2

Multiscale Computational Materials Science

One might wonder why one does not derive all physical behavior of matter from an as small as possible set of fundamental equations, e.g. the Dirac equation of relativistic quantum theory. However, the quest for the fundamental principles of physics is not yet finished; thus, the appropriate starting point for such a strategy still remains unclear. But even if we knew all fundamental laws of nature, there is another reason, why this strategy does not work for ultimately predicting the behavior of matter on any length scale, and this reason is the growing complexity of fundamental theories – based on the dynamics of particles – when they are applied to systems of macroscopic (or even microscopic) dimensions.

The idea that matter is made of small particles, called atoms[1], goes back to the ideas of Leucippus and Democritus (460–370 B.C.) in classical Greek philosophy of the fifth century B.C., see e.g. [68, 69], and has been very successful in the development of modern concepts in physics. Introduced as a working hypothesis in chemistry by John Dalton[2] (1766–1847) at the beginning 19th century for explaining the stoichiometry in chemical reactions, the reality of atoms was finally generally accepted among the scientific community roughly 100 years later due to overwhelming experimental evidence and theoretical achievements, e.g. Boltzmann's kinetic gas theory which is based on the pre-condition that atoms exist, or Einstein's famous 1905 paper [72] in which he developed a statistical theory of the Brownian motion which allowed to calculate the size of molecules and atoms.

Despite these theoretical achievements, there were famous opponents such as Ernst Mach (1838–1916) who was captured in his philosophy of positivism, only accepting empirical data as basic elements of a physical theory. As atoms at that time could not be directly observed, he attributed them to the realm

[1] From Greek "άτομο" (indivisible).

[2] The original publication is [70]. For a review, originally published in 1856 and available today as unabridged facsimile, see [71].

of metaphysical nonsense[3]. Albert Einstein (1879-1955) later wrote about his 1905 paper [74]:

> "My principal aim was to find facts that would guarantee as much as possible the existence of atoms of definite size. [...] The agreement of these considerations with Planck's determination of the true molecular size from the laws of radiation (for high temperatures) convinced the skeptics, who were quite numerous at that time (Ostwald, Mach), of the reality of atoms." (Albert Einstein, 1946, p. 45)

In fact, modern experiments with the largest microscopes available, that is particle accelerators of high-energy physics, revealed that even the constituents of atoms, neutrons and protons themselves show an inner structure and are composed of yet even smaller particles, so-called *quarks*[4]. The idea of the existence of ever smaller particles of matter seems to have come to an end with quarks, due to *quark confinement*, a property of quarks which renders it impossible to observe them as isolated particles[5].

At the beginning 20th century it was realized that the classical (pre-quantum) laws of physics could not be valid for the description of systems at the atomic scale and below. Rutherford's scattering experiments with H_2-particles hitting a gold foil in 1911 [78] had shown that atoms could not be elementary. The eventual development of the formulation of a quantum theory during the years 1925–1927 also changed the "scientific paradigm" of what is to be considered "understanding", and also of what a physical theory is expected to accomplish. In a passage of his book "Der Teil und das Ganze"[79], Werner Heisenberg (1901–1976) writes in Chap. 5 of a discussion with Albert Einstein in which Einstein claims that the idea to base a theory only on *observable* elements, i.e. on elements or objects which are measurable and perceptible in experiments, is nonsense. It is interesting to note that Einstein used this "philosophy" himself as a heuristic concept for deriving the special theory of relativity, eliminating such unobservable, metaphysical concepts like "absolute space", "absolute time", and the idea of an "ether", an ominous substance which – in 19th century physics – was supposed to define

[3] In contrast to Ernst Mach – roughly half a century later – Richard Feynman starts the first chapter of his famous lecture series on physics [73] with the remark that the atomic hypothesis, i.e. the idea that matter is made of single small particles, contains the most information on the world with the least number of words.

[4] This peculiar naming of the smallest known constituents of matter after the sentence "Three quarks for Master Mark" that appears in James Joyce's novel "Finnegans Wake", goes back to Murray Gell-Mann (Nobel price 1969).

[5] This property of the strong interaction was discovered by D. Gross, D. Politzer, and F. Wilczek (Nobel prize 2004) and is due to an increase of the strong coupling constant (and along with it an increase of the strong force) with increasing distance of the quarks. That is, if one tries to separate quarks, energy is "pumped" into the force field until – according to $E = mc^2$ – quark-antiquark systems come into being. The original publications are [75, 76, 77].

an inertial system (IS) and thus, an absolute frame of reference in space. According to Einstein – as Heisenberg writes – it is *theory* that determines what is measurable in experiments and not vice versa.

In this context one may ask questions such as: "What actually *is* a "theory" and what are the characteristics of a theory?" "What does "modeling" actually mean?" "What *is* a model?" "Is a model "reality"?" "What is "reality" in the natural sciences?" "What is the difference between a model and a fundamental law of nature?" And dealing with computational physics one could consequently ask the question as to what degree the result of a computer program can be considered "reality". Albert Einstein (1879–1954) writes about the relevance of such epistemological questions in the natural sciences in a 1916 memorial lecture for Ernst Mach [80]:

> "How does it happen that a properly endowed natural scientist comes to concern himself with epistemology? Is there no more valuable work in his specialty? I hear many of my colleagues saying, and I sense it from many more, that they feel this way. I cannot share this sentiment. When I think about the ablest students whom I have encountered in my teaching, that is, those who distinguish themselves by their independence of judgment and not merely their quick-wittedness, I can affirm that they had a vigorous interest in epistemology. [...] Concepts that have proven useful in ordering things easily achieve such an authority over us that we forget their earthly origins and accept them as unalterable givens. Thus they come to be stamped as "necessities of thought", "a priori givens", etc. The path of scientific advance is often made impassable for a long time through such errors. For that reason, it is by no means an idle game if we become practiced in analyzing the long commonplace concepts and exhibiting those circumstances upon which their justification and usefulness depend, how they have grown up, individually, out of the givens of experience. By this means, their all-too-great authority will be broken." (Albert Einstein, 1916, pp. 101–102)

Obviously, when applying or deriving physical theories and models of reality, there are certain implicit and usually unspoken assumptions being made. In the following sections it is tried to provide a few suggestions as answers to the above questions and to discuss materials science within this context of model building. Then, in Sect. 2.5, the degree to which physical laws have been unified in modern physics is shortly discussed. Finally, in Sect. 2.6 some fundamentals of computer science and the notions of "algorithm" and "computability" and their eventual formalization in the concept of a *universal Turing machine* are discussed.

2.1 Some Terminology

In science, one seeks after *systems* in which all different notions, concepts, and theoretical constructs are combined into one consistent *theory* which explains the diversity of physical phenomena with very few, general *principles*. Usually nothing can be changed in this system, or else it fails. This fact is the hallmark of a physical theory. A prominent example is the *(strong) principle of equivalence*, which states that the inertial mass m_I (a measure of a body's resistance against acceleration), the passive gravitational mass m_P (a measure of the reaction of a body to a gravitational field) and the active gravitational mass m_A (a measure of an object's source strength for producing a gravitational field) are the same. Thus, one can simply refer to the *mass* of a body, where $m = m_I = m_P = m_A$. If this principle was found to be wrong in any experiment in the future, the whole theory of general relativity – which is a field theory of gravity – would completely break down, because it is based fundamentally on this principle. However, in Newton's theory of gravitation, this principle just appears as a theorem which states an empirical observation – another coincidence. Nothing follows from it, and nothing would change in Newton's theory if it was not valid.

In mathematics, such principles, which form the theoretical basis of all developments, are called *axioms*. Together with a set of rules of inference and theorems, derived according to these rules, they build a theory. In fact, the degree to which a theory or model has been "axiomatized" in the natural sciences can be used as a criterion, as to whether a theory is considered to be as "closed". Examples of *closed theories*, i.e. model systems in which all hitherto known experimental facts can at least in principle be explained and derived from a handful of axioms, are all *classical* theories in physics, based on Newtonian mechanics, such as mechanics, thermodynamics, electrodynamics and also the special theory of relativity. These theories are thought to be on excellent ground in both evidence and reasoning, but each of them is still "just a theory". Theories can never be proved and are subject to tests. They can only be falsified. They are subject to change when new evidence comes in.

In the previous considerations we have repeatedly used the words "system", "theory", and "model". These terms are generally more or less mixed up and used interchangeably in the context of model building and there exists no strict, commonly accepted definition of these terms; however, "theory" usually is used in science as a more general term than "model" in the sense that a theory can combine different models (e.g. particles and waves) within one theory (e.g. quantum field theory). Hence, the term "model" generally refers to a lower level of abstraction, whereas a complete theory may combine several models.

In this volume no particular distinction between "model" and "theory" is made and the two terms are used interchangeably. Finally, a "system" is a theoretical construct which includes all different hierarchical basic axioms,

notions, models, and theories which pertain to some specific phenomenon in a way, which renders it impossible to change anything within the system without making it fail. If, despite numerous possibilities to be falsified, a certain system or theory does not lead to any contradiction to known experimental facts, then it is called a *law of nature*.

2.2 What is Computational Material Science on Multiscales?

Strictly speaking, materials science of condensed matter today is focused on the properties of condensed matter systems on a length scale comprising roughly 10 to 12 orders of magnitude, ranging from roughly 10 Å to a few hundred meters for the largest buildings or constructions. The two extremes, physics at the smallest scales below the atomic dimensions ($\leq$1 Å) or under extreme conditions, e.g. Bose-Einstein condensates, or the properties of neutrons and protons, as well as large-scale structures such as stars or galaxies and galaxy clusters, which are millions of light years in extend, are not an object of study in materials science or engineering. Nevertheless, it is physics, that provides the basic theories and modeling strategies for the description of matter on all scales.

The reason why there is no single, perfect and all-comprising model for calculating material properties on all scales relevant for materials science, is, that nature exhibits complex structural hierarchies which occur in chemistry and engineering devices as well as in biological systems (self-organization of matter), investigated in the life sciences. Remarkably, on the Ångstrømscale there are only atoms but then on larger length scales these basic constituents build complex hierarchical structures in biology and chemistry [81], which are treated with different theories that have a certain range of applicability, see Fig. 2.1. These different theories have to prove successful in comparison with experiments.

Materials of industrial importance such as glasses, ceramics, metals or (bio)polymers, today are increasingly regarded as hierarchical systems, cf. Fig. 2.2 and there has been a focus of research on the investigation of the different structures of components, of classes of materials on various structural hierarchies, and their combination within "process chains". Hence, the typical structural or architectural features of materials on different scales have to be taken into account. Usually one of two possible strategies is pursued. In a *bottom-up* approach the many degrees of freedom on smaller scales are averaged out to obtain input parameters relevant on larger scales. In contrast, a *top-down* approach tries to establish the structure on smaller scales starting from a coarse-grained level of resolution.

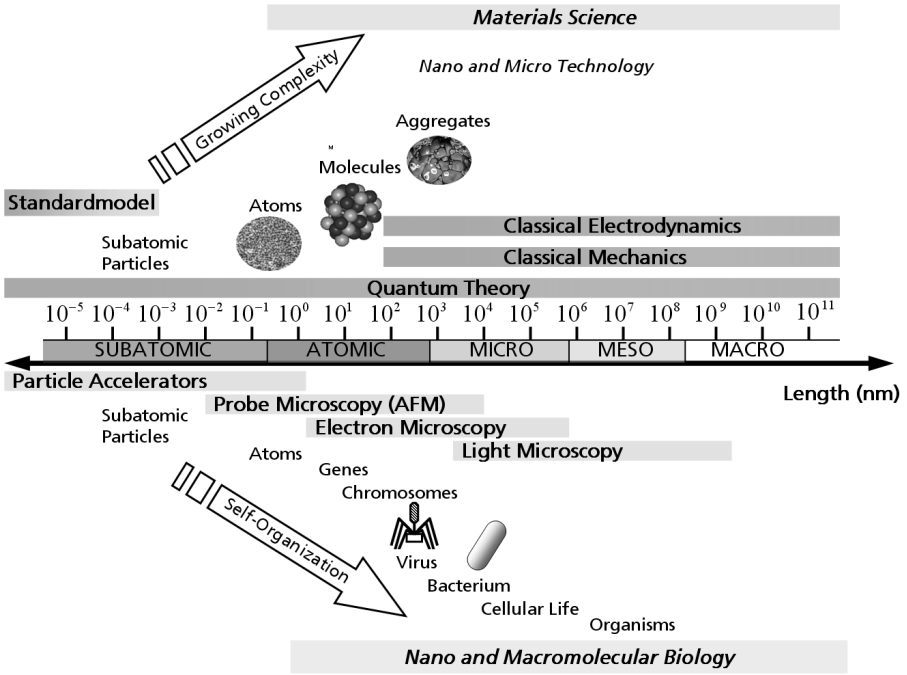


Fig. 2.1. Scope of application of different fundamental physical theories in life sciences (*bottom*) and in the areas of materials science and technology (*top*). For subatomic particles, the *standard model* (cf. Sect. 2.5) is the accepted theory that explains all hitherto observed elementary particles in accelerator experiments. Today, it is generally believed, that *quantum theory* is the most fundamental theory which is in principal valid for the description of material behavior on all length scales, cf. the discussion in the introduction of Chap. 5. However, due to the numerical complexity of many particle systems treated on the basis of the Dirac or Schrödinger equation, *classical mechanics* (instead of quantum mechanics) and *classical electrodynamics* (instead of quantum electrodynamics – the simplest prototype of a quantum field theory – where the electromagnetic field is quantized itself) are useful approximative theories on length scales larger than $\sim$10 Å. The classical theories however are not valid for quantum systems of atomic or subatomic dimensions, for phenomena occurring at speeds comparable to that of light (special relativistic mechanics) and they also fail for the description of large scale structures in the universe in strong gravitational fields (here the general theory of relativity is needed). The typical scopes of important experimental research methods using microscopes are also displayed to scale

2.2.1 Experimental Investigations on Different Length Scales

For large scale structures in the universe, experimental data are collected with telescopes, scanning a broad range of the electromagnetic spectrum. For small structures of fluids and solids in materials science, scattering techniques,

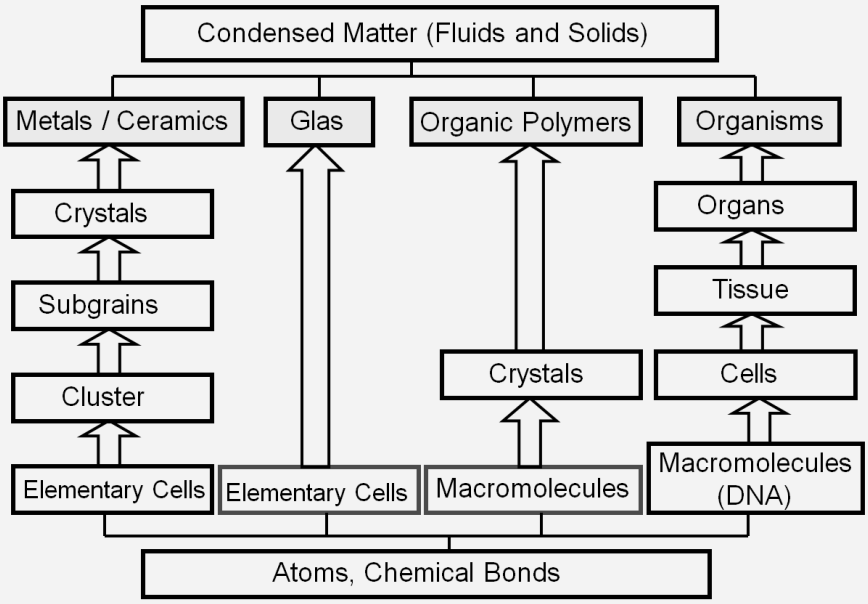


Fig. 2.2. Hierarchical view of structural properties of important classes of materials

such as Brillouin, neutron, Raman or electron scattering as well as different microscopy techniques – as depicted in Fig. 2.1 – are used on different length scales[6]. The minimum separation d that can be resolved by any kind of a microscope is given by the following formula:

$$d = \lambda/(2n\, sin\lambda)\,, \tag{2.1}$$

where n is the refractive index[7] and λ is the wavelength. The *resolution* of a microscope is the finest detail that can be distinguished in an image and is quite different from its *magnification*. For example, a photograph can be enlarged indefinitely using more powerful lenses, but the image will blur together and be unreadable. Therefore, increasing the magnification will not improve resolution. Since resolution and d are inversely proportional, and (2.1) suggests that the way to improve the resolution of a microscope is to use shorter wavelengths and media with larger indices of refraction.

The electron microscope exploits these principles by using the short *de Broglie wavelength* of accelerated electrons to form high-resolution images. The de Broglie wavelength of electrons (e^-) is given by

$$\lambda = \frac{2\pi}{k} = \frac{2\pi\hbar}{p} = \frac{h}{\sqrt{2m_{e^-}E}} \approx \sqrt{\frac{150}{V[Volt]}}\,[\text{Å} = 10^{-10}m]\,, \tag{2.2}$$

[6] Also compare Fig. 7.23 on p. 366.
[7] The refractive index $n = 1$ in the vacuum of an electron microscope.

where k is the wave vector, p is the momentum and energy E is given in electronvolts, that is the acceleration voltage of the electrons.

An electron microscope is an instrument that uses electrons instead of light for the imaging of objects. In 1926, Hans Busch in Germany discovered that magnetic fields could act as lenses by causing electron beams to converge to a focus. A few years later, Max Knoll and Ernst Ruska made the first modern prototype of an electron microscope [82].

There are two types of electron microscopes: the Transmission (TEM) and the Scanning (SEM) (or Scanning Tunneling (STM)) Electron Microscope. In a TEM, a monochromatic beam of electrons is accelerated through a potential of 40 to 100 kilovolts (kV) and passed through a strong magnetic field that acts as a lens. With a TEM one can look at structures in solids and replicas of dead cells after fixation and sputtering with heavy metal, e.g. gold atoms. With this technique, electrons are reflected off the surface of the specimen. The resolution of a modern TEM is about 0.2 nm. This is the typical separation between two atoms in a solid. This resolution is 1,000 times greater than a light microscope and about 500,000 times greater than that of a human eye. The SEM is similar to the TEM except for the fact that it causes an electron beam to scan rapidly over the surface of the sample and yields an image of the topography of the surface. The resolution of a SEM is about 10 nm. The resolution is limited by the width of the exciting electron beam and by the interaction volume of electrons in a solid. As an example, in Fig. 2.3 an SEM picture of the granular surface structure and the fracture surface of Al_2O_3 are shown in different resolutions as displayed in the figure.

Atomic Force Microscopy (AFM) is a form of scanning probe microscopy where a small probe is scanned across the sample to obtain information about

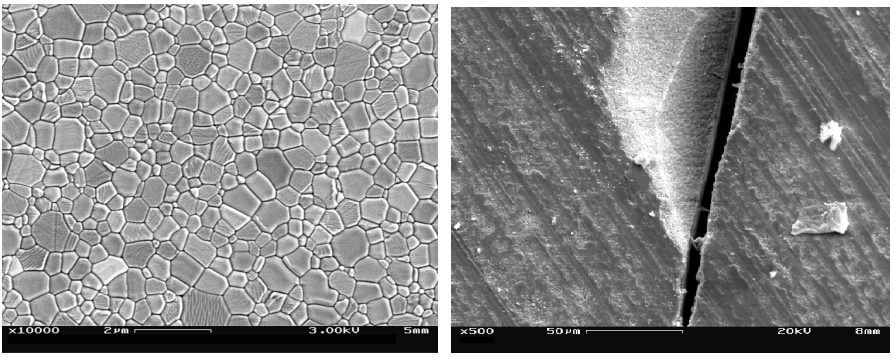


Fig. 2.3. (**a**) SEM micrograph section of an etched Al_2O_3 ceramic surface exhibiting the granular structure on the microscale. (**b**) Fracture surface of Al_2O_3 after an edge-on impact experiment (discussed in Chap. 7) at a speed of $\approx 400m/s$. Photos courtesy Fraunhofer EMI

the sample's surface. The information gathered from the probe's interaction with the surface can be as simple as physical topography or as diverse as the material's physical, magnetic, or chemical properties. These data are collected as the probe is raster-scanned across the sample to form a map of the measured property relative to the X-Y position. Thus, a microscopic image showing the variation in the measured property, e.g. height or magnetic domains, is obtained for the area imaged.

Today, electron microscopy is widely used in physics, chemistry, biology, material science, metallurgy and many other technological fields. It has been an integral part in the understanding of the complexities of cellular structure, the fine structure of metals and crystalline materials as well as numerous other areas of the "microscopic world".

2.3 What is a Model?

The complexity of the world is obviously too large in order to be comprehended as a whole by limited human intellect. Thus, in science, a "trick" is used in order to still be able to derive some simple and basic laws and to develop a mental "picture" of the world. This trick consists in the *isolation* of a system from its surroundings in the first place, i.e. one restricts the investigated system to well-defined, controllable and reproducible conditions. These conditions are called *initial conditions*. After this preparation of a system, one performs experiments and investigates which states the system is able to attain in the course of time. The underlying assumption in this procedure is, that there are certain laws which determine in principle the temporal development of a system, once it has been prepared in an initial state and left to itself.

Usually, model building is lead by the conviction that there exists an "objective reality" around us, i.e. a reality which is independent of the individual observer who performs experiments and makes observations[8].

The idea that there are fundamental laws of nature goes back to Greek philosophy, but until Galileo Galilei (1564–1642), there were only very few experimentally verifiable consequences of this idea. To identify fundamental laws, the considered system has to be isolated from its particular surrounding. Take as an example the ballistic, parabolic curve of a kicked football on Earth. The fact that an object which is thrown away on Earth follows a parabolic path *is* a law of nature; however, it is not a *fundamental* law. One realizes the fundamental law when abstracting from the Earth's atmosphere and then, in a next step, completely abstracting from the special viewpoint on Earth. When throwing away a ball in empty space, far enough away from any gravitating sources, there is no gravity which will force it on a ballistic curve[9], that is,

[8] Solipsism in this context is not a scientific category, as it renders all rational discussions useless and impossible.

[9] Of course, this is only true, if the object has a velocity component parallel to the surface of Earth.

it will simple follow a straight line (a *geodesic*, to be exact). This finding is due to Galilei and is consequently called Galilei's law of inertia. Hence, the fact, that objects thrown on Earth describe ballistic curves (i.e. not straight lines) is only due to the particular circumstances, the special point of view of the observer on Earth, or physically spoken, his frame of reference, which is not the simplest possible one. In the simplest possible frame of reference – a freely falling reference frame – which is "isolated" from both, air resistance and gravity, the ball will follow a simpler path, that is a straight line. Systems in which moving objects that are not subject of any external forces – e.g. due to gravity or electromagnetic fields – follow a straight line, are called *inertial systems*. Mathematically spoken, the inertial systems build an *equivalence class* of an infinite number of systems. The inertial systems are also those frames of reference in which Newton's mechanics and the special theory of relativity are valid.

One remarkable thing about fundamental laws of physics is, that they are *deterministic* and *time-reversible*[10], or time-symmetric (symplectic), i.e. in the example of the flying ball above, each point on the ballistic curve together with its velocity can be viewed as initial state of the system which then – due to the laws of nature – completely determines the future behavior in a *classical sense*. From a fundamental point of view, it is thus amazing that processes in nature seem to occur only in a certain direction such that they can be distinguished in "past" and "future" events[11]. One could say, *because* the fundamental laws of nature do not distinguish any direction of time, in closed systems, eventually all processes die out which are not time reversible, i.e. the system approaches its thermal *equilibrium state*. In equilibrium, no process occurs any more, except remaining time-reversible fluctuations about the equilibrium state which have equal probability. This tendency of the laws of nature are being exploited for example in the computational methods of Molecular Dynamics and Monte Carlo simulations, discussed in Chap. 6.

2.3.1 The Scientific Method

The roots of the scientific method as an interplay between theory (models or systems) and experiment, practiced in the natural sciences today, lie in the philosophy of Plato (427–347) and Aristotle (384–322). An important fundamental idea of Plato is to consider all observable things only as incomplete pictures or reflections of an ideal mathematical world of ideal forms or ideas which he illustrated in his famous *Allegory of the Cave* at the beginning of book 7 of *Republic*, see e.g. [83]. Aristotle however radically discarded Plato's

[10] Time-reversibility (or time-symmetry) is actually broken in certain rare elementary particle physics processes which was shown in 1964 by J.L. Cronin and V.L. Fitch (Nobel Prize 1980) at CERN.

[11] For example, it has never been observed, that the pieces of a broken cup cool off and repair themselves, although this process is not forbidden by any fundamental law of nature.

