Methods of Theoretical Physics: Building a Model of Classical Physics

Lecture in WS 2024/25 at the KFU Graz

Axel Maas

Contents

Chapter 1

Introduction

The aim of this course is to describe how theoretical physics describes the world we are living in using mathematics. Both as an example and as a subject of paramount importance this will be done with constructing all of classical physics. This includes mechanics, electrodynamics, and special relativity. Ultimately, in the last section, the full description of a classical system in a single mathematical framework will be achieved. All classical physics, e. g. thermodynamics and optics, can be derived from this one principle. This will illuminate how (theoretical) physics builds a model of our world. Of course, in research, this is a process which is a continous interplay between experiment and theory. Theory makes a proposal, based on the observations in experiments, how a basic theory could look like. Then predictions are derived from this proposal, which are subjected to experimental tests. Either these confirm it or disagree with it (within experimental uncertainties). If it disagrees, it is necessary to improve the theory, taking the new-found results into account. Or it does agree, then further predictions need to be made. Eventually, we would like to arrive at a theory which describes all we see.

The current best such theory is the combination of the standard model of particle physics joined by quantum gravity, the quantized version of general relativity. As a unified theory building this is still in development, and it is certainly not the last step. That we can already deduce from experiment, specifically from observational astronomy. But is is for sure an important stepping stone. However, integrating quantum physics into the mix raises the complexity substantially, and thus here the focus will remain on classical physics. This is also necessary as classical physics is and remains central to concept formation for all quantum theories.

Classical physics contains, as noted, three ingredients. Classical mechanics, classical electrodynamics, and special relativity. Indeed, electrodynamics cannot exist without special relativity, and thus it is mandatory to include it into the description.

Classical mechanics is, in a sense, a somewhat ancient topic. Essentially completely formulated in its modern form in the 19th century, it has matured into a mathematically consistent and closed theory. Even the advent of special relativity only required a minor modification of the underlying structures and mathematics, and could therefore be technically straightforwardly accommodated. Of course, the interpretational impact has been of much more significance. The basic formulation tools of theoretical mechanics, especially the Lagrangian formulation of chapter 3 and the Hamiltonian formulation of chapter 5, are still mathematical cornerstones of all modern theories of physics. It is not the concept which changed, merely the entities on which it is applied, at least from a mathematical point of view. Of course, again the shifts in physical interpretation had a much larger impact when quantum physics was added.

Thus, theoretical mechanics remains both mathematical and in terms of conception still a cornerstone of even the most modern areas of physics. Understanding theoretical mechanics in these formulations is therefore forming the foundation on which these are build. Of course, not to mention its ubiquitous engineering, and other, applications. Though this is not the focus here.

However, these formulation as Lagrangian and Hamiltonian mechanics in chapters 3 and 5, as powerful as they are, are very dissimilar from the usual concept of Newton's law, even if they embody the same physics. In fact, at first sight it is far from obvious that reformulating mechanics mathematically in this way could be anything but obscuring. It is only when making the transition to quantum physics and general relativity that their conceptual importance becomes readily evident, though the structural simplification can be utilized in itself in applications.

To provide a smooth transition from the experimental view on mechanics to the theoretical formulation, the first step will be to give a more theoretical perspective on Newtonian mechanics in chapter 2, sometimes also called analytical mechanics. In this way, the theoretical approach to physics problems is best outlined, as here the subject of study, ordinary Newtonian mechanics, is already well acquainted from experimental physics, so that one can concentrate on the more abstract theoretical formulation.

Another step is then to introduce the ideas of special relativity in chapter 4. This is somewhat orthogonal, and will therefore be relegated to a suitable spot.

The first version of reformulation of mechanics is then Lagrangian mechanics in chapter 3. There are two versions of this, both having their own advantages. Lagrange's equations of the fist kind are extremely useful when it comes to problems in applications of classical mechanics. They will therefore be skipped here, where conceptual developments are of more interest. Lagrange's equation of the second kind in section 3.4 are the natural way

to formulate problems embodying special relativity and quantum physics simultaneously, which is the arena of particle physics. Afterwards, another reformulation is given in terms of Hamilton's mechanics in chapter 5. This reformulation is especially useful for non-relativistic quantum physics. It may appear odd at first sight that this more special case, as it does not lend itself so easily to special relativity, is treated after the more general case. But this formulation is actually easier to understand from the Lagrangian formulation as then from starting outright towards it.

Especially, here a transition to a field theory will be performed in section 5.10, by considering an infinite number of particles. This mechanical example will be very helpful in grasping the concept of fields, which are essential in electrodynamics, which is a field theory. Electrodynamics is of fundamental importance, as it is also at the root of many day-to-day phenomena. Not only in terms of electric applications, but it is central to anything which has to do with light, as well as with chemistry, and thus the material structure of the world around us.

However, when first exploring electrodynamics, one is usually encountering electric and magnetic phenomena separately. This will be followed by first treating electrostatic phenomena in chapter 6 and then magnetostatic ones in chapter 7. Their unification into a single theory, electrodynamics, has been one of the major achievements of classical theoretical physics in the 19th century, and will be described in chapter 8.

Such a unification of phenomena remains one of the central goals, or perhaps the central goal, of fundamental theoretical physics. However, as will be seen in chapter 9, the unification of special relativity, mechanics, and electrodynamics is far less compelling as in electrodynamics, but at least everything can be discussed in a single set of equations. Further unification is still awaiting discovery.

Chapter 2

Newtonian mechanics

The aim of this first chapter is not necessarily to introduce new physics or new physics concepts. After all, Newtonian physics is familiar from experimental physics. The main aim here is to provide a new perspective on it. So far, the exposure to physics was mainly by means of understanding experiments. The theoretical approach is somewhat different. The basic idea is to start from a set of fixed rules, e. g. Newton's laws. The next step is then to derive consequences of these laws, e. g. the movement of the earth around the sun. Comparison between these derived results and experiments then decides whether the basis of the derivation is actually useful to describe experiments or not.

Of course, the aim of this exercise is to end up with the minimal basis to describe all experimental results. These would then be considered as the basic laws of nature. Whether they are indeed ingrained in reality in some way is a highly non-trivial questions, and not (yet?) resolved. However, to even pose this question requires to be able to connect some limited set of basic principles with experiments. This is the task of theoretical physics. This also includes, of course, to identify this basic set. The resulting basis is called a (standard) model or a theory, while the constructions still awaiting experimental tests are usual considered as hypothesis. There are fine distinctions between these names which, however, play no role in this lecture, nor actually in the day-to-day research.

Note, however, that any experiment can at most falsify a theory. It can never prove it to be correct. Thus, any theory or model can only be considered to be an adequate description for the time being. Though often theories are refuted entirely based on contradiction to experiment, there are some cases where they do contradict experiments, but are not really refuted. In this case, the theory turns out to be the limit of a more general theory in a particular case. E. g., Newtonian mechanics will be the limit of special relativity for small speeds. Only when experiments become sensitive enough they can detect this situation.

A special limit theory is by no means useless. When building a bridge, nobody will do

the statics using special relativity, but ordinary Newtonian mechanics. Thus such a limit theory, more often called an effective theory, is by no means useless. Quite often the more general theory is of less practical use.

2.1 Kinematics

Before starting with physics, sometimes also called dynamics, it is useful to first consider the description of (point) particles, the central entities of mechanics. This is a pure description, and there are no answers to why a particle behaves in a certain way. To separate this from the actual reasons of their movement this is often called kinematics.

The starting point for the description of a particle is the path it follows during an interval of time. In general, this path is described by a vector-valued function $\vec{r}(t)$ in a vector space. This vector space describes the position of the particle in space, while time acts as a parameter to identify the position of the particle along its path, which is also called a trajectory.

The speed $\vec{v}$ of a particle is defined to be

$$\vec{v}(t) = \frac{d\vec{r}(t)}{dt},$$

i. e. the rate of change of the position of the particle along its path. Usually, the path of particles will be infinitely often differentiable, so this is a well-defined quantity. In mechanics, however, usually only the second differential,

$$\vec{a}(t) = \frac{d\vec{v}(t)}{dt} = \frac{d^2\vec{r}(t)}{dt^2}$$

plays also an important role. The others may appear, but are not central quantities.

With these definitions, the behavior of a particle can be described (though not explained).

Given the acceleration, it is possible to obtain the path by twofold integration,

$$\vec{r}(t) = \vec{r}(t_0) + \vec{v}(t_0)(t - t_0) + \int_{t_0}^{t} dt' \int_{t_0}^{t'} dt'' a(t''),$$

where the quantities $\vec{r}(t_0)$ and $\vec{v}(t_0)$ are the position and speed at time t_0, the initial conditions. Similarly, if the speed or higher derivatives are given, the path can be obtained by a single or even more integrations.

2.2 Newton's laws

2.2.1 Preliminaries

While the previous section provided the tools to describe the movement of a particle, it did not give any reason how, e. g., the acceleration comes about. This is the question of dynamics.

As alluded to earlier, the first step in theoretical physics is to define a basis, i. e. a set of laws which describe the dynamics. In analogy to mathematics, these are sometimes also called axioms, but more often nowadays just model, given the insight that all models have their limits.

Classical mechanics is the theory which is based upon Newton's laws. They were found in a long sequence of interplay between theory and experiment. This history of physics will not be traced out here, but it should be noted that the formulation of the model is by no means a trivial exercise, and all laws of physics have been created based on a multitude of experimental insights, no matter how many spectacular things later were predicted by them.

2.2.2 The first law

To formulate Newton's laws, it is necessary to introduce a few more concepts. The first is to define a force as the origin of dynamics. I. e., without forces, the particle will not change its kinematics. To give this statement a more precise meaning requires the notion of an inertial system.

The first law of Newton is that if no forces act on a particle then there exists a coordinate system in which its acceleration and all higher derivatives of its path vanish identically, and thus the particle either remains at rest or moves at constant speed. If the constant speed is zero, this frame is called the rest frame of the particle.

The importance of the concept of an inertial system follows from the following idea (also called gedankenexperiment). A particle can be observed. If now the observer moves with respect to the particle, it will appear to be moving, even if it is the observer, which moves. Thus, the kinematics depend on the relative motion of observer and particle. However, if there are forces, i. e. something changing the kinematics, then there is a source of change not identical to just a change of coordinates, and therefore, no matter how, there is no coordinate system in which the particle behaves as expected.

There is one loophole to be fixed. If the observer would be accelerated, it may look like the kinematics change. Therefore, inertial systems are restricted to such systems which

may at most move relatively with a constant speed with respect to the coordinate system of the particle. If this is not the case, so-called pseudo forces, like the Coriolis force, emerge. This case will be discussed in section 2.9.

The abbreviated, sloppy, version of this law is: If there are no forces, the particle moves at constant speed (which includes zero speed).

2.2.3 The second law

Newton's first law identifies what happens in the absence of forces. Newton's second law describes what happens in the presence of forces. However, this requires first to define forces a little bit more. Note that this is indeed a definition of forces, not an explanation. This is one of the essences of theoretical physics, at least currently: It cannot explain everything in the sense of requiring nothing external. It can at best explain a multitude of experimental observations with a very limited number of external inputs. In this sense, also forces are something motivated and defined by experiment, but without derivation.

The definition is that a force is a vector, which is possibly time-dependent, $\vec{F}(t)$. It may also depend on any other information, e. g. the position and speed of a particle. This also implies that forces add like vectors, and the total force is the sum of all the individual forces. This is sometimes also called the fourth law, since this is also not something derivable.

In addition, every particle is assigned an inertial mass m, which is a property of said particle. There is no explanation of the origin of this mass in mechanics, and for any given problem the value of this mass has to be determined by experiment. This mass is an additive property, that is the mass of two particles M is just the sum of the individual particles' masses

$$M = m_1 + m_2,$$

and so on for more particles. This mass can have any arbitrary, positive value, and it is not quantized (comes in portions). It is an intrinsic property of a particle. In classical mechanics, furthermore, this property is immutable in time, $m(t) = m$.

Having these two concepts, the next step is, for the sake of convenience, to define a new quantity, the momentum of a particle,

$$\vec{p} = m\vec{v} = m\frac{d\vec{r}(t)}{dt}.$$

This will make the analysis especially of mass distributions, i. e. large numbers of particle with small masses, simpler, as well as the generalization beyond classical mechanics.

With this, it is possible to formulate the second law as

$$\vec{F} = \frac{d\vec{p}}{dt} = m\vec{a}(t) \tag{2.1}$$

i. e. the force changes the momentum and equals the acceleration up to a factor of the mass. As said, this is a definition, so there is no reason for it, except that it fits well with experiment.

Though mass is an immutable concept of a particle, it may still be useful to think of a time-dependent mass, e. g. if thinking of an ensemble of particles. When pouring particles on a scale, e. g., the mass on the scale changes. If this is the case, the second law takes the form

$$\vec{F} = \frac{d\vec{p}}{dt} = m(t)\vec{a}(t) + \vec{v}(t)\frac{dm(t)}{dt}, \tag{2.2}$$

i. e. the definition in terms of the momentum is the basic one, not the one in terms of the acceleration or the speed. This also elevates the momentum to be the central kinematic quantity in classical mechanics, and actually far beyond. This equation is also called the (Newtonian) equation of motion, Newton's law, or as the dynamical equation. While the first law and the third law describe only general features of systems with forces, it is this equation which actually describes the impact of forces on particles.

Though not explicitly noted, the force is in general not a constant, but it may (and in general will) depend on the position of the particles, as well as derivatives of the position, like the speed, the acceleration, or, in principle, even higher derivatives. In practice, the majority of cases involve only forces which are position-dependent, and sometimes dependent on the speed. Forces involving the speed are also often called frictional forces, as they usually appear in the context of friction phenomena.

The force may, in addition, depend also explicitly on the time, not only implicitly through the position of the particle as a function of time. This happens especially often if there is an external source of the force.

Because the force can then be seen as a function of other vectors, and has usually a well-defined value for every point in space and time, the force is often also considered a force-field, though the name force is still used for brevity. Only if the force does not depend on the position, but at most explicitly on time, it is strictly speaking not a force field.

2.2.4 The third law

The third law is of a substantially different nature than the two first laws. The two first laws describe how particles are affected by the presence or absence of forces, but do not

make any statement about the origin of the forces. This is changed by the third law. Its statement is that if the force on a particle emanates from another particle, then the target particle acts always with the same but opposite force on the source particle,

$$\vec{F}_s = -\vec{F}_t,$$

i. e. any action induces a, equal in magnitude but opposite in sign, reaction. If there are more than two particles involved, this statement applies pairwise to each possible pairing of particles. Note that still no statement is made about how any of the involved particles creates the forces, but it requires for any possibility that this balance of action and reaction is satisfied. Again, there is no reason for this at the present time, since it is an axiom.

An important approximation is very often that it is assumed that a force is external. I. e., though the back reaction occurs, it is so weak that, for all practical purposes, the origin of the force is not changed, and therefore the force on the particle does not change. An example for this is the movement of the earth around the sun. There, the impact of the third law can be taken exactly into account, but it can also be shown how it becomes irrelevant for the sun being much heavier than the earth.

2.3 Point particle

The simplest possibility is a constant force, $\vec{F}$. The equation of motion (2.2) then becomes for a constant-mass particle[1]

$$\frac{1}{m}\vec{F} = d_t^2\vec{r}(t).$$

This is an ordinary differential equation of second order. It can be solved by integrating twice on both sides,

$$\vec{r}(t) = \int dt' \int dt'' \frac{1}{m}\vec{F} = \int dt' \left(\frac{t'}{m}\vec{F} + \vec{v}_0\right) = \frac{t^2}{2m}\vec{F} + \vec{v}_0 t + \vec{r}_0, \tag{2.3}$$

where the integration constants $\vec{v}_0$ and $\vec{r}_0$ describe the position and speed of the particle at time $t = 0$, and are the so-called initial conditions of the motion. Since the equation of motion is an ordinary second-order differential equation there are always two such initial conditions required to fully specify the solution of the equation.

If the force vanishes, $\vec{F} = \vec{0}$, the particle moves at constant velocity. In this sense, Newton's first law is actually just a special case of Newton's second law, though the definition of an inertial system will play a quite central role later.

[1]For brevity, differential operators d/da will be abbreviated from now on by d_a, and $\partial/\partial a$ by ∂_a.

An interesting observation is the following. Because the equation is linear, adding two forces $\vec{F}_i$ will yield that the movement of the particle will be just the sum of the effects of the two forces,

$$\vec{r}(t) = \frac{t^2}{2m}(\vec{F}_1 + \vec{F}_2) + \vec{v}_0 t + \vec{r}_0.$$

This is called a superposition of the two movements, where the consequences of the initial conditions can be taken to be also the sum of two initial conditions, for each movement separately.

The reason is of mathematical nature. If a differential equation is linear and homogeneous, i. e. is of the type

$$\sum_{i=1}^{N} a_i d_t^i \vec{r} = \vec{0},$$

then any linear combination of two solutions (of which there are N for an ordinary differential equation of order N) again solves this equation. If there is any inhomogeneity on the right-hand side, then an arbitrary solution for this equation is given by any solution of the inhomogeneous equation to which any linear combination of the homogeneous solutions can be added.

Note that this is a particularity of linear differential equations. If, e. g., the prefactors a_i would depend on $\vec{r}$, this is no longer true. However, situations where (2.2) is a linear differential equation, with or without inhomogeneous term, are very common in classical mechanics, and therefore this superposition principle will play an important role.

An interesting example is the situation with friction. Then $\vec{F} = \alpha d_t \vec{r}$, where α is a friction coefficient. The equation of motion (2.2) is then

$$d_t^2 \vec{r} - \frac{\alpha}{m} d_t \vec{r} = 0.$$

This requires that the derivative of the first derivative must again be proportional to the first derivative. This is the typical behavior of the exponential function, and the solution is thus

$$\vec{r}(t) = \frac{m}{\alpha} \vec{v}_0 e^{-\frac{\alpha t}{m}} + \vec{r}_0,$$

where the two initial conditions give again the speed and position of the particle at $t = 0$. The two solutions are now both terms. Either being at rest from the beginning, or its movement being exponentially damped.

2.4 Gravity

As noted before, Newton's laws do not explain the origin of the forces, just how they act on particles. They are therefore sometimes called a dynamical principle, but require

still the force $\vec{F}$ to actually describe motion. It is the realm of fundamental physics to deduce these forces from experiment, and investigate, which of these forces can be deduced from other forces. The most basic ingredients known today are general relativity and the standard model of elementary particles. Though not completely covering all experimental observations, all known forces can, in principle, be deduced from them. However, the actual derivation of, say, the forces involved in standing on a floor from these elementary theories is practically far too involved and too complicated. Therefore, rather than using them, it is much better to use effective forces, which neglect all those aspects which play no role, i. e. are too weak to make any practically measurable difference, rather than the full forces of these theories. This is then considered as an effective theory, rather than a, more or less, fundamental one. Especially, in the realm of classical mechanics only such effective forces play a practical role.

The probably best known of these effective forces is the gravitational force. Given the distance between two bodies, $|\vec{r}_1 - \vec{r}_2|$, the gravitational force between them is

$$\vec{F} = \frac{G m_1^g m_2^g (\vec{r}_1 - \vec{r}_2)}{|\vec{r}_1 - \vec{r}_2|^3}, \tag{2.4}$$

where $G \approx 6.67430(15) \times 10^{-11}$ m^3/(kgs) is Newton's constant, a number obtained from measurement. The quantities m_i^g are the so-called heavy or gravitational mass. Just like the inertial mass, they are a property of particles, and must be deduced by measurement. They act as a so-called charge, i. e. they are features of particles which determine the influence of a force on this particle. It is found experimentally that the gravitational masses are always positive.

It is a remarkable, and highly non-trivial, experimental finding that for a particle of inertial mass m and gravitational mass m^g

$$m = m^g$$

holds, i. e. gravitational and inertial mass are the same. This feature, also known as the equivalence principle, is the basis for general relativity, and thus one of the most basic foundations of modern physics. There is no explanation of it, though, it is again an axiom. However, since experimentally extremely well supported, in the following no more distinction between the gravitational and the inertial mass will be made.

There is an interesting special version of the gravitational force (2.4). Set one of the particles fixed at the coordinate origin, $\vec{r}_2 = \vec{0}$. Then

$$\vec{F} = m_1 G m_2 \frac{\vec{e}_r}{r^2} = m_1 \vec{g}$$

where the so defined vector $\vec{g}$ is called the gravitational acceleration due to the body 2. E. g., on the surface of the earth, $\vec{g}$ always points to the center of the earth and has a value, depending on latitude[2], of about 9.8 m/s. Since the radius of the earth is large, $\vec{g}$ changes only very slowly when moving just a little bit vertically, and therefore can be taken to be roughly constant for the couple of kilometers, from the deepest ocean trench to the traveling altitudes of planes, where human activities usually take place.

2.5 The potential

An interesting concept can be found when considering the movement of a particle under an arbitrary, but only position-dependent, force. It is best to start with a one-dimensional example first. The equation of motion then reads

$$md_t^2 x = F(x).$$

Multiplying this equation by $d_t x$, this yields

$$m(d_t x)(d_t^2 x) = F(x)d_t x. \tag{2.5}$$

It can now be recognized that both sides can be rewritten as derivatives

$$d_t\left(\frac{m}{2}(d_t x)^2\right) = d_t \int^{x(t)} F(x')dx' = -d_t V(x(t)), \tag{2.6}$$

where the quantity $V(x)$, being the primitive of $F(x)$, has been introduced. Note that any integration constant in $V(x)$ does not play a role, as the time derivative removes it immediately. This primitive is called the potential, which generates the force. The appearing minus sign is a matter of convention.

It will happen very often that not the force, but the potential is known. The force can then be obtained by

$$F(x) = -d_x V(x), \tag{2.7}$$

i. e. the derivative of the potential is the force.

This again shows that any constant terms in $V(x)$ do not play a role, and can be chosen at will. Such a kind of arbitrariness seems to be at first quite astonishing, since nature should be somehow uniquely determined. There is, however, a deeper reason behind this arbitrariness, which will become evident later. However, it is best to postpone a deeper discussion of it until more conceptual progress has been made.

[2]Because the earth is not exactly spherical.

Since on both sides of (2.6) total derivatives appear, the equation can be integrated to yield

$$\frac{m}{2}(d_t x)^2 = E - V(x), \tag{2.8}$$

where E is an integration constant. Note that this integration constant, which emerges from a time integration and not a spatial integration, cannot be dismissed. Its relevance will be discussed in more detail below. Before doing this, it is possible to solve this equation by separation of variables, yielding an implicit result

$$t - t_0 = \int_{x(0)}^{x(t)} \frac{dx'}{\sqrt{\frac{2}{m}(E - V(x'))}}, \tag{2.9}$$

which after evaluation of the integral can then be solved for $x(t)$ to get the final solution. This solution is parametrized by the constant E and the initial condition $x(0)$. Thus, these are taking the role of the two initial conditions of position and speed at time zero used before.

While so far only the indefinite integral has been used, it is also possible to use the definite integral

$$W_{12} = \int_{x_1}^{x_2} dx F(x) = V(x_1) - V(x_2). \tag{2.10}$$

This quantity is the necessary integrated force to move something between the two points x_1 and x_2. It is therefore called the work which has (to) be(en) done. It corresponds to the difference in potential. The potential therefore describes the amount of work which can be done, and actually doing some work is reducing the potential. Conversely, investing work raises the potential. Thus, the name potential. The amount of stored work for a particle at a position x with respect to zero potential is called the potential energy. The rate of change of the work

$$P = \frac{dW}{dt} = \frac{d}{dt}\int dx F = \int F \frac{dx}{dt} dt = \frac{d}{dt}\int (Fv) dt = Fv$$

is called power.

If in equation (2.8) the potential is set to zero, the constant E is entirely giving by a quantity obtained from the motion of the particle. On the other hand, solving for E of (2.8) yields

$$E = \frac{m}{2}(d_t x)^2 + V(x). \tag{2.11}$$

For a particle at rest, E is then the potential energy. Thus, in general, E is called the total energy, combining the potential energy $V(x)$ and a contribution from the motion of

the particle $m(d_t x)^2/2 = p^2/(2m)$ which is called the kinetic energy, often abbreviated as T, and where p is again the momentum.

Because of the general solution (2.9), the sum $E - V$ of a particle must be positive at every point in space, since otherwise there is no solution. Thus, a particle needs to have a positive or zero kinetic energy. Furthermore, when $E = V(x)$ for some point x, the kinetic energy has to vanish, as it is a positive quantity. The second derivative may not vanish, and therefore the particle does not necessarily stops there. If the potential further increases to one side, it will actually be deflected there, a so-called inflection point of the movement. Note that the prediction of this behavior did not need the solution of the equation of motion. In fact, the energy will become a very convenient tool to simplify calculations later.

One of the probably most fundamental statements in mechanics is that the energy is x-independent, which follows from (2.8), but also time-independent for conservative, i. e. t-independent, potentials. This follows from (2.5)

$$d_t E = d_t \left(\frac{m}{2} (d_t x)^2 + V(x) \right) = m(d_t x)(d_t^2 x) - F(x) d_t x = 0. \tag{2.12}$$

If the force depends on time and/or speed, i. e., it can be written as

$$F(x) = d_x V(x) + f(x, d_t x, t),$$

the energy changes with a rate of $f(x, d_t x, t) d_t x$.

Before continuing, as a brief remark the situation in more than one dimension is a little more involved. In general, any potential $V(\vec{x})$ will create a force field

$$\left(\vec{F}(\vec{x}) \right)_i = -\partial_{x_i} V(\vec{x}),$$

as can be seen by performing the steps above for every component. In the example of the point particle in section 2.3, the potential for the constant force is $\vec{F}\vec{x}$.

Again, in one dimension any force, which is time and speed independent, has a potential. This is not so simple in more than one dimension. To test, when this is possible, note that if a potential exists (2.10) and its generalization to three dimensions necessarily imply that for any trip starting and ending at the same position $W_{11} = 0$, no matter the path. It can be shown that this is mathematically equivalent to

$$\int_{\mathcal{C}} d\vec{x} \vec{F}(x) = 0,$$

for any closed curve $\mathcal{C}$. There is a powerful theorem in functional analysis, which guarantees

that this is equivalent to the requirement[3]

$$\epsilon_{ijk}\partial_{x_j}F_k = 0$$

for all i, which is much simpler to check than the integral condition. However, this equivalence is only true, if the force is continuously differentiable, or at least if the curve can be contracted to a point without crossing any singularities. This condition also implies

$$\partial_{x_i}\partial_{x_j}V(\vec{x}) = \partial_{x_j}\partial_{x_i}V(\vec{x}),$$

for any combination of i and j, and thus requires that the potential is (at least) twice continuously differentiable. Note that a potential can always be used to find a force, but this force may not everywhere be well-defined, if the potential is not always continuously differentiable.

2.6 Harmonic oscillator

Probably the most important problem in mechanics is the harmonic oscillator. In its simplest form in one dimension, this is the situation if the force is negatively proportional to the distance of the particle from the origin, i. e.

$$d_t^2 r(t) = -\frac{\alpha}{m}r.$$

E. g., a spring, obeying Hooke's law, will create such a force. The associated potential is $\alpha r^2/2$.

The solution of this equation requires two functions which, up to a negative constant, will turn into itself when differentiating twice, to yield the two solutions. Such functions are given by sine and cosine. Therefore, a solution is

$$\begin{aligned} r(t) &= r_0\cos(\omega t) + \frac{v_0}{\omega}\sin(\omega t) \\ \omega &= \sqrt{\frac{\alpha}{m}}, \end{aligned} \tag{2.13}$$

where the two initial conditions r_0 and v_0 are used to fix the position and speed at time $t = 0$, respectively. Thus, after a time $t = \omega/(2\pi)$ the particle is again at the same place and moving at the same speed as at $t = 0$. Such a motion is called periodic, with period $\omega/(2\pi)$.

[3]Note that here and hereafter the Einstein convention is used that over any pairs of indices in a given term a sum is performed over their full range, if not stated otherwise.

Another interesting feature is obtained when considering the more general case of a driven accelerator with friction, described by the equation of motion

$$d_t^2 r(t) + \frac{\beta}{m} d_t r(t) + \frac{\alpha}{m} r = b \sin(\sigma t + \phi). \tag{2.14}$$

To solve this equation, it is best to start with the homogeneous case first, i. e. with $b = 0$. Motivated by the result for friction in section 2.3, it is suggestive to make the ansatz $\exp(ct)$, and insert it into the equation of motion. This yields a quadratic equation for c with the solutions

$$c_\pm = \frac{1}{2m}\left(-\beta \pm \sqrt{\beta^2 - 4\alpha m}\right).$$

In general, $c_\pm$ can be a complex number, but of course r is real. Thus, the real part needs to be taken. The solution for the homogeneous case will then be

$$r_0(t) = \Re\left(\left(\frac{c_- r_0 - v_0}{c_- - c_+}\right) e^{c_+ t} + \left(\frac{c_+ r_0 - v_0}{c_+ - c_-}\right) e^{c_- t}\right).$$

The resulting path then depends on the relative sizes of the involved constants. There are three distinct cases, depending on the value of $c_\pm$.

If $\beta^2 < 4\alpha m$, the argument of the squareroot becomes negative, and the $c_\pm$ are complex. If β is positive, this corresponds to a superposition of an exponential damping, and an oscillation, due to Euler's formula. Thus, after an initial stage, the motion becomes that of an ordinary harmonic oscillator, though with a frequency depending on the interplay of damping and the force yielding a exponentially decreasing amplitude. If β is negative, however, the movement becomes exponentially increasing. This happens, if the friction term enhances motion. Thus, such a term destabilizes the motion.

If the argument exactly vanishes, the movement will be exponentially damped, and degenerate. In this case, the time-dependence can be factored out, and the denominator vanishes.

If the argument is positive, the solution c_+ will again yield an exponentially increasing motion, while c_- an exponentially decreasing one. It depends on the initial conditions, which one will eventually win. However, there is no oscillatory behavior.

Note that the friction force can exponentially increase the motion. But this is no longer really friction, as it provides energy to the system.

If the equation of motion (2.14) b no longer vanishes, the result is

$$\begin{aligned} r(t) &= r_0(t) - d(t) + d(0) \\ d(t) &= b\frac{m}{(m\sigma^2 - \alpha)^2 + \beta^2\sigma^2}\left(\beta\sigma\cos(t\sigma + \phi) + (m\sigma^2 - \alpha)\sin(t\sigma + \phi)\right), \end{aligned}$$

which exhibits again the structure of being the solution to the homogeneous equation plus an explicit solution of the inhomogeneous equation. The important addition is the driving

term d, since for r_0 just the previously stated behavior emerges again. This can be seen to be the necessary inhomogeneous part for the solution by entering it into the solution. It is a reasonable ansatz, as it has the same form as the driving term, but in the end, this is here an educated guess.

