Some additional notes for math 252

Eduardo Sontag, Rutgers (version printed: February 13, 2016)

1 Introduction to modeling with differential equations

A *differential equation* is just an equation which involves "differentials", that is to say, derivatives. A simple example is:

$$\frac{dy}{dt} = 0,$$

where we understand that y is a function of an independent variable t. (We use t because in many examples the independent variable happens to be time, but of course any other variable could be used. It is sometimes convenient to use informal notation, and refer to this example as "$y' = 0$" or as "$\dot{y} = 0$" (the latter notation is favored in engineering and applied mathematics), though such a notation blurs the distinction between *functions* and the *expressions* used to define them.

If $y' = 0$, y must be constant. In other words, the general solution of the given equation is $y \equiv c$, for some constant c.

Another easy example of a differential equation is:

$$\frac{dy}{dt} = -27.$$

This means that $y = y(t)$ has a graph which is a line with slope -27. The general solution of this equation is $y = -27t + c$, for some constant c.

An *initial value problem* is a problem in which we give a differential equation together with an extra condition at a point, like:

$$\frac{dy}{dt} = -27, \quad y(0) = 3.$$

There is a unique solution of this initial-value problem, namely $y(t) = -27t+3$. It can be found by first finding the general solution $y = -27t + c$ and then plugging in $t = 0$ to get $3 = -27(0) + c$, so $c = 3$. This "initial" condition may be specified, of course, at any value of the independent variable t, (not just $t = 0$) for example:

$$\frac{dy}{dt} = -27, \quad y(2) = 3.$$

The solution of this initial-value problem can be also obtained by plugging into the general form $y = -27t + c$: we substitute $3 = y(2) = -27(2) + c$, which gives that $c = 57$, and so the solution is $y(t) = -27t + 57$. Although the word "initial" suggests that we intend to start at that point and move forward in time, the solutions we have found are defined for all values of t. We will not always be so fortunate, but do we expect solutions defined on an interval with the "initial" value in the interior.

A slightly more complicated example of a differential equation is:

$$\frac{dy}{dt} = \sin t + t^2.$$

The general solution is (by taking antiderivatives) $y = -\cos t + t^3/3 + c$. Another example:

$$\frac{dy}{dt} = e^{-t^2}.$$

This equation has a general solution, but it cannot be expressed in terms of elementary functions like polynomials, trigs, logs, and exponentials. (The solution is the "error function" that is used in statistics to define the cumulative probability of a Gaussian or normal probability density.) One of the unfortunate facts about differential equations is that we cannot always find solutions as explicit combinations of elementary functions. So, in general, we have to use numerical, geometric, and graphical techniques in the analysis of properties of solutions.

The examples just given are too easy (even if $y' = e^{-t^2}$ doesn't look *that* easy), in the sense that they can all be solved, at least theoretically, by taking antiderivatives. The subject of differential equations deals with far more general situations, in which the unknown function y appears on both sides of the equation:

$$y' = f(t, y)$$

or even much more general types: systems of many simultaneous equations, higher order derivatives, and even partial derivatives when there are other independent variables (which leads to "partial differential equations" and are the subject of more advanced courses).

One aspect of differential equations is comparatively easy: if someone gives us an alleged solution of an equation, we can *check* whether this is so. Checking is much easier than finding! (Analogy: if I ask you to find a solution of the algebraic equation $10000x^5 - 90000x^4 + 65100x^3 + 61460x^2 + 13812x + 972 = 0$ it may take you some time to find one. On the other hand, if I tell you that $x = 3/2$ is a root, you can check whether I am telling the truth or not very easily: just plug in and see if you get zero.) For example, if someone claims that the function $y = 1 \ / \ (1 + t^2)$ is a solution of the equation $y' = -2ty^2$, we can check that she is right by plugging in:

$$\left(\frac{1}{1+t^2}\right)' = -\frac{2t}{(1+t^2)^2} = -2t\left(\frac{1}{1+t^2}\right)^2.$$

But if someone claims that $y = 1 \ / \ (1 + t)$ is a solution, we can prove him to be wrong:

$$\left(\frac{1}{1+t}\right)' = -\frac{1}{(1+t)^2} \neq -2t\left(\frac{1}{1+t}\right)^2$$

because the two last functions of t are not the same. They even have different values at $t = 0$.

About Modeling

Most applications of mathematics, and in particular, of differential equations, proceed as follows.

