MATH 5525 LECTURE LOG

Lecture 1, 1/23

The exponential function

Covered some material from 1.1. (Here 1.1 refers to Chapter 1, Section 1 in the textbook.[1] Similar convention will be used in what follows.)

Main takeaway from the lecture:

(i) the function $x(t) = e^{at}$ satisfies the differential equation

$$\frac{dx}{dt} = ax, \tag{1}$$

and this includes the case when a is complex (in which case we consider x as a complex-valued function of the real variable t). For any given number C (real or complex), the function $x(t) = Ce^{at}$ also satisfies (1).

(ii) equation (1), in spite of its simplicity, describes some important phenomena (we discussed a couple of examples).

Optional exercise:[2] show that $\lim_{n\to\infty}\left(1+\frac{t}{n}\right)^n = e^t$.

Lecture 2, 2/25

Equations $\frac{dx}{dt} = ax + b$ and $\frac{dx}{dt} = ax + b(t)$

Covered some material from 1.2. We first looked in some detail at the equation

$$\frac{dx}{dt} = ax + b \tag{2}$$

where a, b are constants,[3] and saw that the solution with the initial condition $x(0) = x_0$ is given by

$$x(t) = x_0 e^{at} + \frac{b}{a}\left(e^{at} - 1\right)\ . \tag{3}$$

This is the same as formula (2.7) in the textbook, if we change a to $-a$.

Example

Equation (2) arises in the following situation. Assume we borrow from a bank M_0 dollars at interest rate[4] r and make payments p (per year). Let us make the

[1] The section starts on page 3, ends on page 15.

[2] Additional exercises are in the book. Some of the problems in the book may be parts of future homework assignments. It is not a formal requirement that the students go through the exercises other than those in the official homework assignments, but it is of course good to check how difficult the exercises in the book or optional exercises mentioned in class seem to be.

[3] Note that in the textbook this equation is written in the form $\frac{dx}{dt} + ax = b$, so that our a should be identified with $-a$ in the textbook.

[4] The definition of the interest rate we use here may not be the same as the one used by the bank. First, instead of saying that the interest is, say, 4%, we say that it is 0.04, so in this case we would take $r = 0.04$. Even when taking this way the interest is expressed into account, the bank's numbers might still be slightly different, corresponding perhaps to $e^r - 1$, but the exact definitions by the bank may be more complicated still.

simplifying assumptions that the bank compounds the interest continuously and that the payments are also made continuously. Let $M = M(t)$ be the amount we owe at time t. The equation for M then is

$$\frac{dM}{dt} = rM - p\,, \qquad M(0) = M_0\,, \tag{4}$$

where the time t is measured in years. Using formula (3), we can solve the following problem: suppose we wish to borrow the amount M_0 at interest rate r and we wish to pay the loan off in T years. What will be our payment? To calculate p, we apply (3) with $a = r$ and $b = -p$ to get

$$M(t) = M(t, r, p, M_0) = M_0 e^{rt} - \frac{p}{r}\big(e^{rt} - 1\big) \tag{5}$$

and then solve the equation

$$M(T, r, p, M_0) = 0 \tag{6}$$

for p. We obtain

$$p = \frac{M_0}{T}\;\frac{rT}{1 - e^{-rT}}\,. \tag{7}$$

Note that $\frac{M_0}{T}$ is exactly what the payments would be if there was no interest. The factor $\frac{rT}{1-e^{-rT}}$ expresses the influence of the interest. Our total payments to the bank will be

$$\text{total payments} = p\,T = M_0\;\frac{rT}{1 - e^{-rT}}\;. \tag{8}$$

Let

$$f(\xi) = \frac{\xi}{1 - e^{-\xi}}\,. \tag{9}$$

It is a good exercise to investigate the function f in some detail. The formula might look singular at $\xi = 0$, but the singularity is not "real", f is really a smooth[5] function of ξ, with

$$\lim_{\xi \to 0} f(\xi) = 1\,. \tag{10}$$

As an optional exercise, you can calculate the derivative $f'(0)$.[6] Considering the payment p as a function of T, i. e. $p = p(T)$, note that $\lim_{T\to\infty} p = M_0 r$. This corresponds to the situation that we only pay the interest and the loan is never paid off. In that case the solution of (4) is constant, $M(t) = M_0$.

We next discussed how to solve (2) when $b = b(t)$ is not constant. We obtained

$$x(t) = x_0 e^{at} + \int_0^t e^{a(t-s)} b(s)\,ds\,. \tag{11}$$

[5] and, in fact, analytic, i. e. given by a power series which converges for all ξ

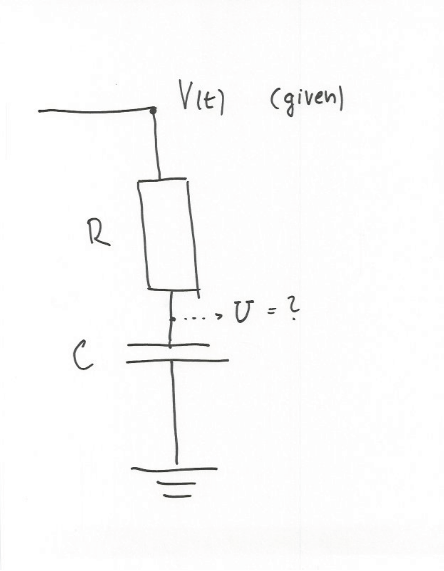
[6] The result is $f'(0) = \frac{1}{2}$. This corresponds to the fact that when rT is small, the bank will make approximately $\frac{1}{2}M_0 rT$ on the loan, about half of what they would make for small rT if they loaned the whole sum M_0 for time T, and the loan would be paid off in one payment after that time.

This is a special case of formula (2.14) in the textbook (again note that we work with different signs), which we will discuss next time.

Lecture 3, 1/28

Interpreting the formulae, equations $\frac{dx}{dt} = a(t) + b(t)$, $\frac{dx}{dt} = ax(1-x)$, separable equations

An important part of understanding solutions of differential question is the interpretation of the formulae which we obtain. Let us look at formula (11) from last lecture in the context of the following electric circuit. The discussion of this example is optional.



The voltage $U = U(t)$ at the point indicated in the picture satifies

$$C\,dU = \frac{V(t) - U}{R}\,dt\ , \tag{12}$$

which is the same as

$$\frac{dU}{dt} = -\frac{U}{RC} + \frac{V(t)}{RC}\ . \tag{13}$$

Letting $U_0 = U(0)$ and using (11), we have

$$U(t) = U_0\ e^{-\frac{t}{RC}} + \int_0^t \frac{1}{RC}\ e^{\frac{(t-s)}{RC}}\ V(s)\,ds\ . \tag{14}$$

From the electric circuit interpretation we expect various properties of the solutions which we would like to confirm from (14). For example, if for a fixed resistance R the capacity C is getting close to 0, then we should have

$$U(t) \sim V(t)\,. \tag{15}$$

More precisely given $0 < \tau < T$ and a fixed $R > 0$, we expect that if $V(t)$ is continuous, then

$$U(t) \to V(t) \quad \text{uniformly for } t \in [\tau, T] \text{ as } C \to 0_+. \tag{16}$$

As an optional exercise you can try to prove it. The key point is to consider the properties of the function

$$\phi_\varepsilon(t) = \frac{1}{\varepsilon}\, e^{-\frac{t}{\varepsilon}}\,. \tag{17}$$

Note that

$$\int_0^\infty \phi_\varepsilon(t)\,dt = 1 \tag{18}$$

and, moreover, for any $\tau > 0$ we have that

$$\int_\tau^\infty \phi_\varepsilon(t)\,dt \to 0 \qquad \text{as } \varepsilon \to 0_+\,. \tag{19}$$

Therefore for small C the main contribution to the right-hand side of (14) will be coming from

$$\int_{t-\tau}^t \frac{1}{RC}\, e^{\frac{(t-s)}{RC}}\, V(s)\,ds \tag{20}$$

for some small $\tau > 0$. As V is continuous, for small $\tau > 0$ it is nearly constant in $(t-\tau, t)$, deviating from $V(t)$ only a little. In view of (18) and (19) we see that (16) should follow. This concludes the discussion of our optional example.

Another optional exercise is the following: verify directly that $x(t)$ given by (11) satisfies (2), or, alternatively, that $U(t)$ given by (14) satisfies (13).[7]

Next we discussed the equation

$$\frac{dx}{dt} = a(t)x + b(t) \tag{22}$$

and derived the formula

$$x(t) = x(t_0)\, e^{(A(t)-A(t_0))} + \int_{t_0}^t e^{(A(t)-A(s))}\, b(s)\,ds\ , \tag{23}$$

[7] The following formula for taking derivatives of an integral is useful for this calculation (and other similar calculations):

$$\frac{d}{dt}\int_0^t f(t,s)\,ds = f(t,t) + \int_0^t \frac{\partial f}{\partial t}(t,s)\,ds\ . \tag{21}$$

In the example above one can in fact avoid using it by pulling the term $e^{\frac{t}{RC}}$ in front of the integral, but it is still useful to know the formula.

where the function $A(t)$ is the primitive of $a(t)$, i. e. $A = \int a$ or, equivalently, $A' = a$, which is the same as formula (2.14) in the textbook.
The formula was derived in two steps: first, we showed how to solve

$$\frac{dx}{dt} = a(t)x \tag{24}$$

by writing the equation as

$$\frac{dx}{x} = a(t)\,dt \qquad \textit{(“separation of variables”)} \tag{25}$$

and integrating both sides, see also section 1.3 in the textbook. Once we know how to solve (24), we seek the solution of (22) as

$$x(t) = C(t)e^{A(t)} \qquad \textit{(“variation of constants”)}\,. \tag{26}$$

Substituting (26) into (22) we obtain

$$C'(t) = e^{-A(t)}b(t) \tag{27}$$

and integration in t now gives (23).

A form of this classical calculation is in the textbook on page 16.

Separation of variables for more general equations

We next discussed the method of separation of variable for more general equations, see section 1.3 in the textbook. As an example, we considered the equation

$$\frac{dx}{dt} = ax(1-x)\,. \tag{28}$$

One interpretation of the equation in terms of “spread of a rumor”. We think of $x = x(t)$ as denoting the fraction of the population who know a rumor. As people meet and share the rumor, equation (28) seems to be a reasonable model for how the knowledge of the rumor evolves. In this interpretation one should have $0 \le x \le 1$, but one can in fact solve the equation also when x takes values outside of this interval. To solve the equation, we will write it as

$$\frac{dx}{x(1-x)} = a\,dt\,. \tag{29}$$

Taking $t_t < t_2$ and letting $x_1 = x(t_1), x_2 = x(t_2)$, we can write

$$\int_{x_1}^{x_2} \frac{dx}{x(1-x)} = \int_{t_1}^{t_2} a\,dt = a(t_2 - t_1)\,. \tag{30}$$

We have

$$\int_{x_1}^{x_2} \frac{dx}{x(1-x)} = \int_{x_1}^{x_2} \left[\frac{1}{x} + \frac{1}{1-x}\right] dx = \log x - \log(1-x) \quad |_{x=x_1}^{x=x_2}\,. \tag{31}$$

Letting $t_1 = 0\,,\ x_1 = x(0) = x_0$ and $t_2 = t\,,\ x_2(t_2) = x(t)$, we obtain after a simple calculation

$$x(t) = \frac{x_0}{x_0(1 - e^{-at}) + e^{-at}}\,. \tag{32}$$

We will look at these solution in more detail next time.

Lecture 4, 1/30

Geometric picture, phase-portraits of 1d equations

Given a differential equation, such as (28), we can try to calculate the general solution (formula (32) in the case of (28)), and then try to understand the properties of the solution by looking at the formula. In practice this approach works only in a limited number of cases; for many situations it is simply not possible to write the general solution in a closed form using elementary functions. However, we can often get a good idea about qualitative properties of solutions of a differential equation just by looking at some simple geometric pictures, without having to perform difficult calculations. This material is discussed in the textbook in the context of more difficult problems in chapters 12 and 13 (you can have a look at the nice pictures there). We will discuss some of these issues in a simple form even at this stage.

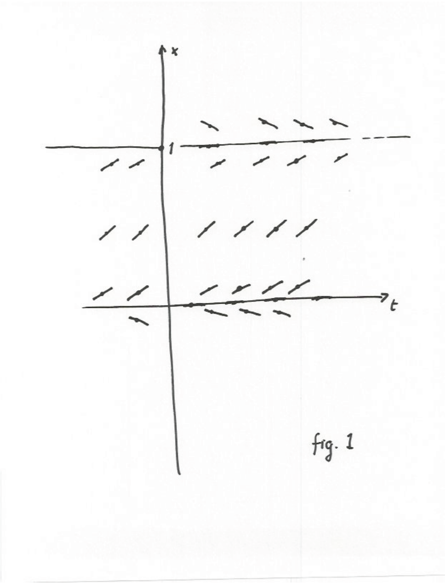
Geometric pictures of (28)

(A) The (t, x)-plane picture

The most straightforward way, which is already useful, is to think about (28) in terms of the (t, x) Cartesian plane and the graphs of the solutions. The equation tells us that if the solution passes through a point (t, x), the slope k of its tangent at this point is given by

$$k = k(t, x) = ax(1 - x)\,. \tag{33}$$

We can imagine drawing a segment with the corresponding slope at each point (t, x) in the plane, so that we get a picture like this:



Our task in solving the equation is to find curves in the (t, x) plane which at each point follow the direction given by the segment at that point. You can note various properties of these segments from formula (33). For example

- the slope at (t, x) is independent of t,
- the slope at $(t, 0)$ is zero,
- the slope at $(t, 1)$ is zero,
- for $0 < x < 1$ the slope at (t, x) is strictly positive,
- for $1 < x$ the slope at (t, x) is strictly negative,
- for $x < 0$ the slope at (t, x) is strictly negative.

Looking at the picture and keeping these properties in mind, we see that

- the functions $x(t) = 0$ and $x(t) = 1$ are solutions,
- the solution $x(t)$ passing through any (t_0, x_0) with $0 < x_0 < 1$ is increasing with t with $\lim_{t\to\infty} x(t) = 1$ and $\lim_{t\to-\infty} x(t) = 0$; this solution can never intersect the constant solutions $x \equiv 0$ or $x \equiv 1$ (or any other solution distinct from itself, for that matter).

With similar observations it is not hard to see without really doing any calculations, that the solutions look something like this:

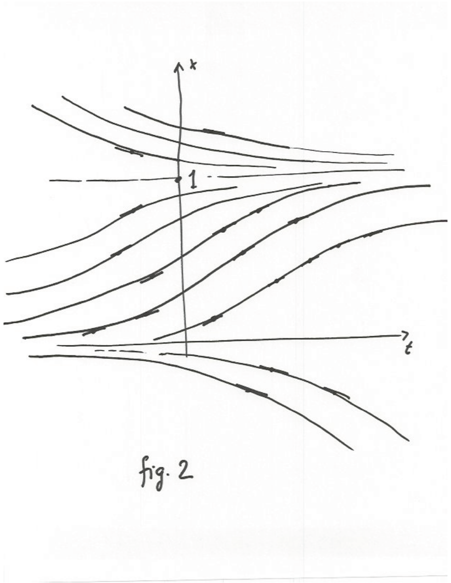


fig. 2

The only “qualitative” question which is not clear without calculation is what the solutions passing through points (t, x) with $x < 0$ do as t increases. It is clear they decrease, but how fast? Looking at (32), you can convince yourself that such solution reach $-\infty$ in finite time. In other words, for each such solution there is T such $x(t)$ is defined only on $(-\infty, T)$ and $\lim_{t\to T} x(t) = -\infty$.

(B) The 1-dimensional “phase portrait”

While the geometric picture above is natural and helpful, there is an even simpler picture which makes the properties of the solutions arguably even easier to see. The key for the viability of this picture is that the expression (33) is independent of t. We can think about (28) in the following way: it is a rule which tells us at which speed we should move along the x axis when we are at a point x. The key point is that the rule does not depend on the time when we are at x, we always move with the same speed $ax(1-x)$, independently of what the time is. Therefore we can represent the situation with the following 1d picture:

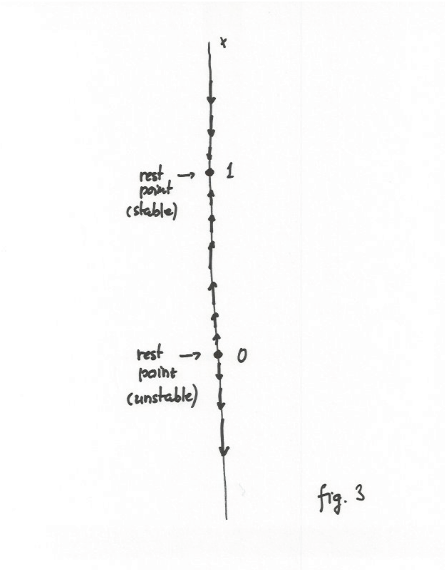


fig. 3

We see that there are two points, $x = 0$ and $x = 1$ where the prescribed velocity vanishes. When we are at these points, we do not move. These points divide the real line R into 3 segments,

$$I_1 = (-\infty, 0), \qquad I_2 = (0, 1), \qquad I_3 = (1, \infty) \ . \tag{34}$$

If we are in I_1, we move towards $-\infty$; if we are in I_2 we move towards 1; if we are in I_3, we also move towards 1 (this time from the other side).

We see that this 1d picture gives us actually a very good idea what the solutions do. As an exercise you can perform the same analysis for the equation

$$\frac{dx}{dt} = \sin x \,, \tag{35}$$

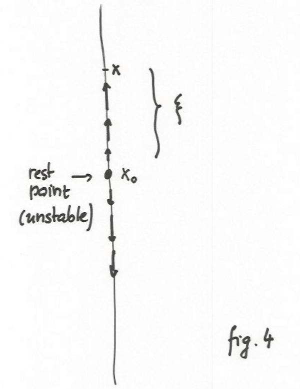
as we did in class. The method works well for the equations of the form

$$\frac{dx}{dt} = f(x) \,. \tag{36}$$

We see the prominent role of the point x with $f(x) = 0$. The solutions starting at those points are trivial, but they tend to separate regions with different behavior, and therefore are important. (The situation in higher dimensions is more complicated, but the "rest points" of the equations still play a very important role.)

Linearization near the rest points

Let us consider (36) with a smooth f and assume that $x_0 \in R$ is such that $f(x_0) = 0$ and $f'(x_0) = a \neq 0$. Let us introduce a new coordinate ξ by $x = x_0 + \xi$. If we look at the phase portrait of (36) near x_0 in the coordinate ξ, we see something like this:



This is very similar to the phase portrait of the equation

$$\frac{d\xi}{dt} = a\xi \tag{37}$$

we studied in Lecture 1, and we suspect that the solutions which are close to x_0 will mostly behave as solution of (37), which are

$$\xi(t) = Ce^{at}\,. \tag{38}$$

We emphasize again that this is expected to be valid only when Ce^{at} is small. A simple calculation supports this heuristics: in the coordinate ξ we can write

$$\frac{d\xi}{dt} = f(x_0 + \xi) = f(x_0) + f'(x_0)\xi + O(\xi^2) = a\xi + O(\xi^2)\ , \tag{39}$$

as $f(x_0) = 0$ and we denoted $f'(x_0) = a$. So in the approximation up to error of order ξ^2 (which is of course much smaller than ξ when ξ is small) we have

$$\frac{d\xi}{dt} = a\xi\,. \tag{40}$$

We see that for equation (36), in the case when f is smooth and has only isolated zeros x_0 with $f'(x_0) \neq 0$ the qualitative behavior of the solutions is simple:

- In any closed interval $J = [a, b]$ which does not contain a rest point, the solution moves from one end to another at some non-zero speed, exceeding a certain minimal value.

- Near a rest point x_0, the solution is either exponentially attracted (in the positive time direction) to it (when $f'(x_0) < 0$) or exponentially repelled (in the positive time direction) from it (when $f'(x_0) > 0$).

We see that the qualitative behavior of the solution of (36) is quite transparent (at least under the assumptions above), even though we may not be able to describe the solution by some elementary formulae in the general case.

As an optional exercise, you can check by direct inspection of the behavior of $x(t)$ given by (23) near a rest point (both $x = 0$ or $x = 1$) that the solution behaves exactly as expected from the considerations above.

Lecture 5, 2/1

Systems of Classical Mechanics with 1 degree of freedom

In this lecture we covered material from 5.1, except for the Projectile problem (page 29). The situation considered here concerns systems of classical mechanics with 1 degree of freedom. There are several levels of generality at which the problem can be considered. The simplest one is

1. *Particle moving along a straight line under influence of a time-independent force*

Here we think literally about motion of a particle with mass m along a straight line. The position of the particle is denoted by x, and we assume that the coordinate x coincides with the length taken along the line (beginning from some fixed point). We assume that the force acting on the particle at point x is $F(x)$. Since we are on a line, we can always write

$$F(x) = -\frac{\partial V}{\partial x}(x)\,. \tag{41}$$

Since our coordinate x is only one-dimensional, we could also write

$$F(x) = -\frac{dV}{dx}(x) = -V'(x)\,. \tag{42}$$

However, we will use notation (41), which is customarily used in this situation. Fixing some point x_0 on the line, we can write

$$V(x) = -\int_{x_0}^{x} F(\tilde{x})\,d\tilde{x}\ , \tag{43}$$

and we see that $V(x)$ is minus the work done by the force when the particle is moved from x_0 to x, which is the same as the work we have to do while moving the particle between x_0 and x. In other words, V can be considered as the potential energy of the particle. The equation of motion (Newton's 1687 *Principia*) is

$$m\frac{d^2x}{dt^2} = F(x)\,. \tag{44}$$

Newton's original notation $\ddot{x}$ is also often used instead of $\frac{d^2x}{dt^2}$ and $\dot{x}$ for $\frac{dx}{dt}$. The kinetic energy of the particle is $\frac{1}{2}m\dot{x}^2$, the total energy is

$$\frac{1}{2}m\dot{x}^2 + V(x)\,. \tag{45}$$

This quantity is preserved during the motion:

$$\frac{d}{dt}\left(\frac{1}{2}m\dot{x}^2 + V(x)\right) = m\dot{x}\ddot{x} + \frac{\partial V}{\partial x}\dot{x} = (m\ddot{x} - F)\dot{x} = 0\,. \tag{46}$$

We can write

$$\frac{1}{2}m\dot{x}^2 + V(x) = E = \text{const.} \tag{47}$$

and this can be considered as an equation for $\dot{x}$:

$$\dot{x} = \pm\sqrt{\frac{2(E - V(x))}{m}}\,. \tag{48}$$

We will explain momentarily how we choose the signs. This is an equation of the form

$$\frac{dx}{dt} = f(x) \tag{49}$$

and can be solved by "separating the variables", e. i. by writing the equation as

$$\frac{dx}{f(x)} = dt \tag{50}$$

and integrating on both sides. We note that the function f can have zeroes, and we have to be somewhat careful.

There are several different scenarios for the behavior of the solutions. For example, an important situation is as follows.

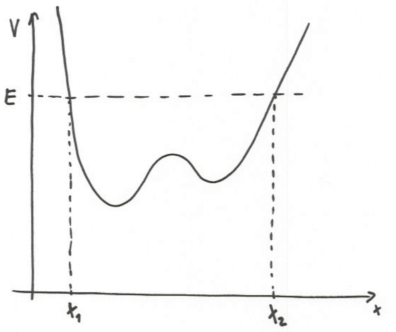


fig 5

In this picture the value of V is plotted against the coordinate x. The total energy of the motion is E, which means that if at some time t the particle is in the interval (x_1, x_2), it can never escape from this interval (if no additional forces act on it), and keeps moving back and forth between x_1 and x_2. At those points $V(x_1) = V(x_2) = E$, and therefore if the particle is at either x_1 or x_2, its velocity $\dot{x}$ vanishes. When the particle moves from x_1 towards x_2, we take the + sign in (48), when it moves from x_2 to x_1, we take the minus sign.

In the open interval (x_1, x_2) equation (48) implies the equation of motion (44), as one can easily see by simply taking the time derivative.

The situation at the "turning points" x_1 and x_2 is more subtle. Note that the function

$$x(t) \equiv x_1 \tag{51}$$

is a solution of (48), while it is not a solution of Newton's equation (44). The "real motion" given by (44) with the initial data $x(0) = x_1$ and $\dot{x}(0) = 0$ will of course start immediately moving towards x_2 and we will have $x(t) > x_1$ for $0 < t < T$, where T is the time when the "real motion will return to x_1. We expect the motion to be periodically oscillating between x_1, x_2, and T will be the period of the oscillation. Therefore (51) represents a "parasitic solution", which is unphysical and should not really be considered. Nevertheless the solution is still interesting, because it demonstrates *non-uniqueness* of the solutions of (48) (taken with the + sign) with the initial value $x(0) = x_1$. The non-uniqueness is possible because the right-hand side of (48) is not smooth at x_1. the situation at x_2 is similar.

The "right solution" starting at some time t_1 at x_1 with $\dot{x}(t_1) = 0$ can be obtained from

$$\int_{x_1}^{x} \frac{d\tilde{x}}{\sqrt{\frac{2(E-V(\tilde{x})}{m}}} = \int_{t_1}^{t} d\tilde{t} = t - t_1 \ , \tag{52}$$

which works until $x(t)$ reaches x_2 for the first time since t_1. From this formula we see that the solution will make the journey from x_1 to x_2 in time

$$\int_{x_1}^{x_2} \frac{dx}{\sqrt{\frac{2(E-V(x))}{m}}} = \int_{x_1}^{x_2} \frac{dx}{\sqrt{\frac{2(V(x_1)-V(x))}{m}}} \ . \tag{53}$$

Therefore the formula for the period T of the motion is

$$T = 2\int_{x_1}^{x_2} \frac{dx}{\sqrt{\frac{2(V(x_1)-V(x))}{m}}} \ . \tag{54}$$

The integrals in these formulae represent various levels of difficulty for various V. Often they may not be expressible in terms of elementary functions. Note that the integrant always approaches $+\infty$ as x approaches the endpoints x_1, x_2, as the particle is slowing down near those points.

An important special case of V is

$$V(x) = \frac{1}{2}\kappa x^2 \, . \tag{55}$$

This potential approximates well generic general potentials in a neighborhood of the point of their minima, see the figure below.