If $\beta = 0$, the amplitude diverges if $m\sigma^2 = \alpha$. Thus, close to this value, the amplitude grows beyond any bounds, and therefore totally dominates the result. This is called a resonant behavior, or just a resonance. Physically, what happens is that the external force applies just exactly such that it always accelerates the movement a little bit further, instead of damping it occasionally. Thus, the motion grows beyond any bound.

If $\beta \neq 0$, there is no divergence, but still a strong enhancement at $m\sigma^2 = \alpha$. This is called a damped resonance. Considering the prefactor as a function of α, the width is defined as the amount of deviation of α from the critical value $\alpha = m\sigma^2$ where the function drops to a half[4]. This width is $2(\beta^2\sigma^2 + 2m^2\sigma^4)^{1/2}$.

Similar (damped) resonances are a feature exhibited by many physical system. Understanding it well for the harmonic oscillator is therefore understanding a prototype for a multitude of physical phenomena, from celestial mechanics to particle physics.

Note that ϕ only acts as a phase, i. e. it modifies only to some extent the relative behavior of the two oscillations, if also r_0 shows an oscillatory behavior.

More interesting is that the maximal amplitude is reached no longer when r_0 reaches it maximum, but delayed by a so-called phase shift γ. The value of this phase shift depends on the value of α when all other quantities are fixed. It is trivially zero for $\alpha = 0$, as then all oscillation is just given by the driving force. It can be shown to be negative for all values of α, becoming $-\pi/2$ at the resonant frequency, and tends to $-\pi$ for α to infinity. If β is zero, it actually jumps from zero to $-\pi$ at the resonant frequency, and this jump becomes smoothed out the larger β becomes.

2.7 Central potential

Generically, forces on particles of the type

$$\vec{F} = f(\vec{r}, d_t\vec{r}, ..., t)\vec{r}, \tag{2.15}$$

i. e. having the same direction as the vector $\vec{r}$, are called central forces. These forces assume that the source of the force is at the center of the coordinate system, and that the direction of the force is towards (or radially away from) this center The further crucial assumption

[4]Sometimes also the value is determined when it drops to $1/e$ or to $\ln 2$, depending on the context. These are then different numbers than for the factor 2.

is that there are no lateral forces, so that the prefactor only involves the absolute values, $f = f(|\vec{r}|, |d_t\vec{r}|, ..., t)$. These forces are probably the most important type, especially in astrophysics, since gravity, (2.4), is of this type. Their structure also entail many further consequences, which are prototypical. They shall therefore be treated in some detail.

As a first step, it is useful to find the condition under which there is a potential for these forces. The first condition is that the function f is at most depending on $\vec{r}$, to avoid any dissipation. A conservative central force must then be of the form $f(r)\vec{r}$, where $r = |\vec{r}|$ for brevity. There is also a second, more technical criterion, which will be skipped here.

2.7.1 Angular momentum

For the following, it is useful to first introduce a further concept, angular momentum. Take the equation of motion (2.2), and form a vector product with $\vec{r}$, yielding

$$(\vec{M})_i = \epsilon_{ijk} r_j F_k = m\epsilon_{ijk} r_j d_t^2 r_k = d_t(m\epsilon_{ijk} r_j d_t r_k) = d_t(\epsilon_{ijk} r_j p_k) = d_t l_i. \tag{2.16}$$

The so defined quantity $\vec{l}$ is called the angular momentum, whose time evolution is determined by the torque $\vec{M}$. This is called the angular momentum law. If there is no torque, because either the force itself vanishes or has a vanishing vector product with the position, the angular momentum is conserved, as its time derivative vanishes.

However, the actual value of the angular momentum is not intrinsic to the system, but depends on the coordinate system. This can be seen from the free particle. Since the force vanishes, so does the torque. But the position vector and the speed need not be parallel, but can be, depending on the coordinate system, though it is always constant. Thus, a statement about the value of the angular momentum requires to also provide the coordinate system. The rate of change, however, is physical, as it is uniquely given by the torque.

For a central force, (2.15), the torque always vanishes, and the angular momentum is conserved. This is the reason why it will be so relevant in this section.

A useful first consequence is the geometrical interpretation of the angular momentum, which can be derived from its magnitude

$$\frac{1}{2m}|\vec{l}| = |\epsilon_{ijk} r_i d_t r_j e_k|. \tag{2.17}$$

The right hand side has dimension area per time. Geometrically, it is the area which the vector pointing from the origin to the particle covers per unit time. If angular momentum is conserved, this area is constant. As will be seen, e. g. the planets fulfill this so-called area theorem, or better known as Kepler's second law.

2.7.2 Effective potential

For a conservative central potential both the energy (2.11) and the angular momentum (2.16), and especially its length $l = |\vec{l}|$, are conserved. This can be used to simplify the solution of the equations of motion.

First, because the angular momentum is conserved the speed and the position are coplanar in a temporally constant plane, since

$$\vec{r}\vec{l} = d_t\vec{r}\vec{l} = \vec{0}, \tag{2.18}$$

since the cross product is perpendicular to its components. Thus, it is possible to restrict the description to a plane. Due to the radial symmetry it is convenient to chose angular coordinates for the description, i. e. a distance from the origin r (on which the potential depends) and an angle ϕ, thus $x = r\cos\phi$ and $y = r\sin\phi$.

This implies

$$\begin{aligned} E &= \frac{m}{2}(d_t\vec{r})^2 + V(r) = \frac{m}{2}((d_t r)^2 + r^2(d_t\phi)^2) + V(r) \\ &= \frac{m}{2}(d_t r)^2 + \frac{l^2}{2mr^2} + V(r) = \frac{m}{2}(d_t r)^2 + V_e(r). \end{aligned} \tag{2.19}$$

This result has a number of interesting implications. First, the energy is entirely determined by the radius r as a function of time, i. e. the distance from the origin. Its angular position is not relevant. Secondly, the situation is analogous for the one-dimensional particle of section 2.5, but with the modified, the so-called effective, potential V_e, rather than the original potential V.

2.7.3 Planetary motion

This previous result is fully general. It is, however, quite interesting to study the case of $V(r) \sim 1/r$ in more detail. This particular case is of special importance as there are two situations in which it arises.

One is the electrostatic force, where it takes the form

$$V(r) = -\epsilon\frac{q_1 q_2}{r}, \tag{2.20}$$

where ϵ is a constant depending on the system of units chosen, and this is the potential between two electric charges having electric charges q_1 and q_2 respectively.

The other one is Newton's law of gravity,

$$V(r) = -\gamma\frac{mM}{r}, \tag{2.21}$$

and describes the potential between two bodies of masses m and M, and γ is again a system-of-units-dependent constant, called Newton's constant.

In both cases it is tacitly assumed that one of the bodies resides at the center of the coordinate system. Otherwise, r has to be replaced by $|\vec{r}_1 - \vec{r}_2|$, the distance between both bodies, and both affect each other as given by Newton's third law. In the following it will be furthermore assumed that the body at the origin will not move. This is an excellent approximation if one of the bodies is much heavier than the other. This is e. g. true if the two bodies are the sun and the earth, a satellite and the earth, or an atomic nucleus and an electron. If this is not true, this becomes a two-body problem which, in this particular case, can actually be solved only with slightly more effort. However, since this only obscures a few things right now, this will be postponed until later, and the approximation will be made.

It should be noted that, though this is called Newton's law of gravity, there is a-priori no casual connection between this law and Newton' three laws (except that they have been discovered by the same person, Newton). The first three laws are about inertial masses. This law is about the gravitational masses. It is, as noted above, only an experimental result that these two masses are identical.

Though this problem can be solved with the methods of the previous section 2.7.2, it is useful to take the opportunity to demonstrate also a different, but equally useful, way of solving the problem: Another replacement of variables.

To start out, define $s = 1/r$. Though s is a function of the time t, it is also possible to consider it rather as a function of the angle ϕ in the same two-dimensional coordinate systems as in section 2.7.2. Then the chain rule yields

$$\frac{ds}{d\phi} = \frac{d\frac{1}{r}}{dt}\frac{dt}{d\phi} = -\frac{d_t r}{r^2}\frac{mr^2}{l}$$

where the angular velocity

$$d_t\phi(t) = \frac{lr^2}{m}, \tag{2.22}$$

has been used, which follows from (2.18). Inserting this into (2.19) yields

$$E = \frac{l}{2m}\left((d_\phi s)^2 + s^2\right) + V(s).$$

Though it may appear tempting to directly solve this equation, it is in this case better to take a detour, born from hindsight and experience.

Differentiating this equation a second time with respect to ϕ yields

$$0 = \frac{l^2}{2m}\left(2(d_\phi s)(d_\phi^2 s) + 2s d_\phi s\right) + (d_s V(s))(d_\phi s) = d_\phi s\left(\frac{l^2}{2m}\left(2(d_\phi^2 s) + 2s\right) + d_s V(s))\right).$$

This equation has one solution with $ds/d\phi = 0$, i. e. r is constant as a function of ϕ. This is a perfect circular orbit. The term in parentheses is also a linear differential equation of the second kind

$$0 = \frac{l^2}{2m}\left(2(d_\phi^2 s) + 2s\right) + d_s V(s).$$

Furthermore, the derivative of V with respect to s is just a constant, and therefore the equation is

$$d_\phi^2 s + s = \frac{\gamma m^2 M}{l^2}. \tag{2.23}$$

This shows the motivation for this approach. Out of a non-linear differential equation the task has been reduced to the task of solving two linear differential equations, which is far simpler in general.

But equation (2.23) is already known. It is one particular version of the harmonic oscillator equation of section 2.6. Thus, the solution can be read off directly by replacing t in section 2.6 by ϕ, and yields

$$s(\phi) = \alpha \sin\phi + \beta \cos\phi + \frac{\gamma m^2 M}{l^2},$$

where the constants α and β have to be fixed by initial conditions. A convenient choice is to select $\alpha = 0$, which implies that the minimal value of r is at $\phi = 0$, if $\beta > 0$. The solution is then given by

$$\begin{aligned} s &= \frac{1}{r} = \beta\cos\phi + \frac{\gamma m^2 M}{l^2} = \frac{1}{k}(1 + \epsilon\cos\phi) \\ k &= \frac{l^2}{\gamma M m^2} \\ \epsilon &= \beta k \end{aligned}$$

where the last rewriting is convenient to characterize the geometric properties of the solution. These are just cone cuts, and for ϵ smaller, equal, or greater than 1 this describes an ellipsoid, a paraboloid, and a hyperboloid. Note that the case $\beta = 0$ returns exactly the circle case. Thus, ϵ is also called the eccentricity of the orbit. Thus, the quality of the solution is entirely characterized by ϵ.

The case of $\epsilon < 1$ (with $\epsilon = 0$ being the circle) is just the ordinary case of a planetary orbit around the sun. This can be seen from the fact that $(1 + \epsilon\cos\phi)$ remains bounded, and hence so does r. At the same time, the motion is periodic.

If $\epsilon = 1$, then the motion is no longer periodic, since now s can vanish for $\phi = \pi$. Still, ϕ can still reach all values, especially zero, and therefore the movement is a parabola. It is an ellipsoid of which one of the main axes goes to infinity.

If $\epsilon > 1$, then there is no longer a positive solution for r for all values of ϕ. It can therefore no longer be an ellipsoid. This implies that not all ϕ values are allowed, and the path of the particle becomes a hyperbola: The particle tries to approach the central one, but is deflected before it can reach it. The point of closest approach is given by $r = k/(1+\epsilon)$.

It is quite instructive to return to the energy of the particle. Since the kinetic energy is always positive, it follows from (2.19)

$$E \geq \frac{l^2}{2mr^2} - \frac{\gamma m M}{r}.$$

If the energy is greater than zero, this equation has a solution for all r greater than some limiting r_0, which vanishes for $l = 0$. This implies that a particle with positive energy corresponds to the hyperbola. A non-vanishing angular momentum increases the point of closest approach r_0. Therefore it is said that the angular momentum creates an angular momentum barrier. If E is exactly zero this is the smallest possible value for which it is still possible to have an infinite distance. It is therefore corresponding to the parabola. If the energy becomes negative, the equation has no solution also for r larger than some value r_1, the movement is therefore bounded. This corresponds again to the closed orbits. Since in general $r_0 \neq r_1$, the orbit is an ellipsoid. For the smallest possible value of the energy, $r_0 = r_1$, and the motion becomes the circular orbit.

Thus, even without solving the system, it would have been possible to determine many qualitative features of the particle's motion just by studying (2.19). Such studies therefore are an extremely important tool. If, e. g., the potential would have taken the value $V(r) = r$, a similar study would immediately yield that the particle can never move to $r = \infty$.

In the whole discussion, the time-dependence was not an issue. If the time-dependence is actually of relevance, it can be obtained from (2.22). However, the actual solution is rather involved.

Still, from the results the famous laws of Kepler can be deduced for planetary, i. e. bounded, motion. Since the heavier mass is at the origin the planets move along ellipsoids with the sun at one of the focus points. The relation (2.17) is actually already Kepler's second law if angular momentum is conserved, as it is here.

Kepler's third law states that the ratio between the time needed for a complete orbit T squared and the third power of the larger axis is constant. It appears that this would require to have the time dependence, but this is not the case. The area of the ellipse is, using (2.17),

$$\pi a b = \frac{T l}{2m}$$

But the value of b is just given by the minimal value of s at $\pi/2$. Thus follows

$$\frac{T^2}{a^3} = \frac{4\pi^2}{\gamma M},$$

which completes the proof.

There is one more feature, which may appear odd. While angular momentum conservation requires the movement to be within the plane, it by no means requires that the generalized ellipsoids, on which the particle moves, are constant in space. In fact, if using a potential like $1/\sqrt{r}$, this would not happen. The reason is that there is a third conserved quantity, besides the angular momentum and the energy, the so-called Runge-Lenz vector,

$$\vec{p} \times \vec{L} - m\gamma \frac{\vec{r}}{r}.$$

While its explicit expression is not particularly helpful, nor the proof for conservation, it can be shown that such a conserved quantity only exists for very few potentials, so-called maximally superintegrable potentials, to which (2.20-2.21) belong. The other in three space dimensions is the harmonic oscillator. Such potentials have the featured that there exists $2(d-1)$ conserved quantities. Note that certain scalar products and vector products between the Runge-Lenz vector and the angular momentum vanish or yield the energy, leaving only five conserved quantities. As a consequence, the six degrees of freedom of position and momentum can be reduced to a single one using the five conserved quantities. The disturbance due to other planets will reduce the number of conserved quantities, yielding (slow) changes of the ellipsoids over time.

2.8 Galileo group

The concept of an inertial system was of central importance in formulating Newton's laws in section 2.2. It is worthwhile to reinvestigating this concept a little more.

The basic starting point is to require that there are coordinate systems in which Newton's first law is valid. Especially, assume that there is one coordinate system in which there are no forces acting on a particle, and it therefore moves with constant speed. If this is the case, then it is possible to define coordinate systems, which are not inertial systems, by having them being accelerated with comparison to the first. There are also other inertial systems, e. g. some which move at constant speed with respect to the first one. In those, also the particle moves at fixed speed.

To systematize this, it is useful to work in the language of linear algebra. Two coordinate systems shall have coordinates $\vec{r}$ and $\vec{R}$. What will be assumed is that the time in

both frames, t and T, transforms into each other trivially

$$t = T. \tag{2.24}$$

Thus also the time differentials are the same, $dt = dT$. It is this part, which will be lifted when introducing special relativity in chapter 4. It is axiomatic in Newtonian' mechanics, and reflects a view of space and time where time is independent of space, sometimes called an absolute time.Especially, this allows for the introduction of a universally valid time coordinate for everyone.

Start with two inertial systems. Then if for a particle of mass m $md_t^2\vec{r} = 0$ is valid, then so must be $md_t^2\vec{R} = 0$. From this follows that, up to a rotation, the most general possibility of transformations between two inertial systems is

$$\vec{r}(t) = \vec{v}_0 t + \vec{r}_0 + \vec{R}(t),$$

parametrized by the two vectors $\vec{v}_0$ and $\vec{r}_0$. This is called a Galileo transformation. Note that the forces on a particle are due to (2.2) then identical in all inertial systems, and all coordinate system reachable by Galileo transformations. However, the explicit form may differ, especially if friction is involved.

If two consecutive Galileo transformation are performed, they are equivalent to a single Galileo transformation with

$$\vec{v}_0 = \vec{v}_0^1 + \vec{v}_0^2 \tag{2.25}$$
$$\vec{r}_0 = \vec{r}_0^1 + \vec{r}_0^2. \tag{2.26}$$

and (2.24). Galileo transformations therefore form an (Abelian) group, the so-called Galileo group.

This group can also be extended by including time-independent rotations and reflections. In this case no longer an equality of the forces is true, but the force are still equal up to rotations and reflections.

2.9 Pseudo forces

An interesting situation arises when the condition of inertial systems is relaxed. Allowing for a time-dependent rotation $\Lambda(t)$ and an acceleration $r_0(t)$ yields that the force observed in the other coordinate system is given by

$$\vec{F}(t) = m\left(\Lambda(t)d_t^2\vec{R}(t) + d_t^2\vec{r}_0(t) + 2(d_t\Lambda(t))d_t\vec{R} + (d_t^2\Lambda(t))\vec{R}\right). \tag{2.27}$$

The first term is the force in the original coordinate system, up to a (time-dependent) rotation, and therefore the same as with transformations between inertial frames. The three other terms are different.

The second term corresponds to an acceleration of the whole coordinate frame without rotations. It is therefore a relative acceleration. As a consequence, the particle experiences an additional force due to this acceleration. Since this force has no physical origin but emerges just from a an accelerated coordinate system it is not a true force. It is therefore called a pseudo force. Nonetheless, it must be included when describing the movement of the particle inside an accelerated coordinate system. Sometimes these are also called inertial forces.

The other two terms come from a time-dependent rotation of the coordinate systems with respect to each other. They also create pseudo forces. The first term is called the Coriolis force and the second one the centrifugal force. These appear, e. g., when a particle is moving on the surface of a rotating body like the Earth. Though small in this case, they can be measured.

The origin of these forces can also be understood differently. When considering a force-free particle, it will move along a straight line in its inertial system. In an accelerated frame, this path is no longer of constant speed, and may be bend. To describe the effect, this can either be done by the coordinate transformation, or by the forces necessary to modify the path of the particle, which is essentially given by (2.27). Alternatively, the bending of the path of the particle in the accelerated coordinate system is not due to a movement of the particle. It is, because the observer is moving along a bended path away from the particle.

2.10 Mass distributions and center of mass

2.10.1 General properties

The previously discussed situation with two particles is a special case of having N particles. It is useful to discuss this situation in some detail, as it appears quite often, e. g. when looking at the total solar system.

Let each of the particles have its own mass m_i. Because of Newton's third law the force $\vec{F}_{ij}$ with which particle i acts on particle j, e. g. by gravitation, must obey $\vec{F}_{ij} = -\vec{F}_{ji}$. Of course, a particle does not act on itself, and thus for simplicity $\vec{F}_{ii} = \vec{0}$. There can be also external forces on the individual particles $\vec{F}_i$. For the solar system, this may be the gravitational pull of the galaxy as a whole.

There are thus dN equations of motion, where d is the number of dimensions[5],

$$m_i d_t^2 \vec{r}_i = \vec{F}_i + \sum_j \vec{F}_{ij}.$$

Consider now the sum of all N equations of motion,

$$\sum_i m_i d_t^2 \vec{r}_i = \sum \vec{F}_i, \tag{2.28}$$

and thus the internal forces do not appear. Define furthermore

$$\begin{aligned} M &= \sum_i m_i \\ \vec{R} &= \frac{1}{M} \sum_i m_i \vec{r}_i \\ \vec{f} &= \sum_i \vec{F}_i. \end{aligned}$$

The equation (2.28) then reads

$$M d_t^2 \vec{R} = \vec{f},$$

which looks just like a single-particle equation of motion. Therefore, the center of mass $\vec{R}$ moves like a single particle having the total mass M only under the influence of the external forces. It does not matter how involved the internal forces are for this, they do not appear.

It may also be useful to define analogously a total momentum and angular momentum,

$$\begin{aligned} \vec{P} &= \sum m_i d_t \vec{r}_i = \sum \vec{p}_i \\ \vec{l} &= \sum m_i \epsilon_{ijk} r_i d_t r_j \vec{e}_k = \epsilon_{ijk} R_i P_j \vec{e}_k + \sum_i \epsilon_{ijk} (r_i - R_i)(p_j - P_j) \vec{e}_k = \vec{L} + \sum \vec{l}_i \end{aligned}$$

where $\vec{L}$ is the angular momentum of the center of mass while the $\vec{l}_i$ are the relative angular momentum with respect to the center of mass. Furthermore, if all forces are conservative then in analogy to the single particle a potential energy can be defined, where it useful to define separately the potential between two particles and of the external forces separately. Especially, for a two-particle potential it holds that

$$\begin{aligned} \vec{F}_{jk} &= \vec{e}_i d_{\Delta \vec{r}_i} V_{jk}(\Delta \vec{r}) \\ \Delta \vec{r} &= \vec{r}_j - \vec{r}_k \end{aligned}$$

and thus the forces are obtained by deriving with respect to the connecting vector. This implies that the potentials are superimposed like the forces, i. e. the total potential acting on a particle is the sum of all potentials acting on it.

[5]There may also exist genuine n-body forces, which involve non-separable functions of more than two coordinates. This only complicates the remainder unnecessarily, but can appear in practice.

2.10.2 Two-particle systems

Of particular practical importance is the situation when $N = 2$, i. e. there are two particles only. The center-of-mass is then

$$\vec{R} = \frac{m_1\vec{r}_1 + m_2\vec{r}_2}{m_1 + m_2}.$$

The vector $\vec{r} = \vec{r}_1 - \vec{r}_2$ gives the relative distance between both particles. It is now possible to rewrite the position of the particles in terms of relative coordinates and the center-of-mass coordinates only,

$$\vec{r}_1 = \vec{R} + \frac{m_2}{M}\vec{r} \tag{2.29}$$

$$\vec{r}_2 = \vec{R} - \frac{m_1}{M}\vec{r}. \tag{2.30}$$

The center-of-mass movement is as before determined by the external forces only. The equation of motion for the relative coordinate is

$$d_t^2\vec{r} = \frac{\vec{F}_1}{m_1} - \frac{\vec{F}_2}{m_2} + \frac{\vec{F}_{12}}{m_1} - \frac{\vec{F}_{21}}{m_2}.$$

The dependence on the masses suggests to define the so-called reduced mass

$$\mu = \frac{m_1 m_2}{m_1 + m_2}.$$

In terms of this mass the equation of motion becomes

$$d_t^2\vec{r} = \frac{\vec{F}_1}{m_1} - \frac{\vec{F}_2}{m_2} + \frac{\vec{F}_{12}}{\mu}.$$

If the external forces vanish, what is called a closed system, the equation of motion for the relative coordinate no longer references the external system. Furthermore, it can be shown that also the relative angular momentum takes the form $\vec{l} = \mu\vec{r} \times d_t\vec{r}$. Thus, the equation of motion for the relative coordinate is the same as for a single particle with the reduced mass.

If one particle is much heavier than the other the reduced mass is essentially identical to the mass of the lighter particle. Thus, in this formulation the approximation made in section 2.7.3 becomes more transparent. Since the planet (or satellite) is so much lighter than the sun (the planet), the reduced mass is essentially that of the planet (the satellite). Thus, the equation of motion for the relative coordinate is essentially the one in the case of the fixed heavier particle. However, with the formulation developed here, it is possible, using equations (2.29-2.30), to determine the actual movement of both bodies, but this

requires to solve only the one-particle problem. By direct comparison, this is much simpler: It now only necessary to substitute in all results of section 2.7.3 the mass of the planet by the reduced mass, and the result is automatically the correct one for the two-body problem.

Of course, the substitution can be done for either of the bodies. Thus, both bodies will perform exactly the same type of movement. In particular, for a bounded planetary motion both celestial bodies will have an orbit around the same point, which is the center of mass. Superimposed to the orbital motion of both bodies can then be a movement of the center of mass.

This is also the way how Newton's third law seems to have no impact on the source of the force, as announced in section 2.2.4: The reduced mass is so close to the earth's mass that setting it equal to the earth's mass does not create an appreciable error. On the other hand, the force on the sun is so tiny that its movement is all but negligible. Thus, while in the closed system Newton's third law holds, its violation by setting the external force constant and immutable is a very good approximation. This also justifies to make this approximation in more involved cases, where the backreaction is known to be small, but its details too involved to be included, a situation arising quite often in practice.

It is a unfortunate insight that a similar reduction of complexity is not possible if there is more than two objects involved. The three(and more)-body problem cannot be reduced to a one-body problem, but a full solution is required. E. g. the movement in the solar system or of the solar system around the center of the galaxy is such a three, or more, body problem. However, in these cases it is possible to simplify the problem, see section 2.11.

2.10.3 Scattering

A situation which is of high relevance in many physical application is a closed system such that the two-body potential (or a force) is short-range. This can happen in either of two cases. One possibility is that the potential has a sharp cut-off, $V(\Delta\vec{r})\ \theta(|\Delta\vec{r}-\vec{r}_0|)$, where the value $\vec{r}_0$ belongs to the definition of the potential. The other is that it decays fast enough. What fast enough means is rather context-dependent, but very often requires at least an exponential decay at large distances, $\exp(-|\Delta\vec{r}|/r_0)$, where the scale r_0 is once more characteristic for the potential, but even maybe some kind of power-law dependence $1/|\Delta\vec{r}|$ is sometimes sufficient.

In such a situation it is possible to define a scattering process in the following way. Start out with two particles, which are initially so far separated that they are not interacting.

They are then send towards each other in such a way as their center of mass is at rest[6]. They afterwards escape again to infinity, and stop interacting.

No matter how the details of this kind of interaction is, there are a number of conservation laws, which are necessarily fulfilled, if the interactions are conservative, which is called elastic scattering. Especially, the linear momentum has to be conserved,

$$\vec{p}_1 + \vec{p}_2 = \vec{p}_1' + \vec{p}_2',$$

where the primes denote the situation after the scattering. By construction, the initial relative momentum is zero, $\vec{p}_1 + \vec{p}_2 = \vec{0}$. Therefore, the final total momentum must also be zero. Furthermore, this implies that the direction is arbitrary, and can therefore be selected at will.

Furthermore the energy is conserved. Since in the beginning there is no potential energy, it follows that

$$T = \frac{\vec{p}_1^2}{2m_1} + \frac{\vec{p}_2^2}{2m_2} = \frac{\vec{p}_1'^2}{2m_1} + \frac{\vec{p}_2'^2}{2m_2} = T'.$$

Since $\vec{p}_1 = -\vec{p}_2$, $\vec{p}_1^2 = \vec{p}_2^2 = \vec{p}_1'^2 = \vec{p}_2'^2$.

These conditions enforce that the incoming and outgoing momenta both lie along respective lines. Hence, all movement is contained in a plane, and the coordinate system can be chosen such that all components in the third direction can be chosen to be zero. Thus, there remain only two components for the final momenta free. One is constrained by the energy conservation, leaving only one component. This can be traded in for the relative angle between the incoming momentum and outgoing momentum, θ. Thus, the only freedom left by the kinematics is the relative angle between one of the incoming momenta and one of the outgoing momenta. This also implies that, no matter how complicated the potential in the interaction area, all its consequences are encoded in this one angle, the so-called scattering angle. This scattering angle may, however, also be influenced by the initial momenta and the masses.

Dropping this condition of elasticity allows to transmute energy and momentum in the potential to an energy loss or gain, a so-called inelastic reaction. Though such a situation can be realized, the system is no longer really closed, as energy and momentum are no longer conserved quantities. In this case, the energy balance is

$$T = T' + Q,$$

where the energy transfer to or from the potential Q can be either negative or positive. Still $\vec{p}_1' = -\vec{p}_2'$ and the momentum conservation holds trivially in the center-of-mass system.

[6] This is not necessary but makes the following calculations much simpler.

The final state of the system is then no longer entirely determined by the scattering angle θ, but also by Q. Again, both quantities may depend on the properties of the initial state.

While the above describes accurately the situation for the scattering of two particles, it is often useful to make more statistical statements. Especially, it is very often interesting, how many particles are scattered into which direction per unit of incoming particles. This quantity is called the (differential) cross section and is defined as

$$\sigma(\Omega)d\Omega = \frac{\text{Particles going into } d\Omega}{\text{Total incident particles}}, \tag{2.31}$$

where $d\Omega$ is the solid angle, in three dimensions e. g. $\sin\theta d\theta d\phi$ in the, for this problem particularly suited, spherical coordinates. Integrating the differential cross-section on $d\Omega$ yield the total cross section.

As an example, consider the situation of the gravitational potential of section 2.7.3. Of course, only the non-closed trajectories are interesting, as for the bounded cases no particles are incident or going away[7]. It is useful to define the impact parameter b as the point of closest approach to the center of the potential. It is connected to the angular momentum by

$$l = b\sqrt{2mE} = mv_0 b,$$

where v_0 is the speed of the incident particle at infinity. Because of the symmetry of the problem, the scattering, as it is just one of the non-closed trajectories, is in a plane. Thus, by suitably choosing the coordinate frame, there will be no dependence on the angle ϕ, and the cross section will only depend on the azimuth angle θ.

The number of particles N which are then scattered into an azimuth angle of θ are uniquely[8] determined by the trajectories of section 2.7.3, and thus by their respective impact parameter. In terms of the impact parameters and the number of incident particles equation (2.31) now reads

$$2\pi N b db = -2\pi\sigma N \sin\theta d\theta,$$

where N is the total number of incident particles. This can be rewritten in terms of the cross section, yielding

$$\sigma = -\frac{b}{\sin\theta}\frac{db}{d\theta}.$$

[7]Gravitational capture requires thus another, dissipative, effect to reduce the energy of a particle coming from infinity such that it becomes bounded.

[8]This determinism is lacking in quantum physics, where the actual direction of a single particle is given by a probability. Detecting such probabilities belong to the fundamental experiments with which quantum effects have been detected.

So far, it was not necessary to specify the detailed form of the central potential. Using the form $-\alpha/r$ for the potential yields

$$b(\theta) = \frac{\alpha}{2E} \cot \frac{\theta}{2},$$

and thus

$$\sigma(\theta) = \frac{1}{16} \left(\frac{\alpha}{E}\right)^2 \frac{1}{\sin^4 \frac{\theta}{2}}.$$

This is Rutherford's scattering cross section. Incidentally, the complete quantum calculation for this potential leads to the same result, but this is due to the special structure of the potential, and not true in general.

At first sight, there is one worrisome feature of this result: At small angles, and thus in forward direction, it diverges. This is an artifact of the approximation that the sun (or the gravitational center) only produces the gravitational field, but is otherwise not present. In reality, the particle would collide with the sun, and therefore resolve the structure of it beyond the existence of the gravitational field, and the approximation of a point-like gravity origin breaks down. Of course, deviations will appear already at slightly large than zero angle, because the sun also has a finite extension. Therefore, Rutherford's formula should not be used at small scattering angles. However, it provides a good description at larger angles.

