sphere, invariant under rotations, has another interesting feature. The evolution of a system can be seen as a continuous transformation, mapping the phase space unto itself. Such a transformation is a group, and for many groups the structure is analyzed in depth. Hence it might be possible to map a dynamical system to a group. However, only for very simple linear dynamical systems such a transformation is interesting as it simplifies the formulation, and yields global properties, such as spectra. The one-dimensional harmonic oscillator can be mapped to the $SO(2)$ (all orthogonal 2×2 matrices with unit determinant) or the $U(1)$ (all unitary 1×1 matrices) group. The $U(1)$ group is the multiplication group of all complex phases:

$$(2.25) \qquad\qquad e^{i\phi}e^{i\psi} = e^{i(\phi+\psi)} \ ,$$

where the generator of rotations is the multiplication with i:

$$(2.26) \qquad\qquad \dot{z} = iz \ ,$$

where the complex variable represents the canonical pair: $z = x_1 + ix_2$, with the radial frequency is $\omega = 2\pi$. In this case the action-angle variables correspond to the modulus-argument variables $(|z|, \operatorname{Im} \ln z)$.

The equivalent $SO(2)$ formulation yield group elements: $X \in so(2)$:

$$(2.27) \qquad\qquad X(\phi) = \begin{pmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{pmatrix} \ ,$$

with the properties:

$$(2.28) \qquad X(-\phi)X(\phi) = X^T(\phi)X(\phi) = 1 = X(0) = X(\phi)X^T(\phi) \ ,$$

$$(2.29) \qquad\qquad X(\phi)X(\psi) = X(\phi + \psi) \ ,$$

$$(2.30) \qquad\qquad \partial_\phi X(\phi) = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} X(\phi) \ .$$

The nicest aspect of a group is that only a single point of the phase space has to be analyzed, all other points are related to that single point by a group transformation X. Furthermore, the coordinate system of the phase space, and the tangential vectors $\nabla_{\mathbf{x}} E(\mathbf{x})$ and $\dot{\mathbf{x}}$, are naturally given by the group generators and the group orbits. These principles can be used to improve the numerical simulation of such system. See Section 3.4.

Another class of groups are the symplectic groups, which are the coordinate transformations $\mathbf{x}'(\mathbf{x})$ which leave the dynamics invariant. Not all coordinate transformations are allowed, as is to be expected. This can most easily be seen from considering the configuration space $(\mathbf{x}, \dot{\mathbf{x}})$. Given a transformation of the position $\mathbf{x}'(\mathbf{x})$ the transformations of the velocity $\dot{\mathbf{x}}'$ is given:

$$(2.31) \qquad\qquad \dot{\mathbf{x}}' = \frac{\partial \mathbf{x}'}{\partial \mathbf{x}} \dot{\mathbf{x}} \ .$$

The matrix $F = \frac{\partial \mathbf{x}'}{\partial \mathbf{x}}$ is the transformation Jacobian, sometimes called the deformation gradient.[Truesdell and Toupin, 1960]

2.4. A bridge problem

Before we arrive at dynamical systems with an infinite number of degrees of freedom, we will first discuss a typical example of a finite-order approximation of such systems. All macroscopic objects have in principle an infinite number of degrees of freedom. A stone, subject to enough force will shatter in many pieces, but for all practical purposes it can be treated as a single extended object, with six degrees of freedom; translations and rotations.

Hence the number of degrees of freedom depends on the functionality; the type of interaction with its environment. One step beyond a completely undeformable object would be a simple deformation based on static applied force. [Cohen and Muncaster, 1988] The greater part of this thesis is devoted to put such approximations on a solid footing, with systematic extensions. Furthermore, if interaction defines the functionality of an object, and thereby its degrees of freedom, it is important to clearly define this interaction. Interaction for structural elements is often the result from identification of points or nodes. A single position can be part of two or more elements, thus joining them together. Implicitly, the mutual forces should cancel each other, according to Newton's action and reaction. In chemical and fluid systems different types of interactions exist. However, they can be seen as the natural partner of the structural collocation interaction. In such systems matter, charge, or another quantity is conserved. The transmission of such conserved quantity can be seen as the analogue of the action-reaction pair, the loss of one component equals the gain of the other. The notion of energy comes into the picture from the effort or potential at which this transmission occurs.

The modular approach consists of two major elements. First, the dynamical model of each part. Secondly, the interaction between two parts. The collocation interaction, identifying configuration points, leads to the simplest modular model; a dynamical model, given by the Hamiltonian, of the whole.

As an example of the modular approach to structural modelling we examine a structure like a bridge, which has horizontal and vertical segments. Each segment and each junction can be analyzed on its own, and the whole, with reduced order
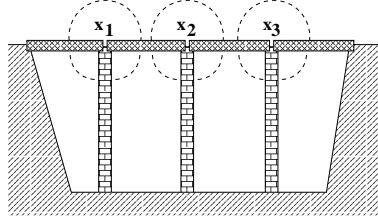
FIGURE 2.5. The analysis of the bridge combines all components. At each junction the position $\mathbf{x}_i$ and the angle between the components are relevant.

segments can be combined and analyzed. Apart from the computational advantage of not having to perform a very large structural calculation, the advantage of decomposing the large structure of a bridge into logical units is the fact that each unit, such as a bridge column, a connection, of a surface element, can be analyzed on its own, exposing its weaknesses.

The vibrational modes of a bridge or a comb arise in many engineering problems. The usual approach is a large structural calculation, using all the elements combined. However, such problems may be regular, having a large amount of repetitive structure, such a row of identical bridge columns and road segments. See Figure 2.5.

Each of the columns we may investigate separately, yielding a response to a force $\mathbf{f}$:

$$(2.32) \qquad\qquad \mathbf{x}(\mathbf{f}) \approx \mathbf{C}\mathbf{f} \quad .$$

For a composite column calculating the response to a stationary force is the easiest approach since for each segment for the column the same force applies, and the sum of all displacements is the total displacements in the linear approximation:

$$(2.33) \qquad\qquad \mathbf{C} = \sum_{\text{segments}} \mathbf{C}_{\text{segment}} \quad .$$

See Figure 2.6 The elastic energy of a column is:

$$(2.34) \qquad\qquad E_{\text{elastic}} = \frac{1}{2}\mathbf{x}^T \mathbf{C}^{-1}\mathbf{x} \quad .$$

However, this does not complete the analysis of the column. Possibly, the columns may possess sufficient inertia, that they yield an effect in lowering the vibrational modes of the bridge as a whole. The displacement $\mathbf{x}(z)$ as function of the distance from the top $z = 0$, to the bottom of the column $z = 1$ is inverse proportional to the integrated stiffness:

$$(2.35) \qquad\qquad \mathbf{x}(z) - \mathbf{x}(0) = \int_0^z \mathbf{c}(z)\mathbf{f}dz \quad .$$

For a homogeneous column the displacement is linear. In that case only a third of the mass of the column contributes to the inertia of the bridge:

$$(2.36) \qquad\qquad E_{\text{kinetic}} = \frac{1}{6}M_{\text{column}}\dot{\mathbf{x}}^2 \quad .$$

In most practical cases the bottom of the column is heavier and stiffer that the top, reducing the inertial column effects further.
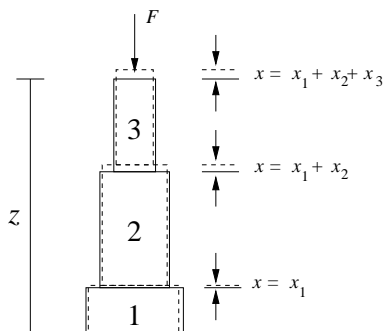
FIGURE 2.6. A single column can be analyzed. Its characteristic is the sum displacement as function of the applied force. The displacements of all segments can be summed to the total displacement.

The inertia of the surface of the bridge contributes in full to the vibrational inertia. The stiffness is however more complicated. Between two columns a surface segment has stretch and bending stiffnesses. The stretch stiffness depends only on the distance between the ends:

$$(2.37) \qquad E_{\text{stretch}} = \frac{1}{2}k(\mathbf{x}_i - \mathbf{x}_{i+1} - l)^2 \quad ,$$

where l is the length of a segment.

The bending stiffnesses depend both on the angle the bridge surface makes between two neighboring segments, and the angle each segment makes with each of the two support columns. Hence there are two tensorial elastic energy contributions depending on the vectors $(\mathbf{x}_i - \mathbf{x}_{i+1})$ and $(\mathbf{x}_i - \mathbf{x}_{i-1})$ for the surface bending, and $(\mathbf{x}_i - \mathbf{x}_{i+1})$ and $(\mathbf{x}_i - \mathbf{x}_{\text{base}})$ for the column-surface angle:

$$(2.38) \qquad E_{\text{bending}} = K(\mathbf{y}^T\mathbf{z}, \mathbf{y}^2, \mathbf{z}^2) \quad ,$$

where the variables of the function K are the invariants, like $\mathbf{y}^T\mathbf{z}$, $\mathbf{y}^2$, and $\mathbf{z}^2$, which can be made of the two relative vectors $\mathbf{y}$ and $\mathbf{z}$. The parameters appearing in the function, will depend on whether it represents the surface bending or the surface-column shear. See Figure 2.5 The position and orientation of the column or a surface element will not influence the elastic energy. The invariants single out the orientation-independent deformations.

Each of these terms can be determined in a limited and more detailed static analysis. Afterward all the elements are combined to a single energy, from which the Hamiltonian is constructed by inverting the mass matrix.

In Chapters 8 and 11 the elements of the finite element methods are constructed along the same lines, where in the latter chapter, Chapter 11 the energy is indeed constructed in terms of invariants. The simplest linear elements are most easily constructed, but elements satisfying some physical laws, such as orientation invariance, may predict the correct behavior much more accurately. The numerical code Spacar, to predict and control the behavior of industrial robots using the orientation of different points on a robot arm, is a good example, although the implementation is not in terms of invariants. [Waiboer et al., 2005]

2.5. Infinite-dimensional systems

Infinite dimensional systems come in four main categories: First, the countable infinite systems, which are made up of a countable set of finite systems. Second, the compact infinite systems, defined by a finite number of functions on a compact domain. For example the complex functions $f(z)$ with on the unit circle $|z| = 1$ is such a compact infinite system, which may be represented by a trigonometric sum $\{e^{inz}\}_{n=-\infty}^{n=\infty}$. Third, the infinite systems on an infinite domain. Such systems will cause all kind of mathematical and physical problems, if the so-called "behavior at infinity" is not specified. The domain $z \in \Omega$ on which the functions $f(z)$ are defined in any of the latter two categories is called the reference space.

Finally, there are ensemble systems, which arise by taking some finite-dimensional system given by the state $x(t)$ and evolving it for an infinite number of initial states $x(0) \in \mathcal{S}$, where $\mathcal{S}$ is a set from the infinite-dimensional state space. The properties of the domain $\mathcal{S}$ make little difference in the treatment of ensemble systems, as usually, for $t \to \infty$, the domain $\{x(t)|x(0) \in \mathcal{S}\}$ will not be compact and often very complicated. Chaotic systems are usually analyzed as ensemble systems. The special domain $\mathcal{S}$ of a lower dimension than the original space in which x resides, such that $\{x(t) \in \mathcal{S}|x(0) \in \mathcal{S} \wedge t \in \mathbb{R}^+\}$ is called an attractor, which, in nonlinear dynamics, can have a very complicated form. It corresponds to the subspace in which a system will end up in many cases. The Poincaré map $\mathcal{S}'_\Delta = \{x(t + \Delta t)|x(t) \in \mathcal{S}\}$ is an important tool to recover attractors. Mathematicians like Arnold [Arnold, 1973] and Smale [Hirsch and Smale, 1974] have investigated dynamics as mappings of a domain onto a space.

The first three systems: countable, compact infinite, and infinite are the systems we will consider here. Using mathematical tools and restrictions one can map systems downward, from infinite, to compact infinite, to countable. In the appropriate setting the numerical results of the countable system will converge fast at finite truncation.

The work of Weierstrass, Schauder, Leray, and others has established a clear link between compact infinite and countable. [Gohberg and Goldberg, 1981] For linear systems, Hilbert or Banach spaces, a basis for a function space, on a compact domain, would link a large class of functions to a countable basis, the Schauder basis. Typically a set of eigenmodes of vibration would be a useful countable basis for the displacements of a compact object. However, such countable basis is the result of proper conditions on the operator underlying the vibration dynamics, which are not always satisfied. In the case of nonlinear operators there are few guarantees for a connection to a countable basis.

Functions on an infinite domain suffer from more mathematical difficulties. Intrinsically complicated physical problems, such as the irreversibility in thermodynamics, the decay of quantum states, or radiation into open space, will manifest itself in the mathematical description, typical for infinite-dimensional systems on an infinite domain. For example, radiation will radiate off to infinity. Any measurement, corresponding to a compact operator, will yield eventually zero, but without energy or mass loss in the system. Irreversible systems connected to a heat bath suffer from the same complications in connection with mathematics. These problems are also related to the problems around the direction of time. [Rosenfeld, 1961] The resulting irreversible systems require different modelling principles. [Breedveld, 1984]

Apart from these, more or less, fundamental distinctions among infinite systems, there are many mathematical distinctions among infinite systems, characterized by the treatment of the system, rather than its nature. It is important to note that, for example for a partial differential equation, there is little to say about the solution, if one does not specify the space in which this solution lies. This space will be considered in this thesis more freely than most approaches in functional analysis prescribe.

In functional analysis, one starts with the notion of "size" in the form of a norm $\| \cdot \|$ of a function. Several norms exist, and different norms may be used in conjunction. For compact infinite systems the norm of a function is usually the integral over this function:

$$(2.39) \qquad \|f\|_n = \left| \int_\Omega |f(x)|^n dx \right|^{\frac{1}{n}} \quad ,$$

which gives rise to Hölder's inequality:

$$(2.40) \qquad \left| \int_\Omega f(x)g(x)dx \right| \le \|f\|_p \|g\|_q \quad ,$$

for any $p, q \ge 0 : p^{-1} + q^{-1} = 1$. Therefore the norm $\| \cdot \|_2$ is special, as one can associate a bilinear inner-product, or dot-product, to it:

$$(2.41) \qquad \langle f, g \rangle = \int_\Omega f^*(x)g(x)dx \quad ,$$

where f^* is the complex conjugate of f (if f is complex), such that $\langle f, f \rangle = \|f\|_2^2 = \|f^*\|_2^2$. This space is called the Lebesgue space; any function in this domain is $f \in \mathcal{L}^2(\Omega)$. [Rudin, 1966]

Important to the notion of norm is the fact that zero norm means the zero-function:

$$(2.42) \qquad \|f\| = 0 \Rightarrow f = 0 \quad .$$

Rather than a result, this implication should be read as a definition of the zero-function in a particular space. For integrals many functions are equivalent, in the sense that their difference is a zero function: $\|f - g\| = 0$.

The norm turns out to be very useful, as most global properties of a system, such as the mass and energy, can be expressed as norms. However, without any sense of "distance" on the reference space Ω, a system will just be a very large collection of independent systems. In Table 2.1 different notions of distance are listed. The notion of distance gives rise to the notion of collective behavior: the behavior $f(z)$ at two points z_1 and z_2 will be similar if these points are close together. The behavior is correlated. Note that this distance is in the reference space Ω, and not the infinite function space. The spatial derivative gives rise to this notion. Higher values for the derivative means less correlation. The notion of collective behavior means that $\partial_z f(z)$ is bounded in some sense. Hence, it might be more appropriate to start with a mathematical function space which is tied closer to a definition of a spatial derivative ∂_z.

Instead of using norms, other mathematical views of functions are a more natural means to consider differentiability. One could consider analytic functions, which are infinitely differentiable on its domain Ω. If the domain is simply connected, such

name	topology	
point set	subsets	$x \in \mathcal{S}' \subset \mathcal{S}$
metric	distance	$\rho(x, y) \geq 0$
normed	norm	$\|x\| \geq 0$ with $\rho(x, 0) = \|x\|$
Banach	linear norm	$\|\lambda x\| = |\lambda| \|x\|$
Hilbert	inner product	$\langle x, y \rangle$ with $\langle x, x \rangle = \|x\|^2$
linear algebra	mass matrix	$\mathbf{M}$ with $\langle x, y \rangle = \mathbf{x}^T \mathbf{M} \mathbf{y}$

TABLE 2.1. Closed and complete spaces in analysis. Completeness means that a converging series $\|x_i - x_{i+1}\| \to 0$ has a limit $\|x_i - x\| \to 0$ within the space. Dynamical systems fall naturally in the category of normed spaces, where the energy $E = \|x\| \geq 0$. For mathematical analysis more structure, like a Banach space or Hilbert space is more useful. For numerical analysis, a finite linear algebra, at least locally, is necessary.

that is has no holes, a small neighborhood of a single point on the domain is required for the evaluation. The value at all other points $z \in \Omega$ result from analytic continuation. Complex functions have many other advantages, since integration and differentiation are naturally defined on complex functions. The consistency relations for complex analyticity, the Cauchy-Riemann relations, are even defined as differential equations. See, for example, Section 5.5, where the boundary conditions for the Cauchy-Riemann equations and the Laplace equation are investigated.

However, analyticity is a very strict condition, many functions will not satisfy this property, and, for example, no function on a compact support, such that it is only nonzero on a compact domain Ω, can be analytic. There are means around this problem, by considering only finite domains, which lead to, e.g., Fourier series and Hardy spaces, however, the class of finite differentiable functions $\mathcal{C}^n(\Omega)$ will be in practice the function space of choice. It only requires the necessary condition needed to perform the calculations. It is a strictly local property: in every neighborhood of a point $z \in \Omega$, the n-th derivative of a function $f \in \mathcal{C}^n(\Omega)$ is continuous. The derivative of a function can be seen as an algorithmic or operational property. Looking at differentiation is such an abstract sense, defining "$\partial_z f(z)$" rather than calculating it, avoids some of the pitfalls of real analysis. The use of differential operators in finite function spaces is a possible alternative. See Sections 2.8 and 11.4.

Real functions of real variables are a possible beehive of stings. Weierstrass has learned us as much. It is possible to define the most horrible functions, which are nowhere continuous, or differentiable. [Gelbaum and Olmsted, 1964] However, the functions in this thesis are the mathematical representations of physical systems, therefore, we assume they behave nicely, at least, at the scale we look at them. Furthermore, they are the result of some external input, and eventually we will discover that the quality of the input, we are free to choose, will determine the quality of the functions representing the system under investigation.

2.6. Notions of distance, approximation, and change

Continuous dynamical systems are based on three notions, related to three forms of continuity: the continuity of the underlying space $z \in \Omega \subset \mathbb{R}^n$, the

continuity of the states, functions or fields, on that space, $f(z) \in \mathbb{R}^m$, and the continuity in time; the dynamics A. These three notions are intimately related. Without the underlying space Ω the dynamics consists of infinitely many copies of finite-dimensional dynamics for each of the positions z. The differential equations couple these systems through the notion of distance: $\|z - z'\| \approx 0$, however, regularity, and collective motion arise only if the states change only slowly with position:

$$(2.43) \qquad \lim_{z \to z'} \frac{\|f(z) - f(z')\|}{\|z - z'\|} = \text{finite} \quad,$$

which is typical for linear, Banach spaces. However, more closely related to the notion of energy, $E = \|f\|$ is the finite variation in the form:

$$(2.44) \qquad \lim_{z \to z'} \frac{\|f(z)\| - \|f(z')\|}{\|z - z'\|} = \text{finite} \quad.$$

From the global norm, $\|f\|$, over the whole domain, for example $\|f\| = \max_{z \in \Omega} \|f(z)\|$, the local norm $\|f(z)\|$ can be reached through the use of either some test, or measure, functions, or a variational principle:

$$(2.45) \qquad g(z) = f(z) + \epsilon h(z) \quad,$$

where $h(z)$ is some localized variation, around z and z', and ϵ an infinitesimal quantity. It uses the Taylor expansion in ϵ and thereby the tools of linear analysis.

However, generally, the energy $E = \|f\|$, and the fields $f(z)$ on the domain Ω have a strained relation. Many approaches are based on some form of regularity. A priori finite approximations are maybe the most successful approaches. [Kantorovich and Krylov, 1958, Funaro, 1992, Creutz, 1983] They form the basis of Galerkin methods and finite element methods.

Sobolev spaces are another method to introduce regularity in the functions. The modern theory of existence and uniqueness of elliptic partial differential equations is based upon the theory of Sobolev spaces. Not only the values of the functions are bounded, but also the variations, given by the derivatives up to order l. The typical Sobolev norm, defining the Sobolev space $W_p^l(\Omega)$ is:

$$(2.46) \qquad \|f\|_{W_p^l} = \left(\sum_{i=0}^{l} \int_\Omega |\partial_z^i f(z)|^p dz \right)^{\frac{1}{p}} < \infty \quad.$$

In particular for $p = 2$ the Banach space has an inner product and defines a Hilbert space, with all its nice consequences. For formal results such as existence and uniqueness [Hörmander, 1990] the Sobolev space has been instrumental. In order to find some finite approximation $f_d(z)$ of a solution $f(z)$, such that $\|f_d - f\| < \epsilon$, these methods have limited use. Eventually, Weierstrass approximation theorem, [Yosida, 1980] which proves that a continuous function on a finite domain can be approximated by a polynomial, in the local sense $\|f_d(z) - f(z)\| < \epsilon$, remains one of the strongest results for those who view functions in the classical sense, i.e., as a graph $(z, f(z))$. The lifted approaches of jet spaces, and fiber and vector bundles can be seen as a geometrical extension of this view. [Abraham et al., 1991] The lifted approach is particularly useful in the study of integrability, if the state variables do not correspond to the configuration variables. [Ivey and Landsberg, 2000] See also Section 4.3.   indexintegrability

If a linear normed space, or Banach space, is separable, i.e., can be expanded in a countable set, it can be mapped, to a continuous function $C[0,1]$ on the domain $[0,1]$, endowed with the standard norm: $\|f\| = \max_z \|f_d(z) - f(z)\|$. [Liusternik and Sobolev, 1961] For Banach spaces, Schwartz spaces and other distribution, or generalized function, spaces, have proved also useful in combination with Fourier-Laplace transformations.[Gel'fand and Shilov, 1964, Shilov, 1968, Schwartz, 1966]

With the change, or dynamics, of a system there seems little formal problem. The integrability condition of the kernel, or evolution operator, A, is very simple, and a wide variety of operators may be integrated without problem, formally or explicitly, considering the correct measure and integration theory, typically Stieltjes integration, sometimes Lebesgue integration for stationary and variational problems which are more easily dealt with in Hilbert space:

$$(2.47) \qquad\qquad x(t) = e^{At}x_0 = T_t x_0 \ \ .$$

The complication arises from the interference between the approximation of the space Ω and, in particular, the approximation of the fields $x = f(z)$. They interfere with the desired features of the time-evolution. For example, in the case of energy conservation $\|x\| = E = $ constant, the time-translation operator T_t must be a norm-preserving unitary operator, $\|x_0\| = \|T_t x_0\|$, in short, $\|T_t\| = 1$ in operator norm. [Gohberg and Goldberg, 1981]

Besides regularity, homogeneity can help to establish results. One of the most successful treatments of linear dynamical systems is based on the time-translational invariance of the system. The system can therefore be described by a semi-group:

$$(2.48) \qquad\qquad T_{t+t'}x = T_t T_{t'}x = e^{A(t+t')}x = e^{At}\left(e^{At'}x\right) \ \ .$$

The continuity in time follows from the basic properties of the semi-group, e.g., $T_0 = \mathbf{1}$. [Yosida, 1980, Curtain and Zwart, 1995]

On the other hand the dynamics of the system dictates a relation between the space discretization and the time discretization of a system. The well-known Courant-Friedrichs-Lewy relation determines that the step size must be smaller than the mesh size divided by the characteristic velocity for stability. [Lax and Wendroff, 1960, Courant and Hilbert, 1961, Morton and Mayers, 1994] Hence at different levels the continuities are intertwined.

From the abstract continuous systems there are many routes to a variety of finite-order numerical models. Not only the approximations, but also the order in which they are made, will determine the results. It is important to structure these approximations, and the different notions of continuity.

2.7. Variational methods

In general, variational methods are a powerful method to find functional solutions. In dynamics, they yield an alternative formulation to the equations of motion. In optimization they yield an explicit implementation of multiple criteria in a single cost function, possibly augmented with Lagrange multipliers. Together with a variational principle, an optimal solution can thus be found. [Carathéodory, 1935, Lanczos, 1966, Mikhlin, 1964, Troutman, 1983]

In this thesis extensive use is made of the energy functional, i.e., the energy as function of the state functions or fields. To link this global property to equations of

motion, variational principles and variational derivatives are required. See, among others, Sections 5.10, 7.5, and 5.3. In this section we will give some background to these methods. Variational methods were introduced by Johann Bernoulli and Euler. They stated that any physical problem can be formulated as an extremal problem of some kind.

Historically, with variational principles the modern analysis of static and dynamical systems started. The variational form, such as the minimum or maximum of a function, allowed one to transform the problem into a convenient form, in order to find the solution. Coordinate transformations, and model reduction, are therefore closely related to the variational form of problems described by differential, and partial differential, equations. Still such formulations prevail in many of the descriptions of physical systems. Some notable exceptions are, for example, Faraday and Gauss, who, each in their own way had a more visual and geometrical view of the physical systems they studied. In Chapter 12 we establish some connections between the variational and the geometric views. In some respect is the variational principle more fundamental than the formulation in terms of equations of motion or equations of stability.

We start with a simple example of a line $x(z)$ as a function of $z \in [-1, 1]$. A variational principle would be the shortest path between $x(-1) = x_-$ and $x(1) = x_+$. It is common knowledge that the shortest path is a straight line:

$$(2.49) \qquad x_0(z) = \frac{1}{2}x_-(1-z) + \frac{1}{2}x_+(1+z) \ .$$

However, to derive this result is less than straightforward.

The path length is the integral:

$$(2.50) \qquad J[x] = \int_{-1}^{1} dz \sqrt{(\partial_z z)^2 + (\partial_z x)^2} = \int_{-1}^{1} dz \sqrt{1 + (\partial_z x)^2} \ .$$

If a path $x(z)$ is the minimum, any variation $x(z) \to x(z) + \delta x(z)$ means an increase of length. The minimum path satisfies that an infinitesimal variation δx, with $\delta x(-1) = \delta x(1) = 0$, vanishes:

$$(2.51) \qquad J[x + \delta x] - J[x] = 0 \ ,$$

which yields, by expanding the integral and partial integration, the differential equation:

$$(2.52) \qquad J[x + \delta x] - J[x] = \int_{-1}^{1} dz \frac{\delta x}{\sqrt{1 + (\partial_z x)^2}} \partial_z^2 x = \frac{1}{\sqrt{1 + (\partial_z x)^2}} \partial_z^2 x = 0 \ .$$

Since is for any $\delta x(z)$ the integral should vanish, the factor must vanish for all z.

Hence, the straight path $x_0(z)$ is a solution to this equation. Such a solution is called a geodesic, and commonly the, rather unseeming, square root is removed to yield the standard action principle $\mathbf{x}(z) \in \mathbb{R}^n$:

$$(2.53) \qquad J[\mathbf{x}] = \frac{1}{2} \int dz \dot{\mathbf{x}}^T \dot{\mathbf{x}} \ .$$

The z coordinate is interpreted as time coordinate, with $\partial_z \mathbf{x} = \dot{\mathbf{x}}$.

Many different variational principles exist. Commonly, not the initial and final positions are given, but just the initial position $x(-1)$ and the initial velocity
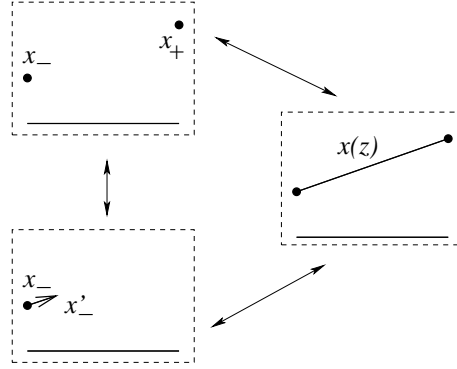
FIGURE 2.7. The equivalent descriptions of a line: the values x_- and x_+ of both ends, the value x_- and the derivative x'_- at one end, and the linear function $x(z)$.

$\partial_z x(-1)$. However, the same variational principle yields the same differential equation, with a different set of boundary conditions:

$$(2.54) \qquad \partial_z^2 x(z) = 0, \quad x(-1) = x_-, \quad \partial_z x(-1) = x'_- \quad ,$$

yielding the same straight line, in terms of different conditions:

$$(2.55) \qquad x_0(z) = x_- + (z+1)x'_- \quad .$$

The mapping between different sets of boundary conditions, such as between x_-, x_+ and x_-, x'_-, is an important aspect of this work. In this case the map is linear and straightforward:

$$(2.56) \qquad \begin{pmatrix} x_- \\ x_+ \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} x_- \\ x'_- \end{pmatrix} \quad .$$

See Figure 2.7.

The geodesic is not always a straight line. The metric tensor g is the geometric equivalent of the mass matrix, which may depend on the position x:

$$(2.57) \qquad J[x] = \frac{1}{2} \int dz\, g_{ij} \dot{x}^i \dot{x}^j \quad .$$

For example, the geometric theory of light only takes into account the refraction index $n(x)$, which yields a change of velocity, due to varying densities of the traversing medium. The functional for a geometric light ray is:

$$(2.58) \qquad J[\mathbf{x}] = \frac{1}{2} \int dz \frac{1}{n(\mathbf{x})^2} \dot{\mathbf{x}}^T \dot{\mathbf{x}} \quad .$$

For example, in the atmosphere light rays are deflected by warm air with a lower refraction index n.

A solution $\mathbf{x}(t)$ which minimizes the time, or in this case the approximate functional $J[\mathbf{x}]$, is called a geodesic. It is a special case $n^{-2}\delta_{ij} = g_{ij}$ of the general metric tensor g_{ij}. The corresponding geodesic differential equation is:

$$(2.59) \qquad \partial_t^2 x^i + \Gamma_{jk}^i \partial_t x^j \partial_t x^k = 0 \quad ,$$

FIGURE 2.8. The two extreme cases of initial conditions, and the corresponding time-evolution along the characteristic directions, given by the arrows. The two functions represent the initial value $u(z, t = 0)$ and the initial time-derivative $\dot{u}(z, t = 0)$. The le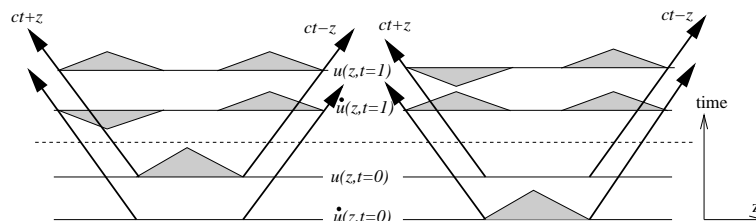ft case corresponds to only a non-zero value, the right case to only a non-zero derivative. Both correspond to the superposition of two triangular wave fronts moving away in opposite directions. Separating the boundary conditions into $u_\pm(z, t = 0) = c\partial_z u(z, t = 0) \pm \dot{u}(z, t = 0)$ leads to left-moving and right-moving parts, related to scattering variables.[John, 1982]

where the Christoffel symbols Γ^i_{jk} are given by:

$$(2.60) \qquad \Gamma^i_{jk} = \frac{1}{2} g^{im} \left(\partial_{x^k} g_{mj} + \partial_{x^i} g_{mk} - \partial_{x^m} g_{jk} \right) \quad ,$$

The geometric theory of light, using only the direction of the ray given by geodesics, is an approximation of the wave theory of light. [Born et al., 1980] In the theory of hyperbolic partial differential equations, or wave equations, the dynamics of the ray correspond to the dynamics of shock waves or singularities, which retain their form during propagation. The path of propagation is described by the characteristics of the wave equation: [Hörmander, 1990]

$$(2.61) \qquad \partial_t^2 y(t, z) - c^2 \partial_z^2 y(t, z) = 0 \Rightarrow \text{characteristics}: ct = \pm z \quad ,$$

where c is the wave velocity. Characteristics have extensive use in the theory of partial differential equations. In the case of linear equations with constant coefficients, the characteristics are simple. See Figure 2.8. In many other cases the construction of characteristics is as complicated as solving the hyperbolic equations themselves. [Carathéodory, 1935, Kamke, 1962]

For a long time it was thought that wave equations, of the form above, were fully characterized by tracing back the rays along the characteristics, but the reality turned out to be more complicated, depending on the dimension of the space. [Gårding, 1998]

2.8. Differential operators

A number of near-equivalent formulations of classical electrodynamics exists. [Barut, 1981] The Lagrangians differ only by a total divergence, a term which can be written as:

$$(2.62) \qquad \mathcal{L}_{\text{total divergence}} = \nabla \cdot \mathbf{S} \quad .$$

In practice, it means that the action integral will only have contributions from such terms at the boundary, through the divergence theorem:

$$(2.63) \qquad \int_\Omega \nabla \cdot \mathbf{S} d^n z = \int_{\partial\Omega} \mathbf{n} \cdot \mathbf{S} d^{n-1} z \to 0 \quad .$$

In electrodynamics the boundary is usually set at infinity, where the term is expected to vanish, giving rise to the equivalence. Rather than dealing with cleverly chosen boundary conditions, we study here boundary control; the effects of boundary conditions and the variations of them. Hence such near-equivalent formulations in terms of Lagrangians differing only a total divergence $\mathcal{L}_{\text{total divergence}}$ are rather ambiguous. The energy of the field will lead to a consistent definition to which the boundary conditions can be related.

The connecting elements are the differential operators. [Lanczos, 1961, Gilles, 1973] Adding a total divergence to a Lagrangian allows one to rewrite the equations, ignoring the boundary terms:

$$(2.64) \qquad (\partial_z \phi)^2 - \partial_z(\phi \partial_z \phi) = -(\phi \partial_z^2 \phi) \quad .$$

The second term on left-hand side is a total divergence, which allows us to rewrite the weak formulation, in terms of the energy or Hamiltonian density, to the strong formulation, in terms of the differential equation, using variational principles.

Another manner to express the relation between the weak and the strong formulation is the use of formal adjoint differential operators. The adjoint operator D^* arises partial integration under the integral ignoring the boundary terms, previously represented by the total divergence:

$$(2.65) \qquad \int dz \phi(z) \partial_z \psi(z) = - \int dz (\partial_z \phi(z)) \psi(z) + \text{boundary terms} \quad .$$

Hence $D^* = -\partial_z$ is the formal adjoint of $D = \partial_z$.

The boundary terms can be seen as a functional on the boundary, the trace operator, or the boundary operator θ. In Table 2.2 we list a number of well-known differential operators and their boundary or trace operators. The difference between the operator D and its formal adjoint D^* is the boundary operator:

$$(2.66) \quad \langle f, Dg \rangle - \langle g, D^* f \rangle = \int_\Omega (fDg - gD^*f) d^n z = \int_{\partial\Omega} f\theta g \, d^{n-1} z = (f, \theta g) \quad .$$

For example, the curl operator $D = \nabla \times$ maps a vector-valued function $\mathbf{f}$ back onto a vector-valued function. A bi-linear functional $J[\mathbf{f}, \mathbf{g}]$ may be constructed with the curl operator:

$$(2.67) \quad J[\mathbf{f}, \mathbf{g}] = \int_\Omega \mathbf{g}^T(\nabla \times \mathbf{f}) = \int_\Omega \epsilon_{ijk} g_i \partial_j f_k = \int_\Omega \epsilon_{ijk}(\partial_j(g_i f_k) - f_k \partial_j g_i) \quad .$$

$$(2.68) \qquad\qquad = \int_{\partial\Omega} \mathbf{n}^T(\mathbf{f} \times \mathbf{g}) + \int_\Omega \mathbf{f}^T(\nabla \times \mathbf{g}) \quad .$$

Hence, for $D = \nabla \times$ the formal adjoint is $D^* = D = \nabla \times$ and the boundary operator is $\theta = \mathbf{n} \times$. The Poynting vector, expressing the electromagnetic energy flow through the boundary is an example of such a boundary operator. [Jackson, 1975]

Combining adjoint operators with variational equations, the differential operator D can be inserted to the variation, leading to the relation:

$$(2.69) \qquad\qquad \delta_\phi H(\phi) = D^* \delta_{D\phi} H \quad ,$$

D	D^*	θ
∂_z	$-\partial_z$	1
∂_z^2	∂_z^2	$\partial_z - \partial_z^\dagger$
∇	$\nabla\cdot$	$\mathbf{n}$
Δ	Δ	$\mathbf{n}^T\nabla - \mathbf{n}^T\nabla^\dagger$
$\nabla\times$	$\nabla\times$	$\mathbf{n}\times$

TABLE 2.2. Differential operators, D, their formal adjoints, D^* and the boundary operators θ, yielding together the Green identities. The dagger $\dagger$ indicates the differential operator acts on the left: $f\partial_z^\dagger g = (\partial_z f)g$

which is equivalent for fields ϕ which do not vanish under differentiation $D\phi \neq 0$. This statement can be made more explicit if we use a particular basis $j = \{0, 1, 2, \cdots\}$:

$$(2.70) \qquad \phi_j(z) = \frac{1}{j!}z^j \quad,$$

and a particular differential operator:

$$(2.71) \qquad D = \partial_z^n \Rightarrow D^* = (-1)^n\partial_z^n \quad,$$

such that $D\phi_j = \phi_{j-n}$ for $j \geq n$ and $D\phi_j = 0$ for $j < n$. Hence, the variational derivative with respect to ϕ_j includes ϕ_j with $j < n$, while the variational derivative with respect to $D\phi_j$ does not:

$$(2.72) \qquad [\delta_{\phi_j}H]_i \Rightarrow [D^*\delta_{D\phi_j}H]_i = \left(\begin{array}{cc} j < n & 0 \\ j \geq n & [\delta_{\phi_j}H]_i \end{array} \right) \quad,$$

where the projection $[f(z)]_i$ of $f(z)$ onto the state ϕ_i is achieved through the use on a dual basis $\{\phi_i^*\}$:

$$(2.73) \qquad [\phi_j]_i = \langle \phi_i^*, \phi_j \rangle = \delta_{ij} \quad,$$

which must be based on the exact integral inner-product, on a fixed domain, for example $z \in [0, 1]$:

$$(2.74) \qquad N_{ij} = \langle \phi_i, \phi_j \rangle = \int_0^1 dz\phi_i\phi_j = \frac{1}{(i+j+1)i!j!} \quad,$$

such that:

$$(2.75) \qquad \phi_i^* = \sum_j N_{ij}^{-1}\phi_j \quad.$$

Simply said, the use of the differential operator in the variational derivative corresponds to the removal of the corresponding terms of the Taylor expansion of the original function $f(z) = \delta_\phi H$:

$$(2.76) \qquad f(z) \rightarrow f(z) - z\partial_z f(z)|_{z=z_0} - \frac{1}{2}z^2\partial_z^2 f(z)|_{z=z_0} \cdots \quad.$$

The arbitrariness in the polynomial subtraction, apparent through the indeterminate z_0, are precisely the degrees of freedom not determined by the variational derivative of $\delta_{D\phi}$, but determined by other means, usually boundary conditions or as constants of motion.

Later, in Chapter 7 we will use these methods to single out the part of the Hamiltonian which should be determined by the input through boundary conditions. The differential operator D is set to be the same as the leading differential operator in the Hamiltonian. In this manner, the arbitrariness, i.e., the null-space, of the Hamiltonian is determining the input and the corresponding boundary conditions, rather than boundary conditions being the starting point of a proper description of continuous dynamical system. For example, this allows us to use the same Hamiltonian for a component in the case it is freely moving, and in the case it is fixed. In the first case, the arbitrariness is a conserved momentum, set by the initial conditions at the starting time. In the second case, the arbitrariness is fixed by the boundary conditions for every time.

2.9. Complex coordinates

The use of complex coordinates can be very practical in a number of cases, due to the inherent structure of analytical functions. The complex variable $z = x + iy$ is two dimensional. [Conway, 1978] An analytical function $f(z)$, is a function that can be differentiated with respect to z infinitely many times, in most of the complex space, excluding possible singular points or singular lines. The function $f(z)$ is also two dimensional:

$$(2.77) \qquad f(z) \equiv f(x, y) = f_1(z) + if_2(z) \ ,$$

where f_i is a real function. Analyticity requires that the $f(x, y) : \mathbf{R}^2 \to \mathbf{R}^2$ satisfies the Cauchy-Riemann equations:

$$(2.78) \qquad \frac{\partial f_1}{\partial x} = \frac{\partial f_2}{\partial y} \quad \text{and} \quad \frac{\partial f_2}{\partial x} = -\frac{\partial f_1}{\partial y} \ .$$

In most cases only one of the two functions, f_1 or f_2, is of interest. Therefore the reduction of the analytical function to this quantity consists of projecting onto the real (Re) or the imaginary (Im) part. In the case of time-dependent results, the time dependence is given by the complex phase $e^{i\omega t}$. In this case the amplitude is given by the modulus of a complex function $|f|$.

It has some interesting consequences, which makes analytical functions useful for describing two dimensional, and $2n$ dimensional, systems. First, each of the $f_i(x, y)$ satisfies the Laplace equation:

$$(2.79) \qquad \Delta f_i = \frac{\partial^2 f_i}{\partial z \partial \overline{z}} = \frac{\partial^2 f_i}{\partial x^2} + \frac{\partial^2 f_i}{\partial y^2} = 0 \ ,$$

for the very simple reason that $f_i(z)$ is not a function of the complex conjugate coordinate $\overline{z} = x - iy$.

Second, any analytical function $f(z)$ can serve as a conformal mapping, or symplectic coordinate transformation from $z \to z' = f(z)$. Such coordinate transformations are angle preserving, and an analytical function $g(z)$ will also be a solution of the Laplace equation in the new coordinate: $\Delta_{z'} g(z') = 0$. Since x and iy are perpendicular, and can be seen as the potential surfaces and the streamlines of a dynamical system, so will x' and iy' be. The presence of a singularity in the transformation $z \to z'$, e.g., $z' = z + z^{-1}$, will cause the presence of a circulating flow in the new coordinate, if none was present in the original coordinates. The principles of conformal mapping are used the solve the Laplace equation and potential problems in two dimensions in all kind of geometries, such as wedges, cylindrical

cavities. Furthermore, Joukowsky (Zhukovsky) and Kutta derived independently an important result for incompressible flows, without friction or viscosity, for wing sections, which basically states that any wing section will do for such limited model. [Tokaty, 1971]

Third, the Cauchy theorem is Stokes theorem in two dimensions, and gives us the means to invert the Laplace operator and solve inhomogeneous differential equations in two dimensions. The Cauchy theorem, also known as the residue theorem, tells us the relation between the line integral clockwise over a closed contour $\Gamma = \partial\Omega$, or the boundary of a domain Ω, and the values at poles in the domain:

$$(2.80) \qquad \int_\Gamma f(z)dz = 2\pi i \sum_{\text{poles } z_j} \partial_z((z - z_j)f(z))|_{z=z_j} \ .$$

For example, for $f(z) < \infty$ for z lying inside the contour $z \in \Omega$:

$$(2.81) \qquad \frac{1}{2\pi i} \int_\Gamma \frac{f(z)dz}{(z - z')} = \left\{ \begin{array}{cc} f(z') & z' \in \Omega \\ 0 & z' \notin \Omega \end{array} \right. \ .$$

Hence it tells us, that if we know a function $f(z)$ on the boundary Γ, we can construct a solution $f(z)$ for $z \in \Omega$, which satisfies the Laplace equation: $\Delta f = 0$, using the integral above. It is known as the Dirichlet problem. The function $(z - z')^{-1}$ is also known as the Green function for this problem, converting a differential equation into an integral equation. Care must be taken to understand the form $f(z)dz$ correctly. This is the complex product, and written out in terms it is the exterior product for the real part of the expression:

$$(2.82) \qquad f(z)dz = (f_1 + if_2)(dx + idy) = f_1dx - f_2dy + i(f_1dy + f_2dx) \ .$$

In the chapter of partial differential equations, Chapter 5, we will discuss the constraints on the function $f(z)$ on the boundary $z \in \Gamma = \partial\Omega$, as and example of problems with boundary conditions.

In quantum mechanics the complex functions are more than a practical tool. The wave function in quantum mechanics is complex. Only our observations are real, and therefore a measurement is a projection on the real axis in the complex plane.

In higher dimensions there is no unique extension of complex spaces. Three important extensions exist. First, the quarternions, which is a three dimensional extension of the skew symmetry, similar to the generators of the rotation group. Second, the projection of a $2n$ dimensional space on n complex planes. In the sense of Darboux theorem such a projection exists for nonsingular symplectic geometry, however, in general such projections are an active field of research. Third, the mapping on an Euclidean norm for the modulus, with some winding number as complex argument, or phase.

2.10. Bipolar coordinates

A typical nontrivial example of a conformal mapping of complex coordinates is the mapping which produces bipolar coordinates. The solution of the Laplace equation $\Delta f = 0$ in the case of a two-center problem, or a problem between eccentric cylinders yield easy solutions in terms of bipolar coordinates.

Bipolar coordinates ξ, η give the Cartesian x, y-position in the plane: [Korn and Korn, 1968]

$$(2.83) \qquad x = \frac{\alpha \sinh \xi}{\cosh \xi - \cos \eta} \quad,$$

$$(2.84) \qquad y = \frac{\alpha \sin \eta}{\cosh \xi - \cos \eta} \quad.$$

It can be seen as the conformal mapping:

$$(2.85) \qquad v = \log \frac{u + \alpha}{u - \alpha} = \log \left| \frac{u + \alpha}{u - \alpha} \right| + i \left(\arctan \frac{y}{x + \alpha} - \arctan \frac{y}{x - \alpha} \right) \quad,$$

with $u = x + iy$ and $v = \xi + i\eta$. [Spiegel, 1964, Magnus et al., 1966] This is also known as a bilinear conformal mapping. The linear conformal map is the Möbius transform which maps lines and circles to lines and circles. See also Figure 12.4.

The Laplacian is

$$(2.86) \qquad \Delta_{x,y} f = \frac{1}{\alpha^2} (\cosh \xi - \cos \eta)^2 \Delta_{\xi,\eta} f \quad.$$

Hence the coordinates ξ and η themselves are solutions of the Laplacian, except at the singularities $\xi = \pm\infty$.

The bipolar coordinates have several uses. For example, the electrostatic field between two charged cylinders is most easily described in terms of bipolar coordinates, and the Newtonian flow between two rotating eccentric cylinders is given by the velocity potential in bipolar coordinates

In Section 12.3 the geometric aspects of bipolar coordinates are used the analyze the material dependency of the transmission of force through a plane.

2.11. Infinite matrices

From continuous systems, through some set of basis functions, one arrives at infinite vectors. These are the common form of pre-numerical representations of continuous systems. Such vectors usually lie in the Hilbert space called l_2. [Gohberg and Goldberg, 1981] The infinite vectors are truncated in the numerical evaluation, where the length of the truncated vector can be varied to test convergence of the result. However, as we shall see in this section, the theory of infinite vectors and the linear operations thereon, in the form of infinite matrices, is not completely trivial.

The simplest class of operators on the infinite space are infinite matrices, which can be infinite in two senses, half-infinite and double-infinite:

$$(2.87) \qquad \begin{pmatrix} a_{00} & a_{01} & \cdots \\ a_{10} & a_{11} & \\ \vdots & \ddots & \ddots \end{pmatrix} \qquad \begin{pmatrix} \ddots & \ddots & \vdots & & \\ \ddots & a_{00} & a_{01} & \cdots \\ \cdots & a_{10} & a_{11} & \ddots \\ & \vdots & \ddots & \ddots \end{pmatrix} \quad.$$

This infinite dimension has a number of unexpected consequences, in the sense that cutting a couple of dimensions of an infinite dimensional space, yields an infinite dimensional space which can be mapped onto the original space.

Considering the half-infinite matrix, the shift operator:

$$(2.88) \qquad S = \begin{pmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & 0 & 1 & \\ & & & \ddots & \ddots \end{pmatrix} \ ,$$

transforms a row $(a_0, a_1, a_2, \cdots)$, into a row $(0, a_0, a_1, \cdots)$. The adjoint operator S^*, does the opposite: $S^*(a_0, a_1, a_2, \cdots) = (a_1, a_2, a_3, \cdots)$. Hence, the kernel of S^* is the space $(a_0, 0, 0, 0, \cdots)$, and the index associated with the operator S is: [Gohberg and Goldberg, 1981]

$$(2.89) \qquad \text{index} S = \dim \ker S - \dim \ker S^* = -1 \ \ .$$

For the reduction to a finite dimensional space it is appropriate to split the spaces such that the infinite dimensional operator has index zero.

For example, the differential ∂_z operator can be written as a shift operator S on the basis:

$$(2.90) \qquad \partial_z \begin{pmatrix} 1 \\ z \\ \frac{1}{2}z^2 \\ \frac{1}{6}z^3 \\ \frac{1}{24}z^4 \\ \vdots \end{pmatrix} = S^* \begin{pmatrix} 1 \\ z \\ \frac{1}{2}z^2 \\ \frac{1}{6}z^3 \\ \frac{1}{24}z^4 \\ \vdots \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ z \\ \frac{1}{2}z^2 \\ \frac{1}{6}z^3 \\ \vdots \end{pmatrix} \ .$$

Note that the S operator on the coefficients vector corresponds to the adjoint operator S^* on the basis elements.

However, if the monomials $z^n/n!$ are a basis of $\mathcal{L}^2([-\pi, \pi])$, the appropriate representation would be the trigonometric series:

$$(2.91) \qquad \partial_z \{\sin(n+1)z, \cos nz\}_{n=0}^\infty = \{(n+1)\cos(n+1)z, -n\sin nz\}_{n=0}^\infty \ \ ,$$

such that the null space $\cos nz = 1$ can be separated from the infinite-dimensional subspace on which the operator ∂_z has an index zero. In the study of boundary conditions, Section 5.5, the index will arise as well.

Another feature of operators on infinite dimensional spaces the more restrictive norm. For example, the interchanging of summations, no problem for finite sums, leads to different answers for the matrix: [Gelbaum and Olmsted, 1964]

$$(2.92) \qquad A = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{4} & \frac{1}{8} & \cdots \\ -\frac{1}{2} & 0 & \frac{1}{4} & \frac{1}{4} & \cdots \\ -\frac{1}{4} & -\frac{1}{2} & 0 & \frac{1}{2} & \cdots \\ -\frac{1}{8} & -\frac{1}{4} & -\frac{1}{2} & 0 & \\ \vdots & \vdots & \vdots & & \ddots \end{pmatrix} \ ,$$

lead to two distinct double sums:

$$(2.93) \qquad \sum_i \left(\sum_j A_{ij} \right) = 2 \neq -2 = \sum_j \left(\sum_i A_{ij} \right) \ \ .$$

As can be seen from the differential operator ∂_z above, many operators associated with partial differential equations (PDE's) are unbounded, i.e., might yield very large results, in some norm $\| \cdot \|$, for the traditional functions and function spaces.

[Gohberg and Goldberg, 1981] For example the function $f(z) = |z|^\alpha$ for $0 < \alpha < 1$ is a well-behaved continuous function of z, but its derivative is singular. A large part of research on PDE's is devoted of constructing function spaces $\mathcal{S}$ appropriate for functions $f(z)$, such that the image $Df(z)$ of the differential operator D still yields well-behaved functions, in the sense $\|f\|_{\mathcal{S}} < \infty \Rightarrow \|Df\|_{\mathcal{D}} < \infty$. Sobolev spaces and Schwartz spaces are examples of such restricted function spaces, with well-behaved images. [Hörmander, 1990]

Quite often, in numerical evaluations and simulations, one has no other choice than to truncate the space, first from continuous to countable, and subsequently to finite. See, for example, Chapter 8. The solution space is thereby regularized. The functions are restricted to some well-behaved set. In nonlinear partial differential equations, discontinuities and singularities may arise, which can no longer be found in the numerical simulation. [Hörmander, 1990] Another problem are spurious solutions, or intruder states. The name arises from the discovery of this phenomena in large matrix calculations in nuclear shell theory. The collective behavior of a number of functions, each separately irregular enough to be discarded from the finite approximation of the system, may yield some smooth and global dynamics. Such states may arise from systems described either by matrices which large off-diagonal coefficients in $\mathbf{A}$, like in Eq. 2.92, or, similar, from systems with small differences between the diagonal coefficients. In the mathematical theory of matrices, such systems are ill-conditioned. [Wilkinson, 1965]

The high density of states, or modes, of the eigenspectrum of systems, which plagues the conditioning of the resulting matrix, is common to all system of two or more spatial dimensions. The density of states in the eigenspectrum for a given frequency grows with the power of dimension: $\sim \omega^{\dim}$. Even for small coefficients in the off-diagonal part of the matrix, like in Eq. 2.92, the sheer number of off-diagonal coefficients can yield a large truncation error. The local form of the partial differential equation, makes a finite difference approximation near diagonal, which has the advantage of a reduced number of off-diagonal coefficients, and therefore better conditioning.

2.12. Krylov bases

From the large matrices of FEM, Chapter 8, or the infinite matrices from an expansion in basis functions, in Section 2.11, there are many different ways to arrive at a small lumped model, which contains the dominant modes. [Antoulas, 2005] In Chapter 9, the physical and modelling aspects of this reduction are discussed. From the numerical perspective, the Krylov bases are an important tool in the reduction. [Freund, 2003]

The Krylov basis $\mathcal{K}_n(\mathbf{A}, \mathbf{X})$ is a basis generated through the successive operations of the operator $\mathbf{A}$ on the vector $\mathbf{x}$, or on the collection of vectors $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_m\}$:

$$(2.94) \qquad \mathcal{K}_n(\mathbf{A}, \mathbf{X}) = \{\mathbf{x}_i, \mathbf{A}\mathbf{x}_i, \mathbf{A}^2\mathbf{x}_i, \cdots, \mathbf{A}^n\mathbf{x}_i\} \ ,$$

which will be denoted by:

$$(2.95) \qquad \mathcal{K}_n(\mathbf{A}, \mathbf{X}) = \{\mathbf{x}_{ji}\}_{j=0,i=1}^{j=n,i=m} \ .$$

It has the main advantage that the operation $\mathbf{A}$ on the basis is near trivial:

$$(2.96) \qquad \mathbf{A}\mathbf{x}_{ij} = \mathbf{x}_{(j+1)i} \ .$$

However, the basis is not orthogonal, except in special cases.

The Krylov basis is closely related to the moment problem for integral operators. An operator or function, for example in L^2 can be expanded into moments:

$$(2.97) \qquad I_i(f,g) = \int \mu(z) dz f(z) z^i g(z) \ \ .$$

The question is how much information of can be recovered of f and g, or f, given g, using only the moments I_i. [Widder, 1941, Akhiezer, 1965] The original question by Stieltjes was whether for $f(z) = g(z) = 1$, the measure $\mu(z) > 0$ could be reconstructed from the sequence I_i. [Stieltjes, 1894-1895]

Many function transformations are based, in some sense, on the moment expansion. These transformations can be inverted in a particular function space. Well known are the class of almost everywhere (a.e.) functions for transformations, like Fourier, Laplace, and Hankel transformations, based on the Lebesgue integral on a compact domain. Such functions are considered identical if they differ only on a discrete set of points.

Krylov space, with a single vector $\mathbf{X} = \mathbf{v}$ is defined as

$$(2.98) \qquad \mathcal{K}_N(\mathbf{A}, \mathbf{v}) = \{\mathbf{v}, \mathbf{A}\mathbf{v}, \mathbf{A}^2\mathbf{v}, \cdots \mathbf{A}^N\mathbf{v}\} \ \ ,$$

which lies at the basis of Lanczos or Arnoldi tridiagonalization. In the case of a symmetric matrix, $\mathbf{A}^T = \mathbf{A}$, it can lie at the basis of tridiagonalization. The corresponding algorithm picks a vector $\mathbf{v}_0$, applies it to the matrix $\mathbf{A}$, and projects back on $\mathbf{v}_0$, the remainder is a new vector $\mathbf{v}_1$, which applied to $\mathbf{A}$, yields three parts: parallel to $\mathbf{v}_0$, and to $\mathbf{v}_1$ and perpendicular to both vectors: $\mathbf{v}_2$. After the next step, we find that $\mathbf{A}\mathbf{v}_2$ is perpendicular to $\mathbf{v}_0$, in the case of a symmetric matrix.

$$(2.99) \qquad \mathbf{v}_1 = \mathbf{A}\mathbf{v}_0 - \frac{\mathbf{v}_0^T \mathbf{A}\mathbf{v}_0}{\|\mathbf{v}_0\|^2}\mathbf{v}_0 \ \ ,$$

such that $\mathbf{v}_0 \perp \mathbf{v}_1$. Similarly for $\mathbf{v}_2$, we subtract the projections onto $\mathbf{v}_0$ and $\mathbf{v}_1$ from $\mathbf{A}\mathbf{v}_1$:

$$(2.100) \qquad \mathbf{v}_2 = \mathbf{A}\mathbf{v}_1 - \frac{\mathbf{v}_1^T \mathbf{A}\mathbf{v}_1}{\|\mathbf{v}_1\|^2}\mathbf{v}_1 - \frac{\mathbf{v}_0^T \mathbf{A}\mathbf{v}_1}{\|\mathbf{v}_0\|^2}\mathbf{v}_0 \ \ .$$

From a lengthy algebra it can be determined that $\mathbf{v}_0\mathbf{A}\mathbf{v}_2 = 0$. However, since $\mathbf{v}_2$ is perpendicular to both $\mathbf{v}_0$ and $\mathbf{v}_1$, by construction, it follows for a symmetric matrix that:

$$(2.101) \qquad \mathbf{v}_0^T \mathbf{A}\mathbf{v}_2 = \mathbf{v}_2^T \mathbf{A}\mathbf{v}_0 = \mathbf{v}_2^T \mathbf{v}_1 + \frac{\mathbf{v}_0^T \mathbf{A}\mathbf{v}_0}{\|\mathbf{v}_0\|^2}\mathbf{v}_2^T \mathbf{v}_0 = 0 \ \ .$$

which is zero by construction, since $\mathbf{v}_2^T \mathbf{v}_1 = \mathbf{v}_2^T \mathbf{v}_0 = 0$. Hence the tridiagonal form of $\mathbf{A}$. The vectors $\mathbf{V} = \{\mathbf{v}_0, \mathbf{v}_1, \cdots \mathbf{v}_n\}$ form an orthogonal basis for the Krylov space $\mathcal{K}_n(\mathbf{A}, \mathbf{v})$, for any n. The conjugate gradient method, described in Section 10.3, uses the same principles. In that case the inner-product is $\mathbf{v}_i^T \mathbf{A}\mathbf{v}_j = 0$ for $i \neq j$. The actual matrix under investigation, in that case, is $\mathbf{A}^{-1}$. [Stoer and Bulirsch, 1980]

In a given basis $\{\mathbf{v}_0, \mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n\}$ the matrix $\mathbf{A}$ is represented as:

$$(2.102) \qquad A_{ij}^{\mathbf{v}} = \mathbf{v}_i^T \mathbf{A}\mathbf{v}_j \ \ .$$

The matrix constructed by successive orthogonalization has a symmetric tridiagonal form:

$$(2.103) \qquad \mathbf{V}^T \mathbf{A} \mathbf{V} = \mathbf{A}^{\mathbf{v}} = \begin{pmatrix} \alpha_0 & \beta_0 & 0 & & \\ \beta_0 & \alpha_1 & \beta_1 & 0 & \\ 0 & \beta_1 & \alpha_2 & \beta_2 & 0 \\ & & \ddots & \ddots & \ddots \end{pmatrix} \; .$$

This is the Lanczos tridiagonal form of a symmetric matrix.

2.13. Some remarks on state space and spatial reference space

In this chapter we were concerned mainly with the variables, the states, of the dynamical model. The properties of continuous, or infinite-dimensional, systems are also determined by the properties of the underlying, or reference, space. The two are linked through differential geometry and partial differential equations, which is discussed in the Chapters 4 and 5. The existence of a state for a given local energy density is, in the language of differential geometry called the integrability. The state of a conserved current is a central concept from differential geometry. Furthermore, the state variables of a component in this chapter are, as yet, not invariant under coordinate transformations, which is also discussed the Chapter 4.

The dynamics of the state space is discussed in more general terms in the following chapter. Were the models of dynamical systems in this chapter mainly isolated, in the following chapter the input and output will play a central role. Furthermore, we will take a closer look at the simulation, or time-integration, of dynamics, i.e., the evolution in time.

CHAPTER 3

Evolution

> The sciences do not try to explain, they hardly ever try to inter-
> pret, they mainly make models.
> [J. von Neumann]

Linear models of dynamical systems, with input and output, are often brought
to a canonical form, before numerical analysis and simulation. In system engineer-
ing, this canonical form is often the state-space model, or input-state-output model,
or the $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$-model. [Polderman and Willems, 1998, Curtain and Zwart, 1995,
Nijmeijer and van der Schaft, 1990] Physical dynamical systems are only a subset
of all systems that can be brought in the input-state-output form. In the more
mathematical literature, the output is deemed less relevant, and the same model is
referred to as the model for system evolution, i.e., the motion in time. [Hirsch and
Smale, 1974]

The states of input-state-output model are the necessary internal degrees of
freedom, which can be the charge on a capacitor in an electric system, or the position
of a mass in a mechanical system. The state is often seen as the configuration,
independent of any input. For example, the input force may cause an acceleration,
but does not determine the start position or velocity. These are states, however,
only if the position or velocity is relevant for the dynamics. Some parts of the
configuration may not be relevant for the dynamics. For example, in the language
of Einstein's relativity, experiments carried out on a moving train will have the
same outcome, as experiments carried out on solid ground. Hence both the global
position and the global velocity of the experimental setup are irrelevant, and they
not required as states.

The generic input-state-output system we will call an evolution system. The
dynamical systems of the previous chapter are a subclass of the broader set of evo-
lution systems. However, the absence of certain physical principles, such as energy
and momentum conservation, will lead to poorly defined models, which might de-
scribe certain behavior well, but it yields less insight in the internal dynamics of a
system. For example, dissipation is an integral part of such models. Prying away
the dissipation part from the oscillatory part is often done a posteriori in order
to be able to analyze stability.[Willems, 1972a] Furthermore, the physical basis of,
for example, dissipation is usually absent as $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$ models do not start with
a detailed analysis of the system at hand, but are postulated and later adjusted
to fit the dynamics. In the extreme case of this limited insight in the model, the
model is treated as a black box, where the number of states and the parameters of
the input-state-output model is fitted to experiments. This process is called system
identification. [Ljung, 1999] The identification is usually with a linear $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$-
model, since nonlinearities will make identification much more difficult.

3.1. Input-state-output, or $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$, model

A general, but linear, model is usually called cryptically the $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$ model. [Brockett, 1970] The matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$ represent the model with the state vector $\mathbf{x}$ and the input and output $\mathbf{u}$ and $\mathbf{y}$:

$$(3.1) \qquad\qquad\qquad \dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} \quad,$$

with the output:

$$(3.2) \qquad\qquad\qquad \mathbf{y} = \mathbf{C}\mathbf{x} + \mathbf{D}\mathbf{u} \quad,$$

where $\mathbf{D}$ is the direct input-to-output part. The matrix $\mathbf{A}$ determines the evolution of the homogeneous system. The input matrix is $\mathbf{B}$ with the input vector $\mathbf{u}$, while the output matrix is $\mathbf{C}$. One essential ingredient of dynamical systems is retained, which is the invariance under time-translation. If an experiment is done today or tomorrow, under invariant conditions, the outcome should be the same, hence $\mathbf{A}$ does not depend on the time t. Hence, we can still use $t = 0$ as the initial time at which the system has to be characterized for a unique evolution. For a set of first order equations, such as the $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$ model, only $\mathbf{x}(0)$ needs to be given.

Linearity is a severe restriction on a dynamical system. In many cases it can be solved, yielding simply the time-dependence of the state $\mathbf{x}(t)$. In the case of numerical solutions of very large systems it is usually the only option. Nonlinearity makes it very hard to solve such systems. Furthermore, it is difficult to select the right solution, from the multiple solutions that arise from nonlinearities. Later, once physical principles are introduced, more may be said about nonlinearities, however, for many numerical implementations the system is analyzed in a small neighborhood around the state $\mathbf{x}(t)$ under consideration, such that one still resorts to an appropriately linearized $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$ form, i.e., the local decomposition. [Isodori, 1985]

The state $\mathbf{x}(t)$, in Eq. 3.1, given the initial conditions $\mathbf{x}(0)$, is easily integrated:

$$(3.3) \qquad\qquad \mathbf{x}(t) = e^{\mathbf{A}t}\left(\int_0^t e^{-\mathbf{A}\tau}\mathbf{B}\mathbf{u}(\tau)d\tau + \mathbf{x}(0)\right) \quad.$$

However, in practice, if $\mathbf{A}$ is a large, or infinite, matrix or some abstract operator, it might be difficult to recover $\mathbf{x}(t)$ from the equation above. Numerical integration methods, described in Section 3.4, are often used to determine the evolution of $\mathbf{x}(t)$.

In terms of the control vector $\mathbf{u}$ several operators can be distinguished. The steady-state operator $\mathbf{A}^{-1}\mathbf{B}$ yields, if it exists, the state corresponding to a stationary input:

$$(3.4) \qquad\qquad\qquad \mathbf{x} = \mathbf{A}^{-1}\mathbf{B}\mathbf{u} \quad.$$

In the more general case of time-dependent control, using the space $\mathcal{L}^2([0,t])$ the end-point adjustment operator yields the optimal control $\mathbf{u}(t')$ for the change of the free evolution of the state: $\mathbf{x}_{\text{free}}(t) = e^{\mathbf{A}t}\mathbf{x}(0)$ to the desired end-point $\mathbf{x}(t)$:

$$(3.5) \qquad\qquad u(t') = \mathbf{B}^T e^{\mathbf{A}^T(t-t')}W^{-1}(t)((\mathbf{x}(t) - e^{\mathbf{A}t}\mathbf{x}(0)) \quad,$$

where the controllability Gramian $\mathbf{W}(t)$ is given by: [Skogestad and Postlethwaite, 1996]

$$(3.6) \qquad\qquad \mathbf{W}(t) = \int_0^t e^{\mathbf{A}(t-t')}\mathbf{B}\mathbf{B}^T e^{\mathbf{A}^T(t-t')}dt' \quad.$$

The input is restricted to the same time-dependence as the states themselves, which made the self-dual formulation possible.

Another common operator associated with the state space representation is the Laplace transformation [Widder, 1941] of the system evolution:

$$(3.7) \quad \mathbf{G}(s) = \mathcal{L}[\mathbf{C}e^{\mathbf{A}t}\mathbf{B}] + \mathbf{D} = \int_0^\infty e^{-st}\mathbf{C}e^{\mathbf{A}t}\mathbf{B}dt + \mathbf{D} = \mathbf{C}(s\mathbf{1} - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D} \quad .$$

In practical applications of control engineering, the right-hand side is commonly referred to as the transfer matrix from input $\mathbf{u}$ to output $\mathbf{y}$. [Ogata, 1990]

Although, the dynamical systems under consideration are more general in the sense that nonlinearities occur, and more restrictive, in the sense that the matrix $\mathbf{A}$ must be Hamiltonian, such that the energy is conserved, it is useful to keep the $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$ description in mind. In a small neighborhood of a particular state $\mathbf{x}$, the system can be approximated by a $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$. The state space can therefore be characterized by properties of the matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$, rather than by "dynamics" in the abstract sense. Regions can be indicated by the ranks of the observability and controllabity matrices, as the typical invariants of the linear system.

Occasionally the matrix $\mathbf{E}$ is introduced:

$$(3.8) \qquad\qquad\qquad \mathbf{E}\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} \quad .$$

In the case the matrix $\mathbf{E}$ is singular, every independent vector $\mathbf{g}$ in the left null-space of $\mathbf{E}$ would corresponds to a constraint $\mathbf{g}^T\mathbf{A}\mathbf{x} + \mathbf{g}^T\mathbf{B}\mathbf{u} = 0$ on the state vector $\mathbf{x}$. Since the system is linear, it should be no problem to solve the constraint explicitly, and use a lower dimensional state. Conceptually, $\mathbf{E}$ is unwanted too. A state $\mathbf{x}$ should define the system. Every component of the state is an initial condition required for the evolution. With a singular $\mathbf{E}$, initial conditions $\mathbf{x}(0)$ with $\mathbf{g}^T\mathbf{A}\mathbf{x}(0) + \mathbf{g}^T\mathbf{B}\mathbf{u} \neq 0$ will not lead to a well-defined evolution. Initial conditions for partial differential equations are further discussed in Section 5.3.

On the other hand, the singular $\mathbf{E}$ or $\mathbf{A}$ will arise from the internal dynamics of distributed, or continuous systems, described by partial differential equations. The null-space corresponds to degrees of freedom not fixed by the internal dynamics. In the case of a discrete approximation of such systems it is important to retain the null-space. Depending on the view the left null-space of $\mathbf{A}$ corresponds to the constants of motion, such as the momentum of an isolated system or the conserved charge, while the right null-space corresponds to boundary conditions, which act as constraints on the internal dynamics. Hence the interaction of the system with its environment determines the view, and it may not always be proper to reduce the dimension such that $\mathbf{A}$ or $\mathbf{E}$ is nonsingular, if the setting is not specified. In the port-Hamiltonian approach, discussed in Chapter 7, the null-space is isolated but retained, such that, depending on the interaction of the system, the variables in the null-space can act as input, i.e., constraints, or output, i.e., constants of motion, or a more complicated interactive and dynamical response, which we call port-variables.

Linear systems can be brought to the $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$ form. Most physical, mechanical dynamical systems are in their native form described by the Euler-Lagrange equations, and are consequently written as second-order equations. They depend on the the position $\mathbf{q}$ and the velocity $\dot{\mathbf{q}}$, to determine the evolution in terms of the acceleration $\ddot{\mathbf{q}}$. To turn the second-order equations into a set of first-order equations the velocity is used as an independent variable. In the simplest approach $\mathbf{v} = \dot{\mathbf{q}}$,

however, other choices of independent variables exist, such as the momentum $p(\dot{\mathbf{q}})$. The state vector is therefore a combination of position and a function of velocity $\mathbf{x} = (\mathbf{q}, f(\dot{\mathbf{q}}))$.

In mechanical systems $\mathbf{E}$ usually incorporates the mass of the system. However, the Hamiltonian formulation, or the Lancaster scheme, makes it possible to bring the mass to the left-hand side. The Lancaster scheme turns the Newton equation:

$$(3.9) \qquad \mathbf{M}\ddot{\mathbf{q}} + \mathbf{D}\dot{\mathbf{q}} + \mathbf{K}\mathbf{q} = \mathbf{F} \quad,$$

into a first-order equations with $\mathbf{x} = (\mathbf{q}, \dot{\mathbf{q}})$:

$$(3.10) \qquad \begin{pmatrix} \mathbf{D} & \mathbf{M} \\ \mathbf{M} & 0 \end{pmatrix} \dot{\mathbf{x}} + \begin{pmatrix} \mathbf{K} & 0 \\ 0 & -\mathbf{M} \end{pmatrix} \mathbf{x} = \begin{pmatrix} \mathbf{F} \\ 0 \end{pmatrix} \quad.$$

In this case the matrix $\mathbf{E}$ is:

$$(3.11) \qquad \mathbf{E} = \begin{pmatrix} \mathbf{D} & \mathbf{M} \\ \mathbf{M} & 0 \end{pmatrix} \quad.$$

To bring the system into the $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$ form, the matrix $\mathbf{E}$ in front of $\dot{\mathbf{x}}$ must be inverted, which is possible as long as the damping $\mathbf{D}$ is much smaller than the mass $\mathbf{M}$. The Lancaster scheme $\mathbf{E}$ matrix is a special case of the $\mathbf{E}$ matrix that arises in constrained minimization, encountered later.

The Hamilton equations are a similar reduction of the second-order equations to a coupled set of first-order equations, however, in a more systematic manner, such that it can be generalized to nonlinear systems. See also Chapter 2. The Hamilton system is automatically the core of the $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$ system:

$$(3.12) \qquad \begin{pmatrix} \dot{\mathbf{p}} \\ \dot{\mathbf{q}} \end{pmatrix} = \dot{\mathbf{x}} = \begin{pmatrix} 0 & -\mathbf{K} \\ \mathbf{M}^{-1} & 0 \end{pmatrix} \mathbf{x} + \begin{pmatrix} \mathbf{F} \\ 0 \end{pmatrix} \quad.$$

In this case the augmented $\mathbf{x}$ vector does not contain the velocity, but the momentum $\mathbf{p} = \mathbf{M}\dot{\mathbf{q}}$.

Hamilton formulated these equations of motion based on the conserved energy, or Hamiltonian. The contribution of Jacobi has been to remove superfluous variables from these equations using coordinate transformations and generalized coordinates. A similar procedure was carried out by Lagrange to arrive at generalized coordinates for Newtonian mechanics. The names of Hamilton and Jacobi are also connected to the equation of the generating functional S for the canonical coordinate transformation: [Goldstein, 1980]

$$(3.13) \qquad H(\mathbf{q}, \nabla_{\mathbf{q}} S, t) + \partial_t S = 0 \quad.$$

The conserved energy in the Hamiltonian dynamics will lead to a natural definition of the output, corresponding to the input force $\mathbf{F}$. In the absence of an applied force, the Hamiltonian H is constant in time:

$$(3.14) \qquad H = \frac{1}{2}\mathbf{p}^T \mathbf{M}^{-1} \mathbf{p} + \frac{1}{2}\mathbf{q}^T \mathbf{K}\mathbf{q} \quad.$$

In the case of an applied force $\mathbf{F}$, the energy varies. The equations of motion can be used to express the change in energy in terms of the state variables $\mathbf{q}$ and $\mathbf{p}$:
$$(3.15)$$
$$\dot{H} = \nabla_{\mathbf{q}} H \dot{\mathbf{q}} + \nabla_{\mathbf{p}} H \dot{\mathbf{p}} = \mathbf{q}^T \mathbf{K} \mathbf{M}^{-1} \mathbf{p} - \mathbf{q}^T \mathbf{K} \mathbf{M}^{-1} \mathbf{p} + \mathbf{p}^T \mathbf{M}^{-1} \mathbf{F} = \mathbf{p}^T \mathbf{M}^{-1} \mathbf{F} = \dot{\mathbf{q}} \mathbf{F} \quad,$$

where is used that both the stiffness matrix $\mathbf{K} = \mathbf{K}^T$ and the mass matrix $\mathbf{M} = \mathbf{M}^T$ are symmetric. Hence the change of energy, or power, supplied to the system

corresponds to the product of the applied force and the velocity. The natural output matrix $\mathbf{C}$ is therefore:

$$(3.16) \qquad \mathbf{C} = \begin{pmatrix} \mathbf{0} \\ \mathbf{M}^{-1} \end{pmatrix} \quad .$$

The spectral method is a common approach to simplify the evolution of a system. It consists of decomposing the system into independent modes, through diagonalization of the matrix $\mathbf{A}$:

$$(3.17) \qquad \mathbf{A} = \mathbf{V}^\dagger \Omega \mathbf{U} \quad ,$$

where, $\mathbf{V}$ and $\mathbf{U}$ are respectively the left and right eigenvectors and Ω the diagonal matrix consisting of the eigenfrequencies. The dagger $\dagger$ represents the Hermitian conjugate of the matrix: $\mathbf{A}^\dagger = \overline{\mathbf{A}^T}$. For example, the 2×2 symplectic matrix has complex eigenvectors $\mathbf{U}$ and eigenvalues:

$$(3.18) \qquad \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ i & -i \end{pmatrix} \cdot \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix} \cdot \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -i \\ 1 & i \end{pmatrix} \quad .$$

In the absence of degenerate eigenfrequencies the normalized left and the right eigenvectors are trivially related:

$$(3.19) \qquad \mathbf{V}^\dagger = \mathbf{U}^{-1} \quad .$$

Furthermore, if $\mathbf{A}$ is symmetric, the matrix $\mathbf{U}$ is orthonormal, representing a rotation. The exponentiation of a symmetric matrix $\mathbf{A}$ boils down to the exponentiation of the eigenvalues of Ω:

$$(3.20) \qquad e^{\mathbf{A}t} = \mathbf{V}^\dagger e^{\Omega t} \mathbf{U} \quad .$$

The eigenvalues ω for stable dynamical systems lie in the complex half plane $\mathrm{Re}\,\omega \leq 0$, while for an energy conserving system, they consist of pairs $\omega = \pm i\lambda$.

This spectral composition does not always exist, which is a common problem for analysis of linear systems. The general matrix $\mathbf{A}$ can be diagonalized up to Jordan blocks: [Wilkinson, 1965]

$$(3.21) \qquad \mathbf{A} = \begin{pmatrix} \mathbf{J}_1 & & \\ & \ddots & \\ & & \mathbf{J}_n \end{pmatrix} ,$$

with a single block

$$(3.22) \qquad \mathbf{J}_i = \begin{pmatrix} \lambda & 1 & 0 & 0 \\ 0 & \lambda & 1 & 0 \\ 0 & 0 & \lambda & 1 \\ 0 & 0 & 0 & \lambda \end{pmatrix} ,$$

where the block can be of arbitrary length. Instead of an evolution given by the sum of exponentials, the evolution of the Jordan block evolution matrix $\mathbf{J}_i$ will be given by a polynomial $P(t)$ times the exponential:

$$(3.23) \qquad \mathbf{x}(t) = \mathbf{V}_i P(t) e^{\lambda t} \mathbf{U}_i^\dagger \mathbf{x}(0) \quad ,$$

where $\mathbf{V}_i$ and $\mathbf{U}_i$ correspond to the projection on the eigenvectors of Jordan block $\mathbf{J}_i$. For example, the four-dimensional Jordan-block arises from the differentiation of the polynomial vector $\{t^i e^{\lambda t}\}_{i=0}^{i=3}$:

$$(3.24) \qquad \frac{\partial}{\partial t} \begin{pmatrix} 1 \\ t \\ \frac{1}{2}t^2 \\ \frac{1}{6}t^3 \end{pmatrix} e^{\lambda t} = \begin{pmatrix} \lambda & 1 & 0 & 0 \\ 0 & \lambda & 1 & 0 \\ 0 & 0 & \lambda & 1 \\ 0 & 0 & 0 & \lambda \end{pmatrix} \begin{pmatrix} 1 \\ t \\ \frac{1}{2}t^2 \\ \frac{1}{6}t^3 \end{pmatrix} e^{\lambda t} \quad .$$

The exponentiation of a Jordan block, in order to explicitly solve the time evolution of an input-state-output system, is less trivial than the exponentiation of the diagonal matrix, Eq. 3.20. The exponential of $\mathbf{J} \in \mathbb{R}^{5 \times 5}$ is given by:

$$(3.25) \qquad e^{\mathbf{J}} = \begin{pmatrix} e^{\lambda} & \lambda e^{\lambda} & \frac{\lambda^2}{2}e^{\lambda} & \frac{\lambda^3}{6}e^{\lambda} \\ 0 & e^{\lambda} & \lambda e^{\lambda} & \frac{\lambda^2}{2}e^{\lambda} \\ 0 & 0 & e^{\lambda} & \lambda e^{\lambda} \\ 0 & 0 & 0 & e^{\lambda} \end{pmatrix} \quad .$$

The typical off-diagonal coefficient, indexed (i, j), is given by: $\lambda^{j-i} e^{\lambda}/(j-i)!$.

In the case of infinite-dimensional systems, such a decomposition is not possible. There might not be a finite set of eigenvalues of the operator A, although in many cases, given certain conditions of A, certain conclusions might be drawn. For the analysis of infinite-dimensional systems the analyticity of the resolvent R_A is often examined:

$$(3.26) \qquad R_A = \oint_{\Gamma} \frac{1}{z-A} dz A \quad ,$$

which appears in all kind of scalar forms, such as in the Gramian above, Eq. 3.6. The complex variable z is more general than the variable of the Laplace transform, Eq. 3.7, since particular singularities of the spectrum of A can be singled out by an integration contour around them. See also Section 5.7. The Cauchy theorem can be applied, as often is for the inverse Laplace transform, for a closed contour Γ:

$$(3.27) \qquad W_{A,B} = \oint_{\Gamma} B^T \frac{1}{z-A} B dz \quad .$$

If a singularity is encircled, or a cut is crossed, by the contour Γ, the resolvent yields the spectral properties of these singular points. [Kato, 1966]

3.2. Continuous spaces and the $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$-system

In the later chapters continuous systems described by partial differential equations are discussed. The state space $\mathbf{x}$ and the operators $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{D}$ are no longer finite vectors in $\mathbb{R}^n$ and matrices in $\mathbb{R}^{m \times n}$. Instead the state space is a function space, with the operators acting on them. [Riesz and Nagy, 1955] Eventually, for numerical evaluation using the methods of linear algebra, the model must be reduced to, or approximated by, a simple form such as the $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$-form.

The partial differential equations have initial conditions and boundary conditions. See Section 5.1. In the finite-dimensional state-space representation, the initial conditions correspond to the state space $\mathbf{x}$, for which an initial state $\mathbf{x}(t = 0)$ is required. The boundary conditions become the input matrix $\mathbf{B}$ and the input-to-output matrix $\mathbf{D}$, where the actual value of $\mathbf{u}$ is the finite-dimensional representation of the boundary value. In some cases there might be a part of the state vector

associated with an input. For example, a force acting on an inertia, chances the velocity and the position, which are the two states associated with a force input. For consistency these states should be brought into the core model; the model directly, algebraically, related to the boundary conditions. See Section 7.6. Depending on the actual choice of boundary conditions; the causality, these states are either algebraically related to the boundary conditions, or state variables of integrating or buffering elements, consequences of the boundary conditions.

The $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$-form of the core model of the finite approximation of a partial differential equation is therefore not a priori fixed. It depends on the type of boundary condition is used. It is the analogue of a lumped mass, which can be controlled either by single position, or velocity, or more naturally by force control. The number of states are respectively: zero; one; or two, i.e., none; the position; or the position and the momentum. For the lumped case, position control is rather unnatural. The continuous system static analogues are for position the Dirichlet boundary condition and for force Neumann boundary condition, which are both quite natural. The velocity control does not correspond to any static situation, it lies in the more complicated domain of hyperbolic system, discussed in Section 5.3. The aspect of the number of states, or equivalently the constraints on states, of continuous systems was already raised in the Introduction, in Section 1.7. In the port-based modelling [Paynter, 1961, Breedveld, 1984, Karnopp et al., 2000] it is more natural to retain these states and associate them with dynamical states, with initial conditions, or algebraically relate them to input, depending on the situation.

3.3. Observability and controllability

The observability and controllabity criteria are the rank conditions on: [Kalman et al., 1969]

$$(3.28) \qquad n = \mathrm{rank}\left[\mathbf{C}^T, \mathbf{A}^T\mathbf{C}^T, \mathbf{A}^{T2}\mathbf{C}^T, \cdots \mathbf{A}^{T(n-1)}\mathbf{C}^T\right] \quad,$$

$$(3.29) \qquad n = \mathrm{rank}\left[\mathbf{B}, \mathbf{AB}, \mathbf{A}^2\mathbf{B}, \cdots \mathbf{A}^{n-1}\mathbf{B}\right] \quad,$$

where n is the dimension of the state space.

The conditions imply that the projections, through the left and right eigenvectors, of the input matrix $\mathbf{B}$ and the output vectors $\mathbf{C}$, should not be perpendicular to any of the Jordan blocks, for the controllability and observability.

The dynamics of a block is given by an arbitrary polynomial times the leading exponential $P(t)e^{-\lambda t}$, Eq. 3.22. The coefficients can be determined by the same number of time-derivatives at a time t, as the order of the polynomial. Hence, the size of the Jordan blocks determines the order of the differential equation.

The observability and controllability conditions are also closely related to the Krylov expansion. In a sense, for a finite system, the Krylov expansion, using $\mathbf{A}$ and the input, or start, matrix $\mathbf{B}$, will only span the controllable subspace of the full dynamics. If the full dynamical subspace is much smaller that the full state space, one can use it to one's advantage to reduce the system. These principles are explored in Chapter 9.

3.4. Integration

Simulation boils down to the time-integration of dynamical equations. For numerical integration many schemes have been developed. From the known schemes we will treat only three very briefly, afterwards we discuss more general problems and approaches to integration.

For Hamiltonian systems, integration methods which conserve energy, i.e., geometric integration, [Hairer et al., 2002] also guarantee stability, as the energy norm of the system serves as the measure of stability. Apart from general integration methods, we will discuss mainly the use of energy conservation to improve the numerical integration schemes. [Krenk, 2006]

For time-integration a differential equation is mapped to a set of coupled first-order equations:

$$\dot{\mathbf{x}}(t) = f(\mathbf{x}(t)) \quad , \tag{3.30}$$

which immediately gives rise to the Euler integration scheme, nowadays known as the forward Euler scheme:

$$\mathbf{x}(t + \delta) = \mathbf{x}(t) + \delta f(\mathbf{x}(t)) \quad \text{forward} \quad . \tag{3.31}$$

The backward Euler scheme is implicit, in the sense that it uses the function evaluation at the forward time: $f(\mathbf{x}(t + \delta))$:

$$\mathbf{x}(t + \delta) = \mathbf{x}(t) + \delta f(\mathbf{x}(t + \delta)) \quad \text{backward} \quad . \tag{3.32}$$

The problem of the implicit scheme is that is requires knowledge of $\mathbf{x}$ at $t + \delta$ in order to determine $\mathbf{x}$ at that point in time.

The error of an integration scheme depends on the application. Say $f(\mathbf{x})$ is a constant vector, independent of $\mathbf{x}$, the result is exact for any step size δ. However, typically one would compare the error to that of a Taylor expansion:

$$\mathbf{x}(t + \delta) = \mathbf{x}(t) + \mathbf{x}_1 \delta + \mathbf{x}_2 \delta^2 + \cdots . \tag{3.33}$$

Hence, both the Euler integration schemes yield a δ^2 error at each step.

The most used, and best-known, integration scheme is the Runge-Kutta [Stuart and Humphries, 1998, Press et al., 1992] fourth-order integration. It requires a number of function evaluations, which map out the integral over the finite domain $[t, t + \delta]$, which is integrated as accurate as possible:

$$\mathbf{x}(t + \delta) = \mathbf{x}(t) + \frac{\delta}{6} \left(f(\mathbf{x}(t)) + 2f(\mathbf{x}(t) + \Delta_1) + 2f(\mathbf{x}(t) + \Delta_2) + f(\mathbf{x}(t) + \Delta_3) \right) \quad , \tag{3.34}$$

where the predictions of the state at intermediate times are generated recursively:

$$\Delta_1 = \frac{\delta}{2} f\left(\mathbf{x}(t) + \frac{\delta}{2} f(\mathbf{x}(t)) \right) \quad , \tag{3.35}$$

$$\Delta_2 = \frac{\delta}{2} f\left(\mathbf{x}(t) + \frac{\delta}{2} f(\mathbf{x}(t) + \Delta_1) \right) \quad , \tag{3.36}$$

with the final step:

$$\Delta_3 = \delta f\left(\mathbf{x}(t) + \delta f(\mathbf{x}(t) + 2\Delta_2) \right) \quad . \tag{3.37}$$

Dynamical systems do not explicitly depend upon time, since they are usually time-invariant. Therefore we have ignored the typical explicit time-dependence which appears in many ordinary differential equations and their integration schemes. The
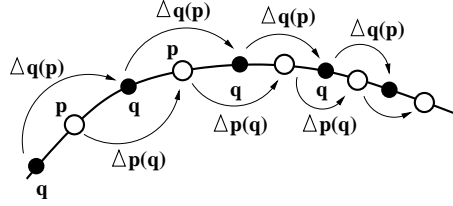
FIGURE 3.1. The leap-frog integration uses successively the values
of the momentum $\mathbf{p}$ to determine the change in position $\mathbf{q}$, and the
new position $\mathbf{q}$ to determine the change in the momentum $\mathbf{p}$.

fourth-order Runge-Kutta integration requires only four function evaluations to
determine both the intermediate steps Δ_i and the values f at these positions, while
higher orders require more function evaluations than the particular order.

Besides basic analytic principles to improve the accuracy of the integration,
there are other concepts which are used in time-integration. For the integration
of the Hamilton equations of motion, which preserve the energy of the system,
symplectic integration schemes has been developed to preserve the energy as well
as possible. One simple, but powerful method is the leap-frog integration. The
symplectic system consists of pairs of canonical variables: positions and momenta
$\mathbf{q}$ and $\mathbf{p}$, or magnetic and electric variables. The leap-frog method consists of
alternating Euler integrations of each set with the new values of the other set:

$$(3.38) \qquad \mathbf{p}(t + \frac{\delta}{2}) = \mathbf{p}(t - \frac{\delta}{2}) - \delta h_q(\mathbf{q}(t), \mathbf{p}(t - \frac{\delta}{2})) \ ,$$

$$(3.39) \qquad \mathbf{q}(t + \delta) = \mathbf{q}(t) + \delta h_p(\mathbf{q}(t), \mathbf{p}(t + \frac{\delta}{2})) \ .$$

The staggered values $t \pm \frac{\delta}{2}$ for the momenta $\mathbf{p}$ are to make apparent the effect
of the alternated integration. One integration uses the middle value of the other
integration. See Figure 3.1.