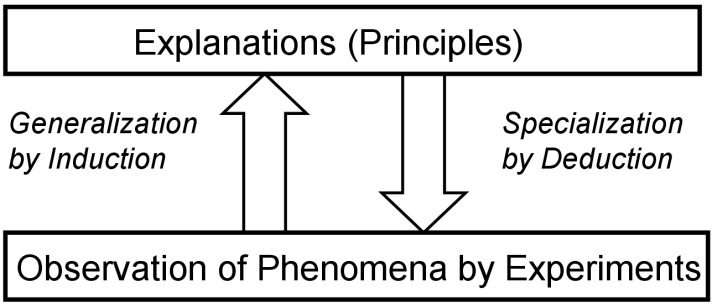


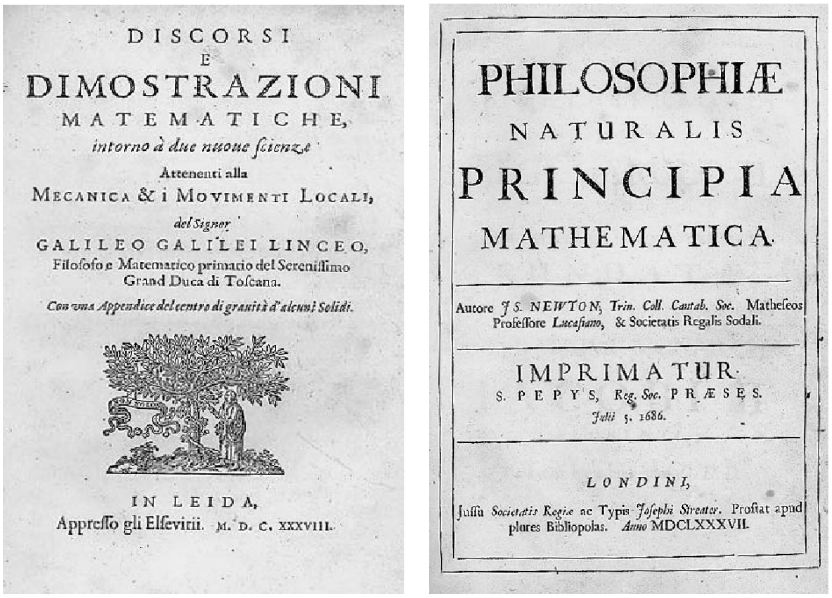
Fig. 2.4. Aristotle's inductive-deductive principle

dual concept of on the one hand, ideal forms, and on the other hand, perceivable phenomena. According to him only the phenomena themselves could be considered as true sources of knowledge about nature. In a process of *generalizing induction*, see Fig. 2.4, explanatory principles are attained. Using general propositions which include those principles, statements on the phenomena are formulated by pure reasoning (deduction). An important point in the Aristotelian method is the fact that his way of modeling does not try to explain things by reduction to higher-ranking mathematical structures (as Plato did). From a modern point of view however, it is astonishing, that Aristotle did not systematically introduce or "invent" the *experiment* as a way to decide about the usefulness and applicability of the explanations of phenomena. This crucial step of artificially *isolating and idealizing* a system from its complex environment was achieved by Galilei [84], which he describes in his *Dialogue* of 1638, cf. Fig. 2.5(a).

He realized the epistemological importance of the experiment which confronts the assumed explanatory principles with the phenomena and thus provides a means of testing and falsifying mathematically formulated hypotheses about nature, cf. Fig. 2.6.

Isaac Newton finally introduced in his *Principia*, cf. Fig 2.5(b), the *axiomatic method* into the natural sciences, where the formulation of mathematical principles in the form of axioms is achieved in a process of intuitive generalization. He was the first one who had the idea that the mechanical behavior of the world might work like a clock that is set; with this idea Newton introduced the crucial partition of the world into *initial conditions* on the one hand, and *laws of nature* on the other hand. In this context Einstein writes in a letter to Maurice Solovine [85]:

> "I see the matter schematically in this way:
> (1) The E's (immediate experiences) are our data.
> (2) The axioms from which we draw our conclusions are indicated by A. Psychologically the A's depend on the E's. But there is no logical route leading from the E's to the A's, but only an intuitive connection (psychological), which is always "re-turning".



(a) (b)

Fig. 2.5. Title pages of the first editions of Galilei's Discorsi **(a)** and Newton's Principia **(b)**. Photos republished with permission. From the original in the Rare Books & special Collections, university of Sydney Library

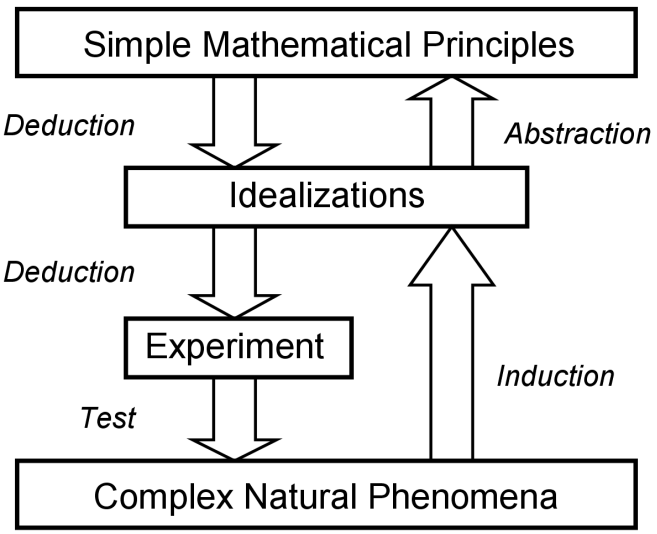


Fig. 2.6. Galilei's method of using experiments to test idealizations of theories which in turn are based on abstract mathematical principles

> (3) *Logically*, specific statements S, S', S" are deduced from A; these statements can lay claim to exactness.
> (4) The A's are connected to the E's (verification through experience). [...] But the relation between S's and E's is (pragmatically) much less certain than the relation between the A's and the E's. If such a relationship could not be set up with a high degree of certainty (though it may be beyond the reach of logic), logical machinery would have no value in the "comprehension of reality"." [accentuations by Einstein] (Albert Einstein, 1952, p. 137)

Thus, according to Einstein, it is impossible to find a strict logical connection between the observed phenomena and the system of axioms by pure induction. Rather, one has to find general principles by *intuition*, using heuristic principles, e.g. symmetry, logical and mathematical simplicity, or keeping the number of independent logical assumptions in a theory as small as possible, cf. [86, 87].

With the axiomatic method one tries to formulate theories on two levels. The first level is the one that states fundamental theorems, principles and the axioms themselves, which summarize experimental facts, e.g Newton's axioms or Maxwell's equations which are "true" or "real" in the sense that they are statements on the behavior of nature. The second level is the level of the theory itself and its interpretation. On this level one may introduce objects and terms such as "potential", "point particle", "wave function", "atom", "quark", "field", etc. which are used to derive and predict properties and states of systems based on the axioms. One might wonder why it is necessary to *interpret* a physical theory when it already has been formulated in axiomatic form. The reason for this need is, that a theory, in order to explain phenomena, needs to contain abstract concepts as elements which go beyond a mere phenomenological description. However, it is not a priori clear to what extend these concepts represent anything in "reality" and can thus be given an explicit meaning. For example, Newton was able to add the notions of an absolute space and time to his theory, which are not defined within his system and which are not used in the formulation of Newton's axioms, i.e. the basic theorems on the first level. Thus, such terms do not really change any consequences or predictions of the theory which could be tested in experiments. Ernst Mach was one of the most prominent critics of such redundancies in Newton's theory which he expresses in the book "Die Mechanik in ihrer Entwicklung – Historisch kritisch dargestellt" [88]. The idea to remove and to avoid redundant elements in theories (sometimes called *Occam's razor*[12]) can sometimes be a very useful heuristic method in the attempts to falsify or improve a theory.

A naïve and primitive notion of "existence" and "reality" is connected with something that can be seen and "perceived" with one's own senses or at least "detected" with some instruments. In modern science however, *every*

[12] After the 14th century monk William of Occam who often used the *principle of unnecessary plurality* of medeval philosophy in his writings, such that his name eventually became connected to it.

element of a theory which leads to consequences that can be tested and falsified is considered to be "as real as a chair"[13]. In this sense, it is theory that determines what "reality" is. An example is the assumption (or hypothesis) of the existence of quarks, the constituents of neutrons and protons in the atomic nuclei. Teller writes about this "philosopher problem" [90]:

> "I take it that the "philosopher problem" refers to the attitude crudely summarized by saying that "if we can't see it, can't see it under any circumstances, then it isn't real". [...] When we see chairs, tables, and any sort of middle-sized macroscopic objects, we do so only via the good offices of a flood of photons, massaged by a lot of optics, interaction with the neurons in the retina, data manipulation in the optic nerve and further cortical processing, until whatever neural processes that ensue count as perception. Now, given that our perception of ordinary chairs and tables is this indirect, what grounds could we have for denying reality (in whatever sense chairs and tables are real) to something we see only slightly more indirectly? [...] In whatever sense you think chairs and tables are real, and once you appreciate the indirectness of our perception of these things, the greater indirectness of seeing smaller things is not going to be an in-principle reason for thinking the smaller things are not real in the same sense. Of course, as the chain involved in indirect perception gets longer, the chance of error may increase; the chances that we have been fooled about quarks may well be larger than the [...] chances that we have been fooled about chairs and tables. But the *kind* of reason we have for thinking that quarks are real differs in degree, not in kind, from the *kind* of reason we have for thinking that chairs and tables are real, always with the same sense of the word "real"." [accentuations by Teller] (Edward Teller, 1997, p. 635)

Despite the fact that it is impossible to detect and virtually "see" quarks separately as individual particles, due to the properties of the interactions between them, these particles are considered to be "real", because their existence helps to establish a consistent theory – the standard model (cf. Sect. 2.5) – which explains all hitherto observed interactions, gravitation excepted.

Pais writes in "*Einstein lived here*" [91] in Chap. 10 that Einstein often distinguished between two kinds of theories in physics: *theories based on principles* and *constructive theories*. This idea was first formulated in a brief article by Einstein in the *Times of London* in 1919 [87]. With a constructive theory, one tries to describe complex observations in relatively simple formalisms, usually based on hypothetical axioms; an example would be kinetic gas theory. With a principle theory, the starting point is not hypothetical axioms, but a set of well-confirmed, empirically found generalized properties – so-called *principles* – of nature; examples include the first and second law of thermodynamics.

[13] See e.g. the comments on p. 54 of Steven Weinberg's book "Dreams of a final theory" [89].

Ultimate understanding requires a constructive theory, but often, according to Einstein, progress in theory is impeded by premature attempts at developing constructive theories in the absence of sufficient constraints by means of which to narrow the range of possibilities. It is the function of principle theories to provide such constraint, and progress is often best achieved by focusing first on the establishment of such principles. In the following, some important general principles are listed that can be found in textbooks on theoretical physics, and which are often used as heuristic guidelines in the development of models in materials science, and for their numerical counterparts.

- Hamilton's principle (Principle of minimization of the integral of action).
- Principle of minimal potential energy (Dirichlet's variational principle).
- Fermat's Principle (Principle of the shortest path of light).
- Principle of virtual work by d'Alembert.
- Ritz' variational principle.
- Galilei's principle of relativity.
- Principle of special relativity.
- Principle of general relativity.
- Principle of general covariance.
- Symmetries of groups and group operations.
- Energy-momentum conservation.
- Angular-momentum conservation.

In this list we haven't mentioned some principles which are applied only on the level of elementary particle physics, such as the principle of charge and parity (CP)-invariance (which is only violated in the decay of the neutral K-meson) or the principle of charge, parity, and time (CPT)-invariance which is assumed to be valid for *all* known fundamental interactions.

As discussed above, there is no general learnable way of guessing or finding physical laws; rather, for want of a logical path often intuition and (sometimes even unconscious) ad-hoc hypotheses finally lead to success[14]. A law of nature that has been found or a theory that has been formulated is just a guess which is then put to the (experimental) test. Some common key features with the development of any model are:

- Simplifying assumptions must be made.
- The number of logically independent elements and heuristic theorems which are not derived from basic notions (axioms) should be as small as possible.
- Boundary conditions or initial conditions must be identified.
- The range of applicability of the model should be understood.

[14] Albert Einstein's quest for a formulation of general relativity during the years 1907–1915 is the classic example, cf. Chap. 3 on p. 171.

It is important to realize, that a theory or a model can only explain phenomena, if it contains abstract concepts as elements which go beyond mere observation. An example for this is Maxwell's displacement current $\partial\vec{E}/\partial t$ (in appropriate units), which he added to Ampère's original law based purely on theoretical considerations, which cannot be obtained from observation.

Example 1 (Maxwell's and Einstein's Field Equations). There is an interesting analogy between Maxwell's development of the field equations of electromagnetism in 1865 [92] and Einstein's development of the field equations of general relativity half a century later. Beginning in the 1850's, Maxwell elaborated ideas of Faraday to give a complete account of electrodynamics based on the concept of continuous fields of force (in contrast to forces acting at a distance). In modern terminology he arrived at the first gauge theory of physics, using the following four partial differential equations which encode observational facts for the electric and magnetic fields $\vec{E}(\vec{x},t)$, $\vec{B}(\vec{x},t)$ and their corresponding sources, charge density and current density $\rho(\vec{x},t)$, $\vec{j}(\vec{x},t)$, directly derived from experiment:

$$\nabla\vec{E} = 4\pi\rho \,, \tag{2.3a}$$

$$\nabla\times\vec{B} = \frac{4\pi}{c}\vec{j} + \frac{1}{c}\frac{\partial\vec{E}}{\partial t} \,, \tag{2.3b}$$

$$\nabla\times\vec{E} = -\frac{1}{c}\frac{\partial\vec{B}}{\partial t} \,, \tag{2.3c}$$

$$\nabla\vec{B} = 0 \,. \tag{2.3d}$$

Maxwell added the extra term $\frac{1}{c}\frac{\partial\vec{E}}{\partial t}$, which was *not* obtained from experiment, to Ampère's law in (2.3b), such that the continuity equation $\nabla\vec{j} = -\partial\rho/\partial t$ is also fulfilled in the case of time dependent fields. The continuity equation then follows from (2.3a) and (2.3b). Additionally, the inclusion of the displacement current leads to transverse electromagnetic waves propagating in a vacuum at the speed of light. Thus, the combination of charge conservation and Coulomb's law implies that the divergence of equation (2.3b) vanishes and renders the set of equations mathematically consistent.

With a similar consideration Einstein arrived at the final field equations of general relativity; the simplest hypothesis involving only the metric coefficients g_{mn} and their first derivatives, is that the Ricci tensor R_{mn} equals the stress energy tensor T_{mn}, (see Chap. 3 for a proper introduction to tensors). It turns out however, that the divergence of R_{mn} does not vanish as it should in order to satisfy local conservation of mass-energy. However, the tensor $R_{mn} - 1/2 g_{mn}R$ does have vanishing convergence due to Bianchi's identity[15]. Thus, when including the additional trace term $-1/2 g_{mn}R$ one yields the complete and mathematically consistent field equations of general relativity:

[15] $R_{ijkl|m} + R_{ijlm|k} + R_{ijmk|l} = 0$, see e.g. [93].

$$R_{mn} - \frac{1}{2} g_{mn} R = -\frac{8\pi G_N}{c^4} T_{mn} \, . \tag{2.4}$$

Einstein commented on this in a letter to Michele Besso in 1918 in which he was chiding Besso for having suggested (in a previous letter) that, in view of Einstein's theory of relativity, "speculation has proved itself superior to empiricism". Einstein disavowed this suggestion, pointing out the empirical base for all the important developments in theoretical physics, including the special and general theory of relativity. He concludes [94]:

> "No genuinely useful and profound theory has ever really been found purely speculatively. The closest case would be Maxwell's hypothesis for displacement current. But there it involved accounting for the fact of the propagation of light (& open circuits)." (Albert Einstein, 1918, p. 524)

The question to what extend a physical theory maps a small part of reality has been answered differently at different times, but one can distinguish at least three different convictions or "paradigms":

1. *Phenomenological*: one seeks an economic description of sensory perceptions and defines them as "reality".
2. *Operational*: one seeks instructions, according to which the descriptive elements of the theory can be measured, and defines them as "reality".
3. *Realistic*: one seeks abstract principles and theorems that go beyond a mere description of observations and defines "reality" by all those elements of the theory the consequences of which can be falsified by experimental tests.

The phenomenological point of view interprets theories as an instrument of describing observations in an economic way; this attitude could be described as positivism of Machian character. The Copenhagen interpretation of quantum theory, cf. Chap. 5, is based on an operational interpretation of a physical theory. Here, everything that can be measured is considered to be "real". A realistic interpretation of theories assumes that the used notions in a theory go beyond a mere description of observations and all elements or objects that are introduced in the theory are considered to be "real" if they lead to any consequences that can be tested in experiment. For example, in the current theory of elementary particles – the standard model – symmetry principles play a fundamental role; they build the basic mathematical ontology of physics. In this context Weinberg writes in "*The rise of the standard model*" [95]:

> "The history of science is usually told in terms of experiments and theories and their interaction. But there is a deeper level to the story – a slow change in the attitudes that define what we take as plausible and implausible in scientific theories. Just as our theories are the product of experience with many experiments, our attitudes are the product of experience with many theories.[...] The rise of the standard model was accompanied by profound changes in our attitudes toward symmetries and field theory." (Steven Weinberg, 1997, p. 36)

2.4 Hierarchical Modeling Concepts above the Atomic Scale

The equations of fundamental theories such as quantum theory become too complex when being applied to macroscopic systems, which involve an astronomical number of constituents. Thus, various approximative theories have been devised, which lead to equations that can be solved at least numerically. Each theory has its range of applicability and this is the main reason why there are so many different computational methods that are used in engineering and materials science, cf. Fig. 2.7.

Classical Newtonian mechanics is a scientific system which is only approximatively valid for the description of the dynamics of condensed matter at small velocities and for lengths larger than $\sim 10^{-10} m$. For smaller distances,

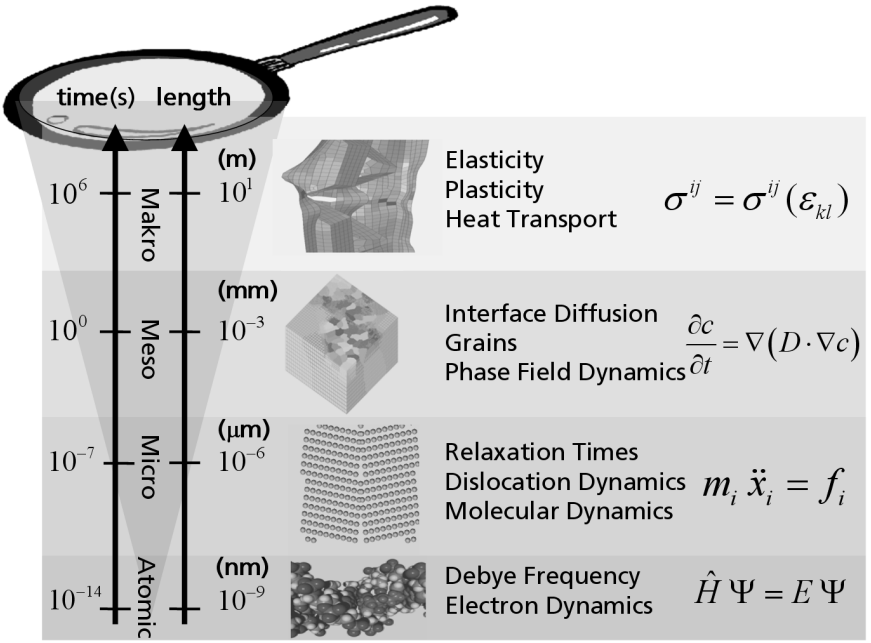


Fig. 2.7. The structural hierarchy of length scales in nature is also reflected in fundamental physical theories based on fields and particles, which have different scopes of validity. Depicted are some typical physical phenomena on different length scales and some basic equations of physical theories used on the respective scale, e.g. the Schrödinger equation at the atomic scale, classical Newtonian particles dynamics, the diffusion equation based on the concept of fields, or the constitutive equation for large scale elasticity and plasticity, connecting stress (force per unit area) and strain in a body, modeled as a continuum with an infinite number of degrees of freedom

classical mechanics has to be replaced by a different system – quantum theory, or sometimes called quantum mechanics, when referring specifically to the description of the motion of particles with mass. For large velocities, special-relativistic mechanics, where space and time are united in spacetime, has to be used, see Sect. 3.6.1 in Chap. 3 on p. 166.

2.4.1 Example: Principle Model Hierarchies in Classical Mechanics

Classical theoretical mechanics is an appropriate theory for the description of phenomena occurring between objects above the spacial dimensions of atoms ($\sim$1Å$= 10^{-10}$ m) and for small velocities. For the description of the dynamics of objects of atomic dimensions, classical physics breaks down and instead quantum theory is a more appropriate model. Today, it is believed that quantum theory is *the* fundamental theory, underlying all subatomic, atomic, microscopic, mesoscopic and macroscopic objects and natural phenomena. For a discussion and references, see Chap. 5.

There are different layers or hierarchies of models in theoretical physics. These could be grouped according to the number of general natural phenomena which can be explained by the respective model system. In *classical mechanics* for example there are different abstraction layers, or hierarchies of the mechanical systems, which we shortly discuss in the following.

Mechanics of Mass Points

The possibly simplest of all model systems is the model of a mass point which is based on the notion of a dimensionless object, i.e. an imaginary object which has no extension in space but contains all mass m of the considered system. This is equivalent with the notion of a "point" in mathematics which can be identified by being assigned some numbers in an appropriately chosen coordinate system. The motion of mass points is then described within the classical Newtonian system. Whether this approximation is useful depends on the considered system (e.g. the motion of earth around the sun).

Mechanics of Point Systems

If there are many mass points one speaks of a *point system*, or a many $N-$particle system. By deriving equations of motion for point systems, some general principles, so called *integral principles* such as d'Alembert's or Hamilton's principle can be derived. The advantage of general principles in the derivation of equations of motion (EOM) is that there is no preferred coordinate system. Thus, using variational calculus one can obtain EOM in general coordinates. Due to the large number of mass points in a macroscopic body of

order $\mathcal{O}(10^{23})$, and the just as many ordinary differential equations, a general solution is only possible when symmetries are present[16].

Mechanics of Rigid Bodies

A rigid body is a special system of mass points in which one can neglect the relative motion of the points with respect to each other. Real systems fulfill this condition only approximately. However, many bodies – under normal conditions – change their shape and volume only marginally. A rigid body has 6 degrees of freedom (3 translations and 3 rotations); thus, for such systems one obtains 6 differential equations of 2nd order.

Mechanics of Continua

In continuum mechanics one only considers motions of bodies, in which neighboring mass points (e.g. atoms) approximately move into the same direction. In this case one can approximate the whole body as a continuum with an infinite number of degrees of freedom, cf. the discussion in Sect. 7.7. One advantage of this approximation is that the number of equations is reduced drastically; instead of having to solve many ordinary differential equations, one has to solve only a few but *partial* differential equations. Usually one distinguishes *elastomechanics* which treats deformable solid media (solid states), where the constituents are still bound strongly together such that small deviations lead to strong forces (inner tensions), and *hydrodynamics* which is applicable to gases and fluids where the constituents are bound only weakly together (see Chap. 7).

It can be useful to employ different models for the same system. For example, when considering the revolution of earth around the sun, it is useful to model the Earth (and the Sun) as a mass point which contains all mass. However, when considering its rotary motion it is useful to use the model of an extended stiff body. However, both of these models will fail when trying to describe tidal effects due to movements of matter within the earth's crust. Here one has to use continuum theoretical concepts. Which model is the best in which situation is not an easy task to decide and no algorithm can help a scientist here in making a useful decision. Sometimes there are some empirical rules (so called *heuristics*, i.e. concepts which have proved successful in the past even if these concepts are rather empirical) which can be used as a guide. In the end, it is the experience and intuition of the scientist which lead to the use of a certain model in some situation.

[16] In fact, the N-body problem is analytically unsolvable for $N \geq 3$, cf. Example 37 on p. 270.

2.4.2 Structure-Property Paradigm

In experimental materials science today one usually breaks a system (e.g. a material such as a metal specimen or a ceramic plate) into smaller pieces and investigates the properties of the obtained smaller structures with the aid of microscopes. The idea, or conviction behind this procedure is based on our hierarchical view of the structure of matter and the believe that if one can understand the mechanics of a small subsystem, one can also understand the whole system. One assumes as a working hypothesis that the observed macroscopic phenomenological properties of a material (e.g. its plastic or elastic reaction to an external load) are determined *in principle* by the properties of its meso-, micro-, and nanostructure. This assumption is called the *structure-property paradigm* of materials science, cf. Table 2.1.

2.4.3 Physical and Mathematical Modeling

As presented in Fig. 2.1 on p. 32, materials science covers roughly 12 orders of magnitude in size, along with the associated time scales of physical processes. While atomistic methods based on quantum mechanics prevail in detailed microstructure simulations at the nanoscale as well as coarse-grained atomistic simulations neglecting the electrons at the microscale, such a detailed numerical treatment of systems cannot be done at the meso- and macroscale. Here, one has to average out the many atomic degrees of freedom and use "superatoms" which represent large clusters of atoms. With this approach, systems on the meso- and macroscale can still be treated with a classical particle approach solving Newton's equations of motion, see e.g. [7, 96]. The physical and mathematical part of model building often go hand in hand, as physical ideas on the materials behavior are usually formulated using mathematical concepts and (differential) equations.