2.10.4 Continuous distribution

So far the situation has been treated that the system is made up of a finite number of individual particles. However, this can actually be generalized to the case where the number of particles becomes infinite, and the system becomes a mass distribution, i. e. a continuous body.

Such a continuous body is characterized by a density $\rho(r)$, a function which determines the amount of mass per unit volume at every point in space. Of course, outside the body this density will be zero. The mass concentrated in some fixed volume V is then given by

$$M_V = \int_V d^3\vec{r}\rho(r).$$

Replacing the finite volume by the whole of space provides the total properties of the body.

The quantities from section 2.10.1 can then be readily generalized to yield

$$\begin{aligned} M &= \int d^3\vec{r}\rho(r) \\ \vec{R} &= \int d^3\vec{r}\vec{r}\rho(r) \\ \vec{P} &= \int d^3\vec{r}(d_t\vec{r})\rho(r), \end{aligned}$$

yielding the total mass, the center of mass and the total momentum of the body, respectively[9]. Other quantities are obtained similarly. Most features of continuous mass distributions are therefore just a straightforward generalization of a system of point particles.

2.10.5 Moment of inertia

There is one aspect where the generalization to a mass distribution leads to an interesting new concept.

Consider a body which rotates, but for the sake of simplicity does not translate, around a fixed axis. Furthermore, the axis should be going through the center of mass of the body. In this case the angular velocity is constant and given by $\vec{\omega} = \omega\vec{e}_z$, where the coordinate system has been chosen to be such that the rotation axis coincides with the z-axis.

The speed of rotation at some point $\vec{r}$ inside the body is then given geometrically by

$$d_t\vec{r} = (\epsilon_{ijk}\vec{\omega}_i\vec{r}_j)\vec{e}_k,$$

and thus lies entirely in the x-y plane. The total kinetic energy is then given by

$$\begin{aligned} T &= \frac{1}{2}\int d^3\vec{r}\rho(r)|d_t\vec{r}|^2 = \frac{1}{2}J\omega^2 \\ J &= \int d^3\vec{r}\rho(\vec{r})\left|\epsilon_{ijk}\frac{\vec{\omega}_i}{\omega}\vec{r}_j\vec{e}_k\right|^2, \end{aligned} \tag{2.32}$$

where the quantity J is called the moment of inertia. Note that its definition is also valid if the z axis is not the rotation axis.

This moment of inertia is a quantity which characterizes the reaction of a body to a rotation around a given axis. E. g., for a homogeneous cylinder of radius R, height h, and density ρ_0 rotating around its symmetry axis it is

$$J = \int d^3\vec{r}\rho(r)(x^2+y^2) = \rho_0\int_0^R d\rho\int_0^{2\pi} d\phi\int_0^h dz\rho^3 = \rho_0\frac{2\pi hR^4}{4} = \frac{1}{2}MR^2,$$

[9] If the mass distribution would be time-dependent, the last formula needs to be generalized and can be obtained from $d_t\vec{R}$.

where in the last step the mass of the cylinder has been used. For comparison, if the cylinder would rotate around an axis orthogonal through its symmetry axis but going through its center of mass, the moment of inertia would be

$$J = \frac{1}{4}MR^2 + \frac{1}{12}Mh^2.$$

Thus, the amount of kinetic energy in the rotation of an extended body is influenced not only be the mass distribution, but also by the relative alignment of the rotation axis with respect to the body.

So far, it was assumed that the rotation axis was going through the center of mass. This will not be the case in general. The consequences of this is described by the theorem of Steiner. In this case, the positions appearing in (2.32) are with respect to the actual rotation axis. Rewriting this position as $\vec{r} = \vec{R} + \vec{s}$, where $\vec{R}$ is the (orthogonal) displacement of the center of mass with respect to the axis of rotation and $\vec{s}$ is the distance to the center of mass. This yields

$$\int d^3\vec{s}\rho(\vec{s}) \left|\epsilon_{ijk}\frac{\vec{\omega}_i}{\omega}(\vec{R}-\vec{s})_j\vec{e}_k\right|^2 = J + \frac{M}{\omega^2}(\vec{\omega}\times\vec{R})^2,$$

where it has been used that

$$|(\vec{R}-\vec{s})\times\vec{\omega}|^2 = (\vec{R}\times\vec{\omega})^2 + (\vec{s}\times\vec{\omega})^2 - 2\left((\vec{R}\vec{s})\omega^2 - (\vec{\omega}\vec{s})(\vec{R}\vec{\omega})\right).$$

The fourth term vanishes since $\vec{R}\vec{\omega} = 0$ by construction. The third term vanishes upon integration, since by construction this is the position of the center of mass for the center of mass being at the displacement position - the vector $\vec{s}$ is measured with respect to the center of mass. I. e. the actual moment of inertia has two components. One describes the rotation of the body as a whole, and the other component the rotation of the body in itself. Thus, again, the two movements separate.

2.11 Perturbation theory

Consider again the one-dimensional case. An interesting situation occurs if a potential $V(x)$ has an extremum at some point x_0. According to (2.7), the force then vanishes.

Expand now the potential around this minimum in a Taylor series, yielding[10]

$$V(x) \approx V(x_0) + d_xV(x)|_{x=x_0}(x-x_0) + d_x^2V(x)|_{x=x_0}(x-x_0)^2 + \mathcal{O}\left((x-x_0)^3\right).$$

[10]If the potential is non-analytic in x_0, things become highly non-trivial. This will not be considered for now.

The first term is just a constant and does not influence the force and hence the motion of the particle. The second term vanishes, since by construction the potential has an extremum. Thus remains the higher-order terms.

In the next step, the situation is considered if the particle is somehow not at x_0, but rather at some place close to x_0. If this is the case, $x - x_0$ is small, and all terms in the Taylor series becoming increasingly smaller. It is thus possible to consider the particle to only be slightly perturbed, and the movement can be analyzed in perturbation theory, i. e. by discussing the effects of the terms in the Taylor series order by order.

To zeroth order, the potential is constant, and therefore the particle remains at the position x, or moves freely if it has a non-zero initial speed. The next order does not change it, as the second term in the Taylor series vanishes as discussed before.

In the next order, only the quadratic term contributes. But such a potential is just the one of the harmonic oscillator in section 2.6. The resulting movement is known and depends on the sign of $d_x^2 V(x)|_{x=x_0}$. If the sign is positive, the motion of the particle to this order in perturbation theory will be an oscillation around the minimum. Higher orders in the perturbative series will not alter this behavior qualitatively, if just the initial x is close enough to x_0. Thus in a potential which has a positive second derivative, and thus a minimum, a particle close to the minimum will perform harmonic oscillations around the minimum. Placing it at x_0 without initial velocity will actually leave it at rest, and any small movement outside the minimum will only yield a small oscillation. Thus the equilibrium position x_0 is called stable. Especially, if there should appear any damping of the movement it will be returning to and staying at the equilibrium position eventually.

If the second derivative is actually negative, however, the particle will move exponentially away from the position x_0. Such an equilibrium position is therefore called unstable. The ultimate fate of the particle is then determined by higher orders in perturbation theory.

There remains the situation if the extremum is actually a saddle point, i. e. the second derivative vanishes. Then again the higher orders will determine the fate of the particle. However, any movement will not be exponential, and therefore the development is much slower. Thus, this situation is called a metastable equilibrium.

Similarly, this implies that a particle which is positioned at the equilibrium position of a potential will react to any infinitesimal external force in one of three ways. If the equilibrium is stable, it will oscillate around it or, if damping is present, return to the equilibrium position. If it is metastable, it will move somewhere, but will eventually be influenced by higher orders of the perturbative series. If it is unstable, it will move away from the equilibrium position.

In this consideration the initial speed has not been an issue. If the speed is so large that the kinetic energy is of a similar size, or larger, than any given order in the perturbative series, the movement is no longer fully described by this order, and higher orders have to be taken into account. If the kinetic energy is small, this will only modify the initial conditions of the movement of the particle to this order. The reason is that the full potential has possibly not the infinite rising barrier of the harmonic oscillator, but may flatten out. Then the kinetic energy is sufficient to escape this potential well (or feel the distortion away from harmonic), and thus a description to this order of perturbation theory makes no sense.

These concepts can be equally well applied to the situation with multiple extrema. When generalizing to more dimensions, a subtlety arises. In this case, an equilibrium position may have different characteristics in different directions. Then the above said has to be weighted with the direction of the external displacement or perturbation. If the potential is stable/metastable/unstable in this direction, the particle will act accordingly, but may show a different behavior in a different direction.

Besides the possibility to apply perturbation theory for equilibrium positions, perturbation theory can also be used if the potential can be Taylor expanded in a different way. E. g. for the case of two planets around the sun, the gravitational interaction between the two planets is small compared to the sun. Thus with the positions $\vec{r}_1$, $\vec{r}_2$ of the two planets and $\vec{r}_s$ of the sun, as well as their respective masses m_1, m_2, and m_s, it is possible to write the potential for one of the planets as

$$V_1(r) = \gamma \frac{m_1 M_s}{|\vec{r}_1 - \vec{r}_s|} + \mathcal{O}\left(\frac{m_1}{m_s}, \frac{m_2}{m_s}\right),$$

that is, the corrections are small since they are suppressed by powers in m_i/m_s. Thus, the dominating part is a two-body problem, and the remainder can be treated in a perturbative fashion rather accurately. the same is true for the movement of the solar system around the milky way.

Chapter 3

Lagrangian mechanics

In the previous chapter the basic concepts of mechanics have been formulated. All of the basic physical mechanisms have been collected. The only thing in which other problems of mechanics would differ is by other versions of the forces or, if existing, potentials. All the conceptual ideas are there.

The following chapters do not aim at any more elaborate problems with more complicated potentials and/or systems. Rather, this and chapter 5 aims at a reformulation of mechanics. While this provides at first sight no new physical systems, this will emphasize concepts which act on a more fundamental level. Furthermore, these two chapters provide the basis for the generalization of mechanics to quantum mechanics. Finally, in more complicated situations the methods to be discussed in the following turn out to be superior to the ones introduced so far.

The downside of this is that the motivation for many of the ideas in the following is not at all obvious at first. In fact, without dealing either with much more complex problems or with quantum physics most of the following will be rather experienced as a complication rather than a simplification. However, would one introduce the following with suitable complicated problems, the sheer complexity of the problems would overlay everything, and therefore the formalism would become obscure. It is therefore a precarious problem to exhibit the reason behind this formalism.

3.1 Constraints

The first basic insight is the realization that most particle movements are not as simply described as before.

Consider, e. g., a particle moving through a winding tube under the influence of gravity. The particle cannot move in all directions equally, but is confined to a one-dimensional

path through the tube. To have such a path requires that the walls of the tube acts with a force on the particle, on top of the gravitational force. Otherwise the particle would fall through the tube walls, following the pull of gravity. To describe this using Newton's equation (2.2) would require to know the particularities of these forces, which are so-called constraining forces. In general these constraining forces are essentially impossible to determine, which appears to make the solution of the problem using the techniques of chapter 2 essentially impossible.

The necessary insight to deal with this problem is that it is actually not necessary to really know these forces in detail, since the consequences of them are known: The particle remains in the tube. The consequence of that is that the three coordinates, x, y, and z, are no longer independent. Since the particle moves along the tube, there is only one way to go (either forward or backward). So, effectively its movement is reduced to one dimension. Hence two of the coordinates are actually functions of the third, e. g. $x(z)$ and $y(z)$.

Of course, the choice of z was arbitrary. Therefore, it is better to formulate the conditions more generally as a so-called set of N constraints

$$f_i(\vec{r}) = 0$$

where the index i runs over the number of restrictions, in the above example it would be from one to two. That the right-hand side is zero can always be achieved by a subtraction, if need be.

A simpler example would be the movement of a particle which is required to move on a circle with radius R. In this case, there would be two constraints,

$$\begin{aligned} f_1(\vec{r}) &= z = 0 \\ f_2(\vec{r}) &= \vec{r}^2 - R^2 = 0. \end{aligned}$$

The first constraint requires the particle to move in the plane of the circle. The second constraint then further enforces it to move along the circle. This also shows that the constraints are not unique. E. g., $f_2 = |\vec{r}| - R$ would equally well be possible. The number of independent constraints is, however, unique.

Such constraints actually eliminate redundant information. In the previous example, the two additional coordinates did not carry any information relevant to the movement of the particle: These information could be completely reconstructed using the one remaining coordinate as well as the constraints. In fact, the final aim will be to find a transformation, constructed in section 5.8, such that the problem becomes trivial. All non-trivial information will then be contained in the transformation.

Before embarking on this process, there are a few more words to be said about the type of constraints. One is that this is not restricted to single particles, and thus many constraints, involving the properties of many particles, can appear. Also, constraints can include explicitly the time. E. g., a particle which moves in an elevator, which itself moves along the z-direction with constant speed v would have the constraint

$$z - z_0 - v(t - t_0) = 0. \tag{3.1}$$

Also, constraints can involve other quantities, especially speeds or acceleration, or inequalities. In such a case the constraints do not reduce the number of degrees of freedom. All of these cases are relevant in practice, but make the treatment substantially more involved.

3.2 Generalized coordinates

Once the set of constraints are known, it is appears reasonable to indeed make the transformation from the ordinary position space to a new space: Dealing with fewer variables and fewer constraints is (very likely) a technical simplification[1]. This space is known as the configuration space. It is a, as noted above, $dn - N$-dimensional space. The $dn - N$ coordinates of this space are called generalized coordinates. Note that the number of generalized coordinates does not correlate with the number of involved particles, which would be the case for ordinary coordinates.

To qualify as a generalized coordinate two conditions must be met. The first condition is the, somewhat redundant, statement that all coordinates are independent in the sense that no constraints involving them exists. Especially all coordinates can range over the full domain of definition. However, they do not need to be a usual coordinate in the sense that they vary from $-\infty$ to $+\infty$. In the above example of a particle moving on a circle a choice for the generalized coordinate is the angle along the circle, and would thus only vary between 0 and 2π. The second condition is that there is a unique relation between the generalized coordinates and the original coordinates: It must be possible to reconstruct the position of the particles in position space from the configuration space at every instance of time unambiguously.

That said, the generalized coordinates are not necessarily unique, just their number. E. g. in the case of the circle it would be equally valid to choose the double angle or half the angle. Furthermore, as this case already indicates, the generalized coordinates are not

[1] Of course, this is not physics. Nature does not care whether a problem is easy to solve. There is no additional knowledge gained by this procedure, just accessibility. But accessibility can be decisive to gain understanding.

necessarily coordinates in the usual sense. In case of the angle, a quantity measured in meters has been transformed into an angle. Much more different coordinates will be found later. E. g. a generalized coordinate may equally well be a speed or acceleration, or even the time. As long as it is an unconstrained and unambiguous description of the system, it is admissible. Though not necessarily useful.

Of course, once such generalized coordinates have been defined, it is possible to derive them with respect to the time, yielding generalized speeds. After a second derivative the generalized accelerations are created.

Since the generalized coordinates uniquely determine all ordinary coordinates, the system is completely described by them. Thus, when rewriting the equations of motions in terms of the generalized coordinates, thereby eliminating a number of them as they are trivially satisfied by the constraints and having only $dn - N$ in the end, the behavior of this system is then fully described by these new equations of motions. These will be in general again second-order differential equations, requiring therefore $2(dn-N)$ initial conditions, e. g. the values of the generalized coordinates and speeds at some time t_0. Then, the behavior of the system is uniquely characterized by the so-obtained trajectory in the configuration space[2].

As an example for generalized coordinates take the movement of a particle on the surface of the earth, approximated by a sphere. Then there is one constraint

$$\vec{r}^2 - R^2 = 0$$

and thus there are two generalized coordinates. A useful choice are the angles θ an ϕ which are defined as

$$\begin{aligned} \theta &= \cos^{-1} \frac{z}{R} \\ \phi &= \tan^{-1} \frac{y}{x} \end{aligned}$$

and from which the original coordinates can be reconstructed as

$$\begin{aligned} x &= R \sin\theta \cos\phi \\ y &= R \sin\theta \cos\phi \\ z &= R \cos\theta \end{aligned}$$

[2]It appears as if less initial conditions are needed. However, in the original problem already not all initial conditions can be chosen independently, but only if the fulfill the constraints, creating the same number of independent ones. Thus, also the choice of initial conditions is simplified to being again arbitrary.

This result also emphasizes an important statement. While for every set of values of the generalized coordinates θ and ϕ there is a unique set of ordinary coordinates x, y, and z, the reverse is not true. At the poles the map to the configuration space is not unique. However, the trajectory both in position space and configuration space remains unambiguous.

It is an interesting observation that a non-trivial trajectory in configuration space in this case, $(\phi, \theta) = (t, 0)$ translates to a trivial one in position space $(0, 0, R)$, though the translation is unique. This emphasizes that the usual intuition of position space may not be easily translated to configuration space, and vice versa.

Though not at first obvious, the process of introducing generalized coordinates can also be useful even if there are no constraints. This is, e. g., the case if the generalized coordinates are better suited for a particular problem if they are adapted to the symmetry of a problem.

3.3 The principle of d'Alembert

The aim is now to find the equations of motions for the generalized coordinates without knowing the constraining forces explicitly. To do so, it is useful to introduce the concept of virtual movement. This is best done by first refining the definition of a real movement.

A real movement is what happens to a particle when it moves along its trajectory. If time proceeds by an infinitesimal amount dt, the coordinates r, generalized or otherwise, change by an amount $dr = (dr/dt)dt$. The precise value of dr is determined by the equations of motion.

In contrast, a virtual displacement δr is defined to be an instantaneous change of the coordinates, i. e. $\delta t = 0$. The virtual displacement have to be compatible with the constraints. The symbol δ is used instead of d to distinguish both cases. Such a virtual displacement is also called a variation. Such variations are also taken to be infinitesimal for now. Especially, these variations are not associated with a speed, and are not constrained by the equations of motion.

Now, any force $\vec{F}$ has two components, $\vec{F} = \vec{K} + \vec{Z}$. The force $\vec{Z}$ is here the constraining force, i. e. the force which acts such that the constraints are fulfilled. The force $\vec{K}$ are all other forces. In the example of the particle in a tube under the influence of gravity in section 3.1, $\vec{K}$ is the gravitational force and $\vec{Z}$ are the forces exerted by the walls of the tube on the particle. If both forces are known, the standard procedure could be performed.

Now comes the basic step. This will be turned around, and a new basic postulate will be made. This, and further steps, will replace in the end Newton's laws, and they will

ultimately be derived from them backwards. This will take some time.

The first new postulate is that the constraining forces do no net work. Of course, they can do work on a part of the system, i. e. a subset of the particles. But the total work done by them is zero. Though this postulate cannot be derived, it makes sense in an intuitive way: If one displaces a particle in a way where it does not violate the constraints, no force has to act, and thus no work to be done, to maintain the constraints.

As an example, consider a particle which moves along some arbitrary curve and which is forced on the curve by constraining forces. E. g. again the example of the particle in a tube. Any displacement along the curve will then not require any work by the constraining forces. This appears again obvious, as the forces to keep the particle on the curve need to act transversely to the curve only, since any move along the curve coincides with the constraints.

An equivalent formulation is the so-called principle of d'Alembert. It states that the effect of the non-constraining forces will create true acceleration of a particle, if they act in a direction which is compatible with the constraints. The formulation of d'Alembert embodies already one of the central goals of the current chapter: It no longer references the constraining forces, while it still a principle, which can be solved. In fact, it is already useful for sufficiently simple problems, but it is hard to generalize to more complex ones. Therefore, it is only an intermediate stage.

3.4 Euler-Lagrange formulation of the second kind

3.4.1 Reformulating d'Alembert's principle

To make progress to a more convenient formulation, it is helpful to recognize the limitations of d'Alembert's principle: The virtual displacements are not independent, but restricted by the constraints. To make it more useful, the next step is to eliminate also the constraints by switching to generalized coordinates. In the following, this will be done concentrating on situations with the simplest kind of time-independent constraints expressed in terms of equalities only, and without dissipation. The other cases can be captured by the Euler-Lagrange formulation of the first kind, which is, however, much more involved, though often relevant in practical problems. Conceptually, it turns out that all fundamental forces known so far are captured in the following formulation, and any deviation from it are effective consequences surfacing at the level of multi-particle systems.

The starting point is that d'Alembert's principle can be written as

$$(\vec{K}_i - d_t\vec{p}_i)\delta\vec{r}_i = 0.$$

The expression in parantheses is Newton's law for the non-constraining forces. Since the virtual displacements do no real work, they need to be orthogonal to constraining forces, but these are exactly what the difference is. Thus, the expression is zero. Starting from there, the equation is transformed into generalized coordinates. After an involved set of transformations, this ultimately leads to a dynamical equation for the generalized coordinates q_i and speeds $d_t q_i$, the virtual displacements δq_i, and the kinetic and potential energy of a system, reading[3]

$$\sum_j \left(d_t \frac{\partial}{\partial d_t q_j}(T-V) - \frac{\partial}{\partial q_j}(T-V) \right) \delta q_j = 0.$$

The quantity

$$L = T - V$$

appearing will be central in the following. Note that the kinetic energy T and potential energy V need to be expressed in terms of the generalized coordinates and generalized speeds as well. It is called the Lagrange function. The Lagrange function is the also the central quantity for bringing together special relativity (and possibly general relativity) and quantum physics. For classical, non-relativistic (quantum) mechanics it will only play a role as an intermediate step to Hamiltonian mechanics in chapter 5. Nonetheless, its central importance for any modern formulation of the fundamental laws of nature cannot be overrated.

Using the Lagrange function, d'Alembert's principle for conservative systems takes the form

$$\sum_j \left(d_t \frac{\partial}{\partial d_t q_j} L - \frac{\partial}{\partial q_j} L \right) \delta q_j = 0, \tag{3.2}$$

and therefore only involves the Lagrange function and the generalized coordinates. It is worthwhile to emphasize that, since the kinetic energy can depend explicitly on the time, so can the Lagrange function, even though the potential energy does not without dissipation.

If the system is conservative, each generalized virtual displacement δq_i in (3.2) can be shown to be independent. That is intuitively already clear, as all generalized coordinates, and thus directions, are independent. Thus, all terms in the sum have to vanish independently, yielding $dn - N$ equations of motion

$$d_t \frac{\partial L}{\partial d_t q_j} - \frac{\partial L}{\partial q_j} = 0, \tag{3.3}$$

[3]While mathematically in detail more subtle, the statement $\partial/(\partial d_t q_j)$ merely signals to derive as if $d_t q_j = x$ would be an ordinary variable x towards which to derive, and not a derivative. E. g. $\partial (d_t q_j)^2/(\partial d_t q_j) = 2 d_t q_j$.

the so-called Lagrange equations of the second kind, or sometimes also Euler-Lagrange equations (of motion). Again, these equations are second-order differential equations, and thus require $2(dn - N)$ initial conditions to solve.

Lagrange's equation of the second kind is what had been sought for, as they do no longer contain the constraints nor the constraining forces. However, they are not yet the simplest form to solve problems in mechanics and, as noted, not the best choice to generalize to non-relativistic quantum mechanics. Since relativistic quantum physics is for many actual problems, e. g. in solid state physics, serious overkill and far too complicated, it is very useful to find also a better formulation to generalize to non-relativistic quantum mechanics. This is the aim of the remainder of this chapter and of chapter 5, though this may at intermediate steps again only be obvious with hindsight.

3.4.2 Equivalence to Newton's law

To see that the formulation is equivalent to Newton's law (2.2), consider a conservative system of a single particle without constraints. The general case can also be done, but is somewhat more involved. In this case, there is a potential $V(\vec{r})$. The Euler-Lagrange equations are then

$$\begin{aligned} L &= \frac{m}{2}(d_t\vec{r})^2 - V(\vec{r}) \\ d_t\frac{\partial L}{d_t r_i} - \frac{\partial L}{\partial r_i} &= m d_t(d_t r_i) + \frac{\partial V}{\partial r_i} = m d_t^2 r_i - F_i = 0. \end{aligned}$$

The three Euler-Lagrange equations are therefore exactly Newton's second law (2.2) in component form. Since Newton's first law and Newton's third law are special cases of the second law - the action-reaction relations appear in the n-body version - they are satisfied as well. Only the mathematical framework postulated around Newton's laws in section 2.2 has to be furthermore carried over as well. But since they are related to the arena of mechanics rather than to the dynamics, this was to be expected.

3.5 Invariances of Lagrange's equation of the second kind

Newton's law (2.2) was written in terms of vectors. It therefore had the same form in any coordinate system, and specifying the actual coordinate system is only necessary if it is needed to evaluate it component-wise. Lagrange's equation of the second kind (3.3) are not given as vectorial equations. Since furthermore the choice of generalized coordinates is

not unique, it is an important question whether when changing from one set of generalized coordinates to another the equations change form, which would seriously impede their usefulness. While quite technical to demonstrate, this holds true.

There is another invariance of the Euler-Lagrange equations, which is very useful. Choose some Lagrange function and an arbitrary (twice differentiable) function $f(q_i, t)$ depending only on the generalized coordinates and the time. Then it can be shown that the Lagrange function

$$L^f = L + d_t f, \tag{3.4}$$

has the same Euler-Lagrange equations. The significance of this will be seen later in section 5.1. Such an addition of a total time derivative is sometimes called a (mechanical) gauge transformation, but this terminology will not be used here, as the more general idea of gauge theories will supersede it in classical field theory and quantum field theory.

3.6 Generalized momenta and cyclic coordinates

To continue, it is helpful to introduce a new concept, the so-called generalized momenta. They are defined as

$$p_i = \frac{\partial L}{\partial d_t q_i}. \tag{3.5}$$

Even for a particle, the generalized momenta will in general not coincide with $m d_t q_i$, though they may. The importance of these generalized momenta is that if the Lagrange function does not depend on a given generalized coordinate then the corresponding p_i must be constant. Thus the quantity p_i is conserved. Such conserved quantities, sometimes also called integrals of motion, are technically very useful. They are also deeply connected to symmetries, as will be explored later. Is this the case the corresponding coordinate q_i is called cyclic. It first appears unlikely that this should happen, but, as will be seen, it is quite common.

In case of the unconstrained particle, see e. g. section 3.4.2, the generalized momenta coincide with the ordinary momenta,

$$\frac{\partial}{\partial d_t r_i}\left(\frac{m}{2}(d_t \vec{r})^2 - V(\vec{r})\right) = m d_t r_i.$$

Cyclic ordinary momenta are then describing if the particle moves in the corresponding direction with constant speed (including zero speed).

3.7 Conservation laws

In section 3.6 conserved quantities, so-called integrals of motions, have been obtained from cyclic coordinates. While cyclic coordinates always yield a conserved quantity, the converse is not true, and there may be (many) more conserved quantities than there are cyclic coordinates in the Lagrange function. This already follows from the fact that generalized coordinates are not unique, and a coordinate may be cyclic for one choice, but not for another. Thus, a good choice of generalized coordinates can simplify the problem tremendously. However, finding them is not always simple. Here, some general rules will be described which allow to construct many conserved quantities, while a much more sophisticated approach will be discussed in section 5.8.

3.7.1 Energy conservation

Assume that the Lagrange function does not depend explicitly on the time. Then it can be shown that the so-called Hamilton function, defined as

$$H = p_i d_t q_i - L, \tag{3.6}$$

is a conserved quantity. This function will reappear also in chapter 5.

At first sight, this is a rather abstract conserved quantity. However, for a conservative system and if the generalized kinetic energy has the property

$$T(a d_t q_i) = a^2 T(d_t q_i)$$

it is possible to give H a very particular meaning. In this case

$$H = T + V,$$

the Hamilton function is just the total energy.

Thus, for such systems, which are by far the most numerous ones and the only ones relevant in fundamental physics, the total energy is conserved if the Lagrange function does not depend explicitly on time. The absence of explicit time-dependence can also be interpreted in a different way. Any transformation $t \to t + \Delta t$ will leave in this case both the Lagrange function and the Euler-Lagrange equations unchanged. Thus, the system is homogeneous (or isotropic) in time, as defining any time as $t = 0$ for the initial conditions is equally good, and there is no distinct time. This also explains that fundamental physics is usually of this type: To the best of our knowledge time has no absolute frame.

3.7.2 Momentum conservation

As may be expected, a very similar argument as in the previous section also holds if the space is homogeneous. Homogeneity in space means that everything depends only on relative distances. An example is the central force problem of section 2.7.3: The potential depended only on the relative distance of two particles, $\vec{r}_1 - \vec{r}_2$, rather than on their absolute positions. Therefore any shift of the coordinate system $\vec{r}_i \to \vec{r}_i + \Delta\vec{r}$ would not have changed anything.

Corresponding generalized coordinates would therefore be relative coordinates. They are not affected by such changes. The center-of-mass coordinate, which is affected, would therefore be cyclic, just as with the time discussed in case of the homogeneity of time. Thus the corresponding generalized momenta, which is just the center-of-mass momentum, will be conserved.

This is therefore a particular case of section 3.6. The same statement could therefore be read that if a system is translationally invariant, i. e. the absolute position does not matter, in a generalized coordinate, its associated generalized momentum is conserved.

3.7.3 Angular momentum conservation

A system is called isotropic in space if it does not change under a rotation. Therefore, the corresponding angular variable will again be cyclic, as the system does not depend on it. If the system is isotropic for all possible rotation axises, the total angular momentum is conserved. As before, the vanishing of the corresponding generalized force implies that the component of the total torque along the rotation axis vanishes.

3.7.4 Noether's theorem

The previous observations can be generalized in the form of Noether's theorem. It is one of the central reasons why symmetries are so important in physics. It shows that for a conservative system any continuous symmetry, i. e. a symmetry where a generalized coordinate is changed by a continuous quantity, entails the existence of a conservation law. The previous cases were special examples of this, as they amounted to a continuous shift in time, space, and rotation.

Such conserved quantities are of central importance for both theoretical and practical reasons. Practical, because exploiting conservation laws is very helpful in solving problems. Theoretical, conserved quantities define the properties of a system.

Essentially, Noether's theorem boils down to the fact that if the Lagrange function is invariant under a variation of a generalized coordinate, the variation can be used to derive a

conserved quantity. In this way, conserved quantities are linked to invariances of a system. Such an invariance is also called a symmetry, as the system looks the same, whether using the coordinates or transformed coordinates. This is a very fundamental insight that symmetries are connected to conservation laws, and will be encountered throughout all of physics.

3.8 Chaos

A particular subclass of physical systems are those which exhibit a so-called chaotic behavior. Chaotic behavior is best defined in phase space. Given two trajectories at some point in phase space at some time t, it is possible to define an Euclidean distance between two trajectories as

$$d(t) = (q_i^1 - q_i^2)^2 + (p_i^1 - p_i^2)^2.$$

A system is said to show chaotic behavior if the distance increases exponential (or faster) as a function of time[4],

$$d(t) \sim e^{at}, \tag{3.7}$$

where a is some constant characteristic of the system.