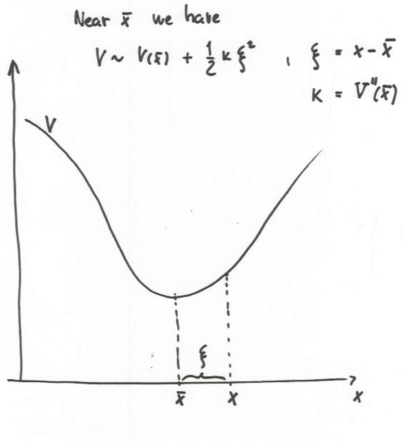

Near $\bar{x}$ we have
$V \sim V(\bar{x}) + \frac{1}{2}\kappa\xi^2$, $\xi = x - \bar{x}$
$\kappa = V''(\bar{x})$
V
ξ
$\bar{x}$
x
x

fig 6

If V attains minimum at $\overline{x}$ and we write $x = \overline{x} + \xi$, then we can write (assuming V is smooth)

$$V(x) = V(\overline{x} + \xi) = V(\overline{x}) + V'(\overline{x})\xi + \frac{1}{2}V''(\overline{x})\xi^2 + O(\xi^3)\,. \tag{56}$$

We know that $V'(\overline{x}) = 0$, as we are at the minimum of V, and we can take $V(\overline{x}) = 0$ without loss of generality. We can then write

$$V(\overline{x} + \xi) = \frac{1}{2}\kappa\xi^2 + O(\xi^3), \qquad \kappa = V''(\overline{x})\,. \tag{57}$$

We see that small oscillations around minima of V should be described quite precisely by potential (55).

The equation of motion given by potential (55) is

$$m\ddot{x} + \kappa x = 0\,. \tag{58}$$

This is a linear equation with constant coefficients and such equations are understood very well - we will study them soon. However, it is of interest to apply the above calculations also to this simple case, even though it can be computed differently. For example, for the period of the oscillations we obtain

$$T = 2\int_{-x_0}^{x_0} \frac{dx}{\sqrt{x_0^2 - x^2}}\,. \tag{59}$$

Letting $x = x_0 s$ in the last integral, we obtain

$$T = \sqrt{\frac{m}{\kappa}}\, 2\int_{-1}^{1} \frac{ds}{\sqrt{1 - s^2}} = 2\pi\sqrt{\frac{m}{\kappa}}\,, \tag{60}$$

where we have used

$$\int_{-1}^{1} \frac{ds}{\sqrt{1 - s^2}} = \pi\,. \tag{61}$$

the integral can be worked out in a number of ways, for example by using the substitution $s = \sin\varphi$.[8]

The following part is optional.

In addition to 1-dimensional pictures such as fig. 4, one can make a plot of the situation in the (x, v) plane, where v is the velocity. It is customary in mechanics to use the momentum $p = mv = m\dot{x}$ rather then the velocity.[9] Expressed in

[8] One can also see it from the formula for the area of the circle: we have $\int_{-a}^{a}\sqrt{a^2 - s^2}\,ds = \frac{1}{2}\pi a^2$ and taking the derivative of this identity with respect to a at $a = 1$ we get (61).

[9] Here we will not discuss the reasons for this. They may not be obvious from what we have learned so far about the system, but at some point in the study of Mechanics it becomes clear that p is the more fundamental quantity.

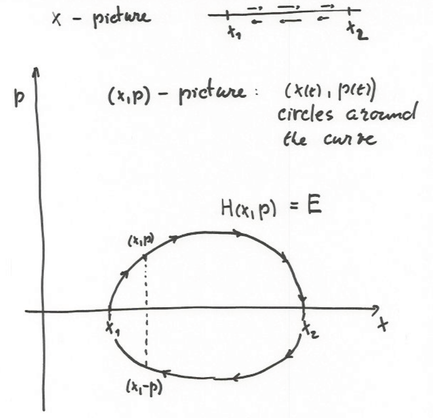
terms of x, p, the total energy is usually denoted by $H = H(x,p)$ and is called the Hamiltonian[10] of the system. In our situation above the formula is

$$H(x,p) = \frac{p^2}{2m} + V(x) \tag{62}$$

The (x,p) plane is called the *phase space* of the system. In the phase space we can plot the curves

$$H(x,p) = E \tag{63}$$

The curve corresponding to the situation on fig. 5 is



The points (x,p) and $(x,-p)$ on the curve tell us that at the point x the particle can have momentum ±p, which is another way of formulating (48). You can find similar pictures in the textbook (for slightly different situations) - see figure 5.1 (page 29) and figure 6.2 (page 34).

2. *Particle moving along a general curve*

In the above considerations we thought of x as a coordinate on a straight line. This is the setup we usually have in mind when we talk about the Newton law (44). However, everything works without any change in the formulae (the only change is in interpretation) if we think about a point mass sliding along some curve in the $n-$dimensional space[11] and think of x as a length parameter along the curve. (The means that the distance along the curve between the

[10] In honor of W. R. Hamilton, 1802 – 1865

[11] in particular, in the plane or the 3d space

points with coordinates x_1, x_2 is $|x_1 - x_2|$, at least when $|x_2 - x_1|$ is small.) Note that if the curve twists and turns in the 3d space, one might be tempted to think three-dimensionally, at least when we think about the force. It turns out, however, that once we know that the motion is constrained to the curve, we can completely forget about the 3-dimensionality of the situation, and pretend that we live in the 1-dimensional space given by the curve. If we take the coordinate x of the curve as length and express the potential energy V in terms of this coordinate, we can pretend that we are on the straight line; the laws of the motion will be the same. This observation is the beginning of the *Lagrangian mechanics*, which very elegantly solves the problem of dealing with motion under constraints. For example, the study rotation of a rigid body about a fixed axis can be fitted into the above scheme. The important point is that when a rigid body is tied to a fixed axis, the configuration space of the system is described by only one parameter – the angle of rotation. This parameter can play the role of the coordinate x above. Although the situation might at first appear quite different from a point-mass sliding along a line, the equations remain the same.

You can have a look at Section 1.6 in the book (starting at page 31). There the above conclusions about the constrained motion (which we only stated but did not prove) are proved in the context of the pendulum, which can be thought of as a motion constrained to a circle, in the presence of gravity.

Lecture 6, 2/4

Motion of a planet (Kepler's problem)

The material covered in this lecture is optional. In the textbook you can find it in Section 4.6. However, the material from differential equations the discussion uses in the end concerns only separable equations of the form (49) and therefore we can discuss it even at this stage. The discussion will use some formulae from multi-variable calculus, which we recall below. If you have not taken a multi-variable calculus, you do not need to worry about it at this point, we really only use the formulae to give a compact appearance to our calculation of the energy conservation. If you prefer to skip the calculations and accept the conservation laws discussed below without proof, it is OK at this stage.

We will consider the classical problem of finding the trajectory of a planet in the gravitational field of the sun. We choose a cartesian coordinate system with the sun located at the origin, and the motion of the planet in the (x_1, x_2)-plane, see fig. 8.

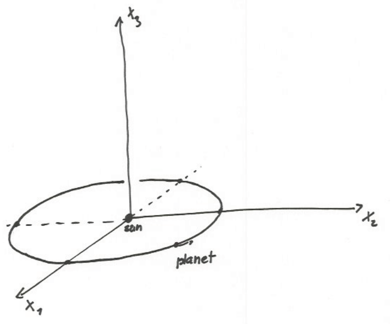


fig. 8

Strictly speaking, the conclusion that the motion can be constrained to a plane may not be completely clear a-priori, it comes as a consequence of Newton's laws, as we will see.

<u>*Some formulae*</u>

We will consider functions and curves in the $n-$dimensional space R^n. You can think of $n = 2$ or $n = 3$ without really losing generality (at least as far as our discussion goes).

Notation

points in R^n	....	$x = (x_1, \ldots, x_n)$
curve in R^n (trajectory)	....	$x(t) = (x_1(t), \ldots, x_n(t))$
euclidean norm in R^n	....	$\lvert x\rvert = \sqrt{x_1^2 + \cdots + x_n^2}$
distance to the origin	....	$r = r(x) = \lvert x\rvert$
scalar product in R^n	....	$x \cdot y = xy = x_1 y_1 + \cdots + x_n y_n$
derivative of a curve $x(t)$ in R^n	....	$\dot{x} = \frac{dx}{dt} = (\dot{x}_1, \ldots, \dot{x}_n) = (\frac{dx_1}{dt}, \ldots, \frac{dx_n}{dt})$

For two curves $x = x(t)$ and $y = y(t)$ we have

$$\frac{d}{dt}(x \cdot y) = \frac{dx}{dt} \cdot y + x \cdot \frac{dy}{dt}\,. \tag{64}$$

In particular,

$$\frac{d}{dt}|x|^2 = 2\,x \cdot \frac{dx}{dt} = 2\,x \cdot \dot{x}\,. \tag{65}$$

For a smooth function $f = f(x) = f(x_1, \ldots, x_n)$ on R^n and a curve $x = x(t)$ we have

$$\frac{d}{dt}f(x(t)) = \frac{\partial f}{\partial x_1}(x)\,\frac{dx_1}{dt} + \cdots + \frac{\partial f}{\partial x_n}(x)\frac{dx_n}{dt} \tag{66}$$

the notation may be shortened in various way, for example, we may just write

$$\frac{d}{dt} f(x) = \sum_i \frac{\partial f}{\partial x_i} \dot{x}_i \,, \tag{67}$$

when it is clear from the context that $x = x(t)$ is a curve. Also, the so-called *Einstein summation convention* is commonly used: we write, for example,

$$\frac{d}{dt} f(x(t)) = \frac{\partial f}{\partial x_i}(x) \frac{dx_i}{dt} \,, \tag{68}$$

where it is understood that we sum over the repeated indices, i. e. $\frac{\partial f}{\partial x_i}(x)\frac{dx_i}{dt}$ really means $\sum_{i=1}^n \frac{\partial f}{\partial x_i}(x)\frac{dx_i}{dt}$. Recalling the definition

$$r = \sqrt{x_1^2 + \cdots + x_n^2} \,, \tag{69}$$

we have, for $x \neq 0$,

$$\frac{\partial r}{\partial x_i} = \frac{x_i}{r} \,. \tag{70}$$

Note that $\frac{x}{r}$ represents a unit vector in the direction of x (unless $x = 0$, when the expression is not well-defined). Looking at the geometric picture, you can convince yourself without doing calculations that the gradient of the function $r = |x|$ (whose graph resembles a funnel) should indeed be given by (70).

Newton's law of gravity and the gravitational potential

Newton's law of gravity says that in the situation on fig. 8, the force on the planet is given by

$$F(x) = -\frac{x}{r} f(r) \,, \tag{71}$$

with

$$f(r) = \frac{\kappa \overline{m} m}{r^2} \,. \tag{72}$$

where $\overline{m}$ is the mass of the sun, m is the mass of the planet and κ is the gravitational constant.[12] This can be also written as

$$F(x) = \kappa \overline{m} m \left(-\frac{x}{r^3}\right) \,. \tag{73}$$

In general dimension $n \geq 2$ the Newton law is

$$F(x) = \kappa_n \overline{m} m \left(-\frac{x}{r^n}\right) \,, \tag{74}$$

where κ_n is the Newton constant in dimension n. (We will not need to speculate about its precise value, the only fact important for us will be that $\kappa_n > 0$.)

[12] $\kappa = 6.6739810^{-11}\mathrm{m}^3\mathrm{kg}^{-1}\mathrm{s}^{-2}$

A crucial fact for our calculation will be that the force $F(x)$ can be written as (negative) gradient of a scalar function, the so called *gravitational potential*, which we will denote by V. More precisely,

$$F_i(x) = -\frac{\partial V}{\partial x_i}(x), \qquad i = 1, \dots n\,, \tag{75}$$

where

$$V(x) = -\frac{\kappa \overline{m} m}{|x|}, \qquad \text{when } n = 3 \tag{76}$$

and

$$V(x) = -\frac{\kappa_n \, \overline{m} \, m}{(n-2)|x|^{n-2}}, \qquad n \geq 3\,. \tag{77}$$

For $n = 2$ we can take

$$V(x) = \kappa_2 \, \overline{m} \, m \, \log \frac{|x|}{r_0} \tag{78}$$

where r_0 is the distance where we wish V to vanish. Note that in dimensions $n \geq 3$ the potential V "vanishes at ∞", i. e. $V(x) \to 0$ as $|x| \to \infty$. Such choice of potential is not possible for $n = 2$. The meaning of the potential is as follows: in dimension $n \geq 3$ the value $-V(x)$ represents the (minimal) amount of work which we will had to do if we wanted to move our planet from x to the spatial ∞, assuming the planet was originally at rest at x. In dimension $n = 2$ we always need infinite amount of work to move the planed to ∞, and we can take replace the infinity in the definition by the circle at distance r_0 from the origin.

Instead of (75) we will also write

$$F(x) = -\nabla V(x)\,. \tag{79}$$

The vector

$$\nabla V = \left(\frac{\partial V}{\partial x_1}, \dots, \frac{\partial V}{\partial x_n} \right) \tag{80}$$

is called the *gradient* of the function V. In general, not every force field F can be written in the form (79). It is a special property of the gravitation force that this is possible.

Newton's equation of motion

The equation of motion of the planet in the above situation[13] is

$$m\ddot{x} = F(x)\,, \tag{81}$$

which is a shorthand for

$$m\ddot{x}_i = F_i(x)\ , \qquad i = 1, \dots n\,. \tag{82}$$

[13] We assume that the sun does not move and no other bodies are present.

The notation is the same as above:

$$\ddot{x} = \frac{d^2x}{dt^2}\,. \tag{83}$$

Trying to solve (81) without any knowledge of various tricks used in similar situations would be a very difficult task. In fact, if you write down what the equation says coordinate-by-coordinate, the problem might look overwhelmingly difficult. Remarkably, the equations can be integrated and the trajectory can be calculated explicitly.[14]

Conservation of energy

The *kinetic energy* of the motion $x(t)$ is an obvious generalization if the 1s formula $\frac{1}{2}m\dot{x}^2$:

$$\text{kinetic energy} = \frac{1}{2}m|\dot{x}|^2\,. \tag{84}$$

The potential energy is

$$\text{potential energy} = V(x). \tag{85}$$

The total energy is

$$\text{total energy} = \frac{1}{2}m|\dot{x}|^2 + V(x)\,. \tag{86}$$

Claim: *The total energy is a constant of the motion.* In other words,

$$\frac{d}{dt}\left(\frac{1}{2}m|\dot{x}|^2 + V(x)\right) \;=\; 0\,. \tag{87}$$

Proof:

$$\frac{d}{dt}\left(\frac{1}{2}m|\dot{x}|^2 + V(x)\right) = m\dot{x}\cdot\ddot{x} + \nabla V(x)\cdot\dot{x} = (m\ddot{x} + \nabla V)\cdot\dot{x} = 0\;, \tag{88}$$

where we have used (67), (79) and (81).

Polar coordinates

We will assume the motion takes place only on the plane (x_1, x_2), i. e. the coordinates $x_3, \dots x_n$ vanish identically during the motion. This assumption will be justified later. It turns out it is advantageous to use the polar coordinates (r, θ) defined by

$$x_1 = r\cos\theta, \qquad x_2 = r\sin\theta\,, \tag{89}$$

see fig. 9.

[14] This was achieved by Newton.

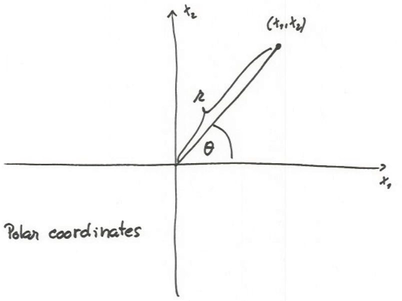


fig. 9

We have

$$\dot{x}_1 = \dot{r}\cos\theta - r(\sin\theta)\,\dot{\theta}\,, \quad \dot{x}_2 = \dot{r}\sin\theta + r(\cos\theta)\,\dot{\theta}\ , \tag{90}$$

and we see that

$$\frac{1}{2}m|\dot{x}|^2 = \frac{1}{2}m\left(\dot{r}^2 + r^2\dot{\theta}^2\right)\ . \tag{91}$$

Hence

$$\text{total energy} = \frac{1}{2}m\left(\dot{r}^2 + r^2\dot{\theta}^2\right) + V(r)\,, \tag{92}$$

where we slightly abuse notation by writing $V(r)$ for $V(x)$ given by (76)– (78) with $|x| = r$. We see that the analogue of the 1d formula (48) is

$$\frac{1}{2}m(\dot{r}^2 + r^2\dot{\theta}^2) + V(r) = E \equiv \text{cost.}\ . \tag{93}$$

In the 1d case formula (48) was enough to integrate the equation. In the case at hand formula (93) is not yet enough, as it contains the term $\dot{\theta}^2$. Luckily, this term can be eliminated by using a second conservation law which we will now discuss.

Conservation of angular momentum and the second law of Kepler

Kepler's laws of planetary motion, derived from astronomical observations, were an important step towards the full theory of Newton. They are as follows:[15]

1. The orbit of every planet is an ellipse with the Sun at one of the two foci.
2. A line joining a planet and the Sun sweeps out equal areas during equal intervals of time.

[15] see `http://en.wikipedia.org/wiki/Kepler's_laws_of_planetary_motion`

3. The square of the orbital period of a planet is directly proportional to the cube of the semi-major axis of its orbit.

Of interest to us at the moment is the second law. In the polar coordinates, it can be expressed as

$$\text{the quantity } r^2\dot\theta \text{ is a constant of motion} \tag{94}$$

or, in other words,

$$\frac{d}{dt}\left(r^2\dot\theta\right) = 0\,. \tag{95}$$

We will now show that this conservation law is a consequence of the equation of motion (81). We can verify by direct calculation that

$$r^2\dot\theta = x_1\dot x_2 - x_2\dot x_1\,. \tag{96}$$

Hence

$$\frac{d}{dt}\left(r^2\dot\theta\right) = \frac{d}{dt}\left(x_1\dot x_2 - x_2\dot x_1\right) = \left(\dot x_1\dot x_2 - \dot x_2\dot x_1\right) + \left(x_1\ddot x_2 - x_2\ddot x_1\right)\,. \tag{97}$$

The first bracket on the right-hand side obviously vanishes. Using (81) and (71) we also see that the second bracket vanishes and the statement is hence proved.[16] Let us set

$$r^2\dot\theta = L \tag{99}$$

We can now express $\dot\theta$ as

$$\dot\theta = \frac{L}{r^2} \tag{100}$$

Substituting into (93), we obtain

$$\frac{1}{2}m(\dot r^2 + \frac{L^2}{r^2}) + V(r) = E\,. \tag{101}$$

Letting

$$e = \frac{E}{m}, \qquad \mu = \kappa\overline{m}\,, \tag{102}$$

we can write (101) as

$$\boxed{\frac{1}{2}\dot r^2 + \frac{1}{2}\frac{L^2}{r^2} - \frac{\mu}{r} = e\,.} \tag{103}$$

This is an equation of the form (47) in the last lecture, with $m = 1$. The problem has been reduced to the 1d situation. We will analyze the solutions next time.

[16] The calculation is even more transparent if we use the cross product. Thinking of the whole situation in 3d, we can write

$$\frac{d}{dt}(x \times \dot x) = \dot x \times \dot x + x \times \ddot x = \dot x \times \dot x + x \times \frac{F(x)}{m}\,. \tag{98}$$

As $a \times a = 0$ for any vector a and $F(x)$ is a multiple of x, we see that the expression on the right-hand side vanishes.

Lecture 7, 2/6

Motion of a planet (Kepler's problem), part 2, (continued from the last lecture)

We will start by analyzing equation (103). Let us set

$$U_{\text{eff}}(r) = \frac{1}{2}\frac{L^2}{r^2} - \frac{\mu}{r}\,. \tag{104}$$

Equation (103) is exactly of the form (47) and we can draw pictures similar to fig. 5, with V replaces by U_{eff}. The graph of U_{eff} is below[17], and we also plot an energy level $e < 0$ in the picture.

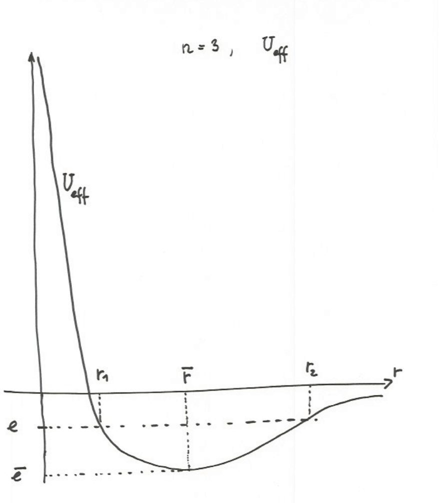


fig. 10

Denoting $0 < r_1 < r_2$ the two roots of the equation $U_{\text{eff}} = e$, we see that the solution of (103) with $e < 0$ will oscillate between r_1 and r_2. By calculating the derivative $\frac{dU_{\text{eff}}}{dr}$ we can see that the minimal possible value of U_{eff} (for a given L and μ) is attained at

$$\bar{r} = \frac{L^2}{\mu}\,, \tag{105}$$

and it is equal to

$$\bar{e} = -\frac{1}{2}\frac{\mu^2}{\bar{r}^2}\,. \tag{106}$$

For a given L, μ, the solution with energy $e = \bar{e}$ corresponds to the circular trajectory at distance $\bar{r}$.

[17] and as an exercise you can check that this picture is qualitatively correct

For $\overline{e} < e < 0$ we know that radius of the trajectory $r(t)$ will stay in $[r_1, r_2]$, so the planet will be "trapped" between the two circles centered at the origin with radii r_1 and r_2, its distance from the sun oscillating between the two radii. At this stage we do not know yet that the trajectory will be "closed". Potentially it could also look as follows

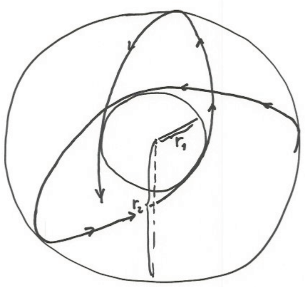


fig. 11

Calculation of the orbit, $n = 3$.

We will now do the calculation which shows that the orbit is an ellipse (for $e < 0$), and hence a closed curve.[18] Using (103) we see

$$\frac{dr}{dt} = \pm\sqrt{2(e - U_{\text{eff}}(r))} \tag{107}$$

and we also have

$$\frac{d\theta}{dt} = \frac{L}{r^2}\,. \tag{108}$$

We can now "eliminate t" from these equations by taking their ratio:[19]

$$\frac{dr}{d\theta} = \pm\frac{r^2}{L}\sqrt{2e + \frac{2\mu}{r} - \frac{L^2}{r^2}}\,. \tag{109}$$

Setting

$$\frac{1}{r} = s \tag{110}$$

[18] If one perturbs V a little to a "generic function" close to the original potential $-\frac{\mu}{r}$, the trajectories may no longer we closed and the situation on fig. 11 becomes relevant. This is also the case in dimension $n = 2$.

[19] This can be justified rigorously, but even at a heuristic level this step may not be completely transparent without some thought. We will return to it later when we discuss systems of equations.

we can write

$$\frac{ds}{d\theta} = \mp\sqrt{\frac{2e}{L^2} + \frac{2\mu}{L^2}s - s^2} = \mp\sqrt{\frac{\mu^2}{L^4} + \frac{2e}{L^2} - (s - \frac{\mu}{L^2})^2}\,. \tag{111}$$

Letting

$$\sigma = s - \frac{\mu}{L^2}, \quad A^2 = \frac{\mu^2}{L^4} + \frac{2e}{L^2}\,, \tag{112}$$

we can write

$$\frac{d\sigma}{d\theta} = \mp\sqrt{A^2 - \sigma^2}\,. \tag{113}$$

Note that $A = 0$ corresponds to e attaining its minimal value for given μ, L, given by (106).
Let us take the coordinates so that

$$r(\theta)|_{\theta=0} = r(0) = r_2, \quad \text{the maximal distance from the sun}\,. \tag{114}$$

Let s_i, σ_i be the values of these variables corresponding to $r_i,\ i = 1, 2$. Then

$$\sigma_1 = -A, \qquad \sigma_2 = A\,. \tag{115}$$

Moreover, as we increase θ from 0 to some small positive value, we expect that σ will increase from $-A$ to some value above $-A$. That means we should take the + sign in (113), at least until sigma will reach A for the first time. We can therefore write

$$\frac{d\sigma}{\sqrt{A^2 - \sigma^2}} = d\theta\,. \tag{116}$$

Integrating both sides, we obtain

$$\arcsin(\frac{\sigma}{A}) - \arcsin(-1) = \theta \tag{117}$$

and recalling $\arcsin(-1) = -\frac{\pi}{2}$, we obtain

$$\sigma = A\sin(\theta - \frac{\pi}{2}) = -A\cos\theta\,. \tag{118}$$

Going back to the variable r, we see that

$$\frac{1}{r} = \frac{\mu}{L^2} - A\cos\theta\,, \tag{119}$$

which is the same as

$$r = \frac{\frac{L^2}{\mu}}{1 - \frac{AL^2}{\mu}\cos\theta} = \frac{\frac{L^2}{\mu}}{1 - \sqrt{1 + \frac{2eL^2}{\mu^2}}\cos\theta} = \frac{\overline{r}}{1 - \varepsilon\cos\theta}\;, \tag{120}$$

where $\overline{r}$ is given by (105) and

$$\varepsilon = \sqrt{1 + \frac{2eL^2}{\mu^2}}\,. \tag{121}$$

This is an equation of an ellipse, with the origin in one of the foci.[20] The quantity ε turns out to be the eccentricity of the ellipse. (Note that $\varepsilon = 0$ corresponds to $e = \overline{e}$.) Note $r(\frac{\pi}{2}) = \overline{r}$, so that $\overline{r}$ coincides with the so-called *latus rectum* of the ellipse.

Also note that the formula makes sense even for $\varepsilon \geq 0$, even though our calculation cannot be taken literally (as (115) is not really well-defined). One can check directly that (120) gives solution to (109) in this case. For $e = 0$ the curve will be a parabola and for $e > 0$ the curve will be a hyperbola, see fig12.

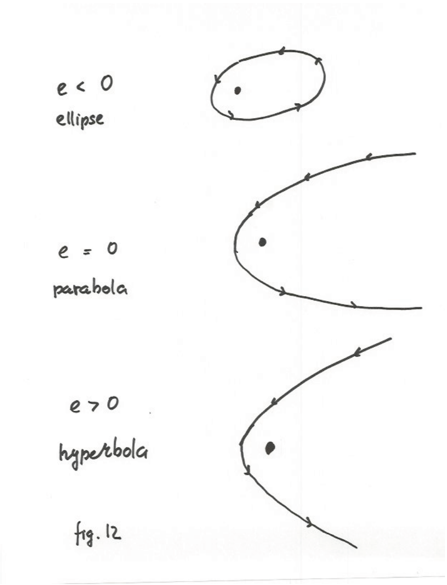


fig. 12

Stability

In the situation of the elliptical orbit, with $L \neq 0$ and $\varepsilon < 0$, if we perturb the parameters slightly[21], the overall situation will not change much. Even with the new values of L, e, if they are not too far away from the original ones, the orbit will change only slightly. We can say that the system is stable.[22] This can also bee seen from fig. 10, where a small change in the graph of U_{eff} will not change r_1, r_2 too much.