Table 2.1. Illustration of the structure-property paradigm of materials science. Molecular structural properties of systems determine their mesoscopic structures and ultimately the observed macrospcopic properties of solids and fluids

Structure			Property
Molecular			Macroscopic
Nano	Micro	Meso	Macro
Electronic structure	Molecule size	Volume ratio	Viscosity
Inter-atomic interaction	Molecular weight	Packing density	Strength
Bond angles	Fiber/matrix interaction	Fiber orientation	Toughness
Bond strength	Grain size distribution	Flexibility	Modulus
Bond failure	Cross-link density	Dispersion	Stress/strain
Chemical sequence	Defects	Heat transport	Plasticity
Unit cell	Crystallinity	Grain orientation	Durability

State Variables and Equations of State for Micro- and Mesostructural Evolution

When investigating fluids or solids using continuum based microstructural simulations, one is usually not interested in the dynamics, i.e. positions and momenta of single fluid particles or superatoms in the solid; rather, one is interested in the average behavior of the system's macroscopic state variables such as temperature T, pressure p, density ρ, displacement u^i, stress σ^{ij} and strain ϵ^{ij}, or free energy F in Ginzburg-Landau type models, etc., which are described as continuum variables. In thermodynamics, *extensive* state variables such as entropy S, volume V, or energy E of a heterogeneous system are additive with respect to the different phases of a system, i.e. they are proportional to the amount of substance, e.g. mass m or number of particles N. In contrast, *intensive* state variables are independent of the amount of substance and may assume different values in different phases of a system. Examples are refraction index, p, ρ, or T, which may be defined *locally*, that is they are parameterized fields with an infinite number of degrees of freedom within the framework of classical field theories. Usually time and position are used for parameterizing field properties, i.e. the state variables are functions of position and time. To determine the spacial dependency of intensive variables, one needs additional conditional equations, for example from hydrodynamics or in the form of other phenomenological equations of state. Examples for equations of state are Hooke's law in dislocation dynamics, nonlinear elasticity laws in polymer dynamics, or the free energy functional in Ginzburg-Landau type microstructural phase field models. The question as to how many state variables are needed to completely characterize a closed system at equilibrium is answered by Gibb's phase rule (see Problem 6.2 on p. 327):

$$f = C + 2 - P \,, \tag{2.5}$$

where C is the number of chemical components, P is the number of phases and f labels the degrees of freedom.

Phenomenological descriptions based on thermodynamic equations of state and continuum mechanics are prevailing in typical engineering applications on the meso- and macroscale, for example in the prediction of material behavior of composite materials such as concrete, metal-matrix composites, polymer fiber-reinforced composites or multi-phase ceramics under various load or processing conditions. A great disadvantage of detailed phenomenological descriptions taking into account structural features observed on the meso- and microscale, is the often large number of state variables that is required. Such an approach, involving many variables, may quickly degrade a transparent physical model based on few assumptions to a mere empirical polynomial fit model, where the state variables just serve as fitting parameters. Too great a number of parameters often reduces the value and the physical significance of a model considerably. In many practical engineering approaches in industry, a many-variable fitting approach may be helpful to gradually optimize materials

in certain manufacturing processes, too complicated for an explicit description, however, it is less desirable in *physically* oriented simulations taking into account micro- and mesostructures.

State variables in mesoscale simulations are usually weighted with certain additional, empirical parameters. Often, these mesoscopic parameters are nonlinearly coupled with other equations of state which renders such models numerically rather complex. By definition, state variables are defined for systems at thermal equilibrium. A system like a polycrystalline microstructure however, is generally not at equilibrium and the systems evolution may occur irreversible; hence, the system's evolution equations are generally path dependent, i.e. they are no total differentials which can be readily integrated. This difficulty gave rise to the increased application of statistical models in recent years, often based on the Monte Carlo Method [97] (cf. Chap. 6.6). MC methods have been used e.g. for the simulation of diffusion behavior or short range ordering [98, 99], recrystallization [100], grain growth [101, 102], or boundary misorientations [103, 104]. Among the most successful mesoscale models for microstructural evolution simulations are vertex models [105], phase field models [106], cellular automata [107, 108] and Potts models [109]. In the beginning 1980s it was realized that Potts domain structures are similar to granular microstructures. As both systems are characterized by a space-filling array of domains which evolve to minimize the boundary area, the Potts model was used for a variety of simulations such as late-stage sintering [110], or grain growth in polycrystals in 2 dimensions (2D) [111, 112, 113, 114], and in 3D [115]. As a result of the above-mentioned modeling approaches for microstructure dynamics, one obtains a set of governing equations, which model the microstructural elements of the considered solid by means of state variables, that are functions of position, time or other parameters, such as the dislocation density or grain curvature.

Kinematic Equations

The mere description of the motion (changes of position) of objects in classical mechanics without reference to their origin, that is forces, is called *kinematics*. Kinematic equations allow for the calculation of certain mechanical properties which are based on coordinates and its derivatives with respect to time, velocity and acceleration, e.g. strains, strain rates, crystal orientations, or rigid body spin[17] to name but a few. It is very important to understand that positions are assigned with reference to arbitrarily chosen coordinate systems. Therefore, positions, that is, coordinates have no intrinsic physical (i.e. metrical) meaning. They are simply numbers that are used to label events in spacetime which change when a different coordinate system is used which may be linearly shifted or rotated with respect to the original one.

[17] In continuum theory this is the antisymmetric part of the tensor of the displacement derivatives.

Spacetime is a unification of space and time into one single (flat) manifold, called *Minkowski space*[18] which simplifies a large amount of physical theory; in particular, it is the underlying structure to be used for the description of events introduced in the special and general theory of relativity (for a formal introduction see Sects. 3.6.1 and 3.6.2 in Chap. 3).

Example 2 (Special Relativistic Kinematics and Lorentz Transformations). An event in spacetime is a point $x^\mu\,, (\mu = 1, 2, 3, 4)$ specified by its time and three spacial coordinates, i.e. $x^\mu = (ct, \vec{x})$, which are called *Minkowski coordinates*

$$x^0 = ct\,,\; x^1 = x\,,\; x^2 = y\,,\; x^3 = z\,. \tag{2.6}$$

The quantity denoted by x^μ is often called *four-vector* (instead of "components of a four-vector" $\vec{x} = x^\alpha \vec{e}_\alpha$ with orthonormal basis $\vec{e}_\alpha \vec{e}^{\,\beta} = \delta_\alpha^\beta = \delta_{\alpha\beta}$). The transformation between contra- (upper) and covariant (lower) components is achieved by the metric tensor of Minkowski space, that is, the Minkowski metric $\eta_{\mu\nu}$, cf. (3.16) on p. 139:

$$x_\mu = \sum_{\nu=1}^{4} \eta_{\mu\nu} x^\nu = (ct, -\vec{x})\,. \tag{2.7}$$

Using a Minkowski metric, the infinitesimal distance ds^2 between two events (two points in Minkowski space) is given by $ds^2 = c^2 dt^2 - d\vec{x}^{\,2}$. The *worldline* of an object, e.g. a particle, is the path that this particle takes in the spacetime and represents its total history. In a sense, physical reality is given by all the events described in spacetime. Applying the special principle of relativity (see Sect. 3.6.1 on p. 166), one obtains for two different inertial frames of reference Σ and Σ', cf. Fig. 2.8, the proper transformation laws of space time coordinates, which are the linear *Lorentz transformations*:

$$x'^\alpha = \Lambda_\beta^\alpha\, x^\beta + a^\alpha\,, \tag{2.8}$$

where a^α represents a time and space translation, and $\Lambda_\beta^\alpha \in SO(3)$. The group of Lorentz-transformations is called Poincaré group and contains (just like the Galilei group) 10 parameters. The translations and rotations build a subgroup of both, the Galilei group and the Lorentz group. In this subgroup one normally excludes a change of handedness, i.e. $\mathrm{Det}(\Lambda) = +1$. In Minkowsik space, the scalar products of four-vectors are invariant under Lorentz-transformations, i.e. $\Lambda_\alpha^\beta \Lambda_\beta^\gamma = \delta_\alpha^\gamma$. For example, for two vectors $A^\mu = (A^0, \vec{A})$ and $B^\mu = (B^0, \vec{B})$ the product $\mathsf{A} \cdot \mathsf{B} = A_\mu B^\mu = A^\mu B_\mu = A^0 B^0 - \vec{A}\vec{B}$. The Lor.entz-transformations are distinguished in that they leave the *proper time interval*

[18] After the mathematician Hermann Minkowski, mathematics professor at the Polytechnikum ETH in Zurich and teacher of Einstein, who introduced this unifying concept in a famous lecture at Cologne in 1908, see pp. 54–71 in [116].

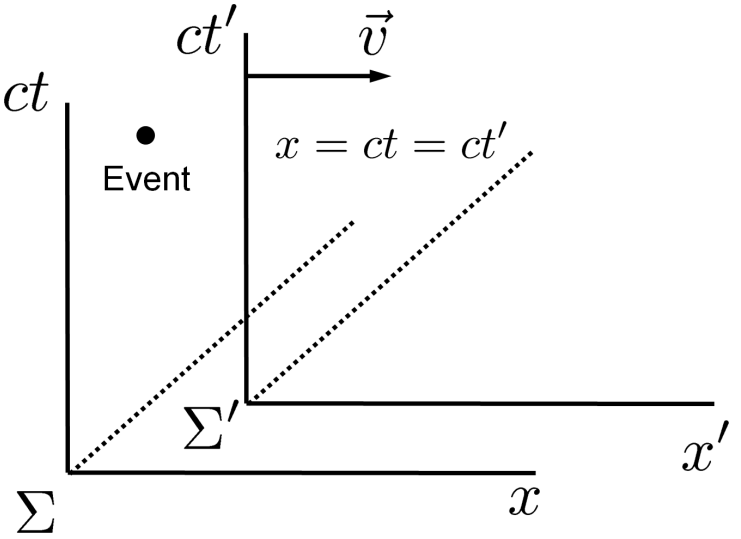


Fig. 2.8. A spacetime diagram in Minkowski space. Two coordinate systems Σ and Σ' move with velocity $\vec{v}$ relative to each other. This configuration is called *standard configuration* in special relativity theory. For simplicity, only two coordinates, x, and ct (time expressed as distance traveled by light in time interval t, with the conversion factor c (velocity of light) are shown. A light signal (*dotted line*), e.g. a flash travels at speed $v = c$, thus its worldline is represented as a bisecting line in both coordinate systems. Let the coordinates of a certain event in Σ be (t, x). In Σ', which moves relative to Σ with velocity $\vec{v}$ in positive x-direction, the very same event is described by coordinates (t', x')

$$d\tau^2 = c^2 dt^2 - (d\vec{x})^2 = \sum_{\mu=0}^{3} dx^\mu dx_\mu \ . \tag{2.9}$$

invariant, cf. Prob. 1. $d\tau$ is a Lorentz invariant and in the rest system of an observer it coincides with the coordinate time. Taking the derivative of the position four-vector with respect to τ is also a four-vector, the four-velocity u^μ.

$$u^\mu = \frac{dx^\mu}{d\tau} = \left(\frac{cdt}{d\tau}, \frac{d\vec{x}}{d\tau}\right) = \left(\frac{cdt}{d\tau}, \frac{d\vec{x}}{dt}\frac{dt}{d\tau}\right) = \frac{dt}{d\tau}(c, \vec{v}) \ . \tag{2.10}$$

To calculate $\frac{dt}{d\tau}$ we go back to (2.9) which can be written as:

$$(d\tau)^2 = (dt)^2 \left(1 - \frac{1}{c^2}\left(\frac{d\vec{x}}{dt}\right)^2\right) = (dt)^2 \left(1 - \left(\frac{\vec{v}}{c}\right)^2\right) \ . \tag{2.11}$$

Hence,

$$\frac{dt}{d\tau} = \frac{1}{\sqrt{1 - \left(\frac{v}{c}\right)^2}} =: \gamma \ , \tag{2.12}$$

and

$$\frac{dx^\mu}{d\tau} = \gamma(c, \vec{v}) \ . \tag{2.13}$$

Multiplying this equation with the invariant mass m yields the four-momentum

$$p^\mu = mc^2 = m\frac{dx^\mu}{d\tau} = m\gamma(c, \vec{v}) = \left(\frac{E}{c}, \vec{p}\right) \ . \tag{2.14}$$

The invariant of the four-momentum is obtained by calculating the inner product of p^μ:

$$\sum_{\mu=0}^{3} p^\mu p_\mu = \frac{E^2}{c^2} - \vec{p}^2 = m^2c^2 \ . \tag{2.15}$$

This is the relativistic energy-momentum relation:

$$E = c\sqrt{(mc)^2 + \vec{p}^2} \ . \tag{2.16}$$

In the rest-system of an observer ($\vec{p} = 0$) the energy is

$$E_0 = mc^2 \ , \tag{2.17}$$

an equation, that has been first derived by Albert Einstein in 1905 [117] and in several later publications [118, 119]. For a photon ($m = 0$), for which no rest system exists, the energy is $E = c\vec{p}$. In some textbooks on relativity, even in the famous 1921 article by W. Pauli [120], a distinction between a *rest mass* m_0 and a *relativistic, velocity-dependent mass* $m_0\gamma$ is made; this however is deceptive, as the mass of a system is a fundamental property of matter and as such an invariant, which does not change with the frame of reference. Thus, it is important to understand that not mass, but rather the *energy* and the *momentum* of a system depend on the state of motion of the observer. In the reference frame in which the system is at rest (in its *rest system*), its energy is not zero, but $E_0 = mc^2$, i.e. proportional to its frame-independent mass m. Energy E and momentum $\vec{p}$ are both components of a four-vector (2.14), and transform together when changing the coordinate system.

A light signal in the standard configuration of Fig. 2.8 travels at a speed of $c = dx/dt$. Therefore, $ds^2 = 0$ for light signals. Thus, it follows

$$ds'^2 = c^2dt'^2 - dx'^2 = \eta_{\alpha\beta}dx'^\alpha dx'^\beta = \eta_{\alpha\beta}\Lambda^\alpha_\gamma\Lambda^\beta_\delta dx^\gamma dx^\delta = ds = \eta_{\gamma\delta}dx^\gamma dx^\delta = 0. \tag{2.18}$$

Hence, by comparison, one obtains

$$\Lambda^\alpha_\gamma\Lambda^\beta_\delta\eta_{\alpha\beta} = \eta_{\gamma\delta}. \tag{2.19}$$

The spacetime translations in (2.8) drop out when taking the differential

$$dx'^\alpha = \Lambda^\alpha_\beta \, dx^\beta \ . \tag{2.20}$$

Because of $x^2 = x'^2$ and $x^3 = x'^3$ in the standard configuration of Fig. 2.8 one can write the transformation as

$$\Lambda = (\Lambda^{\alpha}_{\beta}) = \begin{pmatrix} \Lambda^0_0 & \Lambda^1_0 & 0 & 0 \\ \Lambda^1_0 & \Lambda^1_1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} . \tag{2.21}$$

From (2.19) it follows:

$$(\Lambda^0_0)^2 - (\Lambda^1_0)^2 = 1 \,, \tag{2.22a}$$
$$(\Lambda^1_1)^2 + (\Lambda^0_1)^2 = -1 \,, \tag{2.22b}$$
$$\Lambda^0_0 \Lambda^0_1 - \Lambda^1_0 \Lambda^1_1 = 0 \ . \tag{2.22c}$$

As a solution of (2.22) one may set $\Lambda^0_1 = -\sinh\Theta$ and $\Lambda^1_0 = -\sinh\Theta$. This yields

$$\Lambda = (\Lambda^{\alpha}_{\beta}) = \begin{pmatrix} \Lambda^0_0 & \Lambda^1_0 \\ \Lambda^1_0 & \Lambda^1_1 \end{pmatrix} = \begin{pmatrix} \cosh\Theta & -\sinh\Theta \\ -\sinh\Theta & \cosh\Theta \end{pmatrix} . \tag{2.23}$$

For the origin of Σ' the following equation applies, cf. Fig 2.8:

$$x'^1 = 0 = \Lambda^1_0 \, ct + \Lambda^1_1 \, vt \,, \tag{2.24}$$

and it follows:

$$\tanh\Theta = -\frac{\Lambda^1_0}{\Lambda^1_1} = \frac{v}{c} = \beta \,. \tag{2.25}$$

With (2.21), (2.23) and (2.25) all matrix elements are fixed. Expressed as a function of velocity v, the matrix elements are:

$$\Lambda^0_0 = \Lambda^1_1 = \gamma = (1 - \beta^2)^{-1/2} \,, \tag{2.26a}$$
$$\Lambda^0_1 = \Lambda^1_0 = \frac{-v/c}{\sqrt{1 - \frac{v^2}{c^2}}} \,. \tag{2.26b}$$

As a result, the *Lorentz-transformations* are given by:

$$x' = \gamma(x - vt) \,, \quad y' = y \,, \quad z' = z \,, \tag{2.27a}$$
$$ct' = \gamma(ct - \beta x) \,. \tag{2.27b}$$

These transformations allow for transforming spacetime coordinates from one *locally* defined IS to another IS with its own *local* spacetime coordinates. Thus, special relativity provides a fundamental insight into the structure of spacetime, used in physical theory, in that each single observer has his own set of coordinates, with which spacetime intervals between events are calculated. Hence, there is no *global* coordinate system that can be provided for

all observers in inertial systems and likewise there is no common notion of simultaneity.

A clock, that is at rest in system Σ in Fig. 2.8 displays the so-called proper time $\tau = t$. Thus, the proper time between two events, measured by a clock that is at rest in a frame of reference Σ coincides with the coordinate time and is a directly measurable quantity. In a new (primed) coordinate system Σ' with coordinates $x^{\alpha'}$, the coordinate differentials are given by

$$dx^{\alpha'} = \Lambda^{\alpha}_{\beta} dx^{\beta} \quad . \tag{2.28}$$

Thus, the new coordinate time $d\tau'^2$ will be

$$d\tau'^2 = \eta_{\alpha\beta} dx'^{\alpha} dx'^{\beta} = \eta_{\alpha\beta} \Lambda^{\alpha}_{\lambda} \Lambda^{\beta}_{\delta} dx^{\lambda} dx^{d} = \eta_{\lambda\delta} dx^{\lambda} dx^{\delta} \ , \tag{2.29}$$

and therefore

$$d\tau'^2 = d\tau \ . \tag{2.30}$$

It is easy to see that for $\beta = \frac{v}{c} \ll 1$, prefactor $\gamma = \sum_{n=0}^{\infty} \frac{(-1)^n}{n!} (\beta)^n \approx 1$ in zeroth approximation, thus yielding the approximatively valid *Galilei transformation* (cf. (3.130) on p. 167), where time parameter t does not depend kinematically on the spacial coordinates, that is, on the arbitrarily chosen inertial frame of reference. In this case, the underlying spacetime manifold has the same *global* properties for all observers.

All so-called special relativistic "paradoxa", such as length contraction ($L' = \gamma L$), time dilatation ($t' = \gamma t$)[19], or the relativity of simultaneity are simple consequences of special relativistic kinematics. That is, they arise simply because the numerical values of spacetime coordinates change when switching between inertial systems.

In engineering contexts, similar kinematic properties arise with objects defined in continuum theory such as the deformation or velocity gradients, when they are transformed to a different coordinate system, see e.g. [123]. Usually, the transformations considered in this context, are very special *Euclidean transformations* which include a translation of the system and an orthogonal rotation, and the so-called "non-objectivity"[20] of several quantities does not come as a surprise as they are not defined as covariant[21] tensor equations.

It is important to realize that the changing coordinates (and their transformation rules) are actually not the point of special (and general) relativity theory[22], but rather those properties of spacetime, which are *invariant*

[19] The most prominent application of this kinematic effect is the *twin paradox*, see e.g. [121, 122], which is based on the asymmetry between the twin that stays in the same IS at home, and the one who travels, and has to accelerate, i.e. to switch inertial systems, in order to return. For a recent treatment of the clock paradox, see [122].

[20] Non-invariancy of equations upon coordinate transformations.

[21] See Sect. 3.3.8 on p. 156.

[22] Thus in a sense, the term "relativity theory" is a misnomer and it had probably better be called "theory of invariants".

upon coordinate changes, such as events as such, or the *spacetime interval* $ds^2 = c^2dt^2 - d\vec{x}^{\,2}$, which connects physical events in spacetime through *timelike* ($ds^2 > 0$), *spacelike* ($ds^2 < 0$), and null ($ds^2 = 0$) worldlines. The implications of this theoretical structure underlying classical physics are further considered in Sect. 3.6.1 of Chap. 3 on p. 166.

2.4.4 Numerical Modeling and Simulation

Once a decision is made, a physical model is expressed in mathematical equations which (usually) can be solved in a systematic way, that is, in a way that can be formulated as a finite stochastic or deterministic algorithm[23] and implemented as a computer program. The numerical solutions of the governing equations associated with the physical and mathematical model are then interpreted and provide answers to the specific real system which was transformed into the model system. In Table 2.2, several key classification aspects in physical and mathematical modeling are collocated. A comparison of the answers for a specific problem obtained by mathematically exploiting a specific model, finally provides some ideas about the general validity and the quality of a model system and the derivations and theoretical concepts associated with it. This principal procedure in physical and numerical modeling is illustrated in the flowchart of Fig. 2.9.

2.5 Unification and Reductionism in Physical Theories

In Sect. 2.3 it was stated that on the one hand one has to isolate a system in order to extract the fundamental, universal laws; on the other hand, the natural sciences – and in particular physics – are very much focused on *unifying* different approaches, models and theories, i.e. to reduce existing systems to ever more consistent and fundamental systems that include an ever larger part of observed reality. This paradigm of "reductionism" expresses the idea

Table 2.2. General key aspects to be considered in the development of physical and mathematical models for simulation applications beyond the atomic scale

Classification	**Example**
spatial dimension	1D, 2D, 3D
kind of discretization	particles, super particles, continuum (fields)
spacial scale	macroscopic, mesoscopic, microscopic, nanoscopic
state variables	strain, displacement, dislocation density, temperature
material properties	Hooke's law, Taylor equations, multiparameter plasticity
degree of predictability	ab initio, coarse-grained, phenomenological, empirical

[23] For a definition of "algorithm", see Sect. 2.6.

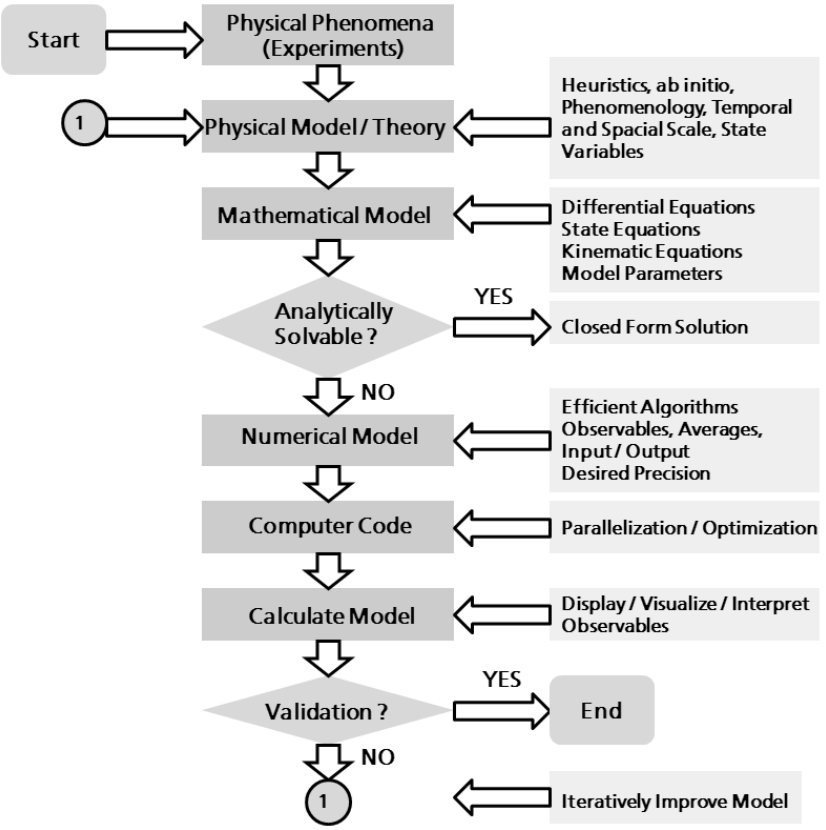


Fig. 2.9. Principal physical, mathematical and numerical modeling scheme illustrated as flow chart. Starting from the experimental evidence one constructs physical theories which determine in principle what can be measured. A mathematical formulation usually leads to differential equations, integral equations, or master (rate) equations for the dynamic (i.e. time dependent) development of certain state variables within the system's abstract state space. Analytic solutions of these equations are usually rarely possible, except for simplified approximation usually due to symmetry. Thus, efficient algorithms for the treated problem have to be found and implemented as a computer program. Execution of the code yields approximate numerical solutions to the mathematical model which describes the dynamics of the physical "real" system. Comparison of the obtained numerical results with experimental data allows for a validation of the used model and subsequent iterative improvement of theory

that it is possible to understand the functioning of arbitrary complex (physical, chemical, biological) systems by reducing them to simpler and smaller systems and by applying the laws of nature to these subsystems. A pessimist could say, that this "fool-proof philosophy" has not led to anything that goes

beyond old Platonist concepts of "ideal forms" that are hidden behind the empirical observations (see e.g. Chap. 4 in "The Presence of the Past" by R. Sheldrake [124]), but its simplicity is very attractive and has led to the successful unification of different theories of physical phenomena during the last 200 years, see Fig. 2.10.