This is still classical mechanics, and therefore completely deterministic. The reason why such a behavior is said to be chaotic is that any error in the determination of the trajectory will therefore be exponentially amplified over time. Thus a prediction becomes essentially impossible in practice, as in any practical (experimental) situation there are always errors. Chaos is therefore a feature of the unavoidable measurement errors rather than a problem coming from theory.

This is quite different for quantum chaos, where quantum effects can introduce additional chaotic behavior, which leads too far beyond the scope of this lecture.

It is actually not necessary that a system shows the behavior (3.7) for all pairs of possible trajectories, i. e. initial conditions, nor for all times. It can well be that the system is chaotic for some cases of initial conditions, but not all. It is also possible that after some time of exponentially diverging trajectories come again close to each other. The simplest such system is actually a double pendulum. For small oscillations it shows no chaotic behavior, but as soon as the kinetic energy exceeds a critical value it becomes chaotic.

Classifying such features of a system is the purview of chaos theory. Interesting features of chaotic systems are, e. g., attractors. These are localized regions in phase space at which

[4]There exists actually a mathematically more precise definition of what a chaotic system is, but for the present purpose this is sufficient.

multiple trajectories, even from completely different initial conditions, coalesce, and stay there. If they do so without forming any kind of periodicity, the attractor is called strange. If the system is dissipative, and therefore the phase space density can change, the attractor can become a point in phase space, which is called a fixed point.

A very simple example of chaotic behavior is given by the equation

$$x_{n+1} = rx_n(1 - x_n),$$

the so-called logistic map, with $0 \leq r \leq 4$ and requiring $0 \leq x_0 \leq 1$. In that case, x_n will be always bounded by the same interval. It appears in many context, but can be viewed as the discretized version of

$$\frac{dp}{dt} = F(p) = (r - 1)p - rp^2,$$

that is of Newton's law with a specific, speed-dependent force. Note that the equation is translationally invariant, and as such that the initial condition for the position does not play a role. There are two contributions, one dissipative, as it reduces the momentum, and one driving force, as it increases the momentum. Solving it yields that for $r < 1$ the dissipation wins, and the particle stops. For $1 < r < 3$, dissipation and driving balance out, and asymptotically the momentum becomes constant. The initial fluctuations before settling down increase with r.

Something more interesting happens when $3 < r < 1 + \sqrt{6}$. In that case, an overshoot effect takes place, which forces a permanent oscillation between two values. This is because the acceleration does not start to vanish when approaching a value as before, but rather remains non-zero towards the other value. Thus, it is not possible to provide a single value for the asymptotic behavior of p anymore. This situation is called a bifurcation.

Increasing beyond $1 + \sqrt{6}$ the system starts to oscillate between more and more values. When reaching a value $r \approx 3.56995$, also this behavior breaks down, and the behavior becomes entirely non-oscillatory. Prediction of the long-time behavior becomes exponentially sensitive to the initial conditions, and the system is chaotic. Only for specific values of r below 4 the behavior becomes regular (though oscillating) again.

As many practically relevant systems, especially those of many degrees of freedom, can show chaotic behavior, this subject is of great importance. Due to its large number of phenomena and features, its detailed study warrants an own lecture, and it will not be continued here.

Chapter 4

Special relativity

So far, classical mechanics was essentially described by Newton's laws of section 2.2, supplemented with forces. Both together form as postulates the framework of classical mechanics. However, it turns out that classical mechanics based on these postulates is not able to describe all situations, no matter how complicated the forces are assumed to be. Of course, also electromagnetic, including optical, phenomena are not described by classical mechanics. By supplementing Newton's laws with Maxwell's laws this is possible, but does not lead to anything conceptually new on the sides of mechanics, as those cases to be discussed here do. Note that thermodynamics and hydrodynamics are actually only the application of classical mechanics to large ensembles of particles, and therefore do not require fundamentally new laws, as can be shown using statistical (classical) mechanics.

The situations where failures arise are, together with their resolutions or combinations of resolutions,

1. At very short distances. This is the purview of quantum mechanics

2. High speeds. This is the purview of special relativity

3. Strong gravitational fields. This is the purview of general relativity, which necessarily induces large speeds and thus includes special relativity

4. At short distances and high speeds. This requires the combination of quantum mechanics and special relativity, which becomes quantum field theory. Note that a full inclusion of electrodynamics in quantum mechanics requires quantum field theory

5. At all of the above. This requires the combination of quantum physics and general relativity in the, not yet fully formulated, quantum gravity

Besides electrodynamics and gravity, there are also two more known sources of forces, the weak and strong nuclear forces. All known forces can actually be derived from these four, though a unifying theory of all of them is not yet available.

Relaxing all of these conditions, and covering everything listed, essentially sums up the complete master study of (theoretical) physics. As a first step, here one of those will be included, those which is the least complicated one: Special relativity, i. e. the modifications necessary to classical mechanics when including large speeds.

4.1 The speed of light and Minkowski space

4.1.1 The equivalence principle

The starting point of special relativity is a reevaluation of Galileo's transformations of section 2.8. Consider a system which moves with a constant speed with respect to another one. This implies that the coordinates in one system are related to the other by

$$\vec{r}' = \vec{r} + \vec{v}t, \tag{4.1}$$

where $\vec{v}$ is the relative speed of the two systems. These are inertial systems, and thus Newton's laws hold equally in both of them. That is readily visible, as differentiating $\vec{r}$ twice it vanishes, thus not changing the acceleration. The force needs also to be invariant under (4.1), and thus many dissipative forces would not appear to fit into the framework. However, this is because they are effective forces, which stem from microscopic processes which individually are Galileo-invariant.

But speeds measured in one inertial system differ from speeds in the other system, since

$$d_t\vec{r}' = d_t\vec{r} + \vec{v}. \tag{4.2}$$

This should then be a true statement for any moving object when described in either coordinate system.

This is where a contradiction to experiments arises. While the relation (4.2) holds to very good accuracy experimentally at low speeds, it does not do so at large speeds. In fact, the deviation is there also at low speeds, but it becomes quickly so tiny, as will be quantified later, that it is below any reasonable experimental uncertainty if the speed is just low enough. Thus, something does not fit at high speeds.

As it turns out, Newton's postulates are the problem. More precisely, their statements about inertial systems and that Newton's second law (2.2) holds equally in all inertial systems connected by Galileo transformations. The reason is that Newton's laws make a

very deep assumption about space-time: That space has an Euclidean structure and time is absolute and independent of the observer. However, as it turns out, nature is not of this form, but has a more complex structure, which only at low speeds looks like this.

In principle, the only necessary thing is to replace the space-time structure accordingly, and then derive the results from this. It is, however, more instructive, to pursue a slightly different path, where the space-time structure stands at the end. As always, there are several equivalent ways of formulating the basic postulates, and each form implies the other.

This different approach will start rather from a physics motivation.

Consider again Newton's laws. It may look like that the second law, equation (2.2), is the most important part. Just because so far this was the one made most reference to. But this is not quite the case. The probably most important statement is that physics is the same in all inertial systems. As the choice of inertial system is done by an observer, this implies that physics is independent of any observer. While this appears obvious at first, this is actually a very deep statement. And a postulate. Changing (2.2) somewhat would just change the resulting trajectories. Changing this statement would give up any notion (or even hope) of something like an objective reality[1].

This will be the first part of the new set of postulates: Physics is observer independent. This implies that the further postulates should have the same form in every inertial frame. In particular, this implies it is impossible to make statements about the frame from within the frame: There is no absolute frame. Only relative differences between frames can be identified. This is known as the equivalence principle.

So far, this seems not to be something different than Newton's requirement that physics should be the same in all inertial frames.

The big difference comes from an experimental observation: All massless particles move at a fixed speed. It is called the speed of light c, as light in the form of photons has been the first observed particles to be massless. Since this speed must therefore be the same in all frames, this implies that (4.2) cannot be correct. Or more aptly put: Galileo transformations are not appropriate to transform between frames. Since this transformation comes from the assumption of an Euclidean space-time, it is this assumption, which must be wrong. It now remains to work out the consequences. This is entirely based on the equivalence principle and the experimental observation of a unique speed of massless particles in every frame[2]. That massless particles may create a problem

[1]Whether such a thing can exist at all is an unsolved question of science philosophy. In physics, it is the description of observations, which is the core duty. And so far all observations available very uniquely point to an observer-independent description.

[2]Historically, it was not the speed of massless particles, but of the speed of light. However, this would

can also be seen from Newton's law (2.2), as it implies that massless particles cannot be described by it. Though this can be alleviated by the formulation using (2.1), it still leaves a certain feeling of uneasiness. Also this will be fixed in special relativity.

4.1.2 Minkowski space

So, it is now clear that something is not right with the way inertial frames, and transformations between them, are defined. This leads to the question how this can be fixed. Consider the distance traveled by a massless particle with its fixed speed[3] c. In a given reference frame, the distance traversed in time t will be

$$\vec{r}^2 = c^2 t^2.$$

Something which would be nice to keep is that space is isotropic - the x, y, and z directions are all the same. The question is, how does this change to a new coordinate system. There are several possibilities. E. g., the left-hand side could change by changing the metric, additional terms could arise, and many more. Experiment would have to decide, which is correct. Though instructive, considering all possibilities would lead far beyond the scope of this lecture. Therefore, with hindsight, the only change will be to give t a non-trivial transformation property, i. e. the time could also differ in both coordinate systems. This deviates from the starting point of the Galileo transformation (2.24).

In another coordinate system then the relation

$$\vec{r}'^2 = c^2 t'^2$$

holds, where the important piece is that c is the same in both equations. Since zero is zero, both equations can be combined to yield the equality

$$\vec{r}^2 - c^2 t^2 = \vec{r}'^2 - c^2 t'^2. \tag{4.3}$$

In the case of a Galileo transformation, a similar equality holds, but there c would be transformed and t would be fixed.

Nonetheless, such a relation as (4.3) looks quite familiar from linear algebra. It reminds of an invariant length, but of a vector space with a non-Euclidean metric. This is actually correctly inferred.

require electrodynamics, which will be avoided here, as considering massless particles lead to the same result

[3]Which is, of course, the speed of light.

Define now space-time as a vector space, the so-called Minkowski space, $\mathbb{M}^4$ as a vector space having a non-trivial metric[4]

$$g^{\mu\nu} = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Then this defines a four-vector[5]

$$x = \begin{pmatrix} ct \\ \vec{x} \end{pmatrix} = \begin{pmatrix} ct \\ x_1 \\ x_2 \\ x_3 \end{pmatrix}.$$

Conventionally, the components of this vector are enumerated with Greek literals starting from zero, $x_0 = ct$, $x_i = (\vec{x})_i$, while the Latin indices still enumerate the spatial components starting at 1.

The scalar product is then

$$x^2 = x^T x = x_\mu g^{\mu\nu} x_\nu = -(ct)(ct) + (\vec{x})(\vec{x}) = \vec{x}^2 - c^2 t^2,$$

and thus in such a space indeed (4.3) is a preserved length. The question is now, whether the standard linear algebra for such a space correctly reproduces physics. Most interesting is, of course, whether rotations in such a space are compatible with the equivalence principle and the speed of light.

As it will turn out that this is indeed correct, it is useful to repeat a few results about vector spaces with non-Euclidean metric. Since it has been the question of coordinate transformations which led to this point, the first important step is that of how they work in Minkowski space.

First of all, a translation works as usual, i. e. a replacement $x_\mu \to x_\mu + a_\mu$ shifts the origin. The only difference is that, in contrast to the Galileo transformations of section 2.8 a_0 also automatically shifts the origin of time.

Leaves rotations. Rotations are operations which leave the scalar product invariant, and thus the generalized length, invariant.

[4]It is equivalently possible to switch the signs of all entries, $g \to -g$, without changing the physics. This will be discussed more in detail below, and is thus a convention.

[5]In older texts this is often done by having an ordinary Euclidean metric but making the time component (or all space-components) purely imaginary. While for special relativity this does not make any difference, this is not well suited for generalizations, which are quite interesting nowadays. It will therefore not be used here.

It turns out that there are six basic rotations, grouped into two sets, which leaves the scalar product invariant. The first three can be recognized as ordinary rotations in three-dimensional space, when ignoring the time. This is not surprising, as the spatial components of the metric g are just an ordinary Euclidean metric, and any invariance there will simply require ordinary rotations. The other three are odd.

First, it is found that they leave the sign, and thus sign of the length, of the scalar product of a vector with itself invariant. Thus, vectors can be classified by the fact whether they have positive, negative, or zero length. Examples for all possibilities are

$$\begin{aligned} v_1 &= \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \\ v_2 &= \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix} \\ v_3 &= \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \end{aligned}$$

where these vectors have negative, zero, and positive norm, respectively. Vectors with negative norm are called timelike, vectors with zero norm are called lightlike, and vectors with positive norm are called spacelike, for reason to become clearer later. In fact, these three additional transformations are the Lorentz transformations in the three directions, e. g. in x-direction

$$\begin{aligned} x' &= \begin{pmatrix} \gamma(ct - \beta x_1) \\ \gamma(x_1 - \beta ct) \\ x_2 \\ x_3 \end{pmatrix} \\ \beta &= \frac{v}{c} \\ \gamma &= \frac{1}{\sqrt{1-\beta^2}}. \end{aligned} \tag{4.4}$$

If the speed is small, $\beta \approx 0$ but $c\beta = v$, this reduces to

$$x' \overset{\beta\approx 0}{=} \begin{pmatrix} ct \\ x_1 - vt \\ x_2 \\ x_3 \end{pmatrix},$$

and therefore reduces to the Galileo transformation, in accordance with the requirement that at low speeds conventional Newtonian physics should reemerge. It should be noted that a shift by a constant (four) vector is satisfying all the above criteria, and is still an admissible transformation. Combining the above Lorentz transformation, the so-called proper Lorentz transformations, with these translation leads to the so-called Poincare transformations. They can be supplemented by operations like space and/or time inversions.

However, all of this is essentially only replacing the arena of mechanics. It is not yet defined how the dynamics, and even the kinematics, should be formulated. This will be done in the following.

4.2 Measurements of time and distance

To define the kinematics requires to give meaning to distances and time differences, to ultimately define speeds and accelerations. This requires to understand how distances are perceived in different coordinate systems.

To understand why this is not obvious, note that in (4.4) not only space changes, but also time. Especially, in relatively boosted systems two events will not take place at the same time - concurrence as known in the Galileo group is lost.

Consider now some rigid object with an extension $l = y_1 - x_1$ in its rest frame. An observer moving along the x-axis will define the length of the stick by the difference in endpoints at the same time in the employed coordinate system. As noted, this does not correspond to the same time in the original coordinate system. It is thus best to use the inverse of (4.4), since

$$x_i = \gamma(x'_i + \beta ct')$$

Calculating the difference yields

$$l' = y'_1 - x'_1 = \gamma^{-1} y_1 - \beta ct' - \gamma^{-1} x_1 + \beta ct' = \gamma^{-1} l.$$

Thus, the stick is shorter in the moving frame, as γ^{-1} is always smaller than 1. This has nothing to do with the finiteness of a maximal speed: Using a Galileo transformation and

a finite speed to transfer the information from the different points of the beginning and ending of the stick will yield a different result. The reason is that spatial distances, due to (4.3), are shorter - a Lorentz transformation only preserves the length element (4.3), but not a spatial length element. That this leads to actual shorter distances is because of the minus sign in the metric. This phenomena is called length contraction.

In the same way time differences measured at the same point change, and become

$$\Delta t' = t'_2 - t'_1 = \gamma(t_2 - t_1) = \gamma \Delta t,$$

but they appear longer. The reason for this is the opposite sign of spatial and temporal parts of the metric. This phenomena is called time dilatation.

It is important to note that this is a reflexive phenomenon: Not only will the observer in the moving frame think sticks are shorter and times longer in the other frame, an observer in the other frame will see the same for the primed frame. This is the equivalence principle at work. In both frames, physics looks the same, as it is not possible to have an absolute speed, merely a relative speed[6].

In the end, both time dilatation and length contraction should not come as a surprise. Lorentz transformations only keep the norm of a 4 vector invariant. Both spatial length and time are only components of a 4 vector. Therefore, they are not separately invariant. This is as for ordinary rotations in Euclidean space: Performing a rotation, the components of a vector are not individually invariant, but only the total length. And just like as with rotation, any reduction in one component's absolute value has to be compensated by an increase in another component's, and thus also some opposite effect has to be expected for both time and space here. Although the indefinite norm has some influence on this naive picture.

Another interesting feature is what happens when performing two boosts after each other. This makes statements about how speeds add to each other. Doing the technical details implies for the total β value

$$\beta = \frac{\beta_1 + \beta_2}{1 + \beta_1 \beta_2} \overset{\beta_i \ll 1}{=} \beta_1 + \beta_2$$

This so-called Einstein's addition theorem of velocities has a number of remarkable properties. First, at small speeds it reduces to its Galileo form (2.26), and once more shows how Galileo transformations emerge as a limiting case of Lorentz transformations. The second is that this implies that the total speed can never exceed the speed of light, as for

[6]Note that these are statements only in frames moving with fixed speed to each other. Accelerated frames, e. g. when starting at the same point, but one accelerated and then coming back at some point in the future, need more to be accurately describe, see section 4.3.

any values of $\beta_i \leq 1$ the total speed also satisfies $\beta \leq 1$. In fact, if both speeds are at the speed of light, the total speed is also the speed of light. This implements the experimental observation that there is no speed possible faster than the speed of light. In fact, the only solutions for $\beta > 1$ require the β_i to be complex, a physically not sensible result.

This observation has far-reaching implications. The probably most profound is that there are events, i. e. points localized in space and time, which cannot influence each other. This is because their distance is so large that no object can reach them quickly enough. They are said to be causally disconnected. In a fixed coordinate system, this happens as two points cannot communicate earlier than a signal traveling at the speed of light can travel from one to the other[7].

This happens, if

$$(ct)^2 < (\vec{x})^2 .$$

But this is nothing than the statement that the connecting vector of both events has a positive norm. Thus, this justifies the name spacelike for these vectors, as there is always an intervening space. On the other hand, two events can interfere if the opposite holds, and thus the connecting vector is timelike. Finally, the borderline is the case of zero length, and thus lightlike vectors. As the length of the vectors are invariant under a Lorentz transformation, by construction, causal connections remain invariant under the change of coordinate systems. This is, as it must be, since causal influence is a physical feature, and thus must be the same in all frames.

This fact can be visualized by the so-called light-cone, which is the part of Minkowski space, for any fixed point, which has a timelike distance to the reference point. The boundary of the cone is defined by events at light-like distances. These are those events, which can be reached when moving at the fastest possible speed. All points in the past to the event can therefore causally influence it, will the event itself can causally influence any event in its future light-cone, where past is defined by $t < t_0$ and future by $t > t_0$, where t_0 is the time coordinate of the event. Also these light-cones are invariant under Lorentz transformations, and the same in all frames.

Events outside the light cone have a space-like distance to the event. The distance cannot be crossed, due to the speed limitation. Every space-like event with the same value of t_0 in the given reference frame can be used to define a hyperplane of perceived simultaneity, i. e. of all events happening at the same instance of time for the observer. While observers using different frames will not agree on these to be simultaneous, this

[7] In a more refined version, it is actually that no information, carried by energy, can be transported faster than the speed of light. While in principle more exact, this will make only a difference when introducing fields.

hyperplane contains all possible causally disjunct points for an instance of time, and thus forms a snapshot of the whole (spatial) universe for the original observer.

This dramatic effect can be seen explicitly in the following. Consider the two events located at $a = 0$ and $b = (0, 0, 0, b_3)$. They are thus, in this frame, simultaneous, but not causally connected, since $(a-b)^2 = b_3^2$. From an observer moving at some velocity $\beta = v/c$ in the three-direction, these vectors looks like

$$\begin{aligned} a' &= 0 \\ b' &= \begin{pmatrix} -\beta\gamma b_3 \\ 0 \\ 0 \\ \gamma b_3 \end{pmatrix} \\ (a'-b')^2 &= -\beta^2\gamma^2 b_3^2 + \gamma^2 b_3^2 = b_3^2. \end{aligned}$$

They are thus still space-like separated by the same distance, but the observer perceives them at different times $a'_0 \neq b'_0$, and thus not simultaneously.

4.3 Relativistic kinematics

With this all necessary ingredients are available to describe the motion of a particle in special relativity, and thus relativistic kinematics and, ultimately, dynamics in section 4.4.

To achieve this aim, it is best to be inspired by the classical case and how it reacted to rotations. The basic ingredient in the formulation of Newton's laws was to require the involved quantities to be vectors. To achieve that physics is invariant under rotations was to formulate all equations such that always both sides transform in the same way under rotations[8]. In the same way, the generalization will be completely in agreement with Lorentz invariance - often called covariant - if particular care is taken to involve only objects transforming in the same way on both sides of every equation.

The first task is to describe the trajectory of a particle. It is given by a sequence of four-vectors x. So far, the position along the trajectory has been identified by the time. This was possible, as time was an immutable concept. Now, however, time is itself part of the trajectory. It is therefore no longer a suitable concept, and a new trajectory parameter seems more appropriate, which uniquely identifies a position along the trajectory.

[8]Note that this is about absolute orientation. A force along the z-axis does not break this absolute rotation invariance. After rotation it will still point into a fixed direction but a different one. However, relatively, there is no rotational invariance, and e. g. angular momentum is not conserved in the plane orthogonal to the force, but only for rotations around the force axis. But in which direction the axis shows absolutely does not matter.

To find one, consider an infinitesimal move along the trajectory. The length of this move is given by

$$\frac{1}{c^2}(dx_\mu)^2 = -(d\tau)^2$$

where the prefactor is chosen by convention. The right-hand side now defines a quantity τ, which is called the eigenzeit, and which by construction exactly describes the change in trajectory, as the advance is uniform in it. It should be noted that the trajectory $x_\mu(\tau)$ is also called a world-line, as it describes the propagation of a particle along a line in the space-time world.

When changing to the rest frame, this reduces to

$$(d\tau)^2 = -\frac{1}{c^2}(dx_\mu)^2 = \frac{1}{c^2}(dct)^2 = (dt)^2.$$

Thus, in the rest frame, sometimes also called eigenframe, the eigentime is the ordinary time, justifying its name and its normalization. Note that by convention the eigenzeit is taken to be the positive value. Note further that for any arbitrary frame

$$d\tau = \sqrt{\frac{1}{c^2}\left((dct)^2 - (d\vec{r})^2\right)} = dt\sqrt{1 - \frac{1}{c^2}\left(\frac{d\vec{r}}{dt}\right)^2} = \gamma^{-1}dt$$

Thus, the time interval dt is the one measured in a system moving with respect to the particle's restframe. This again makes the eigen part of the name eigenzeit clearer: τ is the time in the restframe of the particle, which is used to describe its trajectory.

The eigenzeit can now be used to generalize the concept of speed to the four velocity

$$u_\mu = \frac{dx_\mu}{d\tau} = \gamma\frac{dx_\mu}{dt} = \begin{pmatrix}\gamma c\\ \gamma\vec{v}\end{pmatrix} \overset{\beta_i \ll 1}{=} \begin{pmatrix} c\\ \vec{v}\end{pmatrix}. \tag{4.5}$$

The last equality shows why this is a useful generalization of the concept of speed, as it reduces at low speed to the conventional one.

It is an interesting feature that the length of the four speed is actually constant,

$$u_\mu u^\mu = \gamma^2(\vec{v}^2 - c^2) = -c^2\frac{1-\beta^2}{1-\beta^2} = -c^2.$$

Note that this is also true at low speeds. This also implies that the four-speed is a timelike vector. What is even more remarkable is that therefore the four-speed can never vanish. Thus, in contrast to Newtonian mechanics, no particle can remain stationary in space-time. Rather, even if at rest, it propagates along its world-line in time direction, and always forward. Thus, time always passes, due to the structure of space-time. Therefore, space-time itself dictates that there is a flow of time, and that it is objectively of the same rate for everyone within their own rest frame.

This last observation helps also to understand the so-called twin paradox. If somebody is at rest on a planet, and somebody starts from it, moves some distance away from it, and comes back, less time has passed in the rest frame of the planet than on the space-ship. How is this is possible, if relative motion is irrelevant? The answer lays in a simple, albeit important, subtlety. While relative motion does indeed not alter physics, relative acceleration does. And the space-ship needs at least to accelerate when inverting its direction to come back.

This can be seen more formally as follows. The elapsed time T for an observer in its rest frame is given by

$$T = \frac{1}{c}\int_{\tau_0}^{\tau_1} \sqrt{-u(\tau)^2}d\tau,$$

essentially like in Newtonian mechanics. For the observer at rest, $u(\tau)^2 = -c^2$, and thus it is just the ordinary time-difference. It is worthwhile to work this out in detail.

The person on the planet remains at rest. Thus, the person's trajectory will be governed by the four-velocity $u_1(\tau) = (c, 0, 0, 0)^T$, and thus $x_1(\tau) = (c\tau, 0, 0, 0)$. The person in the space-ship, for simplicity, should perform a parabolic path viewed from the person on the planet,

$$\begin{aligned} x_2(\tau) &= \left(c\tau, 0, 0, 2a\tau - b\tau^2\right) \\ u_2(\tau) &= \left(c, 0, 0, 2a - 2b\tau\right), \end{aligned}$$

where its first passage was at $\tau = 0$ and coming back (and overshooting then) at $\tau = 2$ days[9]. The elapsed time is for the person at rest $\Delta\tau = 2$ days, while for the one in the space ship in their rest frame

$$\Delta\tau = \sqrt{4a^2 + 16b^2 - c^2} + \frac{4a^2 - c^2}{8b} \ln \frac{4a^2 - c^2}{\left(\sqrt{4a^2 + 16b^2 - c^2} - 4b\right)^2}$$

The three-speed is limited by the speed of light, and thus $|2a - 2b\tau| < c$, at least[10] for these 2 days. Putting in numbers, e. g. such that the speed at return is $0.2c$, yields an elasped time of 1.9 days for the person in the spaceship in their restframe. Thus, the time elasped was shorter.

Physically, of course, the important point was the acceleration. The person on the planet traversed their worldline at constant speed, while the person in the space-ship started with a certain speed, then decelerated, and then accelerated again, but was only

[9] Or any other arbitrary time duration

[10] Of course, the formula cannot stay correct forever, as the speed increases arbitrarily.

at rest very briefly. Thus, the second derivative of the person in the space-ship is in every coordinate frame non-zero, while there exist some coordinate frame where this is the case for the person on the planet. This is the distinguishing feature, which allows an asymmetry for the two persons. A calculation where the person in the space-ship really starts from the planet and then stops again is more ambitious, as the calculation requires the integration of a much more complex function, but yields a similar result.

4.4 Relativistic version of Newton's laws

To formulate a relativistic generalization of (2.2), the first postulate is that the mass of a particle, m, is still a well-defined quantity, and especially does not change under Lorentz transformations.

To generalize Newton's law, it is best to start from the more general formulation using the momentum, (2.1). The relativistic momentum is given by $p_\mu = mu_\mu$, as a direct generalization of the non-relativistic one, and which, due to (4.5), reduces to the non-relativistic one at low speeds.

It then remains to postulate the relativistic version of Newton's law such that it is an equation in terms of four-vectors. It is eventually found to be

$$\frac{dp_\mu}{d\tau} = K_\mu = \begin{pmatrix} \gamma\vec{\beta}\vec{F} \\ \gamma\vec{F} \end{pmatrix} \overset{\beta_i \ll 1}{=} \begin{pmatrix} \vec{\beta}\vec{F} \\ \vec{F} \end{pmatrix}. \tag{4.6}$$

By construction, this has the correct low-speed limit. While this is a necessary criterion, this is not sufficient. The extension of Newton's law to the relativistic case cannot be deduced, but can only be experimentally confirmed[11].

However, the fourth component of (4.6) looks strange, and requires further scrutiny. Going back to conservative forces in (2.12), the relation between kinetic energy and power

$$d_t T = \vec{F}\vec{v},$$

allows to make an observation. The fourth component of (4.6) is associated to the change of kinetic energy in the non-relativistic limit. Conversely, this implies that

$$T = \gamma mc^2 \overset{\beta_i \ll 1}{=} mc^2 + \frac{m\vec{v}^2}{2} \tag{4.7}$$

[11] It appears, as if this process would have to be done twice, once for the non-relativistic case, and once again for the relativistic case. That is, of course, a historical artifact, as only small speeds where accessible for a long time. Nature has only one version, (4.6) (or actually its quantum gravity generalization). The low-speed version (2.1) would actually never had been needed to be postulated, if investigations would have started already with access to high speeds. Then directly (4.6) would have been discovered, and (2.1) would have been established as the low-speed limit of (4.6) without extra searches or postulates.

is the relativistic generalization of the kinetic energy. In particular, this kinetic energy is conserved in the absence of forces. Thus, the fourth component of the four momentum is the kinetic energy.

There is however still the question of the constant in (4.7). It seems to make no sense that a particle at rest has kinetic energy. The reason is that it should rather be interpreted as the total energy of the particle, and this is a contribution to this total energy every particle has, even at rest. As this kinetic energy is proportional to the mass, this is also called the rest energy of the particle. It is therefore useful to reinterpret T as the total energy of a particle E.

From this follows

$$p_\mu p^\mu = -m^2c^2 = \vec{p}^2 - \frac{E^2}{c^2}$$

or that the energy of a relativistic particle is given by

$$E = \sqrt{\vec{p}^2c^2 + m^2c^4} \overset{\vec{v}=\vec{0}}{=} mc^2, \tag{4.8}$$

giving a connection between energy and mass. This is a very important statement, as it implies the equivalence of energy and mass, and therefore the possibility to convert one into the other: Mass is no longer a conserved quantity, even if it is a scalar quantity. It can be transformed to energy, which then again can be converted to mass, thereby changing the identity of particles.

What remains in place is Newton's third law of action and reaction. An important consequence of this is

$$d_t p^1 + d_t p^2 = K^1 + K^2 = K^1 - K^1 = 0.$$

Thus, the total four momentum is conserved in a closed system. This also implies that energy is conserved, as this statement holds component-by-component. Four-momentum conservation supersedes energy and momentum conservation of non-relativistic mechanics. This is especially different, as here only the energy is conserved, but this does not need to be in form of mass. Rest mass is no longer individually conserved, but only accidentally, if four momentum conservation implies it. That is of particular importance for the relativistic generalization of scattering processes.

It should also be noted that (4.6) no longer implies that the acceleration $\vec{a}$ needs to be proportional to the force, as it also contributes to the fourth component. As a consequence, a particle can experience lateral accelerations with respect to the force. Such effects can be measured, and represent a genuine effect entirely due to special relativity. Note finally that a free particle will again have a constant four-velocity, and thus move along a straight line in the three-dimensional subspace.