Coordinates as a function of t and the calculation of the period

Above we expressed r as a function of θ. Equation (107) should give r as a function of t. If we separate the variables as usual,

[20] See, for example, `http://en.wikipedia.org/wiki/Ellipse`.

[21] Imagine a disturbance due to, say, another celestial body which arrives from deep space, passing at a relatively large distance, disappearing again, never to return.

[22] There are many notions of stability and here we will not go into precise definitions, keeping the discussion at a heuristic level.

$$\frac{dr}{\sqrt{2(e - U_{\text{eff}})}} = \pm dt\,, \tag{122}$$

it is possible to obtain t as a function of r (on a given arc of the ellipse between r_2 and r_1). However, inverting the relation to obtain $r = r(t)$ cannot be done in terms of elementary functions. Nevertheless, by integrating (122) between r_1 and r_2, one can obtain the value of $\frac{T}{2}$, where T is the period of the orbit (which is well-defined as the orbit is closed). You can check as an optional exercise that

$$T = \frac{2\pi\mu}{\sqrt{-2e^3}}\,. \tag{123}$$

What happens in dimensions $n \neq 3$?

Looking at the situation for $n \neq 3$, we can use formulae (76)– (78). The Kepler law holds in any dimension, with the same proof. The motion will always take place in some two-dimensional plane. One way to see it is as follows: assume the motion takes place in the x_1, x_2 plane. Then our calculations give us some trajectories $x_1(t), x_2(t)$. It is easy to check (and it can be done as an optional exercise) that the curve $(x_1(t), x_2(t), 0, \ldots, 0)$ is then a solution of the full $n-$dimensional problem.

In general, the orbits in dimensions $n \neq 3$ behave differently that for $n = 3$. Perhaps the biggest surprise is in dimensions $n \geq 4$. As we discussed in class, it is easy to check that in those dimensions *there are no stable orbits which stay in a bounded region.* We have of course the circular orbits, but they are unstable; a generic small perturbation of the parameters L, e will change the orbit so that the planet will either fall into the sun, or escape to $+\infty$. This can be easily checked by plotting the potentials U_{eff}.

In dimension $n = 2$ the properties of the orbits are closer to the familiar picture from 3d (and they are stable), but there are still significant differences. First, the orbits will typically not be closed, so that they look similar to fig. 11. Second, $U_{\text{eff}}(r) \to \infty$ for $r \to \infty$, and hence we see that *every orbit will stay in a bounded region.* No matter how fast the planet/sattelite goes away from the star, the gravity of the star will eventually turn it back. It can never escape. There are no analogues of the hyperbolic or parabolic orbits which we can have in 3d.

Lecture 8, 2/8

Second order linear equations with constant coefficients - the homogeneous case

In this lecture we discussed the material in 1.9. The main takeaway from the lecture is that for equations of the form

$$ax'' + bx' + cx = 0 \tag{124}$$

one can find a general solution by the following procedure:[23]

1. Solve the characteristic equation (obtained from (124) by seeking x in the form $x = e^{\lambda t}$)

$$a\lambda^2 + b\lambda + c = 0 \tag{125}$$

2. If equation (125) has two *distinct* roots $\lambda_1 \neq \lambda_2$ (real or complex), then the general (complex-valued) solution is

$$x(t) = C_1 e^{\lambda_1 t} + C_2 e^{\lambda_2 t}\,, \tag{126}$$

where C_1, C_2 are arbitrary complex numbers.

3. The case of a double root ($\lambda_1 = \lambda_2 = \lambda$) will be discussed in Lecture 9, but we include the formula here for completeness:

$$x(t) = C_1 e^{\lambda t} + C_2 t e^{\lambda t}\,. \tag{127}$$

Here C_1, C_2 are again arbitrary complex numbers.

If our task is to find a specific solution of (124), given by, say, the initial conditions

$$x(0) = x_0, \quad x'(0) = x_1\,, \tag{128}$$

we choose the constants C_1, C_2 in the general solution so that the conditions characterizing the particular solution we seek are satisfied. It can happen that the equation and the initial data are real, but the roots λ_1, λ_2 are not real. In this case the coefficients C_1, C_2 will have to be complex, so that the end result can be real.

Lecture 9, 2/11

Second order linear equations with constant coefficients (continued)

We still talked about the homogeneous equation (124) and discussed in some detail the damped pendulum equation (9.32) in the textbook. We then discussed the inhomogeneous equation

$$\ddot{x} + \gamma\dot{x} + \omega^2 x = f(t) \tag{129}$$

for $f = e^{i\omega' t}$, along the lines of 1.10, equation (10.1).

The main points of the lecture: formulae (126) and (127) give the general solution of (124). To calculate a specific solution, such as the one specified by (128), we just need to determine C_1, C_2. This typically leads to a system of two equations for C_1, C_2. For example, in case of the conditions (128) the system is

$$\begin{array}{rcrcl} C_1 & + & C_2 & = & x_0\,, \\ \lambda_1 C_1 & + & \lambda_2 C_2 & = & x_1\,. \end{array} \tag{130}$$

[23] Here and below we assume that a, b, c are any numbers (real or complex) with $a \neq 0$.

You can use your favorite method to solve this system. We used Cramer's rule,[24] which is quite practical for 2×2 systems. We obtained

$$C_1 = \frac{\lambda_2 x_0 - x_1}{\lambda_2 - \lambda_1}, \quad C_2 = \frac{x_1 - \lambda_1 x_0}{\lambda_2 - \lambda_1}\,. \tag{131}$$

This gives

$$x(t) = x_0 \frac{\lambda_2 e^{\lambda_1 t} - \lambda_1 e^{\lambda_2 t}}{\lambda_2 - \lambda_1} + x_1 \frac{e^{\lambda_2 t} - e^{\lambda_1 t}}{\lambda_2 - \lambda_1}\,. \tag{132}$$

For $\lambda_1 \neq \lambda_2$ the formulae (131) give a 1-1 correspondence between C_1, C_2 and x_0, x_1. This means that (132) is another form of the general solution, where we now vary x_0, x_1 instead of C_1, C_2. The last formula also has a good limit when λ_1, λ_2 both approach the same point λ. As an exercise you can do the calculation we briefly discussed in class: when x is given by (132), we have

$$\lim_{\lambda_1, \lambda_2 \to \lambda, \lambda_1 \neq \lambda_2} x(t) = x_0(1 - \lambda t)e^{\lambda t} + x_1 t e^{\lambda t}\,. \tag{133}$$

This calculation explains Rule 3 from the last lecture.

We next discussed the inhomogeneous equation

$$\ddot{x} + \gamma \dot{x} + \omega^2 x = f(t) \tag{134}$$

for a special form of f, namely

$$f(t) = e^{i\omega' t}\,. \tag{135}$$

This is one of the most important examples in applications. In this case we proceed by finding a solution in the form

$$x(t) = A e^{i\omega' t}\,, \tag{136}$$

where A is a constant. This leads to

$$A = \frac{1}{-\omega'^2 + i\gamma\omega' + \omega^2}\,. \tag{137}$$

If we consider $|A|$ as a function of ω', we get the very important *resonance curve*. If you search for this term in google images, you will see many pictures of this curve.

As our equation is linear, once we can find one solution with (135), we can also find solutions for linear combinations of forces of this form. Alsom, for any force $f(t)$, once we find one particular solution $\tilde{x}(t)$, we know the general solution has to be of the form

$$x(t) = \tilde{x}(t) + C_1 e^{\lambda_1 t} + C_2 e^{\lambda_2 t}\ , \tag{138}$$

assuming $\lambda_1 \neq \lambda_2$ are the roots of the characteristic polynomial.

[24] see e. g. `http://en.wikipedia.org/wiki/Cramer's_rule`

Lecture 10, 2/10

2nd order linear equations with constant coefficients and the "variation of constants"; types of equations we have learned to solve so far; change of variables

In this lecture we discussed the material from 1.14, and then briefly the change of variables from 1.3, pages 23-24.

The main point of the lecture was the following calculation in 1.14.[25]
We wish to solve

$$ax'' + bx' + cx = f(t) \tag{139}$$

for a given $f(t)$ and we assume that we know the general solution of the homogeneous equation. We will write the general solution of the homogeneous equation in the form

$$x(t) = C_1\phi_1(t) + C_2\phi_2(t)\,, \tag{140}$$

where $\phi_1(t), \phi_2(t)$ are assumed to be known. For example, if the characteristic equation (125) has two different roots $\lambda_1 \neq \lambda_2$, we can take

$$\phi_1(t) = e^{\lambda_1 t}, \quad \phi_2(t) = e^{\lambda_2 t}\,. \tag{141}$$

In the expression (140) the coefficient C_1 and C_2 are constant. To search for solutions of (139), we will use a trick.[26] We will seek the solutions in the form

$$x(t) = C_1(t)\phi_1(t) + C_2(t)\phi_2(t)\,, \tag{142}$$

where $C_1(t), C_2(t)$ are now some yet unknown functions of t. The quantities which were considered as constants in (140) can vary with time in (142), hence the term "variation of constants".[27] In what follows we will use the shorthand notation

$$x = C_1\phi_1 + C_2\phi_2\,, \tag{143}$$

where we understand that all the quantities may now depend on time. We have

$$x' = C_1'\phi_1 + C_2'\phi_2 + C_1\phi_1' + C_1\phi_2'\ . \tag{144}$$

Now comes the non-obvious step of the method: *impose the condition*

$$C_1'\phi_1 + C_2'\phi_2 = 0\,. \tag{145}$$

[25] See also `http://en.wikipedia.org/wiki/Variation_of_constants` or use www.wolframalpha.com to do a search for $ax'' + bx' + cx = f(t)$

[26] After some more study of the differential equations one starts seeing that it is not really an artificial "trick", but a rather natural step. However, this may not be so clear in the beginning.

[27] Sometimes the term "variation of parameters" is used instead.

on the functions C_1, C_2. The meaning of this step becomes clear in the hindsight. Because of (145), we have

$$x' = C_1\phi_1' + C_2\phi_2' \,. \tag{146}$$

This means that, thanks to (145), the expression for the first derivative is the same as if C_1, C_2 were constant. The second derivative of x is

$$x'' = C_1'\phi_1' + C_2'\phi_2' + C_1\phi_1'' + C_2\phi_2'' \,. \tag{147}$$

Plugging (147), (146) and (143) into equation (139), we obtain

$$aC_1'\phi_1' + aC_2'\phi_2' = f \,, \tag{148}$$

as all the other terms drop out due to the fact that ϕ_1, ϕ_2 satisfy the homogeneous equation. We see that we now have a system of two equations for two unknowns C_1', C_2':

$$\begin{array}{ccccc} \phi_1 C_1' & + & \phi_2 C_2' & = & 0\,, \\ \phi_1' C_1' & + & \phi_2' C_2' & = & \frac{f}{a}\,. \end{array} \tag{149}$$

Solving the system for C_1', C_2' (e. g. by Cramer's rule), we obtain,

$$C_1' = \frac{-\frac{f}{a}\,\phi_2}{\phi_1\phi_2' - \phi_1'\phi_2}, \qquad C_2' = \frac{\frac{f}{a}\,\phi_1}{\phi_1\phi_2' - \phi_1'\phi_2} \,. \tag{150}$$

The function in the denominator is usually called the *Wronskian* and denoted by W.[28] We can write (149) as

$$C_1' = -\frac{f\phi_2}{aW}\,, \qquad C_2' = \frac{f\phi_1}{aW} \,. \tag{151}$$

Therefore

$$C_1 = C_1^0 + \int_{t_1}^{t} -\frac{f(s)\phi_2(s)}{aW(s)}\,ds\,, \quad C_2 = C_2^0 + \int_{t_1}^{t} \frac{f(s)\phi_1(s)}{aW(s)}\,ds\,, \tag{152}$$

where t_1 is some fixed time and C_1^0, C_2^0 are constants (which are really constant). It turns out that the wronskian cannot vanish (as long as ϕ_1, ϕ_2 form a basis of the solutions of the homogeneous equation, which is how we choose them).
In class we used the above method to calculate solutions of

$$x'' + \omega^2 x = e^{i\omega t} \,. \tag{153}$$

which can be thought of as an harmonic oscillator forced at exactly the resonant frequency. In this case everything can be calculated explicitly, and we saw that the solution is unbounded in t, as one can expect.

[28] W depends of the choice of the solutions ϕ_1, ϕ_2, but to a lesser degree that one might naively expect. In fact W satisfies the equation $W' + \frac{b}{a}W = 0$, as can be easily checked by direct calculation.

The above method also works for equations with variable coefficients, it was not important that a, b, c were constant.

We summarize which equations we have learned to solve so far:

1. $\dot{x} = a(t)x + b(t)$ (linear equation of the first order)
2. $\dot{x} = f(x)a(t)$ (separable equation)
3. $a\ddot{x} + b\dot{x} + cx = f(t)$ (second order linear equation, constant coefficients)

There are various classes of equations which can be transformed into these form by a suitable change of variables. See for example the section on the homogeneous equations starting on page 23 in the textbook.

Lecture 11, 2/15

Substitutions and changes of variables (optional) **,**
Linear equations and linear spaces

Substitutions and changes of variables (optional)[29]

Let us consider the equation

$$at^2x'' + btx' + cx = 0\,. \tag{154}$$

This equation can be transformed into (124) by the change of variables

$$t = e^s\,. \tag{155}$$

In the beginning of the lecture we discussed issues concerning notation in situations when we do similar changes of variables. For simplicity we assume that the unknown function x in (154) is defined for $t \in I = (0, \infty)$.[30] When we say that we make a substitution $t = e^s$ (for $t \in I$), we typically mean the following: instead for searching for the function $x(t)$, we will search for the function

$$y(s) = x(e^s)\,, \qquad s \in (-\infty, \infty)\,. \tag{156}$$

[29] This part is a slight extension of comments made in response to a question at the beginning of class.

[30] The interval $(-\infty, 0)$ can also be considered without any problems. However, in intervals containing 0, such as $(-1, 1)$ the issues become more subtle, as the leading coefficient at^2 of the equation vanishes at 0. The point 0 is a *singular point* of the equation. Here we will not consider here the theory differential equations near such points. Although it a topic of significant interest, it would lead us in a different direction...

The function $s \to e^s$ maps $(-\infty, \infty)$ onto $(0, \infty)$ in a one-to-one fashion, the inverse mapping is $t \to \log t$. We rewrite (154) in terms of y. We have

$$\begin{aligned} x(t) &= y(\log t)\,, & (157)\\ x'(t) &= y'(\log t)\,\frac{d\log t}{dt} = y'(\log t)\,\frac{1}{t}\,, & (158)\\ x''(t) &= y''(\log t)\,\frac{1}{t^2} - y'(\log t)\,\frac{1}{t^2}\,. & (159)\end{aligned}$$

Substituting into (154), we obtain

$$ay'' + (b-a)y' + cy = 0\,, \tag{160}$$

which is an equation of the form (124).[31] We say that the substitution (155) transforms eqaution (154) into equation (160). The calculation above was based in the chain rule: if h maps an interval I_1 smoothly into an interval J_1 , we have a smooth function $f\colon J_1 \to R$ and $F = f \circ h$, then $F'(t) = f'(h(t))h'(t)$ In the above example we take $h(t) = \log t$ and $x = y \circ h$.
In spite of the simplicity of the above calculation, there are some interesting conceptual issues at play here. We presented the situation as if there were two different functions here, the function $x\colon (0, \infty) \to R$ and the function $y\colon (-\infty, \infty) \to R$. From the "set-theoretic" point of view this is exactly the case: the functions x and y are quite different objects. Recall that, by the set-theoretical definition, a function is a subset[32] of the set $\{\text{domain of } f\} \times \{\text{range of } f\}$. In this point of view, it does not matter whether we write $x(t)$ or $x(s)$. It is understood that the argument of x will always be an element of the domain of x, which is a part of the definition of x. In particular, the expression $x(3.5201)$, say, is defined uniquely and the notation for the derivative x' in unambiguous. Similar remarks can be applied to y. From the way we defined x, y we know that they are related, but this may look somewhat secondary from the set-theoretic point of view, although it is of course the main point of our calculation. The set-theoretic viewpoint has its advantages, but it also has its downside, which is mainly in a certain rigidity and lack of flexibility. For example, if we think of x as some physical quantity depending on, say, time t, we would like to think of x as the same object, whether we measure t in minutes or in seconds. Even if we decide to measure the time near $t = 0$ on the logarithmic scale, it may still be reasonable to think about x as the same quantity, except it is expressed in a different way each time.

This brings us to a second and slightly different point of view. We can view x as a function on some 1-dimensional manifold M of points (representing time, say). The manifold M can be parametrized in one way by t, but we can choose to parametrize it in a different way by s, (with t and s being related in a definite way). When thinking about x in this way, we can think of x either as a function

[31] The trick of changing (154) into (124) was known already to Euler, who used it for more general equations of the form $a_0 t^m x^{(m)}(t) + a_1 t^{m-1} x^{(m-1)}(t) + \cdots + a_{m-1} t x'(t) + a_0 x = 0$.

[32] with some further properties

of t, or as a function of s. But whether we use s or t for the parametrization of the domain, we are essentially dealing with the same function x, except we express it by different means each time.[33] This interpretation of dependence between quantities is presumably closer to the way things were originally viewed. Today mathematicians use both points of view, but for some reason the first point of view became dominant in math education.[34]

When we adopt the second point of view, notation may become somewhat more ambiguous at times. For example, expression $x(3.5201)$ will no longer be well-defined unless we specify that 3.5201 refers to the $t-$variable. We we should should write $x|_{t=3.5201}$, or something similar, if it is not clear which variable is meant. At the same time all of our quantities x, t, s are now considered as functions on some manifold of points M (which can be parametrized by an interval) and there is no need to introduce the function y above. What was y before is just the same function x expressed through the variable s. In this interpretation the meaning of the expressions such as

$$\frac{dx}{dt}, \quad \frac{dx}{ds}, \quad \frac{ds}{dt}, \quad \frac{dt}{ds} \tag{161}$$

is clear: If we wish to express these quantities at some point $P \in M$ we take points $Q \to P,\ Q \neq P$ and, at the point in question P, we have

$$\frac{dx}{dt} = \frac{dx}{dt}(P) = \lim_{Q\to P,\, Q\neq P} \frac{x(Q) - x(P)}{t(Q) - t(P)}\,, \tag{162}$$

$$\frac{dx}{ds} = \frac{dx}{ds}(P) = \lim_{Q\to P,\, Q\neq P} \frac{x(Q) - x(P)}{s(Q) - s(P)}\,, \tag{163}$$

$$\frac{ds}{dt} = \frac{ds}{dt}(P) = \lim_{Q\to P,\, Q\neq P} \frac{s(Q) - s(P)}{t(Q) - t(P)}\,, \tag{164}$$

$$\frac{dt}{ds} = \frac{dt}{ds}(P) = \lim_{Q\to P,\, Q\neq P} \frac{t(Q) - t(P)}{s(Q) - s(P)}\,. \tag{165}$$

Usually we do not specify the “argument” P, we just write the expression (161), with the understanding that they are considered as functions on the manifold M. They can be expressed by means of the various parameters on M, and different parameters may be useful in different situations. So all these expressions can be considered as functions of t, functions of s, and – on intervals where x is a monotone function of t – functions of x.

With this formalism we can for example write

$$\frac{dt}{ds} = \frac{1}{\frac{ds}{dt}}\,, \tag{166}$$

[33] This is clearly related to some notion of equivalence between the different functions in the set-theoretical definition, but we will not try to analyze this in detail, in the hope that the heuristic description above will be sufficient.

[34] Physicist typically use the second point of view. For example, it would be practically impossible to study Thermodynamics while using the first point of view.

which one can easily recognize as the usual theorem for the derivative of the inverse function: if f maps and interval I into the interval J in a one-to-one way, is smooth and $f' \neq 0$ in I, then the inverse function $h = f^{-1}$ is smooth and satisfies

$$h'(f(t)) = \frac{1}{f'(t)}\,. \tag{167}$$

This is exactly (166), with $f(t) = s(t)$.

If we wish to do the transformation between (154) and (124) from this point of view, we can write

$$\frac{dx}{dt} = \frac{dx}{ds}\frac{ds}{dt}\,, \tag{168}$$

and then express $\frac{ds}{dt}$ as a function of s, either by a direct calculation or by

$$\frac{ds}{dt} = \frac{1}{\frac{dt}{ds}} = \frac{1}{\frac{d\,e^s}{ds}} = e^{-s}\,. \tag{169}$$

Alternatively, we can write

$$\frac{dx}{dt} = \frac{dx}{de^s} = \frac{dx}{e^s\,ds} = e^{-s}\frac{dx}{ds}. \tag{170}$$

If we now wish to calculate $\frac{d^2}{dt^2}$ we can write

$$\frac{d^2x}{dt^2} = e^{-s}\frac{d}{ds}[(e^{-s}\frac{d}{ds})x]\,, \tag{171}$$

and one sees that (154) is transformed into

$$a\frac{d^2x}{ds^2} + (b-a)\frac{dx}{ds} + cx = 0\,, \tag{172}$$

as before (in a slightly different notation).

In our simple example of equation (154) it does not really matter which way we think about the calculation, it should be easy with any choice. In more complicated situations the notation with the differentials such as (161) is usually more efficient. One such example is the optional Problem 8 in Homework Assignment 2.

Note that the integration of

$$\frac{dx}{dt} = f(x) \tag{173}$$

on intervals where f does not vanish is particularly transparent in this notation: if I is one interval where f does not vanish, then the function x can obviously be inverted and the equation is equivalent to

$$\frac{dt}{dx} = \frac{1}{f(x)}\,, \tag{174}$$

which immediately shows us that the derivative of the inverse function $t = t(x)$ is $\frac{1}{f(x)}$.

How can we guess the right substitution? For this it is sometimes good to try to write the equation in terms of some simple operations. For example, looking at (154), we see that we can write the whole equation in terms of the operator $t\frac{d}{dt}$. We note that

$$\left(t\frac{d}{dt}\right)^2 = t^2\frac{d^2}{dt^2} + t\frac{d}{dt}\,, \tag{175}$$

and hence (154) can be written as

$$a\left(t\frac{d}{dt}\right)^2 x + (b-a)\left(t\frac{d}{dt}\right) x + cx = 0\,. \tag{176}$$

It is now clear that we should find a substitution $t = t(s)$ such that

$$t\frac{d}{dt} = \frac{d}{ds}\,, \tag{177}$$

which is the same as

$$\frac{ds}{dt} = \frac{1}{t}\,, \tag{178}$$

which of course leads to (155).

There are of course much more complicated examples of successful changes of variables than our examples above. In general, finding a good change of variables is often a key to integrating an equation, and there are no general recipes. On the other hand, extensive knowledge has been accumulated over the 300+ years during which these methods have been studied.[35]

Linear equations and linear spaces

In the main part of the lecture we discussed the set of solution of a linear (ordinary) differential equation from the point of view of Linear Algebra, discussed in Chapter 2 in the textbook.
Let us consider the differential equation

$$a_0 x^{(m)} + a_1 x^{(m-1)} + \dots a_{m-1}x' + a_m x = 0\,, \tag{179}$$

where $x = x(t)$ and $a_0 \neq 0$. We use the notation

$$x^{(k)} = \frac{d^k x}{dt^k}\,. \tag{180}$$

We will assume that the coefficients a_k are independent of t, although many important facts below remain valid also for variable coefficients.

[35] A book of Erich Kamke *Differentialgleichungen: Gewöhnliche Differentialgleichungen* lists a number of equations and substitutions. For methods concerning the equations of Classical Mechanics, one can consult the book of V.I.Arnold *Classical Mechanics.*

We will consider the equation (179) on some interval $I = (t_{\min}, t_{\max})$, where we allow $t_{\min} = -\infty$ or $t_{\max} = +\infty$. We will consider complex-valued solutions. By definition, a function $x\colon I \to \mathbf{C}$ is a solution of (179) if it is a smooth function which satisfies the equation.[36] The set of all solutions of (179) will be denoted by X.
The basic fact about X is as follows:

Theorem 1. *X is a finite-dimensional linear space.*

We will recall the notions used in the theorem. First, a linear space over $\mathbf{C}$ (complex numbers) is a set where we can add two elements and multiply each element by a complex number and, these operations satisfy some natural properties, see 2.1 in the textbook. In our situation when X is a subset of functions, these operation are the expected ones: addition of function and multiplication of a function by a scalar.

The fact that X is a linear space is clear: if $x_1, x_2 \in X$ and $c_1, c_2 \in \mathbf{C}$, it is obvious that $c_1x_1 + c_2x_2 \in X$.

To recall what *finite-dimensional* means in the theorem we first recall the following notions:
A set $B \subset X$ is said to *span* X (as a linear space) if each $x \in X$ can be written as $c_1b_1 + c_2b_2 + \dots c_kb_k$ for some $b_1, \dots b_k \in B$. If there is a finite set $B \subset X$ which spans X, we say that X is finite-dimensional.

Assume X is finite dimensional and B is a finite set which spans it. We say that B is a *basis* of X, if no proper subset of B spans X.
We now recall the following important result from the theory of linear spaces.

Theorem 2. *Every finite-dimensional linear space has a basis. Moreover, any two bases have the same number of elements.*

The number of elements in a basis of a finite-dimensional linear space is called the *dimension* of the space. The dimension is the most important parameter associated with a finite-dimensional linear space.

At this point we will not go into the proofs of the above theorems. However, we will give some arguments showing that they should be true.

First, it is important to realize that *the space of all smooth functions on the interval I is a linear space, but it is not finite-dimensional.* No finite set of smooth functions can span all other smooth functions by linear combinations. As an optional exercise, you can try to prove this fact.

[36] Instead of using smooth functions, we could also work with, say, functions which are $m-$times continuously differentiable. This only makes a difference if the coefficients a_k depend on t and are not smooth functions of t.

We now indicate why X should be finite dimensional (and, in fact, of dimension m). Let $t_0 \in I$. We note that any solution $x \in X$ with the property that $x(t_0) = 0$, $x'(t_0), \dots, x^{(m-1)}(t_0) = 0$ should vanish identically, i. e. $x(t) = 0$ for each $t \in I$. An argument supporting this although not quite a proof) is as follows: first, the equation implies that $x^{(m)}(t_0) = 0$. Taking the derivative of the equation and expressing $x^{(m+1)}$, we see that $x^{(m+1)}(t_0) = 0$. This can be repeated as many times as necessary, and we see that $x^{(n)}(t_0) = 0$ for each $n = 0, 1, 2, \dots$. In other words, the Taylor series of x at t_0 vanishes. In general, if a Taylor series of a function vanishes at a point, the function may not vanish, but for the solutions of (179) this is indeed the case. While the argument via the Taylor serious can be made fully rigorous under some natural assumptions, there is a simpler and also more natural argument based on the so-called Gronwall inequality.