Starting from a "word problem" description of some observed behavior or characteristic of the real world, we attempt to formulate the simplest set of mathematical equations which capture the essential aspects. This set of equations represents a *mathematical model* of reality. The study of the model is then carried out using mathematical tools. The power of mathematics is that it allows us to make quantitative and/or qualitative conclusions, and predictions about behaviors which may not have been an explicit part of the original word description, but which nonetheless follow logically from the model.

Sometimes, it may happen the results of the mathematical study of the model turn out to be inconsistent with features found in the "real world" original problem. If this happens, we must modify and adapt the model, for example by adding extra terms, or changing the functions that we use, in order to obtain a better match. Good modeling, especially in science and engineering, is often the result of several iterations of the "model/reality-check/model" loop!

Unrestricted Population Growth

When dealing with the growth of a bacterial culture in a Petri dish, a tumor in an animal, or even an entire population of individuals of a given species, biologists often base their models on the following simple rule:

The increase in population during a small time interval of length Δt is proportional to Δt and to the size of the population at the start of the interval.

For example, statistically speaking, we might expect that one child will be born in any given year for each 100 people. The proportionality rule then says that two children per year are born for every 200 people, or that three children are born for each 100 people over three consecutive years. (To be more precise, the rate of increase should be thought of as the "net" rate, after subtracting population decreases. Indeed, the decreases may also assumed proportional to population, allowing the two effects to be combined easily.)

The rule is only valid for small intervals (small Δt), since for large Δt one should also include compounding effects (children of the children), just as the interest which a bank gives us on savings (or charges us on loan balances) gets compounded, giving a higher effective rate.

Let us call $P(t)$ the number of individuals in the population at any given time t. The simplest way to translate into math the assumption that "the increase in population $P(t+\Delta t) - P(t)$ is proportional to Δt and to $P(t)$" is to write

$$P(t+\Delta t) - P(t) = kP(t)\Delta t \tag{1}$$

for some constant k. Notice how this equation says that the increase $P(t+\Delta t) - P(t)$ is twice as big if Δt is twice as big, or if the initial population $P(t)$ is twice as big.

Example: in the "one child per 100 people per year" rule, we would take $k = 10^{-2}$ if we are measuring the time t in years. So, if at the start of 1999 we have a population of 100,000,000, then at the beginning of the year 2001 = 1999+2 the population should be (use $\Delta t = 2$):

$$P(2001) = P(1999) + 10^{-2}P(1999)\Delta t = 10^8 + 10^{-2}10^8(2) = 102,000,000$$

according to the formula. On the other hand, by the end of January 3rd, 1999, that is, with $\Delta t = 3/365$, we would estimate $P(1999 + 3/365) = 10^8 + 10^{-2}10^8(3/365) \approx 100,008,219$ individuals. Of course, there will be random variations, but on average, such formulas turn out to work quite well.

The equation (1) can only be accurate if Δt is small, since it does not allow for the "compound interest" effect. On the other hand, one can view (1) as specifying a step-by-step *difference equation* as follows. Pick a "small" Δt, let us say $\Delta t = 1$, and consider the following recursion:

$$P(t+1) = P(t) + kP(t) = (1+k)P(t) \tag{2}$$

for $t = 0, 1, 2, \ldots$. Then we compute $P(2)$ not as $P(0) + 2kP(0)$, but recursively applying the rule: $P(2) = (1+k)P(1) = (1+k)^2P(0)$. This allows us to incorporate the compounding effect. It has the disadvantage that we cannot talk about $P(t)$ for fractional t, but we could avoid that problem by picking a smaller scale for time (for example, days). A more serious disadvantage is that it is hard to study difference equations using the powerful techniques from calculus. Calculus deals with things such as rates of change (derivatives) much better than with finite increments. Therefore, what we will do next is to show how the problem can be reformulated in terms of a differential equation. This is not to say that difference equations are not interesting, however. It is just that differential equations can be more easily studied mathematically.

If you think about it, you have seen many good examples of the fact that using derivatives and calculus is useful even for problems that seem not to involve derivatives. For example, if you want to find an integer t such that $t^2 - 189t + 17$ is as small as possible, you could try enumerating all possible integers (!), or you could instead pretend that t is a real number and minimize $t^2 - 189t + 17$ by setting the derivative to zero: $2t - 189 = 0$ and easily finding the answer $t = 94.5$, which then leads you, since you wanted an integer, to $t = 94$ or $t = 95$.

