Ludwigs-Maximilians-Universität München
— Department for Physics —

University Observatory

Computational Methods in Astrophysics

Ordinary Differential Equations

—

Cosmological Models

second, completely revised edition (2007)

Joachim Puls/Fabian Heitsch

winter semester 2007/2008

Foreword

This is an *instruction set* for a lab course on Ordinary Differential Equations. As such, it is neither complete nor perfect (and we're not talking about typos here).

From your point of view, the former can be a cause of nuisance and – in most cases – a source for additional work. This is intended. The script cannot replace reading the original literature. Wherever appropriate, we have tried to point to the relevant papers/books.

We would greatly appreciate any feedback on inconsistencies, mistakes, obvious blunders, and outright nonsense in the script. Also, we'd be very thankful for suggestions how to improve the script beyond the afore-mentioned.

Despite these ominous beginnings, we hope you'll enjoy the following. After all, numerics are a never-ending source of surprises!

Contents

Chapter 1

Introduction

1.1. Why Ordinary Differential Equations?

Differential equations (DEs) are omnipresent when it comes to determining the dynamical evolution, the structure, or the stability of physical systems. In many cases, the resulting set of DEs contains several independent variables (e.g., spatial coordinates and time in hydrodynamics), in which case the DEs are called partial differential equations (PDEs). These will be of no concern here.

However, it is often possible to simplify the set of DEs by physical reasoning such that the number of independent variables is reduced to 1, in which case we speak of ordinary differential equations (ODEs). For example, the classical stellar structure models neglect time evolution and non-spherical effects, resulting in coupled differential equations only depending on the radius or the mass coordinate. Likewise, chemical networks are often formulated locally, i.e., the (coupled) rate equations depend only on time.

The solving "strategy" (at least for initial value problems) in most cases boils down to: (0) Choose the appropriate solution method ("solver"), in dependence of the properties of the ODEs (and their solution functions). (1) Determine the initial conditions. (2) Find the appropriate step size. (3) Advance the solution by that step size. (4) Repeat (2) and (3). Of course, the problems arise in the steps (2) and (3). What is the correct step size, and how should we integrate the equations? This entails problems like: how large are the errors we make? And, how much can we trust the solutions? The rest of this script centers around these questions, and attempts to throw some light on possible answers with the help of some examples.

There is a whole wealth of possible applications of ODEs in physics and astronomy. However, instead of (re-)introducing the Kepler problem or the differential equations describing stellar structure, we will (finally) use the Friedmann–Lemaître equation(s) describing the temporal evolution of our cosmos to shed some light on how to integrate ODEs numerically, and to obtain an impression of what can go wrong if one has no theoretical background.

For this purpose (but also in order to tackle different equations underlying different problems – for example, so-called stiff problems), we need the integrators and techniques which are introduced in Chapter 2, after we have briefly recapitulated some basic facts (including the (in?)famous theorem of Picard–Lindelöf which tells us under which conditions ODEs can be solved *uniquely*).

Chapter 3 gives a short introduction into cosmological models, concentrating on the derivation of the Friedmann–Lemaître equations and related problems. As a major outcome, we will formulate the final equation for the temporal evolution of the "cosmic scale factor", which has

to be solved in dependence of the total matter and radiation densities and the cosmological constant, which is interpreted as "dark energy" nowadays.

After all this, we finally get to the experiments themselves in Chapter 4. On the first day of our lab work, we will investigate different techniques to solve ODEs, at hand of three different examples. On the second day then, we will apply our accumulated knowledge to solve the Friedmann–Lemaître equation under various conditions. A highlight will be the reconstruction of the famous Ω_M vs. Ω_Λ diagram (e.g., Perlmutter et al. 1999), which allows us to understand the various possibilities for the future fate of our and other cosmoses, in dependence of total matter density and cosmological constant.

Due to the above layout of the "experimental" work, we suggest the following schedule for a fruitful preparation:

Before day 1: Study the numerical methods (Chapter 2), and have a look into the problems to be solved on the first day (Section 4.1).

Before day 2: Study the introduction into cosmological models (Chapter 3), and refer to the literature in case of problems. Inform yourself on the experiments/problems planned for the second day (Section 4.2). In order to be able to solve some of these problems, we suggest to work on Exercises 5 and 6 also *before* day 2.

Chapter 2

Numerics

2.1. Some general facts about ODEs

A set of ordinary differential equations[1] is defined as

$$\boldsymbol{y}' = \boldsymbol{f}(x, \boldsymbol{y}), \tag{2.1}$$

where x is the independent variable, and the prime denotes derivation w.r.t. x. The vectors shall have the length n. In most of what follows, we will consider the *scalar* ODE (first order)

$$y' = f(x, y) \tag{2.2}$$

for the solution function $y(x)$ with initial condition

$$y(x_0) = y_0 \tag{2.3}$$

for given x_0 and y_0 (*initial value problem*).

Scalar DEs of higher order can be reduced to first order vector DEs (set of DEs, see above), which can be treated as scalar ones.

2.1. Example. $y'' = f(x, y, y')$ with $y(x_0) = y_0$, $y'(x_0) = y_0'$.
Let

$$\boldsymbol{y} := \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \quad \text{and} \quad \boldsymbol{f}(x, \boldsymbol{y}) := \begin{pmatrix} y_2 \\ f(x, y_1, y_2) \end{pmatrix}.$$

Then we have to solve

$$\boldsymbol{y}' = \boldsymbol{f}(x, \boldsymbol{y}) \qquad \text{with} \qquad \boldsymbol{y}(x_0) = \begin{pmatrix} y_0 \\ y_0' \end{pmatrix}$$

Interestingly and fortunately, it can be shown that the above problems for y' have *exactly one solution* y, if f fulfills certain conditions.

2.1.1. Existence and uniqueness

2.2. Definition. *Let $G \subset \mathbb{R} \times \mathbb{R}$ and $f : G \to \mathbb{R}$. Then, f is subject to a* Lipschitz condition *with* Lipschitz constant $L \geq 0$, *if*

$$\big|f(x, y) - f(x, \tilde{y})\big| \leq L|y - \tilde{y}| \qquad \textit{for all } (x, y), (x, \tilde{y}) \in G. \tag{2.4}$$

[1] A helpful summary is given by Stoer and Bulirsch (1990) and Press et al. (1992).

Criterium. Let $f \in \mathcal{C}^1(G)$ with $G \subset \mathbb{R} \times \mathbb{R}$ compact and convex. Then, f is subject to a *Lipschitz condition* with *Lipschitz constant*

$$L = \max_{(x,y)\in G} \left|\partial_y f(x,y)\right|.$$

This is a direct consequence of the mean value theorem of calculus. Try to prove this yourself.

2.3. Theorem (Picard–Lindelöf). *Let $\alpha, \beta > 0$, $(x_0, y_0) \in \mathbb{R}^2$ and*

$$R := \big\{(x,y) : |x - x_0| \le \alpha, |y - y_0| \le \beta\big\}.$$

Moreover, $f \in \mathcal{C}^0(R)$ shall be a function with $0 < \gamma := \max|f| < \infty$, which is subject to a Lipschitz condition.

Then, there is exactly one function $y \in \mathcal{C}^1(I)$ in $I := [x_0 - \delta, x_0 + \delta]$, $\delta := \min\{\alpha, \beta/\gamma\}$ with $y'(x) = f\big(x, y(x)\big)\ \forall\ x \in I$ and $y(x_0) = y_0$.

For a proof, see any textbook on calculus. The generalization to sets of ODEs is straightforward.

The initial value problem $y'(x) = f(x,y)$ with $y(x_0) = y_0$ even has a solution if f is continuous "only". But then the uniqueness of the solution can no longer be warranted.

2.4. Example. $f(x,y) = y^{2/3}$, $(x,y) \in \mathbb{R} \times \mathbb{R}$, is continuous, but *not* subject to a Lipschitz condition, because the requirement $|y^{2/3} - 0| \le L|y - 0|$, or equivalently, $1/|y|^{1/3} \le L$, cannot be fulfilled for any $L \ge 0$ for all y around 0.

Obviously, $y_1(x) = 0$ $(x \in \mathbb{R})$ and $y_2(x) = \frac{1}{27}(x - x_0)^3$ $(x \in \mathbb{R})$ are two different solutions of the initial value problem $y' = f(x,y)$, $y(x_0) = 0$ with $x_0 \in \mathbb{R}$.

Moreover, it can be shown that the solution of an initial value problem depends continuously on the initial value, and that – if $\partial f_i/\partial y_j$ is defined, continuous, and finite on R – the solution even depends continuously differentiably on the initial value.

2.2. Consistency, convergence, and discretization errors

If solving an initial value problem numerically, one has, at specific abscissa values

$$a =: x_0 < x_1 < \ldots < x_N := b \tag{2.5}$$

with step sizes

$$h_n := x_{n+1} - x_n \qquad (n = 0, 1, \ldots, N-1), \tag{2.6}$$

to construct approximations

$$y_n \approx y(x_n) \tag{2.7}$$

for the (exact) values of the desired solution. Often, one chooses equidistant step sizes

$$h := \frac{b-a}{N}. \tag{2.8}$$

Single-step method. By means of a *single-step method*, the approximation y_{n+1} at position x_{n+1} is solely calculated from the approximation (x_n, y_n). (For so-called *multi-step* methods – which are no longer as popular as some time ago – see the literature). The corresponding algorithm has the general form

$$y_{n+1} = y_n + h_n \varphi(x_n, y_n, y_{n+1}, h_n), \tag{2.9}$$

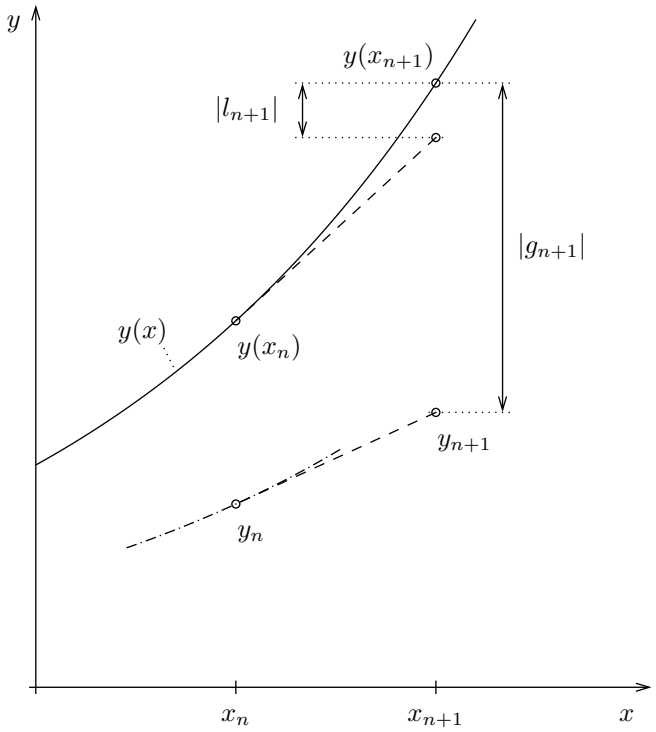

Figure 2.1: Local and global discretization error.

where φ is a specific, algorithm dependent function. If φ does *not* depend on y_{n+1}, we speak of an *explicit single-step method*, otherwise of an *implicit single-step method*.

In the latter case, an *implicit* equation for y_{n+1} has to be solved (numerically). This is the disadvantage of the implicit method. Its advantage is given by its stability (see Section 2.5).

Let us now consider the error (w.r.t. the exact solution) which has accumulated after a certain number of steps. At first, we will neglect rounding errors. In the following, we will restrict ourselves to equidistant step sizes.

Global discretization error. The *global discretization error* at position x_n measures the difference

$$g_n := y(x_n) - y_n \qquad (n = 0, 1, \ldots, N). \tag{2.10}$$

The single-step method is called *convergent*, if

$$\max_{n=0,1,\ldots,N} |g_n| \to 0 \qquad \text{for } h \to 0^+. \tag{2.11}$$

Such a method has a *convergence order* $p > 0$, if the global discretization error can be written in the form

$$\max_{n=0,1,\ldots,N} |g_n| \leq ch^p = \mathcal{O}(h^p), \tag{2.12}$$

with constants $c \geq 0$ and $p > 0$ independent of h. To obtain useful estimates for the global discretization error, we also have to define a

Local discretization error, which is defined (at position x_{n+1}) by

$$l_{n+1} := y(x_{n+1}) - y(x_n) - h\varphi\big(x_n, y(x_n), y(x_{n+1}), h\big). \tag{2.13}$$

This local error describes the deviation of the exact solution from the algorithm function. For an explicit method, l_{n+1} is the difference between exact value $y(x_{n+1})$ and approximation y_{n+1}, if we would start at x_n with the exact value $y(x_n)$ (error of *one step*; see Fig. 2.1).

A single-step method is called *consistent*, if

$$\frac{1}{h} l_{n+1} \to 0 \qquad \text{for } h \to 0^+ \qquad (n = 0, 1, \ldots, N-1). \tag{2.14}$$

Because of

$$\frac{l_{n+1}}{h} = \underbrace{\frac{y(x_{n+1}) - y(x_n)}{h}}_{\text{secant slope of exact solution}} - \underbrace{\varphi\big(x_n, y(x_n), y(x_{n+1}), h\big)}_{\text{approximation of this slope}}, \tag{2.15}$$

we have the equivalence of

$$\frac{1}{h}|l_{n+1}| \xrightarrow{h\to 0^+} 0 \qquad \Longleftrightarrow \qquad \varphi\big(x_n, y(x_n), y(x_{n+1}), h\big) \xrightarrow{h\to 0^+} f\big(x_n, y(x_n)\big). \tag{2.16}$$

The method has a *consistency order* (brief: order) of p, if the local discretization error fulfills the inequality

$$|l_{n+1}| \le ch^{p+1} = \mathcal{O}(h^{p+1}) \tag{2.17}$$

with constants $c \ge 0$ and $p > 0$, independent of h.

We are now able to relate the global with the local discretization error. For a method of consistency order p and an ODE with function y subject to a LIPSCHITZ constant L, we finally obtain after some effort

$$\max_{n=0,1,\ldots,N-1} |g_{n+1}| \le \frac{c}{L}\left(\mathrm{e}^{L(b-a)} - 1\right) h^p \tag{2.18}$$

for the explicit case, and

$$\max_{n=0,1,\ldots,N-1} |g_{n+1}| \le \frac{c}{2L}\left(\mathrm{e}^{2L(b-a)-1}\right) h^p + \mathcal{O}(h^{p+1}) \tag{2.19}$$

for the implicit one. Thus, local errors of $\mathcal{O}(h^{p+1})$ will lead, after $N = \dfrac{b-a}{h}$ steps, to a global error of $\mathcal{O}(Nh^{p+1}) = \mathcal{O}(h^p)$, as to be expected.

Rounding errors. Let us now consider the effect of rounding errors and "faulty" initial values. To this end, we assume an algorithm subject to errors

$$\tilde{y}_{n+1} = \tilde{y}_n + h\varphi(x_n, \tilde{y}_n, \tilde{y}_{n+1}, h) + \delta_{n+1} \qquad (n = 0, 1, \ldots, N-1) \tag{2.20}$$

with

$$\tilde{y}_0 = y_0 + \varepsilon_0. \tag{2.21}$$

Let

$$|\delta_{n+1}| \le \delta \qquad (n = 0, 1, \ldots, N-1) \tag{2.22}$$

and

$$\varepsilon := |\varepsilon_0|. \tag{2.23}$$

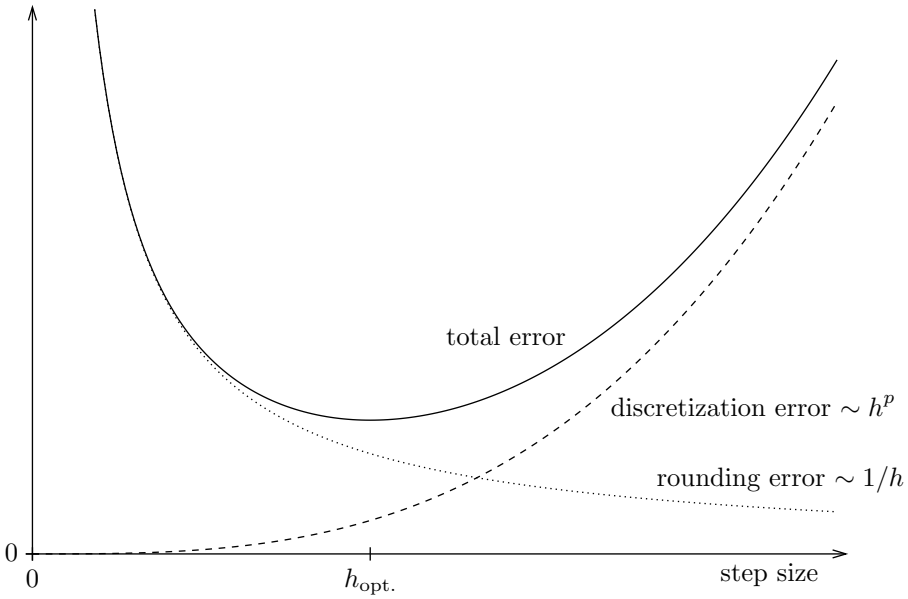

Figure 2.2: Total error due to discretization and rounding.