So far, still the mass appeared throughout, and it is not yet clear, how massless particles should act. However, (4.8) implies for massless particles

$$E = c|\vec{p}|,$$

and thus for a massless particle the momentum is proportional to the energy. Moreover, (4.8) shows that in the massless case the four-momentum becomes light-like rather than space-like. This implies that it describes a movement of a particle at the speed of light, rather than that of a massive particle with a finite speed.

While the formulation (4.6) is certainly a possible formulation of relativistic dynamics, it is actually not a very convenient one. The Lagrangian formalism of chapter 3 is much better suited for it. In fact, the Lagrange function

$$L = -mc^2\sqrt{1 - \frac{m^2p^2}{c^2}}$$

yields (4.6) for $K = 0$, provide that in the Lagrange equations of the second kind (3.3) derivatives with respect to the restframe time t are replaced with derivatives with respect to the eigentime τ. Conservative forces can then be introduced using a potential,

$$L = -mc^2\sqrt{1 - \frac{m^2p^2}{c^2}} - V(x_\mu x^\mu)$$

as long as the potential only depend on the squared four-vector x.

In addition, there is a conceptual problem in applying (4.6). Considering forces like the one of a harmonic oscillator or the gravitational force, they entail the problem that these forces act everywhere simultaneously. It is certainly possible to use the one-particle version in (4.6), as this corresponds to an eternal and immutable force field, which does not change over time. But the situation becomes more complicated when considering the two-body problem with the original gravitational force (2.4). It uses the positions of every particle at a given instance of time. Thus, it assumes that the information about the propagation of a particle is instantaneously transmitted to the other particle, and vice versa. This is inconsistent, as the maximum speed in special relativity is the speed of light. Thus, though its is formally possible to use this force in (4.6), the results will not be compatible with physics. Rather, the force itself needs to have the speed of light as maximum propagation speed build in. Thus, treating dynamical problems is far less trivial, as it requires also new postulates for the forces.

Chapter 5

Hamiltonian mechanics

So far, the formulation of mechanics has been by use of differential equations, especially Lagrange's equation. Though this approach is extremely useful in mechanics, and also of great practical importance, it is not yet the best formulation to extend it to (non-relativistic) quantum physics. The main reason for this is that quantum physics, especially in combination with special relativity, introduces correlations which are not local, i. e. not describable by differential equations, but also requires non-local aspects, i. e. integrals. Though for classical mechanics both a differential and an integral formulation yield the same result, it is helpful to perform the corresponding conceptual formation already in the otherwise much simpler classical mechanics. It will also help eventually to demonstrate that for many cases it is always possible to find generalized coordinates such that the system becomes that of almost trivial mechanical systems in section 5.8.

Unless otherwise noticed, most of the following will apply to the most relevant case of conservative systems.

5.1 Hamilton's principle

5.1.1 Integral formulation

The basic entities to formulate the following are constructed from the configuration space, i. e. the $dn - N$-dimensional space of the independent, generalized coordinates. Any curve inside this space parametrized by a parameter t, no matter whether it is a solution to the equation of motion, is a trajectory describing the movement of the particles of the system.

Consider the solutions of the equations of motion, which are solutions to differential equations which are of second order in time[1], and therefore require $2(dn - N)$ initial

[1]Note that this may appear different when, e. g., a speed becomes a generalized coordinates, but this is

conditions. Though in practice it may be awkward, it is always possible to select two times t_1 and t_2 and specify the initial conditions by the values of all generalized coordinates at these two times, $q_i(t_1)$ and $q_i(t_2)$.

Then it is possible to select among all possible trajectories in the configuration space those which are (twice) differentiable in time and satisfy the initial conditions of the problem at hand. In addition, it will be required that these trajectories must be reachable from a trajectory fulfilling the equations of motions, the so-called physical trajectories, by successive virtual displacements, i. e. the trajectories are dense around the physical trajectories inside the configuration space. However, once this definition is made, any other trajectories inside the set can equally well be used to define it[2].

This subset of configurations will be called $\mathcal{M}$. All such trajectories share a number of characteristics, by construction. First of all, they spend trivially the same time between start point and end point at t_1 and t_2. Secondly, all virtual displacements at the initial time and final time have to vanish, i. e. $\delta q_i(t_1) = \delta q_i(t_2) = 0$, as otherwise the constraints (in form of the initial conditions) would be violated.

It is now possible to evaluate the Lagrange function on any element of $q(t) \in \mathcal{M}$, yielding a function only of time

$$L(q_i(t), d_t q_i(t), t) = L(t),$$

and it is formally possible to integrate this function over time to yield a number

$$S[q] = \int_{t_1}^{t_2} dt L(t), \tag{5.1}$$

which is called the action. This number depends on the choice $q(t)$, i. e. which trajectory is chosen. It therefore maps a function, or more precisely functions, to the real numbers. This is different from functions, which map values of variables to the real numbers. Such an object has therefore a different name, and is called a functional.

After these preparations, it is possible to postulate the principle of Hamilton. There are actually two different ways to postulate it. Again, it can be shown that both yield the same results as the Euler-Lagrange equations in classical mechanics.

---

deceiving. At any rate, there can be no more independent initial conditions than those for all independent coordinates and speeds in the formulation using Newton's law (2.2).

[2]The reason for talking about physical trajectories instead of a single trajectory is to include also cases in which there are two equivalent trajectories, e. g. for moving in both possible directions from one point of the circle to the opposite point. These trajectories are not deformable into each other. Such a situation occurs if there are, like with the circle, holes in a generalized sense. This unusual, but by no means irrelevant, situation will be noted if occurring. Otherwise in the following only a single physical trajectory will be relevant.

The first version is that the physical trajectory is the one trajectory in $\mathcal{M}$, which extremalizes the action. It seems to be odd to define the solution using a manifold which was defined using the solution. However, it will turn out that it is in general much simpler to specify $\mathcal{M}$ even without knowing the actual physical trajectory than to solve the system. This postulate already emphasizes the non-local nature of Hamilton's principle. To know the solution requires not only to know the possible trajectories at all times, but the whole set $\mathcal{M}$. This is very different from the Euler-Lagrange equations or Newton's law, which only needed to know about the function and its derivatives at one point to build the solution. That this is still the more powerful formulation is hence far from obvious.

One advantage is, however, visible: The action does not depend on the coordinate system in configuration space, since it is entirely formulated on paths. It is therefore a coordinate-independent classical mechanics, therefore eliminating the human-made coordinate systems.

5.1.2 Variational formulation

The alternative formulation uses again the concept of virtual displacements. It requires that the physical path is the one where the change of S under virtual displacements of the path in configuration space, $q \to q + \delta q$, vanishes, expressed as

$$\delta S = 0.$$

While this seems to be a very compact notation, it is mathematically far from obvious what it means: How is the variation of a functional calculated? Understanding this is the first part of functional calculus, which is a wide field, which plays an important role in classical as well as quantum field theory. Here, these details will be skipped.

Of course, it is not yet obvious that this formulation is equivalent to the integral version, but also this can be shown. This will amount to the statement that an extremal trajectory has zero variation of the action, which is, not accidentally, reminiscent of the way how extrema of functions are found.

In fact, though technical, it can be shown that if a trajectory satisfies Lagrange's equation then it extremalizes the action. Thus, in a sense, Lagrange's equations are to the action what the derivative is to an ordinary function.

However, this makes no statement yet about whether the extremum is a maximum, minimum, or saddle point. As with ordinary functions, this can be calculated. Also, there may be more than one extrema or turning point in general. However, for physics any case will do. In practice, most mechanical problems lead to minima of the action, some to a maximum. Saddlepoints play almost never a role.

There is also a very deep conceptual change hidden here. With the absence of coordinates, there is no longer the notion of the position of a particle. Rather, the only remaining objects are trajectories. The points along a trajectory can be enumerated using the time-variable, which is thus only a pointer on the trajectory. Thus, particles can likewise be taken to be only pointers onto the trajectories. This would essentially imply that trajectories are the fundamental objects of classical mechanics, or at least equivalent to the particles in Newton's formulation.

5.2 Phase space and state space

Though the concept of trajectories was extremely useful in reformulating mechanics in a more compact way, it has one essential drawback: It requires to know not only the initial state of a system, but also the final state of a system. Though from a mathematical point of view there is no difference between this requirement and the requirement of knowing position and speed at the initial time, in practice this is often quite different. There, it is usually only known how the system starts, and the question is how it will develop with time. This is especially true when it comes to problems which can no longer be solved in a closed, analytic form. And this is true for almost all systems of interest.

It is therefore useful to reformulate Hamilton's principle as an initial value problem. At first sight, it may seem natural to use the generalized coordinates and speeds for this. But, in fact, it is much more useful to rather use the generalized coordinates and momenta, the latter as defined in (3.5). The reason is that cyclic coordinates imply conserved generalized momenta. Formulating the problem thus in these two quantities will make the whole problem trivial for these directions. Thus, they are advantageous, as no such simple relation exists for the generalized speeds.

Thus, the aim in the following is to reformulate the system in terms of the generalized coordinates and generalized momenta. There are again $2(dn - N)$ of these, thus forming again a space of the same dimensionality. To separate this space from the configuration space before, it is called the phase space.

It is quite useful to reiterate the various concepts of spaces encountered so far.

The basic space is the coordinate space, which has dn coordinates, where d is the number of coordinates and n is the number of involved particles. This can be upgraded to space-time with $dn+1$ dimensions by adding time. This is true for the effective low-speed system of classical mechanics but for the special relativity version, which both differ by the metric, being either Euclidean or Minkowski, requires an individual time for every particle. However, the time plays a different role than the other coordinates in classical

mechanics: Here it parametrizes the trajectories of particle in configuration space. In the relativistic case, this role is played rather by the eigentime.

By introducing generalized coordinates using N constraints, the number of coordinates is reduced to $S = dn - N$. This is the configuration space. The paths of all particles are uniquely described by trajectories in this space. If time is included to characterize the progression along the trajectories, this becomes the $S + 1$-dimensional event space: Every event is localized at one point in this space.

However, both in ordinary space-time as well as in event space it is not sufficient to specify a point to specify a system completely. Even in event space, trajectories may cross. Only by adding generalized speeds the trajectory of a given particle is uniquely determined[3].

This problem is resolved by the introduction of phase space. By doubling the degrees of freedom by adding the generalized momenta, a $2S$-dimensional space is obtained. Trajectories do no longer cross. The position along the trajectories can be parametrized by the time. Thus, the position and future movements of any particle of a system is uniquely defined when a position in this space is given. It thus fully describes the state of a system. Conversely, a state is called the minimal collection of information necessary to fully describe a system. Thus, this $2S + 1$-dimensional space is called the state space[4]. Though the state space has been used to define what a state is in the sense of classical mechanics, any other minimal way is equally good. Also, this concept can be extended to any physical system. E. g., in thermodynamics quantities like temperature and pressure may be part of the minimal information defining a state.

Returning to classical mechanics, this conversely implies that measuring all the minimal information, i. e. determining the point in state space at which a system is located, gives all information about a classical mechanics system[5]. Since a single point is sufficient to

[3]That this is sufficient is because Newton's law (2.2), which ultimately underlies any classical mechanics problem, is a differential equation of second order in time. There exists a general mathematical theorem which guarantees that no more initial conditions are needed to uniquely identify a trajectory. If Newton's law would be a third-order differential equation, this would no longer be true, and three initial conditions would be required, and so on. The fact that Newton's law is of second order is nothing which can be derived, but, as noted in section 2.2, is a postulate, derived from observations. Note, however, that there are rare multi-particle systems which are described by higher-order differential equations. However, in these cases this is also only an effective rewriting of initial conditions of multiple particles in the form of a single effective particle.

[4]Consequently, if Newton's law would be rather of mth-order in time, a state space which gives uniquely defined trajectories would be of $mS + 1$ dimensions.

[5]This is the reason why classical mechanics is a deterministic theory. This will no longer be possible in quantum physics.

fully describe a state this implies that a first-order set of $2S$ differential equations in time should exist which describes the evolution of a state. This will be done in the following, leading ultimately to Hamilton's equations in section 5.3. Lagrange's equation are not yet of this type, as they are second order in time.

5.3 Hamilton's equations

This is not entirely straightforward. Since the generalized momenta are known as function of the generalized coordinates and speeds, it is possible to invert the various relations to obtain, say, the Lagrange function as a function of the new quantities. But, this problem can be put into a larger context and a general recipe for this type of transformations can be given. In addition to making the procedure more simple for mechanics, the same approach can be used for a wide range of problems, and will play, e. g. again an important role in thermodynamics. This procedure is called Legendre transformation. A proof of how it works will be skipped here.

As Lagrange's function is still the basic dynamical quantity, the first step is to apply the Legendre transformation to it. Using the generalized momenta

$$p_i = \frac{\partial L}{\partial d_t q_i}$$

yields

$$H(q_i, p_i, t) = p_i d_t q_i(q_j, p_j) - L(q_i, d_t q_i(q_j, p_j), t).$$

However, this function is already known. It is Hamilton's function (3.6), which coincides under certain conditions, homogeneous kinetic energy of order 2 and a conservative system, with the total energy $H = T + V$. Though, at that time the generalized momenta was considered as a function of the generalized coordinates and speeds. It will be the Hamilton function which will take over the role of the central quantity from the Lagrange function for most of the rest of these notes[6].

After transforming the Lagrange function, the next interesting question is how this affects the equations of motion, especially, what the equivalent reformulation of the Euler-Lagrange equations in phase space are. After suitable mathematical transformations these

[6] As discussed previously, this is different in relativistic systems.

are found to be Hamilton's equations

$$d_t q_i = \frac{\partial H}{\partial p_i} \tag{5.2}$$

$$d_t p_i = -\frac{\partial H}{\partial q_i} \tag{5.3}$$

$$-\frac{\partial L}{\partial t} = \frac{\partial H}{\partial t}, \tag{5.4}$$

which are the new equations of motion determining the dynamics as the Euler-Lagrange equations did before. They are also called canonical equations. The last equation is actually not a real equation of motion, but rather a consistency condition, as it is not an equation involving a separate variable. The equations (5.2-5.3) are first order in time, and are therefore describing the evolution of a state in state space with knowledge of a single point, as was aimed at in section 5.2.

The most remarkable result is that Hamilton's equations are now first-order differential equations, rather than the second-order differential equations of Lagrange. In practice, first-order equations are usually easier to solve, and therefore Hamilton's equations are technically a step forward. The disadvantage is that the number of equations of motion to be solved has doubled, but in practice this is still better[7]. The result when solving these equations is then no longer a trajectory in configuration space, but rather a trajectory in phase space.

Though in general the Hamilton function is the Legendre transform of the Lagrange function, under certain conditions discussed in section 3.7.1, it is just the total energy of the system, $H = T + V$. This rather simple relation to the Lagrange function $L = T - V$ comes about as this is true if the kinetic energy is at most quadratic in the speeds. Then the first term in the Legendre transformation taking care of the differential properties is $2T$, and therefore $H = 2T - T + V = T + V$ is the total energy. Though this is the most relevant case, this is by far not always the case.

Similar as there, it can be shown that

$$d_t H = \partial_t H.$$

This implies that if the Hamilton function is not explicitly depending on time, $\partial_t H = 0$, it is a conserved quantity, and actually constant as a function of time. Thus, the time-dependencies of the q_i and p_i have to cancel each other inside the Hamilton function in

[7]Note that ordinary differential equations of second order can always be rewritten as a set of twice as much first-order differential equations. This is not necessarily so for partial differential equations as the equations of motion are. The result is therefore not trivial on mathematical grounds.

this case. If the Hamilton function coincides with the energy, the energy is in this case conserved.

Furthermore, because of (5.3) any cyclic coordinate in the Lagrange function is necessarily also cyclic in the Hamilton function. In case of a cyclic coordinate this also implies that one of Hamilton's equations of motion is solved trivially. The actual value of the corresponding generalized coordinate is then determined by the initial conditions. This is quite different from the Lagrange case, as the Lagrange function depends on the generalized speeds rather than the generalized momenta, and therefore the existence of a cyclic coordinate does not immediately solve any of the equations of motion. This is again a practical advantage of the Hamilton formalism over the Lagrange formalism.

The disadvantage, stated without proof here, of the Hamilton formalism is that it is more complicated in the relativistic case.

Note that (5.2) implies that if a generalized momentum is cyclic, then the corresponding coordinate is constant as well. This case is, however, much rarer then that of a cyclic coordinate.

It is quite useful to study the procedure from the original system up to Hamilton's equations and the solution of Hamilton's equations for an example. To facilitate the comparison to previous results, this will be the harmonic oscillator, though Hamilton's formalism is overkill for it.

In this case, there is a single generalized coordinate $q = x$ describing the position of the oscillator. The kinetic and potential energy of this conservative system is

$$\begin{aligned} T &= \frac{m}{2}(d_t q)^2 \\ V &= \frac{k}{2}q^2. \end{aligned}$$

This yields the Lagrange function

$$L = T - V = \frac{m}{2}(d_t q)^2 - \frac{k}{2}q^2,$$

in which no coordinate is cyclic. However, it does not depend explicitly on the time, and therefore the energy will be conserved. The generalized momentum is

$$p = \frac{\partial L}{\partial d_t q} = m d_t q.$$

The Legendre transformation yields the Hamilton function

$$H = p d_t q - L(q, d_t q(q,p)) = \frac{p^2}{m} - \frac{1}{2m}p^2 + \frac{k}{2}q^2 = \frac{1}{2m}p^2 + \frac{k}{2}q^2 = T + V,$$

which coincides with the conserved energy. This could be used to eliminate either q or p. But it is more instructive to use Hamilton's equations (5.2-5.3), being

$$\begin{aligned} d_t q &= \frac{\partial H}{\partial p} = \frac{p}{m} \\ d_t p &= -\frac{\partial H}{\partial q} = -kq \end{aligned}$$

This set of equations can be solved in multiple ways. Either, it is possible to replace p or q by using the conserved total energy, or by inserting one equation into the other, yielding again a second-order differential equation. At any rate, the solutions are

$$\begin{aligned} q(t) &= q_0 \sin\left(\frac{k}{m}t + \phi_0\right) \\ p(t) &= \frac{mq_0}{k}\cos\left(\frac{k}{m}t + \phi_0\right), \end{aligned}$$

with the two integration constants. The total energy is

$$H = E = \frac{kq_0^2}{2}\cos^2\left(\frac{k}{m}t + \phi_0\right) + \frac{kq_0^2}{2}\sin^2\left(\frac{k}{m}t + \phi_0\right) = kq_0^2.$$

Thus, the energy is completely determined by the initial conditions. The trajectory in phase space is a closed ellipsoid, where the relative size of the major axes depends on the initial conditions as well.

Such a closed path in phase space is typical for periodic systems. If the path is not closed, but both the coordinates as well as the generalized momenta are bounded, this is called an aperiodic system.

All of this can be transferred to the relativistic case, but becomes technically quite involved, even for the free particle. But it allows, e. g., for a consistent description of massless particles. Nonetheless, because of the amount of technicalities involved, this will be skipped here.

5.4 Canonical transformations

One of the important insights in chapter 3 was in section 3.5 that the formulation using Lagrange's equations is invariant of the particular chosen coordinates. The formulation of mechanics in terms of Hamilton's principle gives this a conceptual more important interpretation. Since the objects are now trajectories, rather than the generalized coordinates and speeds, the coordinate-system-invariance is the statement that it does not matter how the trajectories are described. The actual dynamical objects of mechanics are the trajectories, not the coordinates.

The introduction of Hamilton's equations requires to reinvestigate these statements. The question is, whether the trajectories in coordinate space can be replaced by states, and thus whether Hamilton's equations are form-invariant under changes of the so-called canonical coordinates of generalized coordinates and momenta. However, this does not require any deep calculations. When changing the generalized coordinates and speeds to new ones Q and $d_t Q$, which are arbitrary invertible and continuously differentiable functions of time, and reexpressing the Lagrange function in these new coordinates, it was already shown in (3.5) that the equations of motion remain form-invariant. Defining thus new generalized momenta as

$$P_i = \frac{\partial L}{\partial d_t Q_i}, \tag{5.5}$$

the same derivation of Hamilton's equation can be performed as before, exploiting the form-invariance of the Euler-Lagrange equations, and thus arriving at the same form of Hamilton's equations (5.2-5.4), but now for the new coordinates.

Furthermore, in section 3.5 it was said that it is possible to add to the Lagrange function a total time derivative of a function depending only on the generalized coordinates and time, (3.4), and this left the Euler-Lagrange equations also untouched. This has a somewhat different impact in Hamilton's formulation. Deriving from (3.4) the generalized momenta yields

$$p_i^f = p_i + \frac{\partial f}{\partial q_i}. \tag{5.6}$$

That is, adding such a function changes the generalized momenta, while the generalized coordinates remain unchanged, $q_i^f = q_i$. This changes the Hamilton function as

$$H^f = H - \partial_t f.$$

Thus, while the Lagrange function is shifted by a total derivative, the Hamilton function is shifted by a partial derivative. The canonical equation take the form

$$\begin{aligned} \frac{\partial H^f}{\partial p_i^f} &= d_t q_i^f \\ \frac{\partial H^f}{\partial q_i^f} &= -d_t p_i^f. \end{aligned}$$

Thus, the equations are indeed also form-invariant.

While this result is in itself not totally surprising, as the Lagrange and the Hamilton formalism are for any sets of coordinates equivalent as discussed above, this result has some far-reaching consequences. It implies that it is possible to locally redefine the generalized momenta, i. e. change the generalized momenta at every point in space and time almost

arbitrarily, and still get the same physics. At the same time, the trajectories remain fixed, as the generalized coordinates do not change: The particle still move in the same way. And also with the same speed, as also the generalized speeds are not changed. This implies that the generalized momenta are not unique, but can be changed locally. This is a much stronger statement as a change of coordinate system, as this is done for all coordinates in the same way, it is global. This arbitrariness implies that some of the information contained in the generalized momenta is arbitrary. This is not a surprise. As noted above, the basic object is the trajectory. Any solution with the same trajectory will provide the same physics. In the Lagrange formulation, this arbitrariness is not there, as the generalized speeds are uniquely determined once the trajectories are known. This is, as seen here, not true for the generalized momenta.

Such a local arbitrariness is also known as a gauge freedom. Its generalization plays a very important role in physics. Why this is the case can already be seen for the present case in classical mechanics.

This freedom implies that the possible redefinitions of variables in the Hamilton case is much larger than in the Lagrange case. There it was only possible to change the generalized coordinates and speeds in a way which maintained the fact that the generalized speeds were derivatives of the generalized coordinates. Here, there is more freedom. And this freedom can be used to simplify problems. That is the idea behind the following concept.

While it was straightforward than any transformation changing generalized coordinates and speeds left Hamilton's equations invariant, this is actually not possible for an invertible and differentiable point transformation of generalized coordinates and momenta[8]

$$Q_i = Q_i(q_j, p_j, t) \tag{5.7}$$
$$P_i = P_i(q_j, p_j, t). \tag{5.8}$$

Point transformations, which satisfy some Hamilton function's $h(Q_i, P_i, t)$ Hamilton's equations

$$d_t Q_i = \frac{\partial h(Q_i, P_i, t)}{\partial P_i} \tag{5.9}$$
$$d_t P_i = -\frac{\partial h(Q_i, P_i, t)}{\partial Q_i} \tag{5.10}$$

are therefore special, and are called canonical. If also

$$h(Q_i, P_i, t) = H(q_i(Q_i, P_i, t), p_i(Q_i, P_i, t), t),$$

[8]Note that the P_i do not necessarily fulfill the relation (5.5). It will be necessary to specify the conditions on the point transformation when this is the case.

where the Hamilton function satisfies the Hamilton equations in the coordinates q_i and p_i, it is called a proper canonical transformation, but this is not required for a transformation to be canonical. This is a very important distinction. Proper canonical transformations provide a constructive way to create the Hamilton function. Improper canonical transformations only require the existence of some Hamilton function such that (5.9-5.10) is fulfilled. No tool is yet provided to proof the existence of such a canonical transformation, lest alone construct it. Finding (im)proper canonical transformations could therefore be much more complicated.

To get a better idea of how far canonical transformation can go, it is best to consider first an example, before trying to construct general tests for canonicity of transformations.

As a first example, consider a transformation which exchanges generalized coordinates and momenta,

$$Q_i = -p_i \tag{5.11}$$
$$P_i = q_i. \tag{5.12}$$

This transformation is a proper canonical transformation, since for the Hamilton function $h(Q_i, -P_i, t) = H(-p_i, q_i, t)$

$$\frac{\partial h}{\partial P_i} = d_t Q_i$$
$$\frac{\partial h}{\partial Q_i} = -d_t P_i.$$

This is a quite remarkable result. It implies that generalized coordinates and momenta can be exchanged at will, or even partially. Hamilton's formulation does not distinguish between both, emphasizing again that the trajectory is the basic object, and the generalized coordinates and momenta are just description without inherent importance of their own.

So, how far can this be driven? Is it possible to find a canonical transformation which makes coordinates cyclic? Which makes all coordinates cyclic? Then, the problem would become trivial. The answer to this is affirmative, and will be given in section 5.8. Unfortunately, while possible in principle, in practice this is often as hard as solving the original problem. But before this can be achieved, further developments are necessary. The first is finding a criterion for the canonicity of a transformation.

5.5 Poisson brackets

Without proof, a transformation can be most directly be identified to be canonical using a new concept, the so-called Poisson brackets, which will be developed in the following.

Though it is in the end a very compact way of expressing the conditions of canonicality, the Poisson brackets are again a concept with an importance which will only become fully unveiled in quantum physics. The Poisson brackets will also be very useful to formulate many results in a very compact way, though it can be argued whether this can be considered to be an advantage.

While it is possible to just define the Poisson brackets and then show its usefulness, it is arguably better to show that it emerges automatically from a very general question. This question is: Given a classical mechanical system, there may be many quantities of interest. So far, the main focus were on the trajectories of the individual particles. However, other questions may be much more interesting. E. g., how often particles collide, how many particles appear in which area of the system, how often does a particle orbit another particle before escaping to infinity. This list can be extended arbitrarily and infinitely.

Since any mechanical system is uniquely characterized by a state, any such observable f can only depend on the state, i. e. the generalized coordinates, momenta, and the time, $f(p_i, q_i, t)$. An interesting question is, how this quantity changes with time. Of course, solving the system and inserting the results for the generalized coordinates and momenta would answer this question. But it is very often the case that it is already sufficient to know the answer as a function of the generalized quantities. Especially, this provides a result independent of any particular initial conditions, which are needed for any concrete realizations of the trajectories.

The answer to this question is

$$d_t f = \frac{\partial f}{\partial q_i}\frac{\partial H}{\partial p_i} - \frac{\partial f}{\partial p_i}\frac{\partial H}{\partial q_i} + \partial_t f, \tag{5.13}$$

which can be shown using Hamilton's equations of motion (5.2-5.3).

This result motivates to define the Poisson brackets as the following prescription. Given two functions f and g depending on two set of variables p_i and q_i, the Poisson bracket with respect to the two sets of variables is defined as

$$\{f, g\} = \frac{\partial f}{\partial q_i}\frac{\partial g}{\partial p_i} - \frac{\partial g}{\partial q_i}\frac{\partial f}{\partial p_i}.$$

The name brackets originates from the curly braces on the left-hand-side. The Poisson brackets are therefore a mathematical operation, akin to, say, multiplication.

The result (5.13) can therefore be expressed using the Poisson bracket as

$$d_t f = \{f, H\} + \partial_t f. \tag{5.14}$$

Thus, up to any explicit time dependence, the time evolution of a mechanical quantity is determined by its Poisson bracket with the Hamilton function. This gives the Hamilton

function an extremely elevated conceptual importance, as it can therefore be considered as the source of evolution in time.

This is also emphasized by

$$\begin{aligned} d_t q_i &= \{q_i, H\} = \frac{\partial H}{\partial p_i} \\ d_t p_i &= \{p_i, H\} = -\frac{\partial H}{\partial q_i}, \end{aligned}$$

giving a reformulation of Hamilton's equations (5.2-5.3) using the Poisson brackets. Therefore, Hamilton's equations can be considered as special cases of the more general time evolution equation (5.14).

A special case are also the so-called fundamental Poisson brackets

$$\{q_j, q_k\} = 0 \tag{5.15}$$

$$\{p_j, p_k\} = 0 \tag{5.16}$$

$$\{q_j, p_k\} = \delta_{jk}. \tag{5.17}$$

These are the Poisson brackets of the sets of variables itself, therefore the name fundamental, as these are the most elementary Poisson brackets since no further function is involved. Thus, only mixed fundamental Poisson brackets do not vanish. Though these seem to be comparatively simple statements, modifications of the relations (5.15-5.17) can serve as the fundamental postulates of quantum physics.

The time evolution equation (5.14) has a particular interesting implication for constants of motions, i. e. for quantities with $d_t f = 0$. In this case

$$0 = d_t f = \{f, H\} + \partial_t f.$$

This criterion is often much easier to check than any other criteria, besides cyclic coordinates, found so far, as it does not require the explicit solution of the equations of motion. Especially, for any not explicitly time-dependent quantity it follows that

$$\{f(q_i, p_i), H\} = 0 \Rightarrow d_t f = 0, \tag{5.18}$$

and is a very compact test for conserved quantities. For the Hamilton function itself this implies

$$d_t H = \{H, H\} + \partial_t H = \partial_t H,$$

i. e. the statement of section 3.7.1 is recovered that the Hamilton function is conserved if it does not depend explicitly on time. This will be put in a different perspective in section 5.7.

5.6 Poisson brackets and (canonical) transformations

One of the most important features of the equations of motions, as discussed already for the Lagrange equations of the second kind in section 3.5, is that they do not depend on a particular coordinate system, and that they keep the same form in any coordinate system. Studying this feature in detail for the Poisson brackets, and thus especially of the dynamical time evolution formulated in (5.14), will lead back to the canonical transformations of section 5.4.

Consider two sets of generalized variables q_i and p_i, as well as Q_i and P_i, such that the relations

$$\begin{aligned} Q_i &= Q_i(p_j, q_j) \\ P_i &= P_i(p_j, q_j) \end{aligned}$$

are invertible relations, i. e. there are uniquely defined functions

$$\begin{aligned} q_i &= q_i(Q_j, P_j) \\ p_i &= p_i(Q_j, P_j), \end{aligned}$$

just as in section 3.5. In particular,

$$H(p_i, q_i) = H(q_i(Q_j, P_j), p_i(Q_k, P_k)) = H(Q_l, P_l),$$

i. e. the change of variables is canonical.

It can now be shown that

$$\{Q_i, Q_j\} = \{P_i, P_j\} = 0 \tag{5.19}$$

$$\{Q_i, P_j\} = \delta_{ij}. \tag{5.20}$$

This implies that the fundamental Poisson brackets have the same form, no matter the choice of coordinates. In fact, it can be shown that the transformation is canonical if and only if the fundamental Poisson brackets remain unchanged. The proof of this will be skipped.

