Chapter 1

Introduction

Many laws of Nature are expressed through mathematical identities which relate a function to its derivatives. Such identities are called *differential equations*, and this book is about their study.

For example, the mass $x(t)$ of a sample of radioactive material (uranium, say) at each time t obeys a law of the form

$$x'(t) = -cx(t), \tag{1.1}$$

where $x'(t)$ is the derivative of the function x at time t, and c is a positive constant that depends only on the material. This identity means that the rate of radioactive decay at each time is proportional to the amount of material existing at that time, c being the proportionality constant.

Similarly, the position $x(t)$ of a pendulum subject to a constant gravitational field satisfies a relation of the form

$$x''(t) = -\frac{g}{l}\sin x(t), \tag{1.2}$$

where $x''(t)$ is the second derivative of x at time t, and the constants l and g are physical parameters of the pendulum and the gravitational field, respectively.

The study of differential equations is a very classical area of mathematics, with a wealth of theoretical results, and applications that permeate the whole of science and technology. Throughout the book we shall highlight both the conceptual aspects of the theory and its potential for practical applications, trying to keep the prerequisites to a minimum.

We shall deal almost exclusively with the case when the independent variable t is scalar, that is, its values are real numbers. In that context one sometimes talks of *ordinary* differential equations. The case when t is

vectorial, $t = (t_1, \dots, t_m)$, is the subject of the theory of *partial* differential equations, to which we shall only make brief allusions. On the other hand, the function x may be vectorial: we shall consider $x(t)$ with values in any Euclidean space $\mathbb{R}^d$ and, in the final chapters, even in differentiable manifolds.

Our starting point in this chapter will be to define broadly and precisely what we mean by a differential equation. Next, we shall discuss the objectives of this theory, and we shall begin introducing some of its fundamental ideas. For that, we shall resort to several simple yet meaningful examples, including (1.1) and (1.2).

1.1. Differential equations and their solutions

An (ordinary) *differential equation* is an expression of the form

$$x^{(k)} = F\big(t, x, x^{(1)}, \dots, x^{(k-1)}\big), \tag{1.3}$$

where $F : \mathcal{U} \to \mathbb{R}^d$ is a continuous function defined on an open set $\mathcal{U} \subset \mathbb{R}^{1+kd}$. The variable t takes values in $\mathbb{R}$ while $x, x^{(1)}, \dots, x^{(k-1)}$ and $x^{(k)}$ take values in $\mathbb{R}^d$. The integers $k \geqslant 1$ and $d \geqslant 1$ are respectively called the *order* and the *dimension* of the differential equation. Most of the times we shall use x' and x'' instead of $x^{(1)}$ and $x^{(2)}$, respectively.

By definition, a *solution* of equation (1.3) is a C^k function $\gamma : I \to \mathbb{R}^d$ satisfying the following conditions:

(1) I is an open interval;

(2) the vector $v(t) = \Big(t, \gamma(t), \frac{d\gamma}{dt}(t), \dots, \frac{d^{k-1}\gamma}{dt^{k-1}}(t)\Big)$ is in $\mathcal{U}$ for every $t \in I$;

(3) $\frac{d^k\gamma}{dt^k}(t) = F\big(v(t)\big)$ for every $t \in I$.

Generally speaking, *the objective of the theory of differential equations is to find the solutions of a given equation.* We shall see soon that this formulation is too simplistic and needs adjustment. Before this, we shall try to explain the importance of this theory. We shall see that differential equations arise naturally as mathematical models of various phenomena, either natural or artificial, in such a way that the evolution of such phenomena may be understood by studying the behavior of the solutions of the corresponding equations.

Example 1.1 (Radioactive decay)**.** Radioactive isotopes, such as cesium-137 (used in radiotherapy) and uranium-235 (used in nuclear bombs), have unstable nuclei that slowly decay, giving rise to stable isotopes and emitting radiation in the process. This phenomenon is governed by the following law: *the rate of radioactive decay at a given instant* (*or the amount of isotope that*

goes through transmutation per unit of time) *is proportional to the amount of isotope present in the sample at that instant.*

Let us define

$$x(t) = \text{ mass of radioactive isotope present in the sample at time } t.$$

Thus, the rate of radioactive decay is simply the derivative of x with respect to t. The law of radioactive decay translates to the equation

$$x' = -cx, \tag{1.4}$$

where c is a positive constant depending on the isotope in question. The right hand side carries a negative sign because while the mass x is positive, its derivative must be negative: the mass of the radioactive isotope decreases with time due to transmutation into other isotopes.

Observe that (1.4) is a differential equation of order $k = 1$ and dimension $d = 1$. Also observe that in this case the function

$$F(t, x) = -cx$$

is independent of t. In general, whenever $F(t, x, x^{(1)}, \ldots, x^{(k-1)})$ is independent of t, we say that the differential equation given in (1.3) is *autonomous*.

It is easy to verify that every function $\gamma : \mathbb{R} \to \mathbb{R}$ of the form $\gamma(t) = ae^{-ct}$, with $a \in \mathbb{R}$, is a solution of (1.4). It is also true, as we shall see later, that every solution of (1.4) has to be of this form, but this is a bit more difficult to prove. This family of functions has the following interesting property: there exists $T > 0$ (which depends only on c) such that

$$\gamma(t + T) = \gamma(t)/2 \text{ for every } t.$$

To see this, take $T = \log 2/c$. In other words, after every T units of time the mass of the radioactive isotope decreases by half. For this reason T is called the *half-life*[1] of the radioactive isotope.

We shall now describe an example from another area of physics: classical mechanics.

Example 1.2 (Hooke's law)**.** Consider a system as depicted in Figure 1.1: an elastic spring supported by a rigid base and carrying a point particle of mass m at the other end; the spring can be deformed (compressed or elongated) only along its axis. Hooke's law says that the *tension exerted by the spring on the particle is proportional to the deformation with respect to the equilibrium state of the spring.*

[1]The half-life of uranium-235 is approximately 704 million years, while that of caesium-137 is about 30 years. There are other radioactive isotopes that are much more unstable, such as lithium-12 which has a half-life of only 10^{-8} seconds.

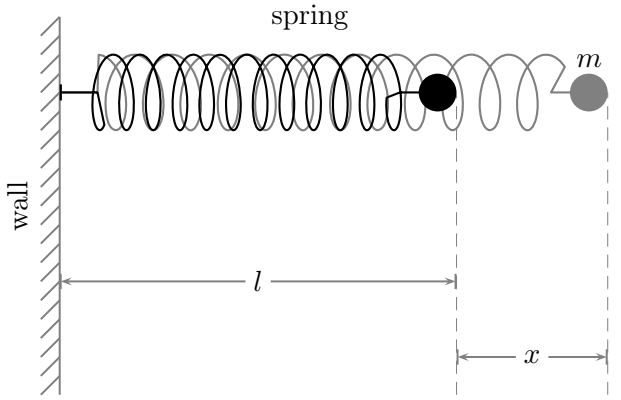

Figure 1.1. Hooke's law: on the left the spring is depicted in its equilibrium position; on the right it experiences a restoring force proportional to the deformation x.

To express this law mathematically, we shall define (see Figure 1.1)

$x(t) =$ deformation of the spring at time t relative to equilibrium.

Hooke's law says that the tension of the spring is given by

$$f = -cx,$$

where c is a positive constant called the *coefficient of elasticity*, which depends on the geometry of the spring and the material used in it. The negative sign corresponds to the fact that the force acts in the direction opposite to that of the deformation: if the spring is elongated, the force tries to shrink it, and if the spring is compressed, the force tries to expand it.

Note that Newton's second law says that the force f is the product of the mass m and the acceleration of the point particle, that is,

$$f = mx'',$$

where x'' is the second derivative of the deformation x. Combining the last two equalities we get

$$x'' = -\frac{c}{m}x, \tag{1.5}$$

which is a differential equation of order 2 and dimension 1, called the *harmonic oscillator*. It is easy to guess some solutions to this equation, for example:

$$\gamma_1(t) = \sin\left(\sqrt{\frac{c}{m}}t\right) \text{ and } \gamma_2(t) = \cos\left(\sqrt{\frac{c}{m}}t\right),\ t \in \mathbb{R}.$$

It is also easy to verify that equation (1.5) has the following properties:

- the sum of any two solutions is also a solution;
- any product of a solution with a real number is also a solution.

In other words, the set of solutions is a vector space. A differential equation having the aforementioned properties is said to be *linear*. Thus, given any $a, b \in \mathbb{R}$, the function

$$\gamma : \mathbb{R} \to \mathbb{R}, \quad \gamma(t) = a \sin\left(\sqrt{\frac{c}{m}}t\right) + b \cos\left(\sqrt{\frac{c}{m}}t\right) \tag{1.6}$$

is a solution of (1.5). In fact, it can be proved that any solution has this form, using results proved later in this book. An interesting consequence is that the motion of the spring is always periodic, with period $T = 2\pi\sqrt{m/c}$.

1.2. Qualitative theory of differential equations

Though our next example also comes from classical mechanics, the corresponding differential equation exhibits very different behavior: finding solutions becomes much more difficult owing to the lack of linearity.

Example 1.3 (Harmonic pendulum)**.** Let us consider the mechanical system depicted in Figure 1.2: a point particle of mass m is suspended from a rigid support by a massless inextensible string of length l, which can only rotate about its point of support. The system is subject to constant gravitational pull.

The equation governing the kinematics of this system can be deduced from Newton's second law as follows. First, let us consider:

$$x(t) = \text{ angle that the string makes with the vertical at time } t.$$

Observe the displacement of the point particle relative to the (vertical) position of equilibrium is given by lx. Hence, Newton's second law gives

$$f = m(lx)'' = mlx'',$$

where f is the force acting on the point mass. Which force is this?

In fact, there is not one but two forces in action. The first one is the weight P, which results from the gravitational pull. It is proportional to the mass m:

$$P = mg,$$

where g is a physical constant denoting the intensity of the gravitational field. The other force, which is called *tension*, is the resistance that is generated in the string due to interactions between its particles, which prevents deformation of its length. Let P_r and P_t be the radial and tangential components of the weight respectively, as described in Figure 1.2. A simple argument using similarity of triangles gives:

$$P_r = P \cos x \text{ and } P_t = -P \sin x.$$

The negative sign in the second equality indicates that P_t points in the opposite direction to the displacement with respect to the vertical line.

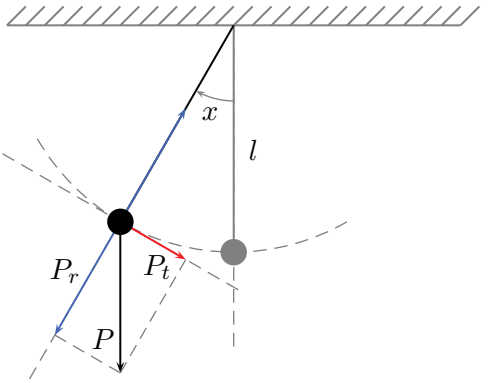

Figure 1.2. Two forces act on a harmonic pendulum: the weight of the particle and the tension of the string; the tension cancels the radial component of the weight.

The tension cancels the component of the weight along the string, thus preventing any changes in its length. This means that the sum of the forces acting on the particle is nothing but P_t. Combining these observations, we get:

$$-mg\sin x = -P\sin x = P_t = mlx''.$$

Hence, the motion of the harmonic pendulum can be modeled by the following second-order differential equation of dimension 1:

$$x'' = -\frac{g}{l}\sin x. \tag{1.7}$$

It is easy to guess a few special solutions to this equation. For example,

$$\gamma_n : \mathbb{R} \to \mathbb{R}, \quad \gamma_n(t) = n\pi \text{ for every } t \in \mathbb{R},$$

is a solution of (1.7) for any $n \in \mathbb{Z}$. Note that these are all constant functions. We say that the values $n\pi$ are *stationary points* of the differential equation.

It is a lot more difficult to exhibit nonstationary solutions, partly because most of them cannot be written using the functions we hear about in calculus or analysis courses: most solutions of (1.7) are not *elementary functions*, meaning that they cannot be defined from polynomial functions, the exponential, trigonometric functions, and their inverses, through a finite number of arithmetic operations ($+$, $-$, $\times$, $\div$) and compositions of functions.

This situation is typical: the solutions of a differential equation tend to be much more complicated than the very function F in the statement of the equation. In some cases, it is possible to express the solutions in terms of F and some elementary functions, through finitely many arithmetic operations, compositions and *integrations*. Then we say that the equation is *integrable*

by quadratures. However, that is also rare: in that sense, most differential equations cannot be solved explicitly.

Though these facts were well-known already in the eighteenth century, the first comprehensive proposal towards a solution came from the great French mathematician Henri Poincaré, at the end of the nineteenth century. In a series of works in the 1880s, he developed a new approach which came to be called the *qualitative theory of differential equations*.

It is insightful to read the original description given by Poincaré himself [2] at the beginning of his paper **[325]**:

> A complete theory of functions defined by differential equations would be very useful in a large number of problems in pure mathematics or mechanics. Unfortunately, it is evident that in a great majority of the cases it is not possible to integrate these equations by means of well-known functions, for example, using functions defined by quadratures. Thus, if we wanted to restrict ourselves to the cases where it is possible to work only using the tools of definite or indefinite integration, the scope of our research would be very restricted and the immense majority of the questions that arise in applications would continue to be unsolvable.
>
> Therefore, it is necessary to study the functions defined by differential equations as they are, without trying to reduce them to simpler functions, as was done for algebraic functions, which we initially tried to reduce to radicals but now study directly, or for the integrals of algebraic differentials, which we tried to express in a finite number of terms for a long time.
>
> Thus, it is of great interest to research the properties of differential equations. We have already taken a first step in this direction by studying the proposed function *in the neighborhood of a point in the plane*. Now it is necessary to go beyond this to study this function *in the whole extension of the plane*. In this research, evidently, our starting point will be the knowledge we already have about the function *in a certain region of the plane*.
>
> The complete study of a function has two parts:
>
> (1) the qualitative part (say), or, the geometric study of the curve defined by the function;
> (2) the quantitative part, or, numerical calculation of the values of the function.
>
> Thus, for example, to study an algebraic equation we start by finding, using Sturm's theorem, the number of real zeros: this is the qualitative part. Later, we calculate the numerical values of these zeros, which constitutes the quantitative study of the equation. Similarly, to study an algebraic curve, we start by *constructing* this curve, that is, by finding the closed branches,

the infinite branches etc. After this qualitative study, we can exactly determine a certain number of its points.

Naturally, the study of any function should start from the qualitative aspect and that is why the first and foremost problem is:

To construct the curves defined by differential equations.

This qualitative study, once complete, will be of great importance for the numerical calculation of the function as, though it is possible to express the function locally as a series, there is no general method to patch together two different series expressions arising in two different regions.

Furthermore, this qualitative theory will be of the greatest importance in itself. Indeed, many very important questions in analysis or mechanics may be reduced to it. Let us consider, for instance, the three-body problem. Is it not natural to ask whether one of the bodies will remain forever in a given region of the sky, or it will, on the contrary, drift away from it indefinitely? Whether the distance between two of the bodies will increase or decrease or remain, on the contrary, between certain bounds? Can you not formulate countless questions of this type, which will all be solved once we know how to construct qualitatively the trajectories of the three bodies? And, considering a larger number of bodies, what is the problem of invariability of the planets' elements, other than a true question in qualitative geometry, given that to prove that the major axis has no secular variations is the same as showing that it oscillates permanently between certain bounds?

This is the vast field of discoveries that opens before the geometers.

The following elementary example illustrates the fact that, even in the few cases when it is possible to find explicit expressions for the solutions, these may turn now to be so convoluted as to be of little help for understanding the actual behavior of the system.

Example 1.4. Let us consider the autonomous differential equation of order 1 and dimension 1 given by:

$$x' = F(t,x) \text{ for } F(t,x) = x(x-1). \tag{1.8}$$

This is a case where the solutions can be obtained explicitly, using the so-called *method of separation of variables* (see Exercise 1.1). This method can be illustrated in a nonrigorous but insightful way as follows. Using the notation $x' = dx/dt$, equation (1.8) can be rewritten as

$$\frac{dx}{dt} = x(x-1) \text{ or, "equivalently", } \frac{dx}{x(x-1)} = dt.$$

Next, "integrating" both sides of the last equality we get:

$$\int \frac{dx}{x(x-1)} = \int dt. \tag{1.9}$$

Also, $\int dt = t + \text{constant}$. On the other hand,

$$\begin{aligned}\int \frac{dx}{x(x-1)} &= \int \left(\frac{1}{x-1} - \frac{1}{x}\right) dx = \log|x-1| - \log|x| + \text{constant} \\ &= \log\left|\frac{x-1}{x}\right| + \text{constant}\,.\end{aligned}$$

Thus, (1.9) implies that

$$\log\left|\frac{x-1}{x}\right| = t + c \text{ or } \left|\frac{x-1}{x}\right| = e^{c+t},$$

where c is an arbitrary real constant. Solving for x, we get the functions

$$x(t) = \begin{cases} 1/(1-e^{c+t}) & \text{for } t \in (-\infty, -c), \\ 1/(1-e^{c+t}) & \text{for } t \in (-c, +\infty), \\ 1/(1+e^{c+t}) & \text{for } t \in \mathbb{R}. \end{cases}$$

Plugging these expressions into (1.8), we can verify that they do satisfy the differential equation. Figure 1.3 shows the graphs of these functions for different values of c. However, there are other solutions (namely, the constant functions equal to 0 and 1) that cannot be found using this method.

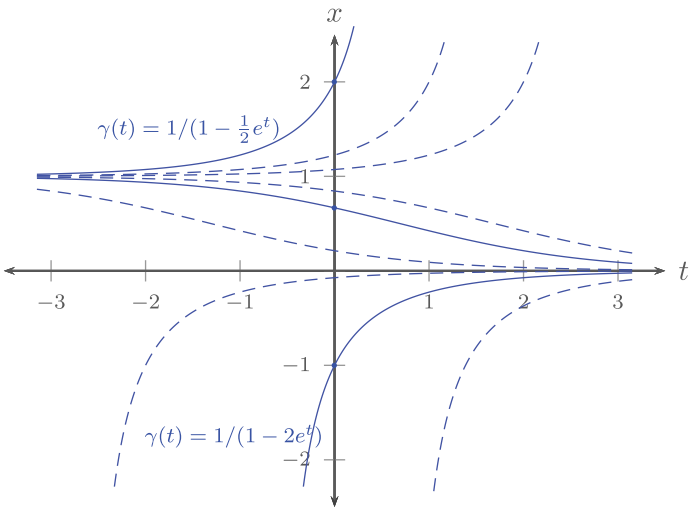

Figure 1.3. Solutions of (1.8) for different values of the parameter c.

However, it is also possible to obtain a very good qualitative description of the behavior of the solutions of (1.9) without any calculation. We start with the following observation:

$$F(t,0) = 0 \text{ for every } t \in \mathbb{R} \quad \text{and} \quad F(t,1) = 0 \text{ for every } t \in \mathbb{R}.$$

This means that the constant functions equal to 0 and 1 satisfy equation (1.8). Next, note that if $x \in (0,1)$, then $F(t,x) < 0$ for every $t \in \mathbb{R}$, and $F(t,x) > 0$ for every $t \in \mathbb{R}$ if $x > 1$ or $x < 0$.