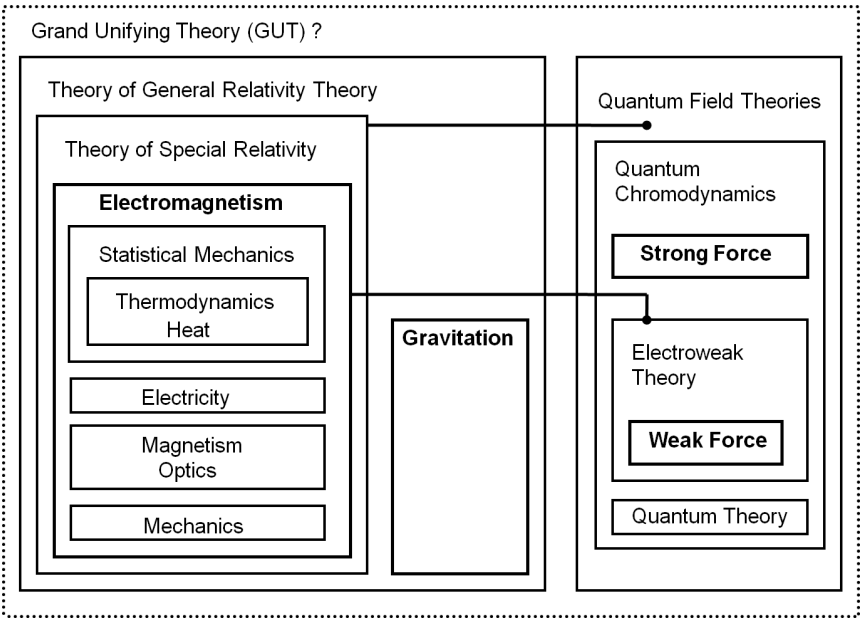


Fig. 2.10. Reductionism in physics. Important unifications of *classical* physical theories based on the concepts of particles and fields are summarized on the left. Unifications based on quantum theory are displayed on the right. By extending the Galileian principle of relativity to *all* processes in physics including electrodynamics, the special theory of relativity (SRT) (1905) could reconcile the apparent disaccord between Newton's system of classical mechanics (1687) and Maxwell's equations (1865). Combining SRT with Quantum Theory (1925) yields Quantum Field Theory (1931) in which fields are treated as operators in Hilbert space, which obey commutator relations and which gives rise to the concepts of antimatter and particle annihilation and creation out of the vacuum (Dirac's hole theory). Within this framework, the fundamental weak force and electromagnetism could be united in the electroweak theory and the strong interactions of hadrons could be explained in a dynamic theory of quarks, called quantum chromodynamics, (QCD) (1970). Thus, three of the four known fundamental interactions have been unified in more general quantum field theories; gravitation, the weakest of the four interactions, which was transformed from a theory with action at-a-distance (1687) into a proper non-instantaneous field theory by the general theory of relativity (1915), might be unified with the other forces in an all-embracing "theory of everything", or GUT, a "Grand Unified Theory"

The paradigm of reductionism is carried to extremes in modern high energy particle physics. Here, it led to the successful development of the *standard model* in the 1960s and 70s, which explains the structure of subatomic particles and all fundamental interactions, gravitation excepted, by the existence of structureless, pointlike elementary spin 1/2 particles, called "quarks" and "leptons" along with their associated symmetries. The model is referred to as "standard", because it provides a theory of all fundamental constituents of matter as an ontological basis.

In this section, for the sake of completeness, we end our discussion of model building with a short (and very superficial) presentation of the ideas pertaining to this fundamental theory and its range of application. The interested reader is referred to specialized literature, see Sect. 2.5.5 on p. 77.

2.5.1 The Four Fundamental Interactions

According to our present knowledge, there are four known fundamental interactions, see Table 2.3 and compare Fig. 2.10:

1. weak interaction (the force related to β-decay);
2. strong interaction (nuclear force);
3. electromagnetic interaction (Coulomb force);
4. gravitational interaction.

The two nuclear forces (1) and (2) do not pertain to any human everyday experience and are very short ranged. They are treated within the framework

Table 2.3. In modern physics, it is assumed that all interactions between particles – and thus, the totality of physical reality – can be described by four fundamental interactions. Each fundamental force has a certain interaction range. In two cases, the range is infinity. The relative strength of the forces can be compared when using *natural units* such that the forces can be assigned dimensionless numbers. The gravitational interaction in these natural units is about 39 orders of magnitude weaker than the strong interaction. Table compiled from [125, 126]

	Weak	Strong	Electromagnetism	Gravitation
Range	$<< 10^{-15}\ m$	$10^{-15}\ m$	∞	∞
Example	β-decay of atomic nuclei	atomic nuclei	forces between charges	forces between astronomic objects
Strength	$G_{Fermi} = 1.02 \times 10^{-5}$	$g^2 \approx 1$	$e^2 = 1/137$	$G_{\text{Newton}} = 5.9 \times 10^{-39}$
Affected particles	quarks/ leptons	quarks	charged particles	all
Exchange particles	vector bosons $W^{\pm}$, Z^0	gluons g_i $(i = 1, ..., 8)$	photon γ	Higgs H (graviton)

of non-Abelian[24] quantum field theories which couple the special principle of relativity with quantum theory. The strong interaction, which keeps protons and neutrons in the nucleus together, has a range of 10^{-15} m. The weak interaction finally even has such a short range ($\leq 10^{-17}$ m) that it only manifests itself in certain particle collisions or decay processes. Weak interactions form the first step in the nuclear chain reaction in the interior of the sun, where two protons fuse and a deuterium nucleus, a positron[25] and a neutrino come into being. These two interactions can be neglected in the atom except within the nucleus and thus they can be completely neglected in conventional engineering applications and in most applications in physics. The so-called *first-principles*, or *ab initio calculations* (discussed in Chap. 5) only take into account electromagnetism (3) and gravitational forces (4) in the framework of *non-relativistic quantum mechanics*. Technically speaking, ab initio methods solve the Schrödinger equation to determine the electron density distribution, and the atomic structures of various materials.

Gravitation is the weakest of all fundamental interactions, see Table 2.3, but plays a dominant role on a cosmic scale, because the planets and stars in galaxies are large agglomerations of electrically neutral masses. Hence, the electromagnetic interaction between electrons and atom cores is in principle responsible for *all* chemical and physical properties of ordinary solids, fluids and gases. Compared to continuum mechanics methods, atomic scale simulations are truly ab initio. However, even in ab initio methods, there are several approximations involved in simulations of the quantum state of many electron systems, e.g. the *Born-Oppenheimer approximation*, discussed in Chap. 5.

2.5.2 The Standard Model

According to the current standard model of elementary particle physics based on quantum field theory, the fundamental ontology of the world is a set of interacting, quantized fields, which arise as two types of fields in the standard model: matter fields and interaction fields, cf. Fig. 2.11.

The quanta of matter fields, called fermions, have half-integral spins. They obey Pauli's exclusion principle which is the basis of structured matter: only a single fermion can occupy a particular quantum state. The quanta of the interaction fields, or *bosons*, have integral spins and thus, many bosons can occupy one quantum state[26]. There are 12 matter fields, organized in three generations, or families, and each has its antifield, cf. Table 2.4.

The higher generations are just replicas of the first generation with short lifetimes and show up only in high energy cosmic rays or in certain particle reactions in accelerators. All stable matter in the universe according to this

[24] For the definition of an Abelian group, see Box 2.1 on p. 65.

[25] The positron e^+ is the antiparticle of the electron e^-.

[26] For example, a coherent laser beam comprises billions of photons oscillating in a single state.

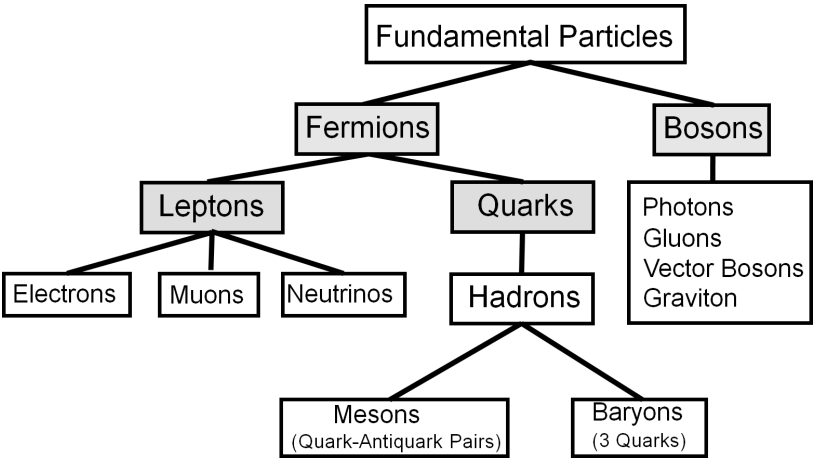


Fig. 2.11. The fundamental particles according to the standard model are the *fermions* (particles with spin 1/2) and the *bosons* (particles with integral spin or spin 0). The fermions split into the quarks, of which the hadrons – particles of the strong interaction – are made, and the leptons, which do not participate in the strong interaction. Baryons are triplets of quarks, such as the proton (uud) or the neutron (udd) and mesons consist of quark-antiquark pairs, e.g. the pion ($\bar{d}u$)

model is made up of only three matter fields of the first generation: electron (e^-), up quark (u), and down quark (d) fields. Protons (uud) and neutrons (udd) are made of quarks and are the constituents of atoms. The neutrinos interact weakly with everything and are not part of stable matter. All particles listed in Table 2.4 are elementary, i.e. they have no known substructure; this, however, does not exclude their decay into other particles, e.g. the myon μ^- may decay into an electron and two neutrinos (the electron antineutrino and the muon neutrino) according to $\mu^- \rightarrow e^- + \bar{\nu_e} + \nu_\mu$. The possible number of decay processes is restricted by empirical conservation laws of quantum mechanical quantities; some examples are listed on p. 67, in Table 2.6 and the examples on p. 68.

The interaction fields are permanently coupled to the matter fields, whose charges are their sources. For example, the electric charge is the source of the electromagnetic field, and the color charges of the quarks are the source of the strong interaction. Fundamental interactions occur only between matter and interaction fields, and they occur at a point. The mathematical form of the point coupling for the nuclear forces is the same as in the case of the electromagnetic field and can be graphically represented as Feynman diagrams, cf. Fig. 2.12. For this classical field, R.P. Feynman (1918–1988) [127], and independently J.S. Schwinger (1918–1994) [128] and S. Tomonaga[27] developed a theory, quantum electrodynamics (QED), which is consistent with quantum theory and which allowed to calculate probability amplitudes, cross-sections

[27] The three shared the Nobel prize in 1965.

Table 2.4. There are 12 known elementary particles called fermions (quarks and leptons) and 4 exchange particles according to the standard model of elementary particle physics. The particles, which are sources of fields, are listed along with their quantum numbers B, L, their charge Q, and their masses m. Each particle also has an antiparticle (not listed); ordinary stable matter is made of particles of the first generation only. The other particles show up only in high-energy experiments for very short time intervals. All exchange particles except the vector bosons are stable. Table compiled from [125, 126]

Generation			Spin	Baryon	Lepton	Charge
1	2	3	S	B	L	Q
Quarks						
u (up)	c (charm)	t (top)	1/2	1/3	0	+2/3
$m = 5\,MeV$	$m = 1.5\,GeV$	$m = 174\,GeV$				
d (down)	s (strange)	b (bottom)	1/2	1/3	0	−1/3
$m = 10\,MeV$	$m = 200\,MeV$	$m = 4.7\,GeV$				
Leptons						
ν_e	ν_μ	ν_τ	1/2	0	1	0
$m \sim 0$	$m \sim 0$	$m \sim 0$				
e^-	μ^-	τ^-	1/2	0	1	−1
$m = 0.511\,MeV$	$m = 105\,MeV$	$m = 1.7\,GeV$				
Exchange Particles						
Photon	(stable)	γ	1	0	0	0
Gluons	(stable)	$g_i\,, i = 1, ..., 8$	1	0	0	0
Vector Bosons	($\sim 10^{-25}\,s$)	$Z, W^\pm$	1	0	0	0,±1
Higgs	(stable)	H	2	0	0	0

and decay rates for the electromagnetic interaction. This theory was the only consistent, that is, renormizable quantum field theory until the t'Hooft publication in the 1970s [129]. It became the prototype of all subsequent quantum field theories. The standard model has been tested to $10^{-18}\,m$ and at present, there is no known contradiction to any experiment. Table 2.5 gives a selected overview of important discoveries in the field of elementary particle physics; several of these were later awarded the Nobel prize, e.g. the discovery of the Ω^--particle, which had been theoretically predicted with correct properties due to the $SU(3)$ symmetries of quark theory, cf. Fig. 2.13.

2.5.3 Symmetries, Fields, Particles and the Vacuum

The four fundamental forces, or interactions emerge from the exchange of particles, so-called *bosons*, which is depicted schematically in Fig. 2.12. The basic objects of this fundamental physical picture of the world are the concepts of *field*, *particle*, *vacuum* and their underlying *symmetries*. These symmetries used in the quantum field theories of the standard model build the mathematical ontology of physics.

At the beginning 20th century it was realized that classical physics is inadequate for the description of quantum structures. In 1925, non-relativistic

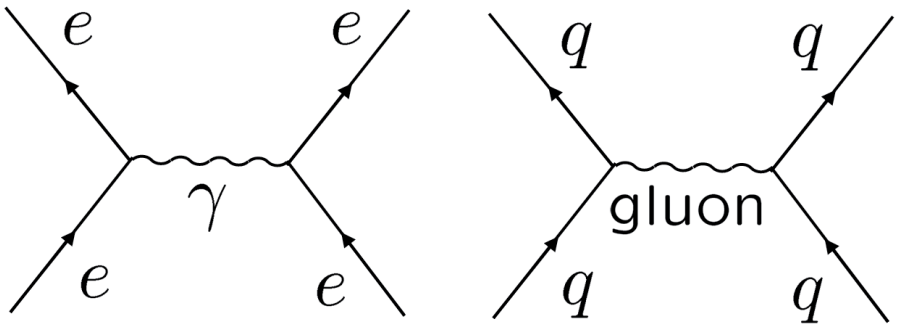


Fig. 2.12. Matter fields are represented in Feynman diagrams as straight lines and the interaction fields by wavy lines. Displayed are examples of the electromagnetic interaction between electrons (e) (**left**) and the strong interaction between quarks (q) (**right**). Feynman diagrams are a graphical way of representing the contributions to the scattering *S-matrix* elements in perturbation theory, which can be applied, if the theory is renormalizable

quantum theory was established and quickly became the basis of a large part of physics, including atomic, molecular and solid-state phenomena. However, in nuclear and high-energy physics, this theory is unsatisfactory because it is incompatible with the principle of special relativity advanced by Einstein in 1905 [154]. Dirac formulated *quantum field theory* (QFT) by uniting quantum mechanics and special relativity in 1930 [137, 138], predicting the existence of a new particle, called "anti-electron", then unknown to experimental physics, having the same mass and opposite charge of an electron. He further predicted that this new stable particle could be produced in a high vacuum by an "encounter between two hard γ-rays (of energy at least half a million volts)", leading "simultaneously to the creation of an electron and an anti-electron" [137]. Quantum field theory introduced into physics the concept of anti-particles and the concept of creation (and annihilation) of new particles from pure energy, thus changing also the scientific notion of the "vacuum"[28]. The anti-particle of the electron (e^+) was experimentally discovered in 1933 by C.D. Anderson [140].

The extension of Dirac's relativistic quantum theory to include *nuclear* interactions took another 25 years. During this process, *gauge fields*, i.e. fields with local symmetries, the idea of which first appeared in the general theory of relativity (in the form of varying orientations of local inertial frames), became dominant. Weyl tried to generalize this idea and suggested that the "scale" of

[28] In Dirac's *hole theory* a "vacuum" is interpreted as the negative energy spectrum of the solutions of his equation. The holes in the "Dirac sea" of negative energies were first interpreted by Dirac as protons [138], but this idea was quickly abandoned under the impression of several arguments put forward by W. Pauli and others.

Table 2.5. A selection of important discoveries in the history of elementary particle physics

Year	Discovery	Reference
5th century BC	4 basic elements in Greek philosophy: earth, air, fire and water	The Presocratics [69]
1789	List of 30 Elements	Lavoisier [130]
1868	Periodic Table of Elements	Mendeleev [131]
1896	Electron	Thomson [132]
1905	Special Theory of Relativity	Einstein [117]
1911	Atomic Nucleus	Rutherford [78]
1911	Nucleus and Shell Model of Atoms	Bohr [133]
1915	General Theory of Relativity	Einstein [134, 135]
1925	Quantum Theory	Heisenberg [136]
1930	Quantum Field Theory	Dirac [137, 138]
1930	Prediction of Neutrino	Pauli [15]
1932	Neutron	Chadwick [139]
1932	Positron	Anderson [140, 141]
1948/1949	Quantum Electrodynamics (QED)	Schwinger [128], Feynman [127], Tomonaga
1956	CP-Violation	Landa et al [142]
1961	Eightfold Way	Gell-Mann [143]
1964	Quark Model	Gell-Mann [144], Zweig [145, 146]
1964	Ω^--particle	Bernes et al [147], Glashow [148],
1961–1969	Electroweak Theory	Weinberg [89], Salam [149]
1972	Quantum Chromodynamics	Fritsch, Gell-Mann [150]
1973	Asymptotic Freedom of Quarks	Politzer [77], Gross, Wilzek [75, 76]
1974	J/ψ-particle	Aubert et al [151]
1974	Renormizability	t'Hooft [129]
1983	Intermediate Vector Bosons $W, Z^{\pm}$	Rubbia et al. [152]
1995	Top Quark	Abe at al. [153]

local frames should also be allowed to vary, so the frames would be enlarged or reduced as one goes about in the manifold[29]. The variation of the frame's scale would be reconciled by the electromagnetic field, just as the variation of their orientation is reconciled by the gravitational field. Weyl called this "Eichinvarianz", which was translated into English in the 1920s as "gauge invarance". Weyl's idea did not work; Einstein pointed out that the proposed

[29] For a discussion of the importance of manifolds in modern physical theory, see Chap. 3.

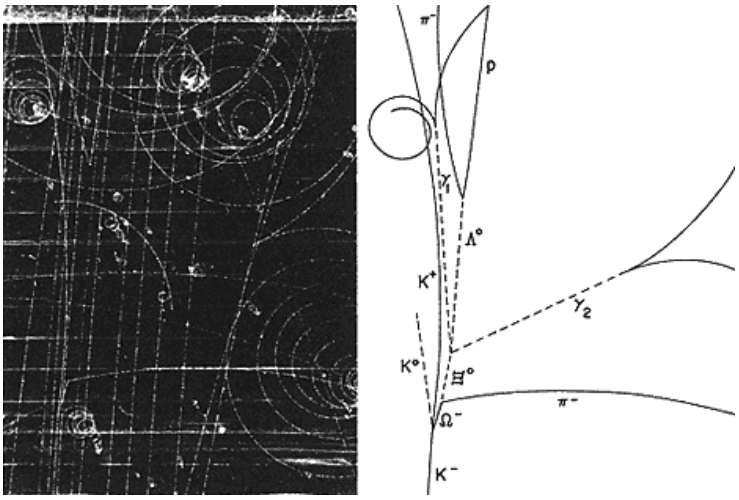


Fig. 2.13. Bubble chamber photograph (**left**) and line diagram (**right**) of event showing the first Ω^--particle. An incoming K^--meson interacts with a proton in the liquid hydrogen of the bubble chamber and produces an Ω^-, a K° and a K^+ meson which all decay into other particles. Neutral particles which produce no tracks in the chamber are shown by dashed lines. The presence and properties of the neutral particles are established by analysis of the tracks of their charged decay products and application of the laws of conservation of mass and energy. Photo courtesy of Brookhaven National Laboratory

scale change renders the rate of a clock dependent on its history, which is not acceptable. With the development of a quantum theory in mid 1920s, it was realized that what varies from point to point is not the scale but the phase of the electron wave function. However, the old names "gauge invariance", "gauge fields" and "gauge theories" are still prevalent[30].

The electromagnetic field is not self-interacting, that is, its field quanta, the photons, do not carry electric charge; thus, the photons do not stick together to form a "light ball". Mathematically, this feature can be seen in the fact that the local symmetry group of electromagnetism is commutative, or Abelian, cf. Box 2.1. In contrast, the symmetry groups of the nuclear interactions are non-commutative, or non-Abelian. The reason for this is that the field quanta of the weak and strong interactions carry coupling charges and thus interact with themselves, whereas the photon is massless and does not carry charge. This is what makes non-Abelian theories more complicated than electromagnetism.

In physical literature, the term "field" has at least two different connotations. First, a field is a continuous dynamical system, i.e. a system with an infinite number of degrees of freedom. Second, a field is also a dynamical

[30] If one were to rename gauge fields today, they would probably be called "phase fields" as the symmetry with matter fields is with respect to their phase, not to some length scale.

Box 2.1 Definition of a Group, Abelian group

A group G is a set of elements $\{g_i\}$ with a single rule of composition $\circ$ such that:

(a) G is closed, i.e. for any two elements $g_1, g_2 \in G$, $g_1 \circ g_2 \in G$,
(b) The composition is associative; i.e. $g_1 \circ (g_2 \circ g_3) = (g_1 \circ g_2) \circ g_3$,
(c) G contains an identity element e such that for all $g \in G$, $g \circ e = e \circ g = g$,
(d) For every $g \in G$ there exists an inverse element $g^{-1} \in G$ such that $g \circ g^{-1} = g^{-1} \circ g = e$.

A group G is commutative or Abelian, if $g_1 \circ g_2 = g_2 \circ g_1$ for all $g_1, g_2 \in G$.

variable characterizing such a system or at least some aspect of a system. The description of field properties is *local*, concentrating on a point entity and its infinitesimal displacement. Literally speaking, the world of fields is "full", whereas in the mechanistic world, particles are separated by empty space across which forces act instantaneously at a distance.

According to the principle of special relativity, no signal can travel at a velocity faster than the velocity c of light; thus, c determines an upper bound at which forces between particles can act. For the conservation laws of energy-momentum to be valid at every moment in time, one assumes that a particle gives rise to a field in its surrounding space which transports energy and momentum. Taking into account that energy and momentum is quantized one is led to the identification of these field quanta as particles. Thus, combining special relativity and quantum mechanics naturally leads to the concept of a field theory in which the fields are quantized themselves and are made up of "exchange particles", the field quanta.

Symmetries

SRT rests on two simple postulates: the principle of special relativity and the constancy of the speed of light. The general theory of relativity (GRT) also has two fundamental postulates: the principle of general relativity and the equivalence principle. Here, we are interested in one aspect of these principles, namely the idea of symmetry. Each principle specifies an equivalence class of coordinate systems which constrains the content of a physical theory. Whatever these constraints may be, the idea of the principles is that certain physical quantities are invariant under certain groups of coordinate transformations; stated in this way the principles are symmetry principles as they state invariance of a system against certain symmetry operations which transform the object back into itself or which leave the object unchanged or invariant. The set of symmetry transformations form a group. A group is an algebraic structure with a single rule of composition, cf. Box 2.1.

The *invariant* features of symmetries are usually the focus of interest. For example, the invariants under the Galilean group of transformations are the time interval as well as the spacial distance and one invariant under the

Box 2.2 $SU(2)$ Symmetry

A set of operators $J_i \in \mathbf{C}^{2,2}$ which obey the relation

$$[J_i, J_k] = \mathrm{i}\epsilon_{ikl} J_l \ \ (i,k,l \in \{1,2,3\}) \tag{2.31}$$

is called an $SU(2)$-algebra. A possible representation of this algebra is given, e.g. by the Pauli matrices. Setting

$$J_1 = \frac{1}{2}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, J_2 = \frac{1}{2}\begin{pmatrix} 0 & -\mathrm{i} \\ \mathrm{i} & 0 \end{pmatrix}, J_3 = \frac{1}{2}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \tag{2.32}$$

one easily shows that the commutator relations of (2.31) are fulfilled. The quantities J_j are called generators of the group. If a Hamilton operator is invariant under the transformations of a group, then it commutes with the generators, i.e. in the case of $SU(2)$ symmetry it commutes with the J_j.

$$U = \exp\left(-\mathrm{i}\sum_{j=1}^{3} \phi_j J_j\right) \tag{2.33}$$

are the special unitary transformations in 2D where ϕ_j is a field variable.

Lorentz group is the spacetime interval $d\tau^2 = c^2dt^2 - d\vec{x}^2$. In mathematics, in general, the profoundity of a concept is associated with its generality; thus, the largest transformation group defines the most fundamental concept. A system, characterized by a large symmetry group retains only the important features. In elementary particle physics, one seeks ever larger symmetry groups in the attempt to ever larger unification. For example, the electromagnetic interaction is characterized by a unitary group of order 1, $U(1)$, the weak interaction by the special unitary group $SU(2)$, see Box 2.2, and their unification in the electroweak interaction is achieved by the larger group $SU(2) \times U(1)$; thus, the standard model is based on the $U(1) \times SU(2) \times SU(3)$ symmetry. In QED, demanding $U(1)$ gauge invariance means that the theory is supposed to be invariant under the transformation $\psi(x) \rightarrow \exp(\mathrm{i}\alpha(x))\psi(x)$, where $\exp(\mathrm{i}\alpha(x))$ is a *local* phase transformation, whereas the transformation is *global* if $\alpha(x) = \text{const}$. For more details on groups and their properties, the reader is referred to more specialized literature, cf. Sect. 2.5.5.

