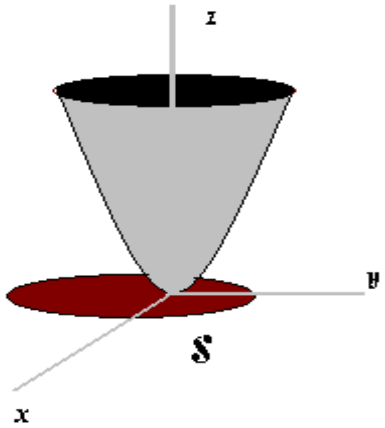


Further information and subscription

For further information, please contact Andonowati at aan@dns.math.itb.ac.id OR aantrav@attglobal.net

To subscribe, please send your cv (due date June 15, 2002) and the name/s of a person/s (with her/his/their contact address/es) who will give you recommendation (if asked) about your academic credibility to

PUSAT PENELITIAN PENGEMBANGAN DAN PENERAPAN MATEMATIKA (P4M)
INSTITUT TEKNOLOGI BANDUNG
Gedung Lab. Tek. III, Jl. Ganesha 10 Bandung, 40132
INDONESIA
Phone/Fax: +62 +22 250 8126
E-mail: p4mitb@bdg.centrin.net.id

Notification for acceptance will be sent before June 22.
Confirmation to attend the course should be submitted by July1, 2002

Target group, Participation

The topics are interesting, and the course is designed, for participants with different background: mathematics, physics, engineering. The division of the material in two levels guarantees that students with different level can profit: S1, S2 and S3 students.

For successful absorption of the material, participants should have basic knowledge of analysis and of ODE's and PDE's. The course will be for the major part be conducted in English, so sufficient level of English is required.

Subscription and selection

Only a LIMITED NUMBER of participants will be accommodated. Selection process will be based on the academic background of the candidates and her/his English proficiency. Please supply us as much information as possible on these matter.

Fee

No registration fee is required.
It is FREE

Awards...

There will be small awards for outstanding performance as well as for enthusiastic participation

Variational Methods in Science

with emphasis on applications from fluid dynamics and optics

ITB, July 15 – July 20, 2002

A one-week course with

- **integrated lectures and exercises**
- **extensive lecture notes**
- **'Maple' and 'Matlab' illustrations**

15/07	Basic Calculus of Variations
16/07	Linear Eigenvalue Problems
17/07	Nonlinear Eigenvalue Problems
18/07	Consistent Variational Approximations
19/07	Variational Numerics
20/07	Summary, epilogue

- 8:00-10:00 Motivation, examples
- 10:00-12:00 General theory
- 13:00-17:00 Class work

organised by
P4M – ITB, Andonowati
in collaboration with
AAMP – UTwente, E. van Groesen

Funded by KNAW through EPAM Industrial Mathematics Project

The subject

Optimisation properties are quite common in every-day life but also in the natural and engineering sciences where many problems are formulated in terms of ordinary or partial differential equations. Well known examples are principles of minimal potential energy, Fermat's principle of least time for light rays, and principle of stationary action for dynamical systems. But also solitons in surface waves or optics can be characterised by the fact that these profiles maximise the momentum at given energy (maximise speed at given cost).

An optimisation principle makes it possible to use special mathematical methods, which then can often be adapted for other problems as well. Besides that, from its origin of study by scientists like Euler, Lagrange, Newton, Huygens, the methods can deal remarkably well with nonlinear problems, and as such it is still one of the most coherent mathematical set of methods for study of nonlinear phenomena.

Since an optimisation property, or more generally a variational structure, is 'special' and leads to specific behaviour of the system, it is clear that when one wants to design simplified models of the system, the model should retain these properties, i.e. inherit the variational structure. In that sense the variational structure is an essential element, and guides the way, to consistent approximations in mathematical modelling. In particular this holds just as well if one looks for numerical discretizations: the numerical code should respect this basic property. Finite element methods are often directly associated with this, but also other methods can be used in 'variational discretizations'.

The course intends to introduce to the basic mathematical methods as well as to recognise variational structures in problems from the natural and engineering sciences.

Topics

The topics will be extensively motivated and illustrated by various examples, where the examples are chosen mainly from the natural and engineering sciences, in particular from problems in fluid dynamics (surface waves) and optics.

- Theory of first and second variation.
- Unconstrained problems: the Euler-Lagrange equations.
- Prescribed and natural boundary conditions.
- Weak formulations and interface conditions
- Direct and inverse problem of the Calculus of Variations
- Constrained problems: Lagrange's Multiplier Rule, and the multiplier as derivative of value function.
- Sturm-Liouville eigenvalue problems, more dimensional and nonlinear extensions.
- Dynamical systems: conservative, dissipative and thermodynamic systems
- Direct optimisation methods, steepest descent method
- Low-dimensional modelling retaining the variational structure
- Numerical codes from variational discretizations: FD and FEM-methods

Applications that will be dealt with include:

- Classical mechanics: Lagrangian and Hamiltonian systems
- Nonlinear oscillators, vibrations of drums and bars
- Variational structure of Maxwell equations and equations for surface waves, like Korteweg – de Vries equation, Nonlinear Schrodinger equation, etc.
- Solitons in surface waves and optics
- Wave propagation through wave guides
- Finite-mode approximations of dispersive wave equations
- FD and FEM numerical schemes for simple wave models.

It is possible that some participants may have specific questions or problems. These problems may be discussed, either included or outside the course week.

Variational Methods in Science

with applications in fluid dynamics and optics

E. (Brenny) van Groesen & Andonowati
Applied Analysis & Mathematical Physics AAMP,
University of Twente, The Netherlands
and

Center of Mathematics P4M,
Institut Teknologi Bandung, Indonesia

10 July 2002; update 18 July 02

Contents

Preface

The present Lecture Notes are prepared for a course given at ITB in summer 2002. The basic material in Chapters 1, 2 and 3 can be found scattered around in several text books; the text here is based on previous lecture notes at UTwente (course: Applied Analytical Methods' of EvG), but with restyled presentation to be attractive for a larger audience. The contents of Chapter 3 and the Appendices were written and assembled for this course.

The authors like to present these Lecture Notes as a gift to the participants of the course as their wedding present and hope that the love in which it is prepared and the enthusiasm for the topic show itself in these notes and in the execution of the course.

We will be grateful to the participants for remarks and criticism.

Bandung, 10 July 2002.

New lecture notes, one-week course...

Introduction

Optimality in the natural sciences

> *"..... je suis convaincu que par tout la nature agit selon quelque principe d'un maximum ou minimum."* (Euler, 1746)

This quotation of one of the greatest scientists that shaped the modern mathematical description and investigation of the natural sciences, expresses clearly the underlying expectation. The belief that optimization was important to describe natural phenomena was verified by Euler for various problems, and exploited to present a more thorough investigation of the problems. More far reaching conclusions were drawn by some other scientists:

> *"..... des loix du mouvement ou l'action est toujours employee avec la plus grande economie, demontreront l'existence de l'Etre supreme ... "*, (Maupertuis, 1757)

but this point of view belongs to metaphysics, and is as such not very fruitful for a deeper investigation[1].

Actually, optimization problems are known already from ancient times; well known is *Dido's problem*: the problem to find the plain domain of largest area given the circumference of the domain. Many other problems can also be formulated as *geodetic problems*, where one investigates those curves (or surfaces) with the property that a functional measuring the length (or the area) is as small as possible. A major example is the following

> **Fermat's principle**, 1662
> The actual trajectory of a light ray between two points in an inhomogeneous medium has the property that the time (or optical length) required to transverse the curve is as small as possible when compared to the time required for any other curve between the points.

In fact, the investigation of this principle led Fermat to the principal mathematical result that will be formulated in the first chapter as Fermat's algorithm. From Fermat's principle, *Snell's law* can be derived about the breaking of light between two media. A dual point of view (looking for the evolution of light

[1] It should be noted, however, that modern theoretical physicists who look for "a theory of everything" (Grand Universal Theory) actually search for functionals (Lagrangians) that produce the desired unified field equations upon optimization, just as Einsteins general theory of relativity is based on a minimality principle.

fronts, the surfaces that can be reached by the light from a point source in a given time) was investigated by Huygens, 1695. Huygens' *principle*, of vital importance for the basic understanding of light propagation, can be considered as a major example of what later has become known as duality methods.

These historical remarks[2] make it clear that practical problems from physics provided the initial motivation for the beautiful mathematical theory that has been developed since then.

Dynamical systems with a variational structure

Except from problems that have by their very nature an "obvious" formulation as a minimization problem (minimum length, minimum costs, etc.), there are many problems for which such an extremizing property exists, but not so obvious. Important examples can be found in dynamical systems.

The *principle of minimum (potential) energy* leads to equilibrium states for which the total energy is minimal (while the kinetic energy vanishes for equilibria). For nontrivial dynamic evolutions in certain systems, a less intuitive quantity, the 'action' (see the quotation of Maupertuis), turns out to be an important functional; actual evolutions correspond to saddle points (not extremizers in general) of this functional. Formulations of such systems were studied by Lagrange, Hamilton etc., and the many results are collected in what is now called *Classical Mechanics*, a well structured set of methods and results to study dynamical systems of collection of mass points, mechanical (rigid) structures etc.. Nowadays, much effort is done to generalize these ideas to partial differential equations for continuous systems such as fluid dynamics and field theories like optics.

The systems referred to above, Lagrangian and Hamiltonian systems, and more generally Poisson systems, are roughly speaking 'conservative' (the energy is conserved), and the dynamic motions have a variational nature. In many cases this structure makes it also possible that special (but important) solutions can be characterized in a variational way. These solutions can be equilibrium solutions or 'steady state solutions'. Often these are called *coherent structures* and are characteristic for such problems; examples are phenomena like 'solitons' and 'vortices' that appear in fluid dynamics and optics. Using their variational nature, these can be found in a systematic way.

Even when the system is not conservative, but (mainly) dissipative, such as in gradient and thermodynamic systems, equilibrium solutions can still be found by exploiting variational structures in the equations.

Modelling and calculating

If a problem has a variational structure, this is a special property. When one wants to make a simpler model or when a numerical algorithm has to be designed for accurate calculations, it is best to take care that this special structure is retained. One reason is that then corresponding properties immediately related to the variational structure are inherited; another reason that in the study of the simplified or discretized model variational methods can again be used. It is not easy to describe in a well-structured way this idea of 'variationally consistent' modelling, but certain basic approaches can be identified. The key method is to restrict the set on which the original functional is considered to a

[2] The interested reader may consult such references like Goldstein, and Newman vol. 2.

'suitable' subset. Probably the best known application of this idea is the Finite Element Method, a numerical method that is obtained by replacing functions from an infnite dimensional space to finite dimensional approximations. But also restrictions to other sets, for instance to functions with a specific profile that are characterized by only a few parameters, can be very useful to obtain simplified models that show the main properties of the full problem. Viewed in this way, 'variational' numerical methods are just variants of more general variationally consistent modelling.

Contents of the Lecture Notes

The lecture notes consist of chapters and a few appendices.
Most chapters start with a motivation and introductory examples; then the basic mathematical method is described in a more or less general setting, which is then illustrated to and specified for various application areas. Exercises finishes this basic material, but is then followed by more advanced methods that can be skipped in first reading.

The main classical *methods of the Calculus of Variations* are described in Chapters 1 and 2.

In Chapter 1 problems 'without constraints' are considered, and the main result is the vanishing of the (variational) derivative at a stationary point of a functional. Essentially this is a direct generalization of results for a finite number of variables, but the infinite dimensional setting will lead to equations that are usually (partial or ordinary) differential equations: the *Euler-Lagrange equations*; the treatment of *boundary conditions* requires special attention.

In Chapter 2 we consider optimization problems for which the functions to be considered do not form a linear space but satisfy certain 'constraints'; it will lead to the infinite dimensional generalization of *Lagrange's Multiplier Rule*. The appearance of a 'multiplier' that is not given in advance but has to be determined as part of the solution of the total problem, resembles the standard Eigenvalue Problem from Linear Algebra, which is why these problems can be called non-linear *Eigenvalue Problems*. Linear eigenvalue problems are a special but important case, which is why this treated in some detail; it will be shown that in many cases complete sets of eigenfunctions can be expected (Fourier theory is a characteristic example). The linear eigenvalue problem gives the best interpretation of infinite dimensional spaces and shows that 'operators' are generalizations of matrices in finite dimensional spaces.

Chapter 3 deals with *variationally-consistent modelling*. The basic idea is to restrict the functional to a subset of the original set; when the subset is simpler, the Euler-Lagrange equation will become simpler too. For instance, as for numerical purposes, when the subset is finite dimensional, this will bring the partial differential equation to an equation in the finite dimensional set, and we have obtained a 'variationally-consistent' discretization of the equation. One of the methods will be the Finite Element Method. Besides we will also show in a very condensed, but complete, way the most well-known method to solve the resulting algebraic equation: the method of steepest descent and its more practical implementation: the conjugate gradient method. The last method is nowadays the prime example of a direct optimization method.

In all these chapters, at several places and in many examples and exercises, problems from the natural sciences are not only used to illustrate the methods, but will also show that *variational structures* appear abundantly. More

elaborate applications from optics and surface waves, exploiting various methods treated in the chapters, are collected in appendices. Appendix A deals with Optics; the basic equations are given and modern ways to treat waveguide modes are described. Appendix B deals with the basic equations from fluid dynamics. Appendix C describes how solitons and wave groups can be obtained for the equations of optics and fluid mechanics: the similar variational structure of the equations leads to the applicability of the fundamental methods to both problems at the same time.

Chapter 1

Basic Calculus of Variations

1.1 Motivation and Basic Notions

1.1.1 Extremal problems in finite dimensions

In real analysis courses at an introductory level, functions of one or more variables are considered. The definition of differentiation of functions is a vital part of such courses, and a standard result is the following

> **Algorithm of Fermat**, for 1-D optimization problems[1].
> If the differentiable scalar function of one variable $f : \mathcal{R} \to \mathcal{R}$ attains a (local) extreme value at the point $\hat{x}$, then the derivative at that point vanishes:
>
> $$f'(\hat{x}) = 0.$$

Viewed as a condition for a point to be an extremal element, this condition is necessary but not sufficient; every point that satisfies this property is called a stationary, or critical, point, so including 'saddle points'.

Knowing the above result for functions of one variable, the generalization to functions of more variables, n dimensional problems, is remarkably simple: the use of partial derivatives reduces the n dimensional problem to n 1-D problems, as follows.

For $F : \mathcal{R}^n \to \mathcal{R}$, recall that at the point x the *directional derivative* in a direction η is found by differentiating the scalar function obtained by restricting F to the line through x in the direction η, i.e. the function $\varepsilon \to F(x + \varepsilon\eta)$,

$$\left.\frac{d}{d\varepsilon}\right|_{\varepsilon=0} F(x + \varepsilon\eta) \equiv DF(x)\eta.$$

Here $DF(x)$ is seen as a map from $\mathcal{R}^n$ into $\mathcal{R}$ as $\eta \to DF(x)\eta$. If x minimizes F on $\mathcal{R}^n$, this point certainly minimizes at $\varepsilon = 0$ the restriction to the line, and hence

$$\left.\frac{d}{d\varepsilon}\right|_{\varepsilon=0} F(x + \varepsilon\eta) = DF(x)\eta = 0.$$

[1] Fermat did not write down the actual equation; he reasoned that small variations near a minimizer produces a higher order variation in the function, the fundamental idea that leads to the result and justifies to adhere his name to the mathematical algorithm. Fermat didn't know the concept of derivative of functions other than polynomials; it was Leibniz who introduced in 1684 the concept of derivative of arbitrary functions.

If x minimizes F on $\mathcal{R}^n$, this should hold for any direction η, and the vanishing of the directional derivative in every direction η can be expressed by writing

$$DF(x) = 0.$$

It is common to rewrite this property by using the notion of gradient as follows.

For $F : \mathcal{R}^n \to \mathcal{R}$, let ∇F be the *gradient of the function*, defined to be the column vector

$$\nabla F(x) = \begin{pmatrix} \partial_{x_1} F(x) \\ ... \\ \partial_{x_n} F(x) \end{pmatrix}.$$

The relation with the directional derivative is simply that for each η

$$DF(x)\eta \equiv \nabla F(x) \cdot \eta$$

where in the rhs the usual inner product of vectors in $\mathcal{R}^n$ is meant. Then, from $DF(x)\eta = \nabla F(x) \cdot \eta = 0$ for all η, the vanishing of the map $DF(x)$ can now be expressed by the vanishing of the gradient (vector)

$$\nabla F(x) = 0.$$

This formulation is the direct generalization of Fermat's algorithm to n dimensional optimization problems.

1.1.2 Generalization to infinite dimensions

In this course we tackle problems for which a scalar function is defined on an infinite dimensional space (the function is then usually called a function*al*). Most times this means that the functional assigns to functions of one or several variables a real number by integrating powers of the function and its derivative: for instance the L_2-norm is an example of such so-called 'density-functionals. Then the above can be generalized as follows:

- by restricting the functional to one dimensional lines the notion of directional derivative can be defined just as easily; it will be called the *first variation* in that case;
- when dealing with density-functionals, a generalization of the gradient can be defined and will lead to the notion of *variational derivative*. The specific expression is related to the choice of the L_2 innerproduct for functions under consideration;
- the fact, which is trivial in finite dimensions, that from $\nabla F(x) \cdot \eta = 0$ for all η, it follows that $\nabla F(x) = 0$, can be generalized to infinite dimensional function spaces with the L_2 innerproduct as a consequence of *Lagrange's Lemma*; this result will enable us to identify the first variation with the variational derivative (not considering boundary contributions).

The infinite dimensional case also brings characteristic differences:

- the functions to which the functional assigns a certain value are defined on a domain; 'variations' of the functions may be restricted in the interior of the domain (for instance if the average of the function is prescribed to vanish) but also because of restrictions on the boundary; boundary conditions may be prescribed or may result for critical points (so-called natural boundary conditions);
- when talking about functions, clearly their smoothness (continuity, differentiability) may also be of importance; it also leads to the most characteristic difference with finite dimensional spaces that in infinite dimensional spaces different (non-equivalent) norms are possible: a function may be square integrable, while the squared derivative may have arbitrary large integral.

The typical notation to be used in the following for the variational derivative is $\delta\mathcal{L}(u)$, and Fermat's algorithm generalizes to

$$\delta\mathcal{L}(u) = 0$$

as the condition for a minimizing element. This equation is most times a differential equation, replacing the algebraic equation $\nabla F(x) = 0$ that is obtained for a minimizer of a function of a finite number of variables, together with boundary conditions.
Just as in finite dimensions, the second derivative may reflect minimization properties, and in general provide insight into the character of a critical point. In the Calculus of Variations these aspect are dealt with in the *theory of first and second variation*.

1.1.3 Notation and General Formulation

Generally speaking, for an optimization problem we have the following basic ingredients:

- a set of admissible elements $\mathcal{M}$, usually some subset of an (infinite dimensional) space $\mathcal{U}$;
- a functional $\mathcal{L}$, defined on $\mathcal{U}$ (or only on $\mathcal{M}$).

The minimization problem of $\mathcal{L}$ on $\mathcal{M}$ concerns questions about an element $\hat{u}$ that minimizes the functional on the set of admissible elements. By definition $\hat{u}$ is the element that, among all admissible elements in the set $\mathcal{M}$ set for which the functional achieves it's lowest value μ :

$$\mu = \mathcal{L}(\hat{u}) \text{ and } \mathcal{L}(\hat{u}) \leq \mathcal{L}(u) \quad \text{for all} \quad u \in \mathcal{M}.$$

We will use in the following the notation

$$\begin{aligned} \hat{u} &\in \text{Min } \{ \mathcal{L}(u) \mid u \in \mathcal{M} \} = \min_{u \in \mathcal{M}} \{ \mathcal{L}(u) \mid u \in \mathcal{M} \} \\ \mu &= \text{Min } \{ \mathcal{L}(u) \mid u \in \mathcal{M} \}. \end{aligned}$$

In principle, the questions could deal with the existence, the uniqueness, and the characterization and computation of the minimizer. In this course, we

will mainly deal with the *characterization* of the minimizer (and more general critical points). We will concentrate on the equation(s) that have to be satisfied by such a critical point: the Euler-Lagrange equation or Lagrange's multiplier rule, boundary conditions, etc..

1.1.4 Functionals

Now we will first deal with the basic ingredients of a variational problem: the functionals and the set of admissible elements.

The functionals we will encounter are mainly *density-functionals*. That means that the functional assigns to each function from a set of admissible functions a number that is found by integrating (powers) of the function and it's derivatives. So, for functions defined on the interval $(0, \ell)$, simple examples are $u(x) \to \int_0^\ell u^2(x)dx$ or $u(x) \to \int_0^\ell \left[(\partial_x u)^2 + u^4 + \sin(u)\right] dx$ etc..

The example of a functional that assigns the value of a (continuous) function in one point, such as $u(x) \to u(\ell/2)$ could be called a 'generalized' density functional since it can be written using Dirac's delta function[2] like $u(x) \to \int u(x)\delta_{Dir}(x - \ell/2)dx$.

The general form of a density functional cannot be given without complicated notation; therefore we just list the main type of functionals that we will encounter in the following with their characteristic names.

Sturm-Liouville type of functionals

For scalar functions u of one variable x, defined on in an interval I, the functional is of the form

$$\mathcal{L}(u) = \int_I \left[p(x)\,(\partial_x u)^2 - q(x)u^2 - f(x)u\right] dx \tag{1.1}$$

where p, q and f are given functions. As is common, the arguments (x) are suppressed in the expression under the integral sign. The functional can be defined for piecewise differentiable functions $u(x)$.

Lagrangian functionals from Classical Mechanics

For vector functions q, say $q \in R^N$, of one variable t (the time), and with $\dot{q}$ denoting the time derivative of q, the functional is of the general form

$$\mathcal{L}(q) = \int L(q, \dot{q}, t)dt \tag{1.2}$$

where the Lagrangian density L is a given, smooth, function of its $2N + 1$ $(R^N \times R^N \times R)$ arguments.

Dirichlet type of integrals

For scalar functions Φ defined on a domain $\Omega \subset R^n$, and with $\nabla\Phi$ the gradient of Φ, the functional integrates over the spatial domain

$$\mathcal{L}(\Phi) = \int \left[p(x)|\nabla\Phi|^2 - q(x)\Phi^2 - f(x)\Phi\right] dx$$

the more-dimensional variant of the SL-type of functionals.

[2] The symbol 'delta' δ will appear abundantly in this course: it will be used to denote a kind of differential operation (variational derivative), but will also appear to denote admissible variations. Therefore, to minimize confusion, Dirac's delta function will be denoted by δ_{Dir}.

Lagrangian functionals for evolution equations

For functions u depending on spatial variables x and the time t , a characteristic example is

$$\mathcal{L}(u) = \int \int_{\Omega} \left[\rho(x) \left(\partial_t u \right)^2 - c^2 \left(\partial_x u \right)^2 - q(x) u^2 \right] dx dt \tag{1.3}$$

and more generally with a Lagrangian density L :

$$\mathcal{L}(u) = \int \left[\int_{\Omega} L(\partial_t u, \partial_x u, u, x, t) dx \right] dt. \tag{1.4}$$

Sometimes one distinguishes between 'single-integral' functionals (when functions depend on only one independent variable, and only integration over that single variable is needed to arrive at a real number) and 'multiple-integral' functionals (when the functions depend on more independent variables), but from the general point of view there is no essential difference.

1.1.5 Bilinear functionals and quadratic forms

Here we recall some general notions that will be used repeatedly in the following.

A functional ℓ defined on a linear function space $\mathcal{U}$ is linear if for all $u, v \in \mathcal{U}$ and all $\lambda \in \mathcal{R}$

$$\ell(u + v) = \ell(u) + \ell(v), \quad \ell(\lambda u) = \lambda \ell(u).$$

A functional $b : \mathcal{U} \times \mathcal{U} \to \mathcal{R}$ is a bilinear functional if it is linear in each of its arguments, so

$$\mathcal{U} \ni v \mapsto b(u, v) \text{ is linear for all } u \in \mathcal{U},$$

and

$$\mathcal{U} \ni u \mapsto b(u, v) \text{is linear for all } v \in \mathcal{U}.$$

A bilinear functional b can have special properties:

$$\begin{aligned}
\text{symmetry} \quad &: \quad b(u, v) = b(v, u) \\
\text{skew-symmetry} \quad &: \quad b(u, v) = -b(v, u) \\
\text{non-degenerate} \quad &: \quad \begin{cases} [b(u, v) = 0 \ \text{ for all } u] \Rightarrow v = 0 \\ [b(u, v) = 0 \ \text{ for all } v] \Rightarrow u = 0 \end{cases} \\
\text{non-negative} \quad &: \quad b(u, u) \geq 0 \\
\text{positive} \quad &: \quad b(u, u) > 0 \ \text{ for all } u \neq 0.
\end{aligned}$$

Note the following fact:

$b(u, u) = 0$ if b is skew-symmetric.

A symmetric bilinear functional is a kind of generalized inner product; when it is positive, it is a true innerproduct. In all cases it defines a quadratic form.

Definition 1 *A functional $a : \mathcal{U} \to \mathcal{R}$ is called a* quadratic form *if there is a symmetric bilinear form b on $\mathcal{U}$ so that*

$$a(u) = b(u, u);$$

when b is positive, a is called a norm, and b is an innerproduct; when b is only non-negative a is called a semi-norm.

Usually a first check to recognize a functional as a quadratic form is to verify that it is homogeneous of degree 2; this is not sufficient, however.

Proposition 2 *Given a quadratic form, the symmetric bilinear functional is given by*

$$b(u, v) = \frac{1}{4}[a(u + v) - a(u - v)];$$

hence there is a one to one relation between quadratic forms and symmetric bilinear functionals.

For the symmetric bilinear functional b and the corresponding quadratic form a, the following relation holds:

$$a(u + \lambda v) = a(u) + 2\lambda b(u, v) + \lambda^2 a(v) \quad \text{for each } \lambda \in R$$

From this we derive the following important consequences:

Proposition 3 *When a is positive semi definite,* Cauchy-Schwartz *inequality holds:*

$$|b(u, v)|^2 \leq a(u)a(v).$$

Exercise.

1. Often a quadratic density functional has the form

   $$a(u) = \int Lu \cdot u$$

   where L is a (differential) operator. Find the corresponding bilinear form.

2. Show that $u \to \sqrt{a(u)}$ with

   $$a(u) = \int_a^b \{\partial_x u\}^2 dx$$

   defines a norm on the set of functions that satisfy $u(0) = 0$; write down the Cauchy-Schwartz inequality. Show that on the set of all smooth functions it is only a semi-norm.

3. Derive the corresponding bilinear functional for the following quadratic form

   $$a(u) = \int_a^b \{\sigma(x)u_x^2 + u^2\} dx.$$

   Compare this bilinear functional with the following one:

   $$c(u, v) = \frac{1}{2} \int_a^b \left[\{-\partial_x (\sigma(x)\partial_x u) + u\}v + \{-\partial_x (\sigma(x)\partial_x v) + v\}u\right] dx.$$

4. Under suitable conditions on the function σ and the space of functions on which it is defined, the quadratic form $\int_a^b \{\sigma(x)u_x^2 + u^2\}dx$ defines a norm; give some examples when it is a norm and then write down the Cauchy Schwartz inequality.

5. The same questions for the more dimensional generalization

$$a(u) = \int_\Omega \{\sigma(x)|\nabla u|^2 + u^2\}dx.$$

■

1.1.6 Admissible variations

Just as is the case that for functions of a finite number of variables its minimal value and the minimizer depend heavily on the domain on which this function is considered, for each variational problem a set of admissible elements should be specified: the set on which the functional is defined, and the functions that are allowed ('admissible') in the competition of looking for the minimizer.

In general, the *set of admissible elements* consists of functions $u(x)$ of one or more variables $x \in R^n$ defined on a certain domain (bounded or not) $\Omega \subset R^n$. Usually these functions are subject to certain conditions, conditions which may be of different character:

smoothness and integrability conditions, at least to assure that the density functional is well defined;

boundary conditions: conditions on the function and/or its derivatives on (parts of) the boundary $\partial\Omega$ of the domain Ω on which the functions are defined;

(internal) *constraints*.

An example of an 'internal constraint' is $\int_\Omega u(x)dx = 0$; it shows that the function over the full domain has to be considered to verify the condition: local variations of the function in the interior can easily make the integral nonzero.

It is essential for the following to describe this more clearly. Therefore we recall that when talking about the 'derivative' of a function at a certain point, we compare the function values at neighbouring points. To be able to conclude that the derivative vanishes at the point where the minimal value is achieved, it should be the case that neighbouring points are in the domain of definition for the optimization problem: a function of one variable that attains its minimal value at the boundary of an interval doesn't need to have vanishing derivative there. Hence, it is important to know in which 'directions' the function values can be compared. That brings us to the notion of (admissible) variation.

Take a function $u(x)$ defined on Ω. A *variation* of that function is a 'small' change of that function, possibly over its full domain Ω. To make 'small' a bit more precise: for any (finite) function $\eta(x)$ on Ω and ε sufficiently small, the function $\varepsilon\eta(x)$ is small and we can consider the function $u(x) + \varepsilon\eta(x)$ to be in the neighbourhood of $u(x)$. Stated differently, and getting close to the same interpretation as a line through a point u in the direction η in finite dimensions,

for given 'direction' $\eta(x)$ the line through the point $u(x)$ in the direction $\eta(x)$ is the set in function space:

$$\varepsilon \to u(x) + \varepsilon\eta(x);$$

it is a family of graphs in which the original graph of $u(x)$ is embedded, and which 'approaches' this function for vanishing ε. PLOTTTTTT
In the classical Calculus of Variations it is common to write $\delta u(x)$ for $\varepsilon\eta(x)$ with ε small and to call it a variation; then $\delta u(x)$ is more interpreted like different elements on the line with ε small than as a single function. To avoid this cumbersome interpretation, we will avoid the use of δu but will write $\varepsilon\eta(x)$ and make the dependence on the parameter ε explicit.

The above holds for any function, and any variation. Now we consider a specified set of admissible elements, to which will correspond at each element a set of admissible variations. The idea can be simply illustrated to a function of three variables that is restricted to the set of admissible points that are points that lie in a plane in the three-dimensional space. Then, for a given point in the plane, only variations 'in the plane' are allowed to be considered, not in the direction perpendicular to the plane for instance. In the example of a plane there are two independent directions such that a full line lies in the plane. When the plane is replaced by a curved surface, a manifold, not two full lines will belong to the manifold in general, but lines in so-called tangent directions (that form the tangent plane at the point of consideration) differ close to the point at the manifold only in higher order. These tangent directions will be the admissible variations. Generally speaking, for a set of admissible elements $\mathcal{M}$ we will define at a point $u \in \mathcal{M}$ the set of admissible variations the functions η such that $u + \varepsilon\eta$ belongs to $\mathcal{M}$ for small ε 'up to higher order'. We will specify this more precisely in Chapter 3, but introduce here already the notation $T_u\mathcal{M}$ for the set of admissible variations at the point u (common notation in geometry to denote the 'tangent'-space).

In this chapter we will deal with admissible sets for which each element (function) can be changed locally (in a neighbourhood of each interior point in its domain of definition) in an arbitrary way and that still the varied element remains admissible.

To give meaning to these statements, first we consider such local variations, i.e. we first introduce test functions. These functions will make it possible to change a given function in the interior of its domain of definition, without altering the behaviour at the boundary.

Definition 4 *Given a domain $\Omega \subset R^n$, the set of* test functions *on Ω will be denoted by $C_0^\infty(\Omega)$ and consist of all functions that are infinitely differentiable (C^∞) and that vanish, together with all derivatives, near the boundary $\partial\Omega$ (C_0^∞).*

Remark. Such test functions really exist: for any interior point $x_0 \in \Omega$ and r_0 such that $x \in \Omega$ if $|x - x_0| < r_0$ an example is

$$\phi(x) = \begin{cases} \exp\left(-\frac{1}{r_0^2 - |x-x_0|^2}\right) & \text{for } |x - x_0| < r_0 \\ 0 & \text{for } |x - x_0| \geq r_0 \end{cases}$$

■

Now we can define two essentially different classes of admissible elements, leading to different variational problems and different methods and results:

Definition 5 *An* unconstrained variational problem *is a problem for which the set of admissible elements $\mathcal{M}$ consists of functions defined on a domain Ω such that all test functions belong to the set of admissible variations:*

$$\text{if } u \in \mathcal{M} \quad \text{then } u + C_0^\infty(\Omega) \subset \mathcal{M} \ ,$$

meaning $u + \varepsilon\eta \in \mathcal{M}$ for each $\eta \in C_0^\infty(\Omega)$.
When this is not the case we will talk about a constrained variational problem*.*

For instance, a problem with admissible set $\mathcal{M}$ of functions that are required to satisfy $\int_\Omega u(x)dx = 0$, will be a problem with constraints, since adding a positive test function the condition will not be satisfied anymore. On the other hand, if $\mathcal{M}$ consists of 'all' smooth functions on Ω but possibly with restrictions on the boundary (boundary conditions), then this will be a problem without constraints.

Example. For the following given sets $\mathcal{M}$ the set of admissible variations $T_u\mathcal{M}$ is specified (in the examples this set does not depend on the point u); when $T_u\mathcal{M}$ contains the test functions (or not) the variational problem is without (or with) constraints. Observe that all these examples are 'planes' in the function space, affine spaces.

1. For $\mathcal{M} = \{\ u \in C^1(0,1) | u(0) = 2, \partial_x u(1) = 0\}$, $T_u\mathcal{M} = \{\eta \in C^1(0,1) | \eta(0) = 0, \partial_x u(1) = 0\} \supset C_0^\infty(0,1)$.
2. For $\mathcal{M} = \{u \in C^0(0,1) | u(0) = 2\}$, $T_u\mathcal{M} = \{\eta \in C^0(0,1) | \eta(0) = 0, \int \eta dx = 0\} \not\supseteq C_0^\infty(0,1)$.

■

1.2 Theory of first variation

In this section we derive the generalization of Fermat's algorithm as announced in the introduction. It must be noted that this is in fact a local result: assuming the existence of a minimizer, we derive the anticipated result; no conditions are stated that guarantee the existence of a minimizer.

1.2.1 First variation and variational derivative

The aim is to consider the 'derivative' of a functional. As stated already, it is natural to use the idea of directional derivative since then the problem is reduced to the differentiation of a scalar function of only one variable.
Hence, let u be a given function, and v an (arbitrary) variation. With this variation the original function u is embedded in a class of "varied" functions (a one-parameter family) of the form

$$\varepsilon \mapsto u + \varepsilon v.$$

Fixing v, and restricting the functional to this line, we get a scalar function of one variable:

$$\varepsilon \mapsto \mathcal{L}(u + \varepsilon v).$$

The derivative of this function is then by definition the directional derivative, the first variation.

Definition 6 First variation
The first variation of a functional $\mathcal{L}$ at u in the direction v is denoted by $\delta\mathcal{L}(u;v)$ and defined as

$$\delta\mathcal{L}(u;v) = \frac{d}{d\varepsilon}\mathcal{L}(u+\varepsilon v)\Big|_{\varepsilon=0}. \tag{1.5}$$

In most cases, the first variation is linear in v (nonlinear in u in general). When it is linear in v (and continuous with respect to a topology on the space), it is also known as the Gateaux-derivative, it is the direct generalization of the directional derivative of a function on a finite dimensional space.

From the definition of first variation above, it follows directly that a linear approximation of $\mathcal{L}(u+\varepsilon v)$ is given as

$$\mathcal{L}(u+\varepsilon v) = \mathcal{L}(u) + \varepsilon\delta\mathcal{L}(u;v) + o(\varepsilon) \tag{1.6}$$

where, here and in the following, $o(\varepsilon)$ means terms that are of higher than first order in ε: $o(\varepsilon)/\varepsilon \mapsto 0$ for $\varepsilon \mapsto 0$.

The definition above applies to all kind of functionals. For density functionals that we will consider mostly, it is usually possible to perform a partial integration and to rewrite $\delta\mathcal{L}(u;v)$ as the $L_2(\Omega)$-innerproduct of v and some function which will be denoted by[3] $\delta\mathcal{L}(u)$ (and which will be the direct generalization of the gradient of a function of a finite number of variables).

This may require the function u to be smooth enough, and usually a contribution consisting of an integration over the boundary appears in addition:

$$\delta\mathcal{L}(u;v) = \int_\Omega \delta\mathcal{L}(u)\cdot v + \int_{\partial\Omega} b(u;v) \tag{1.7}$$

If functions v are considered that vanish on the boundary, the boundary contribution vanishes identically. Therefore, we can use in particular the class of test functions $C_0^\infty(\Omega)$ to avoid these boundary contributions. Then we have the following notion.

Definition 7 *The function $\delta\mathcal{L}(u)$ on Ω defined by the condition*

$$\begin{aligned}\delta\mathcal{L}(u;\eta) &= <\delta\mathcal{L}(u),\eta> \\ &\equiv \int_\Omega \delta\mathcal{L}(u)\cdot\eta\ dx,\ \textit{for all } \eta\in C_0^\infty(\Omega)\end{aligned} \tag{1.8}$$

is called the variational derivative *of the functional $\mathcal{L}$ at the point u.*

[3]For notational convenience we will exploit the notation $\delta\mathcal{L}(u)$, although in much of the literature the notation $\delta\mathcal{L}/\delta u$ is often used:

$$\delta\mathcal{L}(u) \equiv \frac{\delta\mathcal{L}}{\delta u}(u).$$

It will follow from Lagrange's Lemma 11 below that when $\delta\mathcal{L}(u)$ is continuous, (1.8) indeed defines the function $\delta\mathcal{L}(u)$ uniquely. We will give various examples in the following to demonstrate the calculation of the variational derivative.

Remark. Sometimes it is more convenient to extend the notation a little bit, and to interpret the variational derivative $\delta L(u)$ not only as a function, but to interpret $\delta L(u)$ as a convenient notation for the linear functional $\delta L(u;\cdot)$, and so to extend (1.8) to

$$\delta\mathcal{L}(u;v) = \langle \delta\mathcal{L}(u), v\rangle, \text{ for all } v \in T_u\mathcal{M} \tag{1.9}$$

where then $\langle \delta\mathcal{L}(u), v\rangle$ is nothing more than a symbolic way of writing the first variation. But note that then boundary terms are included in the interpretation. In the applications the formulae will be mostly exploited for the variational derivative, i.e. by taking for $\langle\ ,\ \rangle$ the L_2-innerproduct. ■

1.2.2 Characteristic cases

For the characteristic functionals mentioned before, the first variation and the variational derivative will be given here in illustrative notation.

