Mathematics for industry

Lecture notes

Łukasz Płociniczak*

January 21, 2021

Contents

*Faculty of Pure and Applied Mathematics, Wrocław University of Science and Technology, Wyb. Wyspiańskiego 27, 50-370 Wrocław, Poland

Some preliminary information

Literature. The lecture will be self-sufficient, but some textbooks will be very helpful. Below are my suggestions (from the most basic to more advanced).

1. A. Friedman, W. Littman, *Industrial Mathematics - A Course in Solving Real-World Problems*, SIAM, Philadelphia 1994. A compilation of various modelling techniques illustrated with a real-world examples from the industry.

2. S. Howison, *Practical Applied Mathematics*, Cambridge Texts in Applied Mathematics. A highly readable book on modelling in industry and other fields of science and technology. Contains an interesting material concerning thin film flows and lubrication theory.

3. C.C. Lin, L.A. Segel, *Mathematics Applied to Deterministic Problems in the Natural Sciences*, SIAM 1988. A classical text on modelling with differential equations. A must-read for everyone interested in applied mathematics.

1 Introduction

This lecture is about mathematical modelling not only in industry (as its name says) but also in many other branches of science and technology. "Industry" is a vast container and everyone agrees that nowadays there is no clear boundary indicating what is and what is not an industry any more. Therefore, apart from the traditional notion of an industry we will learn about applied mathematics in biology, medicine, chemistry, physics, geophysics, agriculture and so on. In that we way we may appreciate how applied mathematics is useful in describing real-world phenomena.

Note that the material of this lecture concerns only *deterministic* and can be considered as a classical training of undergraduate applied mathematician. There is a huge dual branch of probabilistic modelling which will be covered on other courses (e.g. time series, stochastic processes, statistics, etc.).

The most important philosophical concept that we will learn about is *modelling*. It is something between science and art and one has to learn it in practice (hence the seminar). Mathematical modelling tries to formalize certain natural or industrial phenomena to understand them, forecast, improve, and eventually benefit. Without models the present technology would not be as we can experience it. In industry one usually utilizes mathematics as a optimization tool used in quality and cost control. There are many success stories where an intelligent mathematical model reduced the cost of the whole process. Note that using mathematics is usually very cost-effective - you only need your brain, paper, pencil, and a computer.

It is very hard to define what a model really is and we will defer it for philosophers. A British statistician George Box wrote "all models are wrong, but some are useful." This sentence captures the essence of a good mathematical model - it has to be useful. On the other hand, in the words of Mark Kac - "Models are, for the most part, caricatures of reality, but if they are good, they portray some features of the real world." One usually wants to build up a model that describes the essential features of the considered problem which forms from physical principles by a series of simplifications (but not too much of them). They have to be done systematically in an intelligent way because otherwise, we can loose the predicability of the model and, hence, its usefulness. We are always doing some trade-offs between computational complexity and predictive capabilities. It is an art to find the balance between these two.

A warning is in order. When formulating or analysing real-world models prepare to forget about clean and elegant formulas. Textbook examples are good for learning various techniques however, the reality is much more complex and interesting. Mathematicians are usually very good at reducing the whole complexity of the phenomenon to an isolated subproblem that then can be analysed (a conceptual model). In other case, it would be too cumbersome or even impossible to tackle the complete problem. A good example comes from weather forecast. The equations describing the state of the atmosphere and ocean are nonlinear partial differential equations to be solved on a spherical shell. They are so complicated that this task is virtually impossible even on present day supercomputers. What applied mathematicians do is to focus on really important phenomena and filter out these which are not. For example, sound waves are inessential for meteorology while the whole air flow is shallow (weather happening on 1000 km scales is mostly contained in the troposphere which is about 10 km

thick). Even after many simplification weather forecast is carried over on powerful supercomputers. Everyone knows how useful it is and that is saves lives.

Another textbook habit of undergraduates is anticipation of exact solutions of solved problems. In real life this is hardly the case and even when we are in possession of such, it may be completely useless because of its complexity. Approximate solutions are much better since they allow us to focus on crucial features of the system. The difficulty is to find a systematic way of finding such approximations. In this lecture we will learn of several techniques that are very useful.

Usually mathematical modelling progresses in two stages: formulation, and solution. When formulating a model we use out knowledge about physics, chemistry, biology, economics, ..., to describe what is happening in the analyses phenomenon. Then, we formulate that in terms of equations. This is a difficult part since, by nature, is highly interdisciplinary and requires many skills that cannot be learned from textbooks (such as interpersonal communication with people of different backgrounds). Then, in the solution phase of modelling we reduce the model to be trackable and explicatory. Here, we can use our intuition and many techniques that can be mastered by doing exercises.

Apart from purely deductive models outlined above, there is a large class of the so-called *black box* models. More or less, this is a data trained system of mathematical tools that given an input produces an output. The user does not know what happens in between and we usually loose the information about physics. Lately however, models based on neural networks and machine learning are becoming more popular, effective, and quite useful. This is also an interesting topic for another course.

2 Scaling and dimensional analysis

In this section we will consider a simple yet extremely useful tool of applied mathematics - dimensional analysis. It is very surprising that having only the knowledge of physical quantities that may affect a certain outcome we can infer about their relative combination that has its place in the solution. This works even if we do not know that solution or it is impossible to get an exact one. What we only need to know is a little bit of linear algebra. We will illustrate the method of *dimensional reduction* or *scaling* on a classical example.

Example. (*Projectile problem*) Suppose we throw a ball or shoot a projectile vertically in the x-direction. By $x = x(t)$ denote the position of the particle at time $t > 0$. Suppose also that initially it is on the ground and is thrown at velocity v_0. Then, by Newton's second law and the law of universal attraction it follows that

$$m\frac{d^2x}{dt^2} = -\frac{GMm}{(R+x)^2}, \tag{2.1}$$

with initial conditions

$$x(0) = 0, \quad \frac{dx}{dt}(0) = v_0. \tag{2.2}$$

Here, G is the gravitational constant, while $R = 6371$ km and $M = 5.9 \times 10^{24}$ kg are the radius and mass of Earth. Note that this is really a quite difficult problem! A nonlinear second order equation requires much care to analyse. We can also use numerical methods for solving it however, then we may loose all relevant information about the dependence of the solution on the model parameters[1]. To have a useful model we have to understand how things work and why. Therefore, we will try to find an informative approximate solution.

We can play with the above ODE and use the definition of the gravitational acceleration $g = GM/R^2$ to obtain

$$\frac{d^2x}{dt^2} = -\frac{gR^2}{(R+x)^2}. \tag{2.3}$$

Now, our intuition about throwing balls in the air tells us that most frequently x is much smaller than R and we denote it by $x \ll R$ (this notation will be made precise in later section). Therefore, it is reasonable to think that $R + x \approx R$ and hence,

$$\frac{d^2x}{dt^2} = -g, \tag{2.4}$$

which, along with initial conditions, has the simple high-school solution

$$x(t) = v_0 t - \frac{gt^2}{2}. \tag{2.5}$$

[1]Remember that computer simulations are always solving one concrete example of a set of model parameters at a time. It is very difficult then to learn about qualitative properties of the phenomenon. Therefore, analytical reasoning and human ingenuity is very important and will never be suppressed by computers.

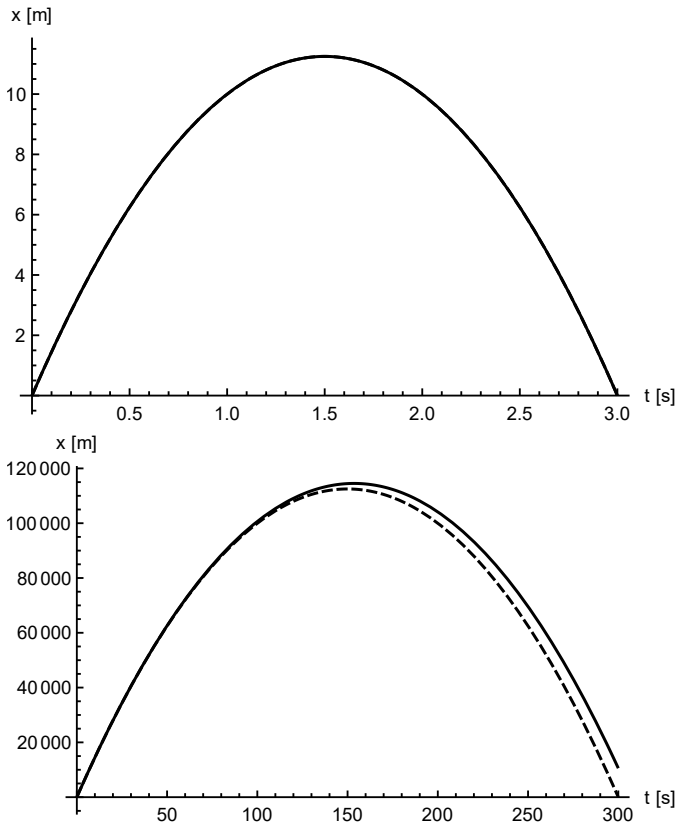

Figure 1: A height of the projectile in time for two different maximal heights. The solid line is the numerical solution while dashed represents approximation.

On Fig. 1 we can see a comparison between numerical and approximate solution that we have obtained. It is clear that for low maximal height, say $\leq$ 30 km, the approximation is extremely accurate. For larger heights it is no longer valid and must not be used in high precision simulations such as space flight.

Numerical simulations are only an experiment. We empirically know, more or less, about the region of applicability of our approximation. Applied mathematician, however, should not be satisfied by this answer. We would like to know exactly about the error that we make, how does the maximal height depend on the model parameters, and how to conduct all the steps systematically in order to improve the approximation.

The validity of the approximation can be inferred by the back-of-the-envelope calculation. From (2.5) we know that the maximal height is $v^2/(2g)$ which implies that we have to assume that

$$\frac{v^2}{2Rg} \ll 1, \tag{2.6}$$

in order to have a consistent approximation. This requirement, however, is vague since we have to be precise what does the "much smaller than" rigorously mean. □

We would like to make the above consideration systematic and precise. More specifically, we would like to know when it is possible to discard a part of the equation and what error we then make. In much more complicated problems (and in real-world

we meet such) this may be not so obvious (and it is not). Further, we would like to improve on our approximation to go further beyond the linear terms (as the uniform gravitational field g in the above example). We will learn about all of these in the sequel.

2.1 Examples of dimensional reduction

We start with the profound idea of dimensional reduction and scaling. Recall that all physical quantities must either be numbers or have a dimension. The types of the latter are length L, mass M, time T, electric current I, and temperature θ. From these *primary quantities* all other physical units can be composed. Note that these are abstract notions indicating various types of a given variable or a space in which it lives. They do not have nothing to do with specific units used in calculations such as meters, kilograms, seconds, amperes, or kelvins in SI system. However, they become specific once we fix the unit system. For example, it does not make any sense to add two quantities representing length and mass nor it is absurd to subtract force from velocity even though they mathematically are vectors. A dimension denotes the vector space in which everything is happening. This is the basic idea of dimensional reduction.

The second observation is that equations describing physical situations have to be *independent* of the specific unit system. Nature does not distinguish between ergs or joules, for it the energy is just energy. Therefore, transforming the equation into a nondimensional form should reveal physics (and usually is much more computationally trackable). The caveat is that in complicated models there is no unique way of doing so and many independent and non-equivalent paths of reasoning exits.

In what follows by the bracket we will denote the dimension of a given quantity, for example for the force F we have $[F] = ML/T^2$.

Example. (*Projectile problem cont'd*) Suppose we would like to find an expression for the maximal height x_m of the flight of the projectile as a function of the remaining parameters. We encapsulate all the information inside some unknown function f. This is the law that interconnects x_m, v, R, and g and can be written abstractly as

$$f(x_m, v, R, g) = 0. \tag{2.7}$$

Now, since the laws of physics laws are independent on the units we must be able to choose such a combination of all present quantities in order to obtain a *nondimensional* result. That is, we look for exponents a, b, c, and d such that

$$[x_m]^a[v]^b[R]^c[g]^d = L^0M^0T^0 = 1. \tag{2.8}$$

Recalling that $[v] = L/T$, $[R] = L$, and $[g] = L/T^2$ we have

$$L^aL^bT^{-b}L^cL^dT^{-2d} = 1. \tag{2.9}$$

Since the fundamental dimensions are independent we obtain a system of linear equations

$$\begin{cases} a + b + c + d = 0, \\ -b - 2d = 0. \end{cases} \tag{2.10}$$

Solving it we obtain a two-parameter family

$$b = -2d, \quad c = d - a, \quad a, d \in \mathbb{R}. \tag{2.11}$$

Therefore, we learn that the following combination of our parameters is nondimensional

$$x_m^a \nu^{-2d} R^{d-a} g^d = \left(\frac{x_m}{R}\right)^a \left(\frac{\nu^2}{Rg}\right)^{-d}. \tag{2.12}$$

Hence, we can infer that due to dimensional homogeneity of physical laws we should have

$$f(x_m, \nu, R, g) = f\left(\frac{x_m}{R}, \frac{\nu^2}{Rg}\right) = 0. \tag{2.13}$$

This tells us a lot! For example, it is almost always the case that we can solve the above implicit equation for one of its variables (Inverse function theorem). In this way we know that there exists a function $h = h(z)$ such that

$$x_m = R\, h\left(\frac{\nu^2}{Rg}\right). \tag{2.14}$$

This is the only way that the maximal height can depend on the remaining parameters. The simplest choice is, of course, h being a linear function (since a constant does not make sense). For example, our approximation (2.5) tells us that $h(z) \approx z/2$ and on Fig. 2 we indeed see that if $x_m \leq 100$ km it is a very good approximation. This shows that even if we do not know h exactly, a first guess can be decently accurate (up to a constant). The exact form of the function h can be found, as we did, by numerical or real-world experiments. Note however, how much we have learned about the projectile without actually solving anything! This is a great aid in modelling. Note that we did not even need to know the exact form of the differential equation as long we knew the relevant physical quantities that entered the formulation. □

Example. (*Drag on a sphere*) We will now investigate a classical problem of determining the drag force on a spherical object submersed in a fluid. This has a profound meaning in motor and aerospace industries. Engineers constantly want to design bicycles, cars, and airplanes to have the smallest drag possible. This is a subject of constant research and investigation even in XIX century. We will see how dimensional analysis can help us to determine the formula for the drag even without knowing the form of the governing differential equations for the flow[2].

Assume that a ball of radius R is immersed inside a flowing fluid of density ρ, velocity U and viscosity μ. By F_D we denote the drag force acting on the sphere. As in the previous example, we assume that the physical law combining all of these quantities is given by a function f, i.e. $f(R, \rho, U, \mu, F_D) = 0$. Now, we collect all the relevant dimensions

$$[R] = L, \quad [\rho] = ML^{-3}, \quad [U] = LT^{-1}, \quad [\mu] = ML^{-1}T^{-1}, \quad [F_D] = MLT^{-2}. \tag{2.15}$$

[2]They are the celebrated Navier-Stokes equations.

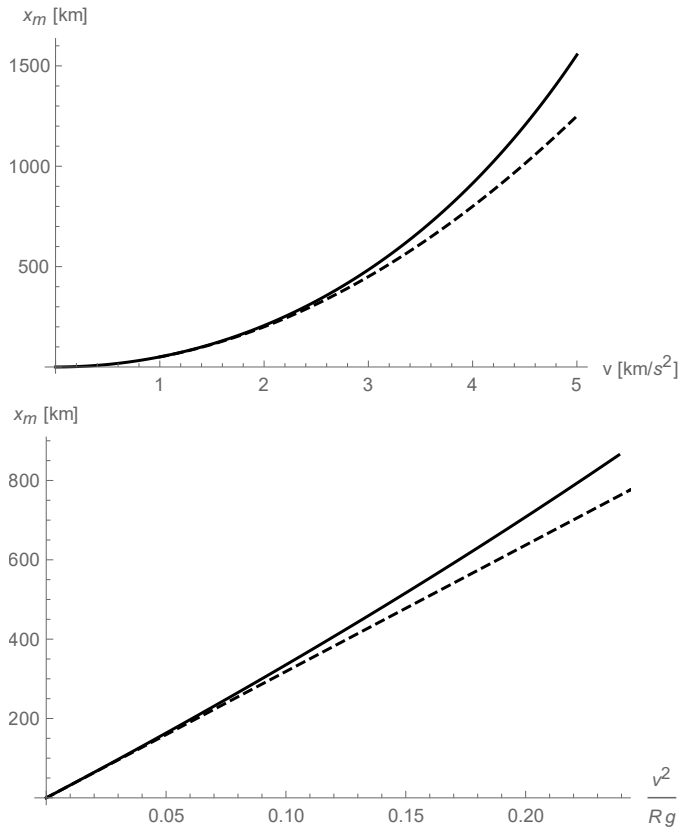

Figure 2: On the left: projectile maximal height (solid line) and its approximation $v^2/(2g)$ (dashed line) as a function of initial velocity. On the right: projectile maximal height as a function of $v^2/(Rg)$ (solid line) and its linear approximation $h(z) = z/2$.

Since all physical laws have to be independent on the specific units chosen, they have to depend only the nondimensional quantities that can be formed from the above. We thus look for a, b, c, d, and e such that

$$R^a \rho^b U^c \mu^d F_D^e = 1. \tag{2.16}$$

Plugging the relevant dimensions we obtain

$$1 = L^a M^b L^{-3b} L^c T^{-c} \mu^d M^d L^{-d} T^{-d} M^e L^e T^{-2e} = L^{a-3b+c-d+e} M^{b+d+e} T^{-c-d-2e}, \tag{2.17}$$

which, due to independence of fundamental dimensions leads to

$$\begin{cases} a - 3b + c - d + e = 0, \\ b + d + e = 0, \\ -c - d - 2e = 0. \end{cases} \tag{2.18}$$

The solution is a two-parameter family given by[3]

$$b = a + e, \quad c = a, \quad d = -a - 2e, \quad a, e \in \mathbb{R}. \tag{2.19}$$

Therefore, we can form a nondimensional combination known as *nondimensional group*

$$R^a \rho^{a+e} U^a \mu^{-a-2e} F_D^e = \left(\frac{R\rho U}{\mu} \right)^a \left(\frac{\rho F_D}{\mu^2} \right)^e. \tag{2.20}$$

Hence, our physical law depends only on two nondimensional parameters in a specific combination instead of five dimensional ones. We can thus write

$$f(R, \rho, U, \mu, F_D) = f\left(\frac{R\rho U}{\mu}, \frac{\rho F_D}{\mu^2} \right) = 0. \tag{2.21}$$

One number that appeared above is of great importance in fluid dynamics, it is the *Reynolds number*

$$Re = \frac{R\rho U}{\mu}. \tag{2.22}$$

This quantity helps to find the properties of the flow and is usually found that for low Reynolds number the flow is laminar (calm, smooth, constant) while for larger values - turbulent. It can be thought as a ratio of inertial to viscous forces. The nonlinear equation $f = 0$ can be solved for one of its variables yielding

$$F_D = \frac{\mu^2}{\rho} F(Re). \tag{2.23}$$

This expression can be conveniently transformed into

$$F_D = \rho R^2 U^2 \frac{\mu^2}{R^2 \rho^2 U^2} F(Re) = \rho R^2 U^2 G(Re), \tag{2.24}$$

[3]Note that it is very beneficial to have the exponent of sought quantity being taken as a parameter of our solution, i.e. $e \in \mathbb{R}$.

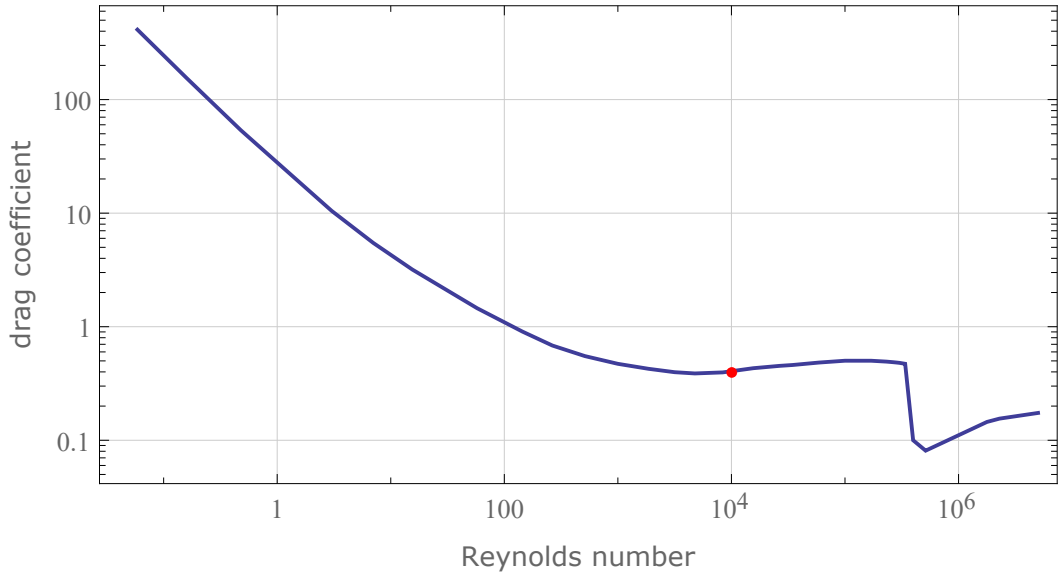

Figure 3: A general shape for the drag coefficient G for a flow over a sphere.

where we defined a new (unknown) function G, the so-called *drag coefficient* modulo some constants, which has to be found either by solving Navier-Stokes equations or doing experiments. An experimentally obtained graph is given on Fig. 3. We immediately notice that for $10^3 < Re < 10^5$ the drag coefficient is almost constant with $G \approx 0.7$. This is the simplest choice. Therefore, we obtain the extremely useful and widely used relationship

$$F_D \propto \rho A U^2, \quad 10^3 < Re < 10^5, \tag{2.25}$$

which means that the drag is proportional to the density, area of a body, and velocity squared. Everyone who rides a bicycle knows that very well - it requires much more energy to accelerate from 30 to 31 km/h than from 20 to 21 km/h in the same wind conditions.

Further, for the low Reynolds number, that is for laminar flow, the graph on Fig. 3 is linearly decreasing on a log-log scale. Translating it backwards it follows that a very good fit can be obtained in the form $G(Re) \propto Re^{-1}$. Therefore,

$$F_D \propto \mu R U, \quad Re < 20, \tag{2.26}$$

which is the famous Stokes Law of drag around a sphere. Stokes solved the dynamical equations analytically and obtained that the drag is proportional to viscosity, radius, and the velocity (not squared!). The constant of proportionality is equal to 6π. This is another triumph of clever analytical reasoning.

Another important application of the found formulas can be found in model testing. The crucial observation follows from the nondimensional units: two flows with the same nondimensional parameters are the same. This means that instead of testing, for example, a large-scale original vessel we can build its small model and choose a particular fluid for the Reynolds numbers to be equal. Evidently, this has great impact on streamlining the process of construction, design and is very economical.

Notice that we were able to deduce an enormous amount of information without even the knowledge of Navier-Stokes equations! Obtained formulas have been improved, tested, and used in industry for over 150 years and certainly will be used much longer. □

Example. (*Pendulum*) We will revisit the high school problem of finding the period P of a mathematical pendulum of mass m, length l, and initial angle θ. The gravitational acceleration is g. Note that it appears that me mentioned all the relevant quantities. The physical law, as before, is $f(P, m, l, \theta, g) = 0$ and we would like to construct a nondimensional combination of given parameters

$$[P]^a[m]^b[l]^c[\theta]^d[g]^e = 1, \tag{2.27}$$

where a, b, c, and d are unknowns while $[P] = T$, $[m] = M$, $[l] = L$, $[\theta] = 1$, and $[g] = LT^{-2}$ (the angle is nondimensional since it is a ratio of two lengths). Whence,

$$T^aM^bL^cL^eT^{-2e} = 1, \tag{2.28}$$

which leads to

$$\begin{cases} a - 2e = 0, \\ b = 0, \\ c + e = 0, \end{cases} \tag{2.29}$$

with solution

$$b = 0, \quad c = -\frac{a}{2}, \quad e = \frac{a}{2}, \quad a, d \in \mathbb{R}. \tag{2.30}$$

Therefore, the mass does not enter the equation what we know very well from physics. We have the nondimensional group

$$P^a l^{-\frac{a}{2}} \theta^d g^{\frac{a}{2}} = \left(P\sqrt{\frac{g}{l}}\right)^a \theta^d, \tag{2.31}$$

and hence our law can be resolved to

$$P = 2\pi\sqrt{\frac{l}{g}}F(\theta), \tag{2.32}$$

for some F. The constant 2π in from of the above expression was added to remind us that for the harmonic oscillator, i.e. pendulum with small initial angle, the period is equal to $2\pi\sqrt{l/g}$. Note that thanks to the dimensional analysis we have determined the unique form of the formula for the period. You may remember from the ODE course that the function F can be found by solving the respective equation which we did not even write here! The answer is

$$F(\theta) = \int_0^\theta \frac{d\varphi}{\sqrt{\cos\varphi - \cos\theta}} = 1 + \frac{1}{16}\theta^2 + \frac{11}{3072}\theta^4 + \ldots \tag{2.33}$$

We see that the simplest guess, that is F constant, gives us a second order correction (since θ^2). □

Example. (*Pulsating stars*) Certain stars change its luminosity in a periodic manner thanks to different mechanisms. For example they can shrink and expand due to thermodynamic processes, or be eclipsed by some smaller object lying on our line of sight. We will use dimensional analysis to find the period of a internally pulsating star of mass m, radius r, period P. The gravitational constant is G. We assume that the physical law governing this phenomenon is $f(m, r, P, G) = 0$, where $[m] = M$, $[r] = L$, $[P] = T$, and $[G] = L^3M^{-1}T^{-2}$. We have

$$M^a L^b T^c L^{3d} M^{-d} T^{-2d} = M^{a-d} L^{b+3d} T^{c-2d} = 0, \tag{2.34}$$

from which we deduce

$$\begin{cases} a - d = 0, \\ b + 3d = 0, \\ c - 2d = 0, \end{cases} \tag{2.35}$$

with a solution

$$a = \frac{1}{2}c, \quad b = -\frac{3}{2}c, \quad d = \frac{1}{2}c, \quad c \in \mathbb{R}. \tag{2.36}$$

Further, we have nondimensional group

$$m^{\frac{c}{2}} r^{-\frac{3c}{2}} P^c G^{\frac{c}{2}} = \left(P\sqrt{\frac{mG}{r^3}} \right)^c. \tag{2.37}$$

We can now invert the physical law to obtain

$$f\left(P\sqrt{\frac{mG}{r^3}} \right) = 0 \to P = C\sqrt{\frac{r^3}{Gm}}. \tag{2.38}$$

Recalling that m/r^3 is proportional to the mean density of the star we can rewrite this as

$$P = D\sqrt{\frac{1}{G\rho}}, \tag{2.39}$$

hence, the period is inversely proportional to the square root of the density. English astronomer - Arthur Eddington - analytically calculated the prefactor with the use of thermodynamics, his result states that $D = \sqrt{3\pi/2\gamma}$ with γ is a ratio of specific heats of stellar material. Note that this result is completely independent on the star size! □

Example. (*Nuclear bomb*) A famous example of dimensional analysis comes from one of the greatest English applied mathematicians - G.I. Taylor. Appointed by the British government he was working on development of nuclear weapons. From security reasons he did not take part in the Manhattan project[4]. Instead, he was armed with a ingenuity in mathematics, physics, and some photographs of the Trinity explosion. Taylor knew that the nuclear explosion creates a spherical shock wave that separates areas of different pressure. By solving fluid flow equations, which was a daunting

[4]Independently, John von Neumann worked on similar topics in US, while Leonid Sedov in USSR.

task, Taylor assumed that the radius r of the explosion depended on the density of air ρ, released energy E, and time after the blast t. Note that guessing the relevant quantities in that case is extremely difficult. The dimensions are $[r] = L$, $[\rho] = ML^{-3}$, $[E] = ML^2T^{-2}$, and $[t] = T$. We have $f(r, \rho, E, t) = 0$, and

$$L^a M^b L^{-3b} M^c L^{2c} T^{-2c} T^d = L^{a-3b+2c} M^{b+c} T^{-2c+d} = 1, \tag{2.40}$$

and hence

$$\begin{cases} a - 3b + 2c = 0, \\ b + c = 0, \\ -2c + d = 0. \end{cases} \tag{2.41}$$

The solution is a one-parameter space

$$b = \frac{a}{5}, \quad c = -\frac{a}{5}, \quad d = -\frac{2a}{5}, \quad a \in \mathbb{R}. \tag{2.42}$$

This gives us that

$$r^5 = C\frac{Et^2}{\rho}, \tag{2.43}$$

and from here we can express the energy of the blast

$$E = C\frac{\rho r^5}{t^2}. \tag{2.44}$$

The constant C follows from thermodynamics and flow equations, it can safely be assumed that it is close to 1 (actually 1.036). Taylor looked at the publicly accessible photographs showing the snapshots of a Trinity explosion and plugged several radii for given times. On this basis, he was able to accurately estimate the energy of detonation. His results, 20 kilotons of TNT, were very close to the official value of 22 kilotons of TNT. Similar dimensional reasoning leads to a results that the volume V of a crater resulted from the explosion is

$$V = C\left(\frac{E}{\rho g}\right)^{\frac{3}{4}}. \tag{2.45}$$

Governments were shocked and Taylor's papers were classified for several years. Observe how far mathematics can take you. □

Example. (*Pythagoras theorem*) Another remarkable example concerns Pythagoras theorem. Let there be a right triangle with sides a, b, and c being its hypotenuse. By α denote one of its acute angles. We assume that the area A can be calculated with knowledge of the hypotenuse and the angle. Since the area has to have dimension L^2 we have

$$A = c^2 f(\alpha). \tag{2.46}$$

Now, if we drop a altitude on the side c we obtain two similar triangles with hypotenuses a and b and areas A_1, A_2 with $A_1 + A_2 = A$. Note also that in both of these triangles one of the acute angles is α. This gives us that

$$A_1 = a^2 f(\alpha), \quad A_2 = b^2 f(\alpha). \tag{2.47}$$

But,

$$A = A_1 + A_2 \to c^2 f(\alpha) = a^2 f(\alpha) + b^2 f(\alpha). \tag{2.48}$$

Cancelling f yields the Pythagoras theorem. What is the exact form of f? □

Example. (*Waves in the ocean*) As with any fluid, the ocean can host many different kind of waves. Let us find the frequency ω of such wave with respect to the wavelength λ, surface tension σ, the density of water ρ, and gravity g. We have $f(\omega, \lambda, \sigma, \rho, g) = 0$ and plugging it respective dimensions we obtain (note that $[\sigma] = MT^{-2}$)

$$T^{-a}L^b M^c T^{-2c} M^d L^{-3d} L^e T^{-2e} = L^{b-3d+e} M^{c+d} T^{-a-2c-2e} = 1. \tag{2.49}$$

It follows that

$$\begin{cases} b - 3d + e = 0, \\ c + d = 0, \\ -a - 2c - 2e = 0, \end{cases} \tag{2.50}$$

with a two-parameter space of solutions

$$b = \frac{a}{2} - 2c, \quad d = -c, \quad e = -\frac{a}{2} - c, \quad a, c \in \mathbb{R}, \tag{2.51}$$

from which it follows that

$$\omega = \sqrt{\frac{g}{\lambda}} F\left(\frac{\sigma}{g\rho\lambda^2}\right), \tag{2.52}$$

for some F. We will learn in the sequel that the correct formula for the frequency of the ocean waves is

$$\omega = \sqrt{2\pi\frac{g}{\lambda} + (2\pi)^3\frac{\sigma}{\rho\lambda^3}} = \sqrt{2\pi\frac{g}{\lambda}}\sqrt{1 + 4\pi^2\frac{\sigma}{g\rho\lambda^3}}, \tag{2.53}$$

from which we can infer about the correct form of the F function. Note also that the first formula above shows that the gravity and capillary effects are separated. Indeed, surface tension is usually very small and hence only observable for very small wavelengths. Do an experiment in your bath tube and on a lake. □

Finally, we state some additional interesting results for Reader's own practice and entertainment. All the below examples can be obtained by similar arguments as above. Here, C always denotes a constant that has to be determined by experiment or theory.

1. The pressure p of a soap bubble of radius r and surface tension σ is given by

$$p = C\frac{\sigma}{r}. \tag{2.54}$$

   Notice that smaller bubbles burst more noisily!

2. The speed of sound c in a medium of density ρ and pressure p is given by

$$c = C\sqrt{\frac{p}{\rho}}. \tag{2.55}$$

3. The tail-beat frequency f of of swimming fish of body length l, muscle strength σ (stress) in a fluid of density ρ is

$$f = \frac{C}{l}\sqrt{\frac{\sigma}{\rho}}. \tag{2.56}$$

4. The time t needed to cool a drink of thermal conductivity κ, and heat capacity c, and temperature θ with ice cubes of side L is

$$t = C\frac{L^2 c}{\kappa}. \tag{2.57}$$

   Notice that t depends on L^2: it is quicker to cool a drink with smaller cubes than larger since the area of heat conduction is greater given the same volume. A striking fact is that this does not depend on the temperature!

5. There are n people in a boat and each of them occupies volume V and puts power P in accelerating it by rowing. Show that the wetted area A of the boat is proportional to $(nV)^{\frac{2}{3}}$. Use the previously determined formula for a drag force, i.e. proportional to $\rho U^2 A$ where U is the velocity, while ρ density, to show that

$$U = C\left(\frac{n^{\frac{1}{3}}P}{\rho V^{\frac{2}{3}}}\right)^{\frac{1}{3}}. \tag{2.58}$$

   Moreover, assume that both V and P are proportional to rower's mass. Is size then important for winning a race?

Finally, note that the success of dimensional analysis lies in the correct identification of relevant parameters and quantities. This comes from trained intuition and knowledge.