The estimate of the total error (due to discretization and rounding)

$$t_n := y(x_n) - \tilde{y}_n \qquad (n = 0, 1, \ldots, N-1) \tag{2.24}$$

can be obtained quite similar to the case without rounding errors, and results in

$$\max_{n=0,1,\ldots,N-1} |t_{n+1}| \leq \mathrm{e}^{L(b-a)}\varepsilon + \frac{1}{L}\left(\mathrm{e}^{L(b-a)} - 1\right)\left(ch^p + \frac{\delta}{h}\right) \tag{2.25}$$

for the explicit case, and

$$\max_{n=0,1,\ldots,N-1} |t_{n+1}| \leq \mathrm{e}^{L(b-a)}\varepsilon + \frac{1}{2L}\left(\mathrm{e}^{2L(b-a)/(1-Lh)} - 1\right)\left(ch^p + \frac{\delta}{h}\right) \tag{2.26}$$

for the implicit one. In particular, we obtain

$$\max_{n=0,1,\ldots,N-1} |t_{n+1}| \leq \frac{1}{L}\left(\mathrm{e}^{L(b-a)} - 1\right)\left(ch^p + \frac{\delta}{h}\right) \tag{2.27}$$

for the explicit, single-step method with *exact* initial values.

Exercise 1: *For the latter case, calculate the optimum step size.*

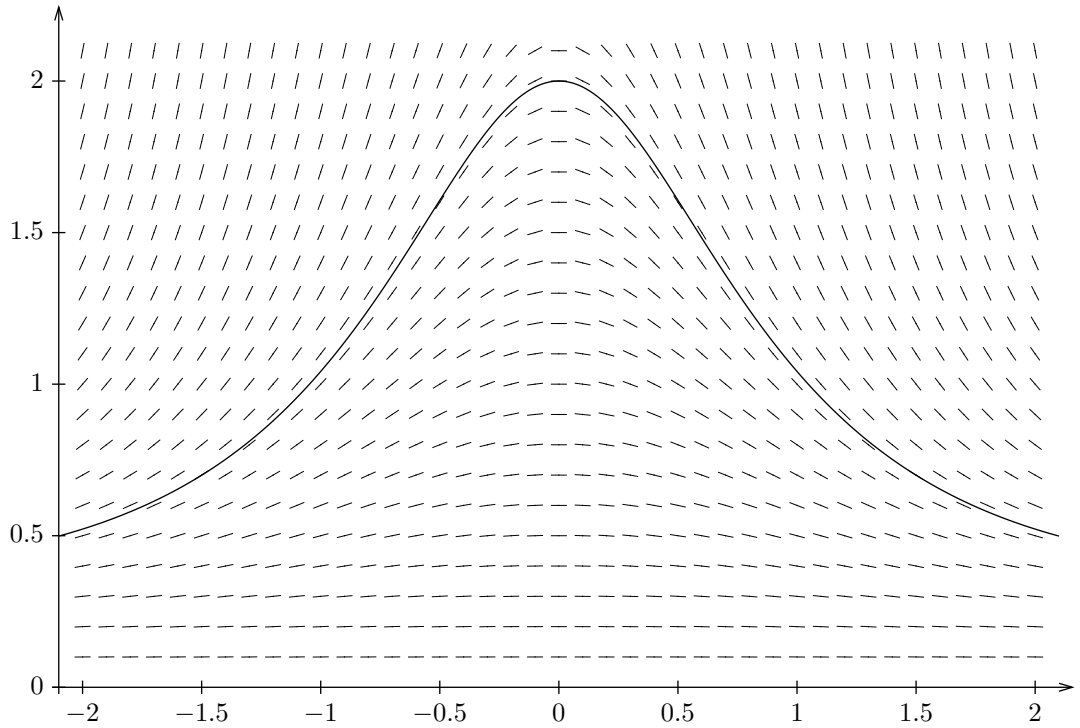

Figure 2.3: Directional field for example 2.5.

2.3. Single-step methods

Let $y(x)$ be the solution of the DE $y' = f(x, y)$. If the graph of this solution passes through a point $(\tilde{x}, \tilde{y})$, the slope of this graph at this point is $f(\tilde{x}, \tilde{y})$: the value $f(\tilde{x}, \tilde{y})$ fixes the slope of the tangent of the solution curve at this point. By drawing the corresponding slopes for *many* points in the xy plane, one can roughly sketch the graph of the solution of the initial value problem $y' = f(x, y)$, $y(x_0) = y_0$, if one follows this *directional* field (also called slope field), starting at the initial value (x_0, y_0).

2.5. Example. $y' = -\sin(x)y^2$ with initial value $y(0) = 2$.
The exact solution is $y(x) = \dfrac{-2}{2\cos(x) - 3}$. The directional field and this solution are displayed in Fig. 2.3.

2.3.1. Euler method

The most simple numerical method to solve the initial value problem (2.2) is to approximate the solution curve by a piecewise linear function (polygon), accounting for the directional field. For each n, the linear piece of the polygon in between x_n and x_{n+1} points to direction $f(x_n, y_n)$ (= slope of the tangent to a solution curve through (x_n, y_n), which usually differs from the exact one through $(x_n, y(x_n))$.

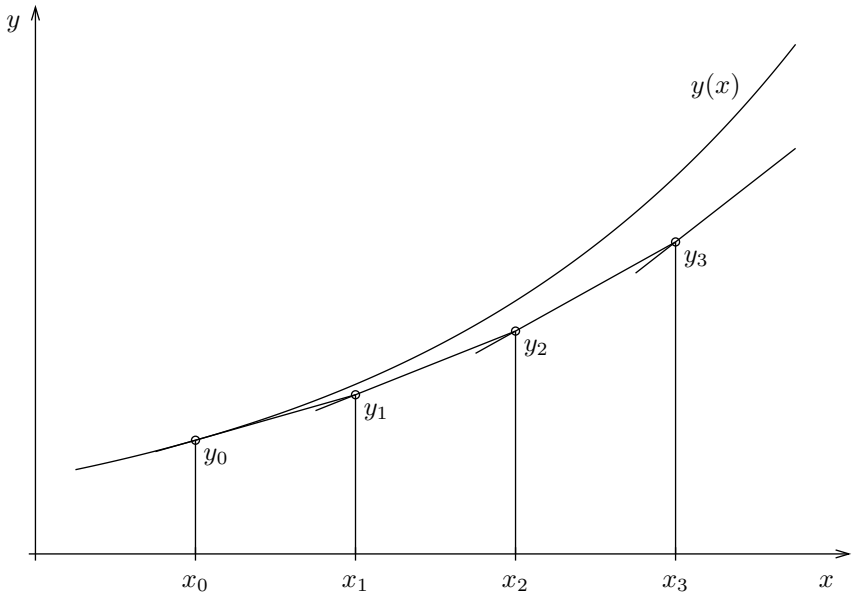

Figure 2.4: EULER method.

By means of this so-called EULER method, the approximations for the solution are thus calculated according to

$$y_{n+1} = y_n + hf(x_n, y_n) \qquad (n = 0, \dots, N-1), \tag{2.28}$$

see Fig. 2.4. The local discretization error is given by

$$\begin{aligned} l_{n+1} &= \underbrace{y(x_{n+1})}_{(*)} - y(x_n) - hf\big(x_n, y(x_n)\big) \\ &= \frac{1}{2}\Big(\partial_x f\big(x_n, y(x_n)\big) + f\big(x_n, y(x_n)\big)\partial_y f\big(x_n, y(x_n)\big)\Big)h^2 + \mathcal{O}(h^3), \end{aligned} \tag{2.29}$$

using the TAYLOR expansion

$$\begin{aligned} (*) &= y(x_n) + y'(x_n)h + \frac{1}{2}y''(x_n)h^2 + \mathcal{O}(h^3) \\ &= y(x_n) + f\big(x_n, y(x_n)\big)h + \frac{1}{2}\Big(\partial_x f\big(x_n, y(x_n)\big) + f\big(x_n, y(x_n)\big)\partial_y f\big(x_n, y(x_n)\big)\Big)h^2 + \mathcal{O}(h^3). \end{aligned}$$

Thus, the consistency order of the EULER method is $p = 1$.

2.6. Example. $y' = -2xy^2$, $y(0) = 1$. Table 2.1 displays the results using the EULER method, for different step sizes (see also Fig. 2.5). Fig. 2.6 shows the corresponding total error, which decreases (roughly) proportional to step size. The exact solution is $y(x) = \dfrac{1}{1+x^2}$.

		$h = 0.1$		$h = 0.01$		$h = 0.001$	
x_n	$y(x_n)$	y_n	t_n	y_n	t_n	y_n	t_n
0.0	1.00000	1.00000	0	1.00000	0	1.00000	0
0.1	0.99010	1.00000	−0.00990	0.99107	−0.00097	0.99020	−0.00010
0.2	0.96154	0.98000	−0.01846	0.96330	−0.00176	0.96171	−0.00018
0.3	0.91743	0.94158	−0.02415	0.91969	−0.00226	0.91766	−0.00022
0.4	0.86207	0.88839	−0.02632	0.86448	−0.00242	0.86231	−0.00024
0.5	0.80000	0.82525	−0.02525	0.80229	−0.00229	0.80023	−0.00023
0.6	0.73529	0.75715	−0.02185	0.73727	−0.00198	0.73549	−0.00020

Table 2.1: Score table of EULER method for example 2.6.

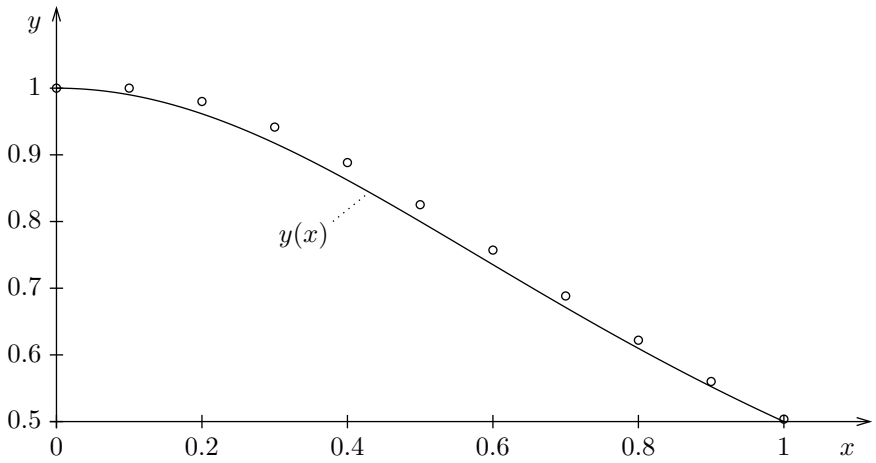

Figure 2.5: EULER method (example 2.6).

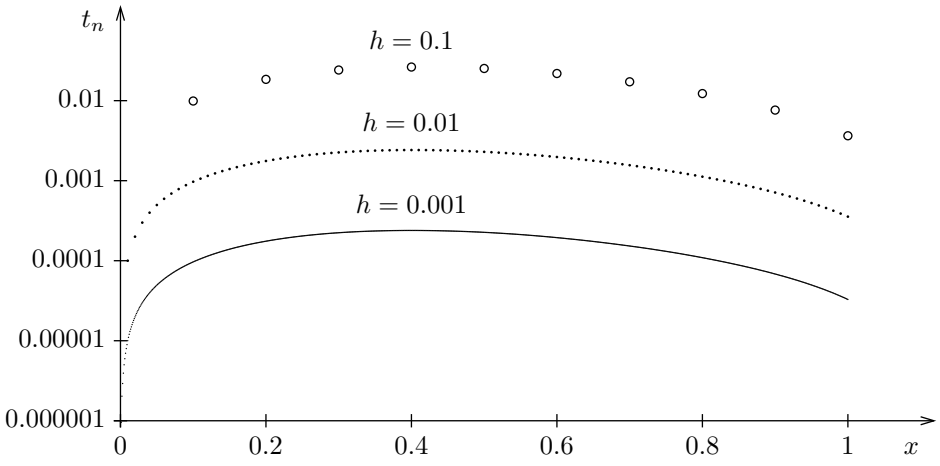

Figure 2.6: Total error of EULER method (example 2.6).

2.7. Example. $y' = y$, $y(0) = 1$. EULER method with $h = 0.1$; the result is displayed in Fig. 2.7, and the exact solution is $y(x) = \mathrm{e}^x$.

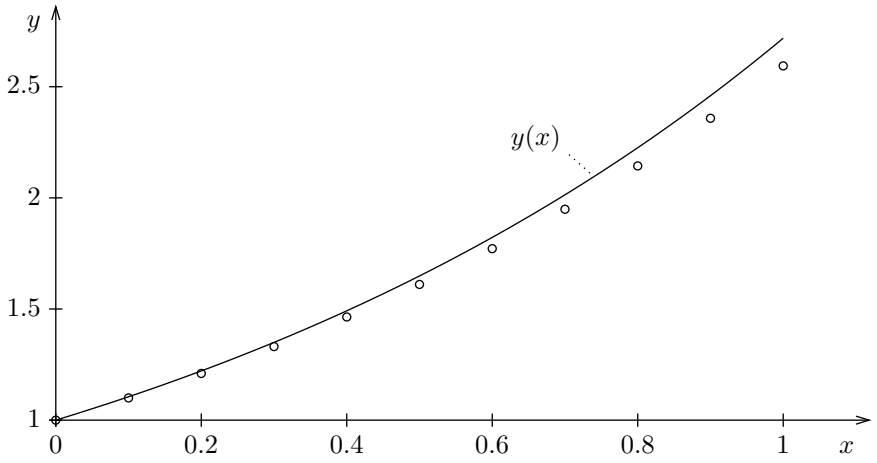

Figure 2.7: EULER method (example 2.7).

For this example, Fig. 2.7 shows that the (absolute) total error increases with increasing x_n, i.e., the numerical solution deviates more and more from the exact one, which is a typical situation. To diminish this effect, one can either choose a smaller step size, i.e., increase the number of steps (global discretization error $\mathcal{O}(h)$). Alternatively, one can use higher-order methods, which is advisable because of rounding errors and computational time.

Note that the Euler method is not symmetric and can even become unstable (cf. Section 2.5, in particular Example 2.10). To see this already here, consider

$$y' = -cy, \tag{2.30}$$

where $c > 0$ is a constant. The Euler scheme with fixed step size h would then be

$$y_{n+1} = y_n + hy'_n = (1 - ch)y_n. \tag{2.31}$$

If $h > 2/c$, the method becomes unstable, since $|y_n| \to \infty$ as $n \to \infty$. For constant c, one could choose the step size small enough to prevent the solution's explosion. However, if c is not constant (and we will meet such cases below), this doesn't work anymore, because for each step we would need to know what the maximum allowed step size h would be. There are two remedies: adaptive step sizes, leading to methods with error control, and implicit or semi-implicit methods.

2.3.2. Generalized RUNGE–KUTTA methods

In general, the directional field changes from one to the next approximation value. This is accounted for by the so-called RUNGE–KUTTA methods, where the idea behind these methods is to take a "trial" step (usually to the midpoint of the interval), and then to use the values of x and y at that point to calculate the "real" step across the whole interval. Note that this is not the same as splitting the integration step in half. A first example is the method by COLLATZ,

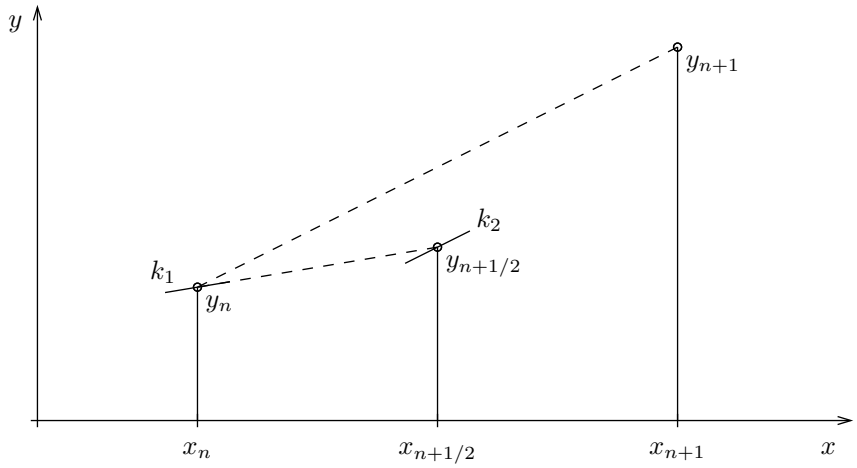

Figure 2.8: COLLATZ method.

which belongs to the class of *extrapolation methods* (see Fig. 2.8):

$$\begin{aligned} k_1 &:= f(x_n, y_n), \\ k_2 &:= f\left(x_n + \frac{h}{2}, y_n + \frac{h}{2}k_1\right), \\ y_{n+1} &= y_n + h \cdot k_2. \end{aligned} \tag{2.32}$$

The local discretization error can be shown to have the structure

$$l_{n+1} = F\big(x_n, y(x_n)\big)h^3 + \mathcal{O}(h^4), \tag{2.33}$$

since the symmetrization cancels out the first and 2nd order terms in h. Thus, this method has a consistency order of $p = 2$.