The leap-frog integration is a large improvement upon the standard Euler in-
tegration for Hamiltonian systems, since it uses, in some variance for the implicit
form, a midpoint scheme, also used for symplectic systems: [Hairer et al., 2002]

$$(3.40) \qquad \mathbf{x}(t + \delta) = \mathbf{x}(t) + \delta f \left(\frac{\mathbf{x}(t + \delta) + \mathbf{x}(t)}{2} \right) \ ,$$

which corresponds to the use of a tangential direction $\mathbf{x}(t + \delta) - \mathbf{x}(t)$ at the point
of evaluation $\frac{\mathbf{x}(t+\delta)+\mathbf{x}(t)}{2}$.

The leap-frog integration uses the additional information known about the sys-
tem. In many cases of higher dimension more information is known. One could, for
example alternate a number of leap-frog steps with a single corrector step which
corrects the state $\mathbf{x}_{\text{old}}$ such that the energy E corresponds again to the starting
energy E_0:

$$(3.41) \qquad \mathbf{x}_{\text{new}} = \mathbf{x}_{\text{old}} - \frac{(E - E_0)\nabla_{\mathbf{x}} H(\mathbf{x}_{\text{old}})}{|\nabla_{\mathbf{x}} H(\mathbf{x}_{\text{old}})|^2} \ .$$

This would be the smallest correction step $\mathbf{x}_{\text{new}} - \mathbf{x}_{\text{old}}$ due to the steepest descent
$\nabla_{\mathbf{x}} H$ direction. Although, no formal justification for such a correction step exists,

the results in the section below, on second-order Hamiltonian integration will lead to the similar results.

3.5. Second-order Hamiltonian integration

For very large systems determining higher-order derivatives is often prohibitively expensive. In the chapter on finite element methods we will discuss iterative approaches for sparse matrices. In many cases it is computationally favorable to restrict the integration to a second-order method. Furthermore, in the case of FEM, the time-evolution does not only depend on the history of a single node $x_i(t)$, but also, and maybe even more so, on the history of the neighboring nodes $x_{i-1}(t)$ and $x_{i+1}(t)$. All the nice properties, which hold for one-dimensional results, i.e., a single function $y(t)$, do not hold for high-dimensional dynamical systems. Generally, the behavior of a component $x_i(t)$, treated as a single function $y(t)$, depends more on the rest of the state vector $\mathbf{x}$, than on the history $t' < t$ of that particular component. In the one-dimensional case, on the contrary, all depends of $x_i(t')$ for $t' < t$.

The integration of Hamiltonian systems is given by the Hamiltonian H, and the symplectic matrix J:

$$(3.42) \qquad\qquad \dot{\mathbf{x}} = J\nabla_{\mathbf{x}}H \quad,$$

we can integrate these equations of a short time δ, assuming $J\nabla_{\mathbf{x}}H$ is constant for the short time-space $[t, t+\delta]$:

$$(3.43) \qquad \mathbf{x}(t+\delta) - \mathbf{x}(t) = \int_t^{t+\delta} J\nabla_{\mathbf{x}}H\,dt \approx \delta J\nabla_{\mathbf{x}}H(\mathbf{x}(t)) \quad.$$

However, we can improve the result by using a Taylor expansion of the variational derivative around $\mathbf{x}(t)$, i.e., by using the equations of motion to approximate the change in $J\nabla_{\mathbf{x}}H(\mathbf{x}(t))$:

$$\int_t^{t+\delta} (J\nabla_{\mathbf{x}}H)\,dt' \approx \delta\,(J\nabla_{\mathbf{x}}H) + J\nabla_{\mathbf{x}}^2 H\dot{\mathbf{x}}\int_t^{t+\delta}(t-t')dt'$$

$$(3.44) \qquad\qquad = \delta J\nabla_{\mathbf{x}}H + \frac{\delta^2}{2}J\nabla_{\mathbf{x}}^2 H J\nabla_{\mathbf{x}}H \quad.$$

Hence, for a finite time the Euler integration $\mathbf{x}(t+\delta) = \delta J_\delta \nabla H(\mathbf{x}(t)) + \mathbf{x}(t)$, the symplectic structure is adjusted, to first order by:

$$(3.45) \qquad J_\delta = J + \frac{\delta}{2}J\nabla_x^2 H J = \begin{pmatrix} -\frac{\delta}{2}\mathbf{M}^{-1} & I \\ -I & -\frac{\delta}{2}\mathbf{K} \end{pmatrix} \quad,$$

where $\mathbf{M}$ and $\mathbf{K}$ are the linearized mass and stiffness matrices. The diagonal part corresponds to a model dependent correction in the direction of the steepest descent: $\nabla_{\mathbf{x}}H$. For a finite step size δ in the numerical integration scheme, the symplectic matrix $J = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$ should be replaced with J_δ in order to attain only second-order errors in integration.

The matrix J_δ is the first correction to change the generator J of symplectic rotations into the rotation matrix J_θ over a finite angle θ, based on the local curvatures of the energy surface, in the neighborhood of a point $(\mathbf{q}, \mathbf{p})$ of the phase space.

The Hamiltonian in the neighborhood of $(\mathbf{q}, \mathbf{p})$ can be written as a quadratic Hamiltonian, up to an additive constant which can be ignored:

$$(3.46) \qquad H = \frac{1}{2}\mathbf{q}^T\mathbf{K}\mathbf{q} + \frac{1}{2}\mathbf{p}^T\mathbf{M}^{-1}\mathbf{p} \;\; .$$

Therefore, a local coordinate system $\mathbf{y}$ can map this surface to the unit sphere:

$$(3.47) \qquad \mathbf{y} = \Omega\mathbf{x} = \begin{pmatrix} \sqrt{\mathbf{K}} & 0 \\ 0 & \frac{1}{\sqrt{\mathbf{M}}} \end{pmatrix}\mathbf{x} \;\; ,$$

yielding a normalized Hamiltonian:

$$(3.48) \qquad H = \frac{1}{2}\mathbf{y}^T\mathbf{y} \;\; .$$

Since $\mathbf{K}^T = \mathbf{K}$ and $\mathbf{M}^T = \mathbf{M}$ are both positive we can take the square root of these matrices. In this coordinate system, any time evolution is a rotation $R(\theta(\delta))$:

$$(3.49) \qquad \mathbf{y}(t + \delta) = \mathbf{R}(\theta)\mathbf{y}(t) \;\; .$$

Hence, inserting an explicit rotation, based on the standard symplectic structure, being the generator of rotation, we find:

$$(3.50) \qquad \mathbf{x}(t + \delta) - \mathbf{x}(t) = \Omega^{-1}\begin{pmatrix} \cos\theta - I & \sin\theta \\ -\sin\theta & \cos\theta - I \end{pmatrix}\Omega^{-1}\Omega^2\mathbf{x} \;\; .$$

Locally, by definition:

$$(3.51) \qquad \Omega^2\mathbf{x} = \nabla_{\mathbf{x}}H \;\; ,$$

such that the symplectic structure for the finite time-step reduces to $\Omega^{-1}(R-1)\Omega^{-1}$. In the limit $\delta \to 0$, we find:

$$(3.52) \qquad \theta = \arcsin\sqrt{\frac{\mathbf{K}}{\mathbf{M}}}\delta \approx \sqrt{\frac{\mathbf{K}}{\mathbf{M}}}\delta \;\; .$$

The finite-step symplectic structure δJ_δ, Eq. 3.45, also reduces to the same result if expanded as a polynomial in δ to the second order:

$$(3.53) \quad \delta J_\delta + \mathcal{O}(\delta^2) = \begin{pmatrix} \frac{1}{\sqrt{\mathbf{K}}}(\cos\sqrt{\frac{\mathbf{K}}{\mathbf{M}}}\delta - 1)\frac{1}{\sqrt{\mathbf{K}}} & \frac{1}{\sqrt{\mathbf{K}}}\sin\sqrt{\frac{\mathbf{K}}{\mathbf{M}}}\delta\sqrt{\mathbf{M}} \\ -\sqrt{\mathbf{M}}\sin\sqrt{\frac{\mathbf{K}}{\mathbf{M}}}\delta\frac{1}{\sqrt{\mathbf{K}}} & \sqrt{\mathbf{M}}(\cos\sqrt{\frac{\mathbf{K}}{\mathbf{M}}}\delta - 1)\sqrt{\mathbf{M}} \end{pmatrix} \;\; ,$$

In practice, calculating the squareroot of a matrix, in Eq. 3.53, may be much more cumbersome than the second order correction of Eq. 3.45.

Hence for quadratic Hamiltonians, the energy conservation can be made exact by using the finite rotation θ, although the evaluation of the analytic functions of the matrices is more expensive than the diagonalization, which allows for an exact solution in terms of independent eigenmodes. This formulation is more useful for general nonlinear Hamiltonians, which already require the evaluation of the mass and stiffness matrix. These evaluations can be used to yield a second-order integration method. Higher-order effects of the nonlinear Hamiltonians can be expressed in the variations along the path of the mass and stiffness matrices, which is discussed in the next section.

3.6. Higher order results

From the previous section it is clear there are several ways to proceed to higher order results. In the case of a quadratic Hamiltonian, the exact rotation can be implemented for diagonal mass and stiffness matrices. For non-diagonal parts, the rotation still yields high-order corrections. In particular for FEM Hamiltonians we can improve upon the result, as we will see later.

For nonlinear Hamiltonians, the mass and stiffness matrices are local approximations. A linear dependence can be fitted to both these matrices $\mathbf{A} = \mathbf{M}^{-1}$ and $\mathbf{A} = \mathbf{K}$:

$$(3.54) \qquad \mathbf{A}(t+\delta) \approx (1 + \delta\mathbf{A}'(t))\mathbf{A}(t) \quad,$$

and

$$(3.55) \qquad \mathbf{A}^{-1}(t+\delta) \approx \mathbf{A}^{-1}(t)(1 - \delta\mathbf{A}'^T(t)) \quad.$$

We no longer rely on the state $\mathbf{x}$ dependence. Generally, we would expect the mass and stiffness matrices, generically denoted by $\mathbf{A} + \mathbf{A}'t$, to vary more in time slowly and weakly than the state $\mathbf{x}(t)$. The linear approximation for these matrices would in most cases yield a higher-order accuracy for the state.

In the simulation the matrices are extrapolated from the previous iteration. Inserting this back into the time integral we find:

$$(3.56) \qquad J_\delta = \begin{pmatrix} -\frac{\delta\mathbf{M}(1-\mathbf{M}'^T t')}{2} & I \\ -I & -\frac{\delta\mathbf{K}^{-1}(1-\mathbf{K}'^T t')}{2} \end{pmatrix} \quad.$$

with the variational derivative contributing:

$$(3.57) \qquad \delta_{\mathbf{x}}H = \begin{pmatrix} (1+\mathbf{K}'t')\nabla_q H \\ (1+\mathbf{M}'t')\nabla_p H \quad. \end{pmatrix}$$

yielding to second order:

$$(3.58) \qquad J_\delta^{(2)} = \begin{pmatrix} -\frac{\delta\mathbf{M}}{2} + \delta^3\frac{\mathbf{M}(\mathbf{K}'-\mathbf{M}'^T)}{4} & 1+\frac{\delta^2\mathbf{M}'}{2} \\ 1+\frac{\delta^2\mathbf{K}'}{2} & -\frac{\delta\mathbf{K}^{-1}}{2} + \delta^3\frac{\mathbf{K}^{-1}(\mathbf{M}'-\mathbf{K}'^T)}{4} \end{pmatrix} \quad.$$

This result is already closer to that of methods such as Runge-Kutta, where the state $\mathbf{x}$ is estimated at later times, to yield an improved mean derivative $H(\mathbf{x}(t)+\Delta)$ over the step size domain $[t, t+\delta]$.

The expense, however, of determining and storing the linear dependence $\mathbf{M}'$ and $\mathbf{K}'$ of the mass and stiffness matrices, might not to be worth the trouble if large systems are studied. Although the mass and stiffness matrices themselves are usually sparse, no guarantee exists that along the path $\mathbf{x}(t)$ these matrices are sparse too. The computational effort increases with the number of nonzero elements of these matrices, if efficient iterative, sparse-matrix routines are used.

3.7. Integration and control

Simulation is about predicting a result based on a model expressed as a differential equation. Control is to generate a desired output of a system. Hence with some knowledge of the system, expressed in a plant model, integration methods can be used to predict the output and adjust the input such that the desired output is achieved. A more competent model of the system will yield a better prediction. The purpose of this simple set-point control, is to give an input, like an applied

force, to change the actual output to a desired output. Even with little knowledge of the plant model, such as inaccurate values for masses and springs, the output can be brought to a desired value. The connection between feedback control theory and numerical integration schemes has been fruitful in many different respects, for example, in audio compression and signal transmission.

In many cases in control the dynamical behavior of a system or plant is approximated by a second-order system. It corresponds to a mass-spring-damper model for system. In discrete time two values $x(t_i)$ and $x(t_{i-1})$ are used to predict the subsequent value $x(t_{i+1})$. Consider the force-actuated, damped oscillator:

$$(3.59) \qquad m\ddot{x} + d\dot{x} + kx = F_{\text{control}} \quad .$$

In the case of integration we can take small time steps $t_{i+1} - t_i = \delta$, and the equation reduces to:

$$(3.60) \qquad m\frac{x(t_{i+1}) - 2x(t_i) + x(t_{i-1})}{2\delta^2} + d\frac{x(t_i) - x(t_{i-1})}{\delta} + kx(t_i) = F(t_i) \quad .$$

For a more advanced discretization and integration, the control can evolve with it. In many cases the physical constants: the mass m, the damping d, and the stiffness k are unknown. System identification tries to recover these constants from responses to particular input $F(t)$, and output $x(t)$, and in the case that identification and control are combined one speaks of adaptive, or learning, control. [Åström and Wittenmark, 1989]

If we wish to control the state variable $x(t)$ it means we wish to determine $x(t_{i+1})$ by adjusting $F(t_i)$ such that $|x(t_{i+1}) - x_{\text{desired}}(t_{i+1})|$ is as small as possible. However, to do so, we should first predict what the result would be without applied force:

$$(3.61)$$
$$x(t_{i+1}) \equiv x_\delta(x(t_i), x(t_{i-1})) = 2x(t_i) - x(t_{i-1}) - d\delta\frac{x(t_i) - x(t_{i-1})}{m} - k\delta^2\frac{x(t_i)}{m} \quad .$$

The simplest control is proportional control. The force proportional to the difference between the desired and the actual output:

$$(3.62) \qquad F_i = K_P(x_{\text{desired}}(t_i) - x(t_i)) \quad .$$

In many cases this will drive $x(t_i)$ to $x_{\text{desired}}(t_i)$. However, a more advanced approach takes into account the velocity which will change $x(t_{i+1})$ with respect to the control and measure point $x(t_i)$:

$$(3.63) \qquad F_i = K_P(x_{\text{desired}}(t_{i+1}) - (1 + \delta)x(t_i) + \delta x(t_{i-1})) \quad .$$

$$= K_P((x_{\text{desired}}(t_{i+1}) - x(t_i)) - K_D(x(t_i) - x(t_{i-1})) \quad .$$

The additional term is called the derivative action.

Finally, the stationary deviation of $x(t_\infty)$ and $x_{\text{desired}}(t_\infty)$ will remain finite, even for a constant set point x_{desired}, due to the reaction force $-kx(t_i)$:

$$(3.64) \qquad k(x_{\text{desired}}(t_\infty) - x(t_\infty)) = F = K_P(x_{\text{desired}}(t_\infty) - x(t_\infty)) \quad .$$

This can be corrected by adding an increasing, integrating, the gain K_I, as the difference persists: [Föllinger, 1974]

$$(3.65) \qquad F_i = K_P(x_{\text{desired}}(t_{i+1}) - x(t_i) + K_I \sum_{j=-\infty} (x_{\text{desired}}(j + i) - x(j + i)) \quad .$$

which allows one to increase the total gain substantially without the fast oscillations of the system blowing up. Without the derivative action, this constitutes a proportional-integrating (PI) controller.

However, the time-dependent behavior will always trail behind the set-point trajectory $x_{\text{desired}}(t)$. These are the typical elements which make up a proportional integrating differential (PID) controller. [Ogata, 1990] The integrating gain K_I and the differential gain K_D are often expressed as their typical time constants τ_I and τ_D times the proportional gain:

$$(3.66) \qquad K_I = \frac{K_P}{\tau_I}, \quad K_D = K_P \tau_D \quad .$$

If we would have had exact values for the system parameters m, d, and k, we could have integrated the equation exactly, yielding:

$$(3.67) \qquad F_i = \frac{m(x_{\text{desired}}(t_{i+1}) - x_\delta(x(t_i), x(t_{i-1})))}{\delta^2} \quad ,$$

which would yield typically very large forces, if $x \neq x_{\text{desired}}$, since the time-step δ is small. The integration method, which yields the result x_δ, however, does not need to be specified for the control purpose. The control based on the exact model, which therefore requires no feedback, if $x = x_{\text{desired}}$ at the start, is called feed-forward control.

3.8. Time and time again

In this chapter, we established a simple model description of dynamical systems, for the purpose to study the dynamical, or time-dependent behavior. Simulation plays an important role in the study of complex interacting, or open systems. Their behavior is not easily characterized, and control of dynamical systems are based on the time-dependent behavior of these systems. For continuous systems, a proper selection of modes, which lie at the basis of the $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$-model, is determined by the behavior as the result of input. This problem is investigated in Chapter 9. The simulation, or time integration, itself, and the implementation together with constraints, linear and nonlinear, is investigated in Chapter 10. The particular structure of the evolution operator $\mathbf{A}$ in the case of FEM, and how it can be used for physical consistency is part of Chapter 8, in particular in Sections 8.5 and 8.7. Eventually, we perceive time through the dynamical systems that surround us. Therefore, dynamical models, time integration, and simulation are closely related, and none of them can exist without the others.

The next two chapters on differential geometry and partial differential equations take a closer look at the mathematics of continuous models. First, aspects of the underlying reference space are examined, in Chapter 4. In the second chapter, Chapter 5, the state space as function with boundary values is studied within the theory of partial differential equations.

CHAPTER 4

Differential geometry

Infinite-dimensional systems only come to life as collective objects through the fact they are continuous systems, i.e., the state variables are defined on some spatial domain, the reference space. The consequences are manifold, and some of them are directly related to the geometric aspects of the domain. They are expressed in terms of differential geometry, linking the properties of the reference space, to the properties of the state space. Furthermore, a port-based approach tries to distill the invariant and conserved global aspects from a system, these aspects are intricately related to transformations and conserved flows, based on the geometry.

In this chapter we aim to introduce some physical concepts related to differential geometry, [Guggenheimer, 1963, Spivak, 1970, Felsager, 1981, Dubrovin et al., 1984, Abraham et al., 1991, Ivey and Landsberg, 2000, Morita, 2001, Frankel, 2004] which will be used later. Differential geometry is a large field of mathematical research. In principle differential geometry deals with differential functions in higher dimensions. Global differential geometry deals with the topology of objects and invariants such as the winding or linking numbers. [Chern, 1989]

In this thesis we deal mainly with local differential geometry; the local, often analytic properties of functions in higher dimensional space. Differential geometry arises in many different aspects of modelling. The integrability condition is probably the most important aspect. For example the elastic energy is given in terms of curvatures and bending. These are the natural variables for the energy, but not for the configuration. Whether for all values of elastic-energy variables a configuration actually exists, is an example of the question of integrability. See for example Section 11.10. The actual construction of appropriate geometric energy variables and Hamiltonians which are invariant, for example, under rotations and translations, are the general study of Chapter 11, where it should be noted that we seek invariant elements and invariant boundary variables. [Kron, 1934, 1965]

From the physical perspective the conservation of energy and the power continuity is an essential part of the modelling strategy. Furthermore, other conservation laws will play an important part to construct consistent models. For that purpose, the energy conservation is expressed as a local law, i.e., the energy continuity, which requires the construction of an energy flux. See Sections 5.3 and 7.5.

4.1. Surfaces

A surface in three dimensional space can be seen as a set of three continuous functions of two variables:

(4.1)
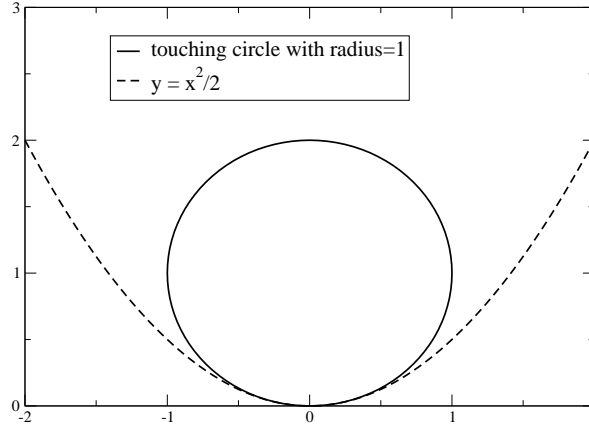$$\begin{pmatrix} x(u,v) \\ y(u,v) \\ z(u,v) \end{pmatrix} = \mathbf{x}(u,v) \quad ,$$

FIGURE 4.1. The curvature of the parabola $y = x^2/2$ at $x = 0, y = 0$ is one. The circle touching the parabola has a radius $r = 1/k = 1$.

which assign to each point u, v of the surface a three-dimensional coordinate $\mathbf{x} = (x, y, z)$.

Likewise, if the surface is not perpendicular to either the x or the y coordinate, the surface coordinates u, v can be replaced by x, y, leaving a single relation for the height $z(x, y)$ as function of the projected surface position.

The invariants associated with a surface are the quantities which do not depend on the orientation of the surface or the choice of local coordinates, like u, v or x, y. The three invariants are the area and two curvatures. The area is given by the integral:

$$(4.2) \qquad A = \int |\partial_u \mathbf{x}(u, v) \times \partial_v \mathbf{x}(u, v)| du dv \quad .$$

The cross product, in more modern mathematical language, is replace by the exterior product $\mathrm{d}x \wedge \mathrm{d}y(u, v)$, where the sign is given by the orientation of x, y in terms of u, v. [Flanders, 1963]

At each point of the surface it also has two curvatures: the maximal curvature and the minimal curvature. The curvature k of a line can be defined as the inverse radius of a circle which touches a curve tangentially. For example, a parabola $y = x^2/(2r)$ has a curvature at $x = y = 0$ of $k = 1/r$. Take a circle of radius $r = 1/k$, touching the x-axis, it can be approximated near $x \approx 0, y \approx 0$ by:

$$(4.3) \qquad 0 = r - \sqrt{x^2 + (y - r)^2} \approx y - \frac{x^2}{2r} \quad .$$

See Figure 4.1.

The two curvatures of a surface are related to the largest and smallest eigenvalue of the Hessian, which describes the parabolic approximation of the surface around a point $\mathbf{x}_0 = \mathbf{x}(u_0, v_0)$ with a tangent surface $\mathbf{V}$:

(4.4)
$$\mathbf{x} = \mathbf{x}_0 + \mathbf{V}_u(u - u_0) + \mathbf{V}_v(v - v_0) + \mathbf{H}_{uu}(u - u_0)^2 + \mathbf{H}_{uv}(u - u_0)(v - v_0) + \mathbf{H}_{vv}(v - v_0)^2 \quad .$$

The tangential directions $\mathbf{V}$ are

$$(4.5) \qquad \mathbf{V}_u = \partial_u \mathbf{x}(u_0, v_0) \quad .$$

The Hessian consists of the second-order terms in the double Taylor expansion:

$$(4.6) \qquad \mathbf{H}_{uu} = \frac{1}{2}\partial_u^2 \mathbf{x}(u_0, v_0) \quad .$$

We choose rectified coordinates with the tangent direction of the surface in the $x - y$-plane such that the $\mathbf{V}_u = \mathbf{e}_x$ and $\mathbf{V}_v = \mathbf{e}_y$. The Hessian is a vector in the z-direction, $\mathbf{H}_{ij} = \mathbf{e}_z z_{ij}$. The eigenvalues of the Hessian, or rather the invariants of the Hessian $H = \det z_{ij} = \lambda_1 \lambda_2$ and $K = \frac{1}{2}\mathrm{Tr} z_{ij} = \frac{1}{2}(\lambda_1 + \lambda_2)$ are not invariant under coordinate transformations, they are defined in this coordinate frame, however, other values for $\mathbf{V}_i = \mathbf{e}_i + \mathbf{e}_z z_i$ with $z_i \neq 0$ are easily accounted for. The invariants are the Gauss curvature:

$$(4.7) \qquad H = \frac{z_{xx} z_{yy} - z_{xy}^2}{(1 + z_x^2 + z_y^2)^2} \quad ,$$

and the mean curvature:

$$(4.8) \qquad K = \frac{(1 + z_y^2)z_{xx} + (1 + z_x^2)z_{yy} - 2z_x z_y z_{xy}}{(1 + z_x^2 + z_y^2)^{\frac{3}{2}}} \quad .$$

A surface with a zero mean curvature everywhere is a minimal surface. Films of soap are the most popular examples of minimal surfaces, however, any system where the energy is dominated by the surface tension has a minimal surface solution as the static solution. [Chern, 1989] There is therefore a connection between the variational problem of minimizing the area A of a surface Ω, given the boundary $\partial\Omega$, and the mean curvature:

$$(4.9) \qquad \delta A = \delta \int_\Omega dxdy |\partial_x \mathbf{x} \times \partial_y \mathbf{x}| = 0 \Leftrightarrow K|_\Omega = 0 \quad .$$

The surface tension γ relates the area of the surface to its energy $E = \gamma A$.

4.2. Frenet-Serret coordinates

For curves, a natural, local coordinate system arises from the curvature. These are called the Frenet-Serret coordinates[Lipschutz, 1969, Guggenheimer, 1963], and are normally denoted by the basis vectors $(\mathbf{n}, \mathbf{b}, \mathbf{r})$. A curve is given by a single variable s, the reference coordinate, which denotes the position $\mathbf{x}(s)$ along the curve. The direction of the curve is given by the change of position as a function of the change of reference position:

$$(4.10) \qquad \mathbf{n}(s) = \frac{\partial_s \mathbf{x}(s)}{|\partial_s \mathbf{x}(s)|} \quad .$$

The direction of the curvature is given by the second derivative:

$$(4.11) \qquad \mathbf{r} = \frac{\partial_s \mathbf{n}(s)}{|\partial_s \mathbf{n}(s)|} \quad ,$$

which is perpendicular to the curve direction $\mathbf{n}$ since:

$$(4.12) \qquad 0 = \partial_s 1 = \partial_s \mathbf{n}^2 = \mathbf{n}^T \mathbf{r} \quad .$$

The third direction is given by:

$$(4.13) \qquad \mathbf{t} = \mathbf{n} \times \mathbf{r} \quad .$$

See also Figure 6.1. The triple $(\mathbf{n}, \mathbf{b}, \mathbf{r})$ form an orthonormal basis, and can be represented as a rotation matrix $\mathbf{R}$: $\mathbf{R}^T \mathbf{R} = 1$

$$(4.14) \qquad\qquad\qquad \mathbf{R} = (\mathbf{n}, \mathbf{b}, \mathbf{r}) \ .$$

The angular representations of the rotation matrix $\mathbf{R}$ can be used in analytic theory of curves. The matrix is represented by a rotation axis $\omega/|\omega|$ and angle $|\omega|$:

$$(4.15) \qquad\qquad\qquad \mathbf{R} = e^{J \cdot \omega(s)} \ ,$$

where the generators are:

$$(4.16) \qquad\qquad J \cdot \omega = \begin{pmatrix} 0 & \omega_3 & -\omega_2 \\ -\omega_3 & 0 & \omega_1 \\ \omega_2 & -\omega_1 & 0 \end{pmatrix} \ .$$

The direction vector $\mathbf{n} = (1, 0, 0)$ yields the curvature $\partial_s \omega_3$ while $\partial_s \omega_2 = 0$ through the definition, and the twist $\partial_s \omega_1$.

A curve is uniquely defined by the curvature $\partial_s \omega_3$ and the twist $\partial_s \omega_1$. Furthermore, the elastic energy of a curve can only be a function of these invariants. On the other hand, it is difficult to recuperate the actual form $\mathbf{x}(s)$ of a curve from these differential quantities. Some simple examples are the straight line for $\partial_s \omega_1 = 0, \partial_s \omega_3 = 0$, circle for $\partial_s \omega_1 = 0, \partial_s \omega_3 = $ constant, and the circular helix for $\partial_s \omega_1 = 0, \partial_s \omega_3 = $ constant.

4.3. Integrability

From the local properties of a function, usually expressed in terms of differential equations, it might be possible to reconstruct the function. However, such result might not always exists. In some cases the differential equations do not allow for a solution, in other cases a solution may exist, but might be difficult to recover in practice. The condition whether the curve or surface may be reconstructed from a differential equation is called the integrability condition. It is a large and diverse field of research, which received a boost after the discovery of integrability of nonlinear systems, which yield an infinite number of conserved quantities. These results are closely related to the inverse scattering problem. [Ablowitz and Clarkson, 1991]

We will discuss only some elementary, classical results, in integrability. A general theory of integrability for nonlinear systems is absent. Furthermore, we take the constructive approach. Rather than defining equations of motion, and asking ourselves the question whether a generating Hamiltonian exists, we construct the Hamiltonian. The state of the system, without dissipation, will stay in surfaces with constant value of the Hamiltonian.

For a curve in three dimensions there is a unique one-to-one mapping from differential invariants, curvature and torsion, to the shape of the curve $\mathbf{x}(s)$. For surfaces in three-dimensional space and in general for differential formulations a global, or even a local, solution may not exists. The Maxwell-Pfaff-Cartan consistency relations are the best known example of a test of the existence of a solution. Given the two differential equations:

$$(4.17) \qquad\qquad \partial_x z(x, y) = p(x, y, z) \ ,$$

and

$$(4.18) \qquad\qquad \partial_y z(x, y) = q(x, y, z) \ ,$$

they will define a surface $z(x, y)$ if they satisfy the hidden, or compatibility equation:

$$(4.19) \qquad \partial_y p + q\partial_z p = \partial_x q + p\partial_z q \quad .$$

In the absence of z dependence, this relation is the Maxwell reciprocity relation in thermodynamics. The local condition at some point x, y, z implies the global condition as well.

Another problem would be if the surface normal $\mathbf{n}$ at each point x, y defines a surface uniquely [Chirgwin and Plumpton, 1964]. Lagrange has studied this question and related the answer to the existence of total level surfaces. Lagrange partial differential equation uses the surface normal $\mathbf{n} = (n_x, n_y, n_z)$ to define the surface implicitly:

$$(4.20) \qquad n_x(x, y)\partial_x z(x, y) + n_y(x, y)\partial_y z(x, y) - n_z(x, y) = 0 \quad .$$

The surface defined by the differential equation can be constructed with the use of two constraint surfaces:

$$(4.21) \qquad u(x, y, z) = \text{constant}, \quad v(x, y, z) = \text{constant} \quad ,$$

which satisfy the equations:

$$(4.22) \qquad \frac{dx}{n_x} = \frac{dy}{n_y} = \frac{dz}{n_z} \quad .$$

For such surfaces it follows that:

$$(4.23) \qquad 0 = du = \partial_x u dx + \partial_y u dy + \partial_z u dz = \partial_x u n_x + \partial_y u n_y + \partial_z u n_z \quad .$$

A general function $F(u, v) = 0$ of the two independent streamline functions u and v satisfies the differential equation including the Lagrange differential equation as a factor:

$$(4.24) \quad (\partial_z u \partial_u F + \partial_z v \partial_v F)(n_x(x, y)\partial_x z(x, y) + n_y(x, y)\partial_y z(x, y) - n_z(x, y)) = 0 \quad ,$$

which follows from combining the $\partial_x F$ and $\partial_y F$. Hence defining a surface from its normal is related to the integrability condition of a vector fields defined by infinitesimals:

$$(4.25) \qquad n_y n_z dx = n_x n_z dy = n_x n_y dz \quad .$$

4.4. Euler and Frobenius integrability

Another integrability condition is associated with the construction of a potential from a vector field. In electrostatics there are two equivalent descriptions of the electrostatic field. One in terms of the electric field $\mathbf{E}$, and one in terms of the potential V. The two are related through:

$$(4.26) \qquad \mathbf{E} = -\nabla V \quad .$$

The question is, if given an arbitrary electric field $\mathbf{E}$, there always is a potential V. Since the electric field is rotation free, i.e., $\nabla \times \mathbf{E} = 0$, this is indeed the case. A slightly more general condition is Euler's condition: [Frankel, 2004]

$$(4.27) \qquad \mathbf{n}^T (\nabla \times \mathbf{n}) = 0 \quad ,$$

where the surface normal $\mathbf{n} = |\mathbf{E}|^{-1}\mathbf{E}$ is based on the vector field $\mathbf{E}$.

The general setting of the Frobenius integrability theorem has many uses in control.[Nijmeijer and van der Schaft, 1990, Slotine and Li, 1991, Pappas et al.,

1998] It is stated in terms of Lie derivatives, and states that a potential V, or a solution of lower dimension of $V = $ constant for multiple vectors fields can only be found if the fields are involutive. A Lie derivative associated with a vector field $\mathbf{E}$ is the differential operator:

$$(4.28) \qquad L_{\mathbf{E}} = \mathbf{E}^T \nabla = E_x \frac{\partial}{\partial x} + E_y \frac{\partial}{\partial y} + E_z \frac{\partial}{\partial z} \quad .$$

A set of differential equations based on the vector fields $\mathbf{E}_i$:

$$(4.29) \qquad L_{\mathbf{E}_1} V = L_{\mathbf{E}_2} V = \cdots = L_{\mathbf{E}_r} V = 0 \quad ,$$

only has a solution V the Lie-derivatives of the vector fields $\mathbf{E}_i$ lie in the subspace spanned by these vector fields. The anti-symmetric product, Lie product or commutator, of two differential operators $L_{\mathbf{E}_i}$ and $L_{\mathbf{E}_j}$ is again a differential operator:

$$(4.30) \qquad [L_{\mathbf{E}_i}, L_{\mathbf{E}_j}] = L_{\mathbf{E}_i} L_{\mathbf{E}_j} - L_{\mathbf{E}_j} L_{\mathbf{E}_i} = \mathbf{E}_{ij}^T \nabla \quad .$$

If for all i, j at every point of space independently holds:

$$(4.31) \qquad \mathbf{E}_{ij} \in \mathrm{span}\{\mathbf{E}_1, \mathbf{E}_2, \cdots \mathbf{E}_r\} \quad ,$$

the vector fields are called involutive, and an integral solution V will exists.

4.5. Conserved flow

The conserved flows are central to stable dynamics. There exist many different notions of stability. [Lyapunov, 1992, Willems, 1972a, Nijmeijer and van der Schaft, 1990, Worden and Tomlinson, 2001] In this thesis the stability is related to the energy norm. Therefore, stability, even in the case of nonlinear systems is guaranteed by the conservation of energy. If the energy flux, the differential geometric version of power in the continuous case, can be defined, it gives one a handle on treating boundary conditions and initial conditions in a systematic manner.

However, conserved flow is more prevalent than only in energy conservation. Incompressible flow is described by a velocity field $\mathbf{v}(\mathbf{x})$ which is divergence free:

$$(4.32) \qquad \nabla \cdot \mathbf{v}(\mathbf{x}) = 0 \quad ,$$

indicating that fluid cannot build up in any region of space; whatever flows in must flow out. The integral version is called the divergence theorem:

$$(4.33) \qquad \int_\Omega \nabla \cdot \mathbf{v}(\mathbf{x}) d^3 x = \int_{\partial \Omega} \mathbf{n} \cdot \mathbf{v}(\mathbf{x}) d^2 x \quad ,$$

where, $\mathbf{n}$ is the surface normal. The theorem states that the total divergence of a flow $\mathbf{v}(\mathbf{x})$ in a domain Ω is equal to the total flow through the surface $\partial\Omega$ of the same domain. The former has little meaning in our macroscopic world, however, the latter is nothing but the bookkeeping of the amount of fluid, associated with the flow, in a domain Ω.

For example, the stress tensor σ is constructed, such that the divergence theorem can be applied. Each of the force components F_x, F_y, and F_z of a stationary object is in balance, which means that the sum of all the forces F_i applied to the surface must add up to zero:

$$(4.34) \qquad 0 = \int_{\partial \Omega} F_i(\mathbf{x}) d^2 x \quad .$$

Hence, in order to describe the forces inside the object it makes sense to use a divergence of a tensor:

$$\sum_j \partial_j \sigma_{ji} = f_i \quad , \tag{4.35}$$

where f_i is the body force, such as gravity $\mathbf{f} = -\mathbf{g}\rho$, or inertial force: $\mathbf{f} = \rho\dot{\mathbf{v}}$.

The stress tensor σ is the microscopic implementation of the force-balance law. Its properties and consistency conditions of the stress tensor have been a matter of debate till far in the twentieth century.

The force balance is just one of the continuity equations. In their generic form they are written as the equation for rate of change of the charge Q equals the divergence of the flow $\mathbf{J}$:

$$\dot{Q} = -\nabla \cdot \mathbf{J} \quad . \tag{4.36}$$

The continuity equation is not affected by rotational flows. Any term $\nabla \times \mathbf{A}$, where $\mathbf{A}$ is an arbitrary vector potential yields the same continuity relation:

$$\dot{Q} = -\nabla \cdot (\mathbf{J} + \nabla \times \mathbf{A}) = -\nabla \cdot \mathbf{J} \quad . \tag{4.37}$$

The flow represented by $\nabla \times \mathbf{A}$ are typical whirls and eddy currents. On the other hand, such whirly flow is conserved by itself. One has to think only of large weather depressions and hurricanes. In the literature a localized whirl is called a Helmholtz flow:

$$\mathbf{v} = \mathbf{n} \times \nabla\psi(\mathbf{x} - (\mathbf{n}^T\mathbf{x})\mathbf{n}) \quad , \tag{4.38}$$

which yields planar flow perpendicular to the normal vector $\mathbf{n}$, where the streamlines follow the levels, or contours of the potential ψ.

In general a vector field can be decomposed into a rotational part, a gradient part, and a harmonic part. This is called the Helmholtz-Hodge decomposition: [Flanders, 1963]

$$\mathbf{v} = \nabla \times \mathbf{A} + \nabla\psi + \mathbf{w} \quad , \tag{4.39}$$

where the harmonic function satisfies the Laplace equation $\Delta\mathbf{w} = 0$. The Green function, or fundamental solution, of the Laplace operator is essential to construct this decomposition:

$$\Delta G(\mathbf{x}) = \delta^n(0) \quad , \tag{4.40}$$

The vector potential generating the rotational part is given by:

$$\mathbf{A}(\mathbf{x}) = \int d^n x' G(\mathbf{x}' - \mathbf{x})(\nabla_{\mathbf{x}'} \times \mathbf{v}(\mathbf{x}')) \quad . \tag{4.41}$$

The scalar potential is given by:

$$\psi(\mathbf{x}) = \int d^n x' G(\mathbf{x}' - \mathbf{x})(\nabla_{\mathbf{x}'} \cdot \mathbf{v}(\mathbf{x}')) \quad . \tag{4.42}$$

The harmonic part is constructed from the remainder, which can be written as a projection operator onto the harmonic part:

$$\mathbf{w} = \mathbf{v} - \int d^n x' G(\mathbf{x}' - \mathbf{x})(\Delta_{\mathbf{x}'}\mathbf{v}(\mathbf{x}')) \quad . \tag{4.43}$$

The Green function in n dimensions is given by: [Gel'fand and Shilov, 1964, Shilov, 1968]

$$(4.44) \qquad G(\mathbf{x}) = \frac{-1}{(n-2)\Omega_n} \frac{1}{|\mathbf{x}|^{n-2}} \quad,$$

where Ω_n is the surface area of the n-dimensional sphere, where $n > 2$. In the two-dimensional case $n = 2$ the Green function of the Laplacian is:

$$(4.45) \qquad G(\mathbf{x}) = \frac{1}{2\pi} \log |\mathbf{x}| \quad.$$

See also Section 12.3. Using the Green functions, the problem of finding a solution to a partial differential equation can be mapped to the, often much simpler, problem of determining an integral.

4.6. Remarks on differential geometry and other differentials

Differential geometry has been useful to distinguish two aspects of a state defined on a domain. First, the possibility to change the state variables and the invariant state variables which transform with the coordinate transformations on the reference space. Second, the relation between the domain and its boundary, which is an essential concept to be able to construct interfaces between interacting, open systems.

For the study of the bulk, i.e., the behavior on the domain, the theory of partial differential equations is more appropriate. This is investigated in the following chapter. The aspects of differential geometry arise in these partial differential equations in two ways. First, the study of boundary conditions, since differential geometry determines the relation between the domain and its boundary. Secondly, the differential operators which appear as building block is partial differential equations and map between different geometric function spaces. As it turns out, the two aspects are intimately related.

Partial differential equations

> In the mid-twentieth century the theory of partial differential equations was considered the summit of mathematics, both because of the difficulty and the significance of the problems it solved and because it came into existence later than most areas of mathematics.
>
> Nowadays many are inclined to look disparagingly at this remarkable area of mathematics as an old-fashioned art of juggling inequalities or as a testing ground of functional analysis.
>
> [V. I. Arnold, Lectures on partial differential equations]

We will not attempt to give an introduction into the theory of partial differential equations. There are many good textbooks, which, together, give an insight in the full breadth of the theory of partial differential equations. The older literature deals with explicit solutions and both linear and nonlinear theory. [Carathéodory, 1935, Courant and Hilbert, 1961, John, 1982, Kamke, 1962, Hadamard, 1923, Arnold, 2004] In intermediate years, the middle of the twentieth century, many new developments arose, [Bers et al., 1964, Gårding, 1998, Garabedian, 1964, Shilov, 1968] but shifted the focus towards operational and weak formulations of the theory of linear partial differential equations, [Hörmander, 1990, Lions, 1970] because the general theory of nonlinear partial differential equations in more than two dimensions, without appropriate structure, is considered prohibitively complicated.[Serre, 2001]

Since our main interest is interacting systems described by partial differential equations, nontrivial boundary conditions are our special interest. The theory of initial and boundary conditions of hyperbolic differential equations is another small niche in the study of partial differential equations. [Sakamoto, 1982, Ladyzhenskaya, 1985] In particular the weak formulations, in terms of norms, are generally set up with vanishing boundary terms. Hence these approaches need to be adapted to make them suitable for nontrivial boundary conditions. The interaction of systems is through the boundary conditions. The corresponding volume norm can be expressed in terms boundary norms, the energy flux, and initial-state and final-state norm, the energy of the respective state.

In this chapter we review and develop some theory concerning boundary conditions in, mainly, hyperbolic partial differential equation. The starting point is the partial differential equation itself, and the energy norm and energy flux, essential for the correct physical interpretation of the boundary condition, are derived. The energy norm is useful for both the mathematical and the physical consistency of the model. Uniqueness and existence follows from the proper energy norm, and energy continuity is an important criterion in modelling.

In this chapter we set up the conventions and tools which we will use in later chapters. The Sturm-Liouville theory (Section 5.2) is considered the paradigm of the modern approach to partial differential equations. It has a clear link to a functional formulation, of which the energy functional is a particular useful physical case.

Partial differential equations are equations in higher order spaces which depend on derivatives with respect to the different variables:

$$(5.1) \qquad P(z, \nabla_z)u(z) = 0 \ ,$$

where u is the unknown function to be determined, and the differential operator P is a function of the variables $z = (z_0, z_1, \cdots, z_n)$ and a polynomial with respect to the different derivatives, or the gradient:

$$(5.2) \qquad \nabla_z = \begin{pmatrix} \partial_{z_0} \\ \partial_{z_1} \\ \partial_{z_2} \\ \vdots \\ \partial_{z_n} \end{pmatrix} \equiv \begin{pmatrix} \partial_0 \\ \partial_1 \\ \partial_2 \\ \vdots \\ \partial_n \end{pmatrix} \ .$$

The standard linear hyperbolic differential equation is the wave equation, which is a second-order polynomial in terms of the derivatives:

$$(5.3) \qquad (\partial_0^2 - \partial_1^2 - \partial_2^2 - \partial_3^2)u(\mathbf{z}) = (\partial_t^2 - \Delta)u(\mathbf{z}) = \Box u(\mathbf{z}) = 0 \ ,$$

where $z_0 = t$ is the time, and z_1, z_2 and z_3 the spatial coordinates. The Laplace operator Δ and the D'Alembertian or wave operator $\Box$ are the standard differential operators for isotropic spaces. [Magnus et al., 1966] This is a hyperbolic differential equation since both signs occur in the quadratic polynomial.

Many types of analysis of PDE's are based on looking at the polynomial $P(x, \xi)$, where the vector ξ replaces the gradient operator. It is closely related to taking the Fourier transform of the equation as a whole:

$$(5.4) \qquad \int dz P(z, \nabla_z)u(z)e^{\xi^T z} = P(\nabla_\xi, \xi)\tilde{u}(\xi) \ .$$

If the dependence on z of P is limited, say $P(z, \xi)$ does not vanish sign for any z given $\xi \neq 0$.

A number of well-known constant differential operators P are listed in Table 5.1. They are used in physical models which are isotropic, which means that the space is homogeneous and the equations do not depend upon orientation and translation of the coordinates.

5.1. Uniqueness and existence

For a well-defined mathematical problem uniqueness and existence are essential. For the linear wave equation, Eq. 5.3, the energy norm will lead directly to these results. [Courant and Hilbert, 1961, Garabedian, 1964, Sakamoto, 1982]

The uniqueness is the simplest condition. Given the partial differential equation, Eq. 5.3:

$$(5.5) \qquad \Box u(t, z) = 0, \ \ z \in \Omega, \ \ t \geq 0$$

name	equation	dimension	phenomena
gradient	$\nabla = (\partial_1, \partial_2, \partial_3)^T$	$n \times 1$	force
curl	$\nabla \times = \epsilon_{ijk}\partial_k$	3×3	rotation
divergence	$\nabla \cdot = (\partial_1, \partial_2, \partial_3)$	$1 \times n$	flux
Laplacian	$\Delta = \partial_1^2 + \partial_2^2 + \partial_3^2$	1	electrostatics
biharmonic	$\Delta\Delta$	1	elasticity, fluid
Laplace-Beltrami	$\Delta = \sum_{ij} \sqrt{g}^{-1}\partial_i(\sqrt{g}g^{ij}\partial_j)$	1	curved spaces
D'Alembertian	$\Box = v^{-2}\partial_t^2 - \Delta$	1	wave equation
parabolic	$\partial_t + \Delta$	1	heat equation
Helmholtz	$\Delta + k^2$	1	scattering equation

TABLE 5.1. Some named differential operators ($\sqrt{g} = \sqrt{\det|g_{ij}|}$).

for an unknown solution u, with initial conditions:

$$(5.6) \qquad \begin{array}{rcl} u(0, z) & = & u_0(z) \\ \partial_t u(0, z) & = & u_1(z) \end{array} \; , \;\; z \in \Omega$$

where u_0 and u_1 are the given initial values, and boundary conditions, for boundary value f:

$$(5.7) \qquad u(t, z) = f(t, z), \;\; z \in \partial\Omega, \;\; t \geq 0 \;\; ,$$

for the boundary $\partial\Omega$ of the domain Ω. The uniqueness means that if two solutions $u = v$ and $u = w$ satisfy the initial and boundary values, u_0, u_1, and f, they are identical $v = w$. Considering their difference $v - w$, the corresponding initial and boundary conditions vanish: $u_0 = 0, u_1 = 0$, and $f = 0$. Hence, if $u = 0$, for vanishing boundary conditions, the solution u for arbitrary boundary conditions is indeed unique. This relies on the linearity of the partial differential equation.

The energy norm is a positive functional of the function u at a particular time t:

$$(5.8) \qquad E[u] = \frac{1}{2}\int_\Omega dz(\partial_t u)^2 + (\nabla_z u)^T(\nabla_z u) \geq 0 \;\; .$$

The vanishing initial conditions also means that at $t = 0$ the energy is zero. The time-derivative of the integral over the energy can be formulated in terms of the wave equation and boundary terms:

(5.9)

$$\partial_t E[u] = \int_\Omega dz\partial_t u\partial_t^2 u + (\nabla_z u)^T(\nabla_z \partial_t u) = \int_\Omega dz\partial_t u(\Box u) + \int_{\partial\Omega} dz\partial_t u\mathbf{n}^T(\nabla_z u) \;\; .$$

The last integral; the boundary integral, vanishes if $\partial_t u = 0$, or $\mathbf{n}^T(\nabla_z u) = 0$ on the boundary. These lead to the Dirichlet boundary condition, described here, and the Neumann boundary conditions. The first term vanishes if u is a solution to the wave equation $\Box u = 0$. Hence $E = 0$ for $t \geq 0$, and thereby $u = 0$, which proves uniqueness.

Existence of a solution u in combination with boundary conditions is a rather complex problem. Often it is solved by construction. If one is able to construct a solution $u(t, z)$, given $u_0(z), u_1(z)$ and $f(t, z)$ the existence follows. The simplest approach is finite forward integration, using the analyticity. Given $\Box u = 0$ the

evolution given by the Euler finite-step δ integration is:
(5.10)
$$\partial_t \left(\begin{array}{c} \partial_t u \\ u \end{array} \right) = \left(\begin{array}{c} \Delta u_0 \\ u_1 \end{array} \right) \rightarrow \left(\begin{array}{c} u_1(t=\delta) \\ u_0(t=\delta) \end{array} \right) = \left(\begin{array}{c} u_1(t=0) + \delta \Delta u_0(t=0) \\ u_0(t=0) + \delta u_1(t=0) \end{array} \right) \quad .$$

The solution at an infinitesimal time $t = \delta$ satisfies the wave equation $\Box u = \Box \partial_t u = 0$, but not necessary the boundary conditions $u|_{\partial\Omega} = f$. This can be corrected by adding at every time t a solution $\Delta u_b = 0$ to the Laplace equation such that $u|_{\partial\Omega} + u_b|_{\partial\Omega} = f$. The Green function on a finite domain, Section 5.8, is constructed along similar lines. The solution u_b does not contribute to the forward evolution in the next Euler step, hence for static boundary conditions the problem reduces to finding boundary conditions for the Laplace equation or its elliptic equivalent, which may still be formidable. [Bitsadze, 1968, Wendland, 1979] In a more modern language, the existence question concerns the stability of the evolution operator, on the right-hand side of Eq. 5.10. In the dynamic case, with time-dependent boundary conditions we use the Hamiltonian dynamics to construct allowed boundary conditions.

A particular interesting feature of hyperbolic differential equations is the time-reversal invariance. If t is replaced by $-t$, the same equation arises. In some cases, such as electrodynamics the dual equation arises, which correspond to a simultaneous spatial reflection $\mathbf{z} \rightarrow -\mathbf{z}$. Time reversal links existence and uniqueness. The forward time evolution yielded the existence, as we saw above. The same, or dual, evolution applied in the backward time direction will give the admissible initial and boundary conditions for that state.

5.2. Sturm-Liouville theory

The description of dynamical system in terms of partial differential equations, such as the wave equation, Eq. 5.3, is only a particular choice of description. In many cases, for analysis of the solution and stability, it is more appropriate to use a weak formulation in terms of norms, such as Sobolev spaces. [Hörmander, 1990] One choice is particular relevant for physical system. This choice is the energy norm. It can also be applied to nonlinear systems. The mapping from the partial differential equation to a minimization, or optimization, problem stems from the days of Bernoulli. However, a practical method, also for solving inhomogeneous equations was developed by Sturm and Liouville.

Sturm and Liouville published in the mid-nineteenth century a series of papers on differential equations in the journal founded by Liouville, [Sturm, 1836, Liouville, 1836] which, together with the work of Heaviside and Laplace-Fourier, modernized the approach to differential equations, in the sense that they constituted a method of solving such equations. One could argue the work was the forerunner of modern functional analysis.

Instead of solving a single differential equation at the time, they studied an associated eigenvalue problem. The expansion into eigenfunctions can be used to solve both inhomogeneous equations and time-dependent problems. The Sturm-Liouville problems are self-adjoint, such that the eigenvalues are real. The boundary conditions must be chosen such that the formally self-adjoint problem, from inspecting the differential equation, remains self-adjoint for a particular choice of boundary conditions.

Many well-known special functions, such as Bessel and Hankel functions, Legendre polynomials, etc., can be expressed as solutions of the Sturm-Liouville problem. [Sommerfeld, 1992, Kamke, 1962, Courant and Hilbert, 1961] Modern theory of functional analysis of differential equations is still very much based on the Sturm-Liouville theory. The boundary conditions are restricted such that the problem is self-adjoint. In practice it means no energy can be transferred in or out of the system. If the self-adjoint operator is positive, a quadratic functional can be constructed which is the energy function of the system. Allowing the energy to change by the boundary conditions breaks the positivity and the self-adjointness. At worse a boundary integral may be added to the total energy. [Sakamoto, 1982] This boundary integral is the power transferred in or out of the system. In the framework of an action integral over space and time, consistent formulations can be made. We will seek extensions of the Sturm-Liouville theory in a different manner. The solution space will be separated in a boundary part, and a Sturm-Liouville part with vanishing boundary conditions.

The Sturm-Liouville differential equation is:

$$(5.11) \qquad -\frac{\partial}{\partial z}\left(p(z)\frac{\partial}{\partial z}f(z)\right) + q(z)f(z) = \lambda f(z) \ .$$

The functions $p(z)$ and $q(z)$ can be interpreted as weights or measures, if they are positive definite. The parameter λ is the eigenvalue. The associated operator is:

$$(5.12) \qquad L = -\frac{\partial}{\partial z}p(z)\frac{\partial}{\partial z} + q(z) \ ,$$

and given an interval $z \in [a,b]$ and boundary conditions, such as, for example, $f(a) = f(b) = 0$, the corresponding weak formulation is:

$$(5.13) \qquad J[f] = \langle f, Lf \rangle = \frac{1}{2}\int_a^b dz\, p(z)\left(\frac{\partial}{\partial z}f(z)\right)^2 + q(z)\left(f(z)\right)^2 \ .$$

In the standard Sturm-Liouville theory $q(z)$ can be negative, we, however, want to use Sturm-Liouville theory for energy functionals and require a positive $q(z)$. The smallest eigenvalue λ_0 in that case is larger than zero:

$$(5.14) \qquad 0 < \min J[f] = \lambda_0 \int_a^b dz\,(f_0(z))^2 = \lambda_0 \|f_0\|^2 \ ,$$

where $f_0(z)$ is the eigenfunction of the Sturm-Liouville operator, corresponding to the eigenvalue λ.

The Sturm-Liouville operator is self-adjoint and generates a spectrum:

$$(5.15) \qquad Lf_i(z) = \lambda_i f_i(z) \ ,$$

where $\lambda_i < \lambda_j$ for $i < j$. The set of eigenfunctions $\{f_i\}$ span the full function space, in the sense that any continuous function can be approximated by a sum of eigenfunctions, with arbitrary accuracy. Therefore in the case of an inhomogeneous equation with the inhomogeneous term $g(z)$ the solution $f(z)$ is easily found:

$$(5.16) \qquad -\frac{\partial}{\partial z}p(z)\frac{\partial}{\partial z}f(z) + q(z)f(z) = g(z) \ .$$

The spectral decomposition yields the solution:

$$(5.17) \qquad f(z) = \sum_{i=0}^{\infty} \frac{\langle f_i, g \rangle}{\lambda_i} f_i(z) \ .$$

The series generally converges due to the increasing eigenvalues λ_i.

This result can also be expressed as the integral over a Green function G:

$$(5.18) \qquad f(z) = \int dz G(z, z') g(z') \quad .$$

where the Green function, or fundamental solution is given in terms of the spectral decomposition consisting of the pairs of eigenfunctions f_i and eigenvalues λ_i:

$$(5.19) \qquad G(z, z') = \sum_{i=0}^{\infty} \frac{f_i(z) f_i(z')}{\lambda_i} \quad .$$

The Green function $G(z, z')$ is the inverse of the Sturm-Liouville differential operator L. See also Sections 5.7 and 5.8.

The Sturm-Liouville theory forms also the basis for time-dependent analysis. The eigenvalue λ may be replaced by time-derivatives:

$$(5.20) \qquad \lambda \to \left\{ \begin{array}{ll} \partial_t & \text{parabolic} \\ -\partial_t^2 & \text{hyperbolic} \end{array} \right. \quad .$$

These systems may be analyzed in space-time, but also in frequency ω or Laplace s variables. The eigensystem corresponds to the the decay modes and vibrational modes for parabolic and hyperbolic systems respectively.

The best known eigenfunctions of a Sturm-Liouville problem are the expansion in term of trigonometric functions, in the case $p = $ constant and $q = $ constant. On the domain $z \in [0, L]$ the eigenfunctions of the differential operators ∂_z^{2j} are:

$$(5.21)$$

$$f_i(z) = \left\{ \begin{array}{llll} \sin \frac{\pi(2i+1)z}{2L}, & f(0) = 0, & \partial_z f(L) = 0, & \lambda_i = (-1)^j \left(\frac{\pi(2i+1)}{2L} \right)^{2j} \\ \cos \frac{\pi(2i+1)z}{2L}, & \partial_z f(0) = 0, & f(L) = 0, & \lambda_i = (-1)^j \left(\frac{\pi(2i+1)}{2L} \right)^{2j} \\ \cos \frac{\pi i z}{L}, & f(0) = 0, & f(L) = 0, & \lambda_i = (-1)^j \left(\frac{\pi i}{L} \right)^{2j} \\ \sin \frac{\pi i z}{L}, & \partial_z f(0) = 0, & \partial_z f(L) = 0, & \lambda_i = (-1)^j \left(\frac{\pi i}{L} \right)^{2j} \end{array} \right. \quad .$$

The different boundary conditions yield different eigenfunctions f_i. Each set is, however, orthogonal on the domain $[0, L]$:

$$(5.22) \qquad \int_0^L dz f_i(z) f_j(z) = \frac{1}{\lambda_i} \int_0^L dz f_i(z) \partial_z^{2j} f_j(z) = \frac{L}{2} \delta_{ij} \quad .$$

The vanishing boundary conditions allowed us to map the strong problem of finding a solution to a second-order differential equation, to the weak problem of finding a minimizing function f_0. The boundary terms vanish under the partial integration. However, more general boundary conditions exist for the Sturm-Liouville theory. The boundary terms from partial integration are:

$$(5.23)$$
$$\theta(h, f) = \langle \partial_z h, p(z) \partial_z f \rangle - \langle h, -\partial_z p(z) \partial_z f \rangle = p(b) h(b) \partial_z f(b) - p(a) h(a) \partial_z f(a) \quad .$$

Hence, in the case of $f(a) = 0$ or $\partial_z f(a) = 0$ and the same for the other boundary condition at b, the boundary terms vanish and the weak and the strong formulation are equivalent on a finite domain. Furthermore, in the case of boundary conditions

of the form $c_a f(a) = \partial_z f(a)$, the weak formulation can be symmetrized:

(5.24)

$$J[f] = \langle f, Lf \rangle = -c_a p(a) f(a)^2 + c_b p(b) f(b)^2 + \int_a^b dz\, p(z) \left(\frac{\partial}{\partial z} f(z) \right)^2 + q(z) \left(f(z) \right)^2 \ ,$$

which yields a positive definite functional $J[f]$ if $c_a \leq 0$ and $c_b \geq 0$.

Hence, in the general theory of systems described by partial differential equations, those which satisfy a positive energy norm $J[f] \geq 0$ form a restricted set. However, the restriction is quite natural for physical systems, and allows one to handle nonlinear systems where the stability is governed be the energy conservation. If the domain of solutions $f \in \mathcal{S}$ for $J[f] < E_0$ is restricted, the system remains in this domain due to energy conservation under evolution.

5.3. Initial conditions and boundary conditions

The study of partial differential equations of a dynamical systems has a long and slightly troubled history as a mathematical problem. [Hadamard, 1923, Bernstein, 1950, Kreiss, 1970, Hörmander, 1990, Sakamoto, 1982, Gårding, 1998, Garabedian, 1964, Highdon, 1986, Lebeau, 1992, Majda and Osher, 1975] The recent results are based mainly on the approaches with linear operators. The evolution of a system in infinite space is called the Cauchy problem. The well-posedness of the Cauchy problem refers to the existence and uniqueness of the forward evolution of a systems for some specified initial conditions, typically at $t = 0$. In a finite space, with boundaries, and boundary conditions one typically speaks of the initial boundary value problem, or the mixed boundary value problem, or even the hyperbolic boundary value problem.

Our particular interest is in the initial boundary value problem, since we want our system to interact with its surroundings. The standard formulation fixes the boundary condition from the start, which leaves a very limited interaction with the surroundings. In the more general case we will not speak of boundary conditions but boundary impedances, so indicate the two-way interaction of the dynamical, evolving system with its surrounding, through the boundary. [Hogan, 1985] The nature of the well-posedness changes drastically, because of the boundary impedance. The energy norm forms the basis of the well-posedness. Originally energy arose as a tool to prove well-posedness, in this case it defines the appropriate forms for boundary impedance.

Given the hyperbolic differential equation, Eq. 5.3:

(5.25)
$$\ddot{\phi} = \Delta \phi \ ,$$

the evolution is not uniquely specified by $\phi(z, t = 0)$. A second kinematical variable $\psi(z, t = 0)$ is required to specify sufficient initial conditions for unique evolution. e.g.:

(5.26)
$$\psi(z, t) = \dot{\phi}(z, t) \ .$$

The system is now given by two first-order equations:

(5.27)
$$\begin{pmatrix} \dot{\phi} \\ \dot{\psi} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ \Delta & 0 \end{pmatrix} \begin{pmatrix} \phi \\ \psi \end{pmatrix} \ .$$

However, likewise we might have chosen as second variable:

(5.28)
$$\Delta \psi(z, t) = \dot{\phi}(z, t) \ ,$$

such that the first order equations of motion are:

$$(5.29) \qquad \begin{pmatrix} \dot{\phi} \\ \dot{\psi} \end{pmatrix} = \begin{pmatrix} 0 & \Delta \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \phi \\ \psi \end{pmatrix} \quad .$$

The two types of differential equations, Eq. 5.27 and Eq. 5.27, corresponding to different choices of ψ are related to the typical choice of boundary conditions. The Laplacian Δ acts on one of the two fields ϕ or ψ, which should be complemented with boundary conditions in the same fields. The choice of boundary conditions determines the form of the equations. In modelling this choice is associated with causality, and therefore not an intrinsic property of the system, or component, but of its interaction. [Paynter, 1961, Karnopp et al., 2000]

In the port-Hamiltonian approach, [Van der Schaft and Maschke, 2002] related to the systems of conservation laws, [Serre, 2001, Dafermos, 1999] usually a symmetric form for the first order equations is chosen:

$$(5.30) \qquad \nabla \cdot \psi(z, t) = \dot{\phi}(z, t) \quad ,$$

such that the equations of motion are similar in structure to the Hamilton equations of motion:

$$(5.31) \qquad \begin{pmatrix} \dot{\phi} \\ \dot{\psi} \end{pmatrix} = \begin{pmatrix} 0 & \nabla \cdot \\ \nabla & 0 \end{pmatrix} \begin{pmatrix} \phi \\ \psi \end{pmatrix} \quad .$$

The differences in the choice of second kinematical variable $\psi(z, t)$ will play mainly a role in the boundary conditions on $\partial\Omega$ if the functions are specified on a domain $z \in \Omega$. The spatial differential operators, like the Laplacian Δ, the gradient ∇ and the divergence $\nabla\cdot$, on the right-hand side of the equations of motion have non-trivial null-spaces. Hence an arbitrary function in the null-space will not affect the forward evolution, and the backward evolution is specified up to solutions in the null-space. In order to arrive at a well-defined set of equations of motion with an existing and unique solution for the initial conditions $\phi(z, t = 0)$ and $\psi(z, t = 0)$, and boundary conditions $B(\phi(z \in \partial\Omega, t), \psi(z \in \partial\Omega, t)) = 0$, must yield a non-singular evolution operator.

In the case of the Laplace operator, Dirichlet and Neumann boundary conditions yield a well-posed problem for the time-independent case:

$$(5.32) \qquad \begin{array}{ll} \Delta f(z) = 0 & z \in \Omega \\ B f(z) = h(z) & z \in \partial\Omega \end{array} \quad ,$$

where B is the boundary operator, and h is the boundary value or function, in the case of Dirichlet boundary conditions $B = 1$, and in the case of Neumann: $B = \mathbf{n}^T \nabla = \partial_\nu$, where $\mathbf{n}$ is the surface normal.

In Section 5.1 we saw that existence could be proven by constructing an explicit forward integration, using for example the Euler integration. If both the forward and the backward evolution must exist, it means the Euler integration yields a bijective mapping. In the Section 5.1 we supplemented the evolution on the domain defined by the Laplacian, by a solution of the Laplace equation $\Delta u_b = 0$, which corrected deviations from the desired boundary conditions. Hence the null-space of the Laplacian supplied additional degrees of freedom in the function space. However, the natural course of action based on evolution operators is to attempt to construct a nonsingular operator. This construction is well-defined for finite dimensions, but rather ambiguous in the infinite-dimensional case. Take for example the domain $z \in [-1, 1]$ and the basis $\{z^j\}_{j=0}^N$. The differential operator ∂_z^2 lowers the power by
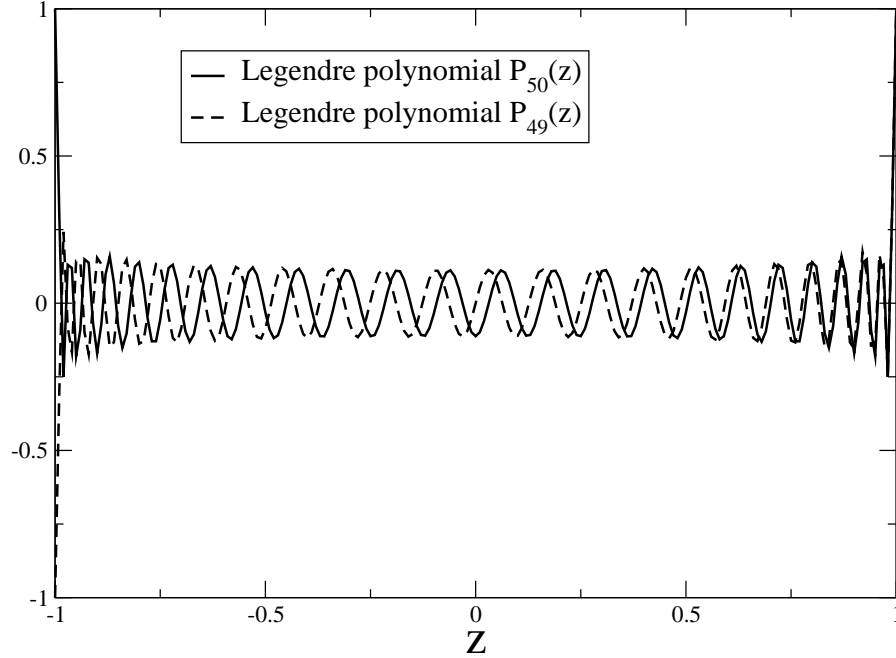
FIGURE 5.1. The two polynomials perpendicular, in Lebesgue inner-product, to the image of the Laplace operator, ∂_z^2, on the polynomial space $\{z^j\}_{j=0}^{50}$. They localize at the boundaries $z = \pm 1$, where they are normalized $P_n(+1) = 1$.

two: $\partial_z^2 \{z^j\}_{j=0}^N = \{z^j\}_{j=0}^{N-2}$. Based on the Lebesgue $\mathcal{L}^2$ norm, the boundary functions, the extensions of the boundary values, are the Legendre polynomials $L_N(z)$ and $L_{N-1}(z)$, which do tend to localize at the boundaries for high values of N. See Figure 5.1.

Hence in the infinite-dimensional function space one can consider a bijective mapping:

$$(5.33) \qquad (\Delta f(z), h(z)) \leftrightarrow_{\text{bijection}} f(z) \quad,$$

where $h(z)$ is defined on the boundary, which may possibly be extended to the whole domain.

This bijection is not a priori evident, and may depend, as noted, on subtleties of the function space we consider. The boundary conditions $h(z)$ augment the range space $\Delta f(z)$, such that for every function $f(z)$ there is a unique decomposition in these functions, which lies at the basis of dynamics.

In the case of dynamical equations with the two initial conditions $\phi(z, t = 0)$ and $\psi(z, t = 0)$, and the appropriate differential equations:

$$(5.34) \qquad \begin{pmatrix} \dot{\phi} \\ \dot{\psi} \end{pmatrix} = \begin{pmatrix} 0 & D \\ -D^c & 0 \end{pmatrix} \begin{pmatrix} \phi \\ \psi \end{pmatrix} \quad,$$

where D is the spatial differential operator, arising from the problem at hand, such as $D = \nabla$ as gradient operator and $D^c = \nabla\cdot$ as divergence operator in Eq. 5.31. The two right-hand sides could be mapped back to the full space, with the use of boundary conditions $h(z)$ and $g(z)$:

$$(5.35) \qquad \begin{pmatrix} (D^c\phi(z), h(z)) \\ (D\psi(z), g(z)) \end{pmatrix} \leftrightarrow_{\text{bijection}} \begin{pmatrix} \psi(z) \\ \phi(z) \end{pmatrix} \quad .$$

The problem is complete in the sense that the kinematical variables $\phi(z)$ and $\psi(z)$ can be expressed in terms of their evolution: $D\psi(z)$ and $D^c\phi(z)$ together with the boundary conditions, $g(z)$ and $h(z)$.

The wave equation on a string and the transmission line are typical one-dimensional ($z \in [0, 1]$) examples of systems where two boundary conditions, say $h(0)$ and $g(1)$, are used to complete the system. In that case $D = D^c = \partial_z$, the one-dimensional null-spaces $\phi(z) = $ constant and $\psi(z) = $ constant each require a single boundary condition. For the transmission line, an applied potential $h(0)$ at the end $z = 0$ will cause a current $g(1)$ at the other end $z = 1$, not only depending on the properties of the transmission line, but also on the impedance at $z = 1$. The two extreme cases of an open end and a short-circuited end at $z = 1$ will both lead to complete reflection at $z = 1$ of the signal from $z = 0$, in the second case with a phase shift: $\cos\pi = -1$.

These mixed boundary conditions, with both nontrivial $h(z)$ and $g(z)$, are typical for the analysis in an operational setting. If the component is part of a greater network, it has a certain functionality, which leads to mixed boundary conditions associated with a signal flow, or Green function or source-and-sink view. [Hirsch and Smale, 1974] Apart from a response view, with general impedance, the actual choice of mixed boundary conditions are not the standard mathematical boundary conditions, which are more likely be associated with either a $h(z)$ or a $g(z)$, not a combination. Of course it is possible to put a potential on each end of the transmission line, $h(0)$ and $h(1)$, it determines its resistance and capacitance, but there is often little purpose to do so. It determines the resistance, and not the transmission properties. For a string these are the typical boundary conditions to determine vibrational modes and frequencies.

The completeness of the stationary problem in terms of the kinematical variables is required for a well-defined dynamical problem, an initial boundary value problem. However, it does not necessarily lead to a well-defined problem. [Serre, 2001, Sakamoto, 1982, Gårding, 1998] Especially the nonlinear hyperbolic problem in more than one space dimension is an open problem, where existence and uniqueness are not a priori clear. On the other hand stability analysis based on contracting maps and Lyapunov functions, typically for discrete systems, yields a well-defined problem in the sense of an energy norm. We will augment the equations of motion with a compatible energy norm. The boundary conditions express the appropriate energy flux, or power flow, for this norm.

Given the energy norm, the possibility to find the associated power-flow boundary impedance depends on the possibility to formulate it as an action integral with a space-time divergence. This approach is similar to the $\mathcal{E}$-well-posedness of [Sakamoto, 1982]. The construction of the energy-momentum tensor goes along similar lines. [Felsager, 1981, Landau and Lifschitz, 1987a] Although, in the general argument of the conserved energy-momentum tensor, one relies on the translational invariance, which, in the case of a domain Ω with boundaries $\partial\Omega$ is not possible, as
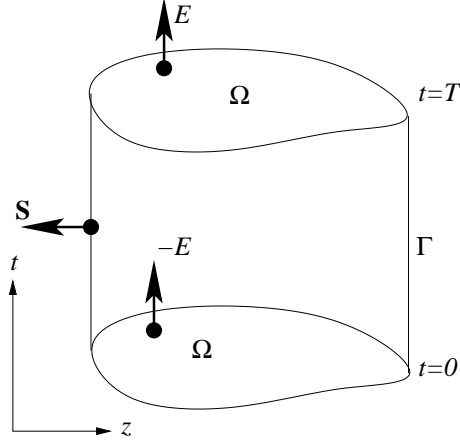
FIGURE 5.2. The space-time tube which combines the initial and boundary conditions. The bottom and top cap $t = 0$ and $t = T$ correspond to the energy in the system, while the boundary $\Gamma = \partial\Omega$, corresponds to the power flow. The hyperbolic partial differential equation inside the tube is the Euler-Lagrange equation.

translation will move states in and out of the domain under consideration. Simply said, the divergence terms and the corresponding boundary terms are ignored under partial integration to construct the energy-momentum tensor.

The zero-energy state $E = 0$ corresponds to the null-state:

$$(5.36) \qquad\qquad E[\phi] = 0 \Leftrightarrow \phi(z) = 0 \;\; .$$

Hence, if we can associate to the partial differential equation the continuity of power, i.e., the conservation of energy, the existence and uniqueness follow from the energy norm $E \geq 0$.[Friedrichs, 1954, Courant and Hilbert, 1961, Garabedian, 1964, Friedrichs, 1973]

On a finite domain Ω, the power, or energy flux though the boundary plays an important role:

$$(5.37) \qquad\qquad E_{\text{final}} - E_{\text{initial}} = P_{\text{boundary}} \;\; .$$

We seek an explicit form of P_{boundary} in terms of the fields ϕ, given the energy E. Such a form already exists in the energy-momentum tensor.

The difference in energy between two times $t = 0$ and $t = T$ can be written as the power integral P over the time-derivative of the energy density:

$$(5.38) \qquad P = \int_0^T \left(\int_\Omega \partial_t E \, dz \right) dt = \int_\Omega E \, dz \Big|_{t=T} - \int_\Omega E \, dz \Big|_{t=0} \;\; .$$

The energy density itself does not depend on time, only via its field variables $\phi(z,t)$. If the power is conserved, the energy density is part of a space-time total divergence, the result of the continuity of energy, [Heaviside, 1891] which can be extended to:

$$(5.39) \qquad\qquad \partial_t E + \nabla \cdot \mathbf{S} = 0 \;\; ,$$

where $\mathbf{S}$ is the spatial power flow, or energy flux. Eq. 5.39 is the local energy continuity, for a conserved energy density. Therefore, given the total space-time

divergence $(E, \mathbf{S})$, the full power integral is:

$$(5.40) \qquad P = \int_\Omega E dz \Big|_{t=T} - \int_\Omega E dz \Big|_{t=0} + \int_0^T \int_{\partial\Omega} \mathbf{n}^T \mathbf{S} dz dt \ .$$

See Fig. 5.2.

All the boundaries of the space-time region $\Omega \times [0, T]$ are now included in the power integral. If the energy is conserved, the integral is zero; $P = 0$. Hence, the pair $(E, \mathbf{S})$ must be a total divergence in the lifted space. [Ivey and Landsberg, 2000] The lifted space is the space not only consisting of the fields $\phi(z, t)$, but also of its derivatives $\dot\phi$ and $\nabla\phi$. For the Euler-Lagrange equations of motion it was recognized that the initial conditions do not only depend on ϕ but also on its time-derivative $\dot\phi$, in the space-time generalization, one includes $\nabla\phi$ in order to yield the boundary conditions. The generating functional, the Lagrangian $\mathcal{L}(\phi, \dot\phi, \nabla\phi)$, for the Euler-Lagrange equations of motion, can be generalized to the energy-energy-flux total divergence:

$$(5.41) \qquad \begin{pmatrix} E \\ \mathbf{S} \end{pmatrix} = \begin{pmatrix} \frac{\delta\mathcal{L}}{\delta\dot\phi}\dot\phi - \mathcal{L} \\ \frac{\delta\mathcal{L}}{\delta\nabla\phi}\dot\phi \end{pmatrix} \ .$$

It is easy to show that the corresponding energy density E and energy flux $\mathbf{S}$ satisfy the energy continuity relation, Eq. 5.39. Taking the time derivative of the energy will yield a total divergence $-\nabla \cdot \mathbf{S}$, with the use of the Euler-Lagrange equations of motion. The explicit time-derivative of the energy E is calculated with the use of the differentiation chain rule:

$$(5.42) \qquad \partial_t E = \partial_t \frac{\delta\mathcal{L}}{\delta\dot\phi}\dot\phi - \frac{\delta\mathcal{L}}{\delta\phi}\dot\phi + \frac{\delta\mathcal{L}}{\delta\dot\phi}\ddot\phi - \frac{\delta\mathcal{L}}{\delta\dot\phi}\ddot\phi - \frac{\delta\mathcal{L}}{\delta\nabla\phi}\nabla\dot\phi \ ,$$

with the use of the Euler-Lagrange equations of motion, Eq. 5.43:

$$(5.43) \qquad \frac{\delta\mathcal{L}}{\delta\phi} - \partial_t \frac{\delta\mathcal{L}}{\delta\dot\phi} - \nabla\frac{\delta\mathcal{L}}{\delta\nabla\phi} = 0 \ ,$$

the first two terms on the right-hand side of Eq. 5.42 are replaced by the gradient term in Eq.5.43. For the wave equation with unit material constants $\rho = \kappa = 1$, the first term is absent, and the last term reduces to the Laplacian: $\Delta\phi$. The Laplacian represents the net force of the left and the right force, i.e, the change in the moment variable.

Furthermore, the third and the fourth term in Eq. 5.42 cancel, and we recover the energy continuity relation, Eq. 5.39, with the explicit form of the energy flux, $\mathbf{S}$:

$$(5.44) \qquad \partial_t E = -\nabla \left(\frac{\delta\mathcal{L}}{\delta\nabla\phi}\dot\phi \right) = -\nabla \cdot \mathbf{S} \ .$$

If we choose the straightforward generalization of the momentum π of finite dynamical systems:

$$(5.45) \qquad \pi(z) = \frac{\delta\mathcal{L}}{\delta\dot\phi} \ ,$$

the energy flux reduces to:

$$(5.46) \qquad \mathbf{S} = -\frac{\delta H}{\delta\nabla\phi}\frac{\delta H}{\delta\pi} \ .$$

name	PDE ($z \in \Omega$)	boundary ($z \in \partial\Omega$)
Dirichlet	$\Delta u(z) = 0$	$u(z) = f(z)$
Neumann	$\Delta u(z) = 0$	$\partial_\nu u(z) = g(z)$
Robin	$\Delta u(z) = 0$	$a(z)\partial_\nu u(z) + u(z) = h(z)$
Poincaré	$\Delta u(z) = 0$	$\mathbf{a}(z) \cdot \nabla u(z) + u(z) = h(z)$

TABLE 5.2. Named elliptic boundary conditions

The Lagrange formulation of dynamics naturally yields the energy flux in this form. However, given the Hamilton equations of motion, more general formulations of the energy flux exists. The differential operator associated with continuity is the divergence operator, with the gradient operator $\nabla\phi$ as adjoint, but more differential operators may be used. This problem is pursued further in Chapter 7.

The initial boundary value problem is the problem of finding an energy norm E and boundary conditions B, like in Eq. 5.32, such that the boundary conditions uniquely specify the energy flux $\mathbf{S}$. In the traditional setting, the boundary conditions yield $\mathbf{n}^T\mathbf{S} = 0$ in the linear case with non-vanishing mass density ρ and stiffness κ:

$$\phi(z)|_{\partial\Omega} = 0 \quad \Rightarrow \quad \rho(z)\tfrac{\delta H}{\delta\pi}\big|_{\partial\Omega} = 0$$

(5.47)

$$\partial_\nu\phi(z)|_{\partial\Omega} = 0 \quad \Rightarrow \quad \kappa(z)^{-1}\mathbf{n}^T\tfrac{\delta H}{\delta\nabla\phi}\Big|_{\partial\Omega} = 0 \ \ .$$

However, more general boundary conditions exist. Especially in a control setting, one likes to specify the boundary condition as relation among the fields $\phi(z)$ and $\pi(z)$ at the boundary $\partial\Omega$.

The stationary Dirichlet boundary condition, in terms of energy conservation, can be rewritten as the rate-of-change condition: $\dot{\phi}(z) = 0$ for $z \in \partial\Omega$, indicating that a prestrained state:

(5.48) $$\phi(z,t) = h(z) \quad \text{for} \quad z \in \partial\Omega \ \text{and} \ t \in [0,T] \ \ ,$$

will lead to $\mathbf{S} = 0$. The energy $E \neq 0$ for $h(z) \neq 0$. In the case that the boundary conditions do not change in time, the energy flux $\mathbf{S}$ is automatically zero. The stationary problem can be mapped to an elliptic problem, [Bitsadze, 1968, Wendland, 1979, Krantz, 1992] without a reference to time. The time-dependent problem may be solved by solving the elliptic eigenvalue problem first, and using the eigenstates, which all satisfy the stationary boundary-value problem, to generate dynamics, similar to the Sturm-Liouville approach of Section 5.2. Typical elliptic boundary conditions are listed in Table 5.2.

For a system in a subspace $H(\phi, \pi) < E_0$ of the full state space can be kept in that subspace by the appropriate choice of boundary values, for particular boundary conditions, consistent with the energy flux. The initial boundary value problem, in terms of the energy H, and the energy flux $\mathbf{S}$ is always conditionally stable. The energy can be extracted from the system through the boundary, but not all forms of boundary impedance will keep the energy in the system within bounds. Take, for example, the flow through a tube. The product of pressure and volume flow is the power at the boundary. For a given flow the pressure at beginning and end may vary, yielding a variation of the energy in the system, which may be stored
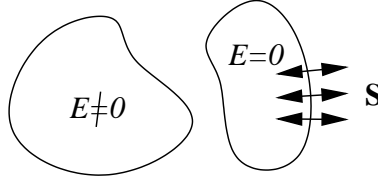
FIGURE 5.3. A simple case where the boundary flux $\mathbf{S}$ may vanish dynamically, although there remains energy on the domain, in an isolated part.

as kinetic energy in the flow, possibly giving rise to turbulence beyond certain boundary values.

5.4. Energy flux and boundary conditions

The energy flux equal zero $\mathbf{S} = 0$ condition, which corresponds to an energy-isolated system, reduces to a variety of well-known boundary conditions in the linear case, as we have seen in the previous section. The more general case, specifying a boundary condition in terms of one of the two boundary efforts:

$$(5.49) \qquad -\frac{\delta H}{\delta \nabla \phi} = h \quad \text{or} \quad \frac{\delta H}{\delta \pi} = g \ ,$$

will also suffice for a well-defined hyperbolic initial boundary value problem, based on Hamiltonian dynamics. The differential form of the energy continuity relation of the space-time volume is the power continuity:

$$(5.50) \qquad E_{\text{final}} - E_{\text{initial}} = P_{\text{boundary}} \Rightarrow \partial_t E = \mathbf{n}^T \mathbf{S} \ .$$

The boundary conditions, in terms of the energy which in turn yields the well-posedness, are proper if either the energy flux $\mathbf{S} = 0$, or if any other value of $\mathbf{S} \neq 0$, for any state (ϕ, ψ) can be achieved by an appropriate boundary value:

$$(5.51) \qquad \forall \phi, \psi, \ \ \forall \mathbf{S} \neq 0 : \quad \exists g, h : \mathbf{S} = \mathbf{S}(\phi, \psi, g, h) \ .$$

This definition allows one to always extract energy from the system if the energy flux is nonzero, driving the energy $E \geq 0$ to zero. If the energy flux is zero $\mathbf{S} = 0$ for any value of f or g, two situations can occur. Either the energy $E = 0$ and consequently the fields $\phi = 0$ and $\psi = 0$, or part of the boundary is isolated from the fields $\phi \neq 0$ and $\psi \neq 0$ in the system causing $E \neq 0$. In the latter case the initial conditions determine the energy retained in the system. See Figure 5.3.

For the case of a string or beam the well-definedness would correspond to the possibility to use boundary values, for the particular boundary conditions, in terms of (g, h), which are absorbing. For example, for position or velocity boundary conditions, the ability to choose the motion of the boundary in the direction of the force exerted by the object on the boundary. Likewise for force boundary conditions to choose the force in the direction of the motion. In both cases, if there is force or motion at the boundary, it is used to extract energy from the system. The extraction of energy means the boundary conditions may yield a contracting map on the state space. This may be the most general boundary-condition well-posedness on an abstract nonlinear energy-norm system.

The product form of the energy flux, Eq. 5.46, yields a set of simple solutions, Eq. 5.49, for the power condition, Eq. 5.51. As long as the nonspecified effort

is nonzero, the boundary condition can be adjusted to extract energy out of the system. The linear results, in the proper formulation as variational derivatives of the Hamiltonian, carry over to the nonlinear theory.

In the case the nonlinear Hamiltonian has several pointlike minima $E_{\min} = H(\phi_{\min}, \psi_{\min}) \geq 0$, the uniqueness is much more complex. In principle, one can define a new local Hamiltonian in the neighborhood of $(\phi_{\min}, \psi_{\min})$ and rely on the local properties:

$$(5.52) \qquad H'(\phi', \psi') = H(\phi_{\min} + \phi', \psi_{\min} + \psi') - H(\phi_{\min}, \psi_{\min}) \ \ .$$

In the case of domain-like minima, like a line or an area, a whole new field of dynamics arises. In the traditional sense, such problems are not well-posed initial boundary value problems.

5.5. A complex boundary-value problem

Not all the problems concerning boundary conditions are discussed in the previous sections. However, physically based boundary condition automatically restrict the problems that may occur. In this section we discuss two problems where the boundary value itself is constrained.

The connection with analytic functions, of complex variables, has been important for the development of the theory of partial differential equations, in particular of elliptic equations, in the nineteenth and the first half of the twentieth century.[Bers et al., 1964] Analytic functions automatically satisfy the Laplace equation, due to the relations between the real and imaginary parts.

The issues concerning boundary-value problems are easily illustrated with a simple elliptic problem in two-dimensions. [Lanczos, 1961] The Cauchy integral equation reconstructs an analytic function $f(z) = u(x+iy) + iv(x+iy)$ from values at the boundary:

$$(5.53) \qquad f(z) = \frac{1}{2\pi i} \int_{\partial\Omega} \frac{f(z')dz'}{z' - z} \ \ ,$$

where each component u and v of f satisfies the Laplace equation: $\Delta u = \Delta v = 0$, since they satisfy the Cauchy-Riemann equations:

$$(5.54) \qquad \frac{\partial u}{\partial x} - \frac{\partial v}{\partial y} = \frac{\partial v}{\partial x} + \frac{\partial u}{\partial y} = 0 \ \ .$$

The Cauchy-Riemann equations can be seen as a linear operator A on a two-dimensional space:

$$(5.55) \qquad A \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} \partial_x & -\partial_y \\ \partial_y & \partial_x \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = 0 \ \ .$$

However, the Cauchy-Riemann equations imply analyticity and therefore integrability of the equations along the boundary $\partial\Omega$, which means the closed integral over the boundary vanishes, for an analytic function f:

$$(5.56) \qquad \int_{\partial\Omega} f(z')dz' = 0 \ \ .$$

Furthermore, for any analytic function $g(z')$ the product $f(z)g(z)$ is again analytic and subsequently these integrals should also vanish:

$$(5.57) \qquad \int_{\partial\Omega} g(z')f(z')dz' = 0 \ \ .$$

Hence the boundary condition $f(z)|_{z \in \partial\Omega}$ must satisfy a condition, Eq. 5.56, itself, actually infinitely many conditions, Eq. 5.57. Part of the equations did not only determine a unique solution on the domain Ω, but also restricted the vector field (u, v) on the boundary $\partial\Omega$. Lanczos, [Lanczos, 1961] in the spirit of linear operators, calls such problems overdetermined. However, it is an overdetermined set of conditions on the boundary, rather than an overdetermined set on the domain. Nowadays, the problem is known as the trace problem. This shows that not all boundary conditions will lead to well-defined problems.

From a single component u or v on the boundary, the other component can be reconstructed, using for example the Hilbert transform. Hence, up to a constant term for the other field, the analytic boundary conditions can be reconstructed, making well-defined boundary condition for the Laplace equation, or the Cauchy-Riemann equations.