The theory that will be developed in Chapters 2 and 3 allows us to show that any solution $\gamma : I \to \mathbb{R}$ such that $\gamma(t_0) \in (0,1)$ for some t_0 always remains in the interval $(0,1)$, that is, it satisfies $\gamma(t) \in (0,1)$ for every $t \in I$. Thus, the earlier remark shows that any solution is decreasing under these conditions. Analogously, any solution $\gamma(t)$ such that $\gamma(t_0) \notin [0,1]$ for some t_0 is increasing. See Figure 1.4.

This example is quite simple, and one should expect that a qualitative theory of equations that is more complicated would need more sophisticated tools. Which tools did Poincaré have, and which ones are available to us today? To answer this is one of the goals of this book. But it is interesting to mention right away a very motivating example which goes back to Poincaré himself: his famous recurrence theorem.

Consider any C^1 function:

$$F : \mathbb{R}^d \to \mathbb{R}^d, \quad (x_1, \dots, x_d) \mapsto (F_1(x_1, \dots, x_d), \dots, F_d(x_1, \dots, x_d)).$$

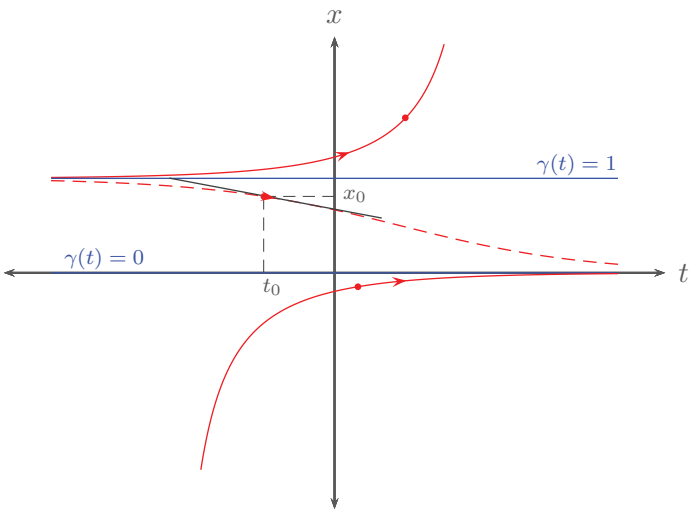

Figure 1.4. Qualitative behavior of solutions of (1.8).

The function $\operatorname{div} F : \mathbb{R}^d \to \mathbb{R}$ defined by

$$\operatorname{div} F = \partial_{x_1} F_1 + \cdots + \partial_{x_d} F_d$$

is called the *divergence* of F. We say that a solution[3] $\gamma : I \to \mathbb{R}^d$ of the first-order autonomous differential equation

$$x' = F(x) \tag{1.10}$$

is *recurrent* if there are $a \in I$ and a sequence $(t_n)_n$ in I converging to the boundary of I (for example, when $I = \mathbb{R}$ this means that $(t_n)_n \to \infty$), such that

$$\big(\gamma(t_n)\big)_n \to \gamma(a).$$

On the other hand, we say that the solution *goes to infinity* if

$$(\|\gamma(t_n)\|)_n \to \infty \text{ for every sequence } (t_n)_n \text{ converging to the boundary of } I.$$

The *Poincaré recurrence theorem* (Theorem 5.21) states that *if the divergence of F is identically zero, then almost every solution $\gamma(t)$ of* (1.10) *is either recurrent or goes to infinity.* By "almost every" we mean that there is a subset of $\mathbb{R}^d$ of full volume such that the conclusion holds for every solution curve that passes through this subset. Thus, the theorem says that almost every solution of (1.10) either keeps returning to the same region or goes to infinity.

One of the reasons this theorem is important is that the condition $\operatorname{div} F \equiv 0$ is very common in differential equations found in classical mechanics and science in general. This fact is illustrated in the next example.

Example 1.5. Initially, remember that equation (1.7) of the harmonic pendulum has order 2. Thus, the recurrence theorem cannot be used directly. However, it is possible to reduce this equation into a system of first-order equations by introducing a new dependent variable y defined by $y = x'$. Then (1.7) can be rewritten as

$$\begin{cases} x' = y \\ y' = -(g/l)\sin x \end{cases} \tag{1.11}$$

or, in vector notation,

$$(x, y)' = F(x, y) \text{ with } F(x, y) = (y, -(g/l)\sin x). \tag{1.12}$$

Observe that we have transformed an equation of order 2 and dimension 1 into an equation of order 1 and dimension 2. This trick is very useful as it

[3] We consider only maximal solutions here, that is, solutions that are not restrictions of other solutions defined in strictly larger intervals. The existence and properties of such solutions will be discussed in Chapter 3.

allows one to transform any equation of order $k \geqslant 1$ to an equation of order 1, at the cost of increasing the dimension. Note that

$$\operatorname{div} F = \partial_x y - \frac{g}{l}\,\partial_y \sin x$$

is identically zero. Thus, the recurrence theorem is applicable in this case.

In fact, we all have practical experience with two types of motions exhibited by a pendulum which correspond to the two types of solutions given by the theorem:

(i) Small oscillations about a stable equilibrium (that is, about the vertical position with the string pointing directly downwards) repeat periodically. This is the principle behind almost all analog clocks; in particular, these movements are recurrent.

(ii) If the pendulum has a sufficiently large velocity, it reaches the unstable equilibrium point (when the string points directly upwards) with nonzero velocity and then continues to rotate in the same direction. In this type of motion, the angle $x(t)$ goes to infinity as $t \to \infty$.

1.3. Numerical analysis of differential equations

As we have seen earlier, for Poincaré the qualitative analysis of a differential equation was not only interesting in itself (as many important questions are qualitative in nature) but also served as a prelude and guide for a quantitative analysis, which is generally numerical. In fact, the subsequent invention and development of computers increased the effectiveness of numerical analysis as an instrument for understanding differential equations much beyond what Poincaré may have predicted. The next two examples are dedicated to a demonstration of the power of numerical methods.

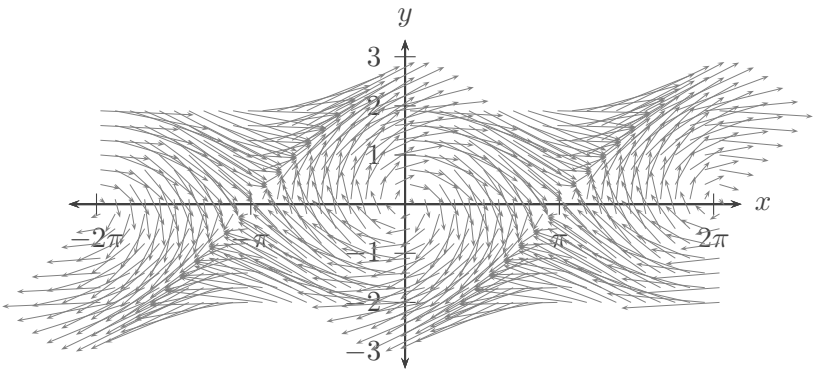

Figure 1.5. Graphical representation of the vector field $F(x, y) = (y, -\sin x)$.

Example 1.6. Figure 1.5 corresponds to equation (1.12) for $g/l = 1$, that is,

$$(x, y)' = F(x, y) \text{ with } F(x, y) = (y, -\sin x). \tag{1.13}$$

The figure shows the direction of the vector field $F(x, y)$ for various points $(x, y) \in [-2\pi,\ 2\pi] \times [-2,\ 2]$. As the solutions $\gamma(t) = (x(t), y(t))$ of (1.13) are curves tangent to the vector field at every point (this is what the equation means!), this representation gives a good idea of the qualitative behavior of the solutions.

Figure 1.5 suggests, for example, that:

(i) the solutions with values close to the origin are periodic curves, that is, curves such that there is a $T > 0$ satisfying $\gamma(t + T) = \gamma(t)$ for every $t \in \mathbb{R}$; in particular, such solutions are recurrent;

(ii) the solutions with large y, either positive or negative, are open curves; $y(t)$ remains bounded but $x(t)$ goes to infinity as $t \to \infty$.

As we shall see, it is possible to prove rigorously that these observations are in fact correct, that is, the solutions of the differential equation indeed behave in this way. This also agrees with our experience with the behavior of the pendulum mentioned before.

Example 1.7. The simplest numerical method to calculate the solutions of an equation is called the *Euler method*. We shall use the following example to illustrate how it works:

$$x' = \cos^2 x. \tag{1.14}$$

The idea is to substitute the derivative x' by a ratio of increments $\Delta x/\Delta t$, where Δt is a small positive increment of the variable t and Δx is the resultant increment of the dependent variable x. Thus, in the place of (1.14), we consider

$$\Delta x = \Delta t \cos^2 x.$$

To calculate the solution, we need to choose a value for Δt: in this case, we shall take $\Delta t = 0.01$. We also need to choose the initial time t_0 in the calculation and the corresponding value x_0 of the solution desired. Thus we obtain successive approximations $x_n \approx x(t_n)$, using the recurrence relations

$$t_n = t_{n-1} + \Delta t \text{ and } x_n = x_{n-1} + \Delta t \cos^2 x_{n-1}. \tag{1.15}$$

Table 1.1 summarizes this calculation for $t_0 = 0$ and $x_0 = \pi/4$.

The columns A, D, and G list successive values of t_n and the columns B, E, and H contain the corresponding values of x_n, calculated from (1.15). In this example, the explicit solution of the equation can be obtained using

Table 1.1. Numerical solution of $x' = \cos^2 x$.

A t_n	B x_n	C $x(t_n)$	D t_n	E x_n	F $x(t_n)$	G t_n	H x_n	I $x(t_n)$
0.00	0.78540	0.78540	0.30	0.91565	0.91510	0.60	1.01301	1.01220
0.01	0.79040	0.79037	0.31	0.91937	0.91880	0.61	1.01581	1.01499
0.02	0.79535	0.79530	0.32	0.92304	0.92246	0.62	1.01859	1.01776
0.03	0.80025	0.80018	0.33	0.92668	0.92609	0.63	1.02134	1.02051
0.04	0.80510	0.80500	0.34	0.93029	0.92969	0.64	1.02407	1.02323
0.05	0.80990	0.80978	0.35	0.93386	0.93325	0.65	1.02677	1.02593
0.06	0.81466	0.81452	0.36	0.93740	0.93677	0.66	1.02945	1.02861
0.07	0.81937	0.81920	0.37	0.94090	0.94027	0.67	1.03211	1.03126
0.08	0.82403	0.82384	0.38	0.94437	0.94373	0.68	1.03474	1.03389
0.09	0.82864	0.82843	0.39	0.94781	0.94715	0.69	1.03735	1.03649
0.10	0.83321	0.83298	0.40	0.95121	0.95055	0.70	1.03993	1.03907
0.11	0.83773	0.83748	0.41	0.95458	0.95391	0.71	1.04249	1.04163
0.12	0.84221	0.84194	0.42	0.95793	0.95724	0.72	1.04504	1.04417
0.13	0.84664	0.84636	0.43	0.96123	0.96054	0.73	1.04755	1.04668
0.14	0.85103	0.85073	0.44	0.96451	0.96381	0.74	1.05005	1.04918
0.15	0.85538	0.85505	0.45	0.96776	0.96705	0.75	1.05253	1.05165
0.16	0.85968	0.85934	0.46	0.97098	0.97026	0.76	1.05498	1.05410
0.17	0.86394	0.86358	0.47	0.97416	0.97343	0.77	1.05741	1.05653
0.18	0.86816	0.86778	0.48	0.97732	0.97658	0.78	1.05983	1.05894
0.19	0.87233	0.87194	0.49	0.98045	0.97970	0.79	1.06222	1.06133
0.20	0.87647	0.87606	0.50	0.98354	0.98279	0.80	1.06459	1.06370
0.21	0.88056	0.88014	0.51	0.98661	0.98586	0.81	1.06694	1.06605
0.22	0.88462	0.88417	0.52	0.98966	0.98889	0.82	1.06927	1.06838
0.23	0.88863	0.88817	0.53	0.99267	0.99190	0.83	1.07158	1.07068
0.24	0.89261	0.89213	0.54	0.99566	0.99488	0.84	1.07387	1.07297
0.25	0.89654	0.89606	0.55	0.99862	0.99783	0.85	1.07615	1.07524
0.26	0.90044	0.89994	0.56	1.00155	1.00076	0.86	1.07840	1.07750
0.27	0.90430	0.90378	0.57	1.00445	1.00366	0.87	1.08063	1.07973
0.28	0.90812	0.90759	0.58	1.00733	1.00653	0.88	1.08285	1.08194
0.29	0.91191	0.91137	0.59	1.01018	1.00938	0.89	1.08505	1.08414

the method of separation of variables:

$$x(t) = \arctan(t + 1).$$

Its values for t_n are displayed in the columns C, F, and I of Table 1.1. This allows us to have an idea of the precision of our numerical calculation by comparing the "approximate" values x_n with the "exact" values $x(t_n)$ obtained by substituting the values of t_n in the analytical expression of the solution. Observe that the difference between x_n and $x(t_n)$ tends to increase for large n, as the errors from individual steps start to accumulate. However, it remains reasonable even after 90 iterations. Also see Exercise 1.22.

1.4. Experiment: population dynamics

Let us consider an animal or plant species thriving in a given ecological environment. As a first approximation, we may assume that the population of the species increases at a rate which is proportional to the size of the population itself. In other words, the number x of individuals of the species[4] satisfies an equation of the form $x' = cx$, where c is a positive constant.

[4]Of course x in reality is a discrete variable, taking only integral values. But for the sake of our analytical method, we shall treat it as a continuous variable.

In practice, the available resources (water, nutrients, oxygen, sunlight, etc.) are limited, so the environment can sustain only a maximum number X of individuals. Thus, it is more realistic to consider an equation of the form

$$x' = cx\left(1 - \frac{x}{X}\right). \tag{1.16}$$

This is known as the *logistic equation*, and despite being rudimentary, it has many applications in ecology and other areas of science.

Next, suppose that the environment supports *two* species that interact with each other, competing for available resources. Let x_1 and x_2 be the respective numbers of individuals. It is reasonable to suppose that their interaction is proportional to the product x_1x_2, which is an approximation for the probability of two individuals belonging to different species coming into contact with each other. The equation describing this situation is called the *Lotka–Volterra equation*:

$$\begin{cases} x_1' = c_1x_1\left(1 - a_{11}x_1 - a_{12}x_2\right) \\ x_2' = c_2x_2\left(1 - a_{21}x_1 - a_{22}x_2\right), \end{cases} \tag{1.17}$$

where $a_{11} = 1/X_1$ and $a_{22} = 1/X_2$ are related to the maximal numbers X_1 and X_2 of individuals of each species that can be sustained by the resources in the absence of the other species, and a_{12} and a_{21} regulate the intensity of the effect of interspecies interaction on each of the species. This equation can be easily generalized for an arbitrary number $d \geqslant 1$ of species:

$$\begin{cases} x_1' = c_1x_1\left(1 - \sum_{j=1}^{d} a_{1j}x_j\right) \\ \quad \cdots \\ x_d' = c_dx_d\left(1 - \sum_{j=1}^{d} a_{dj}x_j\right), \end{cases} \tag{1.18}$$

where $(c_i)_i$ is the *vector of factors* and $(a_{ij})_{i,j}$ is the *interaction matrix*. We shall only discuss the case $d = 2$.

In a competitive situation, as considered before, all the coefficients of (1.17) are positive. Just by changing the coefficients and their signs it is possible to obtain mathematical models for many other problems, not necessarily related to ecology. For example, in a *predator–prey system*, in which one of the species (say the second) eats the other, the factor c_2 should be taken to be negative because, in absence of prey, predators die instead of increasing in numbers. Now, if we consider the resources to be infinite, $(X_1 = X_2 = \infty)$ then equation (1.17) is reduced to

$$\begin{cases} x_1' = c_1x_1\left(1 - a_{12}x_2\right) \\ x_2' = c_2x_2\left(1 - a_{21}x_1\right) \end{cases} \tag{1.19}$$

with $c_2 < 0 < c_1$ and $a_{12}, a_{21} > 0$, which is another well-known form of the Lotka–Volterra equation.

The Euler method introduced earlier can be used to investigate the behavior of the solutions of the Lotka–Volterra equation. In this case, as we are dealing with an equation of dimension 2, instead of (1.15) we have

$$t_n = t_{n-1} + h \text{ and } \begin{cases} x_{1,n} = x_{1,n-1} + hF_1(x_{1,n-1}, x_{2,n-1}) \\ x_{2,n} = x_{2,n-1} + hF_2(x_{1,n-1}, x_{2,n-1}), \end{cases} \tag{1.20}$$

where $(F_1, F_2)(x_1, x_2) = \big(c_1x_1(1 - a_{11}x_1 - a_{12}x_2), c_2x_2(1 - a_{21}x_1 - a_{22}x_2)\big)$ and $h = \Delta t$.

Objectives:

(1) Write a computer program that executes the Euler method for the Lotka–Volterra equation. Fix $d = 2$.

(2) Take $c_1 = 1$, $c_2 = -1$, $a_{11} = 0$, $a_{12} = 1$, $a_{21} = 1$, and $a_{22} = 0$, and numerically integrate the Lotka–Volterra equation for different values of the initial condition $(x_{1,0}, x_{2,0})$ and the step size h.

(3) Repeat step (2) with $c_1 = 1$, $c_2 = -1$, $a_{11} = 1/2$, $a_{12} = 1$, $a_{21} = 1$, and $a_{22} = -1/2$.

(4) Repeat step (2) with $c_1 = 1$, $c_2 = -1$, $a_{11} = 2$, $a_{12} = 1$, $a_{21} = 1$, and $a_{22} = -1/2$.

(5) Compare the conclusions of steps (2), (3), and (4). Does the qualitative behavior of the solutions change significantly when the coefficients are changed?

(6) Interpret these results in terms of the evolution of the ecological system described in each of the cases given above.

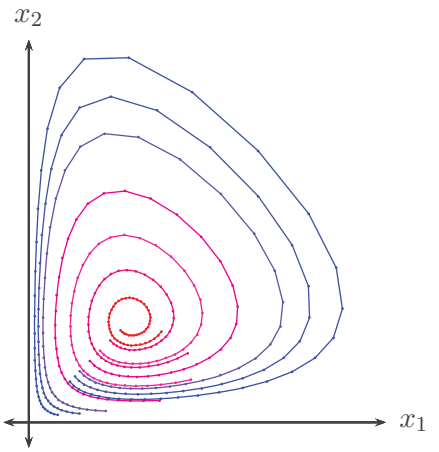

Figure 1.6. Numerical integration of the Lotka–Volterra equation using the Euler method, for $c_1 = 1$, $c_2 = -1$, $a_{11} = 0$, $a_{12} = 1$, $a_{21} = 1$, and $a_{22} = 0$.