2.2 Buckingham Pi Theorem (optional)

It is time to formalize our previous intuitive, but highly efficient, reasoning. As we have noticed, it is all linear algebra. Suppose we have a physical variable q that depends on other quantities p_i for $i = 1, ..., n$. Let D_i with $i = 1, ..., m$ be the fundamental dimensions such as L, M, and T. We can thus write

$$[q] = D_1^{d_1} D_2^{d_2} ... D_m^{d_m}, \tag{2.59}$$

and

$$[p_i] = D_1^{\alpha_{1i}} D_2^{\alpha_{2i}} ... D_m^{\alpha_{mi}}, \tag{2.60}$$

for some fixed α_i and d_i. We start with the physical law $q = f(p_1, p_2, ..., p_n)$, where, for convenience, we have already used the inverse function theorem. In order to dimensionally reduce this we ask if there are numbers a_i for $i = 1, ..., n$, for which

$$[q] = [p_1]^{a_1}[p_2]^{a_2}...[p_n]^{a_n}. \tag{2.61}$$

Plugging (2.59) and (2.60) into (2.61) we obtain a system of linear equations

$$\sum_{j=1}^{n} a_j \alpha_{ij} = d_i, \quad i = 1, ..., m. \tag{2.62}$$

The above can be expressed as a matrix equation

$$A\mathbf{a} = \mathbf{d}, \tag{2.63}$$

for vectors $\mathbf{a} = (a_1, a_2, ..., a_n)^T$ and $\mathbf{d} = (d_1, ..., d_m)^T$ and the *dimension matrix*

$$A = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ \alpha_{m1} & \alpha_{m2} & \cdots & \alpha_{mn} \end{pmatrix}. \tag{2.64}$$

As we can see, the number of rows in A is equal to the number of fundamental dimensions, and the number of columns represents the physical quantities on which q depends. We recall from linear algebra that the solution of (2.63) can be written as

$$\mathbf{a} = \mathbf{a}_p + \sum_{i=1}^{k} c_i \mathbf{a}_i, \quad c_i \in \mathbb{R}, \tag{2.65}$$

where $\mathbf{a}_p$ is the particular solution of (2.63), i.e. any vector satisfying this system, and $\mathbf{a}_i$ are the vectors spanning Ker A, that is

$$A\mathbf{a}_i = 0, \quad i = 1, ..., k, \quad \text{with} \quad \dim \text{Ker } A = k < n. \tag{2.66}$$

Since $\mathbf{a}_i$ constitute a basis, say

$$\mathbf{a}_i = (a_{1i}, a_{2i}, ..., a_{ni})^T, \quad i = 1, ..., k, \tag{2.67}$$

the nondimensional products called *nondimensional gropus*

$$\Pi_i = p_1^{a_{1i}} p_2^{a_{2i}} ... p_n^{a_{ni}}, \quad i = 1, ..., k, \tag{2.68}$$

are independent quantities. Moreover, if

$$\mathbf{a}_p = (a_{1p}, a_{2p}, ..., a_{np})^T, \tag{2.69}$$

then the quantity

$$Q = p_1^{a_{1p}} p_2^{a_{2p}} ... p_n^{a_{np}}, \tag{2.70}$$

has the same physical dimension as q. Therefore, our basic physical law can now be written as

$$q = Q\, F(\Pi_1, \Pi_2, ..., \Pi_k), \quad 0 \leq k < n. \tag{2.71}$$

We have thus proved the famous Buckingham Pi Theorem[5]

Theorem 1 (Buckingham Pi Theorem). *The dimensional expression* $q = f(p_1, p_2, ..., p_n)$ *can be reduced to the form* $q = Q\, F(\Pi_1, ..., \Pi_k)$ *with* Π_i *being the independent nondimensional quantities with* $0 \leq k < n$ *and* Q *having the same physical dimension as* q.

Note that if the fundamental dimensions are independent, then $k = n - m$. In our preceding cases, $k = n - 3$. And thus we have $n - 3$ nondimensional groups.

[5]The word "Pi" comes from (2.68). Edgar Buckingham did not prove this result as a first one. For example, several contributions are due to Joseph Bertrand and Lord Rayleigh. However, Buckingham generalized several special cases and used the letter Π.

2.3 Scaling and nondimensionalization

All we have learned about finding nondimensional quantities has a great importance in mathematical modelling not only die to finding useful formulas. It is an essential tool in simplifying differential equations. When modelling, we usually compare different quantities. Some of them might be neglected due to relative smallness. Dimensional analysis is crucial in determining the respective scales that various terms represent. There is a lot of art and understanding of the physical principles in this process which is the fundamental device in the applied mathematician's repertoire.

Suppose we have a differential equation modelling a certain quantity to change in space and time. The fundamental question is how to choose an appropriate nondimensional system of units in order to cast the equation into the most transparent form. In particular, we would like to see which if the appearing terms play greater role than the other in *considered situation*. Choosing an appropriate system of nondimensional units is called *scaling*.

There is no general theory for scaling. At least when it comes to arriving at physically meaningful results. Since a product of nondimensional quantities is still nondimensional we do not know which of these is the most physically relevant. This is the place where art comes into play. The best way to learn about choosing proper scales is learning them on examples. We are going to look at some of them.

Example. (*Projectile problem revisited*) In the projectile problem we have intuitively assumed that $R + x \approx R$ for $x \ll R$. Now, we would like to see how this can be done systematically. The answer lies in the appropriate scaling of all of the variables appearing in the problem: dependent x and independent t. We introduce a system of nondimensional quantities according to

$$x = \xi x^*, \quad t = \tau t^*, \tag{2.72}$$

where starred variables are nondimensional while ξ and τ are scales (a typical values) to be chosen. These choices are usually the most difficult tasks in dealing with real-world problems. In some situations it is known from the beginning how to proceed, and in other we have to use a general framework of dimensional analysis. As an illustration we will proceed in the latter way.

First, note that the change of variables changes the derivatives in the following way

$$\frac{dx}{dt} = \frac{d(\xi x^*)}{d(\tau t^*)} = \frac{\xi}{\tau}\frac{dx^*}{dt^*}, \tag{2.73}$$

which is a simple application of the chain rule. Then, the projectile equation transforms into

$$\frac{\xi}{\tau^2}\frac{d^2x^*}{d(t^*)^2} = -\frac{gR^2}{(R + \xi x^*)^2}, \tag{2.74}$$

while the initial conditions are

$$x^*(0) = 0, \quad \frac{dx^*}{dt^*}(0) = \frac{\tau}{\xi}v. \tag{2.75}$$

Dimensional analysis tells us that there are three independent nondimensional groups

$$\Pi_1 = \frac{\xi}{gt_c^2}, \quad \Pi_2 = \frac{\xi}{R}, \quad \Pi_3 = \frac{\tau\nu}{\xi}. \tag{2.76}$$

We have to choose the characteristic scales ξ and τ according to one of the above in order to represent the physical situation. Note that, mathematically speaking, this choice is not unique. For example, Π_1 is the relative acceleration of the particle with respect to the gravity, Π_2 is the relative height of the flight with respect to the radius of Earth, and Π_3 measures the relative velocity of the particle with respect to the initial one. Which group to choose as the one relevant to our situation? In general this is far from obvious. There are papers and book written on such topic. In complex situations the modeller has to posses a deep knowledge of the problems and underlying physics, biology, etc. A trial and error method is also frequently implemented.

A good rule to start up is to choose the scale appearing in the initial conditions. That is, we set $\Pi_3 = 1$ and, hence, choose

$$\xi = \tau\nu. \tag{2.77}$$

Second, physically, the projectile does not travel very far and we can choose $\Pi_2 \to 0$. Therefore, we are left with $\Pi_1 = 1$ which along with (2.77) gives us

$$\xi = \frac{\nu^2}{g}, \quad \tau = \frac{\nu}{g}. \tag{2.78}$$

As the correct scales of this problem. Returning to our differential equation we get

$$\frac{d^2x}{dt^2} = -\frac{1}{(1+\epsilon x)^2}, \quad \epsilon = \frac{\nu^2}{gR}, \tag{2.79}$$

where we have dropped asterisks in order not to clutter the notation - this is a common practice once the scales has been chosen (however, this is a slight abuse of notation). The initial conditions now are

$$x(0) = 0, \quad \frac{dx}{dt}(0) = 1. \tag{2.80}$$

Observe how does this look much more pleasant than the dimensional problem. We have chosen the only nondimensional parameter present in the problem to be named ϵ not accidentally. For Earth it is $\epsilon \approx 10^{-8}\nu^2$ which for usual everyday situations is very small. This is the reason we can neglect it. This is the systematic statement of the fact that we previously have simplified the equation by neglecting some terms. Since, in the nondimensional terms, x is of order of unity (because we have scaled it with its characteristic scale), ϵx is a small nondimensional number so we can safely take $\epsilon \to 0$ and know that we neglect a term of small significance. This is a crucial observation for the next section when we develop a method of perturbations where we can utilize the fact that an equation possesses a small parameter.

Notice how the scaling revealed the relative size of various terms appearing in the equation. It is then meaningful and clear to simplify some of them. Observe, however,

that this is a reflection of our initial choice of the scales and choosing them differently would let us focus on different physical situation. Scaling a complex nonlinear equation is usually very difficult and requires a lot of care.

To get a glimpse of what can go wrong suppose we choose our scales differently. For example, we set $\Pi_2 = 1$ and $\Pi_3 = 1$. Then, our problem is

$$\epsilon \frac{d^2x}{dt^2} = -\frac{1}{(1+x)^2}, \tag{2.81}$$

with

$$x(0) = 0, \quad x'(0) = 1. \tag{2.82}$$

This is, of course, equivalent to our original problem. That that, however, that the ϵ not multiplies the derivative. And thus, having it set to 0 results in a superficial contradiction $0 = -1$. We certainly cannot proceed this way. This is a delicate limit and we will learn how to deal with it in further section. The chosen scales are such we choose the typical velocity of the problem to be of the initial one and the height of the projectile to be comparable to Earth's radius. This is a different physical situation and requires a different treatment. □

Example. (*Damped pendulum*) Recall from your ODE course that the evolution of an angle of afree mathematical pendulum of length l is given by

$$l\frac{d^2\theta}{dt^2} + k\frac{d\theta}{dt} + g\sin\theta = 0, \tag{2.83}$$

where k ($[k] = LT^{-1}$) is the damping constant. We assume the following initial conditions

$$\theta(0) = \theta_0, \quad \frac{d\theta}{dt}(0) = \Omega_0. \tag{2.84}$$

In order to nondimensionalize the above we have to prescribe scales for the angle Θ and time τ. For the latter, we have three choices as can be seen by doing dimensional analysis

$$\tau_1 = \sqrt{\frac{l}{g}}, \quad \tau_2 = \frac{l}{k}, \quad \tau_3 = \frac{1}{\Omega_0}. \tag{2.85}$$

The first time scale is the period of small undamped oscillations, the second is the time required for damping to have an effect. The third scale is the inverse of initial angular velocity, i.e. the time required to cover one radian of motion on a circle. The scales for the angle are

$$\Theta_1 = \theta_0, \quad \Theta_2 = \Omega_0\tau, \tag{2.86}$$

where τ is any of the above time scales. There is certainly a lot of possible choices.

Suppose that we want to study the case when the period of oscillation does not depart much from the linearised theory, i.e. we scale the time with τ_1. Suppose also that the initial velocity is small and we would like to scale the angle with Θ_1 (not necessarily small). Then for $\theta = \Theta_1\theta^*$ and $t = \tau_1 t^*$ we have

$$\frac{d\theta}{dt} = \theta_0\sqrt{\frac{g}{l}}\frac{d\theta^*}{dt^*}, \tag{2.87}$$

and

$$l\theta_0 \frac{g}{l} \frac{d^2\theta^*}{d(t^*)^2} + k\theta_0 \sqrt{\frac{g}{l}} \frac{d\theta^*}{dt^*} + g \sin(\theta_0 \theta^*) = 0. \tag{2.88}$$

This can be simplified by dividing and removing asterisks (for convenience)

$$\frac{d^2\theta}{dt^2} + 2\beta \frac{d\theta}{dt} + \omega_0^2 \frac{1}{\theta_0} \sin(\theta_0 \theta) = 0, \quad \beta := \frac{1}{2} \frac{k}{\sqrt{gl}}, \quad \omega_0 = \sqrt{\frac{g}{l}}, \tag{2.89}$$

where we have defined the rescaled damping coefficient β and angular frequency of free oscillations ω_0. The initial conditions are now

$$\theta(0) = 1, \quad \frac{d\theta}{dt}(0) = \gamma, \quad \gamma = \frac{1}{\Omega_0 \theta_0} \sqrt{\frac{g}{l}}. \tag{2.90}$$

Notice that scaling has reduced the number of model parameters from 5 to 3. Moreover, the obtained form of the ODE is convenient to analyse when the oscillations are almost linear. If the initial angle (and due to conservation of energy, all angles) are small, we can approximate the nonlinear term

$$\frac{1}{\theta_0} \sin(\theta_0 \theta) = \theta - \frac{\theta_0^2}{6} \theta^3 + ... \tag{2.91}$$

Leaving only the first term results in the linear oscillator, while retaining the next term gives Duffing's equation. This immediately gives us an important result that the error we make by linearising is $O(\theta_0^2)$. Not bad for just a simple use of algebra. □

Example. (*Piano string*) A rather accurate model of (small) vibrations of the piano string is the following forth order PDE

$$\rho A \frac{\partial^2 y}{\partial t^2} = T \frac{\partial^2 y}{\partial x^2} - EAk^2 \frac{\partial^4 y}{\partial x^4} = 0, \tag{2.92}$$

where ρ is the density of the string, A and area of its cross-section, E is Young's modulus, and k the so-called radius of gyration. Moreover, $y = y(x, t)$ denotes the deflection of the string at time t at position x. Loosely speaking, the second time derivative comes from Newton's second law, spatial derivative represents change in potential energy, while forth derivative describes bending stiffness. Usually, the latter is very small and can be neglected (as for example, in guitar). We would like to assess the relative size of this term and see whether we really can ignore it.

Anticipating the smallness of the fourth order term we scale our equation

$$x = Lx^*, \quad t = \frac{L}{c} t^*, \quad c = \sqrt{\frac{T}{A\rho}}, \tag{2.93}$$

where L is the string length and we have defined the wave velocity c. Note that both of these scales are natural for a freely vibrating string. Moreover, since the PDE is homogeneous we do not need to scale the deflection y (why?). After scaling the equation becomes

$$\frac{\partial^2 y}{\partial t^2} = \frac{\partial^2 y}{\partial x^2} - \epsilon \frac{\partial^4 y}{\partial x^4}, \quad \epsilon = \frac{Ek^2}{\rho L^2 c^2}, \tag{2.94}$$

where we have defined the nondimensional parameter. Observe that now our problem contains only one such. This makes the problem much more trackable.

Let us make a quick assessment of the magnitude of ϵ. For a circular wire of radius $r \approx 1$ mm and length 1m we have $k^2 = r^2/2 \approx 0.5 \times 10^{-6}$. Taking the string to be made of steel we have $E = 2 \times 10^{11}$ Pa and $\rho = 7.8 \times 10^3$ kg m^{-3}. Moreover, let us assume that the tension is 10^3 N. This gives $\epsilon \approx 3 \times 10^{-4}$ which is quite small. As a first approximation we can thus neglect the bending stiffness of the string, however, for more detailed analysis concerning piano tuning we should take it into account. □

The proper use of scaling requires a lot of practice. Especially with complex problems where scales are either not known or too many to just check every possible combination. Being as close as possible to reality is a very good road sign showing the correct route of reasoning. This has a tremendous advantages: we reduce the number of parameters leaving only these combinations that are physically meaningful, and make possible to compare different terms that directly leads to one of the most useful techniques of applied mathematics to be discussed in the next section.

3 Perturbation and asymptotic theory

Convergent series are overadvertised.

In this section we will present an introduction to perturbation theory which proved to be one of the most useful and versatile analytical tool in analysing models. It is indispensable in fluid mechanics, quantum chemistry, gravitation, quantum field theory, and many other - almost all - fields of science. It lets us to study the influence of a small (or large) term in a given equation. In many important cases, the model describing some real-world problem is almost linear. This means that the nonlinearity is relatively small[6]. It acts as a kind of *perturbation* in a sense that we perturb the previously linear system with some small term. Usually this perturbation does not have to be very small for the theory to be effective. There is a systematic way of treating this kind of a problem and we will learn about it.

This is extensively broad subject and itself worth a course of two. There are many very good books that treat many aspects of perturbation theory. Here are some of them.

1. M. Holmes - Introduction to Perturbation Methods, Springer
2. J. Murdock - Perturbations: theory and methods, SIAM
3. C. Bender, S. Orszag - Advanced mathematical methods for scientists and engineers, Springer

3.1 Regular perturbations

We will start with the simplest perturbation possible - found in algebraic equations. Despite the simplicity, they nicely illustrate the general concept and difficulties that may be found in more complex problems.

Example. (*A simple quadratic*) Let us begin with a simple example of finding roots of the following quadratic

$$x^2 + 2\epsilon x - 1 = 0. \tag{3.1}$$

One may ask why to bother in solving this example since every high school student can do it blindly. The solutions, of course, are

$$x_\pm = -\epsilon \pm \sqrt{1+\epsilon^2}. \tag{3.2}$$

Now, this example serves a very important purpose. We can learn what happens to the equation when $\epsilon \to 0^+$ and this is the limit that we will investigate.

First, we can notice that for small ϵ the $2\epsilon x$ term serves as a perturbation. We can write it in a form

$$x^2 - 1 = -2\epsilon x, \tag{3.3}$$

which facilitates the graphical solution: these are the points where $x^2 - 1$ and $-2\epsilon x$ intersect. This is presented on Fig. 17. We immediately see that there are two solutions

[6]Note that thanks to scaling we already know what "relatively" means.

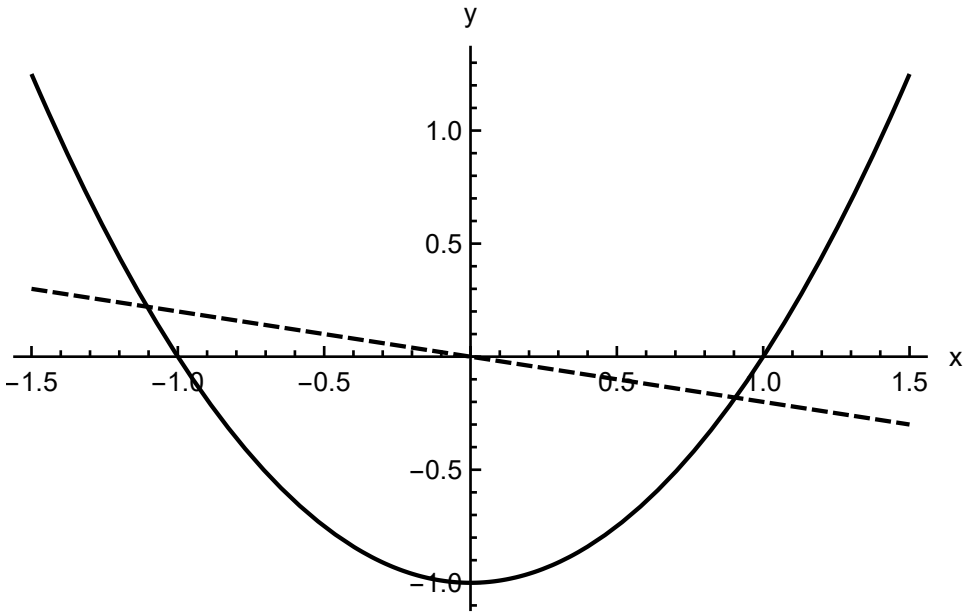

Figure 4: A graphical solution of the quadratic.

for every $\epsilon > 0$ and even this holds for $\epsilon \to 0^+$. This is what makes the case *regular* - the number of solutions of the reduced problem, i.e. for vanishing ϵ, is the same as in the original one.

Note also that for $\epsilon \to 0^+$ the problem is trivial to solve. Usually we want something similar. The reduced problem should be somehow trackable in order to utilize perturbations[7]. We would like to study what happens with solutions when ϵ is a small positive quantity. More specifically, we want to determine the correction to the reduced solution for $\epsilon \to 0^+$. This is easily done by expanding the solution $x_\pm$ into Taylor series

$$x_\pm = \pm 1 - \epsilon \pm \frac{1}{2}\epsilon^2 \mp \frac{1}{8}\epsilon^4. \tag{3.4}$$

This makes sense since ϵ is small and certainly can be made smaller than 1 so the expansion is convergent. However, this is rather a circular argument since usually we do not know the solution. We would not need perturbation theory if we knew it!

Instead, we assume that the solution can be expanded into some appropriate series

$$x = x_0 + \epsilon x_1 + \epsilon^2 x_2 + ..., \tag{3.5}$$

where x_i are unknown numbers independent on ϵ. Notice that it is completely not obvious that the above series has to have integer powers of ϵ. Usually, they should also be found as an part of the solution. When we plug (3.5) into our equation (3.1) we obtain

$$\left(x_0 + \epsilon x_1 + \epsilon^2 x_2 + ...\right)^2 + 2\epsilon(x_0 + \epsilon x_1 + \epsilon^2 x_2 + ...) - 1 = 0. \tag{3.6}$$

Now, we have to collect various powers of ϵ and in order to do that we have to expand the square

$$\left(x_0 + \epsilon x_1 + \epsilon^2 x_2 + ...\right)\left(x_0 + \epsilon x_1 + \epsilon^2 x_2 + ...\right) = x_0^2 + 2\epsilon x_0 x_1 + \epsilon^2\left(2x_0 x_2 + x_1^2\right) + ... \tag{3.7}$$

[7]It does not have to be trivial, though.

Then,

$$x_0^2 - 1 + 2\epsilon\,(x_0 + x_0x_1) + \epsilon^2\left(2x_1 + 2x_0x_2 + x_1^2\right) + \ldots = 0. \tag{3.8}$$

Note the above is a power series that is identically equal to 0. Therefore, all coefficients have to vanish, that is

$$\begin{cases} \epsilon^0: & x_0^2 - 1 = 0, \\ \epsilon^1: & x_0 + x_0x_1 = 0, \\ \epsilon^2: & 2x_1 + 2x_0x_2 + x_1^2 = 0. \\ \ldots & \end{cases} \tag{3.9}$$

A very important observation is that the above infinite system of equations can be solved iteratively one after another. This happens in all regular perturbation problems. Solving, we obtain

$$x_0 = \pm 1, \quad x_1 = -1, \quad x_2 = \pm\frac{1}{2}, \ldots \tag{3.10}$$

and the expansion is

$$x \sim \pm 1 - \epsilon \pm \frac{1}{2}\epsilon^2 + \ldots \quad \text{as} \quad \epsilon \to 0^+, \tag{3.11}$$

which coincides with (3.4). Notice that, here, we do not know anything about the convergence of the above expansion. Hence, instead of writing " = " we use " ∼ " and indicate that $\epsilon \to 0^+$. In the next subsection we will make this precise and rigorous. Finally, note that the error that we make by truncating the approximation after x_2 is proportional to ϵ^3 or higher powers. This is a very useful knowledge. □

There are certain points that follow from the above simple example into general regular perturbation case. We always follow the given steps.

1. Scale the problem into nondimensional one.
2. Identify the small parameter (if there is a large one, say λ, put $\epsilon = \lambda^{-1}$).
3. Expand the unknown solution into a power series with respect to ϵ. Note that here, we may be forced to use Taylor expansions and other techniques.
4. Plug the above into the equation and compare the terms of respective powers of ϵ to obtain a system of equations.
5. Solve the system iteratively as far as it is needed for required accuracy or limited by computational power.

Usually, obtaining very high terms requires a lot of computing power with symbolic manipulation environments. In many cases, however, one or two terms are sufficient to work with.

Example. (*Transcendental equation*) A more interesting example is the following transcendental equation

$$10x = e^x, \tag{3.12}$$

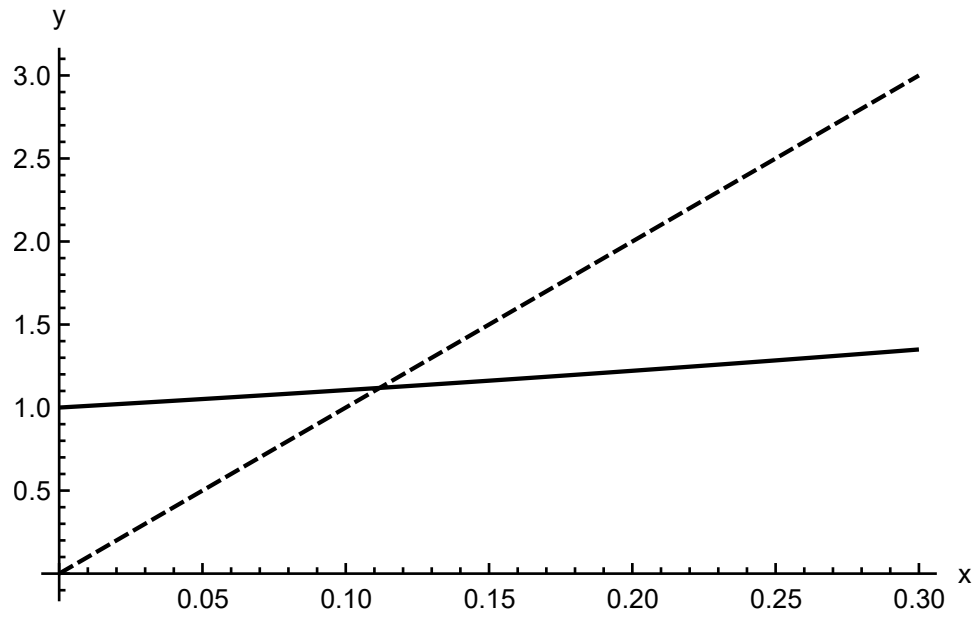

Figure 5: A graphical solution of a transcendental equation.

presented on Fig. 5.

There is no ϵ in the above equation and there seems to be no small parameter. To see what can be done first change the variables $y = 10x$ which gives

$$ye^{-\frac{y}{10}} = 1. \tag{3.13}$$

Now, we can introduce an *artificial* parameter ϵ and consider the following generalization

$$ye^{-\epsilon y} = 1. \tag{3.14}$$

For $\epsilon \to 0^+$ the above has a trivial solution $y = 1$. We would like to find the corrections and expand

$$y = 1 + \epsilon y_1 + \epsilon^2 y_2 + ... \tag{3.15}$$

And by plugging into above we obtain

$$(1 + \epsilon y_1 + \epsilon^2 y_2 + ...)e^{-\epsilon(1+\epsilon y_1+\epsilon^2 y_2+...)} = 1. \tag{3.16}$$

Before we will be able to compare terms with different powers of ϵ we have to expand the exponential

$$e^{-\epsilon y} = e^{-\epsilon(1+\epsilon y_1+\epsilon^2 y_2+...)} = 1 - \epsilon + \left(\frac{1}{2} - y_1\right)\epsilon^2 + ... \tag{3.17}$$

Our equation now becomes

$$1 + (y_1 - 1)\epsilon + \left(\frac{1}{2} - 2y_1 + y_2\right)\epsilon^2 + ... = 1. \tag{3.18}$$

Immediately we have $y_1 = 1$ and $y_2 = 3/2$. We can thus form an approximation

$$y = 1 + \epsilon + \frac{3}{2}\epsilon^2 + ... \tag{3.19}$$

which becomes

$$x = \frac{1}{10}\left(1 + \epsilon + \frac{3}{2}\epsilon^2 + ...\right). \tag{3.20}$$

Now, initially $\epsilon = 0.1$ for which the above approximation gives $x = 0.1115$ while the exact solution is $x = 0.1118$ which differs only at forth decimal place. □

The main application of perturbation theory is in differential equations. Let us, once again, return to the projectile problem.

Example. (*Projectile problem revisited again*) Our scaled problem (2.79) is

$$x'' = -\frac{1}{(1+\epsilon x)^2}, \quad x(0) = 0, \quad x'(0) = 1. \tag{3.21}$$

Let us begin with proposing a *formal* expansion.

$$x(t) \sim x_0(t) + \epsilon x_1(t) + \epsilon^2 x_2(t) + ... \tag{3.22}$$

In order to compare different terms we have to expand the fraction

$$\begin{aligned}\frac{1}{(1+\epsilon x)^2} &= 1 - 2\epsilon x + 3\epsilon^2 x^2 + ... \sim 1 - 2\epsilon\left(x_0 + \epsilon x_1 + \epsilon^2 x_2 + ...\right) + 3\epsilon^2\left(x_0 + \epsilon x_1 + \epsilon^2 x_2 + ...\right)^2 + ... \\ &= 1 - 2\epsilon x_0 + (3x_0^2 - 2x_1)\epsilon^2 + (-4x_0^3 + 6x_0x_1 - 2x_2)\epsilon^3 + ...\end{aligned} \tag{3.23}$$

Thanks to that, the differential equation becomes

$$x_0'' + \epsilon x_1'' + \epsilon x_2'' + ... = -1 + 2\epsilon x_0 - (3x_0^2 - 2x_1)\epsilon^2 + ... \tag{3.24}$$

It is also crucial, and frequently forgotten by beginners, to expand the initial conditions

$$\begin{cases} x_0(0) + \epsilon x_1(0) + \epsilon^2 x_2(0) + ... = 0, \\ x_0'(0) + \epsilon x_1'(0) + \epsilon^2 x_2'(0) + ... = 1. \end{cases} \tag{3.25}$$

Since the above must be valid for every ϵ, we have

$$x_i(0) = 0, \quad x_0'(0) = 1, \quad x_j'(0) = 0, \quad i \geq 0, \quad j \geq 1. \tag{3.26}$$

Therefore, all initial positions vanish while only the zeroth approximation has a non-vanishing derivative. Further, from the ODE, by comparing the terms with powers of ϵ, we infer that

$$\begin{cases} \epsilon^0 : x_0'' = -1, \\ \epsilon^1 : x_1'' = 2x_0, \\ \epsilon^2 : x_2'' = -3x_0^2 + 2x_1, \\ \cdots \end{cases} \tag{3.27}$$

and again a triangular structure is evident. The solution of the leading order term is

$$x_0(t) = t - \frac{1}{2}t^2. \tag{3.28}$$

Hence, the next equation $x_1'' = 2(t - \frac{1}{2}t^2)$ is easily integrated to give

$$x_1(t) = \frac{1}{12}\left(4t^3 - t^4\right). \tag{3.29}$$

Further, doing the analogous steps we have

$$x_2(t) = \frac{1}{360}\left(-90t^4 + 66t^5 - 11t^6\right), \tag{3.30}$$

and we can carry this as far as we want to. The approximation is then

$$x_0(t) \sim t(1 - \frac{1}{2}t) + \frac{\epsilon}{12}t^3\,(4 - t) + \frac{\epsilon^2}{360}\left(-90t^4 + 66t^5 - 11t^6\right) + \ldots \quad \epsilon \to 0^+. \tag{3.31}$$

As we can see, the ϵ-terms are subsequent corrections to the particle position due to nonlinear gravitational field. We can quantitatively see the accuracy of the derived approximation on Fig. 6. We can see that even for rather large $\epsilon = 0.1$ the two-term approximation, i.e. $x_0 + \epsilon x_1$ is very accurate[8]. The graphs are almost indistinguishable with absolute errors 2×10^{-3} for $\epsilon = 0.1$ and 2×10^{-5} for $\epsilon = 0.01$ being even smaller than anticipated. Therefore, for developing an accurate model of low projectiles, this could serve as a great approximation thanks to its simplicity. Notice how much can be said when the exact analytic form of the solution is not known.

Example. (*Space station*) One of the origins of perturbation theory is solving various problems for gravitating particles. Engineers designing spacecrafts, rockets, and shuttles has to accurately determine various trajectories, orbits, and their stability. Consider a satellite or International Space Station orbiting Earth. We know from Physics I that the equation for its orbit in polar coordinates centred at the Earth is

$$r'' - r\,(\theta')^2 = -\frac{GM}{r^2}, \quad \frac{1}{r}\left(r^2\theta'\right)' = 0. \tag{3.32}$$

Here, the prime denotes the derivative with respect to time. The second equation is the consevration of angular momentum which immediately can be integrated to give Kepler's second law

$$r^2\theta' = k. \tag{3.33}$$

We will investigate the stability of a circular orbit. That is, we assume that the space station is perturbed from its original trajectory. Its ultimate fate is what interests us - will it return to its original motion or be fired away to outer space?

Suppose that there is a small perturbation in the radial velocity equal to a constant ϵv (this may happen if an asteroid strikes the hull of the station). We would like to see

[8]Recall that on Earth $\epsilon \approx 10^{-8}v$.