Back to our population problem, in order to use calculus, we must allow P to be any real number (even though, in population studies, only integers P would make sense), and we must also allow the time t to be any real number. Let us see where equation (1) leads us. If we divide by Δt, we have

$$\frac{P(t+\Delta t)-P(t)}{\Delta t} = kP(t).$$

This equation holds for small Δt, so we may let $\Delta t \to 0$. What is the limit of $(P(t+\Delta t)-P(t))\ /\ \Delta t$ as $\Delta t \to 0$? It is, as you remember from Calculus I (yes, you do), the derivative of P evaluated at t. So we end up with our first differential equation:

$$P'(t) = kP(t). \tag{3}$$

This is the differential equation for population growth. We may read it like this:

The rate of change of P is proportional to P.

The solution of this differential equation is easy: since $P'(t)/P(t) = k$, the chain rule tells us that

$$(\ln P(t))' = k,$$

and so we conclude that $\ln P(t) = kt + c$ for some constant c. Taking exponentials of both sides, we deduce that $P(t) = e^{kt+c} = Ce^{kt}$, where C is the new constant e^c. Evaluating at $t = 0$ we have that $P(0) = Ce^0 = C$, and we therefore conclude:

$$P(t) = P(0)e^{kt}.$$

(Actually, we cheated a little, because P'/P doesn't make sense if $P = 0$, and also because if P is negative then we should have used $\ln(-P(t))$. But one can easily prove that the formula $P(t) = P(0)e^{kt}$ is always valid. In any case, for population problems, P is positive.)

Which is better in practice, to use the difference equation (2) or the differential equation (3)? It is hard to say: the answer depends on the application. Mathematically, differential equations are usually easier to analyze, although sometimes, as when we study chaotic behavior in simple one-dimensional systems, difference equations may give great insight. Also, we often use difference equations as a basis of numerical techniques which allow us to find an approximation of the solution of a differential equation. For example, Euler's method basically reverses the process of going from (1) to (3).

Let us now look at some more examples of differential equations.

Limits to Growth: Logistic Equation

Often, there are limits imposed by the environment on the maximal possible size of a population: not enough nutrients for a large bacterial culture, insufficient food for the human population of an island, or a small hunting territory for a given animal species. Ecologists talk about the *carrying capacity* of the environment, a number N with the property that no populations $P > N$ are sustainable. If the population starts bigger than N, the number of individuals will decrease. To come up with an equation that represents this situation, we follow the same steps that we did before, except that now we have that $P(t+\Delta t) - P(t)$ should be negative if $P(t) > N$. In other words, we have $P(t+\Delta t) - P(t) = f(P(t))\Delta t$, where $f(P)$ is not just "kP" but should be instead a more complicated expression involving P, and which has the properties that:

- $f(0) = 0$ (no increase in the population if there is no one around to start with!),
- $f(P) > 0$ when $0 < P < N$ (the population increases while there are enough resources), and
- $f(P) < 0$ when $P > N$.

Taking limits just like we did before, we arrive to the differential equation:

$$P'(t) = f(P(t)).$$

From now on, we will drop the "t" when it is obvious, and use the shorthand notation $P' = f(P)$ instead of the more messy $P'(t) = f(P(t))$. We must still decide what function "f" is appropriate. Because of the properties wanted ($f(0) = 0$, $f(P) > 0$ when $0 < P < N$, $f(P) < 0$ when $P > N$), the simplest choice is a parabola which opens downward and has zeroes at $P = 0$ and $P = N$: $f(P) = -cP(P - N)$, with $c > 0$, or, with $k = cN$, $f(P) = kP(1 - P/N)$. We arrive in this way to the *logistic population model*

$$P' = kP\left(1 - \frac{P}{N}\right). \tag{4}$$

(Remember: this is shorthand for $P'(t) = kP(t)(1 - P(t)/N)$.) The constant k is positive, since it was obtained as cN.