The symmetry $SU(3)$ underlies the strong interaction which treats u, d, and s quarks as equivalent, see Box 2.3. If the symmetry group of a physical system is reduced to one of its subgroups (for example, the $SU(2)$ isospin symmetry is a subgroup of $SU(3)$), one says that the *symmetry is broken*. As a result of a broken symmetry, one obtains more invariants and more features.

Conservation Laws

Another important group of general principles are *conservation laws* which are a consequence of underlying symmetries in a system, usually expressed

Box 2.3 SU(3) Symmetry

Let U be an unitary $n \times n$ matrix, i.e. $U \in U(n)$, $U^\dagger U = 1$ and H an hermitean $n \times n$ matrix. Then U can be written as

$$U = \exp(\mathrm{i}H) \; . \tag{2.34}$$

As H is hermitean, there are n^2 independent, real parameters for H and U. With $det(U) = 1$, the matrices U build the special unitary group $SU(n)$, which, due to (2.34), depends on $n^2 - 1$ parameters λ_ν, which are called *generators*. It can be shown that

$$det(U) = det(\exp(\mathrm{i}H)) = \exp(\mathrm{i}TrH) \; . \tag{2.35}$$

It is

$$U = \exp(-\mathrm{i}\alpha_\nu \lambda_\nu) \; , \tag{2.36}$$

$$H = -\mathrm{i}\alpha_\nu \lambda_\nu \; . \tag{2.37}$$

and

$$Tr(\lambda_\nu) \; . \tag{2.38}$$

Hence, the $n^2 - 1 = 3^2 - 1 = 8$ traceless generators of $SU(n)$ have to be traceless matrices. It is common to introduce the following convention for the generators: $J_j = \frac{1}{2}\lambda_j$

$$\lambda_1 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \; , \quad \lambda_2 = \begin{pmatrix} 0 & -\mathrm{i} & 0 \\ \mathrm{i} & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \; , \quad \lambda_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \; , \tag{2.39a}$$

$$\lambda_4 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \; , \quad \lambda_5 = \begin{pmatrix} 0 & 0 & -\mathrm{i} \\ 0 & 0 & 0 \\ \mathrm{i} & 0 & 0 \end{pmatrix} \; , \quad \lambda_6 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} , \tag{2.39b}$$

$$\lambda_7 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -\mathrm{i} \\ 0 & \mathrm{i} & 0 \end{pmatrix} \; , \quad \lambda_8 = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -2 \end{pmatrix} , \tag{2.39c}$$

The corresponding representation of the group is $\exp(-\frac{\mathrm{i}}{2}\sum_j \phi_i \lambda_j)$. One obtains different representations by finding eight different generators, which obey the same commutator relations as the J_i, i.e. $[J_i, J_k] = \mathrm{i} f_{ikl} J_l$ with the antisymmetric quantities f_{ikl}. Instead of J_3 and J_8, the corresponding quantum numbers in physics are called *isospin* and *hypercharge*.

in the Hamilton operator $\mathcal{H}$ (in quantum systems) or the Hamilton function H for classical systems. For example, if $\frac{\partial H}{\partial t} = 0$, that is if H is time-independent, energy is conserved and $H = E = \text{const}$. Homogeneity of space (H is invariant against translations) leads to conservation of momentum, and isotropy of space leads to conservation of angular momentum. In the following, some important conservation laws are listed.

- Conservation of energy-momentum
- Conservation of angular momentum

- Conservation of baryon number B (nucleons and hyperons[31] are assigned $+1$ and their anti-particles are assigned -1)
- Conservation of lepton number L (ν_{e^-} and e^- are assigned $+1$ and their anti-particles get -1)
- Conservation of isospin I_z (only valid for the strong interaction)
- Conservation of strangeness S (valid for the strong and electromagnetic interaction but not for the weak interaction)
- Conservation of parity P (valid for the strong and electromagnetic interaction but not for the weak interaction)

The general connection between symmetries and conservation laws is provided by two theorems published by Emmy Noether in the article "Invariante Variationsprobleme" in 1918 [155]. Table 2.6 shows a selection of important conserved quantities and the – according to Noether's theorem – associated symmetries.

Example 3 (Some Particle Decays).

- Myon decay:
  The process $\mu^- \rightarrow e^- + \bar{\nu_e} + \nu_\mu$ is allowed. It obeys conservation of myon lepton number $1 \rightarrow 0+0+1$ and electron lepton number $0 \rightarrow 1+(-1)+0$.
  The process $\mu^- \rightarrow e^- + \bar{\nu_e}$ violates conservation of L.
- Proton decay:
  The process $p \rightarrow e^+ \gamma$ violates conservation of B.
- Electron decay:
  The process $e^- \rightarrow \mu^- + \nu_e + \bar{\nu_\mu}$ is forbidden, as $m_e < m_\mu + m_{\nu_e} + m_{\mu_\nu}$.
- Neutron decay: The process $n \rightarrow p + e^- + \bar{\nu}_e$ obeys the conservation of electron lepton number: $0 \rightarrow 0 + 1 + (-1)$.

Exercise 1. Which one of the four neutrinos ($\nu_e, \bar{\nu}_e, \nu_\mu, \bar{\nu}_\mu$) cf. Table 2.4 on p. 61 is involved in the following reactions?

Table 2.6. Some conserved quantities in elementary particle physics and their associated symmetries due to Noether's theorem

Conserved Quantity	Symmetry
Four-momentum $p^\mu = (E_{tot}/c^2, \vec{p}_{tot})$	Spacetime translation
Electric charge Q	Gauge invariance
Baryon B and lepton L numbers	$SU(1)_B$, $SU(1)_L$
Only electromagnetic interaction	
Parity P	Reversion of spacial configuration
Time reversal T	Direction of time
Charge conjugation C	Matter $\leftrightarrow$ Antimatter

[31] Hyperons are baryons with a strangeness quantum number

(a) $(?) + p \rightarrow n + e^+$,
(b) $(?) + n \rightarrow p + \mu^-$,
(c) $(?) + n \rightarrow p + e^-$.

Solution 1. Considering different conservation laws stated above, the answers are:
(a) $\bar{\nu}_e$ (b) ν_μ (c) ν_e.

2.5.4 Relativistic Wave Equations

The canonical starting point for quantum field theories is the variation of the classical integral of action

$$S \equiv \int_{t_1}^{t_2} dt \int d^3x \, \mathcal{L}(\phi, \partial_\mu \phi) \, , \tag{2.40}$$

with *Lagrangian density* $\mathcal{L}(\phi(x), \partial_\mu \phi(x))$, which is a function of the field $\phi(x)$ with an infinite number of degrees of freedom, and its four-gradient $\partial_\mu \phi(x) \equiv \partial/\partial x_\mu = (\partial/\partial t, -\vec{\nabla})$. Performing a variation δ of the integral in (2.40) and applying Hamilton's principle ($\delta S = 0$) as well as the boundary conditions ($\delta\phi(t_1, \vec{x}) = 0 = \delta\phi(t_2, \vec{x})$) of (2.40), one obtains the equation of motion for the fields, i.e. the *Euler-Lagrange equations*:

$$\partial_\mu \frac{\partial \mathcal{L}}{\partial(\partial_\mu \phi(x))} - \frac{\partial \mathcal{L}}{\partial \phi(x)} = 0 \, . \tag{2.41}$$

The field equations (2.41) are covariant, if the Lagrangian density is Lorentz invariant, i.e. a Lorentz-scalar. Due to the boundary conditions of the variation, they are not changed, when a total divergence $\partial_\mu A$ with some field A is added to the Lagrangian density. The symmetries of the field equations and the quantum numbers of the elementary particles are obtained from the symmetries of the Lagrangian density. The same formalism is used in classical mechanics for deriving equations of motion from a classical Lagrange function $L(q^i, \dot{q}^i)$, which is a function of generalized coordinates and velocities, see Box 2.4.

Example 4 (Lagrangian Density of Maxwell's Equations). Introducing the four-potential

$$A^\mu = (\mathbf{\Phi}, \vec{A}) \tag{2.45}$$

in Minkowski space and the antisymmetric field tensor.

$$F^{\mu\nu} \equiv \partial^\nu A^\mu - \partial^\mu A^\nu = -F^{\nu\mu} \, , \tag{2.46}$$

with

$$F^{\mu\nu} = \begin{pmatrix} 0 & E_1 & E_2 & E_3 \\ -E_1 & 0 & B_3 & -B_2 \\ -E_2 & -B_3 & 0 & B_1 \\ -E_3 & B_2 & -B_1 & 0 \end{pmatrix} \, , \tag{2.47}$$

Box 2.4 Lagrangian formulation of Classical Mechanics

The classical equations of motion are obtained by applying Hamilton's principle of least action. A variation δ of the integral of least action

$$S \equiv \int_{t_1}^{t_2} dt\, L(q^i, \dot{q}^i) \,, \tag{2.42}$$

where $L(q^i, \dot{q}^i) = T - V$, the difference between kinetic and potential energy, depends on the generalized coordinates and velocities. Using

$$\delta S = \delta \int_{t_1}^{t_2} dt\, L(q^i, \dot{q}^i) = 0 \,, \tag{2.43}$$

with fixed boundaries t_1 and t_2 of the integral in (2.43) yields the particle trajectories $q^i(t)$ for which the action S is minimized, i.e. stationary. These equations are called *Euler-Lagrange Equations* (see. Prob 3 on p. 106).

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}^i}\right) - \frac{\partial L}{\partial q^i} = 0 \,. \tag{2.44}$$

where a short-form ∂^ν of the four-gradient was used. The co- and contravariant four-gradients and the operator of the wave equation $\Box$ are defined as:

$$\partial_\nu \equiv \frac{\partial}{\partial x^\nu} = \left(\frac{1}{c}\frac{\partial}{\partial t}, \nabla\right) \,, \tag{2.48a}$$

$$\partial^\nu \equiv \frac{\partial}{\partial x_\nu} = \left(\frac{1}{c}\frac{\partial}{\partial t}, -\nabla\right) \,, \tag{2.48b}$$

$$\Box = \partial_\nu \partial^\nu = \frac{1}{c^2}\frac{\partial^2}{\partial t^2} - \triangle \,. \tag{2.48c}$$

Maxwell's equations (2.3)a–d can be derived from the Lagrangian density

$$\mathcal{L} = -\frac{1}{4} F^{\mu\nu} F_{\mu\nu} - j_\mu A^\mu \tag{2.49}$$

in covariant form as

$$\partial^\lambda F^{\mu\nu} + \partial^\mu F^{\nu\lambda} + \partial^\nu F^{\lambda\mu} = \partial^{[\lambda} F^{\mu\nu]} = 0 \,, \tag{2.50a}$$

$$\partial_\mu F^{\mu\nu} = 0 \,, \tag{2.50b}$$

$$\partial_\mu F^{\mu\nu} = -j^\nu \,. \tag{2.50c}$$

In (2.50)a we have used the notation of antisymmetry brackets "[]" which are used to denote antisymmetric components of tensors[32]. Using (2.50)a and (2.50)c, one obtains the covariant form of the continuity equation

[32] For example, $T^{ab}{}_{[cd]} = \frac{1}{2}(T^{ab}{}_{cd} - T^{ab}{}_{dc})$.

$$\partial_\mu j^\nu = -\partial_\nu \partial_\mu F^{\mu\nu} = 0\,, \tag{2.51}$$

with the source four-vector

$$j^{\mu(x)} = (\rho(x), \vec{j}(x))\ . \tag{2.52}$$

Maxwell's equations do not change under the transformation

$$A_\mu(x) \to A_{\mu'}(x) = A_\mu(x) + \partial_\mu \mathbf{\Phi}(x)\ , \tag{2.53}$$

where $\mathbf{\Phi}(x)$ is some scalar field. Due to this gauge invariance, one can impose the Lorentz gauge condition

$$\partial_\mu A^\mu = 0\ , \tag{2.54}$$

which does not change the fields, i.e. the physics.

In Quantum field theories, both, the mass fields (fermions) and the interaction fields (bosons) are described as operators in Hilbert space which obey certain commutation relations and manipulate the fields. The Schrödinger equation reads

$$\mathcal{H}\psi = \mathrm{i}\hbar \frac{\partial \psi}{\partial t}\ , \tag{2.55}$$

and it describes non-relativistic spinless point particles, with the energy operator $\mathcal{H} = \mathcal{H}(\mathcal{P}, x) = \mathcal{H}(\frac{\hbar}{\mathrm{i}}\vec{\nabla}, x)$. In non-relativistic quantum mechanics $\mathcal{H} = \mathcal{T} + V$ where $\mathcal{T} = \frac{\mathcal{P}^2}{2m}$ is the operator corresponding to non-relativistic kinetic energy T and momentum p, respectively, i.e. $v \ll c$, and V is the potential energy. In (2.55) ψ is the wave function describing the single particle amplitude. Usually, in quantum field theory, symbol ψ is reserved for spin 1/2 fermions and the symbol ϕ is used for spin 0 bosons. For relativistic particles ($v \sim c$), the total energy E is given by the Einstein relation $E^2 = \mathcal{P}^2 + m^2$. Thus, the square of the relativistic Hamiltonian H^2 is simply given by promoting the momentum to operator status, i.e.:

$$H^2 \to \mathcal{H}^2 = \mathcal{P}^2 c^2 + m^2 c^4\ . \tag{2.56}$$

Working with this operator in (2.55) and inserting the momentum operator in position space $\mathcal{P} \to -\frac{\mathrm{i}}{\hbar}\vec{\nabla}$, one yields the *Klein-Gordon equation*

$$\left(\Box + \left(\frac{mc}{\hbar}\right)^2\right)\phi(x) = 0\ , \tag{2.57}$$

In (2.57) the box notation was introduced, cf. (2.48) on p. 70:

$$\Box = \partial_\mu \partial^\mu = \partial^2/\partial t^2 - \vec{\nabla}\ , \tag{2.58}$$

Equation (2.57) is the classical homogeneous wave equation for the field $\phi(x)$. The operator $\Box$ is Lorentz invariant, so the Klein-Gordon equation is relativistically covariant, that is, it transforms into an equation of the same form,

provided that ϕ is a scalar function. Thus, under a Lorentz transformation $(ct, \vec{x}) \to (ct', \vec{x}')$,

$$\phi(t, \vec{x}) \to \phi'(t', \vec{x}') \; . \tag{2.59}$$

The Klein-Gordon equation has plane wave solutions of the form:

$$\phi(x) = N \mathrm{e}^{-\mathrm{i}(Et - \vec{p}\vec{x})} \; , \tag{2.60}$$

where N is a normalization constant and $E = \pm\sqrt{c^2\vec{p}^2 + m^2c^4}$ with positive *and* negative energy solutions. The negative solutions for E render it impossible to interpret the Lorentz invariant ϕ as a wave function of a particle (as in non-relativistic quantum theory), because $|\phi|^2$ does not transform like a density. The spectrum of the energy operator is not bounded and one could extract arbitrarily large amounts of energy from the system by driving it into ever more negative energy states. Also, one cannot simply throw away these solutions because they are needed to define a *complete* set of states. These interpretive problems disappear if one introduces the idea of a quantized field and considers ϕ as a a quantum field in the sense of a usual dynamic variable. In this case, the positive and negative energy modes are simply associated with operators that create or destroy particles.

Historically, due to the above mentioned problems in interpreting ϕ as a wave function and to define a probability density (see Problem 4), Dirac tried to find a different equation of first order with respect to time derivatives, hoping that this similarity to the non-relativistic Schrödinger equation would allow such an interpretation. It turned out that Dirac's hopes were in vain, but he did find another covariant equation which allowed for negative solutions, too. His Ansatz was a Hamiltonian of the form

$$\mathcal{H} = \sum_i^3 \alpha_i \mathcal{P}_i + \beta mc^2 \; , \tag{2.61}$$

where $\mathcal{P}_i$ are the three components of the momentum operator $\mathcal{P} = \frac{\hbar}{\mathrm{i}}\vec{\nabla}$. It can be shown that from the requirement $H^2 = \mathcal{H}^2 + m^2c^4$ it follows that α_i and β must be interpreted as 4×4 matrices, and the considered field ψ as a multi-component *spinor* ψ_σ on which these matrices act, thus yielding the *position space Dirac equation*

$$\mathrm{i}\hbar\frac{\partial\psi_\sigma}{\partial t} = -\mathrm{i}\hbar c \sum_{\tau=1}^{N} \left(\sum_{i=1}^{3} \alpha_i \partial_{x^i} + \sum_{\tau=1}^{N} \beta_{\sigma\tau} mc^2 \right) \psi_\tau \tag{2.62}$$

$$= \sum_{\tau=1}^{N} \mathcal{H}_{\sigma\tau}\psi_\tau \; . \tag{2.63}$$

The following combination of matrices is useful for a symmetric formulation of (2.62) in spacetime:

$$\gamma^0 = \beta \, , \gamma^i = \beta\alpha_i \; , (i = 1, 2, 3) \; . \tag{2.64}$$

The quantities γ^μ may be combined to define a four-vector in Minkowski space

$$\gamma^\mu = (\gamma^0, \gamma^1, \gamma^2, \gamma^3)\,, \text{ and } \gamma_\mu = g_{\mu\nu}\gamma^\nu = (\gamma_0, \gamma_1, \gamma_2, \gamma_3)\,, \tag{2.65}$$

thus yielding the *covariant Dirac equation*:

$$(\mathrm{i}\hbar \sum_{i=0}^{4} \gamma^i \partial_i - mc)\psi = (\mathrm{i}\hbar \not{\partial} - mc)\psi = 0\,, \tag{2.66}$$

where in the second term Feynman's "dagger-notation" was used, i.e.:

$$\not{\partial} = \not{\vec{\nabla}} = \sum_{\mu=0}^{4} \gamma^\mu \frac{\partial}{\partial x^\mu} = \frac{\gamma^0}{c}\frac{\partial}{\partial t} + \vec{\gamma}\vec{\nabla}\,. \tag{2.67}$$

The Dirac equation is the equation of motion for the field operator ψ describing spin 1/2 fermions. From (2.67) one can derive a Lorentz-invariant Langrange density for free Dirac particles:

$$\mathcal{L} = \bar{\psi}(\mathrm{i}\hbar \sum_{i=0}^{4} \gamma^i - mc)\psi \tag{2.68}$$

where $\bar{\psi}$ is the adjungated field. Variation $\delta\bar{\psi}$ in (2.68) yields (2.62) and variation $\delta\psi$ yields the adjungated Dirac equation.

Local Gauge Symmetries

Quantum mechanical expectation values of observables

$$\langle O \rangle = \int \psi^\star\, O\, \psi \tag{2.69}$$

and the corresponding Lagrange functions are invariant against phase rotations of the field function $\psi(x)$:

$$\psi(x) \to \psi'(x) = \exp(\mathrm{i}\alpha)\psi(x)\,. \tag{2.70}$$

If α is a constant angle, then the $U(1)$ gauge invariance (2.70) is *global* and, according to Noether's theorem, leads to conservation of electric charge Q.

Exercise 2 (Show that $SU(1)$ symmetry leads to conservation of charge Q).

The global $SU(1)$ symmetry rotates the phase of a field ϕ according to:

$$\phi(x) \to \phi'(x) = \exp(\mathrm{i}Q\alpha)\phi(x)\,, \tag{2.71a}$$

$$\phi^\star(x) \to \phi^{\star\prime}(x) = \exp(-\mathrm{i}Q\alpha)\phi^\star(x)\,. \tag{2.71b}$$

The infinitesimal rotation of the fields is thus:

$$\phi(x) \to \phi'(x) = \phi(x) + \delta\phi(x) = \phi(x) + \mathrm{i}Q(\delta\alpha)\phi(x) \;, \tag{2.72a}$$

$$\phi^\star(x) \to \phi^{\star\prime}(x) = \phi^\star(x) + \delta\phi^\star(x) = \phi^\star(x) - \mathrm{i}Q(\delta\alpha)\phi^\star(x) \,. \tag{2.72b}$$

As $\delta\alpha$ is *not* a function of position $\delta(\partial_\mu\phi) = \mathrm{i}Q(\delta\alpha)\partial_\mu\phi$. Demanding gauge invariance of the Lagrange density $\mathcal{L}(\phi, \phi^\star, \partial_\mu\phi, \partial_\mu\phi^\star)$ and using (2.41) one obtains for arbitrary $\delta\alpha$:

$$\delta\mathcal{L} = \frac{\partial\mathcal{L}}{\partial\phi}\delta\phi + \frac{\partial\mathcal{L}}{\partial(\partial_\mu\phi)}\delta(\partial_\mu\phi) + \frac{\partial\mathcal{L}}{\partial\phi^\star}\delta\phi^\star + \frac{\partial\mathcal{L}}{\partial(\partial_\mu\phi^\star)}\delta(\partial_\mu\phi^\star) \tag{2.73a}$$

$$= \left[\partial_\mu \frac{\partial\mathcal{L}}{\partial(\partial_\mu\phi)}\right] \mathrm{i}Q(\delta\alpha)\phi + \frac{\partial\mathcal{L}}{\partial(\partial_\mu\phi)}\,\mathrm{i}Q(\delta\alpha)\partial_\mu\phi + c.c \,. \tag{2.73b}$$

$$= \mathrm{i}Q(\delta\alpha)\partial_\mu\left[\frac{\partial\mathcal{L}}{\partial(\partial_\mu\phi)}\phi\right] - \mathrm{i}Q(\delta\alpha)\partial_\mu\left[\frac{\partial\mathcal{L}}{\partial(\partial_\mu\phi^\star)}\phi^\star\right] \equiv 0. \tag{2.73c}$$

Hence, there is a continuity equation for the four-current $j^\mu = (\rho, \vec{j})$ of charge $Q = \int d^3x\rho$:

$$\partial_\mu j^\mu = \frac{\partial}{\partial t}\rho + \vec{\nabla}\cdot\vec{j} = 0 \,, \tag{2.74}$$

with

$$j^\mu \equiv -\mathrm{i}Q\left(\frac{\partial\mathcal{L}}{\partial(\partial_\mu\phi)}\phi - \frac{\partial\mathcal{L}}{\partial(\partial_\mu\phi^\star)}\phi^\star\right) \,. \tag{2.75}$$

Hence, in a closed system, charge Q is conserved ($\frac{dQ}{dt} = 0$). Global $U(1)$ gauge symmetries lead to conserved quantum numbers (in this case, charge Q).

If α is a function of space, i.e. $\alpha = \alpha(x)$, (2.70) is a *local* gauge transformation, i.e. one demands that the expectation values (2.69) are invariant against a *local* choice of phase factors, cf. Fig. 2.14. A local, position dependent phase transformation

$$\phi(x) \to \phi'(x) = \exp\left(\mathrm{i}Q\alpha(x)\right)\phi(x) \tag{2.76}$$

yields

$$\partial_\mu\phi(x) \to \partial_\mu\phi'(x) = \exp(\mathrm{i}Q\alpha(x))\left[\partial_\mu\phi(x) + \mathrm{i}Q(\partial_\mu\alpha(x))\phi(x)\right] \;, \tag{2.77}$$

which is not forminvariant. The Lagrange density can be made covariant by substituting the partial derivative by a gauge invariant derivative:

$$\partial_\mu \to D_\mu \equiv \partial_\mu + \mathrm{i}eQA_\mu(x), \tag{2.78}$$

where $q = eQ$ is the electric charge of the field $\psi(x)$ (e: electric charge, Q: quantum number), and A_μ is the four-potential (2.45) which transforms under phase rotations as

$$A_\mu(x) \to A'_\mu(x) = A_\mu(x) - \frac{1}{e}\partial_\mu\alpha(x) \,. \tag{2.79}$$

With this gauge transformation one obtains:

$$
\begin{aligned}
D_\mu\phi(x) &= \partial_\mu + \mathrm{i}eQA_\mu \rightarrow D'_\mu\phi'(x) && \text{(2.80a)}\\
&= (\partial_\mu + \mathrm{i}eQA'_\mu(x))\exp(\mathrm{i}Q\alpha(x))\phi && \text{(2.80b)}\\
&= \exp(\mathrm{i}Q\alpha(x))[\partial_\mu + \mathrm{i}Q\partial_\mu\alpha(x) + \mathrm{i}eQA_\mu(x) && \text{(2.80c)}\\
&- \mathrm{i}Q\partial_\mu\alpha(x)]\phi && \text{(2.80d)}\\
&= \exp(\mathrm{i}Q\alpha(x))\left[\partial_\mu + \mathrm{i}Qe\partial_\mu A_\mu(x)\right]\phi(x) && \text{(2.80e)}\\
&\equiv \exp(\mathrm{i}Q\alpha(x))\, D_\mu\,\phi(x)\ . && \text{(2.80f)}
\end{aligned}
$$

Hence, the invariance of $\phi^\star\, D_\mu\,\phi$ under $U(1)$ phase transformations is achieved by introducing an interaction for the field ϕ, cf. Fig. 2.14.