2.8. Example (continuation of Example 2.6)**.** Score table of COLLATZ method:

	$h = 0.1$		$h = 0.05$	
x_n	y_n	t_n	y_n	t_n
0.0	1.00000	0	1.00000	0
0.1	0.99000	0.00010	0.99007	0.00002
0.2	0.96118	0.00036	0.96145	0.00009
0.3	0.91674	0.00069	0.91727	0.00016
0.4	0.86110	0.00096	0.86184	0.00023
0.5	0.79889	0.00111	0.79974	0.00026
0.6	0.73418	0.00111	0.73503	0.00026
0.7	0.67014	0.00100	0.67091	0.00023
0.8	0.60895	0.00080	0.60957	0.00018
0.9	0.55191	0.00058	0.55236	0.00013
1.0	0.49964	0.00036	0.49992	0.00008

For same step size, the total errors are (absolutely) smaller than for the EULER method. Comparing different step sizes, the consistency order of $p = 2$ is obvious.

In order to develop methods of higher order, some systematic procedure is required. To this end, we integrate the DE $y'(x) = f(x, y(x))$ over the interval $[x_n, x_{n+1}]$ of length $h = x_{n+1} - x_n$ w.r.t. the independent variable x:

$$y(x_{n+1}) = y(x_n) + \int_{x_n}^{x_{n+1}} f\big(x, y(x)\big)\,\mathrm{d}x. \tag{2.34}$$

Note that the initial value problem (2.2) is equivalent to the integral equation

$$y(x) = y_0 + \int_{x_0}^{x} f\big(x', y(x')\big)\,\mathrm{d}x'. \tag{2.35}$$

The integral (2.34) is approximated by quadrature formulas (cf. the Numerical Lab "Integration")

$$\int_{x_n}^{x_{n+1}} f\big(x, y(x)\big)\,\mathrm{d}x \approx h \sum_{r=1}^{m} \gamma_r\, f\big(x_n + \alpha_r h, y(x_n + \alpha_r h)\big) \tag{2.36}$$

with weights $\gamma_r \geq 0$ and abscissa values $x_n + \alpha_r h$ with $0 \leq \alpha_r \leq 1$ $(r = 1, 2, \ldots, m)$, $\alpha_1 := 0$.

The major problem now is that the $y(x_n + \alpha_r h)$ in (2.36) are unknown and have to be replaced by approximations, using the following ansatz:

$$\begin{aligned}
f\big(x_n + \alpha_1 h, y(x_n + \alpha_1 h)\big) &=: k_1\big(x_n, y(x_n)\big),\\
f\big(x_n + \alpha_2 h, y(x_n + \alpha_2 h)\big) &\approx f\Big(x_n + \alpha_2 h, y(x_n) + h\beta_{21} k_1\big(x_n, y(x_n)\big)\Big)\\
&=: k_2\big(x_n, y(x_n)\big),\\
f\big(x_n + \alpha_3 h, y(x_n + \alpha_3 h)\big) &\approx f\Big(x_n + \alpha_3 h, y(x_n) + h\Big(\beta_{31} k_1\big(x_n, y(x_n)\big) + \beta_{32} k_2\big(x_n, y(x_n)\big)\Big)\Big)\\
&=: k_3\big(x_n, y(x_n)\big)\\
&\;\;\vdots\\
f\big(x_n + \alpha_m h, y(x_n + \alpha_m h)\big) &\approx f\left(x_n + \alpha_m h, y(x_n) + h \sum_{s=1}^{m-1} \beta_{ms} k_s\big(x_n, y(x_n)\big)\right)\\
&=: k_m\big(x_n, y(x_n)\big)
\end{aligned} \tag{2.37}$$

with constants β_{rs}. One requires

$$\sum_{s=1}^{r-1} \beta_{rs} = \alpha_r, \qquad (r = 2, 3, \ldots, m), \tag{2.38}$$

so that the approximations (2.37) are exact at least to $\mathcal{O}(h)$. This follows from the TAYLOR expansion in step size:

$$\begin{aligned}
f\big(x_n + \alpha_r h, y(x_n + \alpha_r h)\big) - f\left(x_n + \alpha_r h, y(x_n) + h \sum_{s=1}^{r-1} \beta_{rs} k_s\big(x_n, y(x_n)\big)\right) &=\\
= \left(\alpha_r - \sum_{s=1}^{r-1} \beta_{rs}\right) f\big(x_n, y(x_n)\big)\, \partial_y f\big(x_n, y(x_n)\big)\, h + \mathcal{O}(h^2).&
\end{aligned} \tag{2.39}$$

α_1					
α_2	β_{21}				
α_3	β_{31}	β_{32}			
$\vdots$					
α_m	β_{m1}	β_{m2}	$\cdots$	$\beta_{m,m-1}$	
	γ_1	γ_2	$\cdots$	γ_{m-1}	γ_m

Figure 2.9: BUTCHER tableau for an explicit RUNGE–KUTTA method of m-th degree.

Ansatz (2.37) together with (2.36) yields the algorithm

$$y_{n+1} = y_n + h\sum_{r=1}^{m} \gamma_r k_r\big(x_n, y(x_n)\big). \tag{2.40}$$

From the requirement of consistency,

$$\sum_{r=1}^{m} \gamma_r k_r(x_n, y_n) \xrightarrow{h\to 0^+} f(x_n, y_n),$$

we finally find

$$\sum_{r=1}^{m} \gamma_r = 1, \tag{2.41}$$

since $k_1(x_n, y_n) = k_2(x_n, y_n) = \ldots = k_m(x_n, y_n)$ at $h = 0$.

A method constructed in this way is called an explicit RUNGE–KUTTA method of m-th degree. The degree m denotes the number of evaluations of the function f in one integration step. The coefficients of the algorithm are usually summarized in a tableau following Figure 2.9.

Special cases. In the following, we summarize a few special cases and indicate their correspondence with well-known quadrature formulas.

- $m = 1$:

0	
	1

  $$\begin{aligned} k_1 &= f(x_n, y_n), \\ y_{n+1} &= y_n + hk_1. \end{aligned} \tag{2.42}$$

  EULER Method (2.28): rectangle rule.

- $m = 2$:

0		
1	1	
	$\frac{1}{2}$	$\frac{1}{2}$

  $$\begin{aligned} k_1 &= f(x_n, y_n), \\ k_2 &= f(x_n + h, y_n + hk_1), \\ y_{n+1} &= y_n + \frac{h}{2}(k_1 + k_2). \end{aligned} \tag{2.43}$$

  HEUN method: trapezoid rule.

- $m = 2$:

  Exercise 2:
  Construct the tableau for the COLLATZ *method.*

$$\begin{aligned} k_1 &= f(x_n, y_n), \\ k_2 &= f\left(x_n + \frac{h}{2}, y_n + \frac{h}{2}k_1\right), \\ y_{n+1} &= y_n + hk_2. \end{aligned} \tag{2.44}$$

  COLLATZ method(2.32): mid-point rule.

- $m = 4$:

0				
$\frac{1}{2}$	$\frac{1}{2}$			
$\frac{1}{2}$	0	$\frac{1}{2}$		
1	0	0	1	
	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{6}$

$$\begin{aligned} k_1 &= f(x_n, y_n), \\ k_2 &= f\left(x_n + \frac{h}{2}, y_n + \frac{h}{2}k_1\right), \\ k_3 &= f\left(x_n + \frac{h}{2}, y_n + \frac{h}{2}k_2\right), \\ k_4 &= f\left(x_n + h, y_n + hk_3\right), \\ y_{n+1} &= y_n + \frac{h}{6}\big(k_1 + 2k_2 + 2k_3 + k_4\big). \end{aligned} \tag{2.45}$$

  Classical RUNGE–KUTTA method: SIMPSON's rule.

  To calculate the various "slopes" successively, only the previous value is needed. This *very popular* method has a consistency order of $p = 4$. The generalization to sets of DEs is straightforward:

$$\begin{aligned} \boldsymbol{k}_1 &= \boldsymbol{f}(x_n, \boldsymbol{y}_n), \\ \boldsymbol{k}_2 &= \boldsymbol{f}\left(x_n + \frac{h}{2}, \boldsymbol{y}_n + \frac{h}{2}\boldsymbol{k}_1\right), \\ \boldsymbol{k}_3 &= \boldsymbol{f}\left(x_n + \frac{h}{2}, \boldsymbol{y}_n + \frac{h}{2}\boldsymbol{k}_2\right), \\ \boldsymbol{k}_4 &= \boldsymbol{f}\left(x_n + h, \boldsymbol{y}_n + h\boldsymbol{k}_3\right), \\ \boldsymbol{y}_{n+1} &= \boldsymbol{y}_n + \frac{h}{6}\big(\boldsymbol{k}_1 + 2\boldsymbol{k}_2 + 2\boldsymbol{k}_3 + \boldsymbol{k}_4\big). \end{aligned} \tag{2.46}$$

The maximum possible consistency order $p_{\max}$ of an explicit RUNGE–KUTTA method of m-th degree is given in Table 2.2. To obtain these results, rather complex calculations are required, due to the need for solving large number of non-linear equations in parallel.

m	1	2	3	4	5	6	7	8	9	10	> 10
$p_{\max}$	1	2	3	4	4	5	6	6	7	7	$\leq m - 3$

Table 2.2: Maximum possible consistency order $p_{\max}(m)$ for RUNGE–KUTTA methods. Note that $p_{\max} = 4$ for both $m = 4$ and $m = 5$.

One should not be fooled by the "high order" labels here. Higher order does not necessarily mean higher accuracy. How well the method actually performs, depends solely on the problem to solve. The classical RUNGE–KUTTA method is implemented in routine `rk4`, but be warned: we will break it. The main disadvantage of course lies in the fixed step size, which eventually

brings us back to the problem we encountered with the Euler method. The higher order just postpones the failure of the scheme (see Section 2.5), though not to a considerable extent.

Remark. In addition to the explicit methods outlined, there are also *implicit* RUNGE–KUTTA methods, which might be used for *stiff* problems (cf. Section 2.5).

2.4. Step-size control

Clearly, it would be desirable to have some sort of control over the step size h. After all, the function to be integrated could be smooth over large parts – in which case we would like to integrate quickly over these boring regions, while isolated regions show large variation – for which the step size would have to be small in order to catch the salient details.

The problem is: How do we control the step size? The step size should be linked to an error estimate, i.e., we need an estimate how large the error for a given integration step is. If this estimate is larger than a certain (user-defined) threshold, the step size must decreased, and vice versa. With respect to computational time (and also rounding errors!), the step size should be always as large as possible.

In the following, we consider step $(n+1)$ of an explicit single-step method of consistency order p,

$$y_{n+1} = y_n + h\varphi(x_n, y_n, h). \tag{2.47}$$

The error made in this single step is $\tilde{y}(x_{n+1}) - y_{n+1}$, where $\tilde{y}(x)$ denotes the (exact) solution of the initial value problem

$$\tilde{y}' = f(x, \tilde{y}), \quad \tilde{y}_n = y_n. \tag{2.48}$$

The step size should be chosen such that this *local* error is constrained by

$$|\tilde{y}(x_{n+1}) - y_{n+1}| \lesssim \Delta_0, \tag{2.49}$$

where $\Delta_0 > 0$ is a given tolerance level. Since $\tilde{y}(x_{n+1})$ is unknown, the error has to be estimated. In our case, the local discretization error

$$\tilde{l}_{n+1} := \tilde{y}(x_{n+1}) - \underbrace{\tilde{y}(x_n)}_{=y_n} - h_n\varphi\big(x_n, \underbrace{\tilde{y}(x_n)}_{=y_n}, h_n\big) \tag{2.50}$$

with $\tilde{y}(x)$ instead of $y(x)$ is equal to the corresponding global discretization error

$$\tilde{g}_{n+1} := \tilde{y}(x_{n+1}) - y_{n+1}, \tag{2.51}$$

and because of our assumptions we have (convergence provided)

$$\tilde{g}_{n+1} = \tilde{l}_{n+1} \approx c h_n^{p+1}. \tag{2.52}$$

2.4.1. Error estimate from step doubling

Originally, the error estimate was achieved by step doubling. The integration step is performed twice, once with the full step size, then, independently, twice with the half step size. For a step with step size h_n, we have from above

$$\tilde{g}^{(1)}_{n+1} = \tilde{y}(x_{n+1}) - y^{(1)}_{n+1} \approx c h_n^{p+1}, \tag{2.53}$$

whereas a double step with size $h_n/2$ results in

$$\tilde{g}_{n+1}^{(2)} = \tilde{y}(x_{n+1}) - y_{n+1}^{(2)} \approx 2c\left(\frac{h_n}{2}\right)^{p+1} \approx c\frac{h_n^{p+1}}{2^p} \tag{2.54}$$

(remember that c is independent of h). Subtracting (2.54) from (2.53) yields

$$y_{n+1}^{(2)} - y_{n+1}^{(1)} \approx \left(1 - 2^{-p}\right) c h_n^{p+1}. \tag{2.55}$$

2.9. Example. For a RUNGE–KUTTA method of order $p = 4$, this difference (in terms of step size $h_n/2$) is given by

$$y_{n+1}^{(2)} - y_{n+1}^{(1)} \approx 30c\left(\frac{h_n}{2}\right)^5. \tag{2.56}$$

A final combination again with (2.53) results in

$$\tilde{y}(x_{n+1}) - y_{n+1}^{(1)} \approx c h_n^{p+1} \approx \frac{y_{n+1}^{(2)} - y_{n+1}^{(1)}}{1 - 2^{-p}}. \tag{2.57}$$

Optimum step size. Let $\bar{h}_n$ be the step size which should result in the predefined tolerance level,

$$\left|\tilde{y}(x_n + \bar{h}_n) - y_{n+1}\right| = \Delta_0. \tag{2.58}$$

From (2.52) we have

$$\Delta_0 \approx |c|\,\bar{h}_n^{p+1}, \tag{2.59}$$

and from (2.57) and (2.59) we obtain

$$\bar{h}_n \approx h_n \left(\frac{\left(1 - 2^{-p}\right)\Delta_0}{\left|y_{n+1}^{(2)} - y_{n+1}^{(1)}\right|}\right)^{1/(p+1)} \approx h_n \left(\frac{\Delta_0}{|\Delta y|}\right)^{1/(p+1)}. \tag{2.60}$$

This equation is used in two ways (see also below). If Δy is larger than Δ_0 in absolute value, it tells us how much to *decrease* the step size for a next retry of the present, *failed* step. If Δy is smaller than Δ_0, we can accept the present approximation for y_{n+1}, and the equation tells us how much we can *increase* h for the next step $n + 2$. In so far, we will always integrate close to optimum step size.

2.4.2. Embedded methods

Step-size control can be achieved also in a different way. Instead of calculating two approximations with the same method but different h, we can also calculate two approximations with same h but methods of different consistency order. Indeed, this procedure is favored to date. In step $(n + 1)$, we now have

- explicit single-step method with consistency order p:

$$\tilde{g}_{n+1}^{(1)} = \tilde{y}(x_{n+1}) - y_{n+1}^{(1)} \approx c_1 h_n^{p+1}. \tag{2.61}$$

- explicit single-step method with consistency order $p + 1$:

$$\tilde{g}_{n+1}^{(2)} = \tilde{y}(x_{n+1}) - y_{n+1}^{(2)} \approx c_2 h_n^{p+2}. \tag{2.62}$$

0							
$\frac{1}{5}$	$\frac{1}{5}$						
$\frac{3}{10}$	$\frac{3}{40}$	$\frac{9}{40}$					
$\frac{3}{5}$	$\frac{3}{10}$	$-\frac{9}{10}$	$\frac{6}{5}$				
1	$-\frac{11}{54}$	$\frac{5}{2}$	$-\frac{70}{27}$	$\frac{35}{27}$			
$\frac{7}{8}$	$\frac{1631}{55296}$	$\frac{175}{512}$	$\frac{575}{13824}$	$\frac{44275}{110592}$	$\frac{253}{4096}$		
	$\frac{37}{378}$	0	$\frac{250}{621}$	$\frac{125}{594}$	0	$\frac{512}{1771}$	order 5
	$\frac{2825}{27648}$	0	$\frac{18575}{48384}$	$\frac{13525}{55296}$	$\frac{277}{14336}$	$\frac{1}{4}$	order 4

Table 2.3: Embedded RUNGE–KUTTA method with CASH–KARP coefficients, as used in subroutine `rkck`.

Error estimate $\tilde{y}(x_{n+1}) - y_{n+1}^{(1)}$:

$$\tilde{y}(x_{n+1}) - y_{n+1}^{(1)} \approx y_{n+1}^{(2)} - y_{n+1}^{(1)} + \mathcal{O}(h^{p+2}). \tag{2.63}$$

Optimum step size $\bar{h}_n$ for tolerance level $\Delta_0 > 0$:

$$\left|\tilde{y}(x_{n+1}) - \bar{y}_{n+1}\right| = \Delta_0, \tag{2.64}$$

i.e.,

$$\Delta_0 \approx |c_1|\, \bar{h}_n^{p+1}. \tag{2.65}$$

From (2.61) and (2.63), we find in analogy to (2.60) (neglecting terms of order $\mathcal{O}(h^{p+2})$

$$\bar{h}_n \approx h_n \left(\frac{\Delta_0}{\left| y_{n+1}^{(2)} - y_{n+1}^{(1)} \right|} \right)^{1/(p+1)}. \tag{2.66}$$

Step-size control is implemented in analogy to above, e.g.,

- Calculate $y_{n+1}^{(1)}$, $y_{n+1}^{(2)}$ and $\bar{h}_n$.
- If $h_n \leq \tau \bar{h}_n$: accept $y_{n+1}^{(1)}$ (τ is a safety factor, e.g., $\tau = 0.9$).