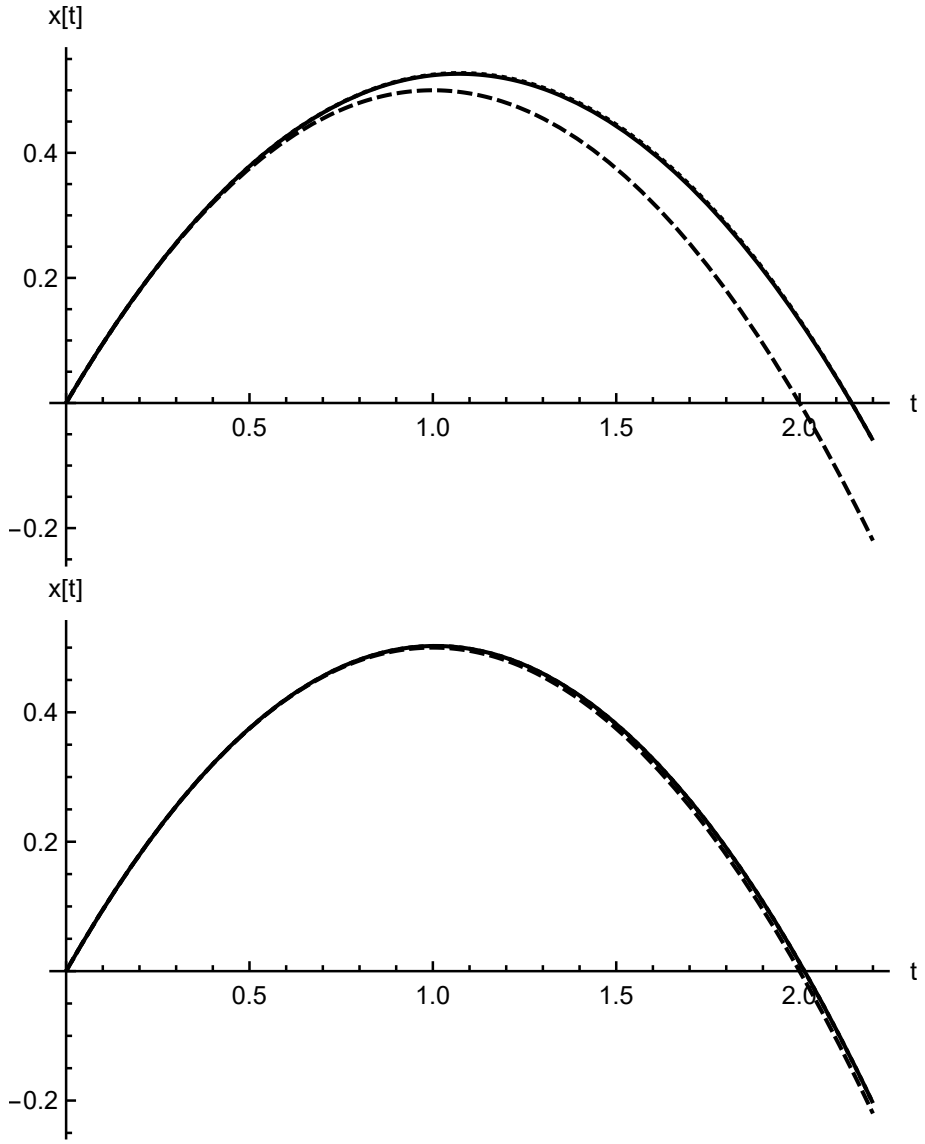

Figure 6: Perturbation approximation of the solution to the projectile problem. On top: $\epsilon = 0.1$, bottom: $\epsilon = 0.01$. Here, solid line is the numerical solution while dashed line represents 1 term and dotted line 2 term approximation.

how the circular orbit, $r(t) = R$ and $\theta(t) = \omega t$ with $R^3\omega^2 = GM$ be affected by this. To this end, suppose that the actual solution expands into

$$r(t) \sim R + \epsilon r_1(t) + ..., \quad \theta(t) \sim \omega t + \epsilon\theta_1(t) + ..., \quad \epsilon \to 0^+. \tag{3.34}$$

Therefore, our ODEs take the form

$$r_1'' + ... - (\epsilon r_1 + ...)(\omega + \epsilon\theta_1' + ...)^2 = -\frac{GM}{(R + \epsilon r_1 + ...)^2}, \tag{3.35}$$

and

$$(R + \epsilon r_1 + ...)^2(\omega + \epsilon\theta_1' + ...) = k. \tag{3.36}$$

By doing the usual expansions in terms of ϵ with the use of $R^3\omega^2 = GM$ we can obtain equations for the correction

$$r_1'' - \omega^2 r_1 = 2R\omega\theta_1', \quad \theta_1' = -\frac{2\omega}{R} r_1. \tag{3.37}$$

Eliminating θ_1 we obtain

$$r_1'' + \omega^2 r_1 = 0. \tag{3.38}$$

This is a good news since we know that the solution of the above is composed of trigonometric functions

$$r_1(t) = C\cos\omega t + D\sin\omega t. \tag{3.39}$$

The amplitude of the above $\sqrt{C^2 + D^2}$ does not change and hence, the space station will oscillate around the circular orbit. The station is thus neutrally stable. This could have been anticipated due to the conservative motion of the system. If the station had some emergency propellers and stabilizers it would return to the original orbit. Since the perturbation is a trigonometric function with the same frequency as the original motion, the new orbit will be elliptical with Earth in its focus (see Fig. 7). □

Example. (*Eigenvalues*) In certain boundary value problems with Robin conditions the following equation determines eigenvalues

$$x\tan x = 1. \tag{3.40}$$

Since $\tan x$ is π-periodic the above has countably many solutions. By inverting we obtain

$$x = n\pi + \arctan\frac{1}{x}, \tag{3.41}$$

where $n \in \mathbb{N}$ and inverse tangent is in its principal branch. We can treat it as a small correction for large x since $|\arctan(1/x)| \leq \pi/2$. Then, we can expect that

$$x \approx n\pi \quad \text{for} \quad x \to \infty. \tag{3.42}$$

This corresponds to a function $1/x$ intersecting the $\tan x$ near its zero (see Fig. 8). We would like to improve this result. A very useful technique is to iterate the algebraic equation. That is, since we have found that $x \approx n\pi$ for large n we can go back to (3.41) and write

$$x = n\pi + \arctan\frac{1}{n\pi} \approx n\pi + \frac{1}{n\pi} \quad \text{for} \quad x \to \infty. \tag{3.43}$$

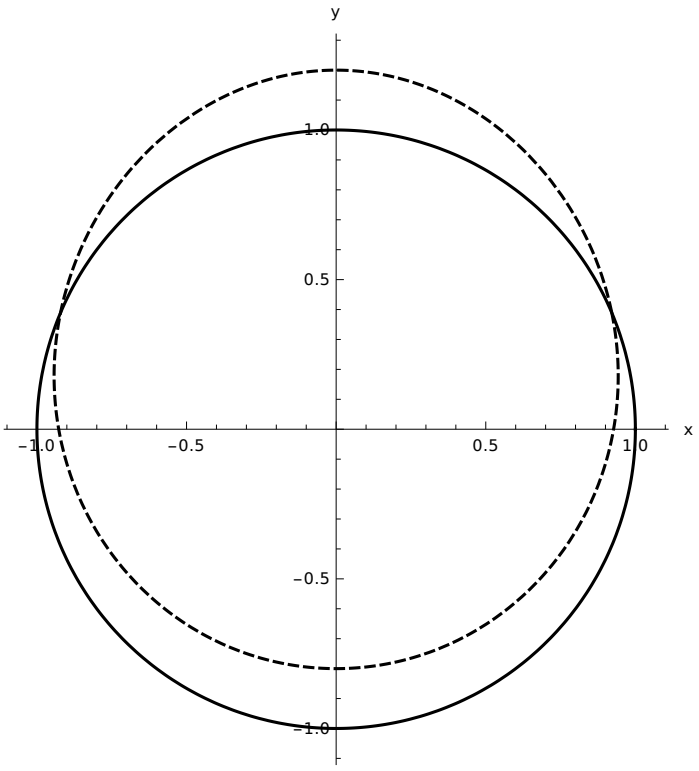

Figure 7: Perturbation approximation of the solution to the space station problem. The solid line is the original circular orbit while dashed line represents perturbed elliptical trajectory.

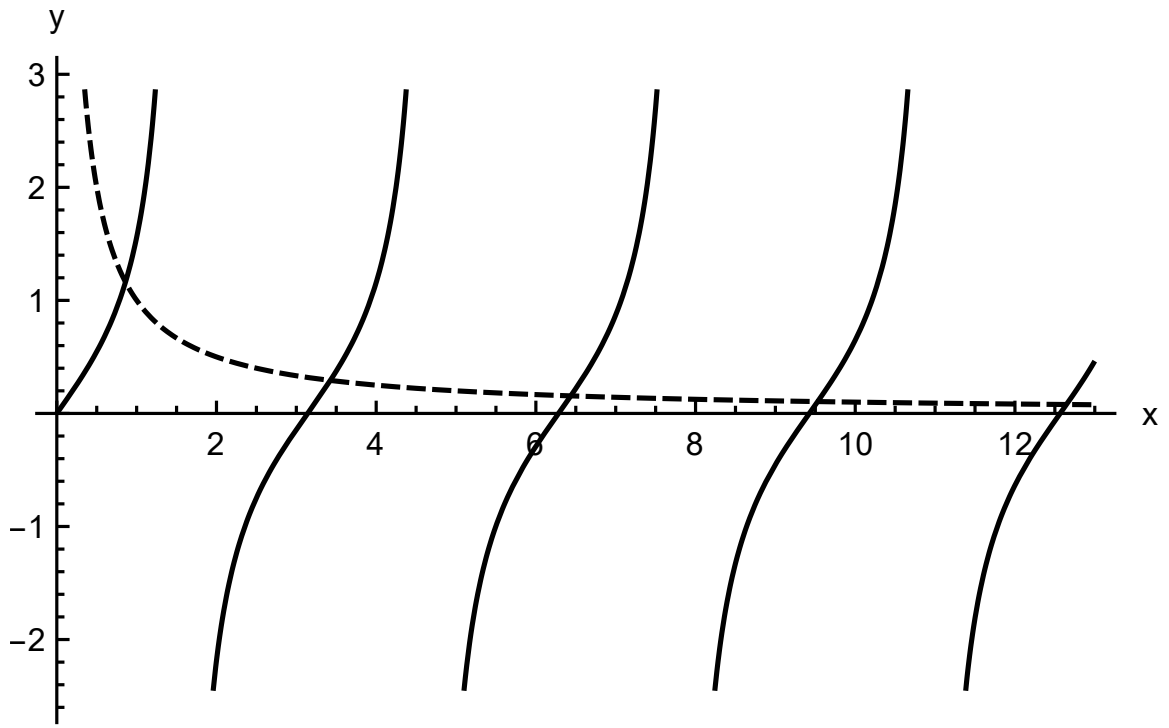

Figure 8: Graphical solution of equation $x \tan x = 1$.

We can carry this further by using Taylor expansion

$$\arctan \frac{1}{x} = \frac{1}{x} - \frac{1}{3x^3} + \frac{1}{5x^5} + ... \tag{3.44}$$

and iterate as long as we want. We would then obtain

$$x = n\pi + \frac{1}{n\pi} - \frac{4}{3(n\pi)^3} + ... \tag{3.45}$$

This is a superb approximation. For example, even for $n = 1$ the exact zero is 3.4256 while the approximation gives 3.4169. Next, $n = 2$ yields exact zero 6.4373 while the approximation is 6.4369 with three correct decimal digits. For $n = 5$ we have an accuracy of 10^{-4} and so on. We remark that all of the above results were obtained under the assumption of large n which these considered are certainly not! We obtained the accuracy for free. □

The perturbation theory us also an indispensable tool in analysis of partial differential equations. We will return to this topic in subsequent sections.

3.2 Asymptotic series

Before we proceed further we have to rigorously define what we mean by various asymptotic objects. In other case, without any mathematical theory, we could make many mistakes without knowing it. We start with recalling the order symbols which denote the relative sizes of different quantities.

We have to formalize the statements such that, for example, ϵ goes to zero at the same rate as $\sin\epsilon^2/\epsilon$, or ϵ^2 vanishes quicker than ϵ, as $\epsilon \to 0^+$. These can be made quantitatively by using the o-notation introduced by Bachmann and popularized by Landau in the very beginning of the XX century. Nowadays, order notation is widely used in applied mathematics and computer science. It has the great advantage to hide unnecessary details yet allow operations and provide insight into the structure of various formulas.

Definition 1. *1. The function $f(\epsilon)$ has the same growth rate as $g(\epsilon)$ as $\epsilon \to 0^+$ when there exist constants M and $\epsilon_0 > 0$ such that*

$$|f(\epsilon)| \leq M|g(\epsilon)| \quad \textit{for} \quad 0 < \epsilon < \epsilon_0. \tag{3.46}$$

We then write

$$f = O(g) \quad \textit{as} \quad \epsilon \to 0^+, \tag{3.47}$$

and say "f is big Oh of g" as $\epsilon \to 0^+$.

2. The function $f(\epsilon)$ vanishes much faster than $g(\epsilon)$ as $\epsilon \to 0^+$ when for every δ there exists $\epsilon_1 > 0$ such that

$$|f(\epsilon)| \leq \delta|g(\epsilon)| \quad \textit{for} \quad 0 < \epsilon < \epsilon_1. \tag{3.48}$$

We then write

$$f = o(g) \quad \textit{as} \quad \epsilon \to 0^+, \tag{3.49}$$

and say "f is small oh of g" as $\epsilon \to 0^+$. We also use the notation $f \ll g$.

Remark 1. *In the above definition we can put $x = \epsilon^{-1}$ and consider orders of magnitude for $x \to \infty$.*

Therefore, since

$$\left|\frac{\sin\epsilon^2}{\epsilon}\right| = \left|\frac{\sin\epsilon^2}{\epsilon^2}\right| \epsilon \leq \epsilon, \tag{3.50}$$

we have $\sin\epsilon^2/\epsilon = O(\epsilon)$ as $\epsilon \to 1$. Here, $M = 1$ and, say, $\epsilon_0 = 1$. Similarly, $\epsilon^2 = o(\epsilon)$ in the same limit because

$$\epsilon^2 \leq \delta\epsilon \tag{3.51}$$

for any $\delta > 0$ and $\epsilon < \epsilon_1 = \delta$. These definitions are useful in showing the relative order is some pathological cases. very frequently, however, we can use simpler criterion based on the limits.

Proposition 1. *Let*

$$\left|\lim_{\epsilon\to 0^+} \frac{f(\epsilon)}{g(\epsilon)}\right| = L, \tag{3.52}$$

where we assume that the limit exists but may be infinite.

1. *If $L > 0$ is finite, then $f = O(g)$ as $\epsilon \to 0^+$. Moreover, if $L = 1$ we can write $f \sim g$ in the same limit (we pronounce "f twiddles g").*

2. *If $L = 0$, then $f = o(g)$ as $\epsilon \to 0^+$.*

3. *If* $L = \infty$, *then* $g = o(f)$ *as* $\epsilon \to 0^+$.

Proof. The proof is elementary and follows from the definition of the limit. We leave it as an exercise in Calculus I. □

Examples.

1. Let $f(\epsilon) = \epsilon \sin(1 + 1/\epsilon)$. Notice that we cannot use the above proposition since the limit does not exist. We have to use the original definition

$$|\epsilon \sin(1 + 1/\epsilon)| \leq \epsilon, \tag{3.53}$$

   and $M = 1$ for $\epsilon_0 = 1$.

2. A function $f = f(\epsilon)$ is bounded as $\epsilon \to 0^+$ in and only if[9] $f = O(1)$ as $\epsilon \to 0^+$. For instance, $f(\epsilon) = (1 + \epsilon^2)^{-1}$.

3. A function $f = f(\epsilon)$ vanishes as $\epsilon \to 0^+$ iff $f = o(1)$ as $\epsilon \to 0^+$.

4. If $f = o(g)$ as $\epsilon \to 0^+$, then $f = O(g)$ as $\epsilon \to 0^+$ (take $M = 1$ in the definition).

5. Let $f(\epsilon) = \exp(-1/\epsilon)$. We have

$$\lim_{\epsilon \to 0^+} \frac{e^{-\frac{1}{\epsilon}}}{\epsilon^\alpha} = 0, \tag{3.54}$$

   for any $\alpha > 0$. Therefore the function f vanishes faster than any power of ϵ (beyond all orders). We say that f is *transcendentally small* with respect to ϵ^α as $\epsilon \to 0^+$.

There is some confusion that arise when using the order notation that may annoy some purists. Note that $\sin \epsilon = O(\epsilon)$ as $\epsilon \to 0^+$ but $O(\epsilon) = \sin \epsilon$ does not make any sense! The equals sign is purely formal and customary and does not imply symmetry. As Donald Knuth pointed out, these are "one-sided equalities". Other examples include: $\epsilon^2 = O(\epsilon^2)$ and $\epsilon^2 = O(\epsilon)$ but $\epsilon \neq \epsilon^2$ for $0 < \epsilon < 1$. To be precise, some mathematicians use the set notation and write $f \in O(g)$ as $\epsilon \to 0^+$ and think about classes of functions. We will not use this device since everyone of us is acquainted with programming where the " = " operator is usually the assignment and a text such as $x = x + 1$ does not seem to be a contradiction.

Having defined the order notation we can give a precise meaning of asymptotic series we have found before. The main motif is: accuracy versus convergence. We would like to answer a question whether we really need convergent series to approximate real-world phenomena. To this end, we need to quantify when a given function is well-approximated by some simpler form. The requirement that the error goes to zero is not enough as the simple example shows. Take $f(\epsilon) = \epsilon^2 + \epsilon^{2020}$. Since $\epsilon^{2020} \ll \epsilon^2$ we can take $f(\epsilon) \approx \epsilon^2$. But also we could have taken $f(\epsilon) \approx \epsilon^2/2$ since the error $f(\epsilon) - \epsilon^2/2 \to 0+$ as $\epsilon \to 0^+$. The second approximation is of course much worse than the first, since we have forgot to account for the relative smallness of the error. In the first case, we have $f(\epsilon) - \epsilon^2 = O(\epsilon^{2020})$ which is extremely small. In the second case, $f(\epsilon) - \epsilon^2 = O(\epsilon^2)$ which is of the same order as the approximation. We can now give a formal definition.

[9] Halmos introduced the abbreviation "iff".

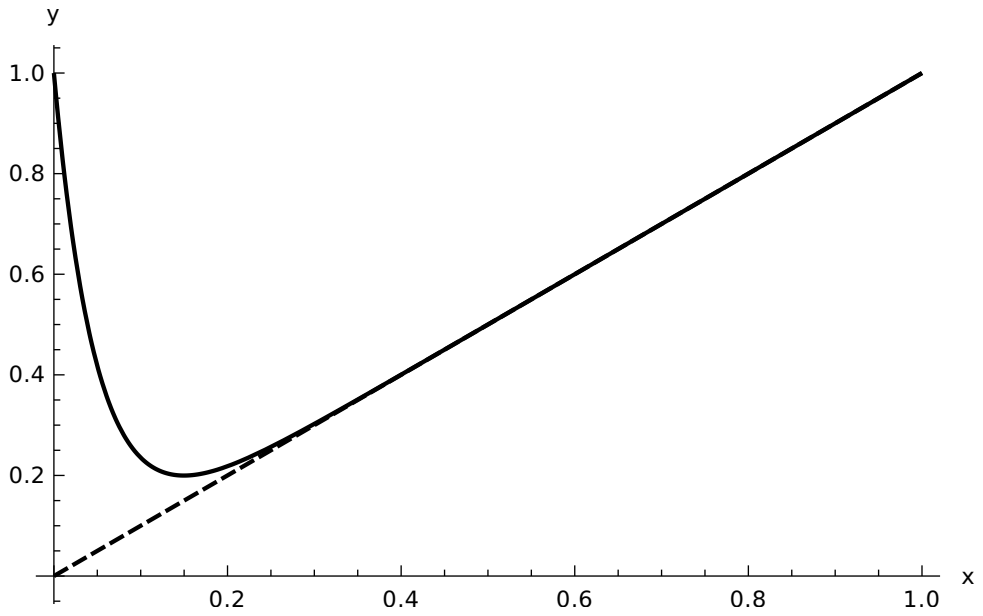

Figure 9: A function x (dashed line) as an asymptotic approximation to $x + e^{-x\epsilon}$ for $\epsilon = 0.05$.

Definition 2. *A function* $\phi(\epsilon)$ *is an* asymptotic approximation *to* $f(\epsilon)$ *as* $\epsilon \to 0^+$ *when* $f \sim \phi$, *that is* $f(\epsilon) = \phi(\epsilon) + o(\phi)$ *as* $\epsilon \to 0^+$.

Examples.

1. A natural source of asymptotic expansions is the Taylor series. For example, let $f(\epsilon) = e^{-\epsilon}$. Then,

$$f(\epsilon) = 1 - \epsilon + \frac{1}{2}\epsilon^2 - \frac{1}{6}\epsilon^3 + O(\epsilon^4), \quad \epsilon \to 0^+. \tag{3.55}$$

   Subsequent terms are asymptotic approximations to the exponential with increasing accuracy. Note that, we also have a strange and lousy looking expression as $f(\epsilon) \sim \cos\epsilon$ since

$$e^{-\epsilon} - \cos\epsilon = -\epsilon + \epsilon^2 + O(\epsilon^3), \quad \epsilon \to 0^+. \tag{3.56}$$

   Nonuniqueness is evident. We will correct it below.

2. Let $f(x) = x + e^{-x/\epsilon}$ where $x \in (0, 1)$ is fixed. For small ϵ we have $f \sim x$. We would like, however, to see how does this approximation is valid for all x. We immediately see that this cannot be a uniform approximation since $f(0) = 1$ for every $\epsilon > 0$. The plot of the situation is depicted on Fig. 9. Note that there is a relatively large region when the approximation is very accurate. It blows up near $x = 0$ for fixed ϵ. This shows that in order to obtain a uniform approximation we somehow have to relate ϵ to x: the closer x to 0 the smaller the ϵ. This is an example of the so-called boundary layer which we will soon meet.

In order to obtain a useful asymptotic approximation of various functions we have to deal with the nonuniqueness and somehow to control the accuracy. The main idea was introduced by Henri Poincare and we follow his definition.

Definition 3. 1. *A set of functions* $\{\phi_n(\epsilon)\}_n$ *forms an* asymptotic sequence *if* $\phi_{n+1} = o(\phi_n)$ *as* $\epsilon \to 0^+$ *for all* n. *We say that* ϕ_n *are* gauge functions.

2. *If* $\{\phi_n(\epsilon)\}_n$ *is an asymptotic sequence, then* $f(\epsilon)$ *has an* asymptotic expansion *with respect to* $\{\phi_n\}_n$ *when for each* n *there exists a constant* a_n *such that*

$$f(\epsilon) = \sum_{i=0}^{n} a_i \phi_i(\epsilon) + o(\phi_n(\epsilon)) \quad as \quad \epsilon \to 0^+. \tag{3.57}$$

In this case we write that f *has an* asymptotic series

$$f(\epsilon) \sim \sum_{i=0}^{\infty} a_i \phi_i(\epsilon) \quad as \quad \epsilon \to 0^+. \tag{3.58}$$

In the above definitions, functions f *and* ϕ_n *can also depend on other variables which are assumed to be* fixed.

There are many possible choices for gauge functions, power $\phi_n(\epsilon) = \epsilon^n$ or exponential $\phi_n(\epsilon) = e^{-n/\epsilon}$ are particular examples. Therefore, an asymptotic series provides an approximation to f when the error of truncation is of smaller order than every retained term for other variables fixed. Note that these expansions does not have to be convergent in the usual sense such that we fix x and let $n \to \infty$. This can be summarized as follows

$$\begin{aligned} &\textbf{Covergent:} \quad \sum_{i=n+1}^{\infty} a_i \phi_i(x, \epsilon) \to 0 \quad \text{as} \quad n \to \infty, \quad \epsilon \text{ fixed}, \\ &\textbf{Asymptotic:} \quad \sum_{i=n+1}^{\infty} a_i \phi_i(x, \epsilon) = o(\phi_n(x, \epsilon)) \quad \text{as} \quad \epsilon \to 0^+, \quad n \text{ fixed}. \end{aligned} \tag{3.59}$$

We can see that the main issue here concerns the exchange of limits. In the asymptotic series the remainder does not have to vanish for $n \to \infty$. As a matter of fact, most useful and accurate series are not convergent. Abel spoke of them as "invention of the devil".

Example. (*Taylor series*) A good example of a convergent and asymptotic series is Taylor series. For if $f = f(x)$ is sufficiently smooth then

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2} f''(x_0)(x - x_0)^2 + \ldots + \frac{1}{n!} f^{(n)}(x_0)(x - x_0)^n + \frac{1}{(n+1)!} f^{(n+1)}(\xi)(x - x_0)^{n+1}, \tag{3.60}$$

where $\xi = \xi(x_0)$ is some constant. Because f is smooth, and hence bounded, we have

$$\frac{1}{(n+1)!} f^{(n+1)}(\xi)(x - x_0)^{n+1} = o((x - x_0)^n) \quad \text{as} \quad x \to 0. \tag{3.61}$$

Therefore, retaining first three terms in the above Taylor series constitutes an asymptotic expansion and we can write

$$f(x) \sim \sum_{i=0}^{n} \frac{f^{(i)}}{i!} (x - x_0)^i \quad \text{as} \quad x \to 0, \tag{3.62}$$

and we further know that if certain conditions on the smoothness of f are met, the above series is also convergent in the classical sense. □

Example. (*Stielties function*) This is probably one of the most famous examples of divergent series that are very accurate. Consider the Stielties function, closely related with exponential integral, which arises in radiative transfer, heat convection, groundwater flow, neutron transport, and many other places in science and engineering

$$S(\epsilon) = \int_0^\infty \frac{e^{-t}}{1+\epsilon t} dt. \tag{3.63}$$

In order to find the asymptotic expansion of the above we use the geometric series valid for arbitrary n

$$\frac{1}{1+\epsilon t} = \sum_{i=0}^{n} (-\epsilon t)^i + \frac{(-\epsilon t)^{n+1}}{1+\epsilon t}, \tag{3.64}$$

which can be put inside the integral and after interchanging order of summation and integration (since we are dealing with a finite series)

$$S(\epsilon) = \sum_{i=0}^{n} (-\epsilon)^i \int_0^\infty e^{-t} t^i dt + E_n(\epsilon), \tag{3.65}$$

where the error is

$$E_n(\epsilon) = (-\epsilon)^{n+1} \int_0^\infty \frac{e^{-t} t^{n+1}}{1+\epsilon t} dt. \tag{3.66}$$

The integral in (3.65) can be computed explicitly, since it is a special value of gamma function[10]

$$S(\epsilon) = \sum_{i=0}^{n} (-1)^i \epsilon^i i! + E_n(\epsilon). \tag{3.67}$$

Note that this is an *exact* formula. We have not done any approximations or limit passages. We immediately notice that the above power series is *divergent* for any $\epsilon > 0$! That is, we cannot let $n \to \infty$ and yet obtain a meaningful result. Why is that so? When expanding $(1+\epsilon t)^{-1}$ into Taylor series we have integrated it outside of its domain of convergence. Dealing with only a finite expansion, this is not a crime. However, we are not justified to pass to the limit.

In order to see if (3.67) is an asymptotic expansion we have to estimate the remainder $E_n(\epsilon)$. We have

$$|E_n(\epsilon)| = \epsilon^{n+1} \int_0^\infty \frac{e^{-t} t^{n+1}}{1+\epsilon t} dt \le \epsilon^{n+1} \int_0^\infty e^{-t} t^{n+1} dt = (n+1)! \epsilon^{n+1} \ll \epsilon^n, \quad \epsilon \to 0^+. \tag{3.68}$$

As an illustration let us numerically assess the accuracy of the asymptotic approximation of, say four terms

$$S(\epsilon) \sim 1 - \epsilon + 2\epsilon^2 - 6\epsilon^3 + 24\epsilon^4. \tag{3.69}$$

[10] Recall that $\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx$.

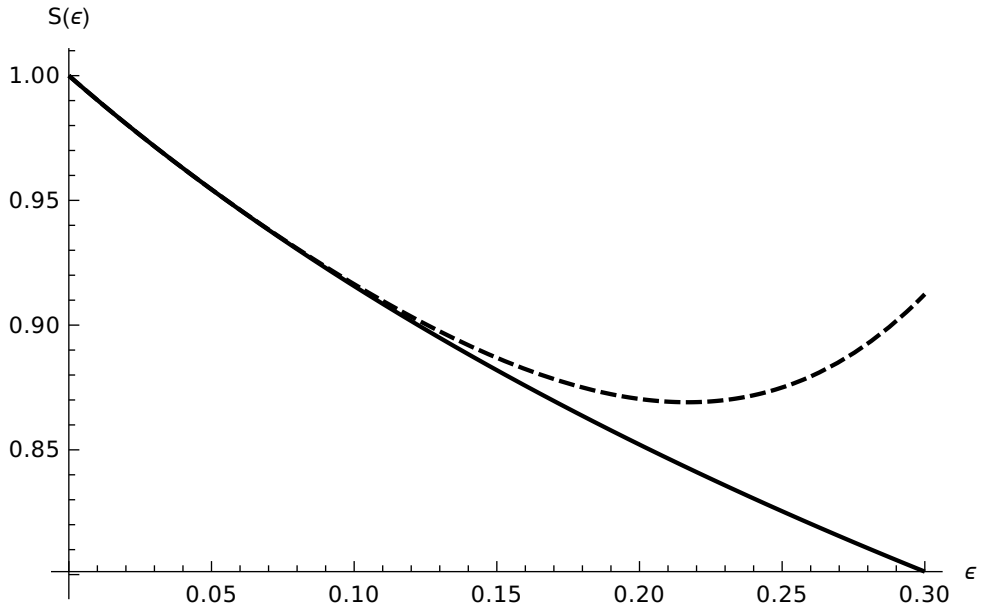

Figure 10: A four term asymptotic approximation to the Stielties function (3.63) for different ϵ.

The graph of the approximation is depicted on Fig. 10. We can see that for small values of ϵ the accuracy is superb. For example with not so small $\epsilon = 1/4$ we have $E_4 \approx 0.05$ while for $\epsilon = 0.1$ we obtain $E_4 \approx 7 \times 10^{-4}$. This is a typical phenomenon meaning that, by definition, taking more terms in the asymptotic expansion we obtain better approximation for $\epsilon \to 0^+$. However, adding these terms contributes to lack of accuracy for fixed ϵ. What usually happens, is that the first terms get smaller for smaller ϵ, then around $O(\epsilon^{-1})$ term they start to increase. This can be seen from the n-th term in the series

$$(-1)^n n! \epsilon^n, \tag{3.70}$$

where with increasing n the factorial will dominate the power of ϵ. It is thus beneficial to truncate the series with respect to ϵ at the largest integer n that is smaller or equal than ϵ^{-1} since the ratio of $n+1$ term to n term in (3.67) is $-n\epsilon$. The resulting expansion is called *superasymptotic*. This can be carried further to *hyperasymptotic* expansions by a careful analysis of the error terms and finding values at which it attains its maximum. This is, however, beyond the scope of our lecture. □

A very useful technique in finding asymptotic expansions is integration by parts.

Example. (*Error function (CDF of normal distribution)*) One of the most useful and important special function in partial differential equations and probability is the error function.

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} \mathrm{d}t. \tag{3.71}$$

This is just a rescaled version of the cumulative distribution function of normal distribution - the most important distribution of all. It is also ever present in investigations of heat conduction in infinite domains. It is thus worth to analyse it thoroughly. Being

a special function it cannot be written as a finite combination of elementary functions leaving asymptotic analysis and numerical methods the only way to find its values with arbitrary accuracy.

The first thing that comes in mind is the Taylor expansion and integration term by term

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \sum_{k=0}^{\infty} (-1)^k \frac{t^{2k}}{k!} dt = \frac{2}{\sqrt{\pi}} \sum_{k=0}^{\infty} \frac{(-1)^k}{k!} \int_0^x t^{2k} dt = \frac{2}{\sqrt{\pi}} \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k+1}}{(2k+1)k!}. \tag{3.72}$$

We anticipate that, due to the nature of Taylor series, the above can be accurate for small x. However, frequently we would like to find the error function for large arguments especially when we would like to estimate the tail of normal distribution. Numerical methods may not be optimal for computing large numbers. It is asymptotic analysis that changes the game.

Note that when $x \to \infty$ the integral in (3.71) converges and $\mathrm{erf}(x) \to 1$. In order to find the rate at which it is approaching that limit we write

$$\mathrm{erf}(x) = 1 - \frac{2}{\sqrt{\pi}} \int_x^{\infty} e^{-t^2} dt. \tag{3.73}$$

Now, we have only to consider the above integral. We would like to integrate by parts but, however, this is not straightforward since each integration would produce a factor of x that will increase the magnitude of the remainder. Then, the expansion would not be asymptotic. Whence, we have to compute the primitive of a function associated with e^{-t^2}. To this end write

$$\int_x^{\infty} e^{-t^2} dt = \int_x^{\infty} \frac{-2te^{-t^2}}{-2t} dt, \tag{3.74}$$

and, integrating by parts, compute the derivative of $1/(-2t)$, and antiderivative of $-2te^{-2t} = (e^{-2t})'$. Thanks to that trick we obtain

$$\int_x^{\infty} e^{-t^2} dt = \frac{e^{-x^2}}{2x} + \frac{1}{2} \int_x^{\infty} \frac{e^{-t^2}}{t^2} dt = \frac{e^{-x^2}}{2x} - \frac{e^{-x^2}}{4x^3} + \frac{3}{4} \int_x^{\infty} \frac{e^{-t^2}}{t^4} dt. \tag{3.75}$$

This can be carried over to obtain

$$\int_x^{\infty} e^{-t^2} dt = \frac{e^{-x^2}}{2x} \sum_{k=0}^{n} (-1)^k \frac{(2k-1)!!}{(2x^2)^k} + R_{n+1}(x), \tag{3.76}$$

where the remainder is

$$R_{n+1}(x) = (-1)^{n+1} \frac{(2n-1)!!}{2^n} \int_x^{\infty} \frac{e^{-t^2}}{t^{2n+1}} dt. \tag{3.77}$$

Each integration by parts brings a odd number to the numerator and a factor of 2 into the denominator of the expansion. Apart from that, the sign alternates. We have to

show that the remainder is asymptotically smaller than the last term in the series as $x \to \infty$. This can be seen by writing

$$\begin{aligned} |R_{n+1}(x)| &\le \frac{(2n-1)!!}{2^n}\frac{1}{x^{2n+1}}\int_x^\infty e^{-t^2}dt \\ &= \frac{(2n-1)!!}{2^{n+1}}\frac{1}{x^{2n+1}}\left(\frac{e^{-x^2}}{x}+\int_x^\infty \frac{e^{-t^2}}{t^3}dt\right) \le \frac{(2n-1)!!}{2^{n+1}}\frac{e^{-x^2}}{x^{2n+1}}\left(\frac{1}{x}+\int_x^\infty \frac{dt}{t^3}\right) \\ &= \frac{(2n-1)!!}{2^{n+1}}\frac{e^{-x^2}}{x^{2n+1}}\left(\frac{1}{x}+\frac{1}{2x^2}\right) = o\left(\frac{e^{-x^2}}{x^{2n+1}}\right) \quad \text{as} \quad x \to \infty. \end{aligned} \tag{3.78}$$

In the second inequality we have moved the power function in front of the integral, then integrated by parts, and finally estimated the exponential of the integrand. We have thus shown that the following asymptotic expansion holds

$$erf(x) \sim 1 - \frac{e^{-x^2}}{x\sqrt{\pi}}\sum_{k=0}^{\infty}(-1)^k\frac{(2k-1)!!}{(2x^2)^k} \quad \text{as} \quad x \to \infty. \tag{3.79}$$

Now, we can compare various numerical results. On Fig. 11 we can see absolute error plots with respect to the number of terms in Taylor (3.72) and asymptotic (3.79). Note that we use two values of x that by no means cannot be classified as large, that is $x = 2$ and $x = 3$. Note the tremendous accuracy of the asymptotic expansion. For $x = 2$ the Taylor series need at least 25 terms in order to match the accuracy of 1 term asymptotic series for which the error is smaller than 10^{-3}. Note also the existence of optimal number of asymptotic terms to provide the least error. For $x = 3$ the performance of asymptotic over Taylor series is even more pronounced by several orders of magnitude. □

We can enumerate several important properties of the asymptotic series. Notice that some of them are not intuitive and have no analogy in convergent power series.