A simpler case with a single condition, instead of the infinitely many in Eq. 5.57 above, is the Neumann boundary condition on a compact simply-connected domain, like a disc. [Arnold, 2004] The Dirichlet boundary condition yield a unique solution u for the real component of a harmonic function:

$$(5.58) \qquad \Delta u = 0|_\Omega \qquad u = f|_{\partial\Omega} \quad .$$

A unique solution u is a point in a function space and therefore zero-dimensional. The Neumann boundary condition, on the other hand, yields a one-dimensional solution space since any constant function u_0, with $\partial_\nu u_0 = 0$, can be added to the solution u:

$$(5.59) \qquad \Delta u = 0|_\Omega \qquad \partial_\nu u = g|_{\partial\Omega} \quad .$$

However, there exists a mapping between the real and imaginary part, u and v, of the harmonic function, through the Cauchy-Riemann equation, such that the normal derivative ∂_ν of u corresponds to the derivative along the boundary for v. Hence the Neumann problem can be mapped to the following Dirichlet problem:

$$(5.60) \qquad \Delta v = 0|_\Omega \qquad \partial_s v = g|_{\partial\Omega} \quad ,$$

where s is the coordinate along the boundary. Hence, a single integrability, or constraint, exists for the boundary condition, if the boundary is simply connected, such that:

$$(5.61) \qquad \oint_{\partial\Omega} g ds = 0 \quad ,$$

and a unique boundary condition f exists.

The Laplace operator maps harmonic functions unto harmonic functions: $\Delta h = g$. The index of the Laplace operator on a compact simply connected domain is zero. The index, introduced in Section 2.11, is the difference between the dimensions of the domain and image space of the operator. The Neumann boundary condition led to an one-dimensional null-space, consequently, the boundary condition g has a one-dimensional constraint, to retain the index = 0 of the Laplacian.

For a domain with a hole, the same mapping between the real and imaginary part is no longer possible. For example, there is no single coordinate s describing the distance along the boundary. The index of a differential operator on a domain is related to the topology of the domain. [Grosche et al., 1995] The consequences of boundary conditions on, for example, an annulus, in the theory of elasticity, have made it impossible, as yet, to formulate general boundary conditions with a unique
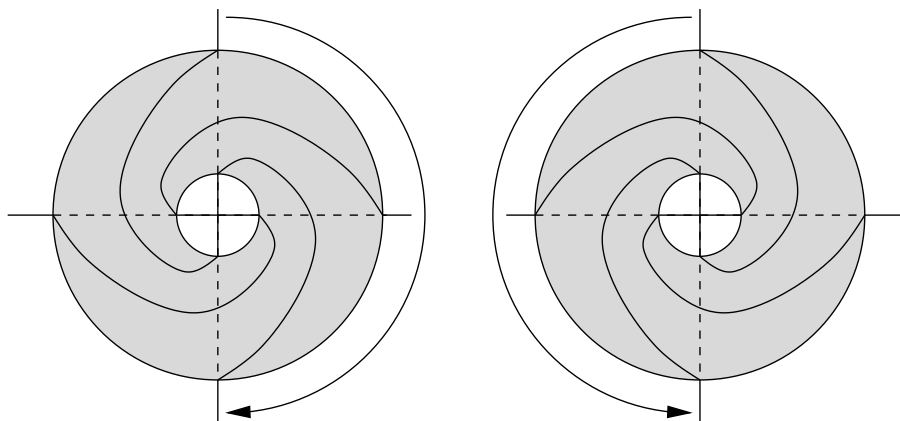
FIGURE 5.4. A simple example where the same boundary conditions may be fulfilled by infinitely many deformation states. For these symmetric boundary conditions, associated with a left and a right-hand turn of 180 degrees, two of the elastic deformation states, with a fixed center and a rotated edge, have the same, lowest energy.

solution for static elasticity problems on domain which is not simply connected. [Truesdell and Noll, 1965, Synge, 1957] See Figure 5.4.

5.6. Lopatinski condition

The best known condition for initial-boundary value problems arises from the theory of elliptic partial differential equations. It has been generalized to the hyperbolic case, but, as much of the older theory on partial differential equations, it relies on a local analytic expansion of the field in terms of the coordinates around a point at the boundary. It is therefore not a global condition, and in most cases the boundaries are rectified to a rectangular grid.

Mixed initial-boundary value problems are hyperbolic partial differential equations with initial data at $t = 0$ on the kinematical domain Ω and boundary data for all times $t \geq 0$ at the domain boundary: $\partial\Omega$. For an n-th order partial differential equation, which means the highest time derivative is ∂_t^n the initial data requires to specify all time derivatives to the order $n - 1$.

The initial-boundary value problem is the problem of finding $f(t, z)$ satisfying the domain condition:

$$(5.62) \qquad L[f] = 0, \quad (t, z) \in \mathbb{R}^+ \otimes \Omega \ ,$$

and the initial value conditions:

$$(5.63) \qquad \partial_t^j f(0, z) = \phi_j(z), \quad j \in \{0, 1, \cdots, n-1\} \ ,$$

and the boundary value conditions:

$$(5.64) \qquad B_j[f] = 0, \quad (t, z) \in \mathbb{R}^+ \otimes \partial\Omega, \quad j \in \{0, 1, \cdots, k\} \ .$$

The number of conditions k the boundary value should satisfy is determined by the Lopatinski condition. For example, for the D'Alembertian; the wave equation in three dimensions, there is a single boundary condition required, for example,

$f(t, z') = 0$ for $z' \in \partial\Omega$. For higher-order differential operators this is no longer trivial. The spatial derivatives should be separated into normal and parallel components, with respect to the boundary $\partial\Omega$. In most cases only the half-space $z_1 \geq 0$, with $\partial\Omega = \{\mathbf{z}|z_1 = 0\}$, is treated. [Hörmander, 1983, Krantz, 1992] Sakamoto [Sakamoto, 1982] treats the unit circle as the standard domain, with the radial coordinate $r = \|\mathbf{z}\|$ as normal coordinate. In general, the boundary should be smooth.

The leading parts, or principal parts, of the differential equation are selected. The highest order derivatives to the time, in each of the terms determine the conditions, just as the n-th order of the operator L, determined the $n-1$ initial conditions. We can express it locally as an analytical function of λ, ξ, κ the Fourier-Laplace transform of $\partial_t, \nabla_\perp, \partial_\parallel$:

$$(5.65) \qquad L_n(t, z_\perp, z_\parallel, \partial_t, \nabla_\perp, \partial_\parallel) \to L_n(t, z_\perp, z_\parallel, \lambda, \xi, \kappa) \quad.$$

The boundary operators can be mapped to analytical functions likewise:

$$(5.66) \qquad B_{j0}(t, z_\perp, z_\parallel, \partial_t, \nabla_\perp, \partial_\parallel) \to B_{j0}(t, z_\perp, z_\parallel, \lambda, \xi, \kappa) \quad.$$

The existence of a solution f to $L_n[f] = 0$ is determined by the solutions of $L_n(t, z_\perp, z_\parallel, \lambda, \xi, \kappa) = 0$. These roots can be separated in two parts, corresponding to characteristics moving away from the boundary, and toward the boundary. The solutions associated to the roots of characteristics moving away from the boundary are related to solutions which need to be determined by the boundary conditions B_j. Hence the number of conditions k should equal the roots $\kappa(t, z_\perp, z_\parallel, \lambda, \xi)$ with $\mathrm{Im}\,\kappa > 0$.

Furthermore, the conditions themselves should not give rise to zeros coinciding with the characteristic roots. For that purpose the function R is constructed which is the determinant of the $k \times k$ matrix with coefficients $R_{j,l}$:

$$(5.67) \qquad R(t, z_\perp, z_\parallel, \lambda, \xi) = \det\left(\oint_\Gamma \frac{B_{j0}(t, z_\perp, z_\parallel, \lambda, \xi, \kappa)\kappa^{l-1}}{L^+(t, z_\perp, z_\parallel, \lambda, \xi, \kappa)}d\kappa\right) \quad,$$

where L^+ is the factor containing only the k positive roots of L, and the contour encircles all these poles. The Lopatinski condition for existence and uniqueness of the solution f to the initial boundary value problem is:

$$(5.68) \qquad \inf_{\substack{t, z_\perp, \mathrm{Im}\,\lambda < 0 \\ |\xi| + |\lambda| = 1}} |R(t, z_\perp, \lambda, \xi)| = c > 0 \quad.$$

The function R for the Lopatinski condition is close to the construction of a local Green function.

5.7. Green function

For a partial differential equation with constant coefficients the Green function approach is often used to solve such equations. In the most appropriate setting distribution theory, or generalized functions are used, together with Fourier-Laplace transform. [Hörmander, 1990, Shilov, 1968, Gel'fand and Shilov, 1964, Schwartz, 1966] Fourier-Laplace transform is, simply said, an extension of ordinary Fourier transform where functions are extended into the complex domain: $(x \in \mathbb{R}^n, \xi \in \mathbb{C}^n)$

$$(5.69) \qquad \mathcal{F}[u](\xi) = \frac{1}{(2\pi)^{\frac{n}{2}}} \int dx\, e^{i\langle \xi, x \rangle} u(x) \quad.$$

In the case of time and space coordinates the conjugate variable of the time component is often denoted by τ. In the case of the wave equation:

$$(5.70) \qquad (\partial_t^2 - \Delta)u(x,t) = 0 \quad,$$

with the initial condition:

$$(5.71) \qquad u(x,0) = g(x), \quad \partial_t u(x,0) = h(x) \quad,$$

the corresponding transformed problem is:

$$(5.72) \qquad (\tau^2 - \xi^2)u(\xi,\tau) = 0 \quad,$$

with

$$(5.73) \qquad u(\xi) = g(\xi), \quad \tau u(\xi) = h(\xi) \quad.$$

The Green functions are the solutions to the fundamental problems:

$$(5.74) \qquad (\partial_t^2 - \Delta)G(x,t) = \delta(x)\delta(t) \quad,$$

and

$$(5.75) \qquad (\partial_t^2 - \Delta)H(x,t) = \delta(x)\theta(t) \quad,$$

where θ is the step function: $\theta(t) = 1$ for $t \geq 0$ and $\theta(t) = 0$ for $t < 0$. The second Green function is also referred to as the Neumann function, or Green function of the second kind, satisfying:

$$(5.76) \qquad (\partial_t^2 - \Delta)\partial_t H = \delta(x)\partial_t\theta(t) = \delta(x)\delta(t) \quad.$$

Instead of treating the initial conditions a part of the solution $u(x,t)$, they are treated as source terms to a homogeneous differential equation in the whole space-time. At $t = 0$ some disturbances occur such that the solution for $t < 0$ is zero, and at $t = 0$ satisfies the initial conditions. These disturbances are therefore singular in nature, but eventually easy to work with as they can be regularized under the integral.

The Cauchy problem can be written as an integral:

$$(5.77) \qquad u(x,t) = \int dx' G(x-x',t)g(x') + H(x-x',t)h(x') \quad,$$

since from standard Heaviside calculus it follows that: $\partial_t\delta(x)\theta(t) = \delta(x)\delta(t)$.

The Green function is easily written down formally:

$$(5.78) \qquad G(x,t) = \frac{1}{(2\pi)^n}\int d\xi d\tau e^{-i\langle\xi,x\rangle - i\tau t}\frac{1}{\tau^2 - \xi^2} \quad.$$

These functions are defined sloppy, especially on the initial surface $t = 0$. Furthermore, the integrals, defining the Green functions, are singular; the poles $\tau = \pm|\xi|$ lie on the real axis. As it turns out, the sloppy definitions hardly hinder the construction of a Green function, and the treatment of the poles determines the choice of homogeneous part, since any function $p(\xi)$ gives a homogeneous solution f, whence integrated with the delta function containing the characteristic equation:

$$(5.79) \qquad f(x,t) = \int d\tau d\xi \delta(\tau^2 - \xi^2)p(\xi)e^{-i\langle\xi,x\rangle - i\tau t},$$

which is a solution to the wave equation in the whole space. The difference between different treatments of the poles are homogeneous functions in the whole space-time. The particular treatment we seek yield solutions only for $t \geq 0$; so-called causal solutions. In the relativistic theory, causality has a different meaning; causality

requires the Green function to be zero outside the light-cone, as no signal can travel faster than the speed of light.[Streater and Wightman, 1964]

The time-derivative part together with causality dictates the treatment of the poles. The simplest way to look at it, is to separate the homogeneous solution into odd and even parts in time. The even part is dictated by $g(x)$, the odd part by $h(x)$. However, odd functions have equal but opposite forward and backward in time, which will requires us to turn the solution causal.

Hence, the odd Green function would given by:

$$(5.80) \qquad \frac{1}{(2\pi)^n} \int d\xi d\tau e^{-i\langle \xi, x-x' \rangle - i\tau t} \frac{1}{\tau(\tau^2 - \xi^2)} \ ,$$

where the singularity at $\tau = 0$ is treated as a principal value integral:

$$(5.81) \qquad \int d\tau \frac{f(\tau)}{\tau} = \lim_{\epsilon \downarrow 0} \int_{-\infty}^{-\epsilon} d\tau \frac{f(\tau)}{\tau} + \int_{\epsilon}^{\infty} d\tau \frac{f(\tau)}{\tau} \ ,$$

which is often more easily expressed as:

$$(5.82) \qquad \int d\tau \frac{f(\tau)}{\tau} = \lim_{\epsilon \downarrow 0} \frac{1}{2} \int d\tau \frac{f(\tau)}{\tau + i\epsilon} + \frac{f(\tau)}{\tau - i\epsilon} \ .$$

Such a small imaginary component $\pm i\epsilon$ to a pole is called a pole prescription.

The pole prescription of the poles $\tau = \pm|\xi|$ follows from the causality criterion: The solution exists for $t \geq 0$. Hence the poles should contribute only to $t \geq 0$ values. For $t \geq 0$, the contour integration must be closed in the $\text{Im}\tau < 0$ plane. Hence, the poles must lie in the lower half plane:

$$(5.83) \qquad \frac{1}{\tau^2 - \xi^2} \rightarrow \frac{1}{(\tau - |\xi| + i\epsilon)(\tau + |\xi| + i\epsilon)} \ .$$

Here we note that the odd Green function would be non-causal, because of the pole in the upper half plane, hence it should be given by poles:

$$(5.84) \qquad \frac{1}{\tau(\tau^2 - \xi^2)} \rightarrow \frac{1}{(\tau + i\epsilon)(\tau - |\xi| + i\epsilon)(\tau + |\xi| + i\epsilon)} \ ,$$

yielding the second Green function:

$$(5.85) \qquad G'(x - x', t) = \frac{1}{(2\pi)^n} \int d\xi d\tau \frac{e^{-i\langle \xi, x-x' \rangle - i\tau t}}{(\tau + i\epsilon)(\tau - |\xi| + i\epsilon)(\tau + |\xi| + i\epsilon)} \ .$$

The complete causal solution is given by:

$$(5.86) \qquad u(x, t) = \int dx' G(x - x', t)g(x') + G'(x - x', t)(h(x') - g(x')) \ ,$$

since:

$$(5.87) \qquad \frac{1}{\tau + i\epsilon} = \frac{1}{2}\left(\frac{1}{\tau + i\epsilon} + \frac{1}{\tau - i\epsilon}\right) + \frac{1}{2}\left(\frac{1}{\tau + i\epsilon} - \frac{1}{\tau - i\epsilon}\right) = \frac{1}{\tau} + \pi i\delta(\tau) \ .$$

5.8. Green function on a finite domain

Above we discussed the construction of a Green function or fundamental solution on an infinite domain. However, in most practical cases, the system will be on a finite domain. Appropriate boundary conditions $h(x,t)$ may not be met by the Green function, similar to the existence argument in Section 5.1:

$$(5.88) \qquad G(x - x', t)|_{x' \in \partial\Omega} \neq h(x', t) \ .$$

However, augmenting the Green function so that it satisfies the boundary condition is an elliptic, stationary problem consisting of two parts: First construct a Green function with vanishing boundary conditions, by adding a homogeneous solution for all times with the same values at the boundary:

$$(5.89) \qquad \Delta g_{x'}(x,t) = 0, \qquad g_{x'}(x,t) = G(x - x', t)|_{x' \in \partial\Omega} \ .$$

Such that the augmented Green function G' has vanishing boundary conditions:

$$(5.90) \qquad G'(x, x', t) = G(x - x', t) - g_{x'}(x,t) \ .$$

Secondly, add to the total solution the quasi-stationary solution to the boundary value problem:

$$(5.91) \qquad \Delta f(x,t) = 0, \qquad f(x',t)|_{x' \in \partial\Omega} = h(x', t) \ .$$

Given a problem with initial values $u(x)_{t=0}$ and boundary values $h(x \in \partial\Omega), t)$, the solution $y(x,t)$ in $\Omega \otimes [0, \infty)$ is given by:

$$(5.92) \qquad y(x,t) = \int_\Omega G'(x, x', t)(u(x') + \partial_t^2 f(x', t))dx' + f(x,t) \ .$$

The time-derivative of the quasi-stationary solution $f(x,t)$ serves as source for the full solution.

5.9. Constrained minimization

In the theory of constrained partial differential equations one often resorts to a method originated by Lagrange, where additional variables, Lagrange parameters or multipliers, are introduced to solve the constraint simultaneously with the constraints. In the modern approach it is called the KKT method. [Karush, 1939, Kuhn and Tucker, 1951, Tröltzsch, 2005] However, in the literature older references to similar approaches exist.[Carathéodory, 1935]

Considering the constrained problem for u:

$$(5.93) \qquad Au = f \quad \text{given} \quad Bu = g \ ,$$

can be rewritten, in the weak formulation, as a Lagrangian:

$$(5.94) \qquad \mathcal{L} = \frac{1}{2}\langle Au, u \rangle - fu + p(Bu - g) \ ,$$

which satisfies the variational criteria:

$$(5.95) \qquad \delta_u \mathcal{L} = \delta_p \mathcal{L} = 0 \ ,$$

where p is the Lagrange multiplier function. The corresponding equations are called the KKT system or the saddle point problem, [Benzi et al., 2005] since it is expected to have positive and negative eigenvalues:

$$(5.96) \qquad \begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \begin{pmatrix} f \\ g \end{pmatrix} \ ,$$

If the operator A is positive definite and, furthermore:

$$(5.97) \qquad \lambda_{\min}(A) > 4\lambda_{\max}(BA^{-1}B^T) \quad,$$

the system is positive definite. The condition above can be seen as the generalized discriminant. In the case of the Stokes flow, the Schur complement $BA^{-1}B^T$ is the identity:

$$(5.98) \qquad BA^{-1}B^T = \nabla \cdot \frac{1}{\Delta}\nabla = 1 \quad.$$

See also Section 8.11, where the Schur complement arises the Guyan projection, in the matrix theory. [Golub and Van Loan, 1996]

5.10. Functional or variational derivative

In dealing with systems described by fields, or functions, the functional derivative plays an important role. See also Section 2.7. However, for nonlinear problems, and the finite approximation of such problems, the functional derivative is not necessarily unique. It depends on the order of the approximations.

A functional J maps a function f to a number:

$$(5.99) \qquad J : f \rightarrow J[f] \in \mathbb{R} \quad.$$

For example, the Lebesgue norm $\mathcal{L}^2_{[-1,1]}$ is a functional:

$$(5.100) \qquad J[f] = \|f\|^2 = \int_{-1}^{1} dz f(z)^2 \quad.$$

The functional derivative links the global functional and the variational equations, such as the Euler-Lagrange or the Hamilton equations for fields. [Felsager, 1981] In the case of stationary problems Sobolev norms are the functionals for elliptic partial differential equations. [Dautray and Lions, 1990] However, careful mathematicians tend to avoid an explicit definition of the variational equations related to the weak problem under investigation. We, on the other hand, keep in mind that our final goal is a consistent, but finite-dimensional approximation, for practical numerical implementation. Consequently, we allow ourselves, in the mathematical description, some latitude.

To define the derivative of a functional, we can proceed in different directions. The directional, variational, or Gateaux, derivative arises from taking a small variation of the functional in the direction g:

$$(5.101) \qquad \delta_g J[f] = \lim_{\epsilon \to 0} \frac{J[f + \epsilon g] - J[f]}{\epsilon} \quad.$$

If such a limit exists, the variational, or functional, derivative can be written as a linear operator on g:

$$(5.102) \qquad \delta_g J[f] = \langle J'[f], g \rangle \quad.$$

The stricter defined Fréchet derivative $J'[f]$, for an arbitrary g, only exists if the norm of the linear operator associated with the directional derivative exists:

$$(5.103) \qquad \lim_{\forall \|g\| \to 0} \frac{\|\delta_g J[f] - J'[f]g\|}{\|g\|} = 0 \quad.$$

The directional derivative exists for many more functionals than the Fréchet derivative will. The same situation applies to ordinary derivatives of multivariate functions:

$$(5.104) \qquad \lim_{\delta x, \delta y \to 0} f(x + \delta x, y + \delta y) \neq f(x, y) + f'_x(x, y)\delta x + f'_y(x, y)\delta y \ \ ,$$

where the vector $F'(x, y) = (f'_x(x, y), f'_y(x, y))$ would be the analogue of the Fréchet derivative. Typical counterexamples are related the integrability condition discussed in Section 4.3.

Besides functional derivatives around particular points, given by the function f, there is a special functional derivative, around the null-function, schematically denoted by:

$$(5.105) \qquad \delta_g J[0] = \delta_g J \ \ .$$

This functional derivative is often encountered in quadratic functionals $J_2[f] = \langle f, Qf \rangle$, where Q is a symmetric linear operator, such that the function is recovered from the variational derivative:

$$(5.106) \qquad J'_2[f] = 2Qf \ \ ,$$

while in general

$$(5.107) \qquad \delta_g J_2[f] = 2gQf \ \ .$$

The corresponding Fréchet derivative is the linear operator: $J'_2 = 2Q$. It should be noted that for nonlinear functionals such relations do not hold.

In the functional derivative we rely heavily on the underlying linear vector space of functions. For example, for the functional which selects the value at a single point $z = z_0$: $J[f] = f(z_0)$, the functional derivative is the distribution:

$$(5.108) \qquad \mathcal{J}' = \delta(z - z_0) \ \ ,$$

in the integral representation, where $\mathcal{J}(z)$ is the integrand corresponding to the functional J.

In most numerical implementations of variational derivatives, the function space is restricted to a finite dimensional space. Even more, for practical purposes a Hilbert space is used, such that the function space and the functional-derivative space are one and the same. As an example we will use the space of normalized Legendre polynomials $\phi_i(z)$ with $i \in \{0, 1, , \cdots, N\}$ on the domain $z \in [-1, 1]$. They satisfy the orthonormality relation:

$$(5.109) \qquad \int_{-1}^{1} dz \phi_i(z)\phi_j(z) = \delta_{ij} \ \ .$$

The symmetric unit operator in the space of Legendre polynomials to order N is a projection operator P on the full space. For any measurable, i.e., integrable function $q(z)$, the projection $g(z)$ on the finite subspace is given by:

$$g(z) = Pq(z) = \sum_{i=0}^{N} \phi_i(z) \int_{-1}^{1} dz \phi_i(z)q(z) \ \ .$$

Hence, any function f and any functional derivative $\delta_g J[f]$ is restricted to an expansion in terms of these polynomials:

$$(5.110) \quad g(z) = \sum_{i=0}^{N} c_i \phi_i(z) \qquad f(z) = \sum_{i=0}^{N} a_i \phi_i(z) \qquad \delta_g J[f] = \sum_{i=0}^{N} b_i(\mathbf{a}, \mathbf{c})\phi_i(z) \ \ ,$$

where $\mathbf{a}, \mathbf{b}$, and $\mathbf{c}$ are the coefficient vectors, associated with $f, \mathcal{J}$, and g. The coefficient vector $\mathbf{b}$ will generally depend of the function f via its coefficient vector $\mathbf{a}$ and the directional derivative direction g through $\mathbf{c}$.

Hence, given a general nonlinear functional, and the functions g and f, the results in the function space must be projected again onto the finite dimensional space. There are two ways to proceed. First, determine the functional in terms of the coefficient vector $\mathbf{a}$, and project the gradient $\nabla_\mathbf{a}$ in terms of this finite dimensional vector $\mathbf{a}$ onto the vector $\mathbf{c}$:

$$(5.111) \qquad J[f] \to J(\mathbf{a}), \qquad \delta_g J[f] \to \mathbf{c}^T \nabla_\mathbf{a} J(\mathbf{a}) \ \ .$$

The second procedure to arrive at the finite-dimensional approximation of the the functional derivative is to calculate the functional derivative first in the infinite-dimensional space, and subsequently project onto the finite function space:

$$(5.112) \qquad \lim_{\epsilon \to 0} \int_{-1}^1 dz \frac{\partial}{\partial \epsilon} \mathcal{J}[f(\mathbf{a}) + \epsilon q(z)] = \int_{-1}^1 dz \delta_{q(z)} \mathcal{J}[f(\mathbf{a})] q(z) \ \ ,$$

where $q(z)$ is an arbitrary measurable function, in the infinite-dimensional space. The projected functional derivative is:

$$(5.113) \qquad P\delta_{q(z)} J[f(\mathbf{a})] \ \ .$$

The function $q(z)$ should also be projected onto the same function space, to yield the coefficient vector $\mathbf{c}$. The two approaches are identical in Hilbert space since the domain of the variational function in the finite function space $Pg = g$, therefore:

$$(5.114) \qquad \int_{-1}^1 dz \delta_{g(z)} \mathcal{J}[f(\mathbf{a})] g(z) = \langle \delta_g \mathcal{J}[f(\mathbf{a})], Pg \rangle = \langle P\delta_g \mathcal{J}[f(\mathbf{a})], Pg \rangle \ \ .$$

Clearly, this relation only holds on a self-dual space, such as the Hilbert space, since then $P^2 = P = P^T$.

From a pointwise comparison it is clear that the results are the same, for sufficiently smooth functionals. In some cases it is easier to take the first route, while in other cases the second route is faster.

For example, for the functional $\mathcal{J}[f] = f^3$, the projected functional is easily determined from the explicit integrals. The Fourier series e^{inz} are even better for such functionals, since the non-zero terms are those with wave numbers summing to zero:

$$(5.115) \qquad \int_{-\pi}^\pi dz e^{inz} e^{imz} e^{ikz} = 2\pi \delta_{k+m+n} \ \ .$$

Similar addition theorems exist for other basis function on compact spaces. [Magnus et al., 1966]

On the other hand, functionals like $\mathcal{J}[f] = e^f$ are more easily evaluated numerically and pointwise:

$$(5.116) \qquad \delta_q J(\mathbf{a}) = \int_{-1}^1 dz \phi(z) e^{\mathbf{a}^T \phi(z)} \approx \sum_i w_i \phi(z_i) e^{\mathbf{a}^T \phi(z_i)} \ \ ,$$

where z_i and w_i are the Gauss-Legendre abscissas and weights. [Abramowitz and Stegun, 1965]

Numerical implementation and the retention of properties of the physical system dictate the approximation. It should be noted that for linear systems many

approximations are equivalent. In the case of nonlinear systems the order in which the different approximations are carried out will affect the results.

Take a simple example of the functional $J[f] = \int_{-1}^{1} dz f^3(z)$, in the subspace $f(z) = a_0 + a_1 z$. We proceed to derive $\nabla_{\mathbf{a}} J$ along two routes. The functional integral in the finite subspace is

$$(5.117) \qquad J[f] \rightarrow J(\mathbf{a}) = \int_{-1}^{1} dz(a_0 + a_1 z)^3 = 2a_0^3 + 2a_0 a_1^2 \ .$$

The variational derivative, along this route, is:

$$(5.118) \qquad \nabla_{\mathbf{a}} J(\mathbf{a}) = \left(\begin{array}{c} 6a_0 + 2a_1^2 \\ 4a_0 a_1 \end{array} \right) \ .$$

The functional derivative, in the abstract function space, is:

$$(5.119) \qquad \partial_{q(z)} J[f] = 3f^2(z) = 3(a_0 + a_1 z)^2 \ .$$

It should be noted that this function lies outside the finite-dimensional subspace, due to the term z^2. The projection onto the two-dimensional function space $\{1, z\}$ is:

$$(5.120) \qquad P\partial_{q(z)} J[f] = \left(\begin{array}{c} \int dz \frac{1}{2} \partial_{q(z)} J[f] \\ \int dz \frac{3}{2} z \partial_{q(z)} J[f] \end{array} \right) = \left(\begin{array}{c} 3a_0^2 + a_1^2 \\ 6a_0 a_1 \end{array} \right) \neq \nabla_{\mathbf{a}} J(\mathbf{a}) \ .$$

Clearly, the two routes to the variational, or functional, derivative in the finite-dimensional function space will lead to different results. In Chapter 11 we will see that consistency conditions, in the general context of the equations of motion, will yield the proper approximation the function space, given the conservation of energy.

5.11. Remarks on partial differential equations

The link between the local differential equation, specifying the solution in a small neighborhood of each point in space, and the global properties is through variational principles. For physical systems these variational principles are in terms of the Lagrangian or the Hamiltonian, or energy density of the system. Some approaches start with field variables and conservation laws and try to construct a second, or closure relation, for a consistent theory, which does not yet have an energy density functional. It might even be difficult to construct such a functional, for example, in the case of chemistry, where systems are often open in terms of the energy flow. [Bird et al., 1960] In this chapter, the approach is followed which keeps a close watch on the energy in the system, for consistency and stability.

Energy conservation, or energy density continuity, guides us to specify general but appropriate boundary conditions, which keep track of the power flowing in and out of the system.

Whether the energy principle restrict the choices of boundary conditions such that inconsistent conditions yielding no solution or multiple solutions for the same conditions is an open question. In many nonlinear cases with multiple local minima in the energy, counter-examples exist for uniqueness. Existence is related to the construction of evolution operators, which in the finite-order approximations studied in the later chapters, seldom is a problem. First, the physical principles and relevant properties of systems described by partial differential equations of this chapter is studied in the next chapter.

CHAPTER 6

Models of infinite-dimensional systems

> According to physics my desk is, for all its seeming fixity and
> solidity, a swarm of vibrating molecules.
> [W.V. QUINE, POSITS AND REALITY]

Infinite-dimensional systems, as defined in Section 2.5, are systems described
by continuous functions, and, typically partial differential equations. None of the
models described here will be done great justice. They serve as advanced examples
of modelling infinite dimensional systems. For a more thorough analysis of each of
the models we refer to the respective literature in these fields.

Many models, whether they describe an elastic beam or the flow in a pipe, have
common features. They satisfy the same physical laws of conservation of energy
and momentum. [Fetter and Walecka, 1980, Landau and Lifschitz, 1987a, Truesdell
and Toupin, 1960] They have similar ways of damping of motion, and they exhibit
similar phenomena, such as wave motion, vibration, and balanced states under ex-
erted forces. The exhibition of these phenomena depends on the regime of external
circumstances. A system with a strong damping will show little vibration, unless
a large driving force at the right frequency is applied. In more complex models,
such as the Navier-Stokes equation, [Landau and Lifschitz, 1987b] different contri-
butions, such as the viscous or damping term, or the nonlinear inertial component,
of the model compete in the force and momentum balance. The domination of one
part, such as viscosity, over another, such as the fluid momentum, can be recovered
from dimensionless constants, in this case the Reynolds number, which combines
fluid properties with system parameters, such as system size.

In a system different effects, such as applied forces, friction, and inertia, are
combined into a force-balance equation:

$$(6.1) \qquad \mathbf{F}_{\text{inertia}} + \mathbf{F}_{\text{damping}} + \mathbf{F}_{\text{potential}} = \mathbf{F}_{\text{applied}} \quad ,$$

where the forces on the left-hand side are determined by the system and the state
it is in. In discrete models these forces are numbers or vectors at worse. In infinite
systems, these forces are generally vector fields; vectors $\mathbf{F}(z)$ at each reference
position z. The force vector fields $\mathbf{F}$ are the dual of the field variables $\phi(z)$, which
describe the state of the system, their product is the work, which yields the local
energy, or the potential part of the Hamiltonian density through the line integral:

$$(6.2) \qquad \mathcal{H}(\phi(z)) = \int_{\phi_0(z)}^{\phi(z)} \mathbf{F}(\phi'(z), z) d\phi'(z) \quad .$$

In this case the reference position z is only an index. The choice of the field
variable ϕ determines the definition of the force, as the energy density is an invariant
quantity. Likewise, the kinetic part of the Hamiltonian defines the momentum field

$\pi(z)$:

$$(6.3) \qquad \mathcal{H}(\dot{\phi}(z)) = \int_{\dot{\phi}_0(z)}^{\dot{\phi}(z)} \pi(\phi'(z), z) d\dot{\phi}'(z) \quad,$$

where the mass density $\rho(z)$ is the local relation, between $\pi(z)$ and $\dot{\phi}(z)$, at the same position z:

$$(6.4) \qquad \pi(z) = \rho(z)\dot{\phi}(z) \quad.$$

Through a variable transformation of field variables $\phi(z)$, the forces $\mathbf{F}(z)$ and the momenta $\pi(z)$ change as well. Their product, however, remains invariant, except under coordinate transformations of the position coordinate z. In infinite systems the coordinate z will also play a role in the differential operators in the Hamiltonian, which intertwines the variable transformations of $\phi(z)$, and the coordinate transformations of z. The invariant power-density products for the change of energy $\mathbf{F}(\phi(z), z)d\phi(z)$ and $\pi(\phi(z), z)d\dot{\phi}(z)$ are retained.

6.1. Beam models

The beam models are important in structural mechanics. There are several beam models in different degrees of sophistication, and for different purposes. The most common beam model is the Euler-Bernoulli beam: [Love, 1944]

$$(6.5) \qquad \rho A \partial_t^2 y(x,t) = \partial_x^2 EI \partial_x^2 y(x,t) \quad,$$

where ρ is the mass density, A is the cross-section area, E is the elasticity modulus, and I the first moment in the bending direction under consideration. The vertical deflection y is a function of the reference position along the beam x and the time t. Beams are structural elements which are typically much longer than they are wide or high. The length l of the beam remains constant, which means that the horizontal position $X(x)$ changes with the vertical position $y(x)$. General deformations are described by the position $\mathbf{X}(\mathbf{x}) = (X, Y)$, as function of a reference coordinate $\mathbf{x} = (x, y)$, where the beam is described by $x \in [0, l]$ and $y = 0$. See also Section 4.2.

In the years 1727 till 1776 Euler made extensive contributions to the theory of beams with several members of the Bernoulli family. [Euler, 1744, Timoshenko, 1953, Truesdell and Noll, 1965] Johann Bernoulli introduced Euler and the rest of the world to variational principles. Also his brother Jakob made significant contributions, but the Bernoulli meant in Euler-Bernoulli equation is the son of Johann: Daniel. Daniel Bernoulli and Euler worked out small displacement statics (the Euler-Bernoulli equation), large-displacement solutions with bifurcations, [Antman and Rosenfeld, 1978] and vibration dynamics with mode shapes.

If the horizontal position is fixed $X(l) = l$, the beam is more likely to act as a string model than a beam model:

$$(6.6) \qquad \rho A \partial_t^2 y(x,t) = \partial_x EA \partial_x y(x,t) \quad,$$

which is an approximation of the elastic energy in the length of the beam. In so-called exact, or Cosserat, models, [Truesdell and Noll, 1965, Cohen and Muncaster, 1988] the beam is approximated by a line $\mathbf{X} = (X(x), Y(x))$ for $0 < x < l$, the second reference coordinate y is absent, and possibly a perpendicular direction $\mathbf{n}$ on that line, which is the orientation. See Figure 6.1. There are only a limited number of invariants of a line, which are the, longitudinal strain, or stretch σ, curvature
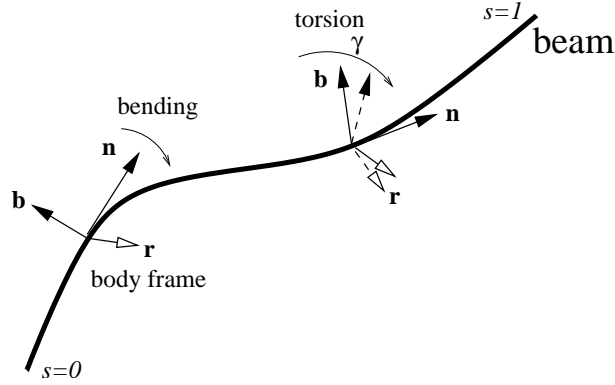
FIGURE 6.1. The beam as a line, function of s, with local body frame coordinates, the Frenet-Serret variables $(\mathbf{n}, \mathbf{b}, \mathbf{r})$. The change of the beam direction $\mathbf{n} = \partial_s \mathbf{X}(s)/|\partial_s \mathbf{X}(s)|$ corresponds to the bending, the body-frame rotation perpendicular to $\mathbf{n}$ yields the torsion. See also Table 6.1.

$\mathbf{k}$, and the torsion γ. There are many different ways to represent these quantities. [Dubrovin et al., 1984] It is important to note that the rate of change of a unit vector $\mathbf{e}(x)$ is perpendicular to that vector. Indeed, since $\mathbf{e}^T \mathbf{e} = 1$,

$$(6.7) \qquad 0 = \partial_x 1 = \partial_x(\mathbf{e}^T \mathbf{e}) = 2\mathbf{e}^T \partial_x \mathbf{e} \quad .$$

One particular elegant way to represent these invariants, except for the stretch, is to use the rotation group $SO(3)$. The change of orientation of the beam, along the beam, can be represented by the rotation of the body-fixed frame:

$$(6.8) \qquad \partial_x R(x) = J(x)R(x) \quad ,$$

where the anti-symmetric $J(x)$ is the generator of rotations. It can be represented by the angle vector $\boldsymbol{\omega}$:

$$(6.9) \qquad [J(x)]_{ij} = \epsilon_{ijk}\omega_k \quad ,$$

where ϵ_{ijk} is the anti-symmetric Levi-Civita tensor. [Frankel, 2004] If the first component of the vector represents the beam direction, $\partial_x \mathbf{X}$, and the second component the beam orientation, $\mathbf{n}$, then ω_2 and ω_3 are the two bending moments in, and perpendicular to, the beam orientation, and ω_1 is the beam torsion. [Lipschutz, 1969] The equations of the different beam models are given in Table 6.1.

One particularly unpleasant feature of the Euler beam model is the absence of a constant or limiting wave velocity, or sound velocity. For most continuous, isotropic media the velocity of sound will reach a limit as the frequency increases. For the beam model, taking a high frequency ω, the wavelength k depends quadratically on this frequency:

$$(6.10) \qquad \rho A \omega^2 = EIk^4 \quad ,$$

which means that the velocity v of a wave increases with frequency:

$$(6.11) \qquad v = \frac{\rho A}{EI}\sqrt{\omega} \quad .$$

name	equations	variables
Euler-Bernoulli	$\partial_t^2 y = \partial_x^4 y$	y: vertical displacement
Timoshenko	$\partial_t^2 W = -\partial_x \psi + \partial_x^2 W$ $\partial_t^2 \psi = \partial_x^2 \psi - \psi + \partial_x W$	W: vertical displacement ψ: rotational displacement
exact, Cosserat	$\mathbf{n}_s + \mathbf{f} = \mathbf{r}_{tt} + \mathbf{b}_{tt}$	$\{\mathbf{n}, \mathbf{b}, \mathbf{r}\}$: orientation
Frenet-Serret	$\mathbf{m}_s + \mathbf{r}_s \times \mathbf{n} + \mathbf{k} = \mathbf{b} \times \mathbf{r}_{tt} + \mathbf{b} \times \mathbf{b}_{tt}$	s: reference coordinate

TABLE 6.1. Beam models

name	condition 1	condition 2
clamped	$y = 0$ (position)	$\partial_x y = 0$ (angle)
free	$\partial_x^2 y = 0$ (torque)	$\partial_x^3 y = 0$ (force)
supported	$y = 0$ (position)	$\partial_x^2 y = 0$ (torque)

TABLE 6.2. Typical boundary conditions for beams. In brackets, the physical variable that is constrained, in the case of position and angle, or has zero input, i.e., is free, in the case of force or torque.

The Timoshenko beam (See also Table 6.1) cures part of this problem by introducing a kinetic energy associated with the bending moment of the beam. [Hansen and Snyder, 1997] However, this leads to two coupled equations, and still the stretch and the shear are not accounted for. For vibrational analysis especially the shear terms will be important if the wavelength approaches the cross-section size of the beam. In that case the bending velocity will become equal to the shear velocity. In the linear theory of elasticity a finite sound velocity arises from a fourth-order time differential equation, leading to two characteristics, i.e., two sound velocities. [Courant and Hilbert, 1961, John, 1982, Taylor and Yau, 2003]

For the modelling of beams the fact that the position of a point along the beam $\mathbf{X}(x)$ depends on the integral over the beam from its fixed position $\mathbf{X}(0) = \mathbf{x}_0$ leads to most problems. In many approximations based on the bending and torsion of the beam, the position will not be given accurately. Furthermore, energy might not be conserved under motion, due to the finite numerical accuracy of the integration over the beam coordinate x. If one has some idea of the possible positions of the beam $\mathbf{X}(x)$, it would be advantageous to use that information and derive the elastic energy and kinetic energy in some finite-dimensional configuration space, i.e., mode shapes. The analysis of the function space $\mathbf{X}(x) \in \mathcal{D}$ is central to the functional system analysis.

The second aspect under investigation in this thesis are the wide variety of boundary conditions. For the Euler-Bernoulli beam there are three well-known boundary conditions. See Table 6.2. In terms of the power products, Eq. 6.2 and Eq. 6.3, these boundary conditions correspond to the absence of two types of energy transfer: the force-velocity and the torque-rotation, in the limited context of the linearized beam model where $y(x) \approx 0$. The beam is therefore isolated from its surroundings, which is possible in a variety of manners.

name	equations	variables
ideal, Euler	$\partial_t \mathbf{u} + (\mathbf{u} \cdot \nabla)\mathbf{u} = -\nabla p$	$\mathbf{u}$: velocity
potential flow	$\nabla(\partial_t \phi + \frac{1}{2}|\mathbf{u}|^2 + w) = 0$	ϕ: flow potential
irrotational	$\mathbf{u} = \nabla \phi$	$w = \int V\, dp$: enthalpy
ideal gas	$\partial_t \rho + \nabla(\rho \mathbf{u}) = 0$	ρ: density
	$\partial_t + (\mathbf{u} \cdot \nabla)\mathbf{u} = -\frac{C}{\rho}\nabla \rho^\gamma$	$\gamma = c_p/c_V$
Stokes,	$\partial_t \mathbf{u} = \nu \Delta u - \nabla p$	$\nu = \mu\rho$: viscosity
viscous	$\nabla \cdot \mathbf{u} = 0$	p: pressure
Navier-Stokes	$\partial_t \mathbf{u} + (\mathbf{u} \cdot \nabla)\mathbf{u} = \nu \Delta u - \nabla p$	
	$\nabla \cdot \mathbf{u} = 0$	
generalized	$\partial_t \rho + \nabla \cdot (\rho \mathbf{u}) = 0$	θ: temperature
Navier-Stokes,	$\partial_t \mathbf{u} + (\mathbf{u} \cdot \nabla)\mathbf{u} = \frac{\mu}{\rho}(\Delta u - \frac{1}{3}\nabla\nabla\cdot)\mathbf{u} - \frac{1}{\rho}\nabla p$	κ: heat conductivity
thermal	$\partial_t \theta + \mathbf{u} \cdot \nabla \theta = \frac{\kappa}{c_V \rho}\Delta\theta + \frac{\theta}{c_V \rho}(\nabla \cdot \mathbf{u})\rho_\theta + \frac{\Psi}{c_V \rho}$	c_V: specific heat
	$\Psi = -\frac{2}{3}\mu(\nabla \cdot \mathbf{u})^2 + \frac{\mu}{2}(\partial_i u_j + \partial_j u_i)^2$	Ψ: dissipation function

TABLE 6.3. Fluid models

6.2. Fluid models

Fluid models are notoriously abstruse. [Mises and Friedrichs, 1971, Landau and Lifschitz, 1987b, Sod, 1985] A wide combination of effects are taking into account in the different models. It already starts with the co-moving coordinate system. The time variation of a quantity at a particular point in the flow, say at a particle that moves along with the stream, depends both on time and the velocity:

$$(6.12) \qquad \frac{d}{dt} = \frac{\partial}{\partial t} + \frac{\partial \mathbf{x}}{\partial t} \cdot \nabla = \frac{\partial}{\partial t} + (\mathbf{u} \cdot \nabla) \ .$$

where $\mathbf{u}$ is the velocity.

Furthermore, fluid has a mass, inertia. Therefore it has the tendency to retain its velocity and direction. Forces on the fluid are either external, through boundaries and pressure, or internal, due to friction and gravity. If the friction is neglected, and the flow dynamics is dominated by the inertia, it leads to the ideal, or Euler, fluid model. This model can also be expressed in an energy relation, which is called the potential model. If, on the other hand, the inertia is neglected, and the forces in the flow only arise through friction, it yields the Stokes flow. If both the friction and the inertia are taken into account, the common model is the Navier-Stokes flow. However, due to the friction, the flow dissipates energy. The mechanical energy is converted into heat. The energy conserving model, including heat generation and flow, for fluid is the generalized, or thermal, Navier-Stokes equation. Table 6.3 gives a list of well-known models for fluids.

Besides the microscopic fluid models, in terms of partial differential equations, fluid systems can also be analyzed globally. The global analysis often leads more quickly to a qualitative answer. One important number which arises in the global analysis is the Reynolds number Re, which indicates if the solution will be laminar, irrotational, or whether eddies and turbulence will arise. The number is a simple

variable	dimensions	typical symbols
length	m	l
time	s	t
mass	kg	m
density	kg m^{-3}	ρ
velocity	m s^{-1}	$v, \mathbf{u}$
acceleration	m s^{-2}	a, g
force	N = kg m s^{-2}	F
pressure	N m^{-2} = kg m^{-1} s^{-2}	p
current	A	I, J
charge	A s	Q, q
temperature	K	T, θ
energy	J = N m = kg m^2 s^{-2}	E, H, W
power	J s^{-1} = kg m^2 s^{-3}	P

TABLE 6.4. dimensional constants

combination of size, velocity and viscosity:

$$(6.13) \qquad \mathrm{Re} = \frac{\text{size } |\mathbf{u}|\rho}{\mu} = \frac{\text{size } |\mathbf{u}|}{\nu} \ .$$

The size in the equation is a typical size of the problem, the diameter of a tube, the width of a channel. Hence the Reynolds number may vary with the choice of typical size. However, this order of magnitude is small compared to the several orders of magnitude which occur in the variety of fluid problems.

In a sense, systems can be mapped to a simple one-dimensional problem of the flow in a pipe, assuming the dimensions, such as length and velocity, can be mapped from one problem to the next. The dynamical regimes of the archetypical system, given by the value of the number, will apply to most systems. However, sometimes care must be taken to choose the right value for the system parameters, such as size or velocity. The flow in a pipe with the radius size = r and a mean velocity u will have typically higher values Re for the regimes as a ball with radius size = r, moving through a stationary fluid with the velocity u. However, these two systems have the same regimes.

The dimensional constants of fluid dynamics, such as the density, viscosity, and conductivity are combined with system parameters, such as size, velocity, and pressure, to yield dimensionless numbers which will give a good indication of the flow solution, not only laminar or turbulence, but also viscosity or inertia dominated, and conduction or convection dominated.

6.3. Dimensional analysis

The construction of an invariant number, such as the Reynolds number, from dimensional constants, as given in Table 6.4, material constants, as given in Table 6.5, and system dimensions can be performed systematically using Buckingham's method. [Welty et al., 1984] Each dimensional constant can be expressed in terms of length (m: meter), mass (kg: kilogram), time (s: seconds), current (A: Ampère), and temperature (K: Kelvin). Expressing the different constants of a problem in terms of the fundamental dimensions, gives a redundancy, which are

variable	dimensions	typical symbols
density	kg m^{-3}	ρ
viscosity	kg m^{-1} s^{-1}	$\mu(\eta)$
kinematical viscosity	m^2 s^{-1}	ν
elasticity modulus	N m^{-2} = kg m^{-1} s^{-2}	E
thermal diffusion	W m^{-1} K^{-1} = kg m s^{-3} K^{-1}	λ
heat capacity	J kg^{-1} K^{-1} = m^2 s^{-2} K^{-1}	c_p

TABLE 6.5. material properties

precisely the number of invariants. Mathematical physicists use this reduction extensively to remove most dimensional constants from the equations. Experimental physicists have come up with a variety of experiments to determine the scaling behavior, i.e., the dimensionless invariants, for the same dimensionless constant, such as Reynolds number. [Van Lammeren, 1944]

In physics and chemistry these dimensions play an important role. They immediately disqualify certain models, due to the incorrect dimensions in the equations. Every term in the force balance equation must have the dimensions of force, Newtons: N = kg m s^{-2}. Likewise, in many cases one seeks the relevant dimension, which causes a physical effect. For example, in deep water the relation between the wave length and the wave velocity of a wave can only be determined by the few physical constants in play: the gravitation and the density, which causes deep water waves to make a fixed angle with a ship of approximately 23 degrees.[Lighthill, 1978] In hadronic physics one still seeks the relevant physical variable of the dimension of length, which causes the strongly interacting elementary particles to be of the size they are. This, yet unknown, process is even referred to as dimensional reduction.

A systematic dimensional analysis of a system is Buckingham's method. It consists of a matrix, where horizontally all the terms of the equations of motion are written, and vertically all the powers of the fundamental dimensions. The rank of the corresponding matrix is the number of *dimensional* constants, the number of columns minus the rank is the number of *dimensionless* constants. Thus equivalently, the null-space of the powers matrix, For example, in fluid dynamics there are two dimensionless constants, the Reynolds and the Euler numbers. The matrix, with force, velocity, density, viscosity, and size, has the rank 3:

$$(6.14) \qquad \begin{pmatrix} & F & v & \rho & \mu & l \\ \hline \text{kg} & 1 & 0 & 1 & 1 & 0 \\ \text{m} & 1 & 1 & -3 & -1 & 1 \\ \text{s} & -2 & -1 & 0 & -1 & 0 \end{pmatrix} \ .$$

The null-space of the matrix gives rise to dimensionless constants. The two dimensionless constants, which can be constructed from these five system dimensions are the Euler number, corresponding to the null-vector $(1, -2, -1, 0, -2)$:

$$(6.15) \qquad \text{Eu} = \frac{F}{l^2 \rho v^2} \ .$$

and the Reynolds number, corresponding to the second independent null-vector $(0, 1, 1, -1, 1)$:

$$(6.16) \qquad \text{Re} = \frac{l \rho v}{\mu} = \frac{l v}{\nu} \ ,$$

name	numerator force	denominator force	equation
Froude	inertia force	gravity force	$\text{Fr} = \frac{u^2}{gl}$
Euler	pressure force	inertia force	$\text{Eu} = \frac{p}{\rho u^2}$
Reynolds	inertia force	viscous force	$\text{Re} = \frac{lu}{\nu}$
Prandtl	momentum diffusion	thermal diffusion	$\text{Pr} = \frac{\nu c_p}{\kappa \rho}$
Nusselt	conductive resistance	convective resistance	$\text{Nu} = \frac{\lambda l}{\kappa}$
Peclet	inertia force	thermal diffusion	$\text{Pe} = \text{Re} \cdot \text{Pr}$
Stanton	viscous force	heat transfer	$\text{St} = \frac{\lambda l}{c_p u \rho}$
Graetz	heat transfer	viscous force	$\text{Gz} = \frac{\rho u c_p}{\lambda l}$
Knudsen	container size	mean free path	$\text{Kn} = \frac{L}{l}$
Grashof	thermal expansion buoyancy	viscous force	$\text{Gr} = \frac{\alpha g T l^3}{\nu^2}$
Schmidt	viscous force	diffusion pressure	$\text{Sc} = \frac{\nu}{D}$
Bond	gravity force	surface tension	$\text{Bo} = \frac{\rho g L^2}{\gamma}$
Biot	surface thermal conductivity	material conductivity	$\text{Bi} = \frac{Rl}{\lambda}$
Fourier	heat flow	heat capacity	$\text{Fo} = \frac{\lambda T}{\rho c_p l^2}$
Mach	speed	sound speed	$\text{M} = \frac{v}{c}$
Rossby	inertial force	Coriolis force	$\text{Ro} = \frac{v}{\Omega l}$
Rayleigh	thermal expansion buoyancy	thermal diffusion	$\text{Ra} = \frac{\alpha c_p \rho_0 g T l^3}{\gamma \lambda}$

TABLE 6.6. dimensionless constants, most of them are empirical relations for particular complex thermofluidic phenomena. The dimensionless constant width over length, d/l, for example for pipe flows, will yield the same constant but for a different effect, in, e.g., St versus Gz.[Holman, 1981]

which makes the introduction of the kinematical viscosity ν natural, to simplify the expression for the Reynolds number. Clearly many other choices of null-vectors are possible, but they may be expressed in terms of the Reynolds and Euler number. The Euler number therefore plays a smaller role in stationary and quasi-stationary situations, which are characterized by the Reynolds number. A small list of well-known dimensionless constants are given in Table 6.6.

The Euler number indicates the systems inertial resistance to motion due to the force. The Reynolds number indicates the balance between viscous force and momentum.

Another, less systematic, but more intuitive way to proceed in the recovering the dimensionless characterizations of a system is by inspecting the equations of motion. Total force in a system are the balance different forces, such as inertial forces, applied forces, frictional forces, and potential forces. The equation of motion is the direct sum of all these terms, and any ratio of two terms give dimensionless number, which indicates the dominance of one term over the other. In Table 6.6 we give a number of common numerator and denominator forces. However, these typical forces may depend on the interpretation. For example, an equation involving pressures and an equation involving forces are fundamentally the same, although in terms of dimensions they differ an area m^2, since pressure times area is force. The interpretation may differ with these dimensions.

6.4. Laminar flow around a cylinder

One of the best known example of dimensional analysis is for the flow around a cylinder. [Feynman et al., 1964] One is not yet able to fully model all the phenomena [Landau and Lifschitz, 1987b] that occur in such flow. Especially the turbulent flow and the transition between turbulent and turbulent boundary flow give rise to many questions. However, the single dimensional constant; the Reynolds number Re determines the different physical regimes, in the sense of typical flow behavior. For the flow around a cylinder the Reynolds number Re is defined as:

$$(6.17) \qquad\qquad \mathrm{Re} = \frac{[\text{density}][\text{velocity}][\text{radius}]}{[\text{viscosity}]} = \frac{\rho v L}{\mu} \quad .$$

The Reynolds number regimes are the following:

< 20 Steady laminar flow, smoothly around the cylinder.
$20 \sim 150$ Laminar eddy flow; behind the cylinder, steady back flow occurs.
$150 \sim 10^5$ Turbulent flow; the eddies separate from the cylinder and move off irregularly.
$< 10^5$ Turbulent boundary layer flow; behind the cylinder the flow is completely irregular like a trailing tail.

Hence, for many different situations, and many different fluids and gases, the same situations occur for the same Reynolds number. Likewise, more complex systems could be characterized by the same regimes. The flow in a pipe or around a corner yields similar flow patterns as the cylinder problem.

6.5. Energy-based discrete fluid model

Different models of systems start with the choice of variables. A different choice of variables may yield a solution where the original equations in the original variables did not. Especially in the case of intrinsically nonlinear systems, such as the full Navier-Stokes equations, the choice of variables may make the world of difference. For example, entropy variables can be used to study the transition from compressible to incompressible fluid dynamics.

In this section, we introduce a different type of discrete model for fluid, based on the energy in the fluid. In many cases, the given variables, such as velocity and pressure, and the given equations, such as the Navier-Stokes equations, are taken for granted as the starting point of the modelling process. Not in all cases these may be the appropriate variables. For example, the standard variables of the Navier-Stokes equation are cumbersome for energy conservation. We show an alternative choice of variables, for which energy conservation can be implemented explicitly, as an example for modelling.

Instead of the pressure we use the energy as extra variable beside the velocity vector $\mathbf{u}$. The velocity field of fluid plays a dual role. First, it determines the geometry; the direction of the flow, and, second, it represents the momentum and the kinetic energy of the flow. Different geometry, such as downstream, or Lagrangian, coordinates are based on the geometric role of the velocity. It is also possible to use a fixed grid with finite cells. Each cell has an associated energy E_i and a mean velocity $\mathbf{v}_i$. The energy is passed on from one cell i to the next j through the mean velocity $\frac{1}{2}(\mathbf{v}_i + \mathbf{v}_j)$. The amount is determined by the area of the common face a_{ij},

the normal $\mathbf{n}_{ij}$, and the volume V_i and V_j:

$$(6.18) \qquad \Delta E_{ij} = \frac{a_{ij}\mathbf{n}_{ij}^T(\mathbf{v}_i + \mathbf{v}_j)}{2V_i}E_i \quad ,$$

where ΔE_{ij} is the energy transferred from cell i to cell j if the velocity is in the direction $\mathbf{n}_{ij}^T(\mathbf{v}_i + \mathbf{v}_j) > 0$.

The energy of the fluid in a cell is mainly the kinetic energy, due to the incompressible nature of fluid:

$$(6.19) \qquad E_i = \frac{1}{2}\rho_i V_i \mathbf{v}_i^T \mathbf{v}_i \quad .$$

In some cases the velocity is reduced in favor of a build up of pressure, or body forces, such as gravitational forces may play a role. The differences between the kinetic energy as given by the velocities $\mathbf{v}_i$ and the energy E_i, which include pressure and potential energy, give rise to forces which will change the momenta.

The viscosity affects the velocities and energies of neighboring cells perpendicular to the direction of velocity. Furthermore, it depends on the differences of the velocities in neighboring cells, instead of the mean, as velocity differences cause shear forces. Here we will not investigate the particular dynamics which arises from these effects. We have showed that energy and velocity, and the basic transfer of energy, Eq. 6.19, due to the velocity, can be expressed in an energy-conserving manner.

In practice separating the different energies, forces, and effects allows us to write a numerical scheme to balance each of this relations. For incompressible flows the compensating pressure should be balanced completely to assure that the energy and velocity are consistent. At each time step an iteration to converge to the incompressibility constraint $\nabla \cdot \mathbf{v} = 0$ may be required. In practice, the incompressibility is never exact, as it would correspond to an infinite sound velocity. However, for all practical purposes, the fluid may be perfectly stiff with respect to the motion considered.

The numerical implementation will play an important role in further details of the model and the integration scheme to be chosen. This goes beyond the scope of this section. It is just to show a simple alternative in the study of fluid dynamics.

6.6. Control of the Voith-Schneider propeller

In the design of continuous components optimization plays an important role. In Section 12.4 shape optimization is discussed as an example of structural optimization. In this section we discuss an example which is closer to control optimization. [Kirk, 2004] In that process, it is useful to characterize different components of a larger system physically. In this case the blade characteristics of a propeller are used to analyze and optimize the control of the Voith-Schneider propeller.

Ever since the mid-nineteenth century it is understood that lift, rather than drag, is the preferred means of ship propulsion. The screw propeller is therefore ubiquitous. However, it is usually designed for a constant linear velocity, with larger, slower propellers, for big ocean ships, and proportionally smaller, faster propellers for more versatile boats. However, the fixed screw propeller is not so useful for boats with variable velocity and trust, and sharp turns, such as ferry boats and tugs. In that case, variable pitch propellers and the Voith-Schneider propeller are often used. The Voith-Schneider propeller consists of a set of rotating
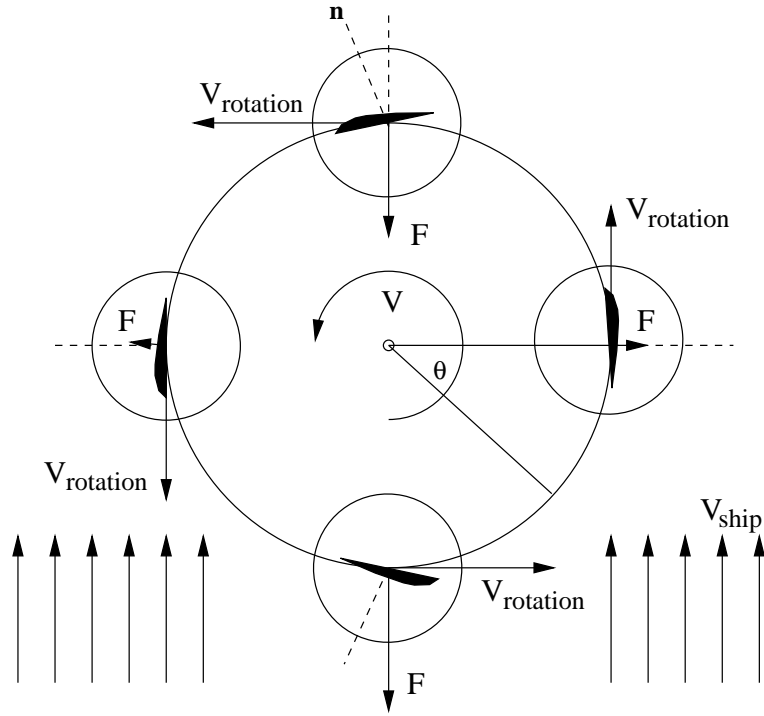
FIGURE 6.2. The Voith-Schneider propeller, viewed from the top.
The blades rotate around in the big circle, while the orientation
$\mathbf{n}(\theta)$ changes along the circle.

vertical blades at the bottom of the hull. [Van Lammeren, 1944] The incident
angles of the blades with respect to the tangential velocity can be altered to create
positive or negative lift, depending on the position of the blade with respect to the
ship's direction. See Figure 6.2.

The lift and drag of a foil, or blade, is a function of the velocity V and the
incidence angle α. The static drag γ acts opposite the direction of the velocity,
while the dynamic drag and the lift are due to the deflection of the momentum of
the flow. A column of water, in the direction of the flow will hit the blade, and
is deflected as light of a mirror, yielding a force. We represent the incidence angle
and the scalar velocity together as a vector velocity $\mathbf{V}$, where the blade direction is
given by the unit normal vector $\mathbf{n}$, perpendicular to the bottom, of which the flow
reflects. The incoming flow has direction $\mathbf{V}$, the outgoing flow is reflected and has
direction $\mathbf{V} - 2\mathbf{n}(\mathbf{V}^T\mathbf{n})$. The difference is the change in momentum and associated
with the force on the blade. The force on the blade is given by:

$$(6.20) \qquad\qquad \mathbf{F} = -\gamma\mathbf{V} - 2A(\mathbf{V}^T\mathbf{n})\mathbf{n} \ \ ,$$

where the static drag γ, the blade normal $\mathbf{n}$, and effective cross section A are the
blade constants for water flow. See Figure 6.3. The flow hits the blade and the
component of the velocity perpendicular to the blade surface is reflected, causing
a change in momentum double to this component. These results are for stationary
laminar flow. [Breslin and Andersen, 1996] In the case of the Voith-Schneider
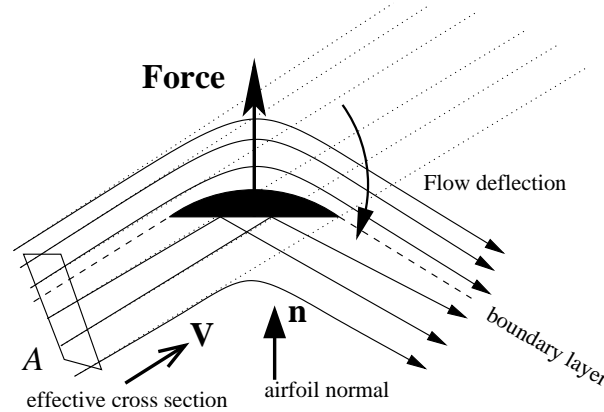
FIGURE 6.3. The typical lift force of a blade in a flow can be seen either in terms of the pressure, or as the momentum change associated with the change of the flow direction. Subtleties at the initial and final edge of the flow separating blade determine the qualities, such as stalling angle and static drag, of the blade shape.

propeller the flow is neither stationary nor laminar, due to the rotation of the multiple blades.

The Voith-Schneider propeller is notoriously difficult to model numerically, due to its extent, and the complex flow around the moving and rotating blades. Instead, we will study the propeller analytically, but in a systematic manner, such that we are convinced that all the effects are named and estimated. At the moment we will not use experimental input to avoid modelling bias, and to test the the strength of the systematic approach, which could lead to new and different insight into old problems.

The Voith-Schneider ship propeller is also a complex control problem. [Jürgens et al., 2006] The control of the orientation of the blades $\mathbf{n}$, the incidence angle, during the revolution in the bigger circle, will increase or decrease the lift. To optimize the thrust, due to the sideway lift, both the incidence angles and the overall rotation velocity ω should be controlled. In many cases, the system is optimized through model identification using experiments. In some cases long and complicated numerical simulations are required to determine the optimal settings for a number of particular cases, ship velocity, rotation velocity, thrust, and so on.

Blades have a static drag, a dynamical drag and a dynamical lift. [Kermode, 1987] The static drag is the minimal drag, γ, for a particular velocity. The dynamical properties are due to the deflection of the water stream, causing both a lift and a drag at the same time. It can be seen as a column of water hitting the blade and being deflected. See Figure 6.3 The momentum rate-of-change of the water equals the force on the blade. The amount of deflected water depends on the effective area, A, of the blade and the angle of incidence. Changing the incidence angle $\mathbf{n}$ during a cycle $\theta \in [0, 2\pi]$ the lift in different direction varies with the angle. The mean lift, or thrust, is the integral over all angles:

$$(6.21) \qquad \mathbf{F}_{\mathrm{mean}} = \frac{1}{2\pi} \int_0^{2\pi} d\theta \mathbf{F}(\theta, \mathbf{V}, \mathbf{n}(\theta)) \ .$$
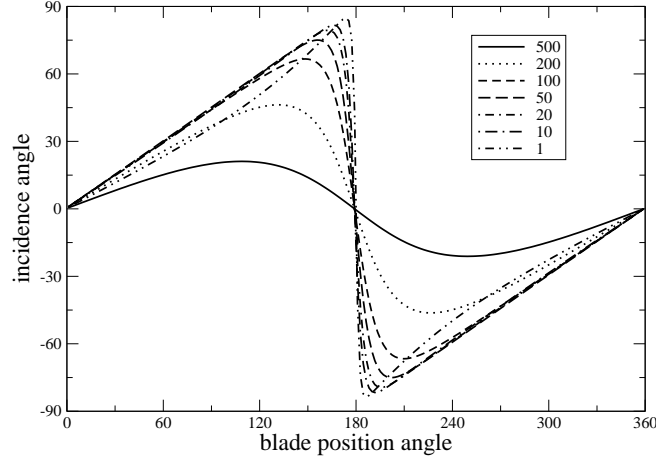
FIGURE 6.4. The optimal incidence angles for the blades are plotted for different values of $\gamma/2A$. The fast change at the angle 180 degrees is associated with the blade moving opposite the thrust direction, where the static drag contributes to the thrust.

The velocity of the water near the blade is the sum of the rotation velocity of the blade $r\omega\mathbf{e}_\theta$ and the ship velocity $\mathbf{V}_{\text{ship}}$. The laminar independent blade approximation of the thrust as function of the rotation velocity ω and the incidence angle function $\mathbf{n}(\theta)$ is:

$$(6.22) \qquad \mathbf{F}_{\text{mean}} = -\frac{1}{2\pi}\int_0^{2\pi} d\theta(\gamma\mathbf{V}_{\text{ship}} + 2A\mathbf{n}(\mathbf{V}_{\text{ship}}^T\mathbf{n} + r\omega\mathbf{e}_\theta^T\mathbf{n})) \ .$$

The static drag averages to zero, but it will contribute to the required torque from the ship motor:

$$(6.23) \qquad T_{\text{axis}} = -\frac{1}{2\pi}\int_0^{2\pi} d\theta(\gamma r^2\omega + 2Ar\mathbf{e}_\theta^T\mathbf{n}(\mathbf{V}_{\text{ship}}^T\mathbf{n} + r\omega\mathbf{e}_\theta^T\mathbf{n})) \ .$$

To first order, the velocity of the ship may be ignored: $\mathbf{V}_{\text{ship}} = 0$. The efficiency can be optimized through the optimization of:

$$(6.24) \ \ \max\frac{r|\mathbf{F}_{\text{mean}}|}{|T_{\text{axis}}|} = \left(\int_0^{2\pi} d\theta 2A(\mathbf{e}_x^T\mathbf{n})(\mathbf{e}_\theta^T\mathbf{n})\right)\left(\int_0^{2\pi} d\theta(\gamma + 2A(\mathbf{e}_\theta^T\mathbf{n})^2)\right)^{-1} \ ,$$

where the force is set in the x-direction $\mathbf{e}_x$. The result is independent of the radius r and the rotation velocity ω. It depends only on the ratio of the static and dynamic drag $\gamma/2A$. The nonlinear optimization problem is easily solved with the methods described in Chapter 10, after a small term $(\partial_\theta\mathbf{n})^2$ is added to remove the solution jumping half a period $\Delta\theta = \pi$. The results of the optimization are shown in Figure 6.4. The optimal blade orientations in a cycle are plotted for different values of $\gamma/2A$.

The Voith-Schneider propeller is a nonideal gyrator. [Karnopp et al., 2000] Given the axis velocity it yields the thrust of the propeller. The drag causes losses in the propeller, independent of any ship velocity generated. Drag can be seen as the internal friction of the propeller. The actual velocity will depend on other factors, such as the ship hull shape and weight.

There are several improvements for this result. First the flow is not really laminar, since the blade changes direction in the flow, which is even enhanced by the changing incidence angle of the blade. Furthermore, the blades disturb the flow, this disturbance increases with the increasing incidence angle. In the worst case the blades cause a vortex which sucks the water up from below and swings it out via the blades. Ventilation systems placed on roofs of buildings sometimes work through this principle. Furthermore, changing the blade angles may cost a considerable amount of force, which should be added to the operating torque T_{axis}.

The disturbances of the flow should be quantified in order to go beyond the independent laminar results. We use the thrust of each blade as the basis of the disturbance it causes. The system itself is a feedback loop.

We investigated only the simplest cases of interference between different aspects of the fluid flow around the blades. However, this approach to modelling the Voith-Schneider propeller allows us to experimentally investigate the actual magnitude of different effects. A consistent model, resulting from experimental input, can be used to investigate and implement more complex incidence angle dependencies.

In the laminar independent blade case we can already see the results of the drag of the asymmetry of the solution. The drag in the direction of the thrust is advantageous, while in the opposite direction it is unwanted. Hence, for the blade angle opposite wanted thrust direction the drag should be minimized, which is at a quarter circle angle of the maximal and minimal lift.

6.7. Elasticity

The standard mechanical engineering formulation of strain and stress starts with the displacement $\mathbf{u} = (u, v, w)$ from which the elongations follow: [Love, 1944, Truesdell and Noll, 1965, Washizu, 1968, Soutas-Little, 1973, Fetter and Walecka, 1980]

$$(6.25) \qquad (\varepsilon_x, \varepsilon_y, \varepsilon_z) = \left(\frac{\partial u}{\partial x}, \frac{\partial v}{\partial y}, \frac{\partial w}{\partial z} \right) \quad .$$

The shear γ is defined as:

$$(6.26) \qquad (\gamma_{xy}, \gamma_{xz}, \gamma_{yz}) = \left(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}, \frac{\partial w}{\partial x} + \frac{\partial u}{\partial z}, \frac{\partial w}{\partial y} + \frac{\partial v}{\partial z} \right) \quad .$$

The strain tensor arises from combining the stretches or elongations ε_i, with the shear $\varepsilon_{ij} = \frac{1}{2}\gamma_{xy}$ such that is yield the symmetric strain tensor:

$$(6.27) \qquad V = \begin{pmatrix} \varepsilon_x & \varepsilon_{xy} & \varepsilon_{xz} \\ \varepsilon_{xy} & \varepsilon_y & \varepsilon_{yz} \\ \varepsilon_{xz} & \varepsilon_{yz} & \varepsilon_z \end{pmatrix} \quad .$$

The odd combinations of the mixed derivatives (i.e., $\partial_x v - \partial_y u$) correspond to rotations, and do not cause any strain.

For the purpose of the tensor, the off-diagonal shear strain ε_{ij} is introduced, with a factor of a half, such that the sum of the squares of the coefficients yield the

linear shear energy. (In some literature, the factor of a half is in the off-diagonal elements of the stress tensor, rather than in the definition of ε_{ij}.)

The change elastic energy is the coefficient-wise, or direct product of the change in strain $d\varepsilon$, with the corresponding force or effort, which is called the stress tensor S:

$$(6.28) \qquad S = \begin{pmatrix} \sigma_x & \tau_{xy} & \tau_{xz} \\ \tau_{xy} & \sigma_y & \tau_{yz} \\ \tau_{xz} & \tau_{yz} & \sigma_z \end{pmatrix} \ ,$$

which is also denoted by σ_{ij}, where $\sigma_{ii} = \sigma_i$ and $\sigma_{ij} = \tau_{ij}$, such that the total elastic energy is given be the integral over the change of strain:

$$(6.29) \qquad W = \int_{\text{Volume}} \left(\sum_i \int^{\varepsilon_i} \sigma_i d\varepsilon_i + \sum_{i \leq j} \int^{\gamma_{ij}} \tau_{ij} d\gamma_{ij} \right) \ .$$

Note that the summation over the shear terms is only over three terms: $(i,j) = \{(1,2),(1,3),(2,3)\}$. In the case of linear elasticity this reduces to:

$$(6.30) \qquad W = \frac{1}{2} \int_{\text{Volume}} \sum_i \sigma_i \varepsilon_i + \sum_{i \leq j} \tau_{ij} \gamma_{ij} \ .$$

The linear relation between stress and strain for isotropic material is:

$$(6.31) \qquad \begin{pmatrix} E\varepsilon_x \\ E\varepsilon_y \\ E\varepsilon_z \\ G\gamma_{xy} \\ G\gamma_{xz} \\ G\gamma_{yz} \end{pmatrix} = \begin{pmatrix} 1 & -\nu & -\nu & & & \\ -\nu & 1 & -\nu & & & \\ -\nu & -\nu & 1 & & & \\ & & & 1 & & \\ & & & & 1 & \\ & & & & & 1 \end{pmatrix} \begin{pmatrix} \sigma_x \\ \sigma_y \\ \sigma_z \\ \tau_{xy} \\ \tau_{xz} \\ \tau_{yz} \end{pmatrix} \ ,$$

where E is the Young modulus, $\nu = 1/m$ is the Poisson ratio, and G the shear modulus. The shear modulus can be expressed in terms of the Young modulus and the Poisson ratio:

$$(6.32) \qquad E = 2G(1 + \nu) \ .$$

The matrix relation for the diagonal parts of the strain and stress tensors is easily inverted:

$$(6.33) \qquad \begin{pmatrix} 1 & -\nu & -\nu \\ -\nu & 1 & -\nu \\ -\nu & -\nu & 1 \end{pmatrix}^{-1} = \frac{1}{(1+\nu)(1-2\nu)} \begin{pmatrix} 1-\nu & \nu & \nu \\ \nu & 1-\nu & \nu \\ \nu & \nu & 1-\nu \end{pmatrix} \ .$$

The inverse is singular for $\nu = \frac{1}{2}$ and $\nu = -1$.

In itself elasticity is a well-defined problem. However, boundary conditions and matching conditions are the complicating factors, since they are not expressed in the stress and strain tensors, but in displacement $\mathbf{u}$ and force $\mathbf{f}$:

$$(6.34) \qquad f_i|_{\text{surface } \mathbf{n}} = \sum_j n_j S_{ji} = \sigma_i + \sum_{j \neq i} \tau_{ij} \ ,$$

which requires one to separate rotations in the displacement $\mathbf{u}$, and combine terms to a force $\mathbf{f}$. Furthermore, it might be possible to construct stresses and strain which do not correspond to conservative force fields and displacements, respectively. Therefore, there are a number of compatibility conditions.

In many cases not the full three-dimensional theory of elasticity is used. For different types of two-dimensional theory, such as plates, and sections of a three-dimensional structure, different specialized results exist. [Soutas-Little, 1973, Love, 1944] The lower-dimensional theory can be derived from the three-dimensional theory by assuming homogeneity, and particular boundary conditions for the dependent directions. The two typical cases are free boundaries, in the case of plates, and fixed boundaries in the case of symmetry. The reduced order models depend only on the remaining reference coordinates (x, y) or only x.

The reduction of the three-dimensional theory to two dimensions can be done in two general manners. First, assuming the third direction to be free. Second, assuming the third direction to be fixed. The first case occurs, for example, if one considers a thin plate where the thickness of the plate can vary freely. The second case occurs for cases with symmetry, the direction in which the system is symmetric can be ignored, but the boundary conditions are fixed. Such conditions occur in very long objects, where only the transverse deformation is considered, like railroad tracks.

The simplest procedure to reduce the dimension is to use the form of the energy relevant for the problem. For the fixed positions in one or two dimensions the energy expressed in terms of the elongations ε_i is the best starting point as $\varepsilon_z = 0$ or $\varepsilon_y = \varepsilon_z = 0$ are simple reductions of the stiffness matrix:

$$(6.35) \qquad W = \frac{1}{2} \begin{pmatrix} \varepsilon_x \\ \varepsilon_y \\ \varepsilon_z \end{pmatrix}^T \begin{pmatrix} E_\| & E_\perp & E_\perp \\ E_\perp & E_\| & E_\perp \\ E_\perp & E_\perp & E_\| \end{pmatrix} \begin{pmatrix} \varepsilon_x \\ \varepsilon_y \\ \varepsilon_z \end{pmatrix} \quad .$$

$$(6.36) \qquad \to W_{\text{2d}} = \frac{1}{2} \begin{pmatrix} \varepsilon_x \\ \varepsilon_y \end{pmatrix}^T \begin{pmatrix} E_\| & E_\perp \\ E_\perp & E_\| \end{pmatrix} \begin{pmatrix} \varepsilon_x \\ \varepsilon_y \end{pmatrix} \to W_{\text{1d}} = \frac{1}{2} \varepsilon_x E_\| \quad , \varepsilon_x$$

where $E_\| = (1 - \nu)E/((1 - 2\nu)(1 + \nu))$, $E_\perp = \nu E/((1 - 2\nu)(1 + \nu))$. Similarly, for free motions, force in the irrelevant direction is zero $\sigma_i = 0$. Hence energy expressed in terms of the stresses σ_x, σ_y, and σ_z is the easy starting point for reduction:

$$(6.37) \qquad W^* = \frac{1}{2E} \begin{pmatrix} \sigma_x \\ \sigma_y \\ \sigma_z \end{pmatrix}^T \begin{pmatrix} 1 & -\nu & -\nu \\ -\nu & 1 & -\nu \\ -\nu & -\nu & 1 \end{pmatrix} \begin{pmatrix} \sigma_x \\ \sigma_y \\ \sigma_z \end{pmatrix} \quad ,$$

yielding the two- and one-dimensional submatrix of the compliance matrix.

Another approach would be to use Guyan reduction. See also section 8.11 on effective Hamiltonians. The boundary condition $\sigma_z = 0$ corresponds to the minimal-energy value for ε_z given ε_x and ε_y:

$$(6.38) \qquad 0 = \frac{\partial W}{\partial \varepsilon_z} = \partial_{\varepsilon_z} \frac{1}{2} \begin{pmatrix} \varepsilon_x \\ \varepsilon_y \\ \varepsilon_z \end{pmatrix}^T \begin{pmatrix} E_\| & E_\perp & E_\perp \\ E_\perp & E_\| & E_\perp \\ E_\perp & E_\perp & E_\| \end{pmatrix} \begin{pmatrix} \varepsilon_x \\ \varepsilon_y \\ \varepsilon_z \end{pmatrix} \quad .$$

This reduces to:

$$(6.39) \qquad \varepsilon_z = -\frac{E_\perp (\varepsilon_x + \varepsilon_y)}{E_\|} \quad ,$$

or as a constraint relation expressing all three elongations in terms of the two independent $\varepsilon_x, \varepsilon_y$:

$$(6.40) \qquad \begin{pmatrix} \varepsilon_x \\ \varepsilon_y \\ \varepsilon_z \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ \frac{-\nu}{1-\nu} & \frac{-\nu}{1-\nu} \end{pmatrix} \begin{pmatrix} \varepsilon_x \\ \varepsilon_y \end{pmatrix} \quad ,$$

with the elastic energy:

$$(6.41) \qquad W_{\text{2d}} = \frac{E}{2} \frac{1}{(1-\nu^2)} \begin{pmatrix} \varepsilon_x \\ \varepsilon_y \end{pmatrix} \begin{pmatrix} 1 & \nu \\ \nu & 1 \end{pmatrix} \begin{pmatrix} \varepsilon_x \\ \varepsilon_y \end{pmatrix} \quad .$$

In the case of two unconstrained motions both ε_y and ε_z are expressed as a function of the independent direction ε_x. The free motion corresponds to the minimal energy for a given ε_x:

$$(6.42) \qquad \frac{\partial W}{\partial \varepsilon_y} = \frac{\partial W}{\partial \varepsilon_z} = 0 \quad ,$$

yielding the two constraints:

$$(6.43) \qquad \begin{pmatrix} \varepsilon_x \\ \varepsilon_y \\ \varepsilon_z \end{pmatrix} = \begin{pmatrix} 1 \\ -\nu \\ -\nu \end{pmatrix} \varepsilon_x \quad ,$$

with the elastic energy:

$$(6.44) \qquad W_{\text{1d}} = \frac{E}{2} \varepsilon_x^2 \quad ,$$

which is Young's result for the elongation of a rod, which radius will decrease appropriately, given by the constraint relation.

In short, inverting the energy function $E = \frac{1}{2} \mathbf{x}^T \mathbf{A} \mathbf{x}$ to a co-energy function $E^* = \mathbf{y}^T \mathbf{A}^{-1} \mathbf{y}$ where $\mathbf{y} = \mathbf{A} \mathbf{x}$ and selecting the appropriate submatrix from the full matrix corresponds to inserting the solution of the variational minimum with respect to the dual variable, since:

$$(6.45) \qquad \mathbf{x}^* = \mathbf{y} = \frac{\partial E}{\partial \mathbf{x}} = 0 \quad ,$$

In some cases it may be convenient to work with strain variables, in other cases with stress variables. In the linear theory there is no fundamental difference, as long as the stiffness is non-singular. The singular stiffness matrix arises in the theory of partial differential equation, for example, if we do not use the stress and the strain but the force and displacement as our variables.

6.8. Viscoelasticity

Viscoelasticity [Christensen, 1971] is the dynamical theory of plasticity. Objects under strain may deform, and part of these deformations can be permanent. Work applied to the system is partly used to change the rest state. Such systems have a hysteresis at best and usually orbits in force-position space which do not close. The result varies with the trajectory in state space taken by the applied force.

Similar to heat transfer in fluids, the theory of viscoelasticity, especially thermo-viscoelasticity, is a phenomenological theory due to the variety of effects. Two simple and well-known characterizations of viscoelastic material are in terms of springs and dampers, either in series in the Maxwell model, or in parallel in the Kelvin, or Voigt, model. See Figure 6.5. The Maxwell model is a viscoelastic
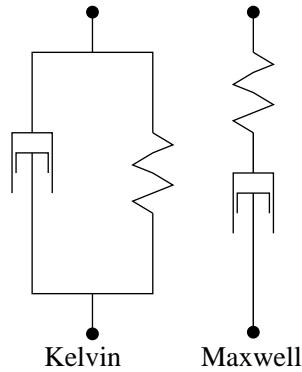
Kelvin   Maxwell

FIGURE 6.5. The Kelvin and the Maxwell models for viscoelastic-
ity. The respective generalized models consists of series and par-
allel combinations of the simple models with a single spring and
damper. The Kelvin, or Voigt, model corresponds to a viscoelastic
solid, while the Maxwell model is a viscoelastic fluid.

fluid, which can not sustain internal shear indefinitely, while the Kelvin model is
a viscoelastic solid, eventually regaining its original shape. In practice, both the
parallel and the series coupling are present, which, however, yield a complicated
network even for homogeneous materials. The generalized Kelvin model consists of
several Maxwell spring-damper combinations in parallel.

More complex models have more parameters to be fitted to experiment. Typ-
ical response results to be fitted to the generalized Kelvin or Maxwell models are
combinations of exponentials related to the damping and the elasticity. Another
typical experiment would be the propagation of waves in viscoelastic material. The
velocity and damping of the waves for different frequencies may be related the
parameters of the model at hand.

Viscoelasticity, especially in the combination with temperature effects, would
benefit from a microscopically based model, similar to the Vanderwaals model for
gas-liquid phases of materials. In applied mathematics the study of polymers and
percolation has led to considerable insight in the phases of polymers, glasses, and
liquid crystals, exhibiting the viscoelastic properties. The bulk properties, however,
have not been of much interest of the modern academic researchers. In the study
of friction dynamical and elastic effects are taken into account in dynamical models
depending on the history of the motion. [Dahl, 1976] Furthermore, the study of
macromolecules, such as DNA, in biochemistry, showed the complex static solu-
tions, corresponding to local minima in energy, slowly deforming through thermal
and mechanical excitations. Biological macromolecules, like enzymes, seem to en-
hance deformation processes through mechanical restrictions, combined with some
electrostatic forces from the polarity of the molecules.

6.9. Electrodynamics

The four Maxwell equations are the start of modern physics as we know it.
[Maxwell, 1891, Jackson, 1975] What were before separate phenomena, such as
magnetism, electricity, currents, and light finally came together in a single model.
Different pieces were already collected by great scientists such as Faraday, Coulomb,

Ampère, and Ohm, but the crown was placed by Maxwell, who added the displacement term to the relation between magnetic flux and current. [Taton, 1965] The Maxwell equations were thereby consistent with the conservation of charge, satisfying the charge continuity relation.

The Maxwell equations themselves yielded a wave equation, which described electromagnetic waves. The presence of the wave solution, and thereby electromagnetic radiation leaving the system, makes the dynamical electromagnetic theory complex, in the sense that it is not always useful to search for stationary or quasistationary solutions to the Maxwell equations.

Despite the seeming simplicity of four Maxwell equations, the construction of an appropriate model for a particular system, with the right independent, or states, and dependent variables, or constitutive or closure relations, and the correct boundary conditions, or input, can be a formidable task.

The four Maxwell equations are respectively, Biot-Savart law with the Maxwell displacement term $\partial_t \mathbf{D}$, Faraday law, Poisson law and the Gauss law:

$$\nabla \times \mathbf{H} = \mathbf{j} + \partial_t \mathbf{D} \quad , \tag{6.46}$$

$$\nabla \times \mathbf{E} = -\partial_t \mathbf{B} \quad , \tag{6.47}$$

$$\nabla \cdot \mathbf{D} = \rho \quad , \tag{6.48}$$

$$\nabla \cdot \mathbf{B} = 0 \quad . \tag{6.49}$$

In many cases, materials have simple direct responses to an external field, such as polarization and magnetization. The microscopic charge densities and eddy currents generated are not considered part of the macroscopic currents $\mathbf{j}$ and charges ρ present in the Maxwell equations. The constants appearing in the Maxwell equations may vary from author to author. There are several units available to define the electromagnetic fields. Here we use SI (MKSA) units, but Gauss and Heaviside units are also very common. [Jackson, 1975]

The magnetic induction $\mathbf{B}$ is related to the magnetic field $\mathbf{H}$ through the induced magnetization $\mathbf{M}$:

$$\mathbf{H} = \frac{1}{\mu_0} \mathbf{B} - \mathbf{M} \quad . \tag{6.50}$$

while the electric displacement $\mathbf{D}$ is related to the the electric field $\mathbf{E}$ through the induced polarization $\mathbf{P}$:

$$\mathbf{D} = \epsilon_0 \mathbf{E} + \mathbf{P} \quad . \tag{6.51}$$

Modelling electrodynamical systems is complex for several reasons. First of all, the Maxwell equations themselves are already a set of complicated equations, mainly in the coupling of the equations, and the constraints $\nabla \cdot \mathbf{D} = \rho$ and $\nabla \cdot \mathbf{B} = 0$, which are difficult to implement on general fields $\mathbf{D}$ and $\mathbf{B}$. Second, the choice of independent and dependent variables, and the associated boundary conditions, is not always clear from the system and its context. Usually, problems are set to be either a static solution of the constraint, like electrostatic and magnetostatic, or a electromagnetic wave trivially satisfying the constraints, like radiation and waveguide problems. The general electromagnetic problem, including light, or radiation, of all frequencies, is often avoided. See also Table 6.7.

name	variable	boundary conditions	sources
conductivity	$\mathbf{j}$	V or $\mathbf{E}$	$\mathbf{E}$
polarization	$\mathbf{D}$	none (local relation)	$\mathbf{E}$
magnetization	$\mathbf{M}$	none (local relation)	$\mathbf{H}$
electrostatics	$\mathbf{E}$ or V	$\mathbf{E}$ or V	ρ
magnetostatics	$\mathbf{B}$	$\mathbf{B}$	$\mathbf{j}$
wave guides	$\mathbf{E} \times \mathbf{B}(\omega)$	$\mathbf{E}$ and $\mathbf{B}$	$\mathbf{E} \times \mathbf{B}$
radiation	$\mathbf{E} \times \mathbf{B}(\omega)$	$\mathbf{E} \times \mathbf{B}$	$\partial_t \mathbf{j}$ or $\ddot{\rho}$
electrodynamics	$\mathbf{E}(t)$ and $\mathbf{B}(t)$	$\mathbf{E}$ and $\mathbf{B}$	$\mathbf{j}(t)$ or $\rho(t)$

TABLE 6.7. Different types of electrodynamical problems

Relativistic invariance of the Maxwell wave equation; it yielded a constant wave velocity in vacuum, independent of the velocity of the observer, led to a juxtaposition in modern physics. For the first time the equations dictated the development rather than the experiments. Special and general relativity followed from the analysis of the consistency of the equations of motion, i.e., the Maxwell equations.