Sturm-Liouville type of functionals

For

$$\mathcal{L}(u) = \int_I \left[\frac{1}{2}p(x)\left(\partial_x u\right)^2 - \frac{1}{2}q(x)u^2 - f(x)u\right] dx$$

we get

$$\begin{aligned} \delta\mathcal{L}(u;v) &= \int_I \left[p(x)\left(\partial_x u\right)\left(\partial_x v\right) - q(x)uv - f(x)v\right] dx, \\ \delta\mathcal{L}(u) &= -\partial_x\left(p(x)\partial_x u\right) - q(x)u - f(x) \end{aligned} \tag{1.10}$$

Lagrangian functionals from Classical Mechanics

For

$$\mathcal{L}(q) = \int L(q,\dot{q},t)dt$$

(using vector notation, and a sloppy, but characteristic, way of writing the derivative of L with respect to the 'variables' that collect the vector $\dot{q}$)

$$\begin{aligned} \delta\mathcal{L}(q;\xi) &= \int \left[\frac{\partial L}{\partial \dot{q}}\dot{\xi} + \frac{\partial L}{\partial q}\xi\right] dt, \\ \delta\mathcal{L}(q) &= -\frac{d}{dt}\left[\frac{\partial L}{\partial \dot{q}}\right] + \frac{\partial L}{\partial q}. \end{aligned} \tag{1.11}$$

Dirichlet type of integrals

For the more-dimensional variant of the SL-type of functionals

$$\mathcal{L}(\Phi) = \int_\Omega \left[\frac{1}{2}p(x)|\nabla\Phi|^2 - \frac{1}{2}q(x)\Phi^2 - f(x)\Phi\right] dx$$

the first variation is given by

$$\delta\mathcal{L}(\Phi,\psi)=\int_\Omega\left[p(x)\nabla\Phi\cdot\nabla\psi-q(x)\Phi\psi-f(x)\psi\right]dx$$

To find the variational derivative we have to perform a 'partial integration' for this multiple integral. This is done by using the following basic elements from differential calculus (often we will write for a vector $\mathbf{a}$ it's divergence as $\mathrm{div}(\mathbf{a})$ or as $\nabla\cdot\mathbf{a}=\mathrm{div}(\mathbf{a})$): first the basic identity:

$$\text{for scalar } \alpha \text{ and vector } \mathbf{a}: \qquad \mathrm{div}(\alpha\mathbf{a})=\nabla\alpha\cdot\mathbf{a}+\alpha\mathrm{div}(\mathbf{a})$$

with which we can write $p(x)\nabla\Phi\cdot\nabla\psi=\mathrm{div}(\psi p(x)\nabla\Phi)-\psi\mathrm{div}(p(x)\nabla\Phi)$. Secondly we recall

$$\textbf{Gausz' theorem}: \int_\Omega \mathrm{div}(\mathbf{a})=\int_{\partial\Omega}\mathbf{a}\cdot\mathbf{n}$$

where $\mathbf{n}$ is the outward pointing normal to the boundary $\partial\Omega$ of Ω.
Then we find:

$$\int_\Omega p(x)\nabla\Phi\cdot\nabla\psi=\int_\Omega-\psi\nabla\cdot[p(x)\nabla\Phi]+\int_{\partial\Omega}\psi p(x)\nabla\Phi\cdot\mathbf{n}$$

Using the common notation: $\partial_n\Phi=\nabla\Phi\cdot\mathbf{n}$ for this *normal derivative*, we have found

$$\delta\mathcal{L}(\Phi,\psi)=\int_\Omega\left[-\nabla\cdot[p(x)\nabla\Phi]-q(x)\Phi-f(x)\right]\psi dx+\int_{\partial\Omega}\psi p(x)\partial_n\Phi \tag{1.12}$$

For finding the variational derivative, we restrict to test functions (which vanish at the boundary) and find

$$\delta\mathcal{L}(\Phi)=-\nabla\cdot[p(x)\nabla\Phi]-q(x)\Phi-f(x) \tag{1.13}$$

Finally, we recall the special notation for the Laplace operator:

$$\textbf{Laplace operator:} \qquad \Delta=\nabla\cdot\nabla=\partial_x^2+\partial_y^2+...$$

so, for instance

$$\delta\int_\Omega|\nabla\Phi|^2=-\Delta\Phi$$

Lagrangian functionals for evolution equations
For the characteristic example

$$\mathcal{L}(u)=\int\int_\Omega\left[\frac{1}{2}\rho(x)\left(\partial_t u\right)^2-\frac{1}{2}c^2\left(\partial_x u\right)^2-\frac{1}{2}q(x)u^2\right]dxdt$$

we find

$$\delta\mathcal{L}(u;v)=\int\int_\Omega\left[\rho(x)\left(\partial_t u\right)\left(\partial_t v\right)-c^2\left(\partial_x u\right)\left(\partial_x v\right)-q(x)uv\right]dxdt$$

and after partial integration (special case of Gausz, but now using repetitive integration may be easier)

$$\delta\mathcal{L}(u) = -\rho(x)\partial_t^2 u + \partial_x\left[c^2\left(\partial_x u\right)\right] - q(x)u;$$

similarly, for the more general functional:

$$\mathcal{L}(u) = \int\left[\int_\Omega L(\partial_t u, \partial_x u, u, x, t)dx\right]dt$$

$$\delta\mathcal{L}(u) = -\frac{\partial}{\partial t}\left[\frac{\partial L}{\partial(\partial_t u)}\right] - \frac{\partial}{\partial x}\left[\frac{\partial L}{\partial(\partial_x u)}\right] + \frac{\partial L}{\partial u} \tag{1.14}$$

(Be careful with confusing notation: the 'partial derivatives' like $\frac{\partial}{\partial t}$ here also differentiate the function u and its derivatives that appear in the arguments.)

Quadratic forms

For any quadratic form a the directional derivative is given by twice the corresponding bilinear functional:

$$\delta a(u; v) = 2b(u, v)$$

and when L is the operator corresponding to $a : a(u) =< Lu, u >$, then the variational derivative is

$$\delta a(u) = 2Lu$$

1.2.3 Stationarity condition

We now consider the basic optimization problem.
Let $\mathcal{M}$ be a smooth manifold, and, as before, let $T_u\mathcal{M}$ denote the set of admissible variations (tangent space to $\mathcal{M}$ at u). Considering $\mathcal{L}(u + \varepsilon v)$ for an admissible variation, in general this value will differ from $\mathcal{L}(u)$ in first order in ε as follows from (1.6). A critical point will be defined by the fact that this difference is of higher (most times second) order.

Definition 8 *A point $\hat{u}$ is called a* critical point, or stationary point, *of the functional $\mathcal{L}$ on the set $\mathcal{M}$ if the following holds:*

$$\delta\mathcal{L}(\hat{u}; v) = 0 \quad \textit{for all } v \in T_{\hat{u}}\mathcal{M}. \tag{1.15}$$

Of course, as in finite dimensions, the notion of 'critical point' is a generalization of a local maximum or minimum:

Proposition 9 *If $\mathcal{L}$ has a local maximal or minimal value at $\bar{u}$, then $\bar{u}$ is a critical point of $\mathcal{L}$.*

This is a basic result in the theory of 'first variation': *(1.15) gives the condition for a point to be a critical point, and this is a necessary condition (usually not sufficient) for a point to be a local maximum or minimum.*

1.2.4 Euler-Lagrange equation

It is possible to translate condition (1.15) into an explicit equation for $\hat{u}$ along the following lines.
Let $\hat{u}$ be a critical point of the *unconstrained* variational problem for $\mathcal{L}$ on $\mathcal{M}$. From the stationarity condition (1.15) and the fact that all test functions are admissible variations ($T_u\mathcal{M} \supset C_0^\infty(\Omega)$), it follows that certainly it must hold that

$$\delta\mathcal{L}(\hat{u};\eta) \equiv \langle \delta\mathcal{L}(\hat{u}), \eta\rangle = 0 \quad \text{for all } \eta \in C_0^\infty(\Omega). \tag{1.16}$$

This leads to the equation for $\hat{u}$:

Proposition 10 Euler-Lagrange equation
If $\hat{u}$ is a critical point of the unconstrained variational problem for $\mathcal{L}$ on $\mathcal{M}$, then (provided $\delta\mathcal{L}(\hat{u})$ is a continuous function) $\hat{u}$ satisfies

$$\delta\mathcal{L}(\hat{u}) = 0. \tag{1.17}$$

This equation for $\hat{u}$ is called the Euler-Lagrange equation of the functional $\mathcal{L}$.

The proof of this result is an immediate consequence of the first order condition (1.16) and the following basic Lemma.

Lemma 11 Lagrange's Lemma
Let f be a continuous function on Ω that is such that

$$\int_\Omega f(x)\eta(x)dx = 0 \text{ for all } \eta \in C_0^\infty(\Omega).$$

Then f vanishes identically on (the interior of) Ω: $f(x) = 0$ for all $x \in \Omega$.

Proof. Suppose that at some interior point x^* of Ω the function f does not vanish, say has value $\alpha > 0$. Then, from continuity of f, there is a small neighbourhood of f such that f doesn't vanish there, and in fact, $f(x) > \alpha/2$ for all $x, |x - x^*| < r_0$ for small enough r_0. Now take a test function, say $\bar{\eta}$, that is nonnegative (positive) inside, and vanishes outside this neighbourhood. Then $\int_\Omega f(x)\bar{\eta}(x)dx > \alpha/2 \int_\Omega \bar{\eta}(x)dx > 0$, contradicting the assumption. ■

1.2.5 Natural boundary conditions

From the vanishing of the first variation for all test functions, the Euler-Lagrange equation is obtained. For unconstrained problems, there may be more admissible variations than only test functions. In that case, for a critical point it should also hold that the boundary contribution in (1.7) vanishes:

$$\int_{\partial\Omega} b(\hat{u}; v) = 0, \quad \text{for all } \; v \in T_{\hat{u}}\mathcal{M}. \tag{1.18}$$

For admissible variations different from test functions this condition will give certain conditions for $\hat{u}$ on the boundary $\partial\Omega$. If this happens, these conditions are called natural boundary conditions: they appear as additional conditions for a critical point, not by the requirement that $\hat{u}$ should belong to $\mathcal{M}$, but from

(1.18), which is a consequence of the stationarity condition (1.15).

Example. The vertical deflection $u(x)$, with $x \in [0,1]$, of a string under the influence of an external force $f(x)$ is governed by the principal of minimal potential energy. That is to say: the potential energy, which is given by

$$\mathcal{L}(u) = \int_0^1 \left[\frac{1}{2} \left(\partial_x u \right)^2 - f(x) u \right] dx,$$

attains the minimal possible value at the actual (physical) deflection, minimal when compared to all other virtual deflections. To specify this more, let us assume the string is fixed at the origin, and 'free' at the other endpoint: $x(1)$ is arbitrary. Then, for this variational principle, the set of admissible elements are all possible deflections that can be collected in the set $\mathcal{M}$:

$$\mathcal{M} = \{ \; u \in C^1(0,1) \mid u(0) = 0 \; \}.$$

The admissible variations are all the functions that vanish at the origin, but are otherwise unrestricted, in particular, have arbitrary value at $x = 1$:

$$T_u \mathcal{M} = \{ \; v \in C^1(0,1) \mid v(0) = 0 \; \};$$

clearly all test functions are admissible variations, so this is an unconstrained variational problem.
Now assume that $\bar{u}$ is the minimal element; then the stationarity result leads to

$$\begin{aligned} \delta \mathcal{L}(\bar{u}; v) &= \int_0^1 \left[\partial_x \bar{u} \cdot \partial_x v - f(x) v \right] dx \\ &= \int_0^1 \left[-\partial_x^2 \bar{u} - f(x) \right] v dx + v \partial_x \bar{u}]_{x=0}^{x=1} \\ &= 0 \text{ for } \textbf{all } v \in T_u \mathcal{M} \end{aligned}$$

Now, first exploit the vanishing of the first variation for test functions: then

$$\int_0^1 \left[-\partial_x^2 \bar{u} - f(x) \right] \eta dx + \eta \partial_x \bar{u}]_{x=0}^{x=1} = \int_0^1 \left[-\partial_x^2 \bar{u} - f(x) \right] \eta dx = 0$$

for all test functions η and therefore (with Lagrange's lemma) we find the Euler-Lagrange equation:

$$-\partial_x^2 \bar{u} - f(x) = 0, \text{ for } x \in (0,1).$$

Now use this fact in the vanishing of the first variation for *any* admissible variation and, using the fact that $v(0) = 0$, find:

$$\int_0^1 \left[-\partial_x^2 \bar{u} - f(x) \right] v dx + v \partial_x \bar{u}]_{x=0}^{x=1} = 0 + v \partial_x \bar{u}]^{x=1} = v(1) \partial_x \bar{u}(1).$$

Then this expression has to vanish (as a consequence of the vanishing of the first variation) for all admissible variations, and therefore for any value of $v(1)$. This then leads trivially to the requirement that

$$\partial_x \bar{u}(1) = 0.$$

Hence, the stationarity condition not only produces the Euler-Lagrange equation, but also a boundary condition, here a Neumann condition at the 'free' endpoint. This boundary condition was *not* prescribed in advance; it is an example of a natural boundary condition. (Observe that it would not have appeared if also at the right endpoint the deflection had been prescribed in advance, for instance $u(1) = 0$.)
From a mathematical point of view, the prescribed boundary condition at $x = 0$, and the natural boundary condition at $x = 1$ together make the problem for the Euler-Lagrange equation a well posed Boundary Value problem (BVP): a unique solution exists.
From a physical point of view, the natural boundary condition means that the actual deflection will be horizontal at the free endpoint. When the right endpoint would have been prescribed, this can not be expected, except for a very special value of the deflection at that point. ■

1.2.6 Weak formulation and Interface conditions

We have seen that the variational derivative of a density functional follows from the first variation by applying a partial integration, restricted to test functions to avoid contributions at the boundary:

$$\delta\mathcal{L}(\hat{u};\eta) \equiv \langle \delta\mathcal{L}(\hat{u}), \eta\rangle \quad \text{for all } \eta \in C_0^\infty(\Omega).$$

Without having explicitly stated that, this *assumes* some additional regularity of the term to be integrated. For instance, looking at

$$\int \partial_x u \partial_x \eta dx = \int -\eta \partial_x^2 u \text{ for each } \eta \in C_0^\infty \tag{1.19}$$

it is seen that the lhs is well defined for (piecewise) differentiable functions u, while a simple interpretation of the rhs requires the function u to be (piecewise) twice differentiable. That implies that when writing down the Euler-Lagrange equation $\delta L(u) = 0$, in general some smoothness assumptions had to be made about the extremal function. On the other hand, we have actually seen that the first variation is the basic result from differentiation in a direction, so without partial integration, that result is valid. So we could just as well *interpret* the result after partial integration, e.g. $-\partial_x^2 u$, even in cases when it is not a continuous function, by requiring (1.19) to hold ! This turns out to be a very fruitful idea to generalize the notion of differentiability of functions that are not differentiable in the classical way. This is then often called generalized derivative, or distributional derivative (in the theory of generalized functions). When we interpret a BVP in this way, we often talk about the variational formulation of the problem, or about the *weak formulation*. It is also the basis of Finite Element methods, where a given equation is not interpreted pointwise, but integrated against spline functions, so that a weak formulation results.

Example.

1. Consider the Heaviside function of unit step:

$$H(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 & \text{for } x > 0 \end{cases}$$

The derivative at the point $x = 0$ cannot be defined in the classical sense, but in the distributional sense it is Dirac's delta function

$$\frac{d}{dx}H(x) = \delta_{Dir}(x)$$

because for integration intervals that contain the origin it holds

$$\int H(x)\partial_x\eta(x) = -\eta(0) = -\int \delta_{Dir}(x)\eta(x) \quad \text{for each } \eta \in C_0^\infty.$$

In the same way, derivatives of Dirac-function can be uniquely defined.

2. The differential equation

$$-\partial_x^2 u = H(x)$$

has the corresponding weak formulation

$$\int \partial_x u \partial_x v = \int H(x)v(x) \quad \text{for each } \eta \in C_0^\infty.$$

3. Now consider the following problem that is characteristic for problems in integrated optics where material properties change abruptly:

$$\partial_x^2 u + (k^2 + \alpha H(x))u = 0$$

with k, α constants. At each side of the origin the solutions are simple, but the jump in the coefficient at the origin makes it at first sight unclear how to connect these solutions. It is natural to require the solution to be continuous at $x = 0$. This jump in coefficient will lead in general (when $u(0) \neq 0$) to a finite jump in the second derivative. And this can only be true if the first derivative is continuous (if it had a jump, the second derivative would be a Dirac delta function). Hence, for a unique interpretation, this differential equation has to be accompanied by so-called *interface conditions* at the point where the equation has to be interpreted in a generalized sense:

$$u \text{ and } \partial_x u \text{ continuous at } x = 0.$$

In solving the equation explicitly, and in the design of discretization schemes, these interface conditions are essential to match the solutions in the two half lines together. It should be noted that the weak formulation

$$\int \left[(\partial_x u)(\partial_x \eta) - (k^2 + \alpha H(x))u\eta\right] = 0 \quad \text{for each } \eta \in C_0^\infty$$

for continuous functions u automatically lead to the continuity of the derivative (split the integral as the sum of integrals over the two half-lines, and observe the contributions at the origin from the two sides). So *continuity of the derivative will be a direct consequence of the variational formulation*, it does not have to be required separately. This can be exploited in the design of numerical schemes that start with the discretization of the corresponding functional

$$\int \left[(\partial_x u)^2 - (k^2 + \alpha H(x))u^2\right].$$

■

1.3 Principle of Minimal Potential Energy

For time independent problems, or for stationary states of time dependent problems, the actual physical state may be described by a *principle of minimum (potential) energy*, which means the following:

- there is a set of admissible, physically acceptable, states $\mathcal{M}$,
- there is a (potential) energy functional $\mathcal{E}$ that assigns a value ("energy-like") $\mathcal{E}(u)$ to each state $u \in \mathcal{M}$,
- the actual physical state is the state $\hat{u}$ that minimizes $\mathcal{E}$ on $\mathcal{M}$.

We present several examples to illustrate the applicability.

Dirichlet's principle

In a domain $\Omega \subset \mathcal{R}^3$ with an electrostatic field E, the potential energy is $\int_\Omega \frac{1}{2}E^2$. Since rot $E = 0$, the field is conservative: $E = -\nabla\phi$ for an electro-magnetic potential ϕ. In the presence of a charge distribution ρ in the domain, the total electrostatic energy is given by

$$\mathcal{E}(\phi) = \int_\Omega \{\frac{1}{2}|\nabla\phi|^2 - \rho(x)\phi\}dx.$$

Dirichlet's *principle* states that the actual field is such that it minimizes the total energy among all potentials that satisfy certain boundary conditions.
Two types of boundary conditions are usually considered. When the boundary consists of two parts $\partial\Omega = \partial\Omega_1 \cup \partial\Omega_2$, they can be described as

- $\partial\Omega_1$ is conducting, i.e. $E \cdot \tau = 0$ for each tangent vector τ; this is achieved by requiring $\phi = 0$ on the boundary;
- $\partial\Omega_2$ is insulating: $E \cdot n = 0$ on the boundary. This implies that the normal derivative of ϕ vanishes on the boundary $\partial_n\phi = 0$.

The minimization problem

$$\hat{\phi} \in \text{Min } \{\ \mathcal{E}(\phi) \mid \phi(x) = 0 \text{ for } x \in \partial\Omega_1\ \}$$

leads to the boundary value problem

$$\begin{cases} -\Delta\phi &= \rho(x) & \text{in } \Omega, \\ \phi &= 0 & \text{on } \partial\Omega_1, \\ \partial_n\phi &= 0 & \text{on } \partial\Omega_2. \end{cases}$$

Observe that the Neumann condition on $\partial\Omega_2$ arises as a natural boundary condition!
Also note that when $\partial\Omega_1$ is empty (only Neumann conditions) a solution can only exist if $\int \rho = 0$. Inhomogeneous Dirichlet and Neumann boundary conditions can be obtained also: the Dirichlet conditions by prescribing the potential, the Neumann condition by adding a suitable boundary functional to the energy.

Exercise.

1. Show that a critical point of

$$\text{Crit } \{ \mathcal{E}(\phi) - \int_{\partial\Omega_2} \psi_2\phi \mid \phi(x) = \psi_1(x) \text{ for } x \in \partial\Omega_1 \}$$

satisfies

$$\begin{cases} -\Delta\phi & = & \rho(x) & \text{in } \Omega, \\ \phi & = & \psi_1 & \text{on } \partial\Omega_1, \\ \partial_n\phi & = & \psi_2 & \text{on } \partial\Omega_2. \end{cases}$$

2. Show that there exists at most one critical point, and that, if it exists, it is in fact a minimizer.

3. When $\partial\Omega_1$ is empty, derive the necessary condition between ψ_2 and ρ for a solution to exist. How is this condition related to the finiteness of the minimum value, i.e. to the boundedness from below, of the functional?

■

Bars and plates, strings and membranes

An elastic medium is characterized by the fact that deformations from a given rest state require a certain amount of energy. In general, the local energy density will depend on the extension as well as on the curvature of the medium.
For simplicity we first restrict to 1D elastic media with a rest state along the x-axis, and deformations in a plane. Two idealizations are

- *strings*: completely flexible, but extension requires energy,
- *bars*: fixed length (inextricable), but bending requires energy.

Theory of bars

For a bar with (fixed) length ℓ it is natural to use the arclength as parameter and to describe its position in the plane as

$$r(s) = (x(s), y(s)), \quad \text{or} \quad r_s(s) = (\cos\theta(s), \sin\theta(s)).$$

The *curvature* $k(s)$ at a point s is defined (up to sign) by

$$k(s)^2 = |r_{ss}|^2 \equiv |\theta_s(s)|^2.$$

The material properties can be described with a local energy density $E = E(s, k(s))$ (depending on position and curvature); in the presence of an external additional potential energy $V = V(r)$, the total energy is then given by

$$\int_0^\ell \{E(s, k(s)) + V(r(s))\} ds.$$

Usually, E is an even function of k and minimal at $k = 0$. Linear elasticity theory (assuming small curvatures) then approximates E like $E(s,k) \approx E(s,0) + \frac{1}{2}\sigma(s)k^2$ leading to an approximate energy functional

$$\int_0^\ell \{\frac{1}{2}\sigma(s)|r_{ss}|^2 + V(r(s))\}ds,$$

with Euler -Lagrange equation

$$\partial_s^2[\sigma(s)r_{ss}] + \nabla_r V(r) = 0$$

When looking for small vertical deformations u from the rest state along the x-axis, x instead of s is used as the independent variable. (Hence a laboratory coordinate x instead of the material coordinate s; note that then, with $x \in [0,\ell]$, the bar slightly extends.) If f is a prescribed vertical force ($\approx \partial_y V(x,0)$), the resulting energy functional becomes

$$\int_0^\ell \{\frac{1}{2}\sigma(x)u_{xx}^2 - f(x)u(x)\}dx.$$

The Euler-Lagrange equation reads

$$\partial_x^2[\sigma u_{xx}] = f(x), \text{ for } x \in (0,\ell).$$

Concerning boundary conditions, two types can be distinguished:

- *supported endpoint*: only the position is prescribed, for instance $u(0) = 0$;
- *inclined endpoint*: a more restricted condition for which both the position and the angle are prescribed, for instance $u(0) = 0, u_x(0) = 1$.

When a bar is supported at one endpoint, say $x = 0$, a natural boundary condition arises in addition to the prescribed position:

$$u(0) = 0, \quad \sigma(0)u_{xx}(0) = 0;$$

at an inclined end point, no natural boundary conditions arise.

Theory of strings

In a string, both longitudinal and transverse displacements of particles will occur. If one restricts to small deflections form a state of rest along the x-axis, the change in length is in lowest order given by

$$[\sqrt{1+u_x^2} - 1] \approx \frac{1}{2}u_x^2.$$

With $\sigma(x)$ the tension in the undeformed state, and f an external vertical force, the approximated potential energy is given by

$$\int_0^\ell \{\frac{1}{2}\sigma u_x^2 - f(x)u\}dx,$$

leading to the Euler-Lagrange equation

$$-\partial_x[\sigma(x)u_x] = f(x) \text{ for } x \in (0,\ell).$$

When the deflection is not prescribed at an end point, a natural boundary condition appears: $\sigma u_x = 0$.

2D-elasticity: plates and membranes

2D elasticity is a direct generalization of 1D elastica; in the linear approximation, the analog of a bar is a plate and has potential energy

$$\int_\Omega \{\frac{1}{2}\sigma(\Delta u)^2 - f(x)u\}dx;$$

the analog of a string is a membrane, with approximate potential energy

$$\int_\Omega \{\frac{1}{2}\sigma|\nabla u|^2 - f(x)u\}dx.$$

Boundary conditions, prescribed and natural boundary conditions, can be of the same type as in the 1D case.

1.4 Dynamical Systems and Evolution Equations

In this section we consider various dynamical systems with some variational structure. We start with the classical systems from Classical mechanics, and end up with Poisson systems, the most general structure that seems easiest applicable to partial differential equations. Also dissipative systems will be considered from a variational point of view.

1.4.1 Classical Mechanics

Problems from Classical Mechanics deal with the motion of a finite number of point-masses, usually with interaction between them and in a force field. These systems are described by ordinary differential equations for the position of each of the masses. Roughly speaking, these equations are in the form of newton's law of force. But, and that is characteristic for Classical Mechanics, the systems are 'conservative', in general: there is no dissipation. In many cases some 'total energy' quantity is conserved. From the conservative nature it follows that the equations of motion themselves have a variational structure: the actual motions are critical points of a certain functional, the 'action functional'. We consider two ways of describing such systems; the Lagrangian and the Hamiltonian way. It may be observed that the integrals are integrals over time, from an initial to a final time. Boundary conditions are NOT considered in general, meaning that at these moments the actual configuration is supposed to be known; note that this (boundary-value formulation, without, however, specifying the positions) is very different from an initial-value problem.

Lagrangian systems

First the general definition, with the characteristic nomenclature, then simple examples

Definition 12 *A dynamical system with position vector $q \in R^N$ is a Lagrangian system if a Lagrangian $L(q, \dot{q}, t)$ can be given such that critical points of the* action functional *(or* Lagrangian functional)

$$\mathcal{L}(q) = \int L(q, \dot{q}, t)dt$$

(the corresponding Euler Lagrange equations) are the dynamic equations of the system. This variational principle is often called the action principle.

Consider the motion of a single mass-point of mass m moving along the x-axis, position at time t denoted by $q(t)$. Consider the Lagrangian functional

$$L(q) = \int \left[\frac{1}{2}m\dot{q}^2 - V(q,t)\right] dt$$

which assigns to a certain trajectory $t \to q(t)$ between initial and final time the value determined by L.

The action-principle then states that the actual, physical trajectory is the one that is a critical point of the Lagrangian functional; more precisely; given initial position $q(t_i) = P$ and final position $q(t_f) = Q$, the admissible trajectories are those that connect these points in the specified time interval and the admissible variations are deformations of the trajectory that vanish at the initial and final time. The Euler-Lagrange equations is given by:

$$m\ddot{q} + \frac{\partial V}{\partial q} = 0.$$

This is a simple form of *Newton's equation* for a system of one degree of freedom: $\ddot{q}$ is the acceleration, and $-\frac{dV}{dq}$ is the 'force', which is here the derivative of the so-called *potential energy* function $V(q)$; such force-fields are called 'conservative'.

Note that the *Lagrangian density is the difference between kinetic energy and potential energy*; this is typically the case for systems from Classical Mechanics..

Also, consider the *total energy* $E(q,\dot{q})$:

$$E(q,\dot{q}) := \frac{1}{2}m\dot{q}^2 + V(q,t)$$

and calculate directly that for solutions of the equations it holds

$$\frac{d}{dt}E(q,\dot{q}) = m\ddot{q}\dot{q} + \frac{\partial V}{\partial q}\dot{q} + \frac{\partial V}{\partial t} = \left[m\ddot{q} + \frac{\partial V}{\partial q}\right]\dot{q} + \frac{\partial V}{\partial t} = \frac{\partial V}{\partial t}.$$

Hence, *if the Lagrangian density does not explicitly depend on time, i.e.* $\frac{\partial V}{\partial t} = 0$, *the total energy is conserved*:

$$\frac{d}{dt}E(q,\dot{q}) = 0 \quad \text{if } \frac{\partial V}{\partial t} = 0.$$

Again, this is a property that holds in more general systems as well, as we shall see.

For such systems of one degree of freedom, energy conservation allows phase-plane analysis: in the phase plane $(q,\dot{q})$ the motion of the particle is restricted to a level set of the total energy $E(q,\dot{q}) = E_0$, where the value E_0 is determined by specifying an initial position and velocity.

Example. *Phase plane analysis for oscillators*[4]
We now consider simple potential energy functions, and find the characteristic behaviour of the solutions form the phase plane analysis. Usually this type of

[4] We will use phse-plane analysis in the investigation of soliton-profiles in KdV and NLS equations later on.

equations are called 'oscillator equations': they describe the motion of a mass point attached to a spring that exerts a force on the particle depending on its deviation form the rest-position at the origin. The simplest case of a linear restoring force leads to periodic motions around the origin; nonlinear effects will disturb these motions a little bit near the origin, but may have large influence for large-amplitude motions; this can be seen from the phase plane analysis.

1. *Harmonic oscillator*: Show that trajectories are ellipses for the well known linear oscillator of which all solutions can be written down explicitly:

$$m\ddot{q} + \omega^2 q = 0$$

2. Nonlinear oscillator with *quadratic nonlinearity*:

$$m\ddot{q} + \omega^2 q + \alpha q^2 = 0$$

3. Nonlinear oscillator with qubic nonlinearity, *Duffing's equation*:

$$m\ddot{q} + \omega^2 q + \gamma q^3 = 0$$

■

Exercise. *Generalization to systems of N degrees of freedom.*
The above example deals with a system of one particle moving along a line: one degree of freedom. Now we consider more degrees of freedom; either because there are more particles with a one-dimensional motion, or one particle but moving in more dimensions.

1. Consider Lagrangian density for position vector $q = q(t) \in R^N$ given by a smooth function such that the Lagrangian functional reads:

$$L = L(q, \dot{q}, t)$$

Show that the Euler Lagrange equations are now N equations that in vector form are written like

$$-\frac{d}{dt}\left[\frac{\partial L}{\partial \dot{q}}\right] + \frac{\partial L}{\partial q} = 0$$

Specialize this result for a system of N mass-points with mutual interactions between them:

$$L = \Sigma_{k=1}^{N} \frac{1}{2} m_k \dot{q}_k^2 - V(q_1, ..., q_N)$$

2. Consider one particle of mass m that moves in the (x, y) plane: $t \rightarrow (x(t), y(t))$, under the influence of a conservative force field from a potential energy function $V(x, y)$.

   (a) Formulate this case in the notation of the exercise above.

   (b) Take the special case that the potential energy does not depend on time; show energy conservation.

(c) Suppose the potential energy depends only on the distance of the particle from the origin

$$V(x,y) = W\left(\sqrt{x^2+y^2}\right),$$

for which ∇V is a so-called 'central' force field. Write down the equations.

(d) Introduce polar coordinates (r,ϕ) in the plane, so that the trajectory is described by $t \to (r(t),\phi(t))$. Write down the equations of motion in terms of $(r(t),\phi(t))$ by transforming the original equations in Cartesian coordinates. Do the same for the total energy.

(e) In the last case, the equations can also be obtained by performing the transformation in the Lagrangian: find the transformed Lagrangian in terms of $(r(t),\phi(t))$ and $(\dot{r}(t),\dot{\phi}(t))$. Take for this Lagrangian the action principle and find the Euler Lagrange equations for $(r(t),\phi(t))$. Conclude that the same equations are found.

(f) Observing the equations of motion in polar coordinates, except the total energy, another constant of the motion is found, the angular momentum; what is the physical interpretation? Note that is a consequence that the variable ϕ does not appear explicitly in the Lagrangian: a missing coordinate in the Lagrangian is called 'cyclic' and automatically gives rise to a corresponding constant of the motion; show this in general.

(g) Take as a comparable case the motion of a spherical pendulum: write down the Lagrangian and find the equations of motion, the energy and angular momentum conservation.

3. Define the *total energy* of a system with Lagrangian $L(q,\dot{q},t)$ by

$$E(q,\dot{q},t) = \dot{q}\cdot\frac{\partial L}{\partial \dot{q}} - L$$

Show that there is *energy conservation* when L does not depend on t explicitly:

$$\frac{d}{dt}E(q,\dot{q},t) = 0 \quad \text{for solutions if } \frac{\partial L}{\partial t} = 0$$

4. Consider as a specific example of an *infinite dimensional Lagrangian system* for functions $u(x,t)$ the Lagrangian density

$$L(u,\partial_t u) = \int\left[\frac{1}{2}(\partial_t u)^2 - \frac{c^2}{2}(\partial_x u)^2 - fu\right]dx;$$

note the notation: $L(u,\partial_t u)$ is, on the one hand, a functional as far it concerns the dependence on x, and therefore it actually depends on two functions, here denoted by u and $\partial_t u$, which makes sense after considering the action functional which really maps the function u of (x,t) into the reals:

$$\int L(u,\partial_t u)dt$$

The Euler-Lagrange equation is a forced wave equation

$$\partial_t^2 u = c^2 \partial_x^2 u - f$$

which is supplemented by prescribed and/or natural boundary conditions, depending on the conditions of the functions in spatial variables. A more dimensional analog, with more spatial dimensions, is $L(u, \partial_t u) = \int \left[\frac{1}{2}(\partial_t u)^2 - \frac{c^2}{2}|\nabla u|^2 - fu \right] dx$ with Euler-Lagrange equation

$$\partial_t^2 u = c^2 \Delta u - f$$

■

Classical Hamiltonian systems

One way to introduce Hamiltonian systems is as an alternative description for of Lagrangian systems, although the correspondence is not one-to-one. The observation is that the Euler Lagrange equations for Lagrangian L are typically second order in time. If a first order in time description is preferred (for which there may good reasons, for instance from a conceptual point of view of the IVP – Initial Value problem –) this can be done by introducing more dependent variables: except the position vector $q \in R^N$ one introduces momentum-type of variables $p \in R^N$ and looks for first order in time system of equations in the pair $(q, p) \in R^N \times R^N$. Starting with a Lagrangian system, this can often be done in a systematic way using Legendre transformation.

Somewhat more general, we define

Definition 13 *A dynamical system is called a classical* Hamiltonian system *if the dynamical equations can be described with pairs of variables* $(q, p) \in R^N \times R^N$ *in the so-called the* phase space*, and with a* Hamiltonian $H(t, q, p) : \mathcal{R} \times R^N \times R^N \to \mathcal{R}$ *in and such that the dynamical equations are found from the* canonical action principle*, which means as the critical points of the* canonical action functional

$$\mathcal{A}_c(q, p) = \int \left[p(t) \cdot \partial_t q(t) - H(t, q(t), p(t)) \right] dt,$$

and hence satisfying the so-called Hamilton equations

$$\begin{cases} \partial_t q &= \frac{\partial H}{\partial p} \\ \partial_t p &= -\frac{\partial H}{\partial q} \end{cases} \tag{1.20}$$

In many problems form classical and continuous mechanics, *the Hamiltonian is the sum of kinetic and potential energy, i.e. the total energy*; this is different from the Lagrangian, which often is the difference between kinetic and potential energy.

Almost immediately seen from the equations is that for autonomous Hamiltonians there is *energy-conservation*:

$$\text{when } \frac{\partial H}{\partial t} = 0 \text{ the Hamiltonian is conserved: } \frac{d}{dt} H(q, p) = 0.$$

Exercise.

1. As a specific example of a finite dimensional system, verify that Newton's equations as given in Lagrangian form by the system of second order equations with mass matrix M and potential energy $V(q)$ by

   $$M\ddot{q} = -\partial_q V(q)$$

   are also obtained for the Hamiltonian

   $$H(q,p) = \frac{1}{2} p \cdot M^{-1} p + V(q)$$

   since then Hamilton's equations read:

   $$\begin{aligned} \partial_t q &= Mp \\ \partial_t p &= -\partial_q V(q). \end{aligned}$$

   Observe that H is indeed the total energy.

2. For plane fluid motions, when the flow is supposed to be irrotational, the Eulerian velocity field $v(x,y)$ can be written with a stream function ψ like

   $$v(x,y) = (\psi_y, -\psi_x).$$

   Observing that the fluid velocity is the particle velocity,

   $$v(x,y) = (\partial_t x, \partial_t y),$$

   it is clear that the particle dynamics is a Hamiltonian system, with the stream function as Hamiltonian. This is an example for which a description with a Lagrangian is not possible in general.

3. Consider as a specific example of an *infinite dimensional Hamiltonian system* for functions $u(x,t), p(x,t)$ the Hamiltonian

   $$H(u,p) = \int \left[\frac{1}{2} p^2 + \frac{c^2}{2} (\partial_x u)^2 + fu \right] dx;$$

   The Hamiltonian is a functional on the space of spatially depending pairs of functions (u,p). The canonical action functional now reads

   $$\int \left\{ \int [p \partial_t u] \, dx - H(u,p) \right\} dt$$

   and Hamilton's equations are

   $$\begin{aligned} \partial_t u &= \delta_p H(u,p) = p \\ \partial_t p &= -\delta_u H(u) = c^2 \partial_x^2 u - f \end{aligned}$$

   describing the same forced wave equation as treated earlier in the Lagrangian setting.

■

1.4.2 Poisson systems

Another way of writing Hamilton's equations will lead the way to a very fruitful genuine generalization. The other way is to recognize that the classical Hamilton's equations can be written like

$$\partial_t \begin{pmatrix} q \\ p \end{pmatrix} = \begin{pmatrix} 0 & I_N \\ -I_N & 0 \end{pmatrix} \begin{pmatrix} \partial_q H \\ \partial_p H \end{pmatrix}$$

where I_N is the identity matrix in R^N. Using the notation $u = (q, p)$, $J = \begin{pmatrix} 0 & I_N \\ -I_N & 0 \end{pmatrix}$, this can be written in the compact form

$$\partial_t u = J \nabla H(u)$$

J is called the standard *symplectic matrix*. It is skew-symmetric $J^* = -J$, and also invertible: $J^{-1} = -J$, $J^2 = -I_{2n}$. Observe that the skew-symmetry immediately implies the conservation of H (when not depending explicitly on time):

$$\frac{\partial}{\partial t} H(u) = \nabla H(u) \cdot \partial_t u = \nabla H(u) \cdot J \nabla H(u) = 0$$

the last equality since for skew-symmetric matrix it holds that $\mathbf{a} \cdot J\mathbf{a} = 0$ for any vector $\mathbf{a}$.

This description motivates a generalization of Hamiltonian systems to so-called Poisson systems. These are systems that have a conservative structure and appear quite regularly is systems form classical mechanics and especially in continuous systems from mathematical physics. The structure of the equation implies that there is an integral, the Hamiltonian of the system (which is most times the total energy). Here we investigate only the simplest properties, and in particular generalize the symplectic matrix J from above to *any* skew-symmetric operator. However, we do restrict to operators that are 'constant', i.e. do not depend on the variables of the system; this makes the theory much simpler, but also less interesting from a geometric point of view[5].

The following definition applies equally well for finite as for infinite dimensional systems; the wording that we use for infinite dimensional case.

Definition 14 *A* Poisson system *in the so-called* state space $\mathcal{U}$ *is a dynamical system for which the equations of motion are of the form*

$$\partial_t u = \Gamma \ \delta H(u) \tag{1.21}$$

with:

- $H : \mathcal{U} \to \mathcal{R}$, *a functional called the* Hamiltonian,
- δH *the variational derivative, defined with the innerproduct in which the operator*

[5] In particular, the famous Jacobi identity will be automatically satisfied for constant operators, and the many Lie-algebraic consequences are not considered here, nor the relation between integrals in (Poisson) involution and the commutation property of their flows. see e.g. [13].