The coupling of the matter field ϕ and the gauge interaction field $A_\mu(x)$ is uniquely determined by demanding *local gauge invariance*, i.e. by introducing the gauge invariant derivation

$$D_\mu\phi(x) = \partial_\mu\phi(x) + \mathrm{i}eQA_\mu x\phi(x)\ , \tag{2.81}$$

also called *minimal gauge invariant coupling*. With the above derivations using $U(1)$ gauge invariance, one obtains for the Lagrangian density of QED, which

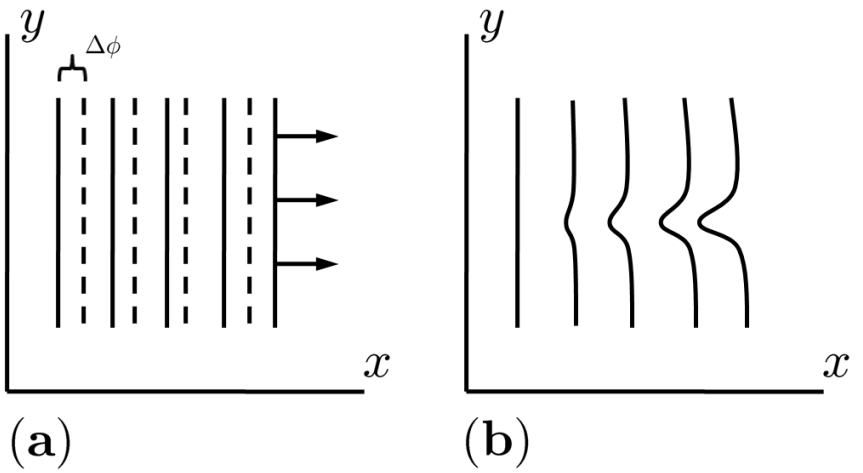


Fig. 2.14. Illustration of a global gauge symmetry vs. local gauge symmetry with plane waves propagating in x direction. In (**a**) a global gauge transformation changes the phase of the plane wave at every position (x, y) by the same amount $\Delta\phi$. Thus, the plane wave fronts move to the dotted lines and there is on average no effect on the shape of the wave propagation. In (**b**) a local phase transformation changes the wave fronts differently at different locations (x, y), i.e. $\Delta\phi = \Delta\phi(x, y)$. The transformed wave is no plane wave any more. This change of shape is explained by the introduction of external interactions

describes the interaction of photons (the electromagnetic interaction field) with fermions:

$$\mathcal{L}_{QED} = \mathcal{L}^{free}_{fermion} + \mathcal{L}^{free}_{photon} + \mathcal{L}_{interaction} \tag{2.82a}$$

$$= -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + \bar{\psi}(\mathrm{i}\hbar\sum_{i=0}^{4}\gamma^i - mc)\psi - ej^\mu A_\mu \tag{2.82b}$$

$$= -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + \bar{\psi}\mathrm{i}\hbar\sum_{i=0}^{4}\gamma^i\partial_\mu\psi - m\bar{\psi}\psi - ej^\mu A_\mu \tag{2.82c}$$

$$= \mathcal{L}^{kin}_{fermion} + \mathcal{L}^{kin}_{photon} + \mathcal{L}^{mass}_{fermion} + \mathcal{L}_{interaction} = T - V\ . \tag{2.82d}$$

For a detailed discussion of non-Abelian $SU(2)$ and $SU(3)$ gauge symmetries within the $U(1) \times SU(2) \times SU(3)$ gauge symmetry of the standard model – usually discussed in the language of differential geometry (local sections of principal fiber bundles) – the interested reader is referred to specialized literature, see Sect. 2.5.5.

The standard model is one of the best tested theories in physics. However, it too, can only be an approximation of a more general theory, due to several problems such as:

- There are more than 30 free parameters (e.g. particle masses and constants of nature) that have to be determined by measurements.
- Why does dark matter exist and why is there more normal matter than antimatter in the known universe?
- The gravitational force is completely excluded.

We end this section with an appropriate quote by Sheldon L. Glashow (Nobel prize 1979), who writes in "The Charm of Physics" about the motivation to do fundamental physics [148]:

> " "Do we do fundamental physics to explain the world about us?" is a question that is often asked. The answer is NO! The world about us was explained 50 years ago or so. Since then, we have understood why the sky is blue and why copper is red. That's elementary quantum mechanics. It's too late to explain how the work-a-day world works. It's been done. The leftovers are things like neutrinos, muons, and K-mesons – things that have been known for half a century, still have no practical application, and probably never will [...] So it is that we are not trying to invent a new toothpaste. What we are trying to do is to understand the birth, evolution, and fate of our universe. We are trying to know why things must be exactly the way they are. We are trying to expose the ultimate simplicity of nature. For it is in the nature of elementary-particle physicists (and some others) to have faith in simplicity, to believe against all reason that the fundamental laws of physics, of nature or of reality are in fact quite simple and comprehensible. So far, this faith has been extraordinarily productive: Those who have it often succeed; those without it, always fail." (Sheldon L. Glashow, 1991, p. 109)

2.5.5 Suggested Reading

For a recent description of latest accelerator experiments at DESY, PETRA and HERA, see e.g. [156]. Elementary introductions into the standard model can be found in Close [157], Halzen and Martin [158], or Nachtmann et al. [159]. Standard references to relativistic quantum theory and gauge theories are Aitchison and Hey [160], Sakurai [161], or Mandl and Shaw [162]. A classic is Bjorken and Drell [163, 164] which is very succinct and quickly advances from chapter to chapter. An excellent book covering the history of elementary particle physics from the 1960s to 1970s is edited by Hoddeson et al. [95]. A very paedagogical introduction to ideas of elementary particle physics is achieved in the classic by Dodd [165]. Some good popular books on elementary particle physics and the principles of field theory are Okun [166] and Fritsch [167]. Lattice Gauge Theory, which was developed in a ground breaking work by M. Creutz [168], in order to be able to perform Monte Carlo simulations of the basic equations of quantum chromodynamics on a lattice, is discussed in Montvay and Münster [169].

2.6 Computer Science, Algorithms, Computability and Turing Machines

Computer science provides a multitude of concepts, methods of description, models, algorithms, or simply ideas which serve to the general purpose of visualizing, organizing and analyzing complex phenomena of reality. In principle, the modeling strategies are basically the same as in the natural sciences. From identifying the most important contents of a real system one derives an abstract model (abstraction). The model might be a formula, an equation, an algorithm, an automata, a graph, etc., cf. Fig. 2.15.

Creating a model of the real complex system allows for saving the model in binary form on a computer with subsequent simulation and analysis based on some algorithm. From the input/output properties of the model system one can make predictions for the behavior of the real system (interpretation). The most important point in this process is the identification of the essential features that characterize the real system in a unique way. Oversimplification is a common weakness of this modeling process.

Computer science is dominated by algorithms. What is an algorithm?

An intuitive notion of an algorithm is a step-by-step procedure (a *finite* set of well-defined instructions) for accomplishing some task which, given an initial state, will terminate in a defined end-state, cf. Fig. 2.16.

An algorithm allows for "mechanically" accomplishing some task by following the step-by-step instructions, even without any intellectual insight into the procedure or the problem to be solved. The computational complexity and efficient implementation of algorithms are very important in scientific computing, and this usually depends on suitable data structures. Algorithms have

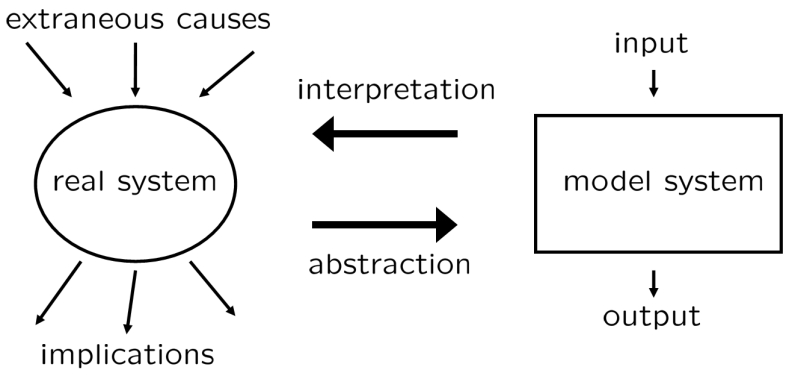


Fig. 2.15. Modeling of a real complex structure in computer science. In computer simulations, one utilizes numerical models. These models can come in many shapes, sizes, and styles. It is important to emphasize that such a model is not the real world but merely a human construct to help one better understand real world systems. In general, all models have an information input, an information processor, and an output of expected results. In an abstraction process the respective model system is extracted from the essential parts of the real system. From the interpretation of the input/output behavior of the model system one draws conclusions on the behavior of the real system

to be written down in a way, such that each step is comprehensible. Hence, the single basic steps have to be a subset of an agreed upon set of *elementary operations*. Algorithms that are implemented on computers are formulated in suitable computer languages and are subsequently translated (compiled) into a machine code which actually consists only of elementary single steps. An algorithm can be written down in several ways, e.g. as written text with *step-by-step directions* as in Algorithm 2.1, as *pseudo-code* revealing the elements of a computer language that have to be used when implementing the algorithm, cf. Algorithm 2.2, or as *explicit* code sample in a computer language, cf. Algorithm 2.3.

We start out with some examples of algorithms, some of which have been known for a long time. The *sieve of Eratosthenes* (see e.g. [170]) is one of the

Fig. 2.16. An algorithm is a set of instructions which – applied to an initial state – develops the system into a defined end-state

Algorithm 2.1 The sieve of Eratosthenes

This is one of the oldest known algorithms which is used for finding all prime numbers up to a specified integer:

1. Write a list (called A) of numbers from 2 to the largest number you want to test for primality.
2. Write the number 2, the first prime number, in another list for primes found. Call this list B.
3. Strike off 2 and all multiples of 2 from list A.
4. The first remaining number in the list is a prime number. Write this number into list B.
5. Strike off this number and all multiples of this number from list A. The crossing-off of multiples can be started at the square of the number, as lower multiples have already been crossed out in previous steps.
6. Repeat steps 4 through 6 until no more numbers are left in list A.

oldest known algorithms. It it used for determining all prime numbers up to some specified integer.

Example 5 (Application of Algorithm 2.1). We want to apply this algorithm to the first 17 integers.

Steps 1 and 2:
list $A = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17\}$, list $B = \{2\}$.
Step 3:
$\underline{2}, 3, \underline{4}, 5, \underline{6}, 7, \underline{8}, 9, \underline{10}, 11, \underline{12}, 13, \underline{14}, 15, \underline{16}, 17$.
Thus $A = \{3, 5, 7, 9, 11, 13, 15, 17\}$.
Step 4:
$B = \{2, 3\}$.
Step 5:
$A = \{\underline{3}, 5, 7, \underline{9}, 11, 13, \underline{15}, 17\} = \{5, 7, 11, 13, 17\}$.
Step 6:
Repeating Steps 4 to 6 finally yields: $B = \{2, 3, 5, 7, 11, 13, 17\}$.

A different example of a well-known algorithm is Euclid's algorithm which determines the largest common divisor of two natural numbers a and b, see the pseudocode in Algorithm 2.2.

After input of the numbers a and b it is tested whether $b = 0$. If this is not true then a is set to the value of b and the new value of b is the rest of

Algorithm 2.2 The largest common divisor (Euclid's algorithm)

```
Read the natural numbers a and b
while b is not 0 do
  Substitute (a,b) by (b, a MOD b)
end return a
```

Algorithm 2.3 Rest-recursive function call

```
long Faculty(long n, long y)
{
  if (n < 0)  return (0);
  if (n == 0) return (1);
  if (n == 1) return (y);
  return ( Faculty(n-1, y * (n-1)) );
}
```

the division of (the previous) a value by b. This is repeated until the rest of the division is 0. Then the searched for result is a.

Sometimes an algorithm which is derived directly from a mathematical definition or a formula is not necessarily the most efficient one. An example of this fact is provided on p. 82 in Example 6.

2.6.1 Recursion

An important principle in computer science for the description of algorithms, functions, or data structures is *recursion*. The basic principle of recursion is that a function calls itself. Recursion is also one of the most important tools for the implementation of efficient search- and sort-algorithms which are also abundant in computational materials science, usually in the form of look-up tables for the determination of interacting particles or finite elements of a system. Almost all search algorithms need a direct access to all elements that are to be sorted. When using a list, there is no such immediate access; thus one uses an index table which contains references to consecutive elements. Generally speaking, search algorithms are based on the strategy that they do not compare *all* elements of a data set S but only certain elements $s_i \in S$ which distinguish this data set from other data sets in a unique way. These specific elements are called "keys". Thus, in a search, only the keys are important.

Exercise 3 (Write a recursive version of Euclid's algorithm. The function may not contain any "FOR", "GOTO" or "WHILE" construct).

Solution 2. We use the key word "PROCEDURE" for calling a subroutine named "Euclid" which in turn calls itself in the following pseudo-code:

```
PROCEDURE Euclid(int a, b)
if b=0 then return a;
  else return Euclid (b, a mod b)
END
```

The concept of recursive functions was introduced by John McCarthy, the inventor of the programming language "LISP" [171]. The first step in writing

a recursive function is a specification of its input/output behavior. When writing the function in the next step one usually uses an if-statement to catch the beginning of the recursion.

Exercise 4 (Write a recursive function that calculates the faculty $n!$).

Solution 3. We use the fact that $n! = n \times (n-1)!$, i.e. the main problem is successively reduced to ever smaller problems until the trivial case ($0! = 1$) occurs. In the language $C++$ or C this reads

```
long Faculty(long n)
{
  if (n < 0)  return (0);
  if (n == 0) return (1);
  return ( n * Faculty(n-1) );
}
```

In Exercise 3, the recursive mathematical formula is identical with the implementation (besides syntactic details). The first *if*-statement makes sure, that recursion is only called with positive numbers. One great disadvantage of recursive functions is the additional overhead in terms of computation time as with each recursive call the function's return address has to be saved and memory for all local parameters has to be provided for. This disadvantage can be avoided by using *rest-recursive functions*, in particular if the used compiler supports "last-call optimization". A rest-recursive function call is implemented in C or $C++$ as displayed in Algorithm 2.3.

This function is illustrated in Fig. 2.17. In Fig. 2.17a the usual recursive function calls are displayed. This function provides at each call the case $(n-1)$ and multiplies it with n. After recursion to the lowest point the function goes upwards again and provides each respective result to the calling function. Thus, in each recursive step both, the arguments and the respective result are important. In Fig. 2.17b the rest-recursive function call is shown.

Here, at each call, the value of n is passed but also the current state of the faculty calculation. Thus, in each recursive step the calculation proceeds by one step until it is finished at the lowest step. Then the final result is simply returned successively to the calling functions. The passed parameters are not needed any more and there is no need to save a return address of the calling function, as the final result can be passed directly to the very first calling function.

2.6.2 Divide-and-Conquer

Divide-and-Conquer is a recursive programming technique which can be applied to some problems, provided that they have a suitable structure. If a problem of size n is to be solved, then this problem is split into two (or more)

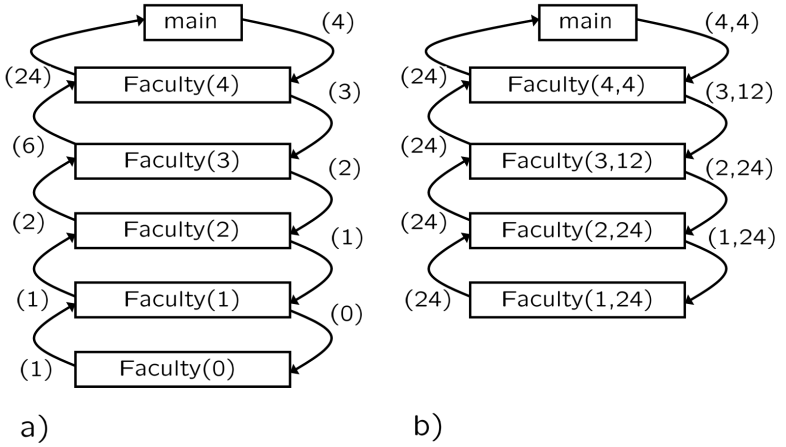


Fig. 2.17. Illustration of a recursive function call (**a**) according to Example 2 vs. a rest-recursive function call (**b**) according to Algorithm 2.3

sub-problems of size $n/2$ which can then be solved by using the same recursive algorithm. The two solutions can then be merged to a solution for the original problem. Examples for applications of this strategy for problem solving can be found in some common search algorithms, such as *merge-sort, bubble-sort, quick-sort* or *binary search*. It can be shown that these algorithms require an effort which is no better than $\mathcal{O}(n \log n)$ in the worst case. With quick-sort in particular, this is only true *on average*. In single cases the effort can be $\mathcal{O}(n^2)$. For an introduction of the $\mathcal{O}$-notation see Sect. 2.6.6.

Example 6 (An algorithm for the power of a number a).

If a power such as a^n is to be calculated for a larger number n, it is advisable, *not* to multiply a n times with a. It is way better to use the following recursion:

$$a^n = \begin{cases} 1 & : \text{ if } n = 0 \,, \\ (a^{n/2})^2 & : \text{ if } n > 0 \text{ and } n \text{ even} \,, \\ a \times (a^{(n-1)/2})^2 & : \text{ if } n < 0 \text{ and } n \text{ uneven} \,. \end{cases} \tag{2.83}$$

Equation (2.83) is provided as pseudo-code in Algorithm 2.4.

The construct "even(n)" in Algorithm 2.4 tests whether the value of n is even and "squared(n)" calculates the square of n. With each recursive function call the number n is halved; thus when calculating the n-th power one needs at the most $\log_2 n$ function calls.

Dynamic Programming

When using recursive function calls, one usually applies a "top-down" principle. A system of size n is split into tasks of smaller size, e.g. $(n-1)$, $(n-2)$ or

Algorithm 2.4 Recursive calculation of a power

```
PROCEDURE Power(a,n)
If n = 0 then return 1
  else if even(n) then return squared(a,n DIV 2)
    else return a * squared( Power(a,(n-1) DIV  2)) END
  END
```

$n/2$ and the corresponding function is called recursively. These subsequent recursive calls are usually independent of each other; thus, it might happen, that the *same* recursive call is done several times within the recursive structure.

In recursion of $f_B(5)$ in (2.84) the value of $f(5)$ is calculated once, $f(4)$ two times, $f(3)$ three times, $f(2)$ five times and $f(0)$ three times as depicted in Fig. 2.18.

This means that the recursion of $f_B(n)$ leads to an exponential effort. In this case it is more efficient to store calculated values and reuse them if needed again. This principle of working with a table in which calculated values are stored for later reuse is called *dynamic programming*. The table is filled "bottom-up", i.e. in the order $i = 0, 1, 2, ..., n$.

Example 7 (The Fibonacci sequence). A standard example for recursion is the calculation of the Fibonacci numbers

$$f_B(n) = f(n-1) + f(n-2),\ f(0) = 0\ . \tag{2.84}$$

A pseudo-code for a corresponding recursive algorithm is shown in Algorithm 2.5.

Example 8 (Pairwise alignment in biological sequence analysis). An important example for the use of dynamic programming techniques is the *editing distance*

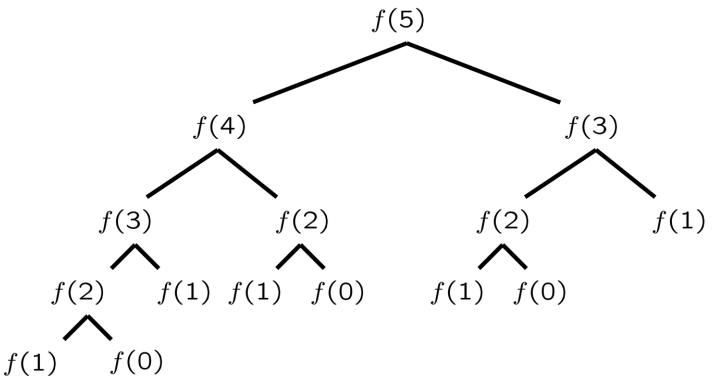


Fig. 2.18. Exponential effort for a recursive calculation of $f_B(5)$

Algorithm 2.5 Recursive calculation of the Fibonacci numbers

```
PROCEDURE f_B(n) if n = 0 then return 0
   else if n = 1 then return 1
      else return f(n-1) + f(n-2) END
   END
```

between two words $(a_1a_2...a_i)$ and $(b_1b_2...b_j)$ which is defined as the minimum number of elementary editing actions ("delete","insert","substitute") necessary to transform the first word into the second one. For example, the words "ANANAS" and "BANANA". Here, the editing distance is two, because:

ANANAS → BANANAS (insert),
BANANAS → BANANA (delete).

The editing distance between words is of great importance in genome research when comparing alignments of sequences (or parts of them) and then deciding whether the alignment is more likely to have occurred because the sequences are related, or just by chance. Box 2.5 shows the sequence alignments to a fragment of human α-globin and β-globin protein sequence (taken from the SWISS-PROT database identifier HBA_HUMAN[33]. The central line in the alignment indicates identical positions with letters, and "similar" positions with a plus sign.

In comparing DNA or protein sequences one is looking for evidence that they have diverged from a common ancestor by a process of mutation and selection. The basic mutational processes that are considered are *substitutions*, which change residues in a sequence, and *insertions* and *deletions*, which add or remove residues. Insertions and deletions are together referred to as *gaps*. Natural selection has an effect on this process by screening the mutations, so that some sorts of change may be seen more than others. The editing distance of sequences can be used to evaluate a scoring for the pairwise sequence alignment. The total score that is assigned to an alignment will be a sum of terms for each aligned pair of residues, plus terms for each gap. In a probabilistic interpretation, this will correspond to the logarithm of the relative likelihood that the sequences are related, compared to being unrelated, cf. [172]. Pairwise alignment constitutes an important application in computational genome research.

Box 2.5 Sequence alignment to a fraction of human α- and β-globin

```
HBA_HUMAN  GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKL
           G+ +VK+HGKKV  A+++++AH+D++ +++++LS+LH  KL
HBB_HUMAN  GNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKL
```

[33] www.expasy.ch/sprot

Greedy Algorithms

With dynamic programming, a table containing all hitherto found optimal solutions is used and updated in each step. *Greedy algorithms* do not make use of such a table; rather, the decision about the next solution component is based solely upon the *locally* available information. The possible next steps for the solution of a problem are compared with an evaluation function. This is to say that a decision for the next solution step among the available alternatives is based on maximizing (or minimizing) some figure of merit which in turn is based on comparing the locally available next steps. There is no *global* criterion as there is no list containing the history of solution steps. This explains the naming of the algorithm: At each step one simply continues into the direction which – from a local perspective – is the most promising one. In many cases, the greedy strategy results in an acceptable, albeit not optimal solution. Thus, this kind of algorithm (also called *Greedy-heuristics*) is mostly used for problems, where no other comparably efficient algorithms are known. A typical application would be the *traveling salesman problem*, (see Example 9), or the reverse Monte Carlo optimization scheme discussed in Sect. 7.2.1 on p. 336.

2.6.3 Local Search

In the previously discussed algorithms the problem solving strategy has always been a partition of the total problem of size $n = (a_1, a_2, ..., a_n)$ into subproblems $i = (a_1, a_2, ..., a_i)$ $(i < n)$ of smaller size. A different strategy is adopted with algorithms that use a *local search* strategy. Here, one starts with solutions of the complete problem of size n and modifies these by small manipulations (in genomics they are called "mutations"), e.g. by modifying only *one* component of the solution:

$$(a_1, a_2, ..., a_{j-1}, a_j, a_{j+1}, ..., a_n) \longrightarrow (a_1, a_2, ..., a_{j-1}, \bar{a_j}, a_{j+1}, ..., a_n) \ . \quad (2.85)$$

This strategy is based on searching for a new solution within the neighborhood of the previous one. The new solution should have improved properties compared with the old one. What "neighborhood" means depends on the context of the problem; e.g. in physics this might simply be some neighborhood of a system at a point in Γ phase space, cf. Chap. 6. Mutations are usually chosen randomly.

Example 9 (Traveling Salesman Problem). Consider a set of n points in $\mathbf{R}^2$. Which one is the graph which yields the *shortest* connection of all points in Fig. 2.19?

As a first attempt one might guess one complete solution, i.e. an arbitrary connection of all points. This temporary solution is then subsequently improved by identifying crossing connection lines (edges) and removing them by local mutations, cf. Fig. 2.19.

The initial configuration can be generated either at random or by a simple Greedy-heuristic.

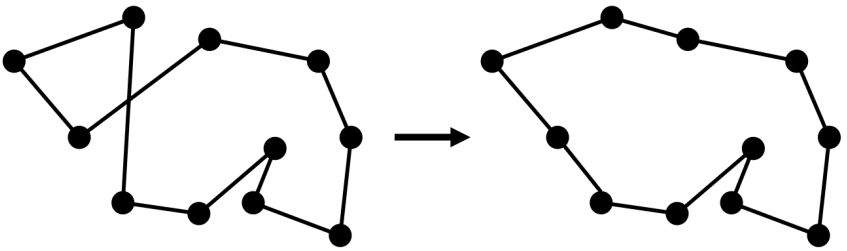


Fig. 2.19. The traveling salesman problem solved by a local search strategy for $n = 11$

A principle problem with this strategy is, that there is no guarantee to find the optimal solution by lining up local improvements, demanding that each single mutation improves some figure of merit (in this case: minimizes the traveled distance).