Figure 1.6 represents some approximate solutions of the Lotka–Volterra equation obtained using the computational algorithm that we just discussed. Observe that these curves seem to be spirals while in reality the exact solutions are closed curves (consult Exercises 1.25 and 5.6). This difference in behavior is caused by the accumulation of errors committed by our method of numerical integration. This error can be controlled to some extent by reducing the step size h. Even better, as we shall see in Chapter 4, it is possible to use methods that give much better approximations of the exact solution than the Euler method (see Figure 4.12).

1.5. Exercises

The first exercises of this chapter are intended as a revision of the principal elementary methods for analytic solution of differential equations, some of which will be used later. Let us consider differential equations of order 1 and dimension 1, written as

$$a(t,x)x' + b(t,x) = 0, \tag{1.21}$$

where a and b are C^1 functions and $a(t,x) \neq 0$ for every (t,x) in the domain $\mathcal{U}$. In terms of the notation used earlier, we have $F(t,x) = -b(t,x)/a(t,x)$. We finish the list with some introductory exercises for numerical analysis of differential equations.

Exercise 1.1 (Separation of variables)**.** We say that (1.21) is *separable* if there are continuous functions ϕ_1, ϕ_2, ψ_1, and ψ_2 such that $a(t,x) = \phi_1(t)\phi_2(x)$ and $b(t,x) = \psi_1(t)\psi_2(x)$. Show that:

(1) If x_0 satisfies $\psi_2(x_0) = 0$, then the constant curve $\gamma(t) = x_0$ is a solution of (1.21).

(2) If x_0 satisfies $\psi_2(x_0) \neq 0$, then the functions

$$g(x) = \int_{x_0}^{x} \frac{\phi_2(y)}{\psi_2(y)}\, dy \text{ and } h(t) = \int_{t_0}^{t} -\frac{\psi_1(s)}{\phi_1(s)} ds$$

are well defined for (t,x) in a neighborhood of (t_0,x_0). Moreover, g is a diffeomorphism and the curve $\gamma(t) = g^{-1}(h(t))$ is a solution of (1.21) with initial condition $\gamma(t_0) = x_0$.

Exercise 1.2. Use the method of separation of variables to solve each of the following problems:

(1) $x' - tx = 0$ with initial condition $(t_0,x_0) = (1,1)$.

(2) $x' + e^x = 0$ with initial condition $(t_0,x_0) = (1,0)$.

(3) $tx' - x^2 = 0$ with initial condition $(t_0,x_0) = (1,1)$.

(4) $x' - e^{2t-x} = 0$ with initial condition $(t_0,x_0) = (0,0)$.

Exercise 1.3 (Change of variables)**.** Transform the differential equation

$$\sqrt{1+t^2}(x+t^2)x' + \big(1+2t(x+t^2)\sqrt{1+t^2}\big) = 0 \tag{1.22}$$

into a separable equation using a change of the dependent variable, and then find the solution satisfying the initial condition $x(0) = 1$.

Exercise 1.4 (Homogeneous equations)**.** Suppose that $\mathcal{U} \subset \mathbb{R}^2$ does not intersect the axis $\{(0,x) : x \in \mathbb{R}\}$ and is invariant under homotheties. We say that equation (1.21) is *homogeneous* if $F(t,x) = -b(t,x)/a(t,x)$ satisfies $F(ct,cx) = F(t,x)$ for all $(t,x) \in \mathcal{U}$ and all $c \neq 0$.

(1) Argue that there exists a continuous function ϕ such that $F(t,x) = \phi(x/t)$ for every $(t,x) \in \mathcal{U}$.

(2) Show that the change of variables $(t,x) \mapsto (t, u = x/t)$ transforms (1.21) into the equation

$$tu' + (u - \phi(u)) = 0. \tag{1.23}$$

(3) Use the fact that (1.23) is separable to find its solutions.

Exercise 1.5. Solve the following differential equations:

(1) $(t+x)x' + (t-x) = 0$ in $\mathcal{U} = \{(t,x) : t > 0 \text{ and } t + x > 0\}$;

(2) $xtx' + (abt^2 - atx - btx) = 0$ in $\mathcal{U} = \{(t,x) : t > 0 \text{ and } x > 0\}$.

Exercise 1.6 (Exact differential equations)**.** We say that equation (1.21) is *exact* if $\partial_t a(t,x) = \partial_x b(t,x)$ for every $(t,x) \in \mathcal{U}$. Consider the function

$$\phi(t,x) = \int_{x_0}^{x} a(t_0,y)\,dy + \int_{t_0}^{t} b(s,x)\,ds,$$

defined in a neighborhood of every point $(t_0,x_0) \in \mathcal{U}$. Show that there exists a C^1 function $x(t)$ in a neighborhood of t_0 such that $\phi(t,x(t)) = 0$ for every t and $x(t_0) = x_0$. Verify that this function satisfies (1.21).

Remark. We can also take $\phi(t,x) = \int_{x_0}^{x} a(t,y)\,dy + \int_{t_0}^{t} b(s,x_0)\,ds$.

Exercise 1.7. Find all the solutions of $(t+1)x' + (t^3+x) = 0$ in the domain $\mathcal{U} = \{(t,x) : t + 1 > 0\}$.

Exercise 1.8 (Integrating factor)**.** Suppose that the functions $a(t,x)$ and $b(t,x)$ are chosen such that there are continuous functions $f(x)$ and $g(t)$ satisfying

$$\partial_t a(t,x) - \partial_x b(t,x) = f(x)b(t,x) - g(t)a(t,x) \text{ for every } (t,x).$$

Find a C^1 function $u(t,x)$ such that $e^{u(t,x)}$ is an *integrating factor* for the differential equation, that is, such that

$$e^{u(t,x)}a(t,x)x' + e^{u(t,x)}b(t,x) = 0$$

is an exact differential equation.

Exercise 1.9. Consider the linear differential equation $x' + a(t)x = b(t)$, where $a(t)$ and $b(t)$ are continuous functions. Find a C^1 function ψ such that the differential equation

$$\psi(t)x' + \psi(t)a(t)x = \psi(t)b(t)$$

is exact. Also, find the general form of the solutions of the given differential equation.

Exercise 1.10. Show that the solutions of a *homogeneous linear differential equation of order* 1,

$$x' + a(t)x = 0, \tag{1.24}$$

are the functions of the form $x(t) = ce^{-\int a(t)\,dt}$, where $\int a(t)\,dt$ represents a primitive of the function $a(t)$ and c is a real number.

Exercise 1.11. Show that, given any continuous function $b(t)$, the solutions of a *nonhomogeneous linear differential equation of order* 1,

$$x' + a(t)x = b(t), \tag{1.25}$$

are functions of the form $x(t) = c(t)e^{-\int a(t)\,dt}$ and determine the conditions that $c(t)$ needs to satisfy for such a function to be a solution of (1.25).

Remark. This is called the *method of variation of the parameter* (compare with the previous exercise) and applies to other classes of differential equations as well.

Exercise 1.12. Respond and justify:

(1) Can the function $\phi(t) = t^2$ defined for $t \in \mathbb{R}$ be a solution for a homogeneous linear differential equation of order 1? What about a nonhomogeneous one?

(2) Can the functions $\phi(t) = e^t$ and $\psi(t) = e^{-t}$ defined for $t \in \mathbb{R}$ satisfy the same first-order homogeneous linear differential equation? What about a nonhomogeneous one?

If yes, give explicit examples.

Exercise 1.13. Let $a, b : \mathbb{R} \to \mathbb{R}$ be two continuous functions such that $a(t) \geqslant c > 0$ for every $t \in \mathbb{R}$ and

$$\lim_{t\to\infty} b(t) = 0.$$

Show that all the solutions of $x' + a(t)x = b(t)$ converge to zero as $t \to \infty$.

Exercise 1.14. Solve the following differential equations:

(1) $x' - (t^2 + tx) = 0$ in $\mathcal{U} = \mathbb{R}^2$.

(2) $x' - \sqrt{1 + x^2} = 0$ in $\mathcal{U} = \mathbb{R}^2$.

(3) $x' - \sin(x/t) = 0$ in $\mathcal{U} = \{(t,x) : t > 0\}$.

(4) $(x-t)x' + (x+t) = 0$ in $\mathcal{U} = \{(t,x) : x > t\}$.

Exercise 1.15. Find the differentiable curve α in $\mathbb{R}^2$ which passes through the point $(1,1)$ and, for any given point $(x,y) \in \alpha$, if we denote by $P(x,y)$ the point of intersection between the tangent line and the horizontal axis, and by $Q(x,y)$ the point of intersection between the normal line and the vertical axis, the distances to $P(x,y)$ and $Q(x,y)$ from the origin are equal. Draw a sketch of this curve.

Exercise 1.16. Let $p, q, r : \mathbb{R} \to \mathbb{R}$ be continuous functions, with $p > 0$. Show that there are continuous functions $a, b : \mathbb{R} \to \mathbb{R}$ such that a is C^1 and the differential equations $p(t)x'' + q(t)x'(t) + r(t)x = 0$ and $(a(t)x')' + b(t)x = 0$ have exactly the same solutions.

Exercise 1.17 (Series expansion)**.** Consider the second-order linear differential equation

$$x'' + a(t)x' + b(t)x = 0, \text{ with } a(t) = \sum_{n=0}^{\infty} a_n t^n \text{ and } b(t) = \sum_{n=0}^{\infty} b_n t^n. \tag{1.26}$$

(1) Determine conditions that a sequence $(c_n)_n$ must satisfy in order for the corresponding infinite series $\sum_{n=0}^{\infty} c_n t^n$ to be a solution of (1.26).

(2) Show that if $(c_n)_n$ satisfies these conditions and the series $\sum_{n=0}^{\infty} c_n t^n$ is convergent, then the sum $x(t)$ satisfies (1.26).

(3) Solve the *Hermite differential equation* $x'' - 2tx + x = 0$.

Exercise 1.18 (Bernoulli equation)**.** Consider the differential equation

$$x' + \varphi(t)x = \psi(t)x^n, \tag{1.27}$$

where φ and ψ are continuous functions and $n \neq 1$. Verify that the change of variable $y = x^{1-n}/(1-n)$ transforms (1.27) into a linear differential equation. Use this observation to find the solutions of (1.27).

Exercise 1.19. Consider the differential equation

$$\xi(t,x)x' + \eta(t,x)(x - tx') + \zeta(t,x) = 0, \tag{1.28}$$

where ξ, η, and ζ are homogeneous functions: there are m and n such that $\xi(ct,cx) = c^m \xi(t,x)$, $\eta(ct,cx) = c^n \xi(t,x)$, and $\zeta(ct,cx) = c^m \zeta(t,x)$ for any t, x, and c. Define a new variable $y = x/t$, and thereafter treat y as the independent variable and t as the dependent variable. Observe that this procedure transforms (1.28) into a Bernoulli equation. Use this observation to find the solutions of (1.28).

Exercise 1.20. Consider the differential equation

$$(a_0 + a_1 t + a_2 x)(tx' - x) - (b_0 + b_1 t + b_2 x)x' + (c_0 + c_1 t + c_2 x) = 0. \tag{1.29}$$

Find constants α and β such that the change of variables $t = s + \alpha$ and $x = y + \beta$ transforms (1.29) into an equation of the form (1.28). Use this observation to find the solutions of (1.29).

Exercise 1.21. A nuclear reactor transforms plutonium-239 into uranium-238 for industrial use. It is observed that, after fifteen years, 0.0043 percent of the initial quantity of plutonium transforms into uranium. Calculate the half-life of plutonium-239.

Exercise 1.22. Reproduce the calculations in Table 1.1, but for $t \in [0, 3]$. Represent the error $E_n = x_n - x(t_n)$ graphically as a function of n and interpret this graph.

Exercise 1.23. Solve the logistic equation (1.16) by the method of separation of variables and use the expression obtained to understand how the solutions behave. Show that every solution $x(t)$ converges to a limit as $t \to +\infty$ and calculate this value.

Exercise 1.24. The evolution of the population of a city is described by the logistic equation

$$x' = 10^{-1} x(1 - 10^{-6} x),$$

with initial condition $x(0) = 10^4$. Investigate, using the Euler method, the behavior of the solution. Represent the results graphically. Compare with the conclusions from the last exercise.

Exercise 1.25. Show that the function

$$V(x_1, x_2) = -c_2 \log x_1 + c_2 a_{21} x_1 + c_1 \log x_2 - c_1 a_{12} x_2$$

is a *first integral* of the equation (1.19), that is, the map $t \mapsto V(x_1(t), x_2(t))$ is constant for any solution $t \mapsto (x_1(t), x_2(t))$ of the equation. Use this fact to comment on the results obtained in Section 1.4, particularly Figure 1.6.

Remark. We shall return to this theme in Exercise 5.6 to conclude that the solutions to equation (1.19) are closed curves.

Exercise 1.26. Use the Euler method to numerically calculate the solutions of the following differential equations in the interval $[t_0, t_0 + 1]$:

(1) $x' - tx = 0$ with the initial condition $(t_0, x_0) = (1, 1)$.
(2) $x' + e^x = 0$ with the initial condition $(t_0, x_0) = (1, 0)$.
(3) $tx' - x^2 = 0$ with the initial condition $(t_0, x_0) = (1, 1)$.
(4) $x' - e^{2t-x} = 0$ with the initial condition $(t_0, x_0) = (0, 0)$.

Compare with the conclusions of Exercise 1.2.

1.6. Notes

The theory of differential equations goes back to Leibniz and Newton, the discoverers of infinitesimal calculus. The general problem of solving such equations was formulated for the first time in a letter from Newton to Leibniz dated October 26, 1676. But some special cases of the *tangent inverse problem*—given the expression of the tangent to a curve in cartesian coordinates, find the curve itself—had been dealt with before. The first known record of the expression *aequatio differentialis* (*differential equation*) was by Jacob Bernoulli **[35]** in 1692.

In *Methodus fluxionum et Serierum Infinitarum* (Method of fluxions and infinite series) **[298]**, Newton classified three types of differential equations:

$$x' = f(x), \quad x' = f(t,x), \quad \text{and} \quad x\frac{\partial u}{\partial x} + y\frac{\partial u}{\partial y} = u$$

(note that the third one is a partial differential equation). His main approach was the method of power series expansion (Exercise 1.17): the unknown $x(t)$ is written as $\sum_{n=0}^{\infty} a_n t^n$, and the coefficients a_n are determined by replacing this expression in the differential equation. The matter of convergence was not raised. Newton observed that a_0 remains undetermined, and so there exists an infinite family of solutions. However, the meaning of this *integration constant* was not properly understood until the second half of the eighteenth century.

Isaac Newton was born[5] on January 4, 1643, in Woolsthorpe, Lincolnshire, England, and died in London on March 31, 1727. He is the most influential scientist of all times. His masterpiece, *Philosophiae Naturalis Principia Mathematica* (Mathematical principles of natural philosophy) **[296]**, published in 1687, set the foundations of classical mechanics and gravitation theory. Newton shares with Leibniz the credit for the discovery of infinitesimal calculus. In 1669 he became the Lucasian Professor at the University of Cambridge, and four years later he was elected the president of the Royal Society, the prestigious academy of sciences of the United Kingdom. He retained both positions for the rest of his life.

Methodus fluxionum was written around 1671, but was not published until 1736, after the author's death. In the meantime, Leibniz made several important contributions to the theory, mostly motivated by the tangent inverse problem. In 1675, he wrote the relation (see Ince **[179**, p. 529**]**)

$$\int t\,dt = \frac{1}{2}t^2,$$

[5] According to the Gregorian calendar, introduced by the Catholic Church in 1583 but not yet adopted in England at that time. By the old Julian calendar, he was born on December 25, 1642 and died on March 20, 1726. All our dates follow the Gregorian calendar.

which contains not only the solution to the differential equation $x' = t$, but also the first written record of the integral sign, not to mention the differential notation dt, which is also due to Leibniz. In a letter to the Dutch scientist Christiaan Huygens on October 5, 1691, Leibniz [**237**] used implicitly the method of separation of variables to solve a differential equation. This was later formalized by Johann Bernoulli [**36**].

Gottfried Wilhelm Leibniz was born on July 1, 1646, in Leipzig, electorate of Saxony, and died on November 14, 1716, in Hannover, electorate of Brunswick-Lüneburg (both electorates were part of the Holy Roman Empire, in the area now recognized as Germany). He was one of the great thinkers of the Age of Enlightenment, and has a prominent position in the history of mathematics. In 1684, he published *Nova methodus pro maximis et minimis* (A new method for maxima and minima) [**236**], which founded differential calculus. Two years later came *De Geometria recondita et Analysi Indivisibilium atque infinitorum* (On the hidden geometry and the analysis of indivisibles and infinites) [**234**], which contains the first rudiments of integral calculus. He was also a productive inventor, especially interested in the construction of calculating mechanical machines. In 1693, he designed a machine (the *integraph*) capable, in theory, of integrating differential equations.

The study of differential equations attracted the interest of the Bernoulli brothers, Jacob and Johann, who started intense correspondence with Leibniz on that subject. The Bernoullis are the most extraordinary family in the history of science, boasting no less than eight renowned mathematicians.[6]

Jacob Bernoulli, the oldest and perhaps greatest of them all, was born on December 27, 1654, in the city of Basel, Switzerland, in whose university he became a professor of mathematics, and where he died on August 16, 1705. In 1690, he published a solution to the *isochrone problem* [**32**]: to find the curve along which a weighted body falls with uniform vertical velocity. Huygens [**178**] claimed in 1673 that this curve is a *cycloid*, that is, a curve described by a point in a circle when the circle rolls along a straight line without slipping. Bernoulli's solution was the first one based on calculus: he expressed the isochronicity condition as an identity between two differentials,

$$dx\sqrt{b^2x - a^3} = dt\sqrt{a^3},$$

concluding that the corresponding integrals must also be equal (*Ergo & horum Integralia aequantur...*), which provides the solution's expression. This work marks the first use of the word *integral* in the modern sense, and originated a new approach to study curves defined by mechanical properties, such

[6]Namesakes are distinguished by numerals: Jacob I (1654–1705); Johann I (1667–1748), brother of Jacob I; Nicolaus I (1687–1759), nephew of Jacob I and Johann I; Nicolaus II (1695–1726), Daniel (1700–1782) and Johann II (1710–1790), sons of Johann I; and Johann III (1744–1807) and Jacob II (1759–1789), sons of Johann II.

as the logarithmic spiral or the lemniscate: to write the differential equations that express those properties, and solve them.

The mathematical contributions of Jacob Bernoulli go much beyond: his major work, *Ars Conjectandi* (The art of conjecture) **[30]**, introduced many of the fundamental ideas in probability theory and combinatorics, including the first version of the *Law of Large Numbers*; he initiated the calculus of variations, together with his brother Johann; and he was the one who discovered the constant e which, nevertheless, is usually named after Euler or Napier. The Bernoulli equation (Exercise 1.18)

$$x' + \varphi(t)x = \psi(t)x^n$$

was proposed by him **[33]** in 1695. Leibniz **[235]** showed that it may be reduced to a linear equation by a change of variables, and Johann Bernoulli **[38]** explained how to transform it into a separable equation.