Assume now we can construct solutions $\phi_1, \phi_2, \dots \phi_m$ of our equation such that

$$\left(\begin{array}{cccc} \phi_1(t_0), & \phi_2(t_0)\,, & \dots & \phi_m(t_0) \\ \phi_1'(t_0), & \phi_2'(t_0)\,, & \dots & \phi_m'(t_0) \\ \phi_1''(t_0), & \phi_2''(t_0)\,, & \dots & \phi_m''(t_0) \\ \dots & \dots & \dots & \dots \\ \phi_1^{(m-1)}(t_0), & \phi_2^{(m-1)}(t_0)\,, & \dots & \phi_m^{(m-1)}(t_0) \end{array}\right) = \left(\begin{array}{cccc} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{array}\right). \tag{181}$$

Given a solution $x \in X$, we can define

$$\tilde{x} = x(t_0)\phi_1 + x'(t_0)\phi_2 + \dots x^{(m-1)}(t_0)\phi^{(m-1)}\,. \tag{182}$$

By definition, the derivatives of the function $\tilde{x} - x$ of order $0, 1, \dots (m-1)$ vanish at t_0 and by the uniqueness result above, we see that we have $\tilde{x} = x$. We see that the functions $\phi_1, \dots, \phi_m$ form a basis of X and hence X should be a finite-dimensional linear space of dimension m. Of course, we have not proved that $\phi_1, \dots, \phi_m$ exist, and hence the argument is incomplete. On the other hand, we tied the finite-dimensionality of X to some plausible statements about solutions of differential equations.

Lecture 12, 2/18

The space of solutions of a linear equation (continued)

We use the same notation as last time: X denotes the space of all solution of equation (179). As last time, let $\phi_1, \dots, \phi_m$ be a basis of X.

The linear structure of X enables us to solve various problems concerning the solutions (assuming the basis is known). Let us for example consider the following question: for given

$$t_{\min} \le t_1 < t_2 \dots < t_m \le t_{\max} \tag{183}$$

can we prescribe values of a solution x of (179)? In other words, given $\beta_1, \dots, \beta_m \in \mathbf{C}$, can we find $x \in X$ with

$$x(t_j) = \beta_j,\ j = 1, \dots, m. \tag{184}$$

This is easily addressed using the linear structure. Writing

$$x = c_1\phi_1 + \cdots + c_m\phi_m \tag{185}$$

we see that the equations (184) reduce to the linear system of equations

$$\begin{array}{rcl} \phi_1(t_1)c_1 + \cdots + \phi_m(t_1)c_m & = & \beta_1 \\ \cdots & \cdots & \cdots \\ \phi_1(t_m)c_1 + \cdots + \phi_m(t_m)c_m & = & \beta_m \end{array} \tag{186}$$

which we can write as

$$Ac = \beta\,, \tag{187}$$

where A is the matrix

$$(a_{i,j}) = (\phi_j(t_i))\ . \tag{188}$$

The equation (187) will be (uniquely) solvable for each vector β if and only if the matrix A is non-singular, i. e. $\det A \neq 0$, see Proposition 5.6 (p. 102) in the textbook. As an example, we considered this problem for the second-order equation (154) in the case of two different roots $\lambda_1 \neq \lambda_2$ of the characteristic polynomial. The functions $e^{\lambda_1 t}, e^{\lambda_2 t}$ then form a basis and the matrix A will be

$$\begin{pmatrix} e^{\lambda_1 t_1} & e^{\lambda_2 t_1} \\ e^{\lambda_1 t_2} & e^{\lambda_2 t_2} \end{pmatrix} \tag{189}$$

and the condition $\det A \neq 0$ is easily seen to be equivalent to

$$\lambda_1 t_1 + \lambda_2 t_2 - \lambda_1 t_2 - \lambda_2 t_1 = 2\pi k i\,, \quad k \in \mathbf{Z}\,. \tag{190}$$

which is the same as

$$(\lambda_2 - \lambda_1)(t_2 - t_1) = 2\pi k i\,, \quad k \in \mathbf{Z}\,. \tag{191}$$

We next discussed some natural linear transformations. If we have two bases $\phi_1, \dots, \phi_m$ and $\tilde{\phi}_1, \dots, \tilde{\phi}_m$, we can write

$$\phi_j = p_{1j}\tilde{\phi}_1 + \cdots + p_{mj}\tilde{\phi}_m\,, \quad j = 1, \dots, m\,. \tag{192}$$

The matrix $P = (p_{ij})$ is called the *transition matrix* between the two bases. And it gives a transformation between the coefficients c_j and $\tilde{c}_j$ in the expressions

$$x = c_1\phi_1 + \dots c_m\phi_m = \tilde{c}_1\tilde{\phi}_1 + \cdots + \tilde{c}_m\tilde{\phi}_m\,. \tag{193}$$

Substituting (192) into (193), we see that

$$\tilde{c} = Pc\,. \tag{194}$$

If P is the transition matrix between $\phi_1, \dots, \phi_m$ and $\tilde{\phi}_1, \dots, \tilde{\phi}_m$ and Q is the transition matrix between $\tilde{\phi}_1, \dots, \tilde{\phi}_m$ and $\overline{\phi}_1, \dots, \overline{\phi}_m$, then the transition matrix between $\phi_1, \dots, \phi_m$ and $\overline{\phi}_1, \dots, \overline{\phi}_m$ is QP with the usual matrix multiplication, see 2.2, pp. 84-85 in the textbook.

For equation (154) with constant coefficients we have natural transformations of the space of solutions X defined by the shifts $x(t) \to x(t+s)$. Here we assume that the interval on which the equation is considered is the whole line R. Note that if a function $t \to x(t)$ belongs to X, then its shift $t \to x(t+s)$ also belongs to X. Hence the shift defines a linear transformation $A(s) \colon X \to X$. Expressed in a given basis, $A(s)$ can be identified with a matrix. As an exercise, you can check that the following identity is satisfied:

$$A(s_1 + s_2) = A(s_1)A(s_2)\,, \quad s_1, s_2 \in R\,. \tag{195}$$

Thus the space X comes with an additional structure: the set of linear transformations (identified with matrices) satisfying (195). Note that if the space X contains non-constant functions, these transformations will be non-trivial, except perhaps for some special values of s.

Lecture 13, 2/19

Matrices, determinants, eigenvalues, eigenvectors

We discussed some basic results about $n \times n$ matrices and determinants. In the textbook the determinants are introduced in Section 2.5. The main points of the lecture:

Theorem 3. *Let $A = (a_{ij})_{i,j=1}^n$ be an $n \times n$ matric (real or complex). Then the following statements are equivalent:*
(i) For each vector b the equation

$$Ax = b \tag{196}$$

(where x is an vector) has a solution.

(ii) The equation

$$Ax = 0 \tag{197}$$

has only the trivial solution $x = 0$.

(iii) For each vector b the equation (196) has a unique solution.

(iv)

$$\det A \neq 0\,. \tag{198}$$

Determinant as volume

For real matrices A the determinant has the following geometric interpretation. Let us write

$$A = (a_1, a_2, \ldots, a_n)\,, \tag{199}$$

where $a_1, a_2, \ldots, a_n$ are column vectors. Consider the set

$$\mathcal{O}_{a_1,a_2,\ldots,a_n} = \left\{ \sum_{i=1}^{n} t_i a_i\,,\ (t_1, t_2, \ldots, t_n) \in [0,1]^n \right\}\,. \tag{200}$$

Then

$$n\text{−dimensional volume of } \mathcal{O}_{a_1,a_2,\ldots,a_n} = \pm \det A\,. \tag{201}$$

See fig. 13.

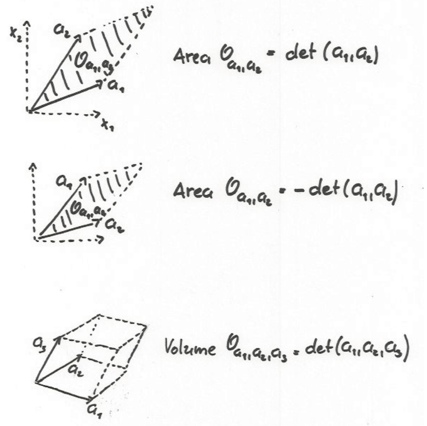

x_2
a_2
$\mathcal{O}_{a_1,a_2}$
a_1
x_1
Area $\mathcal{O}_{a_1,a_2}$ = det (a_1,a_2)
a_1
$\mathcal{O}_{a_1,a_2}$
a_2
Area $\mathcal{O}_{a_1,a_2}$ = − det(a_1,a_2)
a_3
a_2
a_1
Volume $\mathcal{O}_{a_1,a_2,a_3}$ = det(a_1,a_2,a_3)

fig. 13

The sign is chosen according to the following rule: first, if the vectors $a_1, \ldots, a_n$ do not form a basis of R^n, then $\det A = 0$ and there is no problem with the choice of the sign. If $a_1, \ldots, a_n$ do form a basis, then we take the + sign, if the basis has the same *orientation* as the canonical basis $e_1, \ldots, e_n$. This means that for $s \in [0,1]$ we can find a continuous family of bases $a_1(s), \ldots, a_n(s)$ such that for $s = 0$ we have the canonical basis and for $s = 1$ we have our given basis. It is the same as saying that the matrix A can be connected to the identity matrix within the set set of matrices with non-zero determinant. If, on the other hand, the basis $a_1, \ldots, a_n$ can be continuously deformed (within the set of bases) to the basis $-e_1, e_2, \ldots, e_n$, we take the $-$ sign. It can be shown that one of the possibilities will always occur and the two possibilities are mutually exclusive.

As an optional exercise, you can convince yourself at least in dimensions $n = 2$ and $n = 3$ that (201) is true. [37]

Cramer's rule

If $\det A \neq 0$, the solution of (196) is given by

$$x_i = \frac{\det A^{(i,b)}}{\det A}\,, \tag{204}$$

where the matrix $A^{(i,b)}$ is obtained from A by replacing the $i-$th column by b. In other words, if we write

$$A = (a_1, \dots, a_n)\ , \tag{205}$$

then

$$A^{(i,b)} = (a_1, \dots, a_{i-1}, b, a_{i+1}, \dots, a_n)\ . \tag{206}$$

You can derive (204) as an optional exercise. [38]

[37] The proof can be done in many ways. First, it is easy to see that the formula is true if the matrix $A = (a_1, \dots, a_n)$ is triangular. Next, one can check that the rules by which A can be changed without affecting $\det A$ do not affect the volumes either. Finally one can use the fact that A can be brought to a diagonal form by using these rules.
Alternatively, show that $\det(a_1, \dots, a_n)$ is independent of the choice of the positively-oriented orthogonal coordinates in which the vectors a_i are expressed. One way this can be done is as follows: given $a_1, \dots, a_n$ and an orthogonal matrix Q which can be connected to the unit matrix I in the orthogonal matrices, we have

$$\det(a_1, \dots, a_n) = \det(Qa_1, \dots, Qa_n)\,. \tag{202}$$

This follows from $\det(QA) = (\det Q)(\det A)$ and $\det Q = 1$. The last identity follows from taking det of $QQ^t = I$ (which gives $(\det Q)^2 = 1$) and using that Q can be connected to I in the orthogonal matrices. Once (202) is established, then (201) becomes clear, as we can choose the basis in which $a_1, \dots, a_n$ are expresses in such a way that the matrix $A = (a_1, \dots, a_n)$ is triangular.
Formula (202) can also be expressed in the following way. For $1 \leq i_1, \dots, i_n \leq n$ we define $\varepsilon_{i_1 i_2, \dots i_n}$ to be 0 if $i_k = i_l$ for some $k \neq l$, 1 if $i_1, \dots, i_n$ is an even permutation of $1, 2, \dots, n$, and -1 if $i_1, \dots, i_n$ is an odd permutation of $1, 2, \dots, n$. The important fact, closely related to (202), now is that $\varepsilon_{i_1 i_2 \dots i_n}$ is a *pseudo-tensor*, which means that it is invariant up to a sign under the orthogonal change of coordinates. In other words, if $Q = (q_{ij})$ is an orthogonal matrix, then

$$\sum_{j_1, \dots, j_n = 1}^{n} \varepsilon_{j_1 j_2 \dots j_n} q_{j_1 i_1} q_{j_1 i_2} \dots q_{j_n i_n} = \pm \varepsilon_{i_1 i_2 \dots i_n}\ , \tag{203}$$

where the sign depends on whether the transformation Q is orientation-preserving or orientation-reversing.

The main point here is more general: formulae which have some geometric meaning must be invariant under changes of coordinates. Formula (202) or formula (203) expresses exactly that for the determinant. In fact, the determinant is invariant under a bigger group of transformations than orthogonal transformations, but for the purpose of seeing (201) without much calculation, the invariance under the orthogonal group in the sense above is sufficient.

[38] Hint: re-write the equation $Ax = b$ as $x_1 a_1 + x_2 a_2 + \dots, x_n a_n = b$ and show that

$$\det(a_1, \dots, a_{i-1}, x_1 a_1 + \dots + x_n a_n, a_{i+1}, \dots, a_n) = x_i \det A \tag{207}$$

by using $\det(\tilde a_1, \dots, \tilde a_n) = 0$ whenever any two of the vectors $\tilde a_i$, $\tilde a_j$ with $i \neq j$ coincide.

Eigenvalues and eigenvectors

We discussed the definitions in Section 2.6 of the textbook.

Lecture 14, 2/22

Eigenvalues and eigenvectors (continued)

Let A be an $n \times n$ matrix over the real or complex numbers. Recall that $\lambda \in \mathbf{C}$ is an *eigenvalue* of A if there exists a non-zero vector x (which can be complex) such that

$$Ax = \lambda x\,. \tag{208}$$

By Theorem 3 we see that $\lambda \in \mathbf{C}$ is an eigenvalue if and only if

$$\det(\lambda I - A) = 0\,, \tag{209}$$

where I is the identity matrix (characterized by $Ix = x$ for each x). The polynomial

$$p(\lambda) = p_A(\lambda) = \det(\lambda I - A) \tag{210}$$

is called the *characteristic polynomial* of the matrix A. By the fundamental theorem of algebra, we can write

$$p(\lambda) = (\lambda - \lambda_1)(\lambda - \lambda_2)\dots(\lambda - \lambda_n)\,, \tag{211}$$

where $\lambda_1, \dots, \lambda_n$ are the roots of the polynomial p. In general, some of the λ_j can coincide (so we can have for example $\lambda_1 = \lambda_2$). The number of occurrences of a given root λ_j among the roots $\lambda_1, \dots, \lambda_n$ is called the (algebraic) *multiplicity* of the root λ_j. Renumbering the roots, if necessary, we can also write

$$p(\lambda) = (\lambda - \lambda_1)^{k_1}(\lambda - \lambda_2)^{k_2}\dots(\lambda - \lambda_r)^{k_r}\,, \tag{212}$$

where now $\lambda_i \neq \lambda_j$ for $i \neq j$ and k_j is the multiplicity of λ_j. Clearly $k_1 + k_2 + \dots + k_r = n$.

Even if the matrix is real, the roots may be complex, in general.[39]

We have the following important fact (see Proposition 6.2, p. 108 in the textbook):

Lemma 1. *Let A be an $n \times n$ matrix and let $\lambda_1, \dots, \lambda_r$ be any subset of its eigenvalues such that $\lambda_i \neq \lambda_j$ for $i \neq j$, $1 \leq i, j \leq r$. Let $x^{(j)}$ be an eigenvector corresponding to λ_j.*[40] *Then the vectors $x^{(1)}, \dots, x^{(r)}$ are linearly independent. In particular, when $r = n$ the vectors form a basis of $\mathbf{C}^n$.*

[39] There is still one conclusion one can make about the roots related to A being real: if A is real and λ is an eigenvalue, then its complex conjugate $\overline{\lambda}$ is also an eigenvalue. In other words, for real matrices the complex eigenvalues come pairs $\lambda, \overline{\lambda}$. The proof of this statement is left to the reader as an easy exercise.

[40] We recall that, by definition, an eigenvector must be $\neq 0$.

The proof is not hard, you can try to do it as an exercise, or you can check the proof in the textbook (p. 108).

The important point now is that for real or complex $n \times n$ matrices the situation that the characteristic polynomial has n different roots *generically.* We will now explain this notion.

Let us start by a simple example. A quadratic equation

$$a\lambda^2 + b\lambda + c = 0\,, \quad a \neq 0 \tag{213}$$

has roots

$$\lambda_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}\,. \tag{214}$$

We have $\lambda_1 \neq \lambda_2$ unless $b^2 - 4ac = 0$. When $b^2 - 4ac = 0$, equation (213) has only one root. Its multiplicity will be 2. We see the coefficients a, b, c of (213) have to satisfy a non-trivial condition

$$b^2 - 4ac = 0 \tag{215}$$

to have $\lambda_1 = \lambda_2$. In the three-parameter space a, b, c the condition (215) represents a surface. If we choose a, b, c "at random" they will typically not lie on this surface. From this point of view, the "event" that the polynomial (213) has only one root (of multiplicity two) is "exceptional", and very unlikely. If, for a given set of parameters, some event happens only when the coefficients satisfy some fixed (algebraic) equation, we say that the complement of that event happens *generically.* In the example above, we would say that a quadratic polynomial with non-vanishing leading term generically has two different roots. That does not mean that the polynomial always has two different roots. It means that the polynomial fails to have two different roots only for the set of the parameters for which some non-trivial (algebraic) relation is satisfied. If we choose the parameters (a, b, c) "at random" from some open set in $\mathbf{R}^3$ (or $\mathbf{C}^3$), relation (215) will almost never be satisfied, which means that under a random choice of the parameters the polynomial will almost surely have two different roots. This may not be the case when we choose our parameters randomly from some finite set of parameters. For example, if we specify that we will be choosing parameters a, b, c in (213) so that they are integers with values between -10 and 10 and $a \neq 0$, a "random choice" from this finite set will give with non-zero probability polynomials with a single root of multiplicity two.

The situation with higher-order polynomials is similar. For polynomials of order n, which we will write as

$$p(\lambda) = \lambda^n + a_1\lambda^{n-1} + \cdots + a_{n-1}\lambda + a_n\,, \tag{216}$$

there is a non-trivial polynomial $D(a_1, \ldots, a_n)$ in the coefficients (called the discriminant) [41] such that the equation

$$p(\lambda) = 0 \tag{217}$$

[41] see, for example, http://en.wikipedia.org/wiki/Discriminant

has n different roots when $D(a_1, \ldots, a_n) \neq 0$. This means that the coefficients $a_1, \ldots, a_n$ have to fall on some particular "surface" in the space of the coefficients $a_1, \ldots, a_n$ for the polynomial to have multiple roots. If we choose the polynomial "at random" from some open set of the coefficients, then almost surely the discriminant $D(a_1, \ldots, a_n)$ will not vanish, and the polynomial will have exactly n different roots. In other words, the situation when we have n different roots is generic.

When we are dealing with a characteristic polynomial of a matrix $A = (a_{ij})_{i,j=1}^n$ the coefficients $a_1, \ldots, a_n$ of the characteristic polynomial $p_A(\lambda)$ depend in a polynomial way on the coefficients a_{ij} of the matrix. Therefore the condition $D(a_1, \ldots, a_n) = 0$ can be written in terms of the coefficients a_{ij} as

$$\mathcal{D}(a_{11}, a_{12}, \ldots, a_{nn}) = 0\,, \tag{218}$$

where $\mathcal{D}$ is a polynomial of the n^2 coefficients a_{ij}. We emphasize that the polynomial $\mathcal{D}$ may be quite complicated and we will not try to determine it explicitly. For the purposes of our discussion here it is enough to know that the polynomial is non-trivial, as we only wish to illustrate that the case when all the eigenvalues are different is typical (generic), while the other case can be considered as exceptional, al least when we choose matrices from an open subset of all matrices.[42] The matrices for which the characteristic polynomial does not have exactly n different roots are characterized by (218). We see that *unless (218) is satisfied, the matrix A will have exactly n different eigenvalues.* By Lemma 1, in this case the corresponding eigenvectors $x^{(1)}, \ldots, x^{(n)}$ will form a basis of $\mathbf{C}^n$. Note that the linear mapping given by A in the canonical basis, namely,

$$\begin{pmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{pmatrix} \to \begin{pmatrix} a_{11}x_1 + \cdots + a_{1n}x_n \\ \cdot \\ \cdot \\ \cdot \\ a_{n1}x_1 + \cdots + a_{nn}x_n \end{pmatrix} \tag{219}$$

will become very simple in the basis $x^{(1)}, \ldots, x^{(n)}$. Namely, we have

$$Ax^{(j)} = \lambda_j x^{(j)}\,, \qquad j = 1, 2, \ldots, n\ . \tag{220}$$

Hence the matrix of the map (219) in the basis $x^{(1)}, \ldots, x^{(n)}$ will be diagonal:

$$\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n) = \begin{pmatrix} \lambda_1 & 0 & 0 & \ldots & 0 \\ 0 & \lambda_2 & 0 & \ldots & 0 \\ 0 & 0 & \lambda_3 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & \lambda_n \end{pmatrix}\,. \tag{221}$$

[42] Note that if the set of the matrices from which our matrix is chosen is not open in the set of all matrices, it may be a subset of the set given by (218).

We note that the transition matrix P between the basis $x^{(1)}, \ldots, x^{(n)}$ and the canonical basis (see lecture 12, (192)) is given by the matrix

$$P = \left(x^{(1)}, \ldots, x^{(n)} \right) , \tag{222}$$

where we write the eigenvectors $x^{(j)}$ as column vectors. Therefore, from the general facts about transition matrices, the matrix in the basis $x^{(1)}, \ldots, x^{(n)}$ of the map expressed in the canonical basis as $x \to Ax$ is

$$P^{-1}AP \,. \tag{223}$$

At the same time, we know from (220) that this matrix must be equal to Λ. We conclude that

$$\Lambda = P^{-1}AP \,. \tag{224}$$

This is the same as

$$P\lambda = AP \,, \tag{225}$$

which is obvious from the fact that the $j-$th column of P is $x^{(j)}$. Finally, we can also write (225) as

$$A = P\Lambda P^{-1} \,. \tag{226}$$

The important conclusion from the above discussion is that the transformation (224) from a general matrix A to a diagonal matrix is generically possible. The matrices for which it may not be possible must satisfy (218) and hence form a "thin" set in the space of all matrices.[43] The situation in this "thin" set is addressed by the theory of Jordan canonical forms, which we will discuss later.

Let us now illustrate the usefulness of the above notion by a simple example.

Example

Consider two locations, let us call them (1) and (2). We assume that a certain number of people live in these two locations and each week some of them relocate between the two locations according the the following rules:

(i) A person in (1) will relocate to (2) with probability $p \in (0,1)$, and stay at (1) with probability $1-p$.

(ii) A person in (2) will relocate to (1) with probability q, and will stay at (2) with probability $1-q$.

We assume

$$0 < p, q < 1 \,. \tag{227}$$

How many people will be in each location after many weeks?

[43] Even if (218) is satisfied, it may still be possible to diagonalize the matrix, but "generically" *in that set* it is not possible. Note that here we restricted the notion of "genericity" to the set (218), which would strictly speaking need a precise definition, which we omit at this point.

To avoid a discussion of notions from the probability theory which are not very relevant to what we wish to illustrate here, let us assume for simplicity that the number of people at each location does not have to be a natural number and that the relocation is governed by the following deterministic rule: each week $100p$ percent of the people in (1) move to (2) and the rest stays in (1) and, similarly, $100q$ percent of the people in (2) move to (1) and the rest stays at (2).

Let us write down what happens during one round of relocation. If x_1, x_2 give the number of people at (1) and (2) respectively before a relocation, the numbers $x_1^{\text{new}}, x_2^{\text{new}}$ after a relocation are given by

$$x_1^{\text{new}} = (1-p)x_1 + qx_2, \qquad x_2^{\text{new}} = px_1 + (1-q)x_2 \,. \tag{228}$$

Introducing the matrix

$$M = \begin{pmatrix} 1-p & q \\ p & 1-q \end{pmatrix} \tag{229}$$

and writing

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \tag{230}$$

we see that each week the vector x changes from x to Mx. Therefore the sequence

$$x, \, Mx \, , \, M^2x \, , \, M^3x \, , \dots , \, M^kx \, , \dots \tag{231}$$

describes everything we need to know. Note that $x_1 + x_2$ should remain constant in our model, and we see from (228) that this is indeed the case.

The powers of the matrix M are not easily calculated from the form (229). On the other hand, we note that the powers A^k of the matrix A in the form (226) are easily calculated: if $A = P\Lambda P^{-1}$, then $A^k = P\Lambda^k P^{-1}$, and $\Lambda^k = \operatorname{diag}(\lambda_1^k, \dots, \lambda_n^k)$.

Let us find the eigenvalues and eigenvectors of M. The characteristic polynomial is

$$\det(M - \lambda I) = \det \begin{pmatrix} 1-p-\lambda & q \\ p & 1-q-\lambda \end{pmatrix} = \lambda^2 - (2-p-q)\lambda + 1 - p - q \,. \tag{232}$$

The roots are

$$\lambda_1 = 1, \qquad \lambda_2 = 1 - p - q \,. \tag{233}$$

Note that

$$\lambda_2 \in (-1, 1) \,. \tag{234}$$

The corresponding eigenvectors (determined only up to a scalar multiple) can be taken for example as

$$x^{(1)} = \begin{pmatrix} q \\ p \end{pmatrix} , \qquad x^{(2)} = \begin{pmatrix} 1 \\ -1 \end{pmatrix} . \tag{235}$$

Note that the eigenvector $x^{(2)}$ forms a basis of the 1-d space $\{x, x_1 + x_2 = 0\}$ which is preserved by the matrix M, and this can be used to identify it without calculation.

