

# Chapter 1

## Mathematical Models

In this chapter we present a range of dynamical systems from different areas of application and use them as examples to illustrate some typical problems from systems and control theory. Several of the mathematical models we introduce and discuss in the following sections will be taken up as examples in later chapters.

The development of mathematical systems theory starts in the next chapter. Readers who prefer to go directly to Chapter 2 can do so without any difficulty as the mathematical exposition in that chapter is self-contained and independent of following material. On encountering an example based on a dynamic model from Chapter 1, they may wish to look back to its origin here to find more details and get additional background information.

This chapter consists of six sections in which we present dynamical models from the following areas:

- Biology (Population Dynamics)
- Economics
- Mechanics
- Electromagnetism and Electrical Systems
- Digital Systems
- Heat Transfer

The mathematical models in the first three sections are described by *ordinary differential equations* and by *difference equations*. Also in Section 1.4, although the basic equations of electromagnetism are *partial differential equations*, we will only consider so-called *lumped models* of electromagnetic devices which again are described by ordinary differential equations. Different types of models are presented in the remaining two sections. In Section 1.5 we consider digital systems which have only a finite number of different states and are represented as finite automata. In the last section we deal with an example of a distributed parameter system described by partial differential equations.

In all these sections we will not only discuss the mathematical models but also point out some of the problems encountered in determining a mathematical model for a real process. While most of the sections just present a gallery of typical examples, some modelling methods will be sketched out in the sections on mechanical and electrical systems.

## 1.1 Population Dynamics

In order to predict or estimate the growth of a given population one needs a dynamical model. Such models may also be useful if one wants to control the development of a population. For example problems of control arise in fisheries management where one would like to keep fishing at a sustainable level and maximize the average catch over long time periods. In other applications interaction between different populations may be important and one may make use it for control purposes, e.g. in pest control where one introduces predators to reduce the pest. In this section we consider two classical models of population dynamics.

**Example 1.1.1. (Logistic growth model).** The simplest growth model is

$$\dot{x}(t) = ax(t). \quad (1)$$

Here  $x(t)$  is the size, density or biomass of a given population at time  $t$  and the growth parameter  $a$  is the *intrinsic growth rate* (difference between the birth rate and the death rate) of the population. If the initial size of the population is  $x(0) = x_0 > 0$  the development follows the exponential law  $x(t) = e^{at}x_0$ . Thus we have exponential growth if  $a > 0$  (i.e. the birth rate is larger than the death rate) and exponential decay if  $a < 0$ . The idea that human populations when “unchecked by the difficulties of subsistence” have a positive constant natural growth rate goes back to Malthus. In his *Essay on Population (1798)* he contrasted the natural geometric growth of mankind with the linear growth of subsistence resources and drew far reaching conclusions from this which had a profound effect on political economics.

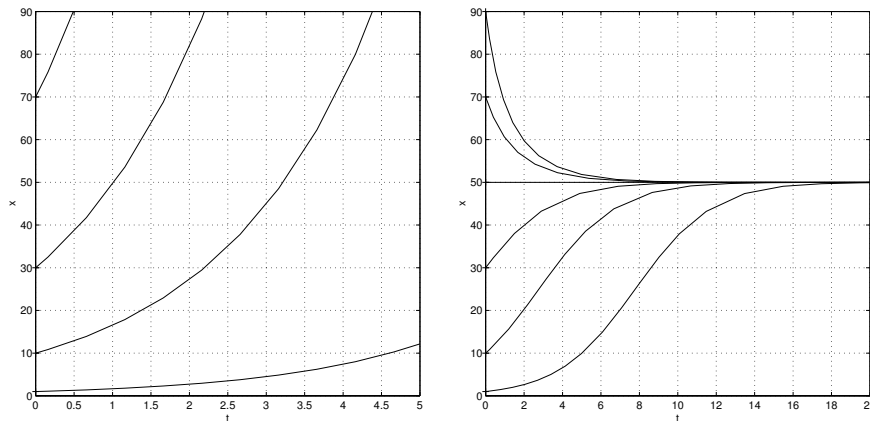


Figure 1.1.1: Exponential and logistic growth models

The exponential growth model, although adequate in many applications over a limited time span becomes unrealistic in the long run since  $e^{at}x_0 \rightarrow \infty$  as  $t \rightarrow \infty$ . The growth rate  $\dot{x}(t)/x(t)$  cannot be constant over arbitrarily long periods of time, since resources are limited. As the population becomes larger and larger, restraining factors will have an increasingly negative effect on population growth (“crowding”). In 1838 Verhulst proposed another growth model which incorporated the limiting factors and accounted for the fact that individuals compete for food, habitat, and other limited resources,

$$\dot{x}(t) = r(K - x(t))x(t). \quad (2)$$

According to this model a small population will initially grow at an exponential rate  $rK$  but as the population increases the growth rate will be diminished.

If the system is initially at  $x_0 = K$  then it will remain at  $x(t) = K$  for all time. Then the population is at an equilibrium  $x(t) \equiv \bar{x} = K$ ,  $t \geq 0$ . If  $0 < x_0 < K$  the population  $x(t)$  will increase continuously and approximate  $K$  as  $t \rightarrow \infty$ . If  $x_0 > K$ , the population size  $x(t)$  will converge towards  $K$  from above. In fact the following formula for the solution is easily obtained by separation of variables

$$x(t) = \frac{K}{1 + (K/x_0 - 1)e^{-rKt}}.$$

The graphs of these solutions are called *logistic curves* and Verhulst's model is also known as the *logistic growth model*. Figure 1.1.1 illustrates that  $x(t) \equiv K$  is a stable equilibrium, i.e. all trajectories with initial state  $x_0 > 0$  converge towards this equilibrium as  $t \rightarrow \infty$ . The saturation level  $K$  is interpreted as the *environmental carrying capacity* of the corresponding ecosystem. Now suppose that we want to describe the dynamics of a fish population under the influence of fishing. If  $u(t) \geq 0$  is the catch rate and we assume the logistic growth model for the undisturbed fish population, we obtain *Schaefer's model*

$$\dot{x}(t) = r(K - x(t))x(t) - u(t). \quad (3)$$

Note that only non-negative solutions  $x(t, u) \geq 0$  make sense. Given an initial state  $x_0 > 0$  and a fixed time period  $[t_0, t_1]$ , a fishing policy  $u(\cdot) : [t_0, t_1] \rightarrow \mathbb{R}_+$  may be called "admissible" if it leads to a non-negative solution  $x(t, u)$  of (3) for  $t \in [t_0, t_1]$  and "optimal" if it maximizes the overall catch during that period. Such an "optimal" fishing policy will, however, lead to depletion at time  $t_1$ . To prevent this one may wish to impose a "terminal constraint"  $x(t_1) \geq x_1$  where  $x_1 > 0$  is a lower bound to an acceptable fish population at the end of the period. Thus we end up with the following *optimal control problem*:

$$\text{Maximize } \int_{t_0}^{t_1} u(t)dt \text{ subject to } u(t) \geq 0, x(t, u) \geq 0, t \in [t_0, t_1], x(t_1) \geq x_1.$$

If  $u(t)$  is required to be constant, the problem is easily solved, see Ex. 2.1.15.

Another optimal control problem which can be solved by elementary means is the *optimal constant-effort harvesting problem*. Here the harvesting rate  $u(t)$  is by definition proportional to  $x(t)$ , i.e.  $u(t) = cx(t)$ . This is a simple example of *feedback control* where the control variable  $u(t)$  is determined as a given function of the instantaneous state  $x(t)$  of the system. Following this control strategy one obtains a Verhulst model in which the parameters have changed

$$\dot{x}(t) = r(K - c/r - x(t))x(t).$$

If  $c < rK$  there is an equilibrium solution  $x(t) = \bar{x} = K - c/r$ ,  $t \geq 0$  corresponding to the constant harvesting policy  $u(t) = c\bar{x}$ ,  $t \geq 0$ . Again one can determine the optimal constant harvesting policy which yields the highest sustainable harvesting rate, see Ex. 2.1.15.  $\square$

**Remark 1.1.2.** Although the logistic model is a widely used and successful model which predicts quite well the growth of various laboratory populations (see *Notes and References*), it is a highly simplified model. It is based on a number of assumptions which are not usually satisfied when the growth of a species in a real ecosystem is considered, e.g.

- (i) The influence of environmental factors on the growth of the species is assumed to be constant in time. But these factors and the behaviour of a species usually vary with the time of the year. Also there are often random variations in the environment.

- (ii) The effects of limited resources are assumed to affect all individuals of the species in an equal manner. A more realistic model would take the spatial distribution of the species and its resources into account (partial differential equations).
- (iii) It is assumed that the birth and death rates of the population respond instantly to the population size, whereas usually there is a delay between birth and the ability to give birth.
- (iv) The age distribution of the population is assumed to be constant or that if it changes it does not influence the growth of the species.

Although the assumptions are not realistic, highly simplified models like that of Verhulst are often of great scientific value. Their purpose is not to give an accurate portrait of an underlying real process but to enhance the understanding of some of its internal mechanisms. As such they can be more important motors for scientific progress than complex “realistic” simulation models<sup>1</sup>.  $\square$

Often the dynamics of a population are strongly influenced by the interaction with other populations in the same ecosystem. Several species may compete for the same natural resources or a species may be predatory on some species while serving as prey for others. In the following example we describe a classical predator-prey model due to Lotka and Volterra<sup>2</sup>.

**Example 1.1.3. (Predator-prey system).** Suppose that an island is populated by goats and wolves. The goats survive by eating the island’s vegetation and the wolves survive by eating the goats. Often oscillations are observed in the development of such predator-prey populations. If, initially, there are only a few wolves but many goats, the wolves have a lot to eat and the number of goats will be diminished while the number of wolves will increase until there are not enough goats to feed them. Then the number of wolves will be reduced so that the goats will be able to recover and this closes the cycle. The classical Lotka-Volterra model for such a predator-prey system is

$$\begin{aligned}\dot{x}_1 &= ax_1 - bx_1 x_2 \\ \dot{x}_2 &= -cx_2 + dx_1 x_2,\end{aligned}\tag{4}$$

where  $x_1$  and  $x_2$  are the densities (number per unit area) of the prey and predator populations respectively, and  $a, b, c, d$  are positive constants. The model mirrors a qualitative feature which has been observed in many real predator-prey systems, the persistence of periodic fluctuations. This is illustrated in Figure 1.1.2.  $\bar{x} = (c/d, a/b)$  is an equilibrium point of (4) and any initial state  $x^0 \neq \bar{x}$ ,  $x_1^0 > 0, x_2^0 > 0$  leads to a periodic trajectory cycling around this equilibrium point in the positive orthant.

Clearly, this is a simplistic model and does not aim at simulating or predicting a real process. The model is based on the following assumptions.

<sup>1</sup>“This work seeks to gain general ecological insights with the help of general mathematical models. That is to say the models aim not at realism in detail, but rather at providing mathematical metaphors for broad classes of phenomena. Such models can be useful in suggesting interesting experiments or data collecting enterprises, or just in sharpening discussion.” (R. M. May, Preface of “Stability and Complexity in Model Ecosystems”).

<sup>2</sup>The story of how Volterra came to design the model (independently of Lotka) is interesting. For many years fishermen had observed periodic fluctuations between sharks and their prey populations in the Adriatic Sea. During World War I, commercial fishing was greatly reduced and so it was expected that there would be plentiful fish stocks for harvesting after the war was over. Instead the catches of commercially valuable fish declined after the war while the number of sharks increased.

- (i) In the absence of predators the prey population grows exponentially with rate  $a$ .
- (ii) In the absence of prey the predator population decreases at the death rate  $c$ .
- (iii) The growth of the predator population depends affinely on the food intake, i.e. on predation.
- (iv) Predation depends on the likelihood that a victim is encountered by a predator and this likelihood is proportional to the product  $x_1x_2$  of the two populations' densities.

An assumption similar to (iv) is made in chemical kinetics where, according to the so-called law of mass action, the rate of molecular collisions of two substances in a given solution is assumed to be proportional to the product of their concentrations.

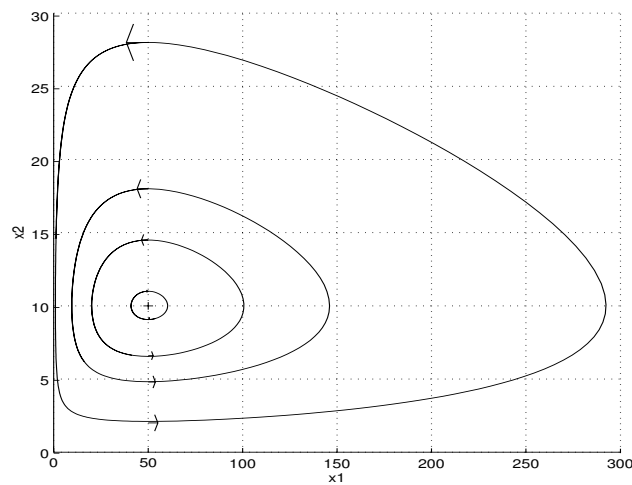


Figure 1.1.2: Predator-prey trajectories

Many “more realistic” models have been obtained from (4) by modifying the predator-free prey growth term  $ax_1$  to include crowding effects or by allowing for saturation effects and lags in the predators' response to increasing prey densities. For instance, in order to eliminate the assumption that the prey grows exponentially in the absence of predators one could introduce a term  $-ex_1^2$  in the first equation of (4) which accounts for the effect of crowding on the growth of the prey (see Example 1.1.1).

$$\begin{aligned}\dot{x}_1 &= ax_1 - bx_1x_2 - ex_1^2 = e(a/e - x_1)x_1 - bx_1x_2 \\ \dot{x}_2 &= -cx_2 + dx_1x_2.\end{aligned}\tag{5}$$

This drastically alters the qualitative behaviour of the predator-prey system. In the absence of predators the prey now evolves according to a logistic growth model with carrying capacity  $a/e$ . Moreover, the new system does not always have an equilibrium with positive coordinates. In fact the equilibrium equations are

$$(a - bx_2 - ex_1)x_1 = 0, \quad (-c + dx_1)x_2 = 0$$

and these equations have a (unique) positive solution  $\bar{x} = (c/d, (da - ec)/bd)$  if and only if  $a/e > c/d$ . Figure 1.1.3 illustrates the changed behaviour of the modified predator-prey system (5). In particular, it has no non-constant periodic solutions and its only

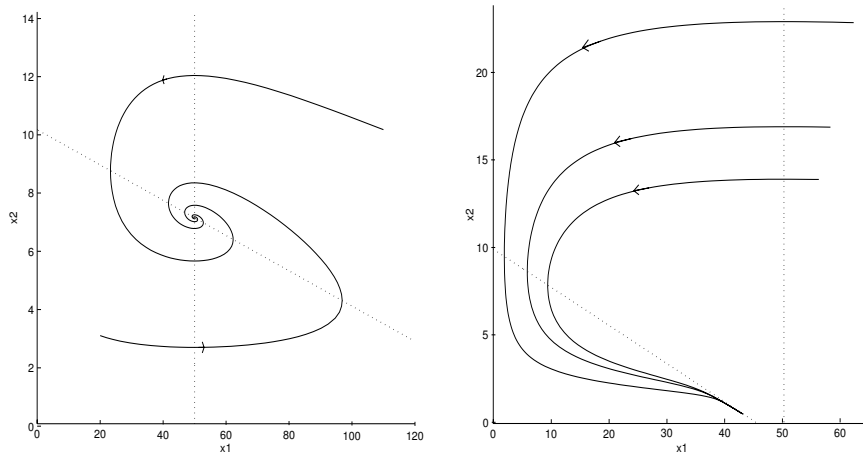


Figure 1.1.3: The effect of crowding

positive equilibrium point  $\bar{x} = (c/d, (ad - ec)/bd)$  is now asymptotically stable. It attracts all trajectories starting at initial states close to it, but not necessarily those starting further away, see Figure 1.1.3. In Chapter 3 we will show how the stability or instability of an equilibrium point can be examined for a given set of parameters. Using these results it is possible to show that the other two equilibrium points  $(0, 0)$  and  $(a/e, 0)$  are unstable. The qualitative changes between (4) and (5) do not depend on the size of  $e > 0$  which can be arbitrarily small. This shows that the classical predator-prey system is not structurally stable in the sense that a perturbation of the model, however small, may exhibit a qualitatively different behaviour.

In spite of their simplicity predator-prey systems and other models of two species are used in a number of control applications, e.g. in the management of renewable resources or in pest control where predators are introduced to control pests feeding on agricultural crops. Consider for instance a predator-prey system of salmon and herring in marine fishing. Choosing a suitable predator-prey model and adding control terms to both equations (catch rates) one may ask what are the optimal sustainable harvesting rates given the prices for salmon and herring, and what is the corresponding equilibrium point of the system, i.e. the stocks of salmon and herring which allow one to realize the optimal rates. If this optimal equilibrium point is found, The problem then arises of how the optimal equilibrium can be attained from a given initial population of salmon and herring by applying a suitable fishing policy. This is a controllability problem which we consider in Vol. II. In order to be of any practical value, the optimal equilibrium must be asymptotically stable since otherwise unavoidable small deviations of the populations from their optimal sizes may lead to large deviations from the optimal equilibrium solution. But asymptotic stability is not enough. It is important that this property is preserved under perturbations which reflect the uncertainties about the model and its parameters. This is a problem of robust stability which we will analyze in Chapter 5.  $\square$

### 1.1.1 Notes and References

#### Modelling in general

There are a number of introductions to dynamical systems which emphasize modelling. In particular, we recommend the book by *Luenberger* (1979) [349]. Many elementary

examples of control systems can be found in a collection of case studies by *MacClamroch* (1980) [369]. Additional information about modelling and a large number of examples can be found in textbooks on the analysis, modelling and design of dynamic systems, see for example *Ogata* (1992) [397], *Burton* (1994) [84], *Close and Frederick* (1995) [105].

The reader who is interested in *modelling techniques* for a variety of physical systems is referred to *Wellstead* (1979) [516], *Shearer et al.* (1967) [462], *MacFarlane* (1970) [355].

### Population Dynamics

A comprehensive textbook discussing dynamic models in various areas of biology and the life sciences is *Murray* (1993) [385], see also *Edelstein-Keshet* (1988) [146], *Hoppensteadt and Peskin* (1992) [263]. The book by Murray also contains an extensive bibliography.

Population Dynamics is one of the core subjects of mathematical biology and was amongst the first areas in life sciences which attracted mathematical methods. Classical references are *Malthus* (1798) [358], *Verhulst* (1938) [506], [347], *Volterra* (1927) [509] [510], *Kostitzin* (1934) [315] and *Kolmogoroff* (1936) [313]. English translations of some of their papers and brief discussions of their work can be found in [489].

Various empirical investigations have shown a good fit between the logistic model and the growth of actual laboratory populations, see e.g. *Lotka* (1924) [347] (*Drosophila*) and *Gause* (1959) [185] (*Paramecium caudatum*). A detailed discussion of the Verhulst model can be found in *May* (1981) [368]. The behaviour of the *discrete time logistic equation*  $x(t+1) = rx(t)(1-x(t))$  has been analyzed by means of cobweb diagrams in *Edelstein-Keshet* (1988) [146]. The qualitative features of the model change drastically at certain critical parameter values (bifurcation values) and for certain values of  $r$  chaotic behaviour is observed, see Ex. 3.1.15. A discussion of this model in the context of Population Dynamics can be found in *May* (1976) [367].

The controlled Verhulst equation (3) was used by *Schaefer* (1954) [449] to study the tuna fisheries of the tropical Pacific. It is probably the simplest dynamical model in Bio-economics (an interdisciplinary field which combines Mathematics, Biology and Economics), and has been used to study the effect of harvesting on growing populations. A standard reference on this subject is *Clark* (1976) [101], see also [102]. Control aspects are also important in bio-technology. A book on modelling bio-reactions and bio-reactors is *Nielsen and Villadsen* (1994) [392].

There is a large variety of models for interacting populations and some of them can already be found in the classical references above. These models play an important role in theoretical Ecology, see *Pielou* (1977) [411], *May* (1981) [368] and [366]. Important areas to be analyzed are the existence and stability of equilibria, the existence and stability of periodic solutions, their dependency on parameters, the effect of lags, the relationship between stability and complexity, the effect of competition, age structure and migration on growth rates, the extinction of species etc. An interesting mathematical discussion of various two species models is given in *Hirsch and Smale* (1974) [258]. The problem of robust stability or “resilience” is of particular interest in Ecology, for a discussion in the context of “complexity versus stability”, see *May* (1974) [366].

Supplemented with a control term population models are also used in the management of renewable resources, see *Clark* (1985) [102]. Other areas of application include Epidemiology *Bailey* (1975) [31]), theories of evolution *Hofbauer and Sigmund* (1988) [259], and pest control (rabies, weed dispersal, foot and mouth, etc.), see e.g. *Evans and Pritchard* (2001) [153] and the references therein.

## 1.2 Economics

In contrast to the previous examples we will now consider dynamic models which evolve in discrete time  $t = 0, 1, 2, \dots$ . The time axis is sub-divided in periods of equal length and  $x(t)$  denotes the value of  $x$  in the period  $t$ . Usually economic data is not available in a continuous way, but is given as a time series accumulated over certain periods (days, months, years,...). So discrete time models are particularly appropriate here.

**Example 1.2.1. (Cobweb model).** Supply and demand of a given commodity depend upon its price. With an increasing price  $p$  the supply  $S(p)$  increases whilst the demand  $D(p)$  decreases. Given the supply and demand curves of a commodity its equilibrium price will be that value  $\bar{p}$  which clears the market, i.e. the supply matches the demand. Thus  $\bar{p}$  is the abscissa of the intersection point of the supply and the demand curves, see Figure 1.2.1. This is a *static* supply and demand model for determining the price of a single commodity in a market. It remains unclear how this equilibrium price is actually realized by the interaction of sellers and buyers in the market place. But the model is not unreasonable if we assume that the commodity is not stored (and will perish if it is not sold). Let us now consider an economy where pork for example is produced for immediate

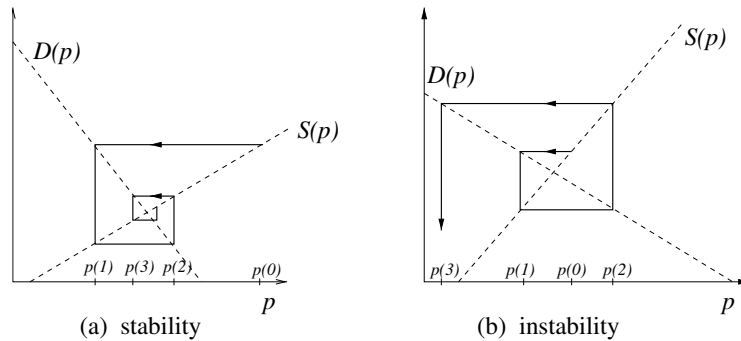


Figure 1.2.1: Cobweb Diagram

consumption and let us take into account the fact that the production (raising pigs) takes time. Choosing the production time as the basic period, the supply of pigs at time  $t \in \mathbb{N}$  will depend on the price  $p(t-1)$  valid at the time  $t-1$  when the decision was taken to produce the pigs for consumption in period  $t$ . On the other hand the actual demand for pork at time  $t$  depends on the current price  $p(t)$ . Let us assume – according to the above static model – that the price  $p(t)$  is determined in such a way that the complete supply is sold at time  $t$ . Assuming that the demand curve is strictly decreasing and its range contains the range of the supply function there will exist a unique value of  $p = \bar{p}$  for which this happens,  $\bar{p} = D^{-1}(S(\bar{p}))$ . Thus, starting with an initial price  $p(0) = p_0$  the prices  $p(t)$  will develop according to the difference equation

$$p(t+1) = D^{-1}(S(p(t))), \quad t \in \mathbb{N}, \quad p(0) = p_0. \quad (1)$$

Using the given supply and demand curves the solution  $p(\cdot)$  of this initial value problem can easily be constructed, see Figure 1.2.1. The initial price  $p(0)$  determines the supply



$S(p(0))$  at period 1 via the supply curve and this supply determines the equilibrium price  $p(1)$  which clears the market at period 1 as the unique solution  $p$  of  $D(p) = S(p(0))$ . Going through the same cycle with  $p(1)$  instead of  $p(0)$  and continuing the process we obtain a sequence  $(p(t))_{t \in \mathbb{N}}$  of prices. The corresponding time series of purchases/sales is given by  $D(p(t)) = S(p(t-1))$ ,  $t \in \mathbb{N}$ . The cobweb-like picture generated in this way led to the naming of the model.

$\bar{p}$  is an equilibrium of the above system i.e. a solution starting in  $\bar{p}$  will always remain there, if and only if, it is a fixed point of the function on the RHS<sup>1</sup> of the difference equation. Equivalently,  $S(\bar{p}) = D(\bar{p})$ . So  $(\bar{p}, D(\bar{p}))$  is just the intersection point of the demand and supply curves. In the situation depicted in Figure 1.2.1(a) the prices  $p(t)$  and purchases/sales  $D(p(t)) = S(p(t-1))$  converge towards the equilibrium values  $\bar{p}$  and  $D(\bar{p})$  as  $t \rightarrow \infty$  (asymptotic stability). The second picture shows that a different configuration of the two curves can lead to a diverging spiral around the equilibrium point (instability). This means that a small initial deviation of  $p(0)$  from  $\bar{p}$  will lead to ever larger oscillations of prices and purchases/sales around their equilibrium values. Here a weakness of the model becomes apparent since in this case prices will eventually become negative.

We now analyze the conditions under which stability and instability may occur. For this let us suppose that the supply and demand curves are linear,

$$D(p) = D_0 - ap, \quad S(p) = S_0 + bp$$

where  $D_0 \geq 0$ ,  $a, b > 0$  and  $S_0 \in \mathbb{R}$  are constants. Then the price  $p$  clearing the market with supply  $S > 0$  is determined by the equation  $D_0 - ap = S$ , i.e.  $p = (D_0 - S)/a$ . Hence the difference equation (1) reads

$$p(t+1) = (D_0 - S_0 - bp(t))/a = (D_0 - S_0)/a - (b/a)p(t), \quad t \in \mathbb{N}, \quad p(0) = p_0. \quad (2)$$

The corresponding equilibrium price is  $\bar{p} = (D_0 - S_0)/(a + b)$ . An easy calculation shows that the deviations from the equilibrium  $x(t) = p(t) - \bar{p}$  satisfy the difference equation

$$x(t+1) = -(b/a)x(t), \quad t \in \mathbb{N}, \quad x(0) = p_0 - \bar{p}.$$

The solution of this equation is  $x(t) = (-b/a)^t x(0)$  and so we have damped oscillations (asymptotic stability) if and only if  $b < a$  i.e. the demand curve is steeper than the supply curve. In economic terms this means that the consumers react more sensitively to changes in the price than the suppliers. Similarly we have instability if and only if  $b > a$ . Equality between the two parameters leads to periodic oscillations around the equilibrium.

The cobweb model assumes that the suppliers do not learn from past experience - in making their production decision they always expect the price in the next period to be equal to the present one. This is rather unrealistic. The following model, due to Goodwin, assumes that price expectations which guide the production decisions are modified by past experiences according to the rule

$$\hat{p}(t) = p(t-1) + \rho[p(t-1) - p(t-2)]$$

where  $\rho \in \mathbb{R}$  is a constant. The case  $\rho = 0$  corresponds to the cobweb model. Usually the value of  $\rho$  is chosen between  $-1$  and  $0$ , in which case the price is expected to move in the opposite direction to that of the previous period, i.e. suppliers expect oscillations in the price. If  $\rho > 0$  the price is expected to move in the same direction as in the previous

---

<sup>1</sup>RHS: right hand side, LHS: left hand side.

period. Assuming the same linear demand and supply curves as before we are led to the following difference equation for the price clearing the market at period  $t + 1$

$$p(t + 1) = (D_0 - S(\hat{p}(t)))/a = (D_0 - S_0)/a - (b/a) \{p(t - 1) + \rho[p(t - 1) - p(t - 2)]\}.$$

Equivalently

$$p(t + 1) = (1 + \rho)(b/a) p(t - 1) - (b/a) \rho p(t - 2) + (D_0 - S_0)/a. \quad (3)$$

This is a difference equation with two time lags and hence *two* initial values, say  $p(0)$  and  $p(1)$ , have to be specified to start up an iterative solution process. In the next chapter we will derive explicit formulas for the solutions of such equations. (3) has the same equilibrium solution  $p(t) \equiv \bar{p} = (D_0 - S_0)/(a + b)$  as (1), but now it is no longer immediately obvious how the asymptotic stability of this equilibrium depends on the parameters  $(a, b, \rho)$  of the system. In Chapter 3 we will develop methods for analyzing stability properties of equilibria of discrete time systems, see Ex. 3.3.16.  $\square$

The cobweb model is concerned with a micro-economic dynamical problem – the price dynamics in a single product market. In contrast we will now consider a model for the dynamics of a whole national economy. One would expect such a model to involve an enormous number of difference equations representing the production, pricing and consumption of a large variety of goods, incomes, saving and investment activities, tax flows and public expenditures etc. In fact such large, “realistic” models have been built in econometrics and have been used for economic forecasting and policy making. On the other hand highly aggregated models are used in theoretical macro-economics in order to gain insight into basic economic mechanisms. The next example deals with a classical model of the business cycle.

**Example 1.2.2. (Samuelson-Hicks multiplier-accelerator model).** We begin with a nonlinear version of the model. The basic variables are

$Y(t)$	the total national income (= national product) in year $t$
$C(t)$	the total consumer expenditure in year $t$
$I(t)$	the total (net) investment in year $t$
$G(t)$	the total government expenditure in year $t$ .

We make the following assumptions:

- (i) the total national product is the sum of consumer, investment and government expenditure,

$$Y(t) = C(t) + I(t) + G(t), \quad t \in \mathbb{N}, \quad (4)$$

- (ii) the consumer expenditure in year  $t + 1$  depends only on the income in the previous two years  $t$  and  $t - 1$ ,

$$C(t + 1) = f(Y(t), Y(t - 1)), \quad (5)$$

- (iii) the investment in year  $t + 1$  only depends on the increase of national income from year  $t - 1$  to year  $t$ ,

$$I(t + 1) = h(Y(t) - Y(t - 1)). \quad (6)$$

Substituting (6) and (5) in (4) gives

$$Y(t + 1) = f(Y(t), Y(t - 1)) + h(Y(t) - Y(t - 1)) + G(t + 1). \quad (7)$$

(7) is an example of a nonlinear second order difference equation. Given future government expenditure  $G(t)$ ,  $t = 2, 3, \dots$  and the national income  $Y(0)$ ,  $Y(1)$  in the initial two years one can solve (7) recursively to determine the future national income  $Y(t)$ ,  $t = 2, 3, \dots$ . Since the government is free to determine its expenditures (within certain constraints)  $G(t)$  represents a control variable.