Johann, brother and rival of Jacob Bernoulli, was born on August 6, 1667, in Basel, Switzerland, where he died on January 1, 1748. Despite having graduated in medicine, he became a professor of mathematics at the university in Groningen, Netherlands. After the death of his brother, he took his position at the university in Basel. Johann contributed more than anyone else to the development of integral calculus and, in particular, of the methods for solving differential equations. In 1694, he formulated explicitly the method of separation of variables (*seperatio indeterminatarum*) **[36]**. The homogeneous linear equation of order 1 (Exercise 1.10) is a special case of a separable equation, but Johann **[38]** went beyond and solved the general linear equation of order 1, which also involves using the method of variation of the parameter (Exercise 1.11). An important case of a separable equation,

$$t\,dx - x\,dt = 0,$$

remained open, because separation leads to $dt/t = dx/x$, and the expression dt/t had not yet been integrated, although Napier's work on logarithms **[293]** had been published already 80 years prior. However, in that same year of 1694, Leibniz solved the *problem of the quadrature of the hyperbola*—calculate the area of the region below the hyperbola on a given interval—and that helped establish the interpretation of $\int dt/t$ as $\log t$.

The role of Johann Bernoulli in the history of mathematics is further enhanced by his being a teacher and mentor to his sons Nicolaus II and Daniel, as well as to Leonhard Euler, the greatest analyst of the eighteenth century and one of the most important mathematicians of all time. Johann was also the author of several results published in the book **[242]** of the Marquis de L'Hôpital, including the famous *L'Hôpital rule* for the calculation of limits. Although the author had paid for the right to use those results and

had made it very clear that they were not his,[7] Johann resented the little credit he got for the success of the book which, in fairness, was due in no small part to L'Hôpital's remarkable writing skills.

In 1696, Johann Bernoulli [**40**] challenged "the brightest mathematicians in the world" with his famous *brachistochrone* (*curve of fastest descent*) *problem*: to find the curve along which a weighted body moves between two given points in the shortest possible time. The solution is the cycloid, just as in the isochrone problem, but Johann believed his brother would give a wrong answer, which was a strong motivation for his setting the challenge in the first place (their relationship was already much deteriorated by then). However, Jacob did give a correct solution [**34**], which was published jointly with those of Newton [**297**], Leibniz [**233**], L'Hôpital [**243**], and Johann [**37**] himself. Newton was the quickest, having solved the problem in a single night. Although he submitted his solution anonymously, Johann is said to have claimed that he could tell "the lion by the mark of his claw".

Jacob Bernoulli's solution [**34**] to the brachistochrone problem inspired the famous work of Euler [**121**] on maxima and minima in function spaces which, together with the work of Lagrange [**112**, **202**], led to the *Euler–Lagrange equation:*

$$\frac{\partial L}{\partial q}(t, q(t), q'(t)) - \frac{d}{dt}\frac{\partial L}{\partial q'}(t, q(t), q'(t)) = 0. \tag{1.30}$$

In honor of Lagrange, who was only 19 years old at the time, Euler [**120**] named this new area of mathematics the "calculus of variations".

The *isoperimetric problem*, another precursor to the calculus of variations, also contributed to worsening the rivalry between the two Bernoulli brothers. The problem consists in finding among all closed curves with a given length the one that bounds the largest possible area. It was formulated in 1696 by Jacob, who solved it five years later [**31**]. Johann would fix and publish his own solution [**41**] only in 1718.

By the first decades of the eighteenth century, the main basic techniques for solving differential equations were already in place: separation of variables, power series expansion, change of variables, and order reduction. This last one applies, for example, to equations of the form $f(x, x', x'') = 0$, which do not depend explicitly on the t variable. Taking x as the new independent variable, and writing $x' = y(x)$, we get that $x'' = (dy/dx)y$, and equation

[7] "I recognize I owe much to the masters Bernoulli, especially the young one, currently a professor at Groningen. I helped myself to their discoveries at will, as I did with those of master Leibniz. That is why I consent to their claiming anything they want, contenting myself with whatever they wish to leave me."

takes the form

$$f\left(x, y, \frac{dy}{dx}y\right) = 0, \tag{1.31}$$

which is of order 1 on the variable y. The holy grail was still to find general methods for solving "all" differential equations, but that task was now being passed on to the next generation, led by Euler. We shall talk more about that in the Notes to Chapter 2. For more information on the early history of the theory of differential equations, check Ince [**179**, Appendix A] and Archibald [**8**].

The works of Poincaré more directly relevant for this chapter are his memoirs *Sur les courbes définies par les équations différentielles* (on the curves defined by the differential equations) [**325**–**328**], published between 1881 and 1886.

The Lotka–Volterra equation was proposed originally in 1910 by the biophysicist Alfred James Lotka [**259**], as a model for certain chemical reactions. Lotka was born on March 2, 1880, in Lviv (in nowadays Ukraine), which was then part of the Austrian-Hungarian empire, and died in New York on December 5, 1949. In the years 1920–1925, he extended his initial model, applying it to predator–prey situations in ecological systems [**260**].

The same equations were published in 1926, independently, by the Italian mathematician and physicist Vito Volterra [**408**, **409**], whose interest in biology was prompted by his future son-in-law, the Italian marine biologist Umberto D'Ancona. D'Ancona had observed that the percentage of predator fishes captured in the Adriatic sea had grown substantially during the First World War (1914–1918), although the fishing activity was much reduced during the conflict. Volterra used the dynamics of the equations to explain this observation.

Vito Volterra was born in Ancona, Italy, on May 3, 1860, and died in Rome on October 11, 1940. Umberto D'Ancona was born on May 9, 1896, in Fiume (nowadays Rijeka, in Croatia), which was then part of the Austrian-Hungarian empire, and died in Ravenna, Italy, on August 24, 1964. The Lotka–Volterra equation is used as a model in several areas of science, including chemistry and economics. Chapter 12 of the book by Hirsch, Smale [**170**] contains a qualitative description of its dynamical behavior.

Chapter 12

Poincaré–Hopf theorem

The Poincaré–Hopf theorem is part of a family of results that reveal surprising connections between topological (global) objects and analytical (local) objects of very different natures. To this family belong, among others, the Gauss–Bonnet theorem of differential geometry, which we shall also discuss here, the Riemann–Hurwitz and Riemann–Roch theorems of complex analysis, and the famous Atiyah–Singer index theorem. Often those connections can be interpreted through ideas from physics.

The Poincaré–Hopf theorem relates the topology of the ambient manifold M, expressed through its Euler characteristic, to the behavior of the vector fields on M in the vicinity of their stationary points. Indeed, let F be any C^1 vector field whose stationary points are isolated. We shall see in Section 12.1 that it is possible to associate to each stationary point p an integer, called the *index* of F at the point p and represented by $\mathrm{ind}(F, p)$, which can be understood as *the number of turns that the vector field makes when we travel once around the stationary point.*

For example, when $M = \mathbb{R}^2$ we can consider a simple closed curve $c : [0, 1] \to \mathbb{R}^2$ whose image is contained in a small neighborhood of the stationary point, and which turns once around p in the counterclockwise direction. Then, $\mathrm{ind}(F, p)$ is the number of turns that the curve

$$F \circ c : [0, 1] \to \mathbb{R}^2 \setminus \{(0, 0)\}, \quad s \mapsto F(c(s))$$

makes around the origin of $\mathbb{R}^2$. The precise definition for vector fields on surfaces will be given in Section 12.1. Note that the index is a local object: it only depends on the behavior of the vector field in the vicinity of the stationary point.

The notion of *Euler characteristic* goes back to Euler's statement that the numbers V of vertices, E of edges, and F of faces of any polyhedron satisfy the equality

$$V - E + F = 2.$$

In fact, Euler considered only convex polyhedra. In general, given a polyhedron P, not necessarily convex, the number $\chi(P) = V - E + F$ is called the *Euler characteristic* of P. We shall explain in Section 12.2 how this notion can be extended to any compact surface. Section A.8 contains a more general discussion, for polyhedra and manifolds of any dimension.

The Euler characteristic is a global topological invariant, that is, it is preserved by homeomorphisms. In fact, in the class of orientable (or nonorientable) compact surfaces it is a *complete* topological invariant: according to the classification theorem for surfaces, *two orientable* (or *nonorientable*) *compact surfaces are homeomorphic if and only if they have the same Euler characteristic.*

The Poincaré–Hopf theorem states that these two very different concepts are intimately related:

Theorem 12.1 (Poincaré and Hopf)**.** *If M is a compact manifold without boundary and F is a C^1 vector field on M with finitely many stationary points, $p_1, \ldots, p_N$, then,*

$$\sum_{i=1}^{N} \operatorname{ind}(F, p_i) = \chi(M). \tag{12.1}$$

The statement remains valid for manifolds with boundary, under the assumption that the vector field F points "outwards" at every point on the boundary. An example is proposed in Exercise 12.7.

We shall prove Theorem 12.1, in Sections 12.3 and 12.4, only in the case when M is a compact orientable surface. It is interesting to note that the proof will involve another important concept, of a metric nature: the *Gauss curvature* of the surface. In fact, we shall also see at the end of Section 12.4 that the same ideas prove another fundamental result from differential geometry, the Gauss–Bonnet theorem (Theorem 12.14), according to which the total curvature of an orientable compact surface M is equal to $2\pi\chi(M)$. In particular, if the Euler characteristic is positive, as is the case for the sphere $\mathbb{S}^2$, the surface cannot be endowed with a metric of nonpositive curvature at every point.

12.1. Index of a stationary point

Throughout the chapter we shall call a *curve* any continuous and piecewise differentiable function defined on a compact interval. Next we shall define

the concept of the index of an isolated stationary point for vector fields on surfaces. Near the end, in Section 12.1.4, we shall sketch how this notion can be extended to higher dimensions.

12.1.1. Winding number. By definition, the *winding number* about the origin of a closed curve $\gamma : [0,1] \to \mathbb{R}^2 \setminus \{(0,0)\}$ is the real number (see Figure 12.1)

$$\frac{1}{2\pi}\int_\gamma d\theta = \frac{1}{2\pi}\int_0^1 d\theta_{\gamma(t)}(\gamma'(t))\,dt, \tag{12.2}$$

where $d\theta$ is the differential 1-form defined on $\mathbb{R}^2 \setminus \{(0,0)\}$ by

$$d\theta_{(x,y)} = \frac{x}{x^2+y^2}dy - \frac{y}{x^2+y^2}dx. \tag{12.3}$$

It is easy to see that this differential form is closed:

$$\begin{aligned} d(d\theta)_{(x,y)} &= \partial_x\left(\frac{x}{x^2+y^2}\right)dx\wedge dy - \partial_y\left(\frac{y}{x^2+y^2}\right)dy\wedge dx \\ &= \frac{y^2-x^2}{(x^2+y^2)^2}dx\wedge dy - \frac{x^2-y^2}{(x^2+y^2)^2}dy\wedge dx \equiv 0. \end{aligned}$$

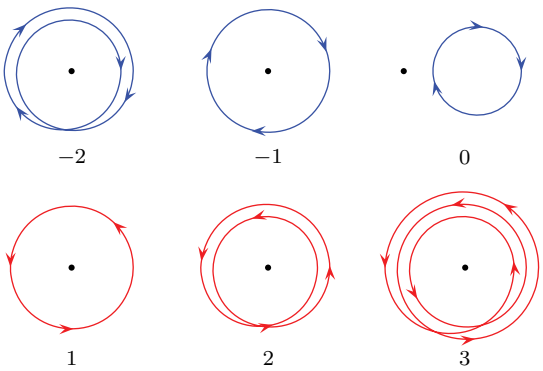

Figure 12.1. Winding number about the origin.

To interpret definition (12.2), it is good to analyze the differential 1-form $d\theta$. Let $X_+ = \{(x,0) : x \geqslant 0\}$ and $X_- = \{(x,0) : x \leqslant 0\}$, and consider the "angle" functions

$$\begin{aligned} &\theta_+ : \mathbb{R}^2 \setminus X_+ \to (0,2\pi) \text{ given by } (\cos\theta_+(z), \sin\theta_+(z)) = \frac{(x,y)}{\|(x,y)\|}, \\ &\theta_- : \mathbb{R}^2 \setminus X_- \to (-\pi,\pi) \text{ given by } (\cos\theta_-(z), \sin\theta_-(z)) = \frac{(x,y)}{\|(x,y)\|}. \end{aligned} \tag{12.4}$$

Note that $\tan\theta_\pm(x,y) = y/x$ if $x \neq 0$ and $\cot\theta_\pm(x,y) = x/y$ if $y \neq 0$. Differentiating these equalities, we obtain that

$$D\theta_\pm(x,y) = \frac{1}{x^2+y^2}(xdy - ydx) = d\theta_{(x,y)} \text{ for all } (x,y) \in \mathbb{R}^2 \setminus X_\pm. \tag{12.5}$$

In other words, the functions θ_+ and θ_- are primitives of the 1-form $d\theta$ on the respective domains. Hence, the integral $\int_\gamma d\theta$ corresponds to the total variation of the "angle" along the curve γ. The proof of the following lemma explains this idea.

Lemma 12.2. *The winding number about the origin of any closed curve $\gamma : [0,1] \to \mathbb{R}^2 \setminus \{(0,0)\}$ is an integer.*

Proof. Recall that $\gamma : [0,1] \to \mathbb{R}^2 \setminus \{(0,0)\}$ is continuous, by assumption. As the domain is compact, it follows that there exists $\delta > 0$ such that $\|\gamma(t)\| \geqslant \delta$ for all $t \in [0,1]$. Moreover, there exists $N \geqslant 1$ such that, for each $i = 0, \ldots, N-1$,

$$\gamma\big([i/N, (i+1)/N]\big) \cap X_+ = \emptyset \quad \text{or} \quad \gamma\big([i/N, (i+1)/N]\big) \cap X_- = \emptyset. \tag{12.6}$$

By (12.5), this implies that the integral of the 1-form $d\theta$ on the curve segment $\gamma \mid [i/N, (i+1)/N]$ is equal to

$$\theta_+\big(\gamma((i+1)/N)\big) - \theta_+\big(\gamma(i/N)\big) \quad \text{or} \quad \theta_-\big(\gamma((i+1)/N)\big) - \theta_-\big(\gamma(i/N)\big),$$

respectively. We shall use these facts to construct a continuous function $\alpha : [0,1] \to \mathbb{R}$ such that

$$\gamma(t) = \|\gamma(t)\|\,(\cos\alpha(t), \sin\alpha(t)) \text{ for all } t \in [0,1]. \tag{12.7}$$

Choose any $\alpha_0 \in \mathbb{R}$ such that $\gamma(0) = \|\gamma(0)\|\,(\cos\alpha_0, \sin\alpha_0)$. Note that α_0 is unique up to addition with an integer multiple of 2π. If $\gamma([0,1/N])$ does not intersect X_+, take $k_0 \in \mathbb{Z}$ such that $\alpha_0 = \theta_+(\gamma(0)) + 2\pi k_0$ and define $\alpha(t) = \theta_+(\gamma(t)) + 2\pi k_0$ for each $t \in [0,1/N]$. If $\gamma([0,1/N])$ does not intersect X_-, take $k_0 \in \mathbb{Z}$ such that $\alpha_0 = \theta_-(\gamma(0)) + 2\pi k_0$ and define $\alpha(t) = \theta_-(\gamma(t)) + 2\pi k_0$ for each $t \in [0,1/N]$. In either case, α is continuous and equality holds in (12.7) for all $t \in [0,1/N]$. Moreover,

$$\int_{\gamma|[0,1/N]} d\theta = \theta_\pm\big(\gamma(1/N)\big) - \theta_\pm\big(\gamma(0)\big) = \alpha(1/N) - \alpha(0).$$

Proceeding by induction, assume that the function α has been constructed in the interval $[0, i/N]$ for some $i = 1, \ldots, N-1$. Repeating the previous construction with $\alpha_i = \alpha(i/N)$ instead of α_0, we obtain a continuous extension of the function α to the interval $[0, (i+1)/N]$, satisfying (12.7) and such that

$$\int_{\gamma|[i,(i+1)/N]} d\theta = \theta_\pm\big(\gamma((i+1)/N)\big) - \theta_\pm\big(\gamma(i/N)\big) = \alpha((i+1)/N) - \alpha(i/N).$$

After N steps of this procedure, we obtain a continuous extension of the function α to all of the interval $[0,1]$, satisfying (12.7) and

$$\int_\gamma d\theta = \sum_{i=0}^{N-1} \int_{\gamma|[i,(i+1)/N]} d\theta = \sum_{i=0}^{N-1} \alpha((i+1)/N) - \alpha(i/N) \tag{12.8}$$
$$= \alpha(1) - \alpha(0).$$

As the curve γ is closed, we have that, $\gamma(0) = \gamma(1)$ and hence

$$(\cos\alpha(0), \sin\alpha(0)) = (\cos\alpha(1), \sin\alpha(1)).$$

This implies that $\alpha(1) = \alpha(0) + 2\pi l$ for some $l \in \mathbb{Z}$. Relation (12.8) gives that this integer l coincides with the winding number of γ about the origin. □

12.1.2. Vector fields on the plane. Let $F : \mathcal{U} \to \mathbb{R}^2$ be a C^1 vector field on an open set $\mathcal{U} \subset \mathbb{R}^2$. Suppose that there exists exactly one point $p \in \mathcal{U}$ such that $F(p) = 0$. Let $c : [0,1] \to \mathcal{U}$ be a simple closed curve about p, that is, such that p is in the inside of c, in the sense of the closed curve theorem. Also assume that c is oriented in the counterclockwise direction, as illustrated in Figure 12.2.

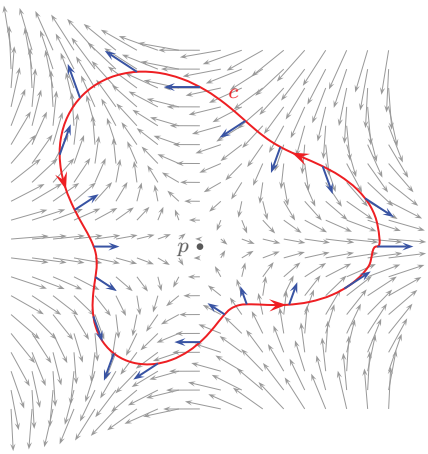

Figure 12.2. Index at an isolated stationary point of a vector field on the plane.

By definition, the *index of the vector field F at the point p* is the winding number of the curve $F \circ c$ about the origin. In other words,

$$\mathrm{ind}(F,p) = \frac{1}{2\pi} \int_{F\circ c} d\theta \tag{12.9}$$

where $d\theta$ is the differential 1-form defined in (12.3). Figure 12.3 describes some examples.