Let us write

$$x = \xi_1 x^{(1)} + \xi_2 x^{(2)} \,, \tag{236}$$

so that ξ_j are coordinates of x in the basis $x^{(1)}, x^{(2)}$. In these coordinates the transformation M is

$$\begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} \to \begin{pmatrix} \lambda_1 \xi_1 \\ \lambda_2 \xi_2 \end{pmatrix} = \begin{pmatrix} \xi_1 \\ \lambda_2 \xi_2 \end{pmatrix} . \tag{237}$$

We see that the coordinate ξ_1 is preserved. This reflects the preservation of $x_1 + x_2$. Indeed, the transition matrix from the canonical basis to the basis $x^{(1)}, x^{(2)}$ is

$$P = \begin{pmatrix} q & 1 \\ p & -1 \end{pmatrix} . \tag{238}$$

The transition matrix between the basis $x^{(1)}, x^{(2)}$ and the canonical basis is

$$P^{-1} = \frac{1}{p+q} \begin{pmatrix} 1 & 1 \\ p & -q \end{pmatrix} . \tag{239}$$

The relation between the coordinates x and ξ is

$$x = P\xi, \qquad \xi = P^{-1} x \,. \tag{240}$$

In particular

$$\xi_1 = \frac{x_1 + x_2}{p+q} \,. \tag{241}$$

The sequence (231) corresponds in the new variables to

$$\begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} , \begin{pmatrix} \xi_1 \\ \lambda_2 \xi_2 \end{pmatrix} , \begin{pmatrix} \xi_1 \\ \lambda_2^2 \xi_2 \end{pmatrix} , \dots , \begin{pmatrix} \xi_1 \\ \lambda_2^k \xi_2 \end{pmatrix} , \dots \tag{242}$$

As we have $|\lambda_2| < 1$ we see that the sequence converges to

$$\begin{pmatrix} \xi_1 \\ 0 \end{pmatrix} . \tag{243}$$

Going back to the variables x, we see that the sequence (231) will converge to the vector

$$\begin{pmatrix} \frac{q}{p+q}(x_1 + x_2) \\ \frac{p}{p+q}(x_1 + x_2) \end{pmatrix} . \tag{244}$$

This vector represents a “dynamical equilibrium” of the system: each week the same number of people depart from (1) to (2) as is the number of people who arrive from (2) to (1), so that the number of people at each location remains the same.

We see that switching to the basis consisting of the eigenvectors makes the situation completely transparent in this case. We will see that often the same is true about differential equations.

Remark (optional)

The above example can be generalized to the situation with n locations, which is related to the so-called Perron-Frobenius Theorem.[44] The matrix M then becomes

$$M = \begin{pmatrix} p_{11} & p_{12} & \dots & p_{1n} \\ p_{21} & p_{22} & \dots & p_{2n} \\ \dots & \dots & \dots & \dots \\ p_{n1} & p_{n2} & \dots & p_{nn} \end{pmatrix}, \tag{245}$$

which, despite the use of p_{ij} in the notation, should not be confused with some transition matrix P between bases. The numbers p_{ij} are all strictly positive[45] and denote the probability that a person will move from location j to location i. Note that if the system is "closed" (i. e. the people can move only between the n locations), then

$$p_{1i} + p_{2i} + \dots + p_{ni} = 1\,. \tag{246}$$

With these assumptions it can be shown that for a vector x with non-negative coordinates, the sequence

$$x, Mx, M^2x, \dots, M^kx, \dots \tag{247}$$

converges to a vector $\overline{x}$ satisfying

$$M\overline{x} = \overline{x}\,, \tag{248}$$

which represents a (unique) "dynamical equilibrium". Clearly $\overline{x}$ is an eigenvector, the corresponding eigenvalue being $\lambda = 1$. All the other eigenvalues λ_j (which can be complex) satisfy $|\lambda_j| < 1$, so that the situation is quite similar to the 2d case calculated above. The proofs are now more complicated, though.

Lecture 15, 2/25

The spectrum of a real symmetric matrix

By definition, the *spectrum* of a matrix (real or complex) is the set of all its eigenvalues.

In this lecture we discussed real symmetric matrices. Recall that a real $n \times n$ matrix $A = (a_{ij})_{i,j=1}^n$ is symmetric if

$$a_{ij} = a_{ji}\,, \qquad 1 \le i, j \le n\,. \tag{249}$$

To each symmetric matrix A there corresponds a quadratic form in $\mathbf{R}^n$ given by

$$x \to \frac{1}{2}\sum_{i,j} a_{ij}x_jx_i = \frac{1}{2}(Ax)\cdot x\,, \tag{250}$$

[44] See e. g. `http://en.wikipedia.org/wiki/PerronFrobenius_theorem`

[45] In fact, the conclusions below are true to all elements of some power M^k of M are strictly positive.

where $x \cdot y$ denotes the scalar product $\sum_{i=1}^{n} x_i y_i$. Vice versa, each quadratic form on $\mathbf{R}^n$ arises in this way (and determines a symmetric matrix). The main point of the lecture was the proof of the following important theorem.

Theorem 4. *Let A be a real symmetric matrix. Then all its eigenvalues are real and there is an orthogonal basis of $\mathbf{R}^n$ in which A becomes diagonal. In other words, there are real numbers $\lambda_1, \dots, \lambda_n$ (not necessarily all different from each other) and an orthogonal matrix Q such that*

$$A = Q\Lambda Q^{-1} = Q\Lambda Q^t\,, \tag{251}$$

where Q^t denotes the transposed matrix to Q and

$$\Lambda = \mathrm{diag}\,(\lambda_1, \dots, \lambda_n)\,. \tag{252}$$

A proof of the theorem can be found in the textbook, Section 2.11 (p. 130).

Later we will discuss also the version of this theorem for Hermitian matrices.

Lecture 16, 2/27

Real symmetric matrices (continued)

1. Ellipses, Ellipsoids, Hyperbolae, Hyperboloids... (optional)

If $\lambda_1, \dots, \lambda_n > 0$, then the equation

$$\lambda_1 x_1^2 + \cdots + \lambda_n x_n^2 = c > 0 \tag{253}$$

describes an ellipsoid with axes parallel to the coordinate axes, of length $\frac{1}{\sqrt{\lambda_j}}$, $j = 1, 2, \dots, m$. In particular, in dimension $n = 2$ it is an ellipse. If A is a symmetric matrix which is positive definite, i. e. $(Ax) \cdot x > 0$ for $x \neq 0$, the equation

$$(Ax) \cdot x = c > 0 \tag{254}$$

also describes an ellipsoid, but in general its axes are not parallel to the coordinate axes. The problem of finding the axes of the ellipsoid is practically the same as the problem of finding the eigenvalues/eigenvectors of A. There are many ways to see this. Let us consider for example the following. Let M denote the surface (254). If we suspect that M should be an ellipsoid, it is natural to try to find its longest/shortes axis by maximizing/minimizing the function

$$x_1^2 + \cdots + x_n^2 \tag{255}$$

over the surface M. This is a typical constraint minimization problem, which can be solved by the method of Lagrange multipliers. Let

$$f(x) = \frac{1}{2}(Ax) \cdot x\,, \qquad g(x) = x_1^2 + \cdots + x_n^2\,. \tag{256}$$

If f attains a maximum/minimum on the surface $\{g(x) = c\}$ at some point $\overline{x}$, then, recalling the Lagrange multiplier method, we have

$$\nabla f(\overline{x}) = \lambda \nabla g(\overline{x}) \tag{257}$$

for some $\lambda \in \mathbf{R}$, where we denote by ∇f the gradient $\left(\frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_n}\right)$. A straightforward calculation gives

$$\nabla f(x) = Ax, \qquad \nabla g(x) = x \tag{258}$$

and we see that (257) is equivalent to

$$Ax = \frac{1}{\lambda} x\,. \tag{259}$$

We see that the eigenvectors correspond to the critical values of the function g on the surface (254). One can also reverse the role of the function f and g and consider the extrema of f on the unit sphere $\{x,\, |x|^2 = 1\}$, which leads to equation

$$Ax = \lambda x\,. \tag{260}$$

If the matrix A is non-singular indefinite (the form $(Ax)x$ attains both positive and negative values), we obtain hyperboloids, rather than ellipsoids.

Example - interacting oscillators (see also Section 3.6 in the textbook)

Assume that we have a mass m on a spring of stiffness κ. For simplicity we will ignore gravity for the moment, assuming the the only (non-inertial) force on the particle is due to the spring. Also, we assume that the motion of the mass is one-dimensional. The equation is

$$m\ddot{x} + \kappa x = 0\,. \tag{261}$$

We know how to solve this equation - the general solution is

$$x = c_1 \cos(\omega t) + c_2 \sin(\omega t)\,, \qquad \omega = \sqrt{\frac{\kappa}{m}}\,. \tag{262}$$

If we have n of such oscillators with masses $m_1, \dots, m_n$ and spring constants $\kappa_1, \dots, \kappa_n$ and the oscillators do not interact with each other, we have an equation of the form (261) for each of the oscillators, and we can solve the equations separately - no oscillator influences any other oscillators.

Suppose now we have several oscillators which interact with each other (perhaps by introducing some extra springs between them). Let us consider the important case of small oscillations of such a system about its equilibrium.

Assume the system is described by coordinates $x_1, \dots, x_n$, with x_j describing the coordinate of the $j-$th particle. To each configuration $x_1, \dots, x_n$ of the system we associate potential energy $V(x_1, \dots, x_n)$. In the system with springs the value of V represents the elastic energy stored in the springs for the given configuration. The force on the particle i under these assumptions is

$$f_i = -\frac{\partial V}{\partial x_i}\,. \tag{263}$$

We expect that the configuration $(\overline{x}_1, \dots, \overline{x}_n)$ with a minimal possible energy[46] V will be a "rest state" of our system, i. e. in this configuration the system is at an equilibrium, without any motion. If we perturb slightly our system from this equilibrium position, it will oscillate around it. Assuming the perturbation is sufficiently small, we can expect to be able to write

$$V(\overline{x}+y) = V(\overline{x}) + \sum_i \frac{\partial V(\overline{x})}{\partial x_i} y_i + \frac{1}{2}\sum_{i,j} \frac{1}{2}\frac{\partial^2 V}{\partial x_i \partial x_j}(\overline{x}) y_i y_j + O(|y|^3)\,. \tag{264}$$

The constant term $V(\overline{x})$ does not affect the force (263), and the linear term $\sum_i \frac{\partial V(\overline{x})}{\partial x_i} y_i$ vanishes as $\overline{x}$ is minimizes V. Therefore the dominant term for small $|y|$ will be $\frac{1}{2}\sum_{i,j}\frac{1}{2}\frac{\partial^2 V}{\partial x_i \partial x_j}(\overline{x}) y_i y_j$. Letting

$$a_{ij} = \frac{\partial^2 V}{\partial x_i \partial x_j}(\overline{x})\,. \tag{265}$$

The equations of motion (expressed in terms of the variables y_i) are

$$m_i \ddot{y}_i + \sum_j a_{ij} y_j = 0. \tag{266}$$

Let us write

$$\sqrt{m_i}\; y_i = z_i\,, \qquad \tilde{a}_{ij} = \frac{a_{ij}}{\sqrt{m_i m_j}}\,. \tag{267}$$

One check easily that under this change of variables (266) becomes

$$\ddot{z}_i + \sum_j \tilde{a}_{ij} z_j\,. \tag{268}$$

We will write this as

$$\ddot{z} = \tilde{A} z\,. \tag{269}$$

The matrix $\tilde{A}$ is symmetric, and hence we know that

$$\tilde{A} = Q^{-1} \Lambda Q\,, \tag{270}$$

where Q is an orthogonal matrix and

$$\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_n)\,. \tag{271}$$

[46] Here we do not go into the questions under which assumptions such minimizers exist, are unique, etc.

Letting

$$w = Qz\,, \tag{272}$$

we see that

$$\ddot{w} + \Lambda w = 0\,, \tag{273}$$

or

$$\ddot{w}_j + \lambda_j w_j = 0\,, \qquad j = 1, 2, \dots, n, \tag{274}$$

which we know how to solve. We see that in the new variables w_i we again have non-interacting harmonic oscillators, at least mathematically. The possibly complicated linear interaction given by $\tilde{A}$ in (269) can be always changed by a suitable change of variables into the simple situation of the non-interacting oscillators. We emphasize that this is only possible for the linear systems.

The eigenvalues λ_j will then give the frequencies at which our system will produce sound, if it is oscillating in a gas which transmits sound.

The above considerations apply to many mechanical systems oscillating with a small amplitude about an equilibrium, as long as the forces acting in the system are given by some potential energy. (Such forces are usually called *conservative.*)

Lecture 17, 3/1

More on linear equations with constant coefficients

Some general considerations (optional, with the exception of formula (290))

Let us return to the linear equation (179), which we will consider with $a_0 = 1$ (without loss of generality), i. e.

$$x^{(m)} + a_1 x^{(m-1)} + \dots + a_{m-1}x' + a_m x = 0\,. \tag{275}$$

The coefficients $a_1, \dots, a_m$ are assumed to be constant.[47] Later we will also consider the non-homogeneous equation

$$x^{(m)} + a_1 x^{(m-1)} + \dots + a_{m-1}x' + a_m x = f(t)\,. \tag{276}$$

There are simple algorithms for calculating the solutions of such equations, which we will discuss soon. We will first discuss the equation from the more theoretical point of view. This part is optional. What you should really know is the practical algorithm for solving the equation.

[47] The case when the coefficient a_j depend on t is also important. Although some of the general properties of the solutions in this case are similar to the constant coefficient case, the investigation of more detailed properties of the solutions can be quite more subtle.

Let us consider the homogeneous equation (275) on the real line $\mathbf{R}$. Let X be the space of the solutions of the equation.[48]
As we discussed in lecture 11, X is a linear space of dimension m. Let us consider complex-valued solutions, so the linear space X is considered over the complex numbers $\mathbf{C}$. The coefficients a_j can then also be complex.

Note that the space X comes without some kind of a "canonical basis". We can of course choose a basis, and some choices are more natural than others, but there is more than one good choice.[49]

In lecture 12 we defined the shift operator A on the space X, see the paragraph just before (195). Recall that

$$[A(s)x](t) = x(t+s)\,. \tag{279}$$

Note that this is an example where we defined a linear mapping in another way than by specifying its matrix in a given basis. The map is defined "intrinsically". As we discussed in lecture 11, A satisfies

$$A(s_1+s_2) = A(s_1)A(s_2). \tag{280}$$

[48] Strictly speaking, we should specify what exactly we mean by a solution. A natural definition is that x is a solution of (275) if it is a function with m continuous derivatives which satisfies (275). It can be shown (and it is not too hard, you can try to do it as an optional exercise) that any such solution is in fact infinitely smooth. That means, even though we initial require only m continuous derivatives, the solution actually has continuous derivatives of any order. The reason behind this is that the equation expresses the $m-$th derivative in terms of the derivatives of order $\leq m-1$. Let us show how to prove that the solution (originally defined as having m continuous derivatives actually has $m+1$ continuous derivatives. We show the argument for the equation

$$x' = ax\,, \tag{277}$$

the argument for (275) being essentially the same. Assume x satisfies (277) and has one continuous derivative. Let us consider $D_s x$ defined by

$$D_s x(t) = \frac{x(t+s)-x(t)}{s}\,. \tag{278}$$

Then clearly $D_s x$ is again a solution of (277) and as $s \to 0$ the functions $D_s x$ converge locally uniformly to x'. Now from the equation we get that the functions $D_s(x')$ also converge locally uniformly to some continuous function. This means that x is twice differentiable and $D_s(x') \to x''$ as $s \to 0$. This step can now be repeated to obtain that x has in fact 3 continuous derivatives. Then it can again be repeated to show that x has 4 continuous derivatives, etc., leading to the conclusion that x is infinitely smooth.

[49] The situation is somewhat similar to the following example. Let us consider the linear space $\mathbf{C}^n$. This space does have a "canonical basis", namely $e_1 = (1,0,\dots,0),\dots,e_n = (0,\dots 0,1)$. There are of course many other bases, but most people will probably agree that the canonical basis is the simplest one to choose. Let us now consider the linear $(n-1)-$dimensional subspace Y of $\mathbf{C}^n$ given by the equation $z_1+\dots+z_n = 0$. One can of course choose a basis in Y, for example $e_1-e_n, e_2-e_n, \dots e_{n-1}-e_n$ but this choice is kind of "arbitrary", there are many other ways to choose basis which are just as good and perhaps even more natural. Choosing a basis of the space X of solutions of (275) is somewhat like choosing a basis in Y. There is not a clear "canonical choice". In some sense, the choice of the canonical basis $e_1,\dots,e_n$ in $\mathbf{C}^n$ is also somewhat "arbitrary", but in some sense the most "economical".

This says that A is a homomorphism from the additive group $\mathbf{R}$ into the multiplicative group of invertible linear maps of the space X. If we choose a basis, we can identify A with a matrix, and $s \to A(s)$ can be considered as a map from $\mathbf{R}$ into the set of $n \times n$ non-singular complex matrices, which is usually denoted by $GL(n, \mathbf{C})$. The set $GL(n, \mathbf{C})$ is a group under the matrix multiplication, and (280) again says that $s \to A(s)$ is a group homomorphism.

The property (280) puts strong restrictions and A and, in fact, maps $s \in \mathbf{R} \to A(s) \in GL(n, \mathbf{C})$ which have this property can be completely characterized. We will return to this point later.

The shift operator $x(t) \to x(t+s)$ is closely related to the derivative operator $x \to x'$, as we have

$$x'(t) = \lim_{s \to 0} \frac{x(t+s) - x(t)}{s} \,. \tag{281}$$

Note that $x \in X$ implies that $x' \in X$. Therefore

$$x \to x' \tag{282}$$

can be considered as a linear operator on X. Let us write

$$x' = Bx \,. \tag{283}$$

Now B is a linear map of the finite-dimensional linear space X into itself, and therefore can be identified with a matrix. The matrix representation which we will obtain for B will depend on our choice of a basis. If $\phi_1, \dots, \phi_n$ is a basis of X, then we have

$$\phi_j' = b_{1j}\phi_1 + \cdots + b_{nj}\phi_n \qquad j = 1, 2, \dots, m \tag{284}$$

and the matrix of the map B in this basis is $(b_{ij})_{i,j=1}^m$. From general principles concerning linear maps between finite-dimensional spaces we know that the map B has a non-trivial eigenvector with eigenvalue λ. This means that there must be a function $x \in X$ such that

$$x' = \lambda x \,. \tag{285}$$

This means that

$$x(t) = Ce^{\lambda t} \,. \tag{286}$$

If the mapping B has m different eigenvalues $\lambda_1, \dots, \lambda_m$, we know that then the corresponding eigenvectors, which in this case are functions $C_j e^{\lambda_j t}$ form a basis of X. This is why we should search the solutions of the form $e^{\lambda t}$. If we substitute the function $e^{\lambda t}$ into (275), we obtain

$$\lambda^m + a_1\lambda^{m-1} + \cdots + a_{m-1}\lambda + a_m = 0 \,. \tag{287}$$

The polynomial on the left-hand side is called the *characteristic polynomial* of (275). It turns out that the characteristic polynomial of the equation coincides

with the characteristic polynomial of the linear mapping B.[50]
If the characteristic polynomial (287) has m different roots $\lambda_1, \lambda_2, \dots, \lambda_m$, then the functions

$$e^{\lambda_1 t}, e^{\lambda_2 t}, \dots, e^{\lambda_m t} \tag{289}$$

form a basis of X and the general solution of the (homogeneous) equation is

$$C_1 e^{\lambda_1 t} + C_2 e^{\lambda_2 t} + \dots + C_m e^{\lambda_m t}\,. \tag{290}$$

In practical terms this can be easily verified directly, without the more abstract considerations above: clearly the functions (280) solve the equation (this is how we chose λ_j) and one only needs to show that the functions (289) are linearly independent, which is not hard (and it is a good exercise).

What happens in the case of multiple roots? Here we just state the rule, without analyzing it. Assume that (287) has a root λ of multiplicity k. Then the functions

$$e^{\lambda t}, te^{\lambda t}, \dots, t^{k-1}e^{\lambda t} \tag{291}$$

all solve (275). This can be verified by a direct calculation. It is easy to verify that there functions are linearly independent. Therefore, for each root λ of multiplicity k we add to the general solution an expression

$$C_1 e^{\lambda t} + C_2 t e^{\lambda t} + \dots + C_k t^{k-1} e^{\lambda t} \tag{292}$$

and this way we again get a space of solution of dimension m.

Higher-order equation as a first-order system (Section 3.3 in the textbook)

One can arrive also arrive at an equation similar to (283) by the following more "concrete" way. We again consider (275). We set

$$x = y_1\,,\ x' = y_2\,,\ x'' = y_3\,,\ \dots, x^{(m-1)} = y_m\,. \tag{293}$$

Clearly

$$y_1' = y_2,\ y_2' = y_3,\ \dots y_{m-1}' = y_m \tag{294}$$

[50] Recall that that polynomial is defined as $p(\lambda) = \det(\lambda I - B)$. We remark that our definition of determinant was based on a matrix representation of a map. That is, we defined determinant for a matrix, not for a linear map. We can of course define it for a linear map by saying that the determinant of a linear map is the determinant of its matrix (in some basis). One has to check that this definition does not depend on the choice of the basis, which you can do as an optional exercise.
To see that the characteristic polynomial of B is (287) we note that (275) implies that

$$B^m + a_1 B^{m-1} + \dots + a_{m-1}B + a_m = 0\,. \tag{288}$$

We claim that B cannot satisfy a polynomial relation $B^k + \tilde{a}_1 B^{k-1} + \dots + \tilde{a}_{k-1}B + \tilde{a}_k = 0$ for $k < m$. If this were the case, then any function in X would satisfy $x^{(k)} + \tilde{a}_1 x^{(k-1)} + \dots + \tilde{a}_{k-1}x' + \tilde{a}_k = 0$, and therefore the dimension of X would have to be $\leq k$, a contradiction. We can now recall the Caley-Hamilton Theorem from linear algebra (see `http://en.wikipedia.org/wiki/Cayley-Hamilton_theorem`) to conclude that not only is the left-hand side of (287) a characteristic polynomial of B, but it is in fact the so called *minimal polynomial* of B. We have not defined this notion yet. It is relevant in the case of multiple roots.

by definition of y and equation (275) gives

$$y'_m = -a_1 y_{m-1} - a_2 y_{m-2} \cdots - a_m y_1 \,. \tag{295}$$

In other words, the vector-valued function y satisfies

$$y' = Ay \,, \tag{296}$$

where the matrix A is given by

$$A = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 \\ -a_m & -a_{m-1} & -a_{m-2} & \dots & -a_1 \end{pmatrix} . \tag{297}$$

As an optional exercise you can calculate the characteristic polynomial of A. The calculation gives

$$\det(\lambda I - A) = \lambda^m + a_1 \lambda^{m-1} + \cdots + a_{m-1}\lambda \, + a_m \,, \tag{298}$$

the same as the characteristic polynomial of the equation.
Assume that the characteristic polynomial has m different roots $\lambda_1, \dots, \lambda_m$ and let $b_1, \dots, b_m$ be the corresponding eigenvectors of A. By Lemma 3, the eigenvectors form a basis of $\mathbf{C}^n$. Let us write

$$y = z_1 b_1 + \dots z_m b_m \,, \tag{299}$$

i. e. z_j are coordinates of the vector y in the basis $b_1, \dots, b_m$. In these coordinates equation (296) becomes

$$\begin{array}{ccl} z'_1 & = & \lambda_1 z_1 \,, \\ z'_2 & = & \lambda_2 z_2 \,, \\ \dots & \dots & \dots \,, \\ z'_m & = & \lambda_m z_m \,. \end{array} \tag{300}$$

Hence the general solution of (296) is

$$y = C_1 e^{\lambda_1 t} b_1 + C_2 e^{\lambda_2 t} b_2 + \dots C_m e^{\lambda_m t} b_m \,, \tag{301}$$

which again leads to expression (290) for the general solution of (275).

General linear systems $x' = Ax$ for vector-values x and constant A.

Let us consider functions $x\colon \mathbf{R} \to C^n$. Let A be an $n \times n$ (complex) matrix and let us consider the equation

$$x' = Ax \,. \tag{302}$$

The space of solutions of this equation is a linear space of dimension n. Assume that $\phi_1, \dots, \phi_n$ are (vector-valued) functions which solve (302). If these functions are linearly independent, then the expression

$$x = C_1 \phi_1 + C_2 \phi_2 + \cdots + C_n \phi_n \tag{303}$$

is a general solution of (302), in the same sense as we considered for the scalar equation (275): each solution x of (275) can be expressed in the form (303) for some set of constants $C_1, \dots, C_n$.

If the matrix A has n different eigenvalues $\lambda_1, \dots, \lambda_n$, then, letting

$$\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_n) \tag{304}$$

we have

$$A = P\Lambda P^{-1}\,, \tag{305}$$

where P is the matrix with the $j-$the row being the $j-$the eigenvector, see also (222). Setting $z = P^{-1}x$, we see that the equation (302) is transformed to

$$z' = \Lambda z\,, \tag{306}$$

which has the same from as (300).

Systems of the form (302) can be used to model many phenomena. We see from the considerations above that in the "generic case" they can be transformed to the diagonal systems (306). The diagonal system consists of n equations of the form $z_j = \lambda_j z_j$ which "do not interact" with each other. Each of those equations can be solved separately, and we know how to do from lecture 1. In general , system (302) can describe complicated interaction between the variables x_j (as long as the interaction remains linear). Mathematicians have know for about 200 years that in the generic case such systems can be "diagonalized", i. e. all the non-trivial interactions can be "removed" by a suitable change of variable and today this is of course well-understood. Nevertheless, it is still remarkable that such a simplification can be achieved.