5.7 Poisson brackets and symmetries

Poisson brackets give another conceptual view on invariances and symmetries, and thus the conservation laws of section 3.7.

Consider an infinitesimal canonical transformation of the coordinates and momenta

$$\begin{aligned} Q_i &= q_i + \delta q_i \\ P_i &= p_i + \delta p_i, \end{aligned}$$

where δx should for now only indicate that the quantity δx is small, and is not connected to displacements. It can be shown that there exists some function $G(q_i, p_i)$ such that

$$\begin{aligned} P_i - p_i &= \delta p_i = -\epsilon \frac{\partial G}{\partial q_i} \\ Q_i - q_i &= \delta q_i = \epsilon \frac{\partial G}{\partial p_i} \end{aligned}$$

where ϵ is now a sufficiently small number.

So far, this is general. It is interesting to consider the case that $G = H$ and $\epsilon = dt$ a small interval in time. This leads to

$$\begin{aligned} \delta q_i &= dt \frac{\partial H}{\partial p_i} = \frac{dq_i}{dt} dt = dq_i \\ \delta p_i &= -dt \frac{\partial H}{\partial q_i} = \frac{dp_i}{dt} dt = dp_i. \end{aligned}$$

Thus, this canonical transformation actually pushes the generalized coordinates and momenta forward in time. This implies time evolution is equivalent to an (infinitesimal) canonical transformation. By consecutive such infinitesimal transformations ultimately the whole time evolution can be build. It is therefore valid to say that the Hamilton function creates the evolution in time, which is quite similar to the statement (5.14).

Consider now some arbitrary function of u, which can be shown to change as

$$\delta u = \epsilon \{u, G\}. \tag{5.21}$$

Thus, the change of a quantity u under some (infinitesimal) coordinate transformation G is determined by the Poisson bracket of this quantity with the coordinate transformation. Especially,

$$\delta H = \epsilon \{H, G\}, \tag{5.22}$$

and thus also the Hamilton function itself changes in general.

If the function G is a constant of motion then, because of (5.18), its Poisson bracket with the Hamilton functions vanishes. In the present context, this implies that any coordinate transformation generated by an integral of motion leaves the Hamilton function unchanged.

This statement is actually a very deep one. As has been discussed in section 3.7 an integral of motion is usually associated with some symmetry of the Hamilton function. Thus, (5.21) together with (5.22) states that the integral of motion associated with a

symmetry generate changes of quantities, but leave the Hamilton functions invariant. In fact, the function G will be seen to generate the transformation of any quantity under the corresponding symmetry transformation. In quantum physics, this will be the key to identify symmetries of the systems and all the dynamical features of the theories. It is one of the key properties in the connection of symmetries and physics.

However, a full general proof of the connection is beyond the present scope.

An interesting consequence can be obtained from these insights. As has been shown, the time-evolution along any trajectory in phase space is generated by the Hamilton function. Since trajectories are continuous functions, this is equivalent to an infinitesimal rotation and translation in phase space. Note that similar considerations do not apply to either position space or configuration space. Thus, phase space is the fundamental space of (classical) mechanics.

5.8 Hamilton-Jacobi theory

It can be shown that it is possible to find for any mechanical problem, which is formulated by a Hamilton function, a canonical transformation such that the problem becomes trivial. Trivial in this case means that it is either transformed to a known problem, or to a situation in which all coordinates becomes cyclic and the Hamilton function is explicitly time-independent or finally such that all generalized coordinates become constant, and thus yielding trivial Hamilton's equations.

The drawback is that, while this is possible in principle, it is often as complicated or more complicated than the original problem. The advantage is that, even if practically tedious, the knowledge that it is possible in principle helps often substantially in gaining conceptual insights. Thus, it is an important development.

While this provides further important insights, especially it shows that all trajectories can be transformed either into the closed trajectories of the harmonic oscillator or the straight trajectories of a free particles, the details become exceedingly technical. Overall, it can be considered to be a local deformation of phase space such that all trajectories are deformed to become of either of two types.

Thus, proofing it shows that classical mechanics (in absence of dissipation or similar emergent phenomena) can sustain only trajectories which are continuously deformable to either of these two kinds of trajectories. Thus, mechanics knows only two movements, free and periodic, and the rest are non-trivial, but smooth, deformations of them.

5.9 The geometry of phase space

Classical mechanics has actually an interesting geometric interpretation, which warrants some thoughts. This is especially the case as geometric structures are prevalent in physics, and reappear in one way or another in all fundamental theories known.

This will become most evident when using phase space, and thus Hamiltonian mechanics. Note that phase space is always even-dimensional, with dimension $2n$ in the following. In such spaces there exist a particular group of matrices, the so-called symplectic matrices, $\mathrm{Sp}(2n)$. These are the matrices A with the property

$$\begin{aligned} A^T J A &= J \\ J &= \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \end{aligned} \tag{5.23}$$

where the elements of J are either zero matrices or (negative) unit matrices, each of dimension n. Note for further use that $J^{-1} = J^T = -J$ and $J^2 = -1$. It can be shown that for the submatrices of the matrix A

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

holds that

$$\begin{aligned} (a^T c)^T &= a^T c \\ (b^T d)^T &= b^T d \\ a^T d - c^T b &= 1. \end{aligned}$$

Consider now a not explicitly time-dependent Hamilton function H, and the two vectors

$$\begin{aligned} \vec{v} &= \begin{pmatrix} q_i \\ p_i \end{pmatrix} \\ \vec{\partial} &= \begin{pmatrix} \partial_{q_i} \\ \partial_{p_i} \end{pmatrix}, \end{aligned}$$

where the subvectors are n-dimensional, in phase space. Construct the vector

$$J\vec{\partial}H = \begin{pmatrix} \partial_{p_i} H \\ -\partial_{q_i} H \end{pmatrix}, \tag{5.24}$$

which contains Hamilton's equations. Especially, Hamilton's equation (5.2-5.3) take the compact form of a vector equation

$$\partial_t \vec{v} = J\vec{\partial}H.$$

Thus, the expression (5.24) describes the trajectory in phase space.

As a first example how this formulation could be advantageous, consider the total time derivative of H,

$$d_t H = (\vec{\partial} H)^T J \vec{\partial} H = 0.$$

Thus, the time-independence of the Hamilton function becomes now a geometric feature of phase space, as this hinged only on the features of the matrix J, rather than any other derivatives. It is a statement about the trajectories generated by the function H. This is quite similar in spirit to what the potential does - there it is a gradient in ordinary space which creates by Newton's equations the movement of the system. Here, the gradient of the Hamilton function in phase space, supplemented by the matrix J, creates by Hamilton's equation the movement, and the conservation of the Hamilton function along the trajectories is a geometric consequence of this.

Now, combine these insights with those of section 5.7. There it was shown that movement along a trajectory is created by the Hamilton function,

$$\begin{aligned} dq_i &= \frac{\partial H}{\partial p_i} dt \\ dp_i &= -\frac{\partial H}{\partial q_i} dt, \end{aligned}$$

which was actually a canonical transformation. This implies that a trajectory develops in phase space as

$$d\vec{v} = J \vec{\partial} H dt.$$

In addition, any canonical transformation is given similarly by

$$d\vec{v} = J \vec{\partial} G,$$

where the function G was defined in 5.7. As has been noted there, time evolution was anyhow only a special case of such a transformation. Thus, the symplectic velocity field $J\vec{\partial}G$ creates all (infinitesimal) canonical transformations in phase space.

It is even possible to state more. Consider the matrix of derivatives of the canonical transformations

$$M = \begin{pmatrix} \frac{\partial q_i}{\partial Q_j} & \frac{\partial q_i}{\partial P_j} \\ \frac{\partial p_i}{\partial Q_j} & \frac{\partial p_i}{\partial P_j} \end{pmatrix}.$$

Because the canonical transformations are invertible, it follows that

$$M^{-1} = \begin{pmatrix} \frac{\partial Q_i}{\partial q_j} & \frac{\partial P_i}{\partial p_j} \\ \frac{\partial Q_i}{\partial q_j} & \frac{\partial P_i}{\partial p_j} \end{pmatrix}.$$

It then follows by explicit evaluation that

$$-JM = (JM^{-1})^T$$

holds. But this just defines a symplectic transformation, as this is equivalent to (5.23),

$$J = M^T J M.$$

Thus, a finite canonical transformation can be characterized by a symplectic matrix. As a consequence, the scalar product

$$\vec{v} \cdot \vec{w} \equiv \vec{v}^T J \vec{w},$$

with v and w two vectors in phase space is invariant under any canonical transformations. A space endowed with such a metric is called a symplectic space or symplectic manifold.

Also the Poisson brackets can be integrated into this formulation,

$$\{F, G\} = -(\partial F)^T J (\partial G). \tag{5.25}$$

Thus, it is a scalar product in this symplectic space. Because such a scalar product is invariant under canonical transformations, as noted above, the invariance of Poisson brackets under canonical transformations, obtained in section 5.6, follows here trivially from the geometry of phase space. Especially,

$$\{v_i, w_j\} = -J_{ij}.$$

Therefore, the fundamental Poisson brackets (5.15-5.17) in fact define the metric of phase space. Since in (5.25) the expression $J\partial G$ defines a flow along a canonical transformation, the Poisson bracket acts as directional derivative with respect to the flow ∂F.

5.10 Continuum mechanics

Many systems in mechanics contain a very large number of point particles. In many of these cases it becomes suitable to deal with these particles rather in terms of density than of the individual particles, as already hinted at in section 2.10. In this case, matter is considered as a continuum, described by a mass density $\rho(\vec{r})$. This is the purview of continuum mechanics.

5.10.1 Systems of many oscillators

Before actually treating matter as a continuum system, it is very helpful to first consider a very special system with many degrees of freedom, that of many (coupled) oscillators. Such systems in themselves are relevant, as many physical systems can be described by them, especially in solid state physics. Furthermore, taking the limit of an infinite number

of these oscillators already provides a good description of many continuum systems, and it is therefore technically useful to first study a finite, but large, such system. Moreover, according to section 2.11, any system close to a stable equilibrium is also described by a system of many oscillators.

Taking for N oscillators with coordinates q_i the values q_i^0 to be the equilibrium position, it is then useful to introduce the fluctuation coordinates

$$\eta_i = q_i - q_i^0.$$

Moreover, if the system should remain close to equilibrium also the kinetic energy needs to be close to zero. Considering only conservative systems with the kinetic energy a quadratic function in the generalized speeds, the total Lagrange function reads

$$\begin{aligned} L &= m_{ij}\frac{d\eta_i}{dt}\frac{d\eta_j}{dt} - V_{ij}\eta_i\eta_j \\ V_{ij} &= \left.\frac{\partial^2 V}{\partial q_i \partial q_j}\right|_{q_i=q_i^0, q_j=q_j^0}, \end{aligned}$$

where the constant term in the potential has been dropped and it was permitted that the kinetic term has also off-diagonal elements in the generalized coordinates. The Euler-Lagrange equations are

$$m_{ij}\frac{d^2\eta_j}{dt^2} + V_{ij}\eta_j = 0, \tag{5.26}$$

and therefore form a set of N coupled differential equations. Fortunately, this type of coupled differential equations can be solved in closed form.

Because each individual equation has the same form as an harmonic oscillator, an ansatz of type

$$\eta_i = \Re\left(a_i e^{-i\omega t}\right)$$

appears as a reasonable starting point. Intentionally here only the same frequency is used for the whole set of oscillators. This is based on the insight from section 2.6 that an harmonic oscillator under an external influence starts to behave like the external influence. Thus, it appears possible that there are common oscillation of all oscillators, as all the other oscillator act on each one like an external influence. The different possible solutions will eventually be found to be different values for ω, which then can be superposed to find arbitrary solutions.

This turns the system of differential equations (5.26) into a set of linear equations for the amplitude factors a_i,

$$V_{ij}a_j - m_{ij}\omega^2 a_j = 0. \tag{5.27}$$

Such a system of equations has only a solution if the determinant of the matrix

$$B_{ij} = V_{ij} - m_{ij}\omega^2 \tag{5.28}$$

vanishes, which follows from the general theory of linear systems of equations. This gives an Nth order polynomial for the frequency ω, which thus yields $2N$ solutions, every time $\pm\omega$, producing all the solutions. Especially, also the complex solutions are admissible, as these will be exponentially damped or enhanced (unstable) oscillations[9]. The amplitudes are the solutions to this homogeneous system, which implies that one of the (non-vanishing) amplitudes will be arbitrarily selectable, and the remaining follow from the solutions of the linear system.

An interesting reinterpretation is possible if the mass matrix is diagonal, $m_{ij} = m_i\delta_{ij}$, with no summation implied. Rescaling the η_i by m_i, the masses drop out, and the problem takes the form of an eigenvalue problem,

$$V\vec{a} = \lambda\vec{a},$$

where $\lambda = \omega^2$ are the eigenvalues. Furthermore, Newton's third law requires that the V are symmetric. This implies that all eigenvalues are real, and thus the so-called eigenfrequencies ω^2 are either purely real or purely imaginary. The vectors of amplitudes are then the eigenvectors of the matrix V, and form a complete basis and are orthogonal with respect to each other. These amplitudes are also called eigenoscillations of the system. Note that this also implies that there exists an orthogonal transformation, which is necessarily also a canonical transformation, which decouples all oscillators, as any symmetric matrix can be diagonalized.

If the matrix m is not diagonal, it is possible to construct a very similar line of reasoning, ultimately leading to equivalent results. In this case, it is possible to reinterpret the problem as a problem in a space which is not Euclidean, but rather has a metric m. In the end, the result are again $2N$ eigenvalues $\pm\omega_k$ and N eigenvectors $\vec{a}$ with arbitrary norm in this metric.

It is useful to consider an explicit example. Take e. g. three particles, connected by two springs with the same strength k, all confined to a linear motion in one dimension. The particles at the ends should have a mass m and the one in the center a mass M. If the length of the springs is a, the potential energy is

$$V = \frac{k}{2}(x_2 - x_1 - a)^2 + \frac{k}{2}(x_3 - x_2 - a)^2.$$

[9]The latter will not appear, if the initial assumption of a system in equilibrium is not violated.

As the equilibrium position[10] will be that the springs are at the normal length a, the difference between the equilibrium positions is also a, and in terms of the fluctuations η_i the potential energy takes the form

$$V = \frac{k}{2}(\eta_2 - \eta_1)^2 + \frac{k}{2}(\eta_3 - \eta_2)^2.$$

The corresponding kinetic energy is

$$T = \frac{m}{2}\left((d_t\eta_1)^2 + (d_t\eta_3)^2\right) + \frac{M}{2}(d_t\eta_2)^2.$$

The matrix B from (5.28) is then

$$B = \begin{pmatrix} k - \omega^2 m & -k & 0 \\ -k & 2k - \omega^2 M & -k \\ 0 & -k & k - \omega^2 m \end{pmatrix}.$$

This form, elements on the diagonal and otherwise only elements on directly adjacent subdiagonal, is characteristic for a one-dimensional system of springs, where only nearest neighbors are connected. Thus, the behavior obtained in the following is qualitatively already rather characteristic of such systems. Such systems play a certain role in many physical contexts, especially in solid state physics and optical lattices.

The next step is to solve the equation

$$0 = \det B = \omega^2(k - \omega^2 m)(k(M + 2m) - \omega^2 Mm)$$

for ω^2. There are three solutions,

$$\begin{aligned} \omega_1 &= 0 \\ \omega_2 &= \sqrt{\frac{k}{m}} \\ \omega_3 &= \sqrt{\frac{k}{m}\left(1 + \frac{2m}{M}\right)}. \end{aligned}$$

The first frequency appears somewhat odd, as this would be no oscillation at all. However, there is a comparatively simple explanation for it. This is a uniform motion of the whole system in either direction. Since the constraints do not forbid it, this is a consistent solution. This could have been excluded by adding a constraint that, say, the center-of-mass coordinate needs to be fixed. This would have reduced the degrees of freedom from three to two, leaving only non-vanishing frequencies.

[10] In the present case, strictly speaking, there would be no need to consider equilibrium positions, as the potential is already of harmonic oscillator form. For the sake of sticking with the developed procedures this will be done nonetheless.

The next step is to solve (5.27) to find the relative sizes, i. e. solve $B\vec{a}^k = 0$. This is an ordinary system of linear equations, which yields the (normalized) solutions

$$\vec{a}^1 = \frac{1}{\sqrt{2m+M}}\begin{pmatrix}1\\1\\1\end{pmatrix} \qquad \vec{a}^2 = \frac{1}{\sqrt{2m}}\begin{pmatrix}1\\0\\-1\end{pmatrix}, \qquad \vec{a}^3 = \begin{pmatrix}\frac{1}{\sqrt{2m\left(1+\frac{2m}{M}\right)}}\\ \frac{-2}{\sqrt{2M\left(2+\frac{M}{m}\right)}}\\ \frac{1}{\sqrt{2m\left(1+\frac{2m}{M}\right)}}\end{pmatrix}.$$

where the normalization is chosen for later convenience. The first case is indeed a collective motion in one direction. The second is an oscillation of the endpoints with respect to a fixed center. The third is a relative oscillation of the center mass with respect to the two other masses, which leaves also the center of mass fixed.

Generalizing to masses connected by effective springs in three dimensions, one lesson is that always the motion of the center of mass will appear, if not constrained. Thus, there will always be $3N-3$ non-trivial eigenfrequencies. Further degeneracies will occur if the system shows a symmetry, as then the symmetry transformation needs to leave the motion unchanged. The non-trivial, non-degenerate eigenfrequencies then fully characterize the system. However, in general it will become more involved to determine them, as it requires to solve an $3N-3$ order polynomial, which is in general not possible in closed form.

It is possible to extent the present treatment to generalize the full case with dissipation and external driving forces. This leads to no new conceptual insights, and will therefore not be done here.

5.10.2 Continuous systems of oscillators

After this preliminary investigation, it is now possible to go to continuous systems.

As a next step consider again the linear chain, but now admit an (arbitrarily large) number N of particles, but which should all have the same mass. The Lagrange function is then given by

$$L = \sum_i \left(\frac{m}{2}(d_t\eta_i)^2 + \frac{k}{2}(\eta_{i+1}-\eta_i)^2\right).$$

In principle, there need to be some additional terms to take care of boundary terms. However, they will not play a role, especially as later the limit $N \to \infty$ will be taken. If now the equilibrium distance between any two particles is equal and of size a, this can be rewritten as

$$L = \sum_i a\left(\frac{m}{2a}(d_t\eta_i)^2 + \frac{ka}{2}\left(\frac{\eta_{i+1}-\eta_i}{a}\right)^2\right) = \sum_i aL_i$$

The quantity $m/a = \mu$ is the mass density of the system. The quantity ka has no direct interpretation. If the connection is indeed in the form of springs, it turns out that for usual springs this product is constant, as a short spring has a larger spring constant, the so-called module of Young Y. Here, it will be required by definition that $ka = Y$, just like m/a, is held constant for any value of a.

Now label the particles not by a discrete index, but rather by their position, i. e. $\eta(x)$ is the displacement of the particle with equilibrium position x against its equilibrium position. Then

$$\frac{\eta_{i+1} - \eta_i}{a} = \frac{\eta(x+a) - \eta(x)}{a} \overset{a\to 0}{=} \frac{d\eta}{dx},$$

i. e. the difference in fluctuations becomes in the limit of smaller and smaller distance a of the particles the derivative, or rate of change, in the displacement η. At the same time the sum becomes a Riemann sum, leading to

$$L = \int dx \left(\mu(d_t\eta)^2 - Y(d_x\eta)^2\right) = \int dx \mathcal{L}(x),$$

where $\mathcal{L}$ is called the Lagrangian density. Likewise, the equation of motion turns from

$$m d_t^2 \eta_i - k(\eta_{i+1} - 2\eta_i + \eta_{i-1}) = 0$$

to

$$\mu d_t^2 \eta(x) - Y d_x^2 \eta = 0 \tag{5.29}$$

which is now an equation for the function $\eta(x,t)$ of two variables, rather than for the finite number of quantities $\eta_i(t)$.

5.10.3 Continuous systems

The previous section outlined everything necessary to define the Lagrange formulation for continuous system. Given a field $\eta(\vec{r},t)$, the Lagrange function is given by a volume integral over the Lagrange density $\mathcal{L}$, sometimes called just Lagrangian,

$$L = \int d^3\vec{r}\mathcal{L}(\eta, \partial_i\eta, \partial_t\eta, \vec{r}, t)$$

which can now depend explicitly on both time and the space coordinate. For the Lagrange function to be scalar, the Lagrange density needs to be a density. At the current time, η is itself also a scalar under rotations. However, it would also be possible to define vector fields or fields of even higher tensors, and indeed this will already in classical electromagnetism be necessary. However, for simplicity here only scalar fields will be considered.

The action is still a time-integral over the Lagrange function,

$$S = \int dt L = \int dt d^3\vec{r}\mathcal{L},$$

and thus a space-time integral over the Lagrange density. This already suggests that an extension to special relativity is rather straightforward, and assigns in such a context the Lagrange density a more fundamental role than the Lagrange function.

It can then be shown that the continuum Euler-Lagrange equations take the form

$$\frac{d}{dt}\frac{\partial\mathcal{L}}{\partial\partial_t\eta} + \frac{d}{dr_i}\frac{\partial\mathcal{L}}{\partial\partial_i\eta} - \frac{\partial\mathcal{L}}{\partial\eta} = 0, \tag{5.30}$$

which are the dynamical equations governing continuum mechanics. In this context, the derivative with respect to a function can be defined as

$$\frac{\partial\mathcal{L}(\eta(x))}{\partial\eta} = \left.\frac{\partial\mathcal{L}(y=\eta(x))}{\partial y}\right|_{y=\eta(x)}, \tag{5.31}$$

and likewise for derivatives with respect to derivatives of functions. Functions and their derivatives are considered as independent variables, e. g.

$$\frac{\partial\mathcal{L}(\eta(x,t),\partial_t\eta(x,t))}{\partial\eta\partial\partial_t\eta} = \left.\frac{\partial\mathcal{L}(y=\eta(x,t),z=\partial_t\eta(x,t))}{\partial y\partial z}\right|_{y=\eta(x,t),z=\partial_t\eta(x,t)}. \tag{5.32}$$

5.10.4 Applications of the Lagrange formulation

Consider again the problem of section 5.10.2. Its Lagrangian density was

$$\mathcal{L} = \frac{1}{2}\left(\mu(\partial_t\eta)^2 - Y(\partial_x\eta)^2\right).$$

The corresponding derivatives are

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\partial_t\eta} &= \mu\partial_t\eta\\ \frac{\partial\mathcal{L}}{\partial\partial_x\eta} &= -Y\partial_x\eta\\ \frac{\partial\mathcal{L}}{\partial\eta} &= 0.\end{aligned}$$

Its Euler-Lagrange equation is therefore

$$\mu\partial_t^2\eta - Y\partial_x^2\eta = 0, \tag{5.33}$$

which coincides with (5.29). This is a one-dimensional wave equation with the solutions

$$\begin{aligned} \eta &= \Re\left(c_1 e^{i\frac{vt-x}{x_0}} + c_2 e^{-i\frac{vt-x}{x_0}}\right) \\ v &= \sqrt{\frac{Y}{\mu}}, \end{aligned}$$

which describe forward and backward propagating waves. Thus, fluctuations of the individual particles from their equilibrium position travel as waves along the chain. This is indeed already a simple model of propagating lattice vibrations in crystals, so-called phonons.

To study a more complex problem consider a fluid or gas. Consider now it to be slightly disturbed, on which it reacts locally with some movement in the three spatial directions. This movement field, what is perceived as sound, is given by a vector field $\vec{\eta}$, which gives at every point in space (and time) the displacement of the medium. The kinetic energy is therefore

$$T = \frac{\mu}{2}(\partial_t \vec{\eta})^2,$$

where μ is again the density. In Principle, this density would need to be also a field. Here, however, the case of only small distortions will be considered, such that the mass density can be taken to be approximately constant. Without proof, the potential energy, which is essentially sustained by the presence of the medium, and thus how the other particles oppose the movement of a particle, is given by

$$V = \gamma P(\vec{\partial}\vec{\eta})^2 - 2P\vec{\partial}\vec{\eta}$$

where P is the equilibrium pressure and γ characterizes how the medium's pressure reacts to changes of volume. Here, this only motivates why this particular case could be of practical interest. But the focus is how to treat this system, not where it comes from. This is the purview of a thermodynamic lecture.

The Lagrangian density is therefore

$$\mathcal{L} = T - V = \frac{\mu}{2}(\partial_t \vec{\eta})^2 - \gamma P(\vec{\partial}\vec{\eta})^2 + 2P\vec{\partial}\vec{\eta}.$$

To obtain the equations of motions requires the following quantities

$$\begin{aligned}\frac{\partial(\partial_t\vec{\eta})^2}{\partial\partial_t\eta_i} &= 2\partial_t\eta_i\\ \frac{\partial\vec{\partial}\vec{\eta}}{\partial\partial_i\eta_j} &= \delta_{ij}\\ \frac{\partial(\vec{\partial}\vec{\eta})^2}{\partial\partial_i\eta_j} &= 2\delta_{ij}\vec{\partial}\vec{\eta}\\ \frac{\partial\mathcal{L}}{\partial\eta_i} &= 0.\end{aligned}$$

Interestingly, this result implies that the term linear in $\vec{\partial}\vec{\eta}$ does not contribute to the movement. This is a generic statement: Terms linear in derivatives of the fields do not alter the movement. However, e. g., if the energy stored in the system should be determined, it cannot be ignored.

The equations of motion are thus

$$\mu\partial_t^2\vec{\eta} - \gamma P\vec{\partial}\left(\vec{\partial}\vec{\eta}\right) = 0. \tag{5.34}$$

This is not yet a simply solvable system. To make progress, define the scalar quantity

$$\sigma = -\vec{\partial}\vec{\eta},$$

which is the directed rate of change. As the system is essentially isotropic, it make sense that there will be only one relevant direction, and therefore a single scalar field should actually be sufficient to solve the problem. To rewrite the problem in this form, act with $\vec{\partial}$ on the equation of motion (5.34), yielding

$$\partial^2\sigma - \frac{\mu}{\gamma P}\partial_t^2\sigma = 0.$$

Just as before, (5.33), this is a wave equation, with a speed depending on the properties of the gas and its pressure. The rate of change of the movement of the gas therefore forms waves. This is as expected: Sound is an oscillatory wave.

Chapter 6

Electrostatics

Electrodynamics, and its time-independent version electrostatics, seems always to be quite different from mechanics. The primary reason for this is that its basic concept is a field, rather than a particle. In addition, the mathematics surrounding electrodynamics is quite a bit more involved than for ordinary mechanics. While the latter cannot be helped, the former is, on a conceptual level, quite less different than seeming at first sight. In fact, as will be seen, there are a lot of parallels between electromagnetism and continous mechanical systems, as introduced in section 5.10.

6.1 Coulomb's law

But, as always, the formulation of a new physical theory starts from an observation. This observation was that particles exist, which can attract each other, and do so much stronger than gravitationally, and there are even particles, which can repel each other. This is gravitationally impossible. Thus, this effect needs to happen because of a different interaction, which is electrodynamics[1]. It was found that the force obeys Coulomb's law

$$\vec{F}_C = \gamma q_1 q_2 \frac{\vec{r}_1 - \vec{r}_2}{|\vec{r}_1 - \vec{r}_2|^3}. \tag{6.1}$$

In this formula, γ is a constant, whose value depends on the system of units. In SI, this is Coulomb's unit with value $\gamma \approx 8.99 \times 10^9$ $\mathrm{Nm^2C^{-2}}$, where C is the unit Coulomb, which is essentially defined by this law. The quantities q_1 and q_2 are the electric charges, measured in Coulomb, of these particles. Like the mass, these are properties of particles, which

[1] In fact, almost everything encountered in everyday life is electrodynamics, from the fact that a chair is carrying you, to the fact that you are able to read this. Or breath. All of this boils down eventually to electromagnetic interactions.

need to be measured and cannot be deduced. Finally, $\vec{r}_1$ and $\vec{r}_2$ are the positions of the particles. Thus, the force acts along the connection of both particles, and is repulsive if the sign of the charges of both particles is different, while it is attractive, if the the charges have the same sign. The force thus obeys Newton's third law.

It is a remarkable fact that the functional form of (6.1) is the same as the gravitational force (2.4). This can be understood from the underlying theories, though at the classical level, this is an empirical finding. The second issue are the charges. They are found to be immutable features of particles, very much like masses. Thus, every particle has now two, instead of previously one, immutable feature, mass and charge. However, in contrast to mass, charge can have both signs. Classically, no relation between both exists. Also at the classical level, charge is a continous quantity, and can have any value. Just like for mass, this includes zero: There are particles, which do not interact electrically. Again, in the underlying theories, this changes.

Given this force, the behavior of the two charged particles can then be calculated using Newton's laws, i. e. the Coulomb law gives new origins of the forces, but does not change the dynamical equations for the motion of particles. However, there is a very interesting new feature. Consider three particles in a row. If the middle one has the opposite charge than the ones at the two ends, it will reduce the total force acting on the particles at the end. It is thus possible to screen electric charges, by careful arrangements even to completely cancel them. This is very different from gravity, where this is not possible.

6.2 The electric potential

Because Newton's laws remain active, it immediately follows that the Coulomb force is, just as the gravitational force, a conservative one. Thus, it originates from an electric potential,

$$\vec{F}_C = -\vec{\partial}\Phi = -\vec{\partial}\frac{\gamma q_1 q_2}{|\vec{r}|}$$

where $\vec{r} = \vec{r}_1 - \vec{r}_2$ is now the connecting vector between both charges. The electrostatic potential works in the same way as the potential of any other conservative force. Especially, it determines the electrostatic work which is done by moving a charged particle in the presence of another one. It is also a central force, and thus it has the same features as those listed in section 2.7. This also implies that the Coulomb force preserves angular momentum. Especially, this implies that for two particles of opposite sign the solutions are again described by Kepler's laws. This is not true for same-sign electric charges. The only difference is that the ratio of charge to mass will be relevant, rather than only the mass.

This idea led to the early models of atoms, as it became clear that atoms are made up of electrically charged particles. However, as it will be seen in section 9.3, it is not that simple. That led ultimately to the need of quantum physics.

6.3 The electric field

Given the structure of forces like (6.1) and (2.4), it is a possibility to define a field by setting one of the charges, with charge Q, at the origin as

$$\vec{E} = \gamma Q \frac{\vec{r}}{|\vec{r}|^3}, \tag{6.2}$$

the so-called electric field. Another charge q experiences then the Coulomb force $\vec{F}_C = q\vec{E}$.

Likewise, it is possible to introduce a gravitational field. While it would require general relativity to make full sense of the gravitational field, it will turn out that this electric field is a concept which is extremely helpful already at the classical level, as will be seen in the following.