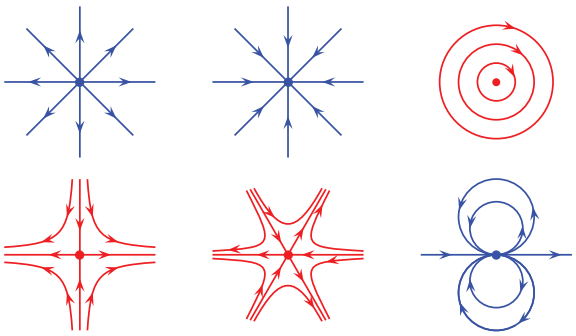

Figure 12.3. Stationary points of a vector field on the plane with indices equal to 1 (source), 1 (sink), and -1 (center) on the first row, and -1 (saddle), -2 (*generalized saddle*), and 2 (*dipole*) in the second.

Lemma 12.3. *The definition of the index* $\operatorname{ind}(F,p)$ *is independent of the choice of the simple closed curve* c *about* p.

Proof. For each $r>0$, consider the parameterization $e_r:[0,1]\to\mathbb{R}^2$ of the circle of radius $r>0$ and center p given by $e_r(t)=p+(r\cos t, r\sin t)$. Given any simple closed curve $c:[0,1]\to\mathcal{U}\setminus\{p\}$, we have that the images of c and e_r are disjoint as long as r is sufficiently small. Consider the domain D bounded by the images of the two curves (see Figure 12.4). By Stokes' theorem and Exercise 12.18,

$$\int_{F\circ c} d\theta - \int_{F\circ e_r} d\theta = \int_c F^*d\theta - \int_{e_r} F^*d\theta = \int_D d\big(F^*d\theta\big) = 0,$$

since the differential form $d\theta$ is closed and, hence, so is $F^*d\theta$. Finally, given any two simple curves $c_1, c_2 : [0,1]\to\mathbb{R}^2\setminus\{(0,0)\}$, we can choose $r>0$

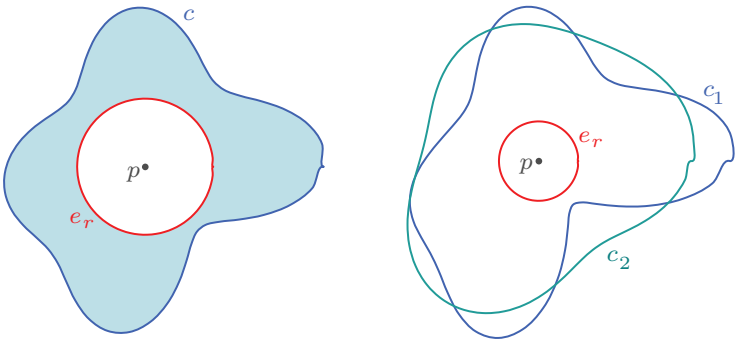

Figure 12.4. The definition of the index does not depend on the choice of the simple closed curve around the stationary point.

sufficiently small for the image of e_r to be disjoint from the images of c_1 and c_2. Then, by the previous argument,

$$\int_{F\circ c_1} d\theta = \int_{F\circ e_r} d\theta = \int_{F\circ c_2} d\theta,$$

and this proves the statement. □

12.1.3. Vector fields on surfaces. Now let F be a C^1 vector field on a compact surface without boundary, M, and let p be an isolated stationary point of F. Initially, we shall assume that M is oriented and equipped with a Riemannian metric. In fact, these two conditions are superfluous, as will be explained at the end.

Consider any local chart $\varphi : U \to \mathbb{R}^2$ compatible with the orientation of M and such that p is the unique stationary point of the vector field in U. Let $c : [0,1] \to U$ be a simple closed curve containing p in its inside and oriented in the counterclockwise direction. What we mean by this is that the simple closed curve $\varphi \circ c$ on the domain $\varphi(U) \subset \mathbb{R}^2$ contains $\varphi(p)$ in its inside and is oriented in the counterclockwise direction. Note that this condition does not depend on the choice of the local chart, only on the orientation chosen on M.

Next, choose a unit vector field $\tilde{e}_1$ on U, and let $\tilde{e}_2$ be the unit vector field orthogonal to $\tilde{e}_1$ and such that the basis $(\tilde{e}_1, \tilde{e}_2)$ has positive orientation at every point of U. We associate to the vector field F the map $\Phi = (\phi_1, \phi_2) : U \to \mathbb{R}^2$ defined by

$$F(q) = \phi_1(q)\tilde{e}_1(q) + \phi_2(q)\tilde{e}_2(q) \text{ for all } q \in U. \tag{12.10}$$

Observe that $\Phi(p) = 0$. By definition, the *index* $\operatorname{ind}(F,p)$ of the vector field F at the point p is the winding number of the curve $\Phi \circ c$ about the origin:

$$\operatorname{ind}(F,p) = \frac{1}{2\pi}\int_{\Phi\circ c} d\theta = \frac{1}{2\pi}\int_c \Phi^* d\theta. \tag{12.11}$$

Proposition 12.4. *The definition of* $\operatorname{ind}(F,p)$ *does not depend on the choices of the simple closed curve c, the unit vector field $\tilde{e}_1$, and the Riemannian metric on M.*

Proof. Consider $\hat{c} = \varphi \circ c$ and $\hat{\Phi} = \Phi \circ \varphi^{-1}$ (check Figure 12.5). It is clear that $\hat{\Phi} \circ \hat{c} = \Phi \circ c$ and, in particular, the corresponding winding numbers about the origin are equal. On the other hand, the same argument as in Lemma 12.3 shows that the winding number

$$\frac{1}{2\pi}\int_{\hat{\Phi}\circ\hat{c}} d\theta = \frac{1}{2\pi}\int_{\hat{c}} \hat{\Phi}^* d\theta$$

of $\hat{\Phi} \circ \hat{c}$ about the origin does not depend on the curve $\hat{c}$. Thus, $\operatorname{ind}(F,p)$ does not depend on the curve c.

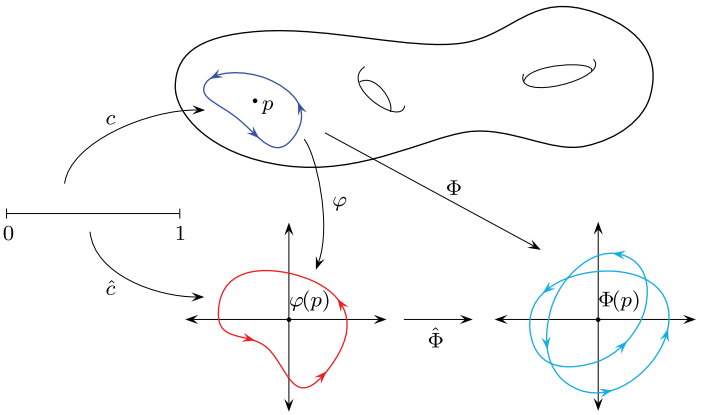

Figure 12.5. The definition of the index does not depend on the choice of the curve c.

Next we show that $\operatorname{ind}(F,p)$ does not depend on the choice of $\tilde{e}_1$. Indeed, consider any other unit vector field $\tilde{f}_1$ defined on U, and let $\tilde{f}_2$ be the unit vector field orthogonal to $\tilde{f}_1$ such that the basis $(\tilde{f}_1,\tilde{f}_2)$ has positive orientation. Take $\tau_0\in[0,2\pi)$ such that

$$\tilde{f}_1(p)=\cos\tau_0\,\tilde{e}_1(p)+\sin\tau_0\,\tilde{e}_2(p).$$

Then there exists a continuous function $\tau:U\to\mathbb{R}$ such that $\tau(p)=\tau_0$ and

$$\begin{aligned}\tilde{f}_1(q)&=\cos\tau(q)\,\tilde{e}_1(q)+\sin\tau(q)\,\tilde{e}_2(q)\ \text{and}\\ \tilde{f}_2(p)&=-\sin\tau(q)\,\tilde{e}_1(q)+\sin\tau(q)\,\tilde{e}_2(q)\end{aligned}$$

for all $q\in U$. Consider the map $\Psi=(\psi_1,\psi_2):U\to\mathbb{R}^2$ defined by

$$F(q)=\psi_1(q)\tilde{f}_1(q)+\psi_2(q)\tilde{f}_2(q)\ \text{for all } q\in U. \tag{12.12}$$

Comparing with (12.10), we see that

$$\Phi=\begin{pmatrix}\cos\tau & -\sin\tau\\ \sin\tau & \cos\tau\end{pmatrix}\Psi. \tag{12.13}$$

Writing

$$\frac{\Psi(c(t))}{\|\Psi(c(t))\|}=\big(\cos\beta(t),\sin\beta(t)\big),$$

relation (12.13) gives that $\|\Phi\|=\|\Psi\|$ and

$$\frac{\Phi(c(t))}{\|\Phi(c(t))\|}=\big(\cos\alpha(t),\sin\alpha(t)\big)\ \text{with}\ \alpha(t)=\beta(t)+\tau(c(t)).$$

Then $\alpha(1)-\alpha(0)=\beta(1)-\beta(0)+\tau(c(1))-\tau(c(0))=\beta(1)-\beta(0)$. In other words, the winding numbers of $\Phi\circ c$ and $\Psi\circ c$ about the origin are equal, as claimed.

It remains to verify that $\mathrm{ind}(F,p)$ is also independent of the Riemannian metric. We shall argue by continuity, as follows. Given any two Riemannian metrics $q \mapsto A_q$ and $q \mapsto B_q$ on M, and any $s \in [0,1]$, the function $q \mapsto A^s_q$ defined by

$$A^s_q(u,v) = (1-s)A_q(u,v) + sB_q(u,v) \text{ for } u,v \in T_qM$$

is a Riemannian metric on M. Indeed, it is clear from the definition that $q \mapsto A^s_q$ is differentiable and that each A^s_q is an inner product on the tangent space T_qM. Given any vector field $\tilde{e}$ on U, with norm 1 for the Riemannian metric A, and any $s \in [0,1]$, consider

$$\tilde{e}^s(q) = \frac{\tilde{e}(q)}{\sqrt{A^s_q(\tilde{e}(q),\tilde{e}(q))}} \text{ for } q \in U$$

and let $\tilde{e}^s_2$ be the vector field on U such that the basis $(\tilde{e}^s_1, \tilde{e}^s_2)$ is positively oriented and orthonormal relative to the Riemannian metric A^s. By construction, these bases vary continuously with the parameter $s \in [0,1]$. Define $\Phi^s = (\phi^s_1, \phi^s_2) : U \to \mathbb{R}^2$ by means of the equality

$$F(q) = \phi^s_1(q)\tilde{e}^s_1(q) + \phi^s_2(q)\tilde{e}^s_2(q) \text{ for } q \in U.$$

Then, Φ^s varies continuously with the parameter s. It follows that the winding number of the curve $\Phi^s \circ c$ about the origin also varies continuously with s. As this number is an integer, this means that the winding number of $\Phi^s \circ c$ about the origin is the same for all $s \in [0,1]$. In particular, considering $s = 0$ and $s = 1$, the indices of the vector field F calculated from the Riemannian metrics $q \mapsto A_q$ and $q \mapsto B_q$ are equal. □

We end the definition of the index on surfaces with two quick comments.

The first one concerns the role of the orientation. If we reverse the orientation of the surface M, then we need to reverse the orientation of the curve c as well, so that it remains counterclockwise. In addition, we need to replace the vector field $\tilde{e}_2$ with $-\tilde{e}_2$, so that basis $(\tilde{e}_1, \tilde{e}_2)$ remains positively oriented. Then, $\Phi = (\phi_1, \phi_2)$ is replaced with $(\phi_1, -\phi_2)$. The combined effect of these two modifications is that the winding number of $\Phi \circ c$ about the origin remains unchanged. Thus, the definition of the index does not depend on the orientation on M. It also follows that the definition makes sense even if M is not orientable: it suffices to choose an orientation in the vicinity of each stationary point, which is always possible, and we have just seen that the index does not depend on this choice.

The second comment has to do with the role of the Riemannian metric. Note that every compact surface (in fact, every manifold) admits some Riemannian metric. The fastest way to verify this fact is by using Whitney's embedding theorem, which states that every compact manifold may be seen

as a submanifold of some Euclidean space $\mathbb{R}^m$. More precisely, for every compact C^r manifold M there exists some C^r embedding $f : M \to \mathbb{R}^m$. Then

$$\langle u, v\rangle_p = Df(p)u \cdot Df(p)v, \text{ for } p \in M, u \in T_pM, \text{ and } v \in T_pM, \tag{12.14}$$

defines a Riemannian metric on M. Thus, we may use any Riemannian metric on M to define the index: as observed previously, the definition does not depend on the choice. A celebrated theorem of Nash (check the Notes in Section 12.8) asserts that, conversely, every Riemannian metric on a compact manifold is of the form (12.14).

12.1.4. Index in higher dimensions. Finally, we shall sketch the definition of the index for stationary points and vector fields on manifolds of any dimension $d \geqslant 2$. This extension uses ideas from Section A.5.

For each $d \geqslant 2$, let $\mathbb{S}^{d-1}$ be the sphere of dimension $d-1$, that is,

$$\mathbb{S}^{d-1} = \{(x_1, \dots, x_d) \in \mathbb{R}^d : x_1^2 + \cdots + x_d^2 = 1\}.$$

The differential form $d\theta$ defined by

$$d\theta_p(v_1, \dots, v_{d-1}) = \det(v_1, \dots, v_{d-1}, p)$$

for $p \in \mathbb{S}^{d-1}$ and $v_1, \dots, v_{d-1} \in T_p\mathbb{S}^{d-1}$ is a volume form on $\mathbb{S}^{d-1}$. It may be shown that if $f : \mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$ is a C^1 map, then there exists an integer $\deg(f)$, called the *degree* of f, such that

$$\int_{\mathbb{S}^{d-1}} f^* d\theta = \deg(f) \int_{\mathbb{S}^{d-1}} d\theta.$$

The map $f \mapsto \deg(f)$ is a group homomorphism:

$$\deg(f_1 \circ f_2) = \deg(f_1)\deg(f_2) \tag{12.15}$$

for any C^1 maps $f_1, f_2 : \mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$. Moreover, if $g : \mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$ is close to f in the C^1 topology, then the two maps have the same degree: $\deg(f) = \deg(g)$.

Let $p \in \mathcal{U}$ be an isolated stationary point of a C^1 vector field $F : \mathcal{U} \to \mathbb{R}^d$ on an open set $\mathcal{U} \subset \mathbb{R}^d$. Let $c : \mathbb{S}^{d-1} \to \mathcal{U}$ be a C^1 embedding whose image is contained in a small neighborhood of p and contains the stationary point on its inside. By definition, the index $\operatorname{ind}(F, p)$ is the degree of the map

$$\mathbb{S}^{d-1} \to \mathbb{S}^{d-1}, \quad x \mapsto \frac{F(c(x))}{\|F(c(x))\|}.$$

It can be shown that the definition does not depend on the choice of the embedding c, and it coincides with the definition of index in Section 12.1.2 in the special case $d = 2$.

Finally, let $p \in M$ be an isolated stationary point of a C^1 vector field F on a manifold M of dimension $d \geqslant 2$. Let $c : \mathbb{S}^{d-1} \to M$ be a C^1 embedding

whose image is contained in a small neighborhood V of p and contains the stationary point in its inside. Consider linearly independent vector fields $\tilde{e}_1, \dots, \tilde{e}_d$ on the neighborhood V, and also the function

$$\Phi : V \mapsto \mathbb{R}^d,\ \Phi(x) = (\phi_1(x), \dots, \phi_d(x)) \text{ given by } F(x) = \sum_{i=1}^{d} \phi_i(x)\tilde{e}_i(x).$$

By definition, the index $\operatorname{ind}(F, p)$ is the degree of the map

$$\mathbb{S}^{d-1} \to \mathbb{S}^{d-1}, \quad x \mapsto \frac{\Phi(c(x))}{\|\Phi(c(x))\|}.$$

The definition does not depend on the choices of the embedding c and the vector fields $\tilde{e}_1, \dots, \tilde{e}_d$, and it coincides with the definition of index in Section 12.1.3 for $d = 2$.

12.2. Euler characteristic

Next, let us recall the notion of Euler characteristic for two-dimensional polyhedra and compact surfaces. Section A.8 contains a broader, though less detailed discussion for polyhedra and manifolds of any dimension.

12.2.1. Polyhedra. The *triangle* defined by noncollinear $v_1, v_2, v_3 \in \mathbb{R}^3$ is the convex hull of the three points, that is, the set

$$T = \{t_1v_1 + t_2v_2 + t_3v_3 : t_1, t_2, t_3 \geqslant 0 \text{ and } t_1 + t_2 + t_3 = 1\}.$$

The points v_1, v_2, v_3 are called *vertices*, the segments connecting the vertices,

$$\{t_iv_i + t_jv_j : t_i, t_j \geqslant 0 \text{ and } t_i + t_j = 1\} \text{ with } (i, j) \in \{(1, 2), (2, 3), (3, 1)\}$$

are called *edges*, and the T itself is called the *face* of the triangle.

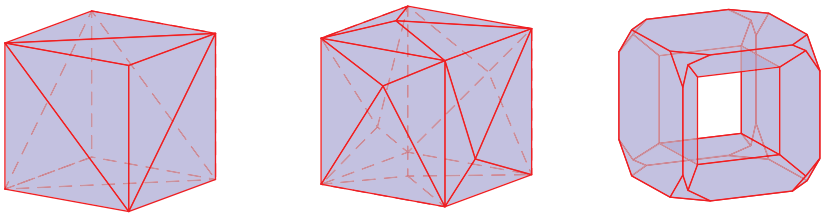

Figure 12.6. Examples of polyhedra: the first two are convex but the last one is not.

A *polyhedron* in the Euclidean space $\mathbb{R}^3$ is any finite collection P of triangles such that any two triangles in P are either disjoint, or have a single edge or a single vertex in common. We shall represent by $|P|$ the union of the triangles in each polyhedron P. Thus $|P|$ is a subset of $\mathbb{R}^3$. See Figure 12.6.

Remark 12.5. In the most usual sense of the word *polyhedron*, faces can be polygons with any number of sides. Here we shall consider only triangular faces, but this does not really constitute a restriction, since every polygon can be decomposed into a finite number of triangles having, at most, one common edge (see Figure 12.7).

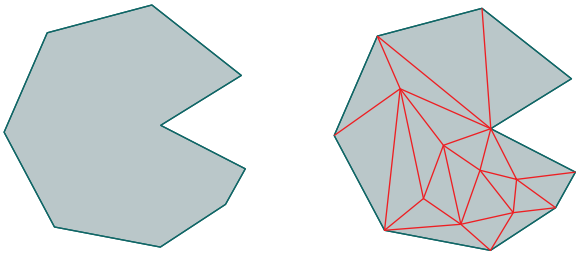

Figure 12.7. Division of a polygon in a finite number of triangles.

Let V be the number of vertices, let E be the number of edges, and let F be the number of faces of a polyhedron P. We define the *Euler characteristic* of P as

$$\chi(P) = V - E + F. \tag{12.16}$$

Example 12.6. In the examples of Figure 12.6, from the left to the right,

$$\chi(P_1) = 8-18+12 = 2,\ \chi(P_2) = 14-36+24 = 2,\ \chi(P_3) = 40-64+24 = 0.$$

Observe that P_1 and P_2 are polyhedra in the same space (a cube), that is $|P_1| = |P_2|$, and they have the same Euler characteristic. This is not a coincidence: a crucial property of the Euler characteristic is that it depends only on the space $|P|$ and not on the polyhedron P itself (Theorem A.31 states an even stronger fact).