1. *Uniqueness.* For a given set of gauge functions $\{\phi_n(\epsilon)\}$ the asymptotic series $f \sim \sum_i a_i\phi_i$ as $\epsilon \to 0^+$ is uniquely given with coefficients

$$a_0 = \lim_{\epsilon\to 0^+}\frac{f(\epsilon)}{\phi_0(\epsilon)}, \quad a_n = \lim_{\epsilon\to 0^+}\frac{f(\epsilon)-\sum_{i=0}^{n-1}a_i\phi_i(\epsilon)}{\phi_n(\epsilon)}, \quad n \ge 1. \tag{3.80}$$

2. *Nonuniqueness.* The asymptotic expansion for a given f may be nonunique by the use of different sequences of gauge functions. For example,

$$\begin{aligned} \sin x &\sim x - \frac{1}{6}x^3 + \ldots \\ &\sim \tan x - \frac{1}{2}(\tan x)^3 + \ldots \quad \text{as} \quad x \to 0^+. \end{aligned} \tag{3.81}$$

3. *Arithmetic.* The asymptotic series can be added, subtracted, multiplied, and divided according to the usual rules for series. This result is a direct consequence of the definition.

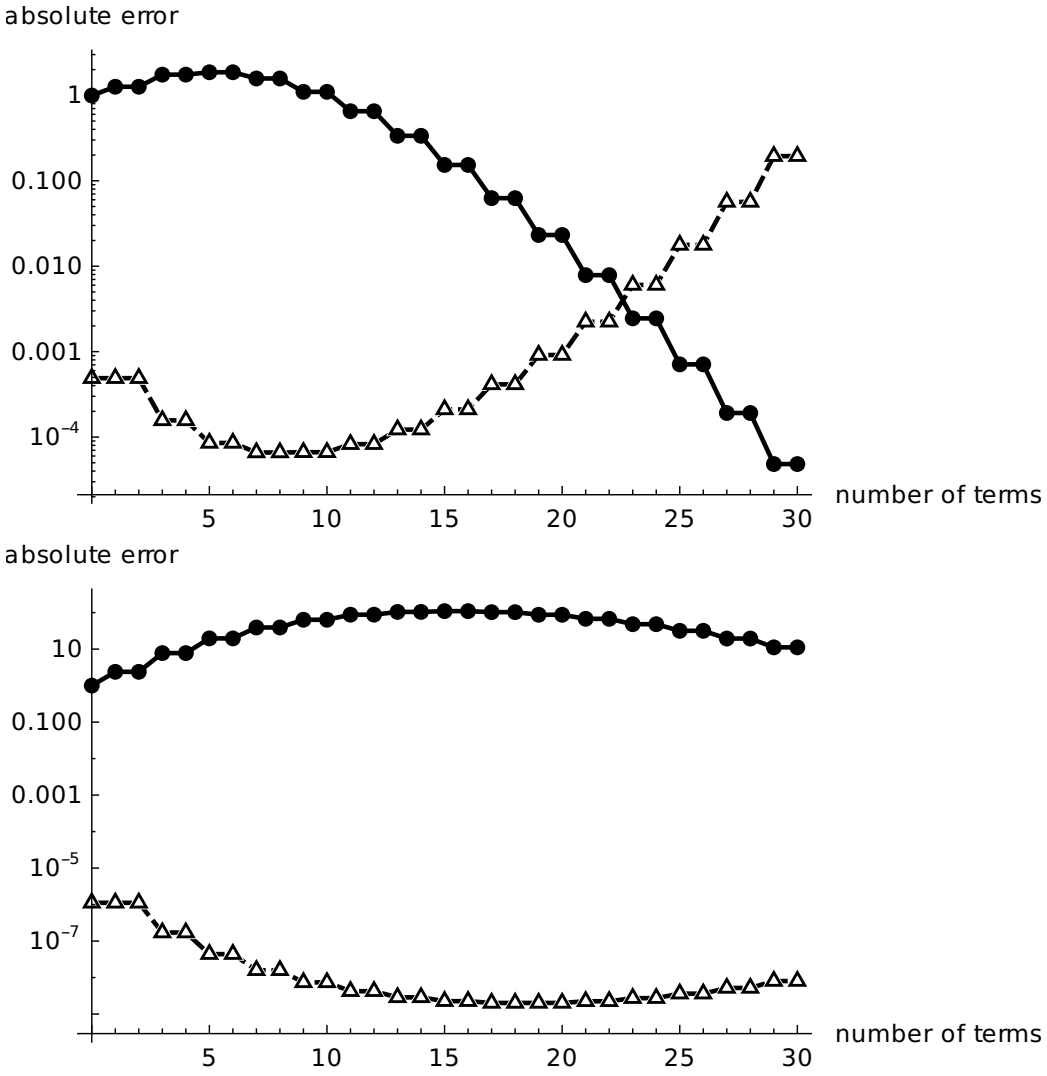

Figure 11: Absolute errors of approximation erf by Taylor (3.72) (solid, circles) and asymptotic (3.79) (dashed, triangles) series for $x = 2$ (top) and $x = 3$ (bottom).

4. *Integration.* The asymptotic *power* series can be integrated term by term with respect to the gauge function argument. That is, when the gauge functions are $\phi_n(\epsilon) = \epsilon^n$ its integral is always $o(\phi_n(\epsilon))$ as $\epsilon \to 0^+$.

   Moreover, the asymptotic series of $f(x, \epsilon)$ can be integrated term by term with respect to x. This immediately follows from the fact that we can integrate finite series.

5. *Differentiation.* In general, asymptotic series *cannot* be differentiated term by term. For instance, let $f(x, \epsilon) = e^{-x/\epsilon} \sin e^{x/\epsilon}$. Then, the expansion in terms of power functions is

$$f(x) \sim 0 + 0 \cdot \epsilon + 0 \cdot \epsilon^2 + \ldots \quad \text{as} \quad \epsilon \to 0^+, \tag{3.82}$$

   since the function is transcendentally small with respect to $\{\epsilon^n\}_n$. However, the derivative is

$$\frac{\partial f}{\partial x}(x, \epsilon) = -\frac{1}{\epsilon} e^{-\frac{x}{\epsilon}} \sin e^{\frac{x}{\epsilon}} + \frac{1}{\epsilon} \cos e^{\frac{x}{\epsilon}}, \tag{3.83}$$

   which does not have the same asymptotic expansion as the zero series given above. This is due to the transcendentally small terms.

   Can we ever differentiate an asymptotic expansion? The answer is - yes - provided the differentiated series is itself asymptotic with respect to the same gauge functions. That is, if

$$f(x, \epsilon) \sim \sum_{i=0}^{n} a_i(x)\phi_i(\epsilon), \tag{3.84}$$

   and

$$\frac{\partial f}{\partial x}(x, \epsilon) \sim \sum_{i=0}^{n} b_i(x)\phi_i(\epsilon), \tag{3.85}$$

   as $\epsilon \to 0^+$ then

$$b_i = \frac{d}{dx} a_i. \tag{3.86}$$

   Wherever we ill differentiate an asymptotic series, we will automatically assume that the series of derivatives is asymptotic.

As we have seen, the optimal number of term needed for an accurate approximation with the asymptotic series depends on the parameter. Moreover, the most useful series are divergent. In applied mathematics we usually use asymptotic expansions in order to learn about physical meaning of the solution. This is almost always the truth - when doing perturbation analysis we obtain several first terms of the asymptotic expansion of the sought solution. Due to the anticipated simplicity of the obtained series, we are able to read physics from the mathematical formulas. Even when the exact analytical solution is available, usually the asymptotic expansion conveys much more meaning and transparency. On the other hand, in real-world examples we are rarely able to obtain more than two or three terms in the expansion due to enormous complexity. Therefore, we cannot even think about determining whether the series is convergent or not. As experience taught generations of applied mathematicians, a solid mathematical reason combined with physical intuition is the best guide in modelling regardless the nature of the asymptotic expansion.

When accuracy is the issue, one usually use high order numerical methods. Usually, asymptotic expansion is frequently used in combination with computer calculations. For example, if we want to compute a value of a certain special function for given argument x we can used the hybrid algorithm. For example, if x is small we use Taylor series when the number of terms is determined by the required accuracy. For large x we set $x = \epsilon^{-1}$ and consider $\epsilon \ll 1$. For medium size x we use numerical methods aided with Padé approximations that are more or less, ratios of Taylor series.

3.3 Asymptotic expansion of integrals

We have seen that asymptotic expanding integrals of some parameters leads to some very useful results. Indeed, many special functions may be represented by integrals and intelligent asymptotic expansion is invaluable aid in their analysis. One way of systematic obtaining such formulas is to use the simple device of integration by parts. There are, as usual, certain situations that it fails. We are going to investigate when it does so and how to repair this.

The class of integrals that we will be analysis is represented by the so-called *Laplace integral*

$$I(x) = \int_a^b f(t)e^{x\phi(t)}dt, \quad x > 0, \tag{3.87}$$

where a, b may be infinite if the integral is uniformly convergent, f and ϕ are given functions, and we consider $x \to \infty$. These integrals arise in waves, optics, special functions, and Laplace transform among other places. Moreover, many integrals can be transformed into (3.87). Taught by the example with error function (3.71) we try to integrate by parts by writing

$$I(x) = \frac{1}{x}\int_a^b \frac{f(t)}{\phi'(t)}\frac{\partial}{\partial t}\left(e^{x\phi(t)}\right)dt = \frac{1}{x}\left[\frac{f(t)}{\phi'(t)}e^{x\phi(t)}\right]_{t=a}^{t=b} - \frac{1}{x}\int_a^b \left(\frac{f(t)}{\phi'(t)}\right)' e^{x\phi(t)}dt. \tag{3.88}$$

By the exactly the same procedure as in the analysis of error function we can show that the integral remainder above is asymptotically smaller than the first term, and hence we prove the following important result.

Lemma 1. *Assume that* $\phi \in C^2[a,b]$, *and* $f \in C[a,b]$. *Moreover, let* $\phi'(t) \neq 0$ *for* $t \in [a,b]$ *and either* $f(a) \neq 0$ *or* $f(b) \neq 0$. *Then, the Laplace integral (3.87) has the following asymptotic expansion*

$$I(x) \sim \frac{1}{x}\left[\frac{f(t)}{\phi'(t)}e^{x\phi(t)}\right]_{t=a}^{t=b} \quad as \quad x \to \infty. \tag{3.89}$$

The assumptions in the theorem are needed to ensure that the remainder integral is well-defined. The above integration by parts algorithm can be continued indefinitely in order to obtain a full expansion with gauge functions $\{x^{-n}\}_n$.

This is not the end of the story since in many important examples it happens that ϕ attains its extreme value in the considered interval and hence, the above theorem cannot be applied.

Example. Consider the exactly solvable integral

$$\int_0^\infty e^{-xt^2}\,dt = \frac{1}{2}\sqrt{\frac{\pi}{x}}. \tag{3.90}$$

Since the result contains an non-integer power of x we expect that the integration by part fails. Here, $\phi(t) = -t^2$ with $\phi'(t) = -2t$ vanishing at $t = 0$ and integrating by parts we obtain an absurd result

$$\int_0^\infty e^{-xt^2}\,dt = \int_0^\infty \frac{-2xte^{-xt^2}}{-2xt}\,dt = \left[\frac{e^{-xt^2}}{-2xt}\right]_0^\infty - \int_0^\infty \frac{e^{-xt^2}}{2xt^2}\,dt. \tag{3.91}$$

The above expression does not even exist. □

The solution of this is due to Laplace. Suppose that smooth ϕ has a maximum at $t = c$, that is $\phi'(c) = 0$ for $c \in [a, b]$. Then, we have the following observation.

Main idea: It is only the immediate neighbourhood of $t = c$ that contributes to the asymptotic expansion of $I(x)$ as $x \to \infty$.

We will prove that this is really the case. We can have three possibilities: $c \in (a, b)$, $c = a$, or $c = b$. We will focus only on the first of these since it is the most interesting. The rest can be dealt in almost the same manner. Fix $\epsilon > 0$. We claim that for $x \to \infty$ independently on ϵ we have

$$I(x) \sim I(x;\epsilon) \quad \text{as} \quad x \to \infty \quad \text{where} \quad I(x;\epsilon) = \int_{c-\epsilon}^{c+\epsilon} f(t)e^{x\phi(t)}dt. \tag{3.92}$$

That is, we want to prove that an arbitrary neighbourhood conveys all the information about the asymptotic behaviour of $I(x)$. To this end, separate the integral into three terms

$$I(x) - I(x;\epsilon) = \int_a^{c-\epsilon} + \int_{c+\epsilon}^b. \tag{3.93}$$

Without any lose of generality we consider only the integral $\int_a^{c-\epsilon}$. Integrating by parts gives us

$$\int_a^{c-\epsilon} f(t)e^{x\phi(t)}dt = \frac{1}{x}\int_a^{c-\epsilon} \frac{f(t)}{\phi'(t)}\frac{\partial}{\partial t}\left(e^{x\phi(t)}\right)dt = \frac{1}{x}\left[\frac{f(t)}{\phi'(t)}e^{x\phi(t)}\right]_{t=a}^{t=c-\epsilon} - \frac{1}{x}\int_a^{c-\epsilon}\left(\frac{f(t)}{\phi'(t)}\right)' e^{x\phi(t)}dt. \tag{3.94}$$

Since $\phi(t)$ attains its maximum at $t = c$ the above is exponentially smaller than $I(x)$. Therefore, $I(x) \sim I(x;\epsilon)$ as $x \to \infty$. As we have this result we can choose ϵ sufficiently small to approximate ϕ by its Taylor series near $t = c$

$$\phi(t) = \phi(c) + \frac{1}{2}\phi''(c)(t-c)^2 + o((t-c)^3), \quad \text{as} \quad t \to c. \tag{3.95}$$

Similarly,

$$f(t) = f(c) + o(1), \quad \text{as} \quad t \to c. \tag{3.96}$$

Therefore,

$$I(x) \sim I(x;\epsilon) = \int_{c-\epsilon}^{c+\epsilon} f(t)e^{x\phi(t)}dt \sim f(c)e^{x\phi(c)} \int_{c-\epsilon}^{c+\epsilon} e^{\frac{x}{2}\phi''(c)(t-c)^2}dt, \quad x \to \infty. \tag{3.97}$$

Now, an important point is that in the above last integral is asymptotic to the integral over the whole $\mathbb{R}$ because $\phi''(c) < 0$ ($t = c$ is a maximum)[11] and a change of variable $s = \sqrt{-x\phi''(c)/2}(t-c)$ leads to

$$I(x) \sim \frac{\sqrt{2}f(c)e^{x\phi(c)}}{\sqrt{-x\phi''(c)}} \int_{-\infty}^{\infty} e^{-s^2}ds, \quad x \to \infty. \tag{3.98}$$

Therefore, we have our main result.

Theorem 2 (Laplace method for integrals). *Let* $\phi \in C^2[a,b]$ *and* $f \in C[a,b]$ *with* $\phi'(c) = 0$ *and* $f(c) \neq 0$ *for* $c \in [a,b]$.

- *If* $c = a$ *we have*
$$I(x) \sim \frac{f(a)e^{x\phi(a)}}{x\phi''(a)}, \quad x \to \infty. \tag{3.99}$$
- *If* $a < c < b$ *we have*
$$I(x) \sim \frac{\sqrt{2\pi}\,f(c)e^{x\phi(c)}}{\sqrt{-x\phi''(c)}}, \quad x \to \infty. \tag{3.100}$$
- *If* $c = b$ *we have*
$$I(x) \sim \frac{f(b)e^{x\phi(b)}}{x\phi''(b)}, \quad x \to \infty. \tag{3.101}$$

The most subtle point of the above proof is the asymptotic equality of and integral over $[c-\epsilon, c+\epsilon]$ to an integral over $\mathbb{R}$. The former has an arbitrarily small measure while the latter - infinite! This is the beauty of asymptotic analysis. The difference of these two integrals is exponentially small due to the form of the integrand.

If ϕ has many extrema inside the interval $[a,b]$ we deal with them by splitting the integral into several parts. Moreover, the method can be simply modified if $\phi''(c) = 0$ by using the appropriate Taylor expansion. This method can also be used in determining the subsequent terms apart from the leading. This is an involved topic which is neatly described in Bender and Orszag's book.

We illustrate the use of Laplace method for one of the most important and useful formulas we have.

Example. (*Stirling's formula*) We will derive the famous approximation of the factorial due to Stirling[12]. We will start with gamma function

$$\Gamma(x+1) = \int_0^{\infty} t^x e^{-t}dt, \tag{3.102}$$

[11] Exercise: prove that $\int_{c-\epsilon}^{c\,\epsilon} \sim \int_{-\infty}^{\infty}$ as $x \to \infty$. Hint: the remainder is exponentially small.

[12] Actually, de Moivre was the first to discover that $n! = O(n^{n+1/2}e^{-n})$ as $n \to \infty$. Stirling found the $\sqrt{2\pi}$ constant of proportionality so that we can write $\sim$.

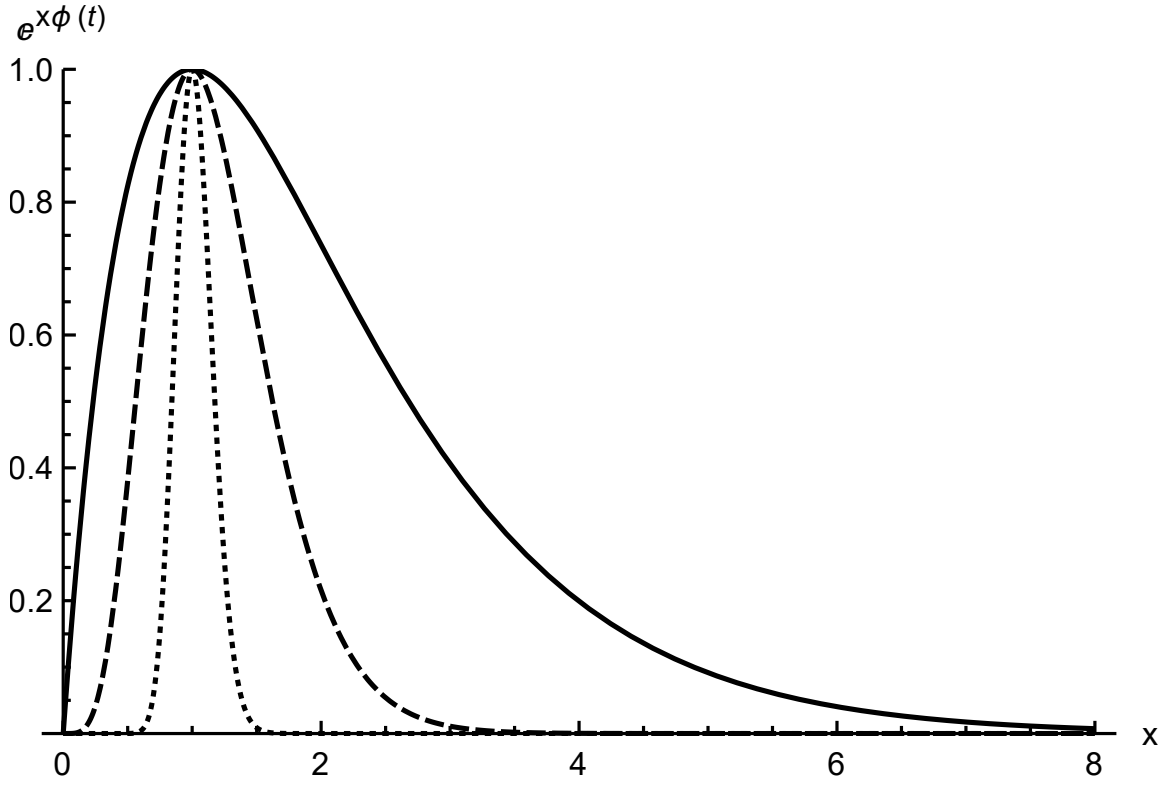

Figure 12: The function $\exp(x\phi(t))/\exp(-x)$ for Striling formula for different: $x = 1$ (solid line), $x = 5$ (dashed line), and $x = 50$ (dotted line). We divide by $\exp(-x)$ in order to normalize the magnitude of all graphs.

and will find its leading order behaviour for $x \to \infty$. By writing

$$\Gamma(x+1) = \int_0^\infty e^{-t+x\ln t} dt, \tag{3.103}$$

we identify that $f(t) = e^{-t}$ while $\phi(t) = \ln t$ having its maximum at infinity! Unfortunately, Laplace method cannot be directly applied. However, we can substitute $t = xs$ which gives

$$\Gamma(x+1) = \int_0^\infty e^{-xs+x\ln x+x\ln s} x ds = x^{x+1}\int_0^\infty e^{x(-s+\ln s)} ds. \tag{3.104}$$

Now, this is exactly a Laplace integral with $f(t) = 1$ and $\phi(t) = -t + \ln t$. We have $\phi'(t) = -1 + 1/t$ and $\phi''(t) = -1/t^2$. And we see that the global maximum occurs at $t = 1$. On the other hand, the function $x\phi(t)$ becomes more and more focused around this maximum which is the general observation for Laplace integrals (see Fig. 12). The value of the integral is thus almost equal to the area under the curve near $t = 1$.

From the Laplace method (3.100) we thus obtain

$$\Gamma(x+1) \sim \sqrt{2\pi} x^{x+\frac{1}{2}} e^{-x} \quad \text{as} \quad x \to \infty. \tag{3.105}$$

As can almost always be anticipated from the asymptotic theory, the accuracy of the approximation is splendid as can be seen on Fig. 13. Even for arguments that cannot be classified as large.

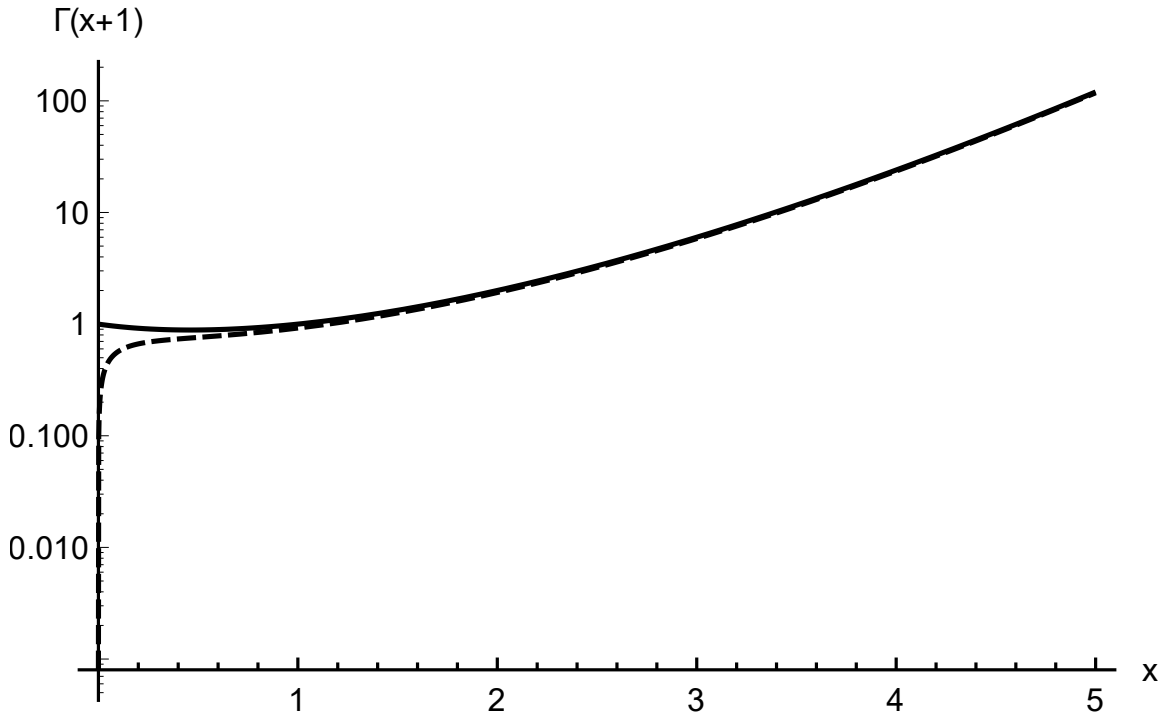

Figure 13: Stirling formula (dashed line) approximating $\Gamma(x+1)$ (solid line).

Laplace method can also be utilized in order to find more terms in the asymptotic expansion by further expanding f and ϕ. The result is

$$\Gamma(x+1) \sim \sqrt{2\pi} x^{x+\frac{1}{2}} e^{-x} \left(1 + \frac{1}{12x} - \frac{1}{288x^3} + \frac{139}{51840x^5} - ...\right) \quad \text{as} \quad x \to \infty, \tag{3.106}$$

and shows even better accuracy. □

There are some interesting generalizations of the Laplace method which enlarge its applicability. The most general case is when we allow the Laplace integral (3.87) to be computed over a path in a complex plane with f and ϕ be complex functions. This leads to the *Steepest descent method*. We will not pursue this issue here, however, a great account can be found in Bender and Orszag. Instead, we will consider the *Method of stationary phase* when ϕ is purely imaginary

$$J(x) = \int_a^b f(t) e^{ix\phi(t)} dt, \tag{3.107}$$

where $x \gg 1$, and a, b can be infinite (and usually are). Here, a significant difference can be quickly spotted, since $|\exp(ix\phi(t))| = 1$, there is no exponential decay that can dominate the growth of f away fr4om the maximum of ϕ. The convergence of that integral is much more subtle. The basic idea comes from an observation that the exponential is a oscillating function with frequency proportional to $x\phi(t)$. Hence, for large x we may expect that this function oscillates such rapidly that the integrand is almost equal to zero due to cancellations of positive and negative contributions from adjacent periods. This can clearly be seen on Fig. 14.

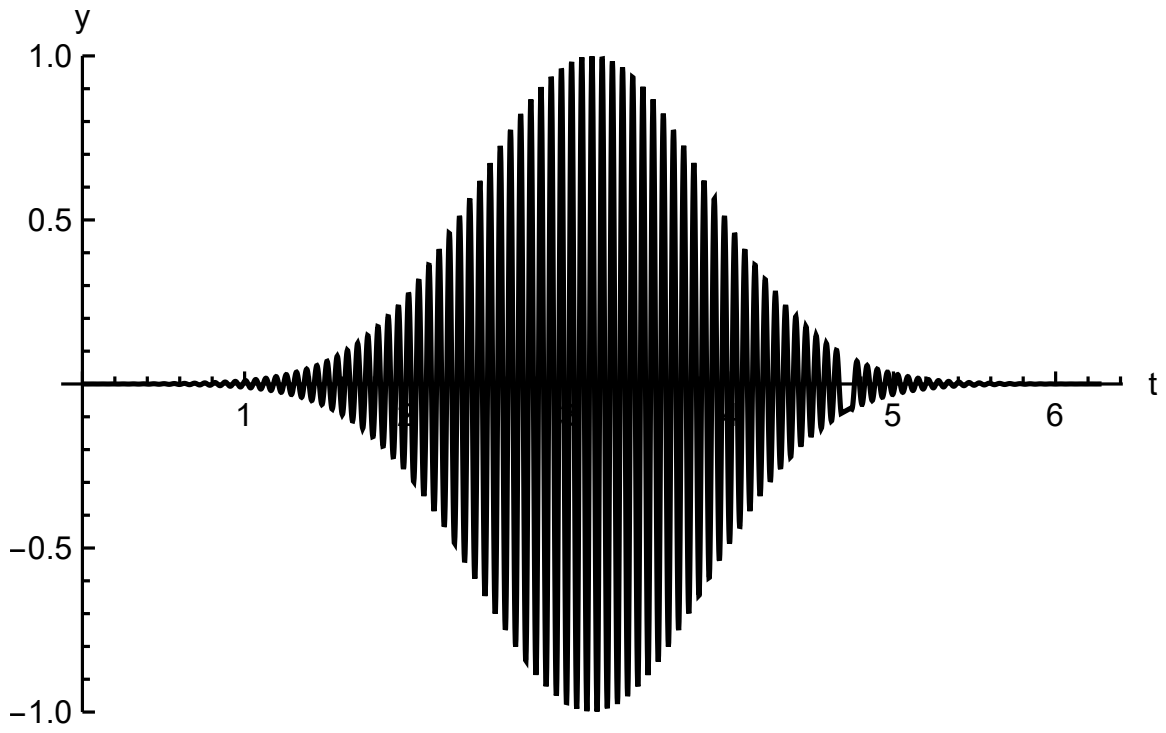

Figure 14: A function $e^{-(t-\pi)^2}\cos(100t)$. We can see that there is almost the same number of positive and negative peaks near a given point. The integral almost cancels out.

The rigorous result is a very well-known theorem in harmonic analysis of Fourier series known as the Riemann-Lebesgue lemma. We will prove its easier version for smooth functions however, the most general result requires only integrability.

Lemma 2 (Riemann-Lebesgue)**.** *Let* f *be a* $C^1[a,b]$ *integrable function along with its derivative. Moreover, assume that* $\phi \in C^1[a,b]$ *with* $\phi'(t) \neq 0$*. Then,*

$$\lim_{x\to\infty} \int_a^b f(t)e^{ix\phi(t)}dt = 0. \tag{3.108}$$

Proof. The proof of the smooth version is simple: integration by parts. We have

$$J(x) = \left[\frac{f(t)}{ix\phi'(t)}\right]_{t=a}^{t=b} + \frac{i}{x}\int_a^b \left(\frac{f(t)}{\phi'(t)}\right)' e^{ix\phi(t)}dt. \tag{3.109}$$

Now, since $\phi'(t) \neq 0$ we have

$$\left|\left[\frac{f(t)}{ix\phi'(t)}\right]_{t=a}^{t=b}\right| \leq \frac{1}{x}\max_{t\in\{a,b\}}\left|\frac{f(t)}{\phi'(t)}\right| \to 0 \quad \text{as} \quad x\to\infty. \tag{3.110}$$

Similarly, due to integrability of f' we can write

$$\left|\frac{i}{x}\int_a^b \left(\frac{f(t)}{\phi'(t)}\right)' e^{ix\phi(t)}dt\right| \leq \frac{1}{x}\int_a^b \left|\left(\frac{f(t)}{\phi'(t)}\right)'\right| dt \to 0 \quad \text{as} \quad x\to\infty. \tag{3.111}$$

Since the two terms resulting from the integration by parts vanish we have $J(x) \to 0$ as $x \to \infty$. The proof is complete. □

In asymptotic analysis we would like to have something more - a precise indication of the order of convergence to zero of $J(x)$. Iterating Riemann-Lebesgue lemma under suitable regularity of the integrand yields the sought result. However, we note that the Riemann-Lebesgue lemma is valid only with a mild assumption of f being integrable and ϕ continuously differentiable but not constant over any subinterval.

Corolary 1. *Let the assumptions of the Riemann-Lebesgue lemma be satisfied. Morever, suppose that* $(f(t)/\phi'(t))'$ *is integrable. Then,*

$$J(x) = O\left(\frac{1}{x}\right) \quad \textit{as} \quad x \to \infty. \tag{3.112}$$

Proof. Integrating by parts we, again, arrive at (3.109). The first term vanishes as $O(x^{-1})$ while the second is $o(x^{-1})$ as $x \to \infty$ because the Riemann-Lebesgue lemma due to the assumption. □

We can illustrate the above results in an example.

Example. Consider the following Fourier integral

$$\int_0^1 \frac{e^{ixt}}{1+t}\,dt. \tag{3.113}$$

If we integrate by parts we obtain

$$\int_0^1 \frac{e^{ixt}}{1+t}dt = -\frac{i}{2x}e^{ix} + \frac{i}{x} - \frac{i}{x}\int_0^1 \frac{e^{ixt}}{(1+t)^2}dt. \tag{3.114}$$

We have to show that the integral above is $O(x^{-2})$ as $x \to \infty$. To this end we can integrate by parts once again to obtain

$$-\frac{i}{x}\int_0^1 \frac{e^{ixt}}{(1+t)^2}\,dt = -\frac{1}{4x^2}e^{ix} + \frac{1}{x^2} - \frac{2}{x^2}\int_0^1 \frac{e^{ixt}}{(1+t)^3}\,dt. \tag{3.115}$$

Now, the last integral can be estimated by

$$\left|\int_0^1 \frac{e^{ixt}}{(1+t)^3}\,dt\right| \le \int_0^1 \frac{1}{(1+t)^3}\,dt = \frac{3}{8}. \tag{3.116}$$

Therefore,

$$\int_0^1 \frac{e^{ixt}}{1+t}dt \sim -\frac{i}{2x}e^{ix} + \frac{i}{x}, \tag{3.117}$$

as $x \to \infty$. □

The above integration by parts results may fail when $\phi'(c) = 0$ for some $c \in [a, b]$. This is frequently the case and brings us to the method of stationary phase. The important remark is the following.

Basic idea. If $\phi'(c) = 0$ for a unique $c \in [a, b]$ then the asymptotic behaviour of $J(x)$ integral is dominated by the integrand in the neighbourhood of $t = c$. The point $t = c$ is called *stationary*.

The proof of the above follows the same route as the one for the Laplace method: show that the immediate vicinity of $t = c$ carries the majority of the mass of the integral, expand ϕ in Taylor series, and compute the asymptotic result explicitly. However, this time the reasoning utilizes complex analysis and, hence, we omit the proof.

Theorem 3 (Method of Stationary Phase)**.** *Let the assumptions of Riemann-Lebesgue lemma be satisfied and* $\phi'(c) = 0$ *with* $c \in (a, b)$*. Then, the integral* $J(x)$ *has the following asymptotic behaviour*

$$J(x) \sim f(c) e^{ix\phi(c) + \frac{i\pi}{4} sgn\phi''(c)} \sqrt{\frac{2\pi}{x|\phi''(c)|}} \quad as \quad x \to \infty. \tag{3.118}$$

From the above result we see that J oscillates with a frequency given by the value of $\phi(c)$ and amplitude decaying as $x^{-1/2}$. If there are many stationary points we have to separate the integral into several ones having only one point each. In contrast with Laplace method obtaining further asymptotic terms is much more problematic due to lack of exponential decay.