Nowadays, general electrodynamics is often treated in relativistic invariant approaches. This means that every space-time point is considered as separate. As an initial boundary value problem (Section 5.3) such approaches will fail, since in that case the initial surface $t = 0$, and the initial conditions, or kinematical variables, are never defined. The kinematical variables, the electromagnetic fields and charge and current densities, at a given time have to satisfy global conditions. See also Section 11.11.

The energy density E of the electromagnetic field is a simple local relation:

$$(6.52) \qquad E = \frac{1}{2}\left(\mathbf{D}^T \mathbf{E} + \mathbf{B}^T \mathbf{H}\right) \quad .$$

This energy density is deceivingly simple. The states $\mathbf{D}$ and $\mathbf{B}$ satisfy the Gauss laws: $\nabla \cdot \mathbf{D} = \rho$ and $\nabla \cdot \mathbf{B} = 0$, and they can therefore not be chosen freely. Therefore, despite its local appearance, i.e., products of fields at the same point, the energy density has a nonlocal nature. The change of energy reveals this partly. The Maxwell equations determine the change of energy:

$$(6.53) \qquad \dot{E} = -\nabla \cdot (\mathbf{E} \times \mathbf{H}) - \mathbf{j}^T \mathbf{E} \quad ,$$

where the first term on the right-hand side is the divergence of the energy flux, given by the vector of Poynting, $\mathbf{E} \times \mathbf{H}$, and the second-term corresponds to the coupling to mechanical energy due to the motion of charge, or current, $\mathbf{j}$ in an electric field $\mathbf{E}$; the power product of effort and flow. [Paynter, 1961] The direction of energy flux, or radiation, is perpendicular to both the electric and magnetic fields $\mathbf{E}(\mathbf{D})$ and $\mathbf{H}(\mathbf{B})$, which are also perpendicular with respect to each other. The Maxwell equations, together with the energy density, Eq. 6.52, suggest that these are first-order Hamilton equations of motion. This is not the case. Only the radiative field has pairs of canonical variables. The remainder of the degrees of freedom are either constraints, or longitudinal fields given by boundary conditions or by charge and current sources. See also Section 11.11.

In Cartesian coordinates only six anti-symmetric combinations $\partial_i E_j H_k$ of the 27 contribute to the radiation in electrodynamics. Likewise, the material losses

are easily described by constitutive relation expressing a current in terms of the fields: $\mathbf{j}(\mathbf{D}, \mathbf{B})$. On the other hand, the change in electromagnetic energy $\dot{E}$ cannot be expressed in terms of applied currents, i.e., $\mathbf{D}(\mathbf{j})$ and $\mathbf{B}(\mathbf{j})$. Electromagnetic radiation arises, from material sources, only due to the acceleration of char ges: $\partial_t \mathbf{j}$. [Jackson, 1975]

6.10. Heat transfer

The separation of different dynamical regimes for heat transfer is a complicated problem, especially in the case of viscous fluids. The expansion of fluids and gases with an increase in temperature yields an upward motion under gravitational forces. The resulting convection occurs in many different types. Some well-known examples are the bubbles in boiling water and flames of a fire, which spell already the complexity of the problem. In the absence of gravity, there is also no convection, and a candle burns in space slow, blue, and spherical, due to the diffusion driven burning process.

The convection processes depend on the material properties of the fluid such as the thermal expansion of the fluid, the viscosity, heat diffusion, the density, and the heat capacity. The circumstances can also cause variations, such as the column height and width and other geometric parameters, the temperature difference, and the presence of disturbances. [McAdams, 1951, Holman, 1981, Vekstein, 1992, Landau and Lifschitz, 1987b]

Besides convection, there is also heat transfer due to diffusion, or conduction, and due to radiation. Radiation plays an important role in the case of large surfaces and large temperature differences. The Stefan law of heat transfer is:

$$(6.54) \qquad j = \frac{\varepsilon}{2 - \varepsilon} \sigma (T_{\text{object}}^4 - T_{\text{surrounding}}^4) \ ,$$

where ε is the emissivity correction for the deviation of the blackbody theory and $\sigma = 5.6696 \ 10^{-8} \text{Wm}^{-2}\text{K}^{-4}$ the Stefan-Boltzmann constant. For example, in the case of two flat surfaces of each a meter square relatively close together, the heat transfer from the hotter surface at 370 degrees to the colder surface at 290 degrees is 761 Watt, assuming the emissivity constant to be $\varepsilon = 1$. The emissivity constant depends on the material, the wavelength, and the surface coating or roughness. The wavelength in its turn depends on the temperature. Take for example glass. For low temperatures it will radiate, and absorbs, very little heat, only close to melting temperatures a glow appears. However, for most materials a typical emissivity is $\varepsilon \approx 0.6 - 0.95$.

In the case of conversion of sunlight into heat, for the purpose of generating useful energy, by a Carnot cycle, [Callen, 1985] the absorption in the visible range must be optimized, while the emissivity in the range of 300-500 Kelvin, the sub-millimeter range of electromagnetic radiation, typical temperatures of sun-heated objects should be minimized. Special coatings and surface treatments exist to achieve this goal. Interestingly, the reciprocity theorem, stating that light will traverse an optical path in both directions, in and out, with equal likelihood due to the time-reversal invariance of the physical laws, is circumvented in this way.

At room temperature $T = 290$K, the conduction due to radiation for a $\varepsilon = 1$ object is approximately:

$$(6.55) \qquad j = 4\sigma (T_{\text{object}} - T_{\text{room}}) T_{\text{room}}^3 = 5.5 (T_{\text{object}} - T_{\text{room}}) \text{Wm}^{-2}\text{K}^{-1} \ .$$

In many case a value of 1. or 2. $\mathrm{Wm^{-2}K^{-1}}$ is used.

Normal conduction, due to diffusion, in solids and fluids is usually much larger. Conduction is expressed in a heat conduction coefficient $\lambda[\mathrm{Wm^{-1}K^{-1}}]$. The length scale $\mathrm{m^{-1}}$ refers to surface^{-1}length. The typical distance length scale, which would yield the same conduction as radiation transfer at room temperature is ten to thirty meters for metals, and couple of centimeters to a meter for minerals and fluids. Aluminum, copper, gold and silver are typical good heat conductors.

Heat transfer is a typical case for dimensional analysis. [McAdams, 1951] The relevance of different phenomena can be estimated by calculating the relevant numbers. However, in practice, experiments are the best test for such systems. In that case, the dimensional analysis can still be used to interpolate between different cases, with different dimensions or material constants. Although the underlying mathematical theories of bulk processes of heat transfer are well known, detailed surface characteristics, boundary layers, and other microscopic processes will yield considerable variations in results. [Holman, 1981, Welty et al., 1984]

6.11. Multi-domain models

In Section 6.3 on dimensional analysis, the five fundamental dimensions: mass, length, time, current, and temperature, of all physical systems were discussed. See Table 6.4. In many fields of research, such as social sciences and economics, it is often difficult to compare or combine one dynamical model with another. The variables may be the same, but the dynamical principles, such as Hamilton equations of motion, if defined, are not of the same type, i.e., the effect one part has on another is not defined. Within the physical domains it is possible to compare and combine elements stemming from different domains, since the same variables and concepts are used within all these domains. However, not all possible types of interactions are defined yet. In particular the dynamical effects of the thermodynamic domain, i.e., non-equilibrium thermodynamics, is mainly studied in linear approximations. [De Groot and Mazur, 1962] These effects can play important roles at small scales and in vacuum.

In particular, the energy, expressed in the Hamiltonian, can bring all these different physical domains together. Each of the physical domains are independent terms in the Hamiltonian. The multi-domain effects arise from interaction terms in the Hamiltonian:

$$(6.56) \qquad H = H_{\mathrm{mechanical}} + H_{\mathrm{electromagnetic}} + Q_{\mathrm{thermal}} + \cdots + H_{\mathrm{interaction}} \quad .$$

The energy in the thermal domain, the heat Q, plays a slightly different role than in the other domains, in the Hamiltonian language used in this thesis. [Breedveld, 1984] Typical interactions between the thermal and mechanical domains are the thermal expansion, the temperature dependence of the stiffness, and the viscoelasticity. [Kovalenko, 1969] Thermal effects are known for most materials. Thermal expansion, changes in the stiffness, variations in magnetic properties and polarizability are only a few of the cross-over effects between the thermal and other domains. In the modelling of modern high-precision machines these effects can no longer be ignored.

The interaction terms in the Hamiltonian represent a consistent manner to include the cross-overs into other domains. In many cases, especially with the thermal domain, the Hamiltonian is treated with a parametric temperature dependence. In

an open thermal environment, with a surrounding temperature, this might be a good manner to proceed, at the loss of energy conservation. It is similar to the electric resistors in electric networks, where energy leaves the system in the resistor port as heat. However, at small scales and in vacuum, the heat capacitance and heat generation will play a dynamical role, with temperatures throughout the system which are never in balance with the surroundings.

The choice to restrict the attention and analysis to a typical physical subdomain, such as mechanical, electrodynamical, or thermal, is to simplify the model. The other domains may yield only small effects in the problem under investigation. The separation in different physical domains is mainly a conceptual one, to focus the thoughts. However, one need to validate the choice to ignore the other physical domains. The respective Hamiltonians H_{domain} consist of all the types of energy available in that domain. The interaction Hamiltonians form mainly the basis for the ports or interfaces between different domains. Especially, the bilinear form of the Hamiltonian, with a linear dependence on each of the variables from a particular domain, has little energy-storage capabilities.

However, it is not always appropriate to express the interaction in terms of an interaction Hamiltonian. The charged particle in a magnetic field will change direction, without a change in energy. The interaction does not involve the exchange of energy, which is therefore not necessarily appropriate described by a port-Hamiltonian, based on energy flux. However, the leading bilinear interaction Hamiltonian between the electromagnetic and the mechanic domain is the potential energy ϕ of charge density ρ:

$$(6.57) \qquad\qquad H_{\text{em,mech}} = \int dz\, \phi(z)\rho(z) \quad ,$$

where $H_{\text{em,mech}}$ is the energy in the mechanical domain, with the matter-charge density ρ as variable, but a co-energy in the electromagnetic domain, with the potential ϕ as effort variable. Matter has besides charge also mass. The power continuity, and the appropriate port-variables, follows from the time-derivative of the interaction Hamiltonian. The total energy does not change, but for each domain the energy change, and the power flow can be reconstructed. The change of mechanical energy due to the interaction with the electromagnetic field is:

$$(6.58) \qquad\qquad \dot{H}_{\text{mech}} = \int dz\, \phi(z)\dot{\rho}(z) = \int dz\, \mathbf{E}^T \mathbf{j} \quad ,$$

where we used the charge continuity equation. If the current $\mathbf{j} = q\mathbf{v}$ is produced by the velocity $\mathbf{v}$ of a particle charge q, the product $q\mathbf{D}$ can be seen as the force $\mathbf{f}$ on the charged particle. Whether the electric field $\mathbf{E}$ or the electric displacement $\mathbf{D}$ is used depends on whether the mechanical properties of the medium are taken into account or not, i.e., whether the current is microscopic or macroscopic. In Eq. 6.58, with $\mathbf{E}$, the microscopic current is used. [Jackson, 1975, de Groot and Suttrop, 1972]

The Maxwell equations say very little about the forces on a mechanical charge q. Lorentz introduced a fifth equation, nowadays known as the Lorentz force equation of the force $\mathbf{f}$ on a charge q: [Lorentz, 1915, Jackson, 1975]

$$(6.59) \qquad\qquad \mathbf{f} = q\mathbf{E} + \mathbf{j} \times \mathbf{H} \quad .$$

The second term will not yield an energy transfer between the electromagnetic and the mechanical domain, since the force is perpendicular to the motion in the
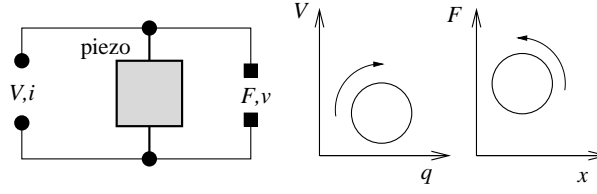
FIGURE 6.6. The main function of a piezo is the couple two physical domains: the electric and the mechanical. It can be used in a Carnot cycle to transform energy from one domain to another.

direction of the current $\mathbf{j}$. However, this second term is the interaction term behind electromotors and dynamos. Actually, in most cases the magnetic field $\mathbf{H}(\mathbf{j})$ is generated by currents in coils itself, and a more appropriate description, without the magnetic field $\mathbf{H}$, is given by the Ampère force law: [Jackson, 1975]

$$(6.60) \qquad \mathbf{f}_{12} = -\frac{\mu_o}{4\pi} \oint \oint \frac{d\mathbf{j}_1^T(\mathbf{x}_1)d\mathbf{j}_2(\mathbf{x}_2)(\mathbf{x}_1 - \mathbf{x}_2)}{|\mathbf{x}_1 - \mathbf{x}_2|^3} \quad .$$

The symmetric geometric form of the line integrals, over the two closed current loops, $\mathbf{j}_1$ and $\mathbf{j}_2$, is often approximated such that the orientation invariance of the interaction is destroyed. [Breedveld, 2001]

The energy extends always across the different domains. Even equivalent descriptions for the same energy may exist in different domains, such as, describing the same electrostatic energy either in terms of the charges of of the fields. The cross-over effects between can however also be small. In other cases there is little point to separate the different domains, as the main function is to interlink the domains. Piezo material is such a case. The electric and mechanical domain are fundamentally coupled in a single energy function. Figure 6.6 shows the lumped representation of a block piezoelectric material. The combination of electric and mechanical capacitance or stiffness, for the same motion, will allow the transfer of energy between the different domains.

It is important to notice that the interaction between domains in continuous systems is often distributed. It is not the change in temperature in time, but the spatial variation in temperature which causes the unwanted deformations in mechanical structures. If the heat is trapped in the object, the temperature spatial variations are smaller, although the temporal variations may be large, the uniform expansion, at typically $\sim 10^{-5}/K$, will not affect the overall shape.

Another peculiar aspect of the physical models is the alternative descriptions of the same phenomena. In the world of applications different phenomena are seen as separated. The underlying microscopic theory is however almost always based on the electric features of the molecules and the atoms of which the material consists. The microscopic peculiarities of these materials will lead to a range of features, from interaction with light, to macroscopic electric properties, such as conductivity, to thermal properties, and elasticity and compressibility. Crystals, or minerals, are, for example generally, hard, brittle and incompressible, non-conducting, and transparent, while metals have another range of features which span the different physical domains. The most interesting aspect of fundamental physics is that these different domains form together a coherent view of matter. The more microscopic the model, the more these features arise from a common source. Unfortunately,

global features arise only with great difficulty from microscopic models, and, as always, a balance between detail and feasibility must be found.

6.12. Some remarks on modelling

In this chapter we studied the models of physical systems within their own right. This is the more traditional form of modelling. The distributed interactions between different physical domains were discussed only in the last section. Our main interest, the modular view of models, their interaction across domain boundaries, is analyzed in the next chapter. The models of continuous systems are generally much more complicated than lumped, low-dimensional models. They are generally very rich in term of the phenomena that can occur. In modelling for engineering purposes, the hope exists that the behavior is limited. Therefore we seek within the full richness of continuous, or infinite-dimensional models, the degrees of freedom relevant for the operations at hand. The typical operation, the typical input and output, determines the relevant function. This is the central question posed and answered in the remainder of the thesis.

CHAPTER 7

Port-based modelling of infinite systems

Global modelling, in terms of lumped parameter systems, can be insufficient to recover all the relevant phenomena. Internal dynamics, such as vibrations, deformations, and delays, may spoil the global model. To improve upon such models detailed local, internal, or distributed theories, of continuous systems, are required. However, many local approaches neglect the global features and interconnections, which have been the main interest.

In this chapter, the global, lumped and local, distributed approaches are linked together, in more than one way. We show the advantages and disadvantages of a collocated approach and the FEM approach, and show that the local theory of the component can be augmented with a local theory of the boundary, in terms of fluxes. These fluxes are the direct implementation of energy, force, or charge balance, and form the natural link between the global, or global, modelling, and the local description.

After a general motivation, the link between the global and local Newton force law is established. Since this is based on general principles, no detailed local theory is required. In the next section the features of local theories which lead to balance laws are investigated, and applied to the inhomogeneous wave equation on a string. In the next section, the interaction with the surrounding through boundary conditions is investigated, first by domain splitting. In the next section by defining appropriate fluxes, the boundary conditions, expressed in the energy flux, yield the core model, which is the local equivalent of the lumped model. The core model may vary with the type of boundary condition and resulting flux. Finally, the leading internal dynamics is discussed. The internal dynamics arises from the coupling with the lumped, or core model, dynamics. We link back to the formal aspects of Hamiltonian dynamics, and we end with some conclusions and an outlook.

7.1. Motivation

Modelling of systems is based on the natural view of the largest independent units of a system. In a human body, for example, we recognize the members separated by joints, although the digits of the hand are much smaller than the upper leg. A medical specialist might, however, see different units, such as muscles, organs, and bones. In principle, units do not need to be spatially distinct. Generally only particular properties, such as electrical, or mechanical, are part of a unit of a model.

Kirchhoff laws and rigid-body dynamics are successful descriptions for lumped parameters of large units. However, for accurate measurement and design of the model units more detailed understanding of the each unit is required. The body, assumed to be rigid, turns out to be elastic, and electrical circuits may have delays, or lags, and mutual interference between components, which are not present in the

dynamics of the Kirchhoff laws. These are both examples of internal dynamics. Instead of a couple of parameters to describe the model, an infinite number, or continuous set, of parameters arise once the unit, or component, no longer acts fully coherently or collectively.

The study and modelling of internal dynamics starts usually with a local theory, expressed in terms of partial differential equations. Successful local theories are consistent with the global, or lumped results. For example, any theory of elasticity should propagate the force from one end of a free object to another, otherwise the micro-balance law, the basis of the force balance, cannot be satisfied. Similarly, local electromagnetic theories, at the basis of local modelling, should, in practice, satisfy local current conservation in order to satisfy Kirchhoff laws at a global scale. If a component is size-extensive, like a homogeneous and isotropic block of material, one might split the block into small pieces, and local blocks will inherit the global laws, yielding a direct link between the integral, of the block as a whole, and the differential, of an infinitesimal part, version of the same conservation law.

Maxwell equations for the local description of electrodynamics were derived to satisfy conservation laws, such as charge conservation. See Section 6.9. Maxwell looked toward the theory of hydrodynamics, with divergence-free vector fields, for inspiration of the final displacement current term in the Biot-Savart law, which guaranteed the global conservation of charge. The local formulation in terms of partial differential equations were only considered complete once the global physical conservation laws were satisfied.

Hence, the local description is connected to the global description via the physical laws, expressed as conserved global quantities, such as charge and momentum, on one hand, and, on the other hand, expressed as properties of the fields, like charge, mass, and current densities. In establishing a close connection between the local theory, expressed in partial differential equations or their FEM approximation, and the global theory, in terms of the lumped parameters, the conservation laws should, and can, play a central role. This is not only for consistency, but also to enable the modeler, or designer, to shift focus smoothly from global system properties to details, which might be critical. Some results, in the more limited context of linear FEM, appeared in [Ligterink et al., 2006].

7.2. Beyond rigid body dynamics

A simple and heuristic example of a bouncing ball should set the mind. In a lumped description of the bounce of a ball, energy is lost at a bounce, to match our observation. However, most balls have a life-span of hundreds of thousands bounces, so it is unrealistic to expect the energy is directly converted through the plasticity of the ball to heat and structure change, like with a lump of clay. In a local description energy is converted from the global, or lumped, motion of the ball, to the vibration of the ball, which is again converted into heat and sound. In principle, the whole conversion process is interesting, however, depending on the zeal of the investigator, he or she might wish to truncate the effort at a certain detail. For example, the amount of conversion from kinetic energy to the dominant elastic mode, and the other elastic energy in higher modes, might already be enough information to determine the lost of global energy within the given accuracy. The dominant elastic mode acts as an ideal spring for the ideal, energy-conserving bounce, while other modes lead to decoherence and the lost of global motion.

Like in the example above, vibrational analysis lies at the core of most efforts to bring to the surface the dominant effects of internal dynamics. This should not be surprising. Vibrational analysis is the physical description of spectral decomposition, which lies at the core of the most powerful mathematical tool for physical systems, namely, functional analysis. However, our mental process might work from the top down: to the global, or lumped, picture we add a couple of degrees of freedom in the form of low-frequency vibrational modes, the mathematical local description starts from the bottom up: from a model with an infinite number of degrees of freedom we should distill a few relevant ones. Spectral decomposition is a hard problem from the bottom up. See Section 5.2. It is only feasible for discretized FEM models, which truncate wave numbers at the inverse mesh spacing. Therefore, they cannot get into the range of vibrational modes, which, for example, lie at the basis of kinetic energy to heat conversion.

In an appropriate and powerful description of the local details of systems the global description should meet the local description at some point. We should ask ourselves the question: what are the local equivalents of the global variables we use? Physical conservation laws should be the guiding principles in such questions. For example, take the motion of an extended object. In the case of a rigid-body description of the extended object the motion of the center-of-mass is a direct consequence of the applied resultant force, also know as Newton force law:

$$(7.1) \qquad \mathbf{F} = M\mathbf{A} \ ,$$

where $\mathbf{A}$ is the acceleration of the center-of-mass and $\mathbf{F}$ the force on the center-of-mass. The capital letters are used for the lumped variables, the small letter for the distributed variables.

The corresponding local description is not immediately evident. Starting with the sum, or lumped, mass M, it could easily be derived from the integration over the mass density ρ:

$$(7.2) \qquad M = \int_{\Omega} \rho(x) dx \ .$$

Furthermore, the acceleration $\mathbf{A}$ can be considered the acceleration of the center-of-mass coordinate $\mathbf{X}$:

$$(7.3) \qquad \mathbf{X} = \frac{1}{M} \int_{\Omega} \mathbf{x}(x) \rho(x) dx \ ,$$

where the positions are normalized with the mass density, which yields:

$$(7.4) \qquad \mathbf{A} = \ddot{\mathbf{X}} \ ,$$

for a constant mass density $\rho(x)$. The position $\mathbf{X}$ is with respect to an origin $\mathbf{x} = 0$, velocity and acceleration are affine and depend only on the orientation of the coordinate system.

The displacement $\mathbf{x}(x)$ is a function of the reference coordinate x, such that the density $\rho(x)$ is invariant. All fields, such as the local forces $\mathbf{f}(x)$, are all functions of the reference coordinates x.

For a rigid body the force law, Eq. 7.1, is the direct sum of all the local forces $\mathbf{f}(x)$:

$$(7.5) \qquad \mathbf{F} = \int_{\Omega} \mathbf{f}(x) dx \ .$$

It is interesting to notice that, unlike the positions $\mathbf{x}$, the forces are not weighted by the mass density. The positions are therefore extensive, while the forces are intensive. In that respect they pair together, which has consequences for the inner-product, as we will see. In the case of a flexible object, internal forces $\mathbf{f}_\perp(x)$, which sum to zero for the rigid-body force, still contributes to the internal dynamics. The meaning of perpendicular is different from what one expect naively. Work and energy are the invariant quantities in mechanics. The invariant inner-product and the geometry should be based on the product of force and displacement, i.e., work. If the force is applied to a particular point of the rigid body, the, as yet unknown, internal force $\mathbf{f}_\perp(x)$ balance will make the object move through the parallel, or sum force, $\mathbf{f}_\parallel$, coherently; as a whole.

In the case of elasticity, the forces should be separated into two parts: the forces that add up to the force causing the global motion, and the remainder. The global forces should not cause any deformation hence they should be coherent with the mass distribution in the system:

$$(7.6) \qquad \mathbf{f}(x) = \mathbf{f}_\parallel(x) + \mathbf{f}_\perp(x) = \mathbf{F}\frac{\rho(x)}{M} + \mathbf{f}_\perp(x) \quad .$$

The global force $\mathbf{F}$ is defined through the local mass and force distributions. The perpendicular part of the distributed force $\mathbf{f}_\perp$ is the remainder of the force, associated with internal force balance.

For the parallel forces, Newton law can be brought inside the integral, to yield an exact local law by definition, for an object without internal stresses. The force yields the corresponding change in momentum, for the collocated inertia $\rho(x)$:

$$(7.7) \qquad \mathbf{f}_\parallel(x) = \rho(x)\mathbf{A} = \dot{\mathbf{P}} \quad ,$$

while the remaining, perpendicular force does not cause any motion of the center-of-mass, but only deformation. This result is most easily understood if the mass integral is considered a mass-weighted inner product:

$$(7.8) \qquad \langle f(x), \rho(x)g(x)\rangle = \int_\Omega f(x)\rho(x)g(x)dx \quad .$$

The center-of-mass $\mathbf{X}$ and the parallel force $\mathbf{F}$ are both projections on the constant and orthogonal unit vector fields $\mathbf{e}(x) = (e_x, e_y, e_z)$:

$$(7.9) \qquad \mathbf{X} = \langle \mathbf{e}, \rho\mathbf{x}\rangle \quad ,$$

and:

$$(7.10) \qquad \mathbf{F} = \langle \mathbf{f}, \mathbf{e}\rangle \quad .$$

Hence the perpendicular force is:

$$(7.11) \qquad \mathbf{f}_\perp(x) = \mathbf{f}(x) - \frac{\rho(x)}{M}\langle \mathbf{f}, \mathbf{e}\rangle \quad ,$$

which inner-product $\langle \mathbf{f}, \mathbf{e}\rangle$ is not the standard self-dual inner-product; it represents global work or energy and therefore is a product of force and displacement.

The associated decomposition of the displacement $\mathbf{x}(x)$ is

$$(7.12) \qquad \mathbf{x}_\perp(x) = \mathbf{x}(x) - \mathbf{X}\mathbf{e}(x) = \mathbf{x}(x) - \frac{1}{M}\langle \mathbf{e}, \rho\mathbf{x}\rangle\mathbf{e}(x) \quad ,$$

such that:

$$(7.13) \qquad \langle \mathbf{f}_\parallel, \mathbf{x}_\perp\rangle = \langle \mathbf{f}_\perp, \mathbf{x}_\parallel\rangle = 0 \quad ,$$

where we used that $\langle e_i, \rho e_j \rangle = M \delta_{ij}$.

Hence it is possible to make a separation between internal dynamics and the rigid dynamics for a local model. The bi-orthogonality is based on the work and energy relations, since, $dE = \mathbf{f} d\mathbf{x}$ is the change in energy density due to the applied force. This work separates into two parts: the rigid-body energy and the internal energy:

$$(7.14) \qquad \int_\Omega dE(x) dx = \langle \mathbf{f}, d\mathbf{x} \rangle = \langle \mathbf{f}_\parallel, d\mathbf{x}_\parallel \rangle + \langle \mathbf{f}_\perp, d\mathbf{x}_\perp \rangle \ ,$$

since the cross terms vanish. Hence, the internal force $\mathbf{f}_\perp$ is perpendicular to the global translations $\mathbf{e}$ associated with $\mathbf{x}_\parallel$.

7.3. Local theories

The local, or distributed, theories in terms of partial differential equations, which would for example yield the specific consequences of elasticity in the example above, are more general than only the separation of rigid-body motion and internal deformation. However, the partial differential equations have one particular property, which led to the recognition of rigid-body motion as a specific form of global dynamics. The rigid-body motion has no associated elastic energy, whatever model one chooses for the elastic, or internal dynamics. The fact that any isolated system has rigid-body modes is the consequence of the translation invariance of physical laws.

If we are guided by the local definition of the energy, as defined by the Hamiltonian density H, we find that the elastic energy vanished for any translation combination $T = (t_x, t_y, t_z)$ of the constant displacements $e_i(x)$:

$$(7.15) \qquad V(\mathbf{x}) - V(\mathbf{x} + T\mathbf{e}) = 0 \ ,$$

where V is the potential part of the Hamiltonian. From the equation of motion it follows that, without applied forces or position constraints the translations yield conserved momenta ($\mathbf{P} = \text{constant}$):

$$(7.16) \qquad \dot{\mathbf{P}} = -\frac{\partial H}{\partial T\mathbf{e}} = 0 \ .$$

In many cases, such as Dirichlet boundary conditions, the boundary conditions fix the values of $\mathbf{P} = 0$. In other cases, such as Neumann boundary conditions, the values of $\mathbf{P} = \text{constant}$ are only partly fixed. Even more general boundary conditions exist, expressed in an input function u, such that $\mathbf{P}(u)$. The link between boundary conditions and constants of motion hints at the possibility of separating the direct consequences of the boundary conditions from the internal dynamics.

The Hamiltonian is positive, so is the elastic energy. By deforming an object energy is added. Some restrictions apply to the form of the elastic energy. Typically one can assume it is a function of a quadratic form, and similarly the differential operators must appear to an even power for positivity. It can be written generically as a weighted form:

$$(7.17) \qquad V((D\mathbf{x})^2, \mathbf{x}) \ ,$$

where D is the differential, or moment operator. See, for example, Sections 2.8 and 5.3. In the case of elasticity the differential operator D could be the strain tensor:

$$(7.18) \qquad \epsilon_{ij} = D\mathbf{x} = \frac{1}{2}(\partial_i x_j(x) + \partial_j x_i(x)) \ .$$

In this case the solutions for which the elastic energy vanishes are besides the translations $\mathbf{e}$, also the rotations around an arbitrary axis $\mathbf{n}$:

$$(7.19) \qquad\qquad \partial_i(\mathbf{n} \times \mathbf{x})_j + \partial_j(\mathbf{n} \times \mathbf{x})_i = 0 \quad .$$

Hence not just the translations but also the rotations are undetermined in the isolated system. The kernel of the differential operator D plays the central role. By introducing the moment variable:

$$(7.20) \qquad\qquad q' = D\mathbf{x} \quad ,$$

the potential energy of the system has a unique minimum $q' = 0$. The indeterminate part, which is either fixed by the boundary conditions, or lead to a conserved quantity, is now explicit in the equations of motion, and does no longer depend on the precise details of the potential, now a function of q'.

In the theory of hyperbolic partial differential equations the existence and uniqueness of a solution for given initial and boundary conditions is sought [Sakamoto, 1982] The proper initial conditions are determined by the Cauchy problem, on an infinite domain or by local analysis. The boundary conditions might be inconsistent with the initial condition. One way to avoid this situation is to choose the stationary, or pseudo-stationary solution for the boundary conditions as initial conditions. A pseudo-stationary solution would, for example, be an accelerating solution for a constant force. Pseudo-stationary solutions only have polynomial time-dependence for the variables of the state for constant boundary values.

The stationary problem for the boundary conditions is called the associated problem. The boundary conditions of the stationary problem are such that the state is stationary. The solutions of the associated problem serve as source to the internal dynamics. Since the solution to the associated problem already satisfies the boundary conditions, the internal dynamics consists of modes with vanishing boundary values. The associated problem, however, is not unique, but may vary with the type of boundary conditions, but as we have seen for the rigid body, a canonical pair of boundary conditions follows from one associated problem, namely, the deformation due to the acceleration of the mass distribution. These pairs are called the port variables.

For example, consider the Hamiltonian of an anisotropic wave equation in one dimension, with a space dependent mass density ρ and stiffness κ:

$$(7.21) \qquad\qquad H' = \frac{p^2}{2\rho(x)} + \frac{\kappa(x)q'^2}{2} \quad ,$$

where $p = \rho(x)\dot{q}(x)$ is the canonical momentum, from Eq. 1.2, and $q' = \partial_x q(x)$ is the moment variable. If a constant force is applied, the object will accelerate. The internal deformation state, however, should be constant in time. The input force is used to generate stationary equations. In these equations the momentum grows linearly, i.e., the solution is polynomial in time. In order to arrive at an oscillatory state, typical for the Hamilton equations of motion, the deformation itself must to give rise to a change of momentum. This would be internal dynamics. States in which internal dynamics is absent are called pseudo-stationary states. They follow from the choice of input and the equations of motion.

The differential operator is $D = -D^* = \partial_x$. Only constant functions vanish for this differential operator. Hence the stationary state, corresponding to constant

input, solution is:

$$(7.22) \qquad v = \begin{pmatrix} F \\ u \end{pmatrix} \quad ,$$

which correspond to a constant moving and compressed, or stressed, state with a velocity u and an internal stress F. These stationary states are, almost trivial, solutions to the algebraic equations:

$$(7.23) \qquad \begin{pmatrix} q'(x) \\ p(x) \end{pmatrix} = \begin{pmatrix} \kappa(x)^{-1}F \\ \rho(x)u \end{pmatrix} \quad .$$

The definition of the moment $q' = Dy$ yields the displacement y:

$$(7.24) \qquad y(x) = y(0) + F \int_0^x dx' \frac{1}{\kappa(x')} \quad ,$$

where the integration constant $y(0) = y_0 + ut$, since no restriction is placed on the time-dependence or initial velocity. Hence, a stationary state can be defined through the force F, the velocity u, and the initial displacement y_0 at a point, e.g., $x = 0$ and $t = 0$.

The solution of constant acceleration, discussed above, would be a pseudo-stationary solution:

$$(7.25) \qquad \begin{pmatrix} 0 \\ p_t \end{pmatrix} = \begin{pmatrix} 0 & \partial_x \\ \partial_x & 0 \end{pmatrix} v = \begin{pmatrix} 0 & \partial_x \\ \partial_x & 0 \end{pmatrix} \begin{pmatrix} F(x) \\ u \end{pmatrix} \quad .$$

$$= \begin{pmatrix} 0 & \partial_x \\ \partial_x & 0 \end{pmatrix} \begin{pmatrix} \kappa(x)q'(x) \\ \rho(x)^{-1}(p(x) + p_t(x)t) \end{pmatrix} \quad .$$

The force $F(x)$ on top row on the right-hand side is the input, therefore $p_t \neq 0$. The second equation must therefore have a zero left-hand side, in order not to generate internal dynamics. The corresponding velocity u is the second port variable; the output. Hence, from the top row it follows $\rho(x)^{-1}p_t(x) = a$ and $\rho(x)^{-1}p(x) = u$ must both be constant, and therefore:

$$(7.26) \qquad F(x) = F(0) + a \int_0^x \rho(x')dx' \quad ,$$

and the displacement $y(x)$ is given by the same integral, where $F(x)$ is no longer constant:

$$(7.27) \qquad y(x,t) = y_0 + ut + \frac{1}{2}at^2 + \int_0^x dx' \frac{F(x')}{\kappa(x')} \quad .$$

A pseudo-stationary solution has a polynomial time-dependence of the states and, in the linear case, also of the energy.

In the general case the equations of motion for the moment variables are the equations of motion of the port-Hamiltonian H': [Van der Schaft and Maschke, 2002]

$$(7.28) \qquad \begin{pmatrix} \dot{q}' \\ \dot{p} \end{pmatrix} = \begin{pmatrix} 0 & D \\ -D^* & 0 \end{pmatrix} \begin{pmatrix} \delta_q H'(q',p) \\ \delta_p H'(q',p) \end{pmatrix} \quad ,$$

where D^* is the formal adjoint of the differential operator D, and p is the canonical momentum for the displacement $q(x)$. The difference between the port-Hamiltonian and an ordinary Hamiltonian is the use of the moment variable q' instead of the position q as variable. The moment variables describe only the internal dynamics. For example, the translation $q = $ constant will yield $\partial_z q = q' = 0$. In the absence

of kinetic energy the translation is irrelevant, however, this is seldom the case. To describe the stationary states, the configuration variables therefore may arise again.

The associated stationary problem would be the solution v_s to:

$$(7.29) \qquad \begin{pmatrix} 0 & D \\ -D^* & 0 \end{pmatrix} v_s = 0 \quad .$$

The pseudo-stationary problems v_p would be of the type:

$$(7.30) \qquad \begin{pmatrix} q_t \\ p_t \end{pmatrix} = \begin{pmatrix} 0 & D \\ -D^* & 0 \end{pmatrix} v_p \quad ,$$

where polynomial solutions $q_t t$ and $p_t t$ are included, to allow for interlinked equations, where not the top and the bottom equation are satisfied independently, but in combination, similar to the Jordan form in Section 3.1. The corresponding components q', p, q_t, p_t, which are constant in time, yield at most linear time-dependent states, arise from solving the algebraic relations:

$$(7.31) \qquad v_p = \begin{pmatrix} \delta_{q'} H'(q', p, q_t t, p_t t) \\ \delta_p H'(q', p, q_t t, p_t t) \end{pmatrix} \quad ,$$

where q_t and p_t are zero, for the stationary equation. The so-called zero modes q_t and p_t play a special role in the Hamiltonian. The Hamiltonian depends parametric on the zero modes. They are the direct consequence of boundary conditions, or constants of motion in the absence of boundary conditions.

The displacement $q_t t$ may not contribute to the elastic energy, but it will contribute to the kinetic energy; it is the constant kinetic energy associated with the overall displacement velocity. Likewise, $p_t t$ will change the kinetic energy. The Hamiltonian is time dependent due to the boundary conditions. If an object accelerates due to the applied force, the kinetic energy changes with the square of the time.

7.4. Boundary conditions

The stationary and pseudo-stationary solutions are not fixed by the equations of motion. They give rise to unknown parameters, such as the force, velocity, and acceleration, which can be seen as the equivalents of constants of motion, in the port-Hamiltonian system. These solutions yield values at the boundary such that the parameters may be fixed by boundary conditions. The boundary-value problem reduces to an algebraic problem.

Decomposition could be an approach to the boundary problems. The integrals of the object domain Ω can be separated in parts $\Omega_1 \cup \Omega_2$:

$$(7.32) \qquad \int_\Omega \rightarrow \int_{\Omega_1} + \int_{\Omega_2} \quad .$$

The definitions of the force $\mathbf{F}$ and the center-of-mass $\mathbf{X}$ hold just as well for each part separately, as for the whole. The linearity of the integral yield the linearity of the global, or global variables. This should not surprise us, as Newton has sought extensive properties of mechanical systems. The work, or power, integral would yield cross-terms between the different domains, related to the power transfer between the domains:

$$(7.33) \qquad \int_\Omega dE dx = \langle \mathbf{f}_1, d\mathbf{x}_1 \rangle_{\Omega_1} + \langle \mathbf{f}_2, d\mathbf{x}_2 \rangle_{\Omega_2} + (\mathbf{f}_1, d\mathbf{x}_2)_{\Omega_1 \cap \Omega_2} + (\mathbf{f}_2, d\mathbf{x}_1)_{\Omega_2 \cap \Omega_1} \quad . \quad ,$$
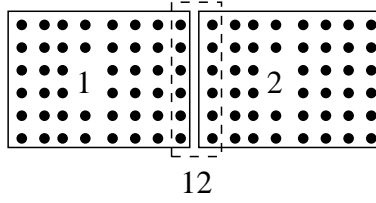
FIGURE 7.1. A larger linear problem, split in two, consists of two internal parts with matrix representations $A^1_{\alpha\beta}$ and $A^2_{\alpha\beta}$ for the energy consisting only of variables of one domain, and an overlap term $A^{12}_{\alpha\beta}$, with variables from both domains.

where the inner-product $(\cdot,\cdot)$ is defined on the common boundary of Ω_1 and Ω_2. The precise definition of the boundary, or trace, integrals $(\cdot,\cdot)$ depends on the local theory. In the absence of spatial derivatives, the boundary terms vanish. They can be seen as the end terms of partial integration, or the displacement flow along force lines. However, without a local theory we can attach little meaning to these boundary terms. Physical theories, however, must satisfy energy conservation and force balance, which translates in this domain decomposition into a constraint on the separation of rigid-body and internal dynamics:

$$(7.34) \qquad (\mathbf{f}_i, d\mathbf{x}_j) = (\mathbf{f}_{\perp i}, d\mathbf{x}_j) \quad,$$

as the result of the locality of the mass density. Discretization will generally destroy this result, as we will see from the example of a string. If a homogeneous string of length L is discretized in n segments with linear displacement, described by $L+1$ end positions we find a stiffness matrix:

$$(7.35) \qquad \int (\partial_x \cdot)^2 \to K = \frac{n}{L} \begin{pmatrix} -1 & 1 & 0 & \cdots & \\ 1 & -2 & 1 & 0 & \\ 0 & 1 & -2 & 1 & 0 \\ \vdots & & \ddots & \ddots & \ddots \end{pmatrix} \quad.$$

The mass matrix is the integral over the density of the linear displaced string:

$$(7.36) \qquad \int \rho(\cdot)^2 \to M = \frac{n\rho}{6L} \begin{pmatrix} 2 & 1 & 0 & \cdots & \\ 1 & 4 & 1 & 0 & \\ 0 & 1 & 4 & 1 & 0 \\ \vdots & & \ddots & \ddots & \ddots \end{pmatrix} \quad.$$

If the mass matrix would have been diagonal, the principles of the continuum mechanics, for the separation of the rigid-body motion and the internal motion and the corresponding forces, would have carried over to the discrete system. However, for the tridiagonal mass matrix this is no longer the case. Moreover, for the time simulation the mass matrix must be inverted, which yields a completely filled matrix M^{-1}. As a consequence, the system can no longer be separated while yielding only additional local boundary terms. Each node of the system is connected to each other node, due to the nonlocal mass matrix inverse M^{-1}.

However, if we consider local, or band-diagonal matrices only, it is possible to split the system in different manners. The larger matrix is split in two similar forms, denoted by $\alpha\beta$. The remainder, consisting of the off-diagonal parts and

diagonal end terms are combined into a 2×2 overlap matrix. See Figure 7.1 If these tridiagonal matrices are split into two equivalent matrices, the remainder is part of the diagonal and the complete off-diagonal part across the split:

$$
(7.37) \qquad A_{\alpha\beta} = \left(\begin{array}{ccc|ccc}
\alpha & \beta & 0 & & & \\
\beta & \alpha & \beta & 0 & & \\
0 & \beta & \alpha & \beta & 0 & \\ \hline
& & 0 & \alpha & \beta & 0 \\
& & & 0 & \beta & \alpha & \beta \\
& & & & 0 & \beta & \alpha
\end{array} \right) \quad ,
$$

which we denote by the Cartesian product in the two subspaces and the boundary matrix:

$$
(7.38) \qquad A_{\alpha\beta} = A_{\alpha\beta}^1 \otimes A_{\alpha\beta}^2 + \left(\begin{array}{c|c} \alpha' & \beta \\ \hline \beta & \alpha' \end{array} \right)_{12} \quad ,
$$

where typically the end point yields $\alpha' = \frac{1}{2}\alpha$. This is a so-called force split, as the two boundary points are not collocated, but adjacent, and free to move independently. Their motion is limited by the kinetic and potential energy given by the boundary matrices K_{12} and M_{12}. A film-layer model to connect two systems would be an example of such a force split.

The other method of splitting systems to study boundary dynamics would be the configuration split, which is more common. A configuration split would correspond to a collocation of the boundary points. Rather than a dynamical boundary connection, it is a constraint relation, with partly overlapping mass and stiffness matrices. The boundary matrix A_{12} should be added one of the subspace matrices $A_{\alpha\beta}^1$ or $A_{\alpha\beta}^2$.

In the case of higher-order differential equations, such as the Euler-Bernoulli beam, or higher dimensions, the splitting of systems, and the variety of boundary terms and effects is even greater. In the continuum theory of elasticity it has led to a number of variational principles, such as the three-field method and the four-field method [Washizu, 1968] to retain the freedom to match boundaries in different ways. Furthermore, the implementation of such splittings will only be feasible for linear systems, where constant matrices are to be decomposed. The general theory of decomposition of such systems, to solve iteratively, dates back to the work of H.A. Schwarz.[Schwarz, 1870] Nowadays domain decomposition is a viable method to generate static solutions of large FEM problems. [Toselli and Widlund, 2004]

7.5. Flux

Clearly, decomposing the linearized and discretized system is hardly unique, or ideal. On the other hand, the rigid-body formulation, with the collocation at the same x of density and displacement is limited. Instead, the boundary integral can be coupled to the volume integrals, through the divergence theorem:

$$
(7.39) \qquad \int_\Omega \nabla \cdot \mathbf{J} dx = \int_{\partial\Omega} \mathbf{J} \cdot \mathbf{n} da \quad .
$$

Therefore, an appropriate formulation should be expressed in terms of fluxes $\mathbf{J}$. The key flux is the energy flux $\mathbf{S}$, since in the case of zero internal energy and zero energy flux the solution, or internal state, is the unique static solution. In general, energy is thus used to determine existence and uniqueness of a solution to a mixed

boundary value problem of a hyperbolic equation. The energy flux is derived from
the canonical Hamiltonian $H(q, p)$ of the system, through the differential operator D
encountered before in the definition of the elastic energy. However, the Hamiltonian
no longer needs to be derived from configuration space. The only link with an
underlying continuous reference space, with boundaries, is through the differential
operator D.

The energy of the system is given by the Hamiltonian density function: $H(q, p)$,
which depends on the canonical position and momentum fields $q(z)$ and $p(z)$. The
energy is conserved, hence the Hamiltonian density satisfies the continuity equation,
Eq. 5.39:

$$(7.40) \qquad\qquad \dot{H} + \nabla \cdot \mathbf{S} = 0 \quad,$$

where $\mathbf{S}$ is the energy flux. [Fetter and Walecka, 1980] The energy flux is the
key guiding principle in the analysis of the interaction between components. The
energy function is a positive operator of the state space, or phase space, and zero
energy corresponds to the lowest, rest, or ground state. Hence stability analysis
and dissipation uses energy and its in- and out-flow.

The energy flux is not uniquely defined. It depends on what quantities are
allowed to flow across boundaries. For example, an ideal balloon might change is
shape but not its content, while a porous paper bag can both change shape and
content. Constraints on the fields might complicate this matter further. The port-
Hamiltonian H' defines the energy-flux. The port-Hamiltonian [Van der Schaft and
Maschke, 2002] arises from the Hamiltonian through a variable substitution:

$$(7.41) \qquad\qquad H(q, p) = H'(Dq, p) \equiv H'(q', p) \quad,$$

where D is an appropriate differential operator. The Hamilton equations, the equa-
tions of motion, are

$$(7.42) \qquad\qquad \begin{pmatrix} \dot{q}' \\ \dot{p} \end{pmatrix} = \begin{pmatrix} 0 & D \\ -D^* & 0 \end{pmatrix} \begin{pmatrix} \delta_{q'} H' \\ \delta_p H' \end{pmatrix} \quad,$$

where D^* is the formal adjoint of the differential operator D.

The energy flux follows from inserting the equations of motion into the time
derivative of the Hamiltonian density:

$$(7.43) \qquad\qquad \dot{H}' = \delta_{q'} H' D \delta_p H' - \delta_p H' D^* \delta_{q'} H' \equiv -\nabla \cdot \mathbf{S} \quad.$$

The expression on the right-hand side is the divergence of the energy flux, which in
many cases can be written as the bilinear product using the boundary operator θ:

$$(7.44) \qquad\qquad \delta_{q'} H' \theta(\mathbf{n}) \delta_p H' = \mathbf{S} \quad.$$

Clearly, appropriate boundary conditions are expressed in terms of $\delta_p H'$ and $\delta_{q'} H'$
rather than the normal state variables q and p.

For the linear wave equation the Hamiltonian density is:

$$(7.45) \qquad\qquad H = \frac{p(z)^2}{2\rho(z)} + \frac{1}{2}\kappa(z)(\nabla q(z))^2 \quad.$$

The associated differential operator is the gradient $D = \nabla$, the adjoint is the
divergence $D^* = -\nabla\cdot$. The directional, or moment, variable is $q' = \nabla q$. Therefore
the boundary operator θ is simply the surface normal:

$$(7.46) \qquad\qquad \theta = \mathbf{n} \quad,$$

and the energy flux follows from the Green identity:

$$(7.47) \qquad \mathbf{S} = \frac{p}{\rho}\mathbf{n} \cdot \kappa \nabla q \ .$$

More general differential operators appear if there are more that one type of energy flux. For example, take the Euler-Bernoulli beam. The elastic energy is given by the second derivative of the deflection $y(z)$:

$$(7.48) \qquad H_{\text{elastic}} = \frac{1}{2}EI(\partial_z^2 y(z))^2 \ ,$$

which may seem like a single bending energy, however, it consists of two parts in terms of boundary conditions. The boundary operator indicates, however, it consists of two parts. The associated differential operator would be $D = \partial_z^2$ such that $q' = \partial_z^2 y$, however, the boundary operator θ is:

$$(7.49) \qquad f\theta g = g\partial_z f - f\partial_z g \ .$$

These two terms correspond to the two boundary conditions, which need to be set for an Euler-Bernoulli beam, at each end, which lead to four possible solutions for $\mathbf{S} = 0$, three of those are familiar boundary conditions: free $(\delta_{q'} H = \partial_z \delta_{q'} H' = 0)$, clamped $(\delta_p H' = \partial_z \delta_p H' = 0)$, and supported $(\delta_p H' = \delta_{q'} H' = 0)$, boundary conditions, where in all cases the initial condition $q = 0$, at the boundaries, is used. The fourth boundary condition corresponds to a fixed orientation, but a free displacement $(\partial_z \delta_p H' = \partial_z \delta_{q'} H' = 0)$, which is difficult to implement physically. One could think of a clamped beam end moving freely along a rail.

The example of the Euler-Bernoulli beam showed that the physical boundary conditions $\mathbf{S} = 0$ are rather unlike the typical mathematical boundary conditions in terms of the state variables q' and p at the boundary. They are, of course, related through the constitutive relations, expressed in term of variational derivatives of the port-Hamiltonian. In the case of isolated problems, associated with $\mathbf{S} = 0$, such distinction is of little relevance, especially in the linear case. The nonlinear, non-isolated case $\mathbf{S} \neq 0$ the distinction is very important. Not all boundary conditions can be implemented straightforwardly.

The pseudo-stationary states, corresponding to the force $\mathbf{F}$ and the velocity $\mathbf{u}$, are, at one hand, related to physical boundary conditions, and, at the other hand, to states q' and p. They form a bridge between the global forces and velocities and the internal states. The model with only the degrees of freedom, such determined uniquely by the boundary conditions, we call the core model.

7.6. Core model

The stationary states are not necessarily static. For example, a system moving frictionless with constant velocity $\mathbf{u}$ is in a stationary state. No energy is transferred in or out of the system, but can be transferred through the system as we will see in the example below. The pseudo-stationary states have a linear time dependence, which may correspond to a complicated energy dependence, if the Hamiltonian is nonlinear. The equations for the quasi-stationary states do not depend on time. The left-hand side, the time derivatives $\dot{q}$ and $\dot{p}$ are replaced by constant states q_t and p_t of linear time-dependent solutions: $\dot{q}_t(z)t = q_t(z)$ and $\dot{p}_t(z)t = p_t(z)$. The energy transfer, or power, can be calculated from the equations of motion:

$$(7.50) \qquad \begin{pmatrix} q_t \\ p_t \end{pmatrix} = \begin{pmatrix} 0 & D \\ -D^* & 0 \end{pmatrix} \begin{pmatrix} \mathbf{F} \\ \mathbf{u} \end{pmatrix} \ ,$$

which yields a much simpler flux and boundary operator than in terms of the states:

$$(7.51) \qquad\qquad \nabla \cdot \mathbf{S} = \mathbf{u} D \mathbf{F} - \mathbf{F} D^* \mathbf{u} \quad .$$

In the case of the differential operator D being the gradient $D = \nabla$, or the divergence $D = \nabla \cdot$, the boundary operators are:

$$(7.52) \qquad\qquad \theta = \left\{ \begin{array}{ll} (\mathbf{n} \cdot \mathbf{u})\mathbf{F}, & D = \nabla \cdot \\ (\mathbf{n} \cdot \mathbf{F})\mathbf{u}, & D = \nabla \end{array} \right. \quad .$$

In the port-Hamiltonian the canonical variables q' and p are geometrically distinct. As a consequence, depending on the choice of differential operator D, the one of two port variables: $\mathbf{u}$ or $\mathbf{F}$, is the boundary-extensive variable; the flux, while the other variable is the boundary-intensive variable; the potential.

In many cases, the port flux is indeed a conserved flux, besides the energy flux. For example, matter flow and charge flow are conserved fluxes, and so is momentum flow in its peculiar way. It might therefore be important to formulate the local theory such that the port flux corresponds to such conserved flux. In many cases the procedure of recovering a Hamiltonian is therefore reversed. The continuity equation is the first equation of motion, the second equation, the closure relation is of less importance, and contains the material properties. Afterwards, the closure relation is integrated to recover the Hamiltonian, or any other total differential. For example, diffusion equations in physical chemistry arise from such approach. Mass is almost mysteriously conserved for any constitutive relation, relating density to pressure, due to the adjoint pair of the gradient operator and the divergence operator.

7.7. Internal dynamics

The core model, rather than the boundary conditions, are the basis of internal dynamics. Since the boundary conditions are satisfied by the states of the core model, the internal dynamics has vanishing boundary conditions. Therefore, the internal dynamics has gained some of independence of the core model state. In some cases, therefore, the internal dynamics may be linearized, while the core model may not.

Inserting the full model back into the equations of motion, taking into account the time-dependence of the core states $q_0 = (q' + q_t t), p_0 = (p + p_t t)$ which depend of the port-variables $\mathbf{F}(t)$ and $\mathbf{u}(t)$, the internal dynamics appears on the right-hand side of the equation:

$$(7.53) \qquad \left(\begin{array}{c} \partial_{\mathbf{F}} q_0 \dot{\mathbf{F}} + \partial_{\mathbf{u}} q_0 \dot{\mathbf{u}} \\ \partial_{\mathbf{F}} p_0 \dot{\mathbf{F}} + \partial_{\mathbf{u}} p_0 \dot{\mathbf{u}} \end{array} \right) = \left(\begin{array}{cc} 0 & D \\ -D^* & 0 \end{array} \right) \left(\begin{array}{c} \delta_{q'} H' - \mathbf{F} \\ \delta_p H - \mathbf{u} \end{array} \right) \quad .$$

Hence, if the input is expanded in terms of polynomials in time, like in a spline approximation, the internal dynamics yield an analogous expansion.

For example, a homogeneous truss, or beam segment, of length L, with elasticity κ and mass density ρ, with the velocity u given as input, and the relative distance determined by the external force F, yields a stationary state with a homogeneous compression $q' = F/\kappa$, and a homogeneous momentum $p = \rho u$. Since the stationary state is determined by two quantities F and u, the forces at the ends are equal but opposite $F_1 = -F_2$, and the velocities are equal $u_1 = u_2$. In order to affect a dynamical model in terms of the four port-variables F_1, F_2, u_1, and u_2, we investigate the leading dynamical modes.

The total energy in the truss is:

$$(7.54) \qquad E = \frac{L\rho}{6}(u_1^2 + u_1 u_2 + u_2^2) + \frac{L}{8\kappa}(F_1 - F_2)^2 \quad ,$$

which is similar to the linear finite-element approximation of a truss, expressed in the applied boundary conditions. The power transfer through the truss is $F_1 u_1$, into the truss at end point u_1 and out again at u_2, with $u_1 = u_2$ and $F_2 = -F_1$. The four variables with the two constraints can be seen as the core model, fully determined by the boundary conditions. The micro-balance determines the force constraint, the power balance determines the velocity constraint. However, more appropriately, the model can be extended to include dynamical relations, rather than constraints between the end points, or port, variables (u_1, F_1) and (u_2, F_2).

The energy of the truss, in combination with the constraints, shows that any dynamical model is incomplete. By a change of force or velocity, the power transfer is nonzero, while for the core model, due to the constraints, the power transfer is always zero. The model can be extended, by adding an approximation to the mean, rigid-body inertial force: $F_1 + F_2$ to the overall, or rigid-body, motion approximated by: $u_1 + u_2$, and for the relative motion $u_1 - u_2$ an internal displacement. Instead we proceed, more systematically, by investigating the dynamical effects of time variation of the input.

If we now have a time-varying force difference $F_1 - F_2$ in the truss, the corresponding state q' will vary, yielding inertial effects. The state will vary with time:

$$(7.55) \qquad \begin{pmatrix} q' \\ p \end{pmatrix} = \begin{pmatrix} \frac{F_1(t) - F_2(t)}{2\kappa} \\ \frac{\rho}{2}(u_1 + u_2) + \rho \frac{\dot{F}_1(t) - \dot{F}_2(t)}{\kappa} \frac{x}{L} \end{pmatrix} \quad ,$$

where $x = [-L/2, L/2]$, the reference coordinate along the truss, which is the solution to the differential equation:

$$(7.56) \qquad \partial_x \frac{p}{\rho} = \dot{q}' = \frac{\partial}{\partial t} \frac{F_1(t) - F_2(t)}{2\kappa} \quad .$$

The time-variation of applied force will cause therefore variation in the relative end-point velocity $u_1 - u_2$.

Only at the next order, $\ddot{F}_1 - \ddot{F}_2$ the effect of the inertia will cause a change in the forces at the end points. The boundary flux Fu based analysis yield a polynomial expansion in modes, due to the homogeneous and isotropic nature, rather than a trigonometric expansion typical for $F_i = 0$ or $u_i = 0$ boundary conditions:

$$(7.57) \qquad q'(x) = d_0 + d_1 \frac{2x}{L} \quad ,$$

where the constant deformation is d_0 and space-dependent deformation is d_1. The corresponding the canonical momenta are

$$(7.58) \qquad p(x) = b_0 + b_1 \frac{2x}{L} \quad .$$

These four amplitudes d_0, d_1, b_0, and b_1 are directly related to the four port variables:

$$(7.59) \qquad \begin{pmatrix} d_0 \\ d_1 \\ b_0 \\ b_1 \end{pmatrix} = \begin{pmatrix} \frac{F_1 - F_2}{2\kappa} \\ \frac{F_1 + F_2}{2\kappa} \\ \frac{\rho(u_1 + u_2)}{2} \\ \frac{\rho(u_2 - u_1)}{2} \end{pmatrix} \ .$$

The energy due to the port-Hamiltonian, restricted to the space of these two modes is given by ($D = \partial_x$):

$$(7.60) \qquad E = \int_{-L/2}^{L/2} dx\, H' = \frac{Lb_0^2}{2\rho} + \frac{Lb_1^2}{6\rho} + \frac{\kappa L d_0^2}{2} + \frac{\kappa L d_1^2}{6} \ .$$

The equations of motion of the stationary system are:

$$(7.61) \qquad \begin{pmatrix} \dot{d}_0 \\ \dot{d}_1 \\ \dot{b}_0 \\ \dot{b}_1 \end{pmatrix} = \begin{pmatrix} \frac{2b_1}{\rho L} \\ 0 \\ -\frac{2\kappa d_1}{L} \\ 0 \end{pmatrix} ,$$

which, once expressed in terms of F_1, F_2, u_1, and u_2, yield two conditions among the four port variables, and two boundary conditions, either for F_1 and u_2, or for F_2 and u_1. The other boundary conditions, like for F_1 and F_2, would arise from the pseudo-stationary state, with, in this case, $p_t \neq 0$.

7.8. Pseudo-symplectic operators

From a formal perspective, the conserved energy of a Hamiltonian is not so much the consequence of the Hamiltonian itself, but of the symplectic matrix J in front:

$$(7.62) \qquad \dot{x} = J\nabla_x H \ ,$$

where $J^T = -J$, which in the canonical form reduces to $J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}$. In the port-Hamiltonian framework the symplectic form is replaced by an arbitrary skew-adjoint differential operator, while the model reduction, based on the core model, changes the differential operator D in shift operators, as shown in Section 2.11, $S_\sigma(x_0, x_1, x_2, \cdots) = (x_\sigma, x_{\sigma+1}, x_{\sigma+2}, \cdots)$ acting on vectors $(x_0, x_1, x_2, \cdots)$ in half-infinite spaces:

$$(7.63) \qquad J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix} \rightarrow \begin{pmatrix} 0 & D \\ -D^* & 0 \end{pmatrix} \rightarrow \begin{pmatrix} 0 & S_\sigma \\ -S_{\sigma'} & 0 \end{pmatrix} \ .$$

The first step is the change of variable from q to $q' = Dq$. The second step is finding the modes for which the differential operators D and D^* reduce to shift operators S_σ and $S_{\sigma'}$. The modes, with amplitudes $(x_0, x_1, x_2, \cdots)$, are successively constructed from the core model states.

Depending on the choice of boundary conditions, the values of σ and σ' change. In the example of the wave equation above, Section 7.3, based on the stationary state, $\sigma = \sigma' = 1$. The sum of the two null spaces $\sigma + \sigma'$ is the dimension of the core model; the states not determined by the equations of motion, but either constants of motion, or determined by the port variables.

In the case of $\sigma = 2$ and $\sigma' = 0$ a constant nonzero q_t exists. Likewise, in the case of $\sigma = 0$ and $\sigma' = 2$ a constant nonzero p_t exists. The first case corresponds to u_1 and u_2 given, the second case corresponds to F_1 and F_2 given.

For the homogeneous string they are linear approximations of the states which satisfy the particular boundary conditions. The index σ, σ' yields a classification of boundary conditions.

In the case of the differential operator D and its adjoint D^* the signatures are the same $\sigma = \sigma'$. Quite often this is not the case. See also Section 5.3. In some cases the boundary conditions dictate a static core model, corresponding to a simple spring model. The Euler-Bernoulli beam is a case where the dynamical modes depend on the choice of boundary conditions. If force or torque is applied, more dynamical degrees of freedom arise, than in the case of position and angle boundary conditions.

The Euler-Bernoulli beam model, Eq. 7.48, will be discussed in detail in Sections 8.2 and 11.1. A segment of the Euler-Bernoulli beam can be modeled with a third-order polynomial:

$$(7.64) \qquad\qquad y(x) = a_0 + a_1 x + \frac{1}{2}a_2 x^2 + \frac{1}{6}a_3 x^3 \quad .$$

Acting on the coefficient vector $\mathbf{a}$ the differential operators D are represented by shift operators:
$$(7.65)$$
$$\partial_x = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \qquad \partial_x^2 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \qquad \partial_x^3 = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad .$$

In the symmetric case with $D = D^* = \partial_x^2$, the signature is $\sigma = \sigma' = 2$. This would for example be the case of a beam position and direction controlled at one end, say $x = 0$, with a_0 and a_1 given by the control, and a mass at the free end $x = 1$, such that the forces and torque at $x = 1$ are associated with acceleration of the mass.

The static case usually studied for the Euler-Bernoulli beam, however, is associated with the signature $\sigma = 4$ and $\sigma' = 0$. The positions and angles are given at both end, and no force comes into play.

The case of $\sigma = 0$ and $\sigma' = 4$ is associated with forces and torques at both end. This could correspond, for example, to a free-falling beam, with masses or springs attached to both ends.

The typical study of the Euler-Bernoulli beam [Beitz et al., 1994] is the loaded beam, where vertical forces are applied to the horizontal beam, clamped or supported. As a simple example, we consider a beam $x \in [0, 2]$ with a vertical force F applied in the middle $x = 1$. The situation is symmetric around $x = 1$, so we only consider one half: $x \in [0, 1]$. The energy is given by:

$$(7.66) \qquad\qquad E = \int_0^1 dx \left(\frac{1}{2}\dot{y}^2 + \frac{1}{2}(\partial_x^2 y)^2 \right) \quad .$$
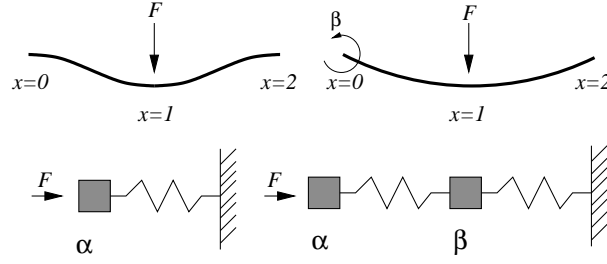
FIGURE 7.2. The two situations under consideration. Left the clamped boundary conditions, right the supported boundary conditions. A force F is applied in the middle. Below the corresponding ideal models, with one and two degrees of freedom.

The mass matrix in terms of the coefficient vector $\mathbf{a}$ is determined by the integrals over the product of modes:

(7.67)
$$
\mathbf{M} = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{6} & \frac{1}{24} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{8} & \frac{1}{30} \\ \frac{1}{6} & \frac{1}{8} & \frac{1}{20} & \frac{1}{72} \\ \frac{1}{24} & \frac{1}{30} & \frac{1}{72} & \frac{1}{252} \end{pmatrix} \quad \mathbf{M}^{-1} = 16 \begin{pmatrix} 1 & -7.5 & 3 & -52.5 \\ -7.5 & 75 & -337.5 & 630 \\ 30 & -337.5 & 1620 & -3150 \\ -52.5 & 630 & -3150 & 6300 \end{pmatrix} .
$$

The Hamiltonian for the finite approximation on $[0, 1]$ is:

(7.68)
$$
H = \frac{1}{2} \dot{\pi}_{\mathbf{a}}^T \mathbf{M}^{-1} \dot{\pi}_{\mathbf{a}} + 2a_2^2 + 12a_2 a_3 + 18a_3^2 \ ,
$$

where $\pi_{\mathbf{a}} = \mathbf{M}\dot{\mathbf{a}}$ is the momentum associated with the coefficient vector $\mathbf{a}$. Two or three degrees of freedom are given by position constraints, depending on supported or clamped boundary conditions; one or two at $x = 0$, and one $\partial_x y = 0$ at $x = 1$, due to the symmetry. See Figure 7.2.

The absence of a_0 and a_1 in the elastic energy means the elastic energy has a two-dimensional null-space. If the values of a_0 and a_1 are fixed by boundary conditions, it means that their time-derivatives $\dot{a}_0$ and $\dot{a}_1$ are also fixed, as secondary constraints, [Dirac, 1964] which has consequences for the momentum $\pi_{\mathbf{a}}$. These are the four boundary conditions which make up $\sigma + \sigma'$.

In the case of clamped boundary conditions $a_0 = y_0$ and $a_1 = y_0'$ there is only a single mode ϕ_{clamped} with coefficient α, with parametric dependence on the boundary conditions:

(7.69)
$$
\phi_{\text{clamped}}(x) = y_0 + y_0' x(1-x)^2 + (3x^2 - 2x^3)\alpha \ .
$$

The corresponding port-Hamiltonian, with $D = \partial_x^3$, yield equations of motion with a pseudo-symplectic operator of rank 4 and signature $\sigma = 3$ and $\sigma' = 1$:

(7.70)
$$
\begin{pmatrix} \dot{\mathbf{a}} \\ \dot{\pi}_{\mathbf{a}} \end{pmatrix} = \begin{pmatrix} 0 & \partial_x^3 \\ \partial_x & 0 \end{pmatrix} \begin{pmatrix} \delta_{\mathbf{a}} H' \\ \delta_{\pi_{\mathbf{a}}} H' \end{pmatrix} \ .
$$

The bottom row is the force equation, the top row is like a constraint equation. Only the dynamics of a single degree of freedom is determined by the top equation, due to the differential operator ∂_x^3, with a null-space of dimension 3.

In the case of the supported beam, the boundary condition y_0' is absent, and a second mode with coefficient β arises:

$$(7.71) \qquad \phi_{\text{supported}}(x) = y_0 + x(1-x)^2\beta + (3x^2 - 2x^3)\alpha \quad .$$

See Figure 7.2. We will not pursue the mapping between the coefficient vector $\mathbf{a}$ and the reduced models variables α and β. In Sections 8.2 and 11.1 the problem is further investigated, especially concerned the matching of variables at the boundaries, and the resulting constraints.

7.9. Conclusion

In this chapter we bring out different aspects of global versus local modelling. Consistent, general, and versatile links between the two approaches are the fluxes, which link micro-balance laws with their global equivalences. Among these fluxes, the energy flux, or power, plays a central role, as the equations of motion are determined by the energy density in canonical form; the Hamiltonian. The internal dynamics, such as vibrations, are no longer a direct consequence of the interaction with the surrounding, but due to the coupling between the core, or lumped, model, and the remainder of the infinite number of degrees of freedom presence in the local model.

Different physical problems and geometries are investigated in later chapters. In some cases the core model might have a nontrivial solution, which has it consequences for the internal dynamics. Furthermore, different control problems follow from the core model. For example, for arbitrary dynamics, the pseudo-stationary problem might be a better candidate of an optimal control problem, than the corresponding full dynamical problem.

CHAPTER 8

Linear finite element methods

In this chapter we will discuss the typical linear Finite Element Methods (FEM), however, from the perspective of physical modelling, and linking the results with typical lumped models. The success of FEM lies mainly in the efficiency that a numerical results can be obtained for a large to very large system. Nowadays, a typical desktop can treat a linear system with tens of thousands of variables easily. The typical simulation packages for lumped systems, which are generally much more versatile and interactive, can handle at most a couple of hundreds of degrees of freedom.

Hence it is not just extending the number of degrees of freedom in lumped approaches, which would make them design and simulation tools for FEM. It is necessary to devise computational methods more closely related to FEM, but with the modularity and versatility of simulation software.

The groundwork to a marriage of FEM and dynamical modelling is the identification of typical concepts and principles of dynamical systems in FEM. For example, the typical stiffness or conductance matrix of modular FEM is singular. The null-space is associated with rigid-body modes and free flow. In a FEM approach the boundary conditions are set to eliminate the null-space from the solution space. In modelling, however, the null-space plays an essential role in determining how elements can be joined together in a complex system.

8.1. Introduction

The relation between the mathematical theory and the numerical solutions of partial differential equations always has been a difficult one. Already is 1950, C. Tompkins, in a study for the Office of Naval Research, mentions

> the electronic digital calculating machines they [the engineers]
> were developing would replace all mathematicians and, indeed,
> all thinkers except the engineers building more machines.... my
> task was frequently one of pointing out that the machines will
> not run by themselves.

The mathematical "background", produced by D. Bernstein [Bernstein, 1950] at the order of C. Tompkins would, however, send most engineers running, and running codes just hoping for the best. The mathematical theory of existence and uniqueness of solutions of partial differential equations is complicated at best. Many a mathematician has steered clear from partial differential equations. The treatment of partial differential equations as operators on some normed space has been the most prevailing and successful approach to uniqueness and existence. The bounds on errors are, however, not always that useful, and certainly yield little feedback on how to improve upon the result.

Fifty years later the uneasy distance between mathematics and numerical code remains, except for the linear theory, which has developed nicely under strict and easy bounds. Antman, noted for his work on nonlinear elasticity, remarks:

> the danger with using sophisticated numerical procedures on powerful computers is that they produce beautiful figures that bear but a tenuous relation to the phenomena they purport to describe, a danger more succinctly stated in Phil Colella's paraphrase of Mark Twain: "There are lies, damned lies, and colored graphics."

For example, in fluid mechanics, the Navier-Stokes equation is a notoriously hard problem. The Clay Institute will award anybody one million dollar who can prove the existence of regular solutions for given initial and boundary conditions. Despite this open problem millions of fluid dynamical problems are studied by the means of FEM, and the engineers trust the numerical answers enough to fly to the moon, with ship designs based on them.

In order to use FEM for other purposes than to generate large, sparse systems of linear equations, we should keep the underlying continuous physical system in mind. In FEM, typically only the state is retained, the flow is defined implicitly, by the equations. For example, energy, and the energy density is useful to retain and use in simulations. Physicists and mathematicians rely on the energy conservation to test the correctness or stability of an approximation. In general, retaining the energy, or any globally conserved quantity, such as charge or mass, is hard, as it arises often as a global constraint on the systems. Local energy continuity, Eq. 5.39, does not always come natural in the FEM formulation. A posteriori error checking, on the generated results, is more easily done. In a later chapter we will turn our eye to the open playing field of nonlinear FEM. Using an explicit energy-dependent method in terms of Hamiltonians, and symplectic structures for simulation, we assure ourselves of the numerical stability and correctness, even in the case of nonlinearities which cause near singular behavior.

Solutions are bounded by energy. In the absence of energy, the solutions are trivially zero. And the conservation of energy may cause the energy to localize in a small region of the system, but even there it is bounded. Especially in FEM the region in which the energy can focus, or localize, has a minimal size determined by mesh, which yields an upper bound on the amplitude of the solution. However, if all the energy focuses in a single element, it may signal the mesh is to coarse for the problem at hand. Several methods of adaptive mesh refinement are based on bounds similar to the energy bounds. Energy is like as set of parcels in a logistic network. Based on inventory, parcels are redistributed. Diffusion equations are only based on the state of the inventory, or state, while wave equations are based on the direction the energy parcel arrived from. However, even with an explicit energy, it remains complex to related the known partial differential equations to such discrete network models, since the energy is typically a quadratic function of the state, to guarantee the lower bound on the energy.

8.2. Spline as beam

Splines were originally a tool for the technical draftsman. It was a flexible metal strip, which allowed the draftsman to draw a smooth curve to a set of points.
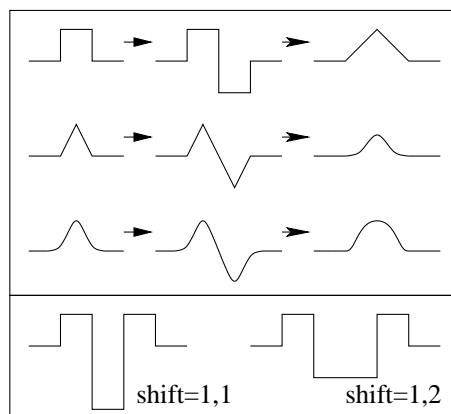
FIGURE 8.1. The successive basis functions for splines can be gen-
erated from the basic block function through shift, flip, end integra-
tion. The shift will determine the shape of the spline, such as in the
case of twice integrated block shapes below. The Euler-Bernoulli
beam is a three times integrated shift-and-flip block function. The
shift = 1,1 (first shift 1, then shift 1) generates a sharper peaked
spline than the shift = 1,2 (first shift 1, then shift 2) block function.

A connection which has almost been severed, since for FEM the spline approxi-
mation of a beam is considered rather cumbersome. Only recently, splines have
returned, from the domain of Computer-Aided Design, in the domain of structural
computations. [Hughes et al., 2005] In order to connect different numerical a phys-
ical aspects we discuss the simple example of cubic splines and the Euler-Bernoulli
beam. See also Figure 8.1.

The spline can serve as an introduction into a nontrivial and physical FEM
modelling. In particular we will discuss the effects of inputs, in terms of the applied
forces, and the elastic energy, which originally led to the Euler-Bernoulli beam
model. [Euler, 1744] A beam connected by both ends can be described by four
degrees of freedom, fixed by the position and angle at both end. Or likewise by the
force and torque at both end. Every point x_i where a force F_i is applied requires
a cut of the beam in two separate segments. Four additional degrees of freedom
are introduced in the model. Three of those are constraint, or algebraic, relations
fixing the segments together. Only the fourth degree of freedom is the dual variable
associated with the applied force. We will see that this is the displacement $y(x_i)$,
based on the elastic beam energy. The product of force and displacement is the
work applied.

The displacement $y(x)$ as function of the horizontal reference position x in the
static case is the solution of the differential equation:

$$(8.1) \qquad\qquad\qquad EI\partial_x^4 y(x) = F(x) \ \ .$$

The cubic spline solution $y(x)$ is a piece-wise cubic function $x_i \leq x \leq x_{i+1}$:

$$(8.2) \qquad\qquad\qquad y(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 \ \ ,$$

where at each boundary x_i the function and the first two derivatives from each
side matches, but the higher derivatives not necessarily. On each of the segments

(x_i, x_{i+1}) the spline is a solution of the force free Euler-Bernoulli equation:

$$(8.3) \qquad\qquad EI\partial_x^4 y(x) = 0 \quad x \notin \{x_i\}_{i=0}^n$$

Only at the boundaries x_i forces occur.

The left higher derivative and the right higher derivative yields a discontinuity, related to a force at x_i:

$$(8.4) \qquad\qquad \lim_{\delta \to 0} \partial_x^3 y(x_i + \delta) - \partial_x^3 y(x_i - \delta) = \frac{F_i}{EI} \quad.$$

Counting the degrees of freedom, each segment has four coefficients a_i, three of those are determined by the boundary conditions between the segments, the fourth is the result of the force F_{i+1}:

$$(8.5) \qquad\qquad F(x) = \sum_i \delta(x - x_i) F_i \quad.$$

Hence, there is a one-to-one map between the displacements $y(x_i)$ and the forces F_i, and endpoint conditions. For n segments, there are $n-1$ forces F_i and displacements $y(x_i)$ and $4n$ coefficients a_j, with $3(n-1)$ relations. Four degrees of freedom are not determined by either the forces or the displacements at the intermediate boundaries between the segments. Between two endpoints these four degrees of freedom are fixed. The boundary, or endpoint, conditions depend on the choice of spline, and the forces used.

Traditionally, the Euler-Bernoulli beam boundary conditions are fixed by either clamped or supported boundary conditions at the endpoints. See Table 6.2. Clamped means both the displacement y and the derivative $\partial_x y$ are fixed, supported means the displacement and the second derivative of the displacement $\partial_x^2 y$ are fixed. Both cases yield a well-defined set of equations for all the coefficients, using either the intermediate displacements $y(x_i)$ or forces F_i. Hence, the spline solution $y(x)$ serves well as the exact solution of the approximated problem of a distributed force $F(x)$:

$$(8.6) \qquad F(x) \to \sum_i \delta(x - x_i) F_i = \sum_i \delta(x - x_i) \int_{\frac{x_i + x_{i-1}}{2}}^{\frac{x_i + x_{i+1}}{2}} dx' F(x') \quad.$$

The determination of the solution, even of the linear map between displacements and forces, is a simple but tedious problem, people prefer to leave to the computer. Each of the forces affects all of the displacements. One can think of a large force in the middle, causing a deflection of the beam as a whole. One way to proceed is the integrate the spline from one end to the other, with the unknown integration constants, to be fixed afterwards by the endpoint boundary conditions. In another approach, for each F_i the response of the displacements of all $y(x_i)$ are easily determined, and the total displacement is a linear combination of all displacements, as long as the endpoint condition is zero, and thereby invariant for the applied forces.

However, there is a different approach to the Euler-Bernoulli spline problem. The pairs of displacements and forces $(y(x_i), F_i)$ are the associated with the power supplied to the system. Each of the segments i has a bending moment, associated with elastic energy:

$$(8.7) \qquad\qquad E_i = \frac{1}{2} EI \int_{x_i}^{x_{i+1}} dx (\partial_x^2 y(x))^2 \quad.$$

For a finite energy $y(x)$ and $\partial_x y(x)$ must be continuous across the boundary. The static solution sought; the solution to the Euler-Bernoulli equation, is the solution which minimizes the bending energy, for the given displacements of forces:

$$(8.8) \qquad \min E \Leftrightarrow \delta E = \delta \sum E_i = 0 \quad ,$$

which means the first variation of the energy, with respect to the independent variables vanishes. Hence, if the forces are given, the solution is minimized for the displacements:

$$(8.9) \qquad \sum_i \int_{x_i}^{x_{i+1}} dx (\partial_x^2 y(x))^2 = \sum_i \partial_x y \partial_x^2 y \Big|_{x_i}^{x_{i+1}} - \sum_i \int_{x_i}^{x_{i+1}} dx \partial_x y(x) \partial_x^3 y(x) \quad .$$

The first term on the right-hand side vanishes of each segment boundary, since the spline is defined to have vanishing first and second derivatives across the boundaries. The second term is easily integrated:

$$(8.10) \qquad -\sum_i \int_{x_i}^{x_{i+1}} dx \partial_x y(x) \partial_x^3 y(x) = -\sum_i y \partial_x^3 y \Big|_{x_i}^{x_{i+1}} + \sum_i \int_{x_i}^{x_{i+1}} dx y(x) \partial_x^4 y(x) \quad .$$

The last term vanishes since $\partial_x^4 y(x) = 0$ on each of the segments. It is not completely trivial. If the sum and the integral would have been interchanged, the boundary effects would have been part of the last term.

Hence the elastic energy can be expressed as:

$$(8.11) \qquad E = -\sum_i \frac{1}{2} y(x_i) F_i \quad ,$$

since the jump of $\partial_x^3 y(x)$ across the segment boundary is the force, and the displacement is continuous across the boundary.

For a given force F_i the displacements $y(x_i)$ are minimized, and for a given displacement the forces are minimized. Without any constraint, the minimization problem would yield a zero solution. With the constraint, the minimization yield the lowest attainable values for both displacement and force.

Later, when we discuss boundary polynomials, we will deal with numerical implementations of spline-like approximations of continuous systems. Solving the stationary situation can be achieved through several means.

8.3. Finite element method

Finite element method is a name for a collection of finite discretizations of continuous systems. Some make the distinction like finite difference and finite volume, and boundary element methods, but that is more a matter of interpretation of the approximation, than fundamental differences.

The simplest interpretation consists of dividing the space in little cells and use a very simple, lumped or discrete, submodel for every cell, such that the total model is a lumped model of cells. The common picture is the approximation of a string by a series of masses and springs, which is a unfortunate analogy, as we will see.

A series of masses and spring will always have a mass *or* a spring at the end, suggesting some breaking of the duality between elastic and kinetic energy, inherent in the infinite dimensional string model, which yields the wave equation. Much more appropriate would be to turn the string into cells with piece-wise linear coordinate dependence. Given the reference, or rest, coordinate of the string s, the

displacement $x(s)$ is a function of that coordinate s. The appropriate finite element approximation would arise from picking abscissas $\{s_i\}_{i=0}^n$, and assigning finite element variables $x_i = x(s_i)$ to these points, with linear dependence between the points:

$$(8.12) \qquad x(s) = x_i \frac{s - s_{i+1}}{s_i - s_{i+1}} + x_{i+1} \frac{s - s_i}{s_{i+1} - s_i} \quad ,$$

which is the function $x(s)$ for cell $i : s \in [s_i, s_{i+1}]$. Whatever infinite dimensional model we had, it simply translates into a finite, discrete model in terms of the variables $\{x_i\}_{i=0}^n$. The kinetic energy of cell i is nothing but the integral over the distributed kinetic energy:

$$(8.13) \qquad E_{\text{kinetic}} = \frac{1}{2} \int_{s_i}^{s_{i+1}} ds \rho \dot{x}(s)^2 = \frac{\rho}{6} (s_{i+1} - s_i)(\dot{x}_i^2 + \dot{x}_i \dot{x}_{i+1} + \dot{x}_{i+1}^2) \quad ,$$

rather than a local mass, which would corresponds to a local model. Similar, the potential, or elastic energy of cell i of a string would be the integral over the distributed elastic energy, limited to the linear function of x_i and x_{i+1}:

$$(8.14) \qquad E_{\text{elastic}} = \frac{1}{2} \int_{s_i}^{s_{i+1}} ds \kappa (\partial_s x(s))^2 = \frac{\kappa}{2} \frac{(x_{i+1} - x_i)^2}{s_{i+1} - s_i} \quad ,$$

where κ is the elasticity modulus.

The total co-energy is the system can be written as:

$$(8.15) \qquad E^* = E_{\text{kinetic}} + E_{\text{elastic}} = \frac{\rho a}{2} \dot{\mathbf{x}}^T \dot{\mathbf{x}} - \frac{\rho a}{12} \dot{\mathbf{x}}^T \mathbf{K} \dot{\mathbf{x}} + \frac{\kappa}{2a} \mathbf{x}^T \mathbf{K} \mathbf{x} \quad .$$

The kinetic term $\dot{\mathbf{x}}^T \mathbf{K} \dot{\mathbf{x}}$ is associated with the internal vibration of a segment, with no associated center-of-mass motion. The mass matrix is non-diagonal: $\mathbf{M} = \rho a (\mathbf{1} - \frac{1}{6} \mathbf{K})$, which will lead to problem in time-integration, since time-integration requires the inverse of the mass matrix, both in Lagrangian and Hamiltonian settings. See Chapter 3.

In terms of matrices, only a single form arises, as here for five segments, of length $s_{i+1} - s_i = a$:

$$(8.16) \qquad \mathbf{K} = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 \\ 0 & 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & -1 & 1 \end{pmatrix} \quad .$$

The problems with the mass matrix are usually avoided by taking only the diagonal part of the mass matrix, ignoring the internal motion given by $\mathbf{K}$. In the approximation of localized masses at the nodes s_i, the energy would have been:

$$(8.17) \qquad E_{\text{mass-spring}} = \frac{\rho a}{2} \dot{\mathbf{x}}^T \dot{\mathbf{x}} + \frac{\kappa}{2a} \mathbf{x}^T \mathbf{K} \mathbf{x} \quad .$$

For the eigenvalues λ_i of the matrix $\mathbf{K}$, the corresponding eigenfrequencies are:

$$(8.18) \qquad \omega^2 = \frac{6 \kappa \lambda}{a^2 \rho (6 + \lambda)} \quad .$$

For small frequencies, the result reduces to the typical stiffness over mass relation, equivalent to the mass-spring result:

$$(8.19) \qquad \omega_{\text{small}} = \sqrt{\frac{\kappa\lambda}{a^2\rho}} \ .$$

However, for large frequencies, each of the segments has a maximal frequency, related to the internal mode of a single segment:

$$(8.20) \qquad \omega_{\text{max}} = \sqrt{\frac{6\kappa}{a^2\rho}} \ .$$

If the wave length of the mode is shorter than the segment length a, the result is no longer reliable in this finite approximation.

Interestingly, by examination of the matrix $\mathbf{K}$ it follows that the null-space displacement $\mathbf{x}$ with $\mathbf{Kx} = 0$ corresponds to the rigid-body motion. For an equal displacement of all the modes $x_i = $ constant the elastic energy is zero:

$$(8.21) \qquad \mathbf{K} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} = 0 \ .$$

Furthermore, except for the endpoints, the linear function $x_i = s_i/a$ also yields a zero result:

$$(8.22) \qquad \mathbf{K} \begin{pmatrix} 0 \\ 1 \\ 2 \\ \vdots \\ n \end{pmatrix} = \begin{pmatrix} -1 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} \ .$$

These results follow easily from $\mathbf{K}$ being the discrete approximation of the second-order differential operator, corresponding to the piece-wise linear approximation:

$$(8.23) \qquad \partial_s^2 \to \frac{1}{a}\mathbf{K} \ .$$

Hence $\partial_s^2 1 = 0$ and $\partial_s^2 s = 0$. The boundary effects arise from the initial and final node, x_0 and x_n, being part of only one segment. Boundary conditions should be implemented to yield the appropriate solution $\partial_s^2 s = 0$. In practice, a linear displacement function may arise as stationary solution, if a force is applied to the boundaries.

For the string, this model using the piece-wise linear approximation is sufficient to first order. Later, more complicated finite-order models will be discussed. For example, for the beam, such a simple piece-wise linear model is insufficient. The elastic energy in this approximation is either infinite or zero depending on the perspective. On each of the segments:

$$(8.24) \qquad E_{\text{bending}} = \frac{1}{2}\int_{s_i}^{s_{i+1}} ds EI(\partial_s^2 x(s))^2 = 0 \ .$$

However, across the segment:

$$(8.25) \qquad E_{\text{boundary}} = \lim_{\delta \to 0} \frac{1}{2} \int_{s_i - \delta}^{s_i + \delta} ds \, EI(\partial_s^2 x(s))^2 = \lim_{\delta \to 0} \infty = \infty \quad ,$$

which shows two important aspects of any finite-order approximation of a continuous model. First, the approximation functions should be appropriate for the differential operator. Second, the boundaries can contribute to the energy, depending on the formulation. In an appropriate function space neither aspect will occur, however, special care must be taken to formulate the model in such a way that the boundaries themselves do not contribute to the energy.

8.4. Stencils

Splines are an example of functions defined on a set of nodes, points, or abscissas. Such a set of points are called a stencil. Stencils are used in all kind of discrete methods, such as finite element methods. [Milne, 1970]

For example, to define a smooth function $f(z)$ through a set of values $f_i = f(z_i)$, one uses at least two points $z_i \leq z \leq z_{i+1}$. The stencil is, in that case, the pair of points $\{z_i, z_{i+1}\}$:

$$(8.26) \qquad f(z) = \frac{z_{i+1} - z}{z_{i+1} - z_i} f(z_i) + \frac{z - z_i}{z_{i+1} - z_i} f(z_{i+1}) \quad .$$

Such an approximation can serve again to an approximation for the derivative:

$$(8.27) \qquad \partial_z f(z) = \frac{f(z_{i+1}) - f(z_i)}{z_{i+1} - z_i} \quad ,$$

for $z_i \leq z \leq z_{i+1}$.

The time-integration of an ordinary differential equation can be seen as a stencil. For a n-th order differential equation at least n points $\{t_{i-n+1}, t_{i-n+2}, \cdots t_i\}$ are required. However, more advanced methods of time-integration use more nodes. A smoother function $f(t)$ is the result, which will improve the forward extrapolation. For spatial functions the similar principles are used for other purposes, such as the construction of a Laplacian $\Delta f(z_i)$ on a grid, using neighbor points.