Example

Let us start by considering 2 simple equations

$$\begin{aligned} \dot{x}_1 &= a\,x_1\,, \\ \dot{x}_2 &= b\,x_2\,. \end{aligned} \tag{307}$$

Let us assume $a, b > 0$, and we can think of two entities owning some money (or debt), with x_1 the balance of the first entity and x_2 the balance of the second entity. Either balance can be positive or negative. The number $a > 0$ represents interest rate for entity (1) and $b > 0$ represents interest rate for entity (2).[51] In the model (307) there is no connection between x_1 and x_2, the two quantities do not interact with each other. The solution of (307) is simple:

$$x_1(t) = \overline{x}_1 e^{at}\,, \qquad x_2(t) = \overline{x}_2 e^{bt}\,, \tag{308}$$

where $\overline{x}_j = x_j(0)\,,\ j = 1, 2$. Assume now that entity (2) discovers that it can extract money from entity (1) and the amount they can extract is proportional

[51] It is of course an over-simplification of the real situation to assume that the interest rate remains the same whether the balance is positive or negative, the model can still be useful.

to x_2. (It is easy to imagine that the influence of (2) is proportional to the capital it owns.) Then the equations become

$$\begin{aligned} \dot{x}_1 &= a\,x_1 - cx_2\,, \\ \dot{x}_2 &= (b+\tilde{c})\,x_2\,, \end{aligned} \tag{309}$$

where $0 < \tilde{c} < c$. The quantity cx_2 is the amount of money extracted by (2) from (1) per unit of time. We assume that there is some cost of the extraction, so the money which (2) receives (per unit of time) is not cx_2, but rather $\tilde{c}x_2$, with $0 < \tilde{c} < c$. Letting

$$\tilde{b} = b + \tilde{c} \tag{310}$$

we will write (309) as

$$\begin{aligned} \dot{x}_1 &= a\,x_1 - cx_2\,, \\ \dot{x}_2 &= \tilde{b}\,x_2\,, \end{aligned} \tag{311}$$

or

$$\dot{x} = Ax\,, \tag{312}$$

with

$$A = \begin{pmatrix} a & -c \\ 0 & \tilde{b} \end{pmatrix}. \tag{313}$$

The solutions of (311) can be calculated easily without diagonalizing the matrix A: we note that the solution of the second equation is

$$x_2(t) = \overline{x}_2 e^{\tilde{b}t}\,, \tag{314}$$

with $\overline{x}_2 = x_2(0)$, and once x_2 is known we can calculate x_1 by the variation of constants, which leads to

$$x_1(t) = \overline{x}_1 e^{at} - c\int_0^t x_2(s)e^{a(t-s)}\,ds\,. \tag{315}$$

A simple integration now gives

$$x_1(t) = \begin{cases} \left(\overline{x}_1 - \frac{c\overline{x}_2}{a-\tilde{b}}\right) e^{at} + \frac{c\overline{x}_2}{a-\tilde{b}} e^{\tilde{b}t}\,, & a \neq \tilde{b} \\ \overline{x}_1 e^{at} - c\overline{x}_2 t e^{at}\,, & a = \tilde{b}\,. \end{cases} \tag{316}$$

Note that when $0 < \tilde{b} < a$, the quantity $\overline{x}_1$ can still grow exponentially for suitable $\overline{x}_1, \overline{x}_2$, in spite of the fact that some funds are being diverted to entity (2). The exponential growth at the rate a is sufficient to keep entity (1) in the positive as long as $\frac{\overline{x}_2}{\overline{x}_1} < \frac{a-\tilde{b}}{c}$. However, once $\tilde{b} \geq a$, entity (1) will always eventually go into debt as long as $\overline{x}_2 > 0$.

It is instructive to derive (316) by diagonalizing the matrix A. As we shall see, this is only possible for $a \neq \tilde{b}$.

Let us look at the eigenvalues and eigenvectors of A. The vector $x^{(1)} = e_1 = (1,0)$ is clearly an eigenvector with eigenvalue a. As $\det(\lambda I - A) = (a-\lambda)(\tilde{b}-\lambda)$, we see that $\tilde{b}$ is also an eigenvalue. The corresponding eigenvector is easily determined from $(A - \tilde{b}I)x = 0$, which is the same as

$$(a - \tilde{b})x_1 - cx_2 = 0. \tag{317}$$

If $a \neq \tilde{b}$, we can take for the second eigenvector the vector

$$x^{(2)} = (c, a - \tilde{b})\,. \tag{318}$$

When $a \neq \tilde{b}$, the vectors $x^{(1)}, x^{(2)}$ form a basis of $\mathbf{R}^2$ and the general solution of (312) can be written as

$$x(t) = C_1 x^{(1)} e^{\alpha t} + C_2 x^{(2)} e^{\tilde{b} t}\,. \tag{319}$$

Note that this coincides with formulae (314), (315) when we choose

$$C_1 = \overline{x}_1 - \frac{c\overline{x}_2}{a - \tilde{b}}\,, \qquad C_2 = \frac{\overline{x}_2}{a - \tilde{b}}\,. \tag{320}$$

In fact, the eigenvectors can be already seen from (314), (315) as the directions of the maximal resp. minimal growth as $t \to \infty$ (when $a > \tilde{b}$).

In the case $a = \tilde{b}$ the matrix A does not have two linearly independent eigenvectors. It is an example of a "non-generic" situation. We will discuss such situations more systematically soon.

At some point entity (1) may also discover that it can extract some money from entity (2), and the situation becomes symmetric. We will end up with equations

$$\begin{aligned} \dot{x}_1 &= \tilde{a}x_1 - cx_2\,, \\ \dot{x}_2 &= -dx_1 + \tilde{b}x_2\,, \end{aligned} \tag{321}$$

which can again be analyzed completely by finding eigenvalues and eigenvectors. The model (321) is of course too simple for practical considerations concerning real-world competing entities,[52] but it is already interesting.

Lecture 18, 3/4

Jordan canonical form

We discussed the Jordan canonical form, Section 2.A in the textbook (p. 140).

Lecture 19, 3/6

Jordan canonical form (continued)

Let X_1 and X_2 be two finite dimensional linear spaces defined over the same field of scalars. (We will usually consider complex spaces although real spaces will also be occasionally considered.) We define a new linear space

$$X = X_1 \oplus X_2 \tag{322}$$

[52] The real-world situation rarely allows for an indefinitely sustained exponential growth, and therefore more realistic models should be non-linear.

called the *direct sum* of the two spaces as follows. As a set, X is the set of all pairs (x_1, x_2) with $x_1 \in X_1$ and $x_2 \in X_2$. The linear space structure on X is given by

$$(x_1, x_2) + (y_1, y_2) = (x_1 + y_1, x_2 + y_2)\,, \qquad c(x_1, x_2) = (cx_1, cx_2)\,. \tag{323}$$

Often we write (perhaps with slight abuse of notation) $x_1 + x_2$ instead of (x_1, x_2). Here the notation $x_1 + x_2$ is "abstract" in the sense that originally we do not assume that x_1, x_2 belong in the same linear space. However, the direct some $X_1 \oplus X_2$ provides exactly such a space. It is just a way of saying that we will consider the formal sums $x_1 + x_2$ as elements a space which contains both X_1 and X_2 in such a way that the two spaces X_1, X_2 do not interact with each other. The sum $x_1 + x_2$ cannot be zero if one of the x_j is non-zero. If $e_1, \dots, e_m$ is a basis of X_1 and $f_1, \dots, f_n$ is a basis of X_2, then $e_1, \dots, e_m, f_1, \dots f_n$ is a basis of X. The elements of X can be expressed as

$$x_1 e_1 + \dots x_m e_m + y_1 f_1 + \dots y_n f_n\,. \tag{324}$$

The statement that every finite-dimensional space (over $\mathbf{C}$, say) has a basis can is equivalent to saying that every finite dimensional space X over C is of the form

$$X = \underbrace{\mathbf{C} \oplus \mathbf{C} \oplus \dots \oplus \mathbf{C}}_{n = \dim X \text{ copies}}, \tag{325}$$

where we think of $\mathbf{C}$ as a one-dimensional vector space (over itself).

Assume now that a linear space X is given and let $X_1 \subset X$ be its linear subspace. Assume that X_2 is a subspace of X such that the linear span of the set $X_1 \cup X_2$ is X and $X_1 \cap X_2 = \{0\}$. Then each element $x \in X$ has a unique decomposition

$$x = x_1 + x_2 \qquad x_1 \in X_1,\ x_2 \in X_2\,, \tag{326}$$

and we can identify X with $X_1 \oplus X_2$. In this situation we simply write

$$X = X_1 \oplus X_2 \tag{327}$$

although, strictly speaking, from the set-theoretic point of view, it may be more isomorphism rather than equality. In practice it is not necessary to dwell on these distinctions, unless we are really interested in some subtle set-theoretical issues with the definitions.

When $X = X_1 \oplus X_2$, we have natural projections

$$P_1 \colon X \to X_1\,, \qquad P_2 \colon X \to X_2 \tag{328}$$

defined by

$$P_1(x_1 + x_2) = x_1\,, \qquad P_2(x_1 + x_2) = x_2\,. \tag{329}$$

Note that

$$P_j^2 = P_j, \qquad P_j(X) = X_j, \qquad j = 1, 2\,. \tag{330}$$

A simple but important property of linear spaces is as follows:

Lemma 2. *If X is a finite-dimensional linear space and $X_1 \subset X$ is a subspace, then there exist a subspace $X_2 \subset X$ such that $X = X_1 \oplus X_2$.*

Let $A\colon X \to X$ be a linear map. If $X_1 \subset X$ is a linear subspace and $A(X_1) \subset X_1$, we say that X_1 is invariant under A. The analogue of Lemma 2 is not necessarily true in the category of spaces invariant under a given map A: if $X_1 \subset X$ invariant under A, there may not be a subspace $X_2 \subset X$ invariant under A such that $X = X_1 \oplus X_2$. (As an optional exercise you can construct an example of such a situation.)

If X_1, X_2 are finite-dimensional linear spaces and $A_1\colon X_1 \to X_1\,,\ A_2\colon X_2 \to X_2$ are linear mappings, we define the mapping

$$A = A_1 \oplus A_2\colon \quad X_1 \oplus X_2 \to X_1 \oplus X_2 \tag{331}$$

by

$$A(x_1 + x_2) = A_1 x_1 + A_2 x_2\,. \tag{332}$$

Let $e_1, \dots, e_m$ be a basis of X_1 and let $f_1, \dots, f_n$ be a basis of X_2. If $\tilde{A}_1$ is the matrix of A_1 in the basis $e_1, \dots, e_m$ and $\tilde{A}_2$ is the matrix of A_2 in the basis $f_1, \dots, f_n$, then the matrix of $A_1 \oplus A_2$ in the basis $e_1, e_2, \dots, e_m, f_1, \dots f_n$ is the "block-diagonal" matrix

$$\begin{pmatrix} \tilde{A}_1 & 0 \\ 0 & \tilde{A}_2 \end{pmatrix}. \tag{333}$$

For $\lambda \in \mathbf{C}$ and a natural number $k \geq 1$ we consider the $k \times k$ matrix

$$J_k(\lambda) = \begin{pmatrix} \lambda & 1 & 0 & \dots & \dots & 0 \\ 0 & \lambda & 1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & 0 & \lambda & 1 & 0 \\ \dots & \dots & \dots & 0 & \lambda & 1 \\ 0 & 0 & 0 & \dots & 0 & \lambda \end{pmatrix} \tag{334}$$

with the convention that

$$J_1(\lambda) = (\lambda)\ . \tag{335}$$

Let us say that a linear map is of type $J_k(\lambda)$ if it can be represented by $J_k(\lambda)$ in some basis.

Theorem 5. *Let X be a finite-dimensional linear space over $\mathbf{C}$ and let $A\colon X \to X$ be a linear mapping. Then there exists a decomposition $X = X_1 \oplus X_2 \oplus \cdots \oplus X_r$ and maps $A_j\colon X_j \to X_j$ of type $J_{k_j}(\lambda_j)$ such that $A = A_1 \oplus A_2 \oplus \cdots \oplus A_r$.*

Note that the theorem does not exclude the possibility that some of the λ_j or k_j occur multiple times.

We can interpret the theorem as follows: we have a list of simple objects (maps of type $J_k(\lambda)$) and we have a simple construction which can put these objects together (direct sum, in our case here). The theorem above says that any linear mapping can be obtained in this way.

Another way of stating Theorem 5 is as follows: for every linear map $A\colon \mathbf{C}^n \to \mathbf{C}^n$ there is a basis in which the matrix of A is of the "block form"

$$\begin{pmatrix} J_{k_1}(\lambda_1) & 0 & \dots & 0 \\ 0 & J_{k_2}(\lambda_2) & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & J_{k_r}(\lambda_r) \end{pmatrix} . \tag{336}$$

Yet another way is as follows: given any matrix A over the complex numbers, there exists a non-singular (complex) matrix P such that the matrix PAP^{-1} is of the form (336). The matrix P can be interpreted as a transition matrix to a basis in which A has the form (336).

We will now discuss the main points of the proof of the theorem. This part is optional. A proof based on a slightly different approach can be found in the textbook, see Sections 2.7 and 2.A.

Let us consider an $n \times n$ matrix A. From the fundamental theorem of algebra we know that the characteristic polynomial has a non-trivial root λ , and hence the matrix

$$M = A - \lambda I \tag{337}$$

is not invertible. In particular its kernel

$$\mathrm{Ker}(M) = \{x \in \mathbf{C}^n\, ,\, Mx = 0\} \tag{338}$$

is non-trivial, i. e. its dimension is ≥ 1. The *range* of M is defined by

$$\mathcal{R}(M) = \{Mx\, ,\, x \in \mathbf{C}^n\}\,. \tag{339}$$

Both $\mathrm{Ker}(M)$ and $\mathcal{R}(M)$ are linear subspaces of $\mathbf{C}^n$, and by one of the fundamental theorems of Linear Algebra[53] we know that

$$\dim \mathrm{Ker}(M) + \dim \mathcal{R}(M) = n\,. \tag{340}$$

Let us first assume that

$$\mathrm{Ker}(M) \cap \mathcal{R}(M) = \{0\}\,. \tag{341}$$

Due to (340), this is the same as saying

$$\mathbf{C}^n = \mathrm{Ker}(M) \oplus \mathcal{R}(M)\,. \tag{342}$$

[53] See the textbook, Proposition 3.6 on page 90

We note that from the definitions it is obvious that both $\mathrm{Ker}(M)$ and $\mathcal{R}(M)$ are invariant under the mapping M, and therefore also under the mapping A. As M vanishes on $\mathrm{Ker}(M)$, we know that that the restriction of A to $\mathrm{Ker}(M)$ is λI_r, where I_r is the $r \times r$ identity matrix and $r = \dim \mathrm{Ker}(M)$. Hence we obtain

$$A = \lambda I_r \oplus A' \,, \tag{343}$$

where A' is an $(n-r)\times(n-r)$ matrix. We can now repeat the whole procedure with A' and try to decompose A in several steps as

$$A = \lambda_1 I_{r_1} \oplus \lambda_2 I_{r_2} \oplus \dots \oplus \lambda_l I_{r_l} \,. \tag{344}$$

This almost works, except that the conditions of the type (341) which we need at each inductive step may not be always satisfied. Two subspaces of $\mathbf{C}^n$ with the sum of their dimensions equal to n will "generically" intersect only at the origin, but we would like also to cover the "non-generic" situations when the intersection is bigger.

Let us then drop the assumption (341) and assume that $\mathrm{Ker}(M)$ and $\mathcal{R}(M)$ can intersect in a non-trivial subspace.

The key step in overcoming the difficulties when the intersection $\mathrm{Ker}(M) \cap \mathcal{R}(M)$ is non-trivial is the following: we consider the subspaces

$$L_j = \mathrm{Ker}(M^j), \qquad Y_j = \mathcal{R}(M^j) \,, \qquad j = 1, 2, \dots \tag{345}$$

Note that all these subspaces are invariant under M. Clearly

$$L_1 \subset L_2 \subset L_3 \subset \dots \tag{346}$$

and

$$Y_1 \supset Y_2 \supset Y_3 \supset \dots \tag{347}$$

As $0 \le \dim Y_j \le n$ for each j, there exists a first j such that

$$Y_j = Y_{j+1} \,. \tag{348}$$

Let us denote this first j by j_0 in what follows. We note that (348) means that the restriction of M to Y_j is invertible, and therefore of we will have

$$Y_{j_0} = Y_{j_0+1} = Y_{j_0+2} = \dots \tag{349}$$

and using

$$\mathrm{Ker}(M^j) + \mathcal{R}(M^j) = n \tag{350}$$

we see that also

$$L_{j_0} = L_{j_0+1} = L_{j_0+2} = \dots \tag{351}$$

The key point is the following:

Lemma 3.

$$\mathbf{C}^n = L_{j_0} \oplus Y_{j_0}\,. \tag{352}$$

Proof: Let $N = M^{j_0}, L = L_{j_0}, Y = Y_{j_0}$. From the definitions we have $\mathrm{Ker}(N) = L$ and $\mathcal{R}(N) = Y$. We recall that M restricted to Y is invertible, and hence so is $N = M^{j_0}$. In particular, if $x \in Y$ and $Nx = 0$, then $x = 0$. Our goal is to show that $L \cap Y = \{0\}$. Arguing by contradiction, assume that there exists a non-zero vector $y \in L \cap Y$. Then $Ny = 0$ because $y \in L$ but as $y \in Y$ the equation $Ny = 0$ means that $y = 0$ as we have seen above. This gives the required contradiction and the proof is finished.

The space L_{j_0} is called the *generalized eigenspace* of the eigenvalue λ. It is the set of all vectors $x \in \mathbf{C}^n$ such that

$$(A - \lambda I)^k x = 0 \tag{353}$$

for some natural number k. As we have seen above, this space coincides with $\mathrm{Ker}\,(A - \lambda I)^{j_0}$.

We can now apply the idea mentioned above concerning the inductive decomposition of $\mathbf{C}^n$ into the eigenspaces. For the eigenspaces it did not quite work in the general situation because condition (341) might fail. However, we see from Lemma 3 that the idea will work in the general case if we replace the eigenspaces with generalized eigenspaces.
Hence we reach the following conclusion.

Lemma 4. *Let A be a $n \times n$ matrix. Let $\lambda_1, \dots, \lambda_r$ be the set of eigenvalues of A. Let $E_{\mathrm{gen}}(\lambda_j)$ be the generalized eigenspace corresponding to λ_j. Then*

$$\mathbf{C}^n = E_{\mathrm{gen}}(\lambda_1) \oplus E_{\mathrm{gen}}(\lambda_2) \oplus \cdots \oplus E_{\mathrm{gen}}(\lambda_r)\,. \tag{354}$$

To complete the proof of Theorem 5, we only need to analyze the restriction of $A - \lambda_j I$ to the space $E_{\mathrm{gen}}(\lambda_j)$. This amounts to analyzing the so-called *nilpotent matrices*, i. e. the matrices M with $M^l = 0$ for some l. What we need to show is that each nilpotent linear mapping is a direct sum of mappings of type $J_k(0)$ considered above. This is easy to see in the case when $\dim \mathrm{Ker}(M) = 1$. Indeed, let M be an $m \times m$ matrix with $M^l = 0$ and $\dim \mathrm{Ker}(M) = 1$. We note that the dimension of the subspaces in the sequence

$$\mathcal{R}(M) \supset \mathcal{R}(M^2) \supset \mathcal{R}(M^3), \dots \tag{355}$$

can drop at most by 1, as $\dim \mathrm{Ker}(M) = 1$. Hence we have $M^k \neq 0$ for $k < m$. At the same time, if $M^m \neq 0$, the sequence would stabilize before reaching $\{0\}$ and the matrix would not be nilpotent. Hence the minimal l with $M^l = 0$ is $l = m$ (again, in the case $\mathrm{Ker}(M)$ is one-dimensional). If we choose a vector x such that $M^{m-1}x \neq 0$, then the vectors

$$x, Mx, M^2x, \dots, M^{m-1}x \tag{356}$$

are easily seen to form a basis in which the matrix M has the form (334) with $\lambda = 0$.

The general case is quite similar in spirit, except the proof requires some more "bookkeeping". You can consult the textbook, page 141.

Lecture 20, 3/10/2013

The matrix exponential
We discussed e^{tA} for matrices, see Section 3.1 in the textbook.

Lecture 21, 3/13/2013
Midterm 1

Lecture 22, 3/15/2013
Hermitian and anti-hermitian matrices, Section 2.11.

Lecture 23, 3/25/2013
General solutions for systems with multiple eigenvalues; convergence of solutions for $t \to \infty$.

Let us consider the system

$$x' = Ax\,, \tag{357}$$

where $x = (x_1, \dots, x_n)$ and A is a $n \times n$ matrix. (Strictly speaking, we should write x as a column vector,

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{pmatrix}\,, \tag{358}$$

but we will sometimes slightly abuse notation and write it as a row vector, if there is no danger of confusion.)[54]
One way to write the general solution is the following:

$$x(t) = e^{tA}x^{(0)}, \quad x^{(0)} \in \mathbf{C}^n\,. \tag{359}$$

This is the same as writing

$$x(t) = C_1 e^{tA} e_1 + \dots C_n e^{tA} e_n\ , \tag{360}$$

[54] In the orthodox notation, the row vectors are really co-vectors, i. e. , linear functionals or elements of the dual space to the space of the vectors. There are situations where it can be important to make carefully this distinction, but in our case it is not necessary.

where $e_1, \dots, e_n$ denotes the canonical basis of $\mathbf{C}^n$, and $C_1, \dots, C_n$ are constants. (Comparing (359) and (360) we see that $C_1 = x_1^{(0)}, C_2 = x_2^{(0)}, \dots, C_n = x_n^{(0)}$.)

Assume now that we chose our coordinates so that A is already in its Jordan canonical form, see Theorem 5. Of course, typically the matrix of a system we deal with is not in the Jordan canonical form, we have to perform a change of coordinated to bring it to that form. However, we know that such a change of coordinates always exists, and hence for theoretical considerations we can assume that A is in such form, keeping in mind that x may not be the original variable in which the system is given, but a new variable given by $x = P^{-1}\tilde{x}$, where P is a transition matrix and $\tilde{x}$ is the original variable. We recall that for one Jordan block

$$J_k(\lambda) = \begin{pmatrix} \lambda & 1 & 0 & \dots & \dots & 0 \\ 0 & \lambda & 1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & 0 & \lambda & 1 & 0 \\ \dots & \dots & \dots & 0 & \lambda & 1 \\ 0 & 0 & 0 & \dots & 0 & \lambda \end{pmatrix} \tag{361}$$

the exponential is

$$e^{tJ_k(\lambda)} = \begin{pmatrix} e^{\lambda t} & te^{\lambda t} & \frac{t^2}{2!}e^{\lambda t} & \dots & \dots & \frac{t^{k-1}}{(k-1)!}e^{\lambda t} \\ 0 & e^{\lambda t} & te^{\lambda t} & \frac{t^2}{2!}e^{\lambda t} & \dots & \frac{t^{k-2}}{(k-2)!}e^{\lambda t} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & 0 & e^{\lambda t} & te^{\lambda t} & \frac{t^2}{2!}e^{\lambda t} \\ \dots & \dots & \dots & 0 & e^{\lambda t} & te^{\lambda t} \\ 0 & 0 & 0 & \dots & 0 & e^{\lambda t} \end{pmatrix} \tag{362}$$

Assuming that $J_k(\lambda)$ is the first Jordan block of A in our coordinate system, we see that the part of the expression (360) corresponding to this block is

$$\begin{aligned} &\left(C_1 e^{\lambda t} + C_2 t e^{\lambda t} + C_3 \frac{t^2}{2!} e^{\lambda t} + \dots C_k \frac{t^{k-1}}{(k-1)!} e^{\lambda t} \right) e_1 \\ &\qquad + \left(C_2 e^{\lambda t} + C_3 t e^{\lambda t} + C_3 \frac{t^2}{2!} e^{\lambda t} + \dots C_k \frac{t^{k-2}}{(k-2)!} e^{\lambda t} \right) e_2 \\ &\qquad\qquad + \dots \\ &\qquad\qquad\qquad + C_k e^{\lambda t} e_k \,. \end{aligned} \tag{363}$$

The contributions from other Jordan blocks will be similar and the general solution will be a sum of expression of this form, where the constants C_j are chosen independently for each Jordan block. We emphasize that that, in general, there can be several Jordan blocks $J_{k_1}(\lambda), \dots, J_{k_r}(\lambda)$ corresponding to the same eigenvalue λ. If we work in coordinates where the matrix A is not in the Jordan canonical form, we can still write a similar expression, if we replace the vectors

e_j of the canonical basis by suitably chosen vectors in the invariant subspace of A corresponding to the Jordan block. Namely, we can find linearly independent vectors $x^{(1)}, x^{(2)}, \dots x^{(k)}$ so that

$$(A-\lambda I)x^{(1)}=0, (A-\lambda I)x^{(2)}=x^{(1)}, \dots, (A-\lambda I)x^{(k)}=x^{(k-1)}\,, \tag{364}$$

and the expression

$$\begin{aligned}\left(C_1e^{\lambda t}+C_2te^{\lambda t}+C_3\frac{t^2}{2!}e^{\lambda t}+\dots C_k\frac{t^{k-1}}{(k-1)!}e^{\lambda t}\right)x^{(1)}\\ +\left(C_2e^{\lambda t}+C_3te^{\lambda t}+C_3\frac{t^2}{2!}e^{\lambda t}+\dots C_k\frac{t^{k-2}}{(k-2)!}e^{\lambda t}\right)x^{(2)}\\ +\dots\\ +C_ke^{\lambda t}x^{(k)}\,.\end{aligned} \tag{365}$$

is a part of the general solution. The Jordan theorem guarantees that we can find a basis of $\mathbf{C}^n$ consisting of strings of vectors satisfying (364) (for suitable k and λ depending on the string), and the general solutions can be written as a sum of expressions similar to (365). Of course, finding the basis consisting of such strings is equivalent to finding a coordinate system in which the matrix has the Jordan canonical form.

From these considerations we can deduce the following important theorem. Recall that the *spectrum* $\sigma(A)$ of a matrix A is simply the set of its eigenvalues.

Theorem 6. *Let A be an $n\times n$ matrix and consider the system*

$$x'=Ax\,. \tag{366}$$

The following conditions are equivalent:
(i) For each solutions $x(t)$ of the system we have

$$\lim_{t\to\infty}x(t)=0\,. \tag{367}$$

(ii) The spectrum $\sigma(A)$ is contained in $\{z\in\mathbf{C}, \operatorname{Re}z<0\}$.
If any of these conditions is satisfied, then the decay is in fact exponential: there exists $\varepsilon>0$ such that for each solution $x(t)$ there exists $C>0$ such that

$$|x(t)|\le Ce^{-\varepsilon t}\,. \tag{368}$$

The exponent ε can be taken as any number satisfying $\sigma(A)\subset\{z\in\mathbf{C},\,\operatorname{Re}z<-\varepsilon\}$.

The proof of the theorem follows easily from expression (359) for the general solution and from the fact that

$$\lim_{t\to\infty}t^me^{-\varepsilon t}=0 \tag{369}$$

for any m and any $\varepsilon>0$. If the spectrum σA contains an eigenvalue λ with $\operatorname{Re}\lambda>0$, then the equation has a solution of the form $Cye^{\lambda t}$, which grows exponentially. For each λ in the spectrum with $\operatorname{Re}\lambda = 0$ there is a solution of the form $Cye^{\lambda t}$ which is either a non-zero constant (when $\lambda = 0$ or a non-trivial periodic solution (when $\lambda = i\kappa$ for $\kappa \neq 0$.)

The case of a single equation (275) of order m discussed in lecture 17 is subsumed in the above considerations as we can write the equation in the form (357). In particular, Theorem 6 remains true if we replace $\sigma(A)$ by the set of roots of the characteristic polynomial. A useful fact which is sometimes useful to keep in mind is that the matrix (297) coming from equation (275) cannot have more than one Jordan block corresponding to each eigenvalue.[55]

Lecture 24, 3/27/2013

Inhomogeneous equations $x' = Ax + f(t)$

We discussed material in Section 3.4 (pp. 163–165).