  Else: replace h_n by $\tau \bar{h}_n$. Recalculate $y_{n+1}^{(1)}$, $y_{n+1}^{(2)}$ and $\bar{h}_n$.
  If necessary, reduce step size again.
- Use $\tau \bar{h}_n$ as initial step size for the next step.

$y_{n+1}^{(2)}$ should be calculated in parallel to $y_{n+1}^{(1)}$ with almost no additional effort. This can be achieved by so-called *embedded* RUNGE–KUTTA schemes, or RUNGE–KUTTA–FEHLBERG integrators. FEHLBERG used the fact that for RK schemes of order $p > 4$, more than p function

evaluations are needed (though never more than $p+3$, cf. Table 2.2). Fehlberg found a 5th-order method with $m = 6$ function evaluations, while another combination of those six functions (i.e., identical α_r and β_{rs} but different γ_r) yields a 4th-order method. Thus, the method of lower order is *embedded* into the higher order one. Table 2.3 shows the coefficients for such a $p = 4, 5$ method with coefficients as derived by Cash & Karp, which are somewhat advantageous compared to the original coefficients from Fehlberg.

2.4.3. Defining the tolerance level

With all this, at least the structure is set up. However, one question remains: How do we define the tolerance level, Δ_0, especially if we have a system of ODEs? This depends on the application. A first guess would be to choose a fractional error. However, this is bound to fail if we integrate oscillating functions, or simply quantities which are not positive definite (like, e.g., velocities!). So, should we use absolute errors? But then imagine you plan to integrate the trajectory of a particle in a gravitational field of a star, let's say. If the error in the radial coordinate r is absolute, the integration will get less and less accurate the closer to the star the particle passes. One possibility is to use a scaling array with an entry for each ODE, such that for a fractional error ϵ, the i-th equation would get a desired accuracy of

$$\Delta_0 = \epsilon y_{\mathrm{scal},i}, \tag{2.67}$$

where $y_{\mathrm{scal},i}$ is set to y_i for fractional errors, and to some absolute value for absolute errors. A useful "trick" to obtain constant fractional errors except near zero crossings is

$$y_{\mathrm{scal},i} = |y_i| + |h\partial_x y_i|. \tag{2.68}$$

This error scaling is done by the routine `odeint`.

2.5. Absolute Stability. Stiff sets of differential equations

Inappropriate use of numerical methods for solving initial value problems can lead to instabilities. In the following, we will discuss how to avoid them. Let us firstly consider the test initial value problem

$$y' = \lambda y, \qquad y(0) = 1, \tag{2.69}$$

which has, for $\Re(\lambda) < 0$ (denoting by $\Re$ the real part), the well known solution

$$y(x) = \mathrm{e}^{\lambda x}. \tag{2.70}$$

Let us solve this problem with the classical Runge–Kutta method (see (2.45)):

$$\begin{aligned}
k_1 &= \lambda y_n, \\
k_2 &= \lambda\left(y_n + \frac{h}{2}k_1\right), \\
k_3 &= \lambda\left(y_n + \frac{h}{2}k_2\right), \\
k_4 &= \lambda\left(y_n + hk_3\right), \\
y_{n+1} &= y_n + \frac{h}{6}\big(k_1 + 2k_2 + 2k_3 + k_4\big).
\end{aligned} \tag{2.71}$$

Thus, we have

$$y_{n+1} = F(\lambda h) y_n \tag{2.72}$$

with

$$F(\lambda h) = 1 + \lambda h + \frac{\lambda^2 h^2}{2} + \frac{\lambda^3 h^3}{6} + \frac{\lambda^4 h^4}{24}. \tag{2.73}$$

The exact solution, on the other hand, follows

$$y(x_{n+1}) = \mathrm{e}^{\lambda h} y(x_n). \tag{2.74}$$

Obviously, a 4th-order TAYLOR expansion of $\mathrm{e}^{\lambda h}$ in λh recovers the factor $F(\lambda h)$. This is consistent with the fact that the local discretization error is of $\mathcal{O}(h^5)$.

The exact solution always decays with $|y(x)| \to 0$ for $x \to \infty$. The numerical approximation, in contrast, decays ($y_n \to 0$ for $n \to \infty$) only if

$$|F(\lambda h)| < 1. \tag{2.75}$$

Because of $|F(\lambda h)| \to \infty$ for $\Re(\lambda) h \to -\infty$, this is not fulfilled for all λh. But for sufficiently small h, the condition (2.75) is warranted though.

The set

$$\big\{\mu \in \mathbb{C} : |F(\mu)| < 1\big\}$$

is called the *region of absolute stability* of the method. A measure for its size is the so-called

$$\textit{stability interval} := \text{stability region} \cap \text{real axis}.$$

The stability region of the classical RUNGE–KUTTA method is located symmetric to the real axis (see Fig. 2.10), and the corresponding stability interval is $]-2.78529, 0[$.

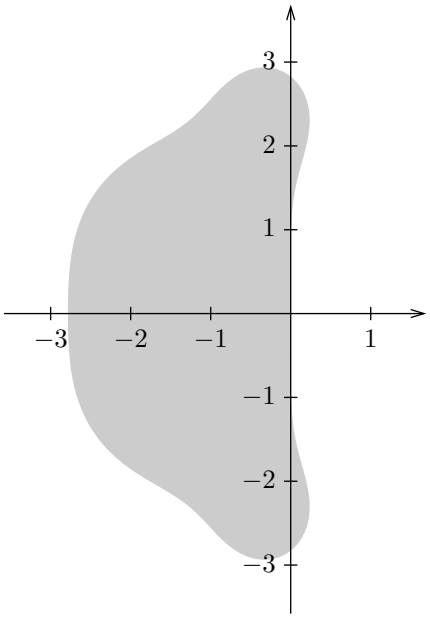

Figure 2.10: Stability region of the classical RUNGE–KUTTA method.

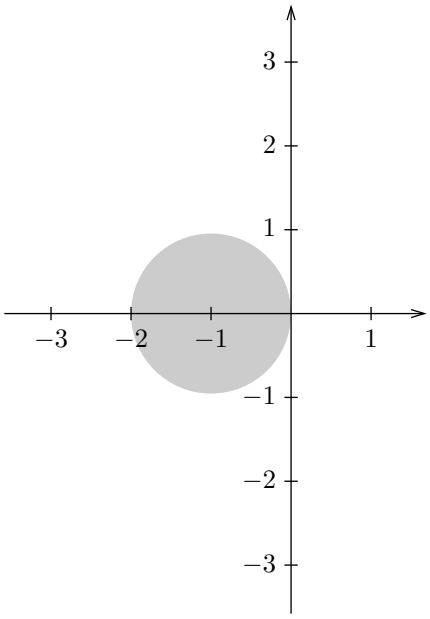

Figure 2.11: Stability region of the EULER method.

2.10. Example (Stability region of the EULER method)**.** For the EULER method, we have

$$F(\lambda h) = 1 + \lambda h, \tag{2.76}$$

i.e., the stability region is $\{\mu \in \mathbb{C} : |\mu + 1| < 1\}$. The stability interval is $]-2; 0[$ (see Fig. 2.11). For real λ, this corresponds to a maximum step size of $h_{\max} = 2/|\lambda|$, cf. also the discussion below Eq. (2.31) with $c \equiv -\lambda$.

The step size $h > 0$ must be chosen in such a way that λh is located within the region of absolute stability. Otherwise, the method will give incorrect results and might even "explode".

If the region of absolute stability comprises the entire half-plane $\Re(\mu) < 0$, the method is called *absolutely stable*.

Remark. *Implicit* RUNGE–KUTTA methods are absolutely stable, whereas typical multi-step methods have a finite stability region.

Solutions of sets of differential equations, which describe physical (or chemical or biological) processes, often have the property that they exponentially reach a stationary solution, partly coupled with (damped) oscillations (e.g., transient phenomena). The individual components of the solution can reach their final, constant value with *different speed*. This is typical for, e.g., chemical reaction networks (molecule formation).

To solve such sets with not too small step sizes, one has to use methods which are either absolutely stable or at least have a large region of absolute stability. For an illustration, we consider the test problem

$$\boldsymbol{y}' = \boldsymbol{A} \cdot \boldsymbol{y} + \boldsymbol{b}, \qquad \boldsymbol{y}(x_0) = \boldsymbol{y}_0 \tag{2.77}$$

with a $d \times d$ matrix $\boldsymbol{A}$ with eigenvalues λ_i $(i = 1, \ldots, d)$ which all have a negative real part.

If these real parts differ considerably, the initial value problem is called *stiff*. A measure for the *stiffness* is the so called *stiffness coefficient*

$$S := \frac{\max_i |\Re(\lambda_i)|}{\min_i |\Re(\lambda_i)|}. \tag{2.78}$$

If the coefficients of the ODE (i.e., the matrix elements) comprise several orders of magnitude, S can reach values of $\mathcal{O}(10^6)$ or even more. **To obtain meaningful numerical solutions, all products $h\lambda_i$ have to be located in the region of absolute stability.**

2.11. Example.

$$\begin{pmatrix} y_1' \\ y_2' \\ y_3' \end{pmatrix} = \begin{pmatrix} -21 & 19 & -20 \\ 19 & -21 & 20 \\ 40 & -40 & -40 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}, \qquad \begin{pmatrix} y_1(0) \\ y_2(0) \\ y_3(0) \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}. \tag{2.79}$$

The exact solution is

$$\begin{pmatrix} y_1(x) \\ y_2(x) \\ y_3(x) \end{pmatrix} = \begin{pmatrix} \frac{1}{2}\mathrm{e}^{-2x} + \frac{1}{2}\left(\cos(40x) + \sin(40x)\right)\mathrm{e}^{-40x} \\ \frac{1}{2}\mathrm{e}^{-2x} - \frac{1}{2}\left(\cos(40x) + \sin(40x)\right)\mathrm{e}^{-40x} \\ -\left(\cos(40x) - \sin(40x)\right)\mathrm{e}^{-40x} \end{pmatrix}. \tag{2.80}$$

Numerical solution with the EULER method: initial value $(y_1(0.1), y_2(0.1), y_3(0.1))^\intercal$ for $x_{\text{start}} = 0.1$, step size $h = 0.04$. The result is displayed in Fig. 2.12. The approximations at the end of the displayed interval are no longer acceptable, and the situation becomes even worse, if we consider the larger interval from 0.1 to 1.0 in Fig. 2.13 (note the different scale).

Since the eigenvalues of matrix $\boldsymbol{A}$ (2.79) are $\lambda_1 = -2$, $\lambda_2 = -40 + 40\mathrm{i}$, and $\lambda_3 = -40 - 40\mathrm{i}$, the stiffness coefficient is $S = 20$, i.e., the problem is not particularly stiff. With a step size of $h = 0.04$ the product $h\lambda_1 = -0.08$ is located within the stability region $\{\mu \in \mathbb{C} : |\mu + 1| < 1\}$ of the EULER method (cf. Fig. 2.11), whilst $h\lambda_2 = -1.6 + 1.6\mathrm{i}$ and $h\lambda_3 = -1.6 - 1.6\mathrm{i}$ are located outside. Thus, the strong deviation of the numerical solution from the exact one is due to the violation of the stability condition by eigenvalues λ_2 and λ_3, *although* their contribution to the solution has almost vanished for $x \gtrsim 0.1$. In order that *all* products $h\lambda_i$, $i = 1, 3$ are located within the stability region, the step size must be $h < 0.025$. In this case, then, the numerical solution becomes satisfactory.

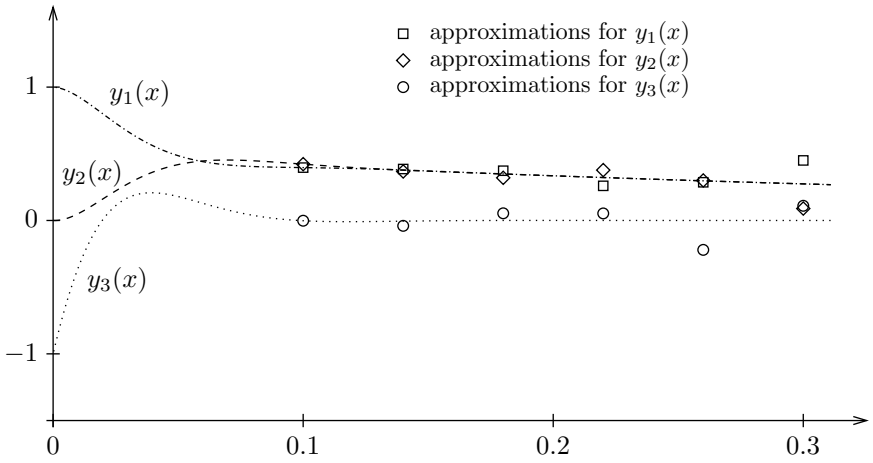

Figure 2.12: Analytical and numerical solution of example 2.11 for small x-values.

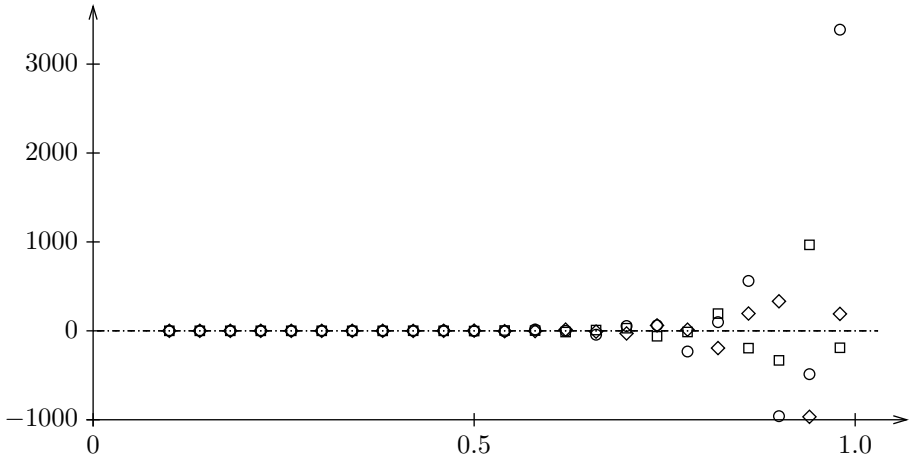

Figure 2.13: As Fig. 2.12, but for larger x-values.

An alternative approach to derive the stability condition for the EULER-method is as follows. We consider a set of ODEs

$$\boldsymbol{y}' = \boldsymbol{A} \cdot \boldsymbol{y}, \tag{2.81}$$

where $\boldsymbol{A}$ is a real matrix with eigenvalues λ_i $(i = 1, \ldots, d)$ which all have a negative real part (in order to assure that the exact solutions are exponentially decaying). Explicit differencing (see (2.31)) gives

$$\boldsymbol{y}_{n+1} = (\boldsymbol{1} + \boldsymbol{A}h) \cdot \boldsymbol{y}_n = \boldsymbol{C} \cdot \boldsymbol{y}_n. \tag{2.82}$$

Now, a matrix $\boldsymbol{C}^n \to 0$ only if the absolutely largest eigenvalue $|\lambda_{\max}| < 1$. Let us denote the eigenvalues of $\boldsymbol{A}$ by $\mu_i = -c_i + \mathrm{i}\Im_i$, where $c_i > 0$ and $\Im_i$ is the imaginary part of μ_i. Thus, the

eigenvalues of $\boldsymbol{C}$ are $(1 - c_i h + \mathrm{i}\Im_i\, h)$ and the maximum step size follows as

$$h_{\max} < \frac{2}{\max\limits_i \left(c_i + \dfrac{\Im_i^2}{c_i} \right)}. \tag{2.83}$$

Exercise 3: *Prove Eq. (2.83) and show that the maximum step size in example 2.11 indeed is* $h_{\max} = 0.025$.

Implicit differencing evaluates the RHS of the ODE not at position n, but $n+1$, i.e.,

$$\boldsymbol{y}_{n+1} = \boldsymbol{y}_n + h\boldsymbol{y}'_{n+1}, \tag{2.84}$$

which is equivalent to

$$\boldsymbol{y}_{n+1} = (\mathbf{1} - \boldsymbol{A}h)^{-1} \cdot \boldsymbol{y}_n. \tag{2.85}$$

If we denote the eigenvalues of $\boldsymbol{A}$ as above, the eigenvalues of $(\mathbf{1} - \boldsymbol{A}h)^{-1}$ are $(1 + c_i h - \mathrm{i}\Im_i\, h)^{-1}$, and their absolute value is $< 1\ \forall h$. Thus, the implicit scheme is stable for all step sizes h (and for the example above, it converges to the correct solution even for very large h). The downside is that each integration step requires a matrix inversion, and that the accuracy (for small x) is rather low, if h is significantly larger than in the corresponding explicit scheme.