Example. (*Bessel functions*) One of the most important special functions arising in problems of circular and cylindrical symmetries are Bessel functions. We meet them frequently in heat conduction, fluid dynamics, acoustics, electromagnetism, etc. The n-th Bessel function can be represented as an integral

$$J_n(x) = \frac{1}{\pi} \int_0^\pi \cos(x \sin t - nt)\, dt. \tag{3.119}$$

In order to use the Method of Stationary Phase we have to write the above in an appropriate form

$$J_n(x) = \mathrm{Re}\, \frac{1}{\pi} \int_0^\pi e^{-int} e^{ix \sin t}\, dt, \tag{3.120}$$

where Re is the real part of a complex number. Here, $f(t) = \exp(-int)$ and $\phi(t) = \sin t$. We have $\phi'(t) = \cos t$ for which $c = \pi/2$. The function $\cos(x \sin t)$ for large x oscillates very rapidly far from $\pi/2$ as can be seen on Fig. 15. Most of these oscillations cancel out leaving only the immediate neighbourhood of $\pi/2$ to be meaningful. Our formula (3.118) gives then

$$J_n(x) \sim \mathrm{Re}\, \frac{1}{\pi} e^{-\frac{in\pi}{2}} e^{ix - \frac{i\pi}{4}} \sqrt{\frac{2}{\pi x}} = \sqrt{\frac{2}{\pi x}} \cos\left(x - \frac{n\pi}{2} - \frac{\pi}{4}\right) \quad \text{as} \quad x \to \infty. \tag{3.121}$$

The accuracy of such an approximation is very decent (see Fig. 15). Note that we probably can safely use the asymptotic formula for $x \geq 4$ which is much more straightforward than the superposition of all oscillatory modes given by (3.119). □

The method of stationary phase, apart from being extremely useful for resolving complicated oscillatory integrals and Fourier transforms, is an indispensable aid for

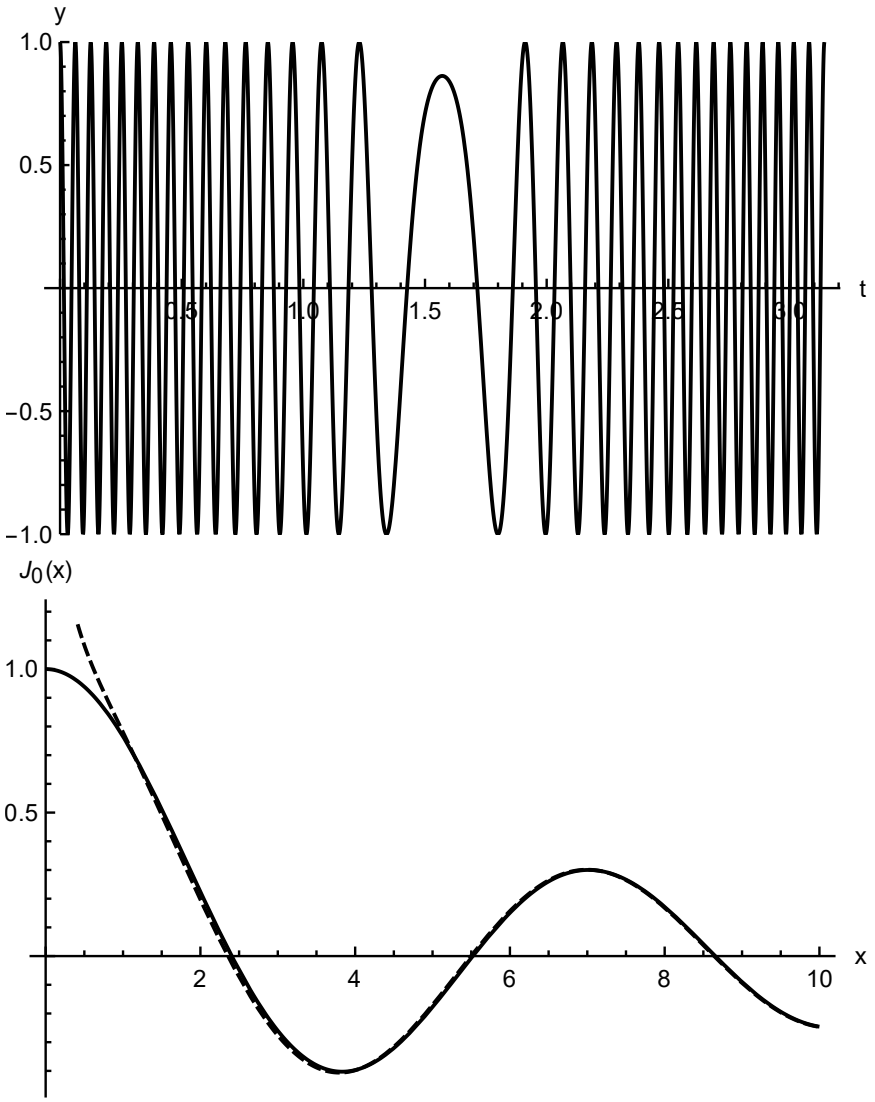

Figure 15: On the top: a function $\cos(100 \sin t)$. On the bottom: Bessel function $J_0(x)$ (solid) and its asymptotic approximation (dashed).

numerical analysis. Using straightforward quadratures to compute integrals of the form (3.107) is very demanding on computer power for large x. Using the Method of Stationary Phase helps us to see what are really the most essential frequencies and how to resolve them. This trait is crucial in optics or water wave analysis on which we will have something to say in further sections.

3.4 Asymptotic expansion of sums. Euler-Maclaurin formula

We will now derive a very elegant formula for finding approximate form of the sums

$$\sum_{k=1}^{n} f(k). \tag{3.122}$$

When $n \to \infty$ we will also be able to derive asymptotic expansions. The result is due to Euler and Maclaurin who discovered it in the first half of the XVII century. Euler used it to approximate the sum $\sum_{k=1}^{n} k^{-2}$ to 20 decimal places. This lead him to a rigorous solution of the famous Basel problem while he was 27 years old. It was Poisson who later provided the formula for the remainder of their approximation. In XX century it appeared that the Euler-Maclaurin formula is an essential tool in analysing algorithms and numerical methods.

The main idea for the method is based on successive integration by parts in a clever way. We assume that f is sufficiently smooth and consider the integral

$$\int_0^n f(x)dx = \sum_{k=0}^{n-1} \int_k^{k+1} f(x)dx. \tag{3.123}$$

We would like to integrate by parts in such a way that the antiderivative of $g(x) = 1$ is the same in each interval $(k, k+1)$. Thanks to that, we will be able to sum these contributions in a neat way. This is the main idea of the method. By this requirement, this has to be a periodic linear function with unit period with

$$P_1(x) = x - [x] - \frac{1}{2} \quad \text{for} \quad x \notin \mathbb{N} \quad \text{and} \quad \lim_{x \to k^{\pm}} P_1(x) = \pm\frac{1}{2}, \tag{3.124}$$

where $[x]$ is the integer part of x, and $k \in \mathbb{N}$. The above is called the *first Bernoulli function* and is a periodic version of the *first Bernoulli polynomial*

$$B_1(x) = x - \frac{1}{2}. \tag{3.125}$$

Of course $P_1'(x) = 1$ for x that are not integers. Consider only the integral over one subinterval and integrate by parts to obtain

$$\begin{aligned} \int_k^{k+1} f(x)dx &= [f(x)P_1(x)]_k^{k+1} - \int_k^{k+1} f'(x)P_1(x)dx \\ &= \frac{1}{2}\left(f(k) + f(k+1)\right) - \int_k^{k+1} f'(x)P_1(x)dx. \end{aligned} \tag{3.126}$$

Now, if we sum the above for $k = 0$ to $k = n - 1$ we obtain

$$\int_0^n f(x)dx = \frac{1}{2}(f(0) + f(n)) + \sum_{k=1}^{n-1} f(k) - \int_0^n f'(x)P_1(x)dx. \tag{3.127}$$

We can now simplify the above form by adding $(f(n) - f(0))/2$

$$\int_0^n f(x)dx + \frac{f(n) - f(0)}{2} = \sum_{k=1}^{n} f(k) - \int_0^n f'(x)P_1(x)dx, \tag{3.128}$$

and hence,

$$\sum_{k=1}^{n} f(k) = \int_0^n f(x)dx + \frac{f(n) - f(0)}{2} + \int_0^n f'(x)P_1(x)dx. \tag{3.129}$$

In this way we have written an integral as a sum of the values of the integrated function at integer points. Notice also the explicit form of the remainder. In numerical analysis this is the trapezoidal rule for quadrature.

To carry this further we have to integrate by parts the remainder term, that is after division into subintervals $(k, k + 1)$ we have

$$\int_k^{k+1} f'(x)P_1(x)dx = \left[\frac{f'(x)P_2(x)}{2}\right]_k^{k+1} - \frac{1}{2}\int_k^{k+1} f''(x)P_2(x)dx, \tag{3.130}$$

where we have to specify the *second Bernoulli function* $P_2(x)$ (note that for convenience we have moved the factor $1/2$ out of definition). First of all, for $x \in (0, 1)$ we should have $P_2'(x) = 2B_1(x) = 2x - 1$. Furthermore, $P_2(x)$ should be continuously periodic with period 1. Therefore, it should be a periodic second order polynomial of the form $x^2 - x + B_2$, where the constant C has to be determined from the periodicity condition (since, as we will see, higher degree Bernoulli functions will also be periodic)

$$\int_0^1 P_2(x)dx = 0 \quad \rightarrow \quad \frac{1}{3} - \frac{1}{2} + B_2 = 0, \tag{3.131}$$

and hence $B_2 = 1/6$. Therefore, $P_2(x)$ is a periodic repetition of the *second Bernoulli polynomial*

$$B_2(x) = x^2 - x + \frac{1}{6}. \tag{3.132}$$

Now, $P_2(x)$ is continuous. Returning to our integral and evaluating $P_2(x)$ we obtain

$$\int_k^{k+1} f'(x)P_1(x)dx = \frac{1}{12}(f'(k+1) - f'(k)) - \frac{1}{2}\int_k^{k+1} f''(x)P_2(x)dx, \tag{3.133}$$

and hence, by summing over k, we arrive at

$$\sum_{k=1}^{n} f(k) = \int_0^n f(x)dx + \frac{f(n) - f(0)}{2} + \frac{1}{2}P_2(0)(f'(n) - f'(0)) - \frac{1}{2}\int_0^n f''(x)P_2(x)dx. \tag{3.134}$$

It is now apparent how to continue this procedure of integration by parts. All is needed to do is to define the *k-th Bernoulli functions*

$$P_k(x) = k\int_0^x P_{k-1}(x)dx + B_k \text{ where } B_k \text{ is chosen to satisfy } \int_0^1 P_k(x)dx = 0. \tag{3.135}$$

which are periodic versions of *n-th Bernoulli polynomials*

$$\begin{aligned} &B_0(x) = 1, \quad B_1(x) = x - \frac{1}{2}, \quad B_2(x) = x^2 - x + \frac{1}{6}, \quad B_3(x) = x^3 - \frac{3}{2}x^2 + \frac{1}{2}x, \\ &B_4(x) = x^4 - 2x^3 + x^2 + \frac{1}{30}, \quad B_5(x) = x^5 - \frac{1}{2}x^4 + \frac{1}{3}x^3 - \frac{1}{30}x, \quad \ldots \end{aligned} \tag{3.136}$$

that is $P_k(x) = B_k(x - [x])$. The constants $B_0 = 1$, $B_1 = -B_1(0) = B_1(1) =$ and $B_k = B_k(0) = B_k(1)$, for $k > 1$, are called *Bernoulli numbers*. They vanish for odd k and are chosen in order to satisfy the periodicity requirement. Bernoulli polynomials and numbers have many interesting properties on their own and arise frequently in calculus. Having all this machinery we can formulate the most general result.

Theorem 4 (Euler-Maclaurin). *If $f \in C^p([1, n])$, then we have*

$$\sum_{k=1}^{n} f(k) = \int_0^n f(x)dx + \frac{f(n) - f(0)}{2} + \sum_{k=1}^{[p/2]} \frac{B_{2k}}{(2k)!}\left(f^{(2k-1)}(n) - f^{(2k-1)}(0)\right) + R_p, \tag{3.137}$$

where the remainder is

$$R_p = \frac{(-1)^p}{p!}\int_0^n f^{(p)}(x)P_p(x)dx. \tag{3.138}$$

The essential result in the Euler-Maclaurin summation formula is the explicit form of the remainder (3.138). Since P_p, from definition, has zero average, we expect that the remainder could actually be really small. Moreover, there is a very sophisticated and difficult result stating that $|B_p(x)| \leq 2p!/(2\pi)^p\zeta(p)$, where ζ is the Riemann Zeta function. For even p this bound is optimal and attained for $x = 0$. Having this, we can obtain a very useful estimate for the remainder

$$|R_p| \leq \frac{2\zeta(p)}{(2\pi)^p}\int_0^n |f^{(p)}(x)|dx. \tag{3.139}$$

Usually, a little simpler formula can be obtained when f along with all of its derivatives vanish at infinity and $f^{(p)}$ is integrable. Letting $n \to \infty$ in (3.137) we obtain

$$\sum_{k=1}^{\infty} f(k) = \int_0^\infty f(x)dx - \frac{f(0)}{2} - \sum_{k=1}^{[p/2]} \frac{B_{2k}}{(2k)!}\left(f^{(2k-1)}(0)\right) + \frac{(-1)^p}{p!}\int_0^\infty f^{(p)}(x)P_p(x)dx. \tag{3.140}$$

One can also let $p \to \infty$ if the function f is infinitely smooth. However, it then rarely happens that the series above is convergent is the classical sense. Instead, from (3.137) we obtain

$$\sum_{k=1}^{n} f(k) \approx \int_0^n f(x)dx + \frac{f(n) - f(0)}{2} + \sum_{k=1}^{\infty} \frac{B_{2k}}{(2k)!}\left(f^{(2k-1)}(n) - f^{(2k-1)}(0)\right). \tag{3.141}$$

In many important cases the integral above can be evaluated explicitly giving a neat way of expressing the sum for large n. We will illustrate the summation formula with several examples.

Example. (Euler-Mascheroni constant) Recall from Calculus 1 that the following limit exists and is called the *Euler-Mascheroni constant*

$$\gamma = \lim_{n\to\infty} \left(\sum_{k=1}^{n} \frac{1}{k} - \ln n \right). \tag{3.142}$$

Take $f(x) = (1+x)^{-1}$, then $f^{(p)}(x) = (-1)^p p!(1+x)^{-1-p}$. Therefore, the Euler-Maclaurin summation formula (3.141) written for $n-1$ instead of n and $p \to \infty$ gives

$$\sum_{k=1}^{n-1} \frac{1}{k+1} - \int_0^{n-1} \frac{dx}{1+x} = \gamma + \frac{1}{2n} - \sum_{k=1}^{\infty} \frac{B_{2k}}{2k} \frac{1}{n^{2k}}, \tag{3.143}$$

where we have put all the constants into $\gamma - 1$. It is precisely the Euler-Mascheroni constant because evaluating the integral and changing the limits of summation yields

$$\sum_{k=2}^{n} \frac{1}{k} - \ln n = \gamma - 1 + \frac{1}{2n} - \sum_{k=1}^{\infty} \frac{B_{2k}}{2k} \frac{1}{n^{2k}}, \tag{3.144}$$

which can be rearranged into

$$\gamma \sim \sum_{k=1}^{n} \frac{1}{k} - \ln n - \frac{1}{2n} + \sum_{k=1}^{\infty} \frac{B_{2k}}{2k} \frac{1}{n^{2k}}, \quad n \to \infty. \tag{3.145}$$

The right-hand side of the equation converges to γ. However, the above asymptotic expansion is much better than a simple statement of a limit. For example, using the seemingly crude approximation with only one Bernoulli number $B_2 = 1/6$

$$\gamma \approx \sum_{k=1}^{n} \frac{1}{k} - \ln n - \frac{1}{2n} + \frac{1}{12n^2} \tag{3.146}$$

and putting $n = 10$ one obtains $\gamma \approx 0.57721$ where all digits are accurate. Note that you can easily do these calculations on a handheld calculator. Computing γ from the definition requires taking $n \geq 10^6$ in order to obtain the same accuracy! This throttles even a decently fast modern computer. Note the genius of Euler. □

Example. (*Riemann Zeta*) The same technique as above can be used in order to find an asymptotic formula for Riemann zeta function which is one of the most profound special functions that has far-reaching applications in computer science and constitute a bridge between analysis and number theory. The result is as follows

$$\sum_{k=1}^{n} \frac{1}{k^s} \sim \zeta(s) - \frac{1}{(s-1)n^{s-1}} + \frac{1}{2n^s} - \sum_{k=1}^{\infty} \frac{B_{2k}}{(2k)!} \frac{(s+2k-2)!}{(s-1)!n^{s+2k-1}} \quad \text{as} \quad n \to \infty, \tag{3.147}$$

where $\zeta(s)$ is the Riemann Zeta function. This, as Euler did, can be inverted to obtain an accurate approximation of the Basel problem for $s = 2$

$$\zeta(2) \sim \sum_{k=1}^{n} \frac{1}{k^2} + \frac{1}{n} - \frac{1}{2n^2} + \frac{1}{6n^3} - \frac{1}{30n^4} + \ldots \quad \text{as} \quad n \to \infty. \tag{3.148}$$

For example, plugging $n = 10$ into the above gives $\zeta(2) \approx 1.64493$ which are all exact digits for the value $\zeta(2) = \pi^2/6$. Remarkably, $n = 2$ gives 3 accurate decimal digits! This approximation led Euler to guess the true solution of the Basel problem and prove it rigorously by different means. □

Example. (*Sums of powers (Faulhaber's formula)*) This example traces the historical origin of Bernoulli numbers. If we put $f(x) = x^m$ in (3.137) the $m + 1$ derivative vanishes yielding a zero remainder and an exact formula. That is,

$$\sum_{k=1}^{n} k^m = \frac{1}{m+1} n^{m+1} + \frac{1}{2} n^m + \sum_{k=1}^{[m/2]} \frac{B_{2k}}{(2k)!} m(m-1)\ldots(m-2k+2) n^{m-2k+1} \tag{3.149}$$

The sum can be simplified by noticing that the binomial coefficient can be simply factored

$$\begin{aligned} &\sum_{k=1}^{[m/2]} \frac{B_{2k}}{(2k)!} m(m-1)\ldots(m-2k+2) n^{m-2k+1} = \frac{1}{m+1} \sum_{k=1}^{[m/2]} \frac{B_{2k}}{(2k)!} \frac{(m+1)!}{(m+1-2k)!} n^{m-2k+1} \\ &= \frac{1}{m+1} \sum_{k=1}^{[m/2]} \binom{m+1}{2k} B_{2k} n^{m-2k+1}. \end{aligned} \tag{3.150}$$

Now, since the odd Bernoulli number vanish we can introduce a new summation variable $j = 2k$ to have

$$\sum_{k=1}^{n} k^m = \frac{1}{m+1} n^{m+1} + \frac{1}{2} n^m + \frac{1}{m+1} \sum_{j=2}^{m} \binom{m+1}{j} B_j n^{m-j+1}. \tag{3.151}$$

The above can be simplified further with recalling that $B_0 = 1$ and $B_1 = 1/2$ to absorb the n^{m+1} and n^m terms into the sum. We finally have the very elegant *Faulhaber's formula* of sums of integer powers

$$\sum_{k=1}^{n} k^m = \frac{1}{m+1} \sum_{j=0}^{m} \binom{m+1}{j} B_j n^{m-j+1} \tag{3.152}$$

For example, with $m = 1, 2, 3$ we obtain the well-known polynomials

$$\sum_{k=1}^{n} k = \frac{n(n+1)}{2}, \quad \sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}, \quad \sum_{k=1}^{n} k^3 = \left(\frac{n(n+1)}{2}\right)^2. \tag{3.153}$$

Note that this has nothing to do with asymptotics and is just a different proof of these exact formulas.

There is an interesting history behind many results concerning sums of integer powers. The idea of summing successive squares and cubes can be dated to ancient Greece (Pythagoras, Archimedes), India (Aryabhata), and medieval Persia (Abu Bakr al-Karaji). Then, in the mid XVII century Faulhaber gave formulas for summing powers up to 17th. However, it was Jakob Bernoulli who found out that the polynomial coefficients constitute a sequence that can be defined in a recursive way. A completely rigorous proof of Faulhaber's formula was given in the mid XIX century by Jacobi. □

Example. In this example we will consider the following sum that is closely related with Jacobi theta function that appears in harmonic analysis, heat conduction, and quantum field theory among others. It has the form

$$\sum_{k=-n}^{n} e^{-\frac{k^2}{n}}. \tag{3.154}$$

This series cannot be summed in an exact form and due to exponential decay the terms may be difficult to compute for large n. This is the limit we would like to study. Note that the above series is a little bit different than the one we considered. However, it is easy to write a generalization of Euler-Maclaurin summation formula for this case

$$\sum_{k=-n}^{n} f(k) = \int_{-n}^{n} f(x)dx + \frac{f(-n)+f(n)}{2} + \sum_{k=1}^{[p/2]} \frac{B_{2k}}{(2k)!}\left(f^{(2k-1)}(n) - f^{(2k-1)}(-n)\right) + R_p. \tag{3.155}$$

Now, taking $f(x) = g(x/\sqrt{n})$ where $g(x) = e^{-x^2}$ and remembering an exercise from Calculus I we have

$$f^{(k)}(x) = n^{-\frac{k}{2}} Q_k\left(\frac{x}{\sqrt{n}}\right) e^{-\frac{x^2}{n}}, \tag{3.156}$$

where Q_k is some[13] polynomial of degree k. Note also that due to the exponential decay, $f^{(k)}(x) \to 0$ when $x \to \pm\infty$. This causes the sum in (3.155) to vanish in that limit. On the other hand, the integral is

$$\int_{-n}^{n} e^{-\frac{x^2}{n}} dx = \sqrt{n}\int_{-\sqrt{n}}^{\sqrt{n}} e^{-y^2} dy \sim \sqrt{n\pi} \quad \text{as} \quad n \to \infty. \tag{3.157}$$

We are only left in investigating the remainder. From (3.138)

$$|R_p| \leq \frac{2\zeta(p)}{(2\pi)^p} \frac{1}{n^{-\frac{k}{2}}} \int_{-\sqrt{n}}^{\sqrt{n}} |P_p(y)| e^{-y^2} dy \sim \frac{2\zeta(p)}{(2\pi)^p} \frac{1}{n^{\frac{p}{2}-1}} \int_{-\infty}^{\infty} |P_k(y)| e^{-y^2} dy \quad \text{as} \quad n \to \infty, \tag{3.158}$$

where the integral above is convergent thanks to the Gaussian. Combining all terms needed for the Euler-Maculaurin formula we arrive at

$$\sum_{k=-n}^{n} e^{-\frac{k^2}{n}} \sim \sqrt{n\pi} + O(n^{1-\frac{p}{2}}) \quad \text{as} \quad n \to \infty, \tag{3.159}$$

[13] It is closely related to *Hermite polynomial*.

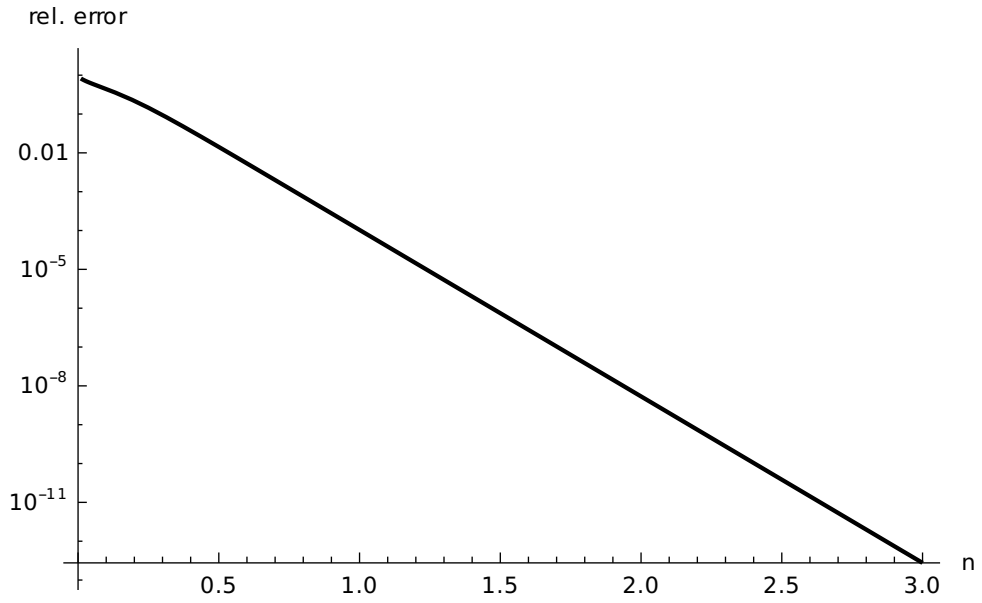

Figure 16: A log plot of the relative error of approximating (3.154) with $\sqrt{n\pi}$.

where $p > 0$ is an *arbitrarily* large number. Therefore, the remainder is transcendentally small with respect to integer powers of inverse n and it can be shown that it is indeed exponentially small. On Fig. 16 we can see the log plot of the relative error of approximating the above series with $\sqrt{n\pi}$. Notice the extreme smallness of it even for $n \approx 0.5$ while the above asymptotic result way obtained only for $n \to \infty$. Behold the true power of asymptotic analysis! □

Many examples above show that having an asymptotic series that may not be convergent is surprisingly useful in accurately calculating many quantities. Asymptotic series are used in almost every field that uses mathematics: from computer science, through astronomy, to mathematical biology. It is not true that having a computer lets you calculate everything with arbitrary accuracy. You have to have many ingenious algorithms in hand unless you have time to wait for the calculations to complete (you do not). Practical algorithms are always constructed with the aid of asymptotic analysis that helps to quantify computational complexity or use divergent series as a superb approximation for computed functions. Notice that understanding some of these ideas requires opening your mind to new material that may have been controversial some time ago. However, this is mathematics and everything is perfectly rigorous when considering theory. In applications one usually is not that lucky and have to rely on intuition, knowledge, and simulations to build and analyse a successful model. For example, in dealing with realistic perturbation series one usually is able to find one or two first terms. Then, there is no question about convergence or even the mathematical meaning about the series. We have to learn to live with that lack of complete knowledge and use our experience and understanding to guide us. With sufficient training this approach is highly successful.

3.5 Singular perturbations and boundary layers

Armed with all we learned on asymptotic analysis we now go back to the beginning of this section in which we analysed the perturbation theory. We have developed an algorithm for finding regular perturbation expansions for which we now know that the are asymptotic with respect to ϵ. In practice, however, one is usually faced with more complex circumstances where the regular perturbations fail. For example, it can happen that when passing with $\epsilon \to 0$ we loose one of the solutions - the one that is relevant. We cave defined the regular perturbations as these which do not change the number of solutions when $\epsilon \to 0$. Those perturbations that are not regular are called *singular* and, not surprisingly, constitute a vast field of applied mathematics. Much broader, voluminous, and rich in interesting examples and prolific applications than regular perturbations. In this part we will only scratch the surface of singular perturbation theory. There are many books written on that subject and the research is still ongoing. We will illustrate the topic on several examples. This is the best way of grasping what is going on in the real-world.

Example. We will start with a simple archetypal algebraic example

$$\epsilon x^2 + 2x - 1 = 0, \quad \epsilon \ll 1. \tag{3.160}$$

Notice that the ϵ multiplies the highest order term and when $\epsilon \to 0$ the quadratic becomes a linear polynomial. We loose one solution. It is helpful to see the problem graphically as presented on Fig. **??**. For any $\epsilon > 0$ the line $-2x + 1$ and the quadratic ϵx^2 have two intersections. One is close to $1/2$ and the other escapes to $-\infty$ when $\epsilon \to 0$. Indeed, the exact solutions are

$$x_+ = \frac{-1 + \sqrt{1+\epsilon}}{\epsilon}, \quad x_- = \frac{-1 - \sqrt{1+\epsilon}}{\epsilon}. \tag{3.161}$$

The positive one is indeed $x_+ = 1/2 - \epsilon/8 + O(\epsilon^2)$ and

$$x_- = -\frac{2}{\epsilon} - \frac{1}{2} + O(\epsilon) \quad \text{as} \quad \epsilon \to 0. \tag{3.162}$$

We can see that the negative solution does not have a standard regular expansion in terms of ϵ. But how to find such an expansion? A good procedure is to magnify the place where the problematic zero occurs and study its behaviour.

In order to do this we introduce a scaling transformation

$$y = \frac{x}{\epsilon^\alpha}, \tag{3.163}$$

where α is to be determined in order to focus on the negative zero. Plugging the above into the quadratic we obtain

$$\epsilon^{1+2\alpha} y^2 + 2\epsilon^\alpha y - 1 = 0. \tag{3.164}$$

Now, we have to think about the above as a balance of terms. We would like to have $\epsilon \to 0$ while keeping $y = O(1)$ in order for the scaling to be effective. Since before

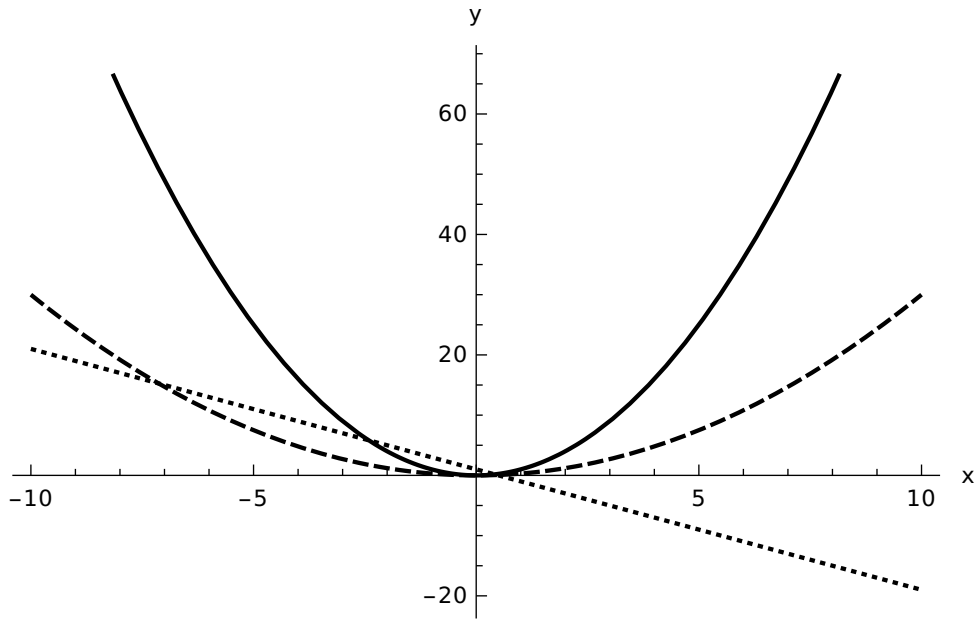

Figure 17: A graphical solution of the singular quadratic. Here, the solid line represents x^2, dashed line $0.3x^2$, and dotted line is the line $-2x + 1$.

the scaling we have lost one of the solutions due to vanishing of the leading term in the equation, we would like to keep it right now. This is why we have to balance the quadratic term with other ones. If we balance quadratic term with a constant, then

$$1 + 2\alpha = 0, \tag{3.165}$$

which gives $\alpha = -1/2$. This leads to

$$\sqrt{\epsilon}y^2 + 2y - \sqrt{\epsilon} = 0, \tag{3.166}$$

what becomes $y = 0$ when $\epsilon \to 0$. This is not consistent with our assumption that $y = O(1)$ and hence we have to discard this balance. On the other habd, of the quadratic balances linear term we should have

$$1 + 2\alpha = \alpha, \tag{3.167}$$

when $\epsilon \to 0^+$. This gives $\alpha = -1$ and

$$y^2 + 2y - \epsilon = 0, \tag{3.168}$$

which gives $y^2 + 2y = 0$ in the limit. Now, this is something more we wanted to have. We can use the regular perturbation theory in a form

$$y = y_0 + \epsilon y_1 + O(\epsilon^2), \tag{3.169}$$

to find out that

$$\begin{cases} \epsilon^0 : & y_0^2 + 2y_0 = 0, \\ \epsilon^1 : & 2y_0y_1 + 2y_1 - 1 = 0. \end{cases} \tag{3.170}$$

From the first equation we have $y_0 = 0$ and $y_0 = -2$. This might be strange a little bit since it seems that we have obtained three solutions. However, when we go back to the scaling the $y_0 = 0$ solution is consistent with x_+, for $x_p \to x_0 = 1/2$ when $\epsilon \to 0$ with $y_0 = \epsilon x_0$. The other solution $y_0 = -2$ is the one we look for. Further, solving the ϵ^1 equation we arrive at

$$y = -2 - \frac{1}{2}\epsilon + O(\epsilon^2), \tag{3.171}$$

when $\epsilon \to 0$. This is completely consistent with Taylor expansion of x_- and constitutes a main idea of the singular perturbation theory. The most significant is the scaling of the initial variable to focus on the singular behaviour. □

Above introductory example shows the essential features of basic singular perturbation theory: a solution is lost in the limit and we have to rescale the problem to track it. In more complex examples this is by no means easy or straightforward and we have to really understand the problem to know how to rescale. Notice that one solution above has been an order of magnitude (in ϵ) larger than the other. A even more vivid example is given by an analysis of a boundary value problem for an ODE.

Example. A classical example of singular perturbation theory is presented by an ODE when the small parameter multiplies the highest order derivative

$$\epsilon y'' + 2y' + y = 0. \tag{3.172}$$

We impose two boundary conditions

$$y(0) = 0, \quad y(1) = 1. \tag{3.173}$$

This example can be solved exactly and is chosen to do so in order to learn what goes on when $\epsilon \to 0$. First, notice that under that limit the equation becomes a first order ODE which, in general, cannot satisfy two boundary conditions! Suppose that we apply our regular perturbation theory. Let

$$y = y_0 + \epsilon y_1 + O(\epsilon^2) \quad \text{as} \quad \epsilon \to 0. \tag{3.174}$$

Then, plugging the above expansion into (3.172) and equating coefficients of ϵ we obtain a system

$$\begin{cases} \epsilon^0 : & 2y_0' + y_0 = 0, \\ \epsilon^1 : & y_0'' + 2y_1' + y_1 = 0, \\ \dots \end{cases} \tag{3.175}$$

The solution of the first one is $y_0(x) = Ce^{-x/2}$. The constant C has to be determined from the boundary conditions which are

$$y_0(0) = 0, \quad y_0(1) = 1. \tag{3.176}$$

Notice that if we impose the first one we should choose $C = 0$ and hence $y_0 \equiv 0$. This is not what we have expected and, hence, have to choose $C = \exp(1/2)$ in order to satisfy the condition at the right boundary. Therefore,

$$y_0(x) = e^{\frac{1}{2}(1-x)}. \tag{3.177}$$

Similarly, the next equation now becomes

$$2y_1' + y_1 = -\frac{1}{4}e^{\frac{1}{2}(1-x)}, \quad y_1(1) = 0. \tag{3.178}$$

Note that we have included only the boundary condition at $x = 1$ in order to be consistent with the leading order solution. We find that the ϵ correction is

$$y_1(x) = \frac{1}{8}(1-x)e^{\frac{1}{2}(1-x)}, \tag{3.179}$$

and our regular perturbation solution is

$$y(x) = e^{\frac{1}{2}(1-x)}\left(1 + \frac{\epsilon}{8}(1-x) + O(\epsilon^2)\right) \quad \text{as} \quad \epsilon \to 0. \tag{3.180}$$

We have succeeded only partially - our solution does not satisfy the condition at $x = 0$ and hence, cannot be taken as a uniform approximation. Since the above is a perturbation expansion away from that boundary we call it *outer solution*. It is analogous to the x_+ zero of our quadratic example.