Let us now suppose that  $Y(t) \equiv \bar{Y}$  is some given equilibrium solution of (7) corresponding to constant government expenditure  $G(t) = \bar{G}$ , i.e.

$$\bar{Y} = f(\bar{Y}, \bar{Y}) + h(0) + \bar{G}. \quad (8)$$

In order to analyze the system's behaviour close to this equilibrium solution let  $y(t) = Y(t) - \bar{Y}$ ,  $u(t) = G(t) - \bar{G}$  and assume that up to first order we have

$$f(\bar{Y} + y_1, \bar{Y} + y_2) \sim f(\bar{Y}, \bar{Y}) + c_1 y_1 + c_2 y_2, \quad h(y) \sim h(0) + ay.$$

The constant  $a$  is called the *acceleration coefficient*,  $c = c_1 + c_2$  the *marginal propensity to consume* and  $s = 1 - c$  the *marginal propensity to save* (it is always assumed that  $0 < c < 1$ ). Subtracting (8) from (7) we obtain to first order

$$y(t+1) = cy(t) + k(y(t) - y(t-1)) + u(t), \quad k = a - c_2. \quad (9)$$

This is the Samuelson-Hicks multiplier-accelerator model. It describes how the deviations  $y(t) = Y(t) - \bar{Y}$  of the actual national product from an equilibrium  $\bar{Y}$  evolves given the initial deviations  $y(0)$ ,  $y(1)$  and the deviation  $u(t) = G(t) - \bar{G}$  of the government expenditure from the constant value  $\bar{G}$ .

In the fifties considerable attention was paid to the possibility of "progressive expansion" of an economy in the presence of constant government expenditures. For the above linear model, this question is easily analyzed.  $y(t) = (1+r)^t y_0$  with  $r \in \mathbb{R}$ ,  $y_0 \neq 0$  solves (9) with  $u(t) \equiv 0$  for all  $t \in \mathbb{N}$  if and only if  $(1+r)^2 = c(1+r) + k(1+r-1)$ , i.e.

$$r^2 - (k - s - 1)r + s = 0, \quad (s = 1 - c).$$

This equation has a positive solution  $r$  (and hence (9) has a solution with constant growth rate  $r > 0$ ) if and only if

$$k - s - 1 > 0 \text{ and } (k - s - 1)^2 \geq 4s, \quad \text{i.e. } k \geq (1 + \sqrt{s})^2.$$

It was concluded from this result that, even with fixed government expenditure, an acceleration coefficient of moderate size could produce enough investment to make a constant growth rate of the national income possible. Although this result seems to be quite satisfactory at first sight, it must be regarded with some scepticism. The linear multiplier-accelerator model (9) is at best an appropriate model for small deviations  $y(t)$  from the equilibrium solution  $\bar{Y}$ . Therefore the solution  $y(t) = (1+r)^t y_0$  will, in the long run, move out of the neighbourhood of the origin where the model is meaningful. Adequate models for long term economic growth cannot be expected to be linear. Assuming the validity of the nonlinear model the significance of the above analysis for the long term behaviour of (7) is that the equilibrium solution  $Y(t) \equiv \bar{Y}$  is unstable if the parameters of the linearization (9) satisfy the inequality  $k = a - c_2 \geq (1 + \sqrt{s})^2$ . We will illustrate this in Chapter 3. Another question which is of obvious importance for the theory of the business cycle is to determine those values of the parameters  $a, c_1, c_2$  for which the solutions of the linear model are oscillatory. This question can be answered by applying the formulas for solutions derived in the next chapter or via the spectral analysis of Chapter 3.  $\square$

### 1.2.1 Notes and References

Standard references for dynamic models in Economics are *Allen* (1959) [9], *Gandolfo* (1980) [181]. The cobweb model can be found in these books and they also discuss models *with stocks* or *inventories* where supply and demand may be different. Goodwin's model which allows for the influence of past price changes on the suppliers' price expectations is described in [200]. In the econometric literature there are reports on single markets of a particular commodity where prices show an oscillatory behaviour similar to that generated by an undamped cobweb model.

The multiplier-accelerator model is discussed in most textbooks on Mathematical Economics and Macro-economics. The model was first described by *Samuelson* (1939) [446] and later elaborated by *Hicks* (1950) [227]. Various stabilization policies for these type of models have been suggested and analyzed by *Phillips* (1954) [409]. As in classical control engineering Phillips distinguishes between *proportional*, *derivative* and *integral* stabilization policies and analyzes their effects on the national income in the presence of constant external disturbances.

## 1.3 Mechanics

In this section we describe some mathematical models of simple mechanical systems. The modelling of such systems is based on the laws of classical mechanics and various techniques have been developed for this over the centuries. These methods have been corroborated by experiments and as a consequence reliable models are available for a great number of mechanical devices. We begin by describing a modelling technique which builds up an approximate *lumped model* of a mechanical system by representing it as an interconnection of ideal translational and rotational *elements* characterized by simple *constitutive laws*. To understand the interaction of these elements within the system, the forces and torques generated by the connection of one element with another must be considered. In a final subsection we briefly describe the variational (Lagrangian or Hamiltonian) approach to modelling which is based on energy considerations. Here the interconnecting forces and torques do not play a role. For this approach an elegant and powerful coordinate free framework has been developed in the general setting of symplectic manifolds, see *Notes and References*. However, an exposition of this framework is beyond the scope of this section. Instead we limit ourselves to a description of the variational method based on local (generalized) coordinates. We emphasize that the purpose of this section is not to give an introduction to classical mechanics, but to present some models of technical mechanical devices and sketch a few modelling techniques.

### 1.3.1 Translational Mechanical Systems

The dynamic behaviour of a mechanical system is described by vectors of displacements, velocities, forces and torques. A common modelling technique is to represent a mechanical system approximately as an interconnection of a finite number of idealized elements (masses,<sup>1</sup> springs, dampers, transformers and their rotational counterparts). The behaviour of each element is governed by a simple law relating the external force to the displacement, velocity or acceleration associated with the element. This law is called the *constitutive* relation or equation of the element.<sup>2</sup> Table 1.3.1 summarizes the constitutive laws for a pure mass, a linear spring and a linear damper. In the table arrows are associated with the forces. This does not mean that the forces are actually in these directions since the magnitude of  $F(t)$  may be negative. For example if for the spring  $y_{12}(t) > \bar{y}_{12}$  then the force required to produce the extension is in the direction shown. However if  $y_{12}(t) < \bar{y}_{12}$ , then one needs to compress the spring, so  $F(t) < 0$  and the force is actually in the opposite direction to the one shown.

For a single particle, *Newton's Second Law of Motion* states that *the sum of the forces acting on the particle is equal to the time rate of change of its linear momentum*. Therefore the constitutive law of a mass element is given by  $\frac{d}{dt}(mv(t)) = F(t)$ . Here

---

<sup>1</sup>It may seem strange to some readers that mass is regarded as a constitutive element of a mechanical system in parallel with springs and dashpots. This is, however, common practice in the modelling of engineering systems. The reader should distinguish between the fundamental concept of *mass* in theoretical mechanics and the notion of a *mass element* as a building block ("pure mass") in the modelling of a mechanical system.

<sup>2</sup>Throughout the present and the following section the predicate *constitutive* will only be used in this terminological sense, see [84], [105].

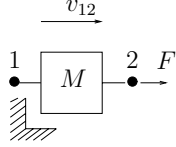
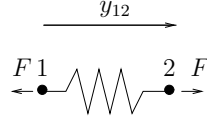
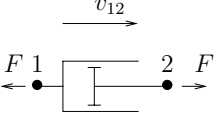
Symbol	Constitutive Law	Variables
	$\frac{d}{dt}(Mv(t)) = F(t)$	$v(t) = v_{12}(t)$ velocity of mass $F(t)$ force applied to mass
	$ky(t) = F(t)$	$y(t) = y_{12}(t) - \bar{y}_{12}$ net elongation $F(t)$ force applied to spring
	$cv(t) = F(t)$	$v(t) = v_{12}(t)$ relative velocity of piston $F(t)$ force applied to damper

Table 1.3.1: Symbols and constitutive laws of mass, spring, and damper

the velocity and acceleration must be measured with respect to an inertial reference frame (in classical mechanics this is usually fixed at the centre of the Sun).

The constitutive law of the linear spring is given by Hooke's law. In reality this linear relation between force and elongation will only be approximately valid within certain bounds on the elongation. Hence the use of a linear spring element in a mechanical model imposes constraints on the variables involved.

Similarly for models involving a damper. A physical realization of a linear damper is a dashpot where a piston moves through an oil-filled cylinder and there are holes in the face of the piston through which the oil passes. If the rates of flow are kept within certain bounds viscous damping results in a linear relation between the force and the relative velocity of the piston with respect to the cylinder. At higher velocities such a dashpot will show nonlinear characteristics.

The spring, damper and mass in the above table are also idealized objects from another point of view. Any real spring has some (albeit comparatively small) inertia and damping. Similarly any damper has some mass and exhibits small spring effects. We may account for the difference between the real devices and idealized objects by lumping all inertias of a given mechanical system together in the masses, all stiffness effects in the springs and all frictional forces in the dashpots ("lumped parameter model"). This lumped parameter approach to modelling a mechanical system is not limited to linear models. Nonlinear relations between stresses and deflections in a mechanical system can be modelled by nonlinear springs, and nonlinear viscous frictions between adjacent bodies can be modelled by nonlinear dampers.

If we describe a mechanical system as an interconnection of a finite number of masses, springs and dampers, a model of the overall system is obtained by combining the constitutive relations of its elements with the *interconnection laws* governing the interaction between them. Throughout this section we will assume that the forces between mechanical elements obey Newton's third law of action and reaction: *Any force of one element on another is accompanied by a reaction force on the first*

element of equal magnitude and opposite direction along the line joining them, see Table 1.3.1 where the forces on the left of the spring and damper symbols are the reaction forces to those on the right. There are various methods of obtaining the equations for the overall mechanical system from the constitutive relations of its elements and the *interconnection laws* e.g. bond graph methods and network methods, see Section 1.4. For more detailed information about this mass-spring-damper modelling approach, see *Notes and References*.

We now give a few examples of mechanical systems.

**Example 1.3.1. (Trolley).** Consider a trolley of mass  $M$  moving on rails under the influence of a force  $\beta u(t)$  as in Figure 1.3.2. Here  $\beta$  is a constant which converts the control variable  $u$  (e.g. a voltage) into a force. We neglect all frictions present in the system –

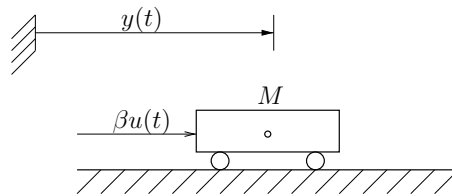


Figure 1.3.2: Pure mass: trolley

friction between wheels and rails, friction in the wheel bearings, drag friction of the trolley moving through the atmosphere. We also neglect the masses of the wheels and assume that the trolley behaves like a rigid body. Finally we assume that the line of action of the force is through the trolley's centre of mass, parallel to the rails. So the trolley does not rotate and, under the influence of gravity, does not lose contact with the rails. Since the mass of the trolley is constant Newton's second law yields the following scalar equation of motion,

$$M\ddot{y}(t) = \beta u(t) \quad (1)$$

where  $y(t)$  is the displacement of the centre of mass of the trolley from a fixed point in an inertial reference frame. In order to determine the motion of the system for  $t \geq 0$  it is necessary to know the initial position  $y(0)$  and the initial velocity  $\dot{y}(0)$  of the trolley. Moreover the exterior force  $\beta u(t)$  must be known as a function of time  $t \geq 0$ . If we consider the force as a control variable and fix a rest position at  $y = 0$  as the set point of the trolley, a typical control problem is to find a feedback control law  $u(t) = f(y(t), \dot{y}(t))$  which brings the trolley back or approximately back to the prescribed rest position from any given pair of initial values  $(y(0), \dot{y}(0))$ . If we assume that the control values are limited by  $|u(t)| \leq c$ ,  $t \geq 0$  where  $c > 0$  is a given constant, a typical optimal control problem is: given the initial conditions  $(y(0), \dot{y}(0))$ , find a control  $u(\cdot) : [0, t_1] \rightarrow [-c, c]$  which steers the trolley back to the rest position  $(y(t_1), \dot{y}(t_1)) = (0, 0)$  in minimal time  $t_1$ . Additionally constraints may be imposed on the trajectory of the trolley (e.g.  $|y(t)| \leq d$ ,  $d > 0$ ) and this leads to an *optimal control problem with state constraints*.  $\square$

In the next example we consider interconnections of mechanical elements. The harmonic oscillator is used as a highly simplified model for many technical systems. We illustrate this by a mass-spring-damper model for an automobile suspension system.

**Example 1.3.2. (Linear oscillator).** Consider the vertical motion of a mass  $M$  sliding in some bearing and suspended to a support by a spring as in the left hand figure in Figure 1.3.3. Besides the exterior forces (gravity and an additional time-depending force

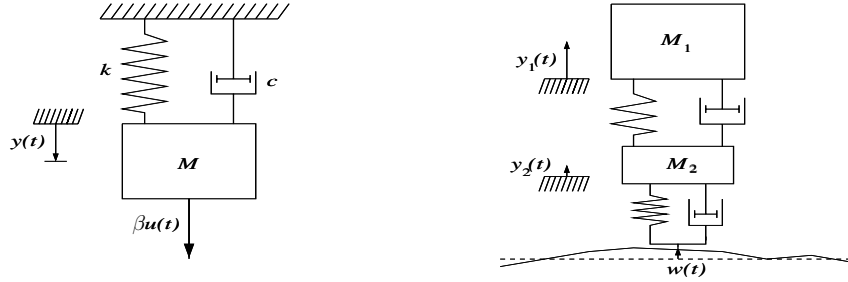


Figure 1.3.3: Mass-spring-damper systems

$\beta u(t)$ ) two types of interior forces act on the mass. These are modelled by a linear spring and a linear damper with coefficients  $k$  and  $c$ , respectively. Let us determine the equation of motion of the above mass-spring-damper system. The behaviour of the system is completely described by the vertical position and velocity of the mass. In order to eliminate the gravitational force we introduce the displacement  $y$  of the centre of mass from its equilibrium position under the influence of gravity. By Newton's second law the sum of the forces acting on  $M$  must equal  $M\ddot{y}$ . Note that the force exerted by the spring and the damper on the mass is opposite to the direction of the displacement and velocity respectively. The resulting equation of motion is

$$M\ddot{y}(t) + c\dot{y}(t) + ky(t) = \beta u(t). \quad (2)$$

We will now construct a simple mass-spring-damper model for an automobile suspension system. The purpose of such a suspension system is to smooth the response of the car body to the irregular ups and downs of the road. We will only consider the vertical movements of the car body and axles and make the non-realistic assumptions that both axles move in the same way so that they can be lumped together and the rotational motion of the car body can be ignored. We first assume that the road is flat. Since the car body and the axle can move independently, we need two position variables  $y_1$  and  $y_2$ . As reference points for these positions we choose the rest positions of the car body (mass  $M_1$ ) and of the axle (mass  $M_2$ ) over the nominal road level under the influence of gravity. The tyres are modelled as springs with comparatively high stiffness  $k_2$  coupled in parallel with a dashpot accounting for the energy dissipation through the tyres. The main suspension mechanism consisting of coil springs, leaf springs and shock absorbers, is modelled in a lumped manner by a linear spring and a linear damper connecting the axle with the car body, see the right hand figure in Figure 1.3.3. Let  $w(t)$  be the displacement of the point of contact between tyre and road from the nominal road level.  $w(t)$  is determined by the profile of the road and the position of the car. The tyre spring force (in an upward direction) corresponding to the deviation of the tyre from the nominal road level is  $k_2 w$ , the corresponding frictional force is  $c_2 \dot{w}$ . Applying Newton's second law to each of the two masses and Newton's third law to the interaction between the two masses we obtain the equations of motion

$$\begin{aligned} M_1 \ddot{y}_1 + c_1 (\dot{y}_1 - \dot{y}_2) + k_1 (y_1 - y_2) &= 0 \\ M_2 \ddot{y}_2 + c_1 (\dot{y}_2 - \dot{y}_1) + k_1 (y_2 - y_1) + c_2 \dot{y}_2 + k_2 y_2 &= c_2 \dot{w} + k_2 w. \end{aligned} \quad (3)$$



Here  $w(t)$  may be considered as a perturbation and an important design objective would be to ensure that the road conditions which the car is likely to encounter, will not generate vibrations of the car body i.e. values of  $y_1(t)$  and  $\dot{y}_1(t)$  which are not acceptable from the point of view of passenger comfort. If the suspension mechanism can be controlled, a typical problem would be to design a feedback control which decouples the vertical velocity of the car body  $\dot{y}_1(t)$  as much as possible from the (largely unknown) perturbations  $w(t)$  generated by the irregular road surface (*disturbance attenuation problem*).  $\square$

The previous two examples deal with translational mechanical systems whose movements are restricted to one direction. Arbitrary motions of a mass in three dimensional space are governed by a vector version of Newton's second law. Here and in the next section all vectors in  $\mathbb{R}^3$  or families of such vectors are written in bold face and we use vector analysis definitions and notations as found, for example, in [362]. We assume that the positions are determined with respect to a cartesian coordinate system which is fixed in an inertial frame. If the position vector of a particle of mass  $m$  at time  $t$  is denoted by  $\mathbf{r}(t)$  and  $\mathbf{F}(t)$  is the vector sum of all individual forces applied to the mass at time  $t$ , then

$$\dot{\mathbf{p}}(t) = (m\ddot{\mathbf{r}})(t) = \mathbf{F}(t), \quad (4)$$

where  $\mathbf{p} = m\dot{\mathbf{r}}$  is the *linear momentum* of the mass point.

Now consider a system of  $N$  particles with masses  $m_i$  at positions  $\mathbf{r}_i$ ,  $i \in \underline{N}$ . The linear momentum of such a system is by definition the sum of the linear momenta of each particle,

$$\mathbf{p}(t) = \sum_{i=1}^N \mathbf{p}_i(t) = \sum_{i=1}^N m_i \dot{\mathbf{r}}_i(t). \quad (5)$$

Applying Newton's second law to each of the particles we must distinguish between the external forces  $\mathbf{F}_i^e(t)$  and the interactive forces  $\mathbf{F}_{ij}(t)$  between the particles of the system. Summing over all particles we obtain from (4)

$$\dot{\mathbf{p}}(t) = \sum_{i=1}^N m_i \ddot{\mathbf{r}}_i(t) = \sum_{i=1}^N \mathbf{F}_i^e(t) + \sum_{i,j=1, i \neq j}^N \mathbf{F}_{ij}(t). \quad (6)$$

By Newton's third law of action and reaction  $\mathbf{F}_{ij}(t) + \mathbf{F}_{ji}(t)$  is zero for all  $t$  and  $i, j \in \underline{N}$ ,  $i \neq j$  and so the second term on the right vanishes. Hence, if we define the *total external force* and the *centre of mass* of the system at time  $t$  by

$$\mathbf{F}^e(t) = \sum_{i=1}^N \mathbf{F}_i^e(t), \quad \bar{\mathbf{r}}(t) = \sum_{i=1}^N \frac{m_i \mathbf{r}_i(t)}{M} \quad \text{where } M = \sum_{i=1}^N m_i \quad (7)$$

equations (5) and (6) can be written in the form

$$\mathbf{p}(t) = M\dot{\bar{\mathbf{r}}}(t) \quad \text{and} \quad \dot{\mathbf{p}}(t) = M\ddot{\bar{\mathbf{r}}}(t) = \mathbf{F}^e(t). \quad (8)$$

In particular, *the centre of mass of the system moves as if the total external force were acting on the entire mass of the system concentrated at the centre of mass.*

In order to describe a rigid body in three-dimensional space, the position of its centre of mass and the orientation of the rigid body with respect to an inertial reference frame must be specified. We therefore need a counterpart of Newton's second law for rotational motions.

### 1.3.2 Mechanical Systems with Rotational Elements

Consider a fixed point  $O$  in an inertial reference frame with origin  $O^*$ . The *angular momentum*  $\mathbf{H}(t)$  of a particle of mass  $m$  about the reference point  $O$  is defined by the vector product

$$\mathbf{H}(t) = \mathbf{r}(t) \times \mathbf{p}(t) = \mathbf{r}(t) \times m\dot{\mathbf{r}}(t)$$

where  $\mathbf{r}(t)$  is the “moment arm”, i.e. the vector from the point  $O$  to the position of the particle, and  $\mathbf{p}(t) = m\dot{\mathbf{r}}(t)$  is the linear momentum of the mass (with respect to the inertial reference frame). The corresponding *moment of force* or *torque*  $\mathbf{N}(t)$  due to the force  $\mathbf{F}(t)$  is defined by  $\mathbf{N}(t) = \mathbf{r}(t) \times \mathbf{F}(t)$ . As a consequence of (4) one obtains the following relation between the net torque applied to the particle and the rate of change of its angular momentum

$$\dot{\mathbf{H}}(t) = \frac{d}{dt}(\mathbf{r}(t) \times m\dot{\mathbf{r}}(t)) = \dot{\mathbf{r}}(t) \times m\dot{\mathbf{r}}(t) + \mathbf{r}(t) \times m\ddot{\mathbf{r}}(t) = \mathbf{r}(t) \times \mathbf{F}(t) = \mathbf{N}(t). \quad (9)$$

Note that the angular momentum and the torque both depend upon the point  $O$  about which moments are taken.

Let us now consider a system of  $N$  particles with the same setup as that which led to (8). The *total angular momentum* of such a system about  $O$  is obtained by summing up the angular momenta of all the particles, i.e.

$$\mathbf{H}(t) = \sum_{i=1}^N \mathbf{r}_i(t) \times m_i \dot{\mathbf{r}}_i(t),$$

so that by (6) and (9)

$$\dot{\mathbf{H}}(t) = \sum_{i=1}^N \mathbf{r}_i(t) \times m_i \ddot{\mathbf{r}}_i(t) = \sum_{i=1}^N \mathbf{r}_i(t) \times \left( \mathbf{F}_i^e(t) + \sum_{j=1, j \neq i}^N \mathbf{F}_{ij}(t) \right).$$

Now  $\mathbf{F}_{ij}(t) = -\mathbf{F}_{ji}(t)$  by Newton’s third law, and the same law implies that the vectors  $\mathbf{r}_i(t) - \mathbf{r}_j(t)$  and  $\mathbf{F}_{ij}(t)$  are linearly dependent. Hence, if

$$\mathbf{N}^e(t) = \sum_{i=1}^N \mathbf{r}_i(t) \times \mathbf{F}_i^e(t) \quad (10)$$

is the *total external torque*, then

$$\dot{\mathbf{H}}(t) = \mathbf{N}^e(t) + \sum_{i=1}^N \sum_{j=i+1}^N (\mathbf{r}_i(t) - \mathbf{r}_j(t)) \times \mathbf{F}_{ij}(t) = \mathbf{N}^e(t). \quad (11)$$

So the rate of change of the total angular momentum of a system of particles about a fixed point  $O$  is equal to the sum of the moments of the external forces about  $O$ .

By (8) the total linear momentum of a system of  $N$  particles is the same as if the entire mass were concentrated at the centre of mass and moving with it. We now develop a counterpart of this result for the angular momentum which includes the possibility that the point about which we take moments is moving. Let us denote this moving point by  $O_t$  and suppose the vector from  $O^*$  to  $O_t$  is  $\mathbf{r}^*(t)$ , the vector from  $O_t$  to the centre of mass is  $\bar{\mathbf{r}}(t)$ , the vector from the centre of mass to the  $i$ -th particle is  $\mathbf{r}'_i(t)$  and  $\mathbf{v}^*(t) = \dot{\mathbf{r}}^*(t)$ ,  $\bar{\mathbf{v}}(t) = \mathbf{v}^*(t) + \dot{\bar{\mathbf{r}}}(t)$ ,  $\mathbf{v}_i(t) = \bar{\mathbf{v}}(t) + \dot{\mathbf{r}}'_i(t)$  are the velocity vectors of  $O_t$ , of the centre of mass and of the  $i$ -th particle (with respect to the inertial frame). The angular momentum about  $O_t$  takes the form<sup>3</sup>

<sup>3</sup>In order to compactify the equations we drop the time variable  $t$  where necessary.

$$\begin{aligned}\mathbf{H} &= \sum_{i=1}^N (\bar{\mathbf{r}} + \mathbf{r}'_i) \times m_i (\bar{\mathbf{v}} + \dot{\mathbf{r}}'_i) \\ &= \sum_{i=1}^N \bar{\mathbf{r}} \times m_i \bar{\mathbf{v}} + \sum_{i=1}^N \mathbf{r}'_i \times m_i \dot{\mathbf{r}}'_i + \left( \sum_{i=1}^N m_i \mathbf{r}'_i \right) \times \bar{\mathbf{v}} + \bar{\mathbf{r}} \times \frac{d}{dt} \sum_{i=1}^N m_i \mathbf{r}'_i.\end{aligned}$$

Since  $\sum_{i=1}^N m_i \mathbf{r}'_i(t) = 0$  by the definition of the centre of mass (7), the last two terms on the right hand side vanish and we obtain

$$\mathbf{H}(t) = \bar{\mathbf{r}}(t) \times M \bar{\mathbf{v}}(t) + \sum_{i=1}^N \mathbf{r}'_i(t) \times m_i \dot{\mathbf{r}}'_i(t). \quad (12)$$

Note that by the above argument  $\mathbf{H}'(t) = \sum_{i=1}^N \mathbf{r}'_i(t) \times m_i \dot{\mathbf{r}}'_i(t) = \sum_{i=1}^N \mathbf{r}'_i(t) \times m_i \mathbf{v}_i(t)$  is the angular momentum of the system about the centre of mass. *Thus the total angular momentum of the system about  $O_t$  is the angular momentum of its total mass concentrated at the centre of mass, plus the angular momentum of the system about the centre of mass.* Only if the centre of mass is at rest (i.e.  $\bar{\mathbf{v}} = 0$ ) will the angular momentum be independent of the reference point  $O_t$  and its velocity. In this case  $\mathbf{H}(t)$  reduces to the angular momentum taken about the centre of mass. Differentiating  $\mathbf{H}(t) - \mathbf{H}'(t) = \bar{\mathbf{r}}(t) \times M \bar{\mathbf{v}}(t)$  we obtain

$$\dot{\mathbf{H}} - \dot{\mathbf{H}}' = \dot{\bar{\mathbf{r}}} \times M \bar{\mathbf{v}} + \bar{\mathbf{r}} \times M \dot{\bar{\mathbf{v}}} = (\bar{\mathbf{v}} - \mathbf{v}^*) \times M \bar{\mathbf{v}} + \bar{\mathbf{r}} \times M \dot{\bar{\mathbf{v}}} = -\mathbf{v}^* \times M \bar{\mathbf{v}} + \bar{\mathbf{r}} \times M \dot{\bar{\mathbf{v}}}.$$

In particular if  $O_t$  is the moving centre of mass we have  $\dot{\mathbf{H}}(t) = \dot{\mathbf{H}}'(t)$ . So *in calculating the rate of change of angular momentum of a particle system about its centre of mass, we may treat the centre of mass as if it were at rest.*

Let  $\mathbf{F}_i^e(t)$  be the external forces,  $\mathbf{F}^e(t) = \sum_{i=1}^N \mathbf{F}_i^e(t)$  the total external force and define the total torque about the moving reference point  $O_t$  by (see (10))

$$\mathbf{N}^e(t) = \sum_{i=1}^N (\bar{\mathbf{r}}(t) + \mathbf{r}'_i(t)) \times \mathbf{F}_i^e(t).$$

Then, if  $\mathbf{N}^{e*}(t)$  is the total torque about  $O^*$ , we get

$$\mathbf{N}^{e*}(t) = \sum_{i=1}^N (\mathbf{r}^*(t) + \bar{\mathbf{r}}(t) + \mathbf{r}'_i(t)) \times \mathbf{F}_i^e(t) = \mathbf{r}^*(t) \times \mathbf{F}^e(t) + \mathbf{N}^e(t) = \mathbf{r}^*(t) \times M \dot{\bar{\mathbf{v}}}(t) + \mathbf{N}^e(t).$$

since  $\mathbf{F}^e(t) = M \dot{\bar{\mathbf{v}}}(t)$ , see (8). The total angular momentum  $\mathbf{H}^*(t)$  about  $O^*$  satisfies

$$\mathbf{H}^*(t) = \sum_{i=1}^N (\mathbf{r}^*(t) + \bar{\mathbf{r}}(t) + \mathbf{r}'_i(t)) \times m_i \mathbf{v}_i(t) = \mathbf{H}(t) + \mathbf{r}^*(t) \times M \bar{\mathbf{v}}(t).$$

Since we have  $\dot{\mathbf{H}}^*(t) = \mathbf{r}^*(t) \times M \dot{\bar{\mathbf{v}}}(t) + \mathbf{N}^e(t)$  by (11) we get

$$\dot{\mathbf{H}} = \mathbf{r}^* \times M \dot{\bar{\mathbf{v}}} + \mathbf{N}^e - \dot{\mathbf{r}}^* \times M \bar{\mathbf{v}} - \mathbf{r}^* \times M \dot{\bar{\mathbf{v}}} = \mathbf{N}^e - \mathbf{v}^* \times M \bar{\mathbf{v}}. \quad (13)$$

In particular if  $O_t$  is the moving centre of mass we have  $\dot{\mathbf{H}}(t) = \mathbf{N}^e(t)$ . Therefore *the rate of change of the angular momentum of a particle system about its centre of mass is the sum of the moments about the centre of mass of all the external forces, irrespective of whether the centre of mass is moving or at rest.*

There is an *angular momentum law* for rigid bodies which complements Newton's second law. However we will not develop this for general rotational motions in  $\mathbb{R}^3$ , since in the following examples we only consider *plane* rotational systems. This means, in particular, that all the elements are rotating around axes which are parallel to each other and all forces are restricted to the plane. This assumption greatly simplifies the analysis. If we describe the motion of a system in an inertial reference frame where the  $z$  axis is parallel to the axes of rotation, then all vector products of vectors in the  $x, y$  plane are parallel to the  $z$  axis. As a consequence only the  $z$ -coordinates of these vector products are nontrivial. Now consider any particle of mass  $m$  rotating about an axis parallel to the  $z$  axis through a fixed point  $O = (x_0, y_0, 0)$  in the  $x, y$  plane and let  $(x_0, y_0, 0) + \mathbf{r}(t) = (x_0, y_0, 0) + (x(t), y(t), 0)$  be the coordinates of the particle at time  $t$ . Since by assumption the distance  $\|\mathbf{r}(t)\| = r = (x(t)^2 + y(t)^2)^{1/2}$  between the particle and the point  $O$  is constant we obtain by differentiation

$$x(t)\dot{x}(t) + y(t)\dot{y}(t) = 0.$$

Hence there exists a real number  $\omega(t)$  satisfying

$$\dot{\mathbf{r}}(t) = (\dot{x}(t), \dot{y}(t), 0) = \omega(t)(-y(t), x(t), 0).$$

Let  $\boldsymbol{\omega}(t) = (0, 0, \omega(t))$ , then  $\boldsymbol{\omega}$  is called the *angular velocity* of the particle about  $O$  and we obtain  $\dot{\mathbf{r}}(t) = \boldsymbol{\omega}(t) \times \mathbf{r}(t)$ . The angular momentum of the particle about  $O$  is  $\mathbf{r} \times m\dot{\mathbf{r}} = m\mathbf{r} \times (\boldsymbol{\omega} \times \mathbf{r}) = m(0, 0, \omega(x^2 + y^2))$ . Hence, for plane rotations, the equation of motion (9) is reduced to the scalar differential equation

$$\frac{d}{dt} [m\omega(t)(x(t)^2 + y(t)^2)] = mr^2\dot{\omega}(t) = N(t). \quad (14)$$

Here

$$N(t) = x(t)F_2(t) - y(t)F_1(t) \quad (15)$$

is the  $z$ -component of the torque generated by a given force  $\mathbf{F}(t) = (F_1(t), F_2(t), 0)$  applied to the particle.