2.6. Semi-implicit methods

Since by far not all ODEs have linear coefficients (the matrix $\boldsymbol{A}$ above), we need to generalize the implicit method for ODEs:

$$\boldsymbol{y}' = \boldsymbol{f}(\boldsymbol{y}) \quad \rightarrow \quad \boldsymbol{y}_{n+1} = \boldsymbol{y}_n + h\boldsymbol{f}(\boldsymbol{y}_{n+1}). \tag{2.86}$$

Generally, this set of equations needs to be solved iteratively at each time step, which in most cases means a prohibitively large computational effort. A way out is to linearize the equations,

$$\boldsymbol{y}_{n+1} = \boldsymbol{y}_n + h \left(\boldsymbol{f}(\boldsymbol{y}_n) + \left.\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{y}}\right|_{\boldsymbol{y}_n} \cdot (\boldsymbol{y}_{n+1} - \boldsymbol{y}_n) \right). \tag{2.87}$$

Here, $\partial \boldsymbol{f}/\partial \boldsymbol{y}$ is the JACOBIAN matrix of partial derivatives of the ODE's RHS w.r.t. y_i, i.e.,

$$\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{y}} \equiv \begin{pmatrix} \frac{\partial f_1}{\partial y_1} & \frac{\partial f_1}{\partial y_2} & \cdots & \frac{\partial f_1}{\partial y_n} \\ \frac{\partial f_2}{\partial y_1} & \frac{\partial f_2}{\partial y_2} & \cdots & \frac{\partial f_2}{\partial y_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial y_1} & \frac{\partial f_n}{\partial y_2} & \cdots & \frac{\partial f_n}{\partial y_n} \end{pmatrix}. \tag{2.88}$$

Eq. (2.87) can be rearranged as

$$\boldsymbol{y}_{n+1} = \boldsymbol{y}_n + h \left(\mathbf{1} - h\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{y}} \right)^{-1} \cdot \boldsymbol{f}(\boldsymbol{y}_n). \tag{2.89}$$

If h is not too big, usually one iteration of NEWTON's method is enough. This means that at each step we have to invert the matrix $\mathbf{1} - h\partial \boldsymbol{f}/\partial \boldsymbol{y}$ to find $\boldsymbol{y}_{n+1}$. Since the derivative (due

to the linearization) is taken at $\boldsymbol{y}_n$, the method is called a *semi-implicit* EULER method. It is not guaranteed to be stable, but it usually is, since the JACOBIAN locally corresponds to the constant matrix $\boldsymbol{A}$ from above. Of course, Eq. (2.89) is only 1st order (as the explicit EULER integrator Eq. (2.31)). Again, going to higher order (and subsequently higher computational effort) in most cases pays off, since generally fewer steps are needed. The most common methods are (a) generalized RK-schemes (ROSENBROCK), an example of which has been implemented in subroutine `stiff`, (b) generalized BULIRSCH–STOER methods[2] (extrapolation methods), see PRESS et al. (1992), and (c) predictor-corrector methods.

The ROSENBROCK methods are close to the embedded RUNGE–KUTTA–FEHLBERG integrator introduced in Section 2.4.2. They are robust and perform well for accuracies of $\epsilon \approx 10^{-4} \ldots 10^{-5}$ and for systems of up to approximately 10 ODEs. For larger systems or higher accuracies, the more complicated alternatives mentioned above are preferable.

A ROSENBROCK method seeks a solution of the form

$$\boldsymbol{y}(x_0 + h) = \boldsymbol{y}_0 + \sum_{i=1}^{s} c_i \boldsymbol{k}_i, \tag{2.90}$$

where the corrections $\boldsymbol{k}_i$ are found by solving s linear equations that generalize the structure in Eq. (2.89):

$$\left(\mathbf{1} - \gamma h \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{y}}\right) \cdot \boldsymbol{k}_i = h\boldsymbol{f}\left(\boldsymbol{y}_0 + \sum_{j=1}^{i-1} \alpha_{ij} \boldsymbol{k}_j\right) + h \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{y}} \cdot \sum_{j=1}^{i-1} \gamma_{ij} \boldsymbol{k}_j, \qquad i = 1, \ldots, s. \tag{2.91}$$

The coefficients (γ, c_i, α_{ij}, γ_{ij}) are fixed constants independent of the problem. For $\gamma = \gamma_{ij} = 0$ the scheme reverts to an explicit RK-scheme. Eq. (2.91) can be solved successively for $\boldsymbol{k}_i$. Again, as above, we are interested in the adaptive step size control, and – again again – there exists an *embedded* scheme which returns solutions accurate to 4th and 3rd order (compared to 5th and 4th order previously) by using the same function evaluations with different coefficients.

The details of the scheme are of lesser interest here, however, one point should be raised. As before, the choice of the "correct" (i.e., most appropriate) error criterion is crucial. This even more so here, because of the stiff nature of the ODEs, since they often have pieces of the solution that decay strongly and aren't of interest, so it wouldn't make any sense to assign them the same accuracy (and spend lots of time on that) as to the "relevant" solution parts. Obviously, these choices depend on the problem. One might control the relative error above some threshold $\boldsymbol{C}$ and the absolute error below the threshold by setting (cf. Section 2.4.3)

$$\boldsymbol{y}_{scal} = \max(\boldsymbol{C}, |\boldsymbol{y}|). \tag{2.92}$$

If the variables are properly non-dimensionalized, the components of $\boldsymbol{C}$ should be of order unity, which should be used as a default value as well.

[2] Conventional BULIRSCH–STOER integrators are very useful for integrating non-stiff problems, which have a rather smooth function $f(x, y)$. The method is similar to ROMBERG integration, i.e., involves an extrapolation to step size zero.

Chapter 3

Physics – Cosmological Models

In this chapter, we will give a brief introduction into the derivation and some simple applications of cosmological models. In particular, we will consider the Friedmann–Lemaître models which describe (in form of an initial value problem) the temporal evolution of the so-called *cosmic scale*, $R(t)$. Solutions for the behaviour of $R(t)$, in dependence of various energy density terms (matter, radiation, "vacuum"/dark energy), will be obtained by numerical methods on the second day of our numerical lab work. There is a vast number of literature covering the corresponding physics (cited, e.g., in Wilms05), where, to some extent, we will follow the text book by Roos (2003). Note that we have included a useful manuscript into the course material provided, following a lecture given by J. Wilms (University Tübingen 2005, http://astro.uni-tuebingen.de/~wilms/teach/cosmo/index.html). This manuscript will be cited as "Wilms05" in the following.[1]

3.1. Cosmological redshift and Hubble's law

In 1929, E. Hubble discovered that the spectral lines emitted from various galaxies of well-known distances are systematically redshifted, by wavelength shifts

$$\frac{\lambda_{\mathrm{obs}} - \lambda_{\mathrm{emit}}}{\lambda_{\mathrm{emit}}} = \frac{\Delta\lambda}{\lambda} =: z \qquad \text{or} \qquad \lambda_{\mathrm{obs}} = \lambda_{\mathrm{emit}}(1+z). \tag{3.1}$$

Interpreted in terms of a Doppler shift as a *recession* velocity,

$$\frac{v}{c} = \frac{\Delta\lambda}{\lambda} = z, \tag{3.2}$$

and combining his measurements with the distances of the line-emitting galaxies, Hubble suggested these redshifts to be a linear function of distance,

$$c\,z = v = H_0 r. \tag{3.3}$$

This relation is meanwhile called Hubble's law, with Hubble parameter[2] H_0, where the subscript "0" refers to our present time, t_0. From the relatively close galaxies studied by Hubble, he could only determine a linear relation, though higher order terms in r cannot be excluded and

[1] Be prepared for a couple of typos in this script, which have (hopefully) been corrected here.

[2] The alternative designation as "Hubble's constant" is somewhat misleading, since the quantity itself is changing with time.

are present indeed (see Eq. 3.86). The message of this law is that the Universe is *expanding*, and from the so-called *cosmological principle* (i.e., the Universe is assumed to be homogeneous and isotropic on large scales), it can be shown that observers located at different positions would always measure such a law: independent of location, astronomical objects recede from the observer at the same rate. Thus, a homogeneous and isotropic universe does not have a center.

With respect to redshift alone, the HUBBLE law reads

$$z = H_0 \frac{r}{c} \tag{3.4}$$

and the inverse of H_0 has the dimensions of a time, which gives a characteristic timescale for the expansion of the Universe (though not necessarily its actual age), called the HUBBLE time,

$$\tau_{\mathrm{H}} = H_0^{-1} = h^{-1} \times 9.78 \cdot 10^9 \text{ yr}, \tag{3.5}$$

where h is a dimensionless quantity, conveniently defined as

$$h = H_0/(100 \text{ km s}^{-1} \text{ Mpc}^{-1}). \tag{3.6}$$

Present measurements can restrict h quite precisely, $h \approx 0.73 \pm 0.05$, though there is a certain possibility that its value might be smaller by roughly 15%.

Though the actual size of our Universe is "unmeasurable", one usually describes distances at different epochs (which will change due to the expansion or contraction) in terms of a cosmic scale $R(t)$ (see below), and its present value is denoted by $R_0 = R(t_0)$. Additionally, one can normalize $R(t)$ by its present value and define a *cosmic scale factor*,

$$a(t) := R(t)/R_0. \tag{3.7}$$

This scale factor affects *all* distances, in particular also the wavelength λ_{emit} emitted at time t and observed as λ_{obs} at t_0,

$$\frac{\lambda_{\mathrm{emit}}}{R(t)} = \frac{\lambda_{\mathrm{obs}}}{R_0}. \tag{3.8}$$

Indeed, this relation can be simply derived from the ROBERTSON–WALKER metric (see Section 3.3), accounting for the fact that for photons $\mathrm{d}s^2 = 0$ and that *comoving* distances (in this case, those traveled by photons) remain preserved in such a metric (see WILMS05, p. 4–19).

By a TAYLOR expansion of $a(t)$ for $t < t_0$, we find, to first order

$$a(t) = a(t_0) - \left.\frac{\partial a}{\partial t}\right|_{t_0} (t_0 - t) = 1 - \dot{a}_0 \, (t_0 - t). \tag{3.9}$$

With source distance $r = c\,(t_0 - t)$ and

$$\frac{\lambda_{\mathrm{obs}}}{\lambda_{\mathrm{emit}}} = \frac{R_0}{R(t)} = a^{-1}, \tag{3.10}$$

the redshift can be expressed as

$$z = \frac{\lambda_{\mathrm{obs}}}{\lambda_{\mathrm{emit}}} - 1 = a^{-1} - 1 \approx \frac{1}{1 - \dot{a}_0 \, (t_0 - t)} - 1 \approx \dot{a}_0 \, (t_0 - t) = \dot{a}_0 \, \frac{r}{c}. \tag{3.11}$$

Thus, both redshift and HUBBLE parameter are related to the cosmic scale factor,

$$1+z=\frac{\lambda_{\rm obs}}{\lambda_{\rm emit}}=\frac{1}{a(t)}, \tag{3.12}$$

$$\dot{a}_0=\dot{a}(t_0)=\frac{\dot{R}_0}{R_0}=\frac{\dot{R}(t_0)}{R(t_0)}=H_0, \tag{3.13}$$

i.e., H_0 is nothing else than the present rate of change in this factor (since we know that H_0 is positive, we know that we have an *expanding* universe *at present*).

Remarks
(i) From the above, it should be clear that the cosmological redshift is a consequence of the expansion of the Universe and not of the velocities of the receding objects. There are, of course, such kinematic effects as well (e.g., peculiar velocities resulting from (gravitationally induced) flows on smaller scales such as the Virgo-centric flow), which have to be corrected for when measuring the cosmological redshift.
(ii) For distant objects (of the order of the HUBBLE radius $r_{\rm H}=c/H_0\approx 3000/h\,$Mpc, i.e., for objects which would recede from us with the speed of light according to the linear HUBBLE law), this law has to be modified for relativistic effects. Indeed, it turns out that the redshift from objects located at $r_{\rm H}$ becomes infinite, i.e., we cannot obtain information from larger distances.

3.2. NEWTONIAN expansion

One of the key questions in cosmology is whether the Universe as a whole is a gravitationally *bound* system in which the expansion will be halted one day. A simple model using NEWTONIAN mechanics will give a first answer. Note already here that the following results can be derived from General Relativity (GR) as well, in the limit of weak gravitational fields.

Consider a galaxy of gravitating mass m_g located at distance r from the center of a sphere of mean mass density ρ. The total mass of the sphere is

$$M=\frac{4\pi}{3}r^3\rho, \tag{3.14}$$

so that the gravitational potential of the galaxy is

$$U=-\frac{GMm_g}{r}=-\frac{4\pi}{3}Gm_g\rho r^2, \tag{3.15}$$

with the gravitational constant G. Thus, the acceleration of the galaxy towards the center of the sphere is given by

$$\ddot{r}=-\frac{GM}{r^2}=-\frac{4\pi}{3}G\rho r, \tag{3.16}$$

which is nothing else than NEWTON's law. In a Universe expanding according to HUBBLE's law, the galaxy has a kinetic energy of

$$T=\frac{1}{2}mv^2=\frac{1}{2}m(H_0r)^2, \tag{3.17}$$

with *inertial* mass m. Using the equivalence principle (inertial mass = gravitating mass), $m=m_g$, the total energy of the galaxy is

$$E=T+U=\frac{1}{2}m(H_0r)^2-\frac{4\pi}{3}Gm_g\rho r^2=mr^2\left(\frac{1}{2}H_0^2-\frac{4\pi}{3}G\rho\right), \tag{3.18}$$

which immediately tells that the expansion will come to a halt ($E \leq 0$) if the mass density inside the sphere (i.e., the mean density of the *Universe*), ρ, is larger than or equal to the *critical density*,

$$\rho_c = \frac{3H_0^2}{8\pi G}. \tag{3.19}$$

If $\rho > \rho_c$, we speak of a *closed*, otherwise of an *open* Universe. Note that ρ_c is the *present* critical density, corresponding to the present HUBBLE parameter.

Exercise 4: *Calculate the critical density, in units of* $\mathrm{g/cm^3}$ *and with respect to* h.

Since distance and density are time-dependent, they change with the expansion. Denoting their present ($t = t_0$) values by the subscript "0", we have

$$r(t) = a(t) \cdot r_0; \qquad \rho(t) = \rho_0 / a^3(t) \tag{3.20}$$

(mass conservation), and NEWTON's law (3.16) yields

$$\ddot{a} = \frac{\ddot{r}}{r_0} = -\frac{4\pi}{3} G \frac{r}{r_0} \frac{\rho}{a^3} = -\frac{4\pi}{3} G \rho_0 a^{-2}. \tag{3.21}$$

Multiplying this equation on both sides with $2\dot{a}$,

$$2\dot{a}\ddot{a} = -\frac{8\pi}{3} G \rho_0 \frac{\dot{a}}{a^2}, \tag{3.22}$$

this is equivalent to

$$\frac{\mathrm{d}}{\mathrm{d}a}\dot{a}^2 = \frac{8\pi}{3} G \rho_0 \frac{\mathrm{d}}{\mathrm{d}a}\left(\frac{1}{a}\right), \tag{3.23}$$

which can be easily integrated from t_0 to t (with $a_0 = 1$)

$$\dot{a}^2(t) - \dot{a}^2(t_0) = \frac{8\pi}{3} G \rho_0 \left(\frac{1}{a(t)} - 1\right). \tag{3.24}$$

By introducing the density parameter

$$\Omega_0 = \frac{\rho_0}{\rho_c} = \frac{8\pi G \rho_0}{3H_0^2} \tag{3.25}$$

(in this scenario, $\Omega_0 = 1$ would denote a Universe at critical density), we obtain

$$\dot{a}^2 = H_0^2 \Omega_0 \left(\frac{1}{a} - 1\right) + \dot{a}^2(t_0) \tag{3.26}$$

and, using the definition for $H_0 = \dot{a}(t_0)$,

$$\frac{\dot{a}^2}{H_0^2} = \left(1 - \Omega_0 + \frac{\Omega_0}{a}\right), \tag{3.27}$$

which is identical with the (first) FRIEDMANN equation for similar conditions (only matter, no radiation, no cosmological constant).

An "empty" universe ($\rho_0 = 0 = \Omega_0$) would expand forever, with constant rate $\dot{a} = H_0$. On the other hand, a steady state universe would imply $H_0 = 0$.

Since $\dot{a}^2 \geq 0$ always, we must have

$$1 - \Omega_0 + \Omega_0 / a \geq 0$$

as well. This implies

a) for $\Omega_0 < 1$, that the universe would be an open, ever-expanding one, since $1-\Omega_0+\Omega_0/a > 0$.

b) for $\Omega_0 = 1$, still an ever-expanding universe, where the expansion rate asymptotically reaches $\dot{a} \to 0$.

c) for $\Omega_0 > 1$, a closed universe, where after reaching a certain maximum in size ($a = a_{\max}$), a must *decrease* (i.e., $\dot{a} < 0$) in order to keep the total expression ≥ 0 always.