Proposition 12.7. *If P_1 and P_2 are polyhedra such that $|P_1| = |P_2|$, then $\chi(P_1) = \chi(P_2)$.*

Let us begin by proving a special case of this proposition. Given polyhedra P_1 and P_2, we say that P_1 is a *subdivision* of P_2 if $|P_1| = |P_2|$ and each face of P_1 is contained in some face of P_2. This situation is illustrated in Figure 12.8.

Lemma 12.8. *If P_1 is a subdivision of P_2, then $\chi(P_1) = \chi(P_2)$.*

Proof. Choose a triangle $T \in P_2$. As T has one face, three edges, and three vertices,

$$\chi(T) = 3 - 3 + 1 = 1. \tag{12.17}$$

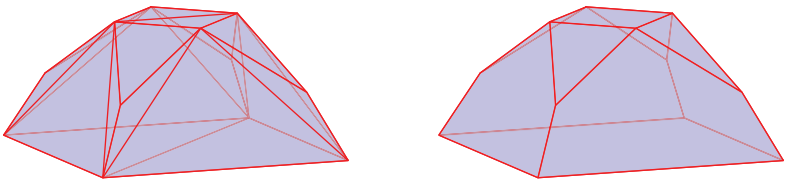

Figure 12.8. Example of subdivision of a polyhedron.

Let $V(T)$, $E(T)$, and $F(T)$ be the number of vertices, edges, and faces of P_1 contained in T. Denote by α_{ij}, $i = 1, \dots, F(T)$ and $j = 1, 2, 3$ the internal angles of the faces of P_1 contained in T. Then,

$$\sum_{i=1}^{F(T)} \sum_{j=1}^{3} \alpha_{ij} = F(T)\pi. \tag{12.18}$$

For each vertex contained in T, the sum of the internal angles of the faces of P_1 contained in T is equal to 2π if the vertex belongs to the interior of T and equal to π if the vertex belongs to the interior of some edge of T. Hence,

$$\sum_{i=1}^{F(T)} \sum_{j=1}^{3} \alpha_{ij} = 2\pi V_i(T) + \pi V_b(T) + \sum_{j=1}^{3} \alpha_j = 2\pi V_i(T) + \pi V_b(T) + \pi,$$

where $V_i(T)$ is the number of vertices of P_1 in the interior of T, $V_b(T)$ is the number of vertices of P_1 in the interior of edges of T, and α_1, α_2, α_3 are the internal angles of the triangle T itself. In other words,

$$\begin{aligned} \sum_{i=1}^{F(T)} \sum_{j=1}^{3} \alpha_{ij} &= 2\pi\big(V_i(T) + V_b(T) + 3\big) - \pi V_b(T) - 5\pi \\ &= 2\pi V(T) - \pi V_b(T) - 5\pi. \end{aligned} \tag{12.19}$$

Each edge of P_1 in the interior of T belongs to the boundary of two faces of P_1 inside T, and each edge of P_1 on the boundary of T belongs to the boundary of a face of P_1 inside T. Thus,

$$3F(T) = 2E_i(T) + E_b(T),$$

where $E_i(T)$ is the number of edges of P_1 in the interior of T, and $E_b(T)$ is the number of edges of P_1 on the boundary of T. Consequently,

$$3\pi F(T) = 2\pi\big(E_i(T) + E_b(T)\big) - E_b(T) = 2\pi E(T) - \pi E_b(T). \tag{12.20}$$

Combining (12.18), (12.19), and (12.20) we obtain

$$2\pi\big(V(T) - E(T) + F(T)\big) = \pi\big(V_b(T) - E_b(T) + 5\big).$$

The number of vertices of P_1 contained in the interior of an edge of T is one unit less than the number of edges of P_1 contained in this edge of T. Hence, $V_b(T) = E_b(T) - 3$. Substituting in the previous relation, it follows that

$$V(T) - E(T) + F(T) = 1.$$

Comparing this equality with (12.17), we see that the Euler characteristic is not affected when each face $T \in P_2$ is substituted with the faces of P_1 contained in it. Repeating this procedure for each of the faces of P_2, we obtain the conclusion of the lemma. □

Lemma 12.9. *If P_1 and P_2 are polyhedra such that $|P_1| = |P_2|$, then there exists a polyhedron Q that is a subdivision for both P_1 and P_2.*

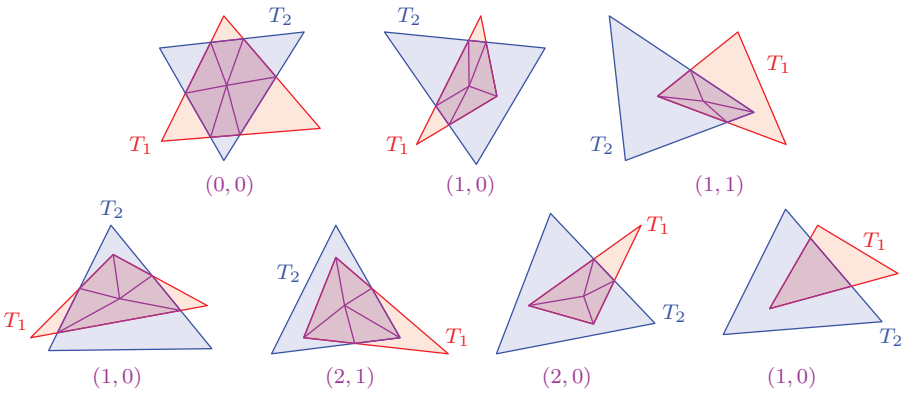

Figure 12.9. Intersections of triangles and their subdivisions in triangles.

Proof. Consider the family $\tilde{Q}$ of the intersections $T_1 \cap T_2$ with $T_1 \in P_1$ and $T_2 \in P_2$. Recalling that T_1 and T_2 are triangles, each element of $\tilde{Q}$ is a convex polygon with 3, 4, 5, or 6 sides. See Figure 12.9. These sides are the segments of edges of one of the polyhedra, P_1 or P_2, determined by points of intersection with edges of the other polyhedron. Subdivide each of these polygons into triangles by choosing a point inside them and connecting it to the vertices by line segments. This way, we get a family Q of triangles with $|Q| = |\tilde{Q}| = |P_1| = |P_2|$.

We claim that Q is a polyhedron. Indeed, by construction, the sides of the triangles in Q are:

(i) the segments in the interior of the polygons of $\tilde{Q}$ that were introduced for the subdivision described above;

(ii) the edge segments of one of the polyhedra, P_1 or P_2, determined by points of intersection with edges of the other polyhedron.

On the one hand, it is clear from the construction that if two triangles in Q are contained in the same polygon of $\tilde{Q}$, then they intersect at a single point, or along one of the edges of type (i). On the other hand, if they are inside distinct polygons of $\tilde{Q}$, then their intersection is either empty, or it consists of a single vertex or a single edge of type (ii). In either case, condition (b) in the definition of polyhedron is satisfied. □

Proof of Proposition 12.7. It follows immediately from Lemmas 12.8 and 12.9: given two polyhedra P_1 and P_2 with $|P_1| = |P_2|$, consider any common subdivision Q. Then $\chi(P_1) = \chi(Q) = \chi(P_2)$. □

A polyhedron P is said to be *convex* if the space $|P|$ is the boundary of a compact convex domain $\Omega(P)$ of $\mathbb{R}^3$. See Figure 12.6. The foundational result for the notion of Euler characteristic is the following theorem:

Theorem 12.10 (Euler)**.** *Every convex polyhedron has Euler characteristic equal to* 2.

Proof. Let P be any convex polyhedron, and let P_0 be a regular tetrahedron. As $\chi(P_0) = 4 - 6 + 4 = 2$, it suffices to show that $\chi(P) = \chi(P_0)$. As the Euler characteristic is not affected by rigid motions or homotheties, we may assume that the origin belongs to the interior of the domain $\Omega(P_0)$ and that $\Omega(P_0)$ is contained in the interior of $\Omega(P)$. Then, by convexity, every ray starting from the origin intersects each of the spaces $|P|$ and $|P_0|$ at a unique point. Hence, the radial projection $\phi : |P_0| \to |P|$ is well defined and is a homeomorphism. In addition, ϕ is piecewise affine in the following sense: given any triangles $T \in P$ and $T_0 \in P_0$, the intersections $T_0 \cap \phi^{-1}(T)$ and $\phi(T_0) \cap T$ are convex polygons (with 3, 4, 5, or 6 sides), and the restriction

$$\phi : T_0 \cap \phi^{-1}(T) \to \phi(T_0) \cap T$$

is the restriction of an affine homeomorphism of $\mathbb{R}^3$. Consider the family of polygons

$$\tilde{Q} = \big\{T_0 \cap \phi^{-1}(T) : T_0 \in P_0 \text{ and } T \in P\big\}.$$

Subdividing each element of $\tilde{Q}$ into triangles, in the same way as in the proof of Lemma 12.9, we obtain a polyhedron Q which is a subdivision of P_0. In particular, $|Q| = |P_0|$. Moreover, as ϕ is affine on each element of $\tilde{Q}$, the image $\phi(Q)$ is a polyhedron and, by construction, is a subdivision of P. Clearly, $\chi(Q) = \chi(\phi(Q))$, as ϕ sends vertices, edges, and faces of Q to vertices, edges, and faces of $\phi(Q)$. Then, Lemma 12.8 gives that $\chi(P_0) = \chi(Q) = \chi(\phi(Q)) = \chi(P)$. □

12.2.2. Surfaces. Now, we shall extend the definition of Euler characteristic to surfaces. A *triangulation* of a surface M is a homeomorphism $\phi : |P| \to M$, where P is a polyhedron. We say that P *triangulates* M. The family

$$\phi(P) = \{\phi(T) : T \in P\}$$

is called a (curvilinear) *polyhedron* in M and its elements are called (curvilinear) *triangles* in M. See Figure 12.10. We call *vertices*, *edges*, and *faces* of the triangulation the images under ϕ of the vertices, edges, and faces of P. Note that any two nondisjoint faces either intersect along a unique edge, or their intersection consists of a unique vertex, as in condition (b) in the definition of polyhedron.

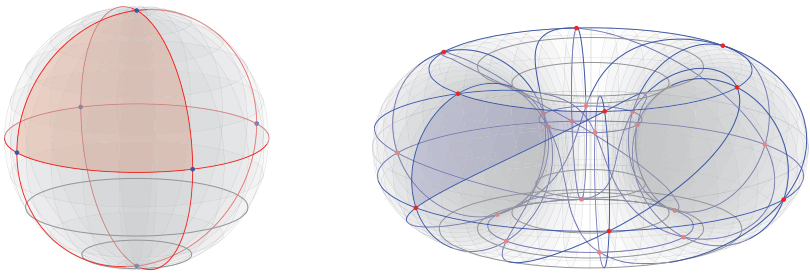

Figure 12.10. Triangulations of the sphere and the torus.

As commented on in Section A.8, in a more general context, every compact surface admits some triangulation. Moreover, if $\phi_1 : |P_1| \to M$ and $\phi_2 : |P_2| \to M$ are triangulations of the same surface, then $\chi(P_1) = \chi(P_2)$. This allows the *Euler characteristic* of a compact surface M to be defined as the Euler characteristic of any polyhedron P that triangulates M. It is clear from the definition that the Euler characteristic is a topological invariant. Indeed, suppose that M and N are homeomorphic compact surfaces. Let $f : M \to N$ be a homeomorphism, and let $\phi : |P| \to M$ be a triangulation of M. Then $f \circ \phi : |P| \to N$ is a triangulation of N and, hence, $\chi(M) = \chi(P) = \chi(N)$.

Remark 12.11. The notions of triangulation and Euler characteristic can be extended naturally to surfaces with boundary, with the extra condition that every connected component of the boundary is formed by vertices and edges. For example, both the cylinder and the Möbius strip admit triangulations with six vertices, twelve edges, and six faces, in which every component of the boundary is formed by three vertices and three edges. See Figure 12.11. Hence, the Euler characteristic of any of these surfaces with boundary is equal to zero.

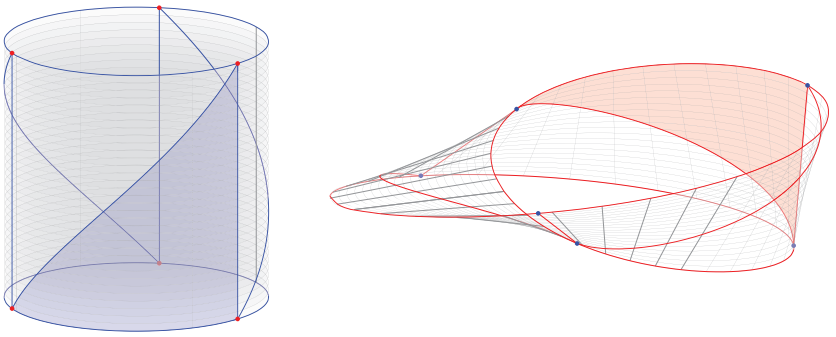

Figure 12.11. Triangulation of surfaces with boundary: cylinder and Möbius strip.

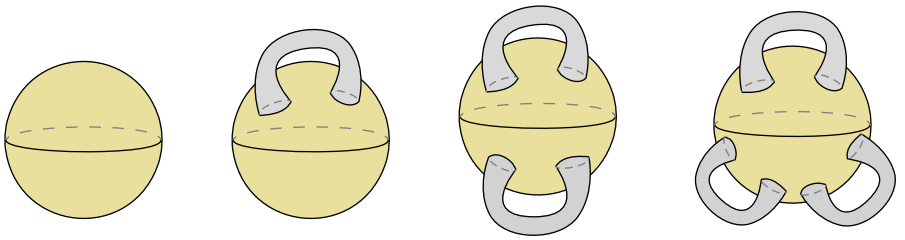

Figure 12.12. Every orientable compact surface may be represented as a "sphere with handles".

Example 12.12. Observing that the regular tetrahedron triangulates the sphere $\mathbb{S}^2$, we obtain immediately that the Euler characteristic of the sphere is equal to $4 - 6 + 4 = 2$. Using this fact, we can calculate the Euler characteristic for any compact orientable surface in the following way. See Figure 12.12.

Choose an integer $g \geqslant 1$. Consider a triangulation of the sphere with a number of triangles much greater than g and such that (at least) $2g$ of those triangles are differentiable discs. Represent by $\tilde{M}_g$ the surface with boundary obtained by removing from the sphere the interiors of these discs. Note that

$$\chi(\tilde{M}_g) = \chi(\mathbb{S}^2) - 2g = 2 - 2g.$$

Next, consider g copies of the cylinder $\mathbb{S}^1 \times [0, 1]$ with a triangulation with six vertices, twelve edges, and six faces, as shown in Figure 12.11. Denote by M_g the compact surface obtained by gluing these cylinders (called *handles*) with $\tilde{M}_g$, along the components of the boundary, in such a way that the corresponding edges and vertices are identified.

By construction, M_g comes with a triangulation, obtained from the triangulations of $\tilde{M}_g$ and the g handles by identifying the edges and the vertices

in the components of the boundary. As the handles have a total of $2g$ boundary components, each with three vertices and three edges, this identification causes the number of vertices to be reduced by $6g$, and similarly for the number of edges. In this way, we conclude that

$$\chi(M_g) = \chi(\tilde{M}_g) + g(6 - 12 + 6) - 6g + 6g = \chi(\tilde{M}_g) = 2 - 2g.$$

The choice of which boundary components of $\tilde{M}_g$ are identified with the components of the same handle is not relevant: it can be shown that different choices produce surfaces M_g that are homeomorphic. The classification theorem for surfaces states that, given any compact orientable surface without boundary M, there exists exactly one value of $g \geqslant 0$, called the *genus* of the surface, such that M is homeomorphic to M_g (with the convention $M_0 = \mathbb{S}^2$). Thus, $\chi(M) = \chi(M_g) = 2 - 2g$.

Similarly, the surface with boundary $M_{g,r}$ obtained by "extracting" disjoint discs from M_g has Euler characteristic $\chi(M_{g,r}) = 2 - 2g - r$. Check Exercise 12.12.

12.3. Indices and curvature

The crucial ingredient in the proofs of the Poincaré–Hopf theorem and the Gauss–Bonnet theorem for compact orientable surfaces is the relationship between the Gauss curvature and the sum of the indices of the vector fields contained in the following proposition. We shall use the characterization of the Gauss curvature given by Theorem A.33 (the reader is recommended to consult Section A.9 before moving on).

Proposition 12.13. *Let M be an orientable compact surface with a Riemannian metric. Let $K : M \to \mathbb{R}$ be the Gauss curvature, and let ω be the area form associated with g. Then, for any C^1 vector field F on M with a finite number of stationary points, $p_1, \ldots, p_N$, we have*

$$\int_M K\omega = 2\pi \sum_{i=i}^{N} \operatorname{ind}(F, p_i). \tag{12.21}$$

Note that the expression on the left side of (12.21) depends only on the Riemannian metric, and not on the vector field. In contrast, the expression on the right hand side depends only on of the vector field, and not on the Riemannian metric. Thus both numbers are independent of both the vector field and the metric.

Proof of Proposition 12.13. Let $\varphi_i : U_i \to X_i$, $i = 1, \ldots, N$, be local charts compatible with the orientation of the surface M, such that $p_i \in U_i$ for each $i = 1, \ldots, N$ and the domains U_i are pairwise disjoint. For each

$i = 1, \dots, N$, let $c : i \to U_i \setminus \{p_i\}$ be a simple closed curve containing p_i in its inside D_i and oriented in the counterclockwise direction.

Let $\tilde{M} = M \setminus \bigcup_{i=1}^{N} D_i$. Then, $\tilde{M}$ is a compact surface with boundary and F does not have stationary points in $\tilde{M}$. Let $p \mapsto \langle \cdot, \cdot \rangle$ be the Riemannian metric on M. Consider the unit vector field defined on $\tilde{M}$ by

$$e_1(p) = \frac{F(p)}{\|F(p)\|_p}, \text{ where } \|F(p)\|_p = \sqrt{\langle F(p), F(p) \rangle_p}. \tag{12.22}$$

Moreover, let e_2 be the unit vector field orthogonal to e_1 and such that $(e_1(p), e_2(p))$ is a positively oriented basis of T_pM for all $p \in \tilde{M}$. Then (e_1, e_2) is a global frame on $\tilde{M}$, in the sense of Section A.9. Let (ω_1, ω_2) be the dual frame, and let ω_{12} be the connection form, given in Theorem A.33:

$$d\omega_1 = \omega_{12} \wedge \omega_2, \ d\omega_2 = \omega_1 \wedge \omega_{12}, \text{ and } d\omega_{12} = -K\omega. \tag{12.23}$$

Using Stokes' theorem, we obtain that

$$\int_{\tilde{M}} -K\omega = \int_{\tilde{M}} d\omega_{12} = \int_{\partial \tilde{M}} \omega_{12} = -\sum_{i=1}^{N} \int_{c_i} \omega_{12} \tag{12.24}$$

(the second minus sign comes from the fact that c_i is negatively oriented with respect to the domain $\tilde{M}$).