Now consider a two dimensional rigid body  $B \subset \mathbb{R}^2$  rotating in the  $x, y$  plane about a perpendicular axis through a fixed point  $O$  with angular velocity  $\omega$ . Suppose that the rigid body has mass density  $\rho(x, y)$ ,  $(x, y) \in B$ . Then for this rigid body the angular momentum law takes the form

$$\frac{d}{dt}(J\omega)(t) = N(t) \quad \text{where} \quad J = \int_B \rho(x, y)(x^2 + y^2) dx dy \quad (16)$$

and  $J$  is the *moment of inertia* of the body about  $O$ . Moreover (16) also holds if  $O$  is a moving centre of mass. For many rigid bodies with uniform mass distribution the moments of inertia about given axes can be found in textbooks on analytic mechanics. The centre of mass  $(\bar{x}, \bar{y})$  and total mass  $M$  of a body  $B$  with mass distribution  $\rho(x, y)$  are given by

$$\bar{x} = \frac{1}{M} \int_B x\rho(x, y) dx dy, \quad \bar{y} = \frac{1}{M} \int_B y\rho(x, y) dx dy, \quad M = \int_B \rho(x, y) dx dy.$$

There is a close relationship between plane rotations and one-dimensional transla-

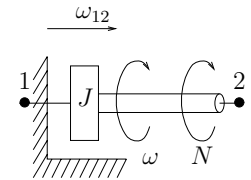
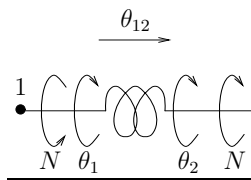
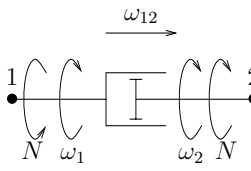
Symbol	Constitutive Law	Variables
	$\frac{d}{dt} (J\omega(t)) = N(t)$	$\omega(t) = \omega_{12}(t)$ angular velocity $N(t)$ torque applied about the axis
	$k\theta(t) = N(t)$	$\theta(t) = \theta_{12}(t) = \theta_2(t) - \theta_1(t)$ relative angular displacement of torsional spring $N(t)$ torque applied to spring
	$c\omega(t) = N(t)$	$\omega(t) = \omega_{12}(t) = \omega_2(t) - \omega_1(t)$ relative angular velocity $N(t)$ torque applied to damper

Table 1.3.4: Symbols and constitutive laws of rotational elements

tional motions. The rotational counterparts of *displacements*, *velocities* and *forces* are *angles*, *angular velocities* and *torques*. The rotational counterpart of mass is, as we have seen, the moment of inertia. Table 1.3.4 summarizes the rotational counterparts of masses, springs and dampers (again the directions indicated by the arrows are arbitrary since the values of the functions may be positive or negative).

Physical devices which may be modelled as rotational springs are, for example, the mainspring of a clock or an elastic rod joining two masses rotating about the same axis. Rotational viscous damping occurs for example if two concentric cylinders separated by an oil film rotate with different angular velocities about a common axis. The interconnection laws for rotational elements are strictly analogous to those for translational systems *if the interacting elements rotate about the same axis*. Then the torque exerted by one element on another is accompanied by a reaction torque of the same magnitude but of opposite direction on the first element. This holds, in general, but is no longer true if the elements rotate about different (albeit parallel) axes. For instance, the contact forces by which two gears act on one another are of equal magnitude and opposite direction, but the corresponding torques will be different if the radii of the gears are different.

In order to decide whether a rotation in the plane is positive or negative we have to fix an orientation of the plane (clockwise or anticlockwise)<sup>4</sup>. In the following examples we will always specify such an orientation. A directed angle (the direction being indicated by an arrowhead) is positive if it coincides with the given orientation of the plane, otherwise it is negative. The next example is a purely rotational

<sup>4</sup>Equivalently we could impose a direction to the axis of rotation and define the orientation of the plane by the right hand screw law.

mechanical system.

**Example 1.3.3. (Pendulum).** Consider a pendulum of length  $l$  suspended from a fixed point  $O$  as shown in Figure 1.3.5. We first model the pendulum as a point mass  $m$

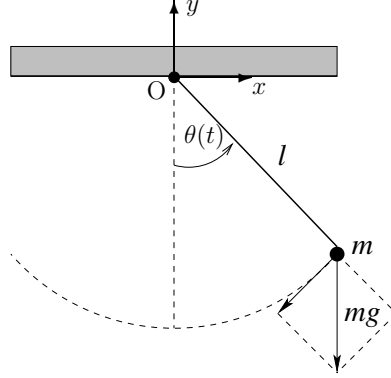


Figure 1.3.5: Pendulum

attached to a mass-less rigid rod of length  $l$  which rotates in the plane without any friction about  $O$ . Suppose that the directed angle from the downward vertical to the rod measured with respect to the anti-clockwise orientation is  $\theta$ . Then the motion of the pendulum is completely described by the angle  $\theta(t)$  as a function of time. Taking moments about  $O$  we obtain from (14) the following equation of motion

$$ml^2\ddot{\theta} = -mgl \sin \theta \quad (17)$$

where  $g$  is the gravitational constant.

Let us now abandon the assumption that the rod is mass-less and the rotation is without friction. Instead we assume that the pendulum is a plane rigid body of total mass  $m$  and there is viscous rotational friction with coefficient  $c$  at the pivot. Since the horizontal component of the gravitational force is zero, the torque about  $O$  exerted by the uniform gravitational field on the rigid body  $B(t)$  at time  $t$  is by (15)

$$\int_{B(t)} x\rho(x, y)g dx dy = mg\bar{x}(t),$$

i.e. the torque is equal to the torque about  $O$  exerted by the gravitational force on a particle of mass  $m$  located at the centre of mass  $(\bar{x}(t), \bar{y}(t))$  of the rigid body at time  $t$ , see Figure 1.3.6 (b). We therefore obtain from (16) the equation of motion

$$\frac{d}{dt}(J\dot{\theta})(t) = -c\dot{\theta}(t) - mg\bar{x}(t) = -c\dot{\theta}(t) - mg\bar{l} \sin \theta(t) \quad (18)$$

where  $J$  is the moment of inertia of the rigid body rotating about  $O$  and  $\bar{l}$  is the distance of the centre of mass from  $O$ . Note that this equation specializes to (17) if the rigid body is a particle and no friction is present. If, for example, the pendulum consists of a slender bar of length  $l$  and mass  $m$  uniformly distributed along the bar then the moment of inertia about  $O$  would be  $J = (1/3)ml^2$  and  $\bar{l} = (1/2)l$ .

Equations (17) and (18) are nonlinear time-invariant equations of second order. Given an initial angle  $\theta(0) = \theta_0$  and an initial angular velocity  $\dot{\theta}(0) = \dot{\theta}_0$  there exist unique solutions of (17) and (18) for all  $t \in \mathbb{R}$  satisfying these initial conditions. The angle  $\theta$  is treated

here as a real variable although only its values modulo  $2\pi$  matter. Both systems (17) and (18) have the same equilibrium solutions corresponding to the pendulum being in a vertical position (either downward or upright) with zero angular velocity: If either of the two systems satisfies the initial conditions  $(\theta(0), \dot{\theta}(0)) = (0, 0)$  or  $(\theta(0), \dot{\theta}(0)) = (\pi, 0)$  it will remain in this position indefinitely. However, the two equilibria exhibit very different behaviour when the initial conditions are slightly perturbed. If the pendulum is initially in the downward rest position a slight perturbation will only lead to small deviations from the equilibrium (see Example 2.1.10), whereas an arbitrarily small initial perturbation of the upper rest position will produce large deviations, because the pendulum will fall down. Hence the first equilibrium position is stable and the other is unstable. Whilst these statements hold for both systems considered here, there is an important difference between them with regard to their behaviour in a neighbourhood of the stable equilibrium point. In the presence of friction the pendulum will gradually return to the downward equilibrium position whereas it will swing with constant amplitude about the equilibrium in the absence of friction. To determine the stability properties of an equilibrium point for a given system is a basic problem in control theory. Since in most applications there are no simple analytic formulas for the solutions of the equation of motion one needs to find a method which allows one to determine the stability or instability of an equilibrium without solving the differential equations. Such a method has been developed by Liapunov whose central idea was to use the energy or an energy-like real valued function for this purpose. This method will be studied in detail in Chapter 3.  $\square$

The unstable upward position of a pendulum can be stabilized by a control mechanism which applies a torque  $N(t)$  to the pendulum depending on the deviation  $\theta(t) - \pi$  from the equilibrium position. A more interesting problem is to stabilize the inverted pendulum by moving its base e.g. in a horizontal or vertical way. This leads to a mechanical system which combines translational and rotational movements.

**Example 1.3.4. (Cart-pendulum system).** Consider a pendulum which rotates about a pivot which is mounted on a cart. The cart has mass  $M$  and is driven on a horizontal rail by a force  $\beta u(t)$  in the same way as the trolley in Example 1.3.1. However, here we allow for viscous friction between the cart and the rail. The centre of mass of the pendulum lies at a distance  $l$  from the pivot and the moment of inertia of the pendulum (modelled as a rigid body) about its centre of mass is  $J$ . We allow for viscous friction at the pivot point. The position of the cart is measured by the horizontal displacement  $r$  of its centre of mass from the origin of an inertial coordinate system. We assume that the centre of mass of the cart is moving along the  $x$ -axis of this coordinate system. The position of the pendulum is measured by the angular displacement  $\theta$  of the line joining its centre of mass with the pivot from the downward vertical (measured in an anti-clockwise direction). Although we view the cart as a rigid body we assume that its motion is one-dimensional, i.e. the torques generated by the totality of forces acting on the cart are in balance. This means that we can neglect the moments about its centre of mass and treat the cart as a point mass. To simplify the notation we assume that the pivot point coincides with the centre of mass of the cart.

In order to obtain a model of the system we will use free-body diagrams for each element, representing all external and interactive forces between the elements by symbols together with arrows which define their “positive senses”, see Figure 1.3.6: The forces are positive if they operate in the directions shown, they are negative if they operate in the opposite

direction. For instance, if the system is at rest in the downward position, the force  $F_2(t)$  acting on the pendulum at the hinge will be directed upwards and hence it will be positive with respect to the direction indicated in Figure 1.3.6. In general, all forces are vectors but since the cart's motion is restricted to one dimension we decompose the forces into their horizontal and vertical components.

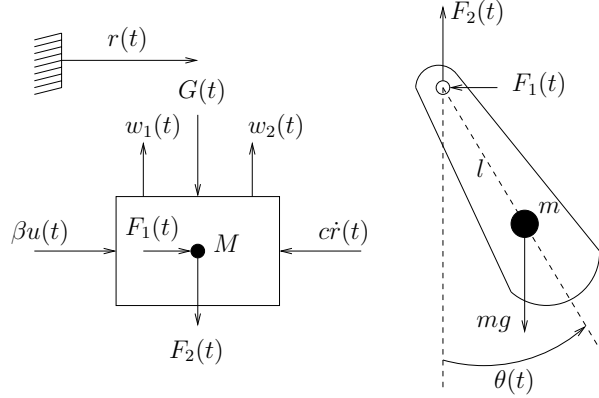


Figure 1.3.6: Free-body diagrams of cart and pendulum

The horizontal forces on the cart are the driving force  $\beta u(t)$ , the viscous friction force  $-c\dot{r}(t)$  and the horizontal component of the (unknown) contact force,  $F_1(t)$  at the pivot. The vertical forces on the cart are the contact forces  $w_1(t)$ ,  $w_2(t)$  through the wheels supporting the cart on the rail, the gravitational force  $G(t)$  and the vertical component of the (unknown) contact force,  $F_2(t)$  at the pivot. Since we assume that the cart is constrained to the one-dimensional motion along the rails, the vertical forces on the cart are in balance. The horizontal motion of the cart is governed by the equation

$$M\ddot{r}(t) = \beta u(t) - c\dot{r}(t) + F_1(t). \quad (19)$$

In order to describe the planar motion of the pendulum it is sufficient to consider the motion of its centre of mass and its rotation about its centre of mass. If  $(x(t), y(t))$  denotes the coordinates of the centre of mass at time  $t$  with respect to the given inertial coordinate system, then (see (8)) the horizontal and vertical motions are determined by

$$\begin{aligned} m\ddot{x}(t) &= m \frac{d^2}{dt^2} [\dot{r}(t) + l(\sin \theta(t))] = -F_1(t) \\ m\ddot{y}(t) &= ml \frac{d^2}{dt^2} (-\cos \theta(t)) = -mg + F_2(t). \end{aligned}$$

Calculating the double derivatives we obtain

$$m [\ddot{r}(t) + l\ddot{\theta}(t) \cos \theta(t) - l\dot{\theta}(t)^2 \sin \theta(t)] = -F_1(t) \quad (20)$$

$$ml [\ddot{\theta}(t) \sin \theta(t) + \dot{\theta}(t)^2 \cos \theta(t)] = -mg + F_2(t). \quad (21)$$

Since the cart-pendulum system is described by two independent variables,  $r(t)$  and  $\theta(t)$ , and since the two contact forces  $F_1(t)$ ,  $F_2(t)$  are unknown we need one more equation of motion. It remains to determine the rotation of the pendulum about its centre of mass.



The gravitational force does not produce any torque on the pendulum about its centre of mass,  $G$ . So the rotation of the pendulum is determined by the torque of the force  $\mathbf{F}(t) = (-F_1(t), F_2(t))$  about  $(x(t), y(t))$ . Now the vector from  $(x(t), y(t))$  to the pivot where the force  $\mathbf{F}(t)$  is applied is given by  $(-l \sin \theta(t), l \cos \theta(t))$  and so the torque of the force  $\mathbf{F}(t)$  about  $(x(t), y(t))$  is  $-F_2(t)l \sin \theta(t) + F_1(t)l \cos \theta(t)$ , see (15). The force of rotational friction  $c_P \dot{\theta}(t)$  acts to oppose the motion. Therefore we obtain from (16) with  $O$  the centre of mass

$$J\ddot{\theta}(t) = -F_2(t)l \sin \theta(t) + F_1(t)l \cos \theta(t) - c_P \dot{\theta}(t). \quad (22)$$

Using (20) and (21) to express the unknown reaction forces  $F_1(t), F_2(t)$  between the cart and the pendulum and replacing  $F_1(t), F_2(t)$  by these expressions in (19), (22) we obtain the following two equations which describe the dynamic behaviour of the cart-pendulum system (we drop the time variable)

$$\begin{aligned} (M+m)\ddot{r} + (ml \cos \theta)\ddot{\theta} + c\dot{r} - ml\dot{\theta}^2 \sin \theta &= \beta u \\ (ml \cos \theta)\ddot{r} + (J + ml^2)\ddot{\theta} + c_P \dot{\theta} + mgl \sin \theta &= 0. \end{aligned} \quad (23)$$

Subtracting suitable multiples of these equations from each other in order to eliminate firstly  $\ddot{\theta}$  and then  $\ddot{r}$  one obtains the equivalent equations

$$\begin{aligned} M(\theta)\ddot{r} &= (J + ml^2)(\beta u - c\dot{r} + ml\dot{\theta}^2 \sin \theta) + ml \cos \theta (mgl \sin \theta + c_P \dot{\theta}) \\ M(\theta)\ddot{\theta} &= -ml \cos \theta (\beta u - c\dot{r} + ml\dot{\theta}^2 \sin \theta) - (M+m)(c_P \dot{\theta} + mgl \sin \theta) \end{aligned} \quad (24)$$

where

$$M(\theta) = (M+m)J + ml^2M + m^2l^2 \sin^2 \theta.$$

Setting

$$x_1(t) = r(t), \quad x_2(t) = \theta(t), \quad x_3(t) = \dot{r}(t), \quad x_4(t) = \dot{\theta}(t)$$

yields the following system of nonlinear first order differential equations for the cart pendulum system

$$\begin{aligned} \dot{x}_1 &= x_3, & \dot{x}_2 &= x_4 \\ \dot{x}_3 &= \frac{1}{M(x_2)} [(J + ml^2)(\beta u - cx_3 + mlx_4^2 \sin x_2) + ml \cos x_2 (mgl \sin x_2 + c_P x_4)] \\ \dot{x}_4 &= \frac{-ml \cos x_2}{M(x_2)} (\beta u - cx_3 + mlx_4^2 \sin x_2) - \frac{(M+m)}{M(x_2)} [c_P x_4 + mgl \sin x_2]. \end{aligned} \quad (25)$$

If  $u(t) \equiv 0$  the system will remain at rest provided that the initial velocities  $x_3(0) = \dot{r}(0)$ ,  $x_4(0) = \dot{\theta}(0)$  are zero and the initial angular displacement  $x_2(0) = \theta(0)$  is either zero or  $\pi$ . Cart pendulum systems which are required to operate close to these equilibrium positions occur in practice. For instance, consider a loading plant (see Figure 1.3.7(a)) where a grab is suspended from a cart rolling on horizontal rails. These plants operate around the downward position of the pendulum and are required to be close to this equilibrium before putting down the load. On the other hand consider the balancing problem illustrated by the inverse pendulum in Figure 1.3.7(b). Such inverse pendulum systems are used in university laboratories for experimentation with controllers which stabilize the system at the upward position. A more practical example of a three dimensional balancing problem is that of the control of a rocket in an upright position in preparation for launch. Another (not so obvious) example is that of maintaining a satellite in a prescribed orbit (see Example 2.1.27). For the inverted pendulum shown in Figure 1.3.7(b) it is usual to

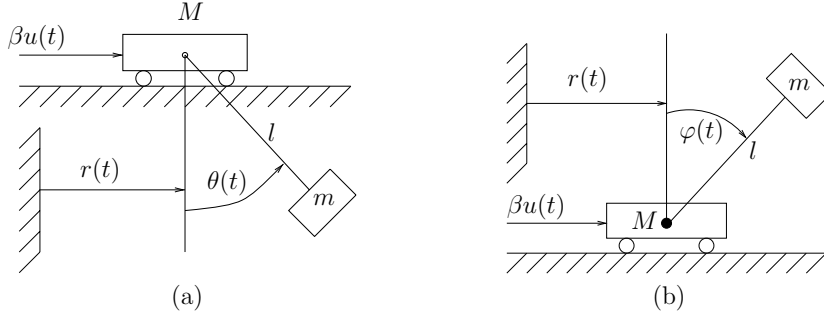


Figure 1.3.7: (a) Loading plant and (b) Inverted pendulum

express the equations of motion in terms of the angle<sup>5</sup>  $\varphi = \theta - \pi$  (the deviation of  $\theta$  from the equilibrium value  $\pi$ ). Setting  $\theta = \pi + \varphi$  in (24) yields

$$\begin{aligned} M(\varphi)\ddot{r} &= (J + ml^2)(\beta u - c\dot{r} - ml\dot{\varphi}^2 \sin \varphi) - ml \cos \varphi (-mgl \sin \varphi + c_P \dot{\varphi}) \\ M(\varphi)\ddot{\varphi} &= ml \cos \varphi (\beta u - c\dot{r} - ml\dot{\varphi}^2 \sin \varphi) - (M + m)(c_P \dot{\varphi} - mgl \sin \varphi). \end{aligned} \quad (26)$$

Let

$$x_1(t) = r(t), \quad x_2(t) = \varphi(t), \quad x_3(t) = \dot{r}(t), \quad x_4(t) = \dot{\varphi}(t)$$

then one obtains a system of nonlinear first order equations similar to (25), but with a different sign pattern.

$$\begin{aligned} \dot{x}_1 &= x_3, & \dot{x}_2 &= x_4 \\ \dot{x}_3 &= \frac{1}{M(x_2)} [(J + ml^2)(\beta u - cx_3 - mlx_4^2 \sin x_2) - ml \cos x_2 (-mgl \sin x_2 + c_P x_4)] \\ \dot{x}_4 &= \frac{ml \cos x_2}{M(x_2)} (\beta u - cx_3 - mlx_4^2 \sin x_2) - \frac{(M + m)}{M(x_2)} [c_P x_4 - mgl \sin x_2]. \end{aligned} \quad (27)$$

Now assume that for the loading plant  $|x_2(t)| = |\theta(t)|$  and  $|x_4(t)| = |\dot{\theta}(t)|$  remain sufficiently small so that

$$\sin x_2(t) \approx x_2(t), \quad \cos x_2(t) \approx 1, \quad x_4(t)^2 \sin x_2(t) \approx 0, \quad \sin^2 x_2(t) \approx 0. \quad (28)$$

Then  $M(x_2) \approx M_0 = (M + m)J + ml^2M$  is approximately constant and we obtain the following approximate linear equation of motion for the loading plant (pendulum down)

$$\dot{x} = Ax + bu, \quad (29)$$

where  $x(t) = [r(t), \theta(t), \dot{r}(t), \dot{\theta}(t)]^\top$  and

$$A = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & a_{32} & a_{33} & a_{34} \\ 0 & a_{42} & a_{43} & a_{44} \end{bmatrix}, \quad b = \begin{bmatrix} 0 \\ 0 \\ b_3 \\ b_4 \end{bmatrix}. \quad (30)$$

With constant entries

$$\begin{aligned} a_{32} &= M_0^{-1}m^2l^2g, & a_{33} &= -M_0^{-1}(J + ml^2)c, & a_{34} &= M_0^{-1}mlc_P, \\ a_{42} &= -M_0^{-1}(M + m)mgl, & a_{43} &= M_0^{-1}mlc, \\ a_{44} &= -M_0^{-1}(M + m)c_P, & b_3 &= M_0^{-1}(J + ml^2)\beta, & b_4 &= -M_0^{-1}ml\beta. \end{aligned} \quad (31)$$

<sup>5</sup>Note that the angle  $\varphi(t)$  as depicted in Figure 1.3.7 (b) is negative.

For the inverted pendulum if  $|\varphi(t)|$  and  $|\dot{\varphi}(t)|$  remain small, then (28) again holds but now  $x_2(t) = \varphi(t)$  and  $x_4(t) = \dot{\varphi}(t)$ . The approximate linear model has matrices of the same form as (30), however some of the matrix entries have different signs

$$\begin{aligned} a_{32} &= M_0^{-1}m^2l^2g, & a_{33} &= -M_0^{-1}(J + ml^2)c, & a_{34} &= -M_0^{-1}mlc_P, \\ a_{42} &= M_0^{-1}(M + m)mgl, & a_{43} &= -M_0^{-1}mlc, \\ a_{44} &= -M_0^{-1}(M + m)c_P, & b_3 &= M_0^{-1}(J + ml^2)\beta, & b_4 &= M_0^{-1}ml\beta. \end{aligned} \quad (32)$$

For the purpose of automatic control, sensors are required which provide continuous information about the current state of the system. Let us consider the balancing problem for the inverted pendulum and suppose that we can measure the current values of  $r(t)$ ,  $\varphi(t)$ . These measurements (“outputs”) are related to the “state”  $x(t) = [x_1(t), x_2(t), x_3(t), x_4(t)]^T$  by the *output* or *measurement* equation

$$y = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} x. \quad (33)$$

The balancing problem consists in designing a *regulator* which keeps the pendulum in an upright position at a fixed value of  $r$ , say 0. The regulator accepts the values  $y(t) \in \mathbb{R}^2$  as input values and produces the values  $u(t) \in \mathbb{R}^1$  as output values. This must be done in such a way that the inherently unstable equilibrium  $x_e = [0, 0, 0, 0]^T$  becomes a stable equilibrium of the feedback system. This *stabilization problem* can be solved using the linearized equations about the equilibrium state  $x_e$ . Linear models are often sufficient in order to design stabilizing controllers even if the underlying system is nonlinear. By keeping the system close to the equilibrium position the controller ensures that the linearized model yields a good approximation of the nonlinear dynamics. This partially explains the surprising success of linear models in feedback control.

Another control problem is best explained for the loading plant. If it is required to position a load accurately at a certain point the question arises whether there exists a control function  $u(\cdot)$  which steers the system from any given initial position to the desired final position in finite time. Additionally, it will be required that the load is at rest at the final position. This is a *controllability* problem. Note that for this problem the use of a linear model is questionable since this is a global problem and its solution requires a model which is accurate for a wide range of values of the system’s state.  $\square$

### 1.3.3 The Variational Method

The previous example illustrates that even for an apparently simple mechanical device it is by no means trivial to find the equations of motion by analyzing the system as an interconnection of masses, springs and dampers. The interconnection of translational and rotational elements poses particular problems. The main difficulty in the modelling process is that the interaction between the elements must be described by introducing “contact forces” or “forces of constraint”, which are not given a priori. They are among the unknowns of the problem and must be eliminated in order to get the system equations. Often the interconnective constraints are quite complicated and if there is a large number of them the above modelling procedure becomes cumbersome. For such cases an alternative procedure is available which is based on energy considerations. As a preparation we need some formulas for the energies of translational masses, springs and dampers and their rotational counterparts which we present in the next example. Additionally we discuss the kinetic and potential energy of a rigid body moving in three-dimensional space.

**Example 1.3.5.** The kinetic energy associated with a point mass  $m$  moving with velocity  $\mathbf{v}(t)$  at time  $t$  is

$$\mathcal{T}(t) = (m/2) \|\mathbf{v}(t)\|^2 = (m/2) \langle \mathbf{v}(t), \mathbf{v}(t) \rangle.$$

For arbitrary motions of a rigid body in three dimensional space the determination of the kinetic energy is more complicated. First consider a rigid body composed of  $N$  point masses  $m_i$ . Let  $\bar{\mathbf{r}}(t)$  be the position of the centre of mass of the body at time  $t$  (with respect to some inertial coordinate system) and fix a coordinate system in the body (moving with the body) whose origin is at the centre of mass. Suppose the body is rotating about an axis through the centre of mass with angular velocity  $\boldsymbol{\omega}(t)$ . So  $\boldsymbol{\omega}(t) \in \mathbb{R}^3$  points in the *direction of the instantaneous axis of rotation* of the body about its centre of mass given by the right hand screw law. If  $\bar{\mathbf{r}}_i$  is the (constant) coordinate vector of the point mass  $m_i$  of the rigid body with respect to the body coordinates then the position vector of this point with respect to the inertial reference system is  $\mathbf{r}_i(t) = \bar{\mathbf{r}}(t) + \bar{\mathbf{r}}_i$  and the velocity vector of the point (with respect to the given inertial coordinate system) is

$$\mathbf{v}_i(t) = \dot{\mathbf{r}}_i(t) = \dot{\bar{\mathbf{r}}}(t) + \boldsymbol{\omega}(t) \times \bar{\mathbf{r}}_i.$$

Hence the kinetic energy is

$$\begin{aligned} \mathcal{T}(t) &= \sum_{i=1}^N (m_i/2) \|\mathbf{v}_i(t)\|^2 = \sum_{i=1}^N (m_i/2) \langle \dot{\bar{\mathbf{r}}}(t) + \boldsymbol{\omega}(t) \times \bar{\mathbf{r}}_i, \dot{\bar{\mathbf{r}}}(t) + \boldsymbol{\omega}(t) \times \bar{\mathbf{r}}_i \rangle \\ &= (M/2) \|\dot{\bar{\mathbf{r}}}(t)\|^2 + \left\langle \dot{\bar{\mathbf{r}}}(t), \boldsymbol{\omega}(t) \times \sum_{i=1}^N m_i \bar{\mathbf{r}}_i \right\rangle + \sum_{i=1}^N (m_i/2) \|\boldsymbol{\omega}(t) \times \bar{\mathbf{r}}_i\|^2, \end{aligned}$$

where  $M = \sum_{i=1}^N m_i$  is the total mass. The middle term in the above expression is zero since  $\bar{\mathbf{r}}_i$  is the vector from the centre of mass to the  $i$ -th point mass. The last term is a quadratic form in  $\boldsymbol{\omega}(t)$ , so we may write

$$\mathcal{T}(t) = (M/2) \|\dot{\bar{\mathbf{r}}}(t)\|^2 + (1/2) \langle \boldsymbol{\omega}(t), \mathbf{J}\boldsymbol{\omega}(t) \rangle,$$

where  $\mathbf{J} = \mathbf{J}^\top \in \mathbb{R}^{3 \times 3}$  is called the *moment of inertia matrix* of the rigid body. This analysis can be extended to continuous distributions and hence the above formula for the kinetic energy holds for arbitrary rigid bodies. So the kinetic energy of a rigid body is the kinetic energy obtained if all the mass of the body were concentrated at the centre of mass, plus the kinetic energy of its motion about the centre of mass.

We now consider *potential energy*. The potential energy stored in a translational or rotational spring displaced from its equilibrium state is equal to the work done to achieve this displacement. If the spring is translational and linear its potential energy at a displacement  $y$  is given by  $(k/2)y^2$ . Similarly the potential energy of a linear torsional spring (where the torque is  $k\theta$ ) at an angular displacement  $\theta$  is  $(k/2)\theta^2$ . Note that an ideal spring does not have kinetic energy since its mass is zero.