In the case of hyperbolic differential equations, Godunov theorem states that any higher-order approximation of the spatial derivatives will lead to oscillatory results. It is the simulation equivalent of Gibbs phenomenon for Fourier series. Nonlinear schemes of constructing derivatives have been developed to reduce the oscillations. [Liu et al., 1994] This can be seen as the nonlinear weighting of several stencils. Such stencils must be higher order, for example, for a point $z: \quad z_i \leq z \leq z_{i+1}$, the stencils $\{z_{i-1}, z_i, z_{i+1}\}$ $\{z_i, z_{i+1}, z_{i+2}\}$ can be used.

In the case of a linear approximation, the stencil function $f_i(z)$ is the dependence of the interpolation $f(z)$ on discrete value $f(z_i)$:

$$(8.28) \qquad f_i(z) = \frac{\partial f(z)}{\partial f(z_i)} \quad .$$

The linear stencil function, or first-order spline, would be:

$$(8.29) \qquad f_i(z) = \begin{cases} \frac{z - z_{i-1}}{z_i - z_{i-1}}, & z_{i-1} \leq z \leq z_i \\ \frac{z_{i+1} - z}{z_{i+1} - z_i}, & z_i \leq z \leq z_{i+1} \end{cases} \quad .$$

The stencil function with a smooth derivative can be multiple, but will always have a larger stencil. Not only at the internal nodes z_i the function must be smooth,

but also at the boundary nodes of the stencil. The cubic spline approximation can be reformulated as stencil function on a stencil of five nodes:

$$(8.30) \qquad f_i(z) = \begin{cases} -\frac{(z-z_{i-2})^2(z_{i-1}-z)}{m_-}, & z_{i-2} \leq z \leq z_{i-1} \\ \frac{(z_{i+1}-z)(z-z_{i-1})}{(z_{i+1}-z_i)(z_i-z_{i-1})}, & z_{i-1} \leq z \leq z_{i+1} \\ -\frac{(z_{i+2}-z)^2(z-z_{i+1})}{m_+}, & z_{i+1} \leq z \leq z_{i+2} \end{cases} \quad ,$$

where the coefficients $m_\pm$ are:

$$(8.31) \qquad m_\pm = \frac{(z_{i\pm1} - z_{i\pm2})^2(z_i - z_{i-1})(z_{i+1} - z_i)}{(z_{i+1} - z_{i-1})} \quad ,$$

such that the derivative at $z = z_{i\pm1}$ is continuous. See also Figure 8.1, in Section 8.2 where this type of spline can be generated by a shift $= 1$, followed by a shift $= 2$.

Most stencil functions are a partition of unity. In the case of a constant function sampled by $f(z_i) = $ constant, the sum of all the stencil functions should reduce to the constant unity function:

$$(8.32) \qquad \sum_i f_i(z) = 1 \quad .$$

However, for example, in the case of positivity bounds $f_i(z) \geq 0$, one could consider a simple stencil function, with a smooth derivative, which is not a partition of unity:

$$(8.33) \qquad f_i(z) = \frac{(z - z_{i-1})^2(z_{i+1} - z)^2}{(z_i - z_{i-1})^2(z_{i+1} - z_i)^2}, \quad z_{i-1} \leq z \leq z_{i+1} \quad .$$

However, typically, a maximal deviation of 12% occurs from the partition of unity. See Figure 11.8.

Higher-order derivatives also require higher order stencils. The simplest approximation of the second derivative need a three-point stencil:

$$\frac{1}{2}\partial_z^2 f(z_i) = \frac{(z_{i+1} - z_i)f(z_{i-1}) - (z_{i+1} - z_{i-1})f(z_i) + (z_i - z_{i-1})f(z_{i+1})}{(z_i - z_{i-1})(z_{i+1} - z_{i-1})(z_{i+1} - z_i)}$$

$$(8.34) \qquad\qquad \rightarrow \frac{f(z_i - \delta) - 2f(z_i) + f(z_i + \delta)}{\delta^2} \quad .$$

In higher dimensions the stencil can extend in different directions. The Laplacian is the typical, isotropic differential operator. The simplest approximation in two dimensions is on a cross stencil $\{z_{i-1,j}, z_{i,j-1}, z_{i,j}, z_{i+1,j}, z_{i,j+1}\}$, which yields the sum of the two second-order derivatives. In the case of the homogeneous square grid the Laplacian is:

$$(8.35) \qquad \Delta_+ f(z_i) = \frac{f(z_{i-1,j}) + f(z_{i,j+1}) - 4f(z_{i,j}) + f(z_{i+1,j}) + f(z_{i,j+1})}{\delta^2} \quad .$$

However, one could also use a diagonal stencil $\{z_{i-1,j}, z_{i,j-1}, z_{i,j}, z_{i+1,j}, z_{i,j+1}\}$ on a square grid:
$$(8.36)$$
$$\Delta_\times f(z_i) = \frac{f(z_{i-1,j-1}) + f(z_{i-1,j+1}) - 4f(z_{i,j}) + f(z_{i+1,j-1}) + f(z_{i+1,j+1})}{2\delta^2} \quad ,$$

where $\sqrt{2}\delta$ is the diagonal grid size.

Combining the cross and the diagonal stencil in the right manner decrease the order of the error. The error is typically defined as the order of the analytical expansion which is no longer correctly represented by the discrete approximation

of the Laplacian.[Milne, 1970] The order-improved Laplace operator is the linear combination:

$$(8.37) \qquad \Delta = \frac{1}{3} \left(2\Delta_+ + \Delta_\times \right) \quad .$$

Besides square grids there are also triangular, honeycomb, and other grids. The second-order analytical expansion can be used to define a proper approximation of the Laplacian. In the case of general geometry, (rectangular, triangular, hexagonal, etc.) the derivative should reproduce the continuous derivative as well as possible. In order to check this we consider a symmetric system with one element center and all his neighbors, of the same shape and size. The Laplacian should be:

$$(8.38) \qquad \Delta f(\mathbf{z}_i) = \gamma \sum_{j:\text{neighbors of } i} \frac{f(\mathbf{z}_j) - f(\mathbf{z}_i)}{\|\mathbf{z}_i - \mathbf{z}_j\|^2} \quad ,$$

where $\mathbf{z}_i$ is the barycenter of center element i, with value $f(\mathbf{z}_i)$. The geometric constant γ we are after. We consider a continuous function $f(\mathbf{z}) = \frac{1}{2}\|\mathbf{z} - \mathbf{z}_i\|^2$. The integral can be calculated in close form:

$$(8.39) \qquad \int_{\text{cell } i} \Delta f d^{\dim} z = A_i \sum_{j:\text{neighbors of } i} \gamma \frac{f(\mathbf{z}_j) - f(\mathbf{z}_i)}{\|\mathbf{z}_j - \mathbf{z}_i\|^2} = A_i \frac{1}{2} \kappa_i \gamma \quad ,$$

where A_i is the area of element i and κ_i is the number of neighbors. Clearly the continuous result should yield the total area times the dimension, since $\Delta \frac{1}{2}\|\mathbf{z} - \mathbf{z}_i\|^2 = \dim$. Hence we find:

$$(8.40) \qquad \gamma = \frac{2\dim}{\kappa_i} \quad .$$

For any grid, this discrete approximation of the Laplacian leads to an error of the order $\mathcal{O}(\delta^3)$ where $\delta = \|\mathbf{z}_j - \mathbf{z}_i\|$.

8.5. Energy stencils

The stencils of the differential operators are related to a particular discretization of the energy functional, the energy stencil, which is a bilinear functional:

$$(8.41) \qquad E[f, g] = \frac{1}{2}\langle Dg, Df \rangle \quad ,$$

where f is the state variable and g is the associated dual, which in most cases where a Hilbert-space inner product is used $\langle x, y \rangle = \overline{\langle y, x \rangle}$, is from the same state space. The energy functional associated with the Laplacian through the variation principle is:

$$(8.42) \qquad \delta_f E = \frac{1}{2}\delta_f \langle \nabla f, \nabla f \rangle \Rightarrow -\Delta f = 0 \quad .$$

The stencil operator Δ_+ corresponds to nearest-neighbor terms in the discretized gradient only:

$$(8.43) \qquad \nabla_{x,\pm} f = \frac{f(z_{i\pm 1,j}) - f(z_{i,j})}{|z_{i\pm 1,j} - z_{i,j}|} \quad ,$$

where the direction of the difference vector $\delta_{x,\pm} = z_{i\pm 1,j} - z_{i,j}$ determines the direction $(x, \pm)$ of the gradient. The cross operator $\Delta_\times$ is related to diagonal difference vectors:

$$(8.44) \qquad \delta_{\pm,\pm} = z_{i\pm 1, j\pm 1} - z_{i,j} \quad .$$
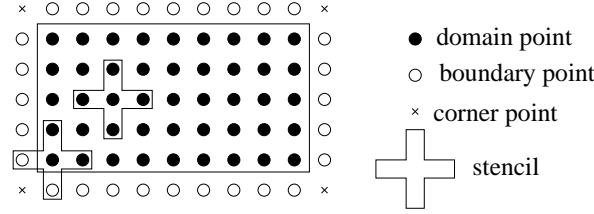
FIGURE 8.2. For each domain point there is a linear relation, defining the values on the domain points uniquely. Hence the set of boundary points are such that all the stencils of Δ_+ are covered. In the case of the $\Delta_\times$ stencils four more points at the corners are required.

The energy stencil function of the second-order improved Laplacian is the weighted sum of both energies:

$$(8.45) \qquad E[f] = \frac{1}{3} \sum_{(i,\pm)} (\nabla_{i,\pm} f)^2 + \frac{1}{6} \sum_{(\pm,\pm)} (\nabla_{\pm,\pm} f)^2 \quad .$$

The factors are determined from the square of the relative distance between an edge length a and a diagonal length $\sqrt{2}a$, which leads to the same result as Eq. 8.37.

8.6. Boundary conditions for stencils

A typical discretization of a partial differential equation on a grid leads to a stencil as we have seen in the previous sections. A stencil is a linear relation among the values on a number of nodes; the discretized version of the partial differential equation, such as $\Delta f = 0$. Hence, the successive application of stencils makes it possible to reconstruct the solution in a domain given the values on the boundary. See Figure 8.2. The stencils already show that the equations cannot be solved successively from one corner to another. At the boundaries four domain points are necessary, at the corner three points, intertwining all the equations.

Therefore not only the type of partial differential equation, but also the finite-order approximation determines the type and number of boundary conditions, for time-dependent systems as well. There is no real difference between the Laplace equation and the wave equation in that respect. Different uses of the same stencil lead to different boundary conditions. In the case that a system can be integrate forwardly, we can denote a node in the stencil as the dependent variable. Using the cross stencil Δ_+ for the integration of the wave equation, the single node with a larger t: $f(z_i, t_{j+1})$ is the normal dependent variable, expressed in terms of the other values:

$$(8.46) \qquad f(z_i, t_{j+1}) = f(z_{i+1}, t_j) + f(z_{i-1}, t_j) - f(z_i, t_{j-1}) \quad ,$$

which is the discretization of the homogeneous wave equation:

$$(8.47) \qquad (\partial_t^2 - \partial_z^2) f(z, t) = 0 \quad .$$

The stencil is symmetric with respect to space. At each end of the domain a single boundary condition is specified. Double boundary conditions only at one end of the domain are also possible. This requires the x-stencil $\Delta_\times$ instead of the cross-stencil Δ_+. See Figure 8.3.
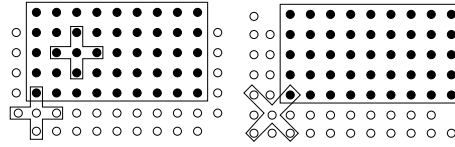
FIGURE 8.3. The boundary conditions corresponding to the simplest wave stencils.

8.7. Finite volume versus finite difference

At lowest order, piece-wise linear approximations, of continuous systems, there is no fundamental difference between finite-volume and finite-difference methods. The equations are the same. At higher orders some differences arise. In general, one could argue that finite volume methods have to match at common boundaries between elements, and finite difference methods have overlapping stencils.

Spectral element methods are an extension to the finite volume method where a set of basis functions on each element are matched at the boundary by neighboring functions. For example, in fluid dynamics, the velocity fields approximations are matched at the boundaries, using Gauss-Lobatto quadrature points. These are associated with orthogonal polynomial solutions with nodes, or abscissas, at the boundary and in the interior. The abscissas at the boundary are associated with the weight function, the orthogonal expansion will be Gegenbauer or Jacobi polynomials of a particular order. For example, on a domain $z \in [-1, 1]$, a weight function $(1 - x^2)^j$ will introduce j-th order zeros at the boundary, associated with j boundary values. The result has an error in the order of the square of the polynomial order, since the inner-products vanish:

$$(8.48) \qquad \int_{-1}^{1} (1 - z^2)^{\alpha - \frac{1}{2}} C_i^{(\alpha)}(z) C_j^{(\alpha)}(z) dz = \frac{\pi 2^{1-2\alpha} \Gamma(i + 2\alpha)}{i!(i + \alpha)\Gamma(\alpha)^2} \quad ,$$

where $\alpha = \frac{5}{2}$. Any solution $f(z) = \sum_{i=0}^{n} a_i (1 - z^2) C_i^{(\alpha)}(z)$ orthogonal to any polynomial to the order n will yield a zero approximate integral, if the abscissas z_i are chosen as the zeros of $C_{n+1}^{(\alpha)}(z_i) = 0$, with the appropriate weights w_i:

$$(8.49) \qquad \int_{-1}^{1} f(z)g(z)dz = \sum w_i f(z_i)g(z_i) \quad .$$

There are the Gauss-Lobatto quadrature abscissas and weights. [Abramowitz and Stegun, 1965, Hildebrand, 1987]

For example, in the case of fluid dynamics, the velocities have to be matched at the element boundaries. The pressure, on the other hand is only needed at the internal points. A different set of abscissas and weights are used for the pressure, in the conjunction with the weights and abscissas for the velocities. The velocity and the pressure are the intensive and the extensive variables of fluid dynamics. The Navier-Stokes equation will tell us as much, under close examination of the dimensions of the different variables. The energy, or the Hamiltonian, likewise separates the state variables and the effort variables, or potentials.

The finite volume method has typically been associated [Saad, 1995] with conservation laws. The flow out of the volume equals the change of a conserved density

such as mass or charge:

$$(8.50) \qquad \int_{\Omega} \dot{Q} d^n z = \int_{\partial\Omega} \mathbf{n} \cdot \mathbf{J} d^{n-1} z = \int_{\Omega} \nabla \cdot \mathbf{J} d^n z \quad .$$

Hence it would be natural use the conserved density Q_i on the finite volumes Ω_i to retain the conservation law exactly. The equations of motion would be the instantaneous exchange of the conserved density:

$$(8.51) \qquad \dot{Q}_i = J_{ij}, \quad \dot{Q}_j = J_{ji} \quad ,$$

where $J_{ij} = -J_{ji}$. More complex function $Q_i(z)$ on the domain Ω_i will yield a more elaborate exchange term:

$$(8.52) \qquad J_{ij} = \int_{\partial\Omega_i \cap \partial\Omega_j} \mathbf{n}_{ij} \cdot \mathbf{J}(z) d^{n-1} z \quad .$$

The closure relation expresses the flow $\mathbf{J}(Q)$ in terms of the obvious state variable $Q(z)$. In practice, the boundary flow J_{ij}, through the common boundary $\partial\Omega_i \cap \partial\Omega_j$, depends on the conserved density on both whole domains $Q_i(z)$ on Ω_i and $Q_j(z)$ on Ω_j.

Hence the finite-volume approach for the conserved density yields a finite-difference approach for the conserved flow. This is a general result. In the case that the closure relation $J(Q)$ is implicitly used, the finite difference part is directly expressed in terms of the finite volume terms. The conservation law will still be exact, but over the quality of the closure-relation approximation there remains little control.

8.8. Model reduction

For the use of FEM in simulation and control model reduction is more or less essential. The state vector $\mathbf{x}$ is too large to be handled and updated within milliseconds, typical for interactive simulations and control on a personal computer.

Hence, from the large space $\mathbf{x} \in \mathbb{R}^N$ we should select a smaller space $\mathbf{c} \in \mathbb{R}^n$ of coefficients such that $n \ll N$:

$$(8.53) \qquad \mathbf{x}(\mathbf{c}) = \sum_{i=1}^{n} c_i(t)\mathbf{x}_i \quad ,$$

represents most of the dynamical behavior. The state vectors $\mathbf{x}_i$ are the selected, normalized modes. The time-dependence is given by the coefficient vector $\mathbf{c}$.

In the process of model reduction two aspects play a role. First, the selection of the modes $\mathbf{x}_i$. Secondly, the projection of the dynamics. The simplest approach would be to project the dynamics on the same set of modes:

$$(8.54) \qquad \mathbf{x}_j^T \mathbf{x}_i \dot{\mathbf{c}} = \mathbf{x}_j^T \mathbf{A} \mathbf{x}_i \mathbf{c} + \mathbf{x}_j^T \mathbf{B} \mathbf{u} \quad ,$$

where the matrix $E_{ji} = \mathbf{x}_j^T \mathbf{x}_i$ is the mass matrix which multiplies the rate-of-change of the coefficients vectors $\mathbf{c}$, and $A_{ji} = \mathbf{x}_j^T \mathbf{A} \mathbf{x}_i$ is the reduced $\mathbf{A}$ matrix.

It is very important to notice that picking a set of modes does not yet introduce an approximation error. The projection determines the loss of accuracy. The

dynamics of the modes may evolve the system outside the subspace spanned by the
modes:

$$(8.55) \qquad \mathbf{x} = \mathbf{x}_{\parallel} + \mathbf{x}_{\perp} = \sum_{i=1}^{n} c_i \mathbf{x}_i + \mathbf{x}_{\perp} \quad,$$

where the inner-product $\langle \mathbf{x}_{\perp}, \mathbf{x}_i \rangle = 0$. Hence the perpendicular part determines
the quality of the reduced order model.

Furthermore, the standard inner-product $\langle \mathbf{y}, \mathbf{x} \rangle = \mathbf{y}^T \mathbf{x}$ may be highly unsuited
to determine the quality of the model. For example, modelling a heat-conductance
problem, with the temperature as state vector $\mathbf{x}$, the air gaps in a system have a
low heat capacity and conductance. Hence, air gaps have little effect on the overall
temperature dynamics, but are slow to respond to changes. Hence, if the temper-
ature inner-product: $\langle \mathbf{y}, \mathbf{x} \rangle$; the direct product of temperature with temperature
summed of all nodes, serves to determine the quality of the model, the air gap
temperature may play an important role in these inner-products. Hence a more
appropriate inner-product would be the one weighted with the heat capacitance
matrix $(\mathbf{y}, \mathbf{x}) = \mathbf{y}^T \mathbf{C} \mathbf{x}$, however, it still has little physical meaning.

8.9. Proper orthogonal decomposition

Proper orthogonal decomposition [Holmes et al., 1996] is a form of model re-
duction which relies on the numerical experiment to select a number of modes.
The large FEM code simulates a long time sequence, with variable input, within
the appropriate range. From the time series a number of high-dimensional states
$\mathbf{x}_i = \mathbf{x}(t_i)$ are selected as modes from the model order reduction.

In order to select an appropriate set $\{\mathbf{x}_{j_i}\}_0^n$ from a larger set $\{\mathbf{x}_i\}_0^N$, one studies
the correlations in the larger set. If two vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ are the same or parallel,
only one has to appear in the smaller set. Two vectors are parallel if:

$$(8.56) \qquad (\mathbf{x}_i^T \mathbf{x}_j)^2 = \mathbf{x}_i^2 \mathbf{x}_j^2 \quad.$$

Constructing a matrix from the correlations, it corresponds to a vanishing deter-
minant:

$$(8.57) \qquad \det \mathbf{C} = \left| \begin{array}{cc} \mathbf{x}_i^2 & \mathbf{x}_i^T \mathbf{x}_j \\ \mathbf{x}_j^T \mathbf{x}_i & \mathbf{x}_j^2 \end{array} \right| = 0 \quad.$$

From these two vectors only one is needed in the reduced set. Hence, selecting
appropriate modes for a proper orthogonal decomposition corresponds to principal
component analysis in statistical analysis. A single representative for vectors which
are nearly parallel is selected, and a number of directions with a large number of
vectors, in each of these directions is selected. This is done by determining the
largest eigenvalues, and the eigenvectors, of the correlation matrix.

In some cases the mean state:

$$(8.58) \qquad \overline{\mathbf{x}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i \quad,$$

is subtracted from each state $\mathbf{x}_i \rightarrow \mathbf{x}_i - \overline{\mathbf{x}}$, such that rate-of-change snapshots
$(\mathbf{x}(t_i) - \mathbf{x}(t_i + \delta))/\delta$ can be included among the state snapshots, to improve the
representation of the transient dynamics.

In the simple case that there are n blocks, with each n_i identical unit-length
vectors, which are orthogonal to all other vectors, the eigenvalues of the large

matrix arise from each block independently. A simple example is a state that jumps between four different domains of the full state space:

$$(8.59) \qquad (\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \cdots, \mathbf{x}_n) = \begin{pmatrix} 0 & ||||||||| & 0 & 0 \\ 0 & 0 & |||| & 0 \\ |||||| & 0 & 0 & 0 \\ 0 & 0 & 0 & || \end{pmatrix} ,$$

where $|||||$ represent a set of n_i nearly identical column vectors. For a single block i of size n_i the eigenvalues are:

$$(8.60) \qquad \lambda_{i_1} = n_i, \qquad \lambda_{i_j} = 0 \;\; (j = 2, \cdots, n_i) \;\; .$$

Hence a single dominant eigenvector represents a mean direction of the block, where the eigenvalue counts the number of states from the large set appearing in the direction of the eigenvector. In a realistic case the same principle holds. Eigenvalue analysis selects dominant directions, while the smaller eigenvectors describe the variations in the dominant directions, or principal components

Three remarks of caution are at place here. The covariance matrix $\mathbf{C}$ depends on the sampling. If a particular state of the time-dependent FEM experiment is sampled a lot, it will be dominant in the POD, while there might not be a physical reason to include this state. For example, the initial state might be dominant if one starts with sampling from the start. Hence often is started with sampling $\mathbf{x}(t_i)$ after a finite time. Moreover, time-sampling will be dominated by the stationary behavior. Fast transient changes, usually of particular interest, might be swamped by the slow variations, with large numbers of sampling points. This problem may be cured in part by using difference vectors: $\mathbf{x}(t_{i+1}) - \mathbf{x}(t_i)$, rather than sampling points $\mathbf{x}(t_i)$.

Furthermore, the inner-product used $\mathbf{x}_i^T \mathbf{x}_j$ in the covariance matrix might be inappropriate. It might be dominated by the part of the states close to the fixed boundary conditions, or irrelevant physics. If possible, the energy, or even better the change of energy; the power, should lie at the basis of the inner-product used in the covariance matrix $\mathbf{C}$.

8.10. Model dimension

The dimension of a competent reduced-order model depends on the details of the driven dynamics. For example, in many thermal cases the reduced-order model which predicts the result with a percentile accuracy may be two or three-dimensional. On the opposite end of the model spectrum are transmission lines, where the dimension is given by the length L, signal velocity c and the input sampling frequency f. The time the signal spent in a transmission line is $\tau = L/c$. If the sampling frequency is f, the amount of information in the transmission line is $f\tau = fL/c$, hence the dimension of the state space must be:

$$(8.61) \qquad \dim \approx \frac{fL}{c} \;\; .$$

The time-dependence of the input signal is directly related to the wave length of the transmission-line signal. The transmission line is like a queue. The number of samples in the queue depends on the sampling frequency and the transmission time.

However, in many cases the dimension may not be a priori clear. A large FEM model lies at the basis and inputs leads to a certain dynamical range, which determines the state space dimension, or model order. The Krylov method based on the input vector $\mathbf{B}$ and the inverse of the evolution matrix $\mathbf{A}$ may break down due to the lack of orthogonality between successive modes. The lack of orthogonality is a physical manifestation of the fact that the dynamics is restricted to a subspace of the full high-order FEM state space.

The angles between different state vectors can be used to determine the dimension of the subspace. If among three state vectors the angle of one pair equals the sum of the other two pairs, the three vectors lie within the same two-dimensional subspace. The two state vectors with the angle closest to $\frac{\pi}{2}$ should remain while the third state vector can be dropped from the reduced-order state space vectors.

8.11. Effective Hamiltonian

A simple way to reduce the number of variables and the dynamics of a FEM model is the introduction of effective Hamiltonians, also known Guyan projection. From the complete set of variables $\mathbf{x}$ a relevant set of configurations and state representatives are selected through a projection operator $\mathbf{P} = \mathbf{P}^T$. For example, for an elastic structure, these could be the variables at the boundary. The remainder is $1 - \mathbf{P} = \mathbf{Q}$. The elastic energy of the remainder is minimized for arbitrary values for $\mathbf{P}\mathbf{x} = \mathbf{x}_{\mathrm{port}}$. Hence for each value of $\mathbf{x}_{\mathrm{port}}$ the values of $\mathbf{x}_{\mathrm{remainder}}$ are chosen such that the energy $\mathbf{x}^T\mathbf{K}\mathbf{x}/2$ is minimal:

$$(8.62) \qquad \mathbf{x}_{\mathrm{remainder}} = -(\mathbf{Q}\mathbf{K}\mathbf{Q})^{-1}\mathbf{Q}\mathbf{K}\mathbf{P}\mathbf{x}_{\mathrm{port}} \quad .$$

The matrix yielding the remainder from the retained part $\mathbf{x}_{\mathrm{port}}$ is known as the Schur complement. [Golub and Van Loan, 1996] The state is described by $\mathbf{x}_{\mathrm{port}}$, which determine the dimension of the subspace. The state however lies in the full state space: $\mathbf{x} = \mathbf{x}_{\mathrm{port}} + \mathbf{x}_{\mathrm{remainder}}(\mathbf{x}_{\mathrm{port}})$.

In terms of matrices we can order the port or boundary configurations such that the first m rows and columns correspond to the retained variables:

$$(8.63) \qquad \mathbf{K} = \left(\begin{array}{cc} \mathbf{P}\mathbf{K}\mathbf{P} & \mathbf{P}\mathbf{K}\mathbf{Q} \\ \mathbf{Q}\mathbf{K}\mathbf{P} & \mathbf{Q}\mathbf{K}\mathbf{Q} \end{array} \right) \quad .$$

The elastic energy as function of the port configurations only is still a quadratic function:

$$(8.64) \qquad E_{\mathrm{elastic}} = \frac{1}{2}\mathbf{x}_{\mathrm{port}}^T\mathbf{K}\mathbf{x}_{\mathrm{port}} + \frac{1}{2}\mathbf{x}_{\mathrm{port}}\mathbf{K}\mathbf{Q}(\mathbf{Q}\mathbf{K}\mathbf{Q})^{-1}\mathbf{Q}\mathbf{K}\mathbf{x}_{\mathrm{port}} \quad .$$

The corresponding kinetic energy is:

$$(8.65) \quad E_{\mathrm{kinetic}} = \frac{1}{2}\dot{\mathbf{x}}_{\mathrm{port}}^T\mathbf{M}\dot{\mathbf{x}}_{\mathrm{port}} + \frac{1}{2}\dot{\mathbf{x}}_{\mathrm{port}}\mathbf{K}\mathbf{Q}(\mathbf{Q}\mathbf{K}\mathbf{Q})^{-1}\mathbf{M}(\mathbf{Q}\mathbf{K}\mathbf{Q})^{-1}\mathbf{Q}\mathbf{K}\dot{\mathbf{x}}_{\mathrm{port}} \quad .$$

This would correspond to the first Krylov mode. The quality can be tested by the overlap between the kinetic and the elastic mode. If these two modes are nearly parallel, the main part of the dynamics is captured. Otherwise more modes or a larger projection $\mathbf{P}$ may be used to yield a better approximation.

8.12. Plaquette

The plaquette is the equivalent of the stencil in quantum field theories. It is used in conjunction with the Maupertius, [Goldstein, 1980] or least action principle, where the fields are integrated over a space-time domain, treating space and time on equal footing. This makes the method especially suited for relativistic theories, but also for time-dependent studies it has the advantage to incorporate the time-integration into the finite-element method. [Creutz, 1983] The energy continuity can be retained in the approximation.

The plaquette and link variables define quantum field theories by numerical means. Most of these approaches are called Euclidean, which means that space and time are treated on equal footing. Even more, the time axis is rotated in the complex plane, which is called a Wick rotation, such that a wave equation in three dimensions yields a Laplacian in four dimensions. Stable solutions of one can be related to the other.

We will discuss the plaquette variable from the Hamiltonian perspective in a single space dimension, without the Wick rotation. However, unlike the true Hamiltonian approach, time is discretized here as well. The energy is given by the Hamiltonian H, together with the corresponding energy flux S it yields a divergence-free vector field in space-time $x = (t, z)$, the two-dimensional version of the energy-continuity relation, Eq. 5.39:

$$(8.66) \qquad\qquad \partial_t H + \partial_z S = 0 \ ,$$

where the space-time vector field on (t, z) is given by:

$$(8.67) \qquad\qquad \mathbf{J} = \begin{pmatrix} H \\ S \end{pmatrix} = \begin{pmatrix} T^{00} \\ T^{01} \end{pmatrix} \ .$$

In terms of the energy-momentum tensor T, the Hamiltonian and the flux are the T^{00} and the T^{01} components respectively, if $t = x^0$ and $z = x^1$. [1]

The divergence theorem relates vector field to its integral over a plaquette $z \in [0, 1]$, $t \in [0, 1]$, such that the values at the boundary satisfy the conservation law:

$$(8.68) \qquad \int dx \nabla \cdot \mathbf{J} = \int dt S(t, 1) - S(t, 0) + \int dz H(1, z) - H(0, z) \ .$$

The integrals over each of the edges are called link variables. The single plaquette is made up of four link variables.

Given a linear approximation for $\mathbf{J}$, the four corners determine the complete field, and given three corners, the fourth corner is determined, similar to the stencils in Section 8.4. In more general approximations the duality between four corners and four links still means a single condition arises from the energy conservation.

The vector field $\mathbf{J}$ is usually derived from the equations of motion, or their generating functional, such as the Hamiltonian or the Lagrangian, for example via the energy-momentum tensor:

$$(8.69) \qquad\qquad T^{ij}(q) = -\frac{\delta \mathcal{L}}{\delta \partial_i q} g^{jk} \partial_k q + g^{ij} \mathcal{L} \ ,$$

---

[1]Some details concerning signs and the Minskowski metric are not treated here, however, different conventions will lead to different signs for both the fluxes and the derivatives. Here the variables are treated as Euclidean. The sign convention for the fluxes is considered Euclidean as well, which means the energy flow in space and time have the same sign.

where $g^{ji} = \left(\begin{smallmatrix} 1 & 0 \\ 0 & -1 \end{smallmatrix}\right)$ is the metric tensor, and $\mathcal{L}$ the Lagrangian. The description is not unique, $T^{ij} \neq T^{ji}$. Hence, one energy flux might differ from another by a total divergence. The single requirement of energy conservation is that it can be written as a rotation field, i.e., an exact 1-form:

$$(8.70) \qquad \mathbf{J} = \left(\begin{array}{c} \partial_z W \\ -\partial_t W \end{array} \right) \quad .$$

However, starting from a general Hamiltonian, the link variables arise naturally during discretization. The variable $q(z,t)$ appears in the Hamiltonian as $\dot{q}$ and $\partial_z q$. These two variables correspond to the link variables on a time-link and space-link respectively:

$$(8.71) \qquad q_z(t) = \int dz \partial_z q(z,t) = q(1,t) - q(0,t) \quad .$$

The links are constrained by the integrability condition: $q_z(1) - q_z(0) + q_t(0) - q_t(1) = 0$.

Maupertius principle, or the Lagrangian variational principle, is the most appropriate form to use the link variables:

$$(8.72) \qquad \delta \int_0^1 dt dz \mathcal{L}(\partial_z q, \dot{q}) \to \delta \int_0^1 dt dz \mathcal{L}_p(q_z(t), q_t(z)) = 0 \quad .$$

The wave equation, for example, reduces to an algebraic variational equation, with the Lagrangian:

$$(8.73) \qquad \mathcal{L}(q_z(t), q_t(z)) = q_t^2 - v^{-2} q_z^2 \quad .$$

Under the assumption that the link variables are linear functions of the remaining degree of freedom, the quadratic Lagrangian is easily integrated over each plaquette:

$$(8.74) \qquad \int_0^1 dz (q_t(0) + (q_t(1) - q_t(0))z)^2 = \frac{1}{3}(q_t(0)^2 + q_t(1)q_t(0) + q_t(1)^2) \quad .$$

The simulation can be done by filling the squares. The complete system for a line consists of a chain of plaquettes. A single boundary time-link is required for a unique solution to the hyperbolic variational problem. The constraints can be used to re-express all the other time-links in terms of the space-links for $t = 0$ and $t = 1$:

$$(8.75) \qquad q_t(1) = q_z(1) - q_z(0) + q_t(0) \quad .$$

The right time-link is determined by the left time-link and the space-links. For the neighboring plaquette, the right time-link is the already determined left time-link of the original plaquette. The time-links of the chain are determined recursively.

The variational problem allows one to reduce the degrees of freedom by half: the space-links for $t = 1$ can be expressed in terms of the space-links for $t = 0$, which is precisely the condition for a unique evolution.
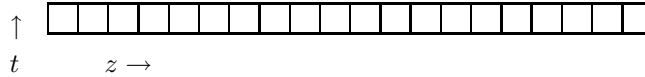
Clearly, the same procedures can be applied to non-linear systems, the adaption lies in the single integrals:

$$(8.76) \qquad \int_0^1 dz \mathcal{T}(q_t(0) + (q_t(1) - q_t(0))z)^2 = T(q_t(0), q_t(1)) \quad ,$$

and, the potential term:

$$(8.77) \qquad \int_0^1 dt \mathcal{V}(q_z(0) + (q_z(1) - q_z(0))t)^2 = V(q_z(0), q_z(1)) \quad .$$

8.13. Simplex

In many modern FEM applications, the underlying mesh consists of triangles. This is mainly the field of mathematicians, who appreciate the simplicity of the simplex, and the field of software engineers, who base their methods on CAD/CAM methods, which allow for a restricted set of smooth curves, which are more easily meshed, or discretized, using triangulations than many other methods. There are several ambitious research programs formulated to make triangular meshes the industry standard, but, probably due to complications at the lowest levels, it has never materialized. The complications for general modelling, with a direct access and control over the mesh and the data, are large and various. Some of the details of mesh refinements, described below, in three dimensions will make this clear.

In two dimensions, the mesh is clear, and some global properties, such a topological properties, are easily determined. For example, the Euler characteristic, χ_E, is an invariant of a closed surface, which is related to the number of genuses, or holes, or handles on a surface. For example a sphere has no handles, a donut one handle, and a solid eight has two handles. The Euler characteristic χ_E is given by the enumeration of all the elements of a surface:

$$(8.78) \qquad \chi_E = \text{vertices} - \text{edges} + \text{faces} = 2 - \text{handles} \quad .$$

For three-dimensional elements such relations do not exist. A characteristic is defined:

$$(8.79) \qquad \chi_3 = \text{vertices} - \text{edges} + \text{faces} - \text{volumes} \quad .$$

But it seems to have no useful topological meaning. One of the problems is the number of ways to define a mesh, or mesh refinement, for a three-dimensional simplex, such as the tetrahedron.

Starting from a cube, there is one easy division in identical simplices, based on the symmetry principle. One can define one simplex by the corners: (0,0,0)-(1,0,0)-(1,1,0)-(1,1,1) By reflections, all other simplices follow, yielding a filling with 6 simplices. One edge is the body diagonal, two edges are face diagonals, and three edges are cube edges. [Maubach, 1994]

Given the positions of the four corners of the simplex $\mathbf{x}_a, \mathbf{x}_b, \mathbf{x}_c,$ and $\mathbf{x}_d$, the geometric properties follow simply:

$$(8.80) \qquad \text{Area}_{\text{face}:abc} = \frac{1}{2} |\mathbf{x}_b \times \mathbf{x}_c + \mathbf{x}_a \times \mathbf{x}_b + \mathbf{x}_c \times \mathbf{x}_a| \quad ,$$

where the direction of the vector is the direction of the surface normal,
(8.81)
$$\text{Volume}_{\text{simplex}:abcd} = |\det(\mathbf{x}_a - \mathbf{x}_b, \mathbf{x}_b - \mathbf{x}_c, \mathbf{x}_c - \mathbf{x}_d)| = |(\mathbf{x}_a - \mathbf{x}_b) \cdot ((\mathbf{x}_b - \mathbf{x}_c) \times (\mathbf{x}_c - \mathbf{x}_d))| \quad .$$

The position of the center is:

$$(8.82) \qquad \mathbf{x}_0 = \frac{\mathbf{x}_a + \mathbf{x}_b + \mathbf{x}_c + \mathbf{x}_d}{4} \quad .$$
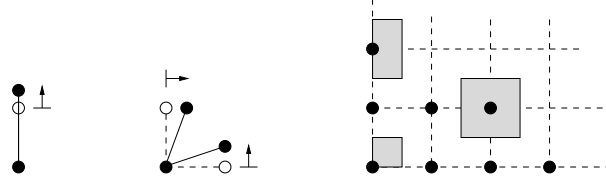
FIGURE 8.4. From left to right, the elongations ε_i, defined using two nodes, the shear γ_{xy} using three nodes, and the masses corresponding the integrals over the patches for center nodes, edge nodes, and corner nodes.

8.14. Transverse motion of a table

As a practical example for comparison we investigate the longitudinal motion and vibration of a table. In several industrial setups large tables are used to position a flat object. Fast motion will cause the table to deform and possibly vibrate. We assume the table to be made of standard steel, with a Young's modulus of $E = 20.0 \ 10^9 \mathrm{Nm}^{-2}$, a Poisson ratio of $\nu = 0.30$, and a specific density of $\rho = 8.0 \ 10^3 \mathrm{kgm}^{-3}$. The table is taken to be $2. \times 2.$ meters. For relative numbers in the linear theory, only the Poisson ratio ν is relevant; the scale can be set arbitrarily.

The three-dimensional elastic energy W of the table is expressed [Washizu, 1968, Landau and Lifschitz, 1959] in its elongations ε_x, ε_y, and ε_z, and the shears γ_{xy}, γ_{xz}, and γ_{yz}:

$$(8.83) \qquad W = \frac{1}{2} \int \left(\begin{pmatrix} \varepsilon_x \\ \varepsilon_y \\ \varepsilon_z \end{pmatrix}^T \begin{pmatrix} E_\| & E_\perp & E_\perp \\ E_\perp & E_\| & E_\perp \\ E_\perp & E_\perp & E_\| \end{pmatrix} \begin{pmatrix} \varepsilon_x \\ \varepsilon_y \\ \varepsilon_z \end{pmatrix} + \quad . \right.$$

$$(8.84) \qquad \left. \begin{pmatrix} \gamma_{xy} \\ \gamma_{xz} \\ \gamma_{yz} \end{pmatrix}^T \begin{pmatrix} G & & \\ & G & \\ & & G \end{pmatrix} \begin{pmatrix} \gamma_{xy} \\ \gamma_{xz} \\ \gamma_{yz} \end{pmatrix} \right) dV \ ,$$

where $E_\| = (1-\nu)E/((1-2\nu)(1+\nu))$, $E_\perp = \nu E/((1-2\nu)(1+\nu))$, and $G = E/(2(1+\nu))$. The elongations and shears can be expressed in terms of the displacements u, v, and w:

$$(8.85) \qquad \varepsilon_x = \frac{\partial u}{\partial x}, \qquad \gamma_{xy} = \frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}, \qquad \text{etc.}$$

The two planar directions x and y are coupled, and the motion in the z direction is free, with the associated inertia neglected:

$$(8.86) \qquad \frac{\delta W}{\delta \varepsilon_z} = \frac{\delta W}{\delta \gamma_{xz}} = \frac{\delta W}{\delta \gamma_{yz}} = 0 \ ,$$

such that the remaining two-dimensional elastic energy is:

$$(8.87) \qquad W_{xy} = \int \left(\frac{E}{2(1-\nu^2)} \begin{pmatrix} \varepsilon_x \\ \varepsilon_y \end{pmatrix}^T \begin{pmatrix} 1 & \nu \\ \nu & 1 \end{pmatrix} \begin{pmatrix} \varepsilon_x \\ \varepsilon_y \end{pmatrix} + \frac{G}{2}\gamma_{xy}^2 \right) dA \ ,$$

which correspond to the energy per unit thickness of the table. See also Section 6.7.
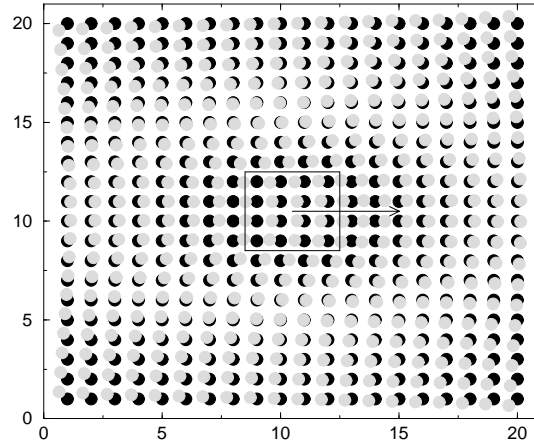
FIGURE 8.5. The static deformation, the black dots are the rest positions of the grid points, and the gray dots are the displaced positions. The scale is set for a good visual effect, the fixture is in the center, indicated by the box.

The table has now three rigid-body modes in terms of the independent displacements u and v, which are the two translations in the x and y direction, and the rotation around an arbitrary axis in the z direction.

Each pair of nearest neighbors is used the calculate the elongations ε_i. The shears require a set of three points: a corner with a point in the x-direction and the y-direction. For the displacement a linear approximation between those points is used. [Hughes, 1987] The mass matrix is taken diagonal, where edge have half and corners have a quarter of the mass of a node away from the edge. See Figure 8.4. The FEM model of the table consists of 20 points in each direction, which yields $2 \times 20 \times 20 = 800$ displacement variables, and $19 \times 19 = 361$ cells. The elongations for the diagonal part of the energy are the $2 \times 20 \times 19 = 760$ elongations, one for each nearest-neighbor pair of grid points. The off-diagonal energy uses the elongations per cell, in order to be able to take the product of x and y elongations. The shear γ_{xy} consists of $4 \times 19 \times 19 = 1449$ terms, one for each corner of the cell. The shear energy consists over the sum over cells with the average over the corner products. If the average shear γ_{xy} would be calculated per cell first before taking the product, it would yield additional zero modes, for $\nu = 0$. The thickness of the table is irrelevant for the longitudinal motion.

Unlike the wave equation for the longitudinal motions in a table, the x and y motion are coupled in elastic theory. The eigenmodes have, therefore, fewer degeneracies, but there are still some, associated with the modes with four-fold symmetries. Furthermore, the shapes of the eigenmodes have more displacements at the boundary, compared to the wave eigenmodes which is the trigonometric series, due to more complex elastic energy. The eigenmodes are the result of the diagonalization of the stiffness matrix:

(8.88) $$\omega_i^2 \mathbf{V}_i = K \mathbf{V}_i \ \ .$$

TABLE 8.1. The lowest eigenfrequencies ω_i, with the corresponding degeneracies. The important modes are indicated with an asterisk (*).

mode	frequency (kHz)	deg.
rigid	0.0	3
1*	1.6965	2
2	1.7989	1
3	1.9000	1
4*	2.0781	1
5*	2.6089	1
6	2.7307	1
7*	2.8066	2
8	3.3069	2
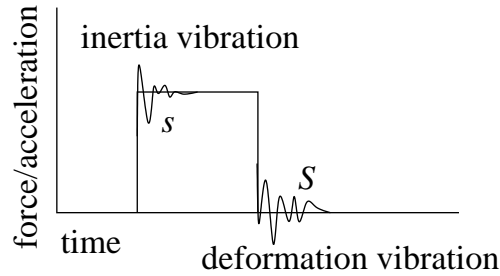
FIGURE 8.6. The two types of spectroscopic factors.

In the case of a degeneracy $\omega_{i_1} = \omega_{i_2} \cdots = \omega_{i_j}$, the actual eigenvectors are irrelevant, only the space spanned by these eigenvectors is relevant:

$$(8.89) \qquad \Omega_{\{i_1,\cdots,i_j\}} = \mathrm{span}\{\mathbf{V}_{i_1},\cdots,\mathbf{V}_{i_j}\} \quad .$$

The results are given in Table 8.1.

The static-deformation mode $\mathbf{x}_0$ is determined from solving the linear equation:

$$(8.90) \qquad K\mathbf{x}_0 = \mathbf{F} - M\mathbf{a} \quad ,$$

where $\mathbf{x}$ is taken perpendicular to the rigid-body modes. The iterative solver SYMMLQ by Paige and Saunders [Paige and Saunders, 1975] can handle indefinite K matrices, however, care must be taken that the right-hand side is perpendicular to the null-space of K. Whether a vibrational mode is excited by the applied force can be determine from the spectroscopic factors S_i, which are the overlap between the normalized eigenvectors and the static deformation mode:

$$(8.91) \qquad S_i = \|\mathbf{x}_0\|_{\Omega_i} \quad ,$$

which is the length of $\mathbf{x}_0$ in the subspace of the eigenvectors of the i-th mode, which reduces to

$$(8.92) \qquad S_i = \sqrt{\mathbf{x}_0^T \cdot \mathbf{V}_i} \quad ,$$

for a non-degenerate eigenmode, where $\|\mathbf{V}_i\| = 1$.

Leading spectroscopic factors
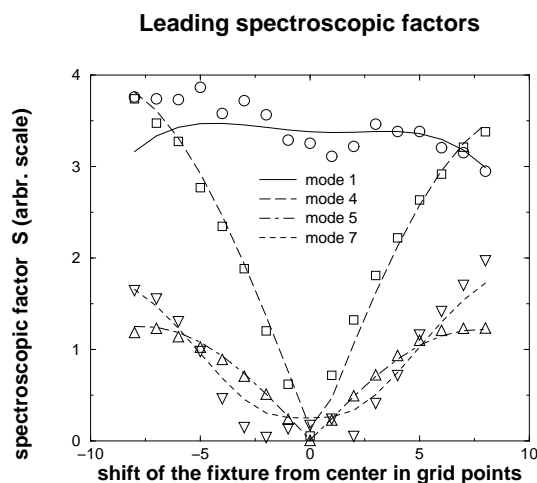
FIGURE 8.7. The spectroscopic factors S_i (lines), for the relaxation of deformation, of the four leading modes. The corresponding calculations of $S_i = \omega_i^2 s_i$ are indicated by the symbols.

A second type of spectroscopic constant is associated with the inertia of the system:

$$(8.93) \qquad s_i = \|\mathbf{F}\|_{\Omega_i} \quad .$$

If we would measure the actual displacement at a certain point if a block-wave force is applied, the oscillations in the rising edge would correspond to the vibrations with the spectroscopic amplitudes s_i, while the vibrations in the falling edge would correspond to the spectroscopic amplitudes S_i. (See Figure 8.6.) In the rising edge the inertia causes the vibrations:

$$(8.94) \qquad m\ddot{\mathbf{x}}(t) + K\mathbf{x}(t) = \mathbf{F} \quad ,$$

with $\mathbf{x}(0) = 0$, while in the falling edge it is the relaxation of the deformation:

$$(8.95) \qquad m\ddot{\mathbf{x}}(t) + K\mathbf{x}(t) = 0 \quad ,$$

with $\mathbf{x}(0) = \mathbf{x}_0$. The initial state is given by: $K\mathbf{x}_0 = \mathbf{F} - M\mathbf{a}$. Hence it follows that:

$$(8.96) \qquad s_i = \omega_i^2 S_i \quad .$$

In practice it is easier to determine S_i than s_i since $\mathbf{x}_0$ is much smoother than $\mathbf{F}$. (See Figure 8.7.) In the case of a linear model, the response to the rising edge and the falling edge is the same. In practice, both numerical as nonlinear effects cause differences between s_i and S_i.

If the fixture is attached to the middle, the coupling between motion and vibration is mainly to mode 1, which has a twofold degeneracy. The symbols represent the spectroscopic factors $\omega_i^{-2}s_i$, which quality is much poorer, due to the sharp behavior of $\mathbf{F}$, leading to numerical instabilities in the overlap. To determine the eigensystem, the eigenvalues and eigenstates, the generic LAPACK [Anderson et al., 1999] routine DSYEV was used.

We attach a fixture to the center of the table in the y-direction, to a square of $4 \times 4 = 16$ grid points. The force is applied in the x direction, and the square region is shifted along the x-axis. The spectroscopic factors depend on the position of the fixture, and the leading factors are plotted (see Figure 8.7). However, the spectroscopic factors do not diminish for higher modes, on average, they stay at about 5% of the leading factor, throughout the whole spectrum. This is the result the sharp edges of the fixture. Using only the lowest vibrational modes will only be sufficient if the high modes damp out quickly, or if they are not excited due to the slow changes of the applied force.

Another way (See Figure 8.8) to test the quality of the position of the fixture, is to check the displacement, both the maximal and the mean values. The maximal value is 3 to 5 times the mean value, and again the center position of the fixture has the smallest values.

8.15. Simulation of dynamics

From the analysis of the spectroscopic factors it is clear that the lowest vibrational mode dominates the dynamics, if the fixture is placed in the center. We compare the single mode result of the degenerate, vibrational mode, with the static-deformation mode, and with the full theory. The input signal is a simple force bump, with a time-scale which is comparable to the eigenfrequency of the table:

$$(8.97) \qquad \mathbf{F}(t) = \mathbf{F}_0 t^2 (T - t)^2; \quad 0 < t < T \ ,$$

where $T \approx 3.3$ periods.

We use a simple Euler stepping algorithm, Section 3.4, with step size δ:

$$(8.98) \qquad \begin{pmatrix} x_{t+\delta} \\ v_{t+\delta} \end{pmatrix} = \begin{pmatrix} x_t \\ v_t \end{pmatrix} + \delta \begin{pmatrix} 0 & 1 \\ -K & 0 \end{pmatrix} \begin{pmatrix} x_t \\ v_t \end{pmatrix} + \delta \begin{pmatrix} 0 \\ \mathbf{F}_t \end{pmatrix} \ ,$$

where we have set the mass $m = 1$, yielding a unit mass matrix. No damping is introduced. As output we follow a corner of the table, and its deviation from the motion it would make if the table was a rigid body. The situation should be in the clear advantage for vibrational modes, often the time-scale of an applied force is ten times or more the typical period, rather than the 3.3 periods used here. Furthermore, the absence of other low-lying vibrational modes means the dynamics is dominated by this single mode. The design has already been optimized to reduced vibrations.

The one-dimensional vibration reduced model:

$$(8.99) \qquad \ddot{x}_v + \omega_1^2 x_v = \|\mathbf{F}(t)\|_{\Omega_1} \ ,$$

with the output:

$$(8.100) \qquad y = \frac{\mathbf{C}^T \mathbf{V}_{1_1} \mathbf{V}_{1_1}^T \mathbf{F} + \mathbf{C}^T \mathbf{V}_{1_2} \mathbf{V}_{1_2}^T \mathbf{F}}{\|\mathbf{F}\|_{\Omega_1}} x_v \ ,$$

where $\mathbf{V}_{1_1}$ and $\mathbf{V}_{1_2}$ are two orthonormal vectors in the first eigenmode.

Instead we can also use the static deformation under a constant acceleration as a mode to study the dynamics. The static-deformation mode leads to a similar reduced model:

$$(8.101) \qquad \ddot{x}_m + \frac{\mathbf{x}_0^T \mathbf{F}_0}{\mathbf{x}_0^2} x_m = \frac{\mathbf{x}_0^T \mathbf{F}(t)}{\|\mathbf{x}_0\|} \ ,$$
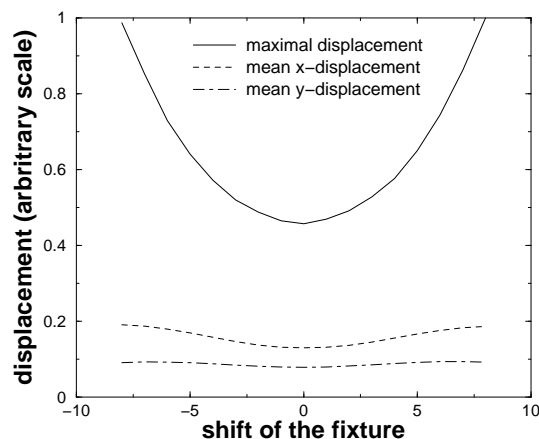
FIGURE 8.8. The maximal displacement, and the mean displacements in x and y direction of the table for the static-deformation, depending on the position of the fixture.

where $K\mathbf{x}_0 = \mathbf{F} - M\mathbf{a}$, and with the output:

$$(8.102) \qquad\qquad y = \frac{\mathbf{C}^T \mathbf{x}_0}{\|\mathbf{x}_0\|} x_m \quad.$$

The computational simplicity of the static-deformation approach, which requires only the solution of a linear equation, is clear. For the vibrational mode, the large eigenvalue equation had to be solved. The two-mode approximation is a straight-forward implementation of the general results. The second mode was projected perpendicular to the first mode to retain a unit mass matrix.

The results, in Figure 8.9, show the deviation of the lowest vibrational mode, which is in practice a two-dimensional approximation, due to the degeneracy. The x-motion is underestimated. The remaining vibration is tracked accurately. The static-deformation mode alone gives accurate results, except for the remaining vibration, after the force pulse. The two-dimensional input-based model yield the most accurate results, both for the motion and the vibration. In the final vibration there remains a small deviation. Including higher-order modes will not increase the accuracy necessarily, because the higher modes have very large overlap with the lower modes, and they will generate thereby numerical instabilities. Note, further-more, in the absence of dynamics; if the mass of the static-deformation mode is neglected, the shape of the force pulse already predicts 90% of the motion of the table, the vibrational effects causes oscillations around that result.

8.16. Conclusion

The method of using the deformations of the table under a constant acceleration as modes for a lumped dynamical model are further discussed in the Chapter 9. The study of finite elements as invariant components, or port Hamiltonian elements, in a network for the use in dynamics automatically leads to a more general, and often nonlinear, formulation. This question is studied in Chapter 11.
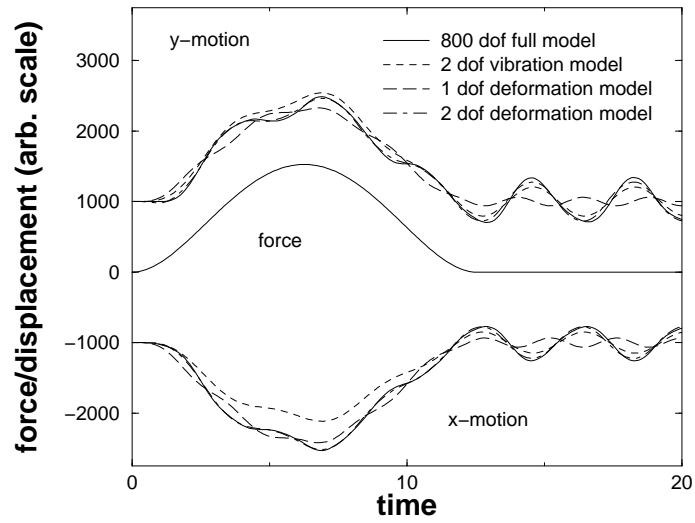
FIGURE 8.9. The deviation of the rigid-body motion of a corner of the table. The two-dimensional deformation model yields the best results; its short-long-dashed line is almost indistinguishable from the full results.

This chapter dealt with the more traditional approaches to dynamical FEM. The specialized methods developed for fluid dynamics are not mentioned explicitly, although being a large field of research. The connection between energy-based port-Hamiltonian methods and general fluid dynamics is tenuous at best.

CHAPTER 9

Analytic input

The displacement of a string at a certain moment betrays some history of the input. If the string is straight and at rest, there has not been any input for at least as long as it takes a wave to travel from one end to another. Furthermore, signals that traveled through the string before that time must have been canceled by appropriate boundary conditions, at the opposite end as from where they emerged.

Hence, to some extent, for general continuous systems, it must be possible to map the state of the system to the history of the input. This mapping is limited; different histories may lead to the same state. Furthermore, the reverse may also be true; the system could be described in terms of input variables. Rather than to give a formal answer to this mapping problem, we seek a practical answer. In many practical cases the detail of the input is limited to a discrete set. In our case we use an analytic input, polynomial functions of time, as a discrete set.

State control has been about finding the state variables that described the system uniquely. In the case of distributed or infinite systems, this is an impossible task. The states of actual interest are those excited by the input and operational circumstances. The input dictates the relevant state variables. Analytic input space is a particular generic input for finding a corresponding state space. As seen before, the stationary state, which nature depends on the type of input, is an important state which characterizes the dominant quality, such as resistive, inertial, or capacitance, in that setting. The stationary state has a natural place in the analytic input based formulation, as the lowest order state. Instead of analytic input, one could choose impulse response as basis for mapping out the state space. It would have three drawbacks. First, it is dominated by the higher frequencies, which is not typically the operational state of interest. It uses intrinsically the linear approximation of the system, which may be a false assumption. Thirdly, the dimension of the state space would in many cases be too high for a useful characterization of the system in a small set of states.

For systems with an energy description, the magnitude of the states is bounded by the integral of all the power supplied to the system. In a behavioral, finite dimensional setting this has been formalized in the theory of dissipative systems. [Willems, 1972a,b]

The use of real-time computing in control made elaborate smooth time-dependent input possible. Usually such input is represented by splines, which are piece-wise polynomial functions with arbitrary smooth boundaries between the pieces. The spectral, or frequency, characteristics of such an input is not necessarily obvious. However, it is clear that it is dominated by the low frequencies. The method based on analytic input focuses therefore on the achievable control.

In this chapter the common results of control theory are established for analytic, and piece-wise analytic input. Furthermore, the focus is shifted from the state to the input. The general benefits of such a shift should be obvious, since the little one knows about the state of a system, the more one knows about the input. In many approaches a step input yields a state that converges to a constant state after a finite time. Although the stationary state does not occur in a finite time window, they are as limiting state relevant for the dynamics. The approach advocated in this chapter associates a constant state with the constant input, and deviations from constancy for input and state simultaneous. Such a framework naturally arrives at analytic theory and polynomial expansion.

Finally, an input-centered approach is the right starting point to extend methods to nonlinear systems, where different state-space representations may be equivalent but not equally practical, and the finite-dimensional input-centered approach distinguishes the relevant subspaces, thereby reducing the complexity to the minimum.

Some subtle differences between the standard control formulation and the analytic input control will arise. These differences are due to restrictions on the input. However, these restrictions are quite natural; infinite slopes do not occur in most physical system. Most singularities in the state arise from, for example, the neglect of inertial forces. The focusing of waves, however, in nonlinear waves or position-dependent dispersion may cause a singular shock waves, or caustic waves. [Courant and Hilbert, 1961]

9.1. Stationary states

The stationary state that arises from constant input is in most cases the most relevant nontrivial state. For a constant current or voltage, that leads to a constant voltage or current as state, it corresponds to the state which determines the Ohmic resistance. In the process of cooling, it is the analogous state of heat conductance, or resistance. Likewise, for a constant force, it is an experimental verification of the inertia, or total mass of the system, for a state with constant acceleration. A state with constant acceleration is only a stationary state in a broad sense, also referred to as a quasi-stationary state. Hence the stationary state for a given input can be seen as the constitutive relation for the global or lumped response of the system; resistance for flow, inertia or capacity for force. Furthermore, the stationary state often gives a lower or upper bound for the dynamic response. For example, the stationary flow is the maximal flow for a particular pressure or potential. From rest, before that flow establishes, the flow is typically smaller because of the buffering effects, and it reaches the stationary flow from below, in the case of resistance, or more general, overcritically damped systems.

Altogether, a careful analysis of the stationary states, like in the linear case, Eq. 3.4, for the particular input and output we choose, is the basis for a lumped or modal view of distributed systems. A careful energy and power analysis determines the appropriate matching of the lumped variables with modal shapes and mean values. For example the total heat W in a system is the product of temperature distribution $T(z)$ as function of the position, and the capacity distribution, $C(z)$, also as function of the position. The total heat in the system is:

$$(9.1) \qquad\qquad W = \int dz T(z) C(z) \ .$$
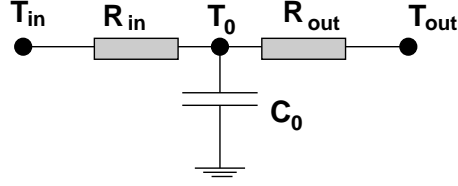
FIGURE 9.1. The simplest network associated with a stationary flow J and a distributed temperature $T(z)$ and capacity. The mean capacity C_0 should be placed such that it gives the correct mean temperature T_0.

The total capacity is given by the integral:

$$(9.2) \qquad C_0 = \int dz C(z) \ .$$

The mean temperature T_0 for the total capacity C_0 with the same total heat W follows from the definition:

$$(9.3) \qquad W = T_0 C_0 = \left(\int dz T(z) \frac{C(z)}{C_0} \right) \left(\int dz C(z) \right) \ .$$

The mean temperature weighted with the relative capacity $C(z)/C_0$, and the total capacity is the integral over the distributed capacity, such that if the temperature is constant, the mean temperature is the constant temperature.

In the case of nonlinear constitutive relations, e.g., $C(T, z)$, the product of the mean values is no longer equal to the work W. With the modal approximation of $T(z)$ also a modal approximation of $W(T_0)$ is required for nonlinear systems. The stationary state is a good starting point for such approximations. Even though the original dynamics maybe highly nonlinear, small temporal variations around the stationary state might be well-approximated by a linear model. For example, consider a system which radiates of its heat. The heat transfer depends on the fourth order of the temperature difference, but around a particular operational, stationary state, with a temperature difference $T_{in} - T_{out}$, the variations in $T_{in} + \Delta T$ will lead to variations ΔJ in the heat transfer J:

$$(9.4) \qquad \Delta J \sim (T_{in} + \Delta T - T_{out})^4 - (T_{in} - T_{out})^4 \approx 4(T_{in} - T_{out})^3 \Delta T \ ,$$

where T_{in} and T_{out} are given by the stationary state $T(z)$.

A simple network, Fig. 9.1, given the mean temperature T_0 and the flow J, and input and output temperatures T_{in} and T_{out} would yield a heat resistance $R = (T_{in} - T_{out})/J$. In a lumped network the total capacitance C_0 should be placed at the position where the mean temperature occurs, in between the incoming resistance R_{in} and the outgoing resistance R_{out}:

$$(9.5) \qquad R_{in} = R \frac{T_{in} - T_0}{T_{in} - T_{out}}, \quad R_{out} = R \frac{T_0 - T_{out}}{T_{in} - T_{out}}$$

Hence, the heat flow and heat conservation dictates the transition from distributed to lumped, depending on the situation at hand. The stationary flow is an appropriate situation, however, many other choices can be made. For example, instead of the mean temperature T_0, one could choose the temperature at a sensor T_{sensor}. The sensor temperature T_{sensor} together with the total capacitance C_0 will not yield

the total heat in the system $W \neq T_{\text{sensor}} C_0$. The appropriate manner to do so is the split the resistance in which the sensor resides R_{in} if $T(z_{\text{sensor}}) > T_0$ or R_{out} if $T(z_{\text{sensor}}) < T_0$, and rely on the energy-conserving network, but often another capacitance C_{sensor} is chosen to match T_{sensor}, i.e., $W = T_{\text{sensor}} C_{\text{sensor}}$. Sensor output dictates many input-output networks, where component consistency is lost. In principle, for such simple networks, it makes no difference, but once it is placed in a greater network, and possibly extended with other properties such as inductances, these choices matter.

Typical input and output temperatures, like T_{out} and T_{in}, are not the mean, or lumped, temperature T_0, used to determine the total heat W in the system. If they are localized at, for example, z_{out} and z_{in}, they follow from the modal shape: $T(z_i)/T_0$. However, in many cases the temperature at the port itself is not constant, but the temperature $T(z)$ over an area of the boundary $z \in \partial\Omega$.

Again the stationary state will yield a consistent definition of the heat flow in and out of the system. The heat flow $j(z)$ through the boundary sums to the total heat flow in and out of the system:

$$(9.6) \qquad\qquad J = \int_{\partial\Omega} dz\, j(z) \ .$$

The particular entropy generation in the system is associated with the temperature difference between input and output:

$$(9.7) \quad S = \int_{\partial\Omega} dz\, (T(z))^{-1} j(z) = \left(\int_{\partial\Omega} dz\, (T(z))^{-1} \frac{j(z)}{J} \right) \left(\int_{\partial\Omega} dz\, j(z) \right) = T^{-1} J \ .$$

In this case, like with $C(z)/C_0$ in Eq. 9.3, the flow provides a measure $j(z)/J$, for a consistent mean temperature at the boundary:

$$(9.8) \qquad\qquad \int_{\Omega} dz\, \frac{C(z)}{C_0} = \int_{\partial\Omega} dz\, \frac{j(z)}{J} = 1 \ .$$

The stationary state of the system gives us appropriate lumped parameters and variables, which are consistent with the conserved quantities in the system, such as energy and entropy. The state at rest, i.e., $T = $ constant, would not have been able to provide this linkage to lumped models.

Furthermore, the stationary state can be of a nonlinear model, it changes very little in the definition of total and mean variables. The product, such as the power product of effort and flow, in the integrand remains a product by definition. Non-linear relations are constitutive relations, or closure relations, [Bird et al., 1960] which link efforts and flows to the state variables.

To extend this notion of stationary states to lumped model beyond the simplest approximation, described above, we view the system as the response to an analytical input function of time. The stationary state is the response to a constant input, at the next orders linear $\mathbf{u}(t) = \mathbf{a} + \mathbf{b}t$ and quadratic input $\mathbf{u}(t) = \mathbf{a} + \mathbf{b}t + \mathbf{c}t^2$ are considered. These functions may seem unnatural for large times $t \to \pm\infty$, however, the modal shapes, e.g., $\{t_i(z)\}$ for a response $T(z,t) = t_0(z) + t_1(z)t + t_2(z)t^2 + \cdots$, can be used to define a fast converging basis for time-variations of the input, which are typically of the same order or slower than the longest time-scales of the response.

9.2. Serial network model through tridiagonalization

Before considering time-dependence of a system as a response to an input, we will examine the linkage between Lanczos tridiagonalization, described in Section 2.12, serial networks, and extension of the simple network of the previous section. The intermediate step between a distributed system, depending on a continuous position z, and a lumped model is often a large FEM model, which is subsequently reduced to a lumped model with a small number of degrees of freedom. [Davies, 1978] This last process is called model-order reduction and has received a lot of attention in recent years. [Freund, 2003] The link between the reduced-order, or lumped model, and the distributed model is tentative at best. Tridiagonalized models, however, can be viewed as lumped, network models. In the matter of truncating, or reducing, the tridiagonal model, the network representation helps to retain charge conservation, from the continuous model to the lumped model.

Complete diagonalization is an iterative procedure which breaks to model into isolated submodels, each corresponding to an eigenmode. However, the procedure of diagonalization does not easily allow for the inclusion of damping or multiple inputs and outputs. The algorithm for tridiagonalization is, on the other hand, well-defined in terms of a finite number of steps. It also allows one to make a direct link a network model.

Some details, such as conservation laws of the distributed system, are retained in the proper network model, but are easily lost, if not the proper care is taken, to consider the network model under reduction. [Freund, 2003] In this section we investigate two cases of network models, which may correspond to a symmetric matrix $\mathbf{A}$.

A general symmetric matrix $\tilde{\mathbf{A}}$ can be brought to a tridiagonal form:

$$(9.9) \qquad \mathbf{A} = \begin{pmatrix} \alpha_1 & \beta_1 & & & & \\ \beta_1 & \alpha_2 & \beta_2 & & & \\ & \beta_2 & \alpha_3 & \beta_3 & & \\ & & \beta_3 & \alpha_4 & \beta_4 & \\ & & & \beta_4 & \alpha_5 & \beta_5 \\ & & & & \ddots & \ddots \end{pmatrix} \, .$$

Such a matrix can be mapped onto a serial network with capacitances and resistances; a RC-network, or capacitances and inductances; a LC-network, depending on the differential equation:

$$(9.10) \qquad \dot{\mathbf{x}} = \mathbf{A}_{RC}\mathbf{x}, \qquad \ddot{\mathbf{x}} = \mathbf{A}_{LC}\mathbf{x} \ ,$$

where x_i is a voltage. The resistors R_i and inductors L_i are in between the voltage points x_i and x_{i+1}, connected to the ground via a capacitor C_i. See Figure 9.2. The consequences of conservation laws, in this case charge conservation, are not a priori clear from the coefficients α_i and β_i. In the RC-network and the LC-network, these conditions are related to the capacitances C_i, as we will see.

In the case of a RC-network the variables are:

$$(9.11) \qquad \mathbf{A}_{RC} = \begin{pmatrix} \frac{-1}{R_1 C_1} & \frac{1}{R_1 C_{12}} & & & \\ \frac{1}{R_1 C_{12}} & \frac{-1}{R_{12} C_2} & \frac{1}{R_2 C_{23}} & & \\ & \frac{1}{R_2 C_{23}} & \frac{-1}{R_{23} C_3} & \frac{1}{R_3 C_{34}} & \\ & & \ddots & \ddots & \end{pmatrix} \, ,$$
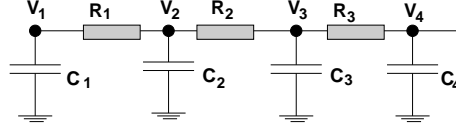
FIGURE 9.2. The serial RC-network which can represent a tridiagonal matrix $\mathbf{A}_{RC}$. In the case of a LC-network, the resistors R are replaces by inductors L.

where $R_{ij} = R_i R_j/(R_i + R_j)$ and $C_{ij} = \sqrt{C_i C_j}$. The replacement of the resistors by inductors changes very little in the network, and the corresponding matrix has the same structure:

$$(9.12) \qquad \mathbf{A}_{LC} = \begin{pmatrix} \frac{-1}{L_1 C_1} & \frac{1}{L_1 C_{12}} & & \\ \frac{1}{L_1 C_{12}} & \frac{-1}{L_{12} C_2} & \frac{1}{L_2 C_{23}} & \\ & \frac{1}{L_2 C_{23}} & \frac{-1}{L_{23} C_3} & \frac{1}{L_3 C_{34}} \\ & & \ddots & \ddots \end{pmatrix} \, ,$$

where analogous: $L_{ij} = L_i L_j/(L_i + L_j)$.

It is important to notice that charge conservation means that:

$$(9.13) \qquad \mathbf{A} \begin{pmatrix} \sqrt{C_1} \\ \sqrt{C_2} \\ \sqrt{C_3} \\ \sqrt{C_4} \\ \vdots \end{pmatrix} = \begin{pmatrix} \sqrt{C_1} & \sqrt{C_2} & \sqrt{C_3} & \sqrt{C_4} & \cdots \end{pmatrix} \mathbf{A} = 0 \, ,$$

which determines the relevant null-space of $\mathbf{A}$. The transformation with the diagonal matrix consisting of $\sqrt{C_i}$, such that $\mathbf{A} \to \mathrm{diag}(\sqrt{C_i})\mathbf{A}\,\mathrm{diag}(\sqrt{C_i})$ transforms the voltage state $x_i = V_i$ into the charge state $x_i = Q_i$. The charge conservation occurs both in the RC and the LC network, the same null-space arises.

Each of the voltages $x_i = V_i$ in the serial network is a coefficient of a mode $V_i(z)$. Hence a lumped model can arise from a numerical model order reduction of a large $\mathbf{A}$ to $\mathbf{A}_{RC}$ or $\mathbf{A}_{LC}$. However, in many cases the lumped parameters have little meaning in the underlying model. The input is connected to all the nodes since, typically, for every mode $V_i(z)$, for every node z_j, the overlap is nonzero: $V_i(z_j) \neq 0$. Hence if an external connection is made at a single node z_j, it is connected to all the modes, and all the nodes of the lumped model, if in the tridiagonalization only the internal dynamics and not the input and output are considered.

In the truncation of the tridiagonal matrix to a smaller matrix with the dominant modes, the last diagonal coefficient $(C_n R_{(n-1)n})^{-1}$ or $(C_n L_{(n-1)n})^{-1}$ must be adapted to respectively $(C_n R_{n-1})^{-1}$ or $(C_n L_{n-1})^{-1}$, in order to retain charge conservation. More generally, numerical errors in the tridiagonalization procedure can be corrected to retain global charge conservation, lost in numerical errors. This procedure can be based on the serial-network analogy.

Likewise, the complete diagonalization of $\mathbf{A}$ turns every state x_i into the coefficient of a vibrational mode $V_i(z)$. In both cases the input and output are distributed, while the dynamics is, more or less, localized. The stationary state connects input and output by a continuous flow, hence the stationary state lies the

basis for distributed dynamics and localized input and output. It is a fundamental difference in perspective.

More appropriate is to rely on the input to determine the lowest mode, such that the interconnection between the system and the environment is consistent for every lumped, low-order approximation of the system. The stationary state is an important aspect in this consistency.

9.3. Analytic input

The construction of modes, or relevant state vectors, from diagonalization, will not include the stationary state for a static response. The dominant characteristic of a system, being a resistor to heat flow, or an inertia under acceleration, is not recognized, and given a place. Furthermore, the next-to-leading-order behavior, such as the elasticity of the accelerating inertia is not singled out by a standard modal analysis. Using a polynomial expansion in time of the input does, on the other hand, separates and singles out the separate characteristics of the system under the influence of particular input.

A state-space representation of the evolution (see Chapter 3) of a linear system is given by:

$$(9.14) \qquad \dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} \ ,$$

and with the output:

$$(9.15) \qquad \mathbf{y} = \mathbf{C}\mathbf{x} + \mathbf{D}\mathbf{u} \ ,$$

where the input $\mathbf{u} \in \mathbb{R}^m$, the state $\mathbf{x} \in \mathbb{R}^n$, and the output $\mathbf{y} \in \mathbb{R}^k$ are functions of time only. If we insert an analytical expansion for the input:

$$(9.16) \qquad \mathbf{u}(t) = \mathbf{u}_0 + \mathbf{u}_1 t + \mathbf{u}_2 t^2 + \cdots + \mathbf{u}_N t^N \ ,$$

with similar expansions for the state $\mathbf{x} \in \mathbb{R}^{n \times m}$:

$$(9.17) \qquad \mathbf{x}(t) = \mathbf{x}_0 + \mathbf{x}_1 t + \mathbf{x}_2 t^2 + \cdots + \mathbf{x}_N t^N \ ,$$

and the output $\mathbf{y} \in \mathbb{R}^{k \times n}$. The typical dimension N, for a linear system does not have to exceed the dimension of the state space. The projection onto the state space determines the maximal dimension of both input richness and observability.

The input $\mathbf{u}_i t^i$ is mapped to the state space $\mathbf{x}_j$. If this mapping of a polynomial in t to a subspace of the state space retains the dimension N of the polynomial, we speak of a rich input. The full richness of the input are all the states possibly excited by an arbitrary time-dependent input.

The evolution now yields a hierarchical set of linear equations for each order t^j:

$$(9.18) \qquad (j+1)\mathbf{x}_{j+1} = \mathbf{A}\mathbf{x}_j + \mathbf{B}\mathbf{u}_j \ ,$$

where the equation for $j = 0$ is the stationary equation.

The first question to address is the input richness. Analytic input yields a full input richness if for $N = n - 1$ and $\exists \mathbf{u}_i$:

$$(9.19) \qquad \mathrm{rank}\{\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_{n-1}\} = n \ .$$

Such a definition is equivalent to the statement that there is a time t for which a state $\mathbf{x}$ can be achieved, with the analytic control $\mathbf{u}$.

The equation for $\mathbf{x}_j$ is easily solved successively, starting from $j = n - 1$:

$$(9.20) \qquad \mathbf{x}_j = \mathbf{A}^{-1}((j+1)\mathbf{x}_{j+1} - \mathbf{B}\mathbf{u}_j) \ .$$

The states can be expressed in terms of the control $\mathbf{u}(t)$:

$$(9.21) \qquad \mathbf{x}_{n-i} = -\sum_{j=1}^{i} \frac{(n-j)!}{(n-i)!} \mathbf{A}^{j-i-1} \mathbf{B} \mathbf{u}_{n-j} \quad .$$

The linear independence of the different $\mathbf{u}_i$ and the upper-diagonal structure of the coefficient matrix yield the input richness condition.

The input richness is equivalent to:

$$(9.22) \qquad \text{rank}\{\mathbf{A}^{-1}\mathbf{B}, \mathbf{A}^{-2}\mathbf{B}, \cdots, \mathbf{A}^{-n}\mathbf{B}\} = n \quad .$$

This input richness is similar but not identical to the standard controllability condition, Eq. 3.29. Controllability is typically associated with low-dimensional systems. Distributed systems have many more degrees of freedom than of interest for standard operation, hence is never controllable. The question is reversed: how many states are participating in the response to typical, slow, i.e., analytic input.

The difference lies in the inverse of $\mathbf{A}$. If $\mathbf{A}$ is singular, the stationary state, associated with the right null-space of $\mathbf{A}$, cannot be maintained with a single stationary input, as would be expected from the $j = 0$ equation.

9.4. Delay systems

Many distributed systems are, as input-output systems, characterized by a delay τ. In terms of states it means that different states have time-shifted algebraic relations: $x_j(t + \tau) = x_i(t)$. Transmission cables are typical examples: the signal enters one end x_i, and arrives τ time later at the other end x_j. In terms of differential equations, delay corresponds to a differential equation of infinite order:

$$(9.23)$$
$$e^{\tau \partial_t} x_j(t) = x_j(t) + \tau \partial_t x_j(t) + \frac{\tau^2}{2!} \partial_t^2 x_j(t) + \frac{\tau^3}{3!} \partial_t^3 x_j(t) + \cdots = x_j(t + \tau) = x_i(t) \quad ,$$

where ∂_t is the differentiation with respect to time as operator, which relies on an analytic expansion of $x_j(t)$. Such delay equations are truncated at a particular order to analyze the effects of a feedback control loop. [Minorsky, 1942]

In terms of analytic input, delay equations are upper triangular matrices. The shift of a polynomial $u(t) = u_0 + u_1 t + u_2 t^2 + \cdots u_n t^n$ will be a polynomial of the same order:

$$(9.24) \quad u(t + \tau) = (1, t, t^2, t^3, \cdots, t^n) \begin{pmatrix} 1 & \tau & \tau^2 & \tau^3 & \cdots & \\ 0 & 1 & 2\tau & 3\tau^2 & 4\tau^3 & \cdots \\ 0 & 0 & 1 & 3\tau & 6\tau^2 & \cdots \\ & & & \ddots & \cdots & \vdots \end{pmatrix} \begin{pmatrix} u_0 \\ u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix} ,$$

where the coefficients follow from the binomial expansion.

If the output is the delayed input, the upper-diagonal matrix, Eq. 9.24, will appear in the system matrix $\mathbf{A}$. For large delays the order of the input t^n, rather than the number of sampling points, determines whether the system is well-conditioned and stable.

9.5. Finite time

A free system has a unique evolution given by its initial state $\mathbf{x}(0)$, which is equivalent to $\mathbf{x}_0$ in the analytic case. The series $\{\mathbf{x}_i\}_{i=0}^{\infty}$ expands to infinity, however, only the first n terms are relevant to define $\mathbf{A}$ and thereby the free system uniquely. In the case of input $\mathbf{u}$ the dimensionality of the input determines the dimensionality of the system that can be identified or controlled. In the standard case an infinite number of derivatives of the input are assumed available to reconstruct the state space. In practice, this is never the case.

In general, the input dimension is much lower than the dimension of any physical system, although many states correspond to irrelevant, damped high-frequency eigenstates. The constructive analysis to optimize an input profile consisting of a piece-wise polynomial function is a version of Heaviside calculus.

Piece-wise polynomial solutions can be constructed in many different ways. For example, splines could lie at the basis of piece-wise polynomial solutions. Higher order splines themselves can be constructed from successive convolutions of block functions, which indicate the low frequency effect of splines. The k-th order spline has a Fourier transform:

$$(9.25) \qquad S^{(k)}(\omega) = \left(\frac{\sin T\omega}{\omega} \right)^k \quad ,$$

where T is the length of the block of unit height. Since a piece-wise polynomial function can be written as the convolution of a Dirac comb with a spline, it is clear that the k-power of the the spline yields a low-frequency dominance of the piece-wise polynomial input data, depending both on the domain T, and on the order of k.

For analytic input the state is a slave of the input, for the first n orders determined by the input. In the case of a piece-wise polynomial input this is no longer the case. The state may be arbitrary.

The study of finite time is also part of the theory of input shaping. [Singer and Seering, 1990] The typical step response is replaced by a smoother, typical spline, function of time. The function is optimized by minimizing some cost function of the defined output. There are many variations in how define the output and a cost function to minimize. Furthermore, the shaping functions can have a large variety, for example in fixed or free abscissas, order, etc.. Most of the approaches rely on linearity of the system, and the problem shifts from input shaping toward system identification. Here we start at completely the opposite end of the spectrum, we know the system, but it is too large. The limitations of the input help us to define a relevant subspace of the total state space.

9.6. Failing input richness

Unlike the controllability condition, which will simply fail for insufficiently rich input, the richness condition will break down more progressively, in many cases. The state vectors $\mathbf{x}_i$ will still have components which span an additional subspace with every new state vector $\mathbf{x}_{i+1}$, associated with a part of the vector perpendicular to all the previous vectors. This is due to the very large dimension of the state space, based on FEM or some other high-order approximation. However, this component will decrease. Within some tolerance ϵ a particular state vector $\mathbf{x}_{i+1}$ will no longer

be perpendicular to the space spanned by: $\{\mathbf{x}_0, \mathbf{x}_1, \cdots, \mathbf{x}_i\}$:

$$(9.26) \qquad \min_{\alpha} \frac{\|\alpha_0 \mathbf{x}_0 + \alpha_1 \mathbf{x}_1 + \cdots + \alpha_i \mathbf{x}_i - \mathbf{x}_{i+1}\|}{\|\mathbf{x}_{i+1}\|} < \epsilon \quad .$$

The input spans thereby a space of i dimensions; the full input richness is i for the tolerance ϵ. In the theory of Krylov spaces a small input richness is an unwanted feature; no longer the full state space is spanned by the input. For control purposes, the small input richness is advantageous. Only a limited number of states are excited, and, therefore, only a limited number of states are required to describe the system dynamics. If, however, more states are included in the state-space model, the dynamics will fail, with singular behavior associated with near-parallel state vectors.

In the construction of the dynamics in the finite-dimensional subspace, Section 9.8, we will consider the effects of failing input-richness, its causes and consequences, further. The failing input richness can be seen as a lack of structure in the system. For example, consider a force exerted on an elastic body in free motion. The stationary response is the acceleration of the inertia as a whole. The next term is the delay in acceleration, due to the elastic deformation of the body. If the object is more or less homogeneous, it can be represented as a series of masses and springs with a homogeneous spectrum. However, if the mass is located in one place of the object and the elastic deformation has no associated mass, the system is well-represented by a single mass-spring pair, and the input richness will fail. It will fail, however, for physical reasons, the remainder of the spectrum, after the dominant mode, occurs at much higher frequencies.

9.7. Smooth matching

The smoothness of the spline input $u(t)$ determines the frequency band in which the input occurs. Hence in terms of a piece-wise polynomial input, it is not the order of the polynomial itself, but the smoothness of the transition from one part to the next that determines the decay rate of the higher frequencies in the signal.

The matching means that at the end of each time slice $t \in [0, T]$ the state $\mathbf{x}(T)$ is determined and used as initial state $\mathbf{x}(0)$ for the next time slice. Rather than a state fully determined by the analytic input, the piece-wise analytic domain will have the initial state as inhomogeneous term, and the equations of motion must be integrated explicitly:

$$(9.27) \qquad \mathbf{x}(T) = e^{\mathbf{A}T} \int_0^T e^{-\mathbf{A}t} \mathbf{B}\mathbf{u}(t)dt + e^{\mathbf{A}T}\mathbf{x}(0) \quad .$$

The state $\mathbf{x} \in \mathbb{R}^n$ is a long vector, which we expanded, in the analytic input approach, in the $\{t^i\}_{i=0}^N$ polynomial dependence on the input. Hence, rather than working with the very large dimension n, we will map the problem on the N-dimensional space of analytic responses.

9.8. Projection on the input space

Previously, we used analytic input to construct a number of modes, or states, which are relevant for the input we are considering. These modes are a particular Krylov space. However, the modes themselves do not determine the dynamics, or the reduced $\mathbf{A}, \mathbf{B}, \mathbf{C}$, and $\mathbf{D}$ matrices. The dynamics of the full space does not guarantee that the state remains in the subspace spanned by the Krylov vectors.

The dynamics out of the Krylov space must be projected back into this subspace under consideration. Many different choices for the constructed of the reduced dynamics in the reduced subspace exists. We will give the general procedure to construct reduced dynamics.

In numerical mathematics the controllability condition is similar to the span of a Krylov space. The dimension of the Krylov space spanned in the rank of controllability matrix. Hence, in the case of model order reduction, based on the input vectors $\mathbf{B}$, the controllabity tells us how well the system is approximated. The difference between the standard controllability:

$$(9.28) \qquad \tilde{\mathbf{X}} = \{\mathbf{B}, \mathbf{A}\mathbf{B}, \mathbf{A}^2\mathbf{B}, \cdots \mathbf{A}^{n-1}\mathbf{B}\} \ ,$$

and the analytic controllability:

$$(9.29) \qquad \mathbf{X} = \{\mathbf{B}, \mathbf{A}^{-1}\mathbf{B}, \mathbf{A}^{-2}\mathbf{B}, \cdots \mathbf{A}^{-n+1}\mathbf{B}\} \ ,$$

has more significance in the numerical setting where the dimension of $\mathbf{A}$ is high, than in the pure control setting.

In the original input dependence, states were constructed algebraically by collecting the same orders of t and selecting the set of equations which would yield a one-to-one map under general controllability conditions. In general the state can be arbitrary. Furthermore, in physical control, the number of states is likely to be infinite. However, from rest $\mathbf{x} = 0$ or any constant state $\mathbf{x} = -\mathbf{A}^{-1}\mathbf{B}u$ for constant input, the time-dependent input determines the dominant subspace of the state space. Eigenstates, not initially excited by the analytic input, are not very likely to be excited at all. In order for such a situation to occur the diagonalization of the matrix $\mathbf{A}$ should contain Jordan blocks larger than the input space dimension.

Hence we seek an approximated, or reduced-order, model determined by the input $\mathbf{u}(t)$. The states $\mathbf{x}_i$ are already determined from the analytic input-to-state mapping, but a projection onto an n-dimensional dual $\mathbf{x}_i^*$ is needed to close the set of equations and yield a n-dimensional linear set:

$$(9.30) \qquad \mathbf{X}^*\dot{\mathbf{X}} = \mathbf{X}^*\mathbf{A}\mathbf{X} + \mathbf{X}^*\mathbf{B}\mathbf{u} \ ,$$

where $\mathbf{X}$ is the matrix constructed from the columns $\mathbf{x}_0, \mathbf{x}_1, \cdots$, and $\mathbf{X}^*$ is likewise constructed from the rows $\mathbf{x}_0^*, \mathbf{x}_1^*, \cdots$. Four obvious choices for the dual $\mathbf{X}^*$ are easily given. The state dual $\mathbf{X}^* = \mathbf{X}^T$ based on the l_2-norm of the state space, the state inverse $\mathbf{X}^* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ which diagonalizes the left-hand side, the weighted dual $\mathbf{X}^* = \mathbf{X}^T\mathbf{A}^{-1}$, and the weighted inverse $\mathbf{X}^* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{A}^{-1}$ which diagonalizes the subspace $\mathbf{A}$ matrix.

All of these dual spaces yield basic block-Hankel matrices for evaluation:

$$(9.31) \qquad H_r = \begin{pmatrix} \mathbf{B}^T\mathbf{A}^{-r}\mathbf{B} & \mathbf{B}^T\mathbf{A}^{-r-1}\mathbf{B} & \mathbf{B}^T\mathbf{A}^{-r-2}\mathbf{B} \\ \mathbf{B}^T\mathbf{A}^{-r-1}\mathbf{B} & \mathbf{B}^T\mathbf{A}^{-r-2}\mathbf{B} & \mathbf{B}^T\mathbf{A}^{-r-3}\mathbf{B} \\ \mathbf{B}^T\mathbf{A}^{-r-2}\mathbf{B} & \mathbf{B}^T\mathbf{A}^{-r-3}\mathbf{B} & \cdot \\ \mathbf{B}^T\mathbf{A}^{-r-3}\mathbf{B} & \cdot & \end{pmatrix} \ ,$$

which are known to lead to poor stability after the fast initial convergence. However, this poor stability is related to higher frequencies in the system, which are, with sufficiently smooth input, not necessary of interest, as pointed out in Section 9.6.