For each $i = 1, \dots, N$, choose a unit vector field $\tilde{e}_1^i$ on U_i. Moreover, let $\tilde{e}_2^i$ be the unit vector field orthogonal to $\tilde{e}_1^i$ such that $(\tilde{e}_1^i, \tilde{e}_2^i)$ is a positively oriented basis on all of the domain U_i. Let $\tilde{\omega}_{12}^i$ be the connection form of the frame $(\tilde{e}_1^i, \tilde{e}_2^i)$: by definition, it satisfies

$$d\tilde{\omega}_1^i = \tilde{\omega}_{12}^i \wedge \tilde{\omega}_2^i, \ d\tilde{\omega}_2^i = \tilde{\omega}_1^i \wedge \tilde{\omega}_{12}^i, \text{ and } d\tilde{\omega}_{12}^i = -K\omega. \tag{12.25}$$

Using Stokes' theorem once more,

$$\int_{D_i} -K\omega = \int_{D_i} d\tilde{\omega}_{12}^i = \int_{c_i} \tilde{\omega}_{12}^i \text{ for } i = 1, \dots, N. \tag{12.26}$$

Adding (12.24) and (12.26), we obtain

$$\int_M K\omega = \sum_{i=1}^{N} \int_{c_i} \big(\omega_{12} - \tilde{\omega}_{12}^i\big). \tag{12.27}$$

To calculate the integrals on the right hand side, we need to relate the bases (e_1, e_2) and $(\tilde{e}_1^i, \tilde{e}_2^i)$. Let us consider the map $\Phi^i = (\phi_1^i, \phi_2^i) : U_i \to \mathbb{R}^2$ defined by

$$F(p) = \phi_1^i(p)\tilde{e}_1^i(p) + \phi_2^i(p)\tilde{e}_2^i(p) \text{ for } p \in U_i.$$

By definition (recall (12.10) and (12.11)),

$$\operatorname{ind}(F, p_i) = \frac{1}{2\pi} \int_{\Phi^i \circ c_i} d\theta = \frac{1}{2\pi} \int_{c_i} (\Phi^i)^* d\theta. \tag{12.28}$$

According to definition (12.4),

$$\Phi^i(p) = \|\Phi^i(p)\|_p \,(\cos\theta_\pm(\Phi^i(p)), \sin\theta_\pm(\Phi^i(p))) \text{ whenever } \Phi^i(p) \in \mathbb{R}^2 \setminus X_\pm.$$

Recalling that $e_1(p) = F(p)/\|F(p)\|_p = \Phi^i(p)/\|\Phi^i(p)\|_p$, it follows that

$$e_1(p) = \cos\theta_\pm(\Phi^i(p))\,\tilde{e}_1^i(p) + \sin\theta_\pm(\Phi^i(p))\,\tilde{e}_2^i(p)$$

and this implies that

$$e_2(p) = -\sin\theta_\pm(\Phi^i(p))\,\tilde{e}_1^i(p) + \cos\theta_\pm(\Phi^i(p))\,\tilde{e}_2^i(p).$$

Thus, the dual frames satisfy

$$\begin{aligned} \tilde{\omega}_{1,p}^i &= \cos\theta_\pm(\Phi^i(p))\,\omega_{1,p} - \sin\theta_\pm(\Phi^i(p))\,\omega_{2,p} \text{ and} \\ \tilde{\omega}_{2,p}^i &= \sin\theta_\pm(\Phi^i(p))\,\omega_{1,p} + \cos\theta_\pm(\Phi^i(p))\,\omega_{2,p}. \end{aligned} \tag{12.29}$$

Using Exercise 12.18 and the equality (12.5), we see that

$$D(\theta_\pm \circ \Phi^i) = (\Phi^i)^* D\theta_\pm = (\Phi^i)^* d\theta.$$

Then, differentiating (12.29) and using (12.23), we obtain

$$\begin{aligned} d\tilde{\omega}_1^i &= \cos(\theta_\pm \circ \Phi^i)\, d\omega_1 - \sin(\theta_\pm \circ \Phi^i)\, d\omega_2 \\ &\qquad - \sin(\theta_\pm \circ \Phi^i)\,(\Phi^i)^* d\theta \wedge \omega_1 - \cos(\theta_\pm \circ \Phi^i)\,(\Phi^i)^* d\theta \wedge \omega_2 \\ &= \cos(\theta_\pm \circ \Phi^i)\,\omega_{12} \wedge \omega_2 - \sin(\theta_\pm \circ \Phi^i)\,\omega_1 \wedge \omega_{12} \\ &\qquad - (\Phi^i)^* d\theta \wedge (\sin(\theta_\pm \circ \Phi^i)\omega_1 + \cos(\theta_\pm \circ \Phi^i)\omega_2) \\ &= \big(\omega_{12} - (\Phi^i)^* d\theta\big) \wedge \big(\sin(\theta_\pm \circ \Phi^i)\,\omega_1 + \cos(\theta_\pm \circ \Phi^i)\,\omega_2\big) \\ &= \big(\omega_{12} - (\Phi^i)^* d\theta\big) \wedge \tilde{\omega}_2^i \end{aligned}$$

and, analogously,

$$d\tilde{\omega}_2^i = \tilde{\omega}_1^i \wedge \big(\omega_{12} - (\Phi^i)^* d\theta\big).$$

Comparing these equalities with (12.25) and using the uniqueness of the connection form, we conclude that $\tilde{\omega}_{12}^i = \omega_{12} - (\Phi^i)^* d\theta$. Consequently,

$$\int_{c_i} (\omega_{12} - \tilde{\omega}_{12}^i) = \int_{c_i} (\Phi^i)^* d\theta = 2\pi \operatorname{ind}(F, p_i). \tag{12.30}$$

Substituting in (12.27), we obtain the claim (12.21) in the statement. □

12.4. Proof of the theorem

We are now ready to prove Theorem 12.1 for the case when M is an orientable compact surface. Proposition 12.13 implies that the sum of the indices of the stationary points does not depend on the vector field. Hence, it is enough to construct *some* vector field for which the sum of the indices satisfies the equality in the statement of the theorem.

We begin with any triangulation of the surface M. On each face of this triangulation we introduce four new vertices, as described in Figure 12.13:

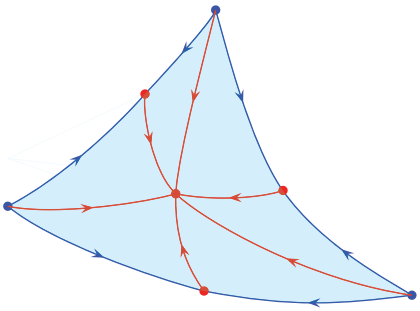

Figure 12.13. Refining a triangulation of the surface.

one vertex in the interior of the face and one more on each of the three boundary edges. Each of these edges is divided into two, oriented to point towards the new vertex. We introduce six new edges, connecting the three original vertices and the three new ones on the boundary to the vertex in the interior and oriented to point towards the latter. Thus, the original face is divided into six new ones.

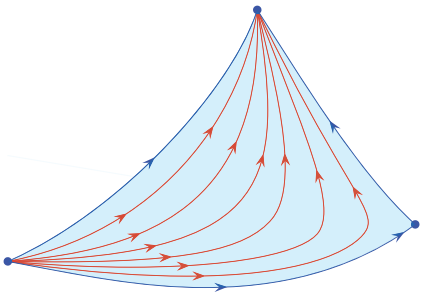

Figure 12.14. Vector field on the triangulation.

Now we can define a vector field F as described in Figure 12.14: on each of the new faces there are three stationary points, that coincide with the vertices of the face, and the edges are regular trajectories, with directions compatible with the orientations previously assigned to the edges. Hence, F has one stationary point at each vertex, on each edge and in each face of the original triangulation. The stationary points at the vertices are repellers (index $= +1$), those in the faces are attractors (index $= +1$), and those on the edges are saddles with multiplicity 1 (index $= -1$). This means that the sum of all the indices is equal to $V - E + F = \chi(M)$, as claimed.

This concludes the proof of Theorem 12.1. As an immediate consequence of this and Proposition 12.13, we also obtain:

Theorem 12.14 (Gauss and Bonnet)**.** *Let M be an orientable compact surface equipped with a Riemannian metric. Then*

$$\int_M K\omega = 2\pi\chi(M). \tag{12.31}$$

Observe that the left hand side depends on the Riemannian metric, in principle, but the right hand side depends only on the topology of the ambient space.

Remark 12.15. One can extend the Gauss–Bonnet theorem to the nonorientable case, using the fact every nonorientable compact surface M admits an *orientable double cover*: a compact orientable surface $\tilde{M}$ and a local diffeomorphism $\pi : \tilde{M} \to M$ such that $\#\pi^{-1}(y) = 2$ for every $y \in M$. Their Euler characteristics are related by $\chi(\tilde{M}) = 2\chi(M)$. Given any Riemannian metric on M there exists exactly one Riemannian metric on $\tilde{M}$ such that π is a local isometry and, in particular, preserves the Gauss curvature and the area form:

$$K(\pi(x)) = \tilde{K}(x) \text{ and } \omega_{\pi(x)}(D\pi(x)u, D\pi(x)v) = \tilde{\omega}_x(u, v)$$

for any $x \in \tilde{M}$ and $u, v \in T_x\tilde{M}$. Applying Theorem 12.14 to the surface $\tilde{M}$, and recalling that π is a two-to-one map, we see that

$$2\pi\chi(\tilde{M}) = \int_{\tilde{M}} \tilde{K}\tilde{\omega} = \int_{\tilde{M}} (K \circ \pi)\tilde{\omega} = 2\int_M K\omega.$$

Dividing by 2, we obtain the statement of the theorem for the surface M.

12.5. Comments on Mayer's theorem

In this section we shall complement the statement of Theorem 11.21 by proving that under its assumptions the Euler characteristic $\chi(M) = 2 - 2g(M)$ of the surface is necessarily nonpositive, and the total number N of flow components, periodic and minimal, is 1 if $\chi(M) = 0$ and does not exceed $-2\chi(M) = 4g(M) - 4$ if $\chi(M) < 0$.

Recall (from Section A.5) that the existence of an area form implies that the surface is orientable. It is also useful to keep in mind that the compact orientable surfaces are classified up to diffeomorphism by the genus $g(M)$ or, equivalently, by the Euler characteristic $\chi(M)$. Check Example 12.12.

Our starting point is the *Euler–Poincaré formula*

$$\sum_{i=1}^{k} m_i = -\chi(M). \tag{12.32}$$

that relates the number k and the multiplicities $m_1, \ldots, m_k$ of the generalized saddles to the Euler characteristic of the surface. In Exercise 12.2 we invite

the reader to check that this formula is a particular case of the Poincaré–Hopf theorem.

As the multiplicities are positive integers, (12.32) implies that vector fields in the conditions of the theorem do not exist when $\chi(M) > 0$, that is, when M is diffeomorphic to the sphere $\mathbb{S}^2$. Moreover, when $\chi(M) = 0$ the Euler–Poincaré formula implies that $k = 0$. In other words, vector fields as in the theorem on surfaces diffeomorphic to the torus $\mathbb{T}^2$ cannot have stationary points. Indeed (Exercise 11.17), their flows are necessarily differentiably equivalent to the rigid flow (11.14) for some value of a.

According to Corollary 11.31 and Lemma 11.34, the boundary of each component consists of saddle connections and stationary points. If $\chi(M) = 0$, then there are no stationary points, and so there are no saddle connections either. Hence, by connectedness, there exists a unique component, which coincides with the whole M.

Hereafter we shall assume that $\chi(M) < 0$. It follows from the Euler–Poincaré formula that the number k of stationary points satisfies $1 \leqslant k \leqslant -\chi(M)$. Corollary 11.31 and Lemma 11.34 also ensure that each saddle connection belongs to the boundary of no more than two components of the flow. Hence, the number N of components cannot exceed twice the number of saddle connections, which does not exceed twice the number of stable (or unstable) separatrices. In other words, using (12.32) once more,

$$N \leqslant 2\sum_{i=1}^{k}(m_i + 1) = 2k - 2\chi(M). \tag{12.33}$$

With a little more work, one can improve this estimate by a factor of half. Let us briefly sketch that argument.

Firstly, the Euler–Poincaré formula (12.32) continues to hold for the restriction of the flow to any open domain D bounded by periodic trajectories:

$$\sum_{z_i \in D} m_i = -\chi(D). \tag{12.34}$$

This is because the Poincaré–Hopf theorem also holds for manifolds with boundary, as long as the vector field is tangent to the boundary and has no stationary points on it. More generally, and for similar reasons, if the boundary D is formed by regular trajectories and stationary points, then

$$\sum_{z_i \in D} m_i + \frac{1}{2}\sum_{z_j \in \partial D} s_j(D) = -\chi(D), \tag{12.35}$$

where $s_j(D)$ represents the number of separatrices of each stationary point $z_j \in \partial D$ contained in D.

As a consequence, *a homoclinic connection γ is never homotopically trivial*, that is, it is never the boundary of a disc contained in M. Indeed, suppose that there exists a homoclinic connection that bounds a disc $D \subset M$ and let z_j be the stationary point associated to such a homoclinic connection. As the Euler characteristic of the disc is equal to 1, the identity (12.35) means that

$$\sum_{z_i \in D} m_i + \frac{1}{2} s_j(D) = -1,$$

and this is impossible because $m_i \geqslant 1$ and $s_j(D) \geqslant 0$.

We can now deduce that the boundary of every component of the flow contains at least *two* saddle connections. Indeed, suppose there is some component whose boundary contains a single saddle connection. If the connection is heteroclinic, that is, if it joins distinct stationary points, then it obviously does not disconnect M. If the connection is homoclinic, its union with the stationary point is a simple closed curve. The only way it could disconnect M would be if it were homotopically trivial, and we just saw that this is impossible. Therefore, this situation cannot occur.

Combining this observation with the argument used to obtain (12.33), we see that the total number N of components of the flow satisfies

$$N \leqslant \sum_{i=1}^{\kappa} (m_i + 1) = k - \chi(M) \leqslant -2\chi(M). \tag{12.36}$$

See also the remark in Exercise 11.19.

12.6. Experiment: oxygen–ozone cycle

Interesting chemical processes consist of several simultaneous chemical reactions interacting among themselves, some much faster than others. Thus the corresponding differential equations involve quantities that vary over very different timescales, which often causes the differential equation to be stiff (recall Section 4.6). We shall illustrate this with the following example.

The *Chapman cycle* models the production and destruction of ozone O_3 in the atmosphere, under the action of the solar ultraviolet rays. It consists of four stages:

(1) Molecules of oxygen (O_2) are broken up into pairs of isolated atoms of O under the action of ultraviolet rays: $O_2 + uv \to O + O$. This is a slow reaction, with reaction rate $\ell_1 = 3 \times 10^{-12}$.

(2) In the presence of a catalyst M, isolated atoms combine with oxygen molecules to form ozone molecules: $M + O + O_2 \to M + O_3$. This reaction is very slow: its reaction rate is $\ell_2 = 1,22 \times 10^{-33}$.

(3) Under the action of ultraviolet rays, the ozone molecule is broken into one molecule of oxygen plus one isolated atom: $O_3 + uv \to O + O_2$. This reaction is fast, with reaction rate $\ell_3 = 5,5 \times 10^{-4}$.

(4) Additionally, ozone molecules can combine with isolated atoms, producing two oxygen molecules: $O + O_3 \to O_2 + O_2$. This is also a slow reaction, with reaction rate $\ell_4 = 6,86 \times 10^{-16}$.

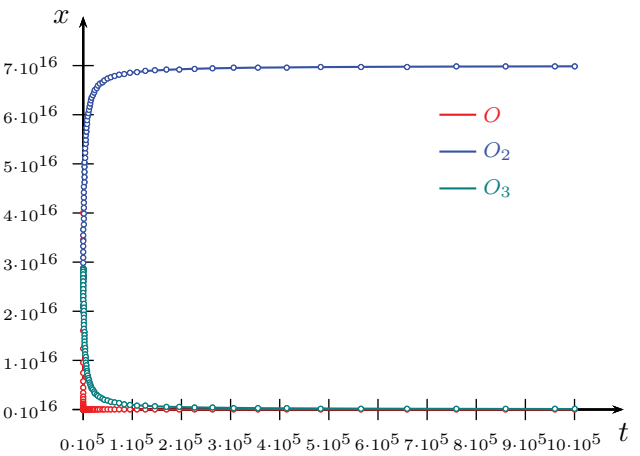

Figure 12.15. Solutions of the Chapman cycle (12.37).

Let $x = [O]$ be the concentration of the isolated atoms O, let $y = [O_2]$ be the concentration of molecular oxygen O_2, let $z = [O_3]$ be the concentration of ozone O_3, and let $m = [M]$ be the concentration of the catalyst M. Evidently, this last concentration is constant over time. It strongly influences the presence of ozone, since the second reaction (creation of O_3) is much slower than the last two (destruction of O_3). According to the laws of chemical kinetics, the other concentrations evolve according to the following differential equation:

$$\begin{cases} x' = 2\ell_1 y - \ell_2 mxy + \ell_3 z - \ell_4 xz, \\ y' = -\ell_1 y - \ell_2 mxy + \ell_3 z + 2\ell_4 xz, \\ z' = \ell_2 mxy - \ell_3 z - \ell_4 xz. \end{cases} \tag{12.37}$$

Objectives:

(1) Write computer codes for the following implicit numerical integration methods: 2-step and 3-step Adams–Moulton, backward differentiation (BDF) with $k = 2$ and $k = 3$, and implicit Runge–Kutta with $k = 2$ and $k = 3$.

(2) Consider $m = 9 \times 10^{17}$. Find approximate solutions of equation (12.37) using the methods listed in the previous item. Represent the results graphically (check Figure 12.15).

(3) Compare the performance of these methods in terms of the execution time. Can you identify the phenomenon of stiffness? Note that the differences become clearer when we integrate the equation over long intervals of time.

(4) Verify that the solution with initial condition $x(0) = 4\times10^{16}$, $y(0) = 2 \times 10^{16}$, and $z(0) = 2 \times 10^{16}$ tends to a "stationary" state of the system.

(5) Experiment with different initial conditions and catalyst concentrations. Look for interesting scenarios.

Use these tasks to reflect on the advantages and disadvantages of the different numerical methods used in computational experiments throughout the book. Under what conditions would you choose each of the methods presented?

12.7. Exercises

Exercise 12.1. For each of the following vector fields $F : \mathbb{R}^2 \to \mathbb{R}^2$, calculate the index of the origin and sketch its integral curves:

(1) $F(x, y) = (-y, x)$,

(2) $F(x, y) = (ax, by)$, $a, b \neq 0$,

(3) $F(z) = z^n$, $z = x + iy$ and $n \geqslant 1$,

(4) $F(z) = \overline{z}^n$, $z = x + iy$ and $n \geqslant 1$.