- Γ *is a linear and skew-symmetric operator (sometimes called* structure map*)*

$$\langle \zeta, \Gamma\eta \rangle = -\langle \Gamma\zeta, \eta \rangle, \quad \textit{for each pair } \zeta, \eta \in \mathcal{U}$$

Related to the specific operator Γ *one defines the* Poisson bracket *of two functionals like*

$$\{\ F, G\ \} := \langle \delta F,\ \Gamma \delta G \rangle$$

In the cases that we will consider, the innerproduct is the standard L_2-innerproduct of the function space $\mathcal{U}$ or the standard innerproduct when is finite dimensional; in the last case, $\delta H(u) = \nabla H(u)$.

We will now consider the simplest dynamic properties of Poisson systems.

Equilibria

If Γ is not degenerate (invertible), the only equilibria are elements $\hat{u}$ with

$$\delta H(\hat{u}) = 0, \tag{1.22}$$

i.e. the critical points of H are equilibrium solutions:

$$\hat{u} \in \text{Crit } \{\ H(u) \mid u \in \mathcal{U}\ \}. \tag{1.23}$$

Remark. Note that these are the same as for the gradient system $\partial_t u = -\delta H(u)$, which equation may be used to construct the special critical points that are the minimizers of H. ■

If Γ is degenerate, elements $\bar{u}$ such that $\delta H(\bar{u}) \in \ker(\Gamma)$ will be other equilibria.

Diagnostics

Any functional F evolves according to

$$\partial_t F(u) = \{\ F, H\ \}(u).$$

In particular, the next result holds.

Proposition 15 *For the Poisson system* $\partial_t u = \Gamma \delta H(u)$*, a functional* I *is a first integral iff* I *Poisson commutes with* H*, meaning* $\{I, H\} = 0$*:*

$$\partial_t I(u) = 0 \quad \textit{iff} \quad \{I, H\} = 0.$$

Since $\{H, H\} = 0$ *from skew-symmetry, the Hamiltonian* H *itself is a first integral:*

$$\partial_t H(u) = 0.$$

Canonical Hamiltonian systems

The standard example of a Poisson system is a classical Hamiltonian system we started with as motivation; it is often called a canonical Hamiltonian system. The *canonical Poisson bracket* with J as structure map is given by

$$\{F, G\} = \nabla F \cdot J \nabla G,$$

Complex canonical structure

Associated to the real canonical structure described by (**??**) there is a natural *complex structure.* The essential relation is that the symplectic matrix J in real space corresponds to multiplication with the imaginary unit i in complex space.

Briefly, for $z \in C^n$, let $\bar{z}$ denotes the complex conjugate:

$$\text{if } \; z = q + ip \in C, \; (q,p) \in \mathcal{R}^n \times \mathcal{R}^n, \quad \text{then } \; \bar{z} = q - ip.$$

The inner product of z_1, and z_2, with $z_k = q_k + ip_k$, reads

$$\langle z_1, z_2 \rangle_C = \mathrm{Re}(z_1 \cdot \bar{z_2}) = q_1 \cdot q_2 + p_1 \cdot p_2$$

where Re denotes the real part. Functions (real valued) $\hat{F}$ on C^n are related to (real valued) functions F on $\mathcal{R}^{2n}$ by $\hat{F}(z) = \hat{F}(q+ip) = F(q,p)$ and for the derivative it holds

$$d\hat{F}(z) = \partial_q F + i\partial_p F.$$

Hence the Poisson bracket

$$\{\hat{F}, \hat{G}\}(z) := \langle d\hat{F}(z), -id\hat{G}(z) \rangle_C \tag{1.24}$$

is naturally related to the real canonical bracket (**??**):

$$\{\hat{F}, \hat{G}\} = \partial_q F \cdot \partial_p G - \partial_p F \cdot \partial_q G = \{F, G\}.$$

The state equation for a Poisson system with Hamiltonian $\hat{H}$ is

$$\partial_t z = -id\hat{H}(z). \tag{1.25}$$

Example. A system of n uncoupled harmonic oscillators with (real) frequencies $\omega_1, \ldots, \omega_n$ is described in complex variables with a Hamiltonian

$$H(z) = \Sigma_k \; \frac{1}{2}\omega_k |z_k|^2, \; z = (z_1, \ldots, z_n) \in C^n$$

as

$$\partial_t z_k = -i\omega_k z_k, \quad 1 \le k \le n.$$

■

1.4.3 Evolution equations (Nonlinear Wave equations)

We now briefly describe the particular Poisson structure that is found for various types of non-linear wave equations. More details are given in the Appendices A and B.

Boussinesq Equations

Equations for surface waves on a layer of fluid, and optical pulses, that depend on one spatial variable but allowing waves running in both directions often have the form

$$\partial_t \, u = -\partial_x \; \delta_\eta \, H(u, \eta), \qquad \partial_t \, \eta = -\partial_x \, \delta_u \, H(u, \eta),$$

i.e. written in characteristic Poisson-form:

$$\partial_t \begin{pmatrix} u \\ \eta \end{pmatrix} = - \begin{pmatrix} 0 & \partial_x \\ \partial_x & 0 \end{pmatrix} \begin{pmatrix} \delta_u H(u,\eta) \\ \delta_\eta H(u,\eta) \end{pmatrix}$$

where u is a velocity-type of variable, and η is the surface elevation. The Hamiltonian usually consists of a part that determines the linear dispersive properties, and a part to account for the nonlinearity in the equations.

KdV (Korteweg - de Vries) Equation

Equations for surface waves on a layer of fluid, and optical pulses, that depend on one spatial variable and are restricted to waves running mainly in one direction often have the form

$$\partial_t \eta = -\partial_x \delta_u H(\eta),$$

where η is the surface elevation. Again, the Hamiltonian usually consists of a part that determines linear dispersive properties, and a part to account for the nonlinearity in the equations.

NLS (Non-Linear Schrodinger) Equation

When modulations of a linear monochromatic wave are studied for KdV-type of equations, the complex amplitude $A(x,t)$ satisfies an NLS-type of equation that is of the form of a complex infinite dimensional Hamiltonian system:

$$\partial_t A = i\delta H(A),$$

with Hamiltonian that accounts for linear dispersive and nonlinear effects.

1.4.4 Gradient systems (Steepest decent)

Gradient systems usually describe systems with some dissipative character. These systems can also be used in a constructive way to calculate minimizers of a given smooth functional.

Definition 16 *A* gradient system *is a dynamical system in the state space $\mathcal{U}$ of the form*

$$\partial_t u = -\delta H(u)$$

where $H : \mathcal{U} \to \mathcal{R}$ a given functional.

Remark. It is possible to consider, somewhat more general, equations of the form

$$\partial_t u = -S\delta H(u)$$

with: S a linear and self-adjoint operator $\langle \zeta, S\eta \rangle = \langle S\zeta, \eta \rangle$, that is positive-definite: $\langle \zeta, S\zeta \rangle > 0$. Then much of the following is easily generalized to this case. Observe the essential difference with Poisson systems: instead of a skew-symmetric structure map Γ we now have a symmetric operator S. ■

Observe that equilibrium solutions of a gradient system arise from a variational principle:

Proposition 17 *For a gradient system $\partial_t u = -\delta H(u)$, the dynamic* equilibrium solutions $\hat{u}$ *are precisely the critical points of the functional H:*

$$\hat{u} \in Crit_u H(u): \quad \delta H(\hat{u}) = 0.$$

The role that the functional H plays for the dynamics can be understood by studying the evolution of H on trajectories and explains why such systems are called *dissipative:* it is a fact that H decreases monotonically outside equilibria:

$$\partial_t H(u) = \langle \delta H(u), -\delta H(u) \rangle = -|\delta H(u)|^2 \quad \begin{cases} \leq 0 \\ = 0 \quad \text{iff} \quad \delta H(u) = 0 \end{cases} .$$

This shows in particular that if $\bar{u}$ is an isolated local minimizer of H, trajectories starting close eventually approach the point $\bar{u}$; it is said that the point $\bar{u}$ is an *asymptotically stable equilibrium solution*; the rate of convergence to the minimizer depends on the geometry of the level sets of H near $\bar{u}$.

Proposition 18 *Local minimizers of H are asymptotically stable equilibrium solutions of the gradient system.*

The decrease of H on solutions also clearly shows that the solutions define trajectories of *steepest descent*: at each point the trajectory is along the direction of steepest descent of the functional H. They define the most 'efficient' way to reach lower values of H. This observation can be exploited in a constructive way. Using a discretization for the time derivative, these equations are often used to find a (possibly local) minimizer of H in a numerical way.
The dynamic behaviour near a local minimizer (where H decreases to its lowest possible value) is completely determined by the topological properties of the level sets of H.

Exercise.

1. In $\mathcal{R}^n$ consider

   $$\partial_t x = -\nabla H(x)$$

   with behaviour of the function H like $H(x) \approx |x|^\alpha$ near 0 for some $\alpha > 1$. Suppose that $x = 0$ is a (local) minimizer for H. Then determine the rate of convergence to 0 depending on the value of α.

2. The simple linear diffusion equation

   $$\partial_t u = u_{xx}, \quad u(0) = u(\pi) = 0$$

   is of the form of a gradient system with $H(u) = \int \frac{1}{2} u_x^2$. Determine the rate of convergence to the zero state of any solution; compare this with the exact general solution that can be written in a Fourier series with respect to the spatial variable.

3. The non-linear diffusion equation (with $\alpha \in \mathcal{R}$)

   $$\partial_t u = u_{xx} + \alpha u(1-u), \quad u(0) = u(\pi) = 0.$$

is a gradient system too:

$$\partial u = -\delta H(u) \quad \text{with} \quad H(u) = \int [\frac{1}{2}u_x^2 - \alpha(\frac{1}{2}u^2 - \frac{1}{3}u^3)]dx.$$

Investigate for which values of α the trivial solution $u \equiv 0$ is the minimizer, and find the rate of convergence in that case.

■

1.5 Exercises

1. *Calculus for variational derivatives*
   Since functionals map functions into $\mathcal{R}$, functionals can be added and multiplied. Verify the following rules of calculation that are well known for functions on finite dimensional spaces:

$$\begin{aligned}
\textit{linearity} \quad &: \quad \delta(\mathcal{L}_1 + \mathcal{L}_2) = \delta\mathcal{L}_1 + \delta_2; \\
\textit{product rule} \quad &: \quad \delta(\mathcal{L}_1.\mathcal{L}_2) = \mathcal{L}_2\delta\mathcal{L}_1 + \mathcal{L}_1\delta_2; \\
\textit{quotient rule} \quad &: \quad \delta\frac{\mathcal{L}_1}{\mathcal{L}_2} = \frac{\mathcal{L}_2\delta\mathcal{L}_1 - \mathcal{L}_1\delta_2}{\mathcal{L}_2^2} \\
\textit{for } g : \mathcal{R} \to \mathcal{R} \quad &: \quad \delta g(\mathcal{L}) = g'(\mathcal{L})\delta\mathcal{L}
\end{aligned}$$

   Derive the corresponding expressions for the second variation.

2. Calculate the first variation and the variational derivative of the following functionals. Below we use the notation with subscripts to denote the derivative: $u_x = \partial_x u$, $u_{xx} = \partial_x^2 u$.

   (a) $\mathcal{L}(u) := \int_0^1 \left[x\, u(x)^2 + u_x(x)^2\right]\, dx$

   (b) $\mathcal{L}(u) := \int_0^1 \left[sin(x)\, u(x)^2 + x^3\, u_x(x)^2\right]\, dx$

   (c) $\mathcal{L}(u) := \int \left[u(x)^2 + u_x(x)^2\right]\, dx$

   (d) $\mathcal{L}(u) := \int_0^1 \left[sin(u(x)) + u_{xx}(x)^2\right]\, dx$

   (e) $\mathcal{L}(u) := \int_0^1 \left[u(x)^4 + u_x(x)^7\right]\, dx$

   (f) $\mathcal{L}(u) := \int_0^1 n(x)\, \sqrt{1 + u_x(x)^2}\, dx$

   (g) $\mathcal{L}(q) := \int \left[\frac{1}{2}\dot{q}(t)^2 - \frac{1}{2}q(t)^2 + q(t)^3\right]\, dt$

   (h) $\mathcal{L}(u) := \int \left[\frac{1}{2}u_x(x)^2 + x^3 \sin(u(x)) + u(x)^5\right]\, dx$

   (i) $\mathcal{L}(u) := \int L(x, u, u_x)dx$, with L a given smooth function of its arguments.

   (j) $\mathcal{L}(u) := \int L(x, u, u_x, u_{xx})dx$, with L a given smooth function of its arguments.

3. *Conservative force fields and calculation of the potential.*
A differentiable vector field $F : \mathcal{R}^n \to \mathcal{R}^n$ is called *conservative* if there exists a scalar function $f : \mathcal{R}^n \to \mathcal{R}$ (the so-called potential) such that $F(x) \equiv \nabla f(x)$.
Given a conservative filed F and then to find the potential f is the more-dimensional analog of finding the primitive of a function of one variable. It is often called the 'inverse' problem.

   (a) Find the conditions on F that guarantee that it is conservative.

   (b) To solve the inverse problem, observe that if f is the potential, then its derivative along a curve $\xi(s)$ can be written like:

$$\frac{d}{ds} f(\xi(s)) = F(\xi(s)) \cdot \frac{d\xi}{ds}$$

   Show from this that the potential is uniquely defined (up to its value at one point, here taken to be the point x^*) by F and can be found by integrating along an *arbitrary* curve $\xi(s)$ from $x^* = \xi(0)$ to $x = \xi(1)$

$$f(x) - f(x^*) = \int_0^1 F(\xi(s)) \cdot \frac{d\xi}{ds} ds$$

   Since the path is arbitrary, a simple path (linear line through x^* and x) can be taken:

$$f(x) - f(x^*) = \int_0^1 F(x^* + s(x - x^*)) \cdot (x - x^*) ds$$

   (c) Consider the following vector fields F on $\mathcal{R}^n$; determine which ones are conservative, and which ones are not. If conservative, write down the potential.

$$\begin{aligned} n &= 2 : F(x, y) = (2x \sin(xy) + x^2 y \cos(xy), x^3 \cos(xy) + y^3) \\ n &= 2 : F(x, y) = (x^2 \sin(y), x^2 \cos(y)) \\ n &= 3 : F(x, y, z) = (2xy \sin(z) + x^3, x^2 \sin(z) + z, x^2 y \cos(z) + y) \end{aligned}$$

4. ** *Inverse problem of the Calculus of Variations*
The variational derivative (and Euler-Lagrange equation) of a given functional can be written down. How is it possible to see for a given equation (bvp) if it is the Euler-Lagrange equation of some functional, and if it is, how can we find the functional? This is the 'inverse problem' of the Calculus of Variations, and a generalization of 'conservative vector fields' of the previous exercise.
For simplicity, we restrict to investigate the equation only, forgetting about boundary values, but these can be included.
The operator $u \mapsto E(u)$ is called *conservative* if there exists some functional $\mathcal{L}$, again called the potential, such that

$$E(u) \equiv \delta \mathcal{L}(u).$$

Let $E'(u)$ denote the formal derivative at the point u. Prove the following result.

(a) **Proposition 19** *The operator E is conservative if its derivative defines a symmetric bilinear form, i.e. if*

$$\langle E'(u)\xi, \eta\rangle = \langle \xi, E'(u)\eta\rangle$$

for all functions ξ, η.
If that is the case, the potential is given (up to a constant) by

$$\mathcal{L}(u) - \mathcal{L}(0) = \int_0^1 \langle E(su), u\rangle \, ds.$$

5. Find the variational formulation of each of the following boundary value problems:

   (a) $-u_{xx} = \sin(u) + e^x u^2,\ u(0) = 0, u_x(1) = 7$. (Make sure the Neumann condition arises as a natural boundary condition by introducing a simple boundary functional to the density functional that produces the correct equation as E-L-equation.)

   (b) $-\frac{1}{r}\partial_r(r\partial_r u) = f(r),\ u_r(0) = u(1) = 0$. (It may be helpful to interpret r as the radial coordinate in a description with polar coordinates.)

   (c) $-div\,[\sigma(x,y)\nabla u(x,y)] + u(x,y) = 0,\ u(x,0) = u(x,1) = 0; u_x(0,y) = 1; u_x(1,y) = 0$.(The Neumann conditions as natural boundary conditions.)

6. *Linear two-point boundary value problem*
   For given $f \in C^0([0,1])$ consider

$$\mathcal{L}(u) = \int_0^1 \{\frac{1}{2}u_x^2 - f(x)u\}dx.$$

   (a) Prove: $\hat{u} \in C^2$ is a solution of the bvp

$$\begin{cases} -u_{xx} = f & on\ (0,1) \\ u(0) = u_x(1) = 0 \end{cases}$$

   iff $\hat{u}$ is the only critical point of $\mathcal{L}$ on

$$M_0 = \{\ u \ \text{ piecewise differentiable} \mid u(0) = 0\ \};$$

   in fact it is a minimizer for $\mathcal{L}$ on this set. (Concerning additional regularity for a critical point, see also the next Chapter, the Exercise on "Lemma DuBois-Reymond, Integrated Euler-Lagrange equation".)

   (b) Show that for the Neumann problem

$$-u_{xx} = f, \quad u_x(0) = u_x(1) = 0,$$

   there exists a solution iff $\int_0^1 f(x)dx = 0$. If it exists, the solution is not unique. Moreover show that

   - if $\int_0^1 f(x)dx = 0$, $\hat{u}$ is a solution iff it is a minimizer (not isolated) of $\mathcal{L}$ on the set of piecewise differentiable functions (no restrictions on the boundary);

- if $\int_0^1 f(x)dx \neq 0$, $\mathcal{L}$ does not have a critical point on the set of piecewise differentiable functions (no restrictions on the boundary); the infimum of this functional is $-\infty$.

7. *Light rays, Fermat's principle*
According to Fermat, the trajectory of a light ray between two points is such that the required time is as small as possible. The propagation speed of light depends on material properties, which is expressed by c_0/n where c_0 is the speed in vacuum (which is maximal), and $n > 1$ is the so-called index of refraction, characteristic for the material.
For trajectories, for simplicity described as graphs of functions $x \to y(x)$, the total time between points is

$$\int n(x,y)\sqrt{1+y_x^2}dx$$

This is also often called the *optical pathlength*. Note that this functional can also be given very different interpretations, depending on the meaning of n (for instance: the cost of a road between points when the local costs are given by n).

(a) Write down the Euler-Lagrange equation.

(b) Determine the optimal trajectory in case n does not depend on y explicitly. Then use the 'conservation'-property expressed by the E-L equations to study the trajectories.

(c) Determine the optimal trajectory in case n does not depend on x explicitly. Then use 'energy-conservation' to study trajectories. [[Alternatively: describe the trajectories as functions $x(y)$ and transform the functional.]]

(d) Consider the special cases $n = y$ and $n = \frac{1}{y}$ for which the trajectories can be expressed explicitly.

8. *Boussinesq type of equations*
Surface waves (in one horizontal direction x) that decay at infinity ($|x| \mapsto \pm\infty$) can be described in terms of the wave height $\eta(x,t)$ and a velocity $u(x,t)$ in the following form (a Hamiltonian system):

$$\partial_t u = -\partial_x \{\delta_\eta H(u,\eta)\}, \tag{1.26}$$
$$\partial_t \eta = -\partial_x \{\delta_u H(u,\eta)\}. \tag{1.27}$$

for a suitable functional (the Hamiltonian) $H(u,\eta)$.

(a) Describe the equations in full detail when the Hamiltonian is given by the following functional

$$H(u,\eta) = \int \{\frac{1}{2} g\eta^2 + \frac{1}{2}(u^2 - \frac{1}{3}u_x^2)\}\, dx.$$

(This set of equations are the 'linearized' equations.)

(b) In another case (shallow water, no dispersion, but nonlinear), the equations are of the form

$$\begin{aligned} \partial_t u &= -\partial_x \{ g\,\eta + \frac{1}{2} u^2 \}, \\ \partial_t \eta &= -\partial_x \{ u + \beta\,\eta\, u \}, \end{aligned}$$

where β is a constant. Determine the value of β such that this system of equations is a Hamiltonian system of the form (3, 4) given above.

(c) Show that the equations have the horizontal momentum as constant of the motion:

$$\int u(x)\eta(x)dx$$

9. The *KdV-eqn* in normalized form is given by

$$\partial_t u = -\partial_x \left[u + \partial_x^2 u + \frac{1}{2} u^2 \right].$$

(a) Show that it can be written as a generalized Hamiltonian system by determining the Hamiltonian H such that

$$\partial_t\, u(t) = \partial_x\, \delta H_{KdV}(u)$$

(b) Show that the following functionals are constants of the motion for KdV: $\int u(x)\,dx$, $\int u(x)^2\,dx$, $H_{KdV}(u)$.

10. The *BBM-eqn* in normalized form is given by

$$\partial_t\, u - \partial_t \partial_x^2 u = -\partial_x (u + \frac{1}{2} u^2).$$

(a) Show that it can be written as a Hamiltonian system by determining the Hamiltonian H and L that is the inverse of a suitable differential operator such that it is given by

$$\partial_t\, u(t) = L\, \partial_x\, \delta H_{BBM}(u)$$

(b) Show that the following functionals are constants of the motion for BBM: $\int u(x)\,dx$, $\int u(x)^2\,dx$, $H_{BBM}(u)$.

11. Show that *Burgers' eqn*

$$\partial_t u + u \partial_x u = \partial_x^2 u$$

can be written as a combination of a conservative and dissipative structure by determining functionals D and H such that it gets the form

$$\partial_t\, u = \partial_x\, \delta H(u) + \delta D(u)$$

12. *Periodic motions and boundary conditions*
We have already remarked that in general the dynamic variational principles are not well suited to prove existence; usually dynamic evolutions

are saddle points of the action functional. In particular cases existence can be proved with variational methods. The most successful results deal with period solutions, the reason being that then the problem can be formulated as a boundary value problem. We will show that in this exercise. Consider a Lagrangian dynamical system with Lagrangian L. L may depend on t, but if it does, it is in a periodic way, say with period T. Then one may look for motions that are periodic with period T.

(a) Show that the evolution $t \mapsto q(t)$ is T-periodic iff it is the periodic continuation of the function defined on $[0, T]$ that satisfies the *periodic boundary conditions*:

$$q(0) = q(T), \quad \dot{q}(0) = \dot{q}(T).$$

(b) Show that (under mild assumptions) these boundary conditions arise partly as natural boundary conditions from the action functional with prescribed boundary condition for q only: $q(0) = q(T)$.

(c) Formulate the periodic boundary conditions for a Hamiltonian system; show that they arise from the canonical action principle when only conditions on q are prescribed as above.

13. ** *Stationary states of a nonlinear diffusion equation*
Consider the stationary solutions of a nonlinear diffusion equation on a domain Ω for which the diffusion coefficient D may depend on u:

$$\begin{cases} \text{div } [D(u)\nabla u] + f(x,u) = 0 & in \Omega \\ u(x) = \varphi(x) & on \partial\Omega_1, \\ D(u)\frac{\partial u}{\partial n} = \psi(x) & on \partial\Omega_2. \end{cases}$$

(a) When D is constant derive the variational formulation of this boundary value problem (including the boundary conditions).

(b) Observe that when D depends on u there is no obvious variational formulation.

(c) Suppose that D is positive and a monotone function of u. Consider the transformation of the dependent variable $u \to v$ such that

$$\nabla v = D(u)\nabla u.$$

Show that v can be expressed directly in terms of u.

(d) Derive the governing boundary value problem for v.

(e) Show that the bvp for v has a variational structure; denote the governing functional by $\mathcal{L} = \mathcal{L}(v)$.

(f) Define uniquely the functional $\bar{\mathcal{L}}$ of u as

$$\mathcal{L}(v) \equiv \bar{\mathcal{L}}(u).$$

Find the critical points of $\bar{\mathcal{L}}(u)$. Verify the transformation between the two formulations of the boundary value problem from relations between $\delta_v \mathcal{L}(v)$ and $\delta_u \bar{\mathcal{L}}$.

(g) Conclusion?

14. ** *Variations of the boundary*
Consider (for simplicity, on the plane) a given density function ρ and the total "mass" in a region Ω:

$$M(\Omega) = \int_\Omega \rho(x,y)dxdy.$$

We want to see how M depends on Ω. (Assume that the regions are "convex-like" and can be deformed smoothly without introducing intersections.)

(a) First take the special case that Ω is the area between the x-axis and the graph of a function $\eta = \eta(x)$:

$$\Omega = \{\ (x,y)\ \mid\ a \le x \le b,\ 0 \le y \le \eta(x)\ \},$$

and consider the corresponding functional

$$\mathcal{L}(\eta) = M(\Omega).$$

Determine the first variation and show that the variational derivative of $\mathcal{L}$ for variations of the domain described by a variation of the function η is given by

$$\delta\mathcal{L}(\eta) = \rho\mid_{y=\eta(x)} \equiv \rho(x,\eta(x)).$$

(b) Now, more generally, describe a variation of the boundary $\partial\Omega$ by a "normal" displacement σ (defined on the boundary). Determine the first variation of M.
Can you find an expression for the variational derivative of M?
Verify the formula for the case of a radial deformation of a circular domain (and $\rho = 1$).

(c) Show that the more general result specializes to the case of changing the graph that determines the boundary. (Relate a variation η and the normal displacement σ in this case.)

15. ** *Jacobi functional* in Classical Mechanics
For a Lagrangian system for which the energy is conserved, one may look for solutions of prescribed total energy E.

(a) Consider the Jacobi functional on the set of functions $[0,1] \ni \tau \mapsto x(\tau) \in \mathcal{R}^n$ with $x(0) = x(1)$:

$$J(x) = \int_0^1 \sqrt{E - V(x)}|x_\tau|d\tau$$

Derive the equation for its critical points, and the boundary conditions.

(b) Show that for a suitable scaling of the parameter τ to t and a related transformation $x(\tau) \equiv q(t)$, a standard second order Newton equation for q and potential V results; show that the solution has indeed energy E. What about the boundary conditions?

(c) How can the Jacobi functional be obtained by constraining the action principle to motions that satisfy the energy constraint?

(d) Show that the following *modified Jacobi functional* can serve the same purposes:

$$[\int_0^1 \dot{x}^2 d\tau] \; . \; [E - \int_0^1 V(x) d\tau].$$

How can this functional be obtained form the action principle and an energy constraint?

1.6 ** Extensions

1.6.1 Theory of second variation

When for fixed v the function $\varepsilon \mapsto \mathcal{L}(u + \varepsilon v)$ is twice differentiable, its second derivative leads to the following notion.

Definition 20 *The* second variation *of a functional $\mathcal{L}$ at u in the direction v is denoted by $\delta^2\mathcal{L}(u; v)$ and is defined as*

$$\delta^2\mathcal{L}(u; v) = \frac{d^2}{d\varepsilon^2}\mathcal{L}(u + \varepsilon v)\Big|_{\varepsilon=0} . \tag{1.28}$$

Hence we have

$$\mathcal{L}(u + \varepsilon v) = \mathcal{L}(u) + \varepsilon\delta\mathcal{L}(u; v) + \frac{1}{2}\varepsilon^2\delta^2\mathcal{L}(u; v) + o(\varepsilon^2). \tag{1.29}$$

From this the following second order condition for an extremal element is obvious.

Proposition 21 *If $\hat{u}$ is a local extremal for $\mathcal{L}$, the second variation is sign-definite for all directions v in the tangent space.*
Specifically, if $\mathcal{L}$ has a (local) minimum at $\bar{u}$:

$$\mathcal{L}(\bar{u}) \leq \mathcal{L}(u) \text{ for all } u \in \mathcal{M} \quad \text{in a neighbourhood of } \bar{u}, \tag{1.30}$$

then

$$\delta^2\mathcal{L}(\bar{u}; v) \geq 0 \text{ for all } \; v \in T_{\bar{u}}\mathcal{M}. \tag{1.31}$$

In most cases, the second variation $\delta^2\mathcal{L}(u; v)$ is quadratic in v. When it is, it can also be obtained as a repeated differentiation of the first variation. In fact, a bilinear form can be defined as follows:

$$Q(u; v, w) := \frac{d}{d\rho}\frac{d}{d\varepsilon}\mathcal{L}(u + \varepsilon v + \rho w)\Big|_{\varepsilon=0,\rho=0} . \tag{1.32}$$

When the order of differentiation can be interchanged, this form is in fact *symmetric* in v and w:

$$Q(u; v, w) = Q(u; w, v) \tag{1.33}$$

and leads to the introduction of a symmetric mapping $Q(u)$ such that

$$Q(u; v, w) = \int_\Omega v \cdot Q(u)w \equiv \langle v, Q(u)w \rangle. \tag{1.34}$$

This mapping $Q(u)$ is the generalization of the *Hessian matrix* of functions on Euclidian space. It is referred to as the *second variation operator*. Its relation to the second variation is explicitly given by

$$\delta^2 \mathcal{L}(u; v) = \langle v, Q(u)v \rangle, \tag{1.35}$$

and in fact, this relation, together with the requirement that Q is symmetric, can serve to define the operator Q.

All these notions can also be translated in statements about the variational derivative $\delta\mathcal{L}$. This is made more precise in the next lemma which will be used frequently in the following.

Lemma 22 *For a functional $\mathcal{L}$ on $\mathcal{M}$, with $\delta\mathcal{L}$ its variational derivative, denote the formal Frechet derivative of $\delta\mathcal{L}$ by $D\delta\mathcal{L}$:*

$$D\delta\mathcal{L}(u)\xi := \frac{d}{d\varepsilon}\delta\mathcal{L}(u + \varepsilon\xi)\Big|_{\varepsilon=0}. \tag{1.36}$$

Then $D\delta\mathcal{L}(u) : T_u\mathcal{M} \to T_u^*\mathcal{M}$ is a symmetric *map, the* second variation operator*, satisfying*

$$\delta^2 \mathcal{L}(u; \xi) =< D\delta\mathcal{L}(u)\xi, \xi >$$

Proof. For arbitrary ξ and η from T_uM it holds

$$\langle D\delta\mathcal{L}(u)\xi, \eta \rangle = \frac{d}{d\varepsilon}\langle \delta\mathcal{L}(U + \varepsilon\xi), \eta \rangle\Big|_{\varepsilon=0}.$$

Using the definition of variational derivative, this can be rewritten like

$$\langle D\delta\mathcal{L}(u)\xi, \eta \rangle = \frac{d}{d\varepsilon}\frac{d}{d\rho}\mathcal{L}(u + \varepsilon\xi + \rho\eta)\Big|_{\rho=0,\varepsilon=0}.$$

Assuming smoothness of the function $(\varepsilon, \rho) \to \mathcal{L}(u + \varepsilon\xi + \rho\eta)$, the order of differentiation at the right hand side can be interchanged and one obtains the symmetry as stated. ■

1.6.2 Legendre transformation

1.6.3 Convexity Theory

1.6.4 Hamilton Jacobi equations

1.6.5 Exercises

1. *Nonlinear two-point boundary value problem*
   For given $f \in C^1([0, 1] \times \mathcal{R}, \mathcal{R})$ consider the non linear bvp

$$\begin{cases} -u_{xx} = f(x, u) & \text{on } (0, 1) \\ u(0) = u(1) = 0 \end{cases}$$

(a) Give the variational formulation, i.e. the functional $\mathcal{L}$ such that its critical points on $M_0 = \{u \mid u(0) = u(1) = 0\}$ correspond to the solutions of the bvp.

(b) Determine the second variation: $\eta \mapsto \delta^2\mathcal{L}(u;\eta) \equiv Q_u(\eta)$.

(c) Write down the Euler-Lagrange equation for $\eta \mapsto Q_u(\eta)$.

(d) Compare the result with the *linearization* of the bvp:

$$\begin{cases} -\eta_{xx} = f'(x,u)\eta & \text{on } (0,1) \\ \eta(0) = \eta(1) = 0 \end{cases}.$$

(e) Show that if the linearized bvp has a nontrivial solution $\hat{\eta}$, then $Q_u(\hat{\eta}) = 0$.

(f) Prove the general result:

Proposition 23 *The linearization of the Euler-Lagrange equation of a functional $\mathcal{L}$ around a solution u is the Euler-Lagrange equation of the second variation $\delta^2\mathcal{L}(u;\cdot)$.*

2. *Euler buckling*
Reconsider the transversal deflections of a bar, written with the arclength s, and angle $\theta = \theta(s)$. Assume that in the rest state the bar is along the x-axis, length ℓ, and fixed at $x = 0$. The other end point is free. Take for the bending energy (related to the curvature θ_s)

$$\int_0^\ell \frac{1}{2}\theta_s^2 ds.$$

If at the free end point a force μ is acting in the direction of the negative x-axis, then $\mu[\ell - x(\ell)]$ is the work executed by the force. For given force, the deflection is described by the principle of minimal potential energy, i.e. of the functional

$$\mathcal{L}(\theta) = \int_0^\ell \{\frac{1}{2}\theta_s^2 - \mu(1 - \cos\theta)\} ds.$$

(a) Determine the bvp for a critical point. What is the meaning of the (natural) boundary condition at $s = 0$ and $s = \ell$.

(b) Relate the equation to that for the pendulum equation

$$\ddot{x} = -\sin x;$$

which dynamic solutions correspond to the desired deflection of the bar? Use "energy-conservation" to write down the solution implicitly.

(c) To investigate the Euler-buckling problem directly, observe that with a solution $\theta(s)$, also $-\theta(s)$ is a solution, and hence $\theta \equiv 0$ is a solution for all μ.

(d) Determine the second variation around the trivial state, and show that only for specific values of $\mu = \mu_k$, $k \in \mathbb{N}$, the linearized equation has nontrivial solutions, and determine these solutions. Verify that all these solutions correspond to the same physical oscillation of the linearized pendulum equation.

(e) Conclude from the phase plane analysis of the pendulum equation that for the Euler buckling there is a *bifurcation value* μ_1 such that for $\mu < \mu_1$ there is no nontrivial buckled state, while for any $\mu > \mu_1$ there is precisely one buckled state that is positive.

Chapter 2

Constrained Problems

2.1 Motivation and Introductory Examples

Example. *Finite dimensional constrained extremal problems*
For a function of one variable, at a minimizer $\bar{x}$ it holds $f'(\bar{x}) = 0$.
Now consider a function of two variables, $F(x, y)$. At a minimizer $(\bar{x}, \bar{y})$ it holds

$$\nabla F(\bar{x}, \bar{y}) = 0, \text{ i.e. } \partial_x F(\bar{x}, \bar{y}) = \partial_y F(\bar{x}, \bar{y}) = 0,$$

$\partial_x F(\bar{x}, \bar{y}) = 0$ meaning that the function vanishes for variations in the x-direction, and $\partial_y F(\bar{x}, \bar{y}) = 0$ meaning that it vanishes for variations in the y-direction. From this we conclude that it vanishes for variations in any direction, which leads to $\nabla F(\bar{x}, \bar{y}) = 0$. Of course, these relations hold when we are allowed to take variations in these directions. In this Chapter we will consider constrained problems, i.e. problems which will lead to restrictions on the admissible variations. In a simple example: if we would not be allowed to vary in the y-direction, we could not conclude that in a minimizer it holds that $\partial_y F(\bar{x}, \bar{y}) = 0$. This would, for instance, be the case for the minimization problem:

$$\text{Min } \{ \ F(x, y) \mid (x, y) \in M \ \}, \quad \text{with } M = \{ \ (x, y) \mid y = 0 \ \},$$

i.e. if we constrain the domain of definition. Then in a minimizer $(x^*, 0)$ of this problem we can only conclude $\partial_x F(x^*, 0) = 0$. With $\boldsymbol{\tau}$ the tangent direction to M, and $\mathbf{n}$ the normal, we can interpret this as $\nabla F(x^*, 0) \cdot \boldsymbol{\tau} = 0$, but have no information about $\nabla F(x^*, 0) \cdot \mathbf{n}$: this last component is at yet undetermined (just like we cannot determine the y-derivative of the function $(x, y) \to x^2 + y$ when we restrict it to the line $y = 0$), and we could write

$$\nabla F(x^*, 0) = \lambda \mathbf{n}$$

for some number λ, which is usually called a *multiplier*. This interpretation becomes even more appealing if we consider the minimization of the function on a given curve in the (x, y)-plane, say

$$\text{Min } \{ \ F(x, y) \mid (x, y) \in M \ \}, \text{ with } M = \{ \ (x, y) \mid y = \phi(x) \ \}.$$

Then the constraint can easily be 'eliminated???', and we find that this is the same as minimization of $x \to F(x, \phi(x))$, which at a minimizer x^* leads to

$$\frac{d}{dx} F(x, \phi(x)) = \partial_x F + \partial_y F . \frac{d\phi}{dx} = \nabla F \cdot \boldsymbol{\tau} = 0$$

where now $\boldsymbol{\tau} = (1, \partial_x \phi)$ is tangent to the curve. And hence, again we conclude that at the minimizer (x^*, y^*) we can write

$$\nabla F(x^*, y^*) = \lambda \mathbf{n}$$

for some number λ.
The same conclusions will be obtained if the curve is given implicitly by some relation $\Phi(x, y) = 0$ or when we consider extrema of a function of three variables restricted to a plane or to a 2-dimensional manifold in that space.

■

Example. *Geometric LMR*
More generally, consider a function F of, say, n variables, restricted to some set $\mathcal{M}$, which at each point has $n - p$ admissible directions, which means that there will be p relations between the n components of each point. Considering the directional derivative at a minimizing point x^* in various directions η: $DF(x^*)\eta = \nabla F(x^*) \cdot \eta$, the minimizing property implies that this is zero if η is a direction that is admissible at x^*, but we have no information for the p directions that are 'perpendicular' to the set at that point and which are non-admissible variations. So, $\nabla F(x^*)$ is undetermined in p non-admissible directions, and we can write

$$\nabla F(x^*) = \lambda_1 \mathbf{n}_2 + \lambda_2 \mathbf{n}_2 + ... \lambda_p \mathbf{n}_p$$

for undetermined multipliers $\lambda_1, \lambda_2, ... \lambda_p$. This is called the *Lagrange Multiplier Rule.*
Of course, the fact that we know that the minimizing element x^* is in the set $\mathcal{M}$ will actually reduce the number of 'free' components of x^* to $n - p$ which should be found from the $n - p$ conditions from the vanishing derivative in admissible directions. Stated differently, when concerning the problem to find the minimizer, consider the p multipliers and the n components of x^* as unknowns; then these should be determined from the p relations between the components of x^* and the n equations of LMR, again the same number of equations as the number of unknowns.

■

The above examples show already the way how to generalize to infinite dimensions when we would be able to find the 'tangent space', or the set of 'normals' to it. In finite dimensions, recall that the gradient of a function at a point defines the direction in which this function increases most (it is the direction of steepest ascent), and is perpendicular to the level set of the function through that point. Hence, if the constraint is given as a level set of an explicit function, a normal direction is given by the gradient of the function at that point. More generally, if the constrained set consists of the intersection of levelsets of p functions, the p normals are found from the gradients of the p functions. All these directions will be perpendicular to the tangent space. If the p gradients are independent, the tangent space will be $n - p$ dimensional; this is usually the case, and such points are therefore called 'regular' points of the manifold. Stated differently; at a regular point of a manifold, the n-dimensional space decomposes into a number of p normal directions and remaining $n - p$ directions from the tangent space. A manifold with only regular points is then called a manifold of dimension $n - p$, or a manifold of co-dimension p.

Example. In finite dimensions $R^n, n > 1$ consider the following examples.