9.9. Laguerre projection

Analytic input has no generic time-scale, analytic functions scale proportionally with the scaling of the time. This approach is very useful for hyperbolic systems, which are not damped, but remain oscillating. Every input, however long ago, adds to the state dynamics, depending on the phase, However, in many cases systems are damped, and they have an inherent time-scale, given by the distance of the poles in the transfer function to the real ω-axis, or imaginary s-axis. Hence, of the time-dependent input only a window, the size of the minimal damping rate, is relevant for the state of the system. The input richness is limited to the behavior in that window. Analytic input cannot separate the behavior inside and outside the window. More appropriately in such situations is to consider a Laguerre projection, which is an orthogonal basis on the window function based on the exponential decay in time of input.

Instead of using the state space, and the projection on the dual space, the time-dependence of input and states can be used to determine to determine the state as function of the input. In the case of the lowest mode having the characteristics of a resistance, it yields a time constant λ, which can serve as the basic scale of an orthogonal basis expansion in terms of Laguerre polynomials $L_i(t)$: [Weniger, 1985]

$$(9.32) \qquad c_i = \lambda \int_{-\infty}^{0} e^{\lambda t} \mathbf{u}(\lambda t) L_i(-\lambda t) dt \quad .$$

The Laguerre polynomials form an orthogonal basis: [Magnus et al., 1966]

$$(9.33) \qquad \int_{-\infty}^{0} e^{t} L_j(-t) L_i(-t) dt = \delta_{ij} \quad .$$

The normalization varies from author [Abramowitz and Stegun, 1965] to author, we use [Magnus et al., 1966] as the standard. In a more general setting, with smooth input, one can use the generalized Laguerre polynomials, where $L_i(t) = L_i^{(0)}(t)$ with the weight function $(-t)^{\alpha} e^{t}$, where $\alpha > -1$ is to be optimized:

$$(9.34) \qquad \int_{-\infty}^{0} (-t)^{\alpha} e^{t} L_j^{(\alpha)}(-t) L_i^{(\alpha)}(-t) dt = \delta_{ij} \Gamma(\alpha+1) \begin{pmatrix} i+\alpha \\ i \end{pmatrix} \quad .$$

The Laguerre polynomials satisfy the second-order differential equation:

$$(9.35) \qquad t\partial_t^2 L_n^{(\alpha)}(t) + (\alpha+1-t)\partial_t L_n^{(\alpha)}(t) + n L_n^{(\alpha)}(t) = 0 \quad .$$

In Section 9.3 the modes or states $\mathbf{x}_i$ were constructed by solving algebraically the evolution equation for every power of time, t^i, necessitating a truncation of the highest order in the analytic expansion of the state, $\mathbf{x}_n t^n$. The Laguerre projection is an alternative form of constructing similar, but not identical, states. The Fourier transform of Laguerre functions $e^{-t} L_n(t)$ have interesting properties, with a single pole and zero of $(n+1)$-th and n-th order. [Wiener, 1949] The stationary state arises from the lowest-order expansion in Laguerre polynomials with a weighted time-average input $\mathbf{u}_0$:

$$(9.36)$$
$$0 = \int_{-\infty}^{0} dt \left(\dot{\mathbf{x}}(t) e^{\lambda t} - \mathbf{A}\mathbf{x}(t) e^{\lambda t} - \mathbf{B}\mathbf{u}_0 e^{\lambda t} \right) = \mathbf{x}(0) + (\mathbf{1} - \lambda^{-1}\mathbf{A})\mathbf{x}_0 - \lambda^{-1}\mathbf{B}\mathbf{u}_0 \quad .$$

Hence there are four directions in state space: the state at $t = 0$: $\mathbf{x}(0)$, the average state: $\mathbf{x}_0$, the average change: $\mathbf{A}\mathbf{x}_0$, and the control input average: $\mathbf{B}\mathbf{u}_0$. The system can be solved yielding the stationary state as average state: $\mathbf{x}_0 = -\mathbf{A}^{-1}\mathbf{B}\mathbf{u}_0$, with $\mathbf{x}(0) = \mathbf{x}_0$. The equation can be balanced, given equal weight to the different parts for $\lambda = \|\mathbf{A}\|$.

In order to construct a set of modes in state space, similar to the analytic expansion $\mathbf{x}_i$, we project the equation of motion onto the different Laguerre polynomials, which define the weighted time averages $\mathbf{x}_i$ and $\mathbf{u}_i$:

$$(9.37) \quad 0 = \int_{-\infty}^{0} dt \left(\dot{\mathbf{x}}(t) L_i(-\lambda t) e^{\lambda t} - \mathbf{A}\mathbf{x}(t) L_i(-\lambda t) e^{\lambda t} - \mathbf{B}\mathbf{u}(t) L_i(-\lambda t) e^{\lambda t} \right) \quad .$$

We use the following properties of Laguerre polynomials: [Magnus et al., 1966]

$$(9.38) \qquad \partial_t L_i^{(\alpha)}(t) = \sum_{j=0}^{i-1} (-1)^{i-j} L_j^{(\alpha)}(t) \quad ,$$

and

$$(9.39) \qquad L_i^{(\alpha)}(0) = \left(\begin{array}{c} n + \alpha \\ n \end{array} \right) \quad ,$$

which yields the j-th equation to solve $\mathbf{x}_i$ and $\mathbf{x}(0)$ from the input $\mathbf{B}\mathbf{u}$ for $\alpha = 0$:

$$(9.40) \qquad 0 = \mathbf{x}(0) + \sum_{i=0}^{j} (-1)^{j-i} \mathbf{x}_i - \lambda^{-1} \mathbf{A}\mathbf{x}_j - \lambda^{-1} \mathbf{B}\mathbf{u}_j \quad .$$

In matrix form the equations for the modes are:
(9.41)

$$\left(\begin{array}{c} \mathbf{x}(0) \\ \mathbf{x}(0) \\ \vdots \\ \mathbf{x}(0) \end{array} \right) + \left(\begin{array}{ccccc} 1 & 0 & 0 & 0 & \cdots \\ -1 & 1 & 0 & 0 & \cdots \\ 1 & -1 & 1 & 0 & \cdots \\ & \ddots & \ddots & & \\ \cdots & \cdots & 1 & -1 & 1 \end{array} \right) \left(\begin{array}{c} \mathbf{x}_0 \\ \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_n \end{array} \right) = \left(\begin{array}{c} \lambda^{-1}\mathbf{A}\mathbf{x}_0 \\ \lambda^{-1}\mathbf{A}\mathbf{x}_1 \\ \vdots \\ \lambda^{-1}\mathbf{A}\mathbf{x}_n \end{array} \right) + \left(\begin{array}{c} \lambda^{-1}\mathbf{B}\mathbf{u}_0 \\ \lambda^{-1}\mathbf{B}\mathbf{u}_1 \\ \vdots \\ \lambda^{-1}\mathbf{B}\mathbf{u}_n \end{array} \right) \quad .$$

Setting $\mathbf{x}(0) = 0$ for a moment, the modes $\mathbf{x}_i$ can be solved as functions of $\mathbf{B}\mathbf{u}_i$:

$$(9.42) \qquad \mathbf{x}_i = \sum_{j=0}^{i} (\lambda - \mathbf{A})^{i-j-1} \mathbf{B}\mathbf{u}_j \quad .$$

In the general case the solution $\mathbf{x}(0)$ is sought such in the subspace spanned by $\{\mathbf{A}^{-i}\mathbf{B}\mathbf{u}\}_{i=0}^{\infty}$, arising from the analytic expansion of $(\text{constant} - \mathbf{A})^{-1}$. Unlike the analytic case the vector $\mathbf{x}_{-1} = \mathbf{B}\mathbf{u}$ plays a role in the Laguerre projection.

Depending on the particular application, different procedures can be followed. For example, one might optimize the value of λ for a particular time-dependent input, $\mathbf{u}_i$. It is important to notice the different linkages between the time-dependent, and the restrictive input signal, and the subspaces of the large state space arising from a FEM model. Input is central in restricting to the relevant dynamical subspace of the state space. This is a different approach from mapping a transfer function between a significant, i.e., physical, input and a, usually rather arbitrary, output, as function of the frequency. [Antoulas, 2005, Worden and Tomlinson, 2001]
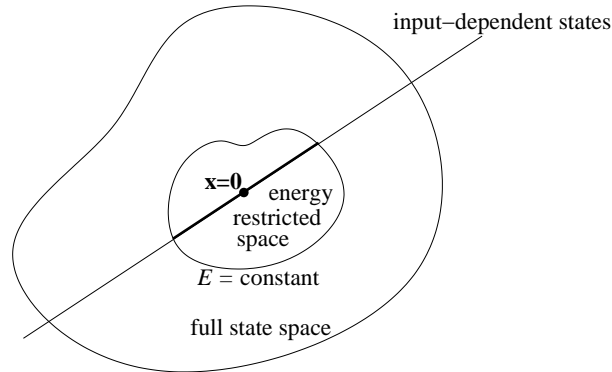
FIGURE 9.3. The energy and the input together restrict the full
state space to a small subspace of dynamical states, which are rel-
evant for the driven, or operational, dynamics. The input restricts
the dimension of the subspace, while the energy restricts the norm
$E \leq$ constant. These can be the operational settings of the partic-
ular system due to the restricted input.

9.10. Remarks on the limitations of linearity

In this chapter we restricted our attention mainly to linear models, typical
of large FEM applications. In this particular field physical knowledge is buried
deep in the large-matrix representations. However, time-dependent behavior may
still pry some of the relevant physics from these representations. For example, the
stationary state is particularly relevant for determining the main function of the
systems under investigation, e.g., being a resistive or capacitive element, and the
scales associated with both characteristics.

Extending this view we note that the operational setting of a system is usually
much more restrictive than the full state space of the FEM model. The time-
dependence of the input is used to determine the smaller, and more versatile, dy-
namical subspace of the full state space. The use of analytic input allows one to
include the stationary state in this subspace, a result commonly not achieved with
impulse response approaches, with a preset input signal, such as a step function or
oscillatory input. [Lee et al., 1999]

After the modes are selected, still some relevant information of the system
at hand is necessary to construct the dynamical equations in the subspace of the
selected modes. In this chapter we restricted ourselves to a noting different projec-
tions on the relevant dual subspaces. Physical principles, such as conservation laws
can be used to select the proper norm and projection. See Figure 9.3.

In the case of systems with nonlinearities the same physical principles can be
used to retain relevant modes from the full space. The input is to be expressed
terms of conserved quantities, such as energy, charge, or mass, such that the state
space dynamics is automatically restricted to the subspace limited by the same
conservation laws. In the case of nonlinear system, the conservation laws are still
linear. They are the guiding principles to steer clear from the instabilities and mul-
tiple, spurious solutions, typical on nonlinear equations. In the following chapters
these questions are investigated further.

CHAPTER 10

Real and virtual dynamics

> For since the fabric of the universe is most perfect and the work
> of a most wise creator, nothing at all takes place in the universe
> in which some rule of maximum or minimum does not appear.
> [L. Euler]

The term virtual work was coined by Lagrange to determine the internal balance of forces, and the balance of forces between objects. He used the concept in the introduction of generalized coordinates and the study of celestial mechanics in the second half of the eighteenth century. [Gantmacher, 1970] D'Alembert consequently popularized the concept.

Force is not an invariant concept, but the product of force and displacement is. In the continuous setting force could be pressure, or some more complex distributed source, expressed as a stress tensor. The product of any such representation of the force and its dual displacement is the invariant work, independent of the representation.

Work allows for general coordinate transformations. The balance is the configuration, or state, of minimal energy, and a small, virtual, change of configuration will always increase the energy. In other words: the virtual change of the state is at the expense of virtual work. In this chapter we will take to analogy one step further and assign to the iteration procedure a virtual time. The process of approximating a solution successively is treated as a dynamical process. Once the analogy is established, it is taken one step further by combining real dynamics with virtual dynamics to describe dynamical systems with constraint relations.

Virtual work can be associated with many other problems. For example, convergent iterative processes [Saad, 1995] are convergent in some norm, which can be interpreted as the energy of the system $E(\mathbf{x}) = \|\mathbf{x}\|_E$, and the Gateaux derivative with respect to the state $\mathbf{x}$ is the associated virtual force:

$$(10.1) \qquad \mathbf{F} = -\frac{\delta E(\mathbf{x})}{\delta \mathbf{x}} \quad .$$

The iterative process can be identified as a damped dynamical process. Furthermore, constrained systems with dynamics can be treated as a whole: the real dynamics can be combined with virtual dynamics of the constraint relations. The linkage exists at multiple levels. First we investigate the time scale and damping of a variational problem given by a sparse matrix. Second, we analyze the virtual dynamics associated with multiple constraints. Finally, we combine the real and virtual dynamics and reduce the influence of the virtual dynamics on the real dynamics.

Linear variational problems with sparse matrices usually arise from FEM approach to partial differential equations. Such matrices have quite a lot of structure

from the underlying geometry. A paradigmatic example is the simplest approximation of the Laplacian Δ in one dimension. The corresponding matrix operator $\mathbf{A} \in \mathbb{R}^{N \times N}$, associated with a piece-wise linear approximation is given by:

$$
(10.2) \qquad \mathbf{A} = \begin{pmatrix} -1 & 1 & 0 & & & \\ 1 & -2 & 1 & 0 & & \\ 0 & 1 & -2 & 1 & 0 & \ddots \\ & 0 & 1 & -2 & 1 & \\ & & \ddots & \ddots & \ddots & \ddots \end{pmatrix} .
$$

The particular Hilbert space l_2 in which the solution $\mathbf{x}$ lies has a priori no significance. The nature of the problem is given by $\mathbf{A}$ and the inhomogeneous term $\mathbf{b}$ for $\mathbf{Ax} = \mathbf{b}$. In general, the solution $\mathbf{x} = (x_1, x_2, \cdots, x_N)$ will have all coefficients $x_i \neq 0$, in no particular order of size $x_i \sim x_j$. A more natural space is the Krylov space $\mathcal{K}_n(\mathbf{x}_0, \mathbf{A}) = \{\mathbf{x}_0, \mathbf{Ax}_0, \mathbf{A}^2\mathbf{x}_0, \cdots, \mathbf{A}^n\mathbf{x}_0)$. The only candidate for the initial vector $\mathbf{x}_0 = \mathbf{b}$ is the inhomogeneous term $\mathbf{b}$, usually representing an applied force of in flow.

An iterative method, for the solution $\mathbf{x}$ of the equation $\mathbf{Ax} = \mathbf{b}$, is based on the successive multiplication of the matrix $\mathbf{A}$ with the vector $\mathbf{b}$. If the vector $\mathbf{b} = (1, 0, 0, 0, \cdots, 0, 0)$ it takes $N - 1$ multiplications before the whole space is filled, for example:

$$
(10.3) \qquad \mathbf{A}^3\mathbf{b} = \mathbf{A}^3 \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \begin{pmatrix} -5 \\ 9 \\ -5 \\ 1 \\ 0 \\ \vdots \end{pmatrix} .
$$

Only the first four coefficients are non-zero. As in most solutions to such variational problems all the coefficients are nonzero, therefore, at least $N - 1$ multiplications are required to find a solution. The matrix $\mathbf{A}$ (Eq. 10.2) is the linear discretization of the Laplacian in one dimension. See Sect. 8.4 The walk through the space from the input at one end, to the other end requires N steps, equivalent to $N - 1$ multiplications with $\mathbf{A}$. The connectivity is $N - 1$, as it takes $N - 1$ steps to walk through the whole state space. The conjugate gradient method, discussed in Section 10.3, is a method which requires precisely N iterations to find the solution, but the amount of work per iteration is more than the virtual dynamic approach requires, as we will see. Furthermore, the dynamical interpretation allows us to extend the method to nonlinear problems.

Another aspect of iterative methods is also immediately clear from inspecting the vector $\mathbf{A}^3\mathbf{b}$. The coefficients are growing and they are alternating. The series $\mathbf{A}^j\mathbf{b}$ is not converging, since $\|\mathbf{A}\| = 2 > 1$. If $\mathbf{A}$ is multiplied with an appropriate factor $(\Delta t)^{-1}$, the series is convergent. The factor is the reciprocal of the discrete time step Δt, associated with evolution and dynamics, discussed in chapter 3. Iterative procedures have a close link with integration methods for stable dynamics.

The number of steps n determine the subspace in which the approximate solution lies. This space is the Krylov space $\mathcal{K}_n(\mathbf{A}, \mathbf{b})$. The connectivity indicates the

minimal size of the subspace. The larger the dimension of the underlying geometry the smaller the connectivity compared with the dimension N of the matrix. In treating the iterative solving of a numerical problem as a dynamical problem these are all aspects that can be taken into account. Eventually more prior knowledge will improve the convergence.

A last aspect, also clear from inspecting the vector $\mathbf{A}^3\mathbf{b}$, is that the sum of the coefficients add up to zero. This is no accident, but the result of the underlying PDE, the Laplace operator. Any constant function of position vanishes, as does any linear function. These vectors lie in the left null-space of the Laplace operator, Eq. 10.2. The FEM discretization should retain this property as this associated with the force balance or mass conservation depending on the underlying problem. Hence deviations from zero for the sum give some insight in the convergence, and the position of the interacting boundaries and the forces at these boundaries. By treating the iterative process as a numerical algorithm only, such aspects are often forgotten.

Recently, the underlying physical system is been taken into account again, at a limited scale, in methods referred to as structure preserving model-order reduction. [Freund, 2003] For example, the positivity of the mass matrix and the stiffness matrix are retained in the approximation, corresponding to a Krylov space $\mathcal{K}_n$ with its dimension much smaller than the dimension N of the full space.

To recapitulate, real and virtual dynamics refer to two aspects of a dynamical simulation of a system. The real dynamics is the time evolution of a system as determined by the equations of motion. Virtual dynamics refers to the iterative solution of a numerical problem, [Saad, 1995, Antoulas, 2005] and treating it as a dynamical problem. [Stuart and Humphries, 1998] These two problems are intimately related, since they are both based on the series $\mathbf{A}^j\mathbf{b}$. First of all, in the iterative solving of a linear equation an approximate solution converges to the true solution. If each step is seen as a time step, the convergence process is similar to that of a stable damping, usually associated with Lyapunov theory. [Lyapunov, 1992] Hence, turning the problem around, many numerical problems can be solved once they are converted to a stable, virtual, dynamical problem.

On the other hand, constraints, or algebraic equations, may appear in dynamical problems, which are simulated through the time integration of the equations of motion. In each step the constraints need to be solved. The explicit solving yields a corrector step; after each step the constraints are implemented, the implicit solving yields an adjusted step, which is already tangential to the constraint surface. However, constraints may be introduced as virtual dynamics such that, even if the initial state does not satisfy the constraint relations, the state converges to the constraint surface and remains there within some numerical bound. A simple generic corrector step is the Newton iteration. If a function should satisfy $H(\mathbf{q}) = E$, the steepest descent yields the smallest linear correction $\mathbf{q} \to \mathbf{q} + \delta\mathbf{q}$, with min $\|\delta\mathbf{q}\|$ to the solution $H(\mathbf{q} + \delta\mathbf{q}) = E$:

$$(10.4) \qquad \delta\mathbf{q} = \frac{(E - H(\mathbf{q}))\nabla_{\mathbf{q}}H}{(\nabla_{\mathbf{q}}H)^T\nabla_{\mathbf{q}}H} \quad .$$

See Figure 10.1. More elaborate schemes exists.

All of these different cases will be discussed in this chapter. The advantages of using virtual dynamics are manifold. First of all, the methods described here can be applied irrespective of the nature of the constraint, unlike most methods. Moreover,
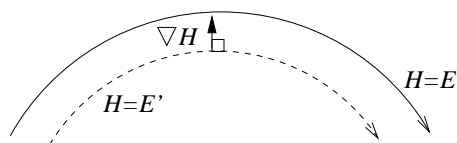
FIGURE 10.1. A corrector step in the direction of the gradient ∇H is the smallest step to move from a trajectory with the incorrect energy E' to the trajectory with the correct energy E.

both the constraints and the equations of motion may be nonlinear, which makes the method versatile. Furthermore, the methods are generic, many problems can be treated in the same manner, and the complexity will not change the general dynamic approach.

10.1. Iterative methods

The relation between symmetric matrices and quadratic Hamiltonians is well-established. The theory of matrices is commonly applied to dynamical systems to derive results of stability and vibrational analysis. However, this relation works both ways. The results above can be set into an explicit Hamiltonian dynamical scheme. Furthermore, the connection with dynamical systems allows one to extend iterative methods to nonlinear problems.

A simple example, related to the problem above, to set the mind, is the positive-definite symmetric matrix $\mathbf{A}$, and the linear equation $\mathbf{A}\mathbf{x} = \mathbf{b}$. If the matrix $\mathbf{A}$ is large and sparse, a direct solver will take prohibitively long.

A corresponding Hamiltonian can be:

$$(10.5) \qquad H(\mathbf{q}, \mathbf{p}) = K + V = \frac{1}{2}\mathbf{p}^2 + \frac{1}{2}\mathbf{q}^T \mathbf{A}\mathbf{q} \ .$$

The Hamiltonian describing the motion of a unit-mass particle is a multi-dimensional ellipsoidal potential V, with a kinetic energy K. The equations of motion are:

$$(10.6) \qquad \begin{pmatrix} \dot{\mathbf{q}} \\ \dot{\mathbf{p}} \end{pmatrix} = \begin{pmatrix} 0 & \mathbf{1} \\ -\mathbf{A} & 0 \end{pmatrix} \begin{pmatrix} \mathbf{q} \\ \mathbf{p} \end{pmatrix} \ ,$$

where the dots in $\dot{\mathbf{q}}$ and $\dot{\mathbf{p}}$ indicate the time derivatives. The solution $\mathbf{q}(t)$ and $\mathbf{p}(t)$ will now oscillate around the minimum $\mathbf{q} = 0$ and $\mathbf{p} = 0$.

The right-hand side $\mathbf{b}$ be added to the equation can be seen as an applied force:

$$(10.7) \qquad \begin{pmatrix} \dot{\mathbf{q}} \\ \dot{\mathbf{p}} \end{pmatrix} = \begin{pmatrix} 0 & \mathbf{1} \\ -\mathbf{A} & 0 \end{pmatrix} \begin{pmatrix} \mathbf{q} \\ \mathbf{p} \end{pmatrix} + \begin{pmatrix} 0 \\ \mathbf{b} \end{pmatrix} \ .$$

The solution $(\mathbf{q}(t), \mathbf{p}(t))$ will oscillate now around the new minimum, which is given by $(\mathbf{q}, \mathbf{p}) = (\mathbf{x}, 0)$, where $\mathbf{x}$ being the solution of $\mathbf{A}\mathbf{x} = \mathbf{b}$. Adding a damping term γ to the equations of motion will make the solution spiral toward the minimum:

$$(10.8) \qquad \begin{pmatrix} \dot{\mathbf{q}} \\ \dot{\mathbf{p}} \end{pmatrix} = \begin{pmatrix} 0 & \mathbf{1} \\ -\mathbf{A} & -\gamma \end{pmatrix} \begin{pmatrix} \mathbf{q} \\ \mathbf{p} \end{pmatrix} + \begin{pmatrix} 0 \\ \mathbf{b} \end{pmatrix} \ .$$

The damping is optimal whenever it is neither too small, which keeps the solution oscillating, nor too large such that the changes are slow. Hence the optimal damping, together with the proper step size Δt, and integration routine will drive the system the quickest to the solution $(\mathbf{x}, 0)$. Such method falls into the class

of direction set methods, like the conjugate gradient method, where $\mathbf{p}$ defines the direction, or improved gradient.

The conditioning of the problem can be carried out by adding a mass term to the momenta, or direction, $\mathbf{p}$:

$$(10.9) \qquad \begin{pmatrix} \dot{\mathbf{q}} \\ \dot{\mathbf{p}} \end{pmatrix} = \begin{pmatrix} 0 & (\mathrm{diag}\mathbf{A})^{-1} \\ -\mathbf{A} & -\gamma \end{pmatrix} \begin{pmatrix} \mathbf{q} \\ \mathbf{p} \end{pmatrix} + \begin{pmatrix} 0 \\ \mathbf{b} \end{pmatrix} \ ,$$

where the unit matrix is replaced by the reciprocals of the diagonal elements of $\mathbf{A}$. If the problem is close to diagonal, all the eigenvalues will be of the same order, which allows for a faster iteration scheme. In the study of vibrational problems the equations of motion contain the mass $\mathbf{M}$ and the stiffness $\mathbf{K}$:

$$(10.10) \qquad \begin{pmatrix} \dot{\mathbf{q}} \\ \dot{\mathbf{p}} \end{pmatrix} = \begin{pmatrix} 0 & \mathbf{M}^{-1} \\ -\mathbf{K} & 0 \end{pmatrix} \begin{pmatrix} \mathbf{q} \\ \mathbf{p} \end{pmatrix} \ .$$

The typical frequency is given by:

$$(10.11) \qquad \omega = \sqrt{\mathbf{K}\mathbf{M}^{-1}} \ .$$

Hence, due to the conditioning with the inverse of the diagonal part of $\mathbf{A}$, the time scale is independent of the problem with all frequencies $\omega \approx 2\pi$. If the step size is too large, a simple Euler integration, associated with standard iteration, will spiral outward in the absence of damping.

Damping and step size are concepts which give intuitive insight into the iterative procedure. Furthermore, these concepts apply to linear systems and nonlinear systems alike. The method can even be illustrated with different material blocks, or balls, in a bowl. The block will move down to the bottom. If the damping, or friction, is too large, it will move very slowly. If the friction is too small, it will not come to rest, at the bottom. See Figure 10.2. Furthermore, the results of numerical time integration can be taken to our advantage to improve the convergence and optimize the memory usage. Beside the theory of time integration, the theory of optimal control can define a fast and smooth path to be taken as ideal for the damping.

In principle the path should straighten during the iteration. If one tries to straighten it to quickly, it might miss flat directions in the potential defined by $\mathbf{A}$. On the other hand if it straightens to slowly, the step size has to remain small to follow the bends in the path. An active, i.e., predictor-corrector, approach to damping and step size may speed up the convergence of the iterative method.

One of the nicest parts of a dynamical interpretation is the strict bounds on the convergence. Even though one might overshoot the solution with a large step, the combination of the potential and the kinetic energy $\frac{1}{2}\mathbf{p}^2$ in the directive vector will decrease, yielding less conservative convergence criteria for a decreasing directive vector $\mathbf{p}$, than the potential alone would be.

For example, take a pendulum with a small damping. In the first swing, the pendulum swings beyond the minimum and the potential energy increases again. The increase in kinetic energy corresponds to an increase in the directive vector, which speeds up the convergence. Although the potential energy does not decrease monotonically, the total energy decreases monotonically all along the path. An adaptive damping will detect the minimum and increase the damping accordingly.
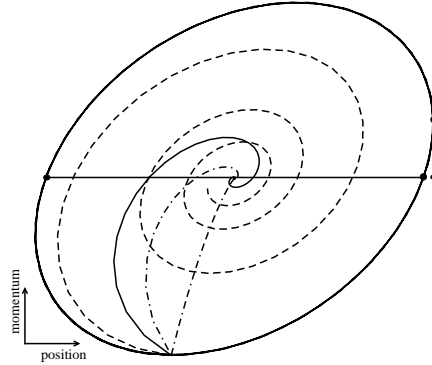
FIGURE 10.2. The trajectories in phase space for different damping. The straightest trajectory is not the fastest, as the damping is high. The ellipse indicate the area of constant energy, in which a dissipative system will remain. A particular horizontal cross section indicates the position. In special cases, the motion might take the system beyond that position, yielding a larger potential energy.

10.2. Adaptive damping and step size

Adaptive damping and step size can arise from tracking a succession of steps:

$$(10.12) \qquad \begin{pmatrix} \mathbf{q}(t_i) \\ \mathbf{p}(t_i) \end{pmatrix}, \begin{pmatrix} \mathbf{q}(t_{i+1}) \\ \mathbf{p}(t_{i+1}) \end{pmatrix}, \begin{pmatrix} \mathbf{q}(t_{i+2}) \\ \mathbf{p}(t_{i+2}) \end{pmatrix}, \ldots \quad .$$

Its basic properties are the Euclidean invariants, invariant under coordinate transformations, such as the length $d_{i+1}^i = |\mathbf{q}(t_{i+1}) - \mathbf{q}(t_i)|$ and the angle $\theta_{i+1}^i = \mathbf{q}(t_{i+1})^T \mathbf{q}(t_i)$.

We use a simple Euler step for the time integration:

$$(10.13) \qquad \mathbf{p}(t_i) = \mathbf{p}(t_{i-1}) - (\mathbf{A}\mathbf{q}(t_{i-1}) + \gamma\mathbf{p}(t_{i-1}) - b)\Delta t_{i-1} \quad .$$

A similar update step is used for $\mathbf{q}$:

$$(10.14) \qquad \mathbf{q}(t_i) = \mathbf{q}(t_{i-1}) + \mathbf{p}(t_i)\Delta t_i \quad .$$

The damping of $\gamma\Delta t_i = 1$, setting the momentum, or directive vector, $\mathbf{p} = 0$, corresponds to the steepest descent, or gradient, method, which is prone to numerical oscillations, [Booth, 1966, Press et al., 1992] between the sides of small valleys, so to say. At each step the directional, or momentum, vector $\mathbf{p}$ is set to zero. A very small damping, $\gamma \approx 0$, will lead to dynamical damping, where the solution oscillates around the minimum $\mathbf{x}$. Somewhere in the middle lies the optimal damping.

The optimal time step is the maximal time step which does not lead to numerical instabilities. For a well-conditioned problem, the norm of the matrix $\mathbf{A}$ is $\|\mathbf{A}\| \approx 1$, such that a convergent series arises if $\Delta t < 1$. With the presence of the directive momentum vector the iteration matrix $\mathbf{B}$ is:

$$(10.15) \qquad \mathbf{B} = \Delta t \begin{pmatrix} 0 & \mathbf{1} \\ -\mathbf{A} & -\gamma \end{pmatrix} \quad .$$

The eigenvectors of $\mathbf{B}$ have a large imaginary part. Furthermore, in the absence of damping $\gamma = 0$, the norm may be smaller than one, but the system will not

converge to a solution, but instead oscillate around the minimum. The stability analysis of the iterative series is therefore more complicated. One can also view the directive vector $\mathbf{p}$ as the weighted summation of successive steepest descent gradients:

$$(10.16) \qquad \|\mathbf{B}\| \approx \sum_j (1 - \gamma\Delta t)^j \|\mathbf{A}\| < \frac{\|\mathbf{A}\|}{\gamma\Delta t} \quad.$$

Hence to avoid erratic behavior, a conservative estimate of the step size Δt is based on the weighted summation:

$$(10.17) \qquad \Delta t < \frac{\|\mathbf{A}\|}{\gamma} \quad.$$

Not all methods developed for large, sparse systems take into account the successive nature of the matrix, arising mainly from FEM and network systems. In order to connect to coefficients of the vector $\mathbf{q}$ a large number n of $\mathbf{A}^n$ are required, typically of half the size of such systems, in the case of one-dimensional models. For higher dimensions of the underlying model, like plates and blocks, the connectivity is higher.

The criteria for tracking a solution in a FEM or network based model, yields a rather trivial result. At each step, the solution progresses a little further through the system, hence the main perpendicular part is the new component added first to $\mathbf{p}$, and then to $\mathbf{q}$. In a dynamical interpretation, it is the wave front of the hyperbolic equation propagating throughout the system. We find that for small damping the convergence of the solution starts to oscillate with the period related to the wave velocity and size of the system.

At every iteration $\mathbf{A}^n\mathbf{b} \to \mathbf{A}^{n+1}\mathbf{b}$, the solution extends only one or two mesh points further away from $\mathbf{b}$. Hence the larger the mesh, the more iterations are required only to fill the vector $\mathbf{x}$. Hence, typically, the number of iterations is proportional to n. This is the period for an optimal time-step. Therefore, the maximal step size $\Delta t \approx \|\mathbf{A}\|^{-1}$. The time step determines the degree of exploration of the solution space, bounded by the energy of the initial guess, while the damping causes the convergence. In practice we find little variation in the optimal damping of a large range of problems. Possibly, convergence can be improved by increasing the damping as the convergence is reached.

Adaptive schemes work best if there is a detailed understanding of the system at hand. The best possible scheme is a single iteration run, where the number of iterations equals the connectivity of the system. This could be considered the discrete equivalent of the direct integration of the inverse $\mathbf{A}^{-1}$, the discrete equivalent of the Green function. This would be the optimal convergence available for a given system. However, in this case the Green function is not known, but implicit in the matrix $\mathbf{A}$. For example, if the equation $\mathbf{A}\mathbf{x} = \mathbf{b}$ represents the displacement $\mathbf{x}$ as the result of the force $\mathbf{b}$, for the stiffness matrix $\mathbf{A}$, the linear equation is the internal force balance. In terms of the force, the system is usually better understood than in terms of the displacement, which depends on the constitutive relations.

Force equations can be integrated. Consider the iteration process as a long line of masses connected by springs. If one mass is pulled with a force $\mathbf{b}$, all the neighboring masses will counteract the force. This counteraction is again counteracted by the next-neighbor masses, and so on. The process described here corresponds to

the conjugate gradient algorithm described in the next section. At every stage the exact solution for in a subspace $\mathcal{K}_n$ for that particular set of variables is sought. The use of damping will yield a slower convergence for the same number of iterations, but less computations per iteration. The dynamical approach makes it possible to combine convergence in one particular subspace with motion in another. The stiffness matrix of an isolated object is singular. Methods based on the positivity of the stiffness matrix may fail for particular boundary conditions. An explicit reference to the dynamics and the force can resolve numerical problems whose origins lie in the problem definition. These principles are explored further in Chapter 12.

The stiffness matrix $\mathbf{A}$ corresponding to a string of masses and springs is singular. The same constant displacement of all the masses will not change the indentation of the springs and consequently will not change the force. The inhomogeneous term, the right-hand side of the equation, representing the applied force, should therefore be balanced; the sum force is zero. Hence the force is balanced locally from source to sink, from the position the force acts, to the position where it is counter-acted. In the case of a system response, the force may also be counter-acted by resistive forces, or by inertial forces caused be acceleration.

Hence the convergence and quality of the iterative solution can be expressed in terms of physical principles, involving local force balance underlying a static solution. The constant displacement $\mathbf{e} = (1, 1, 1, 1, \cdots, 1, 1, 1)$ plays an important role in this analysis, since for the static solution $\mathbf{e}^T \mathbf{x} = 0$. Hence deviations from the static solution during iteration, which we will call force "absorption," represent themselves as $\mathbf{e}^T \mathbf{x}_i$. In the dynamical interpretation of the iteration process, the absorption is related to the inertial and damping forces.

Part of the force is locally balanced, the other force lines run through the whole structure, till fixed boundaries. By the lack of knowledge of the end of the force lines the adaptive method should keep track of the force "transmission" versus the force "absorption." Given the solution $\mathbf{x}_{i+1}$ of a step in the iterative scheme:

(10.18) $$\mathbf{A}\mathbf{x}_i = \mathbf{x}_{i+1} \ .$$

A constant vector $\mathbf{e} = (1, 1, \cdots, 1, 1)$ corresponds to a complete transmission:

(10.19) $$\mathbf{A}\mathbf{e} = \mathbf{e}^T \mathbf{A} = 0 \ .$$

Hence the force absorption can be expressed as a projection on $\mathbf{e}$. The absorption $F_{\text{absorption}}$ is given by the projection on the translation vector:

(10.20) $$F_{\text{absorption}} = \frac{\mathbf{e}^T \mathbf{x}_{i+1}}{\sqrt{N} \|\mathbf{x}_{i+1}\|} \ ,$$

where N is the dimension of the system. The force balance and localization at the inhomogeneous term $\mathbf{b}$, the force input, is an independent test of the correctness and convergence of the iterative solution. In the cases the force is not balanced; the sum force is not zero, the momentum grows due to the acceleration of the system, and the force absorption is associated with the inertial force.

In the absence of knowledge of directions in the system, all the translations are concatenated into one; $\mathbf{e}$. Such knowledge, on the other hand, would make it possible to separate $\mathbf{e}$ into $\mathbf{e}_x$, $\mathbf{e}_y$, etc..

The force transmission T is given by the change from one iteration to the next:

(10.21) $$T = \frac{\|\mathbf{x}_{i+1} - \mathbf{A}\mathbf{x}_i\|}{\|\mathbf{x}_{i+1}\|} \ .$$

Typically, T represents the size of the wave front. The typical instabilities for a large time step occur if the action and counter-action remain local. Behind the leading wave front, filling the state vector successively following the actual connectivity of the underlying grid, the state converges locally, with an exponent determining the quality of the iterative scheme.

Problems described by potential like $V = \mathbf{q}^T \mathbf{A} \mathbf{q}$, where the matrix $\mathbf{A}$ is given by the Laplacian and flow problems with pressure boundary conditions are examples of systems with a high transmission.

10.3. Conjugate gradient method

The standard numerical method closely related to the dynamical approach is the conjugate gradient method. Here we give only a short overview of the method, which consists of two element: the residual vector $\mathbf{r}$, which is a measure for the error of the approximation so far, and its dual $\mathbf{v}$, which introduces a new direction perpendicular to all the old ones in a particular inner product. [Hestenes and Stiefel, 1952, Stoer and Bulirsch, 1980, Golub and Van Loan, 1996, Press et al., 1992, Saad, 1995, Antia, 2002] In this approach the positive definite matrix $\mathbf{A}$ is used as an inner product:

$$(10.22) \qquad \langle \mathbf{v}, \mathbf{x} \rangle_{\mathbf{A}} = \mathbf{v}^T \mathbf{A} \mathbf{x} \ .$$

The iteration procedure resembles a Gram-Schmidt orthogonalization with respect to the inner-product $\langle \cdot, \cdot \rangle_{\mathbf{A}}$. The approach is similar to the Lanczos tridiagonalization, using the Krylov space, described in section 2.12.

The solution $\mathbf{x}$ is expanded in residuals $\mathbf{r}_i$, each the remainder of the equation at that stage: $\mathbf{r}_i = \mathbf{A}\mathbf{x}_{i-1} - \mathbf{b}$, while the next residual vector $\mathbf{v}_i$ in the dual space is orthogonalized with respect to the previous duals $\mathbf{v}_0, \mathbf{v}_1, \cdots, \mathbf{v}_{i-1}$.

For each iteration $\mathbf{x}_i$ the smallest deviation of $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|$ in the subspace span$\{\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_{i-1}\}$ is sought. The residual part $\mathbf{r}_i = \mathbf{A}\mathbf{x}_i - \mathbf{b}$ of the vector $\mathbf{x}_i$ is projected onto the residual dual $\mathbf{v}_i$:

$$(10.23) \qquad \mathbf{v}_{i+1} = \mathbf{r}_i - \frac{\langle \mathbf{v}_i, \mathbf{r}_i \rangle_{\mathbf{A}}}{\langle \mathbf{v}_i, \mathbf{v}_i \rangle_{\mathbf{A}}} \mathbf{v}_i \ .$$

Therefore, two successive vectors are perpendicular: $\langle \mathbf{v}_{i+1}, \mathbf{v}_i \rangle_{\mathbf{A}} = 0$. The algorithm generates an orthogonal set of dual vectors, onto which the equation $\mathbf{A}\mathbf{x} - \mathbf{b}$ can be projected. During the generation of this set, the coefficients α_i of the expansion:

$$(10.24) \qquad \mathbf{x} = \sum_i \alpha_i \mathbf{v}_i = \sum_i \frac{\langle \mathbf{v}_i, \mathbf{r}_i \rangle_{\mathbf{A}}}{\langle \mathbf{v}_i, \mathbf{v}_i \rangle_{\mathbf{A}}} \mathbf{v}_i \ ,$$

can be generated simultaneously. It leads to a set of hierarchical equations, most economically expressed, after some calculations, [Stoer and Bulirsch, 1980] as:

$$(10.25) \qquad \mathbf{r}_{i+1} = \mathbf{r}_i - \frac{\mathbf{r}_i^T \mathbf{r}_i}{\langle \mathbf{v}_{i+1}, \mathbf{v}_{i+1} \rangle_{\mathbf{A}}} \mathbf{A} \mathbf{v}_{i+1} \ ,$$

while the same part is added to the solution vector:

$$(10.26) \qquad \mathbf{x}_{i+1} = \mathbf{x}_i + \frac{\mathbf{r}_i^T \mathbf{r}_i}{\langle \mathbf{v}_{i+1}, \mathbf{v}_{i+1} \rangle_{\mathbf{A}}} \mathbf{v}_{i+1} \ .$$

Furthermore, the iterative generation of the dual vectors can be expressed as:

$$(10.27) \qquad \mathbf{v}_{i+1} = \mathbf{r}_i + \frac{\mathbf{r}_i^T \mathbf{r}_i}{\mathbf{r}_{i-1}^T \mathbf{r}_{i-1}} \mathbf{v}_i \quad .$$

The algorithm resembles the leap-frog integration (see Section 3.4), where $\mathbf{r}_i$ and $\mathbf{v}_i$ are differences in steps in the position: $\Delta \mathbf{x} = \mathbf{x}_i - \mathbf{x}_{i-1}$ and momentum space: $\Delta \mathbf{p} = \mathbf{p}_i - \mathbf{p}_{i-1}$, respectively. Except for the precise value of the step size and damping, the pair of equations read:

$$(10.28) \qquad \mathbf{x}_{i+1} = \mathbf{x}_i + \alpha_i \mathbf{p}_i \quad ,$$

and

$$(10.29) \qquad \mathbf{p}_{i+1} = \mathbf{p}_i - \alpha_i (\mathbf{A}\mathbf{x}_i - \mathbf{b}) \quad .$$

Instead of relying for an exact computation of α_i, which allows for a result in the number of steps identical to the dimension of the matrix, we use some physical principles to estimate two different constants, one for the position space $\mathbf{x}$ and one for the momentum space $\mathbf{p}$. The number of operation per step decreases, but the convergence will decrease as well, since we did not choose the optimal α_i for the particular Krylov space $\mathcal{K}(\mathbf{A}, \mathbf{b})$.

10.4. Dissipative models

The criterion for success of an iterative method can be expressed in terms of the ever decreasing error norm. For a stationary method, using only the solution vector $\mathbf{x}$, the convergence to the solution is expressed as:

$$(10.30) \qquad \|\mathbf{A}\mathbf{x}_i - \mathbf{b}\| > \|\mathbf{A}\mathbf{x}_{i+1} - \mathbf{b}\| \quad .$$

For a symmetric positive definite matrix $\mathbf{A}$ another criterion would be:

$$(10.31) \qquad \frac{1}{2}\mathbf{x}_i^T \mathbf{A}\mathbf{x}_i - \mathbf{x}_i^T \mathbf{b} > \frac{1}{2}\mathbf{x}_{i+1}^T \mathbf{A}\mathbf{x}_{i+1} - \mathbf{x}_{i+1}^T \mathbf{b} \quad ,$$

since the problem can be reformulated using $\mathbf{b}' = \mathbf{A}^{-1}\mathbf{b}$ as the same statement in a weighted norm $\|\mathbf{x} - \mathbf{b}'\|_{\mathbf{A}}$. This norm is used in the conjugate gradient method.

In the dynamical case, the energy, or Hamiltonian, serves as norm for the convergence:

$$(10.32) \qquad H = \frac{1}{2}\mathbf{p}^T \mathbf{p} + \frac{1}{2}\mathbf{x}^T \mathbf{A}\mathbf{x} - \mathbf{x}^T \mathbf{b} \quad .$$

The Hamiltonian is bounded from below if $\mathbf{A}$ is a positive definite matrix. The quadratic term will always dominate the inhomogeneous, or linear, term for large $\|\mathbf{x}\|$. The damping $\gamma > 0$ will always decrease the energy in the system. The energy is dissipated. Lyapunov generalized this property of the dynamics of physical systems to a wider class of systems. [Lyapunov, 1992] It became an anchor for analyzing the stability of motion.

The choice of the step size should be such that the integration will not fail. Extreme accurate time-integration is not required, as the solution converges to a unique stationary value. However, the typical step size, limited by $\Delta t < \|\mathbf{A}\|^{-1}$, will change due to the presence of the momentum $\mathbf{p}$.

10.5. Diffusive-dynamics iteration using sum convergence

Another norm, based on physical systems, could be used to design an iteration scheme. In the case of diffusive processes, the dynamics converges to the stationary flow, given the inflow and outflow in terms of the inhomogeneous vector $\mathbf{b}$. The matrix $\mathbf{A}$ has, as for most partial differential equations, the null-space solution:

$$(10.33) \qquad \mathbf{A}[\text{constant vector}] = \mathbf{A}\mathbf{e} = 0 \quad .$$

For example, for a constant temperature, throughout the system, there will be no heat flow. Likewise, for a constant electric potential there will be no electric current flow. The presence of the sources $\mathbf{b}$ will change that situation, however, the greater part of the matrix yield zero for a constant vector.

If the iteration scheme above is applied to a problem with a null-space $\mathbf{e}$, care must be taken to keep the solution $\mathbf{x}$ perpendicular to the null-space. By reformulating the problem, associating it with a parabolic system rather than a Hamiltonian hyperbolic problem with damping, the numerical stability can be made implicit. The same static force solution can be treated as a conserved charge diffusion, with a distributed inhomogeneous term.

If the rows and columns are reordered such that the mesh points closer to the inflow are the first rows and columns, and the mesh points closer to the outflow are the last row and columns, one can expect the state vector $\mathbf{x}$ to be in an approximate ascending order. This will improve the convergence as we will see.

The symmetric matrix $\mathbf{A}$ can be decomposed into an upper-diagonal part $\mathbf{A}_u$, a lower-diagonal part $\mathbf{A}_l$, and a diagonal part $\mathbf{A}_d$, such that $\mathbf{A} = \mathbf{A}_u + \mathbf{A}_l + \mathbf{A}_d$. A triangular matrix, consisting of an upper or lower part and the diagonal part can be solved directly. This is the basis for some standard iteration schemes, such as Gauss-Seidel, [Press et al., 1992, Antia, 2002] which decompose the matrix in two parts, which can be solved directly:

$$(10.34) \qquad \mathbf{A} = \mathbf{A}_1 + \mathbf{A}_2 \quad ,$$

to use in an iteration scheme like:

$$(10.35) \qquad \mathbf{x}_{i+1} = \mathbf{A}_2^{-1}(\mathbf{b} - \mathbf{A}_1\mathbf{x}_i) \quad ,$$

where $\mathbf{A}_2$ should be invertible. As long as $\|\mathbf{A}_1\|\|\mathbf{A}_2^{-1}\| < 1$ the iteration should converge under general conditions. For Gauss-Seidel $\mathbf{A}_1 = \mathbf{A}_u$ or $\mathbf{A}_1 = \mathbf{A}_l$, and $\mathbf{A}_2 = \mathbf{A}_d + \mathbf{A}_l$ or $\mathbf{A}_2 = \mathbf{A}_d + \mathbf{A}_u$ respectively. For Jacobi iteration, $\mathbf{A}_1 = \mathbf{A}_u + \mathbf{A}_l$, and $\mathbf{A}_2 = \mathbf{A}_d$.

The iteration scheme can be seen as the Euler integration of the dynamical equation:

$$(10.36) \qquad \dot{\mathbf{x}} = \frac{(\mathbf{A}_2^{-1}\mathbf{A}_1 - 1)}{\Delta t}\mathbf{x} + \mathbf{A}_2^{-1}\mathbf{b} \quad .$$

If we use for $\mathbf{A}_1 = \mathbf{A}_u - \mathbf{A}_l$ and for $\mathbf{A}_2 = \mathbf{A}_d - 2\mathbf{A}_u$ the iteration matrix $\mathbf{A}_1$ is skew-symmetric. A skew-symmetric matrix corresponds to a gyroscopic process in dynamics. It redistributes $\mathbf{x}_i$, but leaves the sum of all the coefficients invariant. The values of the coefficients are like the charge density of the system, while the sum is the total conserved charge. Hence only the matrix $\mathbf{A}_2$ affects the charge. The matrix $\mathbf{A}_2$ has typically only negative coefficients, and determines the maximal step size. It would correspond for a one-dimensional second-order differential equation
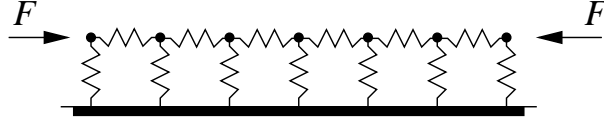
FIGURE 10.3. The one-dimensional Hertzian problem. The vertical springs, connected to the fixed world, have a stiffness α, while the horizontal spring have a unit stiffness.

to rewriting it into an integro-differential equation:

$$(10.37) \qquad \partial_\parallel^2 f(z) + Lf(z) = g(z) \rightarrow \partial_\parallel f(z) = \int_{\text{sink}}^{\text{source}} dz_\parallel (g(z) - Lf(z)) \ ,$$

where the parallel direction $z_\parallel$ is the direction from sink to source as defined by the sign of the inhomogeneous term g. See also Section 12.3. By rewriting the differential equation to an integro-differential equation the null-space of the stiffness matrix $\mathbf{A}$ no longer poses a numerical problem.

10.6. The string model

To illustrate and test the dynamical principles of iterative methods, we examine a string problem in detail. It is similar to an one-dimensional Hertzian contact problem. See Figure 10.3. Such a problem is typical for most stable systems, based on partial differential equations. Also Helmholtz problems yield similar results, for example, for the penetration of electromagnetic fields in a conductor. The parameter α determines the typical penetration depth. The parameter-dependent matrix problem is mapped back to differential equations with an analytical solution. The numerical problem can be studied without this analytic result, however, it yield more insight in the physical aspects of system and solution. For example, the force transmission will vary from complete transmission for $\alpha = 0$ to no transmission for large α. The choice of α, for the numerical example, is in the intermediate regime.

Given an elastic string stretched by an applied, opposite force $\mathbf{b}$, at both ends. The string is also attached to a frame, with a relative spring tension α. The stable solution corresponds to the equation $\mathbf{A}\mathbf{x} = \mathbf{b}$, where $\mathbf{x}$ is the strain, or displacement. The matrix equation is:

$$(10.38) \qquad \mathbf{A}\mathbf{x} = \begin{pmatrix} -1 & 1 & 0 & & & \\ 1 & -2 & 1 & 0 & & \\ 0 & 1 & -2 & 1 & 0 & \ddots \\ & 0 & 1 & -2 & 1 & \\ & & \ddots & \ddots & \ddots & \ddots \end{pmatrix} \mathbf{x} - \alpha\mathbf{x} = \begin{pmatrix} -1 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} = \mathbf{b} \ .$$

The minimum number of iterations required grows linearly with the size of the system, since the effects of the force at the ends has to propagate through the string one step at the time. Multigrid methods are developed to speed up this basic convergence problem, but we will consider a fixed size of the system.

The matrix problem is the discretized version of the differential equation:

$$(10.39) \qquad\qquad \frac{\partial^2 x(z)}{\partial z^2} - \alpha x(z) = 0 \ ,$$
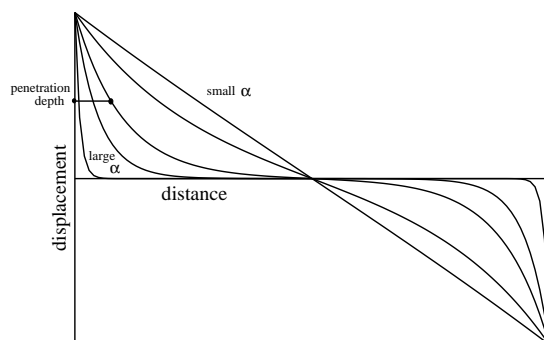
FIGURE 10.4. The normalized analytical results of the Hertzian problem for different values for α. For small values of α the force is "transmitted," from one end to the other, with a linear displacement. For large values of α the force penetrates only a small distance into the material.

with the boundary conditions corresponding to opposite forces at both ends ($\mathbf{b} = (-1, 0, 0, \cdots, 0, 0, 1)$):

$$(10.40) \qquad \left(\frac{\partial x(z)}{\partial z} - \alpha x(z) \right) \Bigg|_{z = \pm N/2} = \pm 1 \quad .$$

The solution to the continuous problem is (see also Figure 10.4):

$$(10.41) \qquad x(z) = \frac{\sinh \sqrt{\alpha} z}{\sqrt{\alpha} \cosh \sqrt{\alpha} N/2 - \alpha \sinh \sqrt{\alpha} N/2} \quad .$$

Due to the boundary conditions, responsible for the inhomogeneous term $\mathbf{b}$, the value at the boundary grows with each iteration. The gradient term propagates the effect inwards.

The step size is usually the largest step size possible for a stable iteration. Since the string problem turns into a wave equation problem with dispersion, by adding the directive vector $\mathbf{p}$, the step size should satisfy criteria that the time step should lie in the forward characteristic cone, determined by the characteristic equation, and such, by the norm of $\mathbf{A}$, the characteristic cone depends only slightly on the considerable damping, therefore the step size may be chosen fixed. These are typical results from the time integration of hyperbolic partial differential equations. [Lax and Wendroff, 1960, Morton and Mayers, 1994]

The damping, on the other hand, should vary with the problem and the stage in the iteration. At first, the damping should be small, such that the solution expands throughout the domain with the maximal, or wave, velocity. Later the damping should increase to damp out oscillations in the solution. A small damping is easily recognized in the solution by an oscillation in the error norm:

$$(10.42) \qquad \varepsilon_i = \| \mathbf{x}_{\text{exact}} - \mathbf{x}_{\text{iterative}}(t_i) \| \quad .$$

The smaller the parameter α the larger the penetration depth, and typically the slower the convergence. The optimal value of $\gamma \approx 0.07 - 0.11$ varies only slowly over the range of parameter α, only when the problem turns into a surface problem,
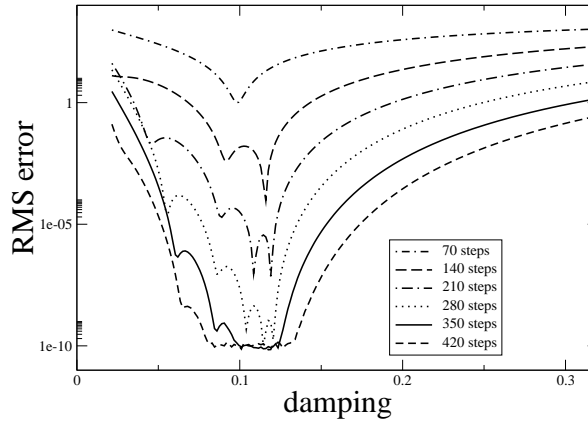
FIGURE 10.5. The convergence to the exact result after different numbers of steps, for a variety of constant damping. The constant $\alpha = 0.003$ is in the intermediate regime and the dimension is $N = 100$.

with a small penetration depth, the optimal damping may be substantially larger $\gamma \approx 0.25$. See Figure 10.5. Even smaller than the range of damping γ is the optimal variation of damping throughout the problem. Typically, the damping should increase as the iteration proceeds by 20% to 30%, to reach the solution the fastest. The size of the vector $\|\mathbf{p}\|$ is a good indication how far quick the convergence is, or how large the system is. The size will grow at first. Once the solution starts to converge, the size will decrease with the error. The typical relation between the parameter α, and the optimal damping is, for $\alpha < 0.02$:

$$(10.43) \qquad\qquad \gamma_{\text{optimal}} \approx 0.07 + 10\alpha \ .$$

As the number of iterations is several times the size of the problem, the domain of convergence is a large region around the optimal damping.

The damping is related to the convergence speed; the quicker the solution convergences, normally for larger values of α, the larger the damping should be. Very large values of $\alpha > 0.1$ yield numerical instabilities, as the penetration depth is in the order of the mesh size. Problems ill-defined by the coarseness of the grid will be less stable for large damping.

10.7. Constraints

Constraints in dynamical systems are usually nonlinear, since most linear relations can be implemented without problems, yielding a lower dimensional dynamical problem. Rigid-body dynamics is well-known for its constraints, which can either by holonomic or non-holonomic, a term coined by Hertz. [Gantmacher, 1970] Eventually we are interested in very large sets of constraints for FEM problems, but for the moment we will look at two simple problems to investigate the implementation of constraints as virtual dynamics. The first case is in the absent of real dynamics. The constraints determine the state almost uniquely for the Stewart platform, as
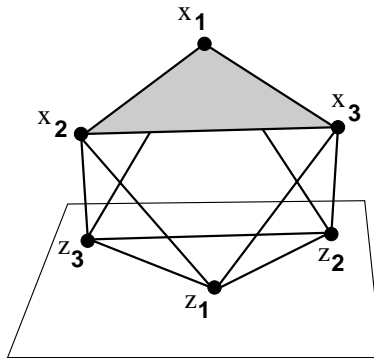
FIGURE 10.6. The Stewart platform, actuated by six trusses, connected to a fixed ground.

discrete local minima are found. In the mathematical pendulum (Section 10.10) we investigate the decoupling of the virtual, or constraint, dynamics and the real dynamics, and the influence of the virtual dynamics on the accuracy of the solution.

The generic methods described here are all local. Around the present position in phase space we examine the constraint direction and the direction of the dynamics. To first order, corresponding to local rectangular coordinates, we are able to decouple the constraint direction, in which the virtual constraint forces act, and the direction of dynamics, in which the system evolves. The nonlinear constraints do not always allow for an analytical estimate of the accuracy and decoupling, however, numerical results are very good.

Eventually these methods are meant to be applied to FEM systems with arbitrary boundary conditions. Many types of boundary conditions cannot be implemented as input, but must be implemented as constraints, as, for example, fixed positions at the boundary. It is possible to reduce the FEM model such that the fixed boundary values are excluded from the model. However, that means the resulting FEM model of the system can not be used again is a component of a larger system in a general setting. For a general setting constrained motion and the associated constraint force are the pair along which a system may interact with neighboring components, or control.

10.8. Stewart platform

In analytical mechanics Cartesian coordinates are traded in for a smaller set of generalized coordinates. For analytical calculations the smaller set of generalized coordinates had clear advantages. In a number of cases it allows for an exact, closed, solution of an isolated system. However, in many cases such a solution is not sought. The complexity of a combined system, consisting of many subsystems, already described in generalized coordinates, may be overwhelming.

In this chapter we advocated the approach of not performing any reduction, using the Cartesian coordinates, and the Euclidean invariants. Any constraint can be solved dynamically. Moreover, during simulation convergence to the constraint surface may be implemented simultaneously with forward simulation.

As a simple example we discuss the Stewart platform, consisting of a plate controlled by six extendable trusses, connected at each end with a ball joint. The

plate can be represented by the three joint positions, $\mathbf{x}_1, \mathbf{x}_2$, and $\mathbf{x}_3$. See Figure 10.6. The geometry and inertia of the plate depends solely of these coordinates. For example, the kinetic energy of the plate can be calculated by:

$$(10.44) \qquad T = \frac{1}{2} \int \rho \dot{\mathbf{x}}^2(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3) = \sum_{ij} M_{ij} \dot{\mathbf{x}}_i \dot{\mathbf{x}}_j \quad .$$

Two simple examples. First, consider the platform mass M located at the center: $\mathbf{x} = \frac{1}{3}(\mathbf{x}_1 + \mathbf{x}_2 + \mathbf{x}_3)$, the mass matrix is therefore:

$$(10.45) \qquad \mathbf{M} = \frac{M}{9} \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} \quad .$$

Secondly, consider the mass to be evenly distributed on a triangle spanned by the three corners $\mathbf{x}_i$, where the triangle has three equal lengths. The center-of-mass position $\mathbf{x}$ is given by the interpolation between the three points:

$$(10.46) \qquad \mathbf{x} = \int_0^{\sqrt{3}} du \int_{-u/\sqrt{3}}^{u/\sqrt{3}} dv \sum \mathbf{x}_i f_i(u, v) \quad ,$$

where the weight function,

$$(10.47) \qquad \mathbf{f}(u, v) = \begin{pmatrix} (1 - \frac{u}{\sqrt{3}}) \\ \frac{1}{2}(\frac{u}{\sqrt{3}} - v) \\ \frac{1}{2}(\frac{u}{\sqrt{3}} + v) \end{pmatrix} \quad ,$$

the kinetic energy is the integral over the distributed mass:

$$(10.48) \qquad \int \frac{1}{2} \rho \dot{\mathbf{x}}^2 = \sum_{ij} \dot{\mathbf{x}}_i \dot{x}_j \int \frac{1}{2} \rho f_i(u, v) f_j(u, v) \quad .$$

The weight functions $f_i(u, v)$ determine the mass-matrix coefficients, independent of orientation or velocity. The weight functions satisfy:

$$(10.49) \qquad f_1(u, v) + f_2(u, v) + f_3(u, v) = 1 \quad ,$$

such that a collective translation is correctly accounted for. The total area of the triangle is $\sqrt{3}$. The mass matrix is ($\sqrt{3}\rho = M$):

$$(10.50) \qquad \mathbf{M} = \frac{M}{12} \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix} \quad .$$

Since the collective motion of all the nodes is associated with the total mass, the elements of the mass matrix should always sum to the total mass M.

The platform itself can be represented by three stiff trusses $V(\mathbf{x}_i - \mathbf{x}_j)$ between each pair of corners. The length at rest is normalized to 2, the same as in the case of the weight functions $f_i(u, v)$ above. The trusses between the platform and the floor have variable lengths l_{ij}, where the truss labelled ij is located at platform corner $\mathbf{x}_i$ and floor corner $\mathbf{z}_j$, such that $i \neq j$.

For all of these trusses we assume a simple elastic energy function:

$$(10.51) \qquad V_l(\mathbf{x} - \mathbf{y}) = \frac{1}{2} k((\mathbf{x} - \mathbf{y})^2 - l^2)^2 \quad .$$

Hence, using the Hamiltonian consisting of all V_l's, the static configuration is the minimal elastic energy. In the case the elastic energy is zero, the configuration is
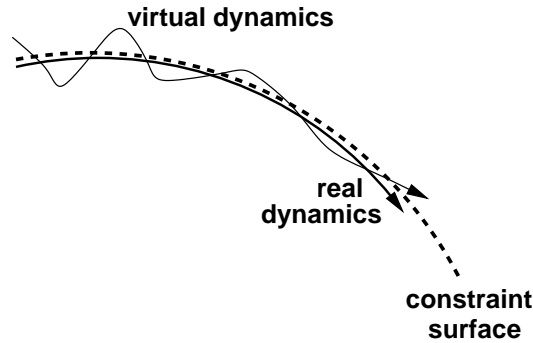
FIGURE 10.7. The virtual dynamics with the appropriate damping converges to the real dynamics, which lies in the constraint surface.

allowed, i.e., all the constraints can be satisfied simultaneously. The mass can be taken into account, even masses of each of the trusses can be added simply with the mass matrix:

$$(10.52) \qquad \mathbf{M}_{\text{truss}} = M_{\text{truss}} \frac{1}{6} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \quad .$$

In order to converge to the static solution, the stiffness must be much larger than the body forces present, for example, the gravity force. Furthermore, in a dynamical context damping can be added which affects only the trusses. Using the standard Rayleigh terms:

$$(10.53) \qquad R(\dot{\mathbf{x}} - \dot{\mathbf{y}}) = \frac{1}{2}\gamma(\dot{\mathbf{x}} - \dot{\mathbf{y}})^2 \quad ,$$

which yields only damping along the trusses. The motion of the truss as a whole is not affected by the damping.

However, in cases with real dynamics and virtual dynamics from constraints, the explicit configuration dependence should be put into the Rayleigh term to guarantee the orthogonality of the damping to the real dynamics:

$$(10.54) \qquad R(\dot{\mathbf{x}} - \dot{\mathbf{y}}, \mathbf{x} - \mathbf{y}) = \frac{\gamma}{2l^2}((\dot{\mathbf{x}} - \dot{\mathbf{y}})^T(\mathbf{x} - \mathbf{y}))^2 \quad .$$

Now the damping is explicitly in the direction of the truss. The damping no longer relies on a small relative velocity $\dot{\mathbf{x}} - \dot{\mathbf{y}}$, which yield damping only in the direction of the constraint due to the local orthogonality of virtual constraint dynamics and real dynamics. This principle can be extended to more general constraints, as we will see in a later section. See Figure 10.7.

The truss Hamiltonian with the truss Rayleigh term do not necessarily have to be virtual terms, with very high stiffness, to yield iterative equations, which converge to the constraint surface. In reality there are no pure rigid elements, all elements have some elasticity, the truss elasticity V_l may be used to examine the vibrational modes of the Stewart platform.

10.9. Virtual dynamics

The Cartesian Hamiltonian dynamics has only constant mass matrices. In the symmetric case, under exchange of platform nodes, the mass matrices even have a
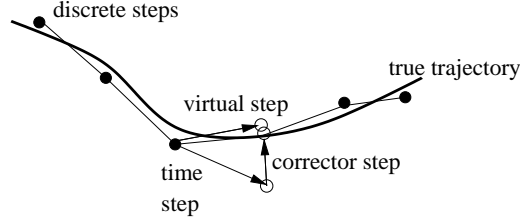
FIGURE 10.8. The virtual dynamics combines the free dynamics
and the constraint in a single integration step. The corrector step
is formulated as damped, or dissipative, dynamics.

very simple form, which is easily inverted:

$$(10.55) \qquad \mathbf{M}^{-1} = \begin{pmatrix} \alpha + \beta & \beta & \beta \\ \beta & \alpha + \beta & \beta \\ \beta & \beta & \alpha + \beta \end{pmatrix}^{-1} .$$

$$(10.56) \qquad = \begin{pmatrix} \alpha^{-1} - b & -b & -b \\ -b & \alpha^{-1} - b & -b \\ -b & -b & \alpha^{-1} - b \end{pmatrix} ,$$

where $b = \beta/(\alpha(\alpha + 3\beta))$. The same matrix occurs for arbitrary dimensions $M_{\alpha,\beta}$, such as in the case of a truss where the matrix is two-dimensional. The inverse is has the same form, except the only the 3 is replaced by the dimension n of the matrix in $b = \beta/(\alpha(\alpha + n\beta))$. Hence in the case of the triangular homogeneous platform the inverse of the mass matrix is:

$$(10.57) \qquad \mathbf{M}^{-1} = \frac{3}{M} \begin{pmatrix} 3 & -1 & -1 \\ -1 & 3 & 1 \\ -1 & -1 & 3 \end{pmatrix} .$$

The equations of motion for the position and momentum vector concatenated from all positions and momenta of each point are given by:

$$(10.58) \qquad \begin{pmatrix} \dot{\mathbf{x}} \\ \dot{\mathbf{p}} \end{pmatrix} = \begin{pmatrix} \mathbf{M}^{-1}\mathbf{p} \\ -\mathbf{MRM}^{-1}\mathbf{p} - \partial_{\mathbf{x}}V \end{pmatrix} ,$$

where the Rayleigh matrix $\mathbf{R}$ is constructed from the Rayleigh term R through:

$$(10.59) \qquad \frac{1}{2}\dot{\mathbf{x}}^T\mathbf{R}\dot{\mathbf{x}} = \frac{1}{2}\sum_{ij} R_{ij}\dot{\mathbf{x}}_i\dot{\mathbf{x}}_j = R .$$

The position-actuated Stewart platform has no real dynamics. The orientation of the platform is fixed by the length of all six trusses l_{ij}, fixing the six degrees of freedom of the platform. Hence there is only virtual dynamics; the convergence to this solution, using the truss energies V_l and damping R.

In the case of force actuation, the truss lengths are the result of the system response to the applied forces. There are at least two approaches to this problem. First, use the configuration information to align the forces parallel to the trusses, and add the corresponding applied resultant force to the Hamilton equations. Second, use the truss length l_{ij} parameter as variable: Rather than assuming a rigid

truss with a high stiffness k, the truss can be a spring with a stiffness corresponding to the length and actuation force:

$$(10.60) \qquad k = \frac{F_{\text{actuator}}}{l} \quad ,$$

where the elastic energy of the truss is:

$$(10.61) \qquad E_{\text{elastic}} = \frac{1}{2}kl^2 \quad .$$

The Hamiltonian can be updated in many different ways, to account for the actuation.

In this study we will take the approach of updating the length, of the stiff truss, according to the force. See Figure 10.8. Given $\mathbf{x}$ and $\mathbf{y}$, the truss force is :

$$(10.62) \qquad F = -\partial_{\mathbf{x}} V_l(\mathbf{x} - \mathbf{y}) = -2k((\mathbf{x} - \mathbf{y})^2 - l^2)|\mathbf{x} - \mathbf{y}| \quad .$$

Hence, by adjusting l, the force F can be set to the actuator force:

$$(10.63) \qquad l^2 = \frac{F}{2k|\mathbf{x} - \mathbf{y}|} + (\mathbf{x} - \mathbf{y})^2 \quad .$$

At each iteration step the truss lengths should be updated. Force actuation does not guarantee stability, also not in the updating case. The power transfer is given by the change of the elastic-truss energy. For stiff trusses the actual length change is limited. However, the consequences for the dynamics of the platform may be large, due to the large forces.

10.10. Pendulum example

In order to test the virtual dynamics, we consider the mathematical pendulum. It consists of point mass in gravity rotating at a fixed length around the origin:

$$(10.64) \qquad \left(\begin{array}{c} \dot{\mathbf{x}} \\ \dot{\mathbf{p}} \end{array} \right) = \left(\begin{array}{c} \mathbf{p} \\ -\mathbf{e}_y \end{array} \right) \quad ,$$

with the constraint $|\mathbf{x}| = 1$. The mass, gravity, and length are all set to unity.

The constraint can be eliminated to reduce the order of the system, retaining a single angle θ:

$$(10.65) \qquad \ddot{\theta} = -\sin\theta \quad ,$$

which is an anharmonic oscillator.

In the case of virtual dynamics, the truss (Eq. 10.51) connecting the mass to the origin is taken to have a large stiffness k and a damping γ:

$$(10.66) \qquad \left(\begin{array}{c} \dot{\mathbf{x}} \\ \dot{\mathbf{p}} \end{array} \right) = \left(\begin{array}{c} \mathbf{p} \\ -\gamma(\mathbf{p}^T\mathbf{x})\mathbf{x} - \mathbf{e}_y - k(\mathbf{x}^2 - 1)\mathbf{x} \end{array} \right) \quad .$$

The elastic and damping forces due to the truss all act in the direction of the truss $\mathbf{x}$. In a decomposition of the coordinates in an angular and a radial coordinate can be achieved by projecting the equations onto $\mathbf{x}$ and $\mathbf{e}_z \times \mathbf{x}$. The equation of motion for the angular coordinate θ is the same as above, except for the presence of the length $\mathbf{x}^2$ in the force term. The radial coordinate $r = \sqrt{\mathbf{x}^2}$ yields a damped nonlinear oscillator around $r = 1$:

$$(10.67) \qquad \left(\begin{array}{c} \dot{r} \\ \dot{p} \end{array} \right) = \left(\begin{array}{c} p \\ -p\gamma r^2 + \cos\theta - k(r^2 - 1)r + r\dot{\theta}^2 \end{array} \right) \quad ,$$
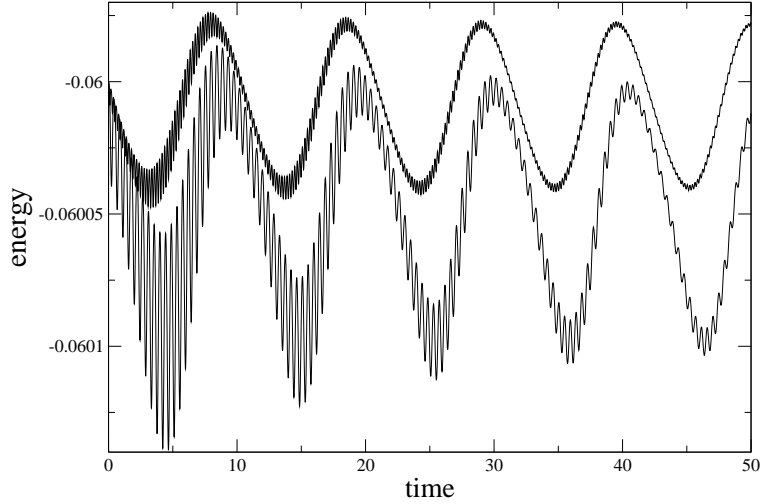
FIGURE 10.9. The energy of the mathematical pendulum, as a
constrained two-dimensional system. The larger deviations are for
the smaller stiffness of $k = 100$, the larger stiffness is $k = 500$.
The damping is small, $\gamma = 0.3$, the slightly larger damping $\gamma = 1$
already damps out the oscillations. The larger oscillations are the
result of the swing. If the truss energy is taken into account, the
energy is constant, up to integration errors.

where $\cos\theta$ is the gravity component in the direction of the truss, and $r\dot\theta^2$ is the
centrifugal force.

In the static case $p = \dot\theta = 0$, the truss is extended due to the gravitational
force:

$$(10.68) \qquad\qquad\qquad r = 1 + \frac{1}{2k} \quad .$$

In the dynamic case, the truss dynamics is driven by the pendulum swing for small
amplitudes θ_0 where: $\theta(t) = \theta_0 \sin t$. The force on the constrained, or virtual,
dynamics is now time-dependent, with a maximum at $\theta = 0$:

$$(10.69) \qquad\qquad \ddot r = -\dot r \gamma r^2 - k(r^2 - 1)r + \cos\theta + r\theta_0^2 \cos^2 t \quad .$$

The driving force will cause damping of the pendulum motion, due to the damping
forces in the constraint dynamics. However, this damping is small after the original
motion around the constraint is damped, and only the driving force remains. The
time-dependence of the driving force, the pendulum frequency, is much slower than
the frequency of the truss. Furthermore, the amplitude is small for large virtual
stiffness of the truss. Hence, the combination of the small amplitude and the slow
times reduces the virtual damping. Numerical investigations show that the energy
loss of the pendulum undergoes a sharp transition for stiffnesses of the order of
$k \approx 150$. For a stiffness of $k = 100$ the pendulum looses about a fraction of

3.0 10^{-6} of its energy per swing. For $k = 200$ the losses are within the numerical precision of the leap-frog integration method (10^{-10}). The integration has been carried out for tens of thousands of pendulum periods. Further improvements are possible as we will see. See Figure 10.9.

10.11. Mixing real and virtual parameters

In the case of the Stewart platform, in the absence of real dynamics, the masses have little meaning. The constraints determine the state fully. Inertial forces are therefore irrelevant. However, in order to yield the fastest convergence either the time-step, or the stiffness, should be adapted, in conjunction with the physical mass. For the iterative methods, typically the time-step is adapted, to match the inverse norm of the **A**-matrix. In physical systems, the mass and the time-step are given by the phenomena under investigation. The stiffness and the damping of the virtual dynamics should be made as large as possible to yield the fastest convergence to the constraint surface.

The virtual stiffness alone does not determine the forces, the expected displacements play also a role. For nonlinear virtual dynamics the distance to the origin would be a reasonable measure for the size of the displacements. The degree of connectivity, for example the two trusses at each joint of the platform, also yield a factor in the bound on the forces. In the case of a discretized PDE, the dimension plays the role of connectivity. A typical conservative estimate, based on numerical experiments, for the maximal stiffness would be:

$$(10.70) \qquad\qquad k \approx \frac{0.1m}{(\Delta t)^2} \quad,$$

where Δt is the step size, and m the total mass of the subsystem in which the stiffness occurs. The damping may be a little larger:

$$(10.71) \qquad\qquad \gamma \approx \frac{0.15m}{(\Delta t)^2} \quad.$$

One would like the state to take half to a quarter step toward the minimum in a single time-step. The typical highest frequency of a harmonic oscillator would be:

$$(10.72) \qquad\qquad \omega_{\max} = \frac{2\pi}{T_{\min}} \approx \sqrt{\frac{nk}{m}} \quad,$$

where n is the connectivity; the number of springs attached to the mass m, and $T_{\min}$ the smallest period.

However, such a large stiffnesses rely on the stability of the integration method. In many cases, the generated forces yield disproportionally large intermediate velocities, with respect to the relevant distances. The problem is therefore ill-conditioned, with large stiffnesses and normal masses. For a generic integration method, the stiffnesses are better approximated reciprocal to the step size:

$$(10.73) \qquad\qquad k \approx \frac{0.1m}{\Delta t} \quad,$$

and a similar approximation for the damping. In the case of virtual dynamics without real dynamics, the mass serves only the convergence. It can also be scaled by the step size, to yield a well-conditioned problem.

10.12. General constraints

The example of the mathematical pendulum (Section 10.10) showed that the constraint dynamics and the real dynamics is not fully decoupled. For particular cases, with analytic results, a full decoupling might be feasible. We, however, focus our attention on schemes that can be implemented numerically for general constraints.

Given the constraint surfaces, the virtual dynamics, i.e., the dynamics perpendicular to the constraint surface can be minimized. The potential dependence on these perpendicular direction are removed by cross terms of real and virtual forces. The effects of the curvature of the constraint surface, leading to typical virtual inertial forces are more difficult to compensate. However, the curvature of the constraint surface may be used to adapt the perpendicular mass, associated to inertial forces perpendicular to the constraint surface, in the system. The inertia associated with the virtual dynamics can be used to minimize the perpendicular motion.

Given the Hamiltonian $H_0(\mathbf{q}, \mathbf{p})$ where $\mathbf{q}, \mathbf{p} \in \mathbb{R}^n$ and the constraints $\chi_j(\mathbf{q}) = 0$ for $j \in \{1, 2, 3, \cdots, r\}$. The augmented Hamiltonian is:

$$(10.74) \qquad H_0(\mathbf{q}, \mathbf{p}) \to H = H_0 + H_c = H_0(\mathbf{q}, \mathbf{p}) + \sum_{j=1}^{r} \frac{k_j}{2} \chi_j(\mathbf{q})^2 \quad .$$

Each constraint is added as a spring with the rest position $\chi_j = 0$. The spring stiffness k_j is to be optimized for the simulation.

The augmented Rayleigh term is given by:

$$(10.75) \qquad R = \frac{\gamma}{2} \sum_j (\dot{\mathbf{q}}^T \nabla \chi_j)^2 = \frac{\gamma}{2} \sum_j (\delta_{\mathbf{p}} H^T \nabla \chi_j)^2 \quad .$$

In the Hamiltonian formulation the Rayleigh term must be expressed in terms of the momenta. The velocity is given as the variational derivative of the Hamiltonian. The equations of motion are the Hamilton equations of motion with the damping and the applied forces on the right-hand side:

$$(10.76) \qquad \begin{pmatrix} \dot{\mathbf{q}} \\ \dot{\mathbf{p}} \end{pmatrix} - \begin{pmatrix} \delta_{\mathbf{p}} H \\ -\delta_{\mathbf{q}} H \end{pmatrix} = \begin{pmatrix} 0 \\ -\delta_{\mathbf{p}} R + \mathbf{F}_{\text{applied}} \end{pmatrix} \quad .$$

The typical deviation from the constraint surface $\chi_j(\mathbf{q}) = 0$ is given by the virtual force; the force perpendicular to the constraint surface. The direction perpendicular to the constraint surface is given by the gradient of the constraint function $\nabla \chi_j$:

$$(10.77) \qquad \mathbf{F}_{\text{constraint}} = -\delta_{\mathbf{q}} H_c = \sum_j \frac{\nabla \chi_j}{|\nabla \chi_j|^2} \nabla \chi_j^T (\delta_{\mathbf{q}} H_0 + \dot{\mathbf{p}} - \mathbf{F}_{\text{applied}}) \quad .$$

The first term in between brackets can be removed by adding a gradient constraint term H_g to the complete Hamiltonian H. This is the potential-energy dependence of the original Hamiltonian H_0 perpendicular to the constraint surface. The correction term H_g to be added to the Hamiltonian compensates the virtual potential-energy dependence:

$$(10.78) \qquad H_g = -\sum_j k_j \chi_j \frac{(\nabla \chi_j)^T \delta_{\mathbf{q}} H_0}{|\nabla \chi_j|^2} \quad .$$
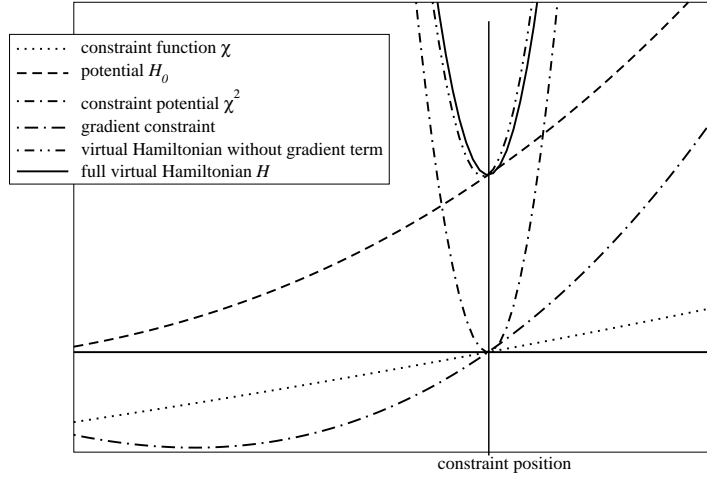
FIGURE 10.10. The potential perpendicular to the constraint direction, with the different additions to the constraint Hamiltonian.

.

Furthermore, the applied force $\mathbf{F}_{\text{applied}}$ is easily projected perpendicular to the constraint surface:

$$(10.79) \qquad \mathbf{F}_{\text{applied}} \to \mathbf{F}_{\text{applied}} - \sum_j \nabla \chi_j \frac{(\nabla \chi_j)^T \mathbf{F}_{\text{applied}}}{|\nabla \chi_j|^2} \quad .$$

See Figure 10.10.

Only the inertial constraint force $(\nabla \chi_j)^T \dot{\mathbf{p}}$ is not that easily compensated for, standing in the way a complete decoupling of real and virtual dynamics. The inertial constraint force is due to the curvature of the constraint surface. The curvature causes the momenta to change direction. If the mass can be compensated for in the direction of the constraint $\nabla \chi_j$, the inertial constraint force will also vanish. However, such terms may give rise to instabilities, since massless objects tend to oscillate fast.

10.13. Remarks on lumped to finite

In this chapter we treated lumped objects, such as trusses and point masses, in nonlinear dynamical models. However, the methods developed here to study dynamics can be applied to large nonlinear models. The construction of such finite elements is investigated in the next chapter. The nonlinearity arises through the consistency. The appropriate degrees of freedom are assigned to the internal dynamics and the general boundary conditions. Whether these boundary conditions

give rise to dynamics or constraint relations depends on the general setting. Eventually, we known that the simulation of such models can be carried out with the methods described in this chapter.

CHAPTER 11

Nonlinear FEM

Don't let them sell you a Turing machine without the tape.
[D. van Dalen, Filosofische grondslagen van de wiskunde]

Nonlinear FEM [Oden, 1972, Hughes, 1987] arises from two sources. First, the constitutive relations might be nonlinear, as is often the case, e.g., in thermodynamical and biological systems. In many other cases a linear constitutive relation is used, valid or not, since it simplifies the methods to solve such problems explicitly. In thermodynamical and biological systems these linear relations are known to fail, For example, the pressure is approximately inversely proportional to the volume, rather than linear. The large molecules in biology do not only have particular chemical characteristics but also rheological and mechanical characteristics.

Secondly, the states considered may be far from the equilibrium state, or rest-state, such that the linear theory, appropriate around the rest state, no longer suffices. Beam models are well-known examples of the latter. The linear Euler-Bernoulli equation (Eq. 6.5) yield correct results only for small deviations, such that even the far end of the beam, with the largest deviation still points approximately in the original direction.

A proper nonlinear model of a beam would take each segment or part independently, and yield a result independent of the overall orientation. As a component, a model can be part of a moving or rotating machine. These global motions will cause inertial effects which must be described with the center-of-mass and moments-of-inertia coordinates, which are not a priori given in a flexible body. The overall orientation can be defined as the collective motion which will not affect the elastic, or internal energy.

The orientation can be given by a rotation $\mathbf{R}$ and a translation $\mathbf{a}$. A proper invariant model of a beam with the position $\mathbf{x}(s)$ as function of the reference coordinate s would have an elastic energy independent of the orientation, that is:

$$(11.1) \qquad E_{\text{elastic}}(\mathbf{x}(s)) = E_{\text{elastic}}(\mathbf{R}\mathbf{x}(s) + \mathbf{a}) \ ,$$

for all rotations $\mathbf{R}$ and translations $\mathbf{a}$.

The corresponding model would necessarily be nonlinear, if the orientation needs to be accounted for too. The change of orientation $(\dot{\mathbf{R}}, \dot{\mathbf{a}})$ is relevant for the kinetic energy of the beam. A finite-dimensional, global part $\mathbf{F}$ of the distributed force $\mathbf{f}(s)$; a finite number of components associated with the center-of-mass and the moments of inertia, will only change the momentum and the angular momentum of the beam.

The potential energy, such as elastic energy, is given by terms in the Hamiltonian which contain partial differential operators. Inherently in these equations are the typical independence of translation, constant changes, and other invariances,
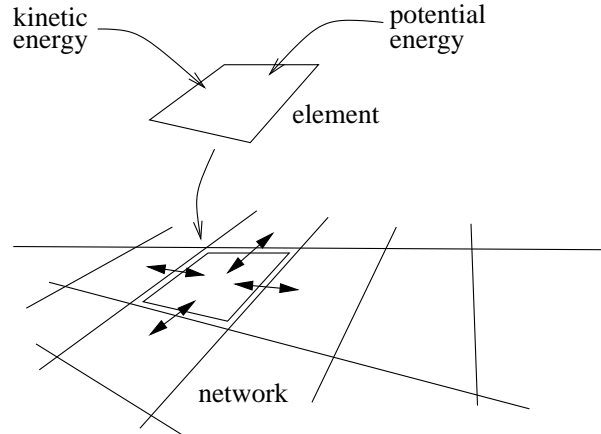
FIGURE 11.1. Several element Hamiltonians make up one for one element, which is combined in a network by variable collocation.

which give rise to constants of motion, and conserved quantities. The finite element approximation should retain these qualities as part of the consistency of the FEM approximation of the system. The most relevant conserved quantity is the energy, as it will yield a stable dynamical FEM model, even for the most abstruse nonlinearities.

In most cases the configuration state is the start of the analysis. In the case of systems described by partial differential equations it is difficult to escape this starting point as it closely ties the potential energy to the geometry of the rest state, or reference state, i.e., the configuration. This is not necessarily an unwanted starting point, as the kinetic energy is also closely tied to the configuration.

Combining the kinetic energy and the potential energy into a single Hamiltonian for an element gives rise to the port-Hamiltonian, due to the fact that the kinetic energy depends on global orientation and the potential energy does not. See Figure 11.1. The momentum and angular momentum are constants of motion for an isolated model, but as a component for a larger system, the states associated with the constants of motion are to be reassigned input via port variables. See Section 7.8. The kinetic energy depends on the change of configuration. Hence the typical variables involved in the potential energy and the kinetic energy are related by a differential operator D. The null-space of this operator is to be augmented with port variables.

For most physical systems similar properties arise, although the relation with the configuration is less stringent. For example charge conservation in electromagnetic systems is a conservation law, which leads to a total divergence as equation of motion. The boundary conditions of the partial differential equation at the basis of the potential energy are the missing variables. These variables are fixed by either input-output considerations, or by connecting elements together in a larger grid and thus reducing the number of variables to a well-defined set.

For rigid bodies the center-of-mass and the moments-of-inertia are static inertia relations between the overall orientation and the associated kinetic energy. For a deformable body, with elastic energy, this is no longer the case. The shape changes with the motion, and causes interaction between the global motion and internal

motion. The boundary conditions, causing both global motion and deformation are therefore not easily decomposed.

From a heuristic viewpoint it is much easier to think of the boundary conditions as global conditions, which restrict the changes in orientation, $(\dot{\mathbf{R}}, \dot{\mathbf{a}})$ in Eq. 11.1, or the changes of free motion associated with momentum and angular momentum. However, such boundary conditions are distributed forces $\mathbf{F}_\parallel(z)$, changing the center-of-mass velocity $\dot{\mathbf{a}}$ and torques τ, with a change in rotation $\dot{\mathbf{R}}$ as response, which are singled out from the boundary forces $\mathbf{F}(z)$ through projections as described in Section 7.2:

$$(11.2) \qquad \mathbf{F}_\parallel(z) = \rho(z)\frac{\int dz' \mathbf{F}(z')}{\int dz' \rho(z')} \ .$$

The torques are also part of the distributed force, $\mathbf{F}(z)$. The torque τ which only changes the overall orientation $\mathbf{R}$ around the rotation axis $\mathbf{n}$, without internal strain, is given by:

$$(11.3) \qquad \tau = (\rho \mathbf{z} \times \mathbf{n}) \left(\int dz'' \mathbf{z}'' \otimes \mathbf{z}'' \rho(z'') \right)^{-1} \int dz' \mathbf{z}' \times \mathbf{F}(z') \ ,$$

where $\mathbf{n}$ is the rotation axis at $\mathbf{z} = 0$, and the inverse of the middle integral is the variable inverse of the moment-of-inertia matrix $\mathbf{J}$:

$$(11.4) \qquad \mathbf{J} = \int dz \mathbf{z} \otimes \mathbf{z} \rho(z) \ .$$

A strained object will deform and the density $\rho(z)$ will vary with the force. Hence the projections do not only depend on the applied force $\mathbf{F}$, but on the object response as well. Lumping together, or projecting, the distributed boundary force $\mathbf{F}(z)$, to a single, but distributed force $\mathbf{F}_\parallel(z)$ and torque $\tau(z)$ these forces are associated with the inertia, from the mass distribution, $\rho(z)$, of the object. Together with the change in orientation, a combination of the translation velocity $\dot{\mathbf{a}}$ and the rotation velocity $\dot{\mathbf{R}}$, the center-of-mass force $\mathbf{F}_\parallel$ and the center-of-mass torque τ, form the typical power-bond pairs. [Paynter, 1961]

If the boundary force is separated in multiple lumped parts, additional lumped variables, besides the center-of-mass variables, typically associated with inertia and elasticity may arise. In the transmission line, and the beam models, one typically thinks of end-point variables, where even combinations correspond to some mean, or global, degree of freedom, similar to the center-of-mass, and the odd combination is some internal degree of freedom, typically associated with its dominant aspect; the elasticity or capacitance, although it also has inertial properties being a continuous system.