3.3. Robertson–Walker metric

A suitable metric describing a curved "three-surface" in Euclidean four-space which is consistent with the cosmological principle (e.g., the spatial part is spherically symmetric) has been introduced 1934 by Robertson & Walker, and can be represented by

$$\mathrm{d}s^2 = c^2\mathrm{d}t^2 - R^2(t)\left(\frac{\mathrm{d}r^2}{1-kr^2} + r^2\mathrm{d}\theta^2 + r^2\sin^2\theta\mathrm{d}\phi^2\right) = g_{\mu\nu}\mathrm{d}x^\mu\mathrm{d}x^\nu. \tag{3.28}$$

Note that this metric depends "only" on three spatial coordinates, because the fourth coordinate is irrelevant (fortunately!) since we are considering "three-surfaces" (notably, the space we inhabit is 3-dimensional) where the fourth coordinate can be expressed by the other three coordinates and a certain constraint.

Before discussing the above metric in some detail, let us consider some analogy which is more common to people who cannot think four-dimensionally (e.g., the author of this manual), namely a "two-surface" in a Euclidean three-space. In particular, we consider the surface of a 3D-sphere, a so-called "two-sphere", where the third spatial coordinate can be expressed in terms of the sphere radius (the constraint from above) and the other two coordinates, ($x_3^2 = R^2 - x_1^2 - x_2^2$). The corresponding length element can be written as

$$\mathrm{d}l^2 = \frac{R^2\,\mathrm{d}r'^2}{R^2 - r'^2} + r'^2\mathrm{d}\theta^2, \tag{3.29}$$

if we use polar coordinates r', θ in the x_3 plane (see Wilms05, p. 4–6). Expressing the radial coordinate in units of sphere-radius, $r = r'/R$, we obtain

$$\mathrm{d}l^2 = R^2\left(\frac{\mathrm{d}r^2}{1-r^2} + r^2\mathrm{d}\theta^2\right). \tag{3.30}$$

Likewise, the metric for a hyperbolic plane (with $x_3^2 = R^2 + x_1^2 + x_2^2$) is identical to the above one, if we exchange the minus sign in the denominator by a plus sign,

$$\mathrm{d}l^2 = R^2\left(\frac{\mathrm{d}r^2}{1+r^2} + r^2\mathrm{d}\theta^2\right). \tag{3.31}$$

Finally, the length element for a conventional plane can be written as

$$\mathrm{d}l^2 = R^2\left(\mathrm{d}r^2 + r^2\mathrm{d}\theta^2\right), \tag{3.32}$$

where R is now an arbitrary scale factor. Summarizing, the length element for all three cases can be written as

$$\mathrm{d}l^2 = R^2\left(\frac{\mathrm{d}r^2}{1-kr^2} + r^2\mathrm{d}\theta^2\right), \tag{3.33}$$

where $k \in \{1, 0, -1\}$ corresponds to a two-sphere, a plane, and a hyperbolic plane, respectively.

In so far, the generalization to three-dimensional (hyper-)surfaces on four-dimensional spheres, i.e., three-spheres, flat three-space, and three-hyperboloids is (almost) straightforward (though the author still has a problem imagining the first and the last case). By allowing the generalized radius/scale factor to become time-dependent and by including the time coordinate into the metric (which is already required in special relativity), we finally obtain the ROBERTSON–WALKER metric, Eq. 3.28.

Comparing the length element in the RW-metric with the corresponding tensor formulation, the components of $g_{\mu\nu}$ are given by (with $x^0 = ct$)

$$g_{00} = 1, \qquad g_{11} = -\frac{R^2}{1-kr^2}, \qquad g_{22} = -R^2 r^2, \qquad g_{33} = -R^2 r^2 \sin^2\theta, \tag{3.34}$$

where $k \in \{1, 0, -1\}$ is called the curvature parameter and corresponds to the three geometries outlined above.

If the Universe is homogeneous and isotropic at a given time and follows the RW-metric, it will always retain these features: a galaxy at coordinates (r, θ, ϕ) will always remain at these coordinates, only the scale $R(t)$ (i.e., the *scale* of distances) will change with time. Since the spatial displacement is $\mathrm{d}r = \mathrm{d}\theta = \mathrm{d}\phi = 0$, the metric equation reduces to $\mathrm{d}s^2 = c^2\mathrm{d}t^2$, and the corresponding frame is called the *comoving frame*. Distances in this frame ("comoving" distances, d, which depend only on the spatial coordinates) remain preserved under expansion, whereas "proper" distances, $D(t) = d \cdot R(t)$, change with time.

Summary: The above metric defines a universal coordinate system *tied* to the expansion of space, whereas the scale $R(t)$ describes its *evolution*.

3.4. FRIEDMANN cosmologies

To obtain a model for an Universe which follows the cosmological principle, we have to combine the RW metric with EINSTEIN's field equations,

$$G_{\mu\nu} = \frac{8\pi G}{c^4} T_{\mu\nu}, \tag{3.35}$$

where

$$G_{\mu\nu} = R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R$$

is the EINSTEIN tensor, derived from the RICCI tensor $R_{\mu\nu}$ and the RICCI scalar R (not to be confused with the cosmic scale!), where the RICCI tensor itself is a contraction of the RIEMANN tensor $R_{\alpha\beta\gamma\sigma}$,

$$R_{\beta\gamma} = R^{\alpha}_{\beta\gamma\alpha}, \qquad R = g^{\beta\gamma} R_{\beta\gamma},$$

and $T_{\mu\nu}$ is the energy–stress tensor. One of the problems in GR is the calculation of the RIEMANN tensor (4th rank, i.e., 256 components, which mostly and fortunately are not independent or vanish), which itself is a combination of the so-called affine connections (in some languages, "CHRISTOFFEL symbols") and their derivatives. The affine connections themselves (no tensors) are derivatives of the metric (here, we have the final relation to our work), and are required to allow for a *covariant* formulation of the physical laws (i.e., formulations in terms of quantities which remain invariant in arbitrary (e.g., accelerated) frames).

To proceed further, we consider a comoving observer in a space described by the RW-metric. Under the assumption of the cosmological principle, the energy–stress tensor becomes purely

diagonal, and the time-time and space-space components T_{00} and T_{11} (only these will be used in the following) can be written as

$$T_{00} = \rho c^2, \qquad T_{11} = \frac{pR^2(t)}{1-kr^2}, \tag{3.36}$$

where ρc^2 is the energy density of the considered "fluid", and p the corresponding pressure (for details, refer to the literature). The corresponding components of the EINSTEIN tensor are calculated from the RW metric ($\rightarrow$ affine connections $\rightarrow$ RIEMANN tensor $\rightarrow$ RICCI tensor/scalar $\rightarrow$ EINSTEIN tensor), with components

$$G_{00} = \frac{3}{c^2R^2}\left(\dot{R}^2 + kc^2\right), \tag{3.37}$$

$$G_{11} = -\frac{1}{c^2(1-kr^2)}\left(2R\ddot{R} + \dot{R}^2 + kc^2\right), \tag{3.38}$$

where, of course, R now denotes the cosmic scale. Using the field equations for the time-time and space-space components, we obtain

$$\begin{aligned} & \frac{\dot{R}^2 + kc^2}{R^2} = \frac{8\pi G}{c^4}\frac{c^2}{3}\rho c^2 \\ \Rightarrow \quad & \frac{\dot{R}^2 + kc^2}{R^2} = \frac{8\pi G}{3}\rho \qquad \text{"FRIEDMANN I"}, \end{aligned} \tag{3.39}$$

$$\begin{aligned} & 2R\ddot{R} + \dot{R}^2 + kc^2 = -\frac{8\pi G}{c^4}c^2 pR^2 \\ \Rightarrow \quad & \frac{2\ddot{R}}{R} + \frac{\dot{R}^2 + kc^2}{R^2} = -\frac{8\pi G}{c^2}p \qquad \text{"FRIEDMANN IIA"}. \end{aligned} \tag{3.40}$$

These equations have been firstly derived by FRIEDMANN in 1922 (and confirmed by an independent derivation by LEMAÎTRE in 1927), i.e., seven years before HUBBLE's detection of the cosmological expansion!!! At that time EINSTEIN had severe doubts in his own theory, because it did not allow for a *static* universe, as it is true for the FRIEDMANN equations as formulated above. (Already in 1917, EINSTEIN tried to "cure" this problem by introducing the cosmological constant, see below.)

Subtracting Eq. I from Eq. IIA results in the alternative formulation

$$\frac{2\ddot{R}}{R} = -\frac{8\pi G}{3c^2}\left(\rho c^2 + 3p\right) \qquad \text{"FRIEDMANN IIB"}. \tag{3.41}$$

Whereas Eq. I shows that the rate of expansion, $\dot{R}$, increases with increasing density, Eq. IIB shows that, because of the negative sign, the expansion actually decelerates, at least within the model discussed so far.

If we evaluate Eq. I at $t = t_0$, we obtain

$$\left(\frac{\dot{R}}{R}\right)_0^2 = \frac{8\pi G}{3}\rho_0 - \frac{kc^2}{R_0^2} \quad \rightarrow \quad H_0^2 = \Omega_0 H_0^2 - \frac{kc^2}{R_0^2} \quad \rightarrow \quad kc^2 = H_0^2 R_0^2(\Omega_0 - 1). \tag{3.42}$$

Thus, the curvature parameter k from the RW-metric implies, for $k \in \{1, 0, -1\}$, density parameters $\Omega_0 > 1$, $= 1$, and < 1, respectively. A spatially flat universe ($k = 0$) is called an EINSTEIN–DE SITTER universe. Rearranging the last equation, we can define an alternative curvature parameter

$$\Omega_\mathrm{K} := -\frac{kc^2}{H_0^2 R_0^2} = 1 - \Omega_0. \tag{3.43}$$

Let us identify, for the moment, the density in Eq. I with mass density alone, such that $\rho(t) = \rho_0/a^3(t)$, similar to our examination of the NEWTONIAN expansion (Section 3.2). Thus, for arbitrary t,

$$\left(\frac{\dot{R}}{R}\right)^2 = \frac{8\pi G}{3}\frac{\rho_0}{a^3(t)} - \frac{kc^2}{R^2}. \tag{3.44}$$

Multiplying by $(R/R_0)^2 = a^2(t)$, we obtain

$$\begin{aligned}
\dot{a}^2 &= \frac{8\pi G}{3}\frac{\rho_0}{a^3}a^2 - \frac{kc^2}{R_0^2} && (3.45)\\
&= H_0^2\Omega_0\frac{1}{a} - \frac{kc^2}{R_0^2} \\
&= H_0^2\Omega_0\frac{1}{a} + (1-\Omega_0)H_0^2 \qquad \text{(from Eq. 3.43)} \\
&= H_0^2\left(\frac{\Omega_0}{a} + 1 - \Omega_0\right), && (3.46)
\end{aligned}$$

which is identical to the result of the NEWTONIAN approach, Eq. 3.27. Another interesting relation follows from a somewhat different manipulation. Again, evaluate Eq. I at arbitrary t,

$$\begin{aligned}
kc^2 &= \frac{8\pi G}{3}\rho R^2 - \dot{R}^2 \\
&= R^2\left(\frac{8\pi G}{3}\rho_c(t)\Omega(t) - \left(\frac{\dot{R}}{R}\right)^2\right) \quad \text{with} \quad \rho_c(t) = \frac{3H^2(t)}{8\pi G}, \quad \Omega(t) = \frac{\rho(t)}{\rho_c(t)}, \quad H(t) = \frac{\dot{R}}{R} \\
&= R^2\left(H^2(t)\Omega(t) - H^2(t)\right) = R^2H^2(t)\big(\Omega(t) - 1\big). && (3.47)
\end{aligned}$$

Equating this expression with the corresponding one from Eq. 3.42 (both being equal to the constant term kc^2),

$$\begin{aligned}
R^2H^2(\Omega - 1) &= H_0^2R_0^2(\Omega_0 - 1) \\
a^2H^2(\Omega - 1) &= H_0^2(\Omega_0 - 1) \qquad \Rightarrow \\
\frac{\Omega - 1}{\Omega_0 - 1} &= \frac{H_0^2}{H^2a^2} = \left(\frac{\dot{R}_0}{\dot{R}}\right)^2 = \left(\frac{\dot{a}_0}{\dot{a}}\right)^2 \ll 1 \quad \text{for} \quad t/t_0 \ll 1, && (3.48)
\end{aligned}$$

since $\dot{a}$ tends towards infinity for small t, as we will see later. From this condition then, Ω must have been *very close* to unity, or, in other words, *the early Universe must have been asymptotically flat*! E.g., the maximum deviation from flatness during the phase of nucleosynthesis ($t \approx 1s$) can be constrained by $\lesssim 10^{-16}$, if the present day value of Ω_0 is of the order of unity.

This is what is called the *flatness* problem. Had Ω been different from unity at its beginning, the Universe would have immediately recollapsed (within one PLANCK time), or expanded too fast to allow for the existence of mankind. Thus, the anthropic point of view requires $\Omega = 1$, i.e., $k = 0$. To generate a universe surviving for many gigayears without Ω being exactly unity would have required an incredible fine-tuning, which is extremely unlikely. Without going into details, *inflation* can cure this problem, by increasing the cosmological scale exponentially (factor of $\approx 10^{43}$) in a (very) early phase ($t \approx 10^{-34}$ s) of the Universe (see below).

3.5. Cosmological constant

As we have already mentioned, EINSTEIN originally believed in a *static* universe with $\dot{R} = \ddot{R} = 0$, $R(t) = R_0$. In this case the FRIEDMANN equations would read

$$\frac{kc^2}{R_0^2} = \frac{8\pi G}{3}\rho_0 = -\frac{8\pi G}{c^2}p_0, \tag{3.49}$$

which implies that a) $k = 1$ in order to allow for a positive present mass density, and b) that the present pressure of matter, p_0, becomes negative then. EINSTEIN cured this problem by introducing a LORENTZ invariant term $g_{\mu\nu}\lambda$, where the "cosmological constant" λ provides a very tiny correction to the geometry of space-time:

$$G_{\mu\nu} := R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R - g_{\mu\nu}\lambda. \tag{3.50}$$

In analogy to Eq. 3.37 and the following ones, this correction implies

$$G_{00} \to G_{00}^{\text{old}} - \frac{\lambda}{c^2} = \frac{3}{c^2R^2}(\dot{R}^2 + kc^2) - 3\frac{\lambda}{3c^2}, \tag{3.51}$$

$$G_{11} \to G_{11}^{\text{old}} + \frac{\lambda}{c^2}\frac{R^2}{1-kr^2} = -\frac{1}{c^2(1-kr^2)}\left(2R\ddot{R} + \dot{R}^2 + kc^2\right) + \frac{\lambda R^2}{c^2(1-kr^2)}, \tag{3.52}$$

and the original FRIEDMANN equations become

$$\frac{\dot{R}^2 + kc^2}{R^2} - \frac{\lambda}{3} = \frac{8\pi G}{3}\rho, \tag{3.53}$$

$$\frac{\ddot{R}}{R} - \frac{\lambda}{3} = -\frac{4\pi G}{3c^2}(\rho c^2 + 3p) \qquad \text{(version B)}. \tag{3.54}$$

Since the pressure of matter is very small, we can approximate $p \approx 0$ (see Section 3.7), and from the second equation we obtain for a static universe

$$\rho_0 = \frac{\lambda}{4\pi G} = 2\rho_\lambda \tag{3.55}$$

if we denote the density corresponding to the λ term by $\rho_\lambda = \lambda/(8\pi G)$.[3] The first equation (with $\dot{R} = 0$) then yields

$$\frac{kc^2}{R^2} = 8\pi G\rho_\lambda \qquad \Rightarrow \qquad R^2 = \frac{kc^2}{\lambda}, \tag{3.56}$$

which, again, makes sense (requiring $\rho_0 > 0 \Rightarrow \lambda > 0$) only for $k = 1$. A positive λ curves space-time in such a way as to counteract gravity and to prevent a collapse!

Thus, a *static* universe with positive matter density has to be closed, and its mean density and "radius" depend exclusively on λ. Remember, however, that "in our case" a static Universe is incompatible with the observational fact that $H_0 = (\dot{R}/R)_0 \neq 0$! Another argument against static universes in general was given by EDDINGTON (1930) (you might have a look into this historical paper): if there is the slightest imbalance between ρ and λ (i.e., a disturbance), $\ddot{R}$ becomes non-zero, and the universe will begin to expand or contract. This leads to still larger deviations of ρ from λ, and so on. In other words, static universes are *unstable* (a fact not realized before), and after EDDINGTON's paper EINSTEIN abandoned his belief in this possibility and also withdrew the cosmological constant from his theory. Only in recent times was it resurrected again, due to additional observational facts.

[3] Previously called "*vacuum density*".