Solution of Logistic Equation

Like $P' = kP$, equation (4) is one of those (comparatively few) equations which can actually be solved in closed form. To solve it, we do almost the same that we did with $P' = kP$ (this is an example of the method of *separation of variables*): we write the equation as $dP/dt = kP(1 - (P/N))$, formally multiply both sides by dt and divide by $P(1 - (P/N))$, arriving at

$$\frac{dP}{P(1 - P/N)} = kdt.$$

Next we take antiderivatives of both sides, obtaining

$$\int \frac{dP}{P(1 - P/N)} = \int kdt.$$

The right-hand side can be evaluated using partial fractions:

$$\frac{1}{P(1 - P/N)} = \frac{N}{P(N - P)} = \frac{1}{P} + \frac{1}{N - P}$$

so

$$\ln P - \ln(N - P) + c_1 = kt + c_2$$

for some constants c_1 and c_2, or, with $c = c_2 - c_1$,

$$\ln\left(\frac{P}{N - P}\right) = kt + c \tag{5}$$

and, taking exponentials of both sides,

$$\frac{P}{N - P} = Ce^{kt} \tag{6}$$

with $C = e^c$. This is an algebraic equation for P, but we can go a little further and solve explicitly:

$$P = Ce^{kt}(N - P) \Rightarrow Ce^{kt}P + P = Ce^{kt}N \Rightarrow P = \frac{Ce^{kt}N}{Ce^{kt} + 1} = \frac{N}{1 + \frac{1}{C}e^{-kt}}.$$

Finally, to find C, we can evaluate both sides of equation (6) at $t = 0$:

$$C = \frac{P(0)}{N - P(0)}$$

and therefore conclude that

$$P(t) = \frac{P(0)N}{P(0) + (N - P(0))e^{-kt}}. \tag{7}$$

Observe that, since $e^{-kt} \to 0$ as $t \to \infty$, $P(t) \to N$, which is not surprising. (Why?)

This formula is also valid for negative values of t with $P(t) \to 0$ as $t \to -\infty$.

Homework assignment: use a computer to plot several solutions of the equation, for various values of N and of $P(0)$.

Some "Small-Print Legal Disclaimers"

(You may want to skip this section in a first reading.)

We cheated a bit when deriving the solution for the logistic equation. First of all, we went a bit too fast over the "divide by dt" business. What is the meaning of dividing by the differential? Well, it turns out that it is OK to do this, because what we did can be interpreted as, basically, just a way of applying (backwards) the chain rule. Let us justify the above steps without using differentials. Starting from the differential equation (4) we can write, assuming that $P \neq 0$ and $P \neq N$ (so that we are not dividing by zero):

$$\frac{P'}{P(1 - P/N)} = k. \tag{8}$$

Now, one antiderivative of $1 \,/\, (P(1 - P/N))$, as a function of P, is the function

$$Q(P) = \ln\left(P \,/\, (N - P)\right)$$

(let us suppose that $N > P$, so the expression inside the log is positive). So, the chain rule says that

$$\frac{dQ(P(t))}{dt} = \frac{dQ}{dP}\frac{dP}{dt} = \frac{1}{P(1 - P/N)}P'(t).$$

Therefore, equation (8) gives us that

$$\frac{dQ(P(t))}{dt} = k$$

from which we then conclude, by taking antiderivatives, that

$$Q(P(t)) = kt + c$$

which is exactly the same as the equation (5) which had before been obtained using differentials. In general, we can always justify "separation of variables" solutions in this manner, but from now on we will skip this step and use the formal method.

There is still a small gap in our arguments, namely we assumed that $P \neq 0$ and that $P \neq N$ (so that we were not dividing by zero) and also $N > P$, so the expression inside the log was positive.

There is a theorem that states that, under appropriate conditions (differentiability of f), solutions are unique. Thus, since $P = 0$ and $P = N$ are equilibria, any solution that starts with $P(0) > N$ will always have $P(t) > N$, and a similar property is true for each of the intervals $P < 0$ and $0 < P < N$. So we can treat each of the cases separately.

If $N < P$, then the antiderivative is $\ln \left| P \,/\, (N - P) \right|$ (that is, we use absolute values). But this doesn't change the general solution. All it means is that equation (6) becomes

$$\left| \frac{P}{N - P} \right| = Ce^{kt}$$

which can also be written as in (6) but with C negative. We can treat the case $P < 0$ in the same way.

Finally, the exceptional cases when P could be zero or N are taken care of once we notice that the general solution (7) makes sense when $P(0) = 0$ (we get $P \equiv 0$) or when $P(0) = N$ (we get $P \equiv N$).