Before continuing with it, it is useful to collect a few features. One is that the electric charge has been set fixed at the origin. It is therefore a static situation, and the field is an electrostatic field. In addition, it is defined everywhere. Thus the name field: It is not a particle, which is localizable, but a field which gives an effect at every point (and at every time, because everything is static). Moreover, the field is defined entirely in terms of a single particle. Hence, the electric field is a property of any electrically charged particle. Every electrically charged particle has such a field. In the truest sense of the meaning, every electric charge creates an electric field, which surrounds and permeates anything else in the universe.

Because of the static nature, every other charged particle will notice the electric field of the charge instantaneously. This is a feature which will need to change later in chapter 9.

It is possible to say that the electric field is created by the electric charge. At the moment, in fact, both concepts appear to be interchangeable. It is thus not clear, what is gained by introducing the concept, except for convenient writing. But it will be found later, in section 8.3, that electric fields can actually exist without the existence of charges. They have therefore an independent reality. This is a new quality, and one of the features which makes electrodynamics quite distinct from the usual classical mechanics.

Before continuing, it is worthwhile to notice a few more consequences.

One is that the electric field can be associated with an electric potential as

$$\vec{E} = -\gamma Q \vec{\partial} \frac{1}{|\vec{r}|} = -\vec{\partial}\phi. \tag{6.3}$$

Hence, the electric field is, for now, associated with such a potential. It will be seen that this becomes more involved when going to dynamical situations, and eventually the electric field is really the fundamentally important quantity, and the potential merely a convenient auxiliary quantity for certain situations. It is also important to note that the potential is only defined up to an arbitrary function g with the property that $\vec{\partial} g = 0$. Also this has a certain significance which will be addressed in section 8.6.

A further empirical fact is that the electric fields of multiple charges add, i. e., the total electric field is given by the sum of the individual fields. Of course, in this case the positions of the charges have to be taken into account, modifying the expressions. At the same time, there exists a mathematical statement

$$\vec{\partial}\vec{E} = \gamma Q \delta(\vec{r})$$

where $\delta(\vec{r})$, the so-called Dirac-δ function, is a function which is only non-zero (and actually infinite) at the position of the charge, and zero elsewhere. It has the feature that

$$\int d^3\vec{r} f(\vec{r})\delta(\vec{r}) = f(\vec{0}). \tag{6.4}$$

Thus, this allows to write the charge density $\rho(\vec{r})$ of a single particle as $\rho(\vec{r}) = Q\delta(\vec{r})$. Because of the additivity of the electric fields, the total electric field therefore obeys

$$\vec{\partial}\vec{E} = \gamma\rho(\vec{r}), \tag{6.5}$$

where the charge density is now the sum of the charge densities of all the individual charges,

$$\rho(\vec{r}) = \sum_i q_i \delta(\vec{r} - \vec{r}_i).$$

Of course, the charge density, just like a matter density, can be taken to be continous, and no longer be created from point-like objects. This can be achieved either in a limiting procedure like in section 5.10 or defined from the outset.

There is an mathematical theorem, which allows to revert expression (6.5), giving the electric field as a function of the charge density as

$$\vec{E}(\vec{r}) = \gamma \int d^3\vec{r}' \rho(\vec{r}') \frac{\vec{r} - \vec{r}'}{|\vec{r} - \vec{r}'|^3}. \tag{6.6}$$

This gives an intuitive picture as the electric field being obtained as the effect of all elements of charge everywhere in linear superposition. This can be explicitly seen by combining (6.3) and (6.4), which demonstrates this for a single point charge, as it yields (6.2).

From this insight, and a few mathematical theorems, follows Gauss' law,

$$Q_V = \gamma \int\limits_V d^3\vec{r}\rho(r) = \int\limits_{S_V} d^2\vec{r}\cdot\vec{E}, \tag{6.7}$$

which requires a bit elaboration. The middle part is a volume integral. It integrates the charge density over the volume, and thus yields the total charge inside the volume, Q_V. The right-hand side is a so-called surface integral. Given that the volume is finite, it has a surface, which is denoted by S_V. The integral on the right-hand side takes the component of the electric field normal to the surface (this is indicated by the $d^2\vec{r}\cdot\vec{E}$ notation), and integrates this component over the whole surface. Such a quantity is called the flux, in this case the flux of the electric field: The amount of field going through a surface, and in the present case through the surface out of the volume. Thus, the physical content of Gauss' law is that the total flux of the electric field going out of a volume is equal to the total charge inside the volume. This emphasizes again the characterization of the charge density as the origin of the electric field.

Note that if there are electric charges of opposite sign they can cancel each other. As a consequence, the total charge in a volume can be zero, even though the electric field inside is not zero. In fact, even the electric field can be non-zero on the boundary. But it will be directed into the volume and out of it in such a manner that the net effect is zero.

A further consequence is that the electric field has no closed field-lines. They always emerge and end on a charge or at infinity. This is mathematically stated by the fact that

$$\vec{\partial}\times\vec{E} = \begin{pmatrix}\partial_2 E_3 - \partial_3 E_2\\ \partial_3 E_1 - \partial_1 E_3\\ \partial_1 E_2 - \partial_2 E_1\end{pmatrix} = \vec{0}. \tag{6.8}$$

This so-called rotation gives the vorticity of a field. Mathematically, it can be shown that from the existence of the electric potential (6.3) this follows. In fact, inserting (6.3) into (6.8) this follows immediately, as then every line is of the type $\partial_i\partial_j\phi - \partial_j\partial_i\phi = 0$.

6.4 Matter and electric fields

The previous sections discussed the situation of isolated charges, which are otherwise in vacuum. At a microscopical level this is quite adequate, as matter is made up of

atoms which themselves separate into electrically charged nuclei and electrons. But at a mesoscopic or macroscopic level this is not very useful. Here, it makes more sense of thinking of matter as an electrically active continuum.

In its simplest form it can be assumed that homogeneous and isotropic matter changes an electric field by a constant factor, the electric permittivity ϵ, yielding the electric displacement field $\vec{D}$,

$$\vec{D} = \epsilon \vec{E} \tag{6.9}$$

Because the field is different inside and outside the matter this implies that there is an effective electric surface charge density, which yields this alteration. It is a, historically motivated, convention that the prefactor has in the vacuum $\epsilon_0 = \epsilon$ units and size $\epsilon_0 \approx 8.854 \times 10^{-12}$ C^2/(Nm^2), and actually is a fixed number $(4\pi \times 10^{-7} \times c)^{-1}$ A^2/N, with A Ampere and c the speed of light. It is called the electric permittivity of the vacuum. In other unit systems, this quantity can have no units, or be even unity, which is usually convenient.

The actual value of the permittivity ϵ is in classical physics an empirical number for each type of matter. It can be calculated, at least in principle, once the atomic structure of the matter is taken into account. Sometimes the relative permittivity $\epsilon = \epsilon_0 \epsilon_r$ is introduced, a quantity which is larger than one for most media, and equal one for the vacuum.

In general, the situation is less trivial. In principle, the matter can have its own source of electric field, a so-called polarization field $\vec{P}$. This can either arise from the reaction of the matter to the external field, or could even be a genuine field carried by the matter itself[2]. This can be taken into account as

$$\vec{D}(\vec{r}) = \epsilon_0 \vec{E}(\vec{r}) + \vec{P}(\vec{r}) = \epsilon_0 \underline{\chi}(r) \vec{E}(\vec{r}) = \underline{\epsilon}(\vec{r}) \vec{E}(\vec{r}). \tag{6.10}$$

The introduction of the polarization tensor $\underline{\chi}$ allows the capture of the effect as a reaction of the medium to the external magnetic field. This allows even for spatial inhomogeneities of the matter. The situation (6.9) is recovered if $\underline{\epsilon}(\vec{r}) = \epsilon \underline{1}$, i. e. if the reaction is homogeneous and (anti-)parallel to the field. Treatment of general matter becomes quite quickly unwieldy, and thus in the following only the simpler situation (6.9) will be considered.

The physical interpretation of the displacement field is that on a charge inside the matter only the displacement field acts, and only it generates a Coulomb force on it. Therefore, depending on whether ϵ is larger or smaller than 1, the electric force is attenuated or enhanced by the medium. Consequently, when switching between two media, this

[2]The latter is rarely encountered, as then usually strains appear which will tend to neutralize this effect.

requires the accumulation of charge on the corresponding interface to create this effect. This electric surface density σ on the boundary between two media is given by

$$(\vec{D}_1 - \vec{D}_2)\vec{n} = \sigma \tag{6.11}$$

where n is a unit vector orthogonal to the boundary. The vacuum as one medium is covered implicitly by setting for the vacuum $\epsilon = \epsilon_0$.

These consideration play important roles in many practical areas of electric matter interactions, starting from capacitors or the optical properties of matter. E. g., the way how light is reflected from a surface follows from (6.11). But this requires first other developments to fully appreciate. Nonetheless, these formulas are of central importance in electrical engineering and optics, and many other branches of applied sciences.

These observations had also another important interpretation. Even if a piece of matter is in itself not electrically charged, i. e. acted on by the Coulomb force of another charge, it can still be electrically active. Microscopically, this can be understood as follows. If matter is build up from charges in such a way that it only becomes zero in total, but not locally, an electric field immersing the matter will act on these. Indeed, atoms are build up from positive charges, the nuclei, and negative charges, the electrons. Thus, the Coulomb force will move both elementary charges in opposite direction, and only the binding force of the matter will prevent them being ripped apart. Anyhow, they are displaced, introducing a field along the line connecting them. This is a directed displacement, a so-called dipole. This will continue until an equilibirum has been reached. Thereby the electric field is modified, giving rise to the displacement field (6.9). If the forces inside the matter do have a preferential direction, which does not align with the direction of the electric field, this will modify the response field's direction relative to the original field, which gives rise to the polarization tensor in (6.10). Therefore, the fact that a displacement field arises supports the idea that neutral matter is build up (at least partly) from charged constituents, a picture confirmed by modern microscopic understanding.

Chapter 7

Magnetostatics

While electrostatics is something not that much accessible to daily experience, Magnetostatics is: Whenever handling magnets, one experiences magnetic fields. It is therefore also a much better way to illustrate the concept of fields, as the experience when trying to bring two magnets together is directly demonstrating the concept of a field. Without physically touching, both already exert forces on each other, with no visible trace. Of course, this is due to the fact that our senses are not created to sense magnetic fields. Some animals can do so, and would perceive the situation quite differently from us. On the other hand, fields which we do perceive, like electromagnetic fields in form of light, show a very different experience than those of magnetostatic fields. Still, both are conceptually, as will be seen, the same thing, irrespective of our very different experience of them. This should always be kept in mind when considering what fields are.

7.1 The Lorentz force

However, the first step will actually not directly address ordinary magnets and the interaction between magnets. The reason is that the microscopic dynamics play an essential role in how the effect comes about, and thus it somewhat involved to understand. Rather the first step in defining the magnetic field, and thereby also to define a measurement process for it, is the so-called Lorentz force.

The Lorentz force is the observation that a magnetic field $\vec{B}$, e. g. of a magnet, exerts a force on an electrically charged particle of charge q, which moves with a speed $\vec{v}$ as

$$\vec{F} = q\vec{v} \times \vec{B}, \tag{7.1}$$

which is given here in SI units. This provides a specific measurement protocol how to map out the magnetic field of, say, a magnet. Thus, a charged particle at rest will not

be affected by a magnetic field. This provides a way to separate the effect of magnetic fields and electric fields, and demonstrates that they are two different concepts. However, they are linked by the fact that the electric charge appears in (7.1). Just as with inertial mass and gravitational mass that was not clear in the beginning when the Coulomb force (6.1) and the Lorentz force (7.1) were discovered. However, postulating that the electric charge appears in both force equations is something which is necessary to formulate the theory of classical electrodynamics, the Maxwell equations, very much like the postulate of the equality of inertial mass and gravitational mass is necessary for the formulation of general relativity. Again, this fact is something experimentally established. In contrast to the gravitational case, however, an explanation will become readily available in section 8.2.

The Lorentz force is quite different from both the gravitational force (2.4) and the Coulomb force (6.1). The structure of (7.1) makes explicit that it is not a central force, like those two. Note that Newton's third law still applies. It is also dependent on speed, and thus appears to be not a conservative force. In the context of classical mechanics, this is indeed not true, because an infinitesimal change of mechanical energy is related to work, (2.10), by

$$dE = \frac{dE}{dt}dt = \frac{dW}{dt}dt = \frac{\vec{F}\cdot d\vec{r}}{dt}dt = \vec{v}\cdot\vec{F}dt \tag{7.2}$$

and thus vanishes for the Lorentz force (7.1). Thus, the Lorentz force is conservative. However, once electrodynamics is fully developed, it will also become clear that energy can be stored in electric and magnetic fields, leading to energy conservation in a more comprehensive way, as long as both particles and fields are taken into account. In a similar manner, also angular momentum will remain an important concept.

The structure of (7.1) makes also manifest that the Lorentz force cannot explain the behavior of two magnets, since they act on each other even when both are at rest. This requires to take the internal structure into account. In fact, there is no reference at all to magnets, just to electric charges, though the existence of magnets seems to suggest there exist genuine magnetic objects. This quagmire will be resolved once the internal structure of magnets becomes clarified. This will show that electric charges moving inside the magnet will be the source of the acting forces. However, the explanation of how exactly this works requires quantum physics, and can thus only be modeled, but not explained, in classical physics. This is also the reason why the Lorentz force is not yet explained in terms of its origin, as the Coulomb force was explained as the electric field of an electric charge in section 6.3.

7.2 Ampere's law

Given that the Lorentz force (7.1) interacts with moving charges, and the possibility to define a charge density in section 6.3, it is natural as a next step to ask what happens in the presence of an electric current,

$$\vec{j}(\vec{r}) = \rho(\vec{r})\vec{v}(\vec{r}),$$

where ρ is the local charge density and $\vec{v}$ is the speed (and thus direction) into which the charges move at that point[1].

While such a current will certainly experience a Lorentz force when subjected to a magnetic field, there is a more important empirical observation: Such a current actually also creates a magnetic field in such a way that the created field satisfies the equation

$$\vec{\partial} \times \vec{B} = \begin{pmatrix} \partial_2 B_3 - \partial_3 B_2 \\ \partial_3 B_1 - \partial_1 B_3 \\ \partial_1 B_2 - \partial_2 B_1 \end{pmatrix} = \mu_0 \vec{j}, \tag{7.3}$$

where $\mu_0 = 4\pi \times 10^{-7}$ N/A^2 is the magnetic permittivity of the vacuum. This is known as Ampere's law. In a way not dissimilar to the electric field in section 6.3 this explains the existence of a magnetic field as a consequence of moving charges. Thus, a magnetic field appears if there are moving charges[2], but then it becomes a property of the charges in the same way as the electric field is a property of electric charges.

Very much in the sense of (6.7) it follows from a mathematical theorem that

$$\int\limits_{C_S} d\vec{r} \cdot \vec{B} = \mu_0 \int\limits_{S} d^2\vec{r} \cdot \vec{j} = \mu I. \tag{7.4}$$

This means that the total current I going through an area S, which is bounded by the curve C_S, is also given by the integral of the component of the magnetic field parallel to S. The total flux of a magnetic field is determined by the total piercing current. Hence, electric and magnetic fields on boundaries are measures for what happens inside the bounded region.

Because of the different structure of the equation (7.3) in contrast to (6.5), there are also some differences. Probably the most important one is that

$$\vec{\partial} \cdot \vec{j} = \sum_i \partial_i j_i = 0, \tag{7.5}$$

[1]This chapter is still on a static situation, and thus such a current is maintained indefinitely, and does not change in time.

[2]What this means in term of the equivalence principle will be explored in chapter 9.

because

$$\vec{\partial} \cdot \vec{\partial} \times \vec{B} = \sum_i \epsilon_{ijk} \partial_i \partial_j B_k = 0.$$

While this result will change once time-dependence is introduced, this is already an important insight. Physically this implies that there are no sources of the electric current due to the magnetic field.

Based on this insight the next natural question is, whether there are magnetic charges, which would yield an equation like (6.5)? At the level of classical physics, it becomes an empirical fact that this is not the case, and thus the equation

$$\vec{\partial} \cdot \vec{B} = 0 \tag{7.6}$$

holds[3]. Magnetic fields are only generated by moving charges.

This has also another important consequence: There is no place a magnetic field line can end. Thus, magnetic field lines exist only as closed lines around currents, which is essentially also the statement of (7.4). Thus, in the static case they are inextricably linked to the existence of currents. This will also change in dynamic situations, giving them an existence of their own. What will not change, however, is that they cannot end somewhere.

7.3 Matter and magnetic fields

In a very similar way as in section 6.4 it is possible to investigate the interplay between magnetic active matter and the magnetic field. By introducing a magnetic permittivity tensor $\underline{\mu}(r)$ in the same way as the electric permittivity tensor $\underline{\epsilon}(\vec{r})$ any reaction of a medium on a magnetic field can be described. Moreover, this also gives how this electric field will now act on a current inside matter, by replacing the field $\vec{B}$ by $\vec{H}(\vec{r}) = \underline{\mu}(\vec{r})\vec{B}(\vec{r})$, which is sometimes called the magnetic field strength, while $\vec{B}$ is called in contrast the magnetic flux strength, due to the relation (7.4). Linear magnetic matter has then $\underline{\mu}(\vec{r}) = \mu\underline{1}$ with a fixed and constant permittivity μ, or relative permittivity μ_r defined by $\mu = \mu_0\mu_r$.

The model which describes how this appears is that the electron of the atoms move around in orbits around the nuclei. As these are moving charges, they create a magnetic field, which is perpendicular to the orbit. If the orbits of the electrons for different atoms are randomly arranged, the total current will vanish when averaged over many atoms, and

[3]This empirical fact is an important ingredient in the construction of the Maxwell equations in chapter 8. It can be shown that you cannot have a theory like the Maxwell equations and at the same time both electric and magnetic charges at the quantum level, though this is beyond the scope of this lecture. A discovery of magnetic charges would therefore fundamentally change the way we understand electrodynamics, and would not be just merely an addition of those charges.

thus the total magnetic field will be zero. When an external magnetic field acts on matter its Lorentz force will tend to align these orbits, and thus create an effective current inside the matter which will yield the alteration of the magnetic field field from $\vec{B}$ to $\vec{H}$.

In such a picture magnets can now be understood. In magnets the electron orbits are aligned without an external field[4]. The magnet can then act on other matter by the Lorentz force of its magnetic field. Because magnetic fields appear as closed lines, because the electrons move in a particular direction around their orbit, magnets exhibit a north pole and a south pole - you cannot just reverse the direction of electron movement. As a consequence, two magnets can attract or repel each other, depending on the relative orientation of the electronic currents and thus magnetic fields and Lorentz forces.

While this picture is appealing, it is unfortunately not correct. In fact, it is not even consistent in classical physics, as will be seen in section 9.3. It requires quantum physics to fully understand how magnetic fields in matter originate, and become stable. This will also explain why some materials can become magnets, some not, and some not even react to magnetic fields at all. However, at the macroscopic, and even mesoscopic, level the idea that some materials just have a magnetic field given by a permittivity tensor is usually sufficient for most cases, especially in applications.

[4]In fact, this usually means that they remain aligned after the magnetic field is removed. This happens usually if the magnetic material was liquid while in the magnetic field of, e. g., the earth during cool down, imprinting the magnetic orientation in the solid. Not all materials can do this, explaining why magnets can only be made from some.

Chapter 8

Electrodynamics

The fact that electric charges played a role in both electrostatics and Magnetostatics is an ominous hint towards a common origin of both. This hinges critically on the appearance of the same charges in Coulomb's law and both Lorentz' law and Ampere's law. This ominous connection will be further strengthened by Faraday's law, which will establish a connection between time-dependent electric fields and magnetic fields.

These various pieces will finally come together in the form of Maxwell's equation, which give a unified picture of electric and magnetic phenomena, which connect to matter via the single property of electric charges. This unification of effects in a single origin is still today a role model in theoretical physics of finding a theory. It has become the driving force, even obsession, of modern theoretical physics to find a similar unification for all of physics.

However, the guiding input in the formulation of Maxwell's equation were indeed the experimentally established universality of the electric charge as well as the independent laws, but especially the connection due to be made using Faraday's law.

8.1 Faraday's law and time-dependence

Because electric charges react to electric and magnetic forces, they will usually change their behavior. Thus, in general the charge density will become time-dependent, $\rho(\vec{r}, t)$, and as a consequence so will the current, $\vec{j}(\vec{r}, t)$. Assuming that electric charges cannot be created or destroyed in the same sense as mass cannot be created and destroyed[1], this has far-reaching consequences.

Given some finite volume V the change of the total charge in it is linked to the current

[1]Both are empirical facts.

through its boundary S_V. This is the only way in which charge could appear or disappear from the volume. Thus

$$\partial_t Q = \int\limits_V d^3\vec{r}\partial_t\rho = \int\limits_{S_V} d^2\vec{r}\cdot\vec{j} \tag{8.1}$$

Taking the volume to be infinitesimally small, it can be shown that this is equivalent to

$$\partial_t\rho = \vec{\partial}\cdot\vec{j}, \tag{8.2}$$

which is known as the continuity equation. This shows that Ampere's law (7.3) can no longer hold in a dynamic situation, as this result contradicts its consequence (7.5).

To understand how it needs to change can be derived from combining (8.1) with (6.7),

$$\partial_t Q = \partial_t \int\limits_{S_V} d^2\vec{r}\cdot\vec{E} = \int\limits_{S_V} d^2\vec{r}\cdot\partial_t\vec{E}.$$

Thus, a change in total charge is accompanied by a change in the electric field. This is not surprising, as the electric field is generated by the electric charge. As a consequence, Ampere's law in a dynamical situation needs to read

$$\vec{\partial}\times\vec{B} = \mu_0\vec{j} + \mu_0\epsilon_0\partial_t\vec{E} = \mu_0\vec{j} + \frac{1}{c^2}\partial_t\vec{E} \tag{8.3}$$

It is not trivial that this is the only change, and no new terms appear besides this minimal addition. The appearance of the speed of light, which emerges from the numerical values of the product $\mu_0\epsilon_0$, is also non-trivial, and will play an important role later one. This was originally an empirical observation. Its significance was only fully appreciated once special relativity had emerged.

This implies that a temporal change of an electric field implies a change of the magnetic field. If the current is localized, this holds still true away from the changing current. Because the electric and magnetic fields are defined everywhere, their changes will influence each other.

This leads immediately to the expectation that a time-dependent magnetic field should somehow yield also to a change of the electric field. Because of the absence of magnetic charges, there can be no similar effect from the current as in (8.3), and hence it can be expected that the effect is entirely given by the time-dependence of the magnetic field. This is indeed experimentally established to be the case, and given by Faraday's law,

$$\vec{\partial}\times\vec{E} = -\partial_t\vec{B}. \tag{8.4}$$

That it takes this form is not trivial, but empirical.

8.2 Maxwell's equations

It is empirical as well that the changes (8.3) and (8.4) together with a trivial addition of time-dependence to (6.5) and (7.6) is sufficient to capture the full time-dependence of all electromagnetic phenomena.

These form together Maxwell's equations,

$$\vec{\partial} \cdot \vec{B} = 0 \tag{8.5}$$

$$\vec{\partial} \times \vec{E} + \partial_t \vec{B} = 0 \tag{8.6}$$

$$\vec{\partial} \cdot \vec{E} = \frac{1}{\epsilon_0}\rho \tag{8.7}$$

$$\vec{\partial} \times \vec{B} - \frac{1}{c^2}\partial_t \vec{E} = \mu_0 \vec{j}. \tag{8.8}$$

The somewhat arbitrary placement of the constants ϵ_0, μ_0, and c is an artifact of the SI system.

This version of the Maxwell equation is usually called 'in the vacuum', though this term is misleading. What it is actually is that all charges are given in terms of ρ and $\vec{j}$, rather than made into effective quantities. Proper vacuum is the situation where $\rho = 0$ and $\vec{j} = 0$, i. e. in absence of charges. The meaning of this situation will be discussed in more detail in section 8.3.

It is also possible to give Maxwell's equation in matter, by introducing $\vec{D}$ and $\vec{H}$, as given in sections 6.4 and 7.3. As long as the matter is linear, this is achieved by direct replacements of the fields with suitable factors of ϵ and μ. The charge density and current then correspond to additional charged matter beyond the active matter. Such charges are sometimes called free charges to separated them from those bound in the active matter. However, if the active matter is not linear, i. e. there are tensorial or even space(-time) dependent permittivites, it is important that they are added inside the derivatives, and product rules may apply. However, these effective Maxwell's equation will not be used hereafter.

There is interesting feature that equations (8.5-8.6) do not involve charges, and only (8.7-8.8) do. There are deep reasons for this, which will be seen once special relativity is made explicit in chapter 9. Then the equations (8.5-8.6) will turn out to have a structural (geometric) reason, and thus encode no physical, but merely mathematical, information, while the physics (dynamics) is contained in the other two equations.

Of course, the system of equations is not yet closed. It describes fully what happens in the presence of an external electric charge and current to the electric and magnetic fields. Thus, they need to be maintained by external influences. To also include the back reaction of the electric and magnetic fields on the charges themselves it is necessary to solve the

corresponding equations of motions for them. In principle, it appears that, as discussed previously, using Newton's equation of motion (2.1) together with the Coulomb force and the Lorentz force should do the trick. That is, however, not entirely true. These forces arose in the static situations. To include time dependence will actually require to treat also the matter relativistically. Thus, this is postponed to chapter 9 as well. Only if the particles move at small speeds, and the electric and magnetic field vary slowly enough that the situation is almost static compared to the speed of the particle, the approach using Newton's equation is sufficiently accurate. Fortunately, many situations are of this type, such that this approach is suitable for many common cases.

8.3 Electromagnetic waves in the vacuum

There is an interesting feature of Maxwell's equation (8.5-8.8): Even if there is no matter, $\rho = 0$ and $\vec{j} = \vec{0}$, the Maxwell's (proper) vacuum equations

$$\begin{aligned} \vec{\partial} \cdot \vec{B} &= 0 && (8.9)\\ \vec{\partial} \times \vec{E} + \partial_t \vec{B} &= \vec{0} && (8.10)\\ \vec{\partial} \cdot \vec{E} &= 0 && (8.11)\\ \vec{\partial} \times \vec{B} - \frac{1}{c^2}\partial_t \vec{E} &= \vec{0} && (8.12) \end{aligned}$$

are differential equations. Just like a particle without a force acting on it still exists and moves, this suggests that electric and magnetic fields could still exist and vary without charges, provided there a solutions to (8.9-8.12) with non-vanishing fields.

This is indeed the case. To see how this works, derive (8.12) once more with respect to time, and combine it with (8.10), yielding

$$\frac{1}{c^2}\partial_t^2 \vec{E} = \vec{\partial} \times \partial_t \vec{B} = -\vec{\partial} \times \vec{\partial} \times \vec{E}. \tag{8.13}$$

This is a non-trivial equation for the electric field alone. A similar equation for the magnetic field is obtained by deriving (8.10) first and then using (8.12). However, in both cases there remains a connection between the electric and magnetic field, giving rise to the notion of the electromagnetic field. In addition, the equations (8.9) and (8.11) remain. They act as constraint equations, i. e. they select out of the possible solutions a subset.

Now, the electric field has three components. This implies that (8.13) are actually three (coupled) differential equations. There are actually equations for coupled harmonic oscillators, as writing them down and comparing with section 5.10.1 shows. Thus, the solutions need to be oscillating.

A suitable ansatz is

$$\vec{E} = \Re\vec{a}_{\pm} \exp(\pm i(\vec{k}\vec{x} \pm ck_0 t))$$

Inserting this into (8.13) yields

$$k_0^2 \vec{E} = -\vec{k} \times \vec{k} \times \vec{E}.$$

Using that the system is rotational invariant, it is useful to select $\vec{k} = k_3 \vec{e}_z$. Then, this yields that $k_0^2 = k_3^2$ and that $\vec{a}$ needs to be perpendicular to $\vec{k}$, but is otherwise free. Because of rotational invariance, this has to hold always, and thus the solution needs to satisfy

$$\begin{aligned} k_0^2 &= \vec{k}^2 \\ \vec{a}_{\pm} &= \vec{n} \times \vec{k}, \end{aligned} \tag{8.14}$$

where $\vec{n}$ is an arbitrary vector. The structure of (8.14) also yields that (8.11) is automatically fulfilled, as the derivative will create a structure $(\vec{n} \times \vec{k}) \cdot \vec{k}$, which identically vanishes.

Thus, the full solution for the electric field to the vacuum Maxwell equations is

$$\vec{E} = \Re(\vec{n} \times \vec{k}) \exp(\pm i(\vec{k}\vec{x} \pm c|\vec{k}|t)).$$

There are quite some interesting observations to be made here. The first is that these are plane waves, which propagate parallel or antiparallel to $\vec{k}$. They do so at the speed of light, and the wave vector $\vec{k}$ determines the frequency of oscillations ω by $\omega = c|\vec{k}|$. Hence, electric waves travel in vacuum at the speed of light. In addition, the electric field is polarized orthogonal to the direction of propagation, it is a purely transverse wave. This is quite distinct from the acoustic waves of section 5.10.4. Choosing the vector $\vec{n}$ to be complex, this adds a phase shift, and therefore also allows the direction of the electric wave to change during propagation, allowing for circular and elliptic polarization. Finally, because the Maxwell equations are linear, any sum of such waves is also a solution. Thus, by superposition of waves also different wave shapes than plane waves can be constructed, very much like in any other wave equation. Thus, the initial conditions determine the wave shape entirely.

The magnetic field can now be determined using either of equations (8.10) and (8.12). This yields that $\vec{B} = \vec{k} \times \vec{E}/(ck_0)$, and the magnetic field is thus uniquely determined by the electric field (or vice versa), and it is always orthogonal to the electric field and the direction of propagation. There is also a relative factor of c, making the magnetic field in these units much smaller than the electric field. This is, however, specific to this

system of units. Also, this implies that electric waves and magnetic waves are inseparable, and accompany each other. Thus, the object should be considered as a single entity, an electromagnetic waves. The origin of this will be found in chapter 9.

The existence of electromagnetic waves independent of the existence of charges implies the independent reality of electromagnetic fields, freeing them from being a property originating with (charged) matter. They thus form a new category of physical entities apart from matter, and actually the prototype of a new class of physical objects, fields. In contrast to a point particle, from which all matter is build up in classical mechanics, which is located to a single point in space at any given time, the electromagnetic field is an entity which fills out the whole of space at any given instance of time. There is also only one electromagnetic field. Even though technically it can be written as a sum of, e. g., the fields of multiple origins, the final physical object, which makes itself felt by the Coulomb force and the Lorentz force, is the one electromagnetic field which emerges as the sum. hence, this is also a new category in physics compared to point particles: All phenomena of electromagnetism are tied to a single entity, the electromagnetic field, rather than to many independent entities, like all of point particles in mechanics. Again, this is a prototype for essentially all of physics at the fundamental level.