The potential energy of any point mass is defined relative to a given conservative field of force to which it is subjected, e.g. the gravitational field of the Earth. If  $\mathbf{F} : \mathbb{R}^3 \rightarrow \mathbb{R}^3$  is a *conservative* field of force then the work done by moving a point mass from  $\mathbf{a} \in \mathbb{R}^3$  to  $\mathbf{b} \in \mathbb{R}^3$  only depends upon the points  $\mathbf{a}, \mathbf{b} \in \mathbb{R}^3$  and not on the path along which the mass has been moved. Fixing a reference point  $O$  the potential energy of a particle positioned at a point  $P$  is, by definition, equal to the work needed in order to move the particle within the force field from  $O$  to  $P$ . The potential energy of a system of  $N$  point masses at positions  $\mathbf{r}_1, \dots, \mathbf{r}_N$  is simply the sum of the individual potential energies. Approximating a rigid body of mass  $M$  by a system of point masses we see that at an altitude  $h$  above

the Earth ( $h$  not too large) its potential energy with respect to the gravitational field of the Earth is approximately  $Mgh$ . Note that the potential energy of a body relative to a conservative force field is only determined up to a constant depending on the reference point. For any system of point masses moving in a conservative field, if no energy dissipation occurs, then the sum of the kinetic and potential energies is constant in time.

Usually, dissipation of energy occurs because kinetic energy is transformed to thermal energy by friction. Frictional forces have to be overcome whenever bodies in contact have a relative velocity. A pure dissipator (damper) is an idealized object in which there is no kinetic or potential energy storage. However there is a dissipation of energy, for example the power absorbed at time  $t$  by a linear translational damper (where the force is  $cv$ ) is  $cv(t)^2$ . Similarly the power absorbed by a linear rotational damper (where the torque is  $c\omega$ ) is  $c\omega(t)^2$ . More generally, suppose that there is a system of  $N$  particles moving with velocities  $\mathbf{v}_i(t) \in \mathbb{R}^3$ ,  $i \in \underline{N}$  and that the particles are subjected to frictional forces which depend linearly on the velocities,  $\mathbf{F}_i(t) = c_i \mathbf{v}_i(t)$ , then the total energy dissipated is  $\sum_{i=1}^N c_i \|\mathbf{v}_i(t)\|^2$ .  $\square$

The variational method has been developed in the context of classical mechanics by, amongst many, Lagrange and Hamilton. We will not explain the derivation of the method here, nor discuss it in detail, but just sketch the essential steps to be followed. For a careful mathematical treatment, see *Notes and References*.

As the previous examples illustrate, the position vectors  $\mathbf{r}_i(t)$  of the point masses of a mechanical system are usually not free to vary independently of each other. The constraints which limit their movements may be classified in various ways. If they can be expressed by equations of the form  $f(\mathbf{r}_1, \dots, \mathbf{r}_N, t) = 0$  they are called *holonomic*. A typical example is given by a rigid body where all the distances between its mass points are constant in time. Another example is given by a particle which moves along a curve (a bead sliding on a wire) or on a surface. Nonholonomic constraints are obtained if not only position but also velocity coordinates enter the constraint equation or if the constraint takes the form of an inequality (for example, gas molecules within a container). We will only consider holonomic constraints. All the constraints in our mechanical examples are of this type.

Now suppose that a system of  $N$  particles is given, together with a number of holonomic constraints of the form

$$f_j(\mathbf{r}_1, \dots, \mathbf{r}_N, t) = 0, \quad j \in \underline{m} \quad (34)$$

where  $\mathbf{r}_i \in \mathbb{R}^3$  denotes the position of the  $i$ -th particle and the  $f_j$  are real-valued smooth functions on  $(\mathbb{R}^3)^N \times \mathbb{R}$ . The set  $\mathcal{M}(t)$  of all possible configurations of the system at time  $t$ , i.e. the set of all the vectors  $\mathbf{r} = (\mathbf{r}_1, \dots, \mathbf{r}_N) \in (\mathbb{R}^3)^N$  satisfying the constraints  $f_1(\mathbf{r}, t) = 0, \dots, f_m(\mathbf{r}, t) = 0$ , is called the *configuration space* of the constrained mechanical system at time  $t$ . Let us fix the time  $t$  for a moment and consider the configuration space  $\mathcal{M}(t)$ . If the gradients of the functions  $f_j(\cdot, t)$  are linearly independent at every point in  $\mathcal{M}(t)$ , the configuration space (at time  $t$ ) carries the structure of an  $\ell$ -dimensional differentiable manifold where  $\ell = 3N - m$ .<sup>6</sup> This implies that  $\mathcal{M}(t)$  is provided with a finite or countable collection of *charts*, so that every point is represented in at least one chart. A chart is an open set  $U$  in  $\mathbb{R}^\ell$

<sup>6</sup>In this case the constrained mechanical system is said to be a system with  $\ell$  degrees of freedom.

with a diffeomorphic mapping  $\phi$  from  $U$  onto some open subset  $V = \phi(U)$  of  $\mathcal{M}(t)$  which we write in the following way.

$$\phi : q \mapsto \mathbf{r}(q, t) = (\mathbf{r}_1(q, t), \dots, \mathbf{r}_N(q, t)) \text{ where } \mathbf{r}_i(q, t) = \mathbf{r}_i(q_1, \dots, q_\ell, t), \quad q \in U. \quad (35)$$

The coordinates  $q_j$ ,  $j \in \underline{\ell}$  of the vector  $q$  are called the *generalized coordinates* of the configuration  $\mathbf{r} = \mathbf{r}(q, t)$  (at time  $t$ ). They yield a parametrization of the open subset  $V = \phi(U)$  of the configuration space. The inverse mapping  $\phi^{-1} : V \rightarrow U$  maps every configuration  $\mathbf{r} \in V$  onto its *generalized coordinate vector*  $q(\mathbf{r}, t)$  (at time  $t$ ). Note that the coordinates of the generalized coordinate vector  $q \in U$  can be varied independently (provided  $q$  remains in the open set  $U$ ) whereas the coordinates of a position vector  $\mathbf{r} = (\mathbf{r}_1, \dots, \mathbf{r}_N) \in V$  cannot be varied independently without violating some of the holonomic constraints (34) (at time  $t$ ). In practice the generalized coordinates are usually obtained by applying the implicit function theorem to the constraint equations. This technique is sometimes called “elimination of the dependent coordinates”. The position vectors  $\mathbf{r}(q, t)$  corresponding to the generalized coordinate vectors  $q \in U$  at time  $t$  automatically satisfy the constraints. The advantage of “getting rid of the constraints” by introducing the generalized coordinates is, however, not obtained without any cost. Firstly, the generalized coordinates describe, in general, only a part of the configuration space. Secondly, they need not have an immediate physical interpretation similar to the original position variables. And often they cannot be related to single elements of the system but are mathematical constructions for the description of the system as a whole.

Now suppose for a moment that the constraints (34) do not depend upon  $t$  (an assumption which is often satisfied in applications) so that the configuration manifold  $\mathcal{M}(t) = \mathcal{M}$ , the charts  $(U, \phi)$  and the corresponding open subsets  $V = \phi(U)$  of  $\mathcal{M}$  are independent of time. Then, as the system moves, the position vector  $\mathbf{r}(t)$  describes a curve in the configuration manifold  $\mathcal{M}$ . If  $\mathbf{r}(t)$  belongs to the scope  $V$  of some chart  $(U, \phi)$  for  $t \in [t_1, t_2]$ , the motion of the system during this time interval can alternatively be described in terms of the position vectors  $\mathbf{r}_i(t)$ ,  $i \in \underline{N}$  (satisfying the constraints) or in terms of the generalized coordinates  $q_j(t)$ ,  $j \in \underline{\ell}$ .

We will now briefly point out how velocities, forces, kinetic and potential energies are expressed in terms of the generalized coordinates. With every family of velocity vectors  $\mathbf{v}_i = \dot{\mathbf{r}}_i = \frac{d}{dt}\mathbf{r}_i$ ,  $i \in \underline{N}$  which is consistent with the given constraints there is an associated generalized velocity vector  $\dot{q} = \frac{d}{dt}q(\mathbf{r}, t) = (\dot{q}_j)_{j \in \underline{\ell}}$  which can be determined by solving the system of  $3N$  linear equations

$$\mathbf{v}_i = \dot{\mathbf{r}}_i = \sum_{j=1}^{\ell} \frac{\partial \mathbf{r}_i}{\partial q_j} \dot{q}_j + \frac{\partial \mathbf{r}_i}{\partial t}, \quad i \in \underline{N}. \quad (36)$$

Similarly, for any family of external forces  $\mathbf{f}_i = (f_{i,1}, f_{i,2}, f_{i,3}) \in \mathbb{R}^3$ ,  $i \in \underline{N}$  applied to the  $i$ -th particle at time  $t$  there is an associated vector  $(F_1, \dots, F_\ell) \in \mathbb{R}^\ell$  called the *generalized force* at time  $t$  defined by

$$F_j = \sum_{i=1}^N \left\langle \mathbf{f}_i, \frac{\partial \mathbf{r}_i}{\partial q_j} \right\rangle_{\mathbb{R}^3} = \sum_{i=1}^N \left( f_{i,1} \frac{\partial x_i(q, t)}{\partial q_j} + f_{i,2} \frac{\partial y_i(q, t)}{\partial q_j} + f_{i,3} \frac{\partial z_i(q, t)}{\partial q_j} \right), \quad j \in \underline{\ell}.$$

where  $\mathbf{r}_i(q, t) = (x_i(q, t), y_i(q, t), z_i(q, t))$ . If the  $i$ -th particle of the system has mass  $m_i$  and is moving with velocity  $\mathbf{v}_i$ , the associated kinetic energy of the system is

$\mathcal{T} = \sum_{i=1}^N (m_i/2) \|\mathbf{v}_i\|^2$ . By means of (36) the kinetic energy can be expressed in terms of the generalized coordinates and velocities,

$$\mathcal{T}(q, \dot{q}, t) = \sum_{i=1}^N (m_i/2) \|\mathbf{v}_i(q, t)\|^2 = \sum_{i=1}^N (m_i/2) \left\| \sum_{j=1}^{\ell} \frac{\partial \mathbf{r}_i}{\partial q_j}(q, t) \dot{q}_j + \frac{\partial \mathbf{r}_i}{\partial t}(q, t) \right\|^2.$$

We see therefore that if the constraints are independent of time, then  $\mathcal{T}$  is a homogeneous quadratic form in the generalized velocities. Now assume for a moment that the mechanical system is *conservative*, i.e. there exists a real valued function  $\mathcal{W}(\mathbf{r}_1, \dots, \mathbf{r}_N, t)$  such that the force  $\mathbf{f}_i$  applied to the  $i$ -th particle is given by the  $i$ -th partial gradient of  $\mathcal{W}$  (i.e. with respect to the coordinates  $x_i, y_i, z_i$  of  $\mathbf{r}_i$ )

$$\mathbf{f}_i(\mathbf{r}_1, \dots, \mathbf{r}_N, t) = -\nabla_i \mathcal{W}(\mathbf{r}_1, \dots, \mathbf{r}_N, t).$$

In this case the generalized force is precisely the negative gradient of  $\mathcal{W}$  viewed as a function of the generalized coordinates:

$$F_j(q, t) = \sum_{i=1}^N \left\langle \mathbf{f}_i, \frac{\partial \mathbf{r}_i}{\partial q_j} \right\rangle(q, t) = - \sum_{i=1}^N \left\langle \nabla_i \mathcal{W}(\mathbf{r}_1, \dots, \mathbf{r}_N, t), \frac{\partial \mathbf{r}_i(q)}{\partial q_j} \right\rangle(q, t) = - \frac{\partial \mathcal{W}(q, t)}{\partial q_j}$$

where  $\mathcal{W}(q, t) = \mathcal{W}(q_1, \dots, q_\ell, t) := \mathcal{W}(\mathbf{r}_1(q, t), \dots, \mathbf{r}_N(q, t), t)$  is called the *generalized potential energy*.

In 1788 Lagrange published in Paris his celebrated *Mécanique Analytique* [325] in which he set out a method for determining the equations of a mechanical system from a knowledge of the kinetic and potential energies. His ideas were developed further by Boltzmann in 1802 and Hamel in 1804 and the form in which we state the equations are essentially due to them, although they are widely referred to as Lagrange's equations. Lagrange [325] introduced what is now known as a *Lagrangian*:

$$L(q, \dot{q}, t) = \mathcal{T}(q, \dot{q}, t) - \mathcal{W}(q, t). \quad (37)$$

Then *Lagrange's equations* of motion take the form

$$\frac{d}{dt} \left( \frac{\partial L}{\partial \dot{q}_j}(q(t), \dot{q}(t), t) \right) - \frac{\partial L}{\partial q_j}(q(t), \dot{q}(t), t) = 0, \quad j = 1, \dots, \ell. \quad (38)$$

In practice most mechanical systems are not conservative, since, they either have significant internal frictions, or external forces are applied which are not derived from a potential. If  $F_j$  are the generalized forces which are not taken into account by the potential energy and  $\mathcal{D}(\dot{q}) = \mathcal{D}(\dot{q}_1, \dots, \dot{q}_\ell)$  is the total energy dissipated by linear dissipators (e.g. dampers), then the equations of motion take the form

$$\frac{d}{dt} \left( \frac{\partial L}{\partial \dot{q}_j}(q(t), \dot{q}(t), t) \right) - \frac{\partial L}{\partial q_j}(q(t), \dot{q}(t), t) + \frac{1}{2} \frac{\partial \mathcal{D}}{\partial \dot{q}_j}(\dot{q}(t)) = F_j(q(t), t), \quad j = 1, \dots, \ell. \quad (39)$$

Note that if the generalized external forces  $F_j$  do not depend on the generalized coordinates  $q$  they can easily be accounted for by modifying the potential energy

$$\mathcal{W} \rightsquigarrow \mathcal{W} - \sum_{j=1}^{\ell} F_j q_j. \quad (40)$$

By suitably modifying the Lagrangian it is also possible to include other generalized forces in Lagrange's equations (38), see *Notes and References*.

If, for a given mechanical system, generalized coordinates can be found, Lagrange's method is a very convenient way to eliminate the forces of constraint from the equations of motion. By this elimination the modelling procedure is greatly simplified. In fact, in order to model a complicated multi-body mechanical system by the free-body diagram approach illustrated in the previous examples, many vector forces and velocities must be handled, whereas whenever a Lagrangian formulation is applicable there is – in principle – a straight forward procedure for deriving the equations of motion. One “only” has to write three *scalar* functions  $\mathcal{T}$ ,  $\mathcal{W}$ ,  $\mathcal{D}$  in generalized coordinates (which may not be so easy), form  $L$ , determine the generalized forces and substitute in (39). Sometimes, of course, one would like to know the contact forces and then it is necessary to resort to free-body diagrams. However, assuming Lagrange's equations have been solved for the generalized coordinates  $q_i(t)$  as functions of time  $t$  and consequently the vector functions  $\mathbf{r}_i(\cdot)$  are known, the equations for the contact forces obtained via free-body diagrams can often be easily resolved.

**Remark 1.3.6.** Lagrange's equations have the following interesting interpretation. Consider any given trajectory  $\mathbf{r}(t)$  of a conservative mechanical system in configuration space from time  $t_0$  to time  $t_1$  and suppose that the trajectory remains inside the scope of a chart so that it can equivalently be described by a curve  $t \rightarrow q(t) = (q_1(t), \dots, q_\ell(t))$ ,  $t \in [t_0, t_1]$  in  $\mathbb{R}^\ell$  (satisfying  $\mathbf{r}(t) = \mathbf{r}(q(t), t)$ ). Hamilton's Principle says: *The motion of a conservative system from time  $t_0$  to time  $t_1$  is such that the action integral*

$$\mathbf{I}(z(\cdot)) = \int_{t_0}^{t_1} L(z(t), \dot{z}(t), t) dt$$

*is an extremum for the actual path of motion  $q(\cdot)$  amongst all other curves  $z(\cdot) : [t_0, t_1] \rightarrow \mathbb{R}^\ell$  connecting  $q(t_0)$  with  $q(t_1)$ .* There are global, coordinate free formulations of this principle which avoid the restriction to parts of the configuration manifold parametrized by a chart, see e.g. [1], [18].

It is shown in the calculus of variations that Lagrange's equations are exactly the necessary and sufficient conditions for the functional

$$\mathbf{I} : \{z(\cdot) \in C^1([t_0, t_1], \mathbb{R}^\ell); z(t_0) = q(t_0), z(t_1) = q(t_1)\} \rightarrow \mathbb{R}$$

to have an extremum at  $z(\cdot) = q(\cdot)$ . In 1766 Lagrange joined Euler as a court mathematician in Berlin under the patronage of Frederick the Great. Euler also developed necessary and sufficient conditions which are equivalent to those of Lagrange and it is usual, at least in the field of the calculus of variations, to refer to the equations as the *Euler-Lagrange equations*. The variational approach is of great importance since variational principles can be used in many fields of physics to express the equations of motion. This makes it possible to transfer the Lagrangian method to other fields and uncover structural analogies between them.  $\square$

Before we consider some examples we briefly outline the *Hamiltonian approach* to classical mechanics which yields another method for deriving the equations of motion of a conservative mechanical system. The result is a transformation of Lagrange's equations (38) which are second order into an equivalent system of *Hamiltonian*



equations which are first order. This is accomplished by applying a Legendre transformation to the Lagrangian, see *Notes and References*. For arbitrary given  $q, t$  this transforms  $L(q, \dot{q}, t)$  viewed as a function of  $\dot{q}$  into a function of the new variable  $p$  where  $\dot{q}$  and  $p$  are related via the formula  $p = \partial L / \partial \dot{q}$ ,

$$H(q, p, t) = \langle p, \dot{q}(q, p, t) \rangle - L(q, \dot{q}(q, p, t), t), \quad (q, p, t) \in \mathbb{R}^\ell \times \mathbb{R}^\ell \times \mathbb{R}. \quad (41)$$

Here the function  $\dot{q} = \dot{q}(q, p, t)$  is defined implicitly by the equation

$$p = \frac{\partial L}{\partial \dot{q}}(q, \dot{q}, t) \quad (42)$$

which is assumed to have a unique solution  $\dot{q}$  for every  $(q, p, t) \in \mathbb{R}^\ell \times \mathbb{R}^\ell \times \mathbb{R}$ .  $H$  is called the *Hamiltonian* and  $p = (p_1, p_2, \dots, p_\ell)$  the *generalized momentum* of the conservative mechanical system. Now the total differential of the Hamiltonian

$$dH = \frac{\partial H}{\partial p} dp + \frac{\partial H}{\partial q} dq + \frac{\partial H}{\partial t} dt$$

is equal to the total differential of  $\langle p, \dot{q} \rangle - L(q, \dot{q}, t)$ ,

$$dH = \langle \dot{q}, dp \rangle + \langle p, d\dot{q} \rangle - \left\langle \frac{\partial L}{\partial q}, dq \right\rangle - \left\langle \frac{\partial L}{\partial \dot{q}}, d\dot{q} \right\rangle - \frac{\partial L}{\partial t} dt.$$

where  $\langle \cdot, \cdot \rangle$  denotes the usual inner product in  $\mathbb{R}^\ell$ . The second and fourth terms cancel because of (42), hence

$$\frac{\partial H}{\partial p} = \dot{q}, \quad \frac{\partial H}{\partial q} = -\frac{\partial L}{\partial q}, \quad \frac{\partial H}{\partial t} = -\frac{\partial L}{\partial t}.$$

Applying Lagrange's equations (38) we obtain *Hamilton's equations*

$$\dot{q} = \frac{\partial H}{\partial p}, \quad \dot{p} = -\frac{\partial H}{\partial q}. \quad (43)$$

We now illustrate the Lagrangian and Hamiltonian approaches by deriving the equations of motion for the cart-pendulum system studied in Example 1.3.4.

**Example 1.3.7. (Cart-pendulum system).** In order to derive the equations of motion for the cart-pendulum system via Lagrange's equations we must determine the kinetic energy  $\mathcal{T}$ , the potential energy  $\mathcal{W}$  and the dissipated energy  $\mathcal{D}$  of this system in terms of its generalized coordinates  $r, \theta$  and the corresponding velocities  $\dot{r}, \dot{\theta}$ . The kinetic energy  $\mathcal{T}$  of the system is the sum of the kinetic energies of the cart and of the pendulum, and the latter is the sum of the kinetic energy of the centre of mass plus the energy of the pendulum rotating about its centre of mass, see Figure 1.3.6. Hence

$$\begin{aligned} \mathcal{T} &= (M/2) \dot{r}^2 + (m/2) \left[ \left[ \frac{d}{dt}(r + l \sin \theta) \right]^2 + \left[ \frac{d}{dt} l \cos \theta \right]^2 \right] + (J/2) \dot{\theta}^2 \\ &= (M/2) \dot{r}^2 + (J/2) \dot{\theta}^2 + (m/2) \left[ (\dot{r} + l \dot{\theta} \cos \theta)^2 + (-l \dot{\theta} \sin \theta)^2 \right]. \end{aligned}$$

The potential energy,  $\mathcal{W}$  is the same as that of a single mass  $m$  located at the centre of mass of the pendulum in a gravitational field, i.e.  $-mgl \cos \theta$  (modulo an additive constant).

Since the time varying external force  $\beta u(t)$  does not depend on the generalized coordinates, we can take it into account by modifying the potential energy as in (40).

$$\mathcal{W} = -mgl \cos \theta - \beta ur.$$

$\mathcal{D}$  is the sum of the dissipated energies due to viscous friction  $c\dot{r}$  between cart and rails and due to viscous friction  $c_P\dot{\theta}$  at the pivot,

$$\mathcal{D} = c\dot{r}^2 + c_P\dot{\theta}^2.$$

The generalized Lagrange equations (39) in terms of the generalized coordinates and associated velocities  $r, \theta, \dot{r}, \dot{\theta}$  are

$$\frac{d}{dt} \left( \frac{\partial \mathcal{T}}{\partial \dot{r}} \right) - \frac{\partial \mathcal{T}}{\partial r} + \frac{\partial \mathcal{W}}{\partial r} + \frac{1}{2} \frac{\partial \mathcal{D}}{\partial \dot{r}} = 0$$

$$\frac{d}{dt} \left( \frac{\partial \mathcal{T}}{\partial \dot{\theta}} \right) - \frac{\partial \mathcal{T}}{\partial \theta} + \frac{\partial \mathcal{W}}{\partial \theta} + \frac{1}{2} \frac{\partial \mathcal{D}}{\partial \dot{\theta}} = 0.$$

Or

$$\frac{d}{dt} [M\dot{r} + m(\dot{r} + l\dot{\theta} \cos \theta)] + c\dot{r} = \beta u$$

$$\frac{d}{dt} [J\dot{\theta} + m(\dot{r} + l\dot{\theta} \cos \theta)l \cos \theta + ml^2\dot{\theta}(\sin \theta)^2]$$

$$+ m(\dot{r} + l\dot{\theta} \cos \theta)l\dot{\theta} \sin \theta - ml^2\dot{\theta}^2 \sin \theta \cos \theta + mgl \sin \theta + c_P\dot{\theta} = 0.$$

A simple calculation yields the nonlinear differential equations

$$(M + m)\ddot{r} + ml\ddot{\theta} \cos \theta - ml\dot{\theta}^2 \sin \theta + c\dot{r} = \beta u$$

$$(J + ml^2)\ddot{\theta} + m\dot{r}l \cos \theta + mgl \sin \theta + c_P\dot{\theta} = 0.$$

Thus the Lagrangian approach leads to the same equations of motion as the approach via free-body diagrams in Example 1.3.4, see (23).

Assuming that frictions can be neglected and the pendulum behaves like a point mass connected to a light rod of length  $l$  (i.e.  $c = c_P = J = 0$ ), the nonlinear equations of motion reduce to

$$(M + m)\ddot{r} + ml\ddot{\theta} \cos \theta - ml\dot{\theta}^2 \sin \theta = \beta u \quad (44)$$

$$l\ddot{\theta} + g \sin \theta + \dot{r} \cos \theta = 0. \quad (45)$$

Setting  $x_1 = r, x_2 = \theta, x_3 = \dot{r}, x_4 = \dot{\theta}$  (resp.  $x_1 = r, x_2 = \varphi, x_3 = \dot{r}, x_4 = \dot{\varphi}$ , see Example 1.3.4) the linearized models of the loading plant and the inverted pendulum, respectively, reduce to

$$\dot{x} = Ax + bu, \quad (46)$$

where (see (31), (32))

$$A = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & a_{32} & 0 & 0 \\ 0 & a_{42} & 0 & 0 \end{bmatrix}, \quad b = \begin{bmatrix} 0 \\ 0 \\ b_3 \\ b_4 \end{bmatrix}, \quad \begin{aligned} a_{32} &= mg/M, & a_{42} &= \mp(M + m)g/(Ml), \\ b_3 &= \beta/M, & b_4 &= \mp\beta/(Ml). \end{aligned}$$

Let us now consider Hamilton's equations for the frictionless case. By (42), the generalized momentum has components

$$p_1 = \frac{\partial L}{\partial \dot{r}} = (M + m)\dot{r} + ml \cos \theta \dot{\theta} \quad (47)$$

$$p_2 = \frac{\partial L}{\partial \dot{\theta}} = ml(\dot{r} \cos \theta + l\dot{\theta}). \quad (48)$$

Hence

$$\dot{r} = (ml^2 p_1 - ml \cos \theta p_2)/(ml^2(M + m \sin^2 \theta)) \quad (49)$$

$$\dot{\theta} = (-ml \cos \theta p_1 + (M + m)p_2)/(ml^2(M + m \sin^2 \theta)). \quad (50)$$

So by (41) the Hamiltonian is

$$H(r, \theta, p_1, p_2) = (ml^2 p_1^2 - 2ml \cos \theta p_1 p_2 + (M + m)p_2^2)/(2ml^2(M + m \sin^2 \theta)) - mgl \cos \theta - \beta ru.$$

Hamilton's equations are, therefore, (49), (50) augmented with

$$\begin{aligned} \dot{p}_1 &= -\frac{\partial H}{\partial r} = \beta u \\ \dot{p}_2 &= -\frac{\partial H}{\partial \theta} = -p_1 p_2 \sin \theta / (l(M + m \sin^2 \theta)) - mgl \sin \theta \\ &\quad + (ml^2 p_1^2 - 2ml \cos \theta p_1 p_2 + (M + m)p_2^2) \sin \theta \cos \theta / (l^2(M + m \sin^2 \theta)^2). \end{aligned} \quad (51)$$

The linearization of equations (49), (50), (51) yields

$$\dot{\tilde{x}} = \tilde{A}\tilde{x} + \tilde{b}u, \quad (52)$$

where  $\tilde{x} = [r, \theta, p_1, p_2]^\top$ ,  $\tilde{b} = [0, 0, \beta, 0]^\top$  and  $\tilde{A}$  is the matrix

$$\tilde{A} = \begin{bmatrix} 0 & 0 & \tilde{a}_{13} & \tilde{a}_{14} \\ 0 & 0 & \tilde{a}_{23} & \tilde{a}_{24} \\ 0 & 0 & 0 & 0 \\ 0 & \tilde{a}_{42} & 0 & 0 \end{bmatrix}, \quad \begin{aligned} \tilde{a}_{13} &= 1/M, & \tilde{a}_{14} &= \tilde{a}_{23} = -1/(Ml), \\ \tilde{a}_{24} &= (M + m)/(mMl^2), & \tilde{a}_{42} &= -mgl. \end{aligned}$$

Equations (46) and (52) for the loading plant are two different mathematical models of the linearized system which are related by the transformations

$$\tilde{x} = Tx, \quad \tilde{A} = TAT^{-1}, \quad \tilde{b} = Tb,$$

where

$$T = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & t_{33} & t_{34} \\ 0 & 0 & t_{43} & t_{44} \end{bmatrix}, \quad t_{33} = M + m, \quad t_{34} = t_{43} = ml, \quad t_{44} = ml^2.$$

Such systems are called *similar* and we will discuss this concept in Section 2.4.  $\square$

We conclude this section by using Lagrange's equations to derive the equations of motion of an inverted double pendulum.