To find the solution near $x = 0$ we have to rescale the variable

$$\xi = \frac{x}{\epsilon^\alpha}, \tag{3.181}$$

when $\alpha > 0$ is to be determined. Let us introduce $Y(\xi) = y(x(\xi))$ as a solution expressed in new, rescaled, variable. We have

$$\frac{dy}{dx} = \frac{d\xi}{dx}\frac{dY}{d\xi} = \frac{1}{\epsilon^\alpha}\frac{dY}{d\xi}, \tag{3.182}$$

and similarly for the second derivative. Our ODE (3.172) now becomes

$$\epsilon^{1-2\alpha}Y'' + 2\epsilon^{-\alpha}Y' + Y = 0. \tag{3.183}$$

Since our aim is to resolve what happens near $x = 0$ we impose only this condition

$$Y(0) = 0. \tag{3.184}$$

Notice that the introduced scaling is a *stretching transformation* that magnifies the vicinity of $x = 0$. The region around $x = 1$ escapes to infinity. This is precisely what we would like to have.

Now, we want to retain the highest order derivative in the limit $\epsilon \to 0$. To this end we have to balance it with one of the other terms in the equation. If the second derivative is of the same order as Y then we have to have $1 - 2\alpha = 0$ which gives $\alpha = 1/2$. The ODE becomes

$$\epsilon^{\frac{1}{2}}Y'' + 2Y' + \epsilon^{\frac{1}{2}}Y = 0, \tag{3.185}$$

which in the limit gives $Y' = 0$ with only a trivial solution satisfying $Y(0) = 0$. Therefore, we have to reject this balance. We are left with balancing Y'' with $2Y'$ and hence $1 - 2\alpha = -\alpha$ which implies $\alpha = 1$. Now,

$$Y'' + 2Y' + \epsilon Y = 0. \tag{3.186}$$

We can now in position to proceed with regular perturbation expansion. However, in order just to illustrate the concept we will compute only the leading order Y_0 by passing with ϵ to zero, i.e.

$$Y_0'' + 2Y_0' = 0, \quad Y_0(0) = 0. \tag{3.187}$$

The solution is

$$Y_0(\xi) = \frac{1}{2}D\left(1 - e^{-2\xi}\right), \tag{3.188}$$

with an arbitrary constant D. Now, we are facing with a completely different situation than before. For every D the above is a function that is called the *inner solution* since we are solving near $x = 0$ and this region is called the *boundary layer*. This nomenclature comes from fluid dynamics where even for flows with a very small viscosity the behaviour near a wall is completely different than in the free stream.

How do we determine D? One of the best ways to obtain a uniform expansion is to require that the inner solution becomes the outer one when leaving the boundary layer. That is, we impose the following condition

$$\lim_{\xi\to\infty} Y(\xi) = \lim_{x\to 0^+} y(x). \tag{3.189}$$

The limit of $x \to 0$ is understandable but why to take $\xi \to \infty$? The reason is the scaling, since for any fixed $x \in (0, 1]$ we have

$$\xi = \frac{x}{\epsilon} \to \infty \quad \text{as} \quad \epsilon \to 0. \tag{3.190}$$

We leave the boundary layer far to the right. The procedure of determining the constant of integration by equating inner and outer solutions is called *matching*. Applying the matching condition (3.189) to the leading order solutions Y_0 and y_0 gives

$$\lim_{\xi\to\infty} \frac{1}{2}D\left(1 - e^{-2\xi}\right) = \lim_{x\to 0^+} e^{\frac{1}{2}(1-x)}, \tag{3.191}$$

which is

$$\frac{1}{2}D = e^{\frac{1}{2}} \to D = 2e^{\frac{1}{2}}. \tag{3.192}$$

Therefore, our inner solution matched with the outer is

$$Y_0(\xi) = e^{\frac{1}{2}}\left(1 - e^{-2\xi}\right). \tag{3.193}$$

We have completed the perturbation theory for finding matching expansions for both regions: outer and boundary layer. We have two pieces of the exact solution but, unfortunately, they work only in each separate region. They agree on a very small *transition region* that can be seen on Fig. 18. Our ultimate goal is thus to find a uniform expansion in ϵ.

There are several ways of finding approximation that is uniformly valid in the whole interval $x \in [0, 1]$. The sought solution is called *composite*. Since the inner and outer solutions match they have a common part. The basic idea of finding a composite expansion is to add two solutions and subtract the part on which they are the same, i.e.

$$y_c(x) = Y_0\left(\frac{x}{\epsilon}\right) + y_0(x) - y_0(0), \tag{3.194}$$

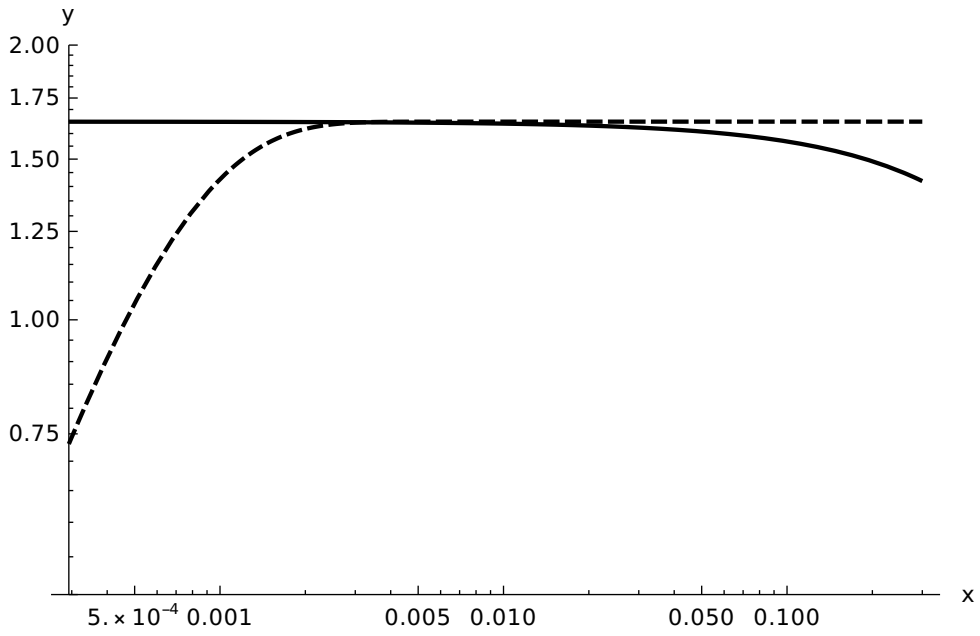

Figure 18: The common part of inner and outer solutions for $\epsilon = 10^{-4}$.

since the $y_0(0)$ is a constant that comes from the matching condition (3.189). Finally, we obtain a uniform composite leading order expansion

$$y_c(x) = e^{\frac{1}{2}} \left(e^{-\frac{x}{2}} - e^{-2\frac{x}{\epsilon}} \right). \tag{3.195}$$

On Fig. 19 we have collected all information that we have obtained concerning the solution of (3.172). Notice the decent accuracy of the composite expansion even though the inner and outer solutions are not so good for this value of ϵ. Note the boundary layer near $x = 0$. This is the place where the solution suddenly jumps from $y(0) = 0$ into the outer region. This is the most important feature of singular perturbation theory applied to boundary layers. The meaning of the second derivative is important only in the close vicinity of $x = 0$ where the derivative is large and it is possible to impose a boundary condition. The stretching transformation is an extremely useful technique to investigate what is happening inside very small regions of the domain. Thanks to this we are able to recast the problem into a shape that supports regular perturbation theory. □

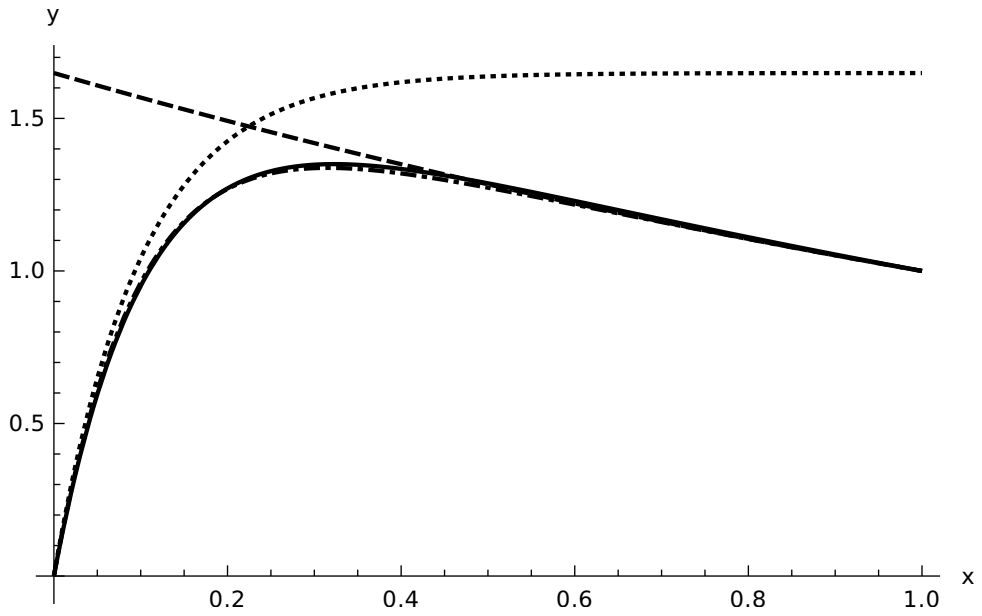

Figure 19: A comparison of various solutions of (3.172): exact (solid line), outer y_0 (dashed line), inner Y_0 (dotted line), and composite y_c (dot-dashed line). Here, $\epsilon = 0.2$.

The overall procedure for many other boundary layer problems can be summarized as follows.

1. By inspection notice the possible boundary layers.
2. Conduct the regular perturbation expansion to find the outer solution.
3. Rescale the dependent variable to magnify the boundary layer.
4. Choose the scaling parameter according to the consistent balance of terms in the equation.
5. Conduct the regular perturbation expansion to find the inner solution.
6. Match inner with outer solution.
7. Form a composite approximation.

The singular perturbation theory is not limited to boundary layers. Similar behaviour can be also found in initial value problems, in systems of equations, and in PDEs. There may also exist several boundary layers or other phenomena in which the solution is strongly dependent on ϵ. In the above example we have seen that inner solution was a function of $x\epsilon$ while the outer solution depended on x. This makes the composite solution really a function of two variables: x and ξ. These distinct scales of the domain are crucial in singular perturbation theory. For example, in analysing nonlinear oscillations one is usually faced with amplitude changing much slower than the period of vibrations. This two-timing causes the regular perturbation theory to fail. Several methods have been developed in order to deal with these problems: strained time coordinates, averaging, and most general - multiple scales method. They all help

to learn crucial facts about oscillations where nonlinearities cause certain exotic phenomena. This is a fascinating topic but, however, is out of scope of this lecture.

4 Kinetics

In this section we will focus on modelling of interactions of various quantities. We will only focus on kinetics, that is, the overall evolution of the investigated objects. We will not delve into the gory details of the considered reactions since, frequently, it is virtually impossible. Rather, we will base our reasoning on empirical laws concerning rates of reactions. This programme proved to be very successful in chemistry, biology, and physics. It constitutes the basis of kinetic modelling in these fields.

Example. (*Radioactive decay*) The simplest example of chemical reaction is the very well-known radioactive decay. For instance, tritium ^3H, the radioactive isotope of hydrogen, decays into Helium when stroke with cosmic radiation in the atmosphere. The full reaction is

$$^3H \longrightarrow {}^3He + e + \nu, \tag{4.1}$$

where e is the electron, while ν is neutrino. Since the masses of electron and neutrino ale magnitudes smaller than the mass of elements, we will focus on modelling their evolution. The main assumption is, of course, that the rate of decay is proportional to the amount of radioactive isotope. That is, if $H = H(t)$ is the amount of tritium, then

$$\frac{dH}{dt} = -kH, \quad t > 0, \tag{4.2}$$

which we all very well know. □

Example. (*Predator-Prey*) The next typical example is the predator-prey model known from the ODE course. If $P = P(t)$ is the number of predators, and $O = O(t)$ number of prey, the simplest Lotka-Volterra equations modelling dynamics are

$$\begin{cases} \dfrac{dP}{dt} = -\alpha P + \beta OP, \\ \dfrac{dO}{dt} = \gamma O - \delta OP, \end{cases} \tag{4.3}$$

with α, β, γ, and δ positive constants. The populations either decay (in case of predators) or grow (in case of prey) exponentially. The interaction is given by the products OP which say that populations can interact only when their both numbers are positive. This is sensible since when there would be no prey, the predators would die out of hunger. In realistic models this interaction term should be limited for larger populations. □

Example. (*SIR epidemic model*) The most famous model of epidemic dynamics is due to Kermack-McKendrick. Suppose that the population is divided between three parts: susceptible for infection S, infected I, and recovered (or dead) R. We assume that an individual once recovered gains a permanent immunity for the infection. The model

has the form

$$\begin{cases} \frac{dS}{dt} = -\alpha S, \\ \frac{dI}{dt} = -\beta I + \alpha SI, \\ \frac{dR}{dt} = \beta I. \end{cases} \tag{4.4}$$

We see that the infection rate is proportional to the product SI - when susceptibles interact with infected. Moreover, individuals recover at a constant rate proportional to their number. Notice that when we add all of the above equations we obtain

$$\frac{dS}{dt} + \frac{dI}{dt} + \frac{dR}{dt} = 0, \tag{4.5}$$

and hence, the total population number is conserved. This is an example of conservation law. □

4.1 Law of mass action

We can pursue the topic of chemical kinetics further. Suppose we have a reservoir of various species that are in constant motion, collide, and react with each other and produce another ones. Symbolically, this can be written

$$\text{reactants} \longrightarrow \text{products}. \tag{4.6}$$

For example, if a species A reacts with B to produce C we can write

$$A + B \longrightarrow C. \tag{4.7}$$

The rate of this reaction, in general, depend on concentrations of all reactant, and more precisely - their collision rate. In turn, this means that it depends on the concentrations of A and B. By using the same letter, which is common in kinetics, to denote concentrations we can write

$$\begin{cases} A' = -r(A, B), \\ B' = -r(A, B), \\ C' = r(A, B). \end{cases} \tag{4.8}$$

Note that the reaction is not possible if one of the species is absent, therefore we have $r(A, 0) = r(0, B) = 0$. Therefore, from Taylor series we know that

$$r(A, B) = kAB + ..., \tag{4.9}$$

where all lower terms vanish due to our observation. This is the reason that the product is the simplest expression of a reaction rate of two species.

The above example is a specific version of the *Law of Mass Action* that governs very general reactions. It is the following.

1. The rate r of reaction is proportional to the product of all reactants with respective powers representing number of molecules involved.

2. The time change of the concentration of each species, i.e. the temporal derivative, is the product of the rate and number of its molecules.

3. If there is a system of reactions, their rate add up.

This means that the reaction rate r is defined to be the *minus rate of consuming one molecule reactant* (or producing one molecule of the product). For example, suppose that we have a reaction

$$nA + mB \longrightarrow pC + qD, \tag{4.10}$$

where n, m, p, and q are number of molecules. According to the Law of Mass Action we have

$$r(A, B) = kA^nB^m, \tag{4.11}$$

and the evolution of the reaction is given by the system

$$\begin{cases} A' = -nkA^nB^m, \\ B' = -mkA^nB^m, \\ C' = pkA^nB^m, \\ D' = qkA^nB^m. \end{cases} \tag{4.12}$$

If we divide each equation by the respective number of molecules and add all equations together we obtain

$$\frac{d}{dt}\left(\frac{1}{n}A + \frac{1}{m}B + \frac{1}{p}C + \frac{1}{q}D\right) = 0, \tag{4.13}$$

which is a conservation law for this reaction. Similarly, we can subtract first two equations, or add the first and third, and still obtain a certain conservation law. This is very useful since it allows for reduction of equations. For instance, first two equations give us

$$\frac{1}{n}A - \frac{1}{m}B = \alpha, \tag{4.14}$$

for some constant α dependent on initial conditions. This gives us $B = m(\alpha - A/n)$ which, plugged into the equation for A, yields

$$A' = -nm^mkA^n(\alpha - A/n)^m, \tag{4.15}$$

which is a single ODE! Solving it, and substituting into other conservation laws gives the final solution. From this we immediately can obtain possible steady-states, that is $A = 0$ and $A = n/\alpha$. This is a very important technique.

Example. Suppose that the specie A reacts with itself to produce C

$$A + A \longrightarrow C. \tag{4.16}$$

Since there are two molecules of A we have

$$\begin{cases} A' = -2kA^2, \\ C' = kA^2. \end{cases} \tag{4.17}$$

The conservation law is $(A/2 + C)' = 0$ and hence,

$$C = \frac{1}{2}(A_0 - A) + C_0, \tag{4.18}$$

where A_0 and C_0 are initial concentrations. The ODE for A can easily be solved to give

$$A(t) = \frac{A_0}{1 + 2kA_0t}, \tag{4.19}$$

and hence

$$C(t) = \frac{1}{2}A_0\left(1 - \frac{1}{1 + 2kA_0t}\right) + C_0. \tag{4.20}$$

Therefore, we see that algebraically A decays to zero while C attains its steady state $A_0/2 + C_0$. Therefore, the amount of C increased by $A_0/2$. □

Example. As a more interesting example consider

$$\begin{aligned} A &\longleftrightarrow C + D, \\ A + B &\longrightarrow 2A + C. \end{aligned} \tag{4.21}$$

The first reaction is *reversible*, that is A changes into C and D and $C + D$ reacts into A. Since we have two reactions there are two rates with constants, say, k_1 and k_2. The reverse reaction proceeds with k_{-1}. We can write the equations

$$\begin{cases} A' &= -k_1A + k_{-1}CD - k_2AB + 2k_2AB = -k_1A + k_{-1}CD + k_2AB, \\ B' &= -k_2AB, \\ C' &= k_1A - k_{-1}CD + k_2AB, \\ D' &= k_1A - k_{-1}CD. \end{cases} \tag{4.22}$$

A possible conservation laws are $B + C - D = \alpha$ and $A + 2B + C = \beta$. This can allow us to reduce the number of equations from four to two and then use the phase plane analysis to check for critical points and their stability. Note the enormous utility of conservation laws. Without spotting them, we would have solved twice as complicated system. □

4.2 Michaelis-Menten kinetics

Many reactions in biochemistry undergo only when a certain enzyme is present. It acts as a key for the reaction to occur. Usually, only a very small concentration of an appropriate enzyme can trigger the reaction. The Michaelis-Menten model provides a useful description of the dynamics of such reaction. We will illustrate the concept on the following archetypal reaction introduced by Brown in the beginning of XX century do describe hydrolysis of sucrose. Later, this reaction was studied mathematically by Michaelis and Menten. It has the form

$$\begin{aligned} S + E &\longleftrightarrow C \\ C &\longrightarrow P + E, \end{aligned} \tag{4.23}$$

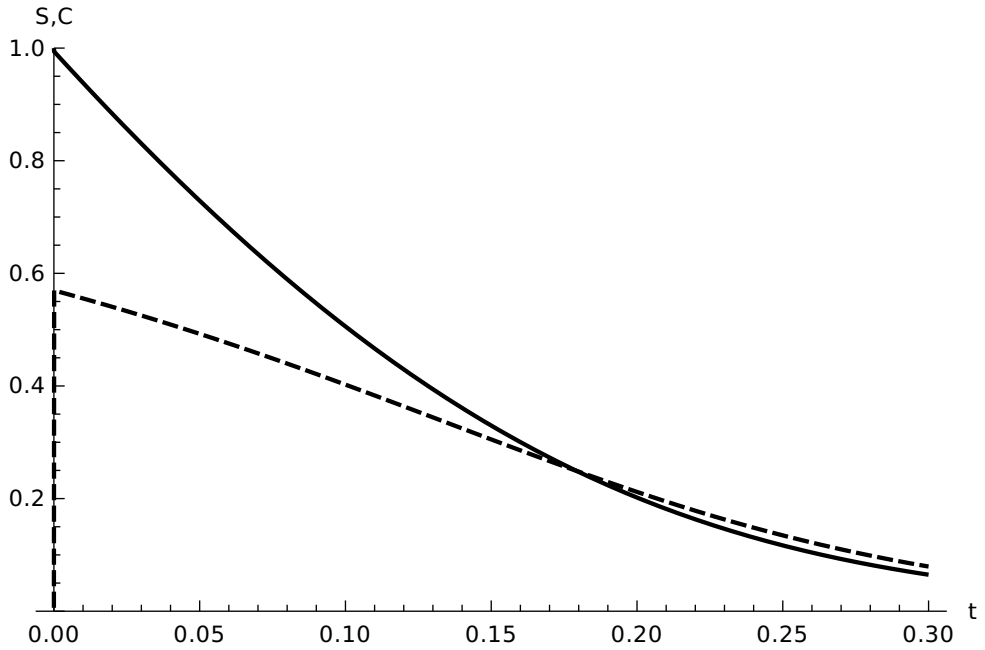

Figure 20: Numerical solutions of (4.25) with typical parameters. The solid line is S/S_0 while dashed C/E_0.

where S is the substrate, E the enzyme, C an intermediate complex, and P the product. The reaction equations are the following

$$\begin{cases} S' &= -k_1 SE + k_1 C, \\ E' &= -k_1 SE + (k_{-1} + k_2)C, \\ C' &= k_1 SE - (k_{-1} + k_2)C, \\ P' &= -k_2 C. \end{cases} \tag{4.24}$$

Two conservations laws are $(E + C)' = 0$ and $(S + C + P)' = 0$ and hence, using initial concentrations $S(0) = S_0$, $E(0) = E_0$, and $C(0) = D(0) = 0$, we can reduce the number of equations to two

$$\begin{cases} S' &= -k_1 E_0 S + (k_{-1} + k_1 S)C, \\ C' &= k_1 E_0 S - (k_2 + k_{-1} + k_1 S)C. \end{cases} \tag{4.25}$$

A typical numerical solution of the above is depicted on Fig. 20. We can see that the compound C instantaneously jumps from zero initial condition to some positive concentration. This is precisely the rapid reaction induced by the enzyme. If we supposed that there are two time-scales: fast on which C jumps, and slower on which S evolves, the compound would essentially be constant on the latter scale. Therefore, we would be tempted to set $C' = 0$ in the above system, solve an algebraic equation for C, and substitute in order to obtain a single ode for S. This premise is the main idea behind the so-called Quasi Steady-State Assumption (QSSA) that is widely used in chemical kinetics. We would like to justify that claim on the basis of perturbation theory.

We start with scaling of all relevant variables. The scaling for the substrate is simple

$$S = S_0 S^*, \tag{4.26}$$

where the nondimensional variable is denoted by S^*. Scaling with initial conditions is usually the correct choice for bounded functions. Next, we cannot use the initial condition for the compound since it is equal to zero. However, from the conservation law $E + C = E_0$ we know that C can be at most E_0 and thus

$$C = E_0 C^*, \tag{4.27}$$

for the new variable. Since no appropriate time-scale is evident, we nondimensionalize the time with yet unknown scale τ, i.e.

$$t = \tau t^*, \tag{4.28}$$

to be chosen in a moment. Plugging everything into (4.25) and omitting the asterisks, we obtain

$$\begin{cases} \dfrac{1}{\tau k_1 E_0} S' = -S + (\mu + S)C, \\ \dfrac{1}{\tau k_1 S_0} C' = S - (\nu + S)C, \end{cases} \qquad \mu = \frac{k_{-1}}{k_1 S_0}, \quad \nu = \frac{k_{-1} + k_1}{k_1 S_0}. \tag{4.29}$$

The equations are indeed simpler. To finish the scaling we have to choose τ. We have two choices, but only one produces the previously observed phenomenon of rapid initial change of C. That is, we take $\tau = 1/(k_1 E_0)$ and obtain

$$\begin{cases} S' = -S + (\mu + S)C, \\ \epsilon C' = S - (\nu + S)C, \end{cases} \tag{4.30}$$

where $\epsilon = E_0/S_0 \ll 1$. The smallness of ϵ is confirmed in reality - the concentration of enzyme is always much smaller than the substrate. Since a small parameter multiplies the derivative, we obtain a singular perturbation problem.

We will perform the leading order asymptotic analysis and retain the same letters to denote all the variables. First, for the outer solution we set $\epsilon = 0$ and obtain the aforementioned QSSA

$$\begin{cases} S' = -S + (\mu + S)C, \\ 0 = S - (\nu + S)C, \end{cases} \tag{4.31}$$

from which

$$C = \frac{S}{\nu + S}, \tag{4.32}$$

and

$$S' = -\frac{\lambda S}{\nu + S}, \quad \lambda = \frac{k_2}{k_1 S_0}. \tag{4.33}$$

This can be integrated to yield

$$\nu \ln S + S = -\lambda t + A, \tag{4.34}$$

where A is the integration constant. Now, we move to the boundary (initial) layer by stretching the time coordinate

$$T = \frac{t}{\epsilon}. \tag{4.35}$$

Equations become

$$\begin{cases} s' = \epsilon(-s + (\mu + s)c), \\ c' = s - (\nu + s)c, \end{cases} \tag{4.36}$$

The leading order is now

$$\begin{cases} s' = 0, \\ c' = s - (\nu + s)c, \end{cases} \tag{4.37}$$

and thus s is constant $s(T) \equiv 1$ from the initial condition. Further, the equation for the compound can be solved explicitly

$$c(T) = \frac{1}{1+\nu}\left(1 - e^{-(1+\nu)T}\right), \tag{4.38}$$

this is the rapid change of c inside the layer. Finally, we match the inner and outer solutions by requiring

$$\lim_{T\to\infty} s(T) = \lim_{t\to 0} S(t), \quad \lim_{T\to\infty} c(T) = \lim_{t\to 0} C(t), \tag{4.39}$$

which gives $A = 1$. The composite solution is thus

$$\nu \ln S(t) + S(t) \sim -\lambda t + 1, \quad C(t) \sim \frac{S(t)}{\nu + S(t)} - \frac{1}{1+\nu} e^{-(1+\nu)t/\epsilon}, \tag{4.40}$$

as $\epsilon \to 0$. Numerical calculations show that the approximation by this leading order is almost perfect. This shows how perturbation theory is useful in finding vert useful approximations of complicated problems.

5 Waves

In this section we will visit the general theory of waves. We can observe then in almost any circumstance both natural and industrial. There are waves on the surface of a pond, in the electromagnetic field that carry our mobile or Wi-Fi signals, in elastic materials that constitute buildings, acoustic waves that we can hear, optical waves that we can see, waves in atmospheres of planets or stars, and gravitational waves in spacetime. Engineers have to be familiar with wave dynamics in order to avoid resonances that may be fatal fro constructions.

Due to all of these distinct fields it is difficult to rigorously define what a wave is. There are surely two characteristics of a typical wave motion:

- the energy is transmitted to a regions away from the initial disturbance,
- the medium of propagation is not disturbed.

Mathematically, waves are almost always described by nonlinear partial differential equations. When displacements are small these can be linearised. And these are the ones that we will focus the most.

5.1 Kinematics of waves

Linear wave dynamics in homogeneous medium is usually governed by a constant coefficient linear PDE of the form

$$L(u) = 0, \tag{5.1}$$

where L is a linear partial differential operator and u is the sought displacement such has amplitude of the water surface, departure from the atmospheric pressure or perturbation of the electromagnetic field. We assume that the wave travels in one spatial dimension described by the variable $x \in \mathbb{R}$ (of course, we can generalize this to higher dimensions). The time is denoted by $t > 0$. As with ODEs we can seek for exponential solutions in the form

$$u(x,t) = \mathrm{Re}\, e^{i(kx-\omega t)}, \tag{5.2}$$

where for convenience we use complex variables. This solution is called the *plane wave*. Here, ω is the *angular frequency* of the wave, i.e. the number measuring how many oscillations there are in a unit of time (times 2π). That is, if T is the *period of oscillations*, then

$$\omega = \frac{2\pi}{T}. \tag{5.3}$$

Moreover, k is the *wave number* stating how many oscillations are in the unit of length (times 2π). That is, if λ is the *wave length* then

$$k = \frac{2\pi}{\lambda}. \tag{5.4}$$

Plugging the above ansatz into the governing PDE produces one of the most important features of wave motion - the *dispersion relation*

$$\omega = \omega(k), \tag{5.5}$$

which says that the frequency depends on the wave number. With these quantities we can associate the *phase speed*

$$c = \frac{\omega(k)}{k} = \frac{\lambda}{T(k)}. \tag{5.6}$$

This means that the phase speed is the distance travelled by the disturbance in one period. When c =const. we say that the wave is *nondispersive* otherwise it is *dispersive*.

Now, in reality waves are rarely plane. Usually, due to linearity, they are superpositions of plane waves of different frequency, that is a general solution of (5.1) can be written as

$$u(x,t) = \mathrm{Re}\int_{-\infty}^{\infty} A(k)e^{i(kx-\omega(k)t)}dk = \mathrm{Re}\int_{-\infty}^{\infty} A(k)e^{i\theta(k,x,t)t}dk. \tag{5.7}$$

where A is the complex amplitude or weighing factor of various plane waves. The above integral has to be convergent in order for this solution to have a meaning. There is a rich and beautiful theory behind that. We have introduced the *wave phase*

$$\theta(k,x,t) = \frac{kx}{t} - \omega(k), \tag{5.8}$$

which brings us to the historical beginning of the Method of Stationary Phase (Theorem 3). In many circumstances we would like to know the asymptotic form of the disturbance for large times, i.e. we would like to know the behaviour of the wave far from the initial time. This means that we look for the $t \to \infty$ limit when x/t is fixed in order to move with the wave. From the Stationary Phase Method we know that the behaviour of the integral is governed by the neighbourhood of the stationary point of the function θ, that is

$$\frac{\partial\theta}{\partial k} = 0 \to \frac{x}{t} = \frac{d\omega}{dk}. \tag{5.9}$$

The quantity $d\omega/dk$ is called *group velocity* and plays a fundamental role in the analysis of wave motion. Since x/t is fixed the above equation may have a solution $k = k_0$. Because of (3.118) we know that in this case

$$u(x,t) \sim \mathrm{Re}\, A(k_0)\sqrt{\frac{2\pi}{t|\omega''(k_0)|}}e^{i(k_0x-\omega(k_0)t)}e^{-i\,\mathrm{sgn}\,\omega''(k_0)\frac{\pi}{4}} \quad \text{as} \quad t\to\infty, \quad \frac{x}{t} \quad \text{fixed}. \tag{5.10}$$

We now see that a general solution of our linear PDE (5.1) for large times behaves as a sinusoidal plane wave moving with velocity $\omega(k_0)/k_0$ where k_0 is satisfies the equation $\omega' = x/t$. Moreover, its amplitude decays as $1/\sqrt{t}$. Further, by some more advanced methodology it can be shown that the energy of the wave moves with the group velocity. In relativity, for example, group velocity must always be smaller than the speed of light even though the phase velocity might exceed it! This phenomenon can happen in waveguides and other similar materials. This is not a contradiction with relativity since the compact superposition of waves, called the *wave packet*, travels at the group speed as we have seen from the method of stationary phase. Further intuition about these properties can be gained by analysing a simple situation of wave superposition.

Example. (*Group velocity*) Let us consider two plane waves with wave numbers k and $k+\Delta k$ with $|\Delta k| \ll |k|$. For convenience choose a real sinusoidal waves with amplitude equal to 1. Their superposition is equal to

$$u(x,t) = \cos(kx - \omega(k)t) + \cos((k+\Delta k)x - \omega(k+\Delta k)t). \tag{5.11}$$

Using the trigonometric formula for a sum of two cosines gives

$$u(x,t) = 2\cos\left(\left(k+\frac{\Delta k}{2}\right)x - \frac{\omega(k)+\omega(k+\Delta k)}{2}t\right)\cos\left(\frac{\Delta k}{2}x - \frac{\omega(k+\Delta k)-\omega(k)}{2}t\right). \tag{5.12}$$

Now, since Δk is small the first wave has a wave number close to k and is a short wave with respect to the other. The other has a very small wave number $\Delta k/2$ which corresponds to a long wave. We can think about it as wave with a wave number k which amplitude is modulated with a long wave as can be seen on Fig. 21. The phase speed of the short wave is

$$\frac{1}{2}\frac{\omega(k)+\omega(k+\Delta k)}{k+\frac{\Delta k}{2}} = \frac{\omega(k)}{k} + O(\Delta k) \quad \text{as} \quad \Delta k \to 0, \tag{5.13}$$

which is close to the original phase speed. The long wave, on the other hand, has

$$\frac{1}{2}\frac{\omega(k+\Delta k)-\omega(k)}{\frac{\Delta k}{2}} = \omega'(k) + O(\Delta k) \quad \text{as} \quad \Delta k \to 0, \tag{5.14}$$

which is very close to the group velocity of the original wave! We have thus found that the superposition of two waves is an amplitude modulating long wave that travels with group velocity. This envelope is then filled with shorter waves travelling at phase velocity. This is a good intuition that generalizes to arbitrary superpositions. In that case the group velocity tells us about the overall speed which the most important wave has. □

5.2 Dispersive waves

In this part we will see several important examples showing how the information can propagate in a dispersive manner. We will meet several partial differential equations however, the reader is not needed to understand them completely. After seeking wave-like solutions the mathematics reduces to algebra.