This chapter builds upon earlier work. [Synge, 1957, Lanczos, 1961] In the early days of numerical computing attempts to link finite-dimensional representations of function spaces and the operators thereon to the infinite-dimensional physical spaces and the differential operators played a larger role than they do nowadays. Representations of operators, other than the ones that appear in the defining functional, say the energy norm, are not considered essential. However, physical principles, such as conservation laws, are written down as auxiliary differential operators, like the finite-dimensional representation of flow or charge. Hence to construct together with the finite-dimensional representation of the problem at hand, the underlying

differential operators, will give further insight in the physics, and may give one a
fighting chance to preserve physical principles.

Conservation laws should also be valid for the discretized models. For example,
take a diffusive problem and its discretization:

$$(11.5) \qquad \dot{f}(z) = \Delta f(z) \quad \Rightarrow \quad \dot{\mathbf{x}} = \mathbf{A}\mathbf{x} \quad .$$

The formulation of the diffusive problem arises from some conservation law. Many
physical chemists [Bird et al., 1960, Sandler, 1977] dealing with such systems will
identify a conservation law, such as the continuity equation:

$$(11.6) \qquad \dot{f}(z) = -\nabla \cdot \mathbf{J}(z) \quad ,$$

and a closure relation of the flow $\mathbf{J}$ in terms of the state f:

$$(11.7) \qquad \mathbf{J} = -\mathbf{D}(\nabla \pi(f)) \quad ,$$

where $\pi(f)$ is the pressure function, and $\mathbf{D}(\cdot)$ the diffusion function, related to
the free path length or a particle that diffuses. Only if the discretization $\mathbf{A}$ of the
Laplacian Δ is constructed from the divergence $\nabla\cdot$ and the gradient operator ∇, it is
easy to reconstruct the conservation law, eq. 11.6, with the diffusion equation 11.5,
also for the case of a nonlinear pressure function or a non-isotropic geometry with
a tensorial $\mathbf{D}$.

Hence, we wish to construct consistent finite-dimensional representations of the
differential operators, and not just for the diffusive example above, that appear in
the models of physical systems and preserve from the start as much of their proper-
ties as possible, such that physical principles are retained in the finite-dimensional
systems.

In this chapter we will start with the spline approximation for each element of
a beam, showing that it suffers from two problems. First, the elastic energy of the
beam is not only given by the energy of the segment, but the boundary between
two segments may yield an infinite contribution to the energy, which can only
be solved by introducing constraints across the segments. Secondly, the standard
spline approximation corresponds to the Euler-Bernoulli beam model, which uses
explicitly a coordinate axis as rest position, with perpendicular displacements. This
is an approach which is not easily extended to an orientation-free description. In the
next section we show how these two problems can be cured, using invariant lines.
The boundary energy is zero and the segment energy is invariant under rotations
and translations.

In the following section we consider explicitly the partial differential operator,
the curl operator $\nabla\times$, and a polynomial approximation of the continuous states on
a three-dimensional cube. The counting of variables, boundary conditions, and the
null-space are the basis of constructing a consistent approximation of the continuous
model. The port-Hamiltonians elements of a triangular plate and electrodynamics
on the simplex are constructed.

11.1. The Euler-Bernoulli spline revisited

Before considering general element Hamiltonians, and the construction of the
port Hamiltonians we study the simple example of the Euler-Bernoulli beam and its
associated approximation in terms of piece-wise cubic splines: ($z \in [z_c - \delta, z_c + \delta] \subset \mathbb{R}$)

$$(11.8) \qquad f(z) = c_0 + c_1(z - z_c) + c_2(z - z_c)^2 + c_3(z - z_c)^3 = \mathbf{c}^T z^{\mathbf{i}} \quad ,$$
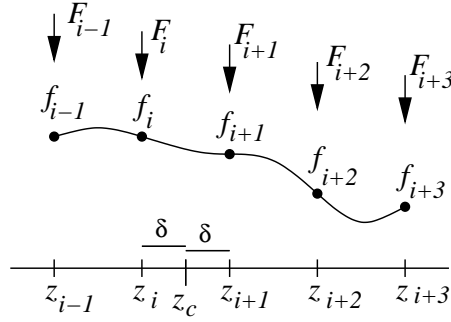
FIGURE 11.2. The spline consists of segments of length 2δ. The center of each segment z_c is used for the expansion of the polynomial.

where z_c is the center of the segment. See Figure 11.2, and also Sections 8.2 and 6.1.

The cubic spline depends on four variables, for which we can make a convenient choice. These four variables can be paired, such that two variables appear on the left-hand side of the domain and two on the right-hand side. The typical variables one could use are the values and derivatives at the ends $f_{i-1} = f(z_{i-1}), f_i = f(z_i)$ and f'_{i-1}, f'_i. The map from these variables to the coefficients of the polynomial on a domain $z \in [z_{i-1}, z_i]$ with the center $z_c = (z_{i-1} + z_i)/2$ of length $2\delta = z_i - z_{i-1}$ is:

$$(11.9) \qquad \mathbf{c} = \begin{pmatrix} 1 & -\delta & \delta^2 & -\delta^3 \\ 0 & -1 & 2\delta & -3\delta^2 \\ 1 & \delta & \delta^2 & \delta^3 \\ 0 & 1 & 2\delta & 3\delta^2 \end{pmatrix}^{-1} \begin{pmatrix} f_{i-1} \\ f'_{i-1} \\ f_i \\ f'_i \end{pmatrix} = \mathbf{T}\mathbf{f}_1 \ ,$$

where $\mathbf{f}_1$ denotes the vector of variables of a single segment. It is comprises of the left-end value and the derivative and the right-end value and derivative. Since the elastic energy is given by the second derivative of the displacement, the elastic energy is a quadratic function of the end point variables:

$$(11.10) \qquad E_{\text{elastic}} = \int_{z_{i-1}}^{z_i} dz \frac{\kappa}{2} (\partial_z^2 f(z))^2 = \frac{\kappa}{2} \mathbf{f}^T \mathbf{T}^T \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 8\delta & 0 \\ 0 & 0 & 0 & 24\delta^3 \end{pmatrix} \mathbf{T}\mathbf{f} \ .$$

Hence the elastic energy is given by a two-dimensional subspace of the configuration state space $\mathbf{f}_i$. In the space of the four variables there is a two-dimensional subspace which leaves the elastic energy unchanged. The two-dimensional null-space consists of the piece-wise linear parts.

The concatenation of piece-wise linear segments will have kinks between them, which yield infinite boundary energy. See Section 11.6. We will show this below. The variables of the spline, the polynomial coefficients on each of the segments define a smooth curve on each segment, but the problems occur at the boundaries. In order to solve this, constraint relations among the coefficients on neighboring segments arise. In the next section we will see that the common variables at the

boundaries, i.e., the port variables, are much more natural to define the shape of the beam.

The spline approximation of the Euler-Bernoulli beam has two further constraints, in the second and third derivative across the successive segments. If the second derivative does not match, it would yield a kink between to successive segments with an infinite bending energy. The value of the third derivative across the segment boundary is associated with a local force F_i applied at z_i. Hence from the four variables $\mathbf{c}$ on each segment only one remains. It can be seen as the position f_i given by the local force F_i. Since these two relations are across the segment boundaries, they link all the segments together, into a collective object. In some cases it might be possible to find the coefficients successively, from one segment to the next. But in many cases the problem has to be solved as a whole, in terms of constraints.

Given the coefficients, these derivatives are easily constructed:

$$(11.11) \qquad \begin{pmatrix} f_i \\ f_i' \\ f_i^{(2)} \\ f_i^{(3)} \end{pmatrix} = \begin{pmatrix} 1 & \pm\delta & \delta^2 & \pm\delta^3 \\ 0 & 1 & \pm2\delta & 3\delta^2 \\ 0 & 0 & 2 & \pm6\delta \\ 0 & 0 & 0 & 6 \end{pmatrix} \mathbf{c} = \mathbf{Q}^\pm \mathbf{c} \quad ,$$

where the sign depends on the evaluation on the left-end or right-end of the segment.

The third and fourth row of the matrix $\mathbf{Q}^\pm\mathbf{T}$ are the basis of the constraint relations. The second derivative vanishes at z_i if:

$$(11.12) \qquad (0, 0, 1, 0) \left(\mathbf{Q}^+\mathbf{T}\mathbf{f}_1^i - \mathbf{Q}^-\mathbf{T}\mathbf{f}_1^{i+1} \right) = 0 \quad .$$

The third derivative yields the force F_i:

$$(11.13) \qquad (0, 0, 0, 1) \left(\mathbf{Q}^+\mathbf{T}\mathbf{f}_1^i - \mathbf{Q}^-\mathbf{T}\mathbf{f}_1^{i+1} \right) = -\frac{F_i}{\kappa} \quad .$$

In this case it is more convenient to have a two-segment vector, corresponding to larger stencils as described in Section 8.4:

$$(11.14) \qquad \mathbf{f}_2 = \begin{pmatrix} f_{i-1} \\ f_{i-1}' \\ f_i \\ f_i' \\ f_{i+1} \\ f_{i+1}' \end{pmatrix} \quad ,$$

such that the constraints are expressed as vector products $\chi_c^T \mathbf{f}_2^i = 0$ and $\chi_f^T \mathbf{f}_2^i = F_i$.

The constraint vectors χ_c and χ_f are versatile. For example, starting from the left, the successive coefficients f_{i+1} and f_{i+1}' can be determined from the four preceding coefficients $\mathbf{f}_1$ and the force F_i. The algorithm is equivalent to the integration scheme of the fourth-order differential equation, which requires five points to determine the sixth:

$$(11.15) \qquad \partial_z^4 f(z) = \frac{F(z)}{\kappa(z)} \quad .$$

See Sections 8.4 and 3.6. If integration is possible, it is a fast method to determine the solution. Not in all cases constraints can be integrated, so to say, from starting at one end and successively solving the equations to the other end. Systems integrable in such a manner we call hierarchical systems, as the equations can

be ordered hierarchically, like an upper or lower triangular matrix equation. In the case of the Euler-Bernoulli beam, the boundary conditions are partially set at both ends. Neither end suffices to solve the system completely. The corresponding matrix equation is not triangular, but banded, where the dimension $\dim\mathbf{f}_2 = 6$ determines the width of the band.

However, constraints can also be implemented dynamically, similar to iterative methods in numerical analysis. See Chapter 10. A large, instead of infinite elastic energy can be associated with the constraint. A Hamiltonian with such elastic energy given by:

$$(11.16) \qquad E_{\text{constraint}} = \sum_i \frac{k_c}{2}(\chi_c^T \mathbf{f}_2^i)^2 + \sum_i \frac{k_f}{2}(\chi_f^T \mathbf{f}_2^i - F_i)^2 \quad ,$$

the first term is the constraint energy, the second term the applied force, also implemented as constraint. With the appropriate damping, it will converge to the constraint solution with $E_{\text{constraint}} = 0$, if possible. In many cases a large constraint problem is converted to an optimization problem.

11.2. Invariant lines

Above we have seen that splines suffer from at least two problems. First, they are not invariant under rotations, as it is tied to the z and f direction for the reference coordinate and the displacement. A rotated beam will have a different energy from the original beam. Secondly, the matching of the segments requires a constraint equation, such that the elastic energy of the boundary remains finite, and automatically zero since the energy is a quadratic function. An invariant function depends only on invariant vector products. The beam is defined by positions of the nodes $\mathbf{x}_i$ and directional vectors at the same node $\mathbf{v}_i$. The position along the curve is given by the reference coordinates s. The lowest order polynomial which gives an initial direction $\mathbf{v}_{i-1}$, a final direction $\mathbf{v}_i$, and a translation $\mathbf{x}_i - \mathbf{x}_{i-1}$ is a third order polynomial. The position $s \in [0,1]$ along the segment is given by:

$$(11.17) \quad \mathbf{x}_i(s) = \mathbf{x}_{i-1} + (\mathbf{x}_i - \mathbf{x}_{i-1})(3s^2 - 2s^3) + \mathbf{v}_{i-1}s(1-s)^2 + \mathbf{v}_i s^2(s-1) \quad .$$

Hence a smooth curve between $\mathbf{x}_{i-1}$ and $\mathbf{x}_i$ depends on the tangential directions $\mathbf{v}_{i-1}$ and $\mathbf{v}_i$ at beginning and end. The size of the tangential vectors determine the flatness at the endpoints. The coordinate s is not the length, i.e., $|\partial_s \mathbf{x}| \neq 1$. The beam may therefore be stretched or compressed along its length, unlike the Euler-Bernoulli beam.

A beam put together from such segments will be invariant under rotations and translations, since it depends only on vectors which all transform in the same manner. Furthermore, the elastic energy is finite. The invariant curvature is given [Guggenheimer, 1963] by:

$$(11.18) \qquad k(s) = \partial_s \mathbf{x}(s) \times \partial_s^2 \mathbf{x}(s) \quad .$$

Inserting the explicit function, Eq. 11.17, in Eq. 11.18, for a segment yields a lengthy expression:

$$(11.19) \qquad k(s) = 6(\mathbf{x}_i - \mathbf{x}_{i-1}) \times \mathbf{v}_{i-1}(1-s)(12s^2 - 9s + 1) \quad .$$

$$+6(\mathbf{x}_i - \mathbf{x}_{i-1}) \times \mathbf{v}_i s(12s^2 - 15s + 4) + 2\mathbf{v}_i \times \mathbf{v}_{i-1}(1 - 2s)(9s^2 - 9s + 1) \quad .$$
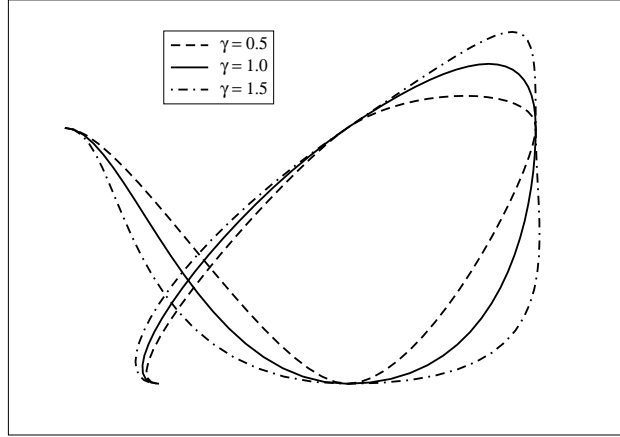
FIGURE 11.3. A smooth curved defined by five points $\mathbf{x}_i$ and two
end point directions $\mathbf{v}_1$ and $\mathbf{v}_5$, of relatively small size compared to
the interposition distance $|\mathbf{v}| < |\mathbf{x}_{i+1} - \mathbf{x}_i|$. For increasing rigidities
the curvature shifts from the set points to the center of the lines.

However, examining the segment boundary, it yields only a finite jump in curvature
across the segments:

$$(11.20) \qquad \lim_{\delta \to 0} k(\pm\delta) = 6\mathbf{v}_i \times (\mathbf{x}_i - \mathbf{x}_{i\pm1}) \quad .$$

Therefore, the elastic energy will have no infinite boundary contribution and no
constraints are required to remove them, as was in the case of the spline approxi-
mation. There is, however, a force acting on the segments, besides the local force
acting on the boundaries. The precise definition of the force depends on the choice
of variables. Was in the case of the spline approximated all the force located at
the segment boundary, with the awkward constraints, here the force is more evenly
distributed, without any constraints, except for possible forces at the boundary. It
should be noted that without force $F(z)$ the beam shape is trivial; the purpose is
to describe the response to force in some distributed sense.

Eventually, one would use the positions and directions, $\mathbf{x}_i$ and $\mathbf{v}_i$, of the nodes
in the discretized theory. A distributed force $\mathbf{F}(s)$ could be averaged over the
segments, like the forces F_i in the Euler-Bernoulli beam. The curvature $k(s)$ and
the stiffness $\kappa(s)$ determine the local, distributed force:

$$(11.21) \qquad \mathbf{F}(s) = \frac{\delta E}{\delta s} = \kappa k(s)\partial_s k(s) \quad .$$

Furthermore, the situation without jump and associated force $\mathbf{F}_i$ arises from:

$$(11.22) \qquad \mathbf{v}_i = \gamma(\mathbf{x}_{i+1} - \mathbf{x}_{i-1}) \quad ,$$

where γ is the rigidity. For small rigidity of the node, the force yielding the curved
beam is located mainly at the nodes $\mathbf{x}_i$; for large rigidity the force acts mainly in be-
tween the nodes. A smooth distribution of the force corresponds to $\gamma \approx 1$. Hence, a

stationary spline like approximation through the points $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \cdots$ arises simply without constraints. See Figure 11.3. Only the first and last directional vectors, at the ends of the line, are not fully determined: $\mathbf{v}_0 = 0$ corresponds to a supported-like boundary, while a nonzero $\mathbf{v}_0$ is like a clamped boundary. The rigidity is the consequence of the polynomial is s. If we would have chosen to construct the segments from piece-wise arc sections, the curvature is automatically homogeneous, and would to depend on the scale introduced by the polynomials and set by the rigidity. The Bezier curves and non-rational uniform B-Splines (NURBS) are only a few of related smooth curves, independent of the orientation. [Foley and van Dam, 1982, Farin, 2001] These curves are mainly used for computer graphics. Only recently they are of interest of the FEM community as a means to bridge the gap between computer aided design (CAD) and FEM analysis and simulation. [Hughes et al., 2005]

The directions $\mathbf{v}_i$ all transform in the same manner under the rotation $\mathbf{R}$ of the affine transformation of Eq. 11.1. The positions $\mathbf{x}_i$ also translate with $\mathbf{a}$ of the full affine transformation. By construction the elastic energy remains the same.

Interestingly, the only contributions to the curvature are by the cross-product of the relative positions $\mathbf{x}_{i+1} - \mathbf{x}_i$ and the directional vectors $\mathbf{v}_i$ and $\mathbf{v}_{i+1}$, and a smaller contribution from $\mathbf{v}_i \times \mathbf{v}_{i+1}$. In fact, the elastic energy can be expressed as weighted vector products:

(11.23)

$$\int_0^1 ds \frac{\kappa k(s)^2}{2} = \frac{\kappa}{2} \begin{pmatrix} (\mathbf{x}_i - \mathbf{x}_{i-1}) \times \mathbf{v}_{i-1} \\ \mathbf{v}_{i-1} \\ (\mathbf{x}_i - \mathbf{x}_{i-1}) \times \mathbf{v}_i \\ \mathbf{v}_i \end{pmatrix}^T \mathbf{K} \begin{pmatrix} (\mathbf{x}_i - \mathbf{x}_{i-1}) \times \mathbf{v}_{i-1} \\ \mathbf{v}_{i-1} \\ (\mathbf{x}_i - \mathbf{x}_{i-1}) \times \mathbf{v}_i \\ \mathbf{v}_i \end{pmatrix} ,$$

where the stiffness matrix $\mathbf{K}$ is positive constant and symmetric. Furthermore, the stretch energy is defined by:

(11.24)
$$E_{\text{stretch}} = \int_0^1 \frac{EA}{2} \partial_s \mathbf{x}^T(s) \partial_s \mathbf{x}(s) ds ,$$

or in the case of a reference or rest length:

(11.25)
$$E_{\text{stretch}} = \int_0^1 \frac{EA}{2} (\partial_s \mathbf{x}^T(s) \partial_s \mathbf{x}(s) - 1)^2 ds .$$

11.3. Postponement and pre-work

There are two main approaches to the numerical implementation of the constraints, which are the consequence of external and internal boundary conditions. They both generate a model, which yield the displacement $f(z)$ as direct result of the boundary conditions and the applied forces F_i. The first approach we call pre-work, and it works mainly in linear theory. For appropriate boundary conditions the effect of the forces decouple from the effects of the boundary conditions. If for each force F_i with clamped zero boundary conditions the displacement is determined the total displacement is a linear combination of forces and boundary values:

(11.26)
$$f(z) = \phi_0(z)f_0 + \psi_0(z)f_0' + \phi_N(z)f_N + \psi_N(z)f_N' + \sum \phi_i(z)F_i ,$$

where the modes ψ_0 and ψ_N correspond to the second boundary condition at each end, associated with the torque force. The four boundary conditions are given by the boundary values f_0, f_0', f_N, and f_N'. Each force F_i at z_i has its own modal function:

(11.27)

$$
\phi_i(z) = \begin{cases} \frac{L(L-z_i)^2}{6\kappa}\left(3\frac{z_i}{L}\left(\frac{z}{L}\right)^2 - \left(1+\frac{2z_i}{L}\right)\left(\frac{z}{L}\right)^3\right), & 0 \le z \le z_i \\ \frac{Lz_i^2}{6\kappa}\left(3\frac{L-z_i}{L}\left(\frac{L-z}{L}\right)^2 - \left(1+\frac{2(L-z_i)}{L}\right)\left(\frac{L-z}{L}\right)^3\right), & z_i \le z \le L \end{cases}.
$$

The four modal functions for the boundaries are given by the transformation matrix $\mathbf{T}$, with the distance $\delta = L$, given in Eq. 11.9.

The disadvantage of such a modal decomposition is three-fold. First, for large models, the number of modes grows out of control. Already the construction and storage of such modes require a lot of computing power. Secondly, the boundary conditions and input forces are set. Using, for example, force actuation at the boundary, or the value of the displacement $f(z_i)$ instead of the force F_i at a particular node requires the partial inversion of the large modal matrix $\phi_i(z_j)$.

The approach described above will work very well in linear models. In nonlinear models it is often prohibitively expensive. In that case every configuration changes the forces and interactions, which need to be calculated at every stage. One seeks to reduce and automate the calculations, and possibly adapt the model that it can be done generically, sequential, or parallel. Such an approach we call postponement. In this case the displacement is written as a general integral with four unknown integration constants:

(11.28)

$$
f(z) = \int_0^z dz' \left(\int_0^{z'} dz'' \left(\int_0^{z''} dz''' \left(\sum_i \frac{F_i\delta(z''' - z_i)}{\kappa(z''')} + C_3 \right) + C_2 \right) + C_1 \right) + C_0 ,
$$

a direct implementation of Eq. 11.15. The numerical implementation will be the numerical scheme to calculate the respective integrals of the different unknown constants.

The four constants C_i are given by the boundary conditions. They match with the four boundary modes $\phi_0(z), \psi_0(z), \phi_N(z)$, and $\psi_N(z)$ above. However, in this case the homogeneous displacement is constructed as we integrate along. Little memory is required; the displacement $f(z)$ is constructed on the fly. Any unknown force F_i given by the constraint $f(z_i)$ is added to the set of integration constants. Nonlinearities will not affect the approach. However, different boundary condition, associated with different causalities, may require a different postponement approach. For example, a beam with one free end, may be integrated from the clamped end to the free end without constraints, while a beam clamped at both end requires to constants C_i is integrated from one side to the other.

11.4. Consistency and conservation laws

Continuous theories are designed to satisfy conservation laws. A theory inconsistent with underlying physical laws, such as mass or energy conservation, is soon rejected. The consistent theories remain. However, continuous theories have, in principle, an infinite number of degrees of freedom and can only be analyzed numerically in a finite order approximation. Within the process of approximating the infinite number of degrees of freedom with a workable finite set, choices have to be

made. The simplest way to make consistent approximations is to resort to the variational principles. If the continuous function satisfies some minimization principle, the appropriate finite-order approximation will also satisfy the same minimization principle where the numerical value approximates the exact continuous result from above. These principles have been the guiding principles in the development of dynamical systems for centuries.

Conservation laws are continuity relations, in their generic form they relate the flow $\mathbf{J}$ to the rate-of-change of a density ρ:

$$(11.29) \qquad\qquad \dot{\rho} + \nabla \cdot \mathbf{J} = 0 \ \ .$$

This equation immediately sums up all the trouble one would have in constructing finite-order approximation of ρ and $\mathbf{J}$. First, the divergence means that the continuous nature of the flow $\mathbf{J}$, as function of position is related to a density ρ at a point. Flow and density are therefore of a different geometric type. The same is true for the time-derivative of the density, $\dot{\rho}$, in the continuity equation. The flow and the density are also of different dynamical types. Furthermore, if the continuity equation holds, so must any derivative, spatial and temporal:

$$(11.30) \qquad\qquad \ddot{\rho} + \nabla \cdot \dot{\mathbf{J}} = \nabla \dot{\rho} - \Delta \mathbf{J} = 0 \ \ ,$$

where we used that typically the flow is defined rotation-free: $\nabla \times \mathbf{J} = 0$. Hence the equations of motion, not resulting directly from conservation laws, must still be consistent with the conservation laws.

A lot of effort is taken to satisfy as many as possible of the underlying physical principles in the numerical approximation of the continuous theory. Conceptually, it is difficult to determine which approach will retain the principles, and the resulting symmetries and conservation laws. Our starting point is retaining as much of the consistency of the equations of motion as possible. Dirac [Dirac, 1964] developed a scheme to implement constraints based on the time dependence of constraints. The constraints are inserted into the equations of motion, yielding possibly additional constraints, and eventually a consistent equations of motion based on the commutator brackets. Our approach is to seek a consistent time dependence of the finite-order approximation. Simply said, even though $\mathbf{J}(z)$ only has limited variation as a function of the spatial coordinate z, the corresponding $\rho(z)$ makes the pair satisfy the continuity relation, Eq. 11.29, exactly. Any conservation law based on the equations of motion will be satisfied as well. Both the states and the Hamiltonian itself are approximated in the process.

The continuous Hamiltonian equations of motion depend on functions of the continuous spatial variable z:

$$(11.31) \qquad\qquad \begin{pmatrix} \dot{q}(z) \\ \dot{p}(z) \end{pmatrix} = \begin{pmatrix} \delta_{p(z)} H \\ -\delta_{q(z)} H \end{pmatrix} + \begin{pmatrix} 0 \\ \mathbf{F}(z) \end{pmatrix} \ \ .$$

The finite-dimensional set of equations of motion with $\mathbf{q}, \mathbf{p} \in \mathbb{R}^N$ are the equations of motion of the coefficients of some linear subspace of the continuous space:

$$(11.32) \qquad\qquad \begin{pmatrix} \dot{\mathbf{q}} \\ \dot{\mathbf{p}} \end{pmatrix} = \begin{pmatrix} \nabla_{\mathbf{p}} H \\ -\nabla_{\mathbf{q}} H \end{pmatrix} + \begin{pmatrix} 0 \\ \mathbf{f} \end{pmatrix} \ \ .$$

The vectors $\mathbf{q}$ and $\mathbf{p}$ are the coefficient vectors, i.e., $\mathbf{q} = (q_0, q_1, q_2, \cdots, q_N)$ of corresponding modes or basis functions in the continuous space:

$$(11.33) \qquad \left(\begin{array}{c} q(z) \\ p(z) \end{array} \right) = \left(\begin{array}{c} \sum_j \phi_j(z) q_j \\ \sum_j \psi_j(z) p_j \end{array} \right) \quad .$$

Like the divergence operator $\nabla\cdot$ in the continuity relation, Eq. 11.29, the Hamiltonian of a continuous system will in general contain differential operators acting on the continuous, or field variables. The differential operator D acting on the functions $q(z)$ and $p(z)$ must have exact representations $\mathbf{D}$ in the coefficient space:

$$(11.34) \qquad Dq(z) = h(z) \Leftrightarrow \mathbf{D}\mathbf{q} = \mathbf{h} \quad ,$$

such that: $h(z) = \sum \phi_j(z) h_j$. For example, polynomial, exponential, and trigonometric functions will be mapped into the same functions under differentiation. These are precisely the functions commonly used the modal approximations, of linear, homogeneous systems.

Consistency of the equations of motion means the right-hand side and the left-hand side of the continuous equation of motion, Eq. 11.31, using the modes $\phi_j(z)$ and $\psi_j(z)$ and the discrete equation of motion, Eq. 11.32 lie in the same subspace of the infinite-dimensional continuous space. If that is the case, the conservation laws derived from the equations of motion are also automatically satisfied. Hence, rather than a metric approach, based on the norm as in the modal approach, we use an algebraic approach, where we identify the operators.

Position-dependent masses and stiffnesses have to be restricted, for example, to piece-wise linear mass densities $\rho_0 + \rho_1 z$ and stiffnesses $\kappa_0 + \kappa_1 z$, such that the symmetric formulation of the one-dimensional wave equation, in a polynomial space of modes [Henrici, 1974] is indeed consistent:

$$(11.35) \qquad \left(\begin{array}{c} \dot{q}(z) \\ \dot{p}(z) \end{array} \right) = \left(\begin{array}{c} \partial_z(\rho_0 + \rho_1 z) p(z) \\ -\partial_z(\kappa_0 + \kappa_1 z) q(z) \end{array} \right) \quad .$$

If $q(z)$ and $p(z)$ lie in a polynomial space of order N, the multiplication with the linear material constant increases the order by one to $N + 1$, while the differentiation lowers the dimension again to N. In the case of nonlinear Hamiltonians this construction is more complicated, as we have seen in Section 5.10, however, luckily the differential operators can be brought outside the Hamiltonian, yielding the port-Hamiltonian formulation, discussed in Section 7.5. The projection $\partial_z[(\rho_0 + \rho_1 z) p(z)]$ from an N-dimensional polynomial space back to the N-dimensional polynomial space should have a nontrivial null-space, associated with a single boundary condition. In this case the one-dimensional null-space is spanned by:

$$(11.36) \qquad p_0(z) = \frac{1}{\rho_0 + \rho_1 z} \quad ,$$

which does not lie in the polynomial subspace. As a result conservation laws will fail for such approximations of position-dependent Hamiltonian with polynomial function spaces. On the other hand, the truncation of $(\rho_0 + \rho_1 z) p(z)$ back to the N-dimensional polynomial space will yield a non-trivial null-space, an approximation of $p_0(z)$, and the requirement of a boundary condition. The particular null-space p_0 is in this case the consequence of the translation invariance, yielding solutions with constant velocity:

$$(11.37) \qquad v = \text{constant} \Rightarrow (\rho_0 + \rho_1 z) p_0 = \text{constant} \quad .$$

In the general dynamical setting of evolution equations, discussed in Chapter 3, the matrix $\mathbf{A}$ is nonsingular otherwise the dimension of the state could be reduced:

$$(11.38) \qquad\qquad \dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} \quad .$$

A typical differential operator $\mathbf{D}$ at the basis of $\mathbf{A}$, however, will be singular: $\mathbf{D}\mathbf{x}_0 = 0$. The resulting $\mathbf{A}$ will also be singular. There are two solutions for consistent dynamics: First, the matrix $\mathbf{A}$ inherits the null-space from the underlying $\mathbf{D}$ operator, and the control vector $\mathbf{B}$ spans in the null-space of $\mathbf{A}$:

$$(11.39) \qquad\qquad \mathbf{A}\mathbf{x}_0 = 0 \Rightarrow \mathbf{B}^T\mathbf{x}_0 \neq 0 \quad .$$

For example, consider a transmission line $\mathbf{A}x_j(t) = x_{j+1}(t+\delta)$, the signal moves to the right, and an input with $\mathbf{B} = (1, 0, 0, \cdots)$ adds a new $x_0(t) = u(t)$ at the input end.

The second solution is a reformulation of the first one: The state $\mathbf{x}$ is restricted to the image space of $\mathbf{A}$, and $\mathbf{B}$ represents the coupling between the boundary conditions, fixing the core model (Section 7.6) in the null-space of $\mathbf{A}$, and the internal states $\mathbf{x}$.

In later sections consistency will appear in different forms. We rely on polynomial spaces to have injective differential operators in the Hamiltonian, using only piece-wise constant material properties. These differential operators are therefore never of full rank. Differential operators may have complicated null-spaces. The case of $\nabla\times$ we analyze in some detail in Section 11.7. The boundary conditions play an important role to yield consistency.

11.5. Element Hamiltonian

In the Hamiltonian nonlinear FEM, each element Hamiltonian is the energy expressed in local variables $\mathbf{q}$. The null-space of the Hamiltonian is the set of variables $\mathbf{q}_0$ under which the energy is invariant:

$$(11.40) \qquad\qquad H(\mathbf{q}) = H(\mathbf{q} + \mathbf{q}_0) \quad .$$

In the case of a quadratic Hamiltonian it corresponds to the eigenspace with eigenvalue 0, i.e., $\mathbf{q}_0^T\mathbf{K}\mathbf{q}_0 = 0$, if $\|\mathbf{K}\| \geq 0$.

This is not invariant under general coordinate transformations. For example consider a Hamiltonian:

$$(11.41) \qquad\qquad H = \frac{1}{2}(q_1^2 + q_2^2) \quad .$$

In this case the null-space is the origin $q_1 = q_2 = 0$, however, under the introduction of polar coordinates r, θ with $r^2 = q_1^2 + q_2^2$ and $\tan\theta = y/x$, the Hamiltonian depends only on r. Such a coordinate transformation is similar to the action-angle coordinates introduced in Section 2.2. The null-space is θ. In the case of discrete approximations, these two sets of variables are distinct approximations, which will lead to different subspaces of the full system. With the polar coordinates it would be difficult to interconnect a grid of segments.

Elasticity Hamiltonians are invariant under translation and rotation of the object as a whole. Hence the translation and rotation transformations should be part of the null-space of the Hamiltonian. The kinetic energy, however, generally depends on translation and rotation velocities.

order	#variables	#boundary	eigenvalues (degeneracies)
0	3	3	0^3
1	24	24	$0^{10}4^68^612^2$
2	81	78	$0^{29}4^68^69^612^213^{12}17^618^622^627^2$
3	192	168	$0^{66}4^68^69^612^213^{12}16^617^618^6$
			$20^{12}22^624^625^{12}27^229^{12}32^634^636^6$

TABLE 11.1. The dimensions and eigenvalues of the quadratic operator $\mathbf{A} \to (\nabla \times \mathbf{A})^2$ in a polynomial space of different low orders.

For example, given a block of elastic material. In the linear approximation it is determined by eight three-dimensional positions: 24 variables. In the null-space of the element Hamiltonian there are three dimensions associated with the translation of the block, and another three with rotations, leaving an 18-dimensional elasticity subspace. The strain tensor is normally only determined by four positions, 12 variables, of which 6 are associated with rotation and translation, leaving three compression and three shear coefficients.

A rotational flow, or magnetic field, which invariant energy is given by:

$$(11.42) \qquad H(\mathbf{A}) = \frac{k}{2}(\nabla \times \mathbf{A})^T(\nabla \times \mathbf{A}) \ \ .$$

If the vector field $\mathbf{A}$ is approximated on a unit cube $[0,1] \otimes [0,1] \otimes [0,1]$ by linear functions:

$$(11.43) \qquad \mathbf{A}(x,y,z) = \mathbf{a}_{000} + \mathbf{a}_{100}x + \mathbf{a}_{010}y + \cdots + \mathbf{a}_{111}xyz \ \ ,$$

there are again 24 variables $\mathbf{a}_{ijk}$, almost identical to the elastic block. The 24 displacement positions can be mapped to 24 terms in a polynomial expansion. The null-space, however, seems much more complicated. The constant function $\mathbf{a}_{000}$, associated with laminar flow, or translation, in three directions is part of the null-space. So are all the symmetric terms: a^x_{100}, a^y_{010}, and a^z_{001}. The three symmetric permutation combinations:

$$(11.44) \qquad a^z_{010} = a^y_{001}, \quad a^x_{001} = a^z_{100}, \quad a^y_{100} = a^x_{010} \ \ ,$$

are also part of the null-space, and so is the combination:

$$(11.45) \qquad a^x_{011} = a^y_{101} = a^z_{110} \ \ .$$

Together they span the 10-dimensional null-space of the curl operator on polynomials to order 1. For polynomials to order 2 the null-space is 29-dimensional, and for order 3 the dimension is 66. See Table 11.1.

11.6. Boundary energy

In order to construct a proper FEM model, using element Hamiltonians, the energy associated with the boundary should vanish. The separate Hamiltonians interact due to their common variables, but any boundary energy is automatically infinite, since it is the product of two delta functions. For example a discontinuous step function $f(z) = 0$ for $z < 0$ and $f(z) = 1$ for $z \geq 0$ will result in a delta function under differentiation: $\partial_z f(z) = \delta(z)$. The boundary energy is therefore infinite:

$$(11.46) \qquad E = \frac{1}{2}\int dz(\partial_z f(z))^2 = \frac{1}{2}\int dz(\delta(z))^2 = \infty \ \ .$$

The approximate solution will be discontinuous for a certain order of the derivative. However, this order should be high enough that it does not yield a contribution to the energy. The spline solution led to a boundary term for the second derivative, while using the invariant-lines approach this contribution was absent.

In the case of first-order operators like $\nabla\times$, only the function itself has to be continuous to yield a consistent result without boundary contributions. The boundary energy for the common boundary $\partial\Omega = \Omega_i \cap \Omega_j$ between domains Ω_i and Ω_j is given by:

$$(11.47) \qquad E_{ij} = \int_{\Omega_1 \cap \Omega_2} dz (\mathbf{n}_{ij}^T \delta_\nabla D(f_i - f_j))^2 \quad,$$

where the boundary operator $\mathbf{n}_{ij}^T \delta_\nabla H$ is defined as the the sum of terms where every time a single differential operator is replaced by the component of the boundary normal $\mathbf{n}_{ij} = -\mathbf{n}_{ji} = (n_{ij}^x, n_{ij}^y, n_{ij}^z)$:

(11.48)
$$(\mathbf{n}_{ij}^T \delta_\nabla \partial_{\alpha_1}\partial_{\alpha_2}\cdots\partial_{\alpha_r}f)^2 = (n_{ij}^{\alpha_1}\partial_{\alpha_2}\cdots\partial_{\alpha_r}f + \partial_{\alpha_1}n_{ij}^{\alpha_2}\cdots\partial_{\alpha_r}f + \partial_{\alpha_1}\partial_{\alpha_2}\cdots n_{ij}^{\alpha_r}f)^2 \quad.$$

The approximate function on the whole domain is the sum of smooth functions f_i, extendable beyond the cell Ω_i, times window functions $w_i(z) = 1$ for $z \in \Omega_i$ and zero outside the cell domain Ω_i:

$$(11.49) \qquad f(z) = \sum w_i(z)f_i(z) \quad.$$

Inserting $f(z)$ in the Hamiltonian results in boundary terms due to the derivative of the window functions:

$$(11.50) \qquad \int_{\Omega_i} dz \nabla w_i(z) = \int_{\Omega_i} dz \mathbf{n}^T \delta(z)|_{\partial\Omega_i} = \int_{\partial\Omega_i} dz \mathbf{n} \quad.$$

The boundary energy contribution only vanishes if $\mathbf{n}_{ij}^T \delta_\nabla D(f_i - f_j) = 0$.

In the case of the energy given by $(\nabla\times\mathbf{A})^2$ the condition reduces to a continuity condition

$$(11.51) \qquad \mathbf{n} \times \mathbf{A}_i|_{\Omega_i \cap \Omega_j} = \mathbf{n} \times \mathbf{A}_j|_{\Omega_i \cap \Omega_j} \quad.$$

The solutions in the null-space of the Hamiltonian automatically satisfy the zero boundary-energy condition.

11.7. Counting degrees of freedom

A system exists on a domain with a boundary. Furthermore, it consists of elements each again with domains and boundaries. The common points of different elements, i.e., the boundary points, are relations which reduce the number of degrees of freedom. Furthermore, from the total set of degrees of freedom a subset will be determined by the equations of motion and the initial state, the remainder must be determined by the boundary conditions, or the input. A consistent set of equations of motion are the result of assigning the right number of variables to the boundary. However, since we make no distinction between internal boundaries, between the elements, and external boundaries, for the boundary conditions, the proper definition of variables on the elements should fix the system completely.

For triangulation, grid consisting of simplices, i.e., tetrahedrons, this extension from elements to the full grid is not trivial. The number of elements with a common vertex or edge is no longer fixed. If, therefore, the relations between degrees of freedom is based on the identification of a common vertex or edge, it is not clear

how many relations will be given by a single identification. The result depends on
the type of triangulation of the domain.

In this section we will restrict ourselves to cubic cells. The number of neighbors
parts are fixed. The corner is shared by eight elements, an edge by four elements,
and a face by two.

Knowledge of the null-space of the element Hamiltonian allows us to address the
question of uniqueness and existence. For example, consider the magnetic problem
expressed in the vector potential $\mathbf{A}$, with $H(\mathbf{A}) = (\nabla \times \mathbf{A})^2$. The static solution
to some boundary conditions $\mathbf{b}$ is:

$$(11.52) \qquad \min_{\mathbf{A}} H(\mathbf{A}), \quad \mathbf{A}|_\partial = \mathbf{b}$$

Uniqueness and existence requires a single minimum to exists. The Hamiltonian
does not have a single minimum, it has a null-space which yield a set of solutions
all with the same minimal energy. The boundary condition $\mathbf{b}$ should select from
this null-space a unique solution. In the continuous case typical boundary condi-
tions would still not yield a unique solution as the null space contains any function
$\mathbf{A}_\parallel = \nabla\phi$, given by an arbitrary ϕ, including functions that vanish at the bound-
ary. In the discrete approximation the typical basis functions do not vanish at the
boundary. In the case of a discrete approximation of a vector field $\mathbf{A}$, it is generally
an unrestricted field $\mathbf{A}$ with $\mathbf{A}_\parallel \neq 0$. We would, however, like to separate the
longitudinal degrees of freedom in $\mathbf{A}$, since they are not associated with dynamics.
Hence the null-space $\nabla\times$ is separated in two parts: the degrees of freedom fixed
by boundary conditions, or port variables, and degrees of freedom associated with
static or constrained longitudinal fields.

In the continuous case, normally, to yield uniqueness a particular gauge choice,
for example the Coulomb gauge $\nabla \cdot \mathbf{A} = 0$ or the axial gauge $A^z = 0$, is used
as auxiliary, or constraint equation. The gauge condition yields a different set of
boundary conditions than the original equations defined in terms of $\mathbf{B} = \nabla \times \mathbf{A}$,
already for the fact that if $\mathbf{A}$ are polynomials of a particular order, the order of the
generated $\mathbf{B}$ is lower. In terms of the magnetic field $\mathbf{B}$ the boundary conditions
should yield $\nabla \cdot \mathbf{B} = 0$. For $\mathbf{A}$ these conditions are naturally fulfilled.

For the polynomial system, the gauge condition should yield a section of the
null-space, such that a unique projection null-space arises. The Coulomb gauge
condition yields the four equations:

$$(11.53) \qquad a^x_{110} + a^z_{011} = a^y_{110} + a^z_{101} = a^x_{110} + a^z_{011} = a^x_{100} + a^y_{010} + a^z_{001} = 0 \ \ .$$

The remainder of the ten-dimensional null-space needs to be fixed by the boundary
conditions. The constant vector potential $\mathbf{a}_{000}$ can clearly only be fixed by bound-
ary conditions. Furthermore, the gauge condition restricts only one dimension in
the three-dimensional symmetric space $\{a^x_{100}, a^y_{010}, a^z_{001}\}$. Hence there are three
nontrivial boundary conditions, such that the intersection of the null-space with
the gauge-condition and the boundary condition $\partial : \mathbf{A}|_\partial = \mathbf{b}$ is unique:

$$(11.54) \qquad \text{rank}\{\nabla\times, \nabla\cdot, \partial\} = 24 \ \ ,$$

and consistent (see Section 11.4):

$$(11.55) \qquad \text{rank}\{\nabla\times\} + \text{rank}\{\nabla\cdot\} + \text{rank}\{\partial\} = 24 \ \ ,$$

where $\nabla\times$, $\nabla\cdot$, and ∂ represent the Hamiltonian differential operator, the gauge
condition operator, and the boundary operator acting on the polynomial vector
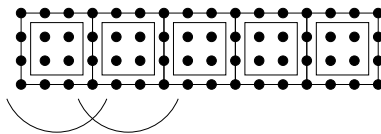
FIGURE 11.4. A third-order polynomial in each variable x, y, z on each cell is determined by the values at $4^{\dim}$ points. The boundary points are shared with neighboring cell, or the boundary of the domain. The internal points are unaffected.

space $\mathbf{A} = \sum \mathbf{a}_{ijk} x^i y^j z^k$. The latter condition is often relaxed in weak formulations. A minimum exists even for inconsistent equations. This means, the equations over specify the field. A unique minimum, however, will still be found. The deviation from the consistent result is due to the balancing of each of the conditions. Hence the $\mathbf{B} = \nabla \times \mathbf{A}$ equation may not hold exactly, but to a particular order, depending on the integration method in the weak approach, such as Galerkin method.

The uniqueness and existence conditions are the discretized version of the Hodge-Helmholtz decomposition. [Hodge, 1941, Flanders, 1963] In the continuum case the field $\mathbf{A}(x, y, z)$ is indeed readily decomposed into (Eq. 4.39):

$$(11.56) \qquad \mathbf{A}(x, y, z) = \mathbf{A}_{\parallel}(x, y, z) + \mathbf{A}_{\perp}(x, y, z) + \mathbf{A}_h(x, y, z) \ ,$$

with $\nabla \times \mathbf{A}_{\parallel}(x, y, z) = \nabla \times \nabla \phi = 0$, $\nabla \cdot \mathbf{A}_{\perp}(x, y, z) = 0$, and $\Delta \mathbf{A}_h(x, y, z) = 0$, where the parallel field $\mathbf{A}_{\parallel}$ is fixed by the gauge condition, the rotational field $\mathbf{A}_{\perp}$ by the original Hamiltonian, and the harmonic field $\mathbf{A}_h$ by the boundary conditions. For a particular finite order polynomial approximation this decomposition no longer holds.

We, however, have explicit information about the null-spaces of the element Hamiltonians. The dimension of the null-space is given by d_n, where n is the polynomial degree. The gauge condition, typical of the continuum theory, is no longer required. Assigning the correct number of variables to the boundary will yield a consistent approximation with a unique result for a given boundary value problem. See Figure 11.4.

A three-dimensional domain approximated by N^3 cells, each with a n-th order polynomial basis functions based on equidistant nodes yields:

$$(11.57) \quad \#\text{domain variables} = c_{nN} = 3(nN + 1)^3 = 3n^3 N^3 + 9n^2 N^2 + 9nN + 3 \ .$$

The number of variables on the boundary of the domain:

$$(11.58) \qquad \#\text{domain boundary variables} = b_{nN} = 18(nN)^2 + 6 \ .$$

The number of cells is:

$$(11.59) \qquad \qquad \#\text{cells} = C_N = N^3 \ .$$

The number of variables on a single cell:

$$(11.60) \qquad \#\text{cell variables} = c_n = 3(n + 1)^3 = 3n^3 + 9n^2 + 9n + 3 \ .$$

The number of variables on the boundary of a cell:

$$(11.61) \qquad \qquad \#\text{cell boundary variables} = b_n = 18n^2 + 6 \ .$$

For each cell, with n-th order polynomials in each direction lead to the conditions:

$$(11.62) \qquad \text{rank}\{\nabla\times\}_{\text{cell}} + \text{rank}\{\partial\}_{\text{cell}} = c_n - d_n + b_n > c_n \quad,$$

and

$$(11.63) \qquad \text{rank}\{\nabla\times, \partial\}_{\text{cell}} = c_n \quad.$$

Since the null-space dimension d_n is smaller than the number of boundary variables b_n, the first condition is not satisfied. For a single cell, with the typical boundary conditions, the number of boundary conditions is larger than the number of undetermined variables, in the null-space of the Hamiltonian.

In the case of N cells, each with n-th order polynomial interpolations, there are c_{Nn} variables, of which b_{Nn} on the boundary. The dimension of the rank$\{\nabla\times\}$ is N^3 times the image space of a single cell operator:

$$(11.64) \qquad \text{rank}\{\nabla\times\}_{\text{domain}} = N^3(c_n - d_n) \quad.$$

The rank of the domain boundary operator is:

$$(11.65) \qquad \text{rank}\{\partial\}_{\text{domain}} = b_{Nn} \quad.$$

Hence for a large number of cells the number of degrees of freedom c_{nN} minus the number of relations from the single-cell Hamiltonians, rank$\{\nabla\times\}_{\text{domain}}$, grows like $N^3 n^2$. If a single cell yields an inconsistent condition, so does a domain of cells:

$$(11.66) \qquad \text{rank}\{\nabla\times\}_{\text{domain}} + \text{rank}\{\partial\}_{\text{domain}} > \text{rank}\{\nabla\times, \partial\}_{\text{domain}} \quad.$$

Besides the cumbersome counting to keep track of the variables and the relations, from the local Hamiltonian and the cell and domain boundaries, the inconsistent result is unwanted. Much more elegant is a treatment based on a single cell, which can be connected consistently with other cells and boundaries. Such a single cell, with port-variables common to adjacent cells, we will call an element port Hamiltonian. The proper choice of port variables covers the null-space of the Hamiltonian. The interconnection of these variables may lead to a network of algebraic relations, depending on the boundary conditions, which connect the system to its surroundings.

11.8. The element Hamiltonian as port Hamiltonian

Constructing full finite-element Hamiltonians from cells, which fill a domain, will lead to consistency problems, as we have seen in the previous sections. The consistency problem manifests itself in an overdetermined set of equations, which therefore are no longer a finite-dimensional subspace of the infinite-dimensional space with exact representations of the differential operators, necessary to retain conservation laws. The polynomial functions make sure the differential operators, in the Hamiltonian and the conservation laws, have exact solutions. The mapping back onto the finite-dimensional coefficient vector must be bijective. The overdetermined set of equations leads to a projection.

Besides consistency problems, the typical formulation of the partial differential equation is the dressed-down version of the partial differential equation that would arise in interacting systems. The boundary conditions are typically in terms of the fields, or state variables. Energy conservation, including the inflow and outflow

through the boundary, will be cumbersome integrals of these fields over the boundary. [Sakamoto, 1982] Introducing physical boundary conditions would simplify the expressions for the energy flow through the boundary. The appropriate, physical boundary condition is some operator $\mathbf{B}$, and the finite-dimensional representation we seek reduces the problem:

$$(11.67) \qquad \min_{\mathbf{A}} H(\mathbf{A}), \;\; \text{given} \;\; \mathbf{BA} = \mathbf{b} \;\; \Rightarrow \;\; \min_{\mathbf{x}} H(\mathbf{b}, \mathbf{x}) \;\; ,$$

where the Hamiltonian is expressed in terms of the boundary condition $\mathbf{b}$ and possibly additional internal variables $\mathbf{x}$, which is a finite-dimensional representation of the fields $\mathbf{A}(\mathbf{b}, \mathbf{x})$.

Some advantages of the form $H(\mathbf{b}, \mathbf{x})$ are immediately clear. First of all, for a given boundary condition $\mathbf{b}$, the Hamiltonian is only a function of $\mathbf{x}$, with a parametric dependence on $\mathbf{b}$. Furthermore, it is an unrestricted function of $\mathbf{x}$; in a time-dependent problem initial conditions are a value for the internal state $\mathbf{x}$ at the initial time. For a domain, consisting of cells, the internal boundaries are implemented by assigning $\mathbf{T}_{ji}\mathbf{b}_i = \mathbf{T}_{ij}\mathbf{b}_j$, where $\mathbf{T}_{ji}$ and $\mathbf{T}_{ij}$ identify the common boundary for adjacent cells i and j. The interaction of the system with its surroundings occurs in terms of the variables which lead to a simple bilinear product, representing the power flow through the boundary.

The typical boundary values $\mathbf{b}$ are the force, the pressure, and the flow through the boundary. These variables have more meaning in the global setting than the state variables, such as density, flux, stress, in which the partial differential equation and consequently the boundary conditions are normally expressed. In order to arrive at these boundary variables, the equations of motion play a central role, both in the derivation as in retaining consistency. The typical equations of motion of a hyperbolic system are more general than the pseudo-symplectic port-Hamiltonian equations (Eq. 7.28) of motion as the pair of operators: D and $D^c \neq D^*$ do not have to be the formal adjoint (Eq. 2.69) of each other. Different boundary conditions lead to different operators D and D^c, as we have seen in the section of initial boundary value problems, Section 5.3.

The general equations of motion for a hyperbolic system are:

$$(11.68) \qquad \begin{pmatrix} \dot{q}(z) \\ \dot{p}(z) \end{pmatrix} = \begin{pmatrix} 0 & D \\ -D^c & 0 \end{pmatrix} \begin{pmatrix} \delta_{D^{c*}q(z)}H \\ \delta_{D^*p(z)}H \end{pmatrix} \;\; ,$$

where D and D^c are positive and negative semi-definite operators associated with the kinetic and the potential energy respectively. The discretized, finite dimensional approximation of these equations is:

$$(11.69) \qquad \begin{pmatrix} \dot{\mathbf{q}} \\ \dot{\mathbf{p}} \end{pmatrix} = \begin{pmatrix} 0 & \mathbf{D} \\ -\mathbf{D}^c & 0 \end{pmatrix} \begin{pmatrix} \nabla_{\mathbf{D}^{c*}\mathbf{q}}H \\ \nabla_{\mathbf{D}^*\mathbf{p}}H \end{pmatrix} \;\; ,$$

where the consistent dimensions $\mathbf{q} \in \mathbb{R}^r$ and $\mathbf{p} \in \mathbb{R}^s$ are yet to be determined from the properties of the matrices $\mathbf{D}$ and $\mathbf{D}^c$. Although in the function space $q(z)$ and $p(z)$ may lead to complicate forms of $\delta_{q(z)}H$ and $\delta_{p(z)}H$ for nonlinear and position-dependent Hamiltonians, the local dimension, in configuration space, of $\nabla_{\mathbf{q}}H$ and $\nabla_{\mathbf{p}}H$ are determined by the dimensions of r, s of $\mathbf{q}$ and $\mathbf{p}$. The mapping back the original function spaces $q(z)$ and $p(z)$ and their finite-dimensional representations $\phi_i(z)$ and $\psi_i(z)$ given in terms of the coefficients $\mathbf{q}$ and $\mathbf{p}$ must be seen and treated as an approximation of the Hamiltonian or the closure relations.

Hence the conserved energy is no longer determined directly by $q(z)$ and $p(z)$, but in the finite-dimensional representation.

The dimensions of the operators $\mathbf{D}$ and $\mathbf{D}^c$ determine the boundary variables $(\nabla_{\mathbf{q}}H, \nabla_{\mathbf{p}}H) \to \mathbf{b}$. The simplest procedure is to determine the time-dependence of the Hamiltonian:

$$(11.70) \qquad \dot{H} = \nabla_{\mathbf{q}}^T H \mathbf{Y} \mathbf{D} \nabla_{\mathbf{D}^*\mathbf{p}} H - \nabla_{\mathbf{D}^{c*}\mathbf{q}}^T H \mathbf{D}^{cT} \mathbf{Y} \nabla_{\mathbf{p}} H \quad,$$

where $\mathbf{Y}$ is the matrix consisting of the integrals over products of basis functions:

$$(11.71) \qquad Y_{ij} = \int dz \phi_i(z) \psi_j(z) \quad.$$

If the ϕ_i and ψ_i are limited respectively to Nth order and Mth order polynomials, the integration must be exact of $(N + M)$th order polynomials to guarantee the exact power conservation.

The finite-dimensional boundary operator is the matrix $\mathbf{B}$ associated with the bilinear power flow $\langle \nabla_{\mathbf{D}^{c*}\mathbf{q}} H, \nabla_{\mathbf{D}^*\mathbf{p}} H \rangle_{\mathbf{B}}$:

$$(11.72) \qquad \mathbf{B} = \mathbf{P}_{\mathbf{D}^c}^T \mathbf{D}^{cT} \mathbf{Y} \mathbf{D} - \mathbf{D}^{cT} \mathbf{Y} \mathbf{D} \mathbf{P}_{\mathbf{D}} \quad,$$

where the projection operators $\mathbf{P}_{\mathbf{D}^c}$ and $\mathbf{P}_{\mathbf{D}}$ project perpendicular to the null-space of $\mathbf{D}^c$ and $\mathbf{D}$ respectively, which arises from the change of variation. This is the difference between the normal variation and the differential variation, related to the formal adjoint (Eq. 2.69):

$$(11.73) \qquad D^* \delta_{Dq} H \subset \delta_q H \quad.$$

The two variations are identical for $Dq \neq 0$.

The special nature of the boundary operator is only clear from the actual representations of the differential operators, D and D^c, and the underlying finite-dimensional subspaces, spanned by ϕ_i and ψ_i.

The image spaces of D and D^c can be embedded in the polynomial domain spaces, with a Cartesian vector space when appropriate, such as in the case of the gradient operator:

$$(11.74) \qquad \nabla \mathcal{P}_N^3 : \nabla \sum_{i,j,k=0}^{N} a_{ijk} x^i y^j z^k = \sum_{i,j,k=0}^{N} \mathbf{c}_{ijk} x^i y^j z^k \subset \mathbb{R}^3 \otimes \mathcal{P}_N^3 \quad,$$

with $\mathbf{c}_{ijk}$ being a vector in the direction of the gradient. Constant differential operators acting on polynomials spaces $\mathcal{P}$ can always be embedded in the original space:

$$(11.75) \qquad D\mathcal{P} \subset \mathbb{E} \otimes \mathcal{P} \quad,$$

where $\mathbb{E}$ is a vector space associated with the operator. Making, a priori a distinction between the domain space and the image space, as has been investigated in the case of the $\nabla \times$ operator, in Section 11.7, turns out to be rather cumbersome with little gain, especially in the generic hyperbolic case $D^c \neq D^*$, investigated here. In the case of the port-Hamiltonian $D^c = D^*$ it would have been naturally to restrict the function spaces accordingly to some bilinear product space, in the port-Hamiltonian formulation called the power product. The two function spaces ϕ_i and ψ_i are the natural dual of each other. Much work has been invested in linking the underlying geometry with the function space using algebraic geometry. [Hodge, 1941, Whitney, 1957, Morita, 2001, Bossavit, 1998] However, as we project the infinite-dimensional function space $(q(z), p(z))$ onto a finite-dimensional space

$(\mathbf{q}, \mathbf{p})$ we will spare ourselves the trouble and deal with matrix-type operators only, after the initial mapping of D and D^c onto $\mathbf{D}$ and $\mathbf{D}^c$.

However, for example, for the most common formulation of hyperbolic systems, in this case the linear wave equation, the operators are not the adjoint of each other, $D = \mathbf{1}$ and $D^c = -\Delta$:

$$(11.76) \qquad \left(\begin{array}{c} \dot{q}(z) \\ \dot{p}(z) \end{array} \right) = \left(\begin{array}{cc} 0 & \mathbf{1} \\ \Delta & 0 \end{array} \right) \left(\begin{array}{c} q(z) \\ p(z) \end{array} \right) \quad ,$$

noting that the Laplacian is a negative operator, without positive eigenvalues. If the boundary operator $\mathbf{B}$ is to be restricted to the part of the equations which are not determined by the equations of motion $\mathbf{Y}$ should not be symmetric, and the subspace ϕ_i and ψ_i not equivalent. Only the projection operator $\mathbf{P}_{\mathbf{D}^c}$ will be nontrivial, while $\mathbf{P}_{\mathbf{D}} = \mathbf{1}$.

The function $q(z)$ needs to be augmented with boundary conditions in order to yield a unique map onto $\dot{p}(z)$. These are the boundary conditions for the elliptic problem $\Delta q(z) = 0$, such as the Dirichlet or Neumann conditions.

However, in a more general setting, instead of fixing the boundary conditions for $q(z)$, the equations can be made complete by adding an applied force to the incomplete force, or Newton, equation, which is the second row:

$$(11.77) \qquad \left(\begin{array}{c} \dot{q}(z) \\ \dot{p}(z) \end{array} \right) = \left(\begin{array}{cc} 0 & \mathbf{1} \\ \Delta & 0 \end{array} \right) \left(\begin{array}{c} q(z) \\ p(z) \end{array} \right) + \left(\begin{array}{c} 0 \\ F(z) \end{array} \right) \quad ,$$

where the force $F(z)$ lies in the left null-space of Δ. Although applied force $F(z)$ may appear reasonable, it will not convergence consistently, in some norm, to the boundary integrals in any function space. The left null-space will include the highest-order terms of the polynomial space, and consequently, a consistent $F(z)$, within the same function space, will be highly oscillatory, if the polynomial space is sufficiently high.

The right null-space of Δ, on the other hand, contains only the constant and linear functions. The null-space solution $q_0(z)$ can be added to the state $q(z)$ without changing the dynamics $(\dot{q}(z), \dot{p}(z))$. However, the time-variation of the null-space solution $\dot{q}_0$ can be traced back to the momentum $p_0(z)$. The core model, Eq. 7.53, contains only $q_0(z)$, the right null-space of Δ. The dimension of the null-space determines the number of boundary, or port, variables. The mapping to the discrete vectors $\mathbf{q}_0$ and $\mathbf{p}_0$ must be such that a network arises.

The geometry of port Hamiltonian elements is usually given by nodes, edges, or faces. On its own is a single element just a blob where the structure has little significance, but connecting elements together yields a geometry. A network arises from the common, or connected, nodes, edges, and faces. If only two elements are connected with another, the joint, or port, is a simple identification of variables. Potentials and densities are the same, flows and forces are opposite if the element normals $\mathbf{n}$ are opposite, which depends on the type of joint. The joint is the geometric aspect of the port. Connecting three or more elements in a single joint requires some joint model. The identification of potentials and densities is straightforward, but the extension of a normal on a boundary of two elements to the common edge of three of more elements is ambiguous.

The right null-spaces of $\mathbf{D}$ and $\mathbf{D}^c$ yield the states of the core model. These should be mapped to port variables, while the remainder of the variables vanish at these joints. For example, in one spatial dimension $z \in [0, 1]$, with $D^c = -\partial_z^2$, the

null-space consists of the span$\{1, z\}$. The typical decomposition would be of a 4-th order polynomial basis, on 5 equidistant nodes $z = \{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1\}$ would be:

$$(11.78) \qquad \begin{pmatrix} \phi_0^1 \\ \phi_0^2 \\ \phi_1 \\ \phi_2 \\ \phi_3 \end{pmatrix} = \begin{pmatrix} 1 - z \\ z \\ (1-z)z\left(z - \frac{1}{2}\right)\left(z - \frac{3}{4}\right) \\ (1-z)z\left(z - \frac{1}{4}\right)\left(z - \frac{3}{4}\right) \\ (1-z)z\left(z - \frac{1}{4}\right)\left(z - \frac{1}{2}\right) \end{pmatrix} \quad .$$

The internal modes ϕ_1, ϕ_2, and ϕ_3 vanish at the port nodes $z = 0$ and $z = 1$, and the boundary, or port, modes ϕ_0^1 and ϕ_0^2 vanish for the D^c operator: $D^c\phi_0^1 = 0$ and $D^c\phi_0^2 = 0$. Furthermore, the null-space of D^c is nicely separated in these two modes each based at a single port node, ϕ_0^1 at $z = 0$ and ϕ_0^2 at $z = 1$. The $\mathbf{D}^c$ matrix has two zero columns, associated with the right null-space:

$$(11.79) \qquad \mathbf{D}^c = - \begin{pmatrix} 0 & 0 & -\frac{13}{4} & -\frac{19}{8} & -\frac{7}{4} \\ 0 & 0 & -\frac{7}{4} & -\frac{19}{8} & -\frac{13}{4} \\ 0 & 0 & 96 & 96 & 96 \\ 0 & 0 & -192 & -192 & -192 \\ 0 & 0 & 96 & 96 & 96 \end{pmatrix} \quad .$$

The boundary operator is:

$$(11.80) \qquad \mathbf{B} = -(\mathbf{1} - \mathbf{P_{D^c}})\mathbf{D}^c\mathbf{Y} \quad ,$$

since $\mathbf{D} = \mathbf{P_D} = \mathbf{1}$.

11.9. From null-spaces to port variables

The null-space of the differential operator D is the null-space of the Hamiltonian based on that operator. If the equations of motion are the Hamilton equations of motion, the dynamics will not depend on degrees of freedom in the left null-space, and degrees of freedom in the dual of right null-space are not determined by the Hamilton equations of motion. The left and the right null space are connected if the dual is based on the inner product of the exact integration in the polynomial space.

In order to uniquely specify the dynamics of the full space, including the null-spaces, the system needs to be augmented with boundary conditions or constants of motion. In case of discrete models, at the basis of a FEM approach, it is useful to assign the null-space to a boundary node, as we have seen in the previous section. The typical approaches will lead to an identification of the left null-space of D, based on the observability argument from Section 3.3. A control vector is added which determines the rate of change of the state vector $\mathbf{x}$ outside the observability range since it lies in the left null-space of $\mathbf{A}$.

However, the left null-space will lead to highly oscillatory functions in a polynomial space, as the highest-order polynomials vanish under differentiation. See also Figure 5.1. The degree of polynomials are lowered under differentiation, hence, the highest degrees lie in the image, or left, null-space. The norm, such as the Lebesgue norm, determines the precise form of the null-space functions. The right null-space is much more smooth and appropriate, and more closely related to lumped approaches which will dominate the reduced-order models. In order to be able to use the right null-space, it is assigned to a boundary node, and the remainder of the function space orthogonalized with respect to that projection.

A node becomes a port by bi-orthogonalization of the remainder rather than the mode itself, typical for multi-component structural analysis, such as Craig-Bampton. [Craig and Chang, 1977, Craig, 1981] The modes associated with the null-space are attached to particular nodes, the port nodes. The other, internal modes have to vanish at the position of these port nodes, in order retain power continuity.

The transformation matrix $\mathbf{T}_0$ from the coefficients of a polynomial expansion $\phi_i(z) = z^i$ of the modes to the values at the nodes is given by:

$$(11.81) \qquad \mathbf{T}_0 = z_{\mathbf{i}}^{\mathbf{j}} = \begin{pmatrix} 1 & z_1 & z_1^2 & \cdots \\ 1 & z_2 & z_2^2 & \cdots \\ 1 & z_3 & z_3^2 & \cdots \\ \vdots & \vdots & & \end{pmatrix} ,$$

which is non-singular for $z_i \neq z_j$. The subscript zero in $\mathbf{T}_0$ indicates the basic modes, not the port modes and the remainder.

The derivative in terms of the node values is given by:

$$(11.82) \qquad \mathbf{D}_0 = \mathbf{T}_0 \begin{pmatrix} 0 & 1 & 0 & \cdots \\ 0 & 0 & 2 & 0 \\ \cdots & 0 & 0 & 3 \\ & & \vdots & \end{pmatrix} \mathbf{T}_0^{-1} = \mathbf{T}_0 \mathbf{d} \mathbf{T}_0^{-1} .$$

The null-space is the constant function, $\phi_0(z) = 1$, which in the node vector is given by $\mathbf{q}_0 = (1, 1, 1, 1, \cdots)$; the first column of $\mathbf{T}_0$, such that $\mathbf{D}_0 \mathbf{q}_0 = 0$. If the value at z_1 is the port variable, the other modes have to vanish at z_1. The transformation matrix $\mathbf{T}$ from the modes to the nodal values is therefore of the following structure:

$$(11.83) \quad \mathbf{T} = \begin{pmatrix} 1 & 0 & 0 & \cdots \\ 1 & z_2 - z_1 & z_2^2 - z_1^2 & \cdots \\ 1 & z_3 - z_1 & z_3^2 - z_1^2 & \cdots \\ \vdots & \vdots & & \end{pmatrix} = \mathbf{T}_0 \begin{pmatrix} 1 & -z_1 & -z_1^2 & \cdots \\ 0 & 1 & 0 & \cdots \\ \vdots & 0 & 1 & \cdots \\ & & & \ddots \end{pmatrix} ,$$

which is the matrix equivalent of subtraction of the Taylor expansion (Eq. 2.76) around the port node z_1.

In the case of a set of ordered nodes $z_1 < z_2 < z_3 \cdots < z_N$, the input port is z_1, the output can be any port, however, typically the last one z_N is picked. The consistent theory of such elements requires two matrices. The first matrix is the differential operator in the space of the particular modes, given by the transformation matrix $\mathbf{T}$:

$$(11.84) \qquad \mathbf{D} = \mathbf{T} \mathbf{d} \mathbf{T}^{-1} .$$

The other matrix is the exact integration of the modes on the space to define the metric:

$$(11.85) \qquad \mathbf{Y} = \mathbf{T} \left(\int_a^b z^{\mathbf{i}+\mathbf{j}} dz \right) \mathbf{T}^{-1} = \mathbf{T} \left(\frac{b^{\mathbf{i}+\mathbf{j}+1} - a^{\mathbf{i}+\mathbf{j}+1}}{\mathbf{i}+\mathbf{j}+1} \right) \mathbf{T}^{-1} ,$$
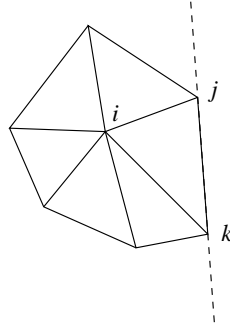
FIGURE 11.5. A patch associated with the central point i is made up of a polynomial of equations of lines constituting the boundary of the domain. The order d of the polynomial depends on the required smoothness.

where $[a, b]$ is the domain. The differentiation and integration carries over from the continuous space:

(11.86)
$$\partial_z \int \phi_i(z)\phi_j(z)dz = \int (\phi_i(z)\partial_z \phi_j(z) + (\partial_z \phi_i(z))\phi_j(z))dz \to [\mathbf{D}^T \mathbf{Y} + \mathbf{Y}\mathbf{D}]_{ij} \quad .$$

11.10. Plate port-Hamiltonian element

For small deflections of a plate from the flat surface $z(x, y) = 0$ the elastic energy of a plate is given by the bi-harmonic form:

(11.87)
$$E_{\text{elastic}} = \int dx dy (\Delta z)^2 \quad ,$$

where the bending stiffness is set to unity, to simplify the equations. Hence the values at the boundary should not only be equal, but also the derivatives across the boundary should be equal. It is the extension of the Euler-Bernoulli beam to two spatial dimensions. The differential operator D is the Laplacian.

The null-space up to second order, in each direction, is five dimensional:

(11.88)
$$\mathcal{N} = \{1, x, y, xy, x^2 - y^2\} \quad .$$

The only nontrivial monomials are $x^2 + y^2$, $x^2 y$, xy^2, and $x^2 y^2$. See Section 11.7.

The smoothness condition for the boundary, such that the boundary energy vanishes (see section 11.6), requires that the derivatives across the boundary are continuous. Similar to the spline (Section 11.1) it is impossible to construct such a solution, with second order polynomials alone. For example a solution could be: $(x, y) \in [0, 1] \otimes [0, 1]$

(11.89)
$$\phi_{\text{internal}}^{\square}(x, y) = (x(1 - x)y(1 - y))^2 \quad ,$$

which requires at least fourth-order terms, x^4, y^4, in the expansion. For a triangular surface the situation is similar, for example:

(11.90)
$$\phi_{\text{internal}}^{\Delta}(x, y) = (xy(1 - x - y))^2 \quad ,$$

for $0 < x < 1$ and $0 < y < 1 - x$, which only have fewer cross terms like $x^i y^i$, with $i + j > 4$.
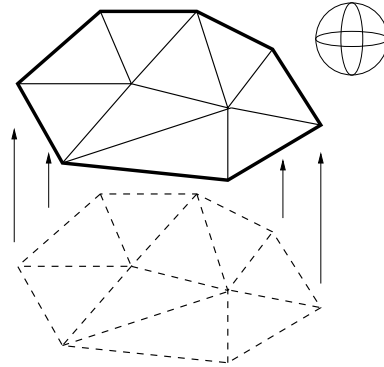
FIGURE 11.6. The triangulation of a sphere can be constructed from matching a triangulated domain with its copy. The vertices and edges at the boundary count for one, all other vertices, edges, and faces occur twice in the sphere, with Euler characteristic $\chi_E = 2$.

Furthermore, even is an element does not have any non-trivial internal modes, which vanish at the boundary, a consistent port-Hamiltonian element may arise. However, in the case of a second-order expansion, the dimension of the null-space is more than half the total dimension. If the simply-connected surface is triangulated, the Euler characteristic, Eq. 8.78, determines:

$$(11.91) \qquad 2 = 2 - \frac{1}{2}V_\text{b} + \frac{1}{2}E_\text{b} = V_\text{i} - E_\text{i} + F \ ,$$

where V, E, and F are vertices, or nodes, edges, and faces respectively, and the subscripts "i" and "b" stand for internal and boundary respectively. For a typical triangulation of a surface a vertex is connected on average to six edges, hence $3V_\text{i} \approx E_\text{i}$ which reduces the internal part of the degrees of freedom to:

$$(11.92) \qquad 2 \times \text{ number of ports} = 2V_\text{i} = F = \text{number of elements} \ .$$

See Figure 11.6.

Every port is a relation between an input and an output. If the state of an element is given by a number of variables, the number of variables in the null-space must be determined by the relations among the port variables. If $z(x, y)$ is fully specified by the values at the vertices, five of the nine degrees of freedom, corresponding to the null-space, are to be determined by the ports, and the remaining four should yield output at the ports, which cannot be.

Starting at degrees of freedom at each vertex (x, y), which we will use as ports, Hence, a priori definitions of low-order modes will lead to inconsistencies in the numbers of degrees of freedom we can assign to ports. The dimension of the null-space is too large compared to the total number of degrees of freedom. For a one-dimensional reference space, like a signal cable, half of the degrees of freedom from the null-space are input, the other half is output. For higher dimensions, the number of input is smaller, therefore the relative dimension of the null-space must be smaller.

A locally smooth surface is defined by its value $z(x, y)$ and a surface unit normal $\mathbf{n}(x, y)$ with $|\mathbf{n}(x, y)| = 1$. In total three degrees of freedom define a locally smooth
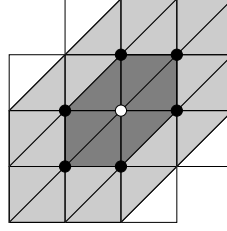
FIGURE 11.7. The white vertex in the center is associated with the dark-grey patch. The stiffness matrix will have non-zero entries due to overlap with the light-grey patches with the six black vertices.

surface, which may act as port variables. The values at the common edge are to be determined solely by the two vertices, at the two ends. However, this is usually not possible.

One procedure to define such interpolations is to define a patch around the vertex, on the domain of faces common to the vertex.

$$(11.93) \qquad \phi_{\text{patch } i}(x, y) = (c_i + a_i x + b_i y) \prod_{\langle j,k \rangle} (\alpha_{\langle j,k \rangle} x + \beta_{\langle j,k \rangle} y + \gamma_{\langle j,k \rangle})^d \quad ,$$

where $d = 2$ for the plate and $\alpha_{\langle j,k \rangle} x + \beta_{\langle j,k \rangle} y + \gamma_{\langle j,k \rangle} = 0$, with $\beta_{\langle j,k \rangle}^2 + \alpha_{\langle j,k \rangle}^2 = 1$, defines a line through the vertices j and k, on an edge $\langle j, k \rangle$ at the boundary of the patch domain. See Figure 11.5. The constants a_i, b_i, and c_i determine the value z and the slope $\mathbf{n}$ at the central vertex $i = (x, y)$. Outside the domain the the function is zero. The patch is d-times differentiable.

Such a patch depends on all the neighboring faces. The port is not an intrinsic quality of a triangular element; it depends on the number of common faces attached to the element. If such elements are used to define a stiffness matrix, it is more closely related to a finite-difference method than a finite-volume method. Furthermore, the order of the polynomial grows with the number of common faces. See also Section 8.7. The stiffness matrix follows from the definition:

$$(11.94) \qquad \frac{\partial^2}{\partial(a_i, b_i, c_i)\partial(a_j, b_j, c_j)} \int dx dy \left(\Delta \sum_k \phi_{\text{patch } k} \right)^2 = \mathbf{K} \quad .$$

See Figure 11.7.

The patches used here will not sum to a constant function for constant coefficients, which is a common requirement. See Figure 11.8. The elastic energy will not vanish if the patches do not sum to a constant. Normally, for higher-order elements splines are used to construct a piece-wise lower-order polynomial which will sum to a constant function. Such splines extend further over a mesh, generating a stiffness matrix with more non-zero elements. Furthermore, for irregular triangulated meshes the procedure to construct the patch is rather complex.

Using only the surface normal $\mathbf{n}(x, y)$ alone is a much simpler approach to defining an element Hamiltonian. In Section 4.3 we have seen that a surface normal determines the surface $z(x, y)$ uniquely up to a constant displacement z_0. Furthermore, the linear interpolation of the surface normal between vertices automatically yields a composite surface, consisting of surfaces on each element which has no
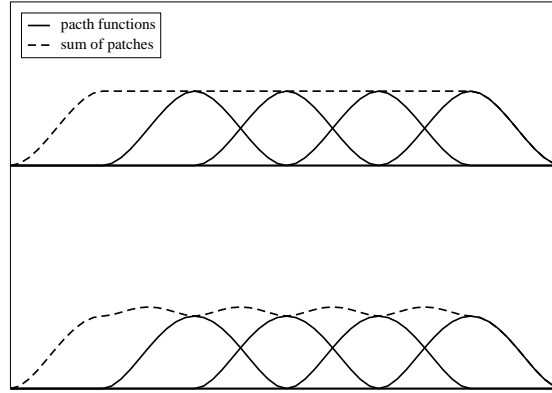
FIGURE 11.8. The patch functions may not sum to a constant function, as is the case for the lower set with $16(z-2)^2 z^2$ patches on $z \in [0, 2]$. The upper set of patches consists of piece-wise $3z^2 - 2z^3$ on $z \in [0, 1]$ and $3(2-z)^2 - 2(2-z)^3)$ on $z \in [1, 2]$. A spline similar to these patches, with a continuous derivative, would consists of three or four polynomial pieces.[Ahlberg et al., 1967]

boundary energy. We only have to define, or approximate, the integration around the edge of the element which is conservative, such that $z(x', y') - z(x, y)$ does not depend on the path of integration between the points (x, y) and (x', y'). This is the integrability condition on the finite-dimensional functional subspace. See Figure 11.9.

For the biharmonic approximation of the plate it is assumed that the deviation from $z = 0$ is relatively small. Hence likewise we can approximate the normal $\mathbf{n}(x, y) \approx \mathbf{n} = (1, n_x(x, y), n_y(x, y))$, such that z is given by the integral:

$$(11.95) \qquad z(x_2, y_2) - z(x_1, y_1) = - \int_{x_1, y_1}^{x_2, y_2} \mathbf{n}^T d\mathbf{l}(x, y) \quad ,$$

where $\mathbf{l}(x, y)$ is a line element along the path from $z(x_1, y_1)$ to $z(x_2, y_2)$. The integrability condition reduces to:

$$(11.96) \qquad \partial_x n_y - \partial_y n_x = 0 \quad ,$$

such that all closed contours are zero.

If $n_x(x, y)$ and $n_y(x, y)$ are piecewise linear functions on a triangle, it means there is a single condition among the six variables defining the normal:

$$(11.97) \quad \partial_x n_y - \partial_y n_x = \partial_x(c_y + c_{yx}x + c_{yy}y) - \partial_y(c_x + c_{xy}y + c_{xx}x) = c_{yx} - c_{xy} = 0 \quad .$$

Hence, the surface $z(x, y)$ on the triangle is defined by six variables in total: the integration constant $z(x_0, y_0)$ at a particular position x_0, y_0, and five coefficients $c_y, c_{yx} = c_{xy}, c_{yy}, c_x, c_{xx}$ for the normal.

The elastic energy in terms of the expansion of the normal $\mathbf{n}$ is given by:

$$(11.98) \qquad E_{\text{elastic}} = \frac{1}{2} \int_\Omega dx dy (\nabla \cdot \mathbf{n})^2 = \frac{1}{2} \text{Vol}_\Omega (c_{xx} + c_{yy})^2 \quad .$$
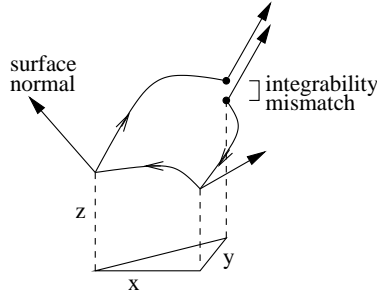
FIGURE 11.9. The implicit definition of a surface, through the surface normal $\mathbf{n}(x,y)$, may not always define a surface. The integrability condition must be satisfied. Otherwise, the integral along a closed contour will not end at the same height z as it started from.

The configuration, the reconstruction of $z(x,y)$ from $\mathbf{n}(x,y)$ and $z(x_0, y_0)$, is given by: $(x_0 = 0, y_0 = 0)$

$$(11.99) \qquad z(x,y) = z(0,0) + c_x x + c_y y + \frac{1}{2}c_{xx}x^2 + \frac{1}{2}c_{yy}y^2 + c_{xy}xy \;\;,$$

which satisfies $\nabla z(x,y) = \mathbf{n}$. This is the small deviation approximation of the mean curvature, Eq. 4.8. A single elastic energy term arises from the plate element, consistent with the null-space, Eq. 11.88. The elastic energy is precisely the single combination of monomials $x^2 + y^2$, which does not lie in the null-space to the Laplace operator. Integrability and consistency are complementary views.

The difference lies in the associated port-variables, in terms of a set of surface normals $n_x(x_1, y_1), n_y(x_1, y_1), n_x(x_2, y_2), n_y(x_2, y_2), n_x(x_3, y_3)$, and $n_y(x_3, y_3)$ at vertices $1, 2$, and 3, where one component of one vertex normal is determined by all the others, and a z-value at a single vertex is the sixth port-variable. Each port is shared by several elements, yielding collocation relations; the same normal is used for to define the surface $z(x,y)$ for several elements. Since, the total number of distinct ports, two per vertex, equals the total number of elements, from Eq. 11.92, up to boundary effects, the one-dimensional state per element of elastic energy is precisely covered by, on average, one distinct port variable per element.

The kinetic energy is associated with the velocity $\dot{z}$. The approximation neglects the horizontal motion in the $x-y$-plane, therefore there is no kinetic energy associated with it, and consequently no rotational energy. The kinetic energy is:

$$(11.100) \qquad E_{\text{kinetic}} = \frac{\rho}{2}\int_\Omega dxdy(\dot{z}(x,y))^2 = \frac{1}{2}\sum_{ij} c_i M_{ij} c_j \;\;,$$

which depends explicitly on the shape of the triangle. The corresponding momenta arise from the derivatives:

$$(11.101) \qquad p_i = \frac{\partial E_{\text{kinetic}}}{\partial c_i} = \sum_j M_{ij} c_j \;\;,$$
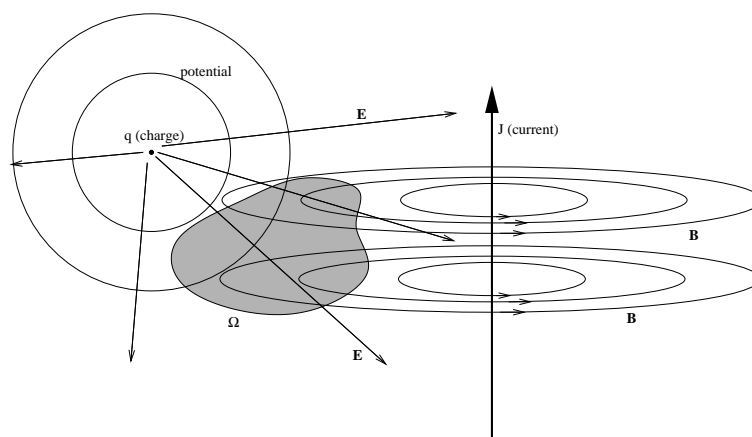
FIGURE 11.10. Static charges and currents outside the domain Ω may cause source-dependent fields $\mathbf{E}$ and $\mathbf{B}$ inside the domain, which are not dynamical, in the sense that the fields depend only on the position and strength of the sources. The field lines are closed in the sense that they enter and leave the domain Ω.

where c_i is any of the coefficients $z(0,0), c_x, c_y, c_{xy}, c_{xx} - c_{yy}$, and $c_{xx} + c_{yy}$. The element port Hamiltonian is:

$$(11.102) \qquad H = \frac{1}{2}\sum_{ij} p_i M_{ij}^{-1} p_j + \frac{1}{2}\text{Vol}_\Omega (c_{xx} + c_{yy})^2 \quad,$$

with the Hamilton equations of motion, and port variables $z(0,0), c_x, c_y, c_{xy}$, and $c_{xx} - c_{yy}$ in the null-space of the elastic energy. The mass matrix M is full rank.

11.11. Electromagnetism on a simplex

Electrodynamics, in terms of the electric and magnetic field $\mathbf{D}$ and $\mathbf{B}$, is a rather complex dynamical problem. The static fields are associated with the static charges and the static currents respectively. These source-dependent fields are longitudinal, i.e. $\nabla \times \mathbf{D} = 0$ and $\nabla \times \mathbf{B} = 0$, inside the simplex, and do not contribute to the dynamics of the electromagnetic fields. See Figure 11.10. Appropriately, the fields corresponding to given charges and currents should be separated from the electrodynamics. They can be treated as ports localized at the position of the charges and currents. This is discussed further in Chapter 12.

The null-space of the $\nabla \times$ operator on the simplex, with a piece-wise linear approximation, gives an alternative view on the separation of dynamical and source-dependent fields. The linear approximation of the fields are given by the twelve values at the vertices; four three-dimensional vectors for $\mathbf{D}$ and likewise for $\mathbf{B}$. The dimension of the null-space of $\nabla \times$ on this space is nine. Therefore only three variables correspond to dynamical, or internal energy on the simplex element. All the other variables will also correspond to energy, but this energy may be determined by just looking at the values on the boundary of the domain.

The fact that only three variables correspond to internal dynamics on a simplex with four faces can be seen heuristically from the fact that the field lines must be

closed on the simplex, hence the flow $\mathbf{E} \cdot \mathbf{n}$ through each of the faces must sum to zero:

$$(11.103) \qquad \sum_{\text{faces}:f} \text{Area}_f \mathbf{E}_f \cdot \mathbf{n}_f = 0 \quad .$$

Therefore one condition exists among the four face flows. In the finite-order approximation, the condition changes in form but not in nature. For the simplicity we consider a simplex with corners $(0,0,0) - (1,0,0) - (0,1,0) - (0,0,1)$. The vector fields at each of the corners are labelled respectively: $\mathbf{E}_0 = \mathbf{E}(0,0,0)$, $\mathbf{E}_x = \mathbf{E}(1,0,0)$, $\mathbf{E}_y = \mathbf{E}(0,1,0)$, $\mathbf{E}_z = \mathbf{E}(0,0,1)$. The linear interpolation yields a vector field inside the simplex:

$$(11.104) \quad \mathbf{E}(x,y,z) = \mathbf{E}_0 + (\mathbf{E}_x - \mathbf{E}_0)x + (\mathbf{E}_y - \mathbf{E}_0)y + (\mathbf{E}_z - \mathbf{E}_0)z = \mathbf{E}_0 + \sum_{i=1}^{3} \mathbf{E}_i x^i \quad ,$$

where $x^1 = x, x^2 = y$, and $x^3 = z$. The surface normals $\mathbf{n}$ are $-\mathbf{e}_x, -\mathbf{e}_y, -\mathbf{e}_z$, and $\frac{1}{\sqrt{3}}(\mathbf{e}_x + \mathbf{e}_y + \mathbf{e}_z)$. The corresponding surface flows are for each of the faces:

$$(11.105)$$
$$\int_{\text{face}} \mathbf{n}^T \mathbf{E}(x,y,z) = \left\{ \begin{array}{ll} \frac{1}{6}\mathbf{e}_z^T(2\mathbf{E}_0 + \mathbf{E}_x + \mathbf{E}_y) & z = 0 \quad (\text{cyclic}\{x,y,z\}) \\ \frac{\sqrt{3}}{6}\mathbf{n}^T(2\mathbf{E}_0 + \mathbf{E}_x + \mathbf{E}_y + \mathbf{E}_z) & x + y + z = 1 \end{array} \right. \quad .$$

Clearly, the total flow in the linear approximation, through the boundary constituted by the four faces, is zero if, and only if, $E_i^i = 0$.

The curl operator $\nabla\times$ maps this vector field onto the constant vector field, the component i is given by:

$$(11.106) \qquad [\nabla \times \mathbf{E}(x,y,z)]_i = \epsilon_{ijk} E_j^k \quad .$$

The components E_0^i and E_i^i are in the null-space of $\nabla\times$. The latter, we have seen, must always be zero, if there is no charge present and the flow is conserved.