3.6. Friedmann–Lemaître cosmologies

If the physics of the vacuum is Lorentz invariant[4], i.e., looks the same to any inertial observer, its contribution to the energy–stress tensor is the same as Einstein's correction $\lambda g_{\mu\nu}$ to the geometry, as noted by Lemaître. The content of Eqs. 3.53 and 3.54 does not change if the corresponding terms are moved from the lhs to the rhs of the equations, though their *interpretation* changes. If put on the rhs, they appear as the density (first equation) and pressure (second equation) of an additional fluid (originally identified with the above *vacuum*), with density $\rho_\lambda = \frac{\lambda}{8\pi G}$ and pressure $p_\lambda = -\rho_\lambda c^2$:

$$\frac{\dot{R}^2 + kc^2}{R^2} = \frac{8\pi G}{3}(\rho + \rho_\lambda) \tag{3.57}$$

$$\frac{2\ddot{R}}{R} + \frac{\dot{R}^2 + kc^2}{R^2} = -\frac{8\pi G}{c^2}(p + p_\lambda) \qquad \text{(version A)} \tag{3.58}$$

$$\frac{\ddot{R}}{R} = -\frac{4\pi G}{3c^2}\Big((\rho + \rho_\lambda)c^2 + 3(p + p_\lambda)\Big) \qquad \text{(version B).} \tag{3.59}$$

In this interpretation then, for $\lambda > 0$ the gravitational effect of this fluid is to counteract, via its negative (!) pressure, the gravitational pull of "ordinary" matter, eventually even leading to an *acceleration* of the scale factor, whereas a negative λ would correspond to an additional attractive term.

Cosmologies as described by Eqs. 3.57 to 3.59 with positive λ are nowadays called Friedmann–Lemaître cosmologies. Since we have to deal with energy densities/pressures from different sources, the *total* density parameter is split into the different contributions by matter (including the dark one), radiation, and cosmological constant

$$\Omega_0 = \Omega_\mathrm{M} + \Omega_\mathrm{R} + \Omega_\Lambda \tag{3.60}$$

with Ω_R and Ω_Λ defined in analogy to Eq. 3.25, i.e.,

$$\Omega_\mathrm{R} = \frac{\rho_\mathrm{r}}{\rho_\mathrm{c}}, \qquad \Omega_\Lambda = \frac{\rho_\lambda}{\rho_\mathrm{c}} = \frac{\lambda}{8\pi G\rho_\mathrm{c}} = \frac{\lambda}{3H_0^2}. \tag{3.61}$$

As you will see (and hopefully confirm) during your lab work, the *present* values of Ω_M and Ω_Λ are quite similar, $\Omega_\mathrm{M} \approx 0.3$ and $\Omega_\Lambda \approx 0.7$. In other words, the density corresponding to the cosmological constant, ρ_λ, must be of similar order as the present critical density, or, more precisely, $\rho_\lambda \approx 0.7\rho_\mathrm{c} \approx 0.7\,h^2 \cdot 1.88 \times 10^{-29}$ g/cm^3 $\approx 6.5 \times 10^{-30}$ g/cm^3.

The vacuum energy problem. Using simple quantum-mechanical arguments, it can be shown that the *actual* vacuum energy of the Universe should be MUCH larger: remember that the ground-state energy of a quantum-mechanical oscillator is not zero by $E_0 = 1/2\hbar\omega$. The generalization to fields is (almost) straightforward. A relativistic field may be thought of as a collection of harmonic oscillators of all possible frequencies. A simple example is provided by a scalar field (e.g., a spinless boson) of mass m. For this system, the vacuum energy is simply a sum of contributions

$$E_0 = \sum_i \frac{1}{2}\hbar\omega_i, \tag{3.62}$$

[4] Assuming that the vacuum state of the Universe is not simply an empty space, but the ground state of some physical system, where this ground state should be independent of coordinate system.

where the sum extends over all possible modes in the field, i.e., over all wave vectors $\boldsymbol{k}$. The corresponding energy density $E_0/L^3 = \rho_{\rm vac}c^2$ can be calculated from integrating over all wavenumbers present in a box of Volume L^3 with periodic boundary conditions ($\lambda_n = 2\pi/k_n = L/n$, $n = 1, 2, \ldots$), and results in

$$\rho_{\rm vac}c^2 = \frac{\hbar c}{16\pi^2}k_{\rm max}^4, \tag{3.63}$$

where $k_{\rm max} \gg mc/\hbar$ is the maximum wavenumber present in the field. Following Carroll et al. (1992), we estimate $k_{\rm max}$ as the energy scale at which our confidence in the formalism no longer holds. For example, it is widely believed that the "Planck energy" $E_{\rm P}$ marks a point where conventional field theory breaks down due to quantum gravitational effects. This energy roughly corresponds to the situation when the wavelength of a particle reaches its corresponding Schwarzschild radius. More precisely, the Planck mass is defined as the mass of a particle for which the Schwarzschild radius is equal to the Compton wavelength divided by π,

$$\frac{1}{\pi}\frac{2\pi\hbar}{mc} = \frac{2Gm}{c^2} \qquad \Rightarrow \qquad m_{\rm P} = \sqrt{\frac{\hbar c}{G}}, \tag{3.64}$$

and the corresponding energy is the Planck mass times c^2. Choosing thus $\hbar k_{\rm max}c = E_{\rm P}$, we obtain

$$k_{\rm max} = \sqrt{\frac{c^3}{G\hbar}}, \tag{3.65}$$

$$\rho_{\rm vac} = \frac{c^5}{16\pi^2G^2\hbar} = 3.3 \times 10^{91}\ {\rm g/cm^3}. \tag{3.66}$$

Thus, the ratio of the expected $\rho_{\rm vac}$ to the "observed value" ρ_λ is $\mathcal{O}(10^{120})$, which indeed is not soooo small.

As in classical mechanics, the absolute value of the vacuum energy has no measurable effect in non-gravitational quantum field theories. Roughly spoken, non-gravitational forces depend on potential *gradients*, i.e., constant terms do not contribute. In GR, however, gravitation couples to *all* energies and momenta, which must include the energy of the vacuum: the only manifestation of vacuum energy will be through its gravitational influence. For a density as high as given by Eq. 3.66, this would mean a dramatic expansion of the Universe: the cosmic microwave background would have cooled below 3 K in the first 10^{-41} s after the Big Bang.

The maximum wavenumber corresponding to the "observed" value of ρ_λ would be $k_{\rm max} \approx 413\ {\rm cm}^{-1} \approx 0.008$ eV, which is way too low, since QM has been tested to be valid at much higher energies, and the reality of a vacuum energy density has been *quantitatively* verified by the Casimir effect.

In physical terms, then, the cosmological constant problem is this: there are independent contributions to the vacuum energy density from the virtual fluctuations of each field (since there is not only a single field but many more, due to the presence of different particle species) and from the potential energy of each field (and maybe even from a "real" cosmological constant itself). Each of these contributions should be much larger than ρ_λ, but they seem to combine to a very small value.[5] Thus, this situation indicates that new, unknown physics must play a decisive role, nowadays called "dark energy" (which usually has an energy density and equation of state which varies through space-time).

[5] At least at the *present* epoch, keeping in mind that the inflationary phase of the Universe requires a rather large value of λ in this very first phase.

3.7. Energy conservation and equation of state

Differentiating Eq. 3.57 with respect to time and combining the result with Eq. 3.54 to cancel the second derivatives, we obtain a new equation for the evolution of the energy density

$$\dot{\rho}c^2 + 3H(t)(\rho c^2 + p) = 0, \tag{3.67}$$

which is valid for both the total energy density and pressure *and* for their individual components, if different energy "forms" are present (following different equations of state). Eq. 3.67, which can be interpreted as a local energy conservation law, shows that the local energy density is changed by the expansion/contraction and by the corresponding volume work. This equation also clarifies the (at first glance somewhat puzzling) fact that a *negative* pressure – which in "normal life" is encountered as a pull – leads to a gravitational *repulsion*, i.e., increased expansion: the larger the ratio of pressure to energy density, the more volume work is done, which has to be compensated by a faster decrease in available gravitating energy density. The lower the energy density, however, the less the expansion ($\dot{R}^2 \propto (\rho R)^2$)! For negative pressure, on the other hand, the volume "work" becomes a *gain*. Loosely formulated, we create energy from the expansion of space, and the decrease in available gravitating energy density is slower than for $p > 0$ (and stops completely for $p = -\rho c^2$). Consequently the expansion remains faster than for positive pressure. Note that a decelerating universe becomes an accelerating one if the *dominating* energy densities/pressures relate via $p < -(1/3)\rho c^2$.

By integration of Eq. 3.67 we obtain

$$\int \frac{\dot{\rho}(t)}{\rho(t) + p(t)/c^2}\,\mathrm{d}t = -3\int \frac{\dot{a}(t)}{a(t)}\,\mathrm{d}t, \tag{3.68}$$

which can be immediately solved when the corresponding equation(s) of state is (are) known.

Equation of state (EOS). The most general form of an EOS in a space-time with RW metric can be shown to be

$$p = w\rho c^2, \tag{3.69}$$

where w depends on the energy form considered. Assuming w to be constant with time, Eq. 3.68 results in

$$\rho(a) \propto a^{-3(1+w)} = (1+z)^{3(1+w)}, \tag{3.70}$$

because $(1+z) = a^{-1}$, cf. Eq. 3.12.

(i) For "ordinary" matter (in the spirit of cosmology, i.e., non-relativistic cold matter: pressureless, non-radiating dust and cold dark matter), we have $p = 0$ (otherwise, galaxies would have random motions similar to that of gas particles under pressure which is not observed), and thus $w = 0$. Accordingly,

$$\rho_\mathrm{m}(a) = \rho_\mathrm{m}(0)a^{-3}. \tag{3.71}$$

(ii) Radiation or relativistic hot gas composed of elastically scattering particles follows

$$p_\mathrm{r} = \frac{1}{3}\varepsilon_\mathrm{r} = \frac{1}{3}\rho_\mathrm{r}c^2$$

(e.g., the ratio of 2nd moment of specific intensity, K, to mean intensity (0th moment), J, is 1/3 for isotropic radiation). Thus, $w = 1/3$, and

$$\rho_\mathrm{r}(a) = \rho_\mathrm{r}(0)a^{-4}. \tag{3.72}$$

(iii) As we have already seen, the pressure corresponding to a cosmological constant is $p_\lambda = -\rho_\lambda c^2$, i.e., $w = -1$ and

$$\rho_\lambda(a) = \rho_\lambda(0), \tag{3.73}$$

assuming that λ and w are constant over the considered expansion/contraction interval. This EOS can be derived from the energy–stress tensor of a perfect fluid required to be LORENTZ-invariant. Note that this follows also from the requirement that, under an adiabatic expansion/compression, $\rho_{\rm vac}$ should remain constant:

$$\mathrm{d}Q = \mathrm{d}U + p\,\mathrm{d}V = \rho_{\rm vac}c^2\,\mathrm{d}V + w\rho_{\rm vac}c^2\,\mathrm{d}V \stackrel{!}{=} 0 \qquad \Rightarrow \qquad w = -1.$$

3.8. Evolution of the scale factor

We are now (finally!) ready to write down the "equation of motion" of the Universe, i.e., the evolution of the cosmic scale factor, $a(t)$. By means of the above relations for the different densities, we obtain from the "first" FRIEDMANN–LEMAÎTRE equation (3.57)

$$\dot{R}^2 = \frac{8\pi G}{3}R^2\left(\frac{\rho_{\rm m}(0)}{a^3} + \frac{\rho_{\rm r}(0)}{a^4} + \rho_\lambda\right) - kc^2, \tag{3.74}$$

which can be easily rephrased in terms of the various density parameters,

$$\dot{a}^2 = a^2H_0^2\left(\frac{\Omega_{\rm M}}{a^3} + \frac{\Omega_{\rm R}}{a^4} + \Omega_\Lambda + \frac{\Omega_{\rm K}}{a^2}\right) = H_0^2\left(\frac{\Omega_{\rm M}}{a} + \frac{\Omega_{\rm R}}{a^2} + a^2\Omega_\Lambda + \Omega_{\rm K}\right), \tag{3.75}$$

again assuming λ to be constant. Convince yourself of the validity of this expression!

Remarks
(i) You might ask yourself why we have always considered the *two* FRIEDMANN(–LEMAÎTRE) equations throughout this chapter, when we finally use only the first one. Remember, however, that the 2nd equation has been used as well (together with the different equations of state) in order to calculate the various densities as a function of a, which otherwise would have remained unspecified!
(ii) Note also that the density parameters in Eq. 3.75 are defined with respect to the *present* critical density (i.e., with respect to H_0). The variation of the density parameters themselves can be calculated from $\Omega(t) = \rho(t)/\rho_{\rm c}(t)$ and results in

$$\Omega_{\rm M}(t) = \frac{H_0^2}{H^2(t)}\frac{\Omega_{\rm M}}{a^3}, \qquad \Omega_{\rm R}(t) = \frac{H_0^2}{H^2(t)}\frac{\Omega_{\rm R}}{a^4}, \qquad \Omega_\Lambda(t) = \frac{H_0^2}{H^2(t)}\Omega_\Lambda, \qquad \Omega_{\rm K}(t) = \frac{H_0^2}{H^2(t)}\frac{\Omega_{\rm K}}{a^2}.$$

Recall that the sum of the first three parameters must have been VERY close to unity for very small t. If λ indeed had been constant, the corresponding density parameter would have been negligible at very early times, and radiation would have dominated throughout the very first epochs.

At $t = t_0$, $\dot{a} = H_0$ and $a = 1$ per definition, such that

$$H_0^2 = H_0^2\left(\Omega_{\rm M} + \Omega_{\rm R} + \Omega_\Lambda + \Omega_{\rm K}\right) \qquad \text{or} \qquad \Omega_{\rm K} = 1 - \left(\Omega_{\rm M} + \Omega_{\rm R} + \Omega_\Lambda\right) = 1 - \Omega_0,$$

cf. Eq. 3.43. Thus, the evolution of the HUBBLE parameter follows

$$\left(\frac{\dot{a}}{a}\right)^2 = H(t) = H_0^2\left(\frac{1-\Omega_0}{a^2} + \frac{\Omega_{\rm M}}{a^3} + \frac{\Omega_{\rm R}}{a^4} + \Omega_\Lambda\right). \tag{3.76}$$

During your lab work, you will solve this equation in the form

$$\left(\frac{\dot{a}}{H_0}\right) = \pm\left(1 - \Omega_0 + \frac{\Omega_\mathrm{M}}{a} + \frac{\Omega_\mathrm{R}}{a^2} + a^2\Omega_\Lambda\right)^{1/2}. \tag{3.77}$$

By integration, we find the time elapsed after the big-bang (requiring $a(0) = 0$),

$$t(z) = \frac{1}{H_0}\int_0^{\frac{1}{1+z}} \mathrm{d}a \left(1 - \Omega_0 + \frac{\Omega_\mathrm{M}}{a} + \frac{\Omega_\mathrm{R}}{a^2} + a^2\Omega_\Lambda\right)^{-1/2}, \tag{3.78}$$

$$t'(a) = \frac{t(a)}{\tau_\mathrm{H}} = H_0\, t(a) = \int_0^{a} \mathrm{d}a \left(1 - \Omega_0 + \frac{\Omega_\mathrm{M}}{a} + \frac{\Omega_\mathrm{R}}{a^2} + a^2\Omega_\Lambda\right)^{-1/2}, \tag{3.79}$$

if we denote by t' the time measured in units of the HUBBLE time $\tau_\mathrm{H} = 1/H_0$, and neglect the inflationary phase. Otherwise, we would have to account for the fact that $\lambda = \lambda(t)$, where λ is considerably larger during inflation than nowadays.

Exercise 5: *Calculate* $t(z = 0)$ (a) *for a flat universe consisting only of matter and* (b) *for an empty universe without vacuum energy.*

Inflation. For the physics of inflation ("inflaton field"), we refer to textbooks. Let us note here only the basic assumption of inflation: during a very early phase in the universe[6], the λ-term dominated the other ones in the FRIEDMANN–LEMAÎTRE equations (i.e., it was much larger than now). Thus,

$$H^2(t) = \left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G(\rho_\mathrm{m} + \rho_\mathrm{r})}{3} - \frac{kc^2}{R^2} + \frac{\lambda}{3} \approx \frac{\lambda}{3}, \tag{3.80}$$

which results, via

$$H(t) = \sqrt{\frac{\lambda}{3}} = \text{const}, \qquad a(t) \propto e^{Ht}, \tag{3.81}$$

in an exponential increase of the scale factor, which then also explains the flatness problem, since

$$\Omega(t) - 1 = -\Omega_\mathrm{K}(t) \propto \frac{1}{a^2} \propto e^{-2Ht}. \tag{3.82}$$

If the inflation lasted for $100\,\tau_\mathrm{H}$, the scale factor increased by $e^{100} \approx 10^{43}$, and $\Omega - 1$ decreased by 10^{-86}, i.e., the Universe became *very* flat indeed, independent of the actual initial conditions!

Remarks
(i) A universe with $\Omega_\Lambda > 0$ but $\Omega_\mathrm{M} = \Omega_\mathrm{R} = 0$ (i.e., a universe where the above relations are valid for all times) is called a DE SITTER universe.
(ii) Due to the different dependencies of the various density parameters on a, the Ω_Λ term, if present, will almost always[7] dominate at later times. Thus, in most cases the final fate of a universe with $\lambda > 0$ is an exponential expansion, sometimes called a *second* inflationary phase.

[6] Starting at $t \approx 10^{-34}$ s and lasting for roughly 100 HUBBLE times ($\tau_\mathrm{H} = 1/H \approx 10^{-34}$ s).
[7] An exception is given for large Ω_M and $\Omega_\Lambda \lesssim 0.2$, as you will see during your lab work.