1. For $n = 2$, the levelsets of a function $K = K(x, y)$ are curves in the plane. A regular point $u = (x, y)$ is one for which $\nabla K(x, y) \neq 0$; the singular points are those for which $\nabla K(x, y) = 0$. At a regular point, the tangent 'space' is the one-dimensional straight line through the point in the tangent direction: the direction vector τ such that $\nabla K(x, y) \cdot \tau = 0$. Hence $\nabla K(x, y)$ is the normal to the tangent line, i.e. normal to the level line.
   For instance, for $K(x, y) = x^2 + y^2$, every point on the level set $K^{-1}(k)$ with $k > 0$ is a regular point; for $k = 0$, the point $(0, 0)$ (which is the only point on the level set) is singular.

2. In $\mathcal{R}^3$, consider the intersection of a sphere with a horizontal plane:

   $$\mathcal{M} = \{\ (x, y, z)\ \mid\ x^2 + y^2 + z^2 = R^2,\ z = \zeta\ \}.$$

   When $|\zeta| < R$, each point is a regular point and the tangent space is one-dimensional; when $\zeta = R$, the point $(0, 0, R)$ is singular, and the tangent space is two-dimensional.

3. *Analytic LMR for levelset-constraints*
   Consistent with the above description is the result for a critical point x^* of

   $$\begin{aligned} & \text{Crit}\ \ \{\ F(x)\ \mid\ x \in \mathcal{M}\ \}, \\ \text{with } \mathcal{M} \quad = \quad & \{\ x \in R^n\ \mid\ K_1(x) = k_1, \ldots, K_p(x) = k_p\ \}. \end{aligned}$$

   Supposing it is a regular point of the manifold, i.e. $\nabla K_1(x^*), ..., \nabla K_p(x^*)$ are independent, then it holds that

   $$\nabla F(x^*) = \lambda_1 \nabla K_1(x^*) + ... + \lambda_p \nabla K_p(x^*)$$

■

All of this will now be generalized to infinite dimensions for the simplest case that the set of admissible elements is the intersection of a finite number of level sets of specific functionals (hence p will be finite, while $n - p$ is infinite in infinite dimensional function spaces): an infinite dimensional manifold of finite co-dimension p.

2.2 Lagrange Multiplier Rule

2.2.1 Constrained to levelsets

When talking about the set $\mathcal{M}$ and the independent admissible variations at a point, we are actually dealing with the tangent space at the point to the 'set', which is usually envisaged as a 'manifold', and which was mentioned already briefly in Chapter 1. The elements v from the tangent space $T_u\mathcal{M}$ are the admissible variations: they are such that with u, the element $u + \varepsilon v$ belongs to $\mathcal{M}$ up to terms of higher order (usually $\mathcal{O}(\varepsilon^2)$) for small real ε; in detail:

$$T_u\mathcal{M} = \left\{ v \ \middle| \text{there is } w(v;\varepsilon) \text{ such that } \left\{ \begin{array}{l} u + \varepsilon v + w \in \mathcal{M} \\ w = o(\varepsilon)\ \ \text{i.e. } \frac{w}{\varepsilon} \to 0 \text{ for } \varepsilon \to 0 \end{array} \right. \right\}$$

In infinite dimensions, for *constrained variational problems* the functions belonging to the set of admissible elements $\mathcal{M}$ have to satisfy certain boundary conditions and certain 'interior' conditions. Then, if u belongs to $\mathcal{M}$, for a variation $\eta \in C_0^\infty(\Omega)$, the function $u + \varepsilon\eta$ does *not* in general belong to $\mathcal{M}$ (up to second order): the set $\mathcal{M}$ is a nonlinear manifold. This has been denoted in Chapter 1 that the tangent space does not contain all test functions:

$$C_0^\infty(\Omega) \not\subset T_u\mathcal{M}.$$

In most problems that we will deal with, the manifold $\mathcal{M}$ is a subset of a function space $\mathcal{U}$ for which the functions satisfy (apart from certain boundary conditions) a finite number of nonlinear *functional constraints*.

First, to deal separately with (linear) inhomogeneous boundary conditions in a simple way in the following we first define a subspace to which the admissible variations will belong. Therefore, denote by $\mathcal{U}$ be the space of functions that satisfy the (linear) boundary conditions, and let $\mathcal{U}_0$ be the tangent space to $\mathcal{U}$, i.e. $\mathcal{U}_0$ consists of elements v such that $u + \varepsilon v \in \mathcal{U}$ whenever $u \in \mathcal{U}$: the functions v are the functions that satisfy the homogeneous boundary conditions. All the admissible elements related to the set $\mathcal{M}$ will certainly belong to this set $\mathcal{U}_0$.

Secondly, we will restrict to the most important sets that will be encountered in the following. These will be sets $\mathcal{M}$ of admissible elements which will be defined as the intersection of the levelsets of certain (density) functionals $\mathcal{K}_1, \ldots, \mathcal{K}_p$:

$$\mathcal{M} \;=\; \{\; u \in \mathcal{U} \;\mid\; \mathcal{K}_1(u) = k_1, \ldots, \mathcal{K}_p(u) = k_p \;\}, \tag{2.1}$$

where $k_1, \ldots, k_p$ are given values. In general this set may be empty, so we assume that for the given values of the constraints $k_1, \ldots, k_p$ this set is non-empty.

Now consider first a single levelset, say for the functional $\mathcal{K}$ the set $\mathcal{K}(u) = k$. Assume that we take a smooth curve in this levelset, through a point $\bar{u}$, say a curve parameterized by ε :

$$\varepsilon \to u(\varepsilon), \quad \text{with } u(0) = \bar{u}, \text{ and } \mathcal{K}(u(\varepsilon)) = k$$

Then, upon differentiating with respect to ε we arrive at:

$$0 = \frac{d}{d\varepsilon}\mathcal{K}(u(\varepsilon)) = \delta\mathcal{K}(u(\varepsilon); \partial_\varepsilon u)$$

When we write $\tau = \partial_\varepsilon u(0)$ for the tangent direction to the curve at $\bar{u}$, we see that at $\varepsilon = 0$,

$$0 = \delta\mathcal{K}(\bar{u}; \tau) =< \delta\mathcal{K}(\bar{u}), \tau > .$$

Thus for any tangent direction τ to the levelset at the point $\bar{u}$ it holds that $< \delta\mathcal{K}(\bar{u}), \tau >= 0$. Therefore it is consistent[1] to consider $n := \delta\mathcal{K}(\bar{u})$ as the

[1] Actually we also have to prove the converse: if $< \delta\mathcal{K}(\bar{u}), \tau >= 0$ then τ is a tangent direction of some curve in the levelset. This can be proved by constructing the curve along that direction; it is clear that in general the actual curve will have a deviation in the normal direction. Taking this as Ansatz, we have that the curve $\varepsilon \to \bar{u} + \varepsilon\tau + \alpha(\varepsilon)n$ belongs to the levelset if $\alpha(\varepsilon)$ can be choosen such that $\mathcal{K}(\bar{u} + \varepsilon\tau + \alpha(\varepsilon)n) = k$ and such that $\alpha(\varepsilon)$ is of higher order than linear in ε. Viewing $\mathcal{K}(\bar{u} + \varepsilon\tau + \alpha n) = k$ as a relation between two real variables that defines α implicitly as a function of ε, the use of the *implicit function theorem* shows the existence of $\alpha(\varepsilon)$ and the required higher order dependence on ε, provided that $\delta\mathcal{K}(\bar{u}) \neq 0$.

normal to the tangent space, provided that $\bar{u}$ is a regular point, i.e. provided that $\delta\mathcal{K}(\bar{u}) \neq 0$. This shows that the finite dimensional picture of tangent space and normal to levelset remains valid in infinite dimensions.

For a set that is the intersection of various levelsets

$$\mathcal{M} \;=\; \{\; u \in \mathcal{U} \;\mid\; \mathcal{K}_1(u) = k_1, \ldots, \mathcal{K}_p(u) = k_p \;\}, \tag{2.2}$$

in the regular points, the p constraints will define a tangent space that contains all but p directions, so that near a regular point the set $\mathcal{M}$ is well approximated by its tangent space (hyper plane) of co-dimension p; this is the analog of a $(n-p)$-dimensional smooth manifold in $\mathcal{R}^n$. Stated differently, at a regular point, there are p independent normal directions to the tangent space.
In a singular point, some of the normal directions to the tangent space coincide: the elements of the tangent space are restricted by less than p conditions. We will make this more precise in the following.

Let $u \in \mathcal{M}$ be a point of the manifold $\mathcal{M}$ at which the linear functionals $\delta\mathcal{K}_1(u;\cdot), \ldots, \delta\mathcal{K}_p(u;\cdot)$ are linearly independent; we will call this a *regular point*. The following result generalizes what has been proved above for one functional constraint.

Lemma 24 *Lyusternik*
The tangent space *to $\mathcal{M}$ at a regular point $u \in \mathcal{M}$ is the set of co-dimension p :*

$$T_u\mathcal{M} := \{\; v \in \mathcal{U}_0 \;\mid\; \delta\mathcal{K}_1(u;v) = \ldots = \delta\mathcal{K}_p(u;v) = 0 \;\}. \tag{2.3}$$

This shows that admissible variations v have to satisfy p linear constraints.

2.2.2 Formulations of LMR

Consider the variational problem

$$\text{Crit}\;\; \{\; \mathcal{L}(u) \mid u \in \mathcal{M} \;\},\; \text{with}\; \mathcal{M} \;=\; \{\; u \in \mathcal{U} \;\mid\; \mathcal{K}_1(u) = k_1, \ldots, \mathcal{K}_p(u) = k_p \;\}$$

Recall the general stationarity condition (1.15) for a critical point $\hat{u}$ of $\mathcal{L}$ on $\mathcal{M}$:

$$\delta\mathcal{L}(\hat{u};v) = 0,\;\; \text{for all } v \in T_{\hat{u}}\mathcal{M}.$$

Using Lyusternik's Lemma for the specific set $\mathcal{M}$ under consideration, this condition for a critical point can be reformulated to

$$\begin{aligned} \delta\mathcal{L}(\hat{u};v) = 0, \quad &\text{for all} \quad && v \in \mathcal{U}_0 \\ &\text{for which} \quad && \delta\mathcal{K}_k(\hat{u};v) = 0, 1 \leq k \leq p. \end{aligned} \tag{2.4}$$

Clearly, (2.4) is satisfied if $\delta\mathcal{L}(\hat{u};\cdot)$ is a linear combination of the $\delta\mathcal{K}_k(\hat{u};\cdot)$, $1 \leq k \leq p$. In fact this is also necessary, as expressed in the next proposition.

Proposition 25 Lagrange's multiplier rule
A regular point $\hat{u} \in \mathcal{M}$ is a constrained critical point of $\mathcal{L}$ on $\mathcal{M}$, i.e. satisfies (2.4), if and only if there are real numbers, called Lagrange multipliers, *$\lambda_1, \ldots, \lambda_p$ such that*

$$\delta\mathcal{L}(\hat{u};v) = \Sigma_k\; \lambda_k \delta\mathcal{K}_k(\hat{u};v), \quad \textit{for all}\;\; v \in \mathcal{U}_0. \tag{2.5}$$

It is possible to formulate this result in a different way; this may be easier to remember, but may also be somewhat misleading.

Proposition 26 LMR with the Lagrangian functional
A regular point $\hat{u} \in \mathcal{M}$ is a critical point of $\mathcal{L}$ on the constrained set $\mathcal{M}$ (2.1) iff for some multipliers $\lambda_1, \ldots, \lambda_p$ the element $\hat{u}$ is an unconstrained *critical point of the* unconstrained *functional*

$$\mathcal{U} \ni u \mapsto \mathcal{L}(u) - \Sigma_m \ \lambda_m \mathcal{K}_m(u). \tag{2.6}$$

This functional is called the Lagrangian functional[2] *of the constrained problem.*

Proof. For a critical point of the Lagrangian functional it holds that

$$\delta[\mathcal{L} - \sum_m \lambda_m \mathcal{K}_m](u; v) = 0, \text{ for all } \ v \in \mathcal{U}_0.$$

This is precisely the equation from the multiplier rule. The other way around is obvious.

■

Remark. The above formulation with the Lagrangian may be misleading in the following respect: it may be possible that $\hat{u}$ is a constrained minimizer, while it is not a minimizer, but only a saddle point of the unconstrained Lagrangian; we will consider this in more detail later. On the other hand, this procedure does lead to the correct set of equations, including possible natural boundary conditions.

■

The results above have obvious consequences for the relation between the variational derivatives since $C_0^\infty \subset \mathcal{U}_0$. When no natural boundary conditions appear, these relations are in fact equivalent to the original result. The investigation of natural boundary conditions, in which the multipliers may appear, should be based on a study of (2.5).

Independence of the linear functionals leads to a result for the variational derivatives (however, not in a one-to-one way). Since $\delta\mathcal{K}(u;\eta) =< \delta\mathcal{K}(u), \eta >$, for $\eta \in C_0^\infty \subset \mathcal{U}_0$, the independence of the linear functionals $\delta\mathcal{K}_1(u;\cdot), \ldots, \delta\mathcal{K}_p(u;\cdot)$ on $\mathcal{U}_0$ implies the independence of the p variational derivatives (as elements of $L_2(\Omega)$)

$$\delta\mathcal{K}_1(u), \ldots, \delta\mathcal{K}_p(u).$$

For admissible variations that are also testfunctions $\eta \in C_0^\infty \subset \mathcal{U}_0$, we get in particular

$$T_u\mathcal{M} \supset \{\ \eta \in C_0^\infty \subset \mathcal{U}_0 \ \mid \ \langle \delta\mathcal{K}_1(u), \eta \rangle = \ldots =< \delta\mathcal{K}_p(u), \eta >= 0\ \}.$$

This makes it clear that the testfunctions from the tangent space satisfy p orthogonality conditions, namely orthogonal to the p normal directions $\delta\mathcal{K}_1, \ldots, \delta\mathcal{K}_p$.

[2] Note that the name "Lagrangian" (functional) appears at various places with a different meaning!

Proposition 27 LMR for variational derivatives
For a constrained critical point $\hat{u} \in \mathcal{M}$ it holds that the variational derivatives are dependent:

$$\delta\mathcal{L}(\hat{u}) = \sum_m \lambda_m \delta\mathcal{K}_m(\hat{u}).$$

Equivalently: the variational derivative of the Lagrangian functional vanishes. Possible natural boundary conditions are overlooked in this formulation.

This restriction to test functions may be incomplete if boundary conditions are involved. The following examples motivate that sometimes one has to work with the first variations of the functionals, instead of with the variational derivatives only.

Exercise. Consider the set

$$\mathcal{M} = \{\ u : [0,1] \to \mathcal{R}\ \mid\ \mathcal{K}(u) = 1, u(0) = 1\ \}.$$

1. For $\mathcal{K}(u) = \int \frac{1}{2}u^2$ the tangent space is

   $$T_u\mathcal{M} = \{\ v; [0,1] \to \mathcal{R}\ \mid\ \int uv = 0, v(0) = 0\ \},$$

   i.e. the functions with $v(0) = 0$ perpendicular to the variational derivative $\delta\mathcal{K}(u)$ $(= u)$, just as is the case for test functions.

2. For $\mathcal{K}(u) = \int \frac{1}{2}u_x^2$ the tangent space is

   $$T_u\mathcal{M} = \{\ v; [0,1] \to \mathcal{R}\ \mid\ \int u_x v_x = 0, v(0) = 0\ \};$$

   relating this to the variational derivative $\delta\mathcal{K}(u) = -u_{xx}$ we find that v has to satisfy

   $$\langle \delta\mathcal{K}(u), v\rangle + u_x(1)v(1) = 0, \text{ and } v(0) = 0.$$

   So, for test functions η this means $\langle \delta\mathcal{K}(u), \eta\rangle = 0$, but there are more functions in the tangent space when $v(1) \neq 0$. These last functions should certainly be considered in the stationarity condition (the multiplier rule) in order to find the correct natural boundary conditions.

3. In more dimensions, an example of a nonlinear boundary condition is

   $$\mathcal{M} = \{\ u : \Omega \to \mathcal{R}\ \mid\ u(x) = \varphi(x) \text{ on } \partial\Omega_1,\ \int_{\partial\Omega_2} u^2(x) = 1\ \}.$$

   The tangent space is given by

   $$T_u\mathcal{M} = \{\ v : \Omega \to \mathcal{R}\ \mid\ v(x) = 0 \text{ on } \partial\Omega_1,\ \int_{\partial\Omega_2} u(x)v(x) = 0\ \}$$

   and clearly contains all test functions.

■

Proof. Proof of the multiplier rule ■

2.2.3 Families of constrained problems

Although the following can be formulated in a much more general way, we restrict to illustrate the general ideas to a simple example: the constrained optimization of one functional H on level sets of another functional I:

$$\text{Crit}\{\ H(u) \mid I(u) = \gamma\ \};$$

we will use the formulation with variational derivatives in the following, not bothering about possible (natural) boundary conditions.

In the previous sections we studied the problem with fixed value of the constraint, i.e. given γ. Now we will treat γ as a parameter, and consider the family of constrained optimization problems. This will enable us to give an interpretation of the multiplier and relate the nature of the constrained critical point to its character as a critical point of the Lagrangian functional.

Suppose therefore that we can find a smooth family

$$\gamma \mapsto U(\gamma) \in \text{Crit}\ \{H(u) \mid I(u) = \gamma\}$$

of constrained critical points of H on level sets $I^{-1}(\gamma)$ for parameter values γ in a neighbourhood of some γ_0. It may happen that, for instance, this family consists only of constrained minimizers, but also that the character of the critical point changes with γ (without violating the smoothness assumption).

A first observation is that the derivative along this family is "normal" to the level sets of I in the following sense. Defining

$$n(U) := \frac{dU}{d\gamma}, \tag{2.7}$$

it is found by differentiating the relation $I(U(\gamma)) = \gamma$ with respect to γ, that $n(U)$ is normal to the level set $I^{-1}(\gamma)$ in the sense that

$$\langle \delta I(U), n(U) \rangle = 1. \tag{2.8}$$

Along this branch, each element satisfies for a multiplier $\lambda = \lambda(\gamma)$ the equation

$$\delta H(U(\gamma)) = \lambda(\gamma)\delta I(U(\gamma)). \tag{2.9}$$

The multiplier can be related to the so-called value function.

Definition 28 *The* value function *of the constrained critical point problem on a branch of critical points is defined as:*

$$h(\gamma) := H(U(\gamma)) = \mathit{Crit}\ \{\ H(u) \mid I(u) = \gamma\ \}. \tag{2.10}$$

The value function can be depicted in a so-called *integral diagram*. With the value of the integrals H and I along the axis, each point represents all states with that value of H and I (a two-dimensional representation of the state space). Assuming that there are branches of constrained critical points parameterized with the value of the constraint functional γ, a schematic representation of these equilibria in the integral diagram may look like Fig. ?????????.

Both the first and second derivative of the value function play a particular role in the understanding of the critical point problem; this will be considered in the next two subsections.

2.2.4 The multiplier as derivative of the value function

Proposition 29 *For the smooth family $\gamma \mapsto U(\gamma)$, the multiplier $\lambda(\gamma)$ appearing in (2.9) is related to the value function according to*

$$\lambda(\gamma) = \frac{dh(\gamma)}{d\gamma}. \tag{2.11}$$

Proof. A direct differentiation with respect to γ leads to the result:

$$\frac{dh(\gamma)}{d\gamma} = \langle \delta H(U(\gamma)), \frac{dU(\gamma)}{d\gamma} \rangle = \lambda \langle \delta I(U(\gamma)), n(U) \rangle = \lambda,$$

the last equality from differentiation of $I(U(\gamma)) = \gamma$. ■

This result clearly shows that only by viewing a single problem for γ_0 as being embedded in a family provides an interpretation of the Lagrange multiplier $\lambda(\gamma_0)$.

2.2.5 Homogeneous functionals

A functional F is called homogeneous of degree α if for each μ

$$F(\mu u) = |\mu|^\alpha F(u).$$

A quadratic functional is homogeneous of degree 2, for instance.

For a homogeneous functional of degree α it holds for the directional derivative in the 'radial' direction that

$$\delta F(u; u) = \alpha F(u)$$

Proposition 30 *If the functional H is homogeneous of degree α, and I is homogeneous of degree β, then the value function is given up to a multiplicative constant by*

$$h(\gamma) = h(0)\gamma^{\alpha/\beta}$$

and the multiplier by

$$\lambda(\gamma) = \lambda(0)\gamma^{\alpha/\beta - 1},$$

where $\lambda(0) = h(0)\alpha/\beta$. By defining the following quotient R that is homogeneous of degree zero

$$R(u) := \frac{H(u)^\beta}{I(u)^\alpha}$$

the critical points of R are up to normalization, in a one to one relation to the constrained critical points of H on levelsets of I.
For quadratic forms, when $\alpha = \beta = 2$, the quotient is known as the Rayleigh quotient*; then the multiplier is independent of the value of the constraint and is given by the value of R at the critical point.*

2.2.6 ** Constrained minimizers and the Lagrangian functional

The character of the constrained critical point at γ_0 can be related to its character as a critical point of $u \mapsto H(u) - \lambda(\gamma_0)I(u)$. This is determined by the second derivative of the value function.

Proposition 31 *The second variation of the constrained problem in the normal direction $n(U)$ defined in (2.7), is precisely the second derivative of the value function:*

$$\langle [D^2H(U) - \lambda D^2I(U)]n(U), n(U)\rangle = \frac{d^2h(\gamma)}{d\gamma^2}. \tag{2.12}$$

In particular it follows that the sign *of the second variation in this direction is determined by the* convexity *or* concavity *of the value function in a neighbourhood of the value of γ.*

Proof. Differentiating the equation (2.9) with respect to γ there results

$$[D^2H(U) - \lambda D^2I(U)]n(U) = \frac{d\lambda}{d\gamma}\delta I(U). \tag{2.13}$$

Since $d\lambda/d\gamma = d^2h(\gamma)/d\gamma^2$, (2.12) results after taking the inner product with $n(U)$. ■

In the foregoing Propositions, U can also be viewed as a critical point of the Lagrangian functional $H - \lambda I$ on the whole space. The expression (2.12) gives information how this functional changes in the direction transversal to level sets of I. This may determine the character of U as an unconstrained critical point of $H - \lambda I$.

Proposition 32 *A constrained minimizer of H on level set of I is an unconstrained (local) minimizer of $H - \lambda I$ (where λ is the multiplier) if the value function is (locally) convex. If the value function is (locally) concave, a constrained minimizer is an unconstrained saddle point of $H - \lambda I$.*

Proof. It would be tempting to use the second variation to provide the proof. Indeed, being constrained minimal, the second variation is non-negative on the tangent space, and the previous result shows the sign in the remaining direction transversal (normal) to the tangent space. This is correct if certain technical conditions are met, viz. non-degeneracy (signs strictly positive or negative), and a property that sign-definiteness of the second variation is sufficient for minimality. However, the following reasoning is elegant and simple, and does not need any of such requirements.
Let U_0 be a constrained minimizer of H on the level set $I = \gamma_0$, with λ_0 the multiplier. Assuming that the value function is (locally) convex (i.e. $d^2h(\gamma)/d\gamma^2(\gamma_0) > 0$), the result that U_0 is in fact an unconstrained (local) minimizer of the functional $H - \lambda_0 I$ follows by comparing its value with arbitrary functions u, with $I(u) = \gamma$, γ in a neighbourhood of γ_0:

$$\begin{aligned} & H(u) - \lambda_0 I(u) \\ & \quad \geq \quad h(\gamma) - \lambda_0\gamma \text{ by definition of } h \\ & \quad \geq \quad h(\gamma_0) - \lambda_0\gamma_0 \text{ from convexity of } h \\ & \quad = \quad H(U_0) - \lambda_0 I(U_0) \text{ since } U_0 \text{ is a constrained minimizer.} \end{aligned}$$

If h is locally concave, $H - \lambda_0 I$ is maximal on the family $\gamma \mapsto U(\gamma)$, and the saddle point character is clear. See Fig. ??????. ■

2.3 Applications

We will now show some main application areas of Constrained Problems.

First we investigate in some detail Linear Eigenvalue Problems; these will be formulated as constrained problems with quadratic functionals, and the eigenvalues are the multipliers.

Further we will consider two applications for dynamical systems; one is a deep result in Classical Mechanics, that can nicely be extended to evolution equations: the theory for Relative Equilibria. The other one is the generalization of the steepest decent method subject to constraints, so-called thermodynamic systems.

2.3.1 Linear Eigenvalue Problems

Basic problem from Linear Algebra

The basic problem in Linear Algebra is to solve the vector equation for a given $n \times m$-matrix A and vector $b \in R^m$

$$x \in R^n \text{ such that } Ax = b$$

For square matrices, $m = n$, the eigenvalue problem is the search for non-trivial eigenvectors φ for which there exists an eigenvalue λ such that

$$A\varphi = \lambda\varphi.$$

The simplest examples show that even for real matrices the eigenvalues and eigenvectors can be complex-valued. Furthermore, a complete set of eigenvectors (the eigenvectors spanning the whole space) cannot be expected to exist. Very different is the situation when the matrix is symmetric:

Definition 33 *A square matrix $S : R^n \to R^n$ is called symmetric (Hermitian) if*

$$Sx \cdot y = x \cdot Sy \text{ for all } x, y$$

Proposition 34 *A symmetric matrix S has the following properties :*

- *all eigenvalues are real; as a consequence, the eigenvectors can be taken to be real,and mutually orthogonal;*
- *there exists a complete set of (real) eigenvectors $\varphi_1, ..., \varphi_n$:*

  $$S\varphi_k = \lambda_k\varphi_k, \quad k = 1...n;$$

  the eigenvalues are expressed in terms of the eigenvectors by the Rayleigh quotient

  $$\lambda_k = R(\varphi_k) \equiv \frac{S\varphi_k \cdot \varphi_k}{\varphi_k \cdot \varphi_k}$$

- *with respect to the set of eigenvectors, the matrix S has the form of a diagonal matrix (it is 'similar' to the diagonal matrix)*

$$S \sim diag[\lambda_1, ..., \lambda_n];$$

- *the null-space of S is the span of all eigenvectors corresponding to eigenvalue zero.*

Using the eigenvectors, the basic vector equation can easily be solved (or investigated) by taking innerproducts with respect to the eigenvectors (normalized for simplicity)

$$Sx = b \Longleftrightarrow b \cdot \varphi_k = Sx \cdot \varphi_k = x \cdot S\varphi_k = \lambda_k x \cdot \varphi_k \text{ for } k = 1...n$$

and hence

$$x = \Sigma_k \frac{b \cdot \varphi_k}{\lambda_k} \varphi_k$$

provided $\varphi_k \neq 0$ for all $k = 1...n$. If S is degenerate, there is only a solution provided the rhs satisfies the orthogonality ('solvability') conditions

$$b \cdot \varphi_j = 0 \text{ for all eigenvectors } \varphi_j \text{ with eigenvalue } = 0;$$

i.e. b should be orthogonal to the null-space of S. In that case the solution is not unique: any vector form the null-space can be added. This is a special case (restricted to symmetric matrices) of the so-called *Fredholm alternative.*

Exercise. Consider a matrix $S = diag[\lambda_1, \lambda_2]$. Sketch the image of the unit-circle under the action of S in the following cases

1. $\lambda_1 = \lambda_2$, either positive or negative;
2. $0 < \lambda_1 < \lambda_2 < 1$,
3. $0 < \lambda_1 < 1 < \lambda_2$.

■

Levelsets of Quadratic Forms

Basically, the characteristic properties of quadratic forms are expressed in the simple examples in $R^2 : a_p(x, y) = x^2 + y^2$ and $a_s(x, y) = x^2 - y^2$ and more generally in $(x, y) \in R^m \times R^{n-m} = R^n$

$$a(x, y) = |x|^2 - |y|^2.$$

Levelsets are either bounded, in which case these have 'elliptic' shape, or unbounded and then these have 'hyperbolic' shape. In the first case the origin is a center, and a minimizer (or maximizer) of the form, in the other case it is a hyperbolic point and a saddle point of the form. This is all that can be expected in finite dimensions:

Proposition 35 *Any quadratic form in R^n can be written like*

$$a(x) = Sx \cdot x$$

where S is a symmetric matrix. With respect to a basis of eigenvectors of S, the matrix has diagonal form

$$S = diag\,[\lambda_1, ..., \lambda_n]$$

and then

$$a(x) = \Sigma_k \lambda_k x_k^2.$$

The form is positive [non-negative] iff all eigenvalues are positive [non-negative]. The number of zero-eigenvalues determines the degeneracy *of the form; the number of negative eigenvalues is the* order of hyperbolicity.

Eigenvalue problem for linear operators

Definition 36 *Let L be a linear mapping (operator) from one function space into another one. The* eigenvalue problem *for L asks for the eigenvalues λ that are the complex numbers for which there exists a non-trivial* eigenfunction *u:*

$$Lu = \lambda u.$$

Clearly the eigenfunctions must be elements that are both in the domain of definition and in the range of the operator.

To define the notion of a symmetric operator, we exploit (as in the previous chapters) the L_2-inner product $< \cdot, \cdot >$ for functions on the domain Ω.

Definition 37 *Let L be an operator defined on a function space $\mathcal{U}$ that contains the test functions. The* formal adjoint *of L is the (linear) operator denoted by L^* such that*

$$\langle Lu, v\rangle =< u, L^*v > \text{ for all } u, v \in C_0^\infty(\Omega).$$

The operator L is called (formally) symmetric *on $\mathcal{U}$ if $L = L^*$ and moreover*

$$\langle Lu, v\rangle = \langle u, Lv\rangle \text{ for all } u \in \mathcal{U}.$$

Associated to L there is a bilinear functional

$$b(u, v) :=< Lu, v > .$$

If L is (formally) symmetric on $\mathcal{U}$, this bilinear functional is symmetric and defines the quadratic form $\mathcal{Q}$ on $\mathcal{U}$:

$$\mathcal{Q}(u) =< Lu, u > .$$

Note that in that case the operator L is obtained as the variational derivative of $\mathcal{Q}$:

$$u \mapsto \mathcal{Q}(u): \quad \delta_u \mathcal{Q}(u) = 2Lu$$

since the first variation is given by $\delta\mathcal{Q}(u;v) = 2b(u,v)$ for all $u, v \in \mathcal{U}$.

Remark. Above we defined the adjoint of an operator. When (homogeneous) boundary conditions are present, one can define the adjoint of the operator and the adjoint boundary conditions from

$$\langle Lu, v\rangle = \langle u, L^*v\rangle \quad for \quad all \quad u \text{ and } v.$$

We will see examples in the following. ■

In the following we will mainly deal with differential operators.

Exercise. Differential operators

For a linear differential operator L, the result applied to a (smooth) function u is a function $Lu(x)$ that depends on u and a finite number of derivatives of u at the point x. The order of the highest derivative is called the order of the differential operator.

1. For U functions on the interval $[0,1]$, and for given functions a, b, c determine the (formal) adjoint of the second order differential operator

$$Lu = a(x)u_{xx} + b(x)u_x + c(x)u.$$

2. Determine conditions on the functions a, b, c that guarantee that L is symmetric.

3. Consider for given function f the following boundary value problem for the operator L above:

$$Lu = f, \quad u(0) = 0, \ u_x(1) = 0.$$

Determine the corresponding adjoint boundary value problem.

4. The generalization to functions of more variables: for functions on the domain $\Omega \subset R^n$, and for given scalar functions $a, b_1, \ldots, b_n, c$, determine the (formal) adjoint of the operator

$$Lu = a(x)\Delta u + \sum_k b_k(x)u_{x_k} + c(x)u.$$

Determine the adjoint boundary value problem for this operator with Dirichlet boundary conditions on part of the boundary $\partial\Omega_1$.

■

General formulation of EVP

There is a one-to-one correspondence between symmetric operators and quadratic forms in finite dimensions as we have seen already; the same is (more or less) true in infinite dimensions: with a symmetric operator L corresponds a quadratic form $\mathcal{Q}(u) =< Lu, u >$ and the variational derivative is $\delta Q(u) = 2Lu$.

A specific example are *Sturm-Liouville operators* in one or more dimensions

$$\begin{aligned} Lu &= -\partial_x[p(x)\partial_x u] + q(x)u, \\ Lu &= -\text{div } [p(x)\ \nabla u] + q(x)u, \end{aligned}$$

with corresponding quadratic forms

$$\begin{aligned}\mathcal{Q}(u) &= \int [p(x)u_x^2 + q(x)u^2]dx,\\ \mathcal{Q}(u) &= \int_\Omega [p(x)|\nabla u|^2 + q(x)u^2]dx.\end{aligned}$$

We will now formulate the eigenvalue problem (EVP) in a somewhat more general way by formulating it using two symmetric quadratic forms instead of with the operators itself.

Let $\mathcal{N}$ be a quadratic form on L_2; we will denote the corresponding operator by N and use the following notation for the corresponding bilinear functional:

$$\mathcal{N}(u) = \langle Nu, u\rangle, \quad \mathcal{N}(u,v) \equiv \langle Nu, v\rangle$$

In the following we want to have $\mathcal{N}$ as a norm, and so we have to assume that $\mathcal{N}$ is positive:

$$\mathcal{N}(u) > 0 \quad \textit{for} \;\; u \neq 0.$$

Note, for $N = Identity$ then $\mathcal{N}(u)$ is just the usual L_2-innerproduct. The more general formulation includes the case when we use so-called weighted L_2-norms. Let $\mathcal{Q}$ be another quadratic form on $\mathcal{U}$, with corresponding symmetric operator L:

$$\mathcal{Q}(u) = \langle Lu, u\rangle, \quad \mathcal{Q}(u,v) \equiv \langle Lu, v\rangle.$$

We will study the eigenvalue problem corresponding to these two operators:

$$Lu = \lambda Nu$$

Note, for $N = Identity$ we recover the standard formulation above.

The eigenvalues corresponding to one eigenvalue λ form a linear space, the *eigenspace* of the eigenvalue, to be denoted by E_λ. These eigenspaces are mutually orthogonal in the following senses.

Proposition 38 *All eigenvalues are real valued, and the eigenfunctions can be assumed to be real.*
Eigenfunctions corresponding to different eigenvalues are "orthogonal" with respect to both quadratic forms:

$$\text{for } \varphi \in E_\lambda,\ \psi \in E_\mu,\ \text{with } \lambda \neq \mu \quad \begin{cases} \mathcal{Q}(\varphi,\psi) &= 0\\ \mathcal{N}(\varphi,\psi) &= 0. \end{cases}$$

We can denote this for the eigenspaces as[3]

$$E_\lambda \perp_{\mathcal{N}} E_\mu, \quad E_\lambda \perp_{\mathcal{Q}} E_\mu \text{ when } \lambda \neq \mu.$$

[3] Be carefull with this (useful) description: since $\mathcal{N}$ is a norm, orthogonality can be understand in the ususal sense; however $\mathcal{Q}$ is not necessarily positive; when it is not positive the use of the word 'orthogonal' may be somewhat misleading.

The eigenvalue problem for $\mathcal{Q}$ and $\mathcal{N}$ can then equivalently be defined as the problem to find eigenfunctions $\varphi \neq 0$, such that for some eigenvalue λ:

$$\mathcal{Q}(\varphi, v) = \lambda \mathcal{N}(\varphi, v) \quad \textit{for all } v \in \mathcal{U}.$$

Since this can be rewritten like

$$\delta\mathcal{Q}(u; v) = \lambda \delta\mathcal{N}(u; v),$$

one interpretation of an eigenfunction with eigenvalue λ is as a critical point of the functional

$$\mathcal{U} \ni u \to \mathcal{Q}(u) - \lambda \mathcal{N}(u).$$

However, since λ is not given, but has to be found, this is not a very useful attack. Much more fruitful is to interpret λ as a multiplier appearing from a constrained problem.

Proposition 39 *(Normalized) Eigenfunctions φ are critical points of:*

$$\varphi \in \mathit{Crit}\ \left\{\mathcal{Q}(u) |\, u \in U,\ \mathcal{N}(u) = 1\right\}.$$

Equivalently,

$$\varphi \in\ \mathit{Crit}\ \left\{\mathcal{R}(u)\ | u \in U\right\}, \textit{with } \mathcal{R}(u) = \frac{\mathcal{Q}(u)}{\mathcal{N}(u)}.$$

where $\mathcal{R}$ is the so-called Rayleigh quotient*. The corresponding eigenvalues are precisely the critical values $\mathcal{R}(\varphi)$.*

This formulation will be most useful, as we will see. It will determine the principal eigenvalue (the largest or smallest one) if $\mathcal{R}$ attains its maximum or minimum. Other eigenfunctions can then be found in a recursive, or non-recursive way, all based on the constrained variational formulation above.

To exploit the variational characterization of the eigenfunctions with the Rayleigh quotient in a constructive way, one has to find out whether the Rayleigh quotient is bounded from above or from below. (In finite dimensions it is bounded both from below and from above, in infinite dimensions this is not the case except in trivial cases).

Definition 40 *We say that $\mathcal{Q}$ is* (strongly) coercive *(or elliptic) with respect to $\mathcal{N}$ if the Rayleigh-quotient is bounded from below (but not from above): for some $\gamma \in \mathcal{R}$*

$$\mathcal{R}(u) \geq \gamma, \textit{ and for some sequence } u_m,\ \mathcal{R}(u_m) \to \infty.$$

Exercise.

1. If $\gamma > 0$, $\mathcal{Q}$ defines a norm itself, and the assumption of ellipticity means that this norm is stronger (not equivalent) than the N-norm. Show that on $U_0 = \{u(0) = u(\pi) = 0\}$ the norm $\mathcal{Q}(u) = \int u_x^2$ is coercive with respect to $\mathcal{N}(u) = \int u^2$.

2. If $\mathcal{R}$ is not definite (then $\gamma < 0$) the quadratic form Q is not a norm; by defining

$$\bar{\mathcal{Q}}(u) := \mathcal{Q}(u) + 2|\gamma|\mathcal{N}(u)$$

it follows that $\bar{\mathcal{Q}}(u) \geq |\gamma|\mathcal{N}(u)$, and so is a norm that is stronger than N. Since the eigenfunctions of $\mathcal{Q}$ and $\bar{\mathcal{Q}}$ are the same, and the eigenvalues just differ the constant shift 2γ, one could just as well study the eigenvalue problem for $\bar{\mathcal{Q}}$ and $\mathcal{N}$. Stated differently, it would be no restriction to assume $\mathcal{Q}$ to be positive definite from the start on.

3. Show that the Sturm-Liouville operator with

$$\mathcal{Q}(u) = \int p(x)u_x^2 + q(x)u^2$$

is coercive on U_0 with respect to $\mathcal{N} = \int \rho u^2$ provided p is non-negative (and non-trivial), q is bounded, and ρ is positive definite.

■

Spectral theorem for differential operators

Theorem 41 Principal eigenfunction and -value
Suppose that $\mathcal{Q}$ is coercive (elliptic) with respect to $\mathcal{N}$: for some $\gamma(>0)$ $\mathcal{Q}(u) \geq \gamma\mathcal{N}(u)$, and assume that the minimization problem for $\mathcal{R}$ has a solution. Then the solution

$$\varphi_1 \in Min\ \{\mathcal{Q}(u)|\ u \in \mathcal{U},\ \mathcal{N}(u) = 1\} \sim Min\,\{\mathcal{R}(u)\,|u \in \mathcal{U}\}$$

is the principal eigenfunction φ_1*, i.e. the eigenfunction corresponding to the smallest eigenvalue, the* principal eigenvalue*, λ_1 that is given by*

$$\lambda_1 = \mathcal{R}(\varphi_1)\ (\geq \gamma).$$

Any other eigenfunction (independent of φ_1) can be assumed to be orthogonal (both in $\mathcal{N}$-, as well as in $\mathcal{Q}$-sense) to φ_1. In the following formulation this will be exploited in a successive way.