**Example 1.3.8. (Inverted double pendulum).** Consider a double pendulum which is mounted on a cart as illustrated in Figure 1.3.8. In a similar way to Example 1.3.4 we assume that the motion of the system is restricted to the vertical plane, the cart is moving on a horizontal rail with viscous friction and the two pendulums behave like rigid bodies with viscous friction at the pivots. Let  $m_i, l_i, J_i, c_i$  ( $i = 1, 2$ ) denote the mass, the distance between the centre of gravity and the lower hinge, the moment of inertia about the centre of mass and the friction coefficient for the lower ( $i = 1$ ) and the upper ( $i = 2$ ) pendulums.  $L$  is the total length of the lower pendulum and  $M, c_0$  denote the mass and the friction coefficient of the cart. As generalized coordinates we choose the distance  $r$  of the cart

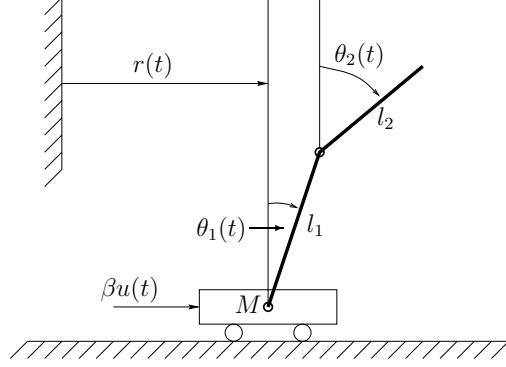


Figure 1.3.8: Double pendulum

from an inertial reference position, and the angles  $\theta_i$ ,  $i = 1, 2$  of the two pendulums to the vertical, with *clockwise* orientation. We apply the method of Lagrange in order to find the equations of motion of the system. The *potential energy* of the system is equal to the sum of the potential energies of the masses  $m_i$  located at the centres of mass of the two pendulums, together with an adjustment for the external force. In terms of the chosen generalized coordinates it is given by

$$\mathcal{W} = m_1 g l_1 \cos \theta_1 + m_2 g (L \cos \theta_1 + l_2 \cos \theta_2) - \beta u r. \quad (53)$$

The energy dissipated by the translational viscous friction between cart and rails and the rotational friction at the two hinges is given by

$$\mathcal{D} = c_0 \dot{r}^2 + c_1 \dot{\theta}_1^2 + c_2 (\dot{\theta}_2 - \dot{\theta}_1)^2. \quad (54)$$

The kinetic energy is the sum of the kinetic energies of the cart plus the kinetic energies of the two pendulums:

$$\begin{aligned} \mathcal{T}_0 &= (M/2) \dot{r}^2 \\ \mathcal{T}_1 &= (J_1/2) \dot{\theta}_1^2 + (m_1/2) \left\{ \left[ \frac{d}{dt} (r + l_1 \sin \theta_1) \right]^2 + \left[ \frac{d}{dt} (l_1 \cos \theta_1) \right]^2 \right\} \\ \mathcal{T}_2 &= (J_2/2) \dot{\theta}_2^2 + (m_2/2) \left\{ \left[ \frac{d}{dt} (r + L \sin \theta_1 + l_2 \sin \theta_2) \right]^2 + \left[ \frac{d}{dt} (L \cos \theta_1 + l_2 \cos \theta_2) \right]^2 \right\}. \end{aligned}$$

A simple calculation yields the *total kinetic energy* of the system

$$\begin{aligned} \mathcal{T} &= (M/2) \dot{r}^2 + (J_1/2) \dot{\theta}_1^2 + (J_2/2) \dot{\theta}_2^2 + (m_1/2) \left\{ \dot{r}^2 + 2l_1 \dot{r} \dot{\theta}_1 \cos \theta_1 + l_1^2 \dot{\theta}_1^2 \right\} + \\ &\quad (m_2/2) \left\{ \dot{r}^2 + L^2 \dot{\theta}_1^2 + l_2^2 \dot{\theta}_2^2 + 2\dot{r} \left[ L \dot{\theta}_1 \cos \theta_1 + l_2 \dot{\theta}_2 \cos \theta_2 \right] + 2Ll_2 \dot{\theta}_1 \dot{\theta}_2 \cos(\theta_1 - \theta_2) \right\}. \quad (55) \end{aligned}$$

Using (53), (54) and (55) we write down Lagrange's equations and obtain by elementary calculations the following equations of motion

$$\begin{aligned} \frac{d}{dt} \left( \frac{\partial \mathcal{T}}{\partial \dot{r}} \right) - \frac{\partial \mathcal{T}}{\partial r} + \frac{\partial \mathcal{W}}{\partial r} + \frac{1}{2} \frac{\partial \mathcal{D}}{\partial \dot{r}} &= 0, \\ (M + m_1 + m_2)\ddot{r} + [(m_1 l_1 + m_2 L) \cos \theta_1] \ddot{\theta}_1 + (m_2 l_2 \cos \theta_2) \ddot{\theta}_2 \\ &\quad - (m_1 l_1 + m_2 L) \dot{\theta}_1^2 \sin \theta_1 - (m_2 l_2 \sin \theta_2) \dot{\theta}_2^2 + c_0 \dot{r} = \beta u. \\ \frac{d}{dt} \left( \frac{\partial \mathcal{T}}{\partial \dot{\theta}_1} \right) - \frac{\partial \mathcal{T}}{\partial \theta_1} + \frac{\partial \mathcal{W}}{\partial \theta_1} + \frac{1}{2} \frac{\partial \mathcal{D}}{\partial \dot{\theta}_1} &= 0, \\ [(m_1 l_1 + m_2 L) \cos \theta_1] \ddot{r} + (m_1 l_1^2 + m_2 L^2 + J_1) \ddot{\theta}_1 + [m_2 L l_2 \cos(\theta_1 - \theta_2)] \ddot{\theta}_2 \\ &\quad + m_2 L l_2 \dot{\theta}_2^2 \sin(\theta_1 - \theta_2) - (m_1 l_1 + m_2 L) g \sin \theta_1 + c_1 \dot{\theta}_1 + c_2 (\dot{\theta}_1 - \dot{\theta}_2) = 0. \\ \frac{d}{dt} \left( \frac{\partial \mathcal{T}}{\partial \dot{\theta}_2} \right) - \frac{\partial \mathcal{T}}{\partial \theta_2} + \frac{\partial \mathcal{W}}{\partial \theta_2} + \frac{1}{2} \frac{\partial \mathcal{D}}{\partial \dot{\theta}_2} &= 0, \\ (m_2 l_2 \cos \theta_1) \ddot{r} + [m_2 L l_2 \cos(\theta_1 - \theta_2)] \ddot{\theta}_1 + (m_2 l_2^2 + J_2) \ddot{\theta}_2 + \\ &\quad - m_2 L l_2 \dot{\theta}_1^2 \sin(\theta_1 - \theta_2) - m_2 g l_2 \sin \theta_2 + c_2 (\dot{\theta}_2 - \dot{\theta}_1) = 0. \end{aligned}$$

Introducing the vector  $z = [r, \theta_1, \theta_2]^\top \in \mathbb{R}^3$ , the equations of motion can be written in the following concise form

$$K_1 \ddot{z} = K_2 \dot{z} + K_3 + k_4 u$$

where

$$\begin{aligned} K_1 &= \begin{bmatrix} m_1 + m_2 + M & (m_1 l_1 + m_2 L) \cos \theta_1 & m_2 l_2 \cos \theta_2 \\ (m_1 l_1 + m_2 L) \cos \theta_1 & J_1 + m_1 l_1^2 + m_2 L^2 & m_2 l_2 L \cos(\theta_1 - \theta_2) \\ m_2 l_2 \cos \theta_2 & m_2 l_2 L \cos(\theta_1 - \theta_2) & J_2 + m_2 l_2^2 \end{bmatrix} \\ K_2 &= \begin{bmatrix} -c_0 & (m_1 l_1 + m_2 L) \dot{\theta}_1 \sin \theta_1 & m_2 l_2 \dot{\theta}_2 \sin \theta_2 \\ 0 & -c_1 - c_2 & -m_2 l_2 L \dot{\theta}_2 \sin(\theta_1 - \theta_2) + c_2 \\ 0 & m_2 l_2 L \dot{\theta}_1 \sin(\theta_1 - \theta_2) + c_2 & -c_2 \end{bmatrix} \\ K_3 &= \begin{bmatrix} 0 \\ (m_1 l_1 + m_2 L) g \sin \theta_1 \\ m_2 l_2 g \sin \theta_2 \end{bmatrix}, \quad k_4 = \begin{bmatrix} \beta \\ 0 \\ 0 \end{bmatrix}. \end{aligned}$$

An equivalent system of first order equation is obtained by setting  $x = \begin{bmatrix} z \\ \dot{z} \end{bmatrix} \in \mathbb{R}^6$

$$\dot{x} = \begin{bmatrix} \dot{z} \\ K_1^{-1}(K_2 \dot{z} + K_3 + k_4 u) \end{bmatrix}. \quad (56)$$

Is it possible to stabilize the double pendulum in the upright position? Most readers will find it difficult to decide this question relying only on their physical intuition. In Vol. II we show that if the deviations from the upright position are small it is possible to find a control which restores this position in finite time. Then we prove that this implies the existence of a regulator which makes the upright position a stable equilibrium point of the feedback system. This regulator accepts as input values  $r(t)$ ,  $\dot{r}(t)$ ,  $\theta_1(t)$ ,  $\dot{\theta}_1(t)$ ,  $\theta_2(t)$ ,  $\dot{\theta}_2(t)$  and so sensors must determine these values for all  $t \geq 0$ . Since sensors are expensive one is interested in reducing the number. In particular the question arises whether or not it is possible to design a regulator which accepts as values, say  $r(t)$ ,  $\theta_1(t)$ . This means

that within the regulator it is necessary to reconstruct the angle  $\theta_2(t)$  and the velocities  $\dot{r}(t)$ ,  $\dot{\theta}_1(t)$ ,  $\dot{\theta}_2(t)$  from the measurements  $r(t)$ ,  $\theta_1(t)$ . This is a typical *observability* problem, see Vol. II.  $\square$

### 1.3.4 Notes and References

Many books on modelling and dynamics contain chapters on the modelling of mechanical systems, see *Ogata* (1992) [397], *Burton* (1994) [84], *Close and Frederick* (1995) [105]. More realistic automobile suspension systems separating, for example, the motions of the front and the rear axles (see Example 1.3.2) can be found in [84] and [103]. In the presence of rotational friction at the hinge *Antman* (1998) [15] has shown that the derivation of the equations of motion of a compound pendulum may be flawed by the assumption that the reactive force at the hinge acts along the pendulum. Modelling the cart-pendulum system of Example 1.3.4 is discussed in more detail in *Clark* (1995) [103]. *Ackermann* (1977) [2] has analyzed the feedback control of the linearized loading plant. The inverted pendulum has been a favourite example for the illustration of modern control methods in textbooks since the sixties, see *Elgerd* (1967) [150]. The balancing problem was solved by various methods on the basis of the linearized model. However, the swinging up problem, i.e. moving the system from the downward to the upright rest position and keeping it there, requires the use of the nonlinear model. This problem has been studied in *Mori et al.* (1976) [382]. The stabilization of double and multiple pendulum systems has been investigated in *Furuta et al.* (1980) [177], *Maletinsky et al.* (1982) [357] and *Kwakernaak and Westdijk* (1995) [323].

An excellent introduction to Newtonian mechanics is contained in the first volume of the *Feynman Lectures on Physics* (1975) [161]. Brief introductions to Lagrangian and Hamiltonian modelling techniques from an engineering point of view are given in *MacFarlane* (1970) [355], *Wellstead* (1979) [516], *Burton* (1994) [84]. More information concerning variational principles, Lagrangian and Hamiltonian mechanics, can be found in standard textbooks on classical mechanics, *Whittaker* (1970) [520], *Gantmacher* (1975) [184], *Landau and Lifshitz* (1976) [329], *Goldstein* (1980) [194], *Chorlton* (1983), [99]. Lagrange's *Mécanique Analytique* is now available in English [325]. It is shown in [194] that by introducing velocity dependent potentials it is possible, under certain conditions, to include non-conservative generalized forces in Lagrange's equations (38). In particular the Lorentz force (4.12), which in general is not conservative, satisfies these conditions.

The Hamiltonian formulation was proposed by Hamilton in a British Association Report in 1834, although in part it had been anticipated by Lagrange and Poisson in 1809/1810. The Legendre transformation maps functions on a vector space to functions on the dual space: Let  $f(x)$  be a convex function of  $x \in \mathbb{R}^n$ , then the Legendre transform is the function  $g : \mathbb{R}^{n*} \rightarrow \mathbb{R}$  defined by

$$g(y) = F(y, x(y)) = \max_x F(y, x), \quad F(y, x) = \langle y, x \rangle - f(x), \quad y = \partial f / \partial x.$$

Details can be found in most references on mathematical physics, see e.g. *Courant and Hilbert* (1953) [112]. Modern advanced mathematical treatments of classical mechanics are given in *Arnold* (1978) [18], *Abraham and Marsden* (1978) [1] and *Marsden and Ratiu* (1999) [361]. These standard references develop the mathematical framework for a coordinate free treatment of the configuration space in the general setting of (symplectic) manifolds. [1] and [361] also contain many instructive historical remarks and comments on the literature.

## 1.4 Electromagnetism and Electrical Systems

This section is divided into two subsections. In the first we give a brief review of some of the historical developments of electromagnetism and describe the basic building blocks of circuits. Then in the second we show how to obtain the equations governing the current flows in networks of circuits.

### 1.4.1 Maxwell's Equations and the Elements of Electrical Circuits

Some of the most outstanding discoveries of the 19th century were connected with electricity and magnetism and their interaction. As a reference we quote *Richard Feynman* (1975) [161]: “From a long view of the history of mankind—seen from, say, ten thousand years from now—there can be little doubt that the most significant event of the 19th century will be judged as Maxwell’s discovery of the laws of electrodynamics”. In this subsection we recall some of the electromagnetic experiments that were carried out and indicate how the conclusions drawn from them can be formulated in a mathematical way, i.e. can be cast in the form of equations. We then use these equations in a number of different examples to obtain mathematical models of various electrical circuits and systems.

#### Maxwell's Equations

If a piece of amber is rubbed with a cloth and then the amber and cloth are separated they are found to attract each other. Such forces are called *electrical forces* and the amber and cloth are said to be electrified or charged with electricity. In 1729 *Gray* [201] discovered that some materials could convey electricity from one place to another. He carried out an experiment with a glass rod connected by a hemp cord of length 400 feet to an ivory ball and was able to electrify the ball by rubbing the glass tube. Further experiments were carried out by *Desaguliers* (1739) [128] who introduced the word *conductors* for those materials which transport electricity easily. *Cavendish* (1776) [93] anticipated Ohm’s law, although much of his work was not published until 100 years later in a collection of his papers put together by *Maxwell* (1879) [365]. In 1821 *Ampère* put forward a workable definition of current and invented a galvanometer to measure it. He thought of voltage as the cause and current as the effect and although he knew that there was a relationship between them, he did not realize that, across a resistor, they are directly proportional. This discovery was made by *Ohm* (1826) [398]. He used as a source a thermoelectric battery with strips of copper and bismuth joined at their two ends. He kept one point of contact in boiling water and the other in ice and thereby obtained a stable current in an external circuit  $C$  which he connected across the two points of contact. Working with this rather deficient apparatus, Ohm performed a series of carefully devised experiments which established, for this circuit, the law of conduction (now known as *Ohm’s law*):

If  $I$  is the current in the circuit,  $C$  and  $V$  the voltage drop between the two points of contact, then  $V = IR$ , where  $R$  is a constant called the *resistance* which varies with the wire which is used to close the circuit, but does not depend on  $V$  or  $I$ .

That electric charges exert forces on each other with a magnitude inversely proportional to the square of the distance between them was suspected early in the 18th century. *Benjamin Franklin* (1755) [106] carried out experiments to determine this law and *Robison* (1769) observed experimentally that the force was proportional to  $r^{-2.06}$  where  $r$  is the distance between the charges. Although, as in the case of *Cavendish* (who gave the law as between  $r^{-1.98}$  and  $r^{-2.02}$ ), the results were not available universally and were published posthumously in 1822. Unaware of their results *Coulomb* (1785) [124] carried out completely different experiments and put forward the inverse square law and now the discovery is usually attributed to him. In 1936 *Plimpton and Lawton* [414] showed, experimentally, that the force deviated from an inverse square law by less than two parts in one billion. Coulomb's experiments led to the formulation:

If a charge of magnitude  $q$  is placed at the origin  $O \in \mathbb{R}^3$ , then the force on a positive unit charge at a point  $\mathbf{r}$  is proportional to  $qr^{-2}\hat{\mathbf{r}}$ , where  $\hat{\mathbf{r}} = \mathbf{r}/\|\mathbf{r}\|$  and  $r = \|\mathbf{r}\|$ . Moreover the force is one of attraction if  $q < 0$  and repulsion if  $q > 0$ .

Hence if we denote this *electrical force* at the point  $\mathbf{r}$  by  $\mathbf{E}(\mathbf{r})$  and write the constant of proportionality in the form  $(4\pi\epsilon_0)^{-1}$ , we have

$$\mathbf{E}(\mathbf{r}) = \frac{q\hat{\mathbf{r}}}{4\pi\epsilon_0r^2} = -\text{grad } \Phi(\mathbf{r}), \text{ where } \Phi : \mathbb{R}^3 \setminus \{O\} \rightarrow \mathbb{R}, \quad \Phi(\mathbf{r}) = \frac{q}{4\pi\epsilon_0r}. \quad (1)$$

$\Phi(\mathbf{r})$  is called the *electrostatic potential* at the point  $\mathbf{r}$  and the constant of proportionality  $\epsilon_0$  the *permittivity* of free space.

The mapping  $\mathbf{r} \rightarrow \mathbf{E}(\mathbf{r}) = q(4\pi\epsilon_0r^2)^{-1}\hat{\mathbf{r}}$  defines a vector field on  $\mathbb{R}^3 \setminus \{O\}$  which assigns to each  $\mathbf{r} \in \mathbb{R}^3 \setminus \{O\}$  the electrical force exerted on a positive unit charge at that point<sup>1</sup>. This vector field is called the electric field of the charge  $q$  placed at  $O$ . In the following most of the vector fields we consider will depend upon time.

There are two important quantities associated with a vector field which are used to describe the results of electromagnetic experiments, namely the *flux* and the *circulation*. These terms have their origins in fluid dynamics where the "flux of velocity" through a surface is the net amount of fluid going through the surface per unit time and the "circulation" around some loop is the net rotational motion around it. More generally for any vector field  $\mathbf{F}$  the flux of  $\mathbf{F}$  through a bounded oriented piecewise smooth surface  $S$  is defined by the surface integral<sup>2</sup>

$$\text{Flux of } \mathbf{F} \text{ through the surface } S = \int_S \langle \mathbf{F}, \mathbf{n} \rangle dS,$$

where  $\mathbf{n}$  is the unit normal defining the orientation of  $S$  and  $\langle \cdot, \cdot \rangle$  is the standard inner product on  $\mathbb{R}^3$ . The circulation of  $\mathbf{F}$  around an orientated piecewise smooth closed curve  $C$  is

$$\text{Circulation of } \mathbf{F} \text{ around the curve } C = \oint_C \langle \mathbf{F}, \mathbf{t} \rangle ds,$$

<sup>1</sup>More generally, a vector field on some region  $\mathcal{D} \subset \mathbb{R}^3$  is a map  $\mathbf{F} : \mathcal{D} \rightarrow \mathbb{R}^3$  that assigns to each point  $\mathbf{r}$  in its domain a vector  $\mathbf{F}(\mathbf{r})$ .

<sup>2</sup>In this section we often suppress the dependency of a vector field on space and time to simplify notation. Where necessary, we use either  $\mathbf{r}$  or  $x$  as a space variable.



where  $\mathbf{t}$  is the unit tangent to the curve. For the definition of the above integrals, see [362]. The line integral of a conservative electric field  $\mathbf{E}$  along an arbitrary piecewise smooth curve  $C$  connecting  $A \in \mathbb{R}^3$  to  $B \in \mathbb{R}^3$ ,  $\mathcal{V}_{AB} = \int_C \langle \mathbf{E}, \mathbf{t} \rangle ds$ , is called the *voltage* (or *potential difference*) between the points  $A$  and  $B$ . We sometimes talk of the voltage of a point and by this we mean the difference in the potential of that point and the potential of an arbitrary established reference point called the *ground state*.

Now suppose that  $S$  is an orientated piecewise smooth closed surface in  $\mathbb{R}^3$ . Then it can be shown that the flux of the electric field  $\mathbf{E}$  (given by (1)) through  $S$  is

$$\int_S \langle \mathbf{E}, \mathbf{n} \rangle dS = \frac{q\omega}{4\pi\epsilon_0}$$

where  $\omega = 4\pi$  if  $O$  is inside  $S$  and  $\omega = 0$  if  $O$  is outside  $S$ .

Let us now consider a continuous distribution of charge with volume charge density  $\rho(x)$  in a bounded region  $\Omega \subset \mathbb{R}^3$  and suppose that  $S$  encloses  $\Omega$ , then the electric field  $\mathbf{E}$  generated by this charge satisfies

$$\int_S \langle \mathbf{E}, \mathbf{n} \rangle dS = \frac{1}{\epsilon_0} \int_{\Omega} \rho(x) dx = \frac{Q}{\epsilon_0}, \quad (2)$$

where  $Q$  is the total charge on  $\Omega$  and  $dx$  is the Lebesgue measure in  $\mathbb{R}^3$ . By the divergence theorem

$$\int_S \langle \mathbf{E}, \mathbf{n} \rangle dS = \int_{\Omega} \operatorname{div} \mathbf{E} dx \quad \text{and so} \quad \int_{\Omega} (\operatorname{div} \mathbf{E} - \rho/\epsilon_0) dx = 0.$$

Since this holds for any closed surface  $S$  it follows that  $\operatorname{div} \mathbf{E} = \rho/\epsilon_0$ . If additionally we suppose the electric field is derived from a potential  $\Phi$ , then

$$\operatorname{div} \mathbf{E} = \rho/\epsilon_0 \quad \text{and we have Poisson's equation} \quad \Delta \Phi = -\rho/\epsilon_0, \quad (3)$$

where  $\Delta$  is the Laplacian defined (in Cartesian coordinates) by

$$\Delta = \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2} + \frac{\partial^2}{\partial x_3^2}. \quad (4)$$

In the case of free space where  $\rho = 0$ ,  $\operatorname{div} \mathbf{E} = 0$  and  $\Phi$  satisfies the Laplace equation  $\Delta \Phi = 0$ .

An awareness of the existence of magnetized materials can be traced back to the Greeks who were familiar with loadstone and its power to attract iron. Indeed the term magnet came into use because loadstone pieces were found near the ancient Greek city called *Magnesia*<sup>3</sup>. Experiments with magnetic materials were of a much older vintage than those with electricity and the first application of magnetism, the compass, was used in Europe at the end of the twelfth century. *Newton* in *Principia* speculated that the law of force between two magnetic poles was proportional to the inverse cube of the distance between them, and *Michell* (1750) [373] was the first to give the correct law as being an inverse square. Thus if there is a magnetic pole of

<sup>3</sup>Plato in the dialogue *Ion* gives Socrates the words "impelling you like the power in the stone Euripides called the magnet....This stone does not simply attract iron rings, just by themselves; it also imparts to the rings a force enabling them to do the same thing as the stone itself".

strength  $m$  at the origin  $O$ , the force on a magnetic pole of positive unit strength at a point  $\mathbf{r}$  is proportional to  $m\mathbf{r}^{-2}\hat{\mathbf{r}}$ . Hence if we denote this *magnetic force* at the point  $\mathbf{r}$  by  $\mathbf{B}(\mathbf{r})$ , we have

$$\mathbf{B}(\mathbf{r}) = \frac{m\hat{\mathbf{r}}}{4\pi\mu_0^{-1}r^2} = -\text{grad } \Psi(\mathbf{r}), \quad \Psi(\mathbf{r}) = \frac{m}{4\pi\mu_0^{-1}r}. \quad (5)$$

$\Psi(\mathbf{r})$  is called the *magnetostatic potential* at the point  $\mathbf{r}$  and the constant of proportionality  $\mu_0$  the *permeability* of free space. The vector field  $\mathbf{B} : \mathbf{r} \rightarrow \mathbf{B}(\mathbf{r})$  on the domain  $\mathcal{D} = \mathbb{R}^3 \setminus \{O\}$  is called the *magnetic field* of the pole of strength  $m$  at  $O$ . The equations in (5) have the same form as those given in (1) and hence one can develop a theory of magnetostatics in parallel with that of electrostatics, see [151]. However there is an important difference. Whereas positive and negative electric charges can exist separately from each other, magnetic poles cannot. In any volume (no matter how small) the density of North poles is always the same as the density of South poles. So the net volume density must be zero and in analogy with the electric case the corresponding equations to (3) are

$$\text{div } \mathbf{B} = 0 \quad \text{and} \quad \Delta \Psi = 0. \quad (6)$$

Now we leave the static case and consider the dynamic case where charges move and hence generate electric currents. In 1820 *Oersted* conducted some experiments which showed that a magnetic field can be generated by an electric current flowing in a wire. *Faraday* (1821) [159] also discovered this and the precise relation as enunciated by *Ampère* takes the form:

The circulation of a magnetic field in a non-magnetic medium around a closed path is equal to  $\mu_0$  times the total current flowing through a surface bounded by the path.

Suppose that at a point  $P$  with position vector  $\mathbf{r}$  the volume charge density of electrons is  $\rho(\mathbf{r})$  and their velocity<sup>4</sup> is  $\mathbf{v}(\mathbf{r})$ , then  $\mathbf{j}(\mathbf{r}) = \rho(\mathbf{r})\mathbf{v}(\mathbf{r})$  is defined to be the *current density* at the point  $\mathbf{r}$ . So if  $S$  is an orientated piecewise smooth surface and  $I$  is the total current through  $S$ , we have

$$I = \int_S \rho \langle \mathbf{v}, \mathbf{n} \rangle dS = \int_S \langle \mathbf{j}, \mathbf{n} \rangle dS.$$

Hence if  $C$  is a closed orientated piecewise smooth curve Ampère's law takes the form

$$\oint_C \langle \mathbf{B}, \mathbf{t} \rangle ds = \mu_0 I = \mu_0 \int_S \langle \mathbf{j}, \mathbf{n} \rangle dS,$$

where the the surface  $S$  is such that  $\partial S = C$ . By Stokes' Theorem

$$\oint_C \langle \mathbf{B}, \mathbf{t} \rangle ds = \int_S \langle \text{curl } \mathbf{B}, \mathbf{n} \rangle dS \quad \text{and so} \quad \int_S \langle \text{curl } \mathbf{B} - \mu_0 \mathbf{j}, \mathbf{n} \rangle dS = 0.$$

And since this holds for any surface  $S$ , we have

$$\text{curl } \mathbf{B} = \mu_0 \mathbf{j}. \quad (7)$$

This is the differential form of *Ampère's law*. Note that since  $\text{div } \text{curl} = 0$ , the above equation implies  $\text{div } \mathbf{j} = 0$  which we will see later is not in general true.

Another major advance was made in 1831 when *Faraday* [159] discovered, experimentally, that a current was induced in a conducting loop when the magnetic field changed. Faraday found that:

---

<sup>4</sup>Strictly speaking,  $\mathbf{v}(\mathbf{r})$  is the *average* velocity of the electrons in a small volume containing  $P$ .

The circulation of the electric field vector around a closed path is equal to the rate of decrease of the magnetic flux flowing through a surface bounded by the path.

The mathematical articulation of this law, now known as *Faraday's law* was first given by *Maxwell*.

$$\mathcal{V} := \oint_C \langle \mathbf{E}, \mathbf{t} \rangle ds = -\frac{d}{dt} \int_S \langle \mathbf{B}, \mathbf{n} \rangle dS, \quad (8)$$

where  $C$  and  $S$  are as above with  $\partial S = C$ . The circulation  $\mathcal{V}$  of the electric field  $\mathbf{E}$  around  $C$  is called the *induced voltage*. Using Stokes' Theorem, the differential form of Faraday's law is

$$\text{curl } \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}. \quad (9)$$

*Maxwell*, when only 24, set out to put Faraday's experimental work on a firm mathematical footing. The work, including a correction of *Ampère's law* (7) (which allowed for the possibility that  $\text{div } \mathbf{j} \neq 0$ ), culminated in his paper "A dynamical theory of the electromagnetic field" published in (1865) [363]. If  $\rho$  is the volume charge density of electrons and  $\mathbf{v}$  their velocity, then given an orientated piecewise smooth closed surface  $S$  enclosing a volume  $\Omega$ , conservation requires that the flux of electrons through  $S$  must be balanced by their rate of decrease in  $\Omega$ , i.e.

$$\int_S \langle \rho \mathbf{v}, \mathbf{n} \rangle dS = -\frac{d}{dt} \int_{\Omega} \rho dx.$$

By the divergence theorem we get

$$\frac{\partial \rho}{\partial t} + \text{div}(\rho \mathbf{v}) = 0.$$

This equation is called the *continuity equation*. Using the first equation in (3) and the fact that  $\mathbf{j} = \rho \mathbf{v}$  yields

$$\text{div} \left( \varepsilon_0 \frac{\partial \mathbf{E}}{\partial t} + \mathbf{j} \right) = 0.$$

We have seen that (7) implies  $\text{div } \mathbf{j} = 0$  which, in general, contradicts this equation. Maxwell saw that if, however,  $\mu_0 \mathbf{j}$  was replaced with  $\mu_0 (\varepsilon_0 \frac{\partial \mathbf{E}}{\partial t} + \mathbf{j})$  in (7), then there would be no contradiction. Therefore his equations consist of the first equations of (3) and (6) together with (9) and the adjustment to (7). Hence they take the form

$$\begin{aligned} \text{div } \mathbf{E} &= \rho / \varepsilon_0, \\ \text{curl } \mathbf{E} &= -\frac{\partial \mathbf{B}}{\partial t}, \\ \text{div } \mathbf{B} &= 0, \\ \text{curl } \mathbf{B} &= \mu_0 \left( \varepsilon_0 \frac{\partial \mathbf{E}}{\partial t} + \mathbf{j} \right). \end{aligned} \quad (10)$$

Maxwell's hypotheses, together with confirmation of the correction term were substantiated experimentally by *Hertz* (1885) eight years after Maxwell's death.

There is a different version of Faraday's law for the case where the magnetic field  $\mathbf{B}$  is constant in time but the wire circuit  $C$  is moving with a velocity  $\mathbf{v}$ . Then the induced voltage,  $\mathcal{V}$ , is given by

$$\mathcal{V} = \oint_C \langle \mathbf{v} \times \mathbf{B}, \mathbf{t} \rangle ds. \quad (11)$$

One can interpret the induced voltage  $\mathcal{V}$  as being caused by an electric field  $\mathbf{E}' = \mathbf{v} \times \mathbf{B}$ , so that  $\mathcal{V} = \oint_C \langle \mathbf{E}', \mathbf{t} \rangle ds$ . This suggests that if a charge of magnitude  $q$  is moving with velocity  $\mathbf{v}$  in both an electrostatic field  $\mathbf{E}$  and a magnetic field  $\mathbf{B}$ , the total force on it,  $\mathbf{F}$ , will be  $q(\mathbf{E} + \mathbf{E}')$ , i.e.

$$\mathbf{F} = q(\mathbf{E} + \mathbf{v} \times \mathbf{B}). \quad (12)$$

This is known as *Lorentz's force law* and its validity has been unquestionably established by experiments.

### The Elements of Electric Circuits

In electrical engineering an important role is played by *circuits* in which power in the form of currents and fields is channelled by slender conductors (wires) connecting discrete elements. To understand the fine detail of the behaviour of these elements it is necessary to solve Maxwell's partial differential equations. Fortunately most elements are amenable to an adequately accurate, approximate treatment which simplifies the situation enormously. This is called the *lumped parameter* approximation and we now illustrate this with a number of examples.

**Example 1.4.1. (Resistor).** Consider a conductor made of homogeneous material in the form of a cylinder of length  $\ell$  and cross section area  $S$ . It is assumed that the current density  $\mathbf{j}$  and the electric field  $\mathbf{E}$  within the conducting material are both constant and in the direction of the axis of the cylinder,  $\hat{\mathbf{z}}$ . A more general version of Ohm's law is  $\mathbf{j} = \sigma \mathbf{E}$ , where  $\sigma$  is called the *conductivity* of the material, see *Notes and References*. The voltage  $\mathcal{V}$  between the ends of the cylinder and the total current  $I$  are

$$\mathcal{V} = \int_0^\ell \langle \mathbf{E}, \hat{\mathbf{z}} \rangle ds = \|\mathbf{E}\|\ell, \quad I = \int_S \langle \mathbf{j}, \mathbf{n} \rangle dS = \|\mathbf{j}\|S.$$

Now since  $\|\mathbf{j}\| = \sigma \|\mathbf{E}\|$ , we have  $\mathcal{V} = (\frac{\ell}{\sigma S}) I = RI$ , where  $R = (\frac{\ell}{\sigma S})$  is the resistance. For example the resistance of a silver wire of length 1.265 m with a circular cross section of radius .048 cm is .0281 ohms.