The curl operators, part of the equations of motion of electrodynamics:

$$(11.107) \qquad \partial_t \mathbf{D} = \frac{1}{\mu}\nabla \times \mathbf{B} \quad .$$

$$(11.108) \qquad \partial_t \mathbf{B} = -\frac{1}{\epsilon}\nabla \times \mathbf{D} \quad ,$$

are mapped onto $\mathbb{R}^{3\times 6}$ matrices for linear approximation on each simplex:

$$(11.109) \quad \nabla \times \begin{pmatrix} E_3^2 \\ E_2^3 \\ E_1^3 \\ E_3^1 \\ E_2^1 \\ E_1^2 \end{pmatrix} = \begin{pmatrix} -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} E_3^2 \\ E_2^3 \\ E_1^3 \\ E_3^1 \\ E_2^1 \\ E_1^2 \end{pmatrix} = \begin{pmatrix} E_0^1 \\ E_0^2 \\ E_0^3 \end{pmatrix} \quad ,$$

where the finite representation of E_i^j is each of the fields $\mathbf{D}$ and $\mathbf{B}$.

The energy in the simplex is a quadratic function of the local vector field, Eq. 11.104:

$$(11.110)$$
$$E = \frac{1}{2}\int_{\text{simplex}} \mathbf{E}(x,y,z)^2 dx dy dz = \frac{1}{12}\mathbf{E}_0^2 + \sum_i \frac{1}{24}\mathbf{E}_0^T \mathbf{E}_i + \frac{1}{120}\mathbf{E}_i^2 + \sum_{i<j}\frac{1}{120}\mathbf{E}_i^T \mathbf{E}_j \quad .$$

$$(11.111) \qquad\qquad\qquad = \frac{1}{2}\mathbf{E}^T\mathbf{Y}\mathbf{E} \ .$$

If the coefficient vector $\mathbf{E}$ is ordered like: $\mathbf{E} = (E_0^1, E_0^2, E_0^3, E_3^2, E_2^3, E_1^3, E_3^1, E_2^1, E_1^2)$, matrix $\mathbf{Y} \in \mathbb{R}^{9\times 9}$ has the following structure:

$$(11.112) \qquad \mathbf{Y} = \begin{pmatrix}
\frac{1}{12} & 0 & 0 & 0 & 0 & 0 & \frac{1}{48} & \frac{1}{48} & 0 \\
0 & \frac{1}{12} & 0 & \frac{1}{48} & 0 & 0 & 0 & 0 & \frac{1}{48} \\
0 & 0 & \frac{1}{12} & 0 & \frac{1}{48} & \frac{1}{48} & 0 & 0 & 0 \\
0 & \frac{1}{48} & 0 & \frac{1}{120} & 0 & 0 & 0 & 0 & \frac{1}{240} \\
0 & 0 & \frac{1}{48} & 0 & \frac{1}{120} & \frac{1}{240} & 0 & 0 & 0 \\
0 & 0 & \frac{1}{48} & 0 & \frac{1}{240} & \frac{1}{120} & 0 & 0 & 0 \\
\frac{1}{48} & 0 & 0 & 0 & 0 & 0 & \frac{1}{120} & \frac{1}{240} & 0 \\
\frac{1}{48} & 0 & 0 & 0 & 0 & 0 & \frac{1}{240} & \frac{1}{120} & 0 \\
0 & \frac{1}{48} & 0 & \frac{1}{240} & 0 & 0 & 0 & 0 & \frac{1}{120}
\end{pmatrix} \ .$$

The discrete energy, Eq. 11.111, and the discrete equations of motion, Eq. 11.109, yield the discrete electromagnetic power flow in and out of the simplex:

$$(11.113) \quad \dot{E} = \frac{1}{\epsilon}\mathbf{D}\mathbf{Y}\dot{\mathbf{D}} + \frac{1}{\mu}\mathbf{B}\mathbf{Y}\dot{\mathbf{B}} = \frac{1}{\mu\epsilon}\mathbf{D}\mathbf{Y}(\nabla \times \mathbf{B}) - \frac{1}{\mu\epsilon}\mathbf{B}\mathbf{Y}(\nabla \times \mathbf{D}) = \mathbf{D}\theta\mathbf{B} \ .$$

The discrete curl matrix has only a non-zero block in the upper right corner. The boundary operator θ has a complicated but symplectic structure, which follows from the definition:

$$(11.114) \ \theta = \frac{1}{\mu\epsilon} \begin{pmatrix}
0 & 0 & 0 & \frac{1}{12} & -\frac{1}{12} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & \frac{1}{12} & -\frac{1}{12} & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{12} & -\frac{1}{12} \\
-\frac{1}{12} & 0 & 0 & 0 & 0 & \frac{1}{48} & -\frac{1}{24} & -\frac{1}{48} & 0 \\
\frac{1}{12} & 0 & 0 & 0 & 0 & 0 & \frac{1}{48} & \frac{1}{24} & -\frac{1}{48} \\
0 & -\frac{1}{12} & 0 & -\frac{1}{48} & 0 & 0 & 0 & \frac{1}{48} & -\frac{1}{24} \\
0 & \frac{1}{12} & 0 & \frac{1}{24} & -\frac{1}{48} & 0 & 0 & 0 & \frac{1}{48} \\
0 & 0 & -\frac{1}{12} & \frac{1}{48} & -\frac{1}{24} & -\frac{1}{48} & 0 & 0 & 0 \\
0 & 0 & \frac{1}{12} & 0 & \frac{1}{48} & \frac{1}{24} & -\frac{1}{48} & 0 & 0
\end{pmatrix} \ .$$

The rank of the boundary operator is $\text{rank}[\theta] = 6$, compatible with the three-dimensional null-space of $\nabla\times$.

Now we have all the ingredients to construct a port-Hamiltonian element. The energy on the element is expressed in two diagonal blocks $\mathbf{Y}$ on the 18 dimensional $(\mathbf{D}, \mathbf{B})$ space. The equations of motion is expressed in two off-diagonal $\nabla\times$, and the boundary operator is θ. In the case of ordinary finite-element method only the $\nabla\times$ operator is needed for the evolution of the system. In that case, for each node value, e.g., $\mathbf{B}_x$, the time-evolution is given by the mean of all the simplices common to this node i:

$$(11.115) \qquad \mathbf{B}_{x,i}(t + \delta t) - \mathbf{B}_{x,i}(t) = \frac{\delta t}{N} \sum_{\text{simplices}\, j} \frac{1}{\epsilon_j}\nabla \times \mathbf{E}_j \ ,$$

where N is the number of simplices. The average due to the summation over simplices destroys the power continuity and the divergence free form of the fields at the level of the numerical finite difference. If the fields are smooth, and vary slowly over the simplices these qualities are retained to a certain extent.

In the port-Hamiltonian element some of the variables D_i^j and B_j^i are dependent variables and some are independent. The bilinear boundary operator tells us we can

construct six pairs of variables, one dependent and one independent, or one input
and one output. The remaining six variables, from the eighteen, are state variables.
Connecting the elements gives rise to a set of algebraic relations, which gives rise
to an interconnection network. Unlike in most FEM approaches, the manner in
which the elements are interconnected is not a priori fixed, and may depend on the
particular boundary conditions, electric or magnetic.

11.12. States from the null-space

In the case of a nonlinear Hamiltonian the state corresponding to a null-space
mode might be highly nontrivial. However, recovering these states will be important
to simplify the actual numerical implementation and simulations. They form the
linkage between the boundary control in terms of the null-space of the differential
operator D preyed from the Hamiltonian, and the internal dynamics, in terms
of local state variables. For example a null-space may correspond to an applied
voltage V, constant on a domain $z \in [0, 1]$. The charge density $q_0(z)$ may be used
as state corresponding to the null-space. It is related implicitly to the applied
voltage through:

$$(11.116) \qquad\qquad V = \delta_q H[q_0] = \text{constant} \;.$$

The state $\mathbf{q}_0$ may be highly nontrivial to recover. However, it forms the linkage
between the input, the boundary conditions in terms of the physical variables, and
the mathematical model of the internal dynamics, in terms of the state variables.

11.13. Final remarks

The construction of energy-consistent elements and their ports for a general
nonlinear FEM has a number of requirements. Consistent equations of motion, to
study dynamical FEM, which is really the purpose of the whole exercise, are based
on non-singular operators with a unique zero state. Therefore the greater part
of constructing these elements is the counting of degrees of freedom, equations of
motion, and conditions. The linkage with the continuous theory and, for example,
continuity equations goes via integrability conditions, which guarantee that the
variables used correspond to existing configurations and unique energy states.

The construction of finite-elements to be used in a large network is very much a
local approach. The global conservation and balance laws rely on the proper imple-
mentation of algebraic relations and accurate time-integration. The next chapter
the approach is taken from the other side. The global properties are linked to a
coordinate system based on input and output. The internal dynamics and material
properties result only in local variations. In a sense, this approach is the simplest
geometric formulation of the core model, described in Section 7.6.

CHAPTER 12

Force lines and velocity fields

> Geometry is the art of correct reasoning on incorrect figures.
> [G. PÓLYA, HOW TO SOLVE IT]

A global or system view of several components in conjunction necessarily has less focus on the details. In a port-based approach the energy flow and the balance relations throughout the system are the first concern. Continuous models, on the other hand, are based on the local, small-scale balance and flow. The global properties in a continuous models follow only from the local analysis. In order retain the global view within the local models an approximate global flow and balance can be established first, where the local details will follow from more detailed analysis. An experienced modeler may pick a dominant mode or static state. In this chapter we use results from harmonic analysis and optimatization to establish global geometric system views. See Figure 12.1.

The geometry of an extended object is usually considered independent from the fields, such as displacements, currents, forces, and potentials, defined on the object. They are, however, intimately related through the constitutive relations, implicit in the Hamiltonian. In some cases, for special geometries, coordinate systems are used which simplify the calculation of static solutions, such as polar coordinates for spherical problems.

The input and output of a system; the sources of the fields, are instrumental in determining the relevant field configurations. Hence a coordinate system on an object, directly based on the fields, will be a geometric representation of the Hamiltonian for the given interactions through the input in output. Simply said, for a distributed system, with a single input and a single output, the parallel coordinate, based on the field, maps the distance from input to output in a physical sense. The surfaces perpendicular to the lines from input to output are the perpendicular coordinates in the system. In electric systems they are the potential surfaces. The lines are the streamlines, or integrals, of the vector field. In general, we make little distinction between the vector fields and the lines.

Take for example a strained object due to two opposite forces applied at different positions. The force is conserved. On each closed surface enclosing only one fixture position, the same amount of total force acts. In the language of Faraday and Maxwell: the force lines are closed. The Poisson ratio determines the spread of the forces lines between the two fixtures. The density of force lines is equal to the local pressure on the surface.

12.1. Electrostatics

The energy of a static electric field can be calculated in different manners. The total energy density of the field $\mathbf{E}$ will give the same result as the charge density ρ
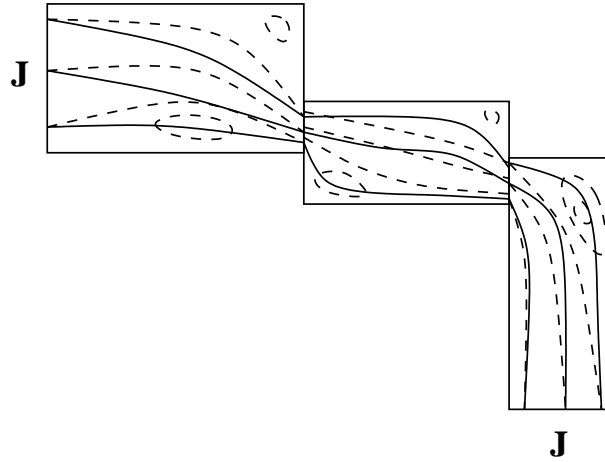
FIGURE 12.1. Given that for a conserved, incompressible flow, the same amount **J** flows in as out, a global solution given by the solid lines is first constructed. The details of the flow in each of the components may change the position of the streamlines, but not the total flow. These are the allowed degrees of freedom of the internal dynamics, for example, those indicated by the dashed lines.

at the given potential V, or the Coulomb law between charges: [Jackson, 1975]
(12.1)
$$E_{\text{electrostatic}} = \frac{1}{2} \int dz \epsilon_0 \mathbf{E}^2(z) = \frac{1}{2} \int dz \rho(z) V(z) = \frac{1}{2} \int dz dz' \frac{\rho(z)\rho(z')}{4\pi\epsilon_0 |z - z'|} \quad .$$

One follows from the other through partial integration and Gauss or Poisson law:

(12.2)
$$\nabla \cdot \mathbf{E} = \frac{1}{4\pi\epsilon_0} \rho \quad .$$

(12.3)
$$-\nabla V = \mathbf{E} \quad ,$$

and the Green function of the Laplacian:

(12.4)
$$f(z) = \Delta_z \int dz' \frac{f(z')}{4\pi |z - z'|} \quad .$$

However, a constant can be added the potential V. This constant is the integration constant from the partial integration, and commonly set such that:

(12.5)
$$\lim_{|z| \to \infty} V(z) = 0 \quad .$$

The surface of the domain is therefore set at infinity: $|z| \to \infty$.

Hence, for a field $\mathbf{E}$ the energy density can be determined by looking only at the charges ρ and the potential at those positions. In most electrostatic problems either the charges ρ or the potential V is given. The energy of the system is the amount of work to bring all the charges from infinity to the given configuration.

The relation between the charge and the potential is given by the capacity C of the system, which is completely determined by the geometry. [Pólya and Szegö, 1951] The potential at the boundary is usually associated with conductors connected

to an external source, and therefore piece-wise constant. The electrostatic co-energy can be given in terms of the potential as well:

$$(12.6) \qquad\qquad E_{\text{electrostatic}} = \frac{1}{2C}V^2 \quad .$$

Since it is a quadratic function the same capacitance C arises as in the energy, expressed in terms of the charges. The capacity C can be determined from the solution of the Laplace equation for piecewise constant boundaries, from the constant potential at the boundary. The Laplace equation is an ideal starting point to characterize the geometry in terms of the boundary.

An arbitrary scalar field f with the appropriate boundary conditions, i.e., port values will give an upper and lower bound for the capacity. [Pólya and Szegö, 1951] The Dirichlet principle is an upper bound for the capacity with $f = V$ at the boundary $\partial\Omega$:

$$(12.7) \qquad\qquad C \le \frac{1}{4\pi V^2} \int_\Omega d^3 z \|\nabla f\|^2 \quad .$$

While, for an arbitrary divergent-free vector field $\mathbf{f}$ the lower bound is given by the Thomson principle:

$$(12.8) \qquad\qquad \frac{1}{C} \le \frac{1}{4\pi Q^2} \int_\Omega d^3 z \|\mathbf{f}\|^2 \quad ,$$

where the charge Q is given by the flow through the boundary $\partial\Omega$:

$$(12.9) \qquad\qquad \frac{1}{4\pi} \int_{\partial\Omega} d^2 z \mathbf{n}^T \mathbf{f} = Q \quad .$$

Hence, by defining divergent free vectors fields $\mathbf{f}$ and scalar fields f the system is already globally characterized, with the use of energy and variational principles. In the next sections we will pursue these principles further. A divergence free vector field, with only in flow and out flow at the boundaries or ports represents any balanced quantity, such as mass, but also force, in the system. Using the conservation law in the analysis will stabilize the numerical schemes. The reason this is normally not pursued is because such conservation laws are global; the whole domain is connected, a change at one place affects changes at other places. Through the use of coordinates based on a particular choice of divergent free field, i.e., the harmonic field, the coordinates can be used as level set for solutions to static, dynamic, and optimization problems.

12.2. Force lines

Maxwell and Faraday introduced the concept of lines of forces, to treat the electrostatic theory in analogy with fluid dynamics. [Maxwell, 1891] The lines of forces are the lines along which a test charge will move in the electric field, perpendicular to the surfaces of equal potential energy. In the more general case forces are still balanced, but there might not be an underlying potential V responsible for the force $\mathbf{F}_{\text{potential}} = q\nabla V$. Euler and Frobenius integrability conditions set rigorous conditions on the fields in order to have an underlying potential. [Frankel, 2004] See also Section 4.4. We will investigate this general case. Our purpose is to decompose the interaction of an object with its environment in a global, transmission part, and local, material effects. The global part is expressed as a harmonic field; solution of

the inhomogeneous Laplace equation, with its useful properties. The local parts of the force field depend on the material.

Given an object with two fixtures, or ports, with equal but opposite force along the line between the fixtures, such that the object is at rest. Through each surface cutting the object in two, separating the two ports goes an equal amount of force lines. The force of the port is distributed in the surface, and depending on the geometry and material properties the force is more or less focused on the shortest path between the two ports. The solution of the Laplacian, using the two ports as sources of force, like charges, the force lines connect the two ports, like the electric field. The conserved force lines are a divergent-free field. They correspond to the stress tensor (Eq. 6.28):

$$(12.10) \qquad \sum_{i=1}^{3} \partial_i S_{ij} = F_j \quad ,$$

where F_j is the applied force, or pressure, at the fixtures.

Hence they can be constructed a force potential $\mathbf{f}$ driven by the input and output forces $\mathbf{F}_{\text{in/out}}$:

$$(12.11) \qquad -\Delta \mathbf{f} = -\nabla \cdot (\nabla \mathbf{f}) = \mathbf{F}_{\text{in/out}} \quad .$$

The surfaces with $\mathbf{f} = $ constant define the parallel coordinates; the relative distance between the ports. The force lines $\mathbf{f}$ associated with the Laplace equation spread out. Near the ports the density of lines, the pressure, is higher, as there is a smaller area of the sphere supporting the total applied force.

The stress tensor can be decomposed in terms of the force lines $\nabla \mathbf{f}$ and the internal balance tensor $\mathbf{G}$, using the Hodge-Helmholtz decomposition (Eq. 4.39):

$$(12.12) \qquad \mathbf{S} = \nabla \mathbf{f} + \nabla \times \mathbf{G} \quad ,$$

or in terms of their components:

$$(12.13) \qquad S_{ij} = \partial_i f_j + \epsilon_{ikl} \partial_k G_{lj} \quad ,$$

where ϵ_{ijk} is the fully anti-symmetric Levi-Civita tensor. The internal balance tensor $\mathbf{G}$ are closed force lines, for each component of the force vector. See Fig. 12.2. The internal balance fields can never contribute to the force at the ports, since $\nabla \cdot (\nabla \times \cdot) = 0$. This makes the force lines $\nabla \mathbf{f}$ with $\Delta \mathbf{f} = \mathbf{F}_{\text{in/out}}$ the natural choice of the principal component of the coordinate system, independent of the material or other constitutive relations. A similar decomposition was also proposed by Weyl . [Weyl, 1940]

The force lines are associated with the harmonic approximation to a problem. In the case of a fixed wave velocity, it would be the solution associated with this wave velocity, propagating along the force lines, for a homogeneous mass density. The internal balance may give rise to auxiliary modes. In the case of the wave equation the force potential generates the static displacement ϕ, to an applied force $\mathbf{F}$:

$$(12.14) \qquad \kappa \Delta \phi = \mathbf{F} \Rightarrow \phi = \frac{\mathbf{e}_z^T \mathbf{f}}{\kappa} \quad .$$

In the case of inhomogeneous elasticity modulus $\kappa(x)$ it should already be taken into account in the Laplacian:

$$(12.15) \qquad \Delta \phi = \nabla \cdot (\kappa(x) \nabla \phi) \quad ,$$

force lines coordinates   internal balance   force field

- - - potential surfaces
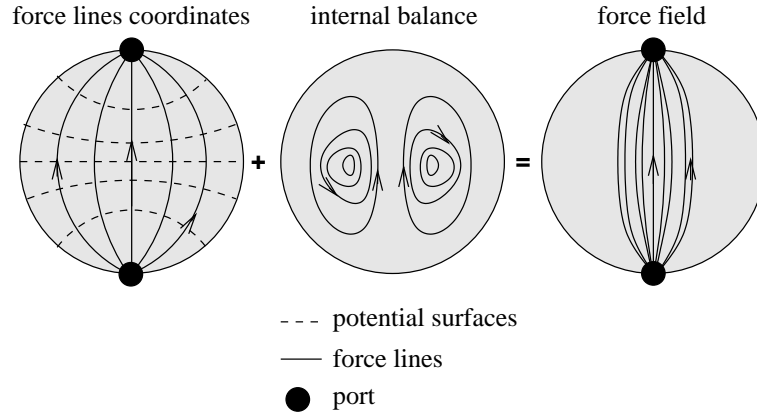
—— force lines

● port

FIGURE 12.2. A sketch of force lines $\nabla \mathbf{f}$, associated with the force potential $\mathbf{f}$, for a material with a small Poisson ratio ν, such that the stress is localized along the axis connecting the ports. The internal balance $\mathbf{G}$ compresses the force lines together. These are only the longitudinal components of the force $\mathbf{f} = f\mathbf{e}_z$.

since the energy is defined as $\kappa(x)(\nabla\phi)^2$, which under partial integration leads to the familiar form.

12.3. Microscopic balance

The force lines are not necessarily parallel to the force direction $\mathbf{f}$, unlike the force lines generated by a potential. The force direction is only the transmission of the input force to the output force. Hence only the force line on the straight line connecting input and output is a balanced force line, in the sense that force line and force direction are parallel:

$$(12.16) \qquad \text{rank}[\nabla \mathbf{f}] = 1 \ .$$

The microscopic balance is automatically satisfied in the formulation in terms of the vector fields $\nabla \mathbf{f}$ and $\nabla \times \mathbf{G}$. The relation between the two is for a greater part determined by the constitutive relations. However, some generic results exist for, for example, elasticity. Such conditions find their basis in the invariant relation between the stress and the strain. For example, results may not depend on the overall rotation of the object, or its reflection. As a consequence, the strain tensor, ϵ_{ij}, is symmetric. If the work W is defined as the scalar product of stress and strain, the stress tensor can be symmetric $S_{ij} = S_{ji}$ too:

$$(12.17) \qquad W = \frac{1}{2} \sum_{ij} \int dz \epsilon_{ij} S_{ij} \ .$$

The asymmetric part will not contribute to the energy. However, this is not a fundamental result, it holds only for isotropic materials. Materials exist with are intrinsically chiral, i.e., left-handed or right-handed, one has only to think of screw-like molecules, typical of biological molecules, and ferrites. For such systems few microscopic theories of the continuous systems exist. [Maugin, 1988]
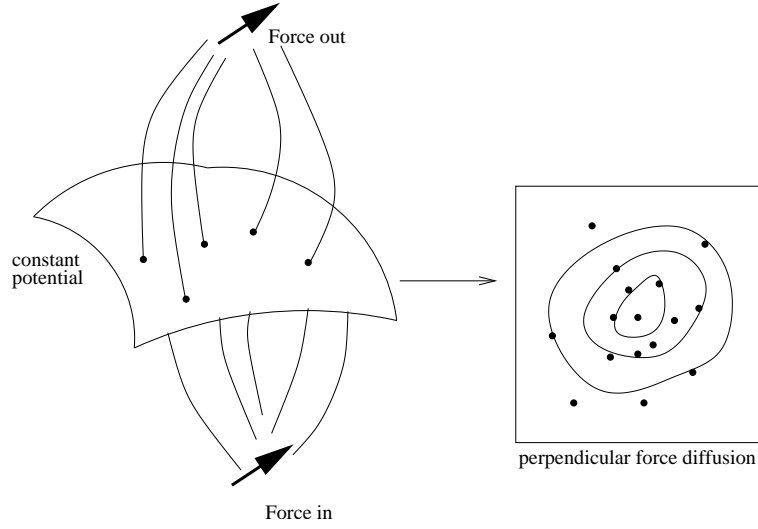
FIGURE 12.3. The force from input to output generates a conserved flow, indicated by the closed force lines. On any surface intersecting the force lines, the material properties will cause a conserved virtual dynamics, i.e., force diffusion, of the force lines. On top of that will internal force balance $\mathbf{G}$ generate forces in perpendicular directions.

On the other hand, energy must be an invariant property, independent of the observer or its coordinate system. Hence the energy must depend on the invariants made with $\mathbf{f}$ and $\mathbf{G}$, where $\mathbf{G}$ is gauge invariant, meaning the energy may not change with the addition of a conservative field $\nabla\chi$ to $\mathbf{G}$:

$$(12.18) \qquad\qquad E(\mathbf{f}, \mathbf{G}) = E(\mathbf{f}, \mathbf{G} + \nabla\chi) \ ,$$

where χ is an arbitrary non-singular scalar vector-valued function.

On a surface $\mathbf{f} = $ constant, the force line $\nabla\mathbf{f}$ is perpendicular to the surface. Furthermore, if the surface encloses the input or the output, the total transmission force $\phi_{\mathbf{f}}^{(0)} = \mathbf{n}^T\nabla\mathbf{f}$ is the same as the input and output force. The transmission force is conserved. The microscopical material properties can only change the distribution of force $\phi_{\mathbf{f}}$ on the surface $\mathbf{f} = $ constant. See Figure 12.3. Hence, although both the surface $\mathbf{f} = $ constant and the harmonic solution $\phi_{\mathbf{f}}^{(0)}$ are global and possibly complex to construct, the variation $\mathbf{G}$ due to material properties is a conserved flow on each surface $\mathbf{f} = $ constant. The redistribution of the force over the surface can be treated as a forced diffusion, where the harmonic force distribution $\phi_{\mathbf{f}}^{(0)}$ and the distribution on neighboring surfaces $\mathbf{f} \pm \delta$ are the source terms. The harmonic solution already yields an appropriate coordinate system for this analysis.

As a simple two-dimensional example we study the force lines in a plane between two point sources at $\mathbf{x}_{\pm} = (\pm 1, 0)$. In the theory of two-dimensional elasticity, different forms of a generating potential, similar to $\mathbf{f}$ are proposed to bring the static deformation equations to a biharmonic form. [Soutas-Little, 1973] In the literature the input is generally not part of the consideration in constructing appropriate fields, the separation in longitudinal and rotational fields is.
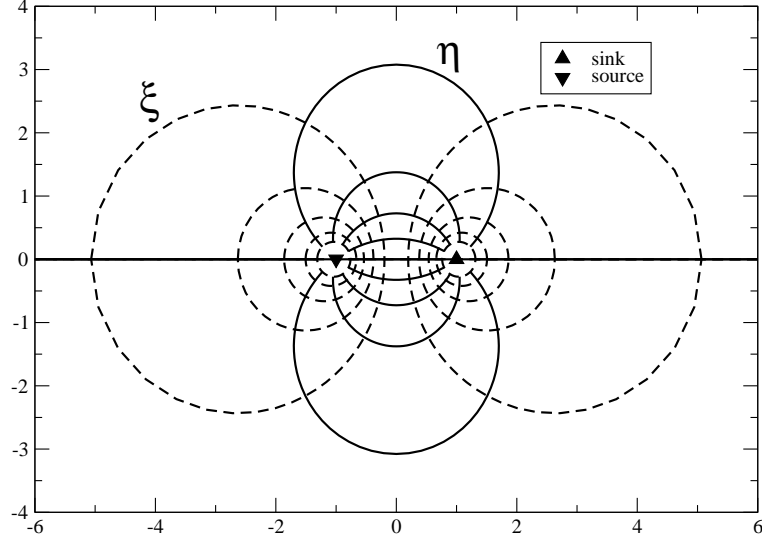
FIGURE 12.4. Bipolar coordinates for $\alpha = 1$. The solid lines $\eta =$ constant are the force lines, integrated from the vector field $\nabla \mathbf{f}$. The dashed lines $\xi =$ constant are the lines with constant potential $\mathbf{f} =$ constant. Both coordinates are circles in the plane.

The input is located at $\mathbf{x}_-$, the output at $\mathbf{x}_+$. The harmonic solution $\mathbf{f}$ yield the bipolar coordinate system. See Section 2.10. The harmonic function is the sum of the two potentials, or Green functions, Eq. 4.45, in two dimensions: [Spiegel, 1964]

$$(12.19) \qquad \mathbf{f} = \text{Re} \frac{\mathbf{e}_i}{2\pi} \log \left(\frac{\mathbf{x} - \mathbf{x}_+}{\mathbf{x} - \mathbf{x}_-} \right) = \frac{\mathbf{e}_i}{2\pi} \log \left(\frac{|\mathbf{x} - \mathbf{x}_+|}{|\mathbf{x} - \mathbf{x}_-|} \right) \quad ,$$

where the unit force is standing in an arbitrary direction $\mathbf{e}_i$. The imaginary part of the complex function can be used as angular coordinates on the lines $\mathbf{f} =$ constant. See Figure 12.4. The density of force lines on every surface is therefore unity. The bipolar coordinates (ξ, η) can be seen as the parallel $x_\parallel = \xi$ and perpendicular coordinates $x_\perp = \eta$ of the system. The parallel coordinate connects input with output, while the perpendicular coordinates, depending on the dimension of the space, is the span of all connections.

The simplest variation $\mathbf{G}$ of the force field is a mapping of the angular coordinate η onto itself. The closed contours of $\mathbf{G}$ are therefore the same as the lines $\eta =$ constant running forward and backward the same amount. The most extreme mapping $-\pi \geq \eta \geq \pi$ to $-\pi \geq \eta_{\text{new}} \geq \pi$ is the step function:

$$(12.20) \qquad \eta_{\text{new}} = 2\pi \left(\theta(\eta) - \frac{1}{2} \right) \quad ,$$

where $\theta(x) = 1$ for $x \geq 0$ and $\theta(x) = 0$ for $x < 0$. This will focus all the force along the straight line connecting the source at $\mathbf{x}_-$ and the sink at $\mathbf{x}_+$. The symmetric mappings $\eta_{\text{new}}(\eta)$ with $\eta_{\text{new}}(0) = 0$ and $\eta_{\text{new}}(-\eta) = -\eta_{\text{new}}(\eta)$ generate a whole
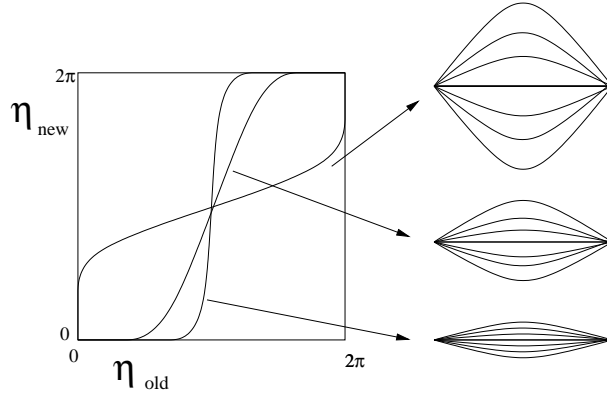
FIGURE 12.5. A sketch of a simple one parameter variation of force lines in a material. The underlying harmonic field ensures force is transmitted from source to sink. The variations are perpendicular to the force lines.

class of force configurations, which focus more or less along the straight line. See Figure 12.5. The interesting part is to derive efficient methods to find $\eta_{\mathrm{new}}(\eta)$ for different materials and input.

The bipolar coordinates may also serve for a near parallel FEM network representation. Take for example the heat flow from $\mathbf{x}_-$ to $\mathbf{x}_+$, or in practice, a small radius around the the source and sink. Since the harmonic field $\mathbf{f}$ suggests the main flow in the direction of the gradient $\nabla \mathbf{f}$, the network may, in first instance, be approximated by the parallel lines $\eta = $ constant. In the case of isotropic material, where the cells can be approximated by a uniform grid in (ξ, η)-space, the capacitance of each cell $\Delta\eta \times \Delta\xi$ is given by:

$$(12.21) \qquad C_{\mathrm{cell}} = \frac{\rho \Delta\eta \Delta\xi}{(\cosh \xi_{\mathrm{cell\ average}} - \cos \eta_{\mathrm{cell\ average}})^2} \quad ,$$

where ρ is the capacity density, $\Delta\eta \Delta\xi$ the surface element, and the denominator follows from the Laplacian, Eq. 2.86. The resistance is in the same manner related to the Laplace operator in the curved, bipolar coordinates. [Magnus et al., 1966] A small cell is a wide as it is long. For a square cell, the resistance is independent of the size. Hence all the resistances between each pair of neighboring cells i and j are the same:

$$(12.22) \qquad \mathbf{J} = -\lambda \nabla T \quad \Rightarrow \quad I_{ij} = y_{ji}\lambda \frac{T_j - T_i}{x_{ji}} = \lambda(T_j - T_i) \quad ,$$

where $\mathbf{J}$ is the heat flux, λ the plane conductivity, x_{ji} the distance between cell centers, and $y_{ji} \approx x_{ji}$ the length of the common edge. As an interesting consequence, the stationary harmonic flow depends only at the temperatures of at the source and sink, analogous to the energy of an electrostatic field in Eq. 12.1. The time-dependent flow can be analyzed on a square grid (ξ, η), with a homogeneous network of resistors λ^{-1}, and a variation in capacitances C_i, given by Eq. 12.21. See Figure 12.6.
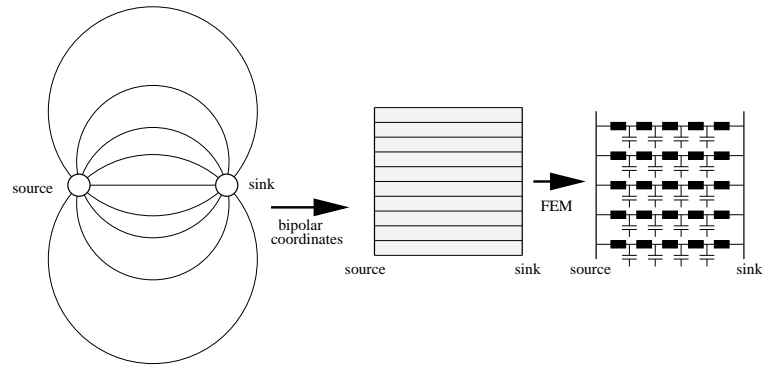
FIGURE 12.6. The near-parallel structure arises through a coordinate transformation to bipolar coordinates. The network representation of the FEM approximation has equal resistors, λ^{-1}, for the square 5×5 geometry. The capacitors depend on the location (ξ, η). The cross-connecting resistors can be ignored for small variations around the static, harmonic solution for heat diffusion or electrostatics.

12.4. Shape optimization

The design of the shape of a domain Ω optimal for the transfer of force is a well-known problem in engineering mathematics, which dates back to the days of Hadamard. [Hadamard, 1908, Courant and Hilbert, 1961, Pironneau, 1984, Cherkaev, 2000, Allaire et al., 2004] The optimization functional to be minimized can be the maximal rigidity:

$$(12.23) \qquad J[\Omega] = \int_\Omega dz \mathbf{u}^T \mathbf{f} + \int_{\partial\Omega} dz \mathbf{u}^T \mathbf{F} \quad ,$$

where $\mathbf{u}$ is the displacement and $\mathbf{f}$ and $\mathbf{F}$ are the applied body force and surface force.

The functional is the sum of the work done by the forces $\mathbf{F}$ applied to the boundary, and the body forces $\mathbf{f}$. Many other functionals exist for a variety of criteria. In the case of vibrations, the structure defined by its outline Ω must also be light to reduce the vibrational modes. This can, for example, be achieved by adding a volume term to the functional

$$(12.24) \qquad J_{\mathrm{volume}}[\Omega] = \alpha \int_\Omega dz \quad ,$$

where α is the relative contribution of the volume term. This is a common procedure, [Allaire et al., 2004] however, it will not take into account critical structure, or an eventual mode shape of the vibration.

The functional $J[\Omega]$ is minimized for the shape Ω of the object:

$$(12.25) \qquad \min_\Omega J[\Omega] \quad ,$$

which will not always lead to a well-defined problem. The functional derivative, necessary for the optimization, with respect to the shape with the use of the surface normal $\mathbf{n}$. A level-set method may be the practical implementation of such derivative.

The boundary $\partial\Omega$ may be defined through a level-set function $\psi(z)$:

$$(12.26) \qquad \psi(z) = \begin{cases} \psi(z) = 0 & z \in \partial\Omega \\ \psi(z) < 0 & z \in \Omega \\ \psi(z) > 0 & z \notin \Omega \end{cases} .$$

The outward surface normal, necessary for variations with changes in shapes, is given by the gradient of the level set ψ at $\psi = 0$:

$$(12.27) \qquad \mathbf{n} = \frac{\nabla\psi}{|\nabla\psi|} .$$

The optimization procedure using the level-set function is a rather abstract numerical method, which uses brute force computation of the functional and its variational, or functional, derivative.

Using force lines in optimization could lead to a more heuristic approach, and, since the force lines are conserved, it is intrinsically stable. The initial force lines in the complete domain may be bundled to smaller areas, decreasing such the volume term, Eq. 12.24, in the minimization functional, Eq. 12.23.

12.5. Following the force lines

Studying microscopically forces and the corresponding fields, which, together, yield the energy density and power continuity, is not a common approach to the study of continuous systems. Forces are not invariant concept in continuous systems. One person may talk about pressure, the other about force, discussing the same phenomena. The energy is, on the contrary, an invariant concept. This is only one of the reasons one has put aside force as central concept in the theory of dynamics.

However, when dealing with actuated systems, force is an important concept. The balancing of force is less trivial than for isolated systems. Most components in a system act to transfer force, or effort, or potential, rather than anything else. This transfer is described by the force lines. Components of systems are elements in chains, and the geometrical concepts like flux lines, such as force lines and stream lines bring this to light. The microscopical balance bring the material properties into play.

CHAPTER 13

Conclusions

> The depth of thought which goes into the formation of mathematical concepts is later justified by the skill with which these concepts are used.
> [E. P. Wigner, The unreasonable effectiveness of mathematics in natural sciences]

> Unfortunately, traditional modes of analysis have often placed constraints upon the forms in which such [system-environment] communication can occur.
> [Henry M. Paynter, Analysis and design of engineering systems]

This thesis addresses a single question:

> *What is an appropriate model of a continuous physical component for analysis and simulation in larger systems?*

This question was answered at many different levels, from conceptual, through mathematical and physical, to numerical. Models are constructed and reconstructed for the purposes at hand. For engineering purposes the function of a component determines its model. As a consequence, the nature of the interaction with the environment plays an important role, since a component's function is established by its interaction.

In this thesis we constructed finite-dimensional physical models of continuous, or infinite-dimensional, components, for the use in larger systems with inputs and outputs. An important part of such models is the interaction, through the common boundary. The relation between the bulk and the boundary has many different aspects which are discussed in this thesis.

First of all, the relation between bulk and boundary is also the relation between the intrinsic properties and the interfaces, which is a system, or operational, relation. If the interface does not transmit a certain quantity or quality, this quality or quantity does not have to appear in the component model, at least not explicitly.

Second, the relation between bulk and boundary is based on the geometric relation between domain and boundary, which, with the help of differential geometry, defines the relation between energy and energy flux, and, more generally, between any conserved quantity and the corresponding conserved flow.

Third, the relation between bulk and boundary, is, in terms of the underlying partial differential equations, the relation, or rather separation, in terms of initial conditions determining the states of the internal dynamics, and the boundary conditions essential for connecting components together. The physical boundary conditions follow from the energy flux through the boundary. The nature of the

boundary conditions in terms of the system description may vary with the circum-
stances. In some cases they will give rise to state variables associated with boundary
values, in other cases the boundary variables are constraints, or even parameters of
the model.

Finally, the relation between bulk and boundary is the relation between the
microscopic and the macroscopic properties, for example, between the charge den-
sity and the potential, and between the stress tensor and the force exerted on the
boundary. The internal and the external variables, used, respectively, in the partial
differential equations and the lumped models connected to the components, have
an uneasy relation to another. The consistency of the microscopic model in the
larger, system setting, follows from isolating the equivalents of the lumped model
parameter, in these models.

Modelling continuous systems remains difficult, mainly because the theories of
continuous systems, such as electrodynamics and elasticity, are difficult due to the
variety of phenomena they bring together. The internal dynamics draws all the
attention. Setting appropriate boundary conditions, for a start, and allowing a
component to interact through the boundary, in a more general system setting, are
complicated aspects, which receive little attention in the standard theory of con-
tinuous systems. For engineering purposes, only limited operations and variations
within these vast fields are relevant. Only the boundary conditions and input can
select or limit the dynamics to the intended function. The initial conditions play a
minor role.

The physical principles such as energy conservation and other conservation laws,
help to guide our attention toward the interface problem. The port-based modelling
is thereby be extended to such components. One nontrivial aspect of this extension
is the typical nonideal nature of continuous systems. Whether the appropriate
description of a component is more like a mass, a damper, or a capacitance, depends
more on the operational setting, i.e., the type and time-dependence of the boundary
conditions, than on the component itself.

The tendency exists to think in the terms of the methods and tools one uses.
Components described by partial differential equations are very much analyzed in
the language and with the methods developed for the intrinsic understanding of
these isolated components themselves. The relevant, modular features of a com-
ponent within greater schemes and systems are not the same as the features of a
component in an isolated setting.

The modular approach has brought electronics great advantages. Building a
large design in terms of functional blocks with well-defined properties seems to be
a successful route for rapid design, analysis, and prototyping. The improvement
and replacement of a part can be seen in the light of the whole, often with little
unknown interference between the components.

The intrinsic properties of a component in terms of its interactions does typi-
cally not exist. It would correspond to combining all the possible generic situations
and boundary conditions which may occur and select from them the largest set of
essential variables for which a core model is be constructed. In this thesis these
intrinsic properties and variables are inferred from the operators which generate

the dynamics. In some practical situations some the degrees of freedom may not be set as initial conditions but arise as boundary conditions, or input.

In setting up a language for components in terms of their interactions we have looked for inspiration towards the lumped models. Simple inputs will generate responses that can be interpreted as the analogues of a lumped characteristic. The stationary states, and some generalizations called quasi-stationary states, are the result. They are non-trivial internal states, which are the direct consequence of how a component interacts with its surroundings. They are not states in the sense that they require some initial condition, to uniquely specify the evolution of the system. Chapter 7 is central to this investigation.

Furthermore, we have tried to set up a network view based on the force balance throughout the system. The conservation, or balance, laws are inherent to a global physical view. Moreover, they are usually the complex physical laws which are traded in for simple microscopical laws, which can be implemented easily in some local finite-element scheme. For example, the whirls and tendrils of fluid motion will draw away attention from the fact that the same amount of incompressible fluid that goes into a container, or piping system, must come out as well. Setting up a quasi-stationary, lumped-equivalent flow satisfying the conservation law as starting point for the microscopical analysis, this analysis can be local, since the global laws are given by the nontrivial starting state based on the interaction with the surroundings. Only a preliminary analysis of a geometric view appears in Chapter 12, while the aspects of these principles, generically referred to as transmission, are scattered throughout the thesis, however, in particular, in Chapters 8 and 10.

The interface between different components is not naturally lumped, unlike the interface between a continuous component and a lumped component. The local product of effort and flow must be separated into a lumped, sum flow, satisfying the balance law, and the consistently lumped, mean effort. The product should still satisfy the power continuity. For FEM this restricts the numbers of internal and boundary variables. In a given finite-dimensional approximation based on states or configurations, the lumped power-continuity across the interface is only satisfied approximately. Some principles exist, such as Saint-Venant principle for elasticity, or the intermediate distance behavior of the Green function, or multi-pole expansion, which justify these approximations. The stationary state forms often the appropriate basis of the mean or lumped interface variable. Some partial answers are provided in this thesis, in particular in Chapters 6, 9, and 10.

The language of most continuous systems is not just based on linear systems, but on static linear systems, except for the vibrational analysis. For driven internal dynamics a modelling language seldom exists, although in operational settings it is the question of interest. Only a small number of well-known phenomena have there own model, such as wind-induced waves, flow-induced vibrations, and thermal convection. The practical use is in these cases obvious. However, also due to the inherent nonlinearities in induced dynamics, such problems are often avoided in modelling. From a functional design and optimization viewpoint, it is hard to understand how they can be avoided. The analysis of FEM models of continuous systems based on input characteristics is discussed in Chapter 9.

In this thesis we have shown that not the vibrational modes, but the core-model modes are the dominant features of continuous components with input and output.

These modes can be linked to the stationary and quasi-stationary states for different types of static input. The input as boundary conditions and boundary values are not given in terms of the states, but in terms of the efforts, or variational derivatives of the Hamiltonian. The consistency between initial conditions, in terms of states, and physical boundary conditions, in terms of the Hamiltonian, is therefore strained. In order to guarantee consistency the model is separated into two parts: the core model for the boundary conditions, and the model for the internal dynamics.

The port-Hamiltonian follows from the ordinary Hamiltonian through a change of variables, from configuration-like variables to energy-like variables, which opens the, initially closed, system for interaction through the boundary. In Chapter 5 this was done from a partial differential equations point of view, while the variational viewpoint was taken in Chapter 7. The differential operator used in the variable transformation determines the type of interface at the boundary. Three aspects of port-Hamiltonian systems should survive in the numerical setting, for power continuity. The exact action of the differential operator on the finite set of modes. The exact integration of the model port-Hamiltonian density, associated with these modes, over the domain. Finally, also the null-space of the differential operator, yielding the degrees of freedom which are necessarily distinct interaction variables, must be retained. This has been discussed in a general setting in Chapter 7, while a number of practical cases were constructed in Chapter 11.

So far, the different integrability conditions, given in Chapter 4 and used in Chapter 11, stood in the way of a uniform description of different port-Hamiltonian models for open systems. The recipe to construct a port-Hamiltonian finite element of arbitrary order and on an arbitrary domain, with a variety of boundary conditions, has been implemented in a number of cases, in Chapter 11. The aspects mentioned in the previous paragraph play an important role. The central place the differential operator in the port-Hamiltonian has taken, in Chapter 7, is one of most important contributions of this thesis to the study of the port-Hamiltonian approach. This central role has been important for the construction, rather than the reconstruction, of power-continuous continuous system components.

> All is grace.
> [GEORGES BERNANOS, DIARY OF A COUNTRY PRIEST]

APPENDIX A

Literature

A.1. Functional analysis

Variational principles, partial differential equations, and numerical methods all have been placed on a solid foundation by the use functional analysis. The shining example of its power is still the classic work by Hilbert and Courant [Courant and Hilbert, 1961]. More mathematical textbooks are by [Riesz and Nagy, 1955], [Yosida, 1980], and [Zeidler, 1995]. In the latter some ground is covered to treat nonlinear theory, and Banach spaces [Berger, 1977]. Applications can be found in [Lebedev and Vorovich, 2000].

More specialized literature can be found on variational methods [Gel'fand and Fomin, 1963, Mikhlin, 1964, Washizu, 1968, Carathéodory, 1935, Oden and Reddy, 1982], approximation theory [Kantorovich and Krylov, 1958, Funaro, 1992], and optimal control [Lasiecka and Triggiani, 2000, Pironneau, 1984, Lions, 1970].

A.2. Partial differential equations

In the study of hyperbolic partial differential equations [Sommerfeld, 1992, Hörmander, 1990, Gårding, 1998, Garabedian, 1964, John, 1982, Sakamoto, 1982] there are many different approaches. We can distinguish a number of main directions. First, the study of two-dimensional problems (one space dimension and one time dimension). This is a mainly mathematical field with many separate results, but not really a field of general methods. Second, the study of linear equation with constant coefficients. Here the theory of Fourier-Laplace transform proved very useful, and many general methods exist. Third, the linear equations with coefficients depending on space and on time. In this case the analysis falls apart in two classes: local analysis and global analysis. The local analysis can prove the existence and uniqueness of a solution in the neighborhood of the initial surface, while the global analysis is restricted to problems with some symmetry. For neither case effective numerical methods for the recovery of a solution exist.

In the numerical field one can separate approaches from strictly numerical which require only little fore work, such as FEM, to almost completely analytical, where only for the representation of the result functions need to be evaluated numerically. Another separation in numerical approaches to PDE's would be to distinguish variational approaches, such as Galerkin, which are inherently stable against direct methods, with limited evaluations which aim at a direct simulation of the evolution of the system.

The study of boundary condition,[Sakamoto, 1982] or the mixed problem, [Hadamard, 1923] or the initial-boundary value problem, of hyperbolic equation is beset with difficulties. Several numerical techniques are applied to such problems, [Lax and Wendroff, 1960, Morton and Mayers, 1994, Sod, 1985] and few results are

251

scattered in the literature. The Lopatinski condition [Krantz, 1992, Gårding, 1998, Hörmander, 1983] arises in many results.

A.3. Numerical methods

The most powerful and versatile of all numerical methods [Press et al., 1992, Hildebrand, 1987, Stoer and Bulirsch, 1980] are those of linear algebra. Solutions of sets of linear equations are easily found, and many problems, such as eigenvalue problems [Wilkinson, 1965, Golub and Van Loan, 1996] fall into clear classes, which specialized numerical methods. It is not surprising that most general problems are approximated in such a way that a numerical linear problem arises.

Closely related to linear problems are Hilbert spaces, of which the space $\mathcal{L}^2$, the Lebesgue space [Rudin, 1966] is the most common one. The norm is defined by the integral, which is a linear functional of the space of functions. Hence, most of the work before arriving at the eventual set of linear equations consists out of the determination of appropriate integrals. There a number of methods for integration, if the functions and the domain are sufficiently smooth and compact. If that is not the case, specialized methods of integration are required, and the discrete, numerical information of the integrand must be adapted to the problem at hand. The smoothing of data is often achieved by interpolation by means of polynomial fitting, such as spline fits. [Oden, 1972, Prenter, 1975, Funaro, 1992]

For nonlinear problems the situation is much less clear, different numerical approaches are set to solve different problems, and are more or less tuned to a particular class of problems. In many cases the problem, or solution, must be augmented with additional information, such as a range of values for the solution, of derivatives of functions, to be minimized or inverted.

For simulation of ODE's and PDE's a wide collection of methods exists. They aim at the maximal accuracy for a particular step size. For PDE's there are a number of bounds on the step size given the quality of initial conditions. A specialized literature on conserved quantities, such as energy, retained during the simulation exists. [Hairer et al., 2002, Serre, 2001] In the physics community the problem is often reformulated to make such conservation laws explicit.

A.4. Geometry

A continuous problem is defined on a space with a notion of distance, given by the metric. Euclid dominated geometry till the middle of the nineteenth century. After the introduction of differentials through variational approaches to optimization problems by Johann Bernoulli, Euler, Gauss, and others in the eighteenth century the road lay open to reformulate geometry in the new language. [Mc-Cleary, 1994, Spivak, 1970] The work of Riemann and Gauss are some highlights of this programme. [Frankel, 2004] This was the birth of differential geometry, or geometry in the small. The infinitesimal distance in differential geometry was and is a concept beset with difficulties. Topology gave a more generic answer to the notion of distance than geometry could. [Lefschetz, 1949] The two fields have up to now still a uneasy relation to another. At the other end of differential geometry the existence of shapes and invariants were questioned. Global solutions [Chern, 1989] of objects with negative curvature were sought, as were embeddings of manifolds, or hypersurfaces, with metric in Euclidean space of higher dimension, such as the surface of the sphere with constant positive curvature is embedded in three

dimensions. Manifolds were characterized by integrals of functions or operators defined on the manifold. Invariant manifolds, group theory, varieties, and index and characteristic numbers play an important role in algebraic geometry. [Hodge, 1941, Whitney, 1957] Algebraic geometry can be seen as the generic study of global characteristic using tools as triangulations of surfaces and polynomial mappings.

A.5. Modelling

To arrive at a PDE describing a system, one requires understanding of the underlying mechanical, physical, or chemical processes. Such information can be found in specialized books and papers in the respective research field. There are, however, attempts to collect and unify this type of modelling. Literature on mathematical methods for physics stems from all ages, and range from very formal to very practical. [Tayler, 2001, Temam and Miranville, 2000, Bossavit, 1998] Attempts to detach the methods, such as operational calculus, [Churchill, 1944] from the actual physical problems have led to abstract theories which survived mainly in extensions to the standard approaches for linear systems. [Åström and Wittenmark, 1989, Ljung, 1999] In practice, such methods are mainly applied in cases where the system is not known, which might manifest itself by a large noise component. For hyperbolic PDE's such black box approach is infeasible, as in many cases, the problem itself will test the limits of computing power, and any overhead should be reduced by the use all the information available on the system at hand.

The general modelling of complex, lumped systems in the field of system engineering, which tries to reach a level of abstraction and generalization in order to model and analyze complex interacting systems in with sufficient knowledge of the parts of the components. [Paynter, 1961, Seireg, 1969, Karnopp et al., 2000, Wellstead, 1979]

The extension to distributed systems and general invariances which associate continuous variables to discrete systems, such as rotation angles, have been a troublesome aspect of modelling. The unclear status of the work of Gabriel Kron is a definite sign of this troubled use of elements of differential geometry in engineering. [Grattan-Guinness, 1994, Paynter, 1961] From the Russian schools, Butkovskiy is a known example of similar status, his work was more along the lines of operational calculus. [Butkovskiy, 1983] However, with robotics, and its limited number of degrees of freedom, differential geometry is making a second entrance into the modelling for the purpose of control. [Pappas et al., 1998, Nijmeijer and van der Schaft, 1990, Slotine and Li, 1991]

A.6. Dynamics

In mechanics, [Arnold, 1989] or even more general dynamical descriptions, models can be formulated in equivalent forms, such as Hamiltonian of Lagrangian. [Goldstein, 1980] However, not in all cases, all formulations exist or are of a feasible form. In the mathematics there is a great interest in integrable systems, [Dubrovin et al., 1984] for which a complete set of trajectories can be determined. Such solutions usually have a underlying symmetry and are closely related to group theory, in particular the theory of Lie groups.

The classical three-body problem is at the edge of the analytically solvable problems. [Whittaker, 1917] More generic questions concerning dynamics involve the periodicity and stability of motion. Chaotic attractors and patterns are the,

mainly visual, manifestations of indeterministic stability, in which mankind likes to classify aspects of nature. [Arnold, 1973, Hirsch and Smale, 1974]

A.7. Classical field theories

The mathematical formulation of multi-physics continuous systems received an impetus with the work of Onsager in 1931, and Truesdell and Toupin in the Encyclopedia of Physics. [Truesdell and Toupin, 1960] Onsager derived relations between flux dependencies on general principles. [Callen, 1985] On separate domains a lot of literature is available, but interconnecting the different domains consistently, taking care of energy conservation and entropy, has been a large and controversial research field of the twentieth century. For example, from cosmology new questions arose concerning the nature of classical field theories. [Soper, 1976] Probably the final words are not yet said on the subject.

The thermodynamics of electromagnetism is discussed in, for example, [de Groot and Suttrop, 1972, Hutter and van der Ven, 1978, Eringen and Maugin, 1998]. Truesdell and Noll give a overview of aspects of elasticity in [Truesdell and Noll, 1965]. Modern mechanical materials with electrodynamic properties are discussed in [Maugin, 1988]. Heat transfer is another topic which interlinks many different physical domains. [McAdams, 1951, Carslaw and Jaeger, 1959, Holman, 1981, Vekstein, 1992]

However, the most diverse and active field of research in classical field theory is the study of fluid dynamics.[Mises and Friedrichs, 1971, Landau and Lifschitz, 1987b, Lighthill, 1978] Turbulence, boundary layers, and free surfaces, such as for droplets, are both from a fundamental and a practical perspective of great interest.[De Gennes et al., 2003, Eggers, 1997]

A.8. Networks of dynamical systems

The bond-graph modelling [Paynter, 1961, Karnopp et al., 2000, Breedveld, 1984] of physical systems is one of the most successful methods to combine graph theory [Mayeda, 1972] with physical modelling. A less successful attempt has been put forward by Gabriel Kron. [Kron, 1965] The purpose of a graph or network is to single out, or focus on, the global system, retaining the relevant parts of the components. The network is the global view. Some more obscure attempts to combine graphs with physical modelling can be found in, for example, [Mead, 1989], extending the integrated circuit design to the physical domain and [Zemanian, 1991], using the results of electrical circuits to study infinite-dimensional networks equivalent to continuous systems.

A.9. Control theory

Control theory is an integral part of the industrial revolution. [Bennett, 1979, Maxwell, 1868] The complexity of the machines required some automated adjustment of settings. The developments in control theory keep pace with the technological developments most of all. The classical system theory and control theory predates the use of electronics and computers in control, both in design of the controllers and in the controllers themselves. [Kalman et al., 1969, Brockett, 1970] The work of H. S. Black on amplifiers in the telephone transmission lines first used the electronic transistor in negative feedback. [Bode, 1960] Modern developments

include all kind of optimization strategies, multivariable systems, and state model identification. [Skogestad and Postlethwaite, 1996, Åström and Wittenmark, 1989, Ogata, 1990, Polderman and Willems, 1998, Antoulas, 2005] Modern equipment such as cd players and jet planes cannot function without a detailed plant model in the controller. In these cases some like to speak of learning control or adaptive control, if the plant model improves during operation. Well-known examples are learning control through neural networks and fuzzy logic. These methods had their predecessor in the theory of dynamic programming, where a solution was constructed recursively or algorithmically rather than explicitly. Dynamic programming is again closely related to the application of variational methods in control design, called optimal control. [Kirk, 2004]

A.10. History of science and engineering

The history of science is relevant for a better understanding of the scientific concepts. Some concepts may seem natural today, but have a troubled journey of existence, we have come to accept. [Dijksterhuis, 1950, Kronig, 1958, Taton, 1965, Bennett, 1979, Grattan-Guinness, 1994] On the theory of classical field theory the contributions of Clifford Truesdell and collaborators are well-known. Furthermore, there are specific contributions of fluid mechanics, [Tokaty, 1971] elasticity, [Timoshenko, 1953].

Bibliography

M. J. Ablowitz and P. A. Clarkson. *Solitons, nonlinear evolution equations and inverse scattering*. Cambridge University Press, Cambridge, 1991.

R. Abraham, J. E. Marsden, and T. Ratiu. *Manifolds, tensor analysis, and applications*, volume 75 of *Applied Mathematical Sciences*. Springer, New York, second edition, 1991.

M. Abramowitz and I. A. Stegun. *Handbook of mathematical functions*. Dover, New York, 1965.

J. H. Ahlberg, E. N. Nilson, and J. L. Walsh. *The theory of splines and their applications*. Academic Press, New York, 1967.

N. I. Akhiezer. *The classical moment problem*. Oliver and Boyd, Edinburgh, 1965.

G. Allaire, F. Jouve, and A.-M. Toader. Structural optimization using sensitivity analysis and a level-set method. *Journal of Computational Physics*, 194:363–393, 2004.

E. Anderson, Z. Bai, C. H. Bishop, S. Blackford, J. W. Demmel, J. J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. Mckenney, and D. C. Sorensen. *Lapack users' guide*. SIAM, Philadelphia, 1999.

H. M. Antia. *Numerical methods for scientists and engineers*. Birkhäuser, Basel, second edition, 2002.

S. S. Antman and G. Rosenfeld. Global behavior of buckled states of nonlinearly elastic rods. *SIAM Review*, 20:513–566, 1978.

A. C. Antoulas. *Approximation of large-scale dynamical systems*. SIAM, Philadephia, 2005.

V. I. Arnold. *Ordinary differential equations*. MIT Press, Cambridge (Mass), 1973.

V. I. Arnold. *Mathematical methods of classical mechanics*. Springer, New York, second edition, 1989.

V. I. Arnold. *Lectures on partial differential equations*. Springer, Berlin, 2004.

K. J. Åström and B. Wittenmark. *Adaptive control*. Addison-Wesley, Reading (MA), 1989.

A. O. Barut. *Electrodynamics and Classical Theory of Fields and Particles*. Dover, New York, 1981.

W. Beitz, K.-H. Küttner, and B.J. Davies. *Dubbel Handbook of mechanical engineering*. Springer, London, 1994.

S. Bennett. *A history of control engineering 1800-1930*. The Institution of Electrical Engineers (Peter Peregrinus Ltd), London, 1979.

M. Benzi, G. H. Golub, and J. Liesen. Numerical solution of saddle point problems. *Acta Numerica*, 14:1–137, 2005.

M. S. Berger. *Nonlinearity and functional analysis*. Academic Press, New York, 1977.

D. L. Bernstein. *Existence theorems in partial differential equations*. Princeton University Press, Princeton (NJ), 1950.

L. Bers, F. John, and M. Schechter. *Partial differential equations*, volume III of *Lectures in applied mathematics*. Interscience, New York, 1964.

C. B. Biezeno and R. Grammel. *Technische dynamik*. Springer, Berlin, 1953.

R. B. Bird, W. E. Stewart, and E. N. Lightfoot. *Transport phenomena*. Wiley, New York, 1960.

A. V. Bitsadze. *Boundary value problems for second order elliptic equations.* North-Holland, Amsterdam, 1968.

H. Bode. Feedback: the history of an idea. In *Proceedings of the symposium on active networks and feedback systems*, Brooklyn, 1960.

A. D. Booth. *Numerical methods.* Butterworths, London, third edition, 1966.

M. Born, E. Wolf, and A. B. Bhatia. *Principles of optics.* Pergamon, Oxford, sixth edition, 1980.

A. Bossavit. *Computational electromagnetism.* Academic Press, San Diego, 1998.

P. C. Breedveld. *Physical system theory in terms of bond graphs.* PhD thesis, University Twente, Enschede, 1984.

P. C. Breedveld. A generic dynamic model of multiphase electromechanical transduction in rotating machinery. In *WESIC 2001*, pages 381–394, Enschede, 2001. University Twente.

J. P. Breslin and P. Andersen. *Hydrodynamics of ship propellers*, volume 3 of *Cambridge ocean technology.* Cambridge, Cambridge, 1996.

R. W. Brockett. *Finite dimensional linear systems.* Wiley, New York, 1970.

A. G. Butkovskiy. *Structural theory of distributed systems.* Ellis-Horwood, Chichester (UK), 1983.

H. B. Callen. *Thermodynamics and an introduction to thermostatistics.* Wiley, New York, 1985.

C. Carathéodory. *Variationsrechnung und partielle differentialgleichungen erster ordnung.* Teubner, Leipzig, 1935.

H. S. Carslaw and J. C. Jaeger. *Conduction of heat in solids.* Clarendon Press, Oxford, second edition, 1959.

A. Cherkaev. *Variational methods for structural optimization.* Springer, New York, 2000.

S. S. Chern. *Global differential geometry*, volume 27 of *Studies in Mathematics.* The Mathematical Association of America, 1989.

B. H. Chirgwin and C. Plumpton. *A course of mathematics for engineers and scientists*, volume 4. Pergamon Press, Oxford, 1964.

R. M. Christensen. *Theory of Viscoelasticity.* Academic Press, New York, 1971.

R. V. Churchill. *Modern operational mathematics in engineering.* McGraw-Hill, New York, 1944.

H. Cohen and R. G. Muncaster. *The theory of pseudo-rigid bodies*, volume 33 of *Springer tracts in natural philosophy.* Springer, New York, 1988.

J. B. Conway. *Functions of one complex variable*, volume 11 of *Graduate Texts in Mathematics.* Springer, New York, second edition, 1978.

R. Courant and D. Hilbert. *Methods of mathematical physics: Partial differential equations*, volume II. Interscience, New York, 1961.

R. R. Craig. *Structural dynamics, an introduction to computer methods.* Wiley, New York, 1981.

R. R. Jr. Craig and C.-J. Chang. On the use of attachment modes in substructure coupling for dynamic analysis. In *Proc. AIAA/ASME 18th Structures, Structural Dynamics and Materials Conference*, volume B, 1977.

M. Creutz. *Quarks, gluons, and lattices.* Cambridge University Press, Cambridge, 1983.

R. F. Curtain and H. Zwart. *An introduction to infinite-dimensional linear systems theory*, volume 21 of *Text in Applied Mathematics.* Springer, New York, 1995.

C. M. Dafermos. *Hyperbolic conservation laws in continuum physics.* Springer, Berlin, 1999.

P. R. Dahl. Solid friction damping of mechanical vibrations. *AIAA Journal*, 14:1675–1683, 1976.

R. Dautray and J.-L. Lions. *Physical origins and classical methods*, volume 1 of *Mathematical analysis and numerical methods for science and technology.* Springer, New York, 1990.

B. Davies. *Integral transforms and their applications*, volume 25 of *Applied Mathematical Sciences*. Springer, New York, 1978.

P.-G. de Gennes, F. Brochard-Wyard, D. Quéré, and A. Reisinger. *Capillarity and wetting phenomena : drops, bubbles, pearls, waves*. Springer, New York, 2003.

S. R. de Groot and P. Mazur. *Non-equilibrium thermodynamics*. North-Holland, Amsterdam, 1962.

S. R. de Groot and L. G. Suttrop. *Foundations of electrodynamics*. North-Holland, Amsterdam, 1972.

E. J. Dijksterhuis. *De mechanisering van het wereldbeeld*. Meulenhoff, Amsterdam, 1950.

P. A. M. Dirac. *Lectures on quantum mechanics*. Belfer Graduate School, Yeshiva University, New York, 1964.

B. A. Dubrovin, A. T. Fomenko, and S. P. Novikov. *Modern geometry -methods and applications, Part I. The geometry of surfaces, transformation groups, and fields*. Springer, New York, 1984.

J. Eggers. Nonlinear dynamics and breakup of free-surface flows. *Reviews of Modern Physics*, 69:865–929, 1997.

A. C. Eringen and G. A. Maugin. *Electrodynamics of continua*, volume I and II. Springer, Berlin, 1998.

L. Euler. *Methodus inveniendi lineas curvas maximi minimive proprietate gaudentes, sive solutio problematis isoperimetrici lattissimo sensu accepti (A method for finding curved lines enjoying properties of maximum or minimum, or solution of isoperimetric problems in the broadest accepted sense)*. Bousquet, Lausanne, 1744.

D. Farin. Shape. In B. Engquist and W. Schmid, editors, *Mathematics unlimited –2001 and beyond*, pages 463–477, New York, 2001. Springer.

B. Felsager. *Geometry, particles and fields*. Odense University Press, Odense, 1981.

A. L. Fetter and J. D. Walecka. *Theoretical mechanics of particles and continua*. McGraw-Hill, New York, 1980.

R. P. Feynman, R. B. Leighton, and M. Sands. *The Feynman lectures of physics*, volume II. Addison-Wesley, Palo Alto, 1964.

H. Flanders. *Differential forms: with applications to the physical sciences*. Academic Press, New York, 1963.

J. D. Foley and A. van Dam. *Fundamentals of interactive computer graphics*. Addison-Wesley, Reading (Mass), 1982.

O. Föllinger. *Lineare abtastsysteme*. Oldenbourg, München, 1974.

T. Frankel. *The geometry of physics*. Cambridge University Press, Cambridge, second edition, 2004.

R. W. Freund. Model reduction methods based on krylov subspaces. *Acta Numerica*, 12: 267–319, 2003.

K. O. Friedrichs. *Energy principle and energy method*. lecture notes, New York University, Institute of Mathematical Sciences, New York, 1954.

K. O. Friedrichs. *Spectral theory of operators in Hilbert space*. Springer, New York, 1973.

D. Funaro. *Polynomial approximation of differential equations*. Springer, Berlin, 1992.

F. Gantmacher. *Lectures in analytical mechanics*. MIR, Moscow, 1970.

P. R. Garabedian. *Partial differential equations*. Wiley, New York, 1964.

L. Gårding. Hyperbolic equations in the twentieth century. *Seminaires et Congres*, 3: 37–68, 1998.

B. R. Gelbaum and J. M. H. Olmsted. *Counterexamples in analysis*. Holden-Day, San Fransisco, 1964.

I. M. Gel'fand and S. V. Fomin. *Calculus of variations*. Prentice-Hall, Englewood Cliffs (NJ), 1963.

I. M. Gel'fand and G. E. Shilov. *Generalized Functions*, volume 1 Properties and operations. Academic Press, Boston, 1964.

E.-D. Gilles. *Systeme met verteilten parametern*. R. Oldenbourg, München, 1973.

I. Gohberg and S. Goldberg. *Basic operator theory*. Birkhäuser, Boston, 1981.

H. Goldstein. *Classical mechanics*. Addison-Wesley, Reading (Mass), second edition, 1980.

G. H. Golub and C. F Van Loan. *Matrix Computations*. John Hopkins, Baltimore, third edition, 1996.

B. R. Gossick. *Hamilton's principle and physical systems*. Academic Press, New York, 1967.

I. Grattan-Guinness. *History and philosophy of the mathematical sciences*, volume 2. John Hopkins University Press, Baltimore, 1994.

G. Grosche, V. Ziegler, D. Ziegler, and E. Zeidler. *Tuebner -Taschenbuch der Mathematik*, volume II. Tuebner, Stuttgart, 1995.

H. W. Guggenheimer. *Differential Geometry*. McGraw-Hill, New York, 1963.

J. Hadamard. Mémoire sur le problème d'analyse relatif à l'équilibre de plaques élastique encastrés. *Mémoires présentés par divers savants á l'Académie des Sciences*, 33:1–128, 1908.

J. S. Hadamard. *Lectures on Cauchy's problem in linear partial differential equations*. Dover, New York, 1923.

E. Hairer, C. Lubich, and G. Wanner. *Geometric numerical integration*, volume 31 of *Springer series in computational mathematics*. Springer, Berlin, 2002.

C. H. Hansen and S. D. Snyder. *Active control of noise and vibration*. Chapman and Hall, London, 1997.

O. Heaviside. On the forces, stresses, and fluxes of energy in the electromagnetic field. *Proceedings of the Royal Society of London*, 50:126–129, 1891.

O. Heaviside. *Heaviside centenary volume*. Institute of Electrical Engineers, London, 1950.

P. Henrici. *Applied and computational complex analysis: Power series, integration, conformal mapping, location of zeros*, volume 1. Wiley, New York, 1974.

M. R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *Journal of Research of the National Bureau of Standards*, 49:409–436, 1952.

R. L. Highdon. Initial-boundary value problems for linear hyperbolic problems. *SIAM Review*, 28:177–217, 1986.

B. F. Hildebrand. *Introduction to numerical analysis*. Dover, New York, second edition, 1987.

M. W. Hirsch and S. Smale. *Differential equations, dynamical systems and linear algebra*. Acedemic Press, San Diego, 1974.

W. V. D. Hodge. *The theory and applications of harmonic integrals*. Cambridge University Press, Cambridge, 1941.

N. Hogan. Impedance control: an approach to manipulation i-iii. *ASME Journal od Dynamic Systems Measurement and Control*, 107:1–24, 1985.

J. P. Holman. *Heat transfer*. McGraw-Hill, New York, fifth edition, 1981.

P. Holmes, J. L. Lumley, and G. Berkooz. *Turbulence, coherent structures, dynamical systems and symmetry*. Cambridge, Cambridge, 1996.

L. Hörmander. *The analysis of linear partial differential operators*, volume 2 Differential operators with constant coefficients. Springer, Berlin, 1983.

L. Hörmander. *The analysis of linear partial differential operators*, volume 1 Distribution theory and Fourier analysis. Springer, Berlin, 1990.

T. J. Hughes, J. A. Cottrell, and Y. Bazilevs. Isogeometric analysis: Cad, finite element, nurbs, exact geometry, and mesh refinement. *Computer Methods in Applied Mechanics and Engineering*, 194:435–458, 2005.

T. J. R. Hughes. *The finite element method*. Prentice-Hall, Englewood Cliff (NJ), 1987.

K. Hutter and A. A.F. van der Ven. *Field matter interactions in thermoelastic solids*, volume 88 of *Lecture Notes in Physics*. Springer, Berlin, 1978.

C. Huygens. *Horologium oscillatorium*. Société Hollandaise des Sciences, Den Haag, 1673.

A. Isodori. *Nonlinear control systems*, volume 72 of *Lecture Notes in Control and Information Sciences*. Springer, Berlin, 1985.

T. A. Ivey and J. M. Landsberg. *Cartan for beginners: differential geometry via moving frames and exterior differential systems*, volume 61 of *Graduate studies in mathematics*. AMS, Providence, Rhode Island, 2000.

J. D. Jackson. *Classical electrodynamics*. Wiley, New York, second edition, 1975.

F. John. *Partial differential equations*, volume 1 of *Applied Mathematical Sciences*. Springer, New York, fourth edition, 1982.

D. Jürgens, M. Palm, S. Singer, and K. Urban. Numerical optimization of the voith-schneider propeller. *Ulm University preprint*, 2006.

R. E. Kalman, P. L. Falb, and M. A. Arbib. *Topics in mathematical system theory*. McGraw-Hill, New York, 1969.

E. Kamke. *Differentialgleichungen*, volume II. Geest und Portig, Leipzig, fourth edition, 1962.

L. V. Kantorovich and V. I. Krylov. *Approximate methods of higher analysis*. Noordhoff, Groningen, 1958.

D. C. Karnopp, D. L. Margolis, and R. C. Rosenberg. *System dynamics: modeling and simulation of mechatronic system*. Wiley, New York, 2000.

W. Karush. *Minima of functions of several variables*. PhD thesis, University of Chicago, Chicago (IL), 1939.

T. Kato. *Perturbation theory for linear operators*. Springer, Berlin, 1966.

A. C. Kermode. *Mechanics of flight*. Longman, London, 1987.

D. E. Kirk. *Optimal control theory*. Dover, New York, 2004.

G. A. Korn and T. M. Korn. *Mathematical handbook for scientists and engineers*. McGraw-Hill, New-York, second edition, 1968.

A. D. Kovalenko. *Thermoelasticity*. Wolters-Noordhoff, Groningen, 1969.

S. G. Krantz. *Partial differential equations and complex analysis*. CRC Press, Boca Raton (FL), 1992.

H. O. Kreiss. Initial boundary value problems for hyperbolic systems. *Communications in Pure and Applied Mathematics*, 23:277–298, 1970.

S. Krenk. State-space time integration with energy control and fourth-order accuracy for linear dynamic systems. *Int. J. Numer. Meth. Engng*, 65:595–619, 2006.

G. Kron. Non-riemannian dynamics of rotating electrical machinery. *Journal of Mathematics and Physics*, 13:103–194, 1934.

G. Kron. Invisible dual (n-1)-networks induced by electric 1-networks. *IEEE Transactions on Circuit Theory*, CT12:464–476, 1965.

R. Kronig. *Leerboek der natuurkunde*. Scheltema en Holkema, Amsterdam, 1958.

H. W. Kuhn and A. W. Tucker. Nonlinear programming. In *Proceedings of 2nd Berkeley Symposium*, pages 481–492, Berkeley, 1951. University of California Press.

O. A. Ladyzhenskaya. *The boundary value problems of mathematical physics*, volume 49 of *Applied Mathematical Sciences*. Springer, New York, 1985.

C. Lanczos. *Linear differential operators*. Van Nostrand, London, 1961.

C. Lanczos. *The variational principles of mechanics*. University of Toronto Press, Toronto, 1966.

L. D. Landau and E. M. Lifschitz. *Theory of elasticity*. Pergamon, Oxford, 1959.

L. D. Landau and E. M. Lifschitz. *Mechanics*, volume 1 of *Course in theoretical physics*. Pergamon, Oxford, third edition, 1976.

L. D. Landau and E. M. Lifschitz. *The classical theory of fields*. Pergamon, Oxford, fourth edition, 1987a.

L. D. Landau and E. M. Lifschitz. *Fluid Mechanics*. Pergamon Press, Oxford, second edition, 1987b.

I. Lasiecka and R. Triggiani. *Control theory for partial differential equations: continuous and approximation theories: II abstract hyperbolic-like systems over a finite time horizon*, volume 75/2 of *Encyclopedia of mathematics and its applications*. Cambridge, Cambridge, 2000.

P. D. Lax and B. Wendroff. System of conservation laws. *Communications on Pure and Applied Mathematics*, 13:217–237, 1960.

G. Lebeau. Control for hyperbolic equations. *Journées Équations aux Dérivées partielles*, 21:1–24, 1992.

L. P. Lebedev and I. I. Vorovich. *Functional analysis in mechanics*. Springer, New York, 2000.

A. Y. Lee, W. S. Tsuha, G. K. Man, and G. A. Macala. *Model reduction methodologies for articulated multi-flexible body systems*, volume 15 (Nonlinear Techniques) of *Structural Dynamic Systems Computational Techniques and Optimization*, pages 195–271. Gordon Breach, 1999.

S. Lefschetz. *Introduction to topology*. Princeton University Press, Princeton (NJ), 1949.

M. J. Lighthill. *Waves in fluids*. Cambridge University Press, Cambridge, 1978.

N. E. Ligterink, P. C. Breedveld, and A. J. Van der Schaft. Internal dynamics around operational fem states. *preprint, submitted for publication*, 2006.

J. L. Lions. *Optimal control of systems governed by partial differential equations*. Springer, Berlin, 1970.

J. Liouville. Démonstration d'un théoreme du a m. sturm et relatif a une classe de fonctions transcendantes. *Journal de mathématiques pures et appliquées*, 2:269–277, 1836.

M. M. Lipschutz. *Differential geometry*. Schaum's outline series. McGraw-Hill, New York, 1969.

X.-D. Liu, S. Osher, and T. Chan. Weighted essentially non-oscillatory schemes. *Journal of Computational Physics*, 115:200–212, 1994.

L. A. Liusternik and V. J. Sobolev. *Elements of functional analysis*. Ungar, New York, 1961.

L. Ljung. *System identification: theory for the user*. Prentice Hall, Upper Saddle River (NJ), 1999.

H. A. Lorentz. *The theory of electrons*. Dover, New York, second edition, 1915.

A. E. H. Love. *A treatise on the mathematical theory of elasticity*. Dover, New York, 1944.

A. M. Lyapunov. *The general problem of the stability of motion*. Taylor & Francis, London, annotated translation of the 1892 classic edition, 1992.

W. Magnus, F. Oberhettinger, and R. P. Soni. *Formulas and theorems for the special functions of mathematical physics*. Springer, New York, third edition, 1966.

A. Majda and S. Osher. Initial-boundary value problems for hyperbolic equations with uniformly characteristic boundary. *Communications on Pure and Applied Mathematics*, 28:607–675, 1975.

J. M. Maubach. Local bisection refinement of n-simplicial grids generated by reflection. *SIAM Journal for Scientific Computing*, 16:210–227, 1994.

G. A. Maugin. *Continuum mechanics of electromagnetic solids*, volume 33 of *Applied mathematics and mechanics*. North-Holland, Amsterdam, 1988.

J. C. Maxwell. On governors. *Proceedings of the Royal Society of London*, 16:270–283, 1868.

J. C. Maxwell. *A treatise on electricity and magnetism*. Clarendon Press, Oxford, third edition, 1891.

W. Mayeda. *Graph theory*. Wiley, New York, 1972.

W. A. McAdams. *Heat Transmission*. McGraw-Hill, New York, second edition, 1951.

J. McCleary. *Geometry from a differentiable viewpoint*. Cambridge University Press, Cambridge, 1994.

C. Mead. *Analog VLSI and neural systems*. Addison-Wesley, Reading (Mass), 1989.

S. G. Mikhlin. *Variational methods in mathematical physics*. Pergamon, Oxford, 1964.

W. A. Milne. *Numerical solution of differential equations*. Dover, New York, second edition, 1970.

N. Minorsky. Self-excited oscillations in dynamical systems possessing retarded action. *Journal of Applied Mechanics*, 9:65–71, 1942.

R. von Mises and K. O. Friedrichs. *Fluid dynamics.* Springer, New York, 1971.

S. Morita. *Geometry of differential forms*, volume 201 of *Translations of mathematical monographs.* American Mathematical Society, Providence, Rhode Island, 2001.

K. W. Morton and D. F. Mayers. *Numerical solution of partial differential equations.* Cambridge University Press, Cambridge, 1994.

H. Nijmeijer and A.J. van der Schaft. *Nonlinear dynamical control theory.* Springer, New York, 1990.

E. Noether. Invariante variationsprobleme. *Nachr. Ges. Wiss. Göttingen*, 171, 1918.

J. T. Oden. *Finite elements of nonlinear continua.* McGraw-Hill, New York, 1972.

J. T. Oden and J. N. Reddy. *An Introduction to the mathematical theory of finite elements.* Wileys, New York, 1982.

K. Ogata. *Modern control engineering.* Prentice-Hall, Englewood Cliffs (NJ), 1990.

C. C. Paige and M. A. Saunders. Solution of sparse indefinite systems of linear equations. *SIAM J. Numer. Anal.*, 12:617–629, 1975.

G. J. Pappas, J. Lygeros, D. Tilbury, and S. Sastry. Exterior differential systems in control and robotics. In J. Baillieul, Sastry S. S., and Sussman H. J., editors, *Essays on mathematical robotics*, volume 104 of *The IMA volumes in mathematics and its applications*, pages 271–372, New York, 1998. Springer.

H. M. Paynter. *Analysis and design of engineering systems.* MIT Press, Cambridge (MA), 1961.

O. Pironneau. *Optimal shape design for elliptic systems.* Springer, New York, 1984.

J. W. Polderman and J. C. Willems. *Introduction to mathematical systems theory*, volume 26 of *Texts in Applied Mathematics.* Springer, New York, 1998.

G. Pólya and G. Szegö. *Isoperimetric inequalities in mathematical physics*, volume 27 of *Annals of mathematical studies.* Princeton University Press, Princeton (NJ), 1951.

P. M. Prenter. *Splines and variational methods.* Wiley, New York, 1975.

W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C: The Art of Scientific Computing.* Cambridge, New York, 1992.

F. Riesz and B. S. Nagy. *Functional analysis.* Frederick Ungar, New York, 1955.

L. Rosenfeld. Questions of irreversibility and ergodicity. In *Teorie ergodiche*, Scuola Internazionale di Fisica –Enrico Fermi– (Societa Italiana di Fisica), pages 1–20, New York, 1961. Academic Press.

W. Rudin. *Real and complex analysis.* McGraw-Hill, New York, second edition, 1966.

Y. Saad. *Iterative methods for sparse linear systems.* PWS Publishing, Boston, 1995.

R. Sakamoto. *hyperbolic boundary value problems.* Cambridge, Cambridge, 1982.

S. I. Sandler. *Chemical and engineering thermodynamics.* Wiley, New York, 1977.

L. Schwartz. *Mathematics for the physical sciences.* Hermann, Paris, 1966.

H. A. Schwarz. Über einen Grenzübergang durch alternierendes Verfahren. *Vierteljahrsschrift der Naturforschenden Gesellschaft in Zürich*, 15:272–286, 1870.

A. Seireg. *Mechanical system analysis.* International Textbook Company, Scranton (PA), 1969.

D. Serre. Systems of conservation laws. In B. Engquist and W. Schmid, editors, *Mathematics unlimited –2001 and beyond*, pages 1061–1080, New York, 2001. Springer.

G. E. Shilov. *Generalized functions and partial differential equations.* Gordon Breach, New York, 1968.

N. C. Singer and W. P. Seering. Preshaping command inputs to reduce system vibration. *Journal of Dynamical Systems, Measurement and Control*, 112:76–82, 1990.

S. Skogestad and I. Postlethwaite. *Multivariable feedback control.* Wiley, Chichester, 1996.

J.-J. E. Slotine and W. Li. *Applied nonlinear control.* Prentice Hall, Englewood Cliffs (NJ), 1991.

G. A. Sod. *Numerical methods in fluid mechanics.* Cambridge, Cambridge, 1985.

A. Sommerfeld. *Partieller differentalgleichungen der physik.* Deutsch, Thun, sixth edition, 1992.

D. E. Soper. *Classical field theory*. Wiley, New York, 1976.

R. W. Soutas-Little. *Elasticity*. Prentice-Hall, Englewood Cliffs (NJ), 1973.

M. R. Spiegel. *Complex variables*. Schaum's Outlines. McGraw-Hill, New York, 1964.

M. Spivak. *A comprehensive introduction to differential geometry*. Publish or Perish, Houston, 1970.

T. Stieltjes. Recherches sur les fractions continues. *Anns. Fac. Sci. Univ. Toulouse*, 8,9: J1–11, A5–47, 1894-1895.

J. Stoer and R. Bulirsch. *Introduction to Numerical Analysis*. Springer, New York, 1980.

R. F. Streater and A. S. Wightman. *CPT, spin and statistics, and all that*. Benjamin, New York, 1964.

A. M. Stuart and A. R. Humphries. *Dynamical systems and numerical analysis*, volume 2 of *Cambridge monographs on applied and computational mathematics*. Cambridge University Press, Cambridge, 1998.

J. C. F. Sturm. Mémoire sur les équations diffrentielles linéaires du second ordre. *Journal de mathématiques pures et appliquées*, 1:106–186,278–289,290–308,, 1836.

J. L. Synge. *The hypercircle in mathematical physics: a method for the approximate solution of boundary value problems*. Cambridge Unversity Press, Cambridge, 1957.

R. Taton. *Science in the nineteenth century*, volume 3 of *General History of Sciences*. Thames and Hudson, London, 1965.

A. B. Tayler. *Mathematical models in applied mechanics*. Oxford University Press, Oxford, 2001.

S. W. Taylor and S. C. B. Yau. Boundary control of a rotating timoshenko beam. *Australian and New Zealand Industrial and Applied Mathematics Journal*, 44:E143–E184, 2003.

R. Temam and A. Miranville. *Mathematical modeling in continuum mechanics*. Cambridge Press, Cambridge, 2000.

D. W. Theobald. *The concept of energy*. Spon, London, 1966.

W. Thirring. *Classical mathematical physics*. Springer, New York, third edition, 1997.

S. P. Timoshenko. *History of strength of materials*. McGraw-Hill, New York, 1953.

G. A. Tokaty. *A history and philosophy of fluidmechanics*. G. T. Foulis and Co, Henley-on-Thames (Oxfordshire), 1971.

A. Toselli and O. Widlund. *Domain decomposition methods - algorithms and theory*, volume 34 of *Springer Series in Computational Mathematics*. Springer, Berlin, 2004.

F. Tröltzsch. *Optimale Steuerung partieller Differentialgleichungen*. Vieweg, Wiesbaden, 2005.

J. L. Troutman. *Variational calculus with elementary convexity*. Undergraduate Texts in Mathematics. Springer, New York, 1983.

C. Truesdell and W. Noll. *Non-linear field theories of mechanics*, volume III/3 of *Handbuch für Physik*. Springer, Berlin, 1965.

C. Truesdell and R. A. Toupin. *The classical field theories*, volume III/1 of *Encyclopedia of Physics*. Springer, Berlin, 1960.

B. van der Pol and H. Bremmer. *Operational calculus*. Cambridge Univesity Press, Combridge, 1959.

A. J. van der Schaft and B. M. Maschke. Hamiltonian formulation of distributed-parameter systems with boundary energy flow. *Journal of Geometry and Physics*, 42:166–194, 2002.

W. P. A. van Lammeren. *Weerstand en voortstuwing van schepen*. Stam, Amsterdam, 1944.

G. E. Vekstein. *Physics of continuous media: A collection of problems with solutions for physics students*. Adam-Hilger, Bristol, 1992.

R. R. Waiboer, R. G. K. M. Aarts, and J. B. Jonker. Application of a perturbation method for realistic dynamic simulation of industrial robots. *Multibody Systems Dynamics*, 13: 323–338, 2005.

K. Washizu. *Variational methods in elasticity and plasticity*. Pergamon Press, Oxford, 1968.

P. E. Wellstead. *Introduction to physical system modelling*. Academic Press, New York, 1979.

J. R. Welty, C. E. Wicks, and R. E. Wilson. *Fundamentals of momentum, heat, and mass transfer*. Wiley, New York, 1984.

W. L. Wendland. *Elliptic systems in the plane*. Pitman, London, 1979.

E. J. Weniger. Weakly convergent expansions of a plane wave and their use in fourier integrals. *Journal of Mathematical Physics*, 26:276–291, 1985.

H. Weyl. The method of orthogonal projection in potential theory. *Duke Mathematical Journal*, 7:411–444, 1940.

H. Weyl. *Space time matter*. Dover, New York, 1952.

H. Whitney. *Geometric integration theory*. Princeton Press, Princeton, 1957.

E. T. Whittaker. *A treatise on the analytical dynamics of particles and rigid bodies*. Cambridge Press, Cambridge, 1917.

D. V. Widder. *The laplace transform*. Princeton Unversity Press, Princeton, 1941.

N. Wiener. *Extrapolation, interpolation and smoothing of stationary time series*. Wiley, New York, 1949.

J. A. Wilkinson. *The algebraic eigenvalue problem*. Orford University Press, Oxford, 1965.

J. C. Willems. Dissipative dynamical systems, part i: general theory. *Archive for Rational Mechanics and Analysis*, 45:321–351, 1972a.

J. C. Willems. Dissipative dynamical systems, part ii: linear systems with quadratic supply rates. *Archive for Rational Mechanics and Analysis*, 45:352–393, 1972b.

K. Worden and G. R. Tomlinson. *Nonlinearity in structural dynamics*. IoP Publishing, Bristol, 2001.

K. Yosida. *Functional analysis*, volume 123 of *Grundlehren der mathematische wissenschaften*. Springer, Berlin, sixth edition, 1980.

E. Zeidler. *Applied functional analysis: Applications to mathematical physics*. Springer, New York, 1995.

A. H. Zemanian. *Infinite electrical networks*. Cambridge University Press, Cambridge, 1991.

Index