Exercise 12.2. Verify that the index of a generalized saddle with multiplicity m (as defined in Section 11.4) is equal to $-m$ and deduce the Euler–Poincaré formula (12.32).

Exercise 12.3. Consider the vector fields $F_1, F_2 : \mathbb{R}^2 \to \mathbb{R}^2$ given by

$$F_1(x, y) = (x^3 - 3xy^2, 3x^2y - y^3) \text{ and } F_2(x, y) = (x^2 - y^2, 2xy).$$

Find the stationary points of F_1 and F_2 and calculate the corresponding indices. For $i = 1, 2$, investigate the existence of a vector field $\tilde{F}$ on the sphere $\mathbb{S}^2$ with a unique stationary point p and such that $\operatorname{ind}(\tilde{F}, p) = \operatorname{ind}(F_i, 0)$.

Exercise 12.4. Let F be a vector field on a compact orientable surface M with an isolated stationary point $p \in M$. Let ω_{12} be the connection form associated to the positively oriented orthonormal frame (e_1, e_2) defined by $e_1(x) = F(x)/\|F(x)\|_x$ for $x \neq p$ in a neighborhood of p. Show that

$$\frac{1}{2\pi} \lim_{r\to 0} \int_{c_r} \omega_{12}\, ds = \operatorname{ind}(F, p),$$

where c_r represents the boundary of the geodesic disc with center at p and radius r.

Exercise 12.5. Verify the steps of the following proof of Theorem 12.10, given by Cauchy in 1811. Let $\phi : |P| \to \mathbb{S}^2$ be any triangulation of the sphere.

(1) Verify that if we remove the interior of a face of P, we obtain a polyhedron P_0 with $\chi(P_0) = \chi(P) - 1$.

(2) Justify that there exists an embedding $\psi : |P_0| \to \mathbb{R}^2$, that is, the polyhedron P_0 triangulates a domain of the plane.

(3) Consider a triangle $T \in P_0$ on the boundary of the polyhedron and let P_1 be the polyhedron obtained by removing T along with its edges and vertices that are not shared by other faces of P_0. Verify that $\chi(P_1) = \chi(P_0)$.

(4) Construct a sequence P_n of polyhedra with a decreasing number of faces and such that $\chi(P_n)$ is constant, and use it to deduce that $\chi(P) = 2$.

Remark. In step (3) there are two cases to consider: T shares with other faces two edges and three vertices, or one edge and two vertices.

Exercise 12.6. Show that if c is a simple closed curve in the plane containing the origin in its inside, then the winding number of c about the origin is equal to ± 1, with sign depending on the orientation of the curve.

Exercise 12.7. Show that the Poincaré–Hopf theorem continues to hold for C^1 vector fields on the closed unit disc $D^2 \subset \mathbb{R}^2$, as long as the vector field points outwards (meaning that $F(x) \cdot x > 0$) at every point on the boundary.

Exercise 12.8. Let $f : D^2 \to D^2$ be a C^1 map on the closed unit disc in the plane. Justify (without using Brouwer's theorem!) that f has some fixed point in D^2.

Exercise 12.9. In the City of Lakes there are seven lagoons, connected by ten canals, so that it is possible to navigate from one lagoon to another through the canals. How many islands does the city have?

Exercise 12.10. Sketch a vector field on the sphere having:

(1) two nodes,

(2) two centers,

(3) a unique stationary point,

(4) three stationary points.

Exercise 12.11. Sketch a vector field on the torus having:

(1) no stationary points,

(2) two centers and two saddles,

(3) one center and one saddle,

(4) two saddles and one dipole (see Figure 12.3).

Exercise 12.12. It is known that every compact orientable surface M with boundary is homeomorphic to a compact surface of genus g with $r \geqslant 1$ disjoint discs "removed" from it. Justify that:

(1) the Euler characteristic of M is given by $2 - 2g - r$;

(2) the Euler characteristic of the closed unit disc is $\chi(D^2) = 1$.

Exercise 12.13. Let M_1 and M_2 be two compact surfaces. The *connected sum* of M_1 and M_2 is defined in the following way. Let $D_i \subset M_i$, $i = 1, 2$, be open discs in each of the surfaces and $f : \partial D_1 \to \partial D_2$ be a diffeomorphism. Consider the surface

$$M_1 \# M_2 = \big((M_1 \setminus D_1) \cup (M_2 \setminus D_2)\big) / \sim,$$

where $\sim$ is the equivalence relation induced by f, that is, the smallest equivalence relation such that $x \sim f(x)$ for every $x \in \partial D_1$. Calculate $\chi(M_1 \# M_2)$ in terms of $\chi(M_1)$ and $\chi(M_2)$. Deduce that $\chi(M \# \mathbb{S}^2) = \chi(M)$ for every compact surface M and interpret this fact.

Exercise 12.14. Show that on a compact orientable surface with positive curvature at all points, two closed geodesics always intersect.

(*Hint*. Use the version of the Gauss–Bonnet theorem for surfaces with boundary formulated in relation (12.40).)

Exercise 12.15. Let M be a compact surface, possibly with boundary, with nonpositive curvature at all points. Show that no closed geodesic can be the boundary of a disc contained in M. In particular, if there exists a closed geodesic, then M cannot be a disc.

Exercise 12.16. Let M be a compact surface, possibly with boundary, with negative curvature at every point. Let γ_1 and γ_2 be two geodesics starting from the same point p and such that they meet again at a point $q \neq p$. Show

that the segments of γ_1 and γ_2 between p and q cannot form the boundary of a disc.

Exercise 12.17. Let F be a C^1 vector field on a manifold of dimension d, and let p be a stationary point. What is the relation between $\mathrm{ind}(F, p)$ and $\mathrm{ind}(-F, p)$?

Exercise 12.18. Let $f : \mathcal{U} \to \mathcal{V}$ be a C^1 map between two open subsets of manifolds. Given any differential 1-form α on $\mathcal{V}$, consider the differential 1-form $f^*\alpha$ defined on $\mathcal{U}$ by $(f^*\alpha)_x = \alpha_{f(x)} \circ Df(x)$. Analogously, given any differential 2-form ω on $\mathcal{V}$, consider the differential 2-form $f^*\omega$ defined on $\mathcal{U}$ by $(f^*\omega)_x = \omega_{f(x)} \circ \big(Df(x) \times Df(x)\big)$. Show that:

(1) If $\alpha = d\varphi$ for some function $\varphi : \mathcal{V} \to \mathbb{R}$, then $f^*\alpha = d(\varphi \circ f)$.

(2) Analogously, if $\omega = d\alpha$, then $f^*\omega = d(f^*\alpha)$.

(3) If α is exact, then $f^*\alpha$ is exact, and if α is closed, then $f^*\alpha$ is closed.

(4) $\int_{f\circ\gamma} \alpha = \int_\gamma f^*\alpha$ for any curve $\gamma : [0, 1] \to \mathcal{U}$.

Exercise 12.19. Let a and b be positive constants. Show that

$$\int_{-\pi/2}^{\pi/2} \frac{ab^2 \cos\theta}{(a^2 \sin^2\theta + b^2 \cos^2\theta)^{3/2}}\, d\theta = 2.$$

(*Hint*. The ellipsoid $S = \big\{(x, y, z) \in \mathbb{R}^3 : x^2/a^2 + y^2/a^2 + z^2/b^2 = 1\big\}$ is relevant here.)

Exercise 12.20. The differential equation $x' = x^2 - x^3$ is a simple model for the combustion of a match. The variable x represents the radius of the flame, which is assumed to be approximately spherical. The equation means that the growth of the flame depends on the difference between the amount of oxygen available to be burned, which is proportional to the area of the outer surface, and the amount of oxygen that is burned, which is proportional to the volume of the flame.

Analyze the behavior of the solutions qualitatively. Solve the differential equation numerically for initial conditions $x(0) = x_0$ close to zero, for example $x_0 = 0.01$, using (a) the RKF45 method and (b) the implicit Euler method. Plot the results graphically. Compare the performances of the two methods, in terms of the number of iterations and the total calculation time.

Exercise 12.21. Use the 2-step and 3-step Adams–Moulton methods to integrate the Lotka–Volterra equation (1.19) with $c_1 = 1$, $c_2 = -1$, $a_{12} = 1$, and $a_{21} = 1$ and various initial conditions. Represent the solutions graphically and compare with the results obtained in Exercise 8.24 through the Crank–Nicolson and RKF45 methods.

Exercise 12.22. The method most commonly used in the numerical solution of stiff equations is called ODE23s. It is a combination of second and third order methods, with interpolation of the points of the trajectory over time. The second order method is given by the relations

$$\begin{cases} F_0 = F(t_n, x_n), & Wk_1 = F_0 + hTd, \\ F_1 = F(t_n + h/2, x_n + hk_1/2), & Wk_2 = (F_1 - k_1) + Wk_1, \\ t_{n+1} = t_n + h, & x_{n+1} = x_n + hk_2, \end{cases}$$

where $T = \partial_t F(t_n, x_n)$, $J = \partial_x F(t_n, x_n)$, $d = 1/(2+\sqrt{2})$, $e = 6+\sqrt{2}$, h is the step size, and $W = \mathrm{Id} - hJd$. It is implicit only in a linear fashion: it suffices to invert the matrix W to be able to extract the values of k_1 and k_2. The third order method is used to improve error control:

$$\begin{cases} F_2 = F(t_{n+1}, x_{n+1}), & Wk_3 = F_2 - e(k_2 - F_1) - 2(k_1 - F_0) + hTd, \\ & \text{error} = h(k_1 - 2k_2 + k_3)/6. \end{cases}$$

This additional step has low computational cost because F_2 may be reused as F_0 in the next step. Finally, one interpolates between x_n and x_{n+1} by means of

$$x_{n+s} = x_n + h\left(\frac{s(1-s)}{1-2d}k_1 + \frac{s(s-2d)}{1-2d}k_2\right) \text{ for } s \in [0,1]. \tag{12.38}$$

This is very useful because stiff problems usually involve phenomena with different time scales, so it may be necessary to integrate over long intervals of time: formula (12.38) enables us to use relatively coarse step sizes, to avoid making the computational cost prohibitive, while still doing a more detailed analysis of the solution at a low cost.

Write a computer program that executes the ODE23s method, and apply it to solve the differential equation $x' = 50(\cos t - x)$ for different initial conditions $x_0 \in [0, 2]$. Compare with the results obtained using the RKF45 method.

Exercise 12.23. Use the 3-step Adams–Moulton, ODE23s, and RKF45 methods to integrate the solutions of the van der Pol equation

$$x'' = -x + 20(1 - x^2)x'$$

with initial conditions $x(0) = 2$ and $x'(0) = 0$. Represent the solutions graphically and compare the performances of the three methods.

Exercise 12.24. Consider the differential equation

$$\begin{cases} x' = -\ell_1 x + \ell_3 yz, \\ y' = \ell_1 x - \ell_2 y^2 - \ell_3 yz, \\ z' = \ell_2 y^2, \end{cases}$$

corresponding to the following model of chemical reaction:

$$\begin{aligned} A \to B \text{ with reaction rate } \ell_1 &= 4 \times 10^{-2}, \\ B + B \to C + B \text{ with reaction rate } \ell_2 &= 3 \times 10^{7}, \\ B + C \to A + C \text{ with reaction rate } \ell_3 &= 10^{4}. \end{aligned}$$

Study the behavior of the solutions with initial condition (x_0, y_0, z_0) close to $(10^{-6}, 0, 0)$, using the ODE23s method.

12.8. Notes

The notion of a winding number of a curve around a point has its origin in complex analysis, more specifically, in Cauchy's residue theorem. Poincaré's interest in the subject arose from his efforts to generalize the residue theorem to dimensions greater than 2. In *Sur les résidus des intégrales doubles* (On the residues of double integrals) [**329**], he gave a suitable formulation of the concept of *surface integral*, characterizing when the integral does not depend on the surface itself but just on its boundary. Thus he rediscovered Stokes theorem,

$$\int_{\partial\Omega} \omega = \int_{\Omega} d\omega,$$

which allowed him to get the desired result for analytic functions with several complex variables. This work led Poincaré to discover homology, in *Analysis situs* [**334**] and its complements, published between 1899 and 1904, and it influenced decisively the work of Alexander and de Rham on cohomology. For more information on the concept of an index, check Milnor [**279**, Chapter 6] or Hirsch [**169**, Chapter 5].

In 1758, Euler [**122**] observed that the quantity $\chi(P) = F - E + V$ is always equal to 2, for any convex polyhedron.[1] Alternative proofs of Euler's theorem were given by mathematicians of the caliber of Legendre, Cauchy, and Steiner. Check the first chapter in the book of Hopf [**175**]. Notwithstanding, the Euler characteristic remained little more than a curiosity until the work of Poincaré [**333**, **335**] revealed its true significance, leading to remarkable generalizations.

Our presentation in Section 12.2 is partly based on Hopf [**175**] and Munkres [**291**, **292**]. Lemmas 12.8 and 12.9 are proved in Munkres [**291**].

[1] The French mathematician and philosopher Réné Descartes had made a related observation, around 1620: *the sum of the angle defects at all vertices of the polyhedron is always equal to* 4π. The original manuscript of Descartes was lost, but a copy that Leibniz made in 1676 and entitled *Progymnasmata de solidorum elementis excerpta ex manuscripto Cartesii* (Preliminary exercises on the elements of solids extracted from a manuscript of Descartes) was preserved [**105**]. The content and history of that text are discussed in Sasaki [**359**, Section 3.3].

Section A.8 contains a broader discussion of the notion of Euler characteristic, including higher dimensions, and we shall also return to the subject in the Notes to the Appendix.

The version of Theorem 12.1 for surfaces was proved by Poincaré [**333**] in 1893. The general statement, for manifolds in any dimension, was proved by Heinz Hopf [**173**] in 1926. The theorem has numerous applications, for example in the results on rigidity of closed convex surfaces in Euclidean space due to Stephan Cohn-Vossen [**92**] (check also [**175**, **279**]).

The German mathematician Heinz Hopf was born on November 19, 1894, in Gräbschen, in Silesia (nowadays Grabiszyn, which is part of the city of Wrocław, in Poland) and died on June 3, 1971, in Zollikon, Switzerland. He spent most of his career at the university in Berlin, Germany, and at the Federal Technical School (ETH) in Zurich, Switzerland. He made numerous contributions to geometry and topology, starting with his thesis, where he proved that every complete simply connected Riemannian 3-manifold with constant sectional curvature is isometric to the Euclidean space $\mathbb{R}^3$, the sphere $\mathbb{S}^3$, or the hyperbolic half-space $\mathbb{H}^3$. Another of his best known results was the discovery of the *Hopf invariant* for maps $S^3 \to S^2$. Hopf was the president of the International Mathematical Union from 1955 to 1958.

Wu [**418**] contains a detailed survey of the history of the Gauss–Bonnet theorem. The classical statement was formulated for a simply connected domain D on a surface, bounded by a finite number of differentiable curves:

$$\int_D K\, d\omega = 2\pi - \int_{\partial D} k_g\, ds - \sum_j (\pi - \alpha_j), \tag{12.39}$$

where k_g is the geodesic curvature of the boundary curves, and the α_j are the inner angles at the boundary vertices. The version for surfaces with boundary is analogous:

$$\int_M K\, d\omega = 2\pi\chi(M) - \int_{\partial M} k_g\, ds. \tag{12.40}$$

The version of (12.39) for geodesic triangles (domains bounded by three geodesic segments),

$$\int_T K\, d\omega = 2\pi - \sum_{j=1}^{3} (\pi - \alpha_j) = \sum_{j=1}^{3} \alpha_j - \pi, \tag{12.41}$$

was found by Gauss [**134**] in 1827. The general case was proved by Bonnet [**49**] em 1848.

Pierre Ossian Bonnet was born on December 22, 1819, in Montpellier, France, and died in Paris on June 22, 1892. He graduated as an engineer but chose to follow a career in research and education, becoming a professor

at the École Polytechnique, the university in Paris (Sorbonne), the École Normale Supérieure, and the Bureau des Longitudes, all in Paris. Among his important contributions to differential geometry is the notion of geodesic curvature, which led him to the result we now know as the Gauss–Bonnet theorem.

Johann Carl Friedrich Gauss was born on April 30, 1777, in Braunschweig, in the electorate of Saxony, and died on February 23, 1855, in Göttingen, in the kingdom of Hannover (both Saxony and Hannover are now part of Germany). The *prince of mathematicians*, as he is often called, is probably the most influential mathematician of all time. His outstanding talent was recognized from very early, and the duke Karl Wilhelm Ferdinand of Brunswick-Wolfenbüttel decided to pay for the studies of the young prodigy of humble origin born in his domain. Together with his fundamental contributions to many areas of mathematics and science in general, such as algebra, analysis, statistics, number theory, astronomy, optics, geodesy, geophysics, and electrostatics, Gauss is justly considered the founder of differential geometry and manifold theory.

The statement of the Gauss–Bonnet theorem for compact orientable surfaces (Theorem 12.14) was obtained by the German mathematician Walther von Dyck [**410**] in 1888. It can also be deduced from (12.39) using a triangulation of the surface. In 1927, Cohn-Vossen [**92**] proved that the inequality

$$\int_M K \, d\omega \leqslant 2\pi\chi(M)$$

holds for any surface, not necessarily compact. In general the inequality is strict: that is the case for surfaces with boundary, for example, since the geodesic curvature k_g of the boundary is necessarily positive.

The Gauss–Bonnet theorem may also be extended to Riemannian manifolds of any even dimension (the case of odd dimension is not interesting, as both sides of the identity are automatically zero). The first result in that direction was obtained by Heinz Hopf [**173**, **174**] in the 1920s, for hypersurfaces of $\mathbb{R}^{2n+1}$. His argument uses the Gauss normal map, and thus is not valid for submanifolds of Euclidean spaces with codimension greater than 1. That difficulty was bypassed in 1940, independently, by Allendoerfer [**4**] and Fenchel [**126**], who generalized the theorem to any submanifold of even dimension of a Euclidean space.

In 1943, Allendoerfer and Weil [**5**] extended the Gauss–Bonnet theorem to any Riemannian manifold with even dimension, not necessarily embedded in a Euclidean space. This extension may be deduced from the result of Allendoerfer and Fenchel using the celebrated isometric embedding theorem of John Nash [**294**, **295**]: *every compact Riemannian manifold may be embedded isometrically in some Euclidean space.* However, such an approach

is unsatisfactory: it should be possible to prove that the theorem holds in abstract manifolds through an intrinsic argument, which does not depend on realizing the manifold inside a Euclidean space. This situation was resolved in 1944 by Chern [**83**]: he found a suitable replacement for the Gauss normal map which allowed him to give an intrinsic proof of the theorem in full generality.

Numerical methods for the integration of stiff problems are presented extensively in the book of Hairer and Wanner [**157**]. The differential equation studied in Exercise 12.20 is attributed to Larry Shampine, see [**281**]. The paper [**417**] discusses several stiff problems, including the model in Exercise 12.24. The Chapman cycle was proposed in 1930 by the British mathematician and physicist Sydney Chapman. The ODE23s method was formulated in 1997 by Shampine, Reichelt [**366**].