Example. (*Traveling wave*) Probably the simplest PDE that gives rise to an interesting wave phenomenon is the travelling wave equation

$$\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} = 0, \tag{5.15}$$

where $u = u(x,t)$, for example, describes contaminant concentration in river, density of cars on a highway, displacement of atoms in elastic material or a density of pack of

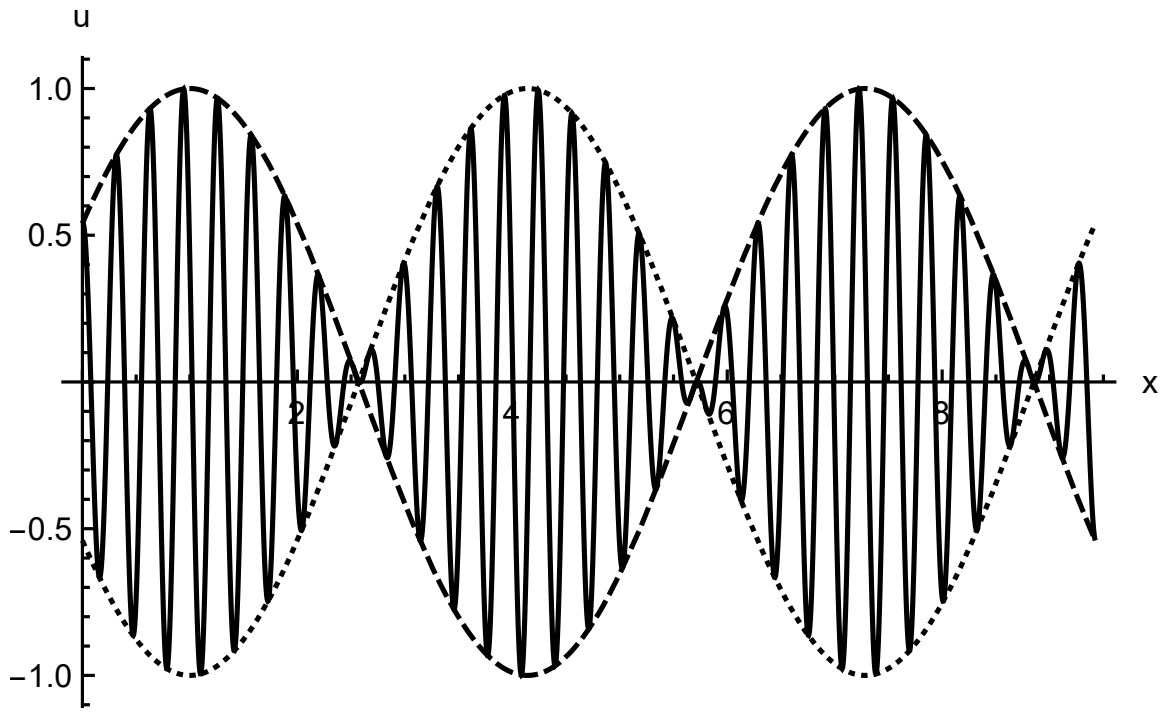

Figure 21: A schematic of a superposition of two waves with close wave numbers. A short wave (solid line) with wave number $k + \Delta k/2$ which amplitude is modulated by w long wave with wave number $\Delta k/2$ (dashed and dotted lines).

animals. Here, c is constant. We look for travelling waves $u(x,t) = \exp(i(kx - \omega t))$. After substitution we can cancel exponentials and obtain

$$-i\omega + ick = 0 \to \omega(k) = ck. \tag{5.16}$$

Therefore, the phase speed is equal to c and the wave is nondispersive. As can easily we calculated the group speed $\omega'(k) = c$ which is the same as phase speed. Since the above is a solution for any $k \in \mathbb{R}$ we can form a superposition

$$u(x,t) = \int_{-\infty}^{\infty} A(k) e^{ik(x-ct)} dk. \tag{5.17}$$

Suppose that the initial disturbance has the shape $f = f(x)$, i.e.

$$f(x) = u(x,0) = \int_{-\infty}^{\infty} A(k) e^{ikx} dk. \tag{5.18}$$

The integral is the same as the general solution with x replaced by $x - ct$. Therefore, our solution is

$$u(x,t) = f(x - ct). \tag{5.19}$$

Whence the name of the equation. The initial profile travels to the right for $c > 0$ and left for $c < 0$ with velocity c since the above formula is just a translation of the function f. Note that the wave does not change its shape. □

Example. (*Wave equation*) Linear waves are very frequently described by the following second order equation known simply as *wave equation*

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}. \tag{5.20}$$

The examples include elasticity, acoustics, and electromagnetism. Let us look for plane waves $u(x,t) = \exp(i(kx - \omega t))$, that is

$$\omega(k)^2 = c^2 k^2, \tag{5.21}$$

which gives $\omega = \pm ck$. These two modes correspond to nondispersive travelling waves going to the left and right with speed c. □

Example. (*Korteweg-de Vries equation*) In many physical situations the following linearised *Korteweg-de Vries* equation arises as a description of many wave phenomena

$$\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} = \alpha \frac{\partial^3 u}{\partial x^3}, \tag{5.22}$$

where c and α are constants. The dispersion relation for this case is

$$\omega(k) = ck + \alpha k^3. \tag{5.23}$$

Therefore, the wave is dispersive with phase velocity $\omega(k)/k = c + \alpha k^2$ while the group velocity is $\omega'(k) = c + 3\alpha k^2$. For example, with $\alpha > 0$ shorter wavelengths (k large) move faster than longer ones. Moreover, $\omega''(k) = 6\alpha k$. We can now look for stationary points that is

$$\omega'(k_0) = 0 \to k_0^2 = -\frac{c}{3\alpha}. \tag{5.24}$$

Therefore, the phase can be stationary for $c\alpha < 0$. □

Example. (*Beam equation*) Transversal vibrations of elastic beam are modelled by

$$\frac{\partial^2 u}{\partial t^2} + \gamma^2 \frac{\partial^4 u}{\partial x^4} = 0. \tag{5.25}$$

The dispersion relation is

$$\omega^2 - \gamma^2 k^4 = 0. \tag{5.26}$$

There are two modes of vibrations

$$\omega_+(k) = \gamma k^2, \quad \omega_-(k) = -\gamma k^2. \tag{5.27}$$

The phase velocities are $\omega_\pm = \pm\gamma k$ while group velocities $\omega'_\pm = \pm 2\gamma k$. The only stationary point is this $k_0 = 0$. □

Knowing the dispersion relation is fundamental in determining the characteristics of waves. In the next subsection we will illustrate all of these ideas on concrete and important examples.

5.3 Water waves

The most familiar, useful, and interesting waves are these produces on the surface of a basin of water. We will study them in detail. Deriving water flow equations is a very educative, however, advanced task and we have to omit some details in this basic treatment. On the other hand, we would like to justify all formulas that we obtain.

5.3.1 Derivation

Let us start by considering a three dimensional flow o water in the $x - y - z$ space. That is, by x and y we denote the horizontal orthogonal coordinates, while z is the vertical (see Fig. 22). As usual, the time is denoted by $t > 0$. Fluid velocity will be denoted by a vector $\mathbf{u}(x, z, t) = (u(x, y, z, t), v(x, y, z, t), w(x, y, z, t))$ where u and v are horizontal components of the flow velocity while w is the vertical. We have to make several assumption of the nature of the vector field $\mathbf{u}$.

- The water has constant density.
- The flow is incompressible.
- The flow is irrotational.

The first assumption above can be stated as the vanishing of divergence

$$u_x + v_y + w_z = 0, \tag{5.28}$$

where subscripts denote partial derivatives with respect to given variable, i.e. $u_x = \partial u/\partial x$. In order to understand that it is needed to utilize some results from multidimensional calculus. However, we will only illustrate this on a one dimensional case. Choose any interval $[a(t), b(t)]$ that moves with the flow, that is $u(a, t) = a'(t)$ and $u(b, t) = b'(t)$. Then, the length of that interval is

$$L(t) = b(t) - a(t). \tag{5.29}$$

Computing the derivative gives $L'(t) = b'(t) - a'(t) = u(b(t), t) - u(a(t), t)$. Now, since the flow is incompressible the length does not change with time

$$0 = u(b(t), t) - u(a(t), t) = \int_{a(t)}^{b(t)} u_x(x, t)dx, \tag{5.30}$$

where we have used Newton-Leibniz theorem. Since the above is valid for all arbitrarily chosen a and b we conclude that $u_x = 0$ which is one-dimensional divergence. The proof for higher dimensions is similar and utilizes Reynolds transport and divergence theorems.

The flow is also irrational meaning that its rotation vanishes[14], that is

$$\nabla \times \mathbf{u} = 0. \tag{5.31}$$

[14]Recall that the *rotation* of a vector field is defined as $\nabla \times \mathbf{u} = (w_y - v_z, u_z - w_x, v_x - u_y)$ and *nabla operator* is $\nabla = (\partial_x, \partial_y, \partial_z)$.

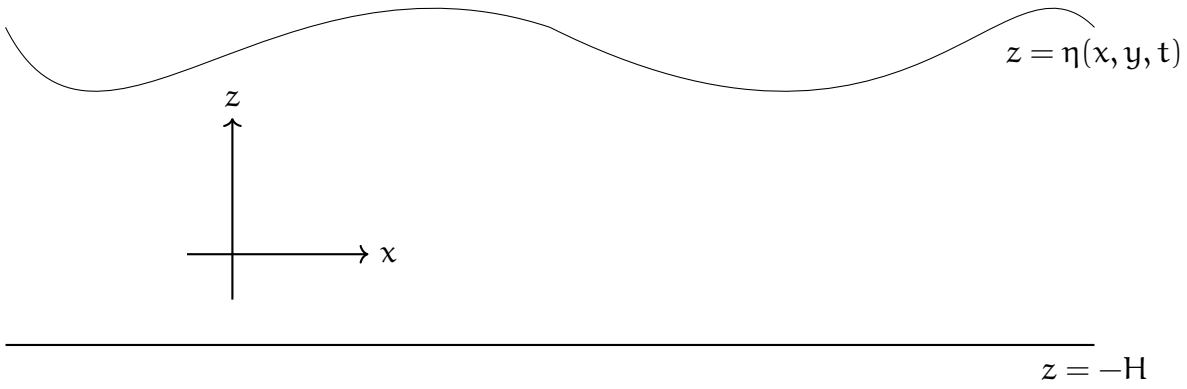

Figure 22: A schematic of the water flow. The y direction is suppressed.

From multidimensional calculus we know that in a simply connected region an irrotational vector field can be written as a gradient of a potential ϕ

$$u = \phi_x, \quad v = \phi_y, \quad w = \phi_z. \tag{5.32}$$

This is very good since we can now investigate only one function. The incompressibility condition (5.28) is now[15]

$$\Delta\phi = \phi_{xx} + \phi_{yy} + \phi_{zz} = 0, \tag{5.33}$$

which is known as Laplace's equation.

Further, we have to add some dynamics to our flow since, so far, it is purely kinetic. The fluid has to conserve momentum and the dynamical equations are just mathematical statements of this principle. We assume that there is only one force acting on the volume (bulk) of the fluid - namely gravity. Also, let the fluid be inviscid so that there is no dissipation of energy by internal friction. We are thus left only with the pressure p representing internal forces. This leads us to

$$\rho\frac{d}{dt}u(x(t), y(t), z(t), t) = -p_x, \quad \rho\frac{d}{dt}v(x(t), y(t), z(t), t) = -p_y \quad \rho\frac{d}{dt}w(x(t), y(t), z(t), t) = -p_z - \rho g \tag{5.34}$$

which says that the change of the momentum of a particle of the fluid with a trajectory $t \mapsto (x(t), y(t), z(t))$ is equal to the (minus) gradient of the pressure and gravity (in the vertical direction). When we use the chain rule to compute the temporal derivative we obtain[16]

$$\begin{cases} u_t + uu_x + vu_y + wu_z = -\dfrac{1}{\rho}p_x, \\ v_t + uv_x + vv_y + wv_z = -\dfrac{1}{\rho}p_y, \\ w_t + uw_x + vw_y + ww_z = -\dfrac{1}{\rho}p_z - g. \end{cases} \tag{5.35}$$

[15]Here, we are dealing with the Laplacian $\Delta = \nabla \cdot \nabla$.

[16]Note that $x'(t) = u(t)$ etc.

These are the celebrated Euler equations of fluid dynamics. They are formidable and very well investigated but still, they pose many different difficulties and challenges. There are books written on them. Now, Euler equations can be cast into some more trackable form thanks to our assumption of irrotational (potential) flow. Plugging in the potential (5.32) into Euler equations yields a vector equation

$$\nabla\phi_t + \frac{1}{2}\nabla(\phi_x^2 + \phi_y^2 + \phi_z^2) = -\frac{1}{\rho}\nabla p - \nabla(gz). \tag{5.36}$$

Since everywhere above we have an equality of gradients it follows that there exist a function $H(t)$ such that[17]

$$\phi_t + \frac{1}{2}(\phi_x^2 + \phi_y^2 + \phi_z^2) + \frac{1}{\rho}p + gz = H(t). \tag{5.37}$$

This is the important *Bernoulli's Law*. Notice that we have obtained a closed system of equations for two unknowns : Laplace's equation for the potential (5.33) and Bernoulli's Law (5.37). By a suitable redefinition of the potential, i.e. $\phi = \widetilde{\phi} + \int H(t)dt$ with $\Delta\widetilde{\phi} = 0$, we can set the integration constant equal to zero and we will make this choice without the loss of generality.

5.3.2 Boundary conditions

So far we have devised a system of differential equations that describe a whole generality of water flow situations. Here, we would like to focus on waves and in order to pursue this task we have to impose some specific boundary conditions. First, we assume that the bottom of our basin is flat, that is there is a impenetrable boundary at $z = 0$. Second, there is a free boundary on which the pressure (usually atmospheric) is prescribed. The word *free* means that we do not know the shape of the surface in advance and it has to be found as a part of the solution. We denote it by $z = \eta(x, y, t)$.

In general we have two types of boundary conditions for the fluid flow: kinematic and dynamic. Kinematic condition means that the fluid cannot penetrate the boundary. In the case of the flat bottom this means that

$$w = \phi_z = 0 \quad \text{on} \quad z = -H, \tag{5.38}$$

where $H > 0$ is the depth of the basin. That is, there is no vertical velocity component on the bottom. The kinematic condition on the surface requires a little bit more more since we do not know η. Let $0 = F(x(t), y(t), z(t), t) = z(t) - \eta(x(t), y(t), t), t)$ denote the particle trajectory along the boundary of the fluid. Since this surface is impenetrable we have $dF/dt = 0$ and hence

$$w = \eta_t + u\eta_x + v\eta_y \quad \text{on} \quad z = \eta(x, y, t), \tag{5.39}$$

which in the language of the potential can be written as

$$\phi_z = \eta_t + \phi_x\eta_x + \phi_y\eta_y \quad \text{on} \quad z = \eta(x, y, t). \tag{5.40}$$

[17] It is a vector analogue of a scalar equation $f'(x) = 0$.

The dynamic boundary condition is the prescription of the pressure. We give such at the free surface

$$p = p_0 \quad \text{on} \quad z = \eta(x, y, t). \tag{5.41}$$

Without the loss of generality we can take $p_0 = 0$. Due to Bernoulli's Law (5.37) we can write

$$\phi_t + \frac{1}{2}(\phi_x^2 + \phi_y^2 + \phi_z^2) + g\eta = 0 \quad \text{on} \quad z = \eta(x, y, t), \tag{5.42}$$

where we have used our remark that we can take $H(t) = 0$. Therefore, the potential has to satisfy two boundary conditions on the free surface (5.40) and (5.42). This seemingly overdetermination of the system is only apparent - the free surface $z = \eta$ is not known and hence we need an additional condition to determine it.

5.3.3 Linearisation

Now, the equations of motion (5.33) and (5.37) with boundary conditions (5.38), (5.40), and (5.42) are nonlinear and describe a multitude of situations. We would like to focus on these phenomena in which the displacement of the free surface along with velocity and its derivatives are small. That is, we would like to linearise the equation. To this end we use perturbation theory with respect to the basic state of the so-called lake at rest, that is we investigate small perturbations to the following situation

$$\phi_0 = 0, \quad \eta_0 = \eta_0(x, y, t). \tag{5.43}$$

Denote the magnitude of the perturbation by $\epsilon > 0$ and consider the expansion

$$\phi = \epsilon\phi_1 + \epsilon^2\phi_2 + O(\epsilon^3), \quad \eta = \eta_0 + \epsilon\eta_1 + \epsilon^2\eta_2 + O(\epsilon^3), \tag{5.44}$$

as $\epsilon \to 0$. We would like to find only equations for the leading order perturbations. First, it is straightforward to see that due to linearity each ϕ_k is a solution of the Laplace's equation

$$\Delta\phi_k = 0. \tag{5.45}$$

Similarly, the kinetic boundary condition at the bottom (5.38) is

$$(\phi_k)_z = 0 \quad \text{on} \quad z = -H. \tag{5.46}$$

On the other hand, by plugging the perturbation expansion into the dynamic condition (5.42) and equating like terms gives a system of equations

$$\begin{cases} \epsilon^0: & \eta_0 = 0, \\ \epsilon^1: & (\phi_1)_t + g\eta_1 = 0, \\ \epsilon^2: & (\phi_2)_t + \frac{1}{2}\left((\phi_1)_x^2 + (\phi_1)_y^2 + (\phi_1)_z^2\right) + g\eta_2 = 0, \\ \dots \end{cases} \quad \text{on} \quad z = \eta_0. \tag{5.47}$$

Therefore, the water surface is indeed at rest in the leading order. And we can evaluate the boundary condition at $z = 0$. This is a great advantage since in the leading order

the boundary condition is imposed on a fixed value of z and not at the unknown η. Further, the kinematic boundary condition (5.40) now becomes

$$\begin{cases} \epsilon^1: & (\phi_1)_z = (\eta_1)_t, \\ \epsilon^2: & (\phi_2)_z = (\eta_2)_t + (\phi_1)_x(\eta_1)_x + (\phi_1)_y(\eta_1)_y, \quad \text{on} \quad z = 0, \\ \ldots \end{cases} \tag{5.48}$$

where we have utilized the fact that $\eta_0 = 0$. We have thus obtained everything we wanted, that is a complete characterisation of equations for ϕ_1 and η_1. Summing up and denoting the perturbation ϕ_1 by ϕ and η_1 by η (this is an abuse of notation however, we will not go further in determining subsequent perturbations) we finally obtain the *linearised equations of water flow*.

1. Governing equation for the potential

$$\Delta\phi = 0. \tag{5.49}$$

2. Surface boundary conditions on $z = 0$

$$\phi_t + g\eta = 0, \quad \phi_z = \eta_t. \tag{5.50}$$

   By differentiating the first equation with respect to t and using the second we can eliminate η

$$\phi_{tt} + g\phi_z = 0 \quad \text{on} \quad z = 0. \tag{5.51}$$

3. Boundary condition at the bottom $z = -H$

$$\phi_z = 0. \tag{5.52}$$

The usual procedure of solving the above is first to solve the Laplace's equation (5.49) along with boundary conditions (5.51) and (5.52). Then, solve the first equation in (5.50) in order to determine the free surface η. Next, the velocity can be found from the definition of the potential (5.32). Finally, the pressure distribution in the water can be found from Bernoulli's Law (5.37)

$$p = -\rho g z - \rho\phi_t. \tag{5.53}$$

Notice that the first term above represents the hydrostatic pressure.

We have thus obtained an enormous simplification of the problem. Equations are linear, decoupled and the domain is simple. That is to say, we reduced the nonlinear system of equation with free boundary into a linear single equation for the potential on the following simple domain

$$(x, y, z) \in \mathbb{R}^2 \times [-H, 0]. \tag{5.54}$$

The domain is so regular that in order to find the general solution we can use separation of variables or similar technique used in solving linear constant coefficient equations. In what follows we will only focus on waves.

5.3.4 Linear water waves

We will look for plane water waves in the $x-y$ direction when their amplitude can change with the depth. Therefore, we propose the following ansatz

$$\phi(x,y,z,t) = \Phi(z)e^{i(\mathbf{k}\cdot\mathbf{x}-\omega t)}, \tag{5.55}$$

where the *wave vector* $\mathbf{k} = (k_x, k_y)$ is a straightforward generalization of the wave number while $\mathbf{x} = (x,y)$. The dot product is denoted by $\cdot$. We also allow for complex solutions in order to facilitate the reasoning. At the end we can always take the real part. We now plug the above into (5.49) to obtain an equation for $\Phi(z)$

$$\Phi'' - K^2\Phi = 0, \tag{5.56}$$

where $K^2 = k_x^2 + k_y^2$. The boundary conditions are the following

$$-\omega^2\phi(0) + g\phi'(0) = 0, \quad \phi'(-H) = 0. \tag{5.57}$$

The solutions of the second order constant coefficient linear ODE are build up from hyperbolic functions from which only the cosine satisfies the boundary condition at the bottom. Therefore,

$$\Phi(z) = C\cosh K(z+H). \tag{5.58}$$

Now, we plug the boundary condition at the surface to obtain

$$-\omega^2\cosh KH + gK\sinh KH = 0, \tag{5.59}$$

That is the dispersion relation for the water waves is

$$\omega = \pm\sqrt{gK\tanh KH}. \tag{5.60}$$

This is a strongly nonlinear formula for the frequency of oscillations with respect to the wave number. The waves that we have obtained are called *surface gravity waves* since they are generated by the gravity. The phase speed is

$$\frac{\omega}{K} = \pm\sqrt{gH}\sqrt{\frac{\tanh KH}{KH}}. \tag{5.61}$$

There are two important limits that have to be considered.

- Deep water waves. These waves arise when the surface is far away from the bottom to feel its presence. Mathematically speaking, the wave length in this situation is much smaller than the depth, that is

  $$KH \gg 1. \tag{5.62}$$

  For these short waves we have $\tanh KH \approx 1$ and hence

  $$\omega \approx \pm\sqrt{gK}, \quad \frac{\omega}{K} \approx \sqrt{\frac{g}{K}}, \quad \frac{d\omega}{dK} = \pm\frac{1}{2}\sqrt{\frac{g}{K}}. \tag{5.63}$$

  Note that the quantities above do not depend on the depth on the basin. The waves are dispersive with long waves travelling faster than short ones. The real-world example of these waves is a throwing a pebble into the pool. First, all wavelengths are present but after few moments, the longer waves overtake short ones and reach distant objects before them.

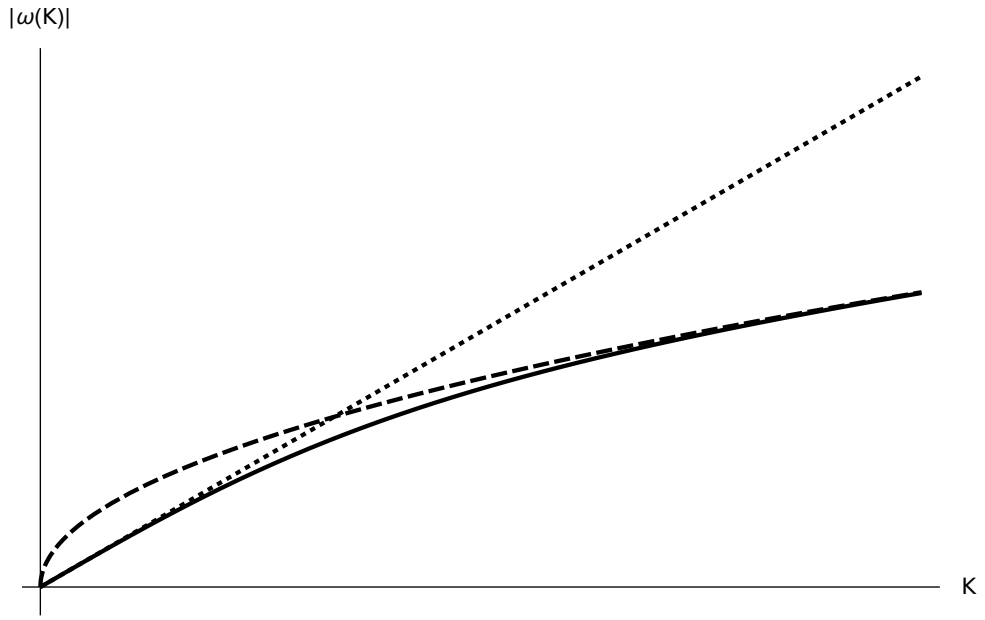

Figure 23: The dispersion relation for water waves with H fixed. The solid line represents the exact formula (5.60) while the dashed line is the deep water approximation and dotted line is shallow water limit.

- Shallow water waves. When the water is shallow, that is the wavelength is larger than the depth

$$KH \ll 1, \tag{5.64}$$

we can approximate $\tanh KH \approx KH$. Then,

$$\omega = \pm K\sqrt{gH}, \quad \frac{\omega}{K} = \pm\sqrt{gH}, \quad \frac{d\omega}{dK} = \pm\sqrt{gH}. \tag{5.65}$$

This means that these waves are nondispersive for which the speed is related to the depth. This approximation is very good for large scale flow in the ocean or atmosphere. Tides or tsunamis are also neatly modelled by the above relations.

The above limits are summarized on Fig. 23. In practice, it is assumed that shallow water waves occur for $k^{-1} > 20K$ while deep water waves have $k^{-1} < 2H$. Then, the relative error that we make is smaller than around 5%. These limiting behaviours of the dispersion relation can explain the common phenomena that ocean waves approach the beach approximately in parallel while far away into the see they are sloping. Fix a wavelength $\lambda = 2\pi/K$. Waves on the deep water have speed proportional to $\sqrt{\lambda}$. On the other hand, near the beach the water is shallow and hence the speed is a multiple of $\sqrt{H}$ where H is the typical shallow water depth. However $H \ll \lambda$ and thus the fraction of the wave near the coast slows with respect to the remaining open see part. The wave thus straightens up. This is, of course, an extremely simple explanation of that phenomenon, however, it is very revealing.

Finally, we would like to find the trajectories of fluid particles carried over along the gravity waves. For simplicity let us assume that we consider motion only in one

plane $y = 0$. Then, according to (5.32) and our solution the velocity has the following coordinates

$$u(x,z,t) = -K\cosh K(z+H)\sin(Kx-\omega(K)t), \quad w(x,z,t) = K\sinh K(z+H)\cos(Kx-\omega(K)t), \tag{5.66}$$

where we have taken the real part of the solution. Note that the velocity vector traces an ellipse with time, that is

$$\left(\frac{u}{K\cosh K(z+H)}\right)^2 + \left(\frac{w}{K\sinh K(z+H)}\right)^2 = 1. \tag{5.67}$$

The horizontal amplitude of the velocity in bounded away from zero, however, the vertical one decreases exponentially and vanishes at $z = -H$. The ellipse becomes squeezed to a line at the bottom.

5.3.5 Gravity-capillary waves

When deriving our equations for water waves we have neglected surface tension σ that contributes to the capillary pressure. When included the dispersion relation describes the *gravity-capillary waves*

$$\omega = \pm\sqrt{gK + \frac{K^3\sigma}{\rho}}\sqrt{\tanh KH}, \tag{5.68}$$

which we have met in (2.53) as a result of dimensional analysis (in the deep water approximation). We immediately see that due to the cubic of K the capillary waves are much more sensitive on the wave length. In particular, for long waves with $K \ll 1$, they are negligible with respect to gravity waves. The typical length of such a wave is usually smaller than 2 centimetres and they are called *ripples*. Observe that next time you visit a lake.

The phase speed on the gravity-capillary waves provides an interesting phenomenon. For simplicity let us focus on deep water approximation. For then, we have

$$c(k)^2 = \left(\frac{\omega}{K}\right)^2 = \frac{g}{K} + \frac{K\sigma}{\rho}. \tag{5.69}$$

When $K \to 0^+$ the above approaches $\pm\sqrt{gH}$ while for $K \to \infty$ the speed becomes infinite. The derivative is then

$$\frac{d}{dk}\left(c(k)^2\right) = -\frac{g}{K^2} + \frac{\sigma}{\rho}. \tag{5.70}$$

Therefore, the square of the phase speed decreases for small K, attains a minimum

$$c_m^2 = 2\sqrt{\frac{g\sigma}{\rho}} \quad \text{for} \quad K_m = \sqrt{\frac{\rho g}{\sigma}}, \tag{5.71}$$

and linearly increases to infinity as $K \to \infty$. The remarkable observation is that there exists a minimum velocity of a gravity-capillary wave. A numerical values for water-air

interface are $c_m = 0.23$ m/s and $\lambda_m = 2\pi/K_m = 1.7$ cm. In principle, a gust of wind slower than this minimal value will not produce any waves over the water. Similarly, a fishing line moving in water will not cause any waves when its velocity is smaller than c_m. However, when it is faster than that, a two sets of waves will be excited: one capillary for $K > K_c$ and one gravity for $K < K_c$.

5.3.6 Ship waves

Wave theory is, of course, extremely important for all nautical applications starting with ship construction and exploration and ending on naval warfare. There is a fascinating universal law that governs all relatively slow swimmers on deep water: beetles, ducks, boats, and ships. An object moving in water generates a *wake* of waves that follow its path (see Fig. 24) It is striking that the spread of that wedge-shaped wake is independent on the size and velocity of the swimmer! (As long the velocity is small enough). We will derive this result with a use of our water waves theory.

Consider an ship moving with velocity v in the given direction in deep water and treat it as a point source. The movement generates waves with wave vector $\mathbf{k}$ that subtends an angle α with the ship (see Fig. 25). In principle the ship produces waves of arbitrary wavelength and direction. However, almost all of them are dissipated by viscous forces. The wave numbers that survive are these that align their crests precisely with the bow of the ship (not seen on our crude point approximation). In other words, the waves in the wake have to be stationary with respect to the hull. Therefore, the phase speed of the wave has to satisfy

$$\frac{\omega}{K} = v \cos \alpha. \tag{5.72}$$

However, for deep waters we have found that $\omega/K = \sqrt{gK}$ and hence the the wave that is created has the number

$$K = \frac{g}{v^2 \sin \alpha}. \tag{5.73}$$

Of course the above is true to any inclination angle α. Therefore, the ship produces a whole spectrum of different waves. However, many of them will cancel with each other leaving only the dominant wave moving with the group speed for which the phase is stationary. This is precisely the Method of Stationary Phase that we have developed previously and is usually called the constructive interference.

In order to find the points of stationary phase let P be an arbitrary point riding on a crest. Let r be the distance of P from the ship and β the angle subtended by the radius and the ship's velocity. The phase at that point is

$$\varphi = \mathbf{k} \cdot \mathbf{x} = -kr \cos(\frac{\pi}{2} - (\alpha - \beta)) = -kr \sin(\alpha - \beta), \tag{5.74}$$

where the minus sign is because of opposite direction of vectors $\mathbf{k} = (k_x, k_y)$ and $\mathbf{x} = -(r \cos \beta, r \sin \beta)$. The phase of such a wave is stationary if $\varphi' = 0$, therefore

$$0 = \varphi'(\alpha) = -\frac{gr}{b^2} \left(\frac{\cos(\alpha - \beta) \sin^2 \alpha - 2 \sin \alpha \cos \alpha \sin(\alpha - \beta)}{\sin^4 \alpha} \right). \tag{5.75}$$

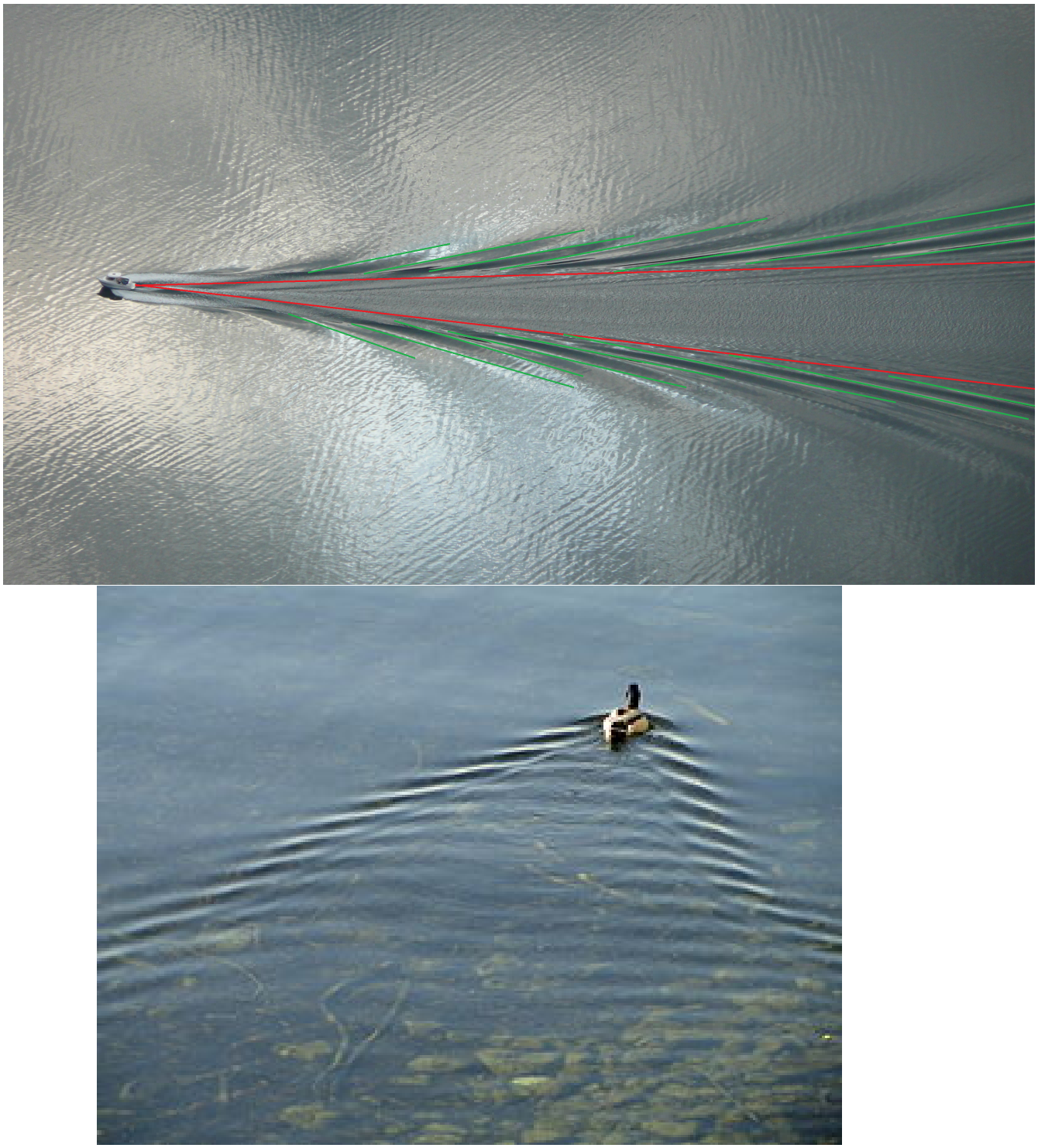

Figure 24: A photograph of a ship and duck wakes. Source: Wikipedia

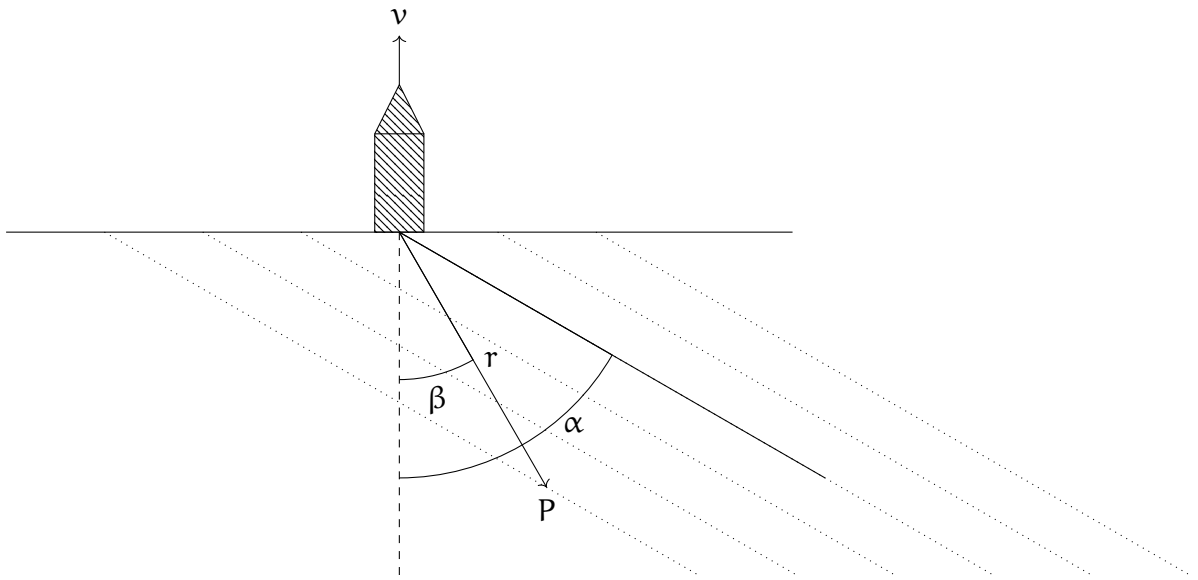

Figure 25: A schematic of the Kelvin wake creation.