*Joule* (1841) [281] reasoned, and then confirmed experimentally, that the energy dissipated as heat when a current  $I$  flows in a metallic conductor of resistance  $R$  is  $RI^2$ .  $\square$

**Example 1.4.2. (Capacitor).** Consider two parallel plates charged with constant charges of equal magnitude but opposite sign. If the distance between the plates is small compared with the size of the plates, the charge will reside almost entirely on the inner surfaces of the plates, the electric field will be zero in the interior of the plates and away from the edges of the plates the electric field between the plates is approximately normal to them. Hence in this region between the plates the potential will only change in a direction  $x_1$  perpendicular to the plates. So Poisson's equation for the potential in Cartesian coordinates reduces to  $\Phi_{x_1 x_1} = 0$ , where  $(\ )_{x_1} = \frac{\partial}{\partial x_1}$ . The solution of this equation has the form  $\Phi(x_1) = \alpha x_1 + \beta$ , where  $\alpha$  and  $\beta$  are constants. Suppose the plates are at  $x_1 = a$  and  $x_1 = b$  and the potentials are constant on each plate and are  $\Phi(a) = V_a$  and  $\Phi(b) = V_b$ , then

$$\Phi(x_1) = \frac{(V_b - V_a)x_1 + bV_a - aV_b}{b - a}.$$

Now consider a closed cylindrical surface  $S$  where the axis of the cylinder is in the  $x_1$  direction and the plane ends of area  $A$  are at  $x_1 = a - \varepsilon$  and  $x_1 = a + \varepsilon$  with  $\varepsilon \ll b - a$ . If  $q$  is the constant surface density of charge (positive on the one at  $x_1 = a$  and negative on the other), then the charge enclosed in  $S$  is  $qA$ . Hence by (2)

$$qA/\epsilon_0 = \int_S \langle \mathbf{E}, \mathbf{n} \rangle dS = - \int_{S_1} \langle \text{grad } \Phi, \mathbf{n} \rangle dS = \frac{V_a - V_b}{b - a} \int_{S_1} dS = \frac{(V_a - V_b)A}{b - a},$$

where  $S_1$  is the plane surface at  $x_1 = a + \varepsilon$ . Thus

$$V_a - V_b = Q/C, \quad C = \frac{A\epsilon_0}{b - a}$$

where  $Q$  is the total charge on the plate at  $x_1 = a$ . So the potential (or voltage) change across the plates is proportional to the charge. The proportionality constant  $C$  is called the *capacitance* and such a configuration is called a *capacitor* or *condenser*. The above result neglects fringing of the electric field at the edges of the plates, for a more accurate expression for the capacitance see *Notes and References*.

Now let us consider the electric energy stored in the capacitor. The capacitor is charged by connecting the plates in a circuit with a battery which has the effect of transferring charge from one plate to the other. If a small charge  $dQ$  is brought from a position  $x_1 = b$  where the potential is  $\Phi(b) = V_b$  to a position  $x_1 = a$  where potential is  $\Phi(a) = V_a$ , then the work done is  $dW = (V_a - V_b)dQ$ . Hence  $dW = (Q/C)dQ$  and so the total work done in charging the capacitor is  $W = Q^2/(2C)$ , where  $\pm Q$  are the final charges on the plates.  $\square$

**Example 1.4.3. (Inductor).** Consider a coil consisting of  $n$  turns of wire which are tightly wound on a toroidal frame of rectangular cross section and permeability  $\mu_0$ . The inner and outer radii of the frame are  $r_1$  and  $r_2$ , respectively, the height of the frame is  $h$  and there is a current of magnitude  $I(t)$ ,  $t \geq 0$  in the conducting wire.

Suppose that cylindrical coordinates are such that the  $z$ -axis is the axis of symmetry

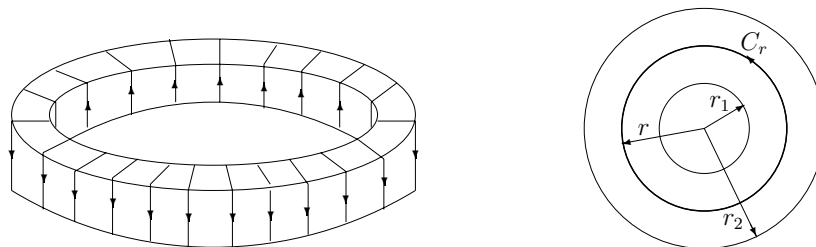


Figure 1.4.1: Toroidal inductor

and the frame is located between  $r = r_1$  and  $r = r_2$ . Let  $C_r$  be a circular path within the toroidal frame of radius  $r$  with  $r_1 < r < r_2$ . We assume axial symmetry so that the magnetic field  $\mathbf{B} = \mathbf{B}(r, z, t)$  only depends on  $r, z$  and  $t$ . By Ampère's law applied to the surface of the disk bounded by  $C_r$ , we have

$$\mu_0 n I(t) = \oint_{C_r} \langle \mathbf{B}, \mathbf{t} \rangle ds$$

where  $\mathbf{t}$  is the unit tangent to  $C_r$ . If  $B_2$  is the magnitude of the magnetic field in the direction  $\mathbf{t}$ , then

$$\oint_{C_r} \langle \mathbf{B}, \mathbf{t} \rangle ds = \int_0^{2\pi} B_2(r, z, t) ds = 2\pi r B_2(r, z, t).$$

So  $B_2(r, z, t) = \mu_0(2\pi r)^{-1}nI(t)$ . Since the coil is tightly wound around the toroidal frame, every loop approximately traces out the perimeter of a surface which is a rectangular cross section of the frame. Let us apply *Faraday's law* to one such surface  $S$ . Then the normal to this surface  $\mathbf{n} = \mathbf{t}$ . So if  $\mathcal{V}(t)$  is the induced voltage

$$\begin{aligned} \mathcal{V}(t) &= -\frac{d}{dt} \int_S \langle \mathbf{B}, \mathbf{n} \rangle dS = -\frac{d}{dt} \int_0^h \int_{r_1}^{r_2} B_2(r, z, t) dr dz = -\frac{\mu_0 n}{2\pi} \frac{dI}{dt}(t) \int_0^h \int_{r_1}^{r_2} r^{-1} dr dz \\ &= -\frac{\mu_0 n h}{2\pi} \ln(r_2/r_1) \frac{dI}{dt}(t). \end{aligned}$$

Since there are  $n$  such coils, if  $V$  is the total voltage dropped, we have

$$V(t) = L\dot{I}(t), \quad L = \frac{\mu_0 n^2 h}{2\pi} \ln(r_2/r_1).$$

The constant  $L$  is called the *inductance* and such a configuration is called an *inductor*.

Now let us consider the magnetic energy stored in the inductor. Each charge in the wire is receiving energy at a rate  $\langle \mathbf{E}, \mathbf{v} \rangle$  where  $\mathbf{E}$  is the force on it and  $\mathbf{v}$  is its velocity. So that if  $\rho$  is the density of charge per unit length the rate of doing work on the coil is

$$\frac{dW}{dt} = \oint_{coil} \langle \mathbf{E}, \mathbf{v} \rangle \rho ds = \oint_{coil} \langle \mathbf{E}, \mathbf{j} \rangle ds = I \oint_{coil} \langle \mathbf{E}, \mathbf{t} \rangle ds = VI = LI \frac{dI}{dt},$$

by (8). So we see that the energy required to build up the current  $I$  in the inductor is  $W = (L/2)I^2$ .  $\square$

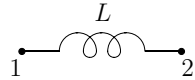
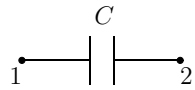
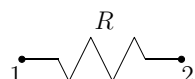
Symbol	Constitutive Law	Variables
	$V_1 - V_2 = LI$	voltage change across an inductor of inductance $L$ with a current $I$
	$V_1 - V_2 = Q/C$	voltage change across a capacitor of capacitance $C$ with a charge $Q$ on one plate and $-Q$ on the other
	$V_1 - V_2 = IR$	voltage change across a resistor of resistance $R$ with a current $I$

Table 1.4.2: Symbols and constitutive laws of a resistor, capacitor and inductor

Inductors, capacitors and resistors are the classical elements of electric circuits. Their symbols and constitutive laws are shown in Table 1.4.2. They are, respectively, the counterparts of masses, springs and dampers in mechanical systems. This correspondence is shown in Table 1.4.3 where  $M, k, c$  are the mass, spring constants and damping coefficient respectively,  $F$  is force,  $y$  displacement,  $v$  velocity and  $\mathcal{T}, \mathcal{W}, \mathcal{D}$  are the kinetic energy of the mass, the potential energy of the spring and the energy

mass	inductor	spring	capacitor	damper	resistor
$M$	$L$	$k$	$1/C$	$c$	$R$
$F$	$V_1 - V_2$	$F$	$V_1 - V_2$	$F$	$V_1 - V_2$
$v$	$I$	$y$	$Q$	$v$	$I$
$F = M\dot{v}$	$V_1 - V_2 = L\dot{I}$	$F = ky$	$V_1 - V_2 = Q/C$	$F = cv$	$V_1 - V_2 = IR$
$\mathcal{T} = (M/2)v^2$	$W = (L/2)I^2$	$\mathcal{W} = (k/2)y^2$	$W = Q^2/(2C)$	$\mathcal{D} = cv^2$	$W = RI^2$

Table 1.4.3: Table of corresponding quantities

dissipated by the damper. For example in the first column mass corresponds to inductance, the force on the mass corresponds to the voltage change across the inductor and the velocity of the mass corresponds to the current in the inductor. The last two rows gives the corresponding constitutive laws and energies of the elements. The correspondence given in Table 1.4.3 is called the Force-Voltage analogy. There is also a Force-Current analogy, see [397]. These analogies suggest that the variational method described for mechanical systems in the previous section can also be applied to electrical systems, and indeed this is the case, see *Notes and References* and the following example.

**Example 1.4.4. (Linear RLC circuit).** Consider the circuit driven by a voltage source  $e(t)$  as in Figure 1.4.4. The corresponding mechanical system is given on the left hand side of Figure 1.3.3. A systematic way of determining the laws of motion for the circuit will be explained in the next subsection. There we will see that by Kirchhoff's

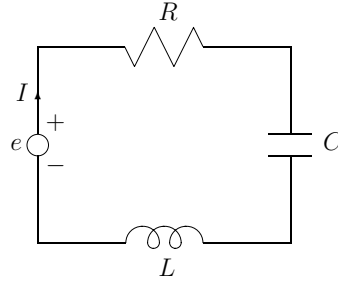


Figure 1.4.4: RLC circuit

voltage law the sum of the voltages around the closed circuit is zero. To be more precise if the current is in the direction indicated in Figure 1.4.4, then there will be a drop in the voltage across each of the elements. And Kirchhoff's law states that the total voltage drop across these elements must be balanced by that supplied by the voltage source. So if the voltage across the resistor, capacitor and inductor at time  $t$  are  $V_R(t)$ ,  $V_C(t)$ ,  $V_L(t)$ , respectively, we have

$$e(t) - V_R(t) - V_C(t) - V_L(t) = 0, \quad t \geq 0.$$

But if the current around the circuit at time  $t$  is  $I(t)$  and the charge on the capacitor is  $Q(t)$ , then

$$V_R(t) = I(t)R, \quad V_C(t) = Q(t)/C, \quad V_L(t) = L\dot{I}(t), \quad I(t) = \dot{Q}(t), \quad t \geq 0.$$

Hence

$$e(t) = L\ddot{Q}(t) + R\dot{Q}(t) + Q(t)/C, \quad t \geq 0.$$

If  $\mathcal{T} = (L/2)\dot{Q}^2$  is the magnetic energy of the inductor,  $\mathcal{W} = Q^2/(2C)$  the electric energy of the capacitor,  $\mathcal{D} = R\dot{Q}^2$  the energy dissipated by the resistor,  $F(t) = e(t)$  and  $L = \mathcal{T} - \mathcal{W}$ , then the above equation can be obtained directly via the variational method by writing down Lagrange's equation (3.39).

$$\frac{d}{dt} \left( \frac{\partial L}{\partial \dot{Q}}(Q(t), \dot{Q}(t)) \right) - \frac{\partial L}{\partial Q}(Q(t), \dot{Q}(t)) + \frac{1}{2} \frac{\partial \mathcal{D}}{\partial \dot{Q}}(\dot{Q}(t)) = F(t).$$

Setting  $x_1 = Q$ ,  $x_2 = \dot{Q}$ , we can re-write the equation of motion as a system of first order equations, namely

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -1/LC & -R/L \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1/L \end{bmatrix} e.$$

Suppose we are interested in determining the charge on the capacitor. It is difficult to measure the charge directly, so we may ask whether or not it is possible to determine the charge by measuring the current  $I = \dot{Q} = x_2$ . Setting  $y = x_2 = [0 \ 1]x$ , this is an observability problem: given the observation  $y(\cdot)$  and the input  $e(\cdot)$  on some time interval, is it possible to determine the state  $x(\cdot)$ ?  $\square$

In the following example we illustrate how the Lorentz force law (12) can be used to describe the interaction between electromagnetic forces and mechanical motion.

**Example 1.4.5. (Loudspeaker).** A loudspeaker is an electromechanical system in which the mechanical part is a loudspeaker diaphragm. Electromagnetic forces are used to make the diaphragm move and the consequent motion generates sound which is then transmitted through the air to the ear. Basically a signal from a tape, record, or disk generates an input voltage  $e(t)$  in a circuit. Part of this circuit is in the form of a coil within a fixed permanent magnet. The motion of the electric charges in the coil interacts with the magnetic field generated by the magnet to produce a Lorentz force as given by (12). Since the speaker diaphragm is rigidly attached to the coil this force on the coil causes the diaphragm to move. The whole idea is that the diaphragm motion which produces the sound should be proportional to the original input signal. An idealized model is given in Figure 1.4.5.

The magnet is cylindrical with an inner solid cylindrical core which is the South pole and an outer concentric cylindrical shell which is the North pole. It is assumed that this configuration results in a radial magnetic field in the air gap between the North and South poles directed to the axis of the magnet. In the figure the magnet is shown as dotted rectangles with small dots. The diaphragm is on the right of the figure and is shown as a rectangle with small circles, whereas the coil, which is rigidly connected to the diaphragm, is situated in the air gap between the North and South poles and is represented by small black circles inside other circles. It consists of  $n$  turns of wire each of which is at a distance  $a$  from the axis of the magnet. The motion of the diaphragm is modelled as an oscillator with mass  $m$ , damping  $c$  and stiffness  $k$  whereas the electric circuit is modelled as one which contains a resistor with resistance  $R$  and an inductor with inductance  $L$ . Suppose  $(r, \theta, z)$  are cylindrical coordinates where the  $z$ -axis is along the central axis of the magnet directed from the magnet to the diaphragm. It is assumed that the diaphragm and coil are constrained so that only motion in the  $z$  direction is allowed. Then if  $F(t)$  is the



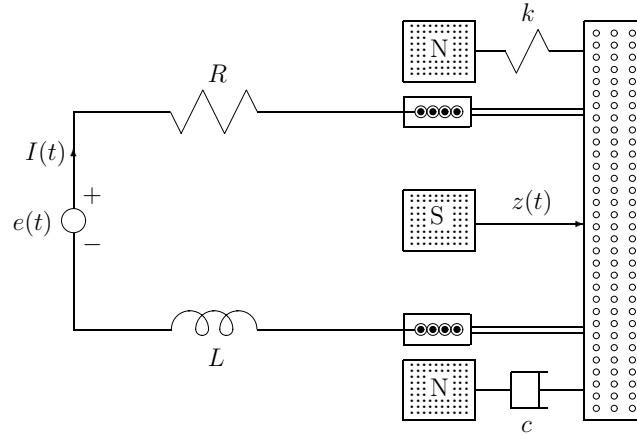


Figure 1.4.5: Loudspeaker

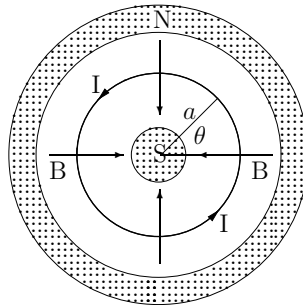


Figure 1.4.6: Magnet-coil geometry

component of the Lorentz force on the coil in this direction, since the coil and diaphragm are rigidly connected, the mechanical equation of motion of the diaphragm is

$$m\ddot{z}(t) + c\dot{z}(t) + kz(t) = F(t).$$

And if  $V(t)$  is the voltage induced as a consequence of the motion of the coil in the magnetic field, then just as in Example 1.4.4, one obtains the following equation of motion for the current  $I(t)$  in the circuit

$$L\dot{I}(t) + RI(t) = e(t) + V(t).$$

In order to complete the picture we find expressions for the terms  $F(t)$  and  $V(t)$ . The magnitude of the magnetic field  $\mathbf{B}$  at  $r = a$  is denoted by  $B$  and it is assumed to be independent of  $z$ ,  $\theta$  and  $t$ . If at a point in the coil parametrized by an angle  $\theta$ ,  $\theta \in [0, 2n\pi)$  there is a charge  $q(t, \theta)$  which has a velocity  $\mathbf{v}(t, \theta)$  the Lorentz force on it is  $q\mathbf{v} \times \mathbf{B}$ . The velocity  $\mathbf{v}$  has a component  $\mathbf{v}_1$  in the direction of the  $z$ -axis, but since  $\mathbf{v}_1 \times \mathbf{B}$  is perpendicular to the  $z$ -axis it will not make any contribution to  $F(t)$ . The other component of the charge's velocity is due to the movement of the charge around the coil. If its magnitude at  $(t, \theta)$  is  $v_2(t, \theta)$ , then the magnitude of the Lorentz force  $F_2$  in the direction of the  $z$ -axis is

$$F_2(t, \theta) = q(t, \theta)v_2(t, \theta)B.$$

Hence

$$\frac{d}{d\theta}F_2(t, \theta) = B \frac{d}{d\theta}(q(t, \theta)v_2(t, \theta)) = aBI(t).$$

So the total force in the direction of the  $z$ -axis is  $F(t) = 2n\pi aBI(t)$ . We see therefore that our mathematical model for the motion of the diaphragm is that of an oscillator driven by a force proportional to the current  $I(t)$  in the coil.

The induced voltage  $V(t)$  in the circuit is due to the motion of the coil in the  $z$  direction. In order to determine it we apply Faraday's law (11) with  $C$  being one loop of the coil in the magnet. If  $\mathbf{t}$  is the unit tangent to  $C$ , we have

$$\mathcal{V} = \oint_C \langle \mathbf{v}_1 \times \mathbf{B}, \mathbf{t} \rangle ds = -aB\dot{z} \int_0^{2\pi} d\theta = -2\pi aB\dot{z}.$$

And since the coil consists of  $n$  turns of wire the total induced voltage is given by  $V(t) = -2n\pi aB\dot{z}(t)$ . Setting  $x_1 = z$ ,  $x_2 = \dot{z}$ ,  $x_3 = I$ , we obtain the following state space system

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ -k/m & -c/m & 2n\pi aB/m \\ 0 & -2n\pi aB/L & -R/L \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} e.$$

Suppose  $y(t) = x_1(t) = z(t)$ , then the design problem is that of choosing some or all of the parameters  $k, c, L, R, n, a, B$  so that  $y(t)$  approximates the input  $e(t)$  for all  $t \geq 0$ .  $\square$

### 1.4.2 Electrical Networks

In this subsection we give a brief account of how graph theoretical methods are used to obtain models of interconnected electrical systems. We will only consider electrical networks consisting of voltage sources, current sources, resistors, inductors and capacitors. For the modelling of more general networks and more detail of network methods, see *Notes and References*.

Determining the differential equations which govern a complicated network can be quite difficult. Nowadays it is common to use computer aided modelling procedures. These are based on a graph theoretical representation of the electrical network. Before describing the details we recall some basic facts from graph theory.

A directed graph  $G = (V, E, \varphi)$  consists of a finite vertex set  $V$ , a finite edge set  $E$  and an *incidence map*

$$\varphi : E \rightarrow V^2, \quad e \rightarrow \varphi(e) = (\varphi_1(e), \varphi_2(e)).$$

If  $\varphi(e) = (v_1, v_2)$ , then one calls  $v_1$  the *initial* vertex and  $v_2$  the *terminal* vertex of the edge  $e$ . Equivalently, edge  $e$  is said to be directed from  $v_1$  to  $v_2$ . For a vertex  $v \in V$  it is useful to define the sets of edges with initial and terminal vertex  $v \in V$ , viz.

$$\begin{aligned} E(v, \cdot) &= \varphi_1^{-1}(v) = \{e \in E; \varphi_1(e) = v\} \\ E(\cdot, v) &= \varphi_2^{-1}(v) = \{e \in E; \varphi_2(e) = v\}. \end{aligned} \tag{13}$$

The cardinalities of these sets are called the *out-degree*,  $d_{out}(v)$ , and *in-degree*,  $d_{in}(v)$ , of  $v$ , respectively. Then  $d(v) = d_{out}(v) + d_{in}(v)$  is the total number of edges incident

on  $v$  and is called the *degree* of  $v$ .

A *path* of length  $r \geq 1$  is a sequence  $\underline{e} = (e_1, e_2, \dots, e_r) \in E^r$  with the property

$$\varphi_1(e_{i+1}) = \varphi_2(e_i) =: v_{k_i}, \quad 1 \leq i \leq r-1.$$

$v_{k_0} := \varphi_1(e_1)$  is called the initial vertex and  $v_{k_r} := \varphi_2(e_r)$  is called the terminal vertex. So one may equally think of  $\underline{e}$  as a path from  $v_{k_0}$  to  $v_{k_r}$ . An *elementary path* is one in which  $v_{k_1}, \dots, v_{k_r}$  are distinct and a *cycle* is an elementary path with  $v_{k_0} = v_{k_r}$ . The directed graph  $G$  is said to be *strongly connected*<sup>5</sup> if for any two distinct vertices  $v, v' \in V$  there exists a path from  $v$  to  $v'$ .

In our application to networks we do not always want the direction associated with an edge to play a role. A succinct way of achieving this with the above set up is to define for every edge  $e \in E$  an additional edge  $-e$  which is directed from the terminal vertex of  $e$  to the initial vertex of  $e$ . Let  $-E$  denote the set of these additional edges. Then  $E \cap -E = \emptyset$ .  $\varphi$  is extended to an incident map  $\tilde{\varphi}$  on  $\tilde{E} = E \dot{\cup} -E$  by setting  $\tilde{\varphi}(-e) = (\varphi_2(e), \varphi_1(e))$  for  $e \in E$ . This results in the graph  $\tilde{G} = (V, \tilde{E}, \tilde{\varphi})$ . A graph  $G' = (V', E', \varphi')$  is called a subgraph of  $G = (V, E, \varphi)$  if  $V' \subset V$ ,  $E' \subset E$  and  $\varphi' = \varphi|_{E'}$ . A *spanning subgraph* of  $G = (V, E, \varphi)$  is a subgraph  $G' = (V, E', \varphi|_{E'})$  with the same vertex set as  $G$ . It is a *proper spanning subgraph* of  $G$  if  $E' \neq E$ . Finally a *cut-set*  $C$  of  $G$  is a set of edges in  $E$  such that if all the edges  $c, -c$  with  $c \in C$  are removed from the graph  $\tilde{G}$ , the resulting graph decomposes into two strongly connected graphs (one of these may consist of a single vertex).

For the graph  $\tilde{G}$  we need the concept of a subgraph which inherits the “undirected” structure of  $\tilde{G}$ . We say that  $G' \subset \tilde{G}$  is a *symmetric subgraph* if it is a subgraph whose edge set  $E'$  has the following property:  $e \in E' \Leftrightarrow -e \in E'$  for all  $e \in E$ . A tree in  $\tilde{G}$  is a minimal symmetric strongly connected subgraph of  $\tilde{G}$  (or equivalently, a symmetric strongly connected subgraph without non-trivial cycles). One can show that a symmetric subgraph of  $\tilde{G}$  is a tree with  $n$  vertices if and only if it has  $2(n-1)$  edges. Moreover, if one adds one edge  $\tilde{e} \in \tilde{E}$  to a tree this creates exactly one cycle. A spanning tree in  $\tilde{G}$  is a spanning subgraph which is a tree in  $\tilde{G}$ . One can show that  $\tilde{G}$  always contains a spanning tree if it is strongly connected.

Now suppose that  $\tilde{G}$  is strongly connected. Given a spanning tree  $T$  of  $\tilde{G}$ , any cycle obtained by adding to  $T$  an edge of the graph  $\tilde{G}$  which is not an edge of the tree is called a *fundamental cycle* of  $\tilde{G}$  (with respect to the given spanning tree).

In electrical networks there are no self-loops, i.e. there are no edges with the property that  $\varphi_1(e) = \varphi_2(e)$ . The constitutive laws of the elements in the network are assumed to be the ones given in Table 1.4.2 and the resistances, capacitances and inductances of the connecting wires are neglected. A directed graph of the network is defined by replacing every element (resistor, inductor, capacitor, voltage and current sources) by an edge and the junction points of the wires (where the elements are connected together) by a vertex. Let  $E$  be the corresponding set of edges (network elements) and  $V$  the set of vertices (junction points). If a network element  $e$  joins the junction points  $v$  and  $v'$ , we may choose the direction of the edge  $e$  arbitrarily

---

<sup>5</sup>The directed graph  $G$  is called *connected* if the extended graph  $\tilde{G}$  is strongly connected. Note that a directed graph consisting of just two vertices and one edge between them, is connected but not strongly connected.

by setting either  $\varphi(e) := (v, v')$  or  $\varphi(e) := (v', v)$ . The incidence map  $\varphi : E \rightarrow V^2$  is defined by choosing one of these two possibilities for each edge  $e \in E$ . With these specifications we obtain a directed graph  $G = (V, E, \varphi)$  representing the electrical network. Associated with each edge  $e \in E$  are two time-varying weighting functions  $I_e(\cdot), V_e(\cdot) : [0, \infty) \rightarrow \mathbb{R}$ ; the current and voltage across the element which is represented by the edge. The direction of each edge is taken as reference direction for the current and the voltage drop. This is not a restriction, since negative values of  $I_e$  and  $V_e$  are allowed. However it does mean that  $I_e$  and  $V_e$  have the same sign. Since we also consider the graph  $\tilde{G}$  we have to associate with each edge  $-e \in -E$  a current and voltage and it is natural to set  $I_{-e} = -I_e$  and  $V_{-e} = -V_e$ , respectively. In assembling an electrical network by interconnecting various elements there are constraints on the currents and voltages given by *Kirchhoff laws*. The *current law* can be expressed in terms of the graph  $G$  whereas we need the extended graph  $\tilde{G}$  in order to state the *voltage law*.

**Kirchhoff's current law** states that the net current flow in and out of every vertex at the time  $t$  is zero, i.e.

$$\sum_{e \in E(\cdot, v)} I_e(t) - \sum_{e \in E(v, \cdot)} I_e(t) = 0, \quad t \geq 0, \quad v \in V. \quad (14)$$

Here  $E(v, \cdot)$  and  $E(\cdot, v)$  are defined by (13).

Since the current in the edge  $-e \in -E$  is by definition  $-I_e$  we could also have expressed the current law for the graph  $\tilde{G}$  with the result that the LHS of (14) would have doubled.

**Kirchhoff's voltage law** states that the total voltage drop around every cycle in  $\tilde{G}$  must be zero, i.e. if  $\underline{e} = (e_1, e_2, \dots, e_r)$  is a cycle in  $\tilde{E}$ , then

$$\sum_{j=1}^r V_{e_j}(t) = 0, \quad t \geq 0. \quad (15)$$

Suppose Kirchhoff's current law is written down for each vertex of  $G$  and we are given a cut-set for this graph. If we sum up the equations for all the vertices in either of the two subgraphs of  $G$  defined by the cut-set, only those currents entering or leaving the subgraph remain since the others cancel. So for the currents in the edges of the cut-set we have:

*Cut-set condition:* The sum of the currents entering one of the two subgraphs of  $G$  defined by a cut-set must equal the sum of the currents leaving it.

This version of Kirchhoff's current law is applied to each cut-set in  $E$ . Then Kirchhoff's voltage law is applied to each cycle in  $\tilde{G}$ . The resulting equations together with the constitutive laws of the elements are used to obtain a dynamical model for the electrical network. However there is a certain amount of redundancy if Kirchhoff's laws are applied to every cut-set and every cycle, in the sense that some of the equations are linearly dependent. Moreover it is not clear which variables should be eliminated and which ones retained in order to get a dynamical system model. Engineers have devised methods for overcoming these problems by means of a judicious choice of cut-sets, cycles and state space variables, see *Notes and References*. They recommend the following:

- (C1) Select a spanning tree of the graph  $\tilde{G}$  so that it contains all resistors, no current sources, and has as many capacitors and as few inductors as possible. In general these last two aims may be contradictory and a compromise must be made. For each edge in  $e \in E$  of the tree, find a cut-set<sup>6</sup> (a subset of  $E$ ) which contains the edge but no other edge in  $E$  of the spanning tree. Then for each such cut-set write down the equation determined by the corresponding *cut-set condition*.
- (C2) For every fundamental cycle obtained by adding to the spanning tree any edge of  $\tilde{G}$  write down Kirchhoff's voltage law (15).
- (C3) For every edge of the graph write down the constitutive law of the corresponding element of the network.
- (C4) Choose the charges on the capacitors and the currents through inductors which appear in the equations obtained by (C1), (C2) and (C3) as state space variables and eliminate all the others.