Theorem 42 Successive characterization
The eigenfunctions and eigenvalues can be obtained in a successive way: if $\varphi_1, \ldots, \varphi_k$ are the eigenfunctions corresponding to the eigenvalues that are ordered like

$$(\gamma \leq)\ \lambda_1 \leq \lambda_2 \ldots \leq \lambda_k,$$

the "next" eigenfunction is found as the solution of

$$\varphi_{k+1} \in Min\ \left\{\mathcal{Q}(u)|\, u \in H, \mathcal{N}(u) = 1, \mathcal{N}(u, \varphi_j) = 0,\ for\ 1 \leq j \leq k\right\};$$

the corresponding eigenvalue $\lambda_{k+1} \equiv \mathcal{R}(\varphi_{k+1})$ "follows" λ_k in the sense that $\lambda_{k+1} \geq \lambda_k$, while, when $\lambda_{k+1} > \lambda_k$, there are no other eigenvalues in between.

Exercise.

1. The orthogonality constraints in the successive characterization are natural constraints: although essential in the definition of the constraint set, there is no effect in the equation for the critical point: the corresponding multiplier vanishes. To verify this, consider the equation for

   $$\psi \in \text{Crit } \{\mathcal{R}(u) \,|u \in H, \mathcal{N}(u,f) = 0\}$$

   where f is any given function. The governing equation is for some multipliers μ, σ

   $$L\psi = \mu N\psi + \sigma f, \quad with \quad \mu = \mathcal{R}(\psi).$$

   Verify that $\sigma = 0$ if f is some eigenfunction, but that in general σ will not vanish.

2. Proof with the methods of this chapter that the EVP from Linear Algebra

   $$Ax = \lambda Bx,$$

   with A and $B > 0$ symmetric matrices, has a complete set of eigenvectors.

3. *Variational accuracy*
   Suppose that in a numerical calculation an approximation $\hat{\varphi}_1$ for the first eigenfunction φ_1 is constructed that is correct up to order ε (in N-norm for instance):

   $$\varphi_1 - \hat{\varphi}_1 = \mathcal{O}(\varepsilon).$$

   Show that then the approximate first eigenvalue $\hat{\lambda}_1$ that is constructed is correct up to order ε^2:

   $$\lambda_1 - \hat{\lambda}_1 = \mathcal{O}(\varepsilon^2).$$

   Investigate the effect of an error ε in the calculation of φ_1 for the approximation of λ_2 and of φ_2. Do the same for higher eigenvalues and eigenfunctions.

■

Remark. By its nature, the above formulation requires the knowledge of the previous eigenfunctions to find the next eigenvalue: the eigenvalue λ_{k+1} follows by investigating the minimizer of $\mathcal{R}$ on the set of functions orthogonal to the eigenfunctions $\varphi_1, ..., \varphi_k$. When one wants to use this formulation in a numerical procedure, for instance, this may lead to serious error-accumulation: in calculating λ_1, an error in the calculation of φ_1 influences the constraint set for λ_2 and induces an additional error in the calculation of λ_2 and of φ_2, and so on. The conclusion must be that the successive characterization as given above is not very suitable for numerical calculation of the successive eigenvalues and eigenfunctions. In subsection .. a non-successive characterization that is free of error-accumulation is described.

■

Just as for symmetric matrices, the eigenfunctions for most differential operators form a complete set; this is a very strong result but requires in infinite

dimensions some additional compactness condition. The proof will be based on the successive characterization, but actually only requires the knowledge that the eigenvalues can be ordered and tend to infinity (are not bounded above). This is usually the case for differential operators when $\mathcal{Q}$ defines a norm that is 'essentially stronger' than $\mathcal{N}$.

Theorem 43 Completeness of the set of eigenfunctions
If each eigenvalue has finite multiplicity, and if the eigenvalues are unbounded:

$$(\gamma \leq)\ \lambda_1 \leq \lambda_2 \ldots \leq \lambda_k \ldots \to \infty,$$

then the set of eigenfunctions is complete, both with respect to the $\mathcal{N}$-norm and with respect to the $\mathcal{Q}$-norm[4]*.*

Generalized Fourier theory

The completeness result implies that any function in $\mathcal{U}$ can be written as a (generalized) *Fourier series*

$$u(x) = \sum_1^\infty u_m \varphi_m(x);$$

using the fact that the eigenfunctions are orthonormal, it follows directly that the *Fourier coefficients* are given by

$$u_m = \mathcal{N}(u, \varphi_m);$$

the infinite sum converges in the sense that

$$\mathcal{N}(u - \sum_1^M u_m \varphi_m) \to 0 \quad \textit{for} \ \ M \to \infty$$

and also in the stronger norm

$$\mathcal{Q}(u - \sum_1^M u_m \varphi_m) \to 0 \quad \textit{for} \ \ M \to \infty.$$

Fredholm alternative

Another interpretation is that the operator L is in *diagonal form* with respect to a basis of eigenfunctions, and hence that the inverse of L can be found easily. For simplicity suppose that $N = Identity$, and consider the inhomogeneous problem

$$Lu = f, \quad u \in \mathcal{U}.$$

Writing $f = \sum f_n \varphi_n$, with f_n the Fourier coefficients of f, the solution is given by

$$u = \sum_m \frac{f_m}{\lambda_m} \varphi_m,$$

at least when

[4] When $\mathcal{Q}$ is not definite, this should be understood with respect to the $\bar{\mathcal{Q}}$-norm, with $\bar{\mathcal{Q}} = \mathcal{Q} + (|\lambda_1| + \varepsilon)\mathcal{N}$, for any $\varepsilon > 0$.

- *either* all eigenvalues λ_m are non-zero (the operator L is invertible),
- *or*, if there is a zero eigenvalue, with eigenspace

  $E_{\lambda=0}$ *(*consisting of the eigenfunctions with eigenvalue 0*),*

  there exists a solution only if the inhomogeneous term satisfies the orthogonality conditions

  $f \perp E_{\lambda=0}$;

  in that case the solution is not unique: any element from $E_{\lambda=0}$ can be added.

These results are just a straightforward generalization of the *Fredholm alternative* for (symmetric) matrices.

Example: EVP for S-L operator on an interval

This example shows that the results for the EVP for specific operators are generalizations of the usual Fourier theory. For given positive functions ρ and p, and a function q on $[0, \pi]$ (all smooth), the Sturm-Liouville eigenvalue problem (with Dirichlet boundary conditions) reads:

$$L\varphi = -\partial_x(p(x)\varphi_x) + q(x)\varphi = \lambda\rho(x)\varphi, \quad \varphi(0) = \varphi(1) = 0,$$

and is obtained in $\mathcal{U}_0 = \{u \in L_2 \mid u(0) = u(\pi) = 0\}$ with the quadratic forms

$$\mathcal{N}(u) = \int \rho(x)u^2, \quad \mathcal{Q}(u) = \int [p(x)u_x^2 + q(x)u^2].$$

Exercise.

1. The special case $\rho \equiv 1, p \equiv 1, q \equiv 0$ provides Fourier theory (for functions that are odd on $[-\pi, \pi]$): then the eigenvalues and (normalized) corresponding eigenfunctions are given by

   $$\lambda_m = m^2, \quad \varphi_m = \sqrt{2/\pi}\sin mx, \quad m \geq 1.$$

   The completeness result in the spectral theorem implies that any function satisfying the boundary conditions can be written as a Fourier-sine series

   $$u(x) = \sqrt{2/\pi}\sum_1^\infty u_m \sin mx,$$

   for Fourier coefficients given by

   $$u_m = \langle u, \varphi_m\rangle = \sqrt{2/\pi}\int u(x)\sin mx dx;$$

   the convergence in the $\mathcal{N}$-norm is just the usual L_2-norm:

   $$\int (u - \sum_1^M u_m\varphi_m(x))^2 dx \to 0, \quad \textit{for } M \to \infty.$$

The convergence in the $\mathcal{Q}$-norm implies a much stronger statement. To investigate that, exploit the Poincaré inequality: for some constant $c_1 > 0$ it holds that

$$|u|_\infty^2 \leq c_1 \int u_x^2 \quad \textit{for all } u,\ u(0 = u(\pi) = 0.$$

Then the convergence in the $\mathcal{Q}$-norm implies the pointwise convergence of the Fourier-sine series:

$$|u - \sum_1^M u_m \varphi_m(x)|_\infty \to 0,\ \textit{for } M \to \infty.$$

2. Changing the boundary conditions to Neumann boundary conditions:

$$u_x(0) = u_x(\pi) = 0$$

provides Fourier cosine series, since then

$$\lambda_m = m^2, \quad \varphi_m = \sqrt{2/\pi}\cos mx, \quad m \geq 0;$$

completeness in L_2-norm, and pointwise, in the same way as above.

3. Observe that in both cases the eigenvalues are "simple": to each eigenvalue there corresponds precisely one eigenfunction; equivalently: the eigenspaces are one-dimensional. This is characteristic for Sturm-Liouville eigenvalue problems on an interval.

4. For functions $u \in U_0$, where U_0 are functions from $C^1([0,1])$ with $u(1) = 0$, show that the following Poincaré -Friedrichs inequality holds for a suitable constant c:

$$\int_0^1 u_x^2 \geq c \int_0^1 u^2.$$

Can you determine the best possible value of c?

■

Example: EVP for S-L operator on a spatial domain

For a domain $\Omega \subset \mathcal{R}^n$, with boundary $\partial\Omega = \partial\Omega_D \cup \partial\Omega_N$, and for given functions $p(x), q(x)$ and $\rho(x)$, the quadratic forms

$$\mathcal{N}(u) = \int_\Omega \rho u^2, \quad \mathcal{Q}(u) = \int_\Omega p(x)|\nabla u|^2 + q(x)u^2$$

on the set

$$\mathcal{U} = \{\ u : \Omega \in \mathcal{R}\ \mid\ u(x) = 0\ \ \textit{for}\ \ x \in \partial\Omega_D\ \}$$

leads to the EVP

$$\begin{aligned} -\text{div}\ (p(x)\nabla\phi) + q(x)\phi &= \lambda\phi \ \text{ in } \ \Omega \\ \phi &= 0 \quad \text{on } \partial\Omega_D \\ p(x)\partial_n\phi &= 0 \quad \text{on } \partial\Omega_N \end{aligned}$$

Sufficient conditions on the functions p, ρ that make it possible to apply the general theory are that they are positive definite:

$$p(x) \geq p_0 > 0, \quad \rho(x) \geq \rho_0 > 0.$$

Then existence and completeness follows.
(It should be noted that in more dimensions ($n \geq 2$) the convergence in $\mathcal{Q}$-norm does *not* imply pointwise convergence; only for functions of one variable the Poincaré inequality holds!)

Exercise.

1. For given smooth and bounded functions $p(x)$ and $q(x)$, consider the quadratic forms

$$\mathcal{Q}(u) = \int_0^1 p(x)u_x^2 + q(x)u^2, \quad \mathcal{N}(u) = \int_0^1 u^2 \tag{2.14}$$

   (a) Write down the Sturm-Liouville eigenvalue problem corresponding to Q and N on the set U_0 of functions from $C^1([0,1])$ with $u(1) = 0$.

   (b) Show that if the function p in (1) is strictly positive on the entire interval $[0,1]$, the Rayleigh quotient

$$\mathcal{R}(u) := \frac{\mathcal{Q}(u)}{\mathcal{N}(u)}$$

   is bounded from below on U_0.

2. In a few specific cases, for special domains Ω, the eigenfunctions can be found explicitly. In all these cases the method of separation of variables is used. This can be done for instance for rectangular domains and for balls. Study the EVP for the Laplace operator with Dirichlet, and then with Neumann, boundary conditions for a domain $\Omega \in \mathcal{R}^2$ in case

   - Ω is the rectangle $[0,a] \times [0,b]$,
   - Ω is the unit disc.

■

Comparison methods for principal eigenvalues

Often we want to compare the eigenvalues of two different eigenvalue problems. When for each problem the eigenvalues are found in a variational way, this may be done in an elegant way. The eigenvalue problems to be compared may differ in three ways (or combinations thereof)

- the operators are different,
- the boundary conditions are different,
- the domain of definition of the functions is different.

Exercise. Consider the vibrations of a linear string governed by

$$u_{tt} = \partial_x(\sigma(x)u_x), \quad u(0,t) = u(\ell,t) = 0,$$

where ℓ is the length, and σ is a material property. When looking for time-harmonic solutions of the form

$$u(x,t) = v(x)\exp[i\omega t]$$

there results the eigenvalue problem for v with $\omega^2 = \lambda$:

$$-\partial_x(\sigma(x)v_x) = \lambda v, \quad v(0) = v(\ell) = 0.$$

The principal eigenvalue λ_1 determines the lowest frequency of vibration of the string; in practise it determines the fundamental tone of a piano etc. Of course, its value depends on the length ℓ, on the material properties described by σ, and in fact also on the boundary conditions.

1. For σ is constant, the principal eigenvalue is given by

   $$\lambda_1 = \sigma(\frac{\pi}{\ell})^2 \text{ with eigenfunction } \varphi_1 = \sin(\frac{\pi x}{\ell}).$$

   Hence, the principal eigenvalue decreases when the length increases, and/or when the tension σ decreases.

2. For a string with a free endpoint at $x = \ell$, the boundary conditions are replaced by $v(0) = 0, v_x(\ell) = 0$. Then the principal eigenvalue μ_1 is given by

   $$\mu_1 = \sigma(\frac{\pi}{2\ell})^2, \quad \text{with eigenfunction } \psi_1 = \sin(\frac{\pi x}{2\ell})$$

   and produces a lower fundamental tone since $\mu_1 < \lambda_1$.

3. The same result holds true for the non-fundamental eigenvalues:

   $$\lambda_m = \sigma(\frac{m\pi}{\ell})^2, \quad \text{with eigenfunction } \varphi_m = \sin(\frac{m\pi x}{\ell}),$$

   $$\mu_m = \sigma(\frac{m\pi}{2\ell})^2, \quad \text{with eigenfunction } \psi_m = \sin(\frac{m\pi x}{2\ell});$$

   note that $\mu_m < \lambda_m$ for all $m \geq 1$.

4. The physical statement of these results is that *upon relaxing the constraints, the eigenvalues decrease.*

■

Using the extremal characterization for the principal eigenvalue, comparison between different problems may be relatively easy. For ease of presentation we will mainly deal with differential operators for which the principal eigenvalue *minimizes* the Rayleigh quotient.

Proposition 44 *Let $\mathcal{U}$ be the linear space, and $\mathcal{R}$ the Rayleigh quotient. Suppose that the principal eigenvalue Λ minimizes $\mathcal{R}$ on $\mathcal{U}$; making its dependence explicitly, we write*

$$\Lambda(\mathcal{R},\mathcal{U}) = Min\{\mathcal{R}(u) \,|u \in \mathcal{U}\}$$

Then Λ depends monotonically on $\mathcal{R}$ and on $\mathcal{U}$ in the following senses:

Proposition 45 •

$$\Lambda(\mathcal{R}_1,\mathcal{U}) \leq \Lambda(\mathcal{R}_2,\mathcal{U}) \text{ if } \mathcal{R}_1(u) \leq \mathcal{R}_2(u) \text{ for all } u \in \mathcal{U}$$

•

$$\Lambda(\mathcal{R},\mathcal{U}_1) \leq \Lambda(\mathcal{R},\mathcal{U}_2) \text{ if } \mathcal{U}_1 \supset \mathcal{U}_2.$$

Proof. The first statement follows from

$$\begin{aligned}
&\mathcal{R}_1(u) \leq \mathcal{R}_2(u) \quad \textit{for all } u \in \mathcal{U} \\
&\Longrightarrow \text{Min } \{\mathcal{R}_1(u) | u \in \mathcal{U}\} \leq \mathcal{R}_2(u) \quad \textit{for all } u \in \mathcal{U} \\
&\Longrightarrow \text{Min } \{\mathcal{R}_1(u) | u \in U\} \leq \text{Min } \{R_2(u) | u \in \mathcal{U}\}.
\end{aligned}$$

The second statement from the fact that the minimizer decreases (or at least does not increase) if the domain of definition is enlarged ("relaxing the constraints ...").

■

Exercise. *Sturm-Liouville comparison results* ■

1. *Different operators*

   (a) Let $p(x) = 1 + x^2$, and $q(x) = \sin x$ in the scalar S-L-operator and consider Dirichlet conditions. Derive a lower bound and an upper-bound for the smallest eigenvalue of the Sturm- Liouville eigenvalue problem.

   (b) Let $\Lambda_{1,2}$ be the principal eigenvalue of respectively

   $$-\text{div } [p_1(x)\nabla u] + q_1(x)u = \Lambda_1 \rho_1 u,$$

   $$-\text{div } [p_2(x)\nabla u] + q_2(x)u = \Lambda_2 \rho_2 u$$

   on a domain Ω with the same boundary conditions. If

   $$p_1 \leq p_2, \quad q_1 \leq q_2, \quad \rho_1 \geq \rho_2 \quad \textit{on } \Omega,$$

   the Rayleigh quotients satisfy

   $$\mathcal{R}_1(u) \equiv \frac{\int [p_1 |\nabla u|^2 + q_1 u^2]}{\int \rho_1 u^2} \leq \frac{\int [p_2 |\nabla u|^2 + q_2 u^2]}{\int \rho_2 u^2} \equiv \mathcal{R}_2(u),$$

   and hence $\Lambda_1 \leq \Lambda_2$.

2. *Different boundary conditions*
   For the same S-L operator on Ω, consider two different boundary conditions

   $$\begin{aligned}
   u &= 0 \text{ on } \partial\Omega_1, \ \& \ \partial_n u = 0 \text{ on } \partial\Omega/\partial\Omega_1 \\
   u &= 0 \text{ on } \partial\Omega_2, \ \& \ \partial_n u = 0 \text{ on } \partial\Omega/\partial\Omega_2
   \end{aligned}$$

   It should be noted now that the Neumann boundary conditions arise as natural boundary conditions; hence the correct boundary conditions are obtained by investigating the Rayleigh quotient on the sets

   $$\mathcal{U}_{1,2} = \{ u \mid u = 0 \textit{ on } \partial\Omega_{1,2} \}.$$

   When $\partial\Omega_1 \subset \partial\Omega_2$ ("relaxing ..."), it holds that $\mathcal{U}_1 \supset \mathcal{U}_2$, and hence $\Lambda_1 \leq \Lambda_2$.

3. *Different domains, Dirichlet boundary conditions*
   Consider the same S-L operator with Dirichlet boundary condition on two domains $\Omega_2 \subset \Omega_1$ (the functions are defined on the largest domain, and so is the Rayleigh quotient). Any function $v_2 \in \mathcal{U}_2 = \{\ v : \Omega_2 \ \mid \ v = 0\ on \partial\Omega_2\ \}$ can be extended to a function v_1 on Ω_1 by assigning it the value zero for $x \in \Omega_1/\Omega_2$; this defines the space of functions $\bar{\mathcal{U}_1} = \{\ v : \Omega_1 \ \mid \ v = 0\ for x \in \Omega_1/\Omega_2\ \}$. Since extension with zero does not change the value of the Rayleigh quotient, $\mathcal{R}(v_1) = \mathcal{R}(v_2)$, and since $\bar{\mathcal{U}_1} \subset \mathcal{U}_1 = \{\ u : \Omega_1 \ \mid \ u = 0\ on \partial\Omega_1\ \}$, it follows that

$$\begin{aligned} \Lambda_2 &= \text{Min}\ \{\mathcal{R}(v_2)|\ v_2 \in \mathcal{U}_2\} = \text{Min}\ \{\mathcal{R}(v_1)|\ v_1 \in \mathcal{U}_1\} \\ &\geq \text{Min}\ \{\mathcal{R}(u)|\ u \in \mathcal{U}_1\} = \Lambda_1 \end{aligned}$$

Remark. It should be noted that the inequalities derived above for the principal eigenvalue cannot so easily be extended to non-principal eigenvalues. The reason is that, for instance for the second eigenvalue, R has to be investigated on the functions orthogonal to the first eigenfunction. When dealing with two problems, the principal eigenfunctions, say Φ_1 and Ψ_1, will differ, and so will their orthogonal complements:

$$\{\ u\ \mid u \in \mathcal{U},\ \mathcal{N}(u, \Phi_1) = 0\ \} \neq \{\ u\ \mid u \in \mathcal{U},\ \mathcal{N}(u, \Psi_1) = 0\ \}$$

while one is not simply included in the other. Hence, no conclusions can be drawn by considering the minimization problems, not even for the same R:

$$\text{Min}\ \{\mathcal{R}(u)|\mathcal{N}(u, \Phi_1) = 0\} = \ ??\ = \text{Min}\ \{\mathcal{R}(u)|\mathcal{N}(u, \Psi_1) = 0\}$$

and hence no conclusions for the second eigenvalue.
This situation motivates the characterization of the second (and higher) eigenvalue without using the first eigenfunction; in the next section we will describe such a non-successive characterization. ■

2.3.2 Algorithm for Relative Equilibrium (Solutions)

Many partial differential equations for problems from the physical and technical sciences are nonlinear, and as such are difficult in the sense that usually no explicit solutions can be written down. Occasionally special solutions can be found, and then usually a who family can be found, a family depending on parameters. These special solutions are often found in an ad-hoc way, using some special Ansatz. But actually, in many cases for dynamical systems, a deep result from Classical Mechanics lies behind this Ansatz. This result can be generalized to Variational Evolution Equations, like wavelike equations and more general continuous Poisson systems. Without proof, it is formulated as an algorithm in the next subsection; from this it also becomes clear that the notion generalizes the idea of 'simple' equilibrium solutions.

Consider a dynamical system with Poisson structure that is written in a general form like

$$\partial_t u = \Gamma \delta H(u)$$

Critical points of the Hamiltonian are then *Equilibrium solutions*:

$$\text{For } \delta H(\bar{u}) = 0, \text{ it holds } \partial_t \bar{u} = 0$$

Assume that the system has one or more 'constants of the motion', meaning function(als) I that do not change in time on solutions:

$$\partial_t I(u) = 0 \text{ for solutions of } \partial_t u = \Gamma \delta H(u);$$

such a functional is often simply called an 'integral' of the dynamical system. Denote the 'flow' of each integrals by Φ^I, meaning that

$$\Phi^I_\tau(v_0) \text{ is the solution of the IVP } \partial_\tau v = \Gamma \delta H(v), \quad v(\tau = 0) = v_0.$$

Then (provided some technical conditions are satisfied, roughly speaking that the integrals are 'independent'), consider the constrained critical point problem:

$$\text{Crit } \{ H(u) \mid I(u) = \gamma \} \tag{2.15}$$

Any constrained critical point is called a *relative equilibrium*; it satisfies for certain multiplier

$$\delta H(\check{u}) = \lambda \delta I(\check{u}). \tag{2.16}$$

Related to this relative equilibrium there is a true solution of the system, called a *relative equilibrium solution* $U(t)$, for which the time evolution is actually given by

$$U(t) = \Phi^I_{\lambda t}(\check{u}); \tag{2.17}$$

for more integrals this has to be interpreted as a superposition of flows (in arbitrary order[5]):

$$U(t) = \Phi^{I_1}_{\lambda_1 t} \circ \Phi^{I_2}_{\lambda_2 t} \circ ... \Phi^{I_N}_{\lambda_N t}(\check{u})$$

In many applications these solutions are rather special and often referred to as *coherent structures* of the dynamical system.

Note that this result means that the relative equilibria can be found from a Nonlinear Eigenvalue Problem of variational form in which the functionals have direct physical relevance for the dynamics. In many cases, the multipliers have a clear relevant meaning also. We will show in Appendix C that 'solitons' in KdV and NLS are special cases of the principle above.

Exercise. *Spherical pendulum*

1. Derive the equation for spherical pendulum using as Lagrangian the difference of kinetic and potential energy, expressed in spherical angle coordinates.

[5] This is one of the basic results for Poisson systems which is not proved in this course: Provided the functionals I_m are 'independent' in the sense that they Poisson-commute, i.e. that they satisfy, using the notation of Poisson bracket, $\{I_k, I_m\} = 0$, their flows commutes : $\Phi^{I_k} \circ \Phi^{I_m} = \Phi^{I_m} \circ \Phi^{I_k}$.

2. Show that the dynamics is a Hamiltonian system with Hamiltonian

$$H = \frac{1}{2}\left(p_\theta^2 + \frac{p_\phi^2}{\sin^2\theta}\right) + \omega_0^2(1-\cos\theta)$$

3. Observe that there is, except from H, an additional first integral I, the so-called angular momentum.

4. Reduce the dynamics by prescribing the value of I, and find equilibria of this reduced dynamics.

5. Show that the equilibria of the reduced dynamics are in fact the relative equilibria: constrained minimizers of H at given I.

6. Determine the relative equilibrium solutions and interpret their motion in space. Relate the angular velocity to (the derivative) of the relevant value function.

■

2.3.3 Thermodynamic systems: constrained steepest descent

We have seen that gradient systems can be used for numerical purposes to find the minimizer of a given functional; the dynamic trajectories are in the direction of steepest descent.

When looking for constrained minimizers, this method has to adapted to take the constraints into account. We briefly describe the modification; for simplicity we restrict ourselves to the case of one functional constraint.
These systems are called thermodynamic systems, since there is one integral that is conserved (the "energy") while another one decreases monotonically (the "entropy")[6].

Definition 46 *A* thermodynamic system *is a dynamical system in the state space $\mathcal{U}$ that is of the form*

$$\partial_t u = -[\delta H(u) - \lambda(u)\delta I(u)], \quad \textit{with } \lambda(u) = \frac{\langle \delta I(u), \delta H(u)\rangle}{\langle \delta I(u), \delta I(u)\rangle}$$

where $H : \mathcal{U} \to \mathcal{R}$ and $I : \mathcal{U} \to \mathcal{R}$ are functionals.

To see the dynamic properties, observe that any functional F evolves according to

$$\partial_t F(u) = \langle \delta F(u), [\delta H(u) - \lambda(u)\delta I(u)]\rangle.$$

Substituting the functional I for F and the expression for $\lambda(u)$, it follows that

$$\partial_t I(u) = \langle \delta I(u), [\delta H(u) - \lambda(u)\delta I(u)]\rangle = 0$$

[6] From a mathematical point of view, the system can also be considered as a simple example of a dynamical system on a manifold: the level set of the conserved functional as the manifold on which a disspative system is defined.

Hence, for the dynamics the functional I is conserved, a constant of the motion, a *first integral*:

$$I(u(t)) = I(u(0)) \text{ for all } t.$$

Geometrically this is seen since the vectorfield $\delta H - \lambda \delta I$ is perpendicular to δI, and so tangent to the level sets of I.

On each level set of I, the system behaves like a dissipative system as treated in Chapter 1. In fact, for the evolution of H:

$$\begin{aligned} \partial_t H(u) &= \langle \delta H(u), [\delta H(u) - \lambda(u)\delta I(u)] \rangle \\ &= \langle [\delta H(u) - \lambda(u)\delta I(u)], [\delta H(u) - \lambda(u)\delta I(u)] \rangle, \end{aligned}$$

so

$$\partial_t H(u) \begin{cases} \leq 0 \\ = 0 \text{ iff } \delta H(u) = \lambda \delta I(u) \end{cases} .$$

From this it follows that H decreases monotonically, except from the points that are the equilibria of the system. Indeed, an equilibrium solution satisfies the equation

$$\delta H(\hat{u}) = \lambda \delta I(\hat{u})$$

for some scalar λ. Recalling Lagrange's multiplier rule, this is the equation for constrained critical points of H on a level set of I:

$$\hat{u} \in \text{Crit } \{ \ H(u) \mid I(u) = \gamma \ \}.$$

From the above observations it is clear that the trajectories are in the direction of *constrained steepest descent*; hence the equation can be used to find constrained minimizers of H on level sets of I in a numerical way.

Exercise.

1. The dynamic system above provides a way to prove Lagrange's multiplier rule in an alternative way, different from the proof as given before. Give the detailed argumentation.

2. Write down the equation of constrained steepest descent to find the solution of

$$\text{Min } \{ \int u_x^2 \mid \int u^2 = 1, u(0) = u(\pi) = 0 \},$$

and investigate the convergence to the minimal element.

■

2.4 Exercises

1. *Lemma DuBois-Reymond, integrated Euler-Lagrange equation*
   We start with a generalization of Lagrange's Lemma, and then show how it can be used to weaken the regularity assumptions for a critical point that were required in the first chapter.

   (a) Use the (linear) multiplier rule to prove the following

   Proposition 47 Lemma DuBois-Reymond
   Let $f \in C^0([0,1])$ *satisfy*

   $$\langle f(x), \xi(x) \rangle = 0, \text{ for all } \xi \in C_0^\infty \text{ with } \int_0^1 \xi(x)dx = 0.$$

   Then f *is constant on* $[0,1]$.

   (b) Give a different proof (based on Lagrange's Lemma) by reformulating the information for f by observing that

   $$\{\eta_x \mid \eta \in C_0^\infty([0,1]) \} \equiv \{ \xi \in C_0^\infty([0,1]) \mid \int_0^1 \xi(x)dx = 0\}.$$

   Observe that this method requires the assumption that f is differentiable!

   (c) *Integrated Euler-Lagrange equation for single integral problems*
   For the single integral functional $\mathcal{L}(u) \equiv \int_a^b L(u, u_x)dx$, and the expression for the first variation

   $$\delta\mathcal{L}(u;\eta) = \int [L_u \eta + L_{u_x} \eta_x]dx$$

   the Euler-Lagrange equation was found in Chapter 1 by partial integration of the last term. This required the assumption that L_{u_x} is differentiable, which usually requires the assumption that $u \in C^2$. Show that, by partial integration of the first (!) term, this assumption can be avoided when using the results of the Lemma of DuBois-Reymond above. The result is then the integrated form of the Euler-Lagrange equation:

   $$-\int^x [L_u] + L_{u_x} = \text{ constant on } [a,b].$$

   Inspecting this integrated form, conclude that at a critical point $\hat{u} \in C^1$ actually L_{u_x} is differentiable. In most cases this implies that actually $\hat{u} \in C^2$. This means that for a critical point *additional regularity is obtained from the stationarity condition!*
   Illustrate this to the simple case $L = \frac{1}{2}u_x^2 - V(u)$.

2. Consider for $\gamma > 0$ the value function of the minimization problem

   $$h(\gamma) = \text{Min } \left\{ \int_0^\pi \left[u_x^2 + u^2\right] dx \;\middle|\; \int_0^\pi u^4 dx = \gamma, u(0) = u(\pi) = 0 \right\}.$$

   Determine, up to a multiplicative factor, the value function by using the homogeneity of the functionals (do not try to calculate the minimizers explicitly).

3. *Geometric problems*
   Following are a few of the classical problems that deal with constrained problems. (Exploit "energy conservation" to help to solve the equations explicitly.)

   (a) *Dido's problem*
   Find the surface of largest area, given the value of its perimeter.

   (b) *Chain line*
   Find (from minimal potential energy) the form of a chain with prescribed length hanging in the (constant) gravitational field between given points.

   (c) *Brachistochrone*
   Determine the shape of a curve of given length in the vertical plane that is such that the time for a particle (starting with zero initial velocity at the highest end point) to reach the other point is as small as possible (friction neglected).

4. *Rotating free fluid surface*
   Consider a standing circular cylinder, with radius R, and with the z-axis pointing in the vertical direction as axis. Assume that the cylinder is partly filled with water (mass density $\rho \equiv 1$) that rotates with constant angular velocity ω around the z-axis. Assume that friction of the fluid with the cylinder wall and the bottom can be neglected.
   Taking the bottom at $z = 0$, the free surface of the fluid can be described by a function $z = \eta(r)$, where r is the radial distance from the axis.

   (a) The volume of the fluid is determined when the free surface is given, i.e. is a functional of η, to be called $V(\eta)$. Determine this functional.

   (b) The kinetic energy of the rotating fluid is given by

   $$K(\eta) = \int_0^R \pi\, \omega^2\, r^3\, \eta(r)\, dr$$

   and the potential energy (with g the gravitational acceleration) by

   $$P(\eta) = \int_0^R \pi\, g\, r\, \eta(r)^2\, dr.$$

      i. Write down the equation for a minimizer of the functional $K(\eta) - P(\eta)$ on the set of functions η that satisfy the constraint

      $$\int_0^R 2\pi\, r\, \eta(r)\, dr = \gamma$$

      where γ is a given constant.

      ii. Determine the minimizing free surface explicitly (for γ sufficiently large).

5. *Free fluid surface in a container*
   Consider a cylinder with axis vertically (in the direction of gravitation,

the z-axis), partly filled with fluid. Assuming that the bottom of the cylinder is described at $z = 0$ by the region $\Omega \in \mathcal{R}^2$, the fluid surface will be described by $u = u(x, y)$, so that the fluid occupies the region

$$\{\ (x, y, z)\ \mid\ (x, y) \in \Omega,\ 0 \leq z \leq u(x, y)\ \}.$$

We are looking for the form of the free surface of the fluid, i.e. the function u, from a minimal energy principle when effects of surface tension and adhesion are taken into account.
To that end, let

- S = the area of the free surface,
- S^* = the area of the wetted part of the cylinder wall,
- V = the volume of the water.

(a) Describe S, S^*, V as functionals of u.

(b) For given $V_0 > 0$ and $\sigma \in \mathcal{R}$ with $|\sigma| < 1$ the minimization problem

$$\text{Min } \{S(u) - \sigma S^*(u) | u \in C^1(\Omega), V(u) = V_0\}.$$

Interpret this optimization problem in physical terms as a minimum energy principle.

(c) Supposing that $\hat{u} \in C^2$, determine the governing boundary value problem for a minimizer $\hat{u}$.

(d) Write $\sigma = \sin\beta$, with $\beta \in (-\pi/2, \pi/2)$. Give the meaning of β. Sketch the form of the free surface for the two different cases that $\sigma < 0$ and $\sigma > 0$. (Can you give examples of fluids with these properties?)

(e) Express the multiplier that appears in the equation for $\hat{u}$ in terms of known quantities (σ, area of Ω, length of $\partial\Omega$).

(f) Approximate the functional S for surfaces for which ∇u is small, by a constant plus a quadratic functional $\bar{S}(u)$. Write down the governing boundary value problem.

(g) Consider the special case of a circular cylinder: $\partial\Omega$ is the circle with radius R. Introduce cylinder coordinates (r, ϕ, z) and write $\bar{S}, S^*, V$ as functionals of $u = u(r, \phi)$. Express the multiplier in terms of σ and R.
Determine explicitly the free surface for given σ and V_0 (sufficiently large) in the approximation considered.

(h) Find a lower bound for V_0 (given σ and R) for which a C^2-solution can be expected. What is the physical interpretation?

6. ** *Periodic oscillations of constrained (pseudo-) potential energy*
When looking for periodic solutions of a second order system with potential energy V

$$-\ddot{q} \ =\ V'(q),\ \ q(0) = q(T),\ \dot{q}(0) = \dot{q}(T),$$

where the period T is not prescribed in advance, one may try to use the constrained critical point problem

$$\begin{aligned} \text{Crit } \{\mathcal{K}(x) \mid \mathcal{V}(x) &= R, x \in X\}, \\ \text{with } \mathcal{K}(x) &= \int_0^1 \frac{1}{2}|\dot{x}|^2 d\tau, \quad \mathcal{V}(x) = \int_0^1 V(x) d\tau \end{aligned}$$

and $X = \{\ x \in C^1([0,1])\ \mid\ x(0) = x(1)\ \}$.

(a) Give sufficient conditions for the potential energy function $\mathcal{V}$ that imply that the multiplier in the equation for the constrained critical points:

$$-\ddot{x} = \lambda V'(x)$$

is positive.

(b) Show that, when $\lambda > 0$, the critical points $x(\tau)$ correspond to the desired periodic solutions up to a scaling of the time variable. Give the physical meaning of the functionals $\mathcal{K}$ and $\mathcal{V}$ expressed in terms of $q(t)$.

(c) The minimization problem Min $\{\mathcal{K}(x) \mid \mathcal{V}(x) = R, x \in X\}$ (assuming the constrained set to be non-empty) has a trivial solution, viz. a constant. Therefore, we have to look for non-minimal critical points.

(d) One case in which non-trivial critical points can be found is when $\mathcal{V}$ is an even function: $\mathcal{V}(x) = \mathcal{V}(-x)$. Show that in that case periodic solutions can be found on an interval $[-1,1]$ by odd continuation of a critical point on $[0,1]$ of

$$\text{Min } \{\mathcal{K}(x) | \mathcal{V}(x) = R, x \in X, x(0) = x(1) = 0\}.$$

(e) Find the solutions when $x = (x_1, x_2)$ and $V(x) = x_1^2 + 3x_2^2$.

2.5 ** Extensions

2.5.1 Theory of Constrained Second Variation

For the manifold $\mathcal{M}$ given in (2.1) we calculate the second variation at a constrained critical point $\hat{u}$ that satisfies (2.5).
To that end, take $v \in T_{\hat{u}}\mathcal{M}$ and investigate for which functions w, depending on ε and v, the curve

$$\varepsilon \mapsto \hat{u} + \varepsilon v + w$$

belongs to $\mathcal{M}$, i.e. satisfies the constraints. Assuming $w = o(\varepsilon)$ from the start (i.e. $w/\varepsilon \mapsto 0$ for $\varepsilon \mapsto 0$), it follows from

$$\begin{aligned} \mathcal{K}(\hat{u} + \varepsilon v + w) &= \mathcal{K}(\hat{u}) + \delta\mathcal{K}(\hat{u}; \varepsilon v + w) + \frac{1}{2}\varepsilon^2\delta^2\mathcal{K}(\hat{u}; v) + o(\varepsilon^2) \\ &= \mathcal{K}(\hat{u}) + \delta\mathcal{K}(\hat{u}; w) + \frac{1}{2}\varepsilon^2\delta^2\mathcal{K}(\hat{u}; v) + o(\varepsilon^2) \end{aligned} \tag{2.18}$$

that w has to satisfy

$$\delta\mathcal{K}_k(\hat{u};w) + \frac{1}{2}\varepsilon^2\delta^2\mathcal{K}_k(\hat{u};v) + o(\varepsilon^2) = 0, \tag{2.19}$$

for $1 \leq k \leq p$. Calculating the functional $\mathcal{L}$ on such a curve, using equation (2.5) for $\hat{u}$, produces

$$\mathcal{L}(\hat{u}+\varepsilon v+w) = \mathcal{L}(\hat{u}) + \varepsilon\delta\mathcal{L}(\hat{u};v) + \delta\mathcal{L}(\hat{u};w) + \frac{1}{2}\varepsilon^2\delta^2\mathcal{L}(\hat{u};v) + o(\varepsilon^2) \tag{2.20}$$

$$= \mathcal{L}(\hat{u}) + \Sigma_k\ \lambda_k\delta\mathcal{K}_k L(\hat{u};w) + \frac{1}{2}\varepsilon^2\delta^2\mathcal{L}(\hat{u};v) + o(\varepsilon^2). \tag{2.21}$$

Inserting the expression for $\delta\mathcal{K}_k(\hat{u};w)$ from (2.19), there results:

$$\mathcal{L}(\hat{u}+\varepsilon v+w) = \mathcal{L}(\hat{u}) + \frac{1}{2}\varepsilon^2\left[\delta^2\mathcal{L}(\hat{u};v) - \Sigma_k\lambda_k\delta^2\mathcal{K}_k(\hat{u};v)\right] + o(\varepsilon^2) \tag{2.22}$$

The expression

$$\delta^2\mathcal{L}(\hat{u};v) - \Sigma_k\ \lambda_k\delta^2\mathcal{K}_k(\hat{u};v) \tag{2.23}$$

for $v \in T_u\mathcal{M}$ is called the *constrained second variation* of $\mathcal{L}$ on the manifold $\mathcal{M}$ at the critical point $\hat{u}$. Note that it is precisely the (unconstrained) variation of the Lagrangian functional (2.6) that leads to the equation for $\hat{u}$, but restricted to variations from the tangent space.