The above is easily solved by

$$2\tan(\alpha-\beta)=\tan\alpha. \tag{5.76}$$

Therefore, the waves interfere constructively for this precise relation between angles. Using the formula for a tangent of a difference we can simplify

$$2\frac{\tan\alpha-\tan\beta}{1+\tan\alpha\tan\beta}=\tan\alpha, \tag{5.77}$$

from which

$$\tan\beta=\frac{\tan\alpha}{2+\tan^2\alpha}. \tag{5.78}$$

The optimal value of β for a given α is depicted on Fig. 26.

We immediately see that there are two extrema which can be found analytically. Set $\beta=\beta(\alpha)$ and differentiate (5.78) with respect to α to obtain

$$\frac{1}{\cos^2\beta}\beta'(\alpha)=\frac{1}{\cos^\alpha}\frac{2-\tan^2\alpha}{(2+\tan^2\alpha)^2}. \tag{5.79}$$

Therefore, the extrema are located at such $\alpha_\pm$ for which $\beta'(\alpha_{pm})=0$, that is

$$\alpha_\pm=\pm\arctan\sqrt{2}\approx\pm54.7^\circ, \tag{5.80}$$

and the corresponding extrema are

$$\beta_\pm=\pm\arctan\frac{\sqrt{2}}{4}=\pm\arcsin\frac{1}{3}=\pm19.5^\circ. \tag{5.81}$$

Note that all the physical model parameters have dropped out and we are left with only numerical values that are universal! The only assumption that we have taken is the

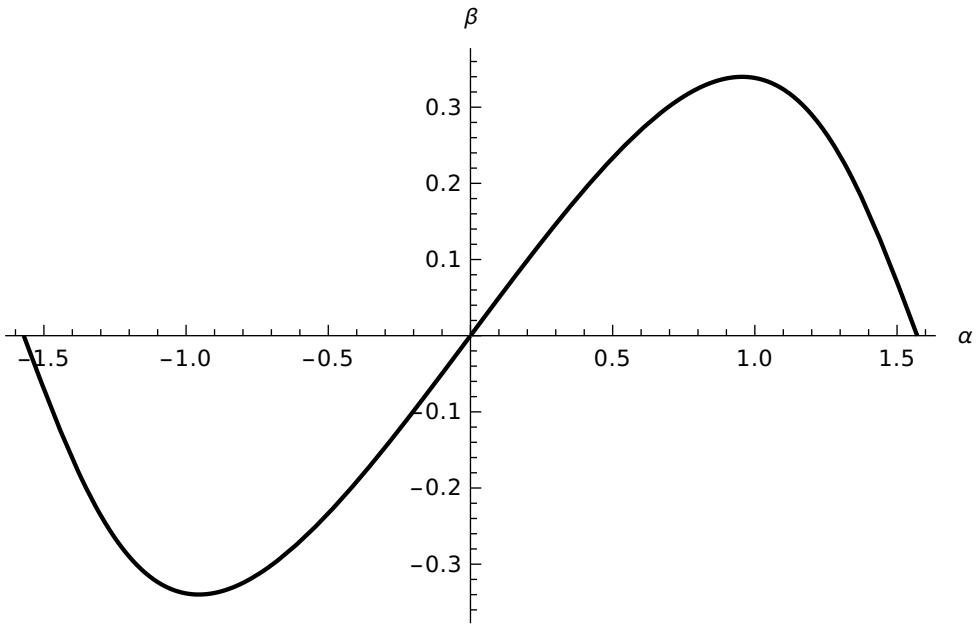

Figure 26: The graph of $\beta = \beta(\alpha)$ satisfying (5.78).

dispersion relation (5.63). Observe how this little of information can have tremendous consequences on large scale behaviour. Henceforth, all points after the ship that are inclined with an angle at most 19.5° will undergo a constructive interference to produce a wake (red line on Fig. 24). From it, the visible portion of waves caused by the passing ship will have a maximal inclination of 54.7° (green lines on Fig. 24). Observe that when you visit a lake next time and try to find some examples on Google Maps!

The above calculations were first given by Lord Kelvin and the solution is known to be the *Kelvin Wake*. However, there is a beautiful geometric reasoning due to Mach that leads to the same conclusion. The main assumption is that on the deep water the group velocity is half the phase velocity. Surprisingly, this is only needed to find out that $\beta_\pm = \pm 19.5°$.

6 Calculus of variations and optimisation

Optimisation is one of the cornerstones of applied and industrial mathematics. We have dealt with simple optimization problems as early as in high school when finding extrema of functions. These method were later generalized to higher dimensions on Calculus. Now, we will provide an ultimate generalization that is not only useful in industrial real-world problems but also is a foundation of many (if not all) physical theories and provides a beautiful results in pure and applied mathematics. We will develop the calculus of variations.

The calculus of variations is based on finding stationary points of functionals which are mappings from a given space of functions into some field such as $\mathbb{R}$. The arguments of functionals are functions themselves which enlarges the space where to look for this critical point. This makes calculus of variations much harder than ordinary calculus. This notion is not purely mathematical. Optimisation of various functionals is ubiquitous in physics starting with Lagrangian and Hamiltonian mechanics, Fermat's principle of least time in optics, "paths of least resistance" in electrostatics, formulating quantum field theory, and deriving Einstein's equations of general relativity. In mathematics some prominent examples include brachistochrone (curve of fastest descent), Plateau's problem (finding a surface of minimal area with a given boundary), geodesics (curves of shortest lengths), eigenvalue problems, and isoperimetric problems (such as finding optimal curves of a given are). Moreover, many industrial problems can be modelled by optimal control, that is controlling a given process in order to minimize the cost. Finally, a paramount numerical methods of solving a large class of PDEs over realistic geometries - finite element method - is based on variational principles. This makes the subject very robust, interesting, and important.

Historically, the first variational problems were solved by Newton. He showed his ingenuity in finding the surface of revolution that experiences the least resistance in a fluid and solved the brachistochrone problem posed by Jakob Bernoulli. Further advancements were made by Lagrange who formulated the mechanics in the language of variations and Euler who greatly developed the theory. Calculus of variations has also focused attention of Jacobi, Gauss, Poisson, Noether, Hilbert, and Lebesgue among others. It happened to be a very fruitful field.

6.1 Euler-Lagrange equations

Let us start with a general functional $J : C^n[a, b] \to \mathbb{R}$

$$J(y) = \int_a^b L(x, y(x), y'(x), ..., y^{(n)}(x))dx. \tag{6.1}$$

The kernel of the above functional, that is L, is called *Lagrangian*. Examples include area functional for which $L(y(x)) = y(x)$ and arc length with $L(y'(x)) = \sqrt{1 + y'(x)^2}$. The most important functionals often involve derivatives only up to order 1 and, hence, we will focus on them.

Following our knowledge from calculus we would like to find necessary (and sufficient) conditions for (6.1) to have an extremum. This is a complicated subject for which a careful and subtle analysis is required. We will follow a physical approach in

which only the knowledge whether a given function is an *stationary* or *critical* point of the functional J is required (that is, it satisfies the necessary condition). Determining whether it is actually an extremum is a different matter. For a function f of one variable a point x_0 is stationary if $f'(x_0) = 0$. This means that f is locally flat. To make this quantitative we fix ϵ and use Taylor series

$$f(x_0 + \epsilon) = f(x_0) + f'(x_0)\epsilon + \frac{1}{2}f''(x_0)\epsilon^2 + O(\epsilon^3), \tag{6.2}$$

as $\epsilon \to 0$. Since x_0 is critical the derivative vanishes and we conclude that for sufficiently small ϵ the values of f near x_0 are either smaller (for $f''(x_0) < 0$) or larger (for $f''(x_0) > 0$) than $f(x_0)$ (provided that $f''(x_0) \neq 0$). We thus have either a local maximum or minimum.

We would like to generalize the above elementary result for a functional. Assume that we are considering functions with fixed endpoints

$$y(a) = y_1, \quad y(b) = y_2. \tag{6.3}$$

Fix ϵ and suppose that $y_c(x)$ is a stationary point of the functional J. We want to add an arbitrary increment $\epsilon h(x)$ to that function. However, the only admissible increments are those which do not alter the boundary conditions. That is, we require

$$h(a) = h(b) = 0. \tag{6.4}$$

Then, the functional has the form

$$J(y_c + \epsilon h) = \int_a^b L(x, y_c(x) + \epsilon h(x), y_c'(x) + \epsilon h'(x))dx. \tag{6.5}$$

Next, we have to expand the Lagrangian

$$\begin{aligned} &L(x, y_c(x) + \epsilon h(x), y_c'(x) + \epsilon h'(x)) \\ &\qquad = L(x, y_c(x), y_c'(x)) + L_y(x, y_c(x), y_c'(x))\epsilon h(x) + L_{y'}(x, y_c(x), y_c'(x))\epsilon h'(x) + O(\epsilon^2), \end{aligned} \tag{6.6}$$

where subscripts denote partial derivatives. When we plug the above into the integral we obtain

$$J(y_c+\epsilon h) = J(y_c)+\epsilon \left[\int_a^b L_y(x, y_c(x), y_c'(x))h(x)dx + \int_a^b L_{y'}(x, y_c(x), y_c'(x))h'(x)dx\right]+O(\epsilon^2). \tag{6.7}$$

Now, the second integral contains a derivative of the perturbation and in order to get rid of if we integrate by parts obtaining

$$\int_a^b L_{y'}(x, y_c(x), y_c'(x))h'(x)dx = [L_{y'}(x, y_c(x), y_c'(x))h(x)]_a^b - \int_a^b \frac{d}{dx}L_{y'}(x, y_c(x), y_c'(x))h(x)dx. \tag{6.8}$$

Due to our boundary conditions (6.4) the term without integral vanishes and we are left with

$$J(y_c+\epsilon h) = J(y_c)+\epsilon \left[\int_a^b \left(L_y(x, y_c(x), y_c'(x)) - \frac{d}{dx}L_{y'}(x, y_c(x), y_c'(x))\right) h(x)dx\right]+O(\epsilon^2). \tag{6.9}$$

Analogously to the one dimensional case we would like to *define* the stationary point of J to be such that the ϵ term above vanishes for all admissible perturbations h. That is,

$$\int_a^b \left(L_y(x, y_c(x), y_c'(x)) - \frac{d}{dx} L_{y'}(x, y_c(x), y_c'(x)) \right) h(x)\,dx \tag{6.10}$$

for all continuous $h = h(x)$ such that (6.4) is satisfied. Since h is arbitrary we conclude that the integrand has to vanish[18] and this happens precisely when

$$\frac{d}{dx} L_{y'}(x, y_c(x), y_c'(x)) = L_y(x, y_c(x), y_c'(x)). \tag{6.11}$$

This partial differential equation is the celebrated *Euler-Lagrange equation* and is a foundation of calculus of variations. We have thus proved the following theorem.

Theorem 5 (Euler-Lagrange)**.** *A necessary condition for $y_c(x)$ to be a stationary point of the functional J defined in (6.1) is that y_c satisfies the Euler-Lagrange equation (6.11).*

6.2 Examples

As we mentioned before, in physical situations it is usually satisfactory to know only that y_c is just stationary rather than an essential extremum. We will illustrate the concept with several examples.

Example. (*Shortest curve on a plane*) We start with probably the simplest variational problem: find the shortest curve joining two points on a place. Let us focus on the shortest route between $(0, 0)$ and $(1, 1)$. Of course, we know that the solution will be a straight line. However, let us see how Euler-Lagrange equations predict this.

Our functional is

$$J(y) = \int_0^1 \sqrt{1 + y'(x)^2}\,dx. \tag{6.12}$$

The partial derivatives of the Lagrangian are

$$L_y = 0, \quad L_{y'} = \frac{y'}{\sqrt{1 + y'^2}}. \tag{6.13}$$

Therefore, Euler-Lagrange equations (6.11) become

$$\left(\frac{y'}{\sqrt{1 + y'^2}} \right)' = 0, \tag{6.14}$$

that is

$$\frac{y'}{\sqrt{1 + y'^2}} = C, \tag{6.15}$$

for constant C. By squaring the above can be solved for y', that is

$$y'^2 = \frac{C^2}{1 - C^2} = a^2, \tag{6.16}$$

[18] This is known as du Bois-Reymond's lemma.

where we defined another constant of integration. Henceforth, the derivative is constant which means that y is a straight line

$$y(x) = ax + b. \tag{6.17}$$

Plugging in boundary conditions yields $b = 0$ and $a = 1$. Our line is then anticipated $y(x) = x$. □

Example. (*Catenary and soap film*) A more elaborate example is finding the surface of revolution that has the smallest area. The generator of the surface is a curve joining two points: (a, c) and (b, d). Physically, this is equivalent to finding a shape of a soap film spanned between two circular rings: the surface tension pulls the soap in order to minimize the area.

For concreteness assume that the generator of evolution joins $(-1, 1)$ with $(1, 1)$. The area of a surface of revolution is then

$$J(y) = 2\pi \int_0^1 y\sqrt{1 + y'(x)^2}dx. \tag{6.18}$$

Calculating partial derivatives of the Lagrangian gives

$$L_y = 2\pi\sqrt{1 + y'^2}, \quad L_{y'} = 2\pi\frac{yy'}{\sqrt{1 + y'^2}}. \tag{6.19}$$

And, hence, the Euler-Lagrange equations become

$$\left(\frac{yy'}{\sqrt{1 + y'^2}}\right)' = \sqrt{1 + y'^2}, \tag{6.20}$$

which might look a little formidable. Fortunately, this is a good place to make a very useful general remark. In this case the Lagrangian does not depend on x and, in general, we can write

$$L_y = (L_{y'})' = L_{yy'}y' + L_{y'y'}y''. \tag{6.21}$$

If we multiply the above by y' we have

$$0 = y'L_y - L_{yy'}y'^2 - L_{y'y'}y'y'' = (L - y'L_{y'})'. \tag{6.22}$$

Therefore, the first integral of the above Euler-Lagrange equation is

$$L - y'L_{y'} = C, \tag{6.23}$$

for some constant C. This is known as *Beltrami's identity*. Returning to our case this is

$$y\sqrt{1 + y'^2} - \frac{yy'^2}{\sqrt{1 + y'^2}} = C. \tag{6.24}$$

With a little manipulation we can get

$$y' = \sqrt{\frac{y^2 - C^2}{C^2}}, \tag{6.25}$$

which is a separated ODE. Integrating it we obtain

$$y(x) = C \cosh \frac{x + D}{C}. \tag{6.26}$$

The shape of our soap film is then given by a revolution of a hyperbolic cosine known as *catenary*. This curve is also the one that dictated the shape of a hanging cable or a chain - hence the name[19]. Our boundary conditions $y(-1) = y(1) = 1$ we obtain $C = 1$ and $D = 0$ which implies

$$y(x) = \tag{6.27}$$

This shape is also used frequently in architecture to construct arches since then no bending moments arise. □

Example. (*Brachistochrone*[20]) This is the classical problem formulated by Jakob Bernoulli. We have to find the curve joining $(0, 0)$ and $(-a, -b)$ on which a particle will slide under the gravity and without friction in the shortest time.

We will use the conservation of mechanical energy which is

$$\frac{1}{2}mv^2 + mgy = 0, \tag{6.28}$$

since initially the potential energy vanishes. Here, $v = v(y)$ is the velocity that can be expressed

$$v(y) = \sqrt{-2gy}, \tag{6.29}$$

where, of course, $y < 0$. Since, from the definition of normal velocity we have $v = ds/dt$ with ds the arc element. Further, $dt = ds/v$ and the total time of descent is

$$T = \frac{1}{\sqrt{2g}} \int_0^{-b} \sqrt{\frac{1 + x'(y)^2}{-y}}\, dx. \tag{6.30}$$

Note that we rather would like to express the sought curve as $x = x(y)$ rather than usual $y = y(x)$. This makes the computations easier. The Lagrangian $L(y, x'(y))$ does not depend on x and hence $L_x = 0$ which gives Euler-Lagrange equation in the form

$$\frac{x'^2}{y(1 + x'^2)} = -\frac{1}{2a}, \tag{6.31}$$

where the form of the constant of integration has been chosen for convenience. Now, we can transform the above

$$\frac{dx}{dy} = \sqrt{\frac{-y}{2a - y}}, \tag{6.32}$$

which immediately can be separated to yield

$$x = \int \sqrt{\frac{-y}{2a - y}}\, dy. \tag{6.33}$$

[19] In Latin *catena* means chain.

[20] From Ancient Greek - "the shortest path".

There are several ways of calculating that integral and, probably, the most educative is to use a rather remarkable substitution

$$y(\theta) = -a(1 - \cos\theta), \tag{6.34}$$

for a new variable θ. Then,

$$x(\theta) = a(\theta - \sin\theta), \tag{6.35}$$

where the integration constant is zero since initially we have to have $x = y = 0$. The unknown constant a has to be chosen in order for the curve to pass through $(-a, -b)$. The parametric formulas we have found define the so-called *cycloid* which is a curve drawn by a valve on a moving bicycle wheel.

We can check our findings numerically. Suppose we have four candidates for the least time curve: straight line (the shortest), parabola, circular arc, and the cycloid (see Fig. 27). Numerically calculating the time of descent (6.30) we arrive at (setting $\sqrt{2g} = 1$)

$$T_{cycl} = 2.582, \quad T_{par} = 2.587, \quad T_{circ} = 2.622, \quad T_{line} = 2.828. \tag{6.36}$$

We see that shortest distance does not mean the least time! This is because due to larger initial slope, the particle can develop larger speed. We can also see that the ime of sliding on a parabola is very close to the least time corresponding to the cycloid. □

Example. (*Shortest curve on a sphere*) A shortest curve that connects two points on a general surface is called a *geodesic*. It is beyond the scope of this lecture to pursue the topic in general, however, we will see what are the shortest route curves on a sphere.

From multidimensional calculus we know that the length of a curve laying on a sphere connecting points 1 and 2 is given by

$$R\int_{\phi_1}^{\phi_2} \sqrt{\left(\frac{d\theta}{d\phi}\right)^2 + \sin^2\theta}\, d\phi, \tag{6.37}$$

where R is the radius, $\theta \in [0, \pi]$ is the latitude, and $\phi \in [0, 2\pi]$ longitude. Our Lagrangian is now $L = L(\theta, \theta')$ and, hence, it does not depend on the independent variable ϕ. We can then use Beltrami's identity to obtain

$$\sqrt{\theta'^2 + \sin^2\theta} - \frac{\theta'^2}{\sqrt{\theta'^2 + \sin^2\theta}} = C, \tag{6.38}$$

for some constant C. After multiplication by the square root we have

$$\sin^2\theta = C\sqrt{\theta'^2 + \sin^2\theta}, \tag{6.39}$$

which can be solved for $d\theta/d\phi$ in a form

$$\frac{d\theta}{d\phi} = \frac{1}{C}\sin^2\theta\sqrt{1 - \frac{C^2}{\sin^2\theta}}. \tag{6.40}$$

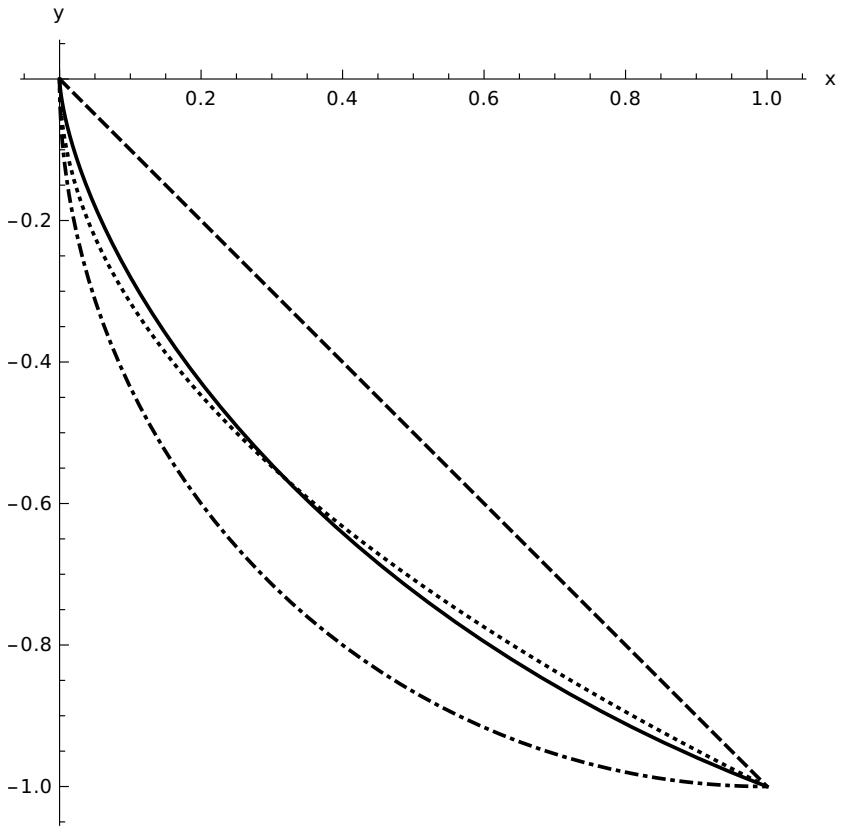

Figure 27: Various curves of descent: cycloid (solid), line (dashed), parabola (dotted), and circular arc (dot-dashed).

This is a separated ODE with an integral

$$\phi = \arcsin\left(\frac{\cot\theta}{\beta}\right) + \alpha, \tag{6.41}$$

which suitably chosen constants of integration α and β. To interpret the above family of curves we rewrite the equation as

$$\cot\theta = \beta\sin(\phi - \alpha), \tag{6.42}$$

and after multiplication by $R\sin\phi$ we obtain

$$A(R\sin\theta\sin\phi) - B(R\sin\theta\cos\phi) = (R\cos\theta), \tag{6.43}$$

for $A = \beta\cos\alpha$ and $B = \beta\sin\alpha$. The terms in the parentheses above are nothing else as relations between Cartesian and spherical coordinates. That is,

$$Ay - Bx = z, \tag{6.44}$$

which is a plane passing through the centre of the sphere! The curves of shortest path, geodesics, are then great circles - intersections of the sphere with a plane passing through the centre. Between many uses, they serve as routes in navigation. □

Example. (*Fermat's least time principle*) We known from high school physics that in a media with constant refraction coefficient light rays are straight. If the medium changes the light will bend according to Snell's law. However, if the coefficient changes continuously the light beam will not be straight any more.

Fermat's discovered that light rays always follow a path for which the time of the travel is the least of all possible routes. Consider a medium with a varying refraction coefficient $n(x, y)$. If the speed of light is denoted by c then the light ray's velocity os given by $v = c/n$. On the other hand, velocity is just $v = ds/dt$ which states that in a short time dt light goes a distance ds representing an increment of a trajectory $y = y(x)$. Then, the total time that is required to travel between two points is

$$T = \int \frac{ds}{v} = \frac{1}{c}\int_a^b n(x, y(x))\sqrt{1 + y'(x)^2}dx. \tag{6.45}$$

Now, the Euler-Lagrange equations give the stationary solution satisfying

$$n_y = \left(\frac{ny'}{\sqrt{1 + y'^2}}\right)_x. \tag{6.46}$$

If, for instance, the refractive index only depends on x (a layered material) then

$$\frac{n(x)y'(x)}{\sqrt{1 + y'(x)^2}} = C, \tag{6.47}$$

for some constant C. From here, we can deduce several important results.

First, suppose that n is piecewise constant, say $n = n_1$ for $x \in [0, 1/2]$ and $n = n_2$ for $x \in (1/2, 1]$. Let θ be the angle subtended between the light ray and the horizontal direction. Since $\sin\theta = \pm y'/\sqrt{1+y'^2}$ we infer from the above formula that the angle is constant in each interval $[0, 1/2]$ and $(1/2, 1]$. Therefore, light travels in straight lines. Moreover,

$$n_1 \sin\theta_1 = n_2 \sin\theta_2, \tag{6.48}$$

which is the Snell's law.

We can actually integrate the resulting ODE for a general refraction index $n = n(x)$. That is,

$$y(x) = \int \frac{C}{\sqrt{n(x)^2 - C^2}} dx, \tag{6.49}$$

which gives the formula for the trajectory (the sign has been incorporated into C). □

Example. (*Lagrangian formulation of mechanics*) One of the most profound and widespread applications of calculus of variations is devising equations modelling various mechanical situations. Many times this is much simpler and straightforward than using Newton's equation.

There is a very general physical law known as *Hamilton's principle of least action* which states that the observable trajectory of a physical system is a stationary point of the following *action* functional

$$S = \int (T - V)\, dx, \tag{6.50}$$

where T is the kinetic energy of the system, while V is the potential. That is, in order to find the equations of motion for any physical system with potential and kinetic energy we have to solve Euler-Lagrange equations corresponding to the above. This can be generalized to higher dimensions and many degrees of freedom and the Lagrangian formulation is independent of the chosen variables. This makes variational framework much more pleasant to work with than Newtonian dynamics.

As an example let us choose a point particle with mass m subject to a one dimensional gravitational field. Then, $T = mx'^2/2$ and $V = mgx$ with x being the height of the particle. Then, the Lagrangian has the form

$$L(t, x, x') = \frac{1}{2} m x'(t)^2 - mgx. \tag{6.51}$$

Therefore, the Euler-Lagrange equations are

$$(mx')' = -mg, \tag{6.52}$$

which is $x'' = -g$. This is the same as Newton's equation. Lagrangian formulation is very frequently used in mechanics of systems with many degrees of freedom, constraints, and curvilinear coordinates. It is much simpler then to obtain a set of meaningful equations. □

6.3 Optimization with constraints

Frequently there is a need of finding an optimal solution under some constraints: limited funds for a project, a given length of a hanging cable or finite time of the process. On multidimensional calculus we have learned about the method of *Lagrange multipliers* to find extrema of scalar functions under some conditions. This method can also be used in variational calculus. There are three important constraints that can be imposed on the optimal solution.

1. Isoperimetric - an integral of the optimal solution is given.
2. Holonomic - the optimal solution has to satisfy an algebraic (geometric) equation.
3. Optimal control - the optimal solution has to be a solution of differential equation.

We will not go into details of constrained optimization problems, however, we will illustrate the concept on the isoperimetric problem. For starters, suppose that we want to find a critical point of $f = f(x, y)$ subject to $g(x, y) = 0$. Suppose we can parameterise the constraint $g = 0$ by $\mathbf{x}(t) = (x(t), y(t))$, then after differentiation we have

$$0 = \frac{d}{dt} g(x(t), y(t)) = \nabla g \cdot \frac{d\mathbf{x}}{dt}, \tag{6.53}$$

and we recover the well-known fact that gradient is orthogonal to level curves. Next, plug the constraint vector into f and define $F(t) = f(x(t), y(t))$. The stationary point of $F(t)$ will be precisely the critical point of f under the condition $g = 0$ since $\mathbf{x}(t)$ holds us always on the curve defined by g. We thus have

$$0 = F'(t) = \nabla f \cdot \frac{d\mathbf{x}}{dt}. \tag{6.54}$$

And we see that the gradient of f is orthogonal to level curves of g. Therefore, there must exist a constant λ, known as Langrange multiplier, such that

$$\nabla f = \lambda \nabla g. \tag{6.55}$$

Therefore, the conditional extremum of f is reduced to finding a local extremum of

$$L(x, y, \lambda) = f(x, y) - \lambda g(x, y). \tag{6.56}$$

Now, suppose we want to find the critical point of the functional

$$S(y) = \int_a^b L(x, y(x), y'(x)) dx, \tag{6.57}$$

under the integral condition

$$G(y) = \int_a^b g(x, y(x), y'(x)) dx = 0. \tag{6.58}$$

Then, analogously to the previous simple example, we can form an *augmented Lagrangian*

$$\mathcal{L}(x, y(x), y'(x), \lambda) = L(x, y(x), y'(x)) - \lambda g(x, y(x), y'(x)), \tag{6.59}$$

and look for stationary points of

$$J(y, \lambda) = \int_a^b \mathcal{L}(x, y(x), y'(x), \lambda) dx. \tag{6.60}$$

This analogy is correct and we will prove it.

Theorem 6. *A stationary point of (6.57) under the isoperimetric condition (6.58) is a solution of Euler-Lagrange equations for the augmented Lagrangian, that is*

$$\frac{\partial \mathcal{L}}{\partial y} = \frac{d}{dx}\frac{\partial \mathcal{L}}{\partial y'}. \tag{6.61}$$

Proof. We will follow the same method as in finding Euler-Lagrange equations for the unconstrained case. That is, we perturb all unknowns

$$y = y_c + \epsilon h(x), \quad \lambda = \lambda_c + \epsilon \kappa, \tag{6.62}$$

where y_c and λ_c are stationary points of our functional. Expanding $J(y_c + \epsilon h(x), \lambda_c + \epsilon\kappa)$ with respect to ϵ exactly as was done in (6.5), conducting integration by parts, and requiring that the $O(\epsilon)$ term vanishes leads us to

$$\int_a^b \left(\frac{\partial L}{\partial y} - \frac{d}{dx}\frac{\partial L}{\partial y'} - \lambda_c \left(\frac{\partial g}{\partial y} - \frac{d}{dx}\frac{\partial g}{\partial y'} \right) \right) h(x) dx + \kappa \int_a^b g(x, y(x), y'(x)) dx = 0. \tag{6.63}$$

Since the above must hold for arbitrary perturbations κ we obtain

$$\int_a^b g(x, y(x), y'(x)) dx = 0, \tag{6.64}$$

which is the isoperimetric constraint. Similarly, the perturbation to the function, that is h, is also arbitrary and, hence,

$$\frac{\partial L}{\partial y} - \frac{d}{dx}\frac{\partial L}{\partial y'} - \lambda_c \left(\frac{\partial g}{\partial y} - \frac{d}{dx}\frac{\partial g}{\partial y'} \right) = 0, \tag{6.65}$$

which is the same as (6.61). □

We will illustrate the above result on a classical problem in calculus of variations.

Example. (*Classical isoperimetric problem*) From all closed curves joining $(0, 0)$ and $(1, 0)$ with a given length P find the one that maximizes the area underneath it. The augmented Lagrangian has the form

$$\mathcal{L}(x, y, y', \lambda) = y(x) + \lambda \left(\sqrt{1 + y'(x)^2} - P + 1 \right), \tag{6.66}$$

where we have accounted for the bottom part of the curve with unit length. Applying (6.61) we arrive at

$$\lambda \left(\frac{y'}{\sqrt{1+y'^2}} \right)' = 1, \tag{6.67}$$

which can be easily solved subject to $y(0) = y(1) = 0$ to yield

$$y(x) = \sqrt{\lambda^2 - \left(x - \frac{1}{2}\right)^2} - \sqrt{\lambda^2 - \frac{1}{4}}. \tag{6.68}$$

Hence, we have obtained a family of circles. Note that for the above to make sense we have to assume that $\lambda \geq 1/2$. We can integrate the found solution in order to prescribe the perimeter

$$P = \int_0^1 \sqrt{1 + y'(x)^2}\,dx = 2\lambda \arcsin\left(\frac{1}{2\lambda}\right) \quad \lambda \geq \frac{1}{2}. \tag{6.69}$$

Solving for λ in terms of P gives the final solution. However, since

$$\frac{1}{2\lambda} \leq \arcsin\frac{1}{2\lambda} \leq \frac{\pi}{2}, \tag{6.70}$$

we have that the possible perimeters are $1 \leq P \leq \pi/2$. For example, when $P = \pi/2$ we have $\lambda = 1$. Note that by considering a parametrised curve $(x(t), y(t))$ we would obtain a greater class of possible curves (no restriction on λ and P). Nevertheless, we have proved that a circle encloses the largest area with a given perimeter. □