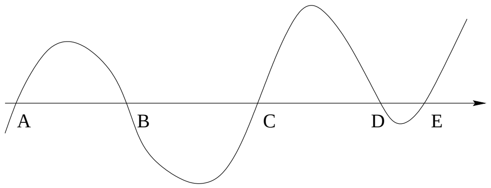
Equilibria

Observe that if, for some time t_0, it happens that $P(t_0) = 0$, then the right-hand side of the differential equation (4) becomes zero, so $P'(t_0) = 0$, which means that the solution cannot "move" from that point. So the value $P = 0$ is an **equilibrium point** for the equation: a value with the property that if we start there, then we stay there forever. This is not a particularly deep conclusion: if we start with zero population we stay with zero population. Another root of the right hand side is $P = N$. If $P(t_0) = N$ then $P'(t_0) = 0$, so if we start with exactly N individuals, the population also remains constant, this time at N. Again, this is not surprising, since the model was derived under the assumption that populations larger than N decrease and populations less than N increase.

In general, for any differential equation of the form $y' = f(y)$, we say that a point $y = a$ is an *equilibrium* if a is a root of f, that is, $f(a) = 0$. This means that if we start at $y = a$, we cannot move away from $y = a$. Or, put in a different way, the constant function $y(t) \equiv a$ is a solution of $y' = f(y)$ (because $y'(t) = a' \equiv 0$ and also $f(y(t)) = f(a) = 0$. One says also that the constant function $y(t) = a$ is an *equilibrium solution* of $y' = f(y)$.

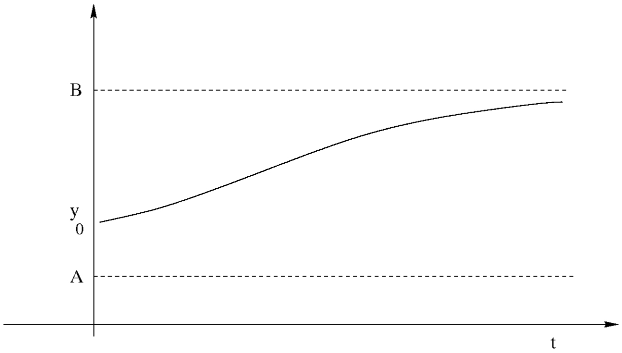
The analysis of equilibria allows us to obtain a substantial amount of information about the solutions of a differential equation of the type $y' = f(y)$ with very little effort, in fact without even having to solve the equation. (For "nonautonomous" equations, when t appears in the right hand side: $y' = f(t, y)$, this method doesn't quite work, because we need to plot f against two variables. The technique of slope fields is useful in that case.) The fundamental fact that we need is that — assuming that f is a differentiable function — *no trajectory can pass through an equilibrium*: if we are ever at an equilibrium, we must have always been there and we will remain there forever. This will be explained later, when covering uniqueness of solutions.

For example, suppose that we know that the plot of $f(y)$ against y looks like this:



where we labeled the points where $f(y)$ has roots, that is to say, the equilibria of $y' = f(y)$.

We can conclude that any solution $y(t)$ of $y' = f(P)$ which starts just to the right of A will move rightwards, because $f(y)$ is positive for all points between A and B, and so $y' > 0$. Moreover, we cannot cross the equilibrium B, so any such trajectory stays in the interval (A, B) and, as t increases, it approaches asymptotically the point B. To summarize, if $y(0) = y_0$ with $y_0 \in (A, B)$, then the graph of the solution $y(t)$ of $y' = f(y)$ must look more or less like this:



Homework assignment: For the same function f shown above, give an approximate plot of a solution of $y' = f(y)$ for which $y(0) \in (B, C)$. Repeat with $y(0) \in (C, D)$ and with $y(0) \in (D, E)$.

Systems

More generally, one considers *systems* of differential equations, such as for example:

$$\begin{aligned}\frac{dx}{dt} &= 2\,x \,-\, 5\,xy \\ \frac{dy}{dt} &= -y \,+\, 1.2\,xy\,.\end{aligned}$$

This example might represent the number of individuals of each of two species of animals, in which the "y" species is a predator of "x". The first species reproduces (at rate "2") if there are no y's present, but when there are y's around, there is a "death rate" for x that is proportional to the number of predators. Similarly, the second population grows in proportion to the population size of x's, but it diminishes when there are no x's (its only source of nutrition).

More Examples

Let us discuss some more easy examples (of single-variable problems).