Proposition 48 *If $\bar{u}$ is a local extremal for $\mathcal{L}$ on the manifold $\mathcal{M}$ given by (2.1), that satisfies the multiplier equation (2.5), the constrained second variation (2.23) is sign-definite in all directions v from the tangent space.*
Specifically, if $\mathcal{L}$ has a (local) minimum at $\bar{u}$, then

$$\delta^2\mathcal{L}(\bar{u};v) - \sum_k \lambda_k\delta^2\mathcal{K}_k(\bar{u};v) \geq 0 \text{ for all } v \in T_{\bar{u}}\mathcal{M}. \tag{2.24}$$

2.5.2 LEVP: Non-successive characterization of eigenvalues

Recall the two reasons we have encountered until now to look for non-successive characterizations for eigenvalues: the error-accumulation when using numerical approximations, and the comparison of non-principal eigenvalues described above.

Min-max and Max-Min formulations

Chapter 3

Variational approximations

3.1 Motivation and Introductory Examples

In the introductory examples in the first Chapter we motivated and illustrated the variational formulations for problems from mathematical physics as generalizations of finite dimensional optimization problems. Now, in a certain sense, we will do the opposite: we will approximate an infinite dimensional variational problems by some simplified, finite dimensional problem. Sometimes the aim is to get an approximation that is as good as possible; sometimes the aim is just to get a simpler formulation that is easier to investigate. The procedure is actually very simple:

Summary. of variational restriction-method:
Starting with the variational formulation of a certain problem (the 'exact' formulation), say

$$\hat{u} \in \text{Crit } \{ \mathcal{L}(u) \mid u \in \mathcal{M} \},$$

a ***consistent simplified model*** *is obtained by restricting the set of competing functions:*

$$\hat{u}_{\mathcal{S}} \in \text{Crit } \{ \mathcal{L}(u) \mid u \in \mathcal{M} \text{ AND } u \in \mathcal{S} \}.$$

■

We call this is a consistent way of approximating the original problem, because by restriction the variational formulation is inherited, and as a consequence characteristic properties remain present which may otherwise be easily lost. Clearly the solution found will depend on the choice of the restriction S, and may or may not be a good approximation.

When dealing with *numerical methods*, the infinite dimensional function space is approximated by a (high-dimensional) finite space, for instance by approximating the competing functions by piecewise linear functions, as in Finite Element Methods.

Different from the direct aim to find explicit or approximate solutions, one can also use a variational structure to obtain a *simplified model* from a complicated model. The restriction will then be specified by the type of phenomena one wants to consider, thereby ignoring other phenomena

that may also be described by the original model. The simplified model will then (approximately) describe the selected type of phenomena, while maybe other phenomena are not at all, or not correctly, described. We will show below, and in Appendix B, how the very complicated full set of equations for free surface waves can be simplified by making further and further restrictions.

We will present some examples.
Example. Trial functions for BVP's

1. Consider the simple BVP

$$-\partial_x^2 u = 1, \quad \partial_x u(0) = u(1) = 0$$

which has a parabola as solution $u(x) = \frac{1}{2}(1-x)(x+1)$.
When we would ask how to approximate the solution by a linear function, which then has to be of the form

$$w(x) = a(1-x)$$

for some value a, we would not know how to do if we don't use the exact solution: inserting such function in the equation doesn't give any useful result (however, see further the Ritz-Galerkin method for a way out).
Now, however, use the exact variational formulation:

$$\text{Min} \left\{ \int_0^1 \left[(\partial_x u)^2 - 2u\right] dx \;\middle|\; u(1) = 0 \right\}$$

The correct equation for the minimizer is found if all smooth functions u with $u(1) = 0$ are taken. But the formulation does makes sense when we restrict to the linear functions above:

$$\text{Min} \left\{ \int_0^1 \left[(\partial_x w)^2 - 2w\right] dx \;\middle|\; w(x) = a(1-x) \right\}$$

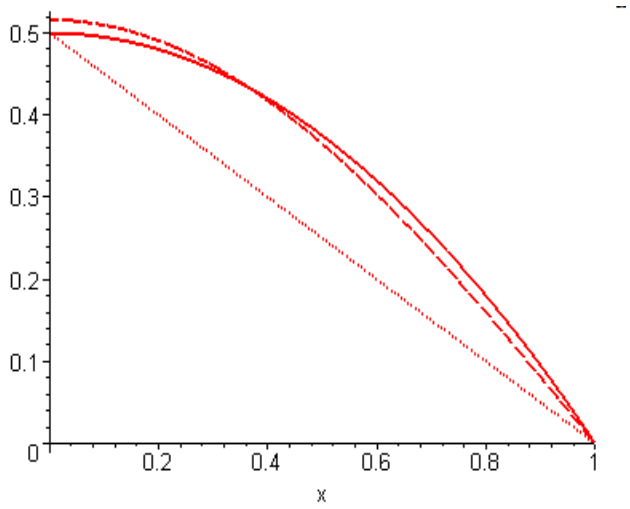
This can easily be calculated and leads to a minimization problem in the single variable a :

$$\text{Min} \quad \left\{ \int_0^1 \left[a^2 - 2a(1-x)\right] dx) \;\middle|\; a \right\} = \text{Min } \{ \; (a^2 - a) \mid a \; \}$$

with solution $a = \frac{1}{2}$. This simple example illustrates the method that can and will be used in much more complicated situations also.
An alternative would be to look for a solution in the form of a harmonic function. If we take it of the form $w(x) = b\cos(x\pi/2)$, this trial function satisfies for each parameter value b both boundary conditions. Inserting in the variational formulation leads to

$$\begin{aligned} & \text{Min} \left\{ \int_0^1 \left[(\partial_x w)^2 - 2w\right] dx) \;\middle|\; w(x) = b\cos(x\pi/2) \right\} \\ = \; & \text{Min} \quad \left\{ \int_0^1 \left[(\frac{b\pi}{2})^2 \sin^2(x\pi/2) - 2b\cos(x\pi/2) \right] dx) \;\middle|\; b \right\} \\ = \; & \text{Min } \{ \; (\frac{1}{2}(\frac{b\pi}{2})^2 - \frac{4}{\pi} b) \mid b \; \} \end{aligned}$$

attained for $b = \frac{16}{\pi^3}$.In the plot below the exact solution (bold line) is compared with the two approximations found here.



2. Now a slight variant:

$$-\partial_x^2 u = 1, \quad u(-1) = u(1) = 0.$$

for which an approximation is sought as a tent-function obeying the boundary conditions: $w(x) = a(1-|x|)$ with obvious result as above with mirror symmetry.

3. Instead of using one 'tent' function, one could use other trial functions, for instance polynomials, harmonics, etc. Or combinations of such functions. When taking combinations of many of such functions, which preferably form a basis when the number of combinations would not be limited, this becomes the area of numerical methods, which will briefly be described in the next section.

■

Accuracy of the restricted solution

By restricting the domain of the functional in the variational restriction method, the critical point of the restricted functional will in general not coincide with that of the functional on the original domain of definition. Stated differently, the restricted critical point will not satisfy the original Euler-Lagrange equation. What can we say about the quality of this restricted critical point as an approximation of the original E-L-equation? Compared to the original set of admissible variations, the restriction method restricts the admissible variations. As a consequence, the directional derivative will vanish only in the restricted

directions. Recalling the ideas of the LMR, this means that only the projection of the variational derivative (i.e. of the E-L-equation) on the restricted set will vanish.

In general notation, assume that the set is given by a parametrized family of functions:

$$S = \{ U(p) \mid p \in P \}$$

where P is the 'parameter space'; for instance, this may be the collection of Fourier coefficients in the procedure to restrict the original functions to finite Fourier-truncations. Then the admissible variations are given by

$$\frac{\partial U}{\partial p}\delta p$$

where δp are the variations in the parameters, and a critical point of the restricted functional $\mathcal{L}$ will satisfy

$$\frac{\partial}{\partial p}\mathcal{L}(U(p)) \equiv \delta\mathcal{L}(U(p); \frac{\partial U}{\partial p}\delta p) = 0.$$

This can be written with the variational derivative in a more appealing way like

$$< \delta\mathcal{L}(U(p)), \frac{\partial U}{\partial p}\delta p >= 0.$$

This shows that, instead of satisfying the full E-L-equation $\delta\mathcal{L}(u) = 0$, the restricted critical point $U(p)$ satisfies the projection in the admissible directions only. For instance, if the parameters are from a linear finite dimensional space (i.e. are 'constants'), δp can be taken out of the innerproduct and it follows (if the parameters are independent)

$$< \delta\mathcal{L}(U(p)), \frac{\partial U}{\partial p} >= 0.$$

Observe that these are just as many 'equations' as their are parameter values p to be determined in the calculation of the restricted critical point; we will return to this in the next section.

When the 'parameters' are actually functions themselves, it cannot be taken out of the inner-product and the expression means that the E-L-equation is satisfied in a weighted sense with weight determined by the admissible directions.

3.2 Variational Numerical Methods

3.2.1 General method

Let S be a finite dimensional subset of an infinite dimensional function space, usually taken as the set consisting of linear superpositions of certain base functions ϕ_k

$$S_N = \{ \Sigma_{k=1}^{N} a_k\phi_k(x) \mid a_1, ...a_N \}$$

To denote such a combination, we often refer only to the vector composed of the coefficients

$$S_N \simeq R^N = \{a | a = (a_{1,}..., a_N)\};$$

the linear combination can then be written when we think of a as a row-vector, and collect the base functions in a column vector like $a \cdot \phi(x)$.

For numerical purposes we often want 'completeness' for $N \to \infty$, meaning that each element from the original function space can be arbitrary well (in some norm) approximated by some element from S_N provided N is taken large enough.

For illustration, for functions of one variable x the following typical choices of base functions are characteristic for so-called global base functions and local base functions respectively:

- **Harmonic functions, Fourier truncation, a global method:**
  Using complex notation, and adjusting the numbering of the index slightly, the base functions are

  $$\phi_k = e^{ikx};$$

  for periodic functions on a given interval of length 2π the index is restricted to the integers, say $m \in [\ -N...N]$, for arbitrary L_2-functions the index is continuous and the summation becomes integration; this is the usual description with a Fourier-basis. Functions can then be represented as usual

  $$u(x) = \Sigma a_{m=-N}^{N} \phi_m(x), \quad \text{with....} a_m = \frac{1}{2\pi} \int u(x) e^{-imx} dx$$

- **FEM, linear elements, a local method**
  Consider an interval on which the problem is defined; make a partition of the interval, for simplicity say equidistant, with mesh-size h and nodes to be called $x_0, x_{1,} x_2, ...x_N$; the interval (x_m, x_{m+1}) is called the $(m+1)$-th element. Consider, so-called, linear splines: continuous linear functions that are confined to two successive elements (and normalized amplitude:

  $$\phi_m(x) = \begin{cases} 0 & \text{for } x \notin (x_{m-1}, x_{m+1}) \\ 1 - |x - x_m|/h & \text{for } x \in (x_{m-1}, x_{m+1}) \end{cases}$$

  For obvious reasons, such functions will be called 'tent-functions'. These form a local base (have only limited extension: any function can thus be uniquely represented by a continuous piecewise linear function, connecting the values of the function at the nodes:

  $$u(x) = \Sigma_m u_m \phi_m(x), \quad \text{with} \quad u_m = u(x_m).$$

  Intuitively, by letting $N \to \infty$ the approximation becomes better and better: the basis is complete.
  Taking the values of a function in nodes is also done in so-called collocation methods, in Finite Difference-methods. There then the behaviour of the function in between is not explicitly stated, Taylor-expansions are used to derive formulas for derivatives etc. Here, in FEM-method, the behaviour of the function in between nodes is prescribed before hand, which will be necessary to approximate integrals that appear in functionals.

Example. A functional of Sturm-Liouville type:

$$\mathcal{L}(u) = \int \left[p(x)(\partial_x u)^2 + q(x)u^2 - 2f(x)u\right] dx$$

reduces to function on R^N by restricting to functions from S_N as follows

$$L(a) := \mathcal{L}(u = \Sigma_{k=1}^N a_k \phi_k) = \Sigma_{m,k} a_m a_k P_{mk} + \Sigma_{m,k} a_m a_k Q_{mk} - 2\Sigma_m a_m F_m$$

in which we have introduced the $N \times N$ matrices P and Q

$$P_{mk} = \int p(x)(\partial_x \phi_m)(\partial_x \phi_k), \qquad Q_{mk} = \int q(x)\phi_m \phi_k$$

and the vector

$$F_m = \int f(x)\phi_m.$$

Then this can be written in matrix notation more elegantly like

$$L(a) = (P + Q)a \cdot a - 2F \cdot a$$

This function $L : R^N \to R$ is called the restriction of the functional to the finite dimensional set.
The SL-BVP results from variations of the functional

$$\text{Min}_{\ u}\ \mathcal{L}(u): \quad -\partial_x p(x)\partial_x u + qu - f = 0;$$

the corresponding equation for the vector a follows analogously from variations of the restricted function

$$\text{Min}_{\ a}\ L(a): \quad (P + Q)a - F = 0$$

In this way we have found a discretization of the differential operator on the set S_N :

$$-\partial_x p(x)\partial_x + q \longleftrightarrow P + Q$$

Observe that the matrices are symmetric (from their definition), and the symmetric operator has been consistently modelled by a symmetric matrix. The finite dimensional equation for a is called the projection of the original equation into the set S_N.
Solving the algebraic equation can be done with linear analysis-methods; however, the variational structure can also be used: a direct method to minimize the function $L(a)$. ■

Example. S-L-Eigenvalue problem
For the S-L-operator above, the constrained formulation for the eigenvalue problem

$$-\partial_x p(x)\partial_x u + q(x)u = \lambda \rho(x)u$$

reads

$$\text{Min}\ \left\{ \int \left[p(\partial_x u)^2 + qu^2\right] \ \middle|\ \int \rho u^2 = 1 \right\}$$

and reduces to

$$\text{Min } \{ (P+Q)a \cdot a \mid Ra \cdot a = 1 \}$$

where the (symmetric) matrix R is defined as $R_{mk} = \int \rho(x)\phi_k\phi_m$
The corresponding eigenvalue problem in R^N :

$$(P+Q)a = \mu Ra$$

will produce approximations for (N) eigenfunctions (forming a complete base inR^N since the eigenvalue problem is symmetric) and of the corresponding (real) eigenvalues. ■

Example. Of course, the precise form of the matrices depend on the set of base functions.
Consider the simple S-L-eigenvalue problem with $p(x) = \rho(x) = 1, q(x) = 0$ on the interval $x \in [0, \pi]$ with Dirichlet boundary values.

- Take $\phi_k(x) = \sin(kx)$. Show that the algebraic problem has only diagonal matrices and produces exactly the first N eigenfunctions and eigenvalues; in this way we reproduce the Fourier-sine-series.
- Using linear elements in FEM, leads to tri-diagonal matrices. For instance for non-boundary matrix-elements:

$$P_{mk} = \int (\partial_x\phi_m)(\partial_x\phi_k)dx = \begin{cases} -1/h & \text{for } k = m \pm 1 \\ 2/h & \text{for } k = m \\ 0 & \text{for } k \neq m, m \pm 1 \end{cases}$$

$$R_{mk} = \int \phi_m\phi_k dx = \begin{cases} h/6 & \text{for } k = m \pm 1 \\ 2h/3 & \text{for } k = m \\ 0 & \text{for } k \neq m, m \pm 1 \end{cases}$$

  In particular, the SL-operator is replaced by a three-point stencil which in an internal node reads

$$\left[-\partial_x^2 u\right]_{x_m} \leftrightarrow \frac{u_{m-1} - 2u_m + u_{m+1}}{h^2},$$

  and which is the same discretization as in FD-methods when using central difference scheme for the second derivative.
  Note, however, that the identity operator in function space is also replaced by a three-point stencil, making a certain average over 3 neighbouring points:

$$[u]_{x_m} \leftrightarrow (u_{m-1} + 4u_m + u_{m+1})/6,$$

  different from standard FD-reasoning[1].

[1] In that case, it is custom to take

$$[u]_{x_m} \leftrightarrow u_m.$$

This would correspond to the gradient of the function Σu_m^2. This can be seen as the 'discretization' of the functional $\int u^2 dx$, if the function u is approximated by a piecewise constant function! Hence, this is then NOT consistent with the reasoning to approximate with piecewise linear functions, which lead to the central difference. Stated differently: by using the discretization $[u]_{x_m} \leftrightarrow u_m$ the function Σu_m^2 is not a restriction of the quadratic functional to the set of piecewise linear functions; the functional has been changed.

■

Exercise. Study the FEM-discretization of the eigenvalue problem on an interval $(0, 2\pi)$ with periodic boundary conditions

$$-\partial_x^2 u = \lambda u;$$

write down the discrete eigenvalue problem, find the eigenvectors and eigenvalues and compare the approximations obtained with the exact results as function of the number of elements. ■

3.2.2 Projection of (variational) equations

Contemplation: If we have a certain problem, and some 'approximate solution', how do we judge the quality of that solution?

In general the answer is difficult to give. Let us write our problem abstractly like $\mathcal{E}(u) = 0$; usually a (set of) ode's or pde's with initial and or boundary values. If we can find exact solutions (rarely), we are done and can verify that these are truly solutions by simply substituting. However, if we substitute an approximate ('trial') solution, the equation will not be satisfied (precisely) and some 'residue' is left:

$$\mathcal{E}(u_{apprx}) = res$$

Intuitively, when the residue is 'small' we will expect the approximation to be good, but it is difficult to make this more precise. (What is small: in which norm, etc.) Using linearization around an exact solution, the error is related to the residue by the linearized equation as follows:

$$\begin{aligned} \mathcal{E}(u_{exact} + err) &= \mathcal{E}(u_{exact}) + \mathcal{E}'(u_{exact})err + ..., \\ \text{so } \mathcal{E}'(u_{exact})err &= res \end{aligned}$$

so, the error is small if the residue is small and the linearized operator boundedly invertible.

When we restrict a functional to find an approximate solution, the situation seems more natural. Indeed, now $\mathcal{E}(u) = \delta\mathcal{L}(u)$, say for $\mathcal{L}$ defined on infinite dimensional space. Suppose we restrict to some linear subspace S. This the defines a function $\bar{L}$ on S, the restricted function

$$\bar{L}(\bar{u}) = \mathcal{L}(\bar{u}) \text{ for } \bar{u} \in S.$$

A critical point of $\bar{L}$ satisfies $\nabla\bar{L}(u_S) = 0$, where ∇ denotes the gradient for differentiation in S. This critical point $u_S \in S$ is then an approximation of the original minimizer. The relation between $\nabla\bar{L}(u_S) = 0$ and the original equation $\delta\mathcal{L}(u) = 0$ is expressed with the first variation like:

$$\nabla\bar{L}(u_S) = 0 \iff \delta L(u_S; \zeta) = 0 \text{ for all } \zeta \in S.$$

This means that the directional derivative is zero for all admissible variations within the subspace S. For instance, when S is the span of a set of base functions $\{\varphi_j\}_{j=1}^N$ and $\bar{u} = \Sigma_j u_j \varphi_j$ then

$$\bar{L}(\bar{u}) = \mathcal{L}(\Sigma u_j \varphi_j), \quad \text{and so } \frac{\partial \bar{L}}{\partial u_j} = < \delta\mathcal{L}(\bar{u}), \varphi_j > .$$

Vanishing of the gradient $\nabla \bar{L}(u_S) = 0$ then implies that u_S satisfies the equation $\delta\mathcal{L}(u) = 0$ not in the full function space, but precisely in the directions φ_j, for $j = 1...N$:

$$< \delta\mathcal{L}(u_S), \varphi_j >= 0 \quad \text{for } j = 1...N.$$

Summarizing, *variational restriction to a subspace gives a solution within the subspace that corresponds to the vanishing of the* projection *of the equation into the subspace.*

This also gives a constructive way to find the solution: there are N equations for the N coefficients u_j, $j = 1...N$ to be determined. Given the base functions, these equations can be written down directly from the equation.

Ritz-Galerkin projection method

The idea above has been generalized to arbitrary equations, not necessarily variational equations.

Consider the general equation $\mathcal{E}(u) = 0$, and consider two sets of base functions: $\{\varphi_j\}_{j=1}^N$ and $\{\psi_j\}_{j=1}^N$. The so-called Ritz-Galerkin method is to find an approximate solution

$$\bar{u} = \Sigma_j u_j \varphi_j \text{ such that } \; < \mathcal{E}(\Sigma_j u_j \varphi_j), \psi_m >= 0 \text{ for all } m = 1..N.$$

Note that again this gives N equations for the N unknowns u_j, $j = 1...N$.

In the variational restriction method the projection is in the same subspace as the solution is sought: $\psi_j = \varphi_j$; this special case is called *Ritz'-method.* (See further Section 'Direct Optimization methods'.)

A motivation to choose different base functions for projecting the equation than the solution is that the function spaces may be very different if $\mathcal{E}$ is a differential operator; the aim of the choice ψ is to capture the main 'directions' of the equation.

3.3 Consistent modelling by restriction

3.3.1 Restriction to suitable families of functions

In the numerical methods treated above the functional was discretized by substituting for the function a linear combination of base functions. In fact, the main reason to do so comes from the completeness argument of getting better approximations of the function by taking larger superpositions. Often a more clever Ansatz is possible when some characteristic property of the solution is known in advance. For instance, if it would be known that the optimal solution is a confined function, or a quickly decreasing function on the interval, Fourier decomposition would require many modes to describe such a function. Then a profile function with a few parameters may give an approximation that can be just as good in a practical sense. The difference depends on the formulation of the set to which the functional is restricted. In the latter case, the choice of this set is important for the quality of the approximation. Besides that, usually the parameters that describe the functions to be varied do not appear linearly: the set is in general not a linear space.

We will illustrate the line of reasoning to specific examples.

At several other places in these notes we will approximate a 'hump-like' solution by simple 'tent-functions' which have as parameters (to be varied and to be determined) the amplitude and width; other trial-functions, usually more elaborate to deal with but possibly more accurate, could be taken. For instance, Gaussian functions; then the parameters appear in a nonlinear way, such as the parameter σ in the Gaussian Ae^{-x^2/σ^2}.

Nonlinear oscillator: Duffing's equation

Consider the equation for an oscillator with third order nonlinearity

$$\ddot{x} + x + x^3 = 0, \quad x(0) = \epsilon, \quad \dot{x}(0) = 0.$$

We are interested in (relatively) small amplitude solutions, so ε is small.

Phase plane analysis of this 2-nd order Lagrangian equation shows that all solutions are periodic, with period that depends on the amplitude of the solution.

Straightforward series expansion in the amplitude will lead to resonance in the third-order, which can be prevented by adjusting the frequency in Poincare-Lindeloff type of way. We will now deal with two (closely related) variational variants of this method.

The first variant starts with the observation that the equation $\ddot{x}+x+x^3 = 0$ transforms after a time scaling $\tau = \omega t \in [0, 2\pi]$ to

$$\omega^2 u'' + u + u^3 = 0.$$

This can be interpreted as a nonlinear eigenvalue problem, with ω^2 (the squared frequency) to be sought as the Lagrange multiplier. Recalling the corresponding theory, the multiplier is found as the derivative of the value function of a constrained variational problem:

$$\omega^2 = \frac{d}{d\gamma}\left[\text{Max } \left\{ \int u^2 + \frac{1}{2}u^4 \;\middle|\; \int u'^2 = \gamma \right\}\right].$$

An approximation for the minimizer and the minimizing value can be obtained by using as trial function the solution of the linear equation, i.e. $U = \varepsilon\cos(\tau)$, with amplitude related to the given constraint, i.e. $\varepsilon = \sqrt{\gamma/\pi}$. Then (using $\int_0^{2\pi}\cos^4(\tau)d\tau = 3\pi/4$) we find

$$\int U^2 + \frac{1}{2}U^4 = \gamma + \frac{1}{2}\gamma^2/\pi^2 \int \cos^4(\tau)d\tau = \gamma + \frac{3}{8\pi}\gamma^2$$

which leads to

$$\omega^2 = \frac{d}{d\gamma}\left[\gamma + \frac{3}{8\pi}\gamma^2\right] = 1 + \gamma\frac{3}{4\pi}, \quad \text{hence } \omega \approx 1 + \gamma\frac{3}{8\pi} = 1 + \frac{3}{8}\varepsilon^2.$$

In the original variables this leads to the first order solution that is asymptotically correct, given by

$$x(t) = \varepsilon\cos\left(\left[1 + \frac{3}{8}\varepsilon^2\right]t\right) + O(\varepsilon^3).$$

The, closely related, second variant is to derive this result without transforming the variables, and apply the same reasoning directly to the Lagrangian functional for the equation:

$$\int \left[\dot{x}^2 - (x^2 + \frac{1}{2}x^4)\right] dt$$

Substituting the trial function $x(t) = \varepsilon \cos \omega t$, and taking the interval of time integration to be one period, $t \in (0, 2\pi/\omega)$ there results

$$\int \left[\varepsilon^2\omega^2 \sin^2(\omega t) - \varepsilon^2 \cos^2(\omega t) - \frac{1}{2}\varepsilon^4 \cos^4(\omega t)\right] = \varepsilon^2\omega\pi - \varepsilon^2\pi/\omega - \frac{3\pi}{8}\varepsilon^4/\omega$$

Taking variations with respect to ε (or ε^2) the same result follows:

$$\omega^2 - 1 - \frac{3}{4}\varepsilon^2 = 0, \quad \text{i.e. } \omega \approx 1 + \frac{3}{8}\varepsilon^2$$

WKB-approximation

The standard second order, linear, non-autonomous ode of mathematical physics is the equation for the pendulum with slowly varying frequency, or, equivalently, the equation that describes the field of an optical pulse in a slowly varying medium:

$$\partial_z^2 u + k^2(z)u = 0.$$

For arbitrary function $k(z)$ the solutions of this equation cannot be written down explicitly. When the given inhomogeneity $k(z)$ varies 'slowly'[2] in z, a good approximation can be found; we will show this now using the variational structure of the equation.

First we remark that it is possible to look for solutions in the form of a *phase-amplitude representation*

$$u(z) = A(z)e^{i\theta(z)}.$$

Then the equation for the (real) amplitude and (real) phase can be obtained by substituting in the ode. However, it is somewhat simpler to do this from the functional. To that end observe that

$$\int \left[|\partial_z u|^2 - k^2|u|^2\right] dz = \int \left[(\partial_z A)^2 + A^2(\partial_z\theta)^2 - k^2A^2\right] dz.$$

[2] This 'slowly' vaying can be made more specific by assuming that the function $k(z)$ is actually given by

$$k(z) = K(\varepsilon z)$$

where K is a given, fixed function; then taking ε small, the function k will be slowly varying: changes of unit order in k will take place on distances $x = O(1/\varepsilon)$, which is large for small ε. Equivalently this can be seen by looking at the derivative:

$$\partial_z k(z) = \varepsilon \partial_\zeta K(\zeta) \text{ for } \zeta = \varepsilon z :$$

with $\partial_\zeta K(\zeta) = O(1)$, it follows that $\partial_z k(z) = O(\varepsilon)$. This explains why the amplitude $A(z)$ in the following is actually a function of εz and that therefore the expression $(\partial_z A)^2$ in the functional, or $\partial_z^2 A$ in the equation, is of second order and will be neglected.

The correct equations are then found (verify!!) by writing down the Euler-Lagrange equations that result from variations in A and θ.

When we assume now that the function $k(z)$ varies slowly, it is reasonable to neglect the term $\int(\partial_z A)^2$ in the functional. Then the equations become:

$$\begin{aligned}\delta_A \int \left[A^2(\partial_z\theta)^2 - k^2A^2\right] dz &= 2A\left[(\partial_z\theta)^2 - k^2\right] = 0,\\ \delta_\theta \int \left[A^2(\partial_z\theta)^2 - k^2A^2\right] dz &= -2\partial_z\left[A^2\partial_z\theta\right] = 0.\end{aligned}$$

These can be solved, for θ up to an an initial phase constant, and for the amplitude up to a multiplicative constant:

$$\theta(z) = \int^z k(\xi)d\xi, \quad A(z) = \frac{c}{\sqrt{k(z)}}$$

The corresponding solution

$$u(z) = \frac{c}{\sqrt{k(z)}} e^{\pm i\int^z k(\xi)d\xi}$$

is called the WKB-approximation (Wentzel, Kramer, Brioullin). It is a remarkably accurate approximation and describes well the main features of the exact solution.

3.3.2 Design of simplified models

Different from the direct aim to find explicit or approximate solutions, one can also use a variational structure to obtain a simplified (usually a restricted, approximate) model. For instance, in Appendix B we show in some detail how the very complicated full set of equations for free surface waves can be simplified by making further and further restrictions:

the full equations are valid for small to large amplitude waves of any wave length,

then by restricting to 'rather small, rather long waves' only – to be made more precise in order to become operational–, the full equations become the much simpler Boussinesq (type of) equations;

while Boussinesq describes waves running in both directions, a further restriction to waves running in only one direction lead to a simpler model of KdV-type;

KdV describes waves with a broad spectrum; by restricting to narrow spectra, an equation for the envelope can be derived, a NLS-type of equation.

ALL these simplified models can be most easily derived by using the variational structure. That is by restricting the functional (Hamiltonian) to smaller and smaller classes of wave phenomena. The validity of the model becomes more restricted in doing so. But the simplified model needs not to be much less accurate, provided it is used for the correct type of phenomena. This variationally consistent way of modelling assures that each more limited model retains a

variational structure, inherited from the full equations.

Example. (See Appendix B for a similar reasoning as described here to derive variationally consistent simplified models for free surface waves.)
Consider the wave equation:

$$\partial_t u = -\partial_x \delta H(u) \text{ with } H(u) = \int \left[\frac{1}{2}u^2 + \beta(\partial_x u)^2 + \gamma u^4\right] dx.$$

When restricting to small amplitude waves, the quartic term in the functional (qubic in the equation) will be neglected. Further, the resulting linear equation has dispersion, determined by the term with β. Looking at the dispersion relation, for waves of small wave length, this β-term doesn't contribute much, and we could neglect this term also.

$$\begin{aligned} & \int \left[\frac{1}{2}u^2 + \frac{1}{2}\beta(\partial_x u)^2 + \frac{1}{4}\gamma u^4\right] dx \\ \rightarrow \quad & \text{small amplitude} \quad \rightarrow \int \left[\frac{1}{2}u^2 + \frac{1}{2}\beta(\partial_x u)^2\right] dx \\ \rightarrow \quad & \text{\& long waves} \quad \rightarrow \int \left[\frac{1}{2}u^2\right] dx. \end{aligned}$$

These restrictions show itself in a different way in the simplification of the successive equations:

$$\begin{aligned} & \partial_t u = -\partial_x \left[u - \beta\partial_x^2 u + \gamma u^3\right] \\ \rightarrow \quad & \text{small amplitude} \quad \rightarrow \partial_t u = -\partial_x \left[u - \beta\partial_x^2 u\right] \\ \rightarrow \quad & \text{\& long waves} \quad \rightarrow \partial_t u = -\partial_x \left[u\right]. \end{aligned}$$

The final equation is the simplest equation, the translation equation: $\partial_t u = -\partial_x u$.
Of course, there is *no justification* in this process: we decide beforehand to what type of phenomena (type of waves) we want to restrict our attention; then we choose the restricted class accordingly.
It should be noted that the simplified model that then results may have, and will have in general, also solutions far out of the restricted set. Indeed, e.g. the translation equation has solutions of arbitrary amplitude, and very short, just as well as very long, waves. But of course, in view of the derivation, this equation, and hence its solutions, are only relevant as a simplified model (with solutions that approximate well the solutions of the original equation) when the solutions belong to the restricted class: waves of small amplitude and long wave length. ■

3.4 Direct optimization methods

One of applications of Variational Methods in Numerics is the cleverly designed numerical scheme called Conjugate Gradient Method (CGM) to solve linear systems, say $Ax = b$, where A is an $n \times n$ -matrix while x and $b \in R^n$. To simplify, without loosing the idea, here we have assumed that A is a symmetric positive definite matrix. Such linear systems are often found in the applications of

other numerical schemes such as Boundary Element Method and Finite Element Methods. We will motivate the description of CGM by first examining a scheme called the steepest descent.

3.4.1 Steepest Descent

Consider the following function

$$\phi(x) =< x, Ax > - < x, b >$$

where A is an nxn symmetric positive definite matrix, x, $b \in R^n$, $<,>$ is an inner product defined by $< x, y >= x^T y$. The equilibrium point, say x_e, of the dynamical system

$$\frac{\partial x}{\partial t} = -\nabla\phi(x)$$

satisfies $\nabla\phi(x) = Ax - b = 0$, so it is a solution of the liner system $Ax = b$. As

$$\frac{\partial \phi}{\partial t} =< \nabla\phi, \frac{\partial x}{\partial t} >=< \nabla\phi, -\nabla\phi >= -|\nabla\phi|^2 \begin{cases} = 0, \text{ if } \nabla\phi(x) = 0 \\ < 0 \text{ , otherwise} \end{cases}$$

then x_e is an isolated local minimizer of $\phi(x)$. Furthermore for a given point $x_0 \neq x_e$, that is the level set $\phi(x) = \phi(x_0)$ does not cross the equilibrium point x_e, the trajectory of x starting from x_0 always points toward x_e.

The steepest descent iterative numerical scheme for solving the linear system $Ax = b$ is designed based on the above observation, namely finding the minimizer of $\phi(x)$ starting from a point, say x_0, taking the direction of the steepest descent $-\nabla\phi(x)$. In this fashion, if $r_c = b - Ax_c = -\nabla\phi(x_c)$, called the residual at the current point x_c, the next point x_{next} to be reached from x_c is by taking the direction r_c, that is

$$x_{next} = x_c + \alpha r_c$$

where α is obtained by minimizing $\phi(x_c + \alpha r_c)$. It is not difficult to see that $\alpha = \frac{<r_c, r_c>}{<r_c, Ar_c>}$.

The scheme thus can be written briefly as follows

x_0 : initial guess
$r_0 = b - Ax_0$
$k = 0$
while $r_k \neq 0$
$k = k + 1$
$\alpha_k = \frac{<r_{k-1}, r_{k-1}>}{<r_{k-1}, Ar_{k-1}>}$
$x_k = x_{k-1} + \alpha_k r_{k-1}$
$r_k = b - Ax_k$
end

Although the method is globally convergent, for some cases when the matrix A has eigenvalues of different orders of magnitude, the convergence is very slow. The following example may provide the idea, GIVE AN EXAMPLE WITH A PICTURE!

3.4.2 Conjugate Gradient Method

Search directions

To overcome the weakness of the steepest descent, the choice of better search directions to move from one level set of ϕ to another level set will be described in the following. As a motivation to understand the concept, let first consider a simple two dimensional linear system written in the matrix form $Ax = b$, where $A = (a_{ij})$, $i, j = 1, 2$ is a 2×2-matrix, while x and $b \in R^2$. Let x_0 be the initial guess and let the first direction $p_1 = r_0$, the residue at $x = x_0$. Let $x_1 = x_0 + \alpha_1 p_1$ be the next point where $\alpha_1 = \frac{<p_1, r_0>}{<p_1, Ap_1>}$, and let r_1 be the residue at $x = x_1$, then $< p_1, r_1 >= 0$ (show this!). If x_e is the exact solution the system, then $Ax_e = b$, then

$$< x_e - x_1, Ap_1 >=< Ax_e - Ax_1, p_1 >=< b - Ax_1, p_1 >=< r_1, p_1 >= 0$$

This shows that $x_e - x_1$ perpendicular to Ap_1 or $x_e - x_1 \in \{Ap_1\}^{\perp}$. To move from the current point x_1 to the exact solution x_e, one must take a search direction that lies on $\{Ap_1\}^{\perp}$. The following picture illustrates this concept.

PICTURE!

To generalize the idea to an n dimensional space, we will employ the following properties.

Proposition 49 *Let A be an $n \times n$ symmetric positive definite matrix, $x, b \in R^n$, and $\phi(x) =< x, Ax > - < x, b >$ .Let x_0 be an initial guess and $\{p_1, p_2, ..., p_k\}$ be the first k general search directions, that is $x_k = x_0 + \sum_{i=1}^{k} \alpha_i p_i$, for some $\alpha_i, i = 1, 2, ..., k$.If $\alpha_i, i = 1, 2, ...k$ are chosen such that x_k minimizes $\phi(x)$ and $r_k = b - Ax_k$ then $< p_i, r_k >= 0$, $i = 1, 2, ..., k$.*

Proof. $\phi(x_k) = \phi(x_0) + < \sum_{i=1}^{k} \alpha_i p_i, Ax_0 > + \frac{1}{2} < \sum_{i=1}^{k} \alpha_i p_i, \sum_{i=1}^{k} A\alpha_i p_i > - < \sum_{i=1}^{k} \alpha_i p_i, b >$.

Then $0 = \frac{\partial \phi}{\partial \alpha_i} =< p_i, Ax_0 > + < p_i, \sum_{i=1}^{k} \alpha_i p_i > - < p_i, b >=< p_i, A(x_0 + \sum_{i=1}^{k} \alpha_i p_i) - b >=< p_i, Ax_k - b >=< p_i, r_k >$, $i = 1, 2, ..., k$ ■

From this property, if x_e is the exact solution of $Ax = b$, it follows that for $i = 1, 2, ...k$,

$$< x_e - x_k, Ap_i >=< Ax_e - Ax_k, p_i >=< b - Ax_k, p_i >=< r_k, p_i >= 0$$

that is $x_e - x_k$ perpendicular to each of $Ap_i, i = 1, 2, ..., k$ or $x_e - x_k \in \{Ap_1, Ap_2, ..., Ap_k\}^{\perp}$. It is then making sense to take the next search direction $p_{k+1} \in \{Ap_1, Ap_2, ..., Ap_k\}^{\perp}$ to move from x_k on the level set $\phi(x) = \phi(x_k)$ to the next point x_{k+1}. Here, $x_{k+1} = x_k + \alpha_{k+1} p_{k+1}$, where α_{k+1} is chosen such that x_{k+1} minimizes $\phi(x)$.

The following proposition shows the possibility of designing an iterative scheme for solving $Ax = b$ that converges at most for n steps.