The improvements are done as long as there are no local moves left which could further improve the current solution. This problem is typical for molecular dynamics simulations, e.g. in a micro-canonical ensemble (i.e. one, that keeps the total energy of the system constant). Under some unfavorable circumstances it might occur that the system is "trapped" in a local energy minimum; thus the "true" equilibrium state, i.e. the global minimum energy state is not attained by the system. This is illustrated in Fig. 2.20.

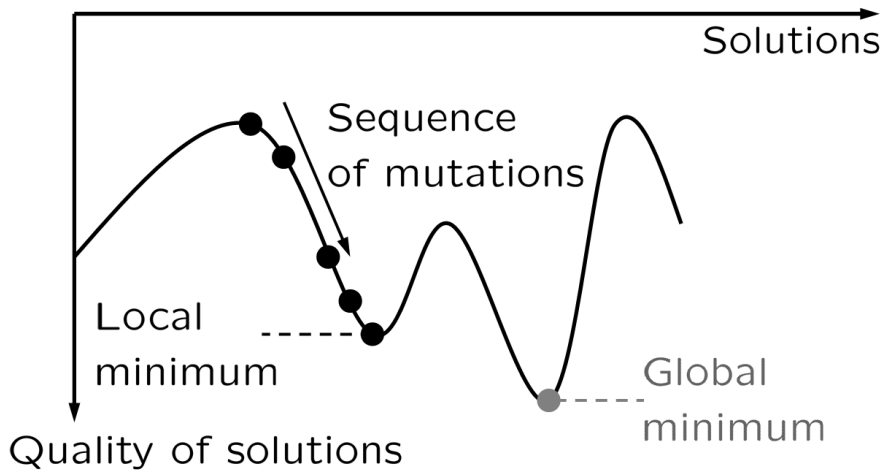


Fig. 2.20. Illustration of the *local search* principle. The sequence of mutations *always* improves some figure of merit in each single step. In physical applications, the quality of successive solutions might be determined by minimizing the total energy of the system. Thus, the system can be "trapped" in a *local* minimum state of energy which however is not the searched (optimal) *global* minimum (*depicted in grey*)

Sometimes it is better to allow for temporary local deterioration on the path to the optimal global solution. Such a strategy is described in the next section and is one of the most widely used principles in computational physics. Yet another strategy is to start the local search process many times with random initial configurations.

2.6.4 Simulated Annealing and Stochastic Algorithms

A method to overcome the discussed fundamental problem of local search in Sect. 2.6.3 is the use of random numbers. The use of random numbers in algorithms is a sometimes surprisingly efficient method for solving problems. Algorithms that make use of random numbers are generally termed *stochastic algorithms*. In contrast to this, algorithms which do not use random numbers and which – at all times – have only *one* possible next step, are called *deterministic algorithms*. Many stochastic algorithms are used in the field of algebra and number theory. In the natural sciences, so-called *Monte-Carlo algorithms* which contain a random element for the acceptance of new system states are very common and are often used in materials science, solid state physics and thermodynamics, cf. our discussion in Chap. 6. A very simple form of a stochastic algorithm would be to randomly shuffle some numbers in a list which are subsequently used as input for the quick-sort algorithm. By statistically permuting the list of numbers which is to be sorted by quick-sort results in an efficiency of $\mathcal{O}(n \log n)$ for *every* function call (instead of an *average* efficiency of $\mathcal{O}(n \log n)$).

A mutation of a system (in physics also called "trial move" or simply "trial") is accepted with a certain probability despite its worse quality in comparison with the solution in the original state. The probability of accepting a worse solution depends on two parameters Δ and T where Δ is the difference of the quality of the old vs. the new solution, that is, if the new solution is much worse than the previous one, then the acceptance probability $P(\Delta, t)$ is accordingly small. The second parameter T is called *temperature*. Initially, one starts at a high temperature T which then gradually decreases and approaches zero. That is, at the beginning, very large deteriorations of the solutions are still accepted with a certain probability P which decreases during the solution process. In the end, the procedure acts as a local search which accepts only quality improvements as a new system state. For the probability of acceptance the function in (2.86) is used.

$$P(\Delta, T) = \begin{cases} 1 & : \text{ new solution is better than the previous one ,} \\ \exp\left(-\Delta/T\right) & : \text{ else ,} \end{cases} \tag{2.86}$$

Algorithm 2.6 provides a generic pseudocode scheme of the Metropolis algorithm, respectively simulated annealing.

The strategy of using a time-dependent (i.e. dependent of the number of mutations) acceptance probability is similar to many physical processes.

Algorithm 2.6 Simulated annealing, Metropolis Algorithm (Version I)

```
Function K is some cost function which has to be minimized.
 tFinal and k are some constants.
 rnd is a uniformly distributed random
number in the interval [0,1].

Set t := initial temperature
Set a := initial solution
while ( t > tFinal ) DO
  for i:= 1 TO k DO
    Choose a mutation b from a at random
      if K(b) < K(a) then a:= b END
    else if rnd <= exp[-(K(b)-K(a)] THEN a:= b END
  END
  t:= t * 0.9
END OUTPUT a
```

One starts at a high temperature and then slowly cools down the system which gradually approaches an equilibrium – e.g. crystalline – state. This crystalline state corresponds to an optimal solution of the system and explains the terminology "simulated annealing" for this simulation strategy.

2.6.5 Computability, Decidability and Turing Machines

Computability is a branch of computer science that deals with principal questions such as:

- How can one formalize the concept of "algorithm"?
- What is a computable function (an algorithmically solvable problem)?
- Are there any non-computable functions?
- Are there any problems that are well-posed but which are not computable by any algorithm?

An *intuitive notion* of an algorithm, respectively "computability" contains the following elements [173]:

- A *finite* description,
- Unique rules for solving a problem, i.e. in every situation the next step is uniquely defined,
- A clear definition of input/output behavior of the algorithm,
- A solution after a *finite* number of steps.

Any description of an algorithm – generally done with some computer language – uses a finite alphabet; thus, there is only a *finite* number of descriptions of algorithms. On the other hand, there is an uncountable number of functions $f : \mathbf{N} \to \mathbf{N}$. Hence, the following question arises: For how many of this uncountable number of functions does no algorithm exist,

i.e. which of these functions are not computable? Based on the above intuitive notion of an algorithm this question cannot be answered.

Automata

In order to find answers to the questions raised above, the notion of an algorithm as a dynamic process that leads from from state to state, was formalized in terms of *deterministic finite automata* (DFAs) which can be represented as directed graphs. Graph theory itself is a powerful modeling tool – abundant in computer science – for the visualization of relations between objects. Each knot of a graph represents a certain state which can be assumed by the considered automata. The knots are connected by edges which are labeled with a symbol from some working alphabet, i.e. the set of symbols that can be read by the automata. By reading a symbol the automata goes from one state to a different one. Any finite, non-empty set can be considered to be a working alphabet. The elements of an alphabet are then called characters, letters or symbols. The sets $\Sigma = \{a, b, c, d, e, f, g, h\}$, $\Sigma = \{0, 2\}$ or $\Sigma = \{$condition, procedure, if, then, else, begin, end, while,...$\}$ are examples of alphabets. A "word" is a combination of elements of an alphabet. For example *bbiaaaacf* is a word on the alphabet Σ. The length $|l|$ of a word, e.g. $|bbiaaaacf| = 9$ is the number of symbols which are included in a word. The set of all words on an alphabet Σ is denoted as $\Sigma^\star$.

Example 10. Let the alphabet $\Sigma = \{a, b\}$. An example for a DFA would be the graph in Fig. 2.21.

This automata has the four states z_0, z_1, z_2, z_3. The start state is indicated by the arrow to the first state z_0. The end state(s) are denoted by a double circle

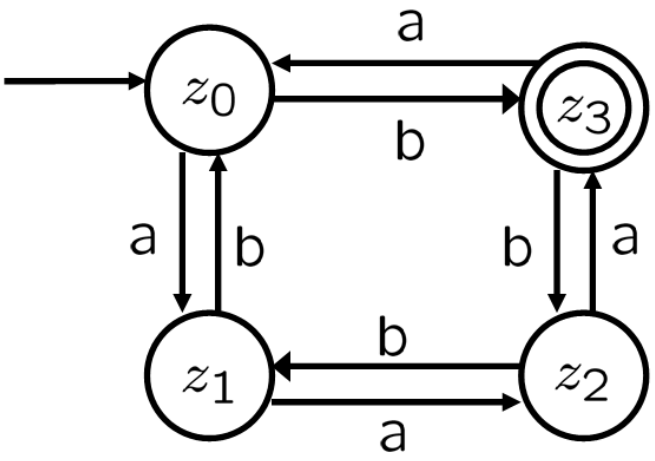


Fig. 2.21. Illustration of a deterministic finite automata as a directed graph

(in this case only z_3). An automata such as in the above example processes a "word" w by consecutively reading the symbols one at a time and performing the corresponding changes of state. If the state after processing the whole word is the end state, then the word is said to be "accepted". For example, the above automata accepts the word *bbbaa* going through states $z_0, z_3, z_2, z_1, z_2, z_3$ but not e.g. *aab*. The set of all accepted words is called the accepted language and is denoted as $T(M)$, i.e.:

$$T(M) = \{w \in A^\star \mid \text{the automata } M \text{ accepts the word } w\} \; . \tag{2.87}$$

For an introduction into the notation of sets such as in (2.87) the reader is referred to Chap. 3.

DFAs are models for simple "computers". They are equipped with a memory in the sense that each respective state of the automata is a piece of finite information. A consequence of this is that some languages such as $\{a^n b^n \mid n \geq 0\}$ cannot be processed. One step to reduce this restriction of automata is the use of a memory which is not restricted by the number of states. Such automata are called *cellar automata*.

Cellar automata have no principal limitation in the number of states they can assume. They are equipped with a stack structure (also called "LIFO" structure – "last in first out") as memory access. This means that the piece of information that has been stored *last* will be accessed *first* in reading mode, cf. Fig. 2.22. A LIFO structure is an example of an abstract data type (see Problem 5), which is abundant in computer science and which is often used for book-keeping of interactions between particles or finite elements in implementations of simulation programs for fluids or solids.

If the cellar memory is empty, there may be no memory access in read mode. This is indicated by the sign "#" at the bottom of the memory. This

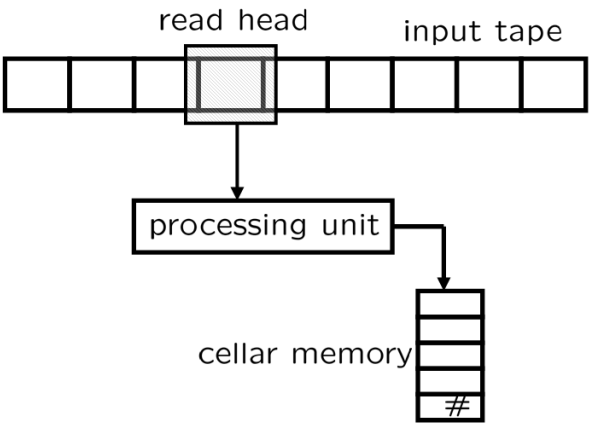


Fig. 2.22. Illustration of a cellar automata

is the lowest sign in the cellar memory and when this is read it indicates that the memory (apart from "#" itself) is empty. One step of a calculation with a cellar automata can be described as follows: The automata reads a sign from the input tape. Depending on the current state of the automata and of the sign read, the uppermost sign in the cellar memory is substituted by a sequence of cellar symbols which also may be empty. Similar to DFAs, a graphical representation of the change of states is used for cellar automata, cf. Fig 2.23. The edge in this case reads: If the cellar automata is in state z and the current input character is a with A being the top character, then the automata switches to state z' in the next step and substitutes A with $B_1B_2...B_k$.

In a mathematical notation one can describe the transition of a cellar machine from state z to state z' with a function δ:

$$\delta(z, a, A) = (z', B_1B_2...B_k) \; . \tag{2.88}$$

Cellar automata can express a larger set of languages than simple DFAs, e.g. the language $S = \{a^n b^n \mid \; n \geq 0\}$ which is not accepted by any finite automata, can be accepted by a cellar automata.

Example 11 (Draw the language $a^n b^n$ accepted by a cellar automata in graphical representation). The solution is depicted in Fig. 2.24b.

The cellar automata in Example 11 works deterministic. Each read a (except the first one) is written as A into the cellar memory. At any one time when b's are read, an A is removed from memory. As soon as no input characters are left on the tape, the cellar memory is empty, i.e. the lowest character "#" in the cellar memory is recognized.

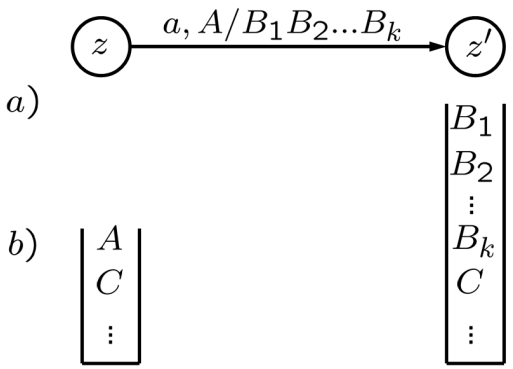


Fig. 2.23. Graphical representation of the change of states in a cellar automata. In (**a**) the notation for a transition from state z to state z' is depicted with A being the top character in the cellar memory and the sequence $B_1, B_2, ...B_k$ being the next characters to be read by the automata. In (**b**) the states of the cellar memory before and after reading the sequence of characters are depicted

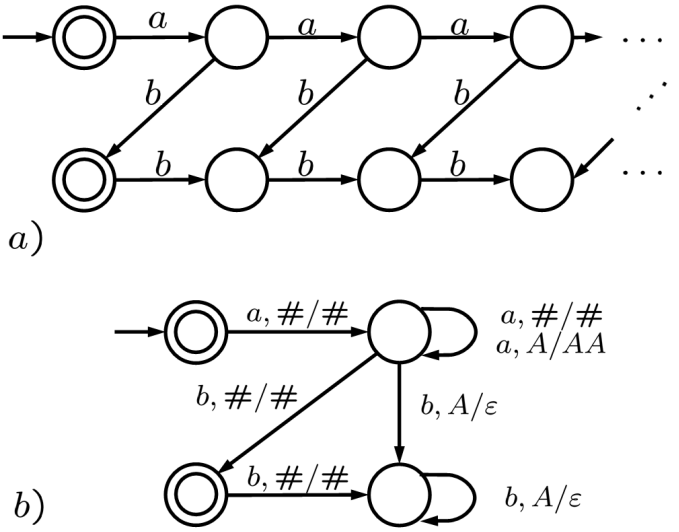


Fig. 2.24. The language $S = \{a^n b^n\}$ which is accepted by a cellar automata. In (**a**) the notation for an infinite DFA is shown and in (**b**) the graphical representation of a cellar automata accepting S is displayed

A cellar automata can only write or read at the end of the cellar memory. The next logical step in the formalization of the notion of a computer removes the restriction of the LIFO structure of cellar automata, i.e one introduces automata with a *random access memory* on all memory cells. Hence, the cellar memory is substituted by a sequential memory in the form of a memory tape which can be read or labeled by a read/write head. There is no additional "computing power" when making a distinction between a working-tape and an input-tape. Thus, input is read from *one* tape and results are written on the same tape by use of a finite read/write head which can move freely and stepwise in both directions along the tape, cf. Fig. 2.25. The tape can only be altered at the current position of the read/write head. Such a generalized model of an automata is called a *Turing machine*.

Turing Machines

The Turing machine is a fundamental concept for a simplest, memory-based computability model. Computation in this model means stepwise modification of the memory content. This can be represented by a transition function δ, e.g.

$$\delta(z, a) = \delta(z', b, x) \; . \tag{2.89}$$

In this case, the Turing machine in state z reads a character a and then assumes state z' substituting a by b where a and b are characters of the working alphabet. After reading of a character, the Turing machines moves

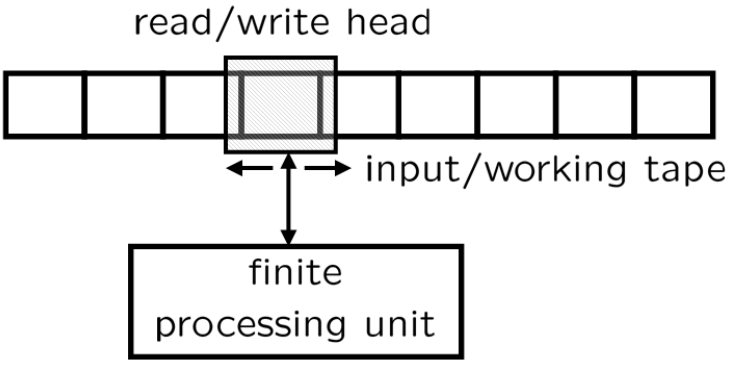


Fig. 2.25. A Turing machine with read/write head that moves freely and stepwise along a working/input tape

the read/write head by one step which is denoted by $x \in M = \{L, R, N\}$. The letters mean "left" (L), "right" (R) and "neutral" (N) (no step at all).

A mathematically precise definition of a Turning machine is the following

Definition 1 (Turing Machine). *A Turing machine (TM) is a 7-tuple* $M = (Z, \Sigma, \Gamma, \delta, z_0, \square, E)$ *, where*
Z *is the finite set of machine states,*
Σ *is the input alphabet,*
$\Gamma \supset \Sigma$ *is the working alphabet,*
$\delta : Z \times \Gamma \to Z \times \Gamma \times \{L, R, N\}$ *is the transition function in the deterministic case,*
$\delta : Z \times \Gamma \to P(Z \times \Gamma \times \{L, R, N\})$*is the transition function in the non-deterministic case,*
$z_0 \in Z$ *is the initial state,*
$\square \in \Gamma - \Sigma$ *is the blank,*
$E \supseteq Z$ *is the set of final states.*

Remark 1. The Turing machine introduces a fundamental notion of a simplest memory-oriented computing model. Computation in this model means stepwise altering the contents of memory cells. The memory of a Turing machine consists of single cells which may each contain one letter of a finite alphabet. The memory cells may be addressed by moving the write/read head stepwise from cell to cell.

In the year 1936 Turing proposed the following formal definition of "computability" by means of a Turing machine [174].

Definition 2 (Computability). *A function* $f : \Sigma^\star \to \Sigma^\star$ *(e.g. a function on* N*) is called* computable *or* Turing computable*, if there exists a Turing*

machine TM which can calculate this function f from any input x, i.e. if a Turing machine reads the binary representation of a natural number x and stops with f(x) (again, in binary representation) as result on the input/working tape after a finite number of steps.

In other words, a function f is called "computable" if there exists an *algorithm* which is able to compute the value of f for any input as argument within a finite number of steps.

Remark 2. Turing's definition marks the end of a development of notions of computability in an attempt to derive a general algorithm (or method) which allows for proving or disproving *all* mathematical theorems by using solely the underlying axioms of the system. This attempt of deriving a complete axiomatic basis of all mathematics along with a proof of its consistency (*Hilbert's Program*) was started by Hilbert in 1900 in a famous lecture [175]. In 1923 *Skolem* [176] considered as a basis for computability the so-called *primitive recursive* functions which are defined inductively. First, one declares a set of functions axiomatically as primitive recursive, i.e. they are computable. For the rest of the definition one provides rules as to how to obtain new – per definition – computable functions from already known computable functions. One rule is *insertion*, i.e. if f and g are computable then also $f(g(x))$ is computable. Another rule is *iteration*, also called called primitive recursion, i.e. if $f(x)$ is a computable function then also $h(n,x) = \underbrace{f(f(...f}_{n-\text{times}}(x)...))$ is computable.

In 1931 *Kurt Gödel* (1906-1978) introduced the primitive recursive function in his work [177] and proved that any sufficiently complex algebraic system is either incomplete or contradictory, i.e. he could prove that Hilbert's idea of a complete axiomatic basis of all mathematics was doomed. Several other proposals for the definition of computability were published, e.g. S.C. Kleene [178] and A. Church [179] in 1936 proposed a definition of computability based on the so-called λ-definable functions. They also showed in the same publications that this definition of computability is equivalent with the one based on primitive recursive functions [178]. In 1937 it was shown that the λ-definition of computability is also equivalent with Turing's definition [174, 180]. Gödel and Turing in essence showed that not all questions that can be asked within an axiomatic system in mathematics, in computer science or in general in some logical system are decidable within the bounds of the system. There are problems in mathematics, e.g. the diophantic equations[34] [181, 182], which are unsolvable within the underlying available system of axioms.

It turned out, that all proposed definitions of "computability" are equivalent to each other. Particularly, there has never been found a computational

[34] Diophantic equations are polynomial equations with integer coefficients for which an integer solution is sought. This problem is also known as Hilbert's 10th problem which was raised by him in the year 1900 [175]. It was not before 1970, when Hilbert's 10th problem could be proved to be unsolvable by Yuri Matiyasevič.

model which could not – in principle – be represented by a Turing machine. This general observation is also true for the so-called Quantum computers or DNA-computers. Based on this observation A. Church in 1936 made the following proposition [180]:

Proposition 1 (Church's thesis). *The notion of computability is adequately defined by the model of a Turing machine.*

Remark 3. Note that this is a proposition, that has been generally accepted, i.e. it cannot be proved. The term "Church's" or "Church-Turing" thesis seems to have been first introduced by Kleene who provides a good survey in Chaps. 12 and 13 of [183].

In summary, every effectively calculable function that has been investigated has turned out to be computable by a Turing machine. All known methods or operations for obtaining new effectively calculable functions from given effectively calculable functions are paralleled by methods for constructing new Turing machines from given Turing machines. All attempts to provide an exact analysis of the intuitive notion of an effectively calculable function, i.e. of the notion of computability, have turned out to be equivalent in the sense that each analysis offered has been proved to pick out the same class of functions, namely those that are computable by a Turing machine. Because of the diversity of the various analyses with this respect, Church's thesis is generally accepted.

Definition 3 (Decidability). *A set $A \subseteq \Sigma^\star$ of a Turing machine (i.e, an accepted language $T(M) \subseteq \Sigma^\star$), is called* decidable *if the characteristic function $\chi_T : \Sigma^\star \to \{0,1\}$ of T can be computed. For all $w \in \Sigma^\star$*

$$\chi_T(w) = \begin{cases} 1 & : \quad w \in T\,, \\ 0 & : \quad w \notin T\,. \end{cases} \tag{2.90}$$

Remark 4. For a definition of $T(M)$ see (2.87) on p. 90.

When it comes to to the question of decidability of formal languages, these languages are also called "Entscheidungsprobleme". With such an algorithm, which is depicted in Fig. 2.26 as a black box, the input is some word over an alphabet. On the other hand, when computing a function, the input is a subset of the natural numbers. Numbers, however, can be represented as words of the alphabet $\{0,1\}$ in binary form, and vice versa, words can be represented as numbers. For this, one simply numbers all symbols of the alphabet A, starting with zero; A word is then interpreted as a number over the number system with basis $|A|$. For a Turing machine, a "natural" form of input/output is given by a binary representation of words over the alphabet $\{0,1\}$.

Example 12 (The Halting Problem). Algorithms are written down in the form of programs, e.g. for a Turing machine. Programs can be written down as

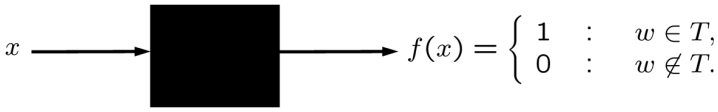


Fig. 2.26. Decidability of a formal language. The corresponding algorithm is depicted as black box. Such problems are also called "Entscheidungsprobleme"

words of a certain alphabet. In this representation, programs can be used as input for another algorithm[35] In the following, a particular language will be considered which has the following alphabet:

$$H = \{A|\ A \text{ is a program, which, when used as input of } A \text{ stops after a finite number of steps.} \tag{2.91}$$

This particular problem is called *Halting problem* and it is undecidable.

Proof: Assuming that the Halting problem was decidable, there is an algorithm which can be depicted as black box, cf. Fig. 2.27.

Using this algorithm, one constructs a new algorithm, which stops exactly, if the black box outputs 0. (The actual output of the algorithm is irrelevant). In the case of 1 as output, the algorithm never stops, i.e. it enters an infinite loop.

Let z the new code of this algorithm. The question is whether z stops or not, i.e. whether $z \in H$,or $Z \notin H$. If z stops after input of z, then by construction of the algorithm, the black box outputs 0 upon input of z. However, the black box is the assumed decidability algorithm for the Halting problem. If the black box outputs 0 after input of z, then this means that the algorithm does not stop upon input of z. Likewise, assuming that z does not stop after input of z, then it follows analogusly that z stops after input of z. Thus, there is a logical contradiction and the Halting problem is not decidable. ■

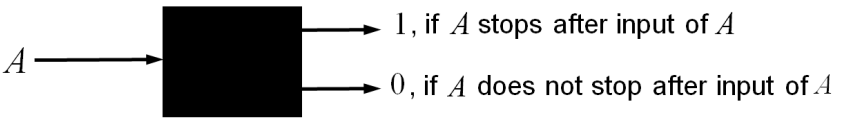


Fig. 2.27. A black box algorithm for the proof of the undecidability of the Halting problem

[35] A simple example for an algorithm that has as input a different algorithm is a compiler.