(a) Populations under Harvesting

Let us return to the population model (4):

$$P' = kP\left(1 - \frac{P}{N}\right)$$

which describes population growth under environmental constraints. Suppose that $P(t)$ represents the population of a species of fish, and that fishing removes a certain number K of fish each unit of time. This means that there will be a term in $P(t + \Delta t) - P(t)$ equal to $-K\Delta t$. When we divide by Δt and take limits, we arrive at the equation for resources under constant harvesting:

$$P' = kP\left(1 - \frac{P}{N}\right) - K.$$

Many variations are possible. For example, it is more realistic to suppose that a certain proportion of fish are caught per unit of time (the more fish, the easier to catch). This means that, instead of a term $-K\Delta t$ for how many fish are taken away in an interval of length Δt, we'd now have a term of the form $-KP(t)\Delta t$, which is proportional to the population. The differential equation that we obtain is now $P' = kP(1 - (P/N)) - KP$. Or, if only fish near the surface can be caught, the proportion of fish caught per unit of time may depend on the power $P^{2/3}$ (do you understand why? are you sure?). This would give us the equation $P' = kP(1 - (P/N)) - KP^{2/3}$.

(b) Epidemics

The spread of epidemics is another example whose study can be carried out using differential equations. Suppose that $S(t)$ counts the number of individuals infected with a certain virus, at time t, and that people mix randomly and get infected from each other if they happen to be close. One model is as follows. The increase in the number of infected individuals $S(t + \Delta t) - S(t)$ during a time interval of length Δt is proportional to the number of close encounters between sick and healthy individuals, that is, to $S(t)H(t)\Delta t$, because $S(t)H(t)$ is the total number of pairs of (sick,healthy) individuals, and the longer the interval, the more chances of meeting.

Taking limits as usual, we arrive to $S'(t) = kS(t)H(t)$, where k is some constant. If the total number of individuals is N, then $H(t) = N - S(t)$, and the equation becomes:

$$S' = kS(t)(N - S(t))$$

which is a variant of the logistic equation. There are many extensions of this idea. For instance, if in every Δt time interval a certain proportion of infected individuals get cured, we'd have a term $-kS(t)$.

(c) Chemical Reactions

Chemical reactions also give rise to similar models. Let us say that there are two reactants A and B, which may combine to give C via $A + B \to C$ (for each molecule of A and B, we obtain a molecule of C). If the chemicals are well-mixed, the chance of two molecules combining is proportional to how many pairs there are and to the length of time elapsed (just like with the infection model, molecules need to get close enough to react). So $c'(t) = ka(t)b(t)$, where $a(t)$ is the amount of A at time t and $b(t)$ the amount of B. If we start with amounts a_0 and b_0 respectively, and we have $c(t)$ molecules of C at time t, this means that $a(t) = a_0 - c(t)$ and $b(t) = b_0 - c(t)$, since one molecule of A and B was used up for each molecule of C that was produced. So the equation becomes

$$c' = k(a_0 - c)(b_0 - c).$$

(d) Air Resistance

Consider a body moving in air (or another fluid). For low speeds, air resistance (drag) is proportional to the speed of the object, and acts to slow down the object, in other words, it acts as a force $k\,|v|$, in a direction opposite to movement, where $|v|$ is the absolute value of the velocity. Suppose that a body is falling towards the earth, and let us take "down" as the positive direction of movement. In that case, Newton's "$F = ma$" law says that the mass times the acceleration v' is equal to the total force on the body, namely mg (its weight) plus the effect of drag, which is $-kv$ (because the force acts opposite to the direction of movement):

$$mv' = mg - kv.$$

For large velocities, drag is often modeled more accurately by a quadratic effect $-kv^2$ in a direction opposite to movement. This would lead to an equation like $mv' = mg - kv^2$ for the velocity of a falling object. Both of these equations can be solved exactly. This allows the validity of the model to be tested by comparing these formulas to experimental results.

(e) Newton's Law of Cooling

The temperature inside a building is assumed to be uniform (same in every room) and is given by $y(t)$ as a function of the time t. The outside air is at temperature $a(t)$, which also depends on the time of the day, and there is a furnace which supplies heat at a rate $h(t)$ (or, for negative h, an air-conditioning unit which removes heat at that rate). What is the temperature in the building? Newton's law of cooling tells us that the rate of change of temperature dy/dt will depend on the difference between the inside and outside temperatures (the greater the difference, the faster the change), with a term added to model the effect of the furnace:

$$mcy' = -k(y - a(t)) + h(t),$$

where the mass of air in the building is the constant m (no windows can be opened, and doors are usually tightly closed, being opened rarely and briefly, so we assume that m is a constant), c is a positive constant (the heat capacity), and k is another positive constant (which is determined by insulation, building layout, etc).