Dark energy. An empirical approach to remove some of the problems with the cosmological constant is to allow for a time-dependence, $\lambda = \lambda(t)$. The corresponding energy component is called "dark energy" then. The initial conditions require a value corresponding to the expected vacuum density, $\rho_\lambda(t_\mathrm{P}) \approx 10^{120}\rho_\lambda(t_0)$ at PLANCK time $t_\mathrm{P} = 10^{-47}$ s, from which it decays to its present value, via matching the inflaton field as well. In this case, the EOS becomes (cf. Eq. 3.70)

$$p_\lambda = -\rho_\lambda \left(1 + \frac{1}{3}\frac{\mathrm{d}\ln\rho_\lambda(a)}{\mathrm{d}\ln a}\right), \tag{3.83}$$

which recovers the limit $w = -1$ for slowly changing ρ_λ. Of course, this approach does not tell us anything about the physics responsible for the effect. Note that enormous effort is presently spent to obtain observational constraints on this EOS, which might allow discriminating between competing physical models or might even trigger setting up new ones.

A popular model for dark energy is the so-called *quintessence*. Quintessence is a scalar field which has an equation of state with $w < -1/3$. As postulated above, quintessence is dynamic, and generally has a density and equation of state that varies through space-time.

Many models of quintessence have a so-called *tracker* behaviour, which partly solves the cosmological constant problem. In these models, the quintessence field has a density which closely tracks (but is less than) the radiation density until matter–radiation equality, which triggers quintessence to start having characteristics similar to the observed ρ_λ, eventually dominating the universe.

The deceleration parameter. In the beginning of this chapter, we derived HUBBLE's law from a *linear* expansion of the cosmic scale factor. Being familiar now with the FRIEDMANN–LEMAÎTRE equations, we are also ready to perform a second order expansion, which will tell us about the *present* acceleration or deceleration of the Universe:

$$\frac{R(t)}{R_0} = a(t) \approx a(t_0) + H_0(t-t_0) + \frac{1}{2}\left.\frac{\ddot{R}(t)}{R_0}\right|_{t_0}(t-t_0)^2, \tag{3.84}$$

where $a(t_0) = 1$ and $\ddot{R}$ can be extracted from the 2nd FRIEDMANN–LEMAÎTRE equation. With $p_\mathrm{M} = 0$ and neglecting ρ_λ and p_λ for the moment, we find

$$\left.\frac{\ddot{R}(t)}{R(t)}\right|_{t_0} = -\frac{4\pi G}{3}\rho_0 = -\frac{\Omega_0}{2}H_0^2, \tag{3.85}$$

$$a(t) = 1 + H_0(t-t_0) - \frac{1}{2}q_0 H_0^2(t-t_0)^2 \tag{3.86}$$

to second order. In this expansion, the term

$$q_0 := \frac{\Omega_0}{2} = -\frac{\ddot{R}(t_0)}{R_0 H_0^2} = -\frac{\ddot{R}(t_0)R_0}{\dot{R}^2(t_0)}$$

measures the *deceleration* of the Universe (i.e., is positive for a decelerating cosmic scale).

Exercise 6: *Generalize the above approach for $\Omega_\Lambda \neq 0$ but $\Omega_\mathrm{R} = 0$ (which is legitimate at the present epoch). What can be concluded from a negative q_0 as derived, e.g., from the distances of SNe Ia?*

Chapter 4

Experiment

4.1. Numerical solution of ODEs: Test problems and integrators

The problems covered in this section should be solved on the first day of your lab work.

4.1.1. The programs

Sources

Everything you need to edit, compile, and run the codes (as well as to create the plots) can be found in (sub-)directory `codes`. Switch to this directory, and have a look into the various subdirectories. All source files (FORTRAN 90, though with the ending `.F`) can be found in directory `src`. A short structure of the program is given in `frame.F`.

In order not to (unintentionally) destroy or corrupt the original files, you will do most of the work in directory `cmp`. Switch to this directory, and have a look into the only file which should reside there, the `Makefile`. Check the various options, and install the source files in `cmp`.

Exercise 7: *Locate all the sources and draw a diagram, specifying briefly what each subroutine does, down to the level of the integrators.*

Compilation, execution, and plotting

Compile all files using `make` (with the appropriate target).

For all your following work, copy the resulting executable into directory `run/problem`N, where N is a number $1 \ldots 4$ corresponding to the specific test problem. Perform all runs (which usually requires updating the particular input files) in this directory. Plots using IDL have to be finally created in the directory `plots`, by using the plotting routines to be found there.

4.1.2. Problem 1 – A first test

We will check whether the integrators are running smoothly by analytically and numerically solving the ODE

$$\partial_x y(x) = a\, y(x) \tag{4.1}$$

with $a = 2.5$ and the initial condition $y(0) = 10^{-3}$ on the interval $x = [0, 10]$.

Exercise 8: *Solve Eq. (4.1) analytically.* `:-)`

Exercise P1: The integrators

All steps which are detailed in this exercise have to be performed analogously when working on the other exercises.

P1.1: *Obviously, something's wrong with* `eulerstep.F`*: there's nothing in there. That's your chance to write your version of an* EULER *integrator. Add the required statements to the corresponding subroutine in* `cmp`*, compile the program, and copy the executable to* `run/problem1`*. Switch to the latter directory and test your integrator using 100 steps. Before running the code, check the corresponding input file* `frame.inp11` *with respect to the various input quantities. Which ones are* not *needed to perform the* EULER *method?*

Update the variables which are *needed and copy* `frame.inp11` *to* `frame.inp`*, which is the file being used by the executable. Hopefully, the program runs. Now inspect the created output file and check whether it makes sense!!!*

Switch to directory `plots`*. At first, study the first part of* `plotode.pro`*, regarding input, possible keywords, and the variables* `run`N *and* `label`N*, where* $N = 1 \ldots 4$*. Plot the result of your first test and the analytical solution with* `plotode.pro` *with options* `/solu` *and* `/points`*. Does the result agree with the analytical solution? Increase the number of steps to 1000.* ***Don't forget to change the label for the output file to*** `"00p1i2"`***!*** *A correct label is important for the plotting routine to work! Gotten any better? (Plot the results of both simulations into one figure, and compare the results.) What are the mean relative errors?*

After you are satisfied with your results, always create a corresponding Postscript file for your lab report (see `plotode.pro` *for the relevant arguments). You may also find it helpful to print out a hardcopy.*

P1.2: *Now use the 4th order RK integrator with 1000 steps. Again, choose a different label (*`"00p1i3"`*) to prevent overwriting of the previous result. Compare the result with the corresponding* EULER *and analytic solution (one plot, again with* `/points`*).*

P1.3: *Repeat P1.2 with* `rkqs` *and* `stiff`*, labels:* `"00p1i4"` *and* `"00p1i5"`*. Convince yourself that the quantities* `eps` *etc. in* `frame.inp` *make sense. How many integration steps does each of the integrators need? Compare the requested accuracy with the achieved accuracy. Are the more "fancy" integrators more accurate? What is their strength in this problem?*

P1.4: *Now use the RK integrator with adaptive step size to find out the number of steps it needs to actually obtain the same accuracy as* `rk4` *with 1000 steps. To this end, change the input parameter* `eps` *until you obtain the desired accuracy. Label:* `"00p1i6"`*.*

P1.5: *Summarize the results you obtained for Problem 1. Which integrator would you trust most, and why?*

4.1.3. Stiff Equations

As discussed above, physical systems which can be described by ODEs often evolve on very disparate scales. Such a set of ODEs is called stiff. In the following, we will test the integrators on two different sets of stiff equations.

Problem 2 – Two coupled ODEs

We begin with the set

$$\partial_x y_1 = \quad 998 y_1 + 1998 y_2, \tag{4.2}$$

$$\partial_x y_2 = -999 y_1 - 1999 y_2. \tag{4.3}$$

This corresponds to `iprob = 2` in `frame.inp`. The initial conditions are $y_1(0) = 1$ and $y_2(0) = 0$, and the ODEs are to be integrated between $x = [0, 4]$.

ADVANCED Exercise 9: *Solve the above problem analytically. Remember the solution ansatz for such homogeneous DEs with constant coefficients:* $\boldsymbol{y} = \boldsymbol{v}\exp(\lambda x)$. *Inserting this ansatz into the DE, you should be able to derive the general solution, and, using the initial condition, the particular one. (Hint: as an intermediate result, you should find the eigenvalues of* $\boldsymbol{A}$ *as* $\lambda_{1,2} = (-1, -1000)$.*)*

Exercise P2: Integrator stability

For this exercise, use directory `run/problem2` *as a working directory.*

P2.1: *Implement Eqs. 4.2 and 4.3 into file* `ododef.F`*, which contains the four subroutines* `setproblem`*,* `initproblem`*,* `derivs`*, and* `jacobian`*. Test all integrators (for the fixed-step integrators, use 10000 steps, for the others an accuracy of* 10^{-5}*), and plot the results with* `plotode`*, with option* `/solu`*. Is there any difference between the results? Plot and compare the relative errors as well (logarithmically, to better display the different scales). How many steps do* `rkqs` *and* `stiff` *need?*

P2.2: *Test the stability conditions derived in Section 2.5, both for the* `euler` *integrator (label* `"00p2i5"`*) and for* `rk4` *(label* `"00p2i6"`*). Which maximum step size is "allowed" in each case? (Eigenvalues provided in exercise 9.) How does this transform into step number? Check the solution with the corresponding (minimum) step numbers, particularly whether the numerical results decay to the analytic solution. Plot the corresponding numerical solutions in comparison to the exact one.*

Reduce the step number successively until the behavior becomes unstable. For comparison, also increase the step number to see how the behavior changes.

Problem 3 – Three coupled ODEs

For this exercise, use directory `run/problem3` *as a working directory. Labels are* `"00p3i1"` *and* `"00p3i2"`*.*

Exercise P3: *Now switch to problem 3 (*`iprob = 3`*). Find its definition in* `ododef.F`*, write down the corresponding set of ODEs and discuss why this is a difficult problem to integrate. Run integrators* `rkqs` *and* `stiff` *over the interval* $[0, 50]$ *and plot the results. Do the results agree? How many iterations does each of the integrators need? (Don't forget to check* `kmax`*.)*

4.1.4. Advanced: Problem 4 – Accuracy and rounding errors

This problem is optional and should be solved only if time allows. Otherwise, contact your supervisor for a discussion of the results. Use directory `run/problem4` *as a working directory.*

In this final test of numerical solutions of ODEs we will check our error analysis with respect to discretization and rounding errors. In particular, we will try to verify the result of **exercise 1** on the existence of an *optimum* step size, which should vary as a function of consistency order, p.

Exercise P4: *To this end, we switch back to problem 1, and investigate the achieved precision as a function of (equidistant) step size, for both the* Euler *and the* Runge–Kutta *integrator.*

P4.1: *At first, we have to modify our driver-routine,* `frame.F`*. We will use the new driver* `frame_error.F`*. Document the differences, and briefly describe their purpose. If you do not know how to find out differences between two files, check the Linux manual page of* `diff`*. Finally, modify* `odeint.F` *by commenting out the output line regarding* `nstp,h,x,y(1)` *in the branch responsible for fixed step size methods. This test performs so many integration steps that the computation time would become dominated by the creation of this output if not commented out.*

P4.2: *Compare the error as a function of step size, for Problem 1 and both integrators as discussed above. Modify the* `Makefile` *in such a way that a new target* `frame_error` *can be built with* `make`*, which should compile* `frame_error.F` *instead of* `frame.F` *and create the executable* `frame_error.x`*.*

Use the `Makefile` *to create the executable, and perform the required test in directory* `problem4`*. Use the corresponding input files. Plot the results with* `ploterror.pro` *(modify to create* `.ps` *files), and discuss them in terms of our theoretical predictions of Section 2.2.*

4.2. FRIEDMANN–LEMAÎTRE cosmologies: numerical solutions

The problems covered in this section should be solved on the second day of our lab work.

4.2.1. Implementation and first tests

Implement the ODE describing the temporal evolution of the scale factor, as derived in Chapter 3, into the program, as a new "problem 5". Neglect the radiation term, since this plays a role "only" in the very first epoch(s) of the Universe, together with inflation, which we will neglect as well, and which is justified as long as we are not interested in the details of these phases and have normalized all quantities to their present values. Remember that all times are in units of $\tau_{\rm H}$ if we solve the equation for $\dot{a}/H_0$.

Perform the required changes as in **P2.1**, and allow the input quantities $\Omega_{\rm M}$ and Ω_Λ to be read in from a file `friedmann.inp` (logical unit = `inpunit`, already defined). As initial value, invert your result from **exercise 5a** (integrated from 0 to a!) to obtain an approximate value $a(t_{\rm start})$. In all what follows, we will adopt $t'_{\rm start} = 10^{-10}$ (in units of $\tau_{\rm H}$). Since we will only use the RK-integrator with adaptive step size, the JACOBIAN does not need to be defined.

For all your following work, use `run/problem5` as a working directory.

P5.1: *Test your program now by comparing with your results from **Exercise 5a** and **5b**. Plot the run of $a(t)$ with* `plotode.pro`*, after appropriate modifications (enable plotting the results for* problem 5*). The IDL routine* `xhair,x,y` *allows measuring the x-y-coordinates in a plot.*

4.2.2. Solutions for various parameter combinations

P5.2: *Now, try the solution for a matter-dominated, closed universe without cosmological constant and $\Omega_{\rm M} = 3$. Calculate until $t'_{\rm max} = 3.0$, observe the problem, and try to cure it by manipulating $t'_{\rm max}$. Plot the result, and try to explain the origin of the problem. At which a does the LIPSCHITZ-constant become infinite?*

If everything you did so far was OK, you should have realized that the temporal resolution is rather coarse (the solution is well-behaved, such that large step sizes are taken.) To cure this "problem", edit `odeint.F`, locate the term "`uncomment`" and do as suggested. By including the appropriate statement, a maximum step width as given by the input variable `dxsav` is ensured. Recompile, and replot the last figure, now with `dxsav=0.01`.

P5.3: *After everything runs smoothly, calculate and plot the following models. Choose an appropriate* `tmax` *– start with a default of* `tmax` $= 4$*, and adapt as necessary to adequately demonstrate the behavior of each universe. Briefly explain/comment on your findings, considering the particular values of $\Omega_{\rm M}$ and Ω_Λ:*

a)	$\Omega_{\rm M} = 3.0$	$\Omega_\Lambda = 0.1$	b)	$\Omega_{\rm M} = 3.0$	$\Omega_\Lambda = 0.2$
c)	$\Omega_{\rm M} = 0.0$	$\Omega_\Lambda = -0.1$	d)	$\Omega_{\rm M} = 1.0$	$\Omega_\Lambda = 1.0$
e)	$\Omega_{\rm M} = 1.0$	$\Omega_\Lambda = 2.55$	f)	$\Omega_{\rm M} = 1.0$	$\Omega_\Lambda = 2.6$

P5.4 – Constraints on $\Omega_{\rm M}$ and Ω_Λ: *Using the observationally well-proven fact that our present Universe is very close to flatness, use your simulations to obtain the $(\Omega_{\rm M}, \Omega_\Lambda)$ pair which is consistent with the present HUBBLE parameter and the age of our universe, $t_0 = 13.7 \pm 0.2$ Gyr (from the WMAP team). Analyze roughly the corresponding errors.*

P5.5 – The $(\Omega_{\mathrm{M}}, \Omega_\Lambda)$ diagram: *The* IDL *procedure* `diag.pro`*, in combination with data from* `diag.sav`*, displays the well-known $(\Omega_{\mathrm{M}}, \Omega_\Lambda)$ diagram which allows visualizing the present constraints for our Universe. At first, run the program, and have a look into the various possibilities for $a(t)$ as a function of $(\Omega_{\mathrm{M}}, \Omega_\Lambda)$, if one integrates the corresponding ODE with initial value $a = 1$ into the past and into the future (which is done here by the executable program* `bs` *which uses the* Bulirsch–Stoer *method).*

Play around a bit, and compare with your previous solutions. In particular, have a look into the "no big bang" and the "loitering[1] *universe" domain. What happens? Note also that there is a (small) region with $\Omega_\Lambda > 0$, where the universe still collapses finally.*

Now, improve the figure as follows, and plot the result.

- *draw the line distinguishing an open from a closed universe, with corresponding captions.*
- *draw the line distinguishing a presently accelerating from a presently decelerating universe.*
- *indicate your solution (with errors) from* exercise P5.4.

[1] In German: "herumlungernd".

Bibliography

S. M. Carroll, W. H. Press, and E. L. Turner: *The cosmological constant.* Annual Review of Astronomy and Astrophysics 30, 499 (1992)

A. S. Eddington: *On the instability of Einstein's spherical world.* Monthly Notices of the Royal Astronomical Society 90, 668 (1930)

S. Perlmutter, G. Aldering, G. Goldhaber, et al.: *Measurements of Omega and Lambda from 42 High-Redshift Supernovae.* The Astrophysical Journal 517, 565 (1999)

W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery: *Numerical Recipes in C. The Art of Scientific Computing.* (2nd ed.) Cambridge University Press (1992)

M. Roos: *Introduction to Cosmology.* (3rd ed.) Wiley (2003)

J. Stoer and R. Bulirsch: *Numerische Mathematik 2.* Springer (1990)

Index