The main points:

- The solution of the system
$$x'(t) = Ax + f(t)\,, \qquad x(0) = x^{(0)} \tag{370}$$
is given by the Duhamel's formula
$$x(t) = e^{tA}x^{(0)} + \int_0^t e^{(t-s)A} f(s)\,ds\,. \tag{371}$$
- In some special cases one can seek the solution in some particular form. For example, if A is real and $f(t) = b\cos\omega t$ for some real vector b, then we can seek a particular solution of the form Re $ze^{i\omega t}$ for a suitable vector z. This leads to
$$i\omega z = Az + b \tag{372}$$
which can be solved if $i\omega$ is not an eigenvalue of A. In that case we have $z = (i\omega - A)^{-1}b$.

We also discussed in some detail the following example:
$$\begin{array}{rcl} \dot{x}_1 & = & -px_1 + \tilde{q}x_2 + f_1(t)\,, \\ \dot{x}_2 & = & \tilde{p}x_1 - qx_2 + f_2(t)\,, \end{array} \tag{373}$$
where $0 < \tilde{p} \leq p < 1$ and $0 < \tilde{q} \leq q < 1$. In this model x_1 represents can be thought of as a number of inhabitants of place (1) and x_2 can be thought of as a number of inhabitants of place (2). The inhabitants move between (1) and (2) according to the following rules: during dt, an infinitesimal time change, a portion $p\,dt$ of the inhabitants of leave (1) and $\tilde{p}\,dt$ of those move to (2),

[55] As an optional exercise, you can try to prove this statement.

while the portion $(p-\tilde{p})\,dt$ leaves our system. Similar consideration apply to place (2) and $q\,dt$, $\tilde{q}\,dt$, $(q-\tilde{q})\,dt$. The right-hand side f can be thought of as representing influx of new inhabitants. (In this interpretation we should take $f_1, f_2 > 0$.) When f is independent of time and one of the numbers $p-\tilde{p}, q-\tilde{q}$ is strictly positive, the system converges as $t \to \infty$ to the equilibrium solution given by $\dot{x} = 0$. If P is the matrix of the system, the equilibrium solution is $x = -P^{-1}f$.[56]

Lecture 25, 3/29/2013

Inhomogeneous scalar equations of higher order. Minimal polynomial of a matrix; Caley-Hamilton theorem

Equation

$$x^{(m)} + a_1 x^{(m-1)} + a_2 x^{(m-2)} + \cdots + a_{m-1}x' + a_m x = f(t) \tag{376}$$

with constant coefficients a_j can re-written as a system of equations and the Duhamel's formula (371) can be applied to obtain solutions. One can also generalize the procedure for the second order we considered in lecture 10 (see also Section 1.14 in the textbook). Let $\phi_1, \dots, \phi_m$ be a basis of the space of solutions of the equation. We will search solutions of (376) in the form

$$x(t) = C_1(t)\phi_1(t) + \cdots + C_m(t)\phi_m(t)\,. \tag{377}$$

We will impose the conditions

$$\begin{array}{ccccc} C_1'\phi_1 + & \dots & +C_m'\phi_m & = & 0\,, \\ C_1'\phi_1' + & \dots & +C_m'\phi_m & = & 0\,, \\ \dots & \dots & \dots & \dots & \dots \\ C_1'\phi_1^{(m-2)} + & \dots & +C_m'\phi_m^{(m-2)} & = & 0\,, \end{array} \tag{378}$$

which guarantee

$$x^{(k)} = C_1\phi_1^{(k)} + \cdots + C_m\phi_m^{(k)}\,, \qquad k = 0, 1, \dots, m-1\,. \tag{379}$$

[56] This can be seen directly without applying the Duhamel's formula. On the other hand, the limit of $x(t)$ for $t \to \infty$ is also given by the Duhamel's formula. Comparing the two ways of expressing the limit, we obtain

$$\int_0^\infty e^{tP}\,dt = -P^{-1}\,. \tag{374}$$

More generally, if A is any matrix with spectrum in $z, \operatorname{Re} z > 0$, then

$$\int_0^\infty e^{-tA}\,dt = A^{-1}\,. \tag{375}$$

For 1×1 matrices, which can be identified with numbers, we know the formula from elementary calculus.

Equation (376) then gives

$$C_1'\phi^{(m-1)} + \cdots + C_m'\phi_m^{(m-1)} = f\,. \tag{380}$$

Letting

$$W(t) = \begin{pmatrix} \phi_1 & \dots & \phi_m \\ \phi_1' & \dots & \phi_m' \\ \dots & \dots & \dots \\ \phi_1^{(m-1)} & \dots & \phi_m^{(m-1)} \end{pmatrix} \tag{381}$$

and

$$C' = \begin{pmatrix} C_1' \\ C_2' \\ \dots \\ C_m' \end{pmatrix}, \qquad F(t) = \begin{pmatrix} 0 \\ 0 \\ \dots \\ f(t) \end{pmatrix} \tag{382}$$

we see that

$$W \cdot C' = F\,. \tag{383}$$

Hence to obtain C', we just invert the matrix W to obtain

$$C' = W^{-1}F\,. \tag{384}$$

One then has

$$C(t) = c + \int_0^t C'(s)\,ds\,. \tag{385}$$

where c is a constant vector. It can be proved that in the situation above the matrix $W = W(t)$ is invertible,[57] and hence (384) is well-defined.

In the second part of the lecture we discussed the minimal polynomial and the Caley-Hamilton Theorem (Sections 2.7 and 2.8 in the textbook).

Lecture 26, 4/1/2013

More on the minimal polynomial

We continued to discuss the minimal polynomial. The main points of the lecture:

- The minimal polynomial can be easily read off the Jordan canonical form. Let A is an $n \times n$ matrix and $\{\lambda_1, \dots, \lambda_r\}$ be its spectrum (i. e. , the set of its eigenvalues, with $\lambda_i \neq \lambda_j$ for $i \neq j$). By Theorem 5 (lecture 19) we know that in a suitable basis the matrix A consists of Jordan blocks $J_k(\lambda)$, where $\lambda \in \{\lambda_1, \dots, \lambda_r\}$ and k is an integer which can change from one Jordan block to another. We also keep in mind that there can be more than one Jordan block associated with any given λ_j. We define k_j as the size of the largest Jordan block associated with λ_j. (By definition, the

[57] The proof is not difficult and it is a good optional exercise.

size of the block $J_k(\lambda)$ is k.) Then the minimal polynomial of the matrix A is

$$p_{\min,A}(\lambda) = (\lambda - \lambda_1)^{k_1}(\lambda - \lambda_2)^{k_2} \dots (\lambda - \lambda_r)^{k_r} \,. \tag{386}$$

Note that this means that the matrix is diagonalizable if and only if its minimal polynomial is $(\lambda - \lambda_1)(\lambda - \lambda_2)\dots(\lambda - \lambda_r)$, or, in other words, $(A - \lambda_1 I)(A - \lambda_2 I)\dots(A - \lambda_r I) = 0$.

- The minimal polynomial always divides the characteristic polynomial. For a generic $n \times n$ matrix the two polynomials coincide, i. e. , the coefficients of the matrices for which the two polynomial do not coincide must satisfy some non-trivial equations $F_j(a_{11}, a_{12}, \dots, a_{nn}) = 0$.

- Genericity can be considered at various levels. For example, a generic $n \times n$ matrix has n different eigenvalues. The matrices which have multiple eigenvalues are characterized by a single (polynomial) equation $F_1(a_{11}, a_{12}, \dots, a_{nn}) = 0$. Let us denote by Σ the surface determined by this equation in the space of the $n \times n$ matrices. The minimal polynomial of matrices away from Σ clearly coincides with their characteristic polynomial. However, even on the surface Σ the fact that the minimal polynomial coincides with the characteristic polynomial is still generic. In other words, for the minimal polynomial not to coincide with the characteristic polynomial additional relations have to be satisfied.

(Optional) **Remark:** It is important to note that there are algorithms for the calculation of the minimal polynomial which are not based on the calculation of the Jordan form, or the roots of the characteristic polynomial. The minimal polynomial of a matrix A can be calculated by relatively simple operations directly on the coefficients of the matrix A.[58] One way to see this is the following. For a fixed matrix A let us consider the (matrix-valued) function $R(\lambda) = R(\lambda, A)$ of the complex variable λ defined by $R(\lambda) = (A - \lambda I)^{-1}$. The function $R(\lambda)$, called the *resolvent* of the matrix A, is well defined for $\lambda \in \mathbf{C} \setminus \{\lambda_1, \dots, \lambda_r\}$, where $\lambda_1, \dots, \lambda_r$ are the eigenvalues of A. We have $R(\lambda) = \frac{\mathrm{Adj}\,(A-\lambda I)}{\det(A-\lambda I)}$, where the matrix $\mathrm{Adj}\,(A - \lambda I)$ the so-called adjugate matrix[59] of the matrix $A - \lambda I$, whose entries are $(n-1) \times (n-1)$ sub-determinants of $A - \lambda I$, which are polynomials of order at most $n - 1$ in λ. We see that $R(\lambda)$ is a rational function of λ, i. e. each entry can be written as a ratio of two polynomial in λ. It is not hard to see[60] that the minimal polynomial is the polynomial $P(\lambda)$ of the minimal possible degree such that $P(\lambda)R(\lambda)$ is a polynomial (for each entry). Therefore finding the minimal polynomial can be reduced to the following task: given a ratio of two polynomials $F(\lambda) = \frac{Q_1(\lambda)}{Q_2(\lambda)}$ what is the polynomial $P(\lambda)$ with minimal possible degree such that $P(\lambda)F(\lambda)$ is a polynomial? This is essentially the same problem as the problem of finding the greatest common divisor of the two

[58] Programs such as Mathematica or Matlab presumably use exactly such algorithms to calculate the minimal polynomial.

[59] see, for example, `http://en.wikipedia.org/wiki/Adjugate_matrix`

[60] e.g., by considering the Jordan form of A

polynomials Q_1 and Q_2. This can be done by the so-called Euclidean algorithm, by simple manipulation of the two polynomials, we do not have to calculate the roots.[61] Applying this procedure to each entry of the matrix $R(\lambda)$, we can in principle find the minimal polynomial of the matrix A by simple operations on its coefficients, without the need to find the eigenvalues of A. This may not be a very efficient algorithm, but it shows that such algorithms exist. For a deeper account of the algebra behind similar considerations you can consult for example the textbook *Algebra* by S. Mac Lane and G. Bikhoff (Chapter X.8), or, for a less abtract account, the textbook *The Theory of Matrices* by F. R. Gantmacher.

As we already mentioned in lecture 25, in our textbook the minimal polynomial is discussed in Section 2.7, where it is used for a construction of generalized eigenspaces via a different method than the one we used in lecture 19.

Lecture 27, 4/3/2013

Real matrices with complex eigenvalues/eigenvectors.

Let us consider a real $n \times n$ matrix A. Let $\lambda_1, \ldots, \lambda_n$ be its eigenvalues. For the rest of the lecture we will assume that all the eigenvalues are different,[62] so that the matrix is diagonalizable when considered as a map $\mathbf{C}^n \to \mathbf{C}^n$. However, over the real numbers the matrix may not be diagonalizable, as one can see from the simple example

$$J = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} . \tag{387}$$

We wish to find the simplest from in which the matrix A can be presented (one a suitable basis) over the real numbers.

Let $a = a_1 + ia_2 \in \mathbf{C}$. The map $z \to az$ can be considered as a map of the 1d complex space $\mathbf{C}$ into itself, and in this interpretation it is represented by the 1×1 matrix (a). Instead of considering $\mathbf{C}$ as a 1d vector space over $\mathbf{C}$, we can consider it as a 2d vector space over $\mathbf{R}$. We can take the basis of this vector space to be 1 and i, and the usual decomposition $z = x_1 + ix_2$ coincides with the representation of z in this basis. In the real coordinates x_1, x_2 the mapping $z \to az$ is represented by the real 2×2 matrix

$$A = \begin{pmatrix} a_1 & -a_2 \\ a_2 & a_1 \end{pmatrix} . \tag{388}$$

Note that the matrix A can be written as

$$A = a_1 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + a_2 \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} = a_1 I + a_2 J \, , \tag{389}$$

[61] see, for example, `http://en.wikipedia.org/wiki/Euclidean_algorithm`

[62] We recall that this is the “generic case”, see lecture 14

corresponding to the decomposition $a = a_1 + ia_2$. The matrices of this type form a (real) algebra and the map $a \to \phi(a) = a_1 I + a_2 J$ is an isomorphism of algebras (over the real numbers) .[63]

Going back to a general real $n \times n$ matrix A above (recall that all eigenvalues of A are assumed to be different) let us now consider a complex eigenvalue λ of A and the corresponding eigenvector z. We assume that λ is not real, e. i. $\overline{\lambda} \neq \lambda$, where the bar denotes the complex conjugate.[64] We have

$$Az = \lambda z, \qquad A\overline{z} = \overline{\lambda}\overline{z}\,. \tag{390}$$

The last equation is obtained by simply taking the complex conjugate of the first and taking into account that $\overline{A} = A$, as A is real. This also reflects the fact that the complex eigenvalues of A come in complex-conjugate pairs $\lambda, \overline{\lambda}$. Let us write

$$z = x + iy, \qquad x, y \in \mathbf{R}^n\,. \tag{391}$$

This is the same as

$$x = \frac{z + \overline{z}}{2}, \qquad y = \frac{z - \overline{z}}{2i}\,. \tag{392}$$

As an optional exercise you can show that the vectors $x, y \in \mathbf{R}^n$ are both non-zero (assuming, as above, $\overline{\lambda} \neq \lambda$) and, moreover, linearly independent over $\mathbf{R}$. Let us denote by V the two dimensional subspace spanned by x, y (over the reals), i. e. $V = \{sx + ty,\, s, t \in \mathbf{R}\} \subset \mathbf{R}^n$. We claim that the subspace V in invariant under A, i. e. , $A(V) \subset V$. To see that, we write $\lambda = \operatorname{Re}\lambda + i \operatorname{Im}\lambda,\ \lambda_j \in \mathbf{R}$, and calculate

$$Ax = \frac{Az + A\overline{z}}{2} = \frac{\lambda z + \overline{\lambda}\overline{z}}{2} = \operatorname{Re}(\lambda z) = \operatorname{Re}(\lambda)x - \operatorname{Im}(\lambda)y \tag{393}$$

and, in a similar way,

$$Ay = \operatorname{Im}(Az) = \operatorname{Re}(\lambda)y + \operatorname{Im}(\lambda)x\,. \tag{394}$$

This shows that $A(V) \subset V$. Moreover, the matrix of the map given by A restricted to V is

$$\begin{pmatrix} \operatorname{Re}\lambda & \operatorname{Im}\lambda \\ -\operatorname{Im}\lambda & \operatorname{Re}\lambda \end{pmatrix} \tag{395}$$

This means that, in suitable coordinates, the restriction of A to V is nothing but the real form of the map given by the multiplication by $\overline{\lambda}$, see the discussion before (388).

Applying the procedure to each complex eigenvalue/eigenvector $\lambda_j, z_{(j)}$ of our matrix A (which – we recall – is assumed to be "generic"), we obtain vectors $x_{(j)}, y_{(j)}$ generating two-dimensional subspaces $V_{(j)}$ which are invariant under A, with $V_{(j)} \cap V_{(k)} = \{0\}$ for $j \neq k$. (As an optional exercise you can verify the last

[63] This means that $\phi(sa + tb) = s\phi(a) + t\phi(b)$ for $s, t \in \mathbf{R}$ and $\phi(ab) = \phi(a)\phi(b)$.

[64] If $z = x_1 + ix_2$, then $\overline{z} = x_1 - ix_2$.

condition.) In addition, we may have one-dimensional invariant subspaces W_k associated with the real eigenvalues. In summary, we obtain a decomposition

$$\mathbf{R}^n = V_{(1)} \oplus V_{(2)} \oplus \dots V_{(r)} \oplus W_{(1)} \oplus \dots W_{(s)} \tag{396}$$

with $\dim V_{(j)} = 2$ and $\dim W_{(k)} = 1$, the subspaces being invariant under A. The restriction of A to $V_{(j)}$ is given by matrices of the form (395), with λ replaced by λ_j. the restriction of A to $W_{(k)}$ is given by multiplication by the corresponding real eigenvalue.

To summarize, if $\lambda_1, \overline{\lambda}_1, \lambda_2, \overline{\lambda}_2, \dots, \lambda_r, \overline{\lambda}_r$ are the complex eigenvalues of A and $\lambda_{2r+1}, \dots, \lambda_n$ are the real eigenvalues (recall that we assume they are all different), then there is a basis in which the matrix can be written as

$$\begin{pmatrix}
\operatorname{Re}\lambda_1 & \operatorname{Im}\lambda_1 & 0 & 0 & \dots & \dots & 0 & 0 & 0 & \dots & 0 \\
-\operatorname{Im}\lambda_1 & \operatorname{Re}\lambda_1 & 0 & 0 & \dots & \dots & 0 & 0 & 0 & \dots & 0 \\
0 & 0 & \operatorname{Re}\lambda_2 & \operatorname{Im}\lambda_2 & \dots & \dots & 0 & 0 & 0 & \dots & 0 \\
0 & 0 & -\operatorname{Im}\lambda_2 & \operatorname{Re}\lambda_2 & \dots & \dots & 0 & 0 & 0 & \dots & 0 \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\
0 & 0 & 0 & \dots & \dots & \operatorname{Re}\lambda_r & \operatorname{Im}\lambda_r & 0 & 0 & \dots & 0 \\
0 & 0 & 0 & \dots & \dots & -\operatorname{Im}\lambda_r & \operatorname{Re}\lambda_r & 0 & 0 & \dots & 0 \\
0 & 0 & 0 & \dots & \dots & 0 & 0 & \lambda_{2r+1} & 0 & \dots & 0 \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\
0 & 0 & 0 & \dots & \dots & 0 & 0 & 0 & 0 & \dots & \lambda_n
\end{pmatrix}. \tag{397}$$

As an example, one can consider an orthogonal 3×3 matrix Q with positive determinant (which then has to be equal to 1, as the matrix preserves the volume in $\mathbf{R}^3$). It is a good exercise to show that such a matrix always has 1 as an eigenvalue. If $x_{(1)}$ is the corresponding eigenvector, the matrix Q represents a rotation about axis $x_{(1)}$. If we choose a positively oriented orthogonal coordinate system so that the x_3 axis coincides with the axis of rotation, the matrix Q will have the form

$$\begin{pmatrix}
\cos\phi & -\sin\phi & 0 \\
\sin\phi & \cos\phi & 0 \\
0 & 0 & 1
\end{pmatrix}, \tag{398}$$

where ϕ is the angle of rotation by about the axis. In this case the invariant space V constructed above coincides with the orthogonal complement of the axis of rotation.

In dimension $n = 4$ an orthogonal matrix Q with positive determinant can be represented in a suitable orthogonal positively oriented basis as

$$\begin{pmatrix}
\cos\phi_1 & -\sin\phi_1 & 0 & 0 \\
\sin\phi_1 & \cos\phi_1 & 0 & 0 \\
0 & 0 & \cos\phi_2 & -\sin\phi_2 \\
0 & 0 & \sin\phi_2 & \cos\phi_2
\end{pmatrix}. \tag{399}$$

Matrices withe nontrivial Jordan blocks over $\mathbf{C}$ can be considered along analogous lines, with the canonical real forms becoming more complicated.

Lecture 28, 4/5/2013

Linear equations with variable coefficients

We started discussing systems

$$x' = A(t)x\,, \tag{400}$$

as in Section 3.8., p. 186 in the textbook. We introduced the map $S(t,t_0)$ (see formula (8.2) in the textbook), discussed the matrix $M(t)$ (formula (8.5)), and Proposition 8.1.

Lecture 29, 4/8/2013

Linear equations with variable coefficients (continued)

We discussed some of the properties of the systems (400) which are similar to the case of the constant coefficients. In particular, the set of all solutions forms a linear space of dimension n (assuming $A(t)$ is $n \times n$ matrix). (By contrast with the constant coefficient case, this time the space of solution is not invariant under the translations $x(t) \to x(t-t_0)$.) The Duhamel's formula remains quite similar, in a suitable interpretation, and we discussed its version (8.19) in the textbook.
For typical 2×2 systems with variable coefficients (or second order scalar equations with variable coefficients) the solutions cannot be expressed in terms of elementary functions and the equations lead to the so-called *special functions*. These functions include for example the Bessel functions, the elliptic integrals, the Airy function, the hypergeometric functions, the confluent hypergeometric functions (Kummer's functions), the parabolic cylinder functions, etc. A Google search for any of these terms will take you to the corresponding Wikipedia pages, with a good overview and additional references. In this course we will not discuss these topics, but it is good to know that this is a very well-studied classical subject (going back for almost 300 years), with extensive literature. The subject has also close connections to modern areas, such as group representations.

Lecture 30, 4/11/2013

Linear equations with periodic coefficients; parametric resonance

In this lecture we first discussed some useful general principles and definitions for linear equations with periodic coefficients, and then we applied them to the well-known example of a (linear) pendulum with a periodically varying length.
Let us consider linear systems

$$x' = A(t)x \tag{401}$$

where $A(t)$ is an $n\times n$ matrix which is a periodic function of t. This means that for some $T>0$ we have

$$A(t+T)=A(t) \tag{402}$$

for each t. An important concept in the study of such systems is the *mapping at a period*, defined as follows. For a given t_0 and $a\in\mathbf{C}^n$ let us find the solution of (401) with $x(t_0)=a$ and let us set $Ba=x(t+T)$. As the equation is linear, the map B (which is the mapping at period mentioned above) is linear and hence is given by an $n\times n$ matrix, which we will also denote by B. The matrix B can be identified with the solution operator $S(t_0,t_0+T)$ which we discussed in previous lectures. Due to the periodicity condition (402) we have (still for $x(t_0)=a$)

$$x(t_0+2T)=B^2a,\quad x(t_0+3T)=B^3a,\dots x(t_0+kT)=B^ka\,. \tag{403}$$

The matrix B is invertible (think of running the equation backwards) with

$$B^{-1}=S(t_0+T,t_0)\,. \tag{404}$$

Therefore

$$x(t_0+kT)=B^ka,\qquad k\in\mathbf{Z}. \tag{405}$$

The matrix B can also be obtained in the following way: we find a matrix-valued solution $X(t)$ of (401), i. e.

$$X'(t)=A(t)X(t)\,, \tag{406}$$

with $X(t_0)=I$. Then $B=X(t_0+T)$.
We recall that the trace of a $n\times n$ matrix A is given by $\mathrm{Tr}A=\sum_j a_{jj}$.

Lemma 5. *For any matrix solution $X(t)$ of (406) we have*

$$\frac{d}{dt}\det X=\det X\ \mathrm{Tr}\,A(t)\,, \tag{407}$$

For a proof of the lemma see exercises 1-3 on page 190 in the textbook. The key point is the formula

$$\frac{d}{dt}\det X=\mathrm{Tr}\,[(\mathrm{Adj}\,X)\dot{X}] \tag{408}$$

where $\mathrm{Adj}\,X$ is the so-called *adjugate matrix* of X, which for an invertible matrix X can be identified with $X^{-1}\det X$, see for example
`http://en.wikipedia.org/wiki/Adjugate_matrix`.
Formula (408) (also called Jacobi's formula) can be derived in many ways. Given (408), it is easy to establish (407):

$$\frac{d}{dt}\det X=\det X\,\mathrm{Tr}\,\left(X^{-1}A(t)X\right)=\det X\,\mathrm{Tr}\,A(t)\,, \tag{409}$$

where we have used that $\mathrm{Tr}\,(PAP^{-1})=\mathrm{Tr}\,A$ for each non-singular matrix P.

Lemma 5 enables us to express the determinant of the matrix of the mapping at period as follows:

$$\det B = \exp \int_{t_0}^{t_0+T} \operatorname{Tr} A(t)\; dt\ . \tag{410}$$

In particular, when $\operatorname{Tr} A(t) = 0$ for each t, then

$$\det B = 1\,. \tag{411}$$

The eigenvalues of the matrix B are often called *characteristic multipliers* or *Floquet multipliers.*

In the class of general systems with coefficients with period T there are no restriction on B, except for the condition $\det B > 0$, which follows from Lemma 5. This can be seen as follows: given B with $\det B > 0$, we can find a smooth curve $X(t)$ in the set of invertible matrices such that $X(t_0) = I$ and $X(t_0 + T) = B$. (The verification of this is a good exercise which can be simple or non-trivial, depending on how much one knows about matrices.) One can also assume that $\dot{X} = 0$ in some neighborhoods of t_0 and $t_0 + T$. Letting $A(t) = X(t)^{-1}\dot{X}(t)$ we see from the definitions that $X(t)$ solves (401). The function $A(t)$ can clearly be extended periodically to all $\mathbf{R}$ and we see that B is the mapping at period of the resulting system.