2 Phase-planes

A technique which is often very useful in order to analyze the phase plane behavior of a two-dimensional autonomous system

$$\begin{aligned}\frac{dx}{dt} &= f(x,y)\\ \frac{dy}{dt} &= g(x,y)\end{aligned}$$

is to attempt to understand the graphs of solutions $(x(t), y(t))$ as the level sets of some function $h(x,y)$.

Some Examples.

Example 1

For example, take

$$\begin{aligned}\frac{dx}{dt} &= -y\\ \frac{dy}{dt} &= x\end{aligned}$$

(that is, $f(x,y) = -y$ and $g(x,y) = x$). If we could solve for t as a function of x, by inverting the function $x(t)$, and substitute the expression that we obtain into $y(t)$, we would end up with an expression $y(x)$ for the y-coordinate in terms of the x coordinate, eliminating t. This cannot be done in general, but it suggests that we may want to look at dy/dx. Formally (or, more precisely, using the chain rule), we have that

$$\frac{dy}{dx} = \frac{dy/dt}{dx/dt} = \frac{x}{-y}$$

which is a differential equation for y as a variable dependent on x. This equation is separable:

$$\int \frac{dy}{y} = \int -\frac{dx}{x}$$

so we obtain, taking antiderivatives,

$$\frac{y^2}{2} + \frac{x^2}{2} = c$$

where c is an undetermined constant, and since c must be nonnegative, we can write $c = r^2$. *In conclusion, the solutions $(x(t), y(t))$ all lie in the circles $x^2 + y^2 = r^2$ of different radii and centered at zero.* Observe that we have **not** solved the differential equation, since we did not determine the forms of x and y as functions of t (which, as a matter of fact, are trigonometric functions $x = r\cos t$, $y = r\sin t$ that draw circles at constant unit speed in the counterclockwise direction). What we have done is just to find curves (the above-mentioned circles) which contain all solutions. Even though this is less interesting (perhaps) than the actual solutions, it is still very interesting. We know what the general phase plane picture looks like.

A variation of this example is:

$$\begin{aligned}\frac{dx}{dt} &= -y\\ \frac{dy}{dt} &= 2x\end{aligned}$$

for which it is easy to see, by a similar reasoning, that the solutions lie in ellipses of the form

$$x^2 + \frac{y^2}{2} = r^2\,.$$

Example 2

Another example is this:

$$\begin{aligned}\frac{dx}{dt} &= y^5 e^x\\ \frac{dy}{dt} &= x^5 e^x.\end{aligned}$$

Here, $dy/dx = x^5/y^5$ so we get again a separable equation, and we see that the solutions all stay in the curves

$$x^6 - y^6 = c.$$

Example 3

More interesting is the general case of predator-prey equations:

$$\begin{aligned}\frac{dx}{dt} &= ax - bxy\\ \frac{dy}{dt} &= -cy + dxy\end{aligned}$$

where a, b, c, d are all positive constants. Then

$$\frac{dy}{dx} = \frac{y(-c + dx)}{x(a - by)}$$

so

$$\int \left(\frac{a}{y} - b\right) dy = \int \left(-\frac{c}{x} + d\right) dx$$

and from here we conclude that the solutions all stay in the sets

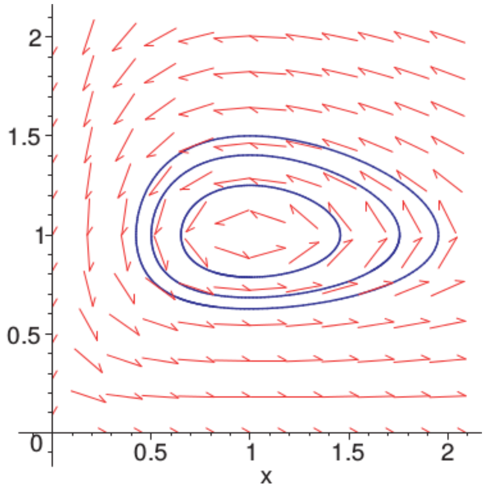
$$a\ln(y) - by + c\ln(x) - dx = k$$

for various values of the constant k.