Testing the implications of the electric and magnetic forces exerted by these fields it is possible to identify visible light, as well as X-rays, infrared radiation, radio waves etc., as electromagnetic waves of different frequencies or wave-vector values. Thus, (generalized) light is indeed identified with this new physical entity. And then it also becomes much more directly accessible to experience.

8.4 Electromagnetic phenomena and matter

Electromagnetic phenomena in electromagnetically linearly active media can, as noted in section 8.2, be well described by using $\vec{H}$ and $\vec{D}$, and appropriate factors of μ and ϵ, while non-linearly active media are more involved.

More interesting, however, are situations which involve boundaries. The first interesting case is, when electromagnetic wave enter a not transparent medium. This can be modeled by introducing complex values of ϵ and/or μ. Solving wave-equations inside the medium then implies that $\vec{k}$ becomes complex. As a consequence, the electromagnetic wave inside the medium, which is the real part of the complex solution, acquires an exponential damping term. Thus, light entering such a medium will very quickly diminish, essentially on a length inverse proportional to the imaginary part of the permittivites. The distance for one e-fold of attenuation is usually called penetration depth. This also

explains why a very thin slab still allows light to go through.

This also allows to understanding the breaking of light at surfaces, if the light does not go into the medium perpendicular. Because the electric field inside the medium is different, the electric (and likewise the magnetic) field cannot be continous across the surface. However, using the Maxwell equations, especially (8.6), it follows that the component of the electric field parallel to the surface needs to be continous. Using the same line of reasoning the perpendicular component of the magnetic field is continous, while the parallel is not. The situation for the fields $\vec{D}$ and $\vec{H}$ are reversed.

Given the special case of two linear media, and an electromagnetic field with wave vector hitting the flat separation surface at an angle α, this is sufficient to deduce that some part of the wave must be reflected, and the reflected plane wave has a wave vector enclosing with the separation surface the angle β, satisfying

$$\sqrt{\frac{\mu_1\epsilon_1}{\mu_2\epsilon_2}} = \frac{\sin\beta}{\sin\alpha},$$

the refraction law of Snellius. Often the combination $n = \sqrt{\mu\epsilon}$ is called the refraction index.

Such kinds of considerations concerning boundaries are useful in many problems. They originate from the fact that the Maxwell equations are linear in the fields and the integral version (6.7), as this implies that any distribution of charges determines alone already the electric field on the boundary. A very common application is the so-called method of image charges to determine the electric field. It applies to cases where there is a grounded object, i. e. one on which no surface charges can form, and hence the potential on the surface needs to vanish. Given the charge density outside ρ_o it determines an electric potential by an integral version of (6.3), essentially by deriving (6.6), as

$$\phi_o(\vec{r}) = \frac{1}{4\pi\epsilon_0}\int d^3\vec{r}'\frac{\rho_o}{|\vec{r}-\vec{r}'|}.$$

On the surface of a grounded object the potential vanishes by definition. This effects can be modeled by finding an imaginary charge distribution ρ_i, such that its potential satisfies

$$(\phi_o(\vec{r}) + \phi_i(\vec{r}))\,|_{\text{on the grounded surface}} = 0.$$

The corresponding electric field can then be derived by (6.3) from $\phi_o + \phi_i$ alone, as the surface of the grounded sphere can be considered to be the boundary of the outside. This invests that the electric field of a charge distribution, which exists only in a finite space volume, falls off to zero at infinity. Hence, the electric field outside needs to be given uniquely by the combination of both potentials.

This demonstrates again a unique feature of the electromagnetic field, highlighting the importance of surfaces and boundaries in dealing with them.

8.5 Dipole radiation and antennae

Once it becomes clear that electromagnetic radiation is equal to light, radio waves etc. it becomes very interesting to understand what it means that there are sources radiating an electromagnetic wave. Or, vice versa, sources receiving an electromagnetic wave as an antennae.

A source for electromagnetic radiation needs to satisfy two conditions. The first is that it is time-dependent. Otherwise, the situation is static, and it becomes no longer sensible to talk about emission and reception. The second is that it is localized. The simplest such construction, which describes, however, many practically important features, is a so-called dipole. This is a charge, which oscillates between two points along a line. This yields also a current, and thus there is no electric radiation without having simultaneously also magnetic radiation.

Of course, movement along a line is not necessary. There are many other possibilities. They are describe by so-called multipoles. They can be ordered by how the source changes under rotation. The simplest is a monopole, a point charge, which is invariant under rotations. In a sense, it pulsates[2]. The dipole is the next symmetric one, having cylindrical symmetry. After that, the next one is a quadrupole, which has even less symmetry, and so on. This can be systematically treated in terms of a multipole description. This is, however, technically too involved to do here.

Assume for the moment an antennae, which is very small, and oriented along the z-direction, which has an oscillatory current with frequency ω, and a maximum current of p_0, defining $\vec{p} = p_0 \vec{e}_z$, located at the origin. The electromagnetic fields, which are obtained from solving Maxwell's equation in a manner not dissimilar from the vacuum case, are then given by

$$\begin{aligned}
\vec{E} &= \frac{1}{4\pi\epsilon_0} \Re \left(\left(\frac{\omega^2 (\vec{r} \times \vec{p}) \vec{r}}{c^2 r^3} + \left(\frac{1}{r^5} - \frac{i\omega}{c r^4} \right) \left(3\vec{r} (\vec{r} \cdot \vec{p}) - r^2 \vec{p} \right) \right) e^{\frac{i\omega r}{c} - i\omega t} \right) \\
\vec{B} &= \frac{\omega^2 \vec{r} \times \vec{p}}{4\pi\epsilon_0 c^3 r} \Re \left(\left(1 + \frac{ic}{\omega r} \right) \frac{e^{\frac{i\omega r}{c} - i\omega t}}{r} \right).
\end{aligned}$$

These are highly involved quantities. The main reason for this is the detailed effects of the changes of directions of the source. This can be somewhat disentangled by looking at the far-field, i. e. at distances $\omega r/c \gg 1$ where many such changes already accumulate. Doing

[2]This would violate charge conservation, and can thus be only thought of as something where an external reservoir provides and removes charge in such a way as to not generate quantitatively relevant currents.

a Taylor expansion of the exact near-field expressions yields the far-field results

$$\begin{aligned}\vec{B} &= \frac{\omega^2}{4\pi\epsilon_0 c^3}\frac{\vec{r}\times\vec{p}}{r^2}\Re e^{i\omega\left(\frac{r}{c}-t\right)} \\ \vec{E} &= c\frac{\vec{B}\times\vec{r}}{r}.\end{aligned}$$

Thus, far away from the source, the emitted electromagnetic waves are orthogonal to each other, just as in the vacuum. Both are also orthogonal to the direction of the source, and propagate radially outwards. Thus, at these distances the antennae looks like a point source, though with a direction attached. Finally, the electromagnetic field diminishes linearly with the distance travelled.

However, the latter gives a slightly distorting view of what happens. For this it is useful to discuss the energy stored in electromagnetic fields. Because electromagnetic fields are fields, it makes often no sense to talk about the whole energy stored in them. After all, this means how much energy is stored in the whole of space. More useful is the energy density e, i. e. the energy per unit volume. It is given by

$$e = \frac{1}{2}\left(\epsilon_0\vec{E}^2 + \frac{1}{\mu_0^2}\vec{B}^2\right) \tag{8.15}$$

This can be determined based on the potential to do work, just as potential energy is determined, which in turn can be derived from the corresponding force laws. It can be obtained directly from the Maxwell equation, but this leads too far here.

Just like for charge and current in equation (8.2), it is possible to show that there exists a kind of continuity equation for the energy density in an electromagnetic situation. This will have two components. One is the energy carried away by the moving charges, and one from the electromagnetic fields. This yields the energy continuity equation

$$\partial_t e = -\vec{\partial}\cdot\vec{S} - \vec{j}\cdot\vec{E} \tag{8.16}$$

$$\vec{S} = \frac{1}{\mu_0}\vec{E}\times\vec{B}, \tag{8.17}$$

where $\vec{S}$ is called the Poynting vector. It is the energy current carried by the electromagnetic field. The existence of such an energy current, and energy density, carried by the electromagnetic field appears necessary, as light is able to incite a reaction, which requires energy. This again emphasizes the independent physical reality of the electromagnetic field. Finally, note that for plane waves the energy current is in the direction of the wave vector $\vec{k}$, and thus energy is carried in the direction of the wave. Note that the term $\vec{j}\cdot\vec{E}$ is the work done by the electric field on electric charges, as derived from (2.10) and the

Coulomb force in terms of the electric field. There is no magnetic term, as the magnetic field does not provide work, as was discussed in section 7.1.

Coming back to the case of a source, the Poynting vector averaged over one period in the far-field is given by

$$\overline{\vec{S}} = \frac{\mu_0 p_0^2 \omega^4}{32\pi^2 c} \frac{\sin^2\theta}{r^3} \vec{r}. \tag{8.18}$$

It points radially outward from the source, but is modulated with respect to the orientation of the dipole. Especially no energy is radiated in the direction of the dipole. The fact that it decays like $1/r^2$ is a necessity from energy conservation. As a energy density current, it provides the amount of energy radiated per unit area and time. The total energy of a surface surrounding the source grows like r^2. With the Poynting vector decaying like $1/r^2$ this yields that the total energy radiated is independent of distance, but becomes spread out more and more the farther away from the source the recipient is. This explains, e. g., why a light source becomes dimmer the further away an observer is from it: Because the area of the eye remains constant, the total energy received diminishes, and so does the perceived brightness of the source. A detector surrounding the whole surface would not see the effect.

This is also true generically for a receiver. The electromagnetic field will act on charges at the receiver with a force, leading to a motion of them. Because this is an oscillatory phenomena, this is very similar to the situation of a driven harmonic oscillator from section 2.6. Just as in this case there is an optimal receiver, which is in resonance with the incoming wave's frequency, yielding the strongest effect. Because the electric charges actually have to be able to move a certain distance, this implies that for any given frequency there is a (minimum) length of the antennae to obtain a signal at a given minimum level. Thus, the physical forms of antennae is an important technical issue. Also, because of the additional angular dependence in (8.18) this implies that the orientation relative to the source matters. All of these aspects have to be taken duly into account in technical situations. They also play an important role in observation, especially in astronomy.

As noted above, different forms will yield different behavior. E. g., monopoles are found to not radiate any energy. Thus, they will not emit detectable radiation. More complicated shapes will yield more involved forms of the electromagnetic waves and of the radiated energy. This can be both hindrance and advantage in technical applications as well as many interesting astronomical problems. Unfortunately, this becomes quickly very technical, and will thus be skipped here.

8.6 Electrodynamics as a gauge theory

There is one more feature, which still appears somewhat mysterious. Why are the electric field and magnetic field so tightly linked? Is there a common origin? The answer to this is affirmative, and it will be again a prototype for more fundamental theories of physics.

Maxwell's equations (8.5-8.8) have the very interesting property that they only connect derivatives of the fields[3]. This allows actually to write the six components of the electric and magnetic field in terms of only four functions, a vector $\vec{A}$ and a single function A_0. As will be seen the latter is intimately connected to the electric potential from section 6.2. However, in the next chapter 9 a much more interesting connection will arise, and thus this connection will not be too strongly emphasized. The vector $\vec{A}$ is often called the vector potential. In the relativistic case, the same name will denote the four-vector build from A_0 and $\vec{A}$.

The connection is the following. Maxwell's equation can be fulfilled, if the electric field and the magnetic field are given by

$$\vec{B} = \vec{\partial} \times \vec{A} \tag{8.19}$$

$$\vec{E} = -\vec{\partial} A_0 - \partial_t \vec{A}. \tag{8.20}$$

The second equation shows that in the static case A_0 is, up to a factor of c, the electric potential, by comparison to (6.3).

That this is indeed a solution to Maxwell's equation can be seen from the fact that (8.5) is trivially fulfilled by (8.19), as any curl of a vector is automatically divergence-free. Likewise with the same argument

$$\vec{\partial} \times \vec{E} = -\vec{\partial} \times \partial_t \vec{A} = -\partial_t \vec{B},$$

and thus (8.6), follows. Thus the homogeneous Maxwell equations are satisfied by construction.

The inhomogeneous ones reduce to

$$\vec{\partial}^2 A_0 + \partial_t \vec{\partial} \cdot \vec{A} = -\frac{\rho}{\epsilon_0} \tag{8.21}$$

$$\left(\vec{\partial}^2 \vec{A} - \frac{1}{c^2}\partial_t^2 \vec{A}\right) - \vec{\partial} \cdot \left(\vec{\partial} \cdot \vec{A} + \frac{1}{c^2}\partial_t A_0\right) = -\mu_0 \vec{j}. \tag{8.22}$$

These are now four equations for four unknowns. Thus, the surplus constraint equations of the original Maxwell equations, after all 8 equations for 6 functions, have been resolved in

[3]The linearity is also an interesting property, but it is found that (specific) non-linearities do not impede the following, though make it more complicated. Such non-linearities actually appear commonly in particle physics.

this way. The questions remains, of course, if for all possible charge densities and currents the solutions of (8.21-8.22) yield the correct electric and magnetic fields.

It turns out that they do. This follows from the fact that the construction explicitly, and identically, resolves the constraint equations of Maxwell's equations, but retain the other degrees of freedom. However, it turns out that for any value of the electric field and magnetic field there is an infinite number of possibilities for both $\vec{A}$ and A_0. This can be seen in the following way. Changing

$$\vec{A} \to \vec{A}' = \vec{A} + \vec{\partial} \cdot \Lambda \tag{8.23}$$

$$A_0 \to A_0' = A_0 + \partial_t \Lambda, \tag{8.24}$$

where Λ is an arbitrary function, will not change either the electric field nor the magnetic one. This follows for the magnetic field because the divergence of a curl is zero, and for the electric field because in (8.20) after exchange of both derivatives the contributions from both terms cancel. As (8.21-8.22) are just Maxwell's equations (8.7-8.8) with (8.19-8.20) inserted, neither will they. Thus, for any set of electric field and magnetic field there is an infinite number of possibilities for $\vec{A}$ and A_0.

Having not a unique solution is nothing fundamentally new. It started with the arbitrary constant which could be added to the mechanical potential in section 2.5. The new aspect is that it is not a constant, but a function, which can be added, without changing anything observable. After all, the observation of electromagnetic fields is from the Coulomb force and the Lorentz force, and they do not change, as they only involve the electric field and the magnetic field.

This has two consequences. One is that this freedom can be used to introduce additional constraints on $\vec{A}$ and A_0, usually such as to simplify calculations. Such constraints are called gauge conditions. These may fix this freedom completely, or only partly. One choice is, e. g., Coulomb gauge, which imposes

$$\vec{\partial} \cdot \vec{A} = 0,$$

which reduces (8.21) to an equation which makes explicit the connection of A_0 and the electric potential. Another one is the Lorentz condition, which imposes

$$\vec{\partial} \cdot \vec{A} + \frac{1}{c^2} \partial_t A_0 = 0,$$

and which thus simplifies (8.22) to an inhomogeneous wave equation with the electric current as a source. Thus, both simplify the solution of Maxwell's equation considerable. In practice, it will depend on the specific form of ρ and $\vec{j}$ which is more advantegous.

Once the solution is obtained, (8.19-8.20) can be used to obtain the electric field and the magnetic field.

It is not obvious that these are valid gauge conditions. To show that, it is necessary that, e. g. for Coulomb gauge,

$$\vec{\partial} \cdot (\vec{A} + \vec{\partial}\Lambda) = 0 \rightarrow \vec{\partial}^2\Lambda = -\vec{\partial} \cdot \vec{A}$$

has a solution to Λ for any $\vec{A}$. That this is the case follows from the theory of partial differential equations. The same is true for the Lorenz gauge condition. Both gauges are not removing all freedom for the vector potential. A transformation depending only on time, $\Lambda(t)$, is still allowed for the Coulomb gauge. For the Lorentz gauge a transformation where Λ satisfies a sourceless wave equation is still possible. That is not a problem, and admissible. The remaining freedom can either be fixed, or initial conditions can be used to select a unique solution.

The other consequence is that this implies that the four degrees of freedom are still too much, and electric and magnetic fields actually have even less. In fact, fully fixing all remaining freedom shows that only two out of the four components are independent. Thus, all six components of the electromagnetic field would be given by only two functions. How this works out can be obtained in the following way. One is that the electric current and the charge density are not independent, but linked by the continuity equation (8.2). This removes one degree of freedom. The other requires special relativity, and is more involved. However, there is a hand-waving explanation for the vacuum case of section 8.3. As was observed there, the electric field and magnetic field were perpendicular to each other and to the direction of propagation. Thus, the electromagnetic field had only the possibility to arrange itself transverse to the direction of propagation, which are exactly two degrees of freedom. The same effect is seen in the radiation of an antennae in section 8.5, underlining that this is not only a feature of the vacuum electromagnetic wave.

Chapter 9

Unification of electrodynamics, special relativity, and mechanics

With Maxwell's equations in place, it is now finally possible to unify classical mechanics, electrodynamics, and, as a necessary glue, special relativity. This gives a unified framework of all of classical physics. Other aspects, e. g. optics, hydrodynamics, and thermodynamics, can then be derived completely out of this underlying description. However, in practice this may be too formidable - just imagine to solve the $\sim 3\times 10^{23}$ equations of motion to describe a mol of gas. Thus, it is often useful to derive effective theories from this underlying single theory. E. g., optics is derived from electrodynamics in the limit of distances much larger than the wave-length of the electromagnetic wave. Thermodynamics is obtained from Newton's equation in the limit of infinitely many particles. And so on. However, it is important to know that all of these myriad of effective theories derive from a single one, the one to be determined now.

But, just as all known theories today, even such a powerful theory as this one carries already the seed of its own failure, both from its mathematical formulation as well as the comparison to experiment. This will be discussed in section 9.3 to motivate the necessity for quantum physics. This will yield another, more fundamental theory, of which the present one can be derived in the limit of sufficiently long distances. Still, the following remains a prototype of a unified description of physics. Just finding one without internal failure or discrepancy to experiment has not been found (yet?).

9.1 Electrodynamics as a relativistic theory

The first step is to understand the vacuum case without electric charges. It is at first sight not obvious how to proceed. After all, special relativity dictates the use of 4-component

objects. Electric and magnetic fields, however, have three components each. There is no obvious fourth component to be added to either, as was the case with time, energy, or power for position, momentum, and force in mechanics.

The reason is involved. To find a solution, first note that there is a fundamental difference between electric and magnetic fields in the vacuum. When performing a parity transformation, i. e. mirroring all vectors, the electric and magnetic fields cannot behave in the same way. This follows from equations (8.10) and (8.12). Because the spatial derivative changes sign under this operation, but not the time derivative, either the electric or the magnetic field have to change sign. Which of them can be deduced from (7.3). This yields that the electric field behaves like a vector, but the magnetic field does not. It is a so-called pseudo-vector. This also implies that something strange is going on, and it is not just finding a fourth component.

The answer to this is that actually neither is extended to a four-vector. What eventually happens is that they form a tensor rather than a vector, the so-called field-strength tensor

$$F_{\mu\nu} = \begin{pmatrix} 0 & E_1/c & E_2/c & E_3/c \\ -E_1/c & 0 & -B_3 & B_2 \\ -E_2/c & B_3 & 0 & -B_1 \\ -E_3/c & -B_2 & B_1 & 0 \end{pmatrix},$$

which is antisymmetric $F_{\mu\nu} = -F_{\nu\mu}$. The factor of c is an artifact of the different units of the electric and magnetic fields in SI units. Such tensors also transform in a prescribed way under Lorentz transformations, which will not be detailed here.

Maxwell's equation can now be written using the field-strength tensor in a relativistic form,

$$\partial^\mu F_{\mu\nu} = 0 \tag{9.1}$$

$$\partial^\mu \epsilon_{\mu\nu\rho\sigma} F^{\rho\sigma} = 0, \tag{9.2}$$

where $\epsilon_{\mu\nu\rho\sigma}$ is the generalization of the usual Levi-Civita tensor to four dimensions. The second equation corresponds to the homogeneous Maxwell equation (8.9-8.10) and the first one to the inhomogeneous ones (8.11-8.12), which can be seen by explicitly writing them out. Note that these are in total eight equations, just as Maxwell's equation.

As a consequence, the change of the electric and magnetic field under a Lorentz transformation is now given entirely in terms of the transformation properties of the field-strength

tensor. Given a speed in x direction, the change is

$$\vec{E}' = \begin{pmatrix} E_x \\ \gamma(E_y - c\beta B_z) \\ \gamma(E_z + c\beta B_y) \end{pmatrix} \tag{9.3}$$

$$\vec{B}' = \begin{pmatrix} B_x \\ \gamma\left(B_y + \frac{\beta}{c} E_z\right) \\ \gamma\left(B_z - \frac{\beta}{c} E_y\right) \end{pmatrix}. \tag{9.4}$$

This is very different from the usual transformation of position and momentum. Especially, the component parallel to the speed is unchanged. Also, the boost mixes electric and magnetic field. In fact, even if one system there has been only either an electric field or a magnetic field, in a boosted system there are both. This underlines that electric field and magnetic field are really just two aspects of the same entity.

This also clarifies the question of what happens when a current is viewed from a frame in which its charges are not moving. Indeed, in this frame there is no magnetic field. But it did not vanish. It was only shifted into the electric one. Which, of course remains, as the charges are still there. How this precisely works is now given explicitly by (9.3-9.4).

The structure becomes even more transparent when rewriting the electric field and magnetic field in terms of the vector potential of section 8.6. The quantities there form now a four vector $(A_0, \vec{A})$, with the zero-component behaving as a time component, and the whole vector potential transforming under a Lorentz transformation like a momentum vector. In terms of these the field-strength tensor is given by

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\mu A_\nu, \tag{9.5}$$

as can be verified by explicitly inserting (8.19-8.20). Then the transformations (9.3-9.4) are no longer so surprising, as the vector potential component in direction of the speed does not appear in the expression for the electric and magnetic fields. The transformation properties then also yield the explicit form, given that both the derivative and the vector potential needs to transform.

Note that the field-strength tensor is gauge-invariant by structure, as the relativistic form of (8.23-8.24) is

$$A_\mu \to A'_\mu = A_\mu + \partial_\mu \Lambda,$$

and then the antisymmetry of the two terms in (9.5) remove the Λ-term.

Moreover, the homogeneous Maxwell equation (9.2) reads now

$$\partial^\mu \epsilon_{\mu\nu\rho\sigma}(\partial^\rho A^\sigma - \partial^\sigma A^\rho) = \epsilon_{\mu\nu\rho\sigma}(\partial^\mu \partial^\rho A^\sigma - \partial^\mu \partial^\sigma A^\rho) = 0.$$

Because $\partial^\alpha \partial^\beta$ is symmetric in the indices, but is contracted with an antisymmetric quantity in $\alpha\beta$, this is identical zero, no matter the value of A_μ. Thus, the homogeneous Maxwell equations are trivial fulfilled, and are so because of the gauge nature and the fact that the electromagnetic field is really from an antisymmetric Lorentz tensor. It is thus a purely geometric feature of the theory.

Of course, this leaves open why the electromagnetic fields are described by such an object. This is unknown so far. But also the appearance of c for the speed of the electromagnetic waves becomes now no longer surprising, as it is exactly this value which is needed to make the inhomogeneous Maxwell equation (9.1) relativistically covariant. The need to be consistent with special relativity thus implies that electromagnetic waves travel at the speed of light, and thus the maximum speed of light, through vacuum.

Interestingly, historically this had been known before the advent of special relativity, as had been the Lorentz transformation. At that time, it was considered to be an oddity of Maxwell's equation and electromagnetism. Special relativity then elevated this oddity to be a feature of space-time itself.

9.2 A unified, relativistic theory of mechanics and electrodynamics

It is now necessary to add matter. Given the form of the inhomogeneous Maxwell equation (9.1), it is necessary that the right hand side is a four-vector. Thus, the charge density and the current need to be combined into one. This is indeed straightforwardly possible as $j = (c\rho, \vec{j})$, which transforms as an ordinary four-vector, like the coordinate. The inhomogeneous Maxwell equation then becomes

$$\partial_\mu F^{\mu\nu} = \mu_0 j^\nu,$$

which can be seen by explicit comparison to (8.7-8.8), and the constants are again an artifact of the SI system. However, this still leaves open how the electric charges are affected by the electromagnetic forces.

It is best to consider for this the relativistic version of Newton's law (4.6). In a fixed reference frame, the force will be the usual Coulomb force and Lorentz force. They can be expressed in terms of the electric and magnetic field. In a frame where this only an electric field this becomes most straightforward

$$\frac{dp_\mu}{d\tau} \overset{\beta_i \ll 1}{=} \begin{pmatrix} q\vec{\beta}\vec{E} \\ q\vec{E} \end{pmatrix}$$

while for a magnetic field

$$\frac{dp_\mu}{d\tau} \overset{\beta_i \ll 1}{=} \begin{pmatrix} 0 \\ \frac{q}{mc}\vec{p} \times \vec{B} \end{pmatrix},$$

because the Lorentz force has no projection into the direction of the movement of the particle. Comparison to the transformation laws for the electric and magnetic fields and the structure of the field-strength tensor yields then

$$\frac{dp^\mu}{d\tau} = \frac{q}{mc} F^{\mu\nu} p_\nu.$$

The corresponding current is then obtained by integrating this equation to get the position of the particle as a function of eigentime, $X(\tau)$. This yields $j = (cq\delta(x - X(\tau)), c\vec{p}\delta(x - X(\tau))/m)$. Thus, this closes the theory completely, and hence describes fully the interaction of a particle and the electromagnetic fields. Matter can then be build from many such particles. Note, however, that this is a highly non-trivially coupled set of differential equations. But it is, in principle solvable.

This finally unifies electromagnetism and mechanics. In fact, it is possible to write this down also in terms of a single Lagrange function, when utilizing again the vector potential,

$$L = -mc^2\sqrt{1 - \frac{m^2p^2}{c^2}} - \frac{q}{c}p_\mu A^\mu - \frac{1}{4}\int d^3x F_{\mu\nu}F^{\mu\nu}.$$

That the second term is a density follows from the fact the electromagnetic fields are defined everywhere in space. Using the techniques from section 5.10.3, it can be shown that this yields the correct equations of motions for all particles, though this requires in detail quite some work. This is even more true when intending to access the Hamilton function, because of its not being relativistically invariant. Thus, this will need to suffice for the purpose of hand, demonstrating that such a unification is indeed possible.

9.3 Radiation of an accelerated charge and the limit of validity of classical physics

As was already indicated in section 8.5, in the presence of a current an electromagnetic field is emitted, and energy carried away.

Consider now a point charge, with a given trajectory $r(\tau)$, and corresponding four-speed $u(\tau) = (c, \vec{\beta})$. Define implicitly an eigentime τ_0 by $|\vec{x} - \vec{r}(\tau_0)| = R$, where $\vec{x}$ will be the point where the electromagnetic field is evaluated. This is the eigentime necessary for something at the speed of light to propagate to a point $\vec{x}$ a distance R away from the position of the particle. This involved construction is necessary, as changes in the

electromagnetic field can only propagate at the speed of light, but the field is defined everywhere. Thus, to understand the field of the point charge, it is necessary to know where the point charge was at a previous time[1]. Solving the corresponding Maxwell equation for such a particle, ignoring the back-reaction of any electromagnetic forces on the particle, yields

$$\begin{aligned}\vec{E}(t,\vec{x}) &= q\left(\frac{\vec{n}-\vec{\beta}(\tau_0)}{\gamma^2(1-\vec{\beta}(\tau_0)\cdot\vec{n})^3R^2}+\frac{\vec{n}\times\left(\left(\vec{n}-\vec{\beta}(\tau_0)\right)\times\partial_\tau\vec{\beta}|_{\tau=\tau_0}\right)}{c(1-\vec{\beta}(\tau_0)\cdot\vec{n})^3R}\right)\\ \vec{B}(t,\vec{x}) &= \vec{n}\times\vec{E}\\ \vec{n} &= \frac{\vec{x}-\vec{r}(\tau_0)}{R}.\end{aligned}$$

This is a very involved expression. There are, however, a few points of importance. The first is that again electric field and magnetic field are orthogonal to each other. The second is that both fields decay as $1/R$, similar to the case of the antenna. At large distances the first term of the electric field becomes irrelevant as it decays like R^2. Then both the electric field and the magnetic field are directed transverse to the direction $\vec{n}$, which points radially away from the point where the particle was. The corresponding Poynting vector (8.17) is thus also directed radially outwards, and there is thus a flux of energy into this direction. This energy flux decays again as $1/R^2$, and thus the total energy radiated integrated over a surface at a given instant in time remains constant[2]. However, the radiated energy is only non-zero, if $\partial_\tau\beta$, i. e. the acceleration of the particle is non-zero[3]. Thus, an accelerated charge radiates off energy.

So far this seems not to be a problem. After all, somehow energy is pumped into the system by accelerating the charge, and some of it is lost in form of radiated electromagnetic waves. In fact, such radiation is very useful, as the example of the antennae shows. There, it is actually this feature which makes transmission of electromagnetic waves possible, and thus a lot of technologies.

But now comes the point where this leads to a conflict with observation. As noted in section 7.3 the existence of magnets indicate the presence of currents inside matter. These currents need now to either push all electric charges outside of any finite piece of matter, which is not observed, or needs to be directed back into the matter eventually. Which

[1]This very involved approach is an artifact of using point charges. More fundamental theories, in which all of matter also become fields, are much less cumbersome.

[2]Here it is glimpsed over a few subtleties, but the message remains the same when doing all details carefully.

[3]At zero acceleration the first term seems to remain, but it has a different direction, and does not yield a net energy flow through a surface surrounding a particle.

would require an acceleration of the charges inside the matter. The same is true for the idea of electrons orbiting a nucleus, as keeping them on the orbit requires acceleration. In all such cases electromagnetic radiation would need to appear. This is not observed. Also, even if it would appear, there would be no source for the required energy. Hence, these observation imply that there needs to be something more than the present unified special relativistic theory of electrodynamics and mechanics of section 9.2. This is indeed the case, and the problem is resolved by quantum mechanics.

Besides this purely experimental evidence for the need to go beyond the classical unified theory there are also other, purely theoretical reasons for this to be necessary. Most notably is the problem of point charges. A point charge has an electric field. Within the electric field energy is stored, as dictated by (8.15). This density behaves like $\vec{E}^2$, but this means it diverges like $1/\vec{r}^4$, due to the electric field of a point charge (6.2). Trying to integrate this to obtain the total energy stored in this static field yields a divergence - an infinity of energy would be stored in the field of a point charge at rest already. Hence, the theory becomes unstable when assuming the existence of point charges. This problem is resolved by a better description of matter in terms of quantum fields. However, this kind of divergency is a general signal for a theory not to be able to describe some regime, in this case very short distances. Thus, electrodynamics is again a prototype for a more general trait. The appearance of such divergencies are one of the key features of modern theories which guarantee that there is yet more to be discovered.