Let us look in more detail at the so-called Hill's equation

$$q'' + a(t)q = 0\,. \tag{412}$$

We assume that $a > 0$ is a periodic function of t with period T. We can think of a (linear) pendulum with length depending on time. We let

$$x_1(t) = q(t), \quad x_2(t) = q'(t) \tag{413}$$

and rewrite the equation as

$$x' = A(t)x\,, \tag{414}$$

with

$$A(t) = \begin{pmatrix} 0 & 1 \\ -a(t) & 0 \end{pmatrix}\ . \tag{415}$$

Note that $\operatorname{Tr} A = 0$ and therefore the matrix B of the mapping at period will satisfy

$$\det B = 1\,. \tag{416}$$

The classical case of a pendulum with nearly (but not exactly) constant length is already interesting. Let us consider

$$a(t) = \omega_0^2 + \varepsilon \cos \omega_1 t\,, \tag{417}$$

where ε is close to 0. Let us denote by B_ε the matrix of the mapping at period for a given ϵ. The matrix B_0 can be easily calculated explicitly:

$$B_0 = \begin{pmatrix} \cos\omega_0 T & \frac{1}{\omega_0}\sin\omega_0 T \\ -\omega_0\sin\omega_0 T & \cos\omega_0 T \end{pmatrix}, \qquad T = \frac{2\pi}{\omega_1}\,. \tag{418}$$

For small ε the matrix B_ε will be close to B_0 and, moreover, $\det B_\varepsilon = 1$. The equation for the eigenvalues of B_ε is

$$\lambda^2 - \lambda\,\mathrm{Tr}\,B_\varepsilon + 1 = 0\,. \tag{419}$$

Writing $\mathrm{Tr}\,B_\varepsilon = 2b_\varepsilon$, we have

$$\lambda_{1,2} = b_\varepsilon \pm \sqrt{b_\varepsilon^2 - 1}\,. \tag{420}$$

For $|b_\varepsilon| < 1$ this represents two complex conjugate roots lying on the unit circle $|\lambda| = 1$. The powers of a matrix with such eigenvalues will stay bounded and the value of the solution at times $t_0 + kT$ (with $k \in \mathbf{Z}$), given by $B_\varepsilon^k a$ (with $a = x(t_0)$, as above) will exhibit an "almost periodic" behavior. We conclude that a sufficiently small periodic perturbation of the length at frequency ω_1 satisfying $-1 < \cos 2\pi\frac{\omega_0}{\omega_1} < 1$ will not have a dramatic effect on the solution. The solution will typically not really be periodic any longer, but will stay bounded and its behavior will be quite regular, in some sense. On the other hand, when $b_0 = \cos 2\pi\frac{\omega_0}{\omega_1} = \pm 1$ any small perturbation of $b_0 = 1$ to b_ε with $|b_\varepsilon| > 1$ will completely change the long-time behavior: once $|b_\varepsilon| > 1$, we will have a real eigenvalue λ of B_ε with $|\lambda| > 1$ and another real eigenvalue $\tilde\lambda$ with $|\tilde\lambda| < 1$ and there will be no bounded solutions in $\mathbf{R}$. The condition $b_0 = \pm 1$ corresponds to

$$\frac{\omega_0}{\omega_1} = \pm\frac{1}{2},\ \pm 1,\ \pm\frac{3}{2}, \dots \tag{421}$$

If the frequency ω_1 of the perturbation is close to the values where one of these relations is satisfied, we can expect that even a small forcing at frequency ω_1 can drastically change the long-time behavior of the solutions. This is called *parametric resonance*. Everyone who has been on a swing has a practical experience with it.

Lecture 31, 4/12/2013

Existence theorems

We started to discuss the existence theorem in Section 4.1.

Lecture 32, 4/15/2013

Existence theorems (continued)

We continued to discuss the existence theorem in Section 4.1.

4/17/2013
Midterm 2

Lecture 33, 4/19/2013
We discussed issues concerning polynomial equations (including those which appeared in Midterm 2).

Lecture 34, 4/22/2013
Existence Theorem; Nonlinear systems

We finished the proof of Proposition 1.1 (p.226) in the textbook. Proposition 1.2 in the book was also briefly discussed, although we did not go into too much details.
We started discussing nonlinear systems of the form

$$x' = f(x,t) \tag{422}$$

starting with the autonomous case $f(x,t) = f(x)$. In (422) we the unknown function $x = x(t)$ is a function from an interval $I = (t_1, t_2)$ into $\mathbf{R}^n$. Although f can be formally described as a function $f\colon \mathbf{R}^n \times I \to \mathbf{R}^n$, it is better to think of it as a vector field depending on the time t.
Let us first look at the case when f is independent of time, i. e. , $f = f(x)$, with no dependence on t. Equation (422) then is

$$x' = f(x)\,. \tag{423}$$

Here we can think of $f(x)$ as an arrow at x which tells us what is the velocity of a trajectory when it passes through the point x. This is an informal definition of a vector field. It is useful to note how system (423) changes when we change the coordinate x. For simplicity we will consider only linear changes of coordinates

$$x = Py\,, \tag{424}$$

where P is a non-singular matrix. Replacing x by Py in (423) we obtain

$$Py' = f(Py)\,, \tag{425}$$

which is the same as

$$y' = P^{-1}f(Py)\,. \tag{426}$$

We see that under the change of coordinates (424) the right-hand side of (423) should be transformed to

$$g(y) = P^{-1}f(Py)\,. \tag{427}$$

Such "transformation rule" under coordinate changes characterizes vector fields. Note that a natural transformation of a scalar function under (425) is

$$(x \to \phi(x)) \to (y \to \phi(Py)) \tag{428}$$

and it would look the same for an $\mathbf{R}^n$ valued function. On the other hand, the transformation of vector fields, given by (427), is different by the factor P^{-1}. This factor exactly describes how the coordinates of the "arrow" described by $f(x)$ change if we use the coordinates y instead of x.

For the autonomous equations (423) one can draw phase portraits which generalize those we were drawing in Lecture 4 (such as fig. 3 and fig. 4 in lecture 4). You can take a look at the phase portraits on figs. 3.2 – 3.6 in the textbook.

Lecture 35, 4/24/2013

Gradient flows

Systems of the form

$$x' = f(x) \tag{429}$$

in dimension $n \geq 2$ can rarely be solved in terms of elementary functions, or integrals of elementary functions. We saw that the same is true for linear systems with time-dependent coefficients

$$x' = A(t)x \tag{430}$$

when $n \geq 2$. When we were dealing with equations with constant coefficients, we might have got used to being able to write down solutions in terms of quite explicit formulae. This is no longer possible (except in some non-typical cases) when dealing with general systems (429) or (430), and instead of trying to get explicit formulae for the solutions, it is often best to try to understand the behavior of the solutions on a qualitative level, without writing down specific formulae.

In the lecture we discussed a special class of non-linear systems of the form (429), the so-called *gradient flows.* These are flows for which the vector field f is given by

$$f(x) = -\nabla\phi(x)\,, \tag{431}$$

where $\phi\colon \mathbf{R}^n \to \mathbf{R}$ is a scalar function. At a heuristic level the equation

$$x' = -\nabla\phi(x) \tag{432}$$

describes a motion in the direction of the steepest descent of the function ϕ. In particular, along each non-trivial trajectory the function $t \to \phi(x(t))$ is decreasing, as one can see from

$$\frac{d}{dt}\phi(x(t)) = (\nabla\phi(x(t)), x'(t)) = -|\nabla\phi(x)|^2\,. \tag{433}$$

If the function ϕ satisfies

$$\lim_{|x|\to\infty} \phi(x) = +\infty \tag{434}$$

and has exactly one critical point[65] $\overline{x}$, then it is easy to see that

$$\lim_{t\to\infty} x(t) = \overline{x} \tag{435}$$

for every solution of the system

$$x' = -\nabla\phi(x)\,. \tag{436}$$

This statement is heuristically obvious, but it is a good (optional) exercise to try to prove it rigorously. You may not find it trivial if you are encountering this type of statement for the first time.[66]

If ϕ has several critical points, the behavior of the solutions can be more interesting. We discussed the example where

$$\phi\colon \mathbf{R}^2 \to \mathbf{R} \tag{437}$$

was given by

$$\phi(x_1,x_2) = \frac{1}{4}\left(x_1^2+x_2^2-1\right)^2 + x_2^2\,. \tag{438}$$

We note that

$$\lim_{|x|\to\infty} \phi(x) = \infty\,. \tag{439}$$

Let us calculate the critical points of ϕ. The equation

$$\nabla\phi(x) = 0 \tag{440}$$

becomes

$$(x_1^2+x_2^2-1)x_1 = 0\,, \qquad (x_1^2+x_2^2-1)x_2 + 2x_2 = 0\,. \tag{441}$$

These represent two equations for two unknowns, and it is not hard to find solutions: they are

$$x^{(1)} = (-1,0)\,, \quad x^{(2)} = (0,0)\,, \quad x^{(3)} = (1,0)\,. \tag{442}$$

We have

$$\phi(x^{(1)}) = \phi(x^{(3)}) = 0\,, \quad \phi(x^{(2)}) = \frac{1}{4}\,. \tag{443}$$

As $\phi \geq 0$ in $\mathbf{R}^2$, the functions ϕ attains its minimum at $x^{(1)}$ and $x^{(3)}$. On the other hand, the point $x^{(2)}$ is a *saddle point*, neither a local maximum nor a local minimum of ϕ. This is easily seen from the expansion

$$\phi(x^{(2)}+\xi) = \phi(\xi) = \frac{1}{4} - \frac{1}{2}\xi_1^2 + \frac{1}{2}\xi_2^2 + O(|\xi|^4)\,. \tag{444}$$

The trajectories of the equation

$$x' = -\nabla\phi(x) \tag{445}$$

[65] recall that a *critical point* of ϕ is any point x where $\nabla\phi(x) = 0$.

[66] Hint: show that $\frac{d}{dt}\phi(x(t)) < 0$ and that for any $\varepsilon > 0$ the trajectory will reach the set $\{\phi(x) < \phi(\overline{x}) + \varepsilon\}$ in finite time.

We note the symmetries

$$\phi(x_1, x_2) = \phi(\pm x_1, \pm x_2)\,. \tag{446}$$

From the properties of the function ϕ it is not hard to see that for the solutions $x = x(t)$ with the initial condition $x(0)$ we can expect

$$x_1(0) < 0 \quad \Longrightarrow \quad \lim_{t\to\infty} x(t) = x^{(1)}\,, \tag{447}$$

$$x_1(0) = 0 \quad \Longrightarrow \quad \lim_{t\to\infty} x(t) = x^{(2)}\,, \tag{448}$$

$$x_1(0) > 0 \quad \Longrightarrow \quad \lim_{t\to\infty} x(t) = x^{(3)}\,. \tag{449}$$

Establishing this completely rigorously requires some work, and the reader is encouraged to do the proof as an optional exercise.

Lecture 36, 4/26/2013

Linearization at equilibria

In this lecture we discussed the linearization of a general system in $\mathbf{R}^n$

$$x' = f(x) \tag{450}$$

at an equilibrium[67] $\overline{x}$. (Recall that $\overline{x}$ is called an equilibrium if $f(\overline{x}) = 0$.) In the textbook this is discussed in Section 4.3, starting on page 247. The key point that when studying the behavior of solutions which are close to $\overline{x}$, one often obtains a good idea about the behavior of the solutions by *linearization* of the field $f(x)$ at $\overline{x}$. Writing

$$x = \overline{x} + \xi\,, \tag{451}$$

where ξ is assumed to be small we can write

$$x' = (\overline{x}+\xi)' = \xi' = f(\overline{x}+\xi) = f(\overline{x}) + f'(\overline{x})\cdot\xi + O\left(|\xi|^2\right) = f'(\overline{x})\cdot\xi + O(|\xi|^2)\,. \tag{452}$$

Letting $A = f'(\overline{x})$, we see that for small ξ the leading part of the equation for ξ is

$$\xi' = A\xi\,, \tag{453}$$

which is an equation with constant coefficients. This way the equations with constant coefficients appear in the analysis of non-linear problems. In the lecture we calculated the matrix A for the gradient system we studied last time. If all the eigenvalues of A lie in the half-plane

$$\{z \in \mathbf{C}\,,\ \operatorname{Re} z < 0\}\,, \tag{454}$$

then all the solutions of the linearized system approach 0 as $t \to \infty$, as we have seen in Theorem 6. Importantly, this conclusion (in its natural modification)

[67] Also called a *critical point* or a "rest point"

is also true for the behavior of the non-linear system: *if the real parts of all the eigenvalues of the matrix $f'(\overline{x})$ are strictly negative, then any solution of the system starting in a sufficiently small neighborhood of $\overline{x}$ converges to $\overline{x}$ as $t \to \infty$.* We will discuss this important result and its generalizations in more detail next time.

Lecture 37, 4/29/2013

Stability

Let us consider an autonomous system

$$x' = f(x) \tag{455}$$

in $\mathbf{R}^n$, and let $\overline{x}$ be an equilibrium of the system, i. e. we have

$$f(\overline{x}) = 0. \tag{456}$$

For simplicity we assume that the function f is smooth, although this assumption can be weakened. We will discuss the notion of *stability* if the equilibrium $\overline{x}$.

We say that $\overline{x}$ is *Lyapunov stable* if for each $\varepsilon > 0$ there exists $\delta > 0$ such that for each $|x_0 - \overline{x}| < \delta$ the trajectory $x(t)$ with $x(0) = x_0$ will satisfy $|x(t) - x_0| < \varepsilon$ for each $t > 0$.

We say that $\overline{x}$ is *asymptotically stable* if there exists $\delta > 0$ such that for each $|x_0 - \overline{x}| < \delta$ the trajectory $x(t)$ with $x(0) = x_0$ satisfies

$$\lim_{t \to \infty} x(t) = \overline{x}\,. \tag{457}$$

We say that $\overline{x}$ is *exponentially stable* if there exists $\delta > 0\,,\ C > 0$ and $\varepsilon > 0$ such that for each $|x_0 - \overline{x}| < \delta$ we have

$$|x(t) - x_0| \le Ce^{-\varepsilon t} \qquad t \ge 0. \tag{458}$$

Various variants of these terms may be used. For example, some authors may use “stable” in place of “Lyapunov stable”, or “exponentially asymptotically stable” in place of “exponentially stable”.

Example 1

Consider the harmonic oscillator described by

$$x' = Ax, \qquad A = \begin{pmatrix} 0 & 1 \\ -\omega^2 & 0 \end{pmatrix}, \qquad \omega > 0\,. \tag{459}$$

The equilibrium $\overline{x} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ is Laypunov stable but not asymptotically stable: once disturbed, the system will not return to the equilibrium, but will oscillate around it. The amplitude of these oscillations will be small if the original disturbance is small.

Example 2
Consider the equation

$$x' = -x^3 \tag{460}$$

for a scalar real-valued function $x(t)$. The general solution on $(0,\infty)$ is

$$x(t) = \pm\frac{1}{\sqrt{2(t+t_0)}}\,, \qquad t_0 \geq 0\,. \tag{461}$$

We see that the the equilibrium $\overline{x} = 0$ is asymptotically stable, but not exponentially stable.

Theorem 6 from lecture 23 can be re-formulated in the following way:
Let A be an $n\times n$ matrix. The equilibrium $\overline{x} = 0$ of the system $x' = Ax$ is exponentially stable if and only if all eigenvalues of A have strictly negative real parts.
There is an important generalization of this theorem to general (not necessarily linear) autonomous systems:

Theorem 7. (Sufficient condition for exponential stability of equilibria)
Let $\overline{x}$ be an equilibrium of a general system $x' = f(x)$ with a smooth[68] f. Let $A = f'(\overline{x})$ be the matrix $\{\frac{\partial f_i}{\partial x_j}(\overline{x})\}_{i,j=1}^n$. The following statements are equivalent.

(i) $\overline{x}$ is exponentially stable.

(ii) The equilibrium $\overline{\xi} = 0$ of the linearized system $\xi' = A\xi$ is exponentially stable.

(iii) All eigenvalues of A have strictly negative real parts.

This is one of the most important results covered in this course.

In the remainder of the lecture we started discussing the predator-prey model found on page 332 in the textbook (the Volterra-Lotka system).

Lecture 38, 5/1/2013

Volterra-Lotka model, Hamiltonian systems
We continued the discussion of the Volterra-Lotka model, p. 332 in the textbook. We noted that the system can be rewritten in the following way

$$\dot{x} = xy\frac{\partial H}{\partial y}\,, \quad \dot{y} = -xy\frac{\partial H}{\partial x}\,, \tag{462}$$

where

$$H(x,y) = \sigma y + \kappa x - a\log y - r\log x\,. \tag{463}$$

[68] This condition can be relaxed, but we are not aiming for the most general formulation.

This is reminiscent of the Hamitonian equations in Classical Mechanics. Recall that for a particle moving along a curve parametrized by its length x, we have kinetic energy $E_{\text{kin}} = \frac{1}{2}m|\dot{x}|^2$, where m is the mass of the particle. The potential energy is assumed to be given by $V(x)$, a smooth function of x. The (generalized[69]) momentum then is $p = m\dot{x}$. The total energy of the system then is

$$H(x,p) = \frac{p^2}{2m} + V(x) \;, \tag{464}$$

the *Hamiltonian function*, and the equations of motion can be written in the *Hamiltonian form*:

$$\dot{x} = \frac{\partial H}{\partial p}\,, \quad \dot{p} = -\frac{\partial H}{\partial x}\,. \tag{465}$$

In this form of the equations the conservation of energy

$$\frac{dH}{dt} = 0 \tag{466}$$

is transparent, as the vector $\begin{pmatrix} \dot{x} \\ \dot{p} \end{pmatrix}$ is obtained by rotating the vector ∇H in by $\frac{\pi}{2}$ in the clock-wise direction, so that it is tangent to the level sets $\{H = \text{const}\}$. The direct evaluation of (466) of course confirms this:

$$\frac{dH}{dt} = \frac{\partial H}{\partial x}\frac{\partial H}{\partial p} + \frac{\partial H}{\partial p}\left(-\frac{\partial H}{\partial x}\right) = 0\,. \tag{467}$$

Therefore in integral curves of (465) are the level sets of H (as the plane (x,p) is two-dimensional). The same considerations of course apply to (462) and we see that the integral lines of the Volterra-Lotka system are level the sets of H. This can be of course arrived at by many other ways.
We note that if we let $\xi = \log$ and $\eta = \log y$ (assuming $x, y > 0$) then (462) becomes

$$\dot{\xi} = \frac{\partial H}{\partial \eta}\,, \qquad \dot{\eta} = -\frac{\partial H}{\partial \xi}\,, \tag{468}$$

where

$$H = \sigma e^{\eta} + \kappa e^{\xi} - a\eta - r\xi\,. \tag{469}$$

Next time we will discuss predator-prey models which are no-longer "Hamiltonian" but are exhibit some "friction" (in the language of Mechanics), so that the solutions converge to an equilibrium. In such systems one can often find a *Laypunov function*, which is a function on the phase-space of the given system which is decreasing along the trajectories (somewhat similarly to the energy decreasing along the solutions of system with friction, although in general the Laypunov function, if it exists, does not have to have a simple physical interpretation). One such model is given by system (13.24) in the textbook (and it is analyzed in the book in some detail).

[69] The curve does not to be a straight line (as long as x is the length parameter). This last assumptions can be also removed, but then our definition of the generalized momentum needs to be adjusted.

Lecture 39, 5/3/2013

Flow maps, Preservation of volume, Poincaré recurrence

Let us consider an autonomous system in $\mathbf{R}^n$ given by

$$\dot{x} = f(x)\,, \tag{470}$$

with a smooth f. Let us assume that for each $x \in \mathbf{R}^n$ the solution $y = y(t)$ of $\dot{y} = f(y)$ with the initial condition $y(0) = x$ exists for all $t \in \mathbf{R}$. We define a family of maps

$$\phi^t \colon \mathbf{R}^n \to \mathbf{R}^n \tag{471}$$

by

$$\phi^t(x) = y(t)\,, \qquad y(0) = x\,. \tag{472}$$

In other words, $\phi^t(x)$ is the point where the solution will be at time t assuming the solution is at x at time $t = 0$. The definitions imply that for the special case of linear systems

$$\dot{x} = Ax\,, \tag{473}$$

where A is an $n \times n$ matrix (independent of time) we have

$$\phi^t(x) = e^{tA}x\,, \qquad x \in \mathbf{R}^n,\ \ t \in \mathbf{R}\ . \tag{474}$$

Under our assumptions the maps ϕ^t are diffeomorphisms of $\mathbf{R}^n$, with

$$\left(\phi^t\right)^{-1} = \phi^{-t}\,. \tag{475}$$

More generally, it is easy to verify that

$$\phi^{t_1} \circ \phi^{t_2} = \phi^{t_1+t_2}\,. \tag{476}$$

The formula

$$e^{t_1A}e^{t_2A} = e^{(t_1+t_2)A}\,. \tag{477}$$

can be thought of as a special case of (476). We say that a diffeomorphism $\phi \colon \mathbf{R}^n \to \mathbf{R}^n$ is *volume preserving* if for each measurable set $E \subset \mathbf{R}^n$ we have

$$|\phi(E)| = |E|\,, \tag{478}$$

$|X|$ denotes the $n-$dimensional Lebesgue measure of the set X (assuming X is measurable). We have the following important result:

Lemma 6. *Let f and ϕ^t be as above. The following conditions are equivalent:*

1. *ϕ^t is volume-preserving for each t ;*
2. *$\det \nabla\phi^t(x) = 1$ for each $x \in \mathbf{R}^n$ and each $t \in \mathbf{R}$;*
3. *$\operatorname{div} f = \frac{\partial f_1}{\partial x_1} + \frac{\partial f_2}{\partial x_2} + \cdots + \frac{\partial f_n}{\partial x_n} = 0$ in $\mathbf{R}^n$.*

If any of these conditions is satisfied, we say that the flow ϕ^t is volume-preserving. It is easy to generalize this notion to the case where the measure we consider is not the Lebesgue masure, but, say, the measure given by the density $\rho(x)\,dx$, where $\rho > 0$ in $\mathbf{R}^n$ is a sufficiently regular function.

Although the proof of the lemma is not difficult, we will not discuss it at this point. As an exercise you can verify the lemma for linear systems with $f(x) = Ax$ for some $n \times n$ matrix A.

An important fact of Classical Mechanics is that the flows given by Hamiltonian systems are volume preserving. More precisely, let us consider a system of Classical Mechanics with n degrees of freedom. We know that such system is governed by the so-called Hamiltonian equations

$$\dot{x}_i = \frac{\partial H}{\partial p_i}\,, \qquad \dot{p}_i = -\frac{\partial H}{\partial x_i}\,. \tag{479}$$

We have the classical Liouville Theorem:

Theorem 8. (Liouville Theorem) *The flow generated by (479) is volume-preserving in the $2n-$dimensional phase-space $x_1, \dots, x_n, p_1, \dots, p_n$.*

The proof follows easily from Lemma 6 and (479) and is left to the reader as an exercise.

The preservation of volume by the flow map has important consequences for the dynamics of the system. For example, one sees easily that a system with volume-preserving flow cannot have an equilibrium which is asymptotically stable.

One of the best-known results for volume-preserving flows is the following:

Theorem 9. (A version of the Poincaré Recurrence Theorem) *Let $\Omega \subset \mathbf{R}^n$ be an open set of finite volume and let $\phi\colon \Omega \to \Omega$ be a volume-preserving diffeomorphism. Then for any measurable set $\mathcal{O} \subset \Omega$ the following statement is true: for almost every $x \in \mathcal{O}$ the sequence $\phi^k(x)\,,\ k = 1, 2, \dots$ will return to $\mathcal{O}$, in the sense that $\phi^{k_1}(x) \in \mathcal{O}$ for some $k_1 = k_1(x) \geq 1$.*

Remarks:
1. One can show under the same assumptions that almost every point of $\mathcal{O}$ will in fact return to $\mathcal{O}$ infinitely often.
2. To appreciate the unexpected nature of the statement, think of ϕ as describing the evolution of a complicated mechanical system in some given period of time, and think of $\mathcal{O}$ as a very small neighborhood of a point in the phase-space.

Sketch of proof of the theorem: Let $E \subset \mathcal{O}$ be a set with $|E| > 0$ such that $\phi^k(E) \cap \mathcal{O} = \emptyset$ for $k = 1, 2, \dots$. The sets all $\phi^k(E)$ have volume $|E| > 0$, and are contained in Ω. As $|\Omega| < +\infty$, we see that $|\phi^{k_1}(E) \cap \phi^{k_2}(E)| > 0$ for some $1 \leq k_1 < k_2$, and this easily leads to a contradiction.

Remark: In some sense, the recurrence theorem can be thought of as a generalization of the following easy fact: Let X be a finite set and let $\phi\colon X \to X$ be a bijective mapping. Then there exists an $m \in \mathbf{N}$ such that $\phi^m(x) = x$ for each $x \in X$.

Lecture 40, 5/6/2013

2d volume-preserving flows; modified Volterra-Lotka

Let $f(x) = \begin{pmatrix} f_1(x_1, x_2) \\ f_2(x_1, x_2) \end{pmatrix}$ be a vector field in $\mathbf{R}^2$ generating a volume-preserving flow. By Lemma 6 from the last lecture this is equivalent to the condition

$$\operatorname{div} f(x) = 0 \quad \text{in } \mathbf{R}^2 \,. \tag{480}$$

Such fields can be characterized as follows:

Lemma 7. *Let f be a smooth vector field in $\mathbf{R}^2$. Then the following conditions are equivalent:*

(i) $\operatorname{div} f = 0$ *in* $\mathbf{R}^2$ *;*

(ii) There exists a smooth function ψ on $\mathbf{R}^2$ such that

$$f = \begin{pmatrix} \frac{\partial\psi}{\partial x_2} \\ -\frac{\partial\psi}{\partial x_1} \end{pmatrix} \,. \tag{481}$$

The function ψ is called the *stream function* of the vector field f.

Proof of the Lemma

The implication (ii) $\Longrightarrow$ (i) is a direct consequence of the identity

$$\frac{\partial}{\partial x_1}\frac{\partial}{\partial x_2}\psi = \frac{\partial}{\partial x_2}\frac{\partial}{\partial x_1}\psi \tag{482}$$

The implication (i)$\Longrightarrow$ (ii) follows from the following statement which is proved in multi-dimensional calculus:

A smooth vector field f in $\mathbf{R}^n$ is a gradient field, i e. $f = \nabla\phi$ for some function $\phi\colon \mathbf{R}^n \to \mathbf{R}$ if and only if

$$\frac{\partial f_i}{\partial x_j} = \frac{\partial f_j}{\partial x_i}\,, \quad i, j = 1, 2, \ldots n \,. \tag{483}$$

If $\operatorname{div} f = 0$ the vector field g defined by $g_1 = -f_2$ and $g_2 = f_1$ satisfies

$$\frac{\partial g_1}{\partial x_2} = \frac{\partial g_2}{\partial x_1} \tag{484}$$

and hence $g = \nabla\psi$ for some function ψ. This is exactly (481).
The stream function ψ is determined by the vector field f uniquely up to a constant.

If f is a vector field in $\mathbf{R}^2$ with $\operatorname{div} f = 0$ and ψ is its steam function and $x(t)$ is a solution of $x' = f(x)$, then

$$\frac{d}{dt}\psi(x(t)) = 0\,. \tag{485}$$

In other words, the steam function is constant on the solutions of $x' = f(x)$. Therefore the level sets $\{\psi = c\}$ give us the trajectories of the system $x' = f(x)$. We conclude that trajectories of 2d autonomous volume-preserving flows must have a relatively simple structure.

In the second part of the lecture we discussed the modification of the Volterra-Lotka model presented by (13.24) in the textbook, following the presentation in the textbook.

Lecture 41, 5/8/2013

Predator Prey models, Long-time behavior of solutions of 2d dynamical systems .

We finished the discussion of the system (13.24) in the textbook. We informally discussed the long-time behavior of the solutions of the 2d autonomous systems (Poincaré-Bendixon theorem, etc.), see Section 4.12 in the textbook.

Lecture 42, 5/10/2013

The possibility of "chaos" in 3d non-linear systems; Lorenz system; Lorenz attractor

We discussed the possible chaotic behavior of 3d systems and one of the famous examples - the Lorenz system.