Proposition 50 *Let $\{p_1, p_2, ..., p_k\}$ be the first k general search directions and $p_{k+1} \in \{Ap_1, Ap_2, ..., Ap_k\}^{\perp}$. Let $x_{k+1} = x_0 + \sum_{i=1}^{k} \alpha_i p_i + \alpha_{k+1} p_{k+1}$ be a minimizer of $\phi(x)$ for $\alpha_i, i = 1, 2, ..., k, k+1$. Then $\alpha_i, i = 1, 2, ..., k$ can be obtained from minimizing $\phi(x_0 + \sum_{i=1}^{k} \alpha_i p_i)$ independently from α_{k+1}. Having obtained $\alpha_i, i = 1, 2, ..., k$ and so defining $x_k = x_0 + \sum_{i=1}^{k} \alpha_i p_i$, $\alpha_{k+1} = \frac{<p_{k+1}, r_k>}{<p_{k+1}, Ap_{k+1}>}$, where $r_k = b - Ax_k$.*

Proof. First observe that $\phi(x_{k+1}) = \phi(x_0 + \sum_{i=1}^{k} \alpha_i p_i + \alpha_{k+1} p_{k+1})$
$= \phi(x_0 + \sum_{i=1}^{k} \alpha_i p_i) + \alpha_{k+1} < p_{k+1}, \sum_{i=1}^{k} \alpha_i Ap_i > + \frac{1}{2}\alpha_{k+1}^2 < p_{k+1}, Ap_{k+1} > -\alpha_{k+1} < p_{k+1}, r_0 > .$
Since $p_{k+1} \in \{Ap_1, Ap_2, ..., Ap_k\}^{\perp}$ then $< p_{k+1}, \sum_{i=1}^{k} \alpha_i Ap_i >= 0$. Thus minimizing $\phi(x_{k+1})$ leads to two independent minimizations $\text{Min}\ \phi(x_{k+1}) = \text{Min}_{\alpha_i, i=1,2,..,k}$ $\phi(x_0 + \sum_{i=1}^{k} \alpha_i p_i) + \text{Min}_{\alpha_{k+1}}\ \frac{1}{2}\alpha_{k+1}^2 < p_{k+1}, Ap_{k+1} > -\alpha_{k+1} < p_{k+1}, r_0 > .$From the second part, we obtain $\alpha_{k+1} = \frac{<p_{k+1}, r_0>}{<p_{k+1}, Ap_{k+1}>}$. We again use the fact that $p_{k+1} \in \{Ap_1, Ap_2, ..., Ap_k\}^{\perp}$ to obtain $< p_{k+1}, r_k >=< p_{k+1}, b - Ax_k >=< p_{k+1}, b - A(x_0 + \sum_{i=1}^{k} \alpha_i p_i) >=< p_{k+1}, b - Ax_0 >=< p_{k+1}, r_0 >$ giving $\alpha_{k+1} = \frac{<p_{k+1}, r_k>}{<p_{k+1}, Ap_{k+1}>}$ ■

Given $\{p_1, p_2, ..., p_k\}$ the first k general search directions, the next question is how to choose $p_{k+1} \in \{Ap_1, Ap_2, ..., Ap_k\}^{\perp}$. The idea is find the right search direction that minimizes the resids. Thus to combine the above scheme with the steepest descent. Here, p_{k+1} is chosen to be the orthogonal projection or $r_k = b - Ax_k$ into $\{Ap_1, Ap_2, ..., Ap_k\}^{\perp}$. In other word $p_{k+1}$is a minimizer of $||p - r_k||_2$, $p \in \{Ap_1, Ap_2, ..., Ap_k\}^{\perp}$. It can be shown that such p_{k+1}can be written as

$$p_{k+1} = r_k + \beta_k p_k$$

where $\beta_k = -\frac{<p_k, Ar_k>}{<p_k, Ap_k>}$.

CGM-Algortihm

The above description leads to an iterative scheme as follows.

x_0 : initial guess
$r_0 = b - Ax_0$
$p_1 = r_0$
$\alpha_1 = \frac{<p_1, r_{k-1}>}{<p_1, Ap1>}$
$k = 0$
while $r_k \neq 0$
$k = k + 1$
$x_k = x_{k-1} + \alpha_k p_k$
$r_k = b - Ax_k$

$$\beta_k = -\frac{<p_k, Ar_k>}{<p_k, Ap_k>}.$$
$$p_{k+1} = r_k + \beta_k p_k$$
$$\alpha_{k+1} = \frac{<p_{k+1}, r_k>}{<p_{k+1}, Ap_{k+1}>}$$

end

Appendix A

Variational Optics

A.1 Basic equations

A.1.1 Macroscopic Maxwell Equations

The Macroscopic Maxwell Equations (MME) in a medium without free charges are given in its standard form by

$$\partial_t \begin{pmatrix} \mathbf{D} \\ \mathbf{B} \end{pmatrix} = \begin{pmatrix} 0 & curl \\ -curl & 0 \end{pmatrix} \begin{pmatrix} \mathbf{E} \\ \mathbf{H} \end{pmatrix}$$

where the basic electromagnetic fields are

$\mathbf{E}$: electric field
$\mathbf{H}$: magnetic field

and the variables

$\mathbf{D}$: dielectric displacement
$\mathbf{B}$: magnetic induction

are expressed in $\mathbf{E}, \mathbf{H}$ by so-called *constitutive relations:*

- for propagation in vacuum, $\mathbf{D} = \varepsilon_0 \mathbf{E}, \mathbf{B} = \mu_0 \mathbf{H}$ with ε_0, μ_0constant ($\varepsilon_0 \mu_0 = \frac{1}{c^2}$ with c the speed of light in vacuum);
- for propagation in material, polarization effects are present because of interaction of fields with molecules and electrons; in these lectures we will assume that the magnetic susceptibility vanishes at the relevant optical frequencies, in which case one has

  $$\begin{aligned} \mathbf{D} &= \varepsilon_0 \mathbf{E} + \mathbf{P}(\mathbf{E}) \\ \mathbf{B} &= \mu_0 \mathbf{H} \end{aligned}$$

  with polarization $\mathbf{P}$ depending on $\mathbf{E}$ in a way determined by the material properties.

- For lossless materials, to which we will restrict in the following, the constitutive relations can be formulated using constitutive functionals[1]. In particular, a functional $\mathcal{H}$ of $\mathbf{E}, \mathbf{H}$ can be found such that

$$\mathbf{D} = \delta_{\mathbf{E}}\mathcal{H} \ , \ \ \mathbf{B} = \delta_{\mathbf{H}}\mathcal{H} \text{ with } \mathcal{H} = \mathcal{C}(E) + \int \frac{1}{2}\mu_0 \mathbf{H}\cdot\mathbf{H}.$$

  For instance, in vacuum, the constitutive functional $\mathcal{C}(E)$ on a domain Ω reads

$$\mathcal{C}(\mathbf{E}) = \int \frac{1}{2}\varepsilon_0 \mathbf{E}\cdot\mathbf{E}$$

- As a consequence of the variational structure of the constitutive relations, and the fact that (with suitable boundary conditions) the matrix operator $\Gamma = \begin{pmatrix} 0 & curl \\ -curl & 0 \end{pmatrix}$ is skew-symmetric (since $curl$ is symmetric), Maxwell equations can be written down in the following variational form[2]:

$$\partial_t \delta\mathcal{H} = \Gamma\delta\mathcal{E} \text{ with } \mathcal{E} = \int \frac{1}{2}(\mathbf{E}\cdot\mathbf{E} + \mathbf{H}\cdot\mathbf{H}).$$

Monochromatic light

In many cases one is interested to investigate time harmonic solutions (often called CW: Continuous Waves, in the physics literature), with frequency ω that may be prescribed or to be found. Then it is custom to exploit complex notation and write fields like $\mathbf{E} = \frac{1}{2}\hat{\mathbf{E}}e^{-i\omega t} + cc$, where here and in the following, cc denotes 'complex conjugate'. Solutions of this type can only be expected to exist provided the polarization of a time-harmonic field is purely harmonic with the same frequency. Then the equations become

$$-i\omega \begin{pmatrix} \hat{\mathbf{D}} \\ \hat{\mathbf{B}} \end{pmatrix} = \begin{pmatrix} 0 & curl \\ -curl & 0 \end{pmatrix} \begin{pmatrix} \hat{\mathbf{E}} \\ \hat{\mathbf{H}} \end{pmatrix}$$

which can be written by eliminating the magnetic field like

$$-\omega^2\mu_0\hat{\mathbf{D}} = curl\ curl\hat{\mathbf{E}}.$$

The variational formulation is then retained by using the related constitutive functional:

$$\delta\left[\int |curl\hat{\mathbf{E}}|^2 + \omega^2\mu_0\mathcal{C}(\hat{\mathbf{E}})\right] = 0.$$

[1] We do not specify here whether these functionals are defined as integrals over the spatial domain or as integrals over the time. We will see below in the case of one spatial dimension that a time-integration may be most natural.

[2] As a dynamical system evolving in time, this is of the form of a Poisson system when the functionals involved are given by integrations over the spatial domain. When they are given by integrals over time, the 'dynamic' interpretation is different, but a variational structure is present. This can be seen by formally writing the equations like

$$\partial_t^{-1}\Gamma \begin{pmatrix} \mathbf{E} \\ \mathbf{H} \end{pmatrix} = \delta\mathcal{H}$$

and, observing that in space-time the operator $\partial_t^{-1}\Gamma$ is symmetric, writing down the Lagrangian for this equation.

A.1.2 Restriction to 2 spatial dimensions

In the following we will restrict to two-dimensional (2D) spatial problems (or to 1D). We will think of structures and variables independent of y, and light propagation in the z-direction. Then the total set of equations for the six field components decouple into two sets of equations for three components only, a splitting is so-called TE-modes (transverse electric) and TM-modes (transverse magnetic):

$$\begin{aligned} \text{TE-case} &: \quad \mathbf{E} = (0, E_y, 0), \quad \mathbf{H} = (H_x, 0, H_z) \\ \text{TM-case} &: \quad \mathbf{E} = (E_x, 0, E_z), \quad \mathbf{H} = (0, H_y, 0) \end{aligned}$$

Restricting to the TE-case, and assuming that the polarization has also only its y-component non-vanishing, MME's become

$$\begin{aligned} \partial_t D_y &= \partial_z H_x - \partial_x H_z \\ \mu_0 \partial_t H_x &= \partial_z E_y \\ \mu_0 \partial_t H_z &= -\partial_x E_y. \end{aligned}$$

These equations can be reduced to a scalar equation for $E \equiv E_y$, with $D \equiv D_y$, the sME (scalar Maxwell Equation):

$$\mathbf{sME}: \quad \mu_0 \partial_t^2 D = \Delta E \equiv \left(\partial_x^2 + \partial_z^2\right) E;$$

in vacuum this leads to the standard wave equation: $\partial_t^2 E = c^2 \Delta E$.

For monochromatic light there results the Helmholtz equation:

$$-\omega^2 \mu_0 \hat{D} = \Delta \hat{E}$$

with variational formulation

$$\delta \left[\int |\nabla \hat{E}|^2 + \omega^2 \mu_0 \mathcal{C}(\hat{E}) \right] = 0.$$

A.1.3 Restriction to 1 spatial dimension

With further restriction, uniformity in the x and y-direction, a further simplification is obtained: the MME's become

$$\partial_t D_y = \partial_z H_x, \quad \mu_0 \partial_t H_x = \partial_z E_y \tag{A.1}$$

and hence

$$\begin{aligned} \text{sME} &: \quad \mu_0 \partial_t^2 D = \partial_z^2 E \\ \text{Helmholtz} &: \quad -\mu_0 \omega^2 \hat{D} = \partial_z^2 \hat{E} \end{aligned}$$

A.1.4 Bidirectional equation for pulse propagation

When the Maxwell equations are restricted to depend on one spatial direction only, the z-direction as above, there result equations for the y-component of the **E**-field and the x-component of the **H**-field; assuming that also the electric polarization has only its y-component non-vanishing, and restricting to non-magnetic materials, the equations (A.1) can be written as a bidirectional equation like

$$\partial_z \begin{pmatrix} E \\ H \end{pmatrix} = \begin{pmatrix} 0 & \partial_t \\ \partial_t & 0 \end{pmatrix} \begin{pmatrix} D \\ \mu_0 H \end{pmatrix} \tag{A.2}$$

which can also be written as the second order scalar equation

$$\partial_z^2 E = \mu_0 \partial_t^2 D$$

In the following we consider lossless material with linear dispersion given by $\hat{\varepsilon}_1(\omega)$ and non-dispersive quadratic and/or qubic nonlinearity[3]; then the dielectric displacement is given by

$$D = \varepsilon_o E + \varepsilon_1 * E + \chi_2 E^2 + \chi_3 E^3$$

and can be written as the variational derivative with respect to E of the constitutive functional[4]

$$\mathcal{C}(E) = \int \left[\frac{1}{2} \left(\varepsilon_0 E^2 + \varepsilon_1 * E \cdot E \right) + \frac{1}{3} \chi_2 E^3 + \frac{1}{4} \chi_3 E^4 \right] dt.$$

The linear dispersion relation for modes $e^{i[kz-\omega t]}$ has two solution branches

$$k = \pm K(\omega) \text{ with } K(\omega) \equiv \frac{\omega}{c} R(\omega) \equiv \frac{\omega}{c} \sqrt{1 + \varepsilon_1(\omega)/\varepsilon_0}$$

with $K(\omega)$ real-valued and skew symmetric for real frequencies. Introducing the 'Hamiltonian'

$$\mathcal{H} = \mathcal{C}(E) + \int \frac{1}{2} \mu_0 H^2 dt$$

the equations can be written as a Hamiltonian system evolving in z as follows

$$\partial_z \begin{pmatrix} E \\ H \end{pmatrix} = \begin{pmatrix} 0 & \partial_t \\ \partial_t & 0 \end{pmatrix} \begin{pmatrix} \delta_E \mathcal{H} \\ \delta_H \mathcal{H} \end{pmatrix}.$$

Using the analogy with wave propagation in fluid dynamics, this variational structure of the equations can be exploited to derive simplified models that describes the envelope equation for waves propagating in one direction. Without derivation, we simply state the results in the next subsections.

A.1.5 Unidirectional Maxwell equation

For weakly dispersive, non-linear equations, as considered here, in a good approximation a splitting can be made between right and left travelling waves. Following the unidirectionalization process described in detail in Van Groesen & De Jager [13], the result is the following unidirectional *Maxwell equation (uni-ME)*

$$\partial_z E + \frac{1}{c} \partial_t \left[R(i\partial_t) E + \tilde{\chi}_2 E^2 + \tilde{\chi}_3 E^3 \right] = 0 \tag{A.3}$$

where we use the notation $\tilde{\chi}_{2,3} = \chi_{2,3}/2\varepsilon_0$.

[3] Actually, the following can be generalised in many ways: higher order dispersion, dispersion in nonlinear terms, higher order non-linearity; only the lossless character is of importance which implies the existence of a constitutive potential for the dielectric displacement. With the same assumption for the magnetic polarization, magnetic properties can be included as well.

[4] Note that here the constitutive functional is given pointwise in space as an integration over time.

The linear part corresponds to the right-travelling branch of the dispersion relation, $k = K(\omega)$, and can be written like $(\partial_z + iK(i\partial_t))\, E = 0$. This can be seen easily from the linear bidirectional equation by writing it as

$$(\partial_z - iK(i\partial_t))\,(\partial_z + iK(i\partial_t))\, E = 0;$$

this also shows that an exact splitting between right- and left-travelling waves is possible.

The unidirectional equation has inherited the Hamiltonian structure of the bidirectional equation, as can be seen by writing

$$\partial_z E = -\frac{1}{c}\partial_t \bar{\mathcal{H}} \text{ with } \bar{\mathcal{H}} = \int \left[\frac{1}{2}RE\cdot E + \frac{1}{3}\tilde{\chi}_2 E^3 + \frac{1}{4}\tilde{\chi}_3 E^4\right] dt. \tag{A.4}$$

The corresponding magnetic field is for this unidirectional propagation given by $H = -\sqrt{\frac{\varepsilon_0}{\mu_0}}E$.

Remark. In the theory of surface waves on a layer of fluid, a similar equation (with z and t interchanged, and for long-wave dispersion approximated by $R = 1 + \partial_t^2$) is known as the Korteweg - de Vries (KdV) equation; it describes unidirectional surface waves in a remarkably good approximation. ■

A.1.6 NLS Envelope equation for pulse propagation

We now present the equation for the envelope of a wave group centered at a central frequency $\bar{\omega}$. The resulting wave group is a modulation of a harmonic mode, represented by a complex-valued amplitude A :

$$u(z,t) = A(z,t)e^{i\bar{\theta}} + cc, \text{ with } \bar{\theta} = K(\bar{\omega})z - \bar{\omega}t.$$

The equation for the amplitude is an NLS-type of equation. To get it in an attractive form, it is custom to eliminate the first order term in the dispersion by introducing a frame moving with the group velocity $1/K'(\bar{\omega})$, i.e. $\tau = t - K'(\bar{\omega})z, \zeta = z$, and to approximate the dispersion relation by a quadratic polynomial at $\bar{\omega}$:

$$K(\bar{\omega}+\nu) \approx K(\bar{\omega}) + K'(\bar{\omega})\nu + \beta\nu^2 \text{ with } \beta = \frac{1}{2}K''(\bar{\omega}).$$

and make the following scaling (restricting for simplicity to the simplest case of qubic nonlinearity[5], $\chi_2 = 0$ and $\chi_3 \neq 0$)

$$z^* = z/c \text{ and } u = \sqrt{\tilde{\chi}_3}E.$$

Then the NLS-equation is obtained:

$$\text{NLS: } \partial_\zeta A + i\beta\partial_\tau^2 A + i\gamma|A|^2 A = 0. \tag{A.5}$$

Again, this equation has variational structure: it is a (complex) Hamiltonian system (evolving in space):

$$\partial_\zeta A = i\delta H(A), \text{ with } H(A) = \int \left[\frac{1}{2}\beta|\partial_\tau A|^2 - \frac{1}{4}\gamma|A|^4\right] d\tau$$

[5] For quadratic nonlinearity, the third order resonance appears through interaction of first with second order effects. The interaction coefficient γ is then much more complicated. See [?, ?, ?] for more details.

Remark. The coefficients γ and β depend on $\bar{\omega}$ and the sign of χ_3; a simple scaling transforms this equation to the normalized form

$$\partial_\zeta A + i\partial_\tau^2 A + i\text{sign}(\beta\gamma)|A|^2 A = 0.$$

Different signs lead to equations with essentially different properties:

CNLS: sign $(\beta\gamma) = 1$, Converging NLS
DNLS: sign $(\beta\gamma) = 1$, Diverging NLS

For instance, for CNLS ($\gamma < 0$ and anomalous dispersion, i.e. $K''(\bar{\omega}) < 0$) soliton solutions exist, but this is not the case for DNLS. ■

The NLS-equations are well known in optics and have been studied extensively; see e.g. [1, 15, 20]

A.1.7 Spatial 2D NLS

Consider the Nonlinear Helmholtz equation in a plane medium with Kerr-nonlinearity (third order: χ_3)

$$\Delta E + \omega^2 \left[n^2 + \chi_3|E|^2\right] E = 0$$

It is then quite common to look for a beam propagating in the z-direction with amplitude-variations in the transversal x-direction and 'slowly' varying in z: Substituting for real propagation constant[6] β and complex-valued amplitude A the Ansatz

$$E = A(x,z)e^{i\beta z}$$

there results

$$\partial_z^2 A + 2i\beta\partial_z A + \partial_x^2 A - \beta^2 A + \omega^2 \left[n^2 + \chi_3|A|^2\right] A = 0$$

The assumed 'slow' variations in the z-direction is exploited by neglecting the second order derivative $\partial_z^2 A$ which then leads to the equation

$$2i\beta\partial_z A + \partial_x^2 A - \beta^2 A + \omega^2 \left[n^2 + \chi_3|A|^2\right] A = 0$$

Assuming the index to be constant (for simplicity), by taking $\beta^2 = \omega^2 n^2$ this simplifies to an NLS-equation in the form

$$2i\beta\partial_z A + \partial_x^2 A + \omega^2\chi_3|A|^2 A = 0, \tag{A.6}$$

which can be further rewritten to normalized form when desired.

A.2 Optical waveguide modes

A.2.1 Preliminaries

Consider a wave guide of width $2w$ in the x-direction. For simplicity, we will consider an index that is symmetric around the z-axis. Although more general

[6] It is custom to use the notation β for the propagation constant, and we follow this custom. Note, however, that this parameter is different from the parameter $\beta = \frac{1}{2}K''(\omega)$ that we used in the previous subsection on pulse propagation.

index variations can be taken, we will consider in particular the case (that can be analyzed most easily) that the index of refraction is a step function

$$n(x) = n_0 + (n_1 - n_0)\chi_{(-w,w)} = \begin{cases} n_1 > n_0 & \text{for } |x| < w \\ n_0 & \text{for } |x| > w \end{cases} .$$

For the TE-case, with time-harmonic electric field, $E = u(x,z)e^{-i\omega t}$ Helmholtz equation for the space dependent field u reads

$$\Delta u + \omega^2 n(x)^2 u = 0.$$

The discontinuity in the index requires to find a 'weak' solution, i.e. a solution that satisfies the interface conditions

$$u \text{ and } \partial_x u \text{ continuous at } x = \pm w.$$

Then, Fourier transformation with respect to z leads one to look for solutions with harmonic z dependence:

$$u(x,z) = \phi(x)e^{i\beta z}.$$

The value β is usually called the 'propagation-constant'; it is the wave number of the travelling wave in the z-direction of the E field:

$$E = \phi(x)e^{i(\beta z - \omega t)}$$

Then the problem for the profile function ϕ becomes

$$\partial_x^2 \phi + \left(\omega^2 n(x)^2 - \beta^2\right)\phi = 0 \tag{A.7}$$

with corresponding interface conditions for ϕ

$$\phi \text{ and } \partial_x \phi \text{ continuous at } x = \pm w.$$

For a step-index the solutions of this problem can be found explicitly. Indeed, both within and outside the wave guide, the index is constant, and the ode for ϕ is a simple equation for which the solution can be written down explicitly. Then the interface conditions should match the interior solution with the outer solution to obtain a valid solution on the whole real line.

In more detail the calculation is as follows.

For constant n the solutions of (A.7) depend on the sign of $\omega^2 n^2 - \beta^2$. If positive, say $k^2 := \omega^2 n^2 - \beta^2 > 0$, the the solutions are harmonic: $\phi = A\cos(kx) + B\sin(kx) = ae^{ikx} + be^{-ikx}$, while for negative values, say $\rho^2 = \beta^2 - \omega^2 n^2$ the solutions are exponentials: $\phi = C\cosh(\rho x) + D\sinh(\rho x) = ce^{\rho x} + de^{-\rho x}$.

In the following we will restrict to the case that the solutions we are looking for should vanish at infinity; these are so-called *guided modes*, defined by the requirement

$$\phi(x) \to 0 \quad \text{for } |x| \to \infty.$$

[[Solutions that are periodic (in x) are called radiation modes, and are referred to as non-guided modes: the light is not 'confined' to, not guided by, the waveguide .]]

Collecting these pieces, we conclude that we can expect guided modes only if the value of β satisfies

$$\omega^2 n_1^2 > \beta^2 > \omega^2 n_0^2.$$

Then the solution will be harmonic in the interior, and exponential outside. Using symmetry, and an (arbitrary) normalization to unity at $x = 0$, the solution in the two regions can be written like

$$\phi(x) = \begin{cases} \cos\left(\sqrt{\omega^2 n_1^2 - \beta^2}\, x\right) & \text{for } 0 < x < w \\ A \exp\left(-\sqrt{\beta^2 - \omega^2 n_0^2}(x - w)\right) & \text{for } x > w \end{cases} .$$

The parameter A is arbitrary as yet, but has to be chosen to satisfy the interface conditions. Requiring continuity, leads to the value $A = \cos\left(\sqrt{\omega^2 n_1^2 - \beta^2}\, w\right)$. To satisfy the continuity of the first derivative requires:

$$\begin{aligned} \sqrt{\omega^2 n_1^2 - \beta^2} \sin\left(\sqrt{\omega^2 n_1^2 - \beta^2}\, w\right) &= A\sqrt{\beta^2 - \omega^2 n_0^2}, \\ \text{i.e. } \sqrt{\omega^2 n_1^2 - \beta^2} \tan\left(\sqrt{\omega^2 n_1^2 - \beta^2}\, w\right) &= \sqrt{\beta^2 - \omega^2 n_0^2}, \end{aligned}$$

which can be written using $\lambda = \sqrt{\omega^2 n_1^2 - \beta^2}$ like

$$\lambda \tan(\lambda w) = \sqrt{\omega^2 \left(n_1^2 - n_0^2\right) - \lambda^2}.$$

Since solutions of this transcendental equation cannot be written down explicitly, we rely on graphical presentation. Plotting the graphs of the right- and left-hand side as function of λ shows that there is at least one solution λ, and hence at least one value of β for which both interface conditions are satisfied, and hence a physical mode profile is obtained.

Stated differently, the problem (A.7) for guided modes is an eigenvalue problem, with ϕ the eigenfunctions to be sought and β^2 the eigenvalues. The possible values of β^2 are discrete and number of modes will depend on the width w and the index difference $n_1 - n_0$. There is always at least one mode, which corresponds to the largest value of β^2; this will be called the *principal mode*, and is symmetric and sign-definite (say positive). For simplicity of exposition, we will restrict the presentation in the following to this principal mode.

Exercise. Investigate the transcendental equation to find the number of modes as depending on the width w and the index-difference. Study in particular the limiting cases, for instance at fixed index-difference, the cases $w \to 0$ and $w \to \infty$. Make a plot (maple) of the principal value β as function of the width. ■

Exercise. Investigate radiation modes: the continuum of solutions (bounded, but non-vanishing at infinity) that correspond to values of $\beta \in (0, \omega n_0)$ (the continuous part of the spectrum). Give also the physical interpretation of these solutions. ■

A.2.2 Variational formulation for guided modes with Transparent BC's

Direct formulation on the unbounded domain

For general index variation, the eigenvalue problem, including the required interface conditions at places where the index has jumps, can be formulated with Rayleigh's quotient or the constrained formulation. When the index profile is assumed to be symmetric (as above), for the normalized, symmetric, principal eigenfunction we can formulate the problem on the positive half-line and the symmetry-boundary condition at $x = 0 : \partial_x \phi = 0$. The variational formulation is then as follows (verify!!):

$$-\beta^2 = \min_{\phi} \left\{ \int_0^\infty \left[\partial_x \phi^2 - \omega^2 n^2 \phi^2\right] dx \;\middle|\; \int_0^\infty \phi^2 = 1 \ , \ \phi \to 0 \ \text{ for } x \to \infty \right\} \tag{A.8}$$

Confined formulation using Transparent Boundary Conditions (TBC)

The problem above is formulated on the unbounded domain. Certainly for numerical methods this is a problem and one would like a confined formulation. We will now derive this formulation directly from the general formulation, but using the knowledge of the solution of the exterior problem. The arguments to find the formulation proceed as follows for the case of a arbitrary index variation within the wave guide of width w, and uniform outside with index n_0.

Take any point outside the wave guide, say $x = B \geq w$; we will split the interval $[0, \infty)$ in a bounded interior and unbounded exterior: $[0, \infty) = [0, B] \cup [B, \infty)$. In the exterior domain we use the fact that we know the solution if the eigenvalue, say $\tilde{\beta}$, and the value Φ of the field at the point $x = B$ would be known. Then we match this exterior solution to the yet unknown solution ψ in the interior, requiring only continuity of the solution at $x = B$, hence we put $\Phi = \psi(B)$. So, trial functions are taken to be of the form:

$$\phi(x) = \begin{cases} \psi(x) & \text{for } x \in [0, B] \\ \psi(B) \exp(-\sqrt{\tilde{\beta}^2 - \omega^2 n_0^2}(x - B)) & \text{for } x > B. \end{cases}$$

The part of the integral over the exterior domain $\int_B^\infty \left[\partial_x \phi^2 - \omega^2 n^2 \phi^2\right] dx$ is first reduced by partial integration and then by using the fact that the function satisfies the correct equation there. This leads to

$$\begin{aligned} \int_B^\infty \left[\partial_x \phi^2 - \omega^2 n_0^2 \phi^2\right] dx &= \int_B^\infty \left[-\partial_x^2 \phi - \omega^2 n_0^2 \phi\right] \phi dx - \left[\phi \partial_x \phi\right]_{x=B} \\ &= -\tilde{\beta}^2 \int_B^\infty \phi^2 dx + \psi(B)^2 \sqrt{\tilde{\beta}^2 - \omega^2 n_0^2} \end{aligned}$$

With the normalization $1 = \int_0^\infty \phi^2 = \int_0^B \psi^2 + \int_B^\infty \phi^2$ we arrive for the integral

over the total domain at

$$\begin{aligned}\int_0^\infty \left[\partial_x\phi^2 - \omega^2 n^2\phi^2\right] dx &= \int_0^B \left[\partial_x\psi^2 - \omega^2 n^2\psi^2\right] dx \\ &\quad +\psi(B)^2\sqrt{\tilde\beta^2 - \omega^2 n_0^2} - \tilde\beta^2\left(1 - \int_0^B \psi^2\right) \\ = \int_0^B \left[\partial_x\psi^2 + (\tilde\beta^2 - \omega^2 n^2)\psi^2\right] dx &+ \psi(B)^2\sqrt{\tilde\beta^2 - \omega^2 n_0^2} - \tilde\beta^2\end{aligned}$$

With this result we can transform the original constrained eigenvalue formulation to the following unconstrained formulation, where now the 'variables' to be varied are the function ψ on the confined interval, and the unknown parameter $\tilde\beta$:

$$\begin{aligned}&\min_\phi \left\{ \int_0^\infty \left[\partial_x\phi^2 - \omega^2 n^2\phi^2\right] \;\middle|\; \int_0^\infty \phi^2 = 1 \ , \ \phi \to 0 \ \text{ for } x \to \infty \right\} = \\ &\min_{\psi,\tilde\beta} \left\{ \int_0^B \left[\partial_x\psi^2 + (\tilde\beta^2 - \omega^2 n^2)\psi^2\right] dx + \psi(B)^2\sqrt{\tilde\beta^2 - \omega^2 n_0^2} - \tilde\beta^2 \right\}\end{aligned} \tag{A.9}$$

Observe that the correct Euler-Lagrange equation and the natural boundary condition at $x = 0$ are found:

$$\begin{aligned}\partial_x^2\psi + \omega^2 n^2\psi &= -\tilde\beta^2\psi \text{ for } x \in (0, B) \\ \partial_x\psi &= 0 \text{ at } x = 0\end{aligned} \tag{A.10}$$

Moreover, at $x = B$, variations of the free end value $\psi(B)$ leads to the natural boundary condition

$$\partial_x\psi = -\sqrt{\tilde\beta^2 - \omega^2 n_0^2}\,\psi \text{ at } x = B_-. \tag{A.11}$$

This is a boundary condition for the 'interior' solution (indicated by writing B_-). For the exterior solution a similar relation holds:

$$\partial_x\phi = -\sqrt{\tilde\beta^2 - \omega^2 n_0^2}\,\phi \text{ at } x = B_+.$$

Since we required $\phi(B) = \psi(B)$, it follows that also the derivatives are the same:

$$\phi(B) = \psi(B) \text{ and } \partial_x\phi(B) = \partial_x\psi(B).$$

This means that the interior and the exterior solution are matched to make one genuine solution on the whole real line since both interface conditions at $x = B$ are satisfied.

Summary. Resuming, we can say that we have reduced the problem on the whole real line (A.8) to the variational formulation (A.9) on a bounded domain. ■

Remark. Observe that the variational formulation (A.9) still has the character of a minimization problem. The optimal value, the eigenvalue to be found $-\beta^2$, is actually equal to $-\tilde\beta^2$ where $\tilde\beta$ is the solution of the variational problem,

i.e. the optimal value $\tilde{\beta}$ is the eigenvalue to be found.

■

Remark. Formulating the results for the differential equation, we observe that in the interior domain we look for a solution and a value $\tilde{\beta}$ such that (A.10) is satisfied together with the boundary condition (A.11). This boundary condition for the interior problem is a condition of mixed Neumann-Dirichlet type, and could be called *a transparent boundary condition*. The remarkable fact is that this condition makes it possible to replace the unbounded problem to a BVP on a bounded interval.

■

Remark. Actually, the position of the point $x = B$ in the above has been 'arbitrary' outside the wave guide, since for $x > B$ the solution was determined by the exterior value n_0. This makes it clear that there is no objection to take $B = w$, just the boundary of the wave guide. For numerical purposes this is the most optimal way to do. ■

A.2.3 Approximations with simple trial profiles

Confinement at 'partly-optimal' Dirichlet boundary

We start with the formulation on the unbounded domain for the case of the step-index. We try a very crude approximation: approximate the principal mode by a confined function identically vanishing for $|x| > W$, where the width W will become a parameter in the trial function.

Remark. An approximation of this kind, to 'confine' the field, is often made: knowing that the field vanishes exponentially, for large enough W the field is very small, and is there, 'far out', replaced by zero: a Dirichlet boundary condition. ■

Continuing with this crude approximation, and anticipating the symmetry and positiveness of the solution, let us simply take as trial profile a function that is piecewise linear, a 'tent-function'. Thus,

$$\phi(x) = a(W - |x|)$$

The approximate constrained formulation then provides an approximate value of the eigenvalue, and becomes (we use additionally symmetry in x)

$$-\beta_{app}^2 = \min_{a,W} \left\{ a^2 \int_0^W \left[1 - \omega^2 n^2 (W-x)^2\right] dx \;\middle|\; a^2 \int_0^W (W-x)^2 dx = 1 \right\}$$

The integrals can be evaluated easily, and the 2-parameter constrained optimization problem can be solved.

To appreciate the value of the variational formulation, and the good result that is obtained for the eigenvalue, in the plots below we present the eigenvalue as function of the width of the waveguide and compare this with the plot of the exact values. Also the value of the 'optimal' Dirichlet-width W as function of the wave guide width is shown.

PLOTS

Exercise. Assume the waveguide is bi-modal, i.e. supports precisely two guided modes. The second mode will then be an odd function of x. Describe a comparable approximation for the eigenvalue of the second mode by using a tent-function vanishing at $x = 0$ and at $x = W$. ■

Using the confined formulation

The approximation with Dirichlet conditions above is rather awkward, since no non-identically vanishing field can smoothly (satisfying interface conditions) be connected to a zero field in an exterior domain. When using the confined formulation derived above this problem can be resolved.

Then again, a simple way to find the principal eigenvalue would be to approximate the function by a tent-like trial function in the interior $x < B$. Taking $B = w$ this is a function of the form

$$\phi(x) = a(w - x) + b$$

The minimization problem then reduces to a 3-parameter minimization problem in the parameters a, b and β. Results of the calculations are plotted below, and compared to the exact value and to the approximation with optimal-Dirichlet boundary conditions derived above.

PLOTS....

Exercise. Observe that for the fundamental mode it is possible to make a much more clever choice for the trial function: a harmonic function of the form

$$\phi(x) = a\cos(px)$$

with a, p parameters. In fact, in this case we know that for solutions in the interior it should hold that $p = \sqrt{\omega^2 n_1^2 - \beta^2}$; using this information the problem reduces to a minimization problem in only two parameters: a and β. Verify that in this way the exact solution is found.

■

Exercise. Use a tent-like approximation to approximate the eigenvalue of the second mode in a bi-modal wave guide. Then use a harmonic trial function.

■

A.2.4 Variational formulation for radiation modes

The (non-guided) modes are solutions that do not vanish at infinity. Restricting to bounded solutions, for $\beta^2 < \omega^2 n_0^2$ the profile functions ϕ behave for $x > w$ like

$$\phi(x) = \phi(w)\exp(\pm i\sqrt{\omega^2 n_0^2 - \beta^2}(x - w)) \ .$$

As stated above, these solutions are called *radiation modes*; their behaviour in the x-direction is oscillatory. In the z-direction the solution is oscillatory for $\beta^2 > 0$ (and hence the electrical field propagating). [[For $\beta^2 < 0$ the solution is evanescent (exponentially decreasing or increasing) in the z-direction.]]

A variational formulation on the whole real line formally doesn't make sense since the exterior integrals will diverge. However, the confined variational formulation, now with β prescribed, is still sensible, just as the differential formulation with the transparent boundary condition.

A.2.5 FEM-numerics for complicated index variations

Of course, the confined formulation is ideally suited to be discretized using FEM: in the interior the function can be approximated by elements, for instance linear elements, leaving the value at the end point of the waveguide free. Actually, arbitrary index-profile within the wave guide can then be calculated; the derived formulation remains valid as long as the exterior problem is not changed.

This remark opens the possibility to deal with much more complicated internal structures, for instance consisting of a number of parallel waveguides with arbitrary index profiles.

Exercise. Using this idea, but taking only a simple trial function, approximate the eigenvalue of the principal guided mode of a system consisting of two separated, parallel wave guides (with the same, or different, index, larger than outside). ■

Appendix B

Variational Fluid Dynamics

B.1 Free Surface Wave Models

The evolution of waves on the surface of a layer of fluid (such as water) remains a challenging task. Assuming the fluid to be incompressible and inviscid, and the flow to be irrotational, the full surface wave equations (FSWE) are well known. The combination of dispersive and non-linear effects present major problems in the numerical simulation as well as in the theoretical analysis of the resulting interesting phenomena. Although FSWE is a well known description for the complete physics, this set of equations is too complicated for a direct investigation. Therefore, to gain insight in interesting characteristic phenomena simplified models are desired that are amenable for theoretical investigations, while, at the same time, should be accurate enough to capture the phenomenon of interest.

In this section we present a unified view based on the basic variational structure and describe models of KdV- and NLS-type of equations and their relation, avoiding for shortness all variational-consistent modelling steps in between[1].

B.1.1 Full surface wave equations

We consider the motion of a layer of fluid under the following simplifying assumptions:

- the fluid is *inviscid*, *incompressible* (density normalized to unity), and no surface tension;
- the bottom is flat, at depth $z = -h$;
- the fluid motion is *irrotational*, assumed to be uniform in the (horizontal) y-direction and unbounded in the x-direction; if the horizontal and vertical velocities are denoted by $U = U(x, z, t)$ and $W = W(x, z, t)$ respectively,

[1] For simplicity we will approximate the dispersive properties as they are relevant for long waves (shallow water), leaving out more precise dispersive properties for short waves (deep water). Formally speaking, the dispersion for long waves leads to the classical KdV-equation, and to the *defocusing (diverging)* NLS-equation for wave packets. For waves with small wave lengths, comparable to the depth of the fluid, the full dispersive properties should be dealt with, which leads to a non-local version of the KdV-equation and the *focusing (convergent)* NLS-equation.

irrotationality allows the introduction of the fluid potential Φ such that $(U, W) = \nabla\Phi$, $U = \Phi_x$, $W = \Phi_z$.Then incompressibility implies

$$\Delta\Phi = 0;$$

- the surface elevation is the graph of a function (no overturning waves) $\eta = \eta(x, t)$.

Then the governing equations are

$$\Delta\Phi \equiv \Phi_{xx} + \Phi_{zz} = 0, \quad -h < z < \eta(x,t) \tag{B.1}$$

$$\Phi_z = 0 \quad \text{at} \quad z = -h, \tag{B.2}$$

$$\partial_t\eta = -\eta_x\Phi_x + \Phi_z \quad \text{at} \quad z = \eta(x,t), \tag{B.3}$$

$$\partial_t\Phi + \frac{1}{2}\left(\Phi_x^2 + \Phi_z^2\right) + g\eta = 0 \quad \text{at } z = \eta(x,t). \tag{B.4}$$

Equation (B.3) is a kinematic condition; equation (B.4) is a dynamic condition, resulting from Bernoulli's equation restricted to the free surface. For a correct interpretation of the FSWE, most important is to observe how the interior Laplace problem is linked to the dynamic equations.