**Example 1.4.6.** Consider the network shown in Figure 1.4.7. The vertices of the asso-

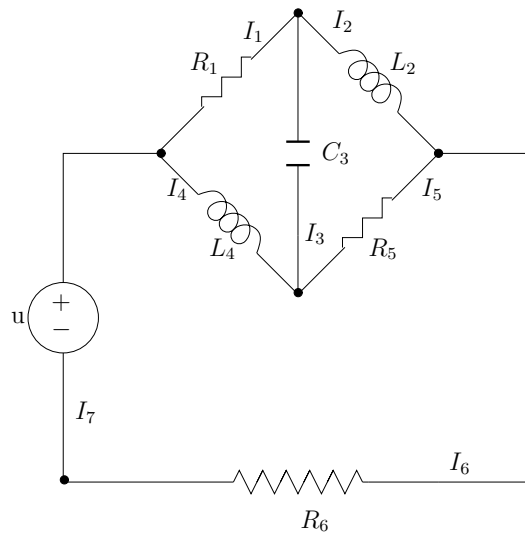
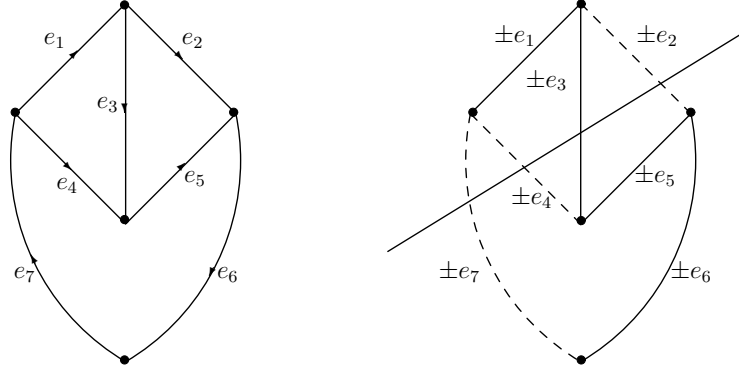


Figure 1.4.7: RLC Network

ciated graph correspond to the junction points marked with a  $\bullet$  in Figure 1.4.8 and the edges correspond to the network elements (1 capacitor, 2 inductors, 3 resistances and a voltage source). Directions for the edges are chosen arbitrarily and the choice we have made is shown in the directed graph on the left of Figure 1.4.8. The extended graph  $\tilde{G}$  is obtained from  $G$  by eliminating the arrowheads on the edges in  $G$ . Thus every line segment in the right hand graph of Figure 1.4.8 stands for a pair of edges  $\{e_i, -e_i\}$  of  $\tilde{G}$ . There are many spanning trees of the graph  $\tilde{G}$ , e.g.  $\{\pm e_1, \pm e_3, \pm e_5, \pm e_6\}$ ,  $\{\pm e_3, \pm e_5, \pm e_6, \pm e_7\}$ , and  $\{\pm e_6, \pm e_7, \pm e_4, \pm e_3\}$ . Guided by (C1) we choose to work with  $\{\pm e_1, \pm e_3, \pm e_5, \pm e_6\}$  since

<sup>6</sup>The cut-set is uniquely determined. It consists of  $e$  together with all those edges in  $E$  which connect a vertex of one of the subgraphs with a vertex of the other.

Figure 1.4.8: Directed graph  $G$  and spanning tree of  $\tilde{G}$ 

it contains all resistors, the capacitor and no inductor. This tree is drawn with continuous edges in the right hand figure of Figure 1.4.8 whereas all other edges of the graph are dashed. The cut-set containing edge  $e_3$ , is  $\{e_3, e_2, e_4, e_7\}$  and by the cut-set condition,

$$I_2 + I_3 + I_4 = I_7. \quad (16)$$

The cut-set containing edge  $e_1$  is  $\{e_1, e_4, e_7\}$ , so

$$I_1 + I_4 = I_7, \quad (17)$$

The cut-set containing edge  $e_5$  is  $\{e_5, e_2, e_7\}$ , so

$$I_5 + I_2 = I_7. \quad (18)$$

Finally the cut-set containing edge  $e_6$  is  $\{e_6, e_7\}$ , so

$$I_6 = I_7. \quad (19)$$

Guided by (C2) we have to find the fundamental cycles in  $\tilde{G}$  associated with the spanning tree  $\{\pm e_1, \pm e_3, \pm e_5, \pm e_6\}$ . These are  $(-e_2, e_3, e_5)$  and the reverse cycle  $(e_2, -e_5, -e_3)$ ,  $(-e_4, e_1, e_3)$  and the reverse cycle  $(e_4, -e_3, -e_1)$ ,  $(e_1, e_3, e_5, e_6, e_7)$  and the reverse cycle  $(-e_1, -e_7, -e_6, -e_5, -e_3)$ . Applying Kirchhoff's voltage law to these cycles we have

$$-V_2 + V_5 + V_3 = 0, \quad (20)$$

$$-V_4 + V_3 + V_1 = 0, \quad (21)$$

$$V_1 + V_3 + V_5 + V_6 = u. \quad (22)$$

The equations (16)-(19) and (20)-(22) are augmented with the constitutive laws

$$V_1 = I_1 R_1, \quad V_2 = L \dot{I}_2, \quad V_3 = Q_3 / C_3, \quad V_4 = L \dot{I}_4, \quad V_5 = I_5 R_5, \quad V_6 = I_6 R_6. \quad (23)$$

We now follow (C4) and choose  $I_2, Q_3, I_4$  as state variables and eliminate all the other variables in (16)-(23). To this end, from (16)-(19) we have

$$\begin{bmatrix} I_1 \\ I_5 \\ I_6 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} I_2 \\ I_3 \\ I_4 \end{bmatrix}$$

and substituting in (20)–(22) and using the expressions for  $V_1, V_5, V_6$  in (23) yields

$$\begin{bmatrix} 1 & -R_5 & 0 \\ 0 & -R_1 & 1 \\ 0 & R & 0 \end{bmatrix} \begin{bmatrix} V_2 \\ I_3 \\ V_4 \end{bmatrix} = \begin{bmatrix} 0 & 1 & R_5 \\ R_1 & 1 & 0 \\ -(R_1 + R_6) & -1 & -(R_5 + R_6) \end{bmatrix} \begin{bmatrix} I_2 \\ V_3 \\ I_4 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} u,$$

where  $R = R_1 + R_5 + R_6$ . But

$$\begin{bmatrix} 1 & -R_5 & 0 \\ 0 & -R_1 & 1 \\ 0 & R & 0 \end{bmatrix}^{-1} = R^{-1} \begin{bmatrix} R & 0 & R_5 \\ 0 & 0 & 1 \\ 0 & R & R_1 \end{bmatrix}$$

and hence

$$R \begin{bmatrix} V_2 \\ I_3 \\ V_4 \end{bmatrix} = \begin{bmatrix} -R_5(R_1 + R_6) & (R_1 + R_6) & R_1 R_5 \\ -(R_1 + R_6) & -1 & -(R_5 + R_6) \\ R_1 R_5 & R_5 + R_6 & -R_1(R_5 + R_6) \end{bmatrix} \begin{bmatrix} I_2 \\ V_3 \\ I_4 \end{bmatrix} + \begin{bmatrix} R_5 \\ 1 \\ R_1 \end{bmatrix} u.$$

Then using (23) and  $I_3 = \dot{Q}_3$  we get

$$R \begin{bmatrix} \dot{I}_2 \\ \dot{Q}_3 \\ \dot{I}_4 \end{bmatrix} = \begin{bmatrix} -R_5(R_1 + R_6)/L_2 & (R_1 + R_6)/L_2 C_3 & R_1 R_5/L_2 \\ -(R_1 + R_6) & -1/C_3 & -(R_5 + R_6) \\ R_1 R_5/L_4 & (R_5 + R_6)/L_4 C_3 & -R_1(R_5 + R_6)/L_4 \end{bmatrix} \begin{bmatrix} I_2 \\ Q_3 \\ I_4 \end{bmatrix} + \begin{bmatrix} R_5 \\ 1 \\ R_1 \end{bmatrix} u.$$

This is the dynamical model for the given RLC network obtained by following the guidelines (C1)–(C4). Note that

$$\begin{bmatrix} I_2 \\ Q_3 \\ I_4 \end{bmatrix} = \begin{bmatrix} 0 & -1 & 1 \\ -R_1 C_3 & -R_5 C_3 & -R_6 C_3 \\ -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} I_1 \\ I_5 \\ I_6 \end{bmatrix} + \begin{bmatrix} 0 \\ C_3 \\ 0 \end{bmatrix} u.$$

Using this transformation we could also write down differential equations for  $I_1, I_5, I_6$  in violation of the guidelines. In this case both  $u$  and its derivative  $\dot{u}$  will appear on the RHS. We will later restrict our analysis to dynamical models which do not contain derivatives of input variables. So without further modification the dynamical model in terms of  $I_1, I_5, I_6$  will not fit this specification.  $\square$

### 1.4.3 Notes and References

A classical reference on electromagnetic theory is the book of *Elliott* (1966) which has been republished as an IEEE reprint, see [151]. Its main features are the historical material in each chapter and the development, via special relativity, of a complete electromagnetic theory. We also recommend the Lecture Notes on Physics by *Feynman* (1975) [161].

The more general version of Ohm's law and the effect of the fringing of the electric field on the capacitor considered in Example 1.4.2 can be found in [151]. A good book on vector fields developed through its application to engineering is *Shercliff* (1977) [463]. As an elementary mathematical introduction to Vector Analysis we recommend the textbook of *Marsden and Tromba* (1996) [362]. For a discussion of the modelling of electrical and electromechanical systems, see *Ogata* (1992) [397], *Burton* (1994) [84], *Close and Frederick* (1995) [105] and for references on electrical circuits see e.g. *Johnson et al.* (1992) [278] and *Wellstead* (1979) [516]. A comprehensive account of graph theory is contained in *Thulasiraman and Swamy* (1992) [495]. A concise description of how to use graph theoretical tools for the modelling of electrical networks can be found in *Zerz* (2000) [545], see also [278] and [516].

## 1.5 Digital Systems

In recent years, due mainly to the simultaneous dramatic improvement and reduction in cost of digital hardware, digital systems have become all pervasive in technology. They form a class of dynamical systems with quite distinctive features and therefore special engineering and mathematical disciplines have been developed for their analysis and design: “Theory of Switching Networks”, “Automata Theory”, “Logic Design”, see *Notes and References*. Although these areas are not subjects of this book, it is appropriate to discuss some examples and special features of digital systems since they are not only an important class of dynamical systems in themselves but are also increasingly used in the control and measurement of analog signals and systems. Indeed many analog devices in signal processing, filtering and control have been replaced by digital counterparts which are often cheaper, more robust and more reliable.

The essential difference between analog and digital systems is that in the former ones input, output and internal state variables take on a continuous range of values whereas in the latter ones there are only a finite number of input, output and state values. Most digital systems are *binary*, i.e. their input, output and state variables take only two different values, “on” and “off”. Physically these values may be encoded by different voltages (e.g. 5 volts versus 0 volts), by the flow or non-flow of an electrical current or by magnetic polarization (North and South). Mathematically the “on” and “off” values are usually represented by 1 and 0, the elements of the simplest nontrivial Boolean algebra  $\mathbb{B} = \{0, 1\}$  or, alternatively, the binary field  $\mathbb{Z}_2 = \mathbb{Z}/(2)$ .

Because of its binary components a digital system is often viewed as a network of switches which operates in discrete time  $t \in \mathbb{N}$  or  $\mathbb{Z}$ . There are two basic classes.

- *Combinational switching networks* are those whose current outputs depend only on the current inputs. Dynamical systems with this property are called *memoryless*, they transform the inputs directly into outputs without intermediate storage of energy or information. Physically, the output changes a short time after the input changes, but this short time delay is neglected in the mathematical description of the digital system. By convention the “current” input at time  $t \in \mathbb{Z}$ ,  $u(t)$ , determines the “current” output,  $y(t)$ .<sup>1</sup> If such a combinational network has  $m$  input and  $p$  output channels its behaviour is completely described by a function  $F$  mapping the  $2^m$  possible input vectors  $u(t) \in \mathbb{B}^m$  into the corresponding output vectors  $y(t) = F(u(t)) \in \mathbb{B}^p$ .
- *Sequential switching networks* or *finite state machines* are those digital systems whose current outputs depend not only on the current inputs but also on the sequence of previous inputs. Such systems (for example a digital clock or a computer) contain memory elements in which information about the history of previous inputs is stored. The (binary) contents of all its, say  $n$ , memory elements form together a binary vector  $x \in \mathbb{B}^n$  which is called the *state* of

---

<sup>1</sup>Alternatively, one could redefine the time dependence of the output function in such a way that the present input  $u(t)$  determines the *next* output  $y(t+1)$ . In fact, this alternative convention is usually chosen in the mathematical description of *sequential* networks, see Example 1.5.2.



the system. The current state and the current input together determine the current output and the next state of the system. The behaviour of a sequential switching network is therefore described by two maps, which determine the current output and the next state as functions of the current input and the current state of the system. This is in contrast with combinational switching networks where there is no need to introduce the notion of state.

Before describing some elementary building blocks of these two types of digital systems we illustrate the difference between combinational and sequential switching networks by two examples.

**Example 1.5.1. (Half and full adder).** Suppose we want to add two binary digits  $A$  and  $B$ . A combinational switching network which performs this addition is called a *half adder*. It accepts two binary digits  $A$  and  $B$  (bits) as inputs and produces two binary digits as outputs, the “sum”  $S = A \cdot (1 - B) + (1 - A) \cdot B$  in  $\mathbb{Z}_2$  (i.e.  $A + B \bmod 2$ ) and the “carry”  $C = A \cdot B$ . The binary number  $CS$  formed by the two outputs is the dyadic representation of the sum of  $A$  and  $B$  in  $\mathbb{Z}$ ,  $A + B = C2^1 + S2^0$ .

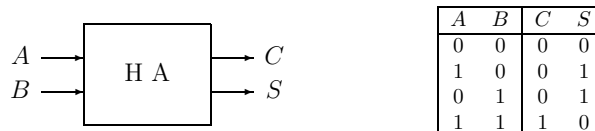


Figure 1.5.1: Block diagram and input-output table of half adder

When two binary numbers are added digit by digit, a third input must be considered, the carry-in from the next lower position. This yields the *full adder*. By a combination of half and full adders one can construct memoryless digital systems for the addition of arbitrary binary numbers of limited length. For instance, one can construct a machine for computing the sum of two binary numbers of 4 digits each by connecting in series one half adder and three full adders.

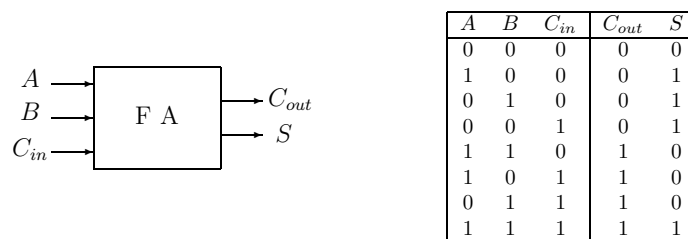


Figure 1.5.2: Block diagram and input-output table of full adder

In the above descriptions of the half and the full adder, time does not play a role since these digital systems are both *memoryless* and *time-invariant*. The current output  $y(t) = (C_{out}, S)$  is completely determined by the *current* input  $u(t) = (A, B)$  (resp.  $u(t) = (A, B, C_{in})$ ); it does not depend upon the previous inputs (the system has no memory). Moreover, identical input vectors always determine the same output vector (independent of the time  $t$  at which they are applied), the input-output relationship does not change with time, it is time-invariant.  $\square$

**Example 1.5.2. (Parity check machine).** Whenever digital systems are used in computing or communication it is necessary to convert numbers and letters into strings of 1's and 0's. A map  $F : U \rightarrow \mathbb{B}^p$  which maps a finite input alphabet (set of characters)  $U$  injectively into the set  $\mathbb{B}^p$  of  $p$ -bit strings (code words) is called a block code of size  $p$ . An arbitrary string of  $p$  bits may or may not be a code word for the code  $F$ . An encoding device can be described as a memoryless time-invariant digital system which accepts inputs  $u$  from the finite input alphabet  $U$  and transforms these into outputs  $y = F(u) \in \mathbb{B}^p$ . A widely used alpha-numerical code is the ASCII code. This is a seven-bit code for the 10 decimal numbers, the 26 lower-case and 26 upper-case characters of the English language and a large number of special characters, such as “+”, “)”, “%” etc. With seven bits it is possible to encode at most  $2^7$  characters.

When information is encoded and transmitted some bits may be changed due to electrical noise or other transient failures. The change of a single bit can be detected by adding one bit to each code word in such a way that after this addition each valid code word has an even number of 1's, e.g. the ASCII code word for  $a$  is 1100001. This word has odd parity and so a 1 would be prefixed to the code word in order to achieve even parity. Thus the enlarged code word permitting error detection would be 11100001. If now one bit is changed in the code word, say by a transmission failure, the error would be detected by examining the parity of the transmitted word.<sup>2</sup> This can be done by a *parity checker*, a device which responds to a finite binary sequence  $(u(0), u(1), \dots, u(t))$  with the output  $y(t+1) = 0$  (in the next time unit) if the number of 1's in the sequence is even (no error), and with a 1 if not (error). The next output of a parity checker clearly depends not only on the current but also on the past inputs. If the number of ones in the past input sequence  $(u(0), u(1), \dots, u(t-1))$  is even the next output  $y(t+1)$  is equal to the current input  $u(t)$ . If, however, the number of ones in the past input sequence is odd, the next output is the complement of the current input,  $y(t+1) = \overline{u(t)} = 1 - u(t)$ . These two cases lead to the idea of constructing a parity checker as a machine with two states, *Even* and *Odd*, which “remember” the parity of the past output sequence and are encoded by 0 and 1. The state transition of the parity checker under the influence of the present input is represented by its *state transition graph* and is explicitly described in the “next state table”, see Figure 1.5.3.

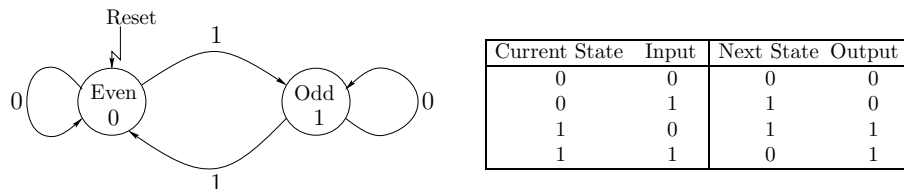


Figure 1.5.3: State transition graph and next state table of a parity checker

The system equations of the parity checking machine are

$$\begin{aligned} x(t+1) &= x(t) + u(t), \quad t \in \mathbb{N}, \quad x(0) = 0 \\ y(t) &= x(t) \end{aligned}$$

<sup>2</sup>Note that if two bits are simultaneously changed in a code word this will *not* be discovered by a parity checker. However, the occurrence of a double error is much less probable than the occurrence of a single error ( $p^2$  instead of  $p$  if  $p$  is the probability of a single error, assuming independence of the transmission errors).

where  $u(t), x(t), y(t) \in \mathbb{Z}_2$  denote the current input, state and output, and  $x(t+1) \in \mathbb{Z}_2$  is the next state. (The addition on the RHS of the first equation is taken in the binary field  $\mathbb{Z}_2$ ). In order that this machine can be used for detecting errors in code words, a reset mechanism is needed which allows one to reset the state of the machine to 0 after the examination of each code word.  $\square$

### 1.5.1 Combinational Switching Networks

In this brief subsection we describe some of the elementary building blocks of combinational networks, the logic gates, and illustrate how simple arithmetic units, like the half and the full adder, can be built from these gates. We also explain by means of an example how the digital input–output behaviour of a gate can be approximately realized by a continuous nonlinearity.

**Example 1.5.3. (Logic gates and half adder).** A logic gate is an electronic device with two (or more) binary inputs and one binary output which performs simple logical operations. Its input–output behaviour can be described by a truth table or in terms of the three basic Boolean operations  $\wedge, \vee$  and complementation. The three logic gates AND, OR, NOT which perform these operations are described in the following table together with a NOR gate which is a cascade connection of an OR-gate and the “inverter” NOT.


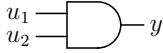
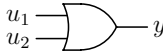
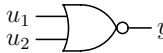
	$y = \bar{u} \quad [= 1 - u]$ NOT	<table border="1" style="border-collapse: collapse; text-align: center;"> <thead> <tr> <th><math>u</math></th> <th><math>y = \bar{u}</math></th> </tr> </thead> <tbody> <tr> <td>0</td> <td>1</td> </tr> <tr> <td>1</td> <td>0</td> </tr> </tbody> </table>	$u$	$y = \bar{u}$	0	1	1	0									
$u$	$y = \bar{u}$																
0	1																
1	0																
	$y = u_1 \wedge u_2 \quad [= u_1 \cdot u_2]$ AND	<table border="1" style="border-collapse: collapse; text-align: center;"> <thead> <tr> <th><math>u_1</math></th> <th><math>u_2</math></th> <th><math>y = u_1 \wedge u_2</math></th> </tr> </thead> <tbody> <tr> <td>0</td> <td>0</td> <td>0</td> </tr> <tr> <td>0</td> <td>1</td> <td>0</td> </tr> <tr> <td>1</td> <td>0</td> <td>0</td> </tr> <tr> <td>1</td> <td>1</td> <td>1</td> </tr> </tbody> </table>	$u_1$	$u_2$	$y = u_1 \wedge u_2$	0	0	0	0	1	0	1	0	0	1	1	1
$u_1$	$u_2$	$y = u_1 \wedge u_2$															
0	0	0															
0	1	0															
1	0	0															
1	1	1															
	$y = u_1 \vee u_2 \quad [= 1 - (1 - u_1) \cdot (1 - u_2)]$ OR	<table border="1" style="border-collapse: collapse; text-align: center;"> <thead> <tr> <th><math>u_1</math></th> <th><math>u_2</math></th> <th><math>y = u_1 \vee u_2</math></th> </tr> </thead> <tbody> <tr> <td>0</td> <td>0</td> <td>0</td> </tr> <tr> <td>0</td> <td>1</td> <td>1</td> </tr> <tr> <td>1</td> <td>0</td> <td>1</td> </tr> <tr> <td>1</td> <td>1</td> <td>1</td> </tr> </tbody> </table>	$u_1$	$u_2$	$y = u_1 \vee u_2$	0	0	0	0	1	1	1	0	1	1	1	1
$u_1$	$u_2$	$y = u_1 \vee u_2$															
0	0	0															
0	1	1															
1	0	1															
1	1	1															
	$y = \overline{u_1 \vee u_2} \quad [= (1 - u_1) \cdot (1 - u_2)]$ NOR	<table border="1" style="border-collapse: collapse; text-align: center;"> <thead> <tr> <th><math>u_1</math></th> <th><math>u_2</math></th> <th><math>y = \overline{u_1 \vee u_2}</math></th> </tr> </thead> <tbody> <tr> <td>0</td> <td>0</td> <td>1</td> </tr> <tr> <td>0</td> <td>1</td> <td>0</td> </tr> <tr> <td>1</td> <td>0</td> <td>0</td> </tr> <tr> <td>1</td> <td>1</td> <td>0</td> </tr> </tbody> </table>	$u_1$	$u_2$	$y = \overline{u_1 \vee u_2}$	0	0	1	0	1	0	1	0	0	1	1	0
$u_1$	$u_2$	$y = \overline{u_1 \vee u_2}$															
0	0	1															
0	1	0															
1	0	0															
1	1	0															

Table 1.5.4: Logic Gates

The table shows the standard symbols of these gates, their truth tables and the expression of their outputs in terms of their inputs (in the Boolean algebra  $\mathbb{B}$ ). Note that these gates can also be described by arithmetic expressions in the binary field  $\mathbb{Z}_2$ , but while the AND gate corresponds to multiplication the OR gate *does not* correspond to addition in  $\mathbb{Z}_2$  (although  $\vee$  is often replaced by  $+$  in textbooks on logic design). The four gates NOT,

AND, OR, NOR correspond, respectively, to the following four operations in  $\mathbb{Z}_2$ :  $X \rightarrow \bar{X} = 1 - X$  (*complementation*),  $(X, Y) \rightarrow X \cdot Y$  (*multiplication*),  $(X, Y) \rightarrow 1 - (1 - X) \cdot (1 - Y)$  and  $(X, Y) \rightarrow (1 - X) \cdot (1 - Y)$ .

These gates can be combined to produce digital networks which perform more complicated logic or arithmetic functions. As an example we show in Figure 1.5.5 the realization of a

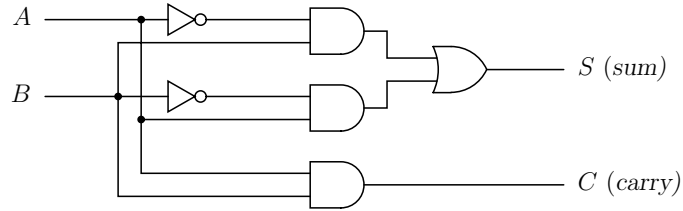


Figure 1.5.5: Realization of a half adder

*half adder* by a network of gates. The half adder is the simplest arithmetic circuit. The full adder (see Example 1.5.1) can be constructed from two half adders and an OR gate.

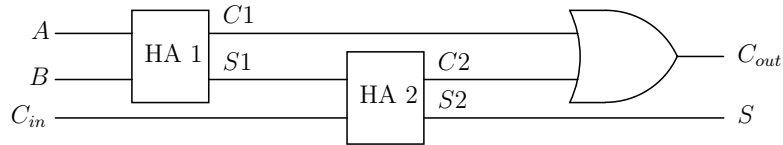


Figure 1.5.6: Realization of a full adder

Half and full adders are simple examples of *composite systems* i.e. systems composed of a number of interconnected subsystems. Very complex systems can be built in this way. In fact any function  $F : \mathbb{Z}_2^m \rightarrow \mathbb{Z}_2$  or, equivalently, any logical/Boolean operation can be realized by the three gates NOT, AND, OR<sup>3</sup>.  $\square$

Usually a given Boolean operation can be realized by an interconnection of gates in many different ways. A basic problem in logic design is that of “minimal realization”: Realize a given Boolean function by a network in which there is a minimum

- number of gates,
- number of gate inputs—this determines the amount of wiring within the network,
- number of cascaded levels of gates (i.e. the number of gates in the largest path from any input to any output). This number determines the overall time delay between inputs and outputs of the network.

Usually the above numbers cannot be minimized simultaneously so one must find a suitable compromise. Problems of minimal realization also arise in systems theory when one wants to realize a given input-output behaviour by a continuous or discrete time linear system with a minimal number of state variables, see Vol. II. Before going on to discuss the notion of a *sequential network (finite state machine)*

<sup>3</sup>For a variety of reasons, NAND and NOR gates (i.e. the inverted AND and OR gates) are preferred in practice to AND and OR gates for realizing logic circuits.

it is useful to comment on some important points related to the physical realization of both kinds of switching networks (combinational and sequential). The physical components of a digital system are constructed from electronic building blocks (resistors, diodes, transistors) – in other words *a digital system is built with analog building blocks* operating in continuous time with real valued input, output and state coordinates. *Natura non facit saltus*. So the physical quantities within the system (voltages, currents) when they move from one of their two values to the other, will vary over a continuous range of transitional values. One must, therefore, distinguish between the digital system as a mathematical model and its physical realization by an electronic circuit. A precise modelling of the latter would be based on ordinary or partial differential equations with a continuous time domain, and these differential equations would describe not only the transition from the current steady state of the circuit to the next one but the whole continuous trajectories of its state and output vectors. In the above, what has been written about digital systems is concerned with their ideal mathematical behaviour and does not exactly apply to their physical realization. In the following example we illustrate how a simple digital system can be *approximately* realized by an analog device.

**Example 1.5.4. (Inverter circuit).** The logic inverter NOT is a digital system whose inputs and outputs are binary digits. It transforms the input 0 into the output 1 and the input 1 into the output 0. However, the circuit which realizes this ideal digital behaviour operates over electrical voltages rather than digits. It accepts arbitrary input voltages in the range of say, 0 to 5 volts and produces output voltages over the same range. The essential property which makes it a good realization of the logic inverter is that it transforms voltages which are “not too far” from 0 volts into voltages very close to +5 volts (representing a logical 1) and voltages which are “not too far” from +5 volts into voltages very close to 0 volts (representing a logical 0). A typical input-output behaviour of

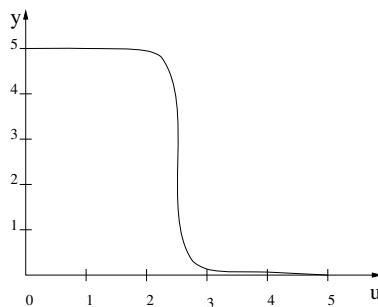


Figure 1.5.7: Input-output behaviour of an inverter circuit

such a circuit is shown in Figure 1.5.7. Here an input in the range of 0 to 2 volts produces an output of approximately 5 volts and an input in the range of 3 to 5 volts produces an output of approximately 0 volts. Thus we may say that the circuit “interprets” an input in the range of 0 to 2 volts as an input 0 producing an output 1 and if the input is in the range of 3 to 5 volts it interprets it as an input 1 and produces the output 0. So minor fluctuations in voltage levels are not misinterpreted by the circuit and have practically no influence on its output. A similar nonlinear behaviour which produces only two different

output values in response to a relatively wide range of inputs values is also exhibited by other building blocks of digital systems.  $\square$

**Remark 1.5.5.** In general, due to unavoidable variations in the manufacturing process the input–output behaviour of an electrical device will differ from its prescribed performance. Furthermore, every interconnection of electrical components is subject to noise and signal degradation along the wires. Hence it is important that the components of a switching network are sufficiently tolerant with respect to input variations. The tolerance of a digital device to deviations of the input signal from the reference voltages is called its *noise margin*. A good noise margin of the components is fundamental for the accuracy and reliability of a digital system. Cascaded digital circuits with a good noise margin (such as the above inverter) can correct signal degradations.  $\square$

## 1.5.2 Sequential Switching Networks

Sequential networks are required if data is to be stored in a network for future use. In this subsection we describe how data can be stored by latches and flip–flops and we discuss the use of clocks in order to synchronize the network elements and thereby enhance the reliability of the network. We outline the main steps in the design of a finite state machine and illustrate this by constructing a three bit counter.

In the next example we describe some basic memory elements of sequential networks.