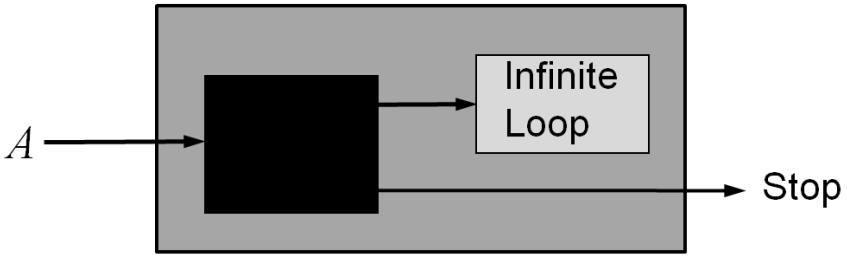


Fig. 2.28. Result of the Halting problem. Either the algorithm enters an infinite loop or it stops, depending on the output

As a result of the Halting problem, one realizes, that there are functions, which are not computable. The characteristic function of the Halting problem is one example, cf. Fig. 2.28.

Example 13 (A Turing machine for the function $f(x) = x + 1$). The Turing machine depicted in Fig. 2.29 calculates the function $f(x) = x + 1$. If one starts this machine with a number x (in binary representation), then it stops after a finite number of steps in a defined end state and the number in the current state of the working tape is $x + 1$ (again, in binary representation). Thus, the function $f(x) = x + 1$ is a "Turing computable" function.

Example 14 (A Turing machine that substitutes all characters). The transition function δ in Fig. 2.29 defines the transition of a Turing machine from one state to the next. This can also be depicted in a *transition table*. Let

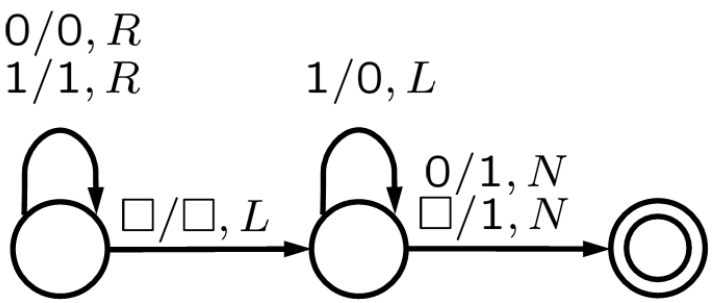


Fig. 2.29. Graphical representation of a Turing machine that calculates the function $f(x) = x + 1$. This machine – without changing the tape content – first moves the read/write head all the way to the right to the position of the lowest bit. If this bit is 0 then it is set to 1 and the machine is done. If the bit is a 1 then the bit is set to 0 and the head moves one step to the left. This procedure is repeated until the leftmost bit is read and changed accordingly

$$Z = \{z_0, z_e\} ,$$
$$\Sigma = \{a, b\} ,$$
$$\Gamma = \{a, b, \square\} .$$

The transition table is a 5-tuple $(z_i, z_j, \gamma_i, \gamma_j, M)$ given in Table 2.7.

A nice Java-program which provides a GUI and which simulates the (binary) input and output of a Turing machine can be obtained in the world wide web [36]. The machine accepts as input words in binary format and a transition table such as in Table 2.7 and checks whether the word is accepted.

We have seen previously that when the "computational power" of a computational model such as the Turing machine is restricted too much, then eventually one reaches a point at which the computational model cannot calculate anymore all possible functions. This is the case, e.g. when a Turing machine is "degraded" to a DFA.

Something similar happens when one restricts the programming language of the underlying model too much, e.g. when only allowing *FOR*-computable functions but no *WHILE*-loops. The main difference between a *WHILE*- and a *FOR*- loop is that with the latter the number of loops is set at the beginning, whereas a *WHILE*-loop decides dynamically after each loop whether a stop-criterion has been reached. Thus, a program which *only* uses *FOR*-loops definitely stops after a finite number of steps. The *Ackermann function* $A(x, y) : \mathbf{N^2} \rightarrow \mathbf{N}$ which is defined for all $(x, y) \in \mathbf{N}$ by the following recursive scheme (see [184] and [185]):

$$A(0, y) = y + 1 , \tag{2.92a}$$
$$A(x + 1, 0) = A(x, 1) , \tag{2.92b}$$
$$A(x + 1, y + 1) = A(x, A(x + 1, y)) , \tag{2.92c}$$

is an example for a function which is only *WHILE*-computable, but not *FOR*-computable, see e.g. [186]. Furthermore, it can be shown that *FOR*-computability and *primitive recursive* computability are equivalent [186]. Thus, $A(x, y)$ is an example of a computable function which is *not* primitive recursive.

Table 2.7. A transition table for a Turing Machine that substitutes each a with b and vice versa. The machine stops when a blank is reached

z_i	γ_i	γ_j	M	z_j
z_0	a	b	R	z_0
z_0	b	a	R	z_0
z_0	$\square$	$\square$	R	z_e

[36] http://ais.informatik.uni-freiburg.de/turing-applet/

2.6.6 Efficiency of Algorithms

While there are some problems which fall into the category of "computable" it turns out that the associated algorithms are useless for all practical purposes because they require astronomical computation times. The *Ackermann function* (2.92)a-c is just one of many examples which are computable *in principle* but which need astronomical computing times, even at small input values. This function grows larger than is possible by substitution or recursion and only for small values of the arguments ($x < 4$ and $y < 4$) an explicit expression of $A(x, y)$ can be given. For the original publication of this function by Ackermann in 1928, which was slightly different from the modern textbook version given above, see [187]; also compare Problem 6.

The computing time for a problem is measured by the number of *elementary steps* (ES) needed by some algorithm until it stops, i.e. until the problem is solved. Examples for elementary steps are:

- Executing one of the elementary operations $(+, -, \times, DIV, MOD)$,
- Assigning a value, i.e. changing the contents of a memory,
- Initializing a loop variable,
- Testing an if-condition.

Example 15 (Number of elementary steps of a very simple sort algorithm). Consider the following piece of pseudocode (2.7) which gets as input an array $a[1, ..., n]$ which is to be sorted.

The kernel of two loops in Algorithm 2.7 consists of one if-condition (1 ES) and – in the positive test case – three assignments (3 ES) which switch a[i] with a[j]. Each i- and j-loop counts 1 ES. Thus, one can directly write down the number of ES:

$$\sum_{k=1}^{n-1}\left(1+\sum_{l=k+1}^{n}(1+4)\right) = (n-1)+\sum_{k=1}^{n-1}\sum_{l=k+1}^{n}5 \tag{2.93}$$

$$= (n-1)+5\times\frac{n\times(n-1)}{2} \tag{2.94}$$

$$= 2.5\times n^2 - 1.5\times(n-1)\,. \tag{2.95}$$

Algorithm 2.7 A simple sort algorithm

```
for i := 1 TO n - 1 DO
  for j := 1 TO n DO
    if a[i] > a[j] then
      h := a[i]; a[i] := a[j]; a[j] := h;
    END
  END
END
```

Assuming that *one* ES on an average computer takes 10^{-9} seconds, one can sort arrays containing 20000 elements within one second. Often, one is only interested in how the run time of an algorithm depends on the number of input elements n, i.e. one only considers the leading term in the computation time. In the example above one would speak of a "*quadratic*", or "order n^2 runtime" and writes symbolically $\mathcal{O}(n^2)$. The meaning of this notation is the following:

Definition 4 ($\mathcal{O}$-notation). *A function $g(n)$ is of order $f(n)$, i.e. $g(n) = \mathcal{O}(f(n))$ if there are constants c and n_0 such that $\forall n \geq n_0$: $g(n) \leq c \times f(n)$.*

The symbol "$\forall$" is short for "for all" in mathematical notation.

Example 16. The function $2n^2 + 5n$ is of order n^2, or in symbolic notation: $2n^2 + 5n = \mathcal{O}(n^2)$, as one can choose $c = 3$. Then $2n^2 + 5n \leq 3n^2 \ \forall n > 5$. Thus, the previous relation is true for e.g. $n_0 = 5$.

To classify the efficiency of algorithms we consider in Table 2.8 five different algorithms A_1, A_2, A_3, A_4, A_5 with corresponding runtimes $n, n^2, n^3, 2^n, n!$, where n is the considered system size, e.g. the number of atoms, particles or finite elements in some simulation program. These runtimes are typical for different applications in materials science. We again assume that one elementary step takes 10^{-9} seconds on a real computer.

It is obvious from Table 2.8 that *exponential* runtimes (algorithms A_4 and A_5) are generally not acceptable for all practical purposes. For these algorithms, even with very small system sizes n one reaches runtimes which are larger than the estimated age of the universe (10^{10} years). Algorithm A_5 could be a solution of the traveling salesman problem (see Sect. 2.6.3). If the first point out of n has been visited, there are $(n-1)$ choices for the second one. This finally results in an exponential runtime of at the least $n!$ steps. A runtime 2^n as in A_4 is typical for problems where the solution space of the problem consists of a subset of a given set of n objects; There are 2^n possible subsets of this basis set. The "efficient" algorithms A_1, A_2, A_3 with runtimes of at the most n^3 are the most commonly used ones in computational materials science.

Usually, in atomistic simulations one assumes the interactions between particles to be pairwise additive. Hence, the interaction of particles (or atoms) in a system depends only on the current position of *two* particles. Sometimes however, three-body interactions have to be included, e.g. when considering bending and torsion potentials in chain molecules, cf. Sect. 6.3.7. These potentials depend on the position of at least three different particles. Solving the Schrödinger equation in ab-initio simulations also leads to a n^3-dependency of the runtime. This is the main reason why ab initio methods are restricted to very small system sizes (usually not more than 1000 atoms can be considered). Solving the classical Newtonian equations of motion with a "brute-force" strategy leads to a n^2-efficiency $\left(\frac{n\times(n-1)}{2}\right)$ of the algorithm

Table 2.8. Overview of typical runtimes of algorithms occurring in materials science applications. Depicted are the number of elementary steps and the corresponding realtimes for the different algorithms under the assumption that one ES takes 10^{-9} seconds

Algorithm	runtime	$n = 10$	$n = 20$	$n = 50$	$n = 100$
A_1	n	10 ES $10^{-8}\,s$	10 ES $2 \times 10^{-8}\,s$	10 ES $5 \times 10^{-8}\,s$	10 ES $10^{-7}\,s$
A_2	n^2	100 ES $10^{-7}\,s$	400 ES $4 \times 10^{-7}\,s$	2500 ES $2.5 \times 10^{-6}\,s$	10000 ES $10^{-5}\,s$
A_3	n^3	1000 ES $10^{-6}\,s$	8000 ES $8 \times 10^{-6}\,s$	10^5 ES $10^{-4}\,s$	10^6 ES $0.001\,s$
A_4	2^n	1024 ES $10^{-6}\,s$	10^5 ES $0.001\,s$	10^{15} ES $13\,days$	10^{30} ES $\sim 10^{13}\,years$
A_5	$n!$	$\sim 10^6$ ES $0.003\,s$	$\sim 10^{18}$ ES $77\,years$	$\sim 10^{64}$ES $10^{48}\,years$	10^{158} ES $\sim 10^{141}\,years$

that calculates the interactions of particles. This is also generally true in finite element codes where special care has to be taken when elements start to penetrate each other. Usually one uses so-called *contact-algorithms* which use a simple spring model between penetrating elements. The spring forces try to separate the penetrating elements again and the core of the contact algorithm is a lookup-table of element knots which is used to decide whether two elements penetrate each other or not. This algorithm in its plain form has an efficiency of n^2. As an n^2 efficiency of an algorithm still restricts the system size to very small systems of a few thousand particles one uses several methods to speed-up the efficiency of algorithms in computer simulations. Usually, this is done by using sorted search tables which can then be processed linearly (and thus reaching an efficiency of $\sim n \log n$). Hence, when it comes to the efficiency of algorithms in materials science, one will always try to minimize the effort to $\mathcal{O}(n)$ (with a remaining prefactor that might still be very large). A discussion of several of the most important speed-up techniques commonly used in MD simulation codes is provided in Sect. 6.4.1.

Another consideration in Table 2.9 shows why algorithms A_1, A_2 and A_3 may be considered to be *efficient*. Assuming that the available computer systems – due to a technology jump – will be 10 or 100 times faster than today, then the efficiency of algorithms A_1, A_2 and A_3 will be shifted by a factor, whereas for the exponential algorithms A_4, A_5 the efficiency will be shifted only by an additive constant, cf. Table 2.9.

Table 2.9. Speedup of the runtime of different algorithms assuming a hardware speedup factor of 10 and 100. The efficiency of polynomial algorithms will be shifted by a factor while exponential algorithms are only improved by an additive constant

Algorithm	runtime	efficiency	speedup factor 10	speedup factor 100
A_1	n	n_1	$10 \times n_1$	$100 \times n_1$
A_2	n^2	n_2	$\sqrt{10} \times n_2 = 3.16 \times n_2$	$\sqrt{100} \times n_2 = 10 \times n_2$
A_3	n^3	n_3	$\sqrt[3]{10} \times n_3 = 2.15 \times n_3$	$\sqrt[3]{100} \times n_3 = 4.64 \times n_3$
A_4	2^n	n_4	$\log_2(10 \times n_4) = n_4 + 3.3$	$\log_2(100 \times n_4) = n_4 + 6.6$
A_5	$n!$	n_5	$\approx n_5 + 1$	$\approx n_5 + 2$

Algorithms A_1, A_2 and A_3 have *polynomial* runtimes. An algorithm is said to be *efficient* if its runtime – which depends on some input n – has a polynomial upper bound. For example, the runtime function $2n^4(\log_2 n)^4+3\sqrt{n}$ has a polynomial upper bound (for large n), e.g. n^5. In $\mathcal{O}$-notation this is expressed as $\mathcal{O}(n^k)$ with k being the degree of the polynomial. Algorithms A_4 and A_5 on the other hand have no polynomial upper limit. Thus, they are called *inefficient*. The class of Problems that can be solved with efficient algorithms – i.e. algorithms that are polynomially bounded – are denoted with the letter **P**, cf. Fig. 2.30 The set of polynomials is closed under addition, multiplication and composition. Thus, **P** is a very robust class of problems. Combining several polynomial algorithms results into an algorithm which again exhibits a polynomial runtime.

Remark 5. Due to the robustness of the definition of the class **P** of efficient algorithms, an inefficient algorithm can have a shorter runtime than its efficient

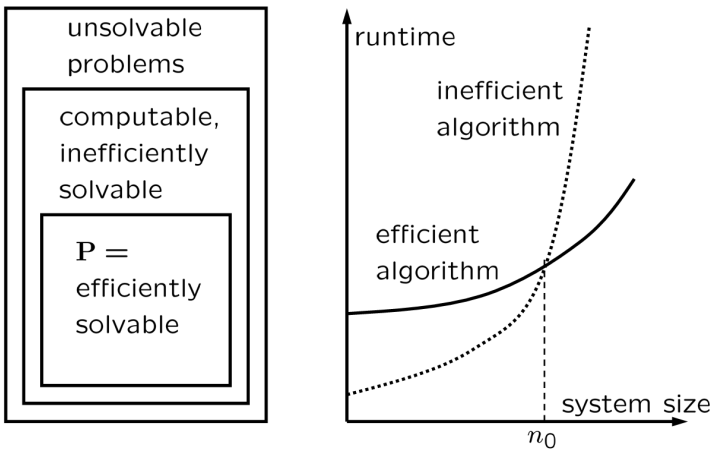


Fig. 2.30. Illustration of the class **P** of efficiently solvable problems, inefficiently solvable and unsolvable problems along with their runtime behavior

counterpart, up to a certain system size n_0. For example, an algorithm with a runtime $1000 \times n^{1000}$ falls into the class **P** whereas an algorithm with a runtime 1.1^n is exponential and thus inefficient. However, the exponential algorithm only exhibits longer runtimes than the efficient one for system sizes $n \sim 123000$, cf. Fig. 2.30.

In Example 15 on p. 99 the "worst-case" runtime for a simple sort-algorithm was considered assuming that the if-condition within the loop of Algorithm 2.7 is true and thus, three elementary steps are *always* executed. In the "best-case" – e.g. if the array has been sorted – this if-condition is not true and there are only $(n-1)+n(n-1) = n^2-1$ elementary steps. For a randomly shuffled array one can show that the expectation value for the number of elementary steps is $\sum_{k=2}^{n}(1/k) \approx \ln n$ [188]. Thus, with a randomly sorted array the total number of ES in this example is roughly $n^2 + 3n \ln n$. Hence, the *actual* runtime of an algorithm lies somewhere between the worst-case and the average-case runtime behavior, cf. Fig. 2.31.

Remark 6 (Cryptology). The fact, that for certain problems no efficient algorithm is known, is the basis of almost all practical cryptographical concepts, e.g. password files $f(x)$ on a computer system are generated from the input x as a new password. However, nobody who can read the password file will be able to calculate the inverse function $f^{-1}(x)$, i.e. to find an x' for which $f^{-1}(x') = f(x)$. Although the password is not known to the system, it can be checked simply by calculating the function $f(y)$ again with the provided password y at login and comparing it with $f(x)$. For the function f, one preferably used so-called *one-way functions*, i.e. functions for which f can be calculated efficiently, but for which no efficient algorithm is known to calculate f^{-1}. For example, a common choice is $f(x) = a^x MOD n$, where $x = 1, 2, ..., n-1$, $a \in \mathbf{N}$ and n is a prime number. The number of necessary steps to calculate

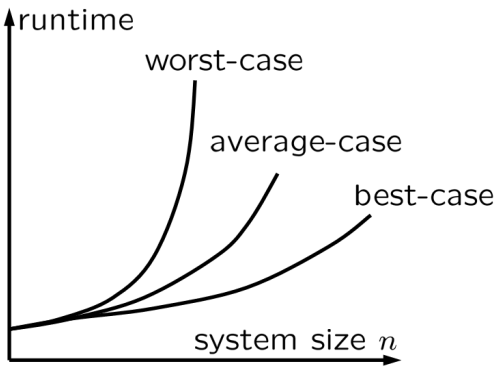


Fig. 2.31. Illustration of the worst-case, average-case and best-case behavior of algorithms. The average-case behavior of an algorithm is usually the most difficult to determine

this function is given by $\mathcal{O}(\log x)$, i.e. *linear* in the binary representation of x and there is no efficient algorithm to calculate $f(x)^{-1}$ [189].

A special class of problems for which no efficient algorithms are known are the so-called **NP**-*complete* problems. The letters **NP** are short for *non-deterministic polynomial runtime*. **NP** problems have the same computability model as **P** problems, but with a non-deterministic algorithm. The NP-complete problems are the most difficult problems in NP in the sense that they are the smallest subclass of NP that could conceivably remain outside of **P**, the class of deterministic polynomial-runtime problems. The reason is that a deterministic, polynomial-runtime solution to any NP-complete problem would also be a solution to every other problem in NP.

We provide in the following a formal definition of a **NP**-problem, following Steven Cook's original article in which the first **NP**-problem was published [190], although in this article the term **NP** was not yet used. A Language L, i.e. a "Entscheidungsproblem" lies within the class **NP**, if there is an efficient algorithm, which works with two input variables, such that the following two conditions are fulfilled:

1. If $x \in L$, then there is a second word y as input (where the length $|y|$ of y has to be polynomial bounded within the length $|x|$) such that the above mentioned efficient algorithm provides 1 as output.
2. If $x \notin L$, then, after input of x, the output of the efficient algorithm is always 0, independent of the arbitrary second input y.

This definition is depicted in Fig. 2.32. The importance of it lies in the fact that no more efficient algorithm for the decision whether $x \in L$ is known except of trying out the complete set of potential input elements $y \in \Sigma$ until the efficient algorithm outputs a 1. When such a y is found then $x \in L$. If the complete search of all elements y was done and the output was always 0 then $x \notin L$. The length of y is bounded by a polynomial p of the length $|x|$; Hence, the search effort is of order $|\Sigma|^{p(|x|)}$ which is exponential, i.e. non-polynomial. The great importance of **NP**-class problems lies in the fact that more than 3000 **NP**-problems are known from such different research areas as number theory, computational geometry, graph theory, sets and partitions, program optimization, automata and language theory and many more, see

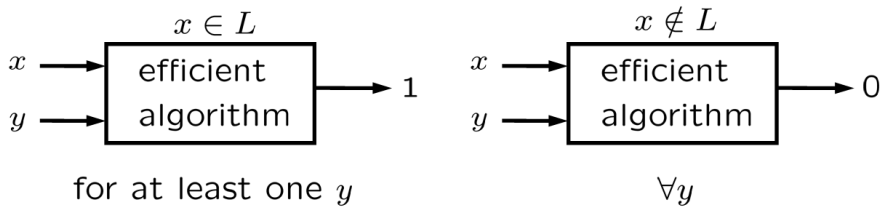


Fig. 2.32. Illustration of a language L that is an element of class **NP**

e.g. [190, 191]. This also has consequences for code optimization in materials science applications. For example, when writing a massive parallel computer program then there is no known optimization strategy for a domain decomposition (a splitting of the considered domain into several parts which are then assigned to different processors) of the considered system which leads a guaranteed optimal solution. In graph theory there exists no efficient algorithm that minimizes the distance between vertices of a graph which is of importance, e.g. when generating a high-quality mesh for finite element applications.

Today, it is generally assumed that all problems in **P** are contained in the class **NP**, cf. Fig. 2.33.

So far, no proof that decides whether $\mathbf{P} = \mathbf{NP}$ or $\mathbf{P} \neq \mathbf{NP}$ is known, i.e. it is unknown whether NP-complete problems are in fact solvable in polynomial time[37].

With these remarks we end our discussion of the attempts to formalize the notions of "computability" and "algorithm".

2.6.7 Suggested Reading

A good starting point to appreciate the developments in the modeling of real systems by means of algorithms would be Chabert [192] and Lee [193]. There is a multitude of excellent textbooks on graph theory, formal languages and automata. Cormen et al. [194] provides a good introduction to algorithms in graph theory. Diestel [195] and Gibbons provide a sophisticated treatment of graph theory. Prömel and Steger [196] treat the Steiner tree problem in depth and provide many useful algorithms for practical problems. Gruska [197] and Hopfcroft [173] provide a solid introduction into the foundations of automata theory. A good treatment of Markov chains can be found in Motwani [189]. The original article by Markov is [198] and for an English translation see e.g. [199]. The Monte-Carlo Method was introduced into physics in 1953 by Metropolis

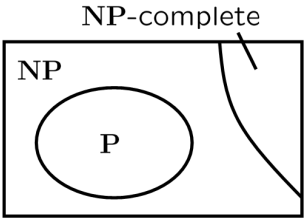


Fig. 2.33. The **NP**-complete problems are the smallest subclass of **NP** for which no efficient algorithms are known

[37] This is one of the great unsolved problems in mathematics. The Clay Mathematics Institute in Cambridge, MA, U.S.A. is offering a 1 million $ reward to anyone who has a formal proof that $\mathbf{P} = \mathbf{NP}$ or that $\mathbf{P} \neq \mathbf{NP}$.

et al. [23]. A nice historical account of the Turing machine is provided in Copeland [200]. A standard introduction into complexity and computability theory is provided by Ausellio et al. [201] or Cooper [202]. Classic textbooks on computability and **NP**-problems are e.g. Garey and Johnson [203] or Homer and Selman [204].

Problems

Problem 1. Proper Time Interval $d\tau$
Show that the Lorentz-transformations Λ^{α}_{β} leave the proper time interval $d\tau$ (see p. 51) invariant.

Problem 2. Conservation Laws
State for each of the following particle reactions whether it is forbidden or not. If applicable, state the conservation law that is violated.

(a) $\bar{p} + p \rightarrow \mu^+ + e^-$,
(b) $n \rightarrow p + e^- + \nu_e$,
(c) $p \rightarrow n + e^+ + \nu_e$.

Problem 3. Euler-Lagrange Equations
Perform the variation of (2.40) and show that (2.41) are the corresponding equations of motion.

Problem 4. Klein-Gordon and Dirac Equation
Show that the field ψ of (2.62) on p. 72 satisfies the correct energy-momentum relation, i.e. it satisfies the Klein-Gordon equation (2.57). Derive from this a set of equations for the α_i and β.

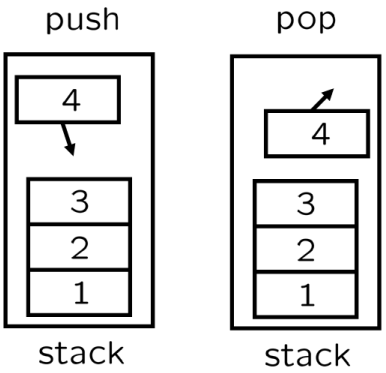


Fig. 2.34. Push **(a)** and pop **(b)** operation with stacks

Problem 5. Abstract Data Types: LIFO Structure
The two basic operations of stacks (LIFO structures) are *push* (putting one data element on the stack) and *pop*, cf. Fig. 2.34.

Write an implementation of the stack with a push and pop functionality in C++ using a modular design, i.e. use a header file "Stack.h" for declarations, a file "Stack.cpp" and a main procedure which tests this implementation.

Problem 6. An implementation of the Ackermann function
Write a recursive implementation of the Ackermann function (2.92)a, (2.92)b, (2.92)c. How long does it take to compute $A(5, 0)$? (You can go and drink a cup of coffee in the mean time). What about $A(5, 1)$?