Notation. In the following we will sometimes use normalized variables, which lead to the same equations but with $g = h = 1$. ■

B.1.2 Variational structure of FSWE

Another way to interpret the dynamical structure is using a variational description. In fact, it has been observed (independently) by Zakharov (1968) and Broer (1974) and Miles (1977) that FSWE can be described as a Hamiltonian system. Summarizing, this can be described by using as variables the fluid potential at the free surface

$$\phi(x,t) = \Phi(x, \eta(x,t), t)$$

and the surface elevation. Then the full surface wave equations can be described as a Hamiltonian system (see for full details [13])

$$\begin{pmatrix} \partial_t\eta \\ \partial_t\phi \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \delta_\eta H(\phi,\eta) \\ \delta_\phi H(\phi,\eta) \end{pmatrix} \tag{B.5}$$

Here, H is the Hamiltonian functional which, just as for Hamiltonian systems from Classical Mechanics, is the sum of kinetic and potential energy

$$H(\phi,\eta) = K(\phi,\eta) + \int \frac{1}{2} g\eta^2 dx.$$

The kinetic energy is given for solutions of the Dirichlet problem for the Laplace problem in the fluid domain:

$$K(\phi,\eta) = \int\int \frac{1}{2}|\nabla\Phi|^2 dxdz = \int \frac{1}{2}\phi\left[\partial_n\Phi\right]_{z=\eta} dx.$$

Since this functional cannot be expressed explicitly in terms of ϕ, η, simplified models are sought for by constructing approximations for this functional. In

doing so, and taking as governing equations the system (B.5) with the approximated Hamiltonian, leads to a model that has retained the basic variational structure, which is not guaranteed in a direct approach.

Remark. The equations can be reformulated by using a velocity type of quantity, the x-derivative of the potential at the free surface:

$$u(x,t) = \partial_x \phi.$$

This could be expected since the potential is determined from physical quantities only up to an arbitrary constant. Then the equations get the form (generalized Hamiltonian, Poisson-structure)

$$\begin{pmatrix} \partial_t \eta \\ \partial_t u \end{pmatrix} = - \begin{pmatrix} 0 & \partial_x \\ \partial_x & 0 \end{pmatrix} \begin{pmatrix} \delta_\eta H(u,\eta) \\ \delta_u H(u,\eta) \end{pmatrix} \tag{B.6}$$

observe the skew-symmetry of the matrix-differential operator. ■

Variational-consistent models: linear, Boussinesq, KdV and NLS

Using the variational formulation of the FSWE above, simplified models can be obtained with variational-consistent modelling by approximating the Hamiltonian, in particular the kinetic energy functional. We will give the main steps and results below, and refer to [13] for full details.

B.1.3 Linearized SW, dispersion

In linearized theory, infinitessimal small surface elevations are considered. Without surface elevation and bottom variations, the problem is on a straight strip and can be solved in closed form by Fourier techniques. For infinitesimally small amplitude waves, the linearized problem can be solved and leads to the basic (linear) dispersion relation between frequency and wavenumber, given by

$$\omega = \pm \Omega(k), \text{ with } \Omega(k) = k\sqrt{g \tanh(hk)/k}. \tag{B.7}$$

Introduce the dispersion operator R such that $\omega = kR(k)$, so

$$R(k) = \sqrt{g \tanh(hk)/k},$$

and observe the limiting behaviour for small wave numbers (long waves)

$$R \sim R_{KdV} \equiv \sqrt{gh}(1 - \frac{1}{6}h^2k^2) \text{ for } k \to 0.$$

Note that this shows that this limiting operator is a simple differential operator:

$$R_{KdV} = \sqrt{gh}(1 + \frac{1}{6}h^2\partial_x^2)$$

Using this dispersion operator, the kinetic energy is approximated by a single integral functional over the horizontal direction

$$K_{lin}(u,\eta) = \int \frac{1}{2g} u R^2 u dx.$$

The linearized equations are then given by (B.6) with linearized Hamiltonian:

$$H_{lin} = \int \left[\frac{1}{2g} u R^2 u + \frac{1}{2} g \eta^2 \right] dx.$$

Observe that when dispersive effects are neglected, this leads to the Hamiltonian for shallow water

$$H_{sh} = \int \left[\frac{1}{2} h u^2 + \frac{1}{2} g \eta^2 \right] dx$$

and the simple equations

$$\partial_t \eta = -\partial_x \left[h u \right], \quad \partial_t u = -\partial_x \left[g \eta \right].$$

This is also true for varying bottom, leading to the second order equation for the wave elevation

$$\partial_t^2 \eta = \partial_x \left[g h(x) \partial_x \eta \right].$$

B.1.4 Boussinesq type of equations

One step further than the linear approximation, is to take for small amplitude solutions a first order nonlinear effect into account. At the same time, assumptions on the characteristic wave length are commonly made; mostly the restriction is to 'long' waves.

A characteristic, often used approximation is the so-called Boussinesq approximation, which corresponds to the specific relation between the wave amplitude ε and wave length λ given by $1/\lambda^2 \sim \varepsilon$, the case of 'rather small, rather long waves'. This is the basic assumption to arrive at what are called Boussinesq-type of equations. These equations describe both waves running to the right and the left. The equations are again given by (B.6), now with Hamiltonian

$$H_{Bous} = \int \left[\frac{1}{2g} u R_{KdV}^2 u + \frac{1}{2} g \eta^2 + \frac{1}{2} \eta u^2 \right] dx.$$

B.1.5 KdV type of equations

Further restricting to waves running mainly in one direction, leads to KdV-type of equations[2]. From two dependent variables u, η the approximation leads to

[2] *Korteweg-de Vries equation (1895)*
Korteweg and de Vries derived in 1895 a model equation for the motion of waves on the surface of a layer of fluid above a flat bottom. Restricting to rather low, rather long waves, they derived the equation (B.8) that now bears their name with $R_{KdV}(k)$. This equation became well known in the sixties since it turned out that from a mathematical point of view it was the first partial differential equation shown to be completely integrable, leading to a huge extension of the theory of nonlinear pde's. It also became clear that many problems in physics and technics are modelled by this equation.
Being an evolution equation, first order in time, the *initial value problem* requires to find the evolution of the surface profile from a given initial profile. This initial value problem for KdV is not easy to solve; for arbitrary initial profiles, numerical calculations have to be used to find the subsequent wave profiles; the complete integrability makes it possible in principle to write down the time-asymptotic profile.

a single variable, say the surface wave elevation η. The Hamiltonian becomes (with some factors taken out)

$$H_{KdV} = \int \left[\frac{1}{2}\eta R\eta + \frac{1}{4}\eta^3\right] dx$$

and the equations becomes a first order in time equation of the form

$$\partial_t \eta = -\partial_x \delta_\eta H(\eta), \text{ i.e. } \partial_t \eta = -\partial_x (R\eta + \frac{3}{4}\eta^2) \tag{B.8}$$

for the normalized surface elevation η. Taking the dispersion operator R_{KdV}, KdV is a partial differential equation:

$$\partial_t \eta = -\partial_x (\eta + \frac{1}{6}\partial_x^2 \eta + \frac{3}{4}\eta^2) \tag{B.9}$$

Often, a moving frame is introduced and an additional scaling in the spatial variable, and the equation gets a 'standard' form given by

$$\partial_t \eta + \partial_x \left[\partial_x^2 \eta + \frac{1}{2}\eta^2\right] = 0.$$

B.1.6 NLS-model

Considering perturbations of a monochromatic wave centered around a certain wavenumber k_0, one looks for the slow and small evolutions of the (complex-valued) amplitude A of the form

$$\eta(x,t) = A(x,t)e^{i(k_0 x - \Omega(k_0)t} + cc$$

For the second nonlinearity as in KdV, the analysis is somewhat complicated (more than for qubic nonlinearities as in optics). The result is easily described: the governing equation for A reads

$$i\left[\partial_t A + V_0 \partial_x A\right] + \beta \partial_x^2 A + \gamma |A|^2 A = 0 \tag{B.10}$$

Here, $V_0 = \Omega'(k_0)$, $\beta = -\frac{1}{2}\Omega''(k_0)$ is the group velocity and the coefficient for the group velocity dispersion respectively. γ is a coefficient from mode generation and also depends on k_0.

In a frame moving with the group velocity, the equation reduces to the form of the NLS-equation:

$$i\partial_\tau A + \beta \partial_x^2 A + \gamma |A|^2 A = 0$$

Performing a simple scaling transforms this equation to the standard form of the NLS-equation

$$i\partial_\tau A + \partial_x^2 A + \text{sign}(\beta\gamma)|A|^2 A = 0, \tag{B.11}$$

a well known equation that has been studied extensively (e.g. [1]).

The sign of the coefficients, or better sign($\beta\gamma$), determines the character of the NLS equation:

- diverging (defocusing) NLS if sign($\beta\gamma$) < 0, and

- converging (focusing) NLS if $\text{sign}(\beta\gamma) > 0$.

The converging NLS has soliton-type of solutions and more 'confined' solutions, as we shall see in the next section. In this case, the dispersive and the nonlinear effects counterbalance each other, while in the defocussing NLS the waves will spread.

For surface wave equations, for all wavelengths β is negative. For the KdV-dispersion, γ is positive for all wavelengths, while when using the full dispersion properties γ changes sign from positive for $k < k_{crit}$ to negative for $k > k_{crit}$. The critical wave number k_{crit} is known as the Davey-Stewartson value, approximately $k_{crit} \approx 1.363$. This distinguishes the two cases that can appear in nature.

The variational formulation for (B.11) is of the form of a complex Hamiltonian system and reads

$$\partial_\tau A = -i\delta H_{NLS}(u), \qquad \text{with} \quad H_{NLS}(A) = \int \left[\frac{1}{2}|\partial_x A|^2 - \frac{\text{sign}(\beta\gamma)}{4}|A|^4\right] dx.$$

Appendix C

Solitons and wave groups

C.1 Coherent structures as relative equilibria

We have come across KdV- and NLS-type of equations in optics and in surface waves:

in optics:
for pulse propagation in dispersive, nonlinear material, the variations of the E-field are described by KdV-type of equation (A.3), and the complex amplitude of wavegroups describing modulations of a monochromatic wave by the NLS-equation (A.5);
in material with Kerr-nonlinearity the amplitude variations for a beam propagating and slowly deforming in one direction is described by the NLS-equation (A.6);

in surface waves:
the KdV for the surface wave elevation for unidirectional waves (B.9), and for the complex amplitude of wavegroups describing modulations of a monochromatic wave, the NLS-equation (B.10).

Depending on the type of application the parameters have a different meaning, just as well as the independent variables which are time- or space-like. Performing a scaling, the parameters can be scaled away, and a 'standard' form of the equations can be written like:

for KdV

$$\partial_t u = -\partial_x \delta H(u), \text{ with } H(u) = \int \left[\frac{1}{2} \left(\partial_x u \right)^2 + \frac{1}{6} u^3 \right] dx$$

for (Converging/Diverging)NLS

$$\partial_t A = -i\delta H(A) \quad \text{with} \quad H(A) = \begin{cases} \int \left[\frac{1}{2} |\partial_x A|^2 - \frac{1}{4} |A|^4 \right] & \text{for CNLS} \\ \int \left[\frac{1}{2} |\partial_x A|^2 + \frac{1}{4} |A|^4 \right] & \text{for DNLS} \end{cases}$$

For ease of presentation we will interpret in the following t as time and x as spatial variable.

Equations like these that are nonlinear and are therefore difficult in the sense that usually no explicit solutions can be written down. Occasionally special solutions can be found, and even a family of solutions depending on parameters. These special solutions are often found in an ad-hoc way, using some special Ansatz, such as a 'travelling wave' Ansatz. In many cases it can be understood in a constructive way using the theory of Relative Equilibria for dynamical systems from Classical Mechanics, when generalized to Variational Evolution Equations like wave equations and more general continuous Poisson systems. We will show in this Appendix how this theory can be applied to find some of the most famous 'coherent structures' in nonlinear wave theory: the soliton solutions of the KdV and NLS equation.

C.2 Solitons of KdV

C.2.1 Motivation from Travelling Wave Ansatz

The motivation for Korteweg and de Vries to study the problem of surface waves, was to settle a dispute that continued throughout the nineteenth century about the existence of *travelling waves: is it possible that a wave exists that doesn't change in time, but is merely translated at a fixed speed?*

They showed, by deriving their (KdV-) equation and analyzing it, that the answer is affirmative. More so, it is possible to write down the wave shapes and speeds explicitly. This is quite unexpected at first sight, since KdV combines nonlinearity (leading to 'breaking'-phenomenon) and dispersion ('spreading' of initial profile). The remarkable property is that these combined effects make it possible that there exist *travelling waves*, waves with a specific profile, say f, that will neither break nor spread (an exact balance between the counteracting effects of breaking and spreading), and that travel undisturbed in shape at a specific speed, say V. That is, a solution of the form

$$\eta(x,\, t) = f(x - V\, t)$$

for specific profiles f and specific related velocity V.

To find the wave profile f and the velocity V, we substitute this form in the KdV-equation, in normalized variables

$$\partial_t \eta + \partial_x \left[\partial_x^2 \eta + \frac{1}{2}\eta^2 \right] = 0.$$

Then the pde becomes an ode for the function f in which V enters as a parameter to be determined together with the profile. We shall see that, in fact, there is a whole family of such waves; the higher the amplitude, the larger the velocity.

Writing $\xi := x - V\, t$, the equation becomes

$$-V\, \partial_\xi f(\xi) + f(\xi)\, \partial_\xi\, f(\xi) + \partial_\xi^3 f(\xi) = 0$$

A solution of this equation, for certain V, produces the wave profile f of the wave that travels undisturbed in shape at speed V.

Analysis of solitary wave profiles

To find the solution we have to distinguish two cases:

- *space-(and time-) periodic solutions*, for which f is a periodic function of ξ, the so-called cnoidal waves (since the profile is expressed with the elliptic cnoidal function), and
- *solitary wave solutions*: wave profiles of a single hump that decay, together with all derivatives, sufficiently fast at infinity ("almost confined", exponentially small outside a certain interval).

We will concentrate on the solitary wave profiles.
Then by integrating the equation above once, noticing that the constant of integration has to vanish as a consequence of the decay at infinity, leads to the second order ode for the profile:

$$-V\,f(\xi) + \frac{1}{2}\,f(\xi)^2 + \partial_\xi^2\,f(\xi) = 0 \tag{C.1}$$

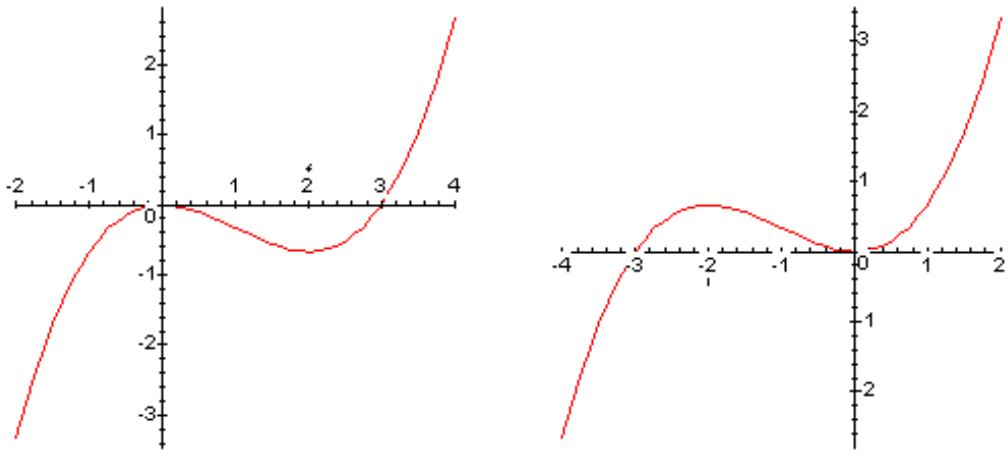
This equation can be solved in a standard way by observing the *mechanical analogue:* when ξ is interpreted as the time, and f as the position, the equation describes the motion of a particle of unit mass subject to a potential force with potential energy U according to Newton's law:

$$\partial_\xi^2\,f(\xi) + \frac{d\,U}{d\,f} = 0 \tag{C.2}$$

with potential energy

$$U(f) = -\frac{1}{2}\,V\,f^2 + \frac{1}{6}\,f^3.$$

The plot of U is qualitatively as shown below, at the left for positive values of V, at the right for negative values:



Looking for a solitary wave profile f that decays to zero for ξ tending to $-\infty$, ∞, we look for the solution that is nontrivial and connects the origin with

itself: a *homoclinic orbit*. Clearly, this can only be achieved for positive values of V.

In more detail, for the profile equation *mechanical-energy conservation* holds and phase plane analysis can be used as is described in the following.

Multiplying (C.2) with $\partial_\xi f$ and integrating the equation again, there results:

$$\frac{1}{2}\,[\partial_\xi f]^2 + U(f(\xi)) = E.$$

Since E should be zero for a solitary wave profile, the equation becomes

$$\frac{1}{2}\,[\partial_\xi f(\xi)]^2 - \frac{1}{2}\,V\,f(\xi)^2 + \frac{1}{6}\,f(\xi)^3 = 0.$$

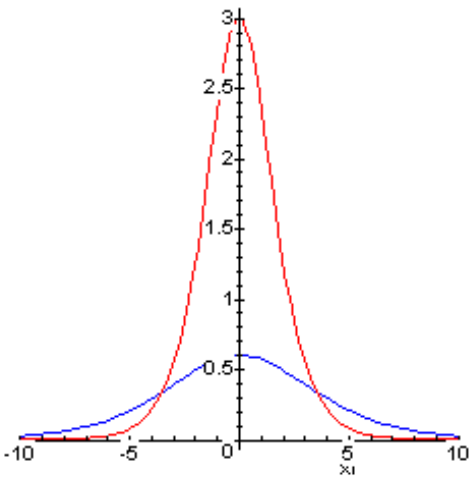
This is a first order equation for the profile function and its solution can be given explicitly. This solution is a *solitary wave profile*: for each V with $V > 0$ it is given by

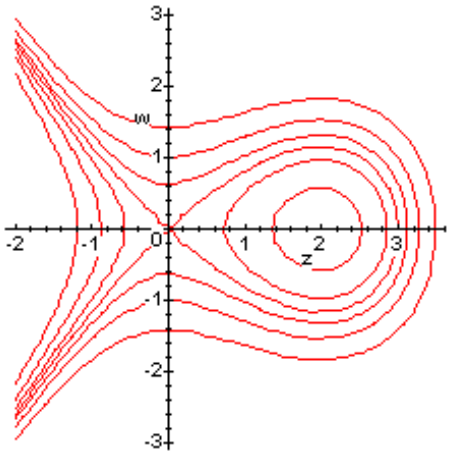
$$f(\xi,\, V) = \frac{3V}{\cosh(\frac{1}{2}\,\sqrt{V}\,\xi)^2}$$

Two profiles, for $V = .2$ and $V = 1$, are shown below. The solution above can be recognized in the *phase plane* $z = f$, $w = \partial_\xi\, f(\xi)$ in the following way. The curves of constant energy, given by

$$\frac{1}{2}\,w^2 + U(z) = E,$$

are sketched in the phase-plane (z, w) below; in this phase portrait the solitary wave corresponds to the homoclinic orbit which is the level curve through the origin (for which $E = 0$).





Remark. *With V the velocity, the amplitude is proportional to V, and the width proportional to $\frac{1}{\sqrt{V}}$: the larger the amplitude, the more confined the wave, and the larger its speed. We will find this same result in a way that doesn't uses the detailed formula in the following.* ■

C.2.2 Solitons as Relative Equilibria

The KdV equation in the standard form has the dynamic structure

$$\partial_t \eta = \partial_x \delta H(\eta), \quad \text{with} \quad H(\eta) = \int \left[\frac{1}{2} (\partial_x \eta)^2 - \frac{1}{6} \eta^3 \right] dx.$$

Restricting to waves that decay at infinity, the following functional is a constant of the motion for the KdV-equation, as can easily be verified:

$$I(\eta) = \int \frac{1}{2} \eta^2 dx.$$

The corresponding flow Φ^I follows from solving the equation

$$\partial_\tau \eta = \partial_x \delta I(\eta) = \partial_x \eta$$

which is a simple translation of any initial profile $u(x)$:

$$\Phi^I_\tau(u)(x) = u(x + \tau).$$

This motivates to call I the *momentum integral*.

The constrained variational problem for RE using the Hamiltonian and the momentum integral reads:

$$\begin{array}{rl} & \text{Crit } \{ H(u) \mid I(u) = \gamma \} \\ = & \text{Crit } \left\{ \int \left[\frac{1}{2} (\partial_x \eta)^2 - \frac{1}{6} \eta^3 \right] \;\middle|\; \int \frac{1}{2} \eta^2 = \gamma \right\} \end{array}$$

with equation from LMR for a critical point f and multiplier λ:

$$-\partial_x^2 f - \frac{1}{2} f^2 = \lambda f$$

This is precisely equation (C.1) for the profile function with (minus) the multiplier the velocity. Then the dynamic solution is given by

$$\Phi^I_{\lambda t}(f) = f(x + \lambda t)$$

just as found above.

Remark. Observe that the soliton, now characterized as a Relative Equilibrium (coherent structure) with the constrained formulation, has the property that it minimizes the energy at given momentum, or reversed, that the momentum is maximized for prescribed energy: nature chooses the soliton profile to transport 'information' with least energy at given momentum, or with maximal momentum at given energy. A nice description and 'discovery' of *optimality in nature*. ■

Remark. Actually the KdV equation is so special that it has infinitely many integrals; therefore many relative equilibria can be constructed by taking more and more integrals as constraints For instance, such formulations will lead to two- and more general N-soliton interactions. The problem is then that the 'flows' of these other integrals are just as difficult to find as the KdV equation itself; the momentum functional above is special in that respect. ■

Scaling argument for KdV solitons

The constrained variational RE-formulation, can also be used to derive in a simple way the relation between amplitude, width and velocity of a soliton by approximating the soliton by a simple confined tent-function. Consider the 'tent-' function as trial function with amplitude a and width W as parameters:

$$\phi(x) = \begin{cases} a\frac{W-|x|}{W} & \text{for } |x| < W \\ 0 & \text{for } |x| \geq W \end{cases}$$

Then, leaving out all numerical constant that appear, the constrained formulation leads to a simple optimization problem in the parameters:

$$\begin{aligned} & \text{Crit} \left\{ \int \left[\partial_x u^2 - u^3\right] dx \;\middle|\; \int u^2 dx = \gamma \right\} \\ \sim\ & \text{Crit}_{W,a} \left\{ W\left[\left(\frac{a}{W}\right)^2 - a^3\right] \;\middle|\; Wa^2 = \gamma \right\} \\ \sim\ & \text{Crit}_W \left[\frac{\gamma}{W^2} - \gamma\sqrt{\gamma/W}\right] \quad (\text{taking } a > 0) \\ \text{attained for } W \ \sim\ & \gamma^{-1/3} \text{ with value } \gamma^{5/3} \end{aligned}$$

From this it follows: $W \sim \gamma^{-1/3}, a \sim \gamma^{2/3}, \lambda \sim \partial_\gamma \gamma^{5/3} \sim \gamma^{2/3}$. Hence, $a \sim \lambda \sim W^{-1/2}$, the same scaling as found above for the exact formula.

Exercise. Consider the so-called BBM eqn. (Benjamin, Bona & Mahony, 1972):

$$(1 - \partial_x^2)\partial_t u = -\partial_x u - u\partial_x u \tag{C.3}$$

This model equation is a variant of the KdV eqn. (normalized variables).

1. Determine the dispersion relation of the linearized equation. What is the relation with the dispersion relation of the linearized FSWE, and with that of the standard-KdV-equation.
2. Looking for travelling waves, $u(x,t) = f(x - V\,t)$, write down the equation for the profile function f; do you recognize this (form of the) equation? Find the solution explicitly.
3. Give the constrained variational formulation for the soliton profile, and explain it as a Relative Equilibrium by recognizing the first integrals.

■

C.3 NLS Wave Groups

Using the Hamiltonian structure of NLS, we will now directly apply the reasoning for Relative Equilibria by using the two most simple integrals. (Just like KdV, NLS is very special and has infinitely many integrals...). Among several other interesting solutions, we will also find the NLS-soliton solution. That is, all these solutions are found for CNLS, while DNLS has none of these solutions. Therefore we restrict to CNLS in the following.

Hamiltonian structure

The CNLS has a Hamiltonian structure of the following form:

$$\partial_\tau A = i\delta H(A), \text{ with } H(A) = \int \left[\frac{1}{2}\beta|A_x|^2 - \frac{1}{4}\gamma|A|^4\right] dx. \tag{C.4}$$

(We retain the parameters β and γ so that we can recognize at each place which term is caused by dispersion and which by nonlinearity.)

First integrals and their flow

The following two quadratic functionals are both constants of the motion. They have a physical meaning and their flow can be written down explicitly from the related equation.

The first integral can be interpreted as the *wave energy (wave power),* and its flow (infinitesimal symmetry) expresses the Gauge invariance of NLS:

$$\begin{aligned} N(A) &= \int \frac{1}{2}|A|^2 dx, \\ \text{with flow} &: \partial_\tau A = i\delta N(A) = iA, \text{ i.e. } A = c.e^{i\tau} \end{aligned}$$

Another quadratic functional is called *Linear momentum* since its flow is translation symmetry:

$$\begin{aligned} L(A) &= \text{Im} \int \frac{1}{2}\bar{A}\partial_x A dx, \\ \text{with flow} &: \partial_\tau A = i\delta L(A) = -\partial_x A, \text{ i.e. } A(x,\tau) = A(x-\tau,0) \end{aligned}$$

C.3.1 Relative Equilibria: soliton- and periodic wave groups

NLS combines diverging/converging effects of dispersion and of nonlinearity. When signs are correct, for the CNLS, the diverging effect of dispersion, and the confining effect of nonlinearity balance each other and 'confined' solutions, like a soliton, exist. CNLS is most famous for its 1,2, .. N - soliton solutions, which can (accidentally) be written down relatively easy. This is related to the completely integrability of NLS, and the related existence of an infinity of conservation laws (first integrals).

Relative equilibria are found as critical points of the Hamiltonian at a given value of one or more other integrals. With the wave energy and linear momentum as constraints, the constrained critical point problem reads

$$\text{Crit } \{ H(A) \mid N(A) = \text{constant}, L(A) = \text{const} \} .$$

and a critical point should satisfy for some multipliers the equation

$$\begin{aligned} \delta H(A) &= \sigma_N \delta N(A) + \sigma_L \delta L(A) \\ -\beta\partial_x^2 A - \gamma|A|^2 A &= \sigma_N A + \sigma_L i\partial_x A. \end{aligned}$$

Actually we can simplify the following a little bit by noticing that the term $\sigma_L i\partial_x A$ can be 'gauged' away from this equation by a transformation $B = Ae^{i\alpha x}$ for a suitable α. So, essentially we consider

$$\text{Crit } \{ H(A) \mid N(A) = \text{constant} \} .$$

with equation (an additional minus sign for the multiplier for convenience)

$$\begin{aligned} \delta H(A) &= -\mu \delta N(A) \\ -\beta \partial_x^2 A - \gamma |A|^2 A &= -\mu A \end{aligned}$$

Clearly we can look for real-valued solutions of this RE-equation, which we will denote by $a(x)$. Having found such a real solution (relative equilibrium), the corresponding dynamic Relative equilibrium solution will be

$$A(x,t) = a(x) e^{-i\mu t}$$

which is a time-harmonic modulation $e^{-i\mu t}$ of the fixed profile $a(x)$, a 'standing wave'.

Basic in the analysis of the RE-equation is the recognition that it has as mechanical analogue Newton's equation: with β the mass of a particle in a conservative force field with potential P:

$$\beta \partial_x^2 a = -\frac{\partial}{\partial a} P(a), \quad \text{with potential} \quad P(a) = -\frac{1}{2}\mu a^2 + \frac{1}{4}\gamma a^4 \tag{C.5}$$

This problem can again be analyzed with phase plane techniques. The sign of μ and γ will be important for the potential profile.

The sign of μ determines the stability of the trivial solution $a \equiv 0$: for $\mu < 0$ the trivial solution is stable, and is unstable for $\mu > 0$. Note that for positive μ, the lowest value of the potential is at $a = \pm\sqrt{\frac{\mu}{\gamma}}$, while the potential is negative for $|a| < \sqrt{2}\sqrt{\frac{\mu}{\gamma}}$.

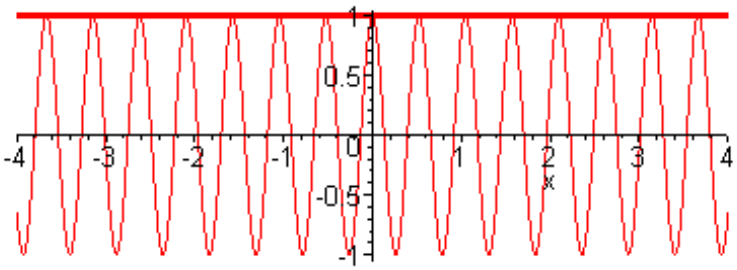
We now describe briefly the various solutions that can exist and can be found form phase plane analysis.

Nonlinear harmonic

This is the solution with constant amplitude:

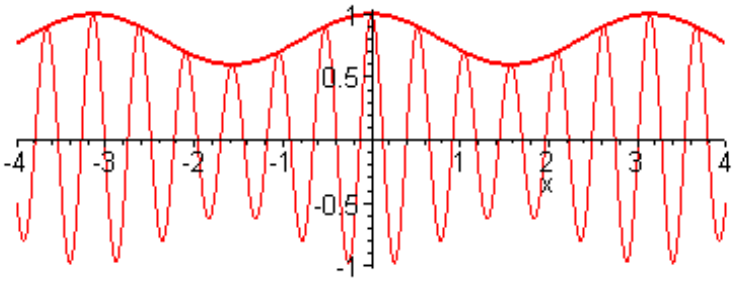
$$A = q e^{-i\gamma q^2 t}$$

which corresponds to the case that $\mu > 0$, and $q = \sqrt{\frac{\mu}{\gamma}}$, is the point of minimal potential. The real part of the NLS-solution is sketched below as function of ζ with the constant amplitude indicated:



Nonlinear modulated harmonic

Also for $\mu > 0$, small amplitude periodic motions around the point of minimal potential energy lead to NLS-solutions that are a modulation of the t-harmonic:

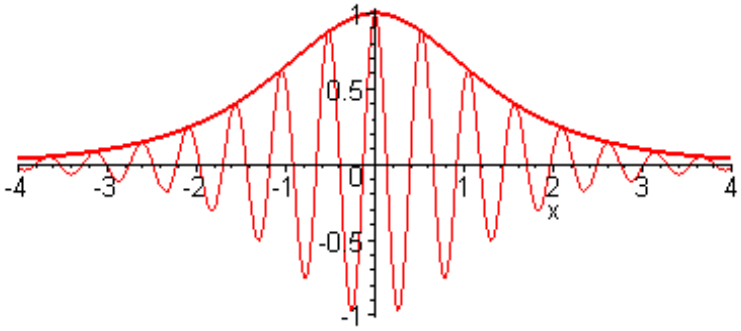


Soliton

For $\mu > 0$ there exists the famous soliton solution as homoclinic orbit. The amplitude

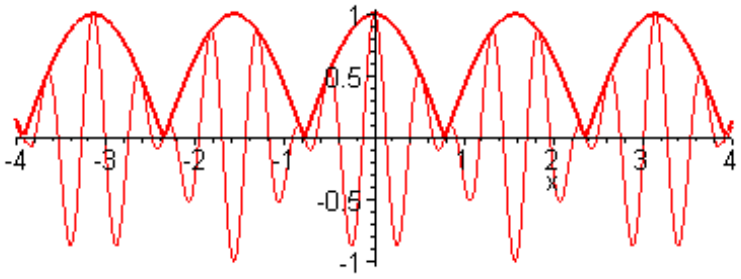
$$a = q sech(q\sqrt{\frac{\gamma}{2\beta}}x)$$

for $\mu = \gamma q^2/2$ modulates the t-harmonic and confines its support:



Nonlinear bi-harmonic

For $\mu < 0$ the potential is convex, and only periodic solutions can exist that cross the origin. This leads to what could be called 'nonlinear bi-harmonic' solutions:



Scaling and (non-) existence of NLS-solitons

Just as for KdV we can get information from the RE-constrained critical point problem about scaling properties of the soliton. However, it is also possible to see from the following simple argument that for DNLS no solitons can be expected to exist. Consider again the 'tent' function as trial function with amplitude a and width W as parameters:

$$\phi(x) = \begin{cases} a\frac{W-|x|}{W} & \text{for } |x| < W \\ 0 & \text{for } |x| \geq W \end{cases}$$

Upon substituting in CNLS (-sign), DNLS (+ sign) there results (up to positive multiplicative factors):

$$\begin{aligned} & \text{Crit } \left\{ \int \left[|\partial_x A|^2 \pm |A|^4\right] dx \;\middle|\; \int |A|^2 dx = \gamma \right\} \\ \sim \; & \text{Crit}_{W,a} \left\{ W\left[\left(\frac{a}{W}\right)^2 \pm a^4\right] \;\middle|\; Wa^2 = \gamma \right\} \\ \sim \; & \text{Crit}_W \left[\frac{\gamma}{W^2} \pm \frac{\gamma^2}{W}\right] \end{aligned}$$

Clearly, with the plus-sign, i.e. for DNLS, no critical points are found. With the minus sign, for CNLS, the minimal value is achieved for $W \sim 1/\gamma$, hence $W \sim 1/a$ and the multiplier μ follows from differentiation the minimal value: $\mu \sim \partial_\gamma \left[\gamma^3\right] \sim \gamma^2 \sim a^2$, just the result found above in the exact formula.

C.4 Exercises

1. *Kink solutions of Sine-Gordon equation*
   The Sine-Gordon equation uses an angle variable u to describe the orientation of spins on a continuous line in an magnetic system. The equation reads (with κ some material constant)

   $$u_{tt} = u_{xx} + \kappa \sin 2u.$$

   (a) Derive the equation for a travelling wave: $u(x,t) = U(x - \lambda t)$.

   (b) Show that there is a *kink-solution*, a travelling wave with $U(\xi) \to 0$ for $\xi \to -\infty$, and $U(\xi) \to \pi$ for $\xi \to \infty$; use phase plane analysis.

   (c) Investigate the variational formulation for the kink solution.

2. *Cnoidal waves for KdV*
   Travelling waves of KdV were investigated on the whole real line before. In this exercise we want to investigate travelling waves that are periodic.

   (a) Show that solutions that are periodic with period 2π on the real line can be found by periodic continuation of functions on $[0, 2\pi]$ that satisfy periodic boundary conditions.

(b) Show that for periodic solutions it is possible to restrict to solutions with zero mass: $\int u = 0$.

(c) Derive the equation for a periodic travelling wave; investigate this equation (phase plane analysis). Derive the solution in an implicit way. Using elliptic functions, the so-called cnoidal function, the solution can be "explicitly" written down; therefore such periodic waves are called *cnoidal waves*.

(d) Show that the cnoidal wave *form* is obtained as a relative equilibrium form the constrained minimal energy problem

$$\text{Min } \left\{ \int (\frac{1}{2}u_x^2 - u^3) \,\middle|\, \int \frac{1}{2}u^2 = \gamma, \int u = 0, u(0) = u(2\pi) \right\},$$

and that the cnoidal wave is the corresponding relative equilibrium *solution*.

3. ** *KdV-cnoidals, cnt'd*
In the rest of this exercise we study the above constrained minimization problem; denote a solution by U (suppressing the dependence on γ that does not play a particular role in this exercise).

(a) Show that $u \equiv 0$ is a critical point, but not the minimizer.

(b) Conclude that for the minimizer $\int U^3 > 0$, and that U cannot be a constant.

(c) Observe that with U, any translate of U is also a minimizer: there is a continuum of minimizers.

(d) Construct the Lagrangian functional; show that this Lagrangian functional is not bounded from below. Hence, U is not the (global) minimizer of the Lagrangian functional.

(e) Now show that U is also not a local minimizer of the Lagrangian functional. To that end, investigate the second variation at U. First show that the second variation at U vanishes in the direction U_x (why?). Then show that the second variation at U in the direction U is negative (use the equation for U; note that the U-direction is not tangent to the level set of the constraint-functional!).

(f) Conclude from the previous result that the value function must be a concave function.

Bibliography

[1] N. Akhmediev & A. Ankiewicz, *Solitons, Nonlinear pulses and beams*, Chapman & Hall, 1997.

[2] V.I. Arnol'd, *Mathematical methods of classical mechanics,* Springer 1989 (revised edition).

[3] C. Carathéodory, *Calculus of Variations and partial differential equations,* Chelsea, New York, 1982.

[4] F.H. Clarke,*Optimization and nonsmooth analysis, Wiley, New York, 1983.*

[5] *R. Courant & D. Hilbert, Methods of Mathematical Physics*, vol 1, 2; Interscience Publishers, New York, 11953, 1962.

[6] I. Ekeland, *Convexity methods in Hamiltonian Mechanics,* Springer, Berlin, 1990.

[7] I. Ekeland and R. Temam, *Convex analysis and variational problems*, North-Holland, Amsterdam, 1976.

[8] B.A. Finlayson, *The method of weigthed residuals and variational principles*, Academic Press, New York, 1972.

[9] A. Friedman, *Variational principles and free boundary problems,* Wiley, New York, 1982.

[10] I.M. Gelfand & S.V. Fomin, *Calculus of Variations*, Prentice Hall, 1963.

[11] H. Goldstein, *Classical Mechanics,* Addison Wesley, 1980.

[12] Goldstein, *History of the Calculus of Variations*, Springer Verlag, 1980

[13] E. van Groesen & E.M. de Jager, *Mathematical structures in continuous dynamical systems,* Elsevier North-Holland, Amsterdam, 1994.

[14] E. van Groesen & J. Westhuis, Modelling and simulation of surface water waves, *Journal Mathematics and Computers in Simulation 2001*

[15] A. Hasegawa and Y. Kodama, *Solitons in optical communications*, Clarendon Press, Oxford, 1995.

[16] M.R. Hestenes, *Optimization theory, The finite dimensional case*, Wiley-Interscience, 1975.

[17] Ioffe & V.M. Tikhomirov, *Theory of Extremal problems*, North-Holland, 1979

[18] C. Lanczos, *The variational principles of mechanics*, Univ. Toronto Press, 1970.

[19] P. Lax, Integrals of nonlinear equations of evolution and solitary waves, *Comm. Pure Appl. Math.* **21**(1968)467-490

[20] A.C. Newell and J.V. Moloney, *Nonlinear optics*, Addison-Wesley, Canada,1992.

[21] C. Sulem & P-L Sulem, *The Nonlinear Schrödinger Equation*, Springer Verlag, 1999.

[22] T. R. Taha & M. J. Ablowitz, Analytical and Numerical Aspects of Certain Nonlinear Evolution Equations II, *J. Comput. Phys.* **55**(1984)203.

[23] V.M. Tikhomirov, *Stories about Maxima and Minima,* American Mathematical Society, 1990.

[24] J.L. Troutman, *Variational calculus with elementary convexity*, Springer, 1983.

[25] Whitham, *Linear and nonlinear waves*, Wiley, 1976.

[26] E. Zeidler, *Nonlinear functional analysis and its applications,* part III, Variational methods and optimization, Springer, 1985.