**Example 1.5.6. (R–S latch and J–K latch).** Broadly speaking a digital system consists of a memory part that stores past data and a combinational part by which new outputs are generated from the stored data and the current inputs. The basic memory elements of a digital system are constructed by feedback interconnection between a (small) number of gates. The most primitive memory devices are *latches*, these are circuits which “latch” onto one bit (0 or 1) and remember it. As an example we consider the R–S latch which is obtained by feedback coupling of two NOR gates. It follows immediately from the definition that a NOR gate acts as an inverter if one of the inputs is set to 0. If one of the inputs is set to 1 its output is always 0. Now consider the cross–coupled NOR

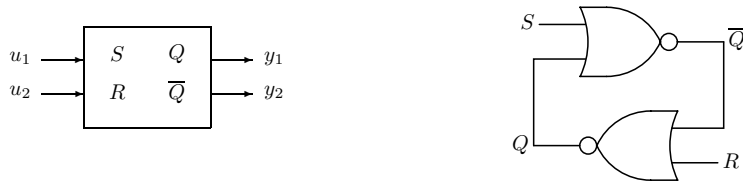


Figure 1.5.8: R–S latch: block diagram and realization by feedback of NOR gates

gates as depicted in Figure 1.5.8. It is required that the outputs of the two NOR gates have complementary values  $Q$  and  $\bar{Q}$ , respectively. The output  $Q$  of the lower NOR gate is said to be the state of the latch. If  $Q = 1$  it is said to be in the *set* state and if  $Q = 0$  it is said to be in the *reset* or *clear* state. Suppose that both inputs  $R$  and  $S$  are set to zero. Since each of the two NOR gates acts as inverter to the signal received from the other, the output values and hence the state remain unchanged (i.e. are stored) as long as both inputs are kept to zero. If  $R = 0$  and  $S = 1$  then the state (output of the lower NOR gate) is set to 1 whereas the output of the upper NOR gate is set to 0. If

$R = 1$  and  $S = 0$  then  $Q$  is reset to 0. Therefore  $S$  is called the *set* input and  $R$  the *reset* input. What happens if both inputs are set to 1, i.e. if the latch is simultaneously set and cleared? In that case the outputs of both NOR gates would necessarily take the value 0 so that the complementarity assumption of the two outputs would be violated. Moreover, if afterwards both inputs were to be simultaneously changed from 1 to 0 at time  $t$  the resulting (next) state and output value  $Q(t + \tau)$  would become unpredictable. If the upper gate switches first, its output  $\bar{Q}$  would switch to 1 and so the next state would be set to  $Q(t + \tau) = 0$ . If the lower gate switches first, then its output would

$u_1(t)$	$u_2(t)$	$x(t)$	$x(t + \tau)$	Comment
0	0	0	0	HOLD
0	0	1	1	
1	0	0	1	SET
1	0	1	1	
0	1	0	0	RESET
0	1	1	0	
1	1	0	?	NOT ALLOWED
1	1	1	?	

Table 1.5.9: Next state table of the R-S latch

switch to  $Q(t + \tau) = 1$  whilst  $\bar{Q}$  would switch to 0. Thus the next state  $Q(t + \tau)$  of the latch would depend upon which gate happens to be faster. Such a situation is referred to as a *race condition*. This unpleasant phenomenon is excluded if the two outputs never have the same value and this is secured if the input pair  $(u_1, u_2) = (1, 1)$  is not allowed. For admissible input pairs the behaviour of the R-S latch is described by the *output map*  $(y_1, y_2) = (x, 1 - x)$  and the *next state map*  $x(t + \tau) = (1 - x(t))u_1(t) + x(t)(1 - u_2(t))$ , see the next state Table 1.5.9. Here  $\tau$  is the propagation delay of the R-S latch, i.e. the time lag before the new steady state (output) is achieved in response to a change in the inputs.

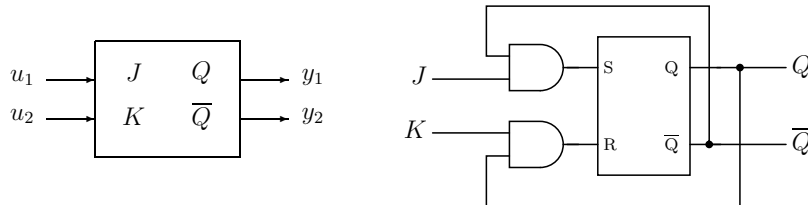


Figure 1.5.10: J-K Latch: Block diagram and circuit

In order to avoid the possibility of inadmissible inputs, the R-S latch can be connected with two additional AND gates as in Figure 1.5.10. By feeding back the outputs in the described manner it is guaranteed that the inputs  $R$  and  $S$  to the R-S latch are never simultaneously 1. The resulting circuit is called a *J-K latch* and is represented by the block diagram shown on the left in Figure 1.5.10. In addition to avoiding the forbidden input combination  $(R, S) = (1, 1)$  at the internal R-S latch the configuration shows a new capability,  *toggling*. If  $J = K = 1$  then the current state  $Q(t) = 0$  will toggle to  $Q(t + \tau) = 1$  and the current state  $Q(t) = 1$  will toggle to  $Q(t + \tau) = 0$ . Thus all possible input combinations lead to useful functions for the J-K latch: hold, reset, set, and toggle, see the next state Table 1.5.11. The behaviour of a J-K latch is described by the output

$u_1(t)$	$u_2(t)$	$x(t)$	$x(t + \tau)$	Comment
0	0	0	0	HOLD
0	0	1	1	
1	0	0	1	SET
1	0	1	1	
0	1	0	0	RESET
0	1	1	0	
1	1	0	1	TOGGLE
1	1	1	0	

Table 1.5.11: J-K Latch: Next state table

function  $(y_1, y_2) = (x, 1 - x)$  and the next state equation

$$x(t + \tau) = (1 - x(t))u_1(t) + x(t)(1 - u_2(t)) \quad x \in \mathbb{Z}_2, u \in \mathbb{Z}_2^2. \quad (1)$$

Note, however, that this equation is not to be understood in discrete time. Both the R-S and the J-K latches are *asynchronous* (or *unclocked*), i.e. they may change their state and outputs at any time in response to changes in the inputs.<sup>4</sup> This leads to a problem which becomes evident when these memory elements are realized by a circuit. For an asynchronous circuit to work properly, the inputs must be (approximately) constant for a sufficiently long time to allow the circuit to reach the corresponding next steady state. Moreover, only one external input should be effective (different from zero) at any given time. The reason for this is that if the two inputs  $u_1(t) = u_2(t) = 1$  for a time interval longer than the propagation delay through the latch, the outputs will toggle an unknown number of times, determined by the length of the interval and the time lag with which a change in the output signal travels, via the feedback loop, through the circuit back to the output. The phenomenon of “oscillating outputs” caused by identical inputs  $u_1(t) = u_2(t) = 1$  is illustrated in the timing diagram shown in Figure 1.5.12. So, although

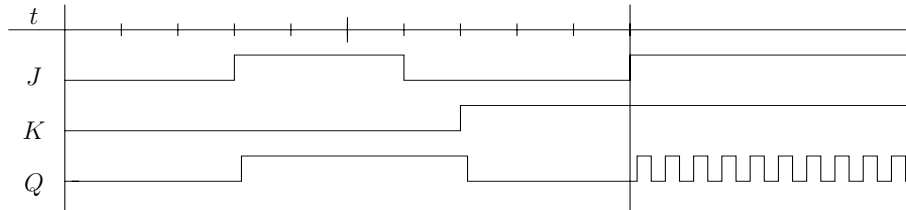


Figure 1.5.12: Timing behaviour of the J-K latch

all input combinations are allowable for the J-K latch, the problem of forbidden inputs reappears in a different form when the memory element is actually realized by a circuit. The input combination  $u_1(t) = u_2(t) = 1$  causes the J-K latch to produce oscillating outputs in continuous time.  $\square$

We have seen in Example 1.5.4 that high reliability can be achieved in spite of unavoidable signal degradation and noise within a network if the network elements

<sup>4</sup>This is why we have denoted the next state by  $x(t + \tau)$  in the above next state tables.



produce signals with only a finite number of steady state/output values (“quantization of signal values”). In order that only these values (representing 0 and 1) determine the behaviour of the network and that the transitional signal values have no effect, time must be discretized as well (“quantization of time”). This is performed by synchronizing the functioning of the network elements. A periodic signal (*clock*) is distributed throughout the circuit in order to ensure that all memory elements change state and output at approximately the same instant. The clock usually generates a square-wave pulse train. By adding for example the clock signal to the inputs of a J–K latch as in Figure 1.5.13 the output and state of this latch will be updated only if the clock is asserted (takes its upper value). When the clock is low,

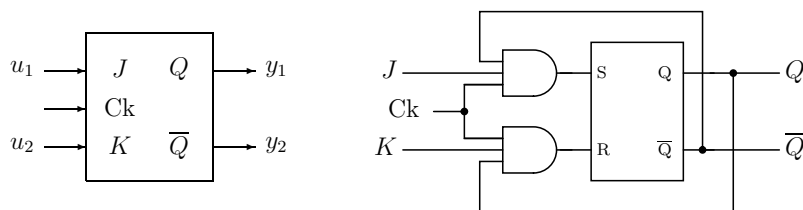


Figure 1.5.13: Clocked J–K Latch: Block diagram and circuit

the steering AND gates are disabled, and the output of the latch remains unaffected by the data inputs J and K. Such a method of synchronization is called *level triggering*, and level triggered storage devices are called *clocked latches*. If the inputs to the network elements do not change during the time when the clock is high and the corresponding steady state and output values are reached within one clock cycle the level triggered network behaves approximately like a digital system. However, level triggering cannot always handle *asynchronous* inputs, i.e. inputs which are changing whilst the clock is high. This may lead to racing problems and unpredictable outputs.

Flip–flops differ from latches in that their outputs change only with respect to the

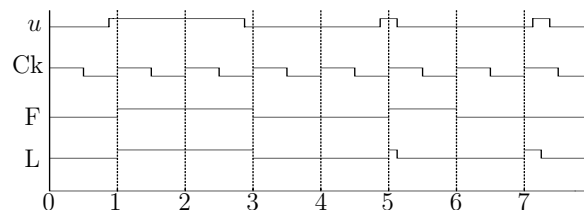


Figure 1.5.14: Time behaviour of a positive edge-triggered flip-flop (F) and a clocked latch (L)

clock whereas clocked latches change output if their inputs change (and the clock is high). *Edge-triggered* flip–flops respond to a rising or falling edge of the clock signal. This is of a very short duration so that racing and oscillating outputs are avoided. A positive (negative) edge-triggered flip–flop samples its inputs on the low–to–high

(resp. high-to-low) transition of the clock and, after a short propagation delay, produces the next state corresponding to the current input and the current state. After this the input may change but the flip-flop will not respond until the next signal from the clock. This is in contrast to the behaviour of a clocked latch as illustrated in Figure 1.5.14. We see that the outputs differ if the input changes when the clock is high. This difference is particularly noticeable between the times 5 and 6 where the clocked latch responds to the decreasing input, but the output of the flip-flop remains at 1.

For reliable operation of flip-flops, the inputs must be “stable” (approximately constant) for a time interval from a *setup time* before, to a *hold time* after the clocking event, see Figure 1.5.15. Proper operation of the circuit requires that the steady

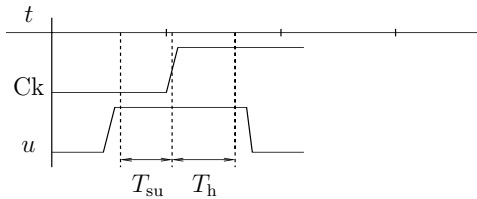


Figure 1.5.15: Setup  $T_{su}$  and hold  $T_h$  times

state value changes only once per clock cycle. In order to guarantee that the correct next state is achieved in spite of varying propagation delays of the input signals the period of the clock should be longer than the worst case propagation delay through the combinational network. If a network is designed in such a way that these constraints are respected, the resulting circuit behaves like a discrete time finite state machine. A careful timing methodology is fundamental for designing reliable sequential networks.

Different types of edge-triggered flip-flops can be created by an interconnection of latches, i.e. by an interconnection of feedback coupled gates. An edge-triggered J-K flip-flop for example, which is one of the most versatile and reliable flip-flops, can be built from 8–10 gates using suitable interconnections and feedback couplings. We do not go into details and refer the interested reader to the literature, see *Notes and References*. Edge triggered flip-flops are represented by block diagrams with a triangle in front of the clock input, see Figure 1.5.16.

In order to illustrate how the above memory elements are used to build a sequential network we conclude this section with an example of a finite state machine design. The main steps in such a design process are listed below.

1. **ABSTRACT REPRESENTATION OF THE MACHINE.** Identify the inputs, outputs, and introduce internal states of the machine which permit an easy description of the desired input–output behaviour. Draw a state diagram, i.e. a graph with vertices representing the states and directed arcs which represent the possible transitions from one state to the next one under the influence of the available inputs. Additionally, a next state table can be established. Describe the outputs associated with given input and state combinations.
2. **STATE MINIMIZATION.** Sometimes the first step results in a description that has a number of redundant states. These states can be eliminated without

affecting the input–output behaviour of the finite state machine. A reduction in the number of states usually reduces the number of logic gates and flip-flops which are needed for the realization of the finite state machine<sup>5</sup>. There are formal procedures and computational algorithms for state minimization, see *Notes and References*.

3. CHOICE OF FLIP–FLOPS for implementing the states.
4. IMPLEMENTATION OF THE FINITE STATE MACHINE. Realize the next state and output mappings by a combinational network connecting inputs, states and outputs.

As an illustration, let us design a synchronous binary counter. Counters are used in many digital systems (e.g. in digital clocks) to count events. They are amongst the simplest possible finite state machines. They typically have only one input (e.g. a square wave signal—the clock) and their outputs are identical with their current state. Their state transition graph consists of a single cycle joining the finitely many binary numbers through which the counter runs successively on each clock pulse.

**Example 1.5.7. (Three bit counter).** We construct a synchronous modulo-8 counter which is driven by a clock. Following the above procedure we begin with an abstract description of the digital system (Step 1). The clock is the only input to the counter. There are three binary output channels corresponding to the three bits  $Q_1, Q_2, Q_3$  which are needed to represent the numbers  $0, \dots, 7$  in the dyadic system. We introduce 8 different states of the counter corresponding to the eight different output combinations and encode the states by the output combination they generate. On each clock pulse the counter advances successively through its 8 states in the following cycle

$$000 \rightarrow 001 \rightarrow 010 \rightarrow 011 \rightarrow 100 \rightarrow 101 \rightarrow 110 \rightarrow 111 \rightarrow 000.$$

In this simple case we may omit the state transition table. The output vector corresponding to the current state  $x(t) = Q_3Q_2Q_1$  is  $(Q_1, Q_2, Q_3)$ . If we want the present output of the counter to be a function of the present state alone, the number of states we have introduced is clearly minimal and we may skip Step 2.

To store the three binary digits  $Q_3, Q_2, Q_1$  three flip-flops are needed. From the state transition graph we see that the digit  $Q_1$  toggles at every clock pulse, the digit  $Q_2$  toggles on every second clock pulse and the digit  $Q_3$  on every fourth clock pulse. This suggests that a toggle flip-flop (T flip-flop) may be most suitable for the implementation of the counter. The T flip-flop has a single input that causes the stored state to remain un-



Figure 1.5.16: Edge-triggered T Flip-Flop: Block diagram and construction from an edge-triggered J-K flip-flop

<sup>5</sup>To realize a machine with  $n$  states at least  $m$  flip-flops are needed where  $2^{m-1} < n \leq 2^m$ .

changed if the input is zero and to be complemented when the input is asserted ( $u = 1$ ). A toggle flip-flop can be constructed from a J-K flip-flop by tying its two inputs together (see Figure 1.5.16). If the input is 0, both J and K are 0 and the flip-flop holds its state; if the input is 1, both J and K are 1 and the flip-flop complements its state, see Table 1.5.11. The state transition of the positive edge triggered T flip-flop takes place on the rising clock edge after the toggle input is set ( $u = 1$ ).

In the final step (Step 4) we express each bit of the next-state<sup>6</sup>  $x(t + 1) = Q_3^+ Q_2^+ Q_1^+$

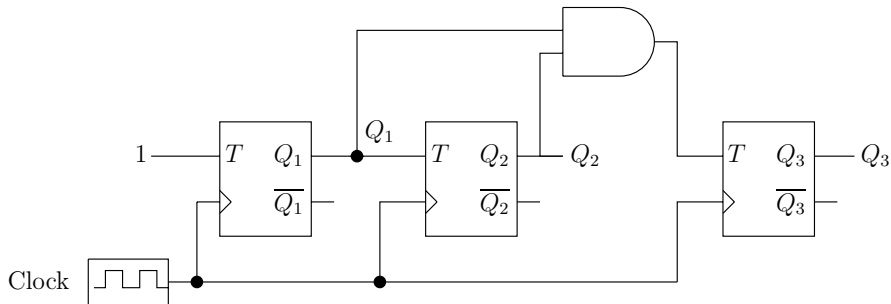


Figure 1.5.17: Three-bit counter circuit

as a combinational logic function of the current state bits and the clock signal. In this simple case the combinational logic for each of the three flip-flops can easily be determined by examining the state transition graph. The flip-flop storing  $Q_1$  toggles on each clock pulse, the flip-flop storing  $Q_2$  toggles at a clock pulse whenever  $Q_1$  is asserted ( $Q_1 = 1$ ) and the flip-flop storing  $Q_3$  toggles at a clock pulse whenever  $Q_2$  and  $Q_3$  are asserted ( $Q_1 = Q_2 = 1$ ). This leads to the circuit shown in Figure 1.5.17.  $\square$

Another simple and important class of sequential networks are shift registers, which play a key role in many finite state machines and communication systems. A simple example of a four bit shift register will be described in Example 2.1.7. For examples of more complicated finite state machines we refer to the literature, see *Notes and References*.

### 1.5.3 Notes and References

The realization of Boolean functions by *combinational* switching networks is based on Boolean algebra and discussed in all textbooks on switching theory and logic design. Important historical references are *Boole* (1849) [66] and *Huntington* (1904) [271].

*Shannon* (1938) [459] was the first to show how Boolean algebra could be applied to digital design. The digital designer wishes to realize a given Boolean function with the minimum number of gates and wires in order to reduce the size, power dissipation and cost of a digital circuit. There are many techniques (and CAD tools) for achieving minimal realizations of a given Boolean function, see *Katz* (1994) [296], *Fabricius* (1992) [154], *Wakerley* (1990) [513] and *Roth* (1985) [437].

The fundamental building blocks of *sequential* switching networks, latches and flip-flops

<sup>6</sup>That is the state in the next clock cycle.

are extensively discussed in most textbooks on digital systems, see the above references. The same holds for registers, memories and counters which can be built from such building blocks, see Example 2.1.7. A more complicated issue is the design of finite state machines for control and decision-making logic in digital systems. The central problem is the realization of a prescribed input-output behaviour by a finite state machine with minimal number of states (*minimal realization*) and efficient state encoding. Good references are Roth (1985) [437], Green (1986) [202], Prosser and Winkel (1987) [422] and Katz (1994) [296]. The latter is especially recommended since it contains many case studies (and two chapters describing how digital design techniques are applied to stored program computers).

The capabilities and behaviours of finite state machines are the subject of Automata Theory which had a strong influence on the early development of mathematical systems theory. Automata Theory studies finite machines as mathematical models of switching and encoding networks in abstraction from specific hardware realizations. For references and further comments on Automata Theory see the *Notes and References* in Section 2.1. In using *discrete* time domains for modelling digital systems one should not overlook the fact that these systems are implemented by electronic circuits in *continuous* time. The resulting timing and synchronization problems are of fundamental importance in the practical design of digital systems. Detailed discussions of these issues can be found in Katz (1994) [296], see also Mead and Conway (1980) [370].

## 1.6 Heat Transfer

Heat transfer is the term used for the exchange of thermal energy. Here we only consider the transfer accomplished by conduction in a solid body and ignore convection and radiation effects. Nowadays it is usual to think of thermal energy or heat as the kinetic energy of the elementary particles of a solid, liquid or gas. The energy levels are a function of temperature with hot regions corresponding to high levels of energy. If a solid is a good electrical conductor there will be a large number of electrons which move freely through the lattice and thermal conduction is a consequence of this motion. In impure metals or in disordered alloys there is also a transfer of energy via lattice vibrations which may be comparable in magnitude to the electronic contribution. For gases the main mechanism is the exchange of kinetic energy from fast moving molecules to slow moving ones caused by collisions amongst themselves. In all cases (including liquids where a variety of mechanisms may be present) there is a flow of energy from regions of high temperature to ones of low temperature.

Given some initial temperature profile within a compact body  $B \subset \mathbb{R}^3$ , our objective is to describe the evolution of the profile in time. We will use the law of conservation of energy to obtain the corresponding differential equations. Let  $V \subset B \subset \mathbb{R}^3$  be an arbitrary fixed, open, connected set with closure in the interior of the body  $B$ . We assume that its boundary  $S$  is orientated and piecewise smooth. If there are no sources of heat in  $V$  the conservation law states that:

The rate of change of the thermal energy in  $V$  with respect to time is equal to the net flow of energy across the surface  $S$  of  $V$ .

We will now translate this law into mathematical formulas and make the statement more precise. Let  $e(x, t)$  be the specific thermal energy (i.e. the energy per unit mass) at position  $x = (x_1, x_2, x_3) \in \mathbb{R}^3$  and time  $t$ . We assume that the density  $\rho = \rho(x)$  of the solid body is independent of time and temperature, then the total thermal energy in  $V$  is

$$\int_V \rho(x) e(x, t) dx$$

where  $dx$  denotes the Lebesgue measure in  $\mathbb{R}^3$ . Assuming that all the functions are continuously differentiable, the time rate of change of the thermal energy in  $V$  is

$$\frac{d}{dt} \int_V \rho(x) e(x, t) dx = \int_V \rho(x) e_t(x, t) dx, \quad e_t(x, t) = \frac{\partial e}{\partial t}(x, t).$$

Let  $\mathbf{q} : B \times \mathbb{R} \rightarrow \mathbb{R}^3$  be the time-dependent vector field which describes the flow of thermal energy in the body  $B$ . The vector  $\mathbf{q}(x, t)$  is called the *heat flux vector* at  $x \in B$  at time  $t$ . Let  $\mathbf{n}(x)$  denote the unit outward normal to the surface  $S$  at the point  $x \in S$ . By the conservation law,

$$\int_V \rho(x) e_t(x, t) dx = - \int_S \langle \mathbf{q}(x, t), \mathbf{n}(x) \rangle dS(x).$$

Applying the divergence theorem to the surface integral over  $S$  we obtain

$$\int_V (\rho(x) e_t(x, t) + \operatorname{div} \mathbf{q}(x, t)) dx = 0.$$

Since the open set  $V$  is arbitrary, we have

$$\rho(x) e_t(x, t) = -\operatorname{div} \mathbf{q}(x, t), \quad x \in \operatorname{int} B, \quad t \in \mathbb{R}. \quad (1)$$

In order to get an equation for the temperature  $\Theta = \Theta(x, t)$ , additional information of an empirical nature is required. For many materials the function  $e$  is linear in the temperature  $\Theta$  over quite large temperature ranges. That is  $e = c\Theta$ , where  $c = c(x)$  is time-invariant and is called the *specific heat* at  $x \in \operatorname{int} B$  (the amount of heat absorbed by the body at the point  $x$  per unit mass per unit rise in temperature). Now the heat energy flows from hot to cold, so the heat flow in any direction  $\mathbf{d} \in \mathbb{R}^3$  will be negative (i.e.  $\langle \mathbf{d}, \mathbf{q} \rangle < 0$ ) if the temperature is rising in that direction (i.e.  $\langle \mathbf{d}, \operatorname{grad} \Theta \rangle > 0$ ), and conversely, if  $\langle \mathbf{d}, \operatorname{grad} \Theta \rangle < 0$  then  $\langle \mathbf{d}, \mathbf{q} \rangle > 0$ . As a consequence there will exist a positive scalar function  $k$  such that  $\mathbf{q} = -k \operatorname{grad} \Theta$ .  $k$  is called the *conductivity* and in general will vary with the medium itself, the position in the medium, the temperature and time. However if the temperature variations are not large, a first approximation which agrees with experiments is to assume that, for a given medium,  $k = k(x)$  is only a function of position. This relationship was postulated by *Fourier* in 1822 and is now known as *Fourier's Law*. With these assumptions (1) becomes

$$c(x)\rho(x) \Theta_t(x, t) = \operatorname{div} (k(x) \operatorname{grad} \Theta(x, t)), \quad (x, t) \in \operatorname{int} B \times (0, \infty). \quad (2)$$

This is the general three-dimensional *heat equation*. If  $k$  does not depend on position, then one obtains the classical form of the heat conduction equation

$$\Theta_t(x, t) = \alpha(x) \Delta \Theta(x, t) \quad (3)$$

where  $\alpha(x) = (c(x)\rho(x))^{-1}k$  is called the *thermometric conductivity* and  $\Delta$  denotes the Laplacian. In order to solve it, an initial temperature distribution must be stipulated and boundary conditions must be specified which describe the way the body interacts with its surroundings. We illustrate this in the following example.

**Example 1.6.1.** Consider a metal rod heated in a furnace. The rod is assumed to be a cylinder of uniform cross sectional radius  $a$  which is heated by jets along its length. The heat from the jets affects the temperature distribution at the surface of the rod which in turn results in changes of the temperature within the rod. Suppose  $(r, \phi, z)$  are cylindrical polar coordinates with the  $z$ -axis along the axis of the cylinder. We will assume that the heat supplied by the jets at point  $z$  along the rod and time  $t$  is the same for all values of  $\phi$  and is given by  $v(z, t)$ . We will also assume the thermometric conductivity  $\alpha$  is constant throughout the rod and the initial temperature distribution at time  $t = 0$  is independent of  $\phi$ . So it is reasonable to seek solutions  $\Theta$  of (3) which have axial symmetry (i.e. are independent of  $\phi$ ), in which case (3) takes the form

$$\Theta_t(r, z, t) = \alpha \Delta \Theta(r, z, t) = \alpha \Theta_{zz}(r, z, t) + \alpha r^{-1} (r \Theta_r(r, z, t))_r. \quad (4)$$

Let

$$\bar{\Theta}(z, t) = (\pi a^2)^{-1} \int_0^{2\pi} \int_0^a \Theta(r, z, t) r \, dr \, d\phi = 2a^{-2} \int_0^a \Theta(r, z, t) r \, dr$$

be the average cross sectional area temperature, then integrating (4) over the cross section at  $z$  we get

$$\bar{\Theta}_t(z, t) = \alpha \bar{\Theta}_{zz}(z, t) + 2\alpha a^{-2} [r \Theta_r(r, z, t)]_0^a = \alpha \bar{\Theta}_{zz}(z, t) + 2\alpha a^{-1} \Theta_r(a, z, t).$$

But  $v(z, t) = -k\Theta_r(a, z, t)$  and hence

$$\bar{\Theta}_t(z, t) = \alpha\bar{\Theta}_{zz}(z, t) + \beta v(z, t),$$

where  $\beta = -2\alpha(ak)^{-1}$ . Let us further assume that the distribution  $b(\cdot)$  of the jets along the rod is fixed, but the magnitude can be varied in time by a control  $u$ , so that  $v(z, t) = b(z)u(t)$ , then

$$\bar{\Theta}_t(z, t) = \alpha\bar{\Theta}_{zz}(z, t) + \beta b(z)u(t) \quad (5)$$

Suppose the temperature at each end of the rod is kept at a constant value  $C$ , and the initial value of  $\bar{\Theta}$  at  $z \in [0, \ell]$  is  $\bar{\Theta}_0(z)$ , so that

$$\bar{\Theta}(0, t) = \bar{\Theta}(\ell, t) = C, \quad t \geq 0, \quad \bar{\Theta}(z, 0) = \bar{\Theta}_0(z), \quad z \in [0, \ell], \quad (6)$$

where  $\ell$  is the length of the rod. Note that if the initial temperature profile is constant with  $\bar{\Theta}_0(z) \equiv C$ , then the corresponding solution of (5) and (6) with  $u(t) = 0, t \geq 0$  is given by the equilibrium solution  $\bar{\Theta}_0(z, t) = C, z \in [0, \ell], t \geq 0$ . For any given solution  $\bar{\Theta}(z, t)$  of the partial differential equation (5) let us denote by  $\theta(z, t)$  the deviation of  $\bar{\Theta}(z, t)$  from the equilibrium solution, i.e.

$$\bar{\Theta}(z, t) = \theta(z, t) + C, \quad \bar{\Theta}_0(z) = \theta_0(z) + C, \quad (z, t) \in [0, \ell] \times \mathbb{R}_+.$$

Then we obtain from (5) and (6) the *one-dimensional controlled heat equation*

$$\begin{aligned} \theta_t(z, t) &= \alpha\theta_{zz}(z, t) + \beta b(z)u(t) \\ \theta(0, t) &= \theta(\ell, t) = 0, \quad \theta(z, 0) = \theta_0(z), \quad (z, t) \in [0, \ell] \times \mathbb{R}_+. \end{aligned} \quad (7)$$

Finally suppose we sense the temperature at a given point  $z_1 \in (0, \ell)$ . In reality the sensor measures a weighted average of the temperature at nearby points. Let us assume that the measurement  $Y(t)$  can be expressed in terms of the average temperature  $\bar{\Theta}(z, t)$  in the form  $Y(t) = \pi a^2 \int_0^\ell c(z)\bar{\Theta}(z, t)dz$ , where the support of the continuous density  $c(\cdot)$  is a small interval around  $z_1$ . If we denote by  $y(t)$  the deviation of  $Y(t)$  from the steady state output  $Y_0(t) = C\pi a^2 \int_0^\ell c(z)dz$  (corresponding to the equilibrium solution  $\bar{\Theta}_0(z, t) = C$ ), then

$$y(t) = \pi a^2 \int_0^\ell c(z)\theta(z, t)dz, \quad t \geq 0. \quad (8)$$

Equations (7) and (8) represent a single input single output system. The state of this system at each time  $t$  is given by the temperature profile  $\theta(\cdot, t)$  which is an infinite dimensional object varying in a function space. Such systems are called *infinite dimensional*.

In applications the above model may be used to determine control laws which drive an initial temperature distribution to some desired final distribution in a given time interval (a controllability problem), see Subsection 2.2.4. Another possible application is to use the model to obtain an estimate of the whole temperature profile  $\theta(\cdot, t)$  from the knowledge of the input and output functions  $u(\cdot), y(\cdot)$  on a given time interval  $[0, T], T > 0$  (an observability problem).  $\square$

### 1.6.1 Notes and References

*J. B. Fourier's* treatise on heat, "Théorie Analytique de la Chaleur", was published in 1822 and an English translation can be found in [171]. There are, of course, whole sections of libraries devoted to heat transfer. One book on the subject is *Ozisik* (1993) [401]. A similar statement is true for books on partial differential equations. We quote *Sobolev* (1964) [469] because some of the material in this section was based on it and because of the influence that Sobolev has had on the mathematical development. A standard reference for the control theory of infinite dimensional systems is *Curtain and Zwart* (1995) [116].