It is not obvious what these sets look like, but if you graph the level sets of the function

$$h(x, y) = a\ln(y) - by + c\ln(x) - dx$$

you'll see that the level sets look like the orbits of the predator-prey system shown, for certain the special values of the parameters. (The initial values for the three shown curves were $(1, 1.25)$, $(0.5, 1)$, and $(1, 1, 5)$.)



(Of course, the scales will be different for different values of the constants, but the picture will look the same, in general terms.) This argument is used to prove that predator-prey systems always lead to periodic orbits, no matter what the coefficients of the equation are.

Problems

In each of the following problems, a system

$$\begin{aligned}\frac{dx}{dt} &= f(x,y)\\ \frac{dy}{dt} &= g(x,y)\end{aligned}$$

is given. Solve the equation

$$\frac{dy}{dx} = \frac{g(x,y)}{f(x,y)}$$

and use the information to sketch what the orbits of the original equation should look like.

Exercise 1

$$\begin{aligned}\frac{dx}{dt} &= y(1+x^2+y^2)\\ \frac{dy}{dt} &= x(1+x^2+y^2)\end{aligned}$$

Exercise 2

$$\begin{aligned}\frac{dx}{dt} &= 4y(1+x^2+y^2)\\ \frac{dy}{dt} &= \tfrac{dy}{dt} = -x(1+x^2+y^2)\end{aligned}$$

Exercise 3

$$\begin{aligned}\frac{dx}{dt} &= y^3e^{x+y}\\ \frac{dy}{dt} &= -x^3e^{x+y}\end{aligned}$$

Exercise 4

$$\begin{aligned}\frac{dx}{dt} &= y^2\\ \frac{dy}{dt} &= (2x+1)y^2\end{aligned}$$

Exercise 5

$$\begin{aligned}\frac{dx}{dt} &= e^{xy}\cos(x)\\ \frac{dy}{dt} &= e^{xy}\end{aligned}$$

3 Matrix Exponentials

Generalities

A system of autonomous linear differential equations can be written as

$$\frac{dY}{dt} = AY$$

where A is an n by n matrix and $Y = Y(t)$ is a vector listing the n dependent variables. (In most of what we'll do, we take $n = 2$, since we study mainly systems of 2 equations, but the theory is the same for all n.)

If we were dealing with just one linear equation

$$y' = ay$$

then the general solution of the equation would be e^{at}. It turns out that *also for vector equations the solution looks like this, provided that we interpret what we mean by "e^{At}" when A is a matrix instead of just a scalar.* How to define e^{At}? The most obvious procedure is to take the power series which defines the exponential, which as you surely remember from Calculus is

$$e^x = 1 + x + \frac{1}{2}x^2 + \frac{1}{6}x^3 + \cdots + \frac{1}{k!}x^k + \cdots$$

and just formally plug-in $x = At$. (The answer should be a matrix, so we have to think of the term "1" as the identity matrix.) In summary, we *define*:

$$e^{At} = I + At + \frac{1}{2}(At)^2 + \frac{1}{6}(At)^3 + \cdots + \frac{1}{k!}(At)^k + \cdots$$

where we understand the series as defining a series for each coefficient. One may prove that:

$$e^{A(t+s)} = e^{At}e^{As} \text{ for all } s, t\,. \tag{9}$$

and therefore, since (obviously) $e^{A0} = I$, using $s = -t$ gives

$$e^{-At} = \left(e^{At}\right)^{-1} \tag{10}$$

(which is the matrix version of $e^{-x} = 1/e^x$). We now prove that this matrix exponential has the following property:

$$\frac{de^{At}}{dt} = Ae^{At} = e^{At}A \tag{11}$$

for every t.

Proof Let us differentiate the series term by term:

$$\begin{aligned}
\frac{de^{At}}{dt} &= \frac{d}{dt}\left(I + At + \frac{1}{2}(At)^2 + \frac{1}{6}(At)^3 + \cdots + \frac{1}{k!}(At)^k + \cdots\right) \\
&= 0 + A + A^2t + \frac{1}{2}A^3t^2 + \cdots + \frac{1}{(k-1)!}A^kt^{k-1} + \cdots \\
&= A\left(I + At + \frac{1}{2}(At)^2 + \frac{1}{6}(At)^3 + \cdots + \frac{1}{k!}(At)^k + \cdots\right) \\
&= Ae^{At}
\end{aligned}$$

and a similar proof, factoring A on the right instead of to the left, gives the equality between the derivative and $e^{At}A$. (Small print: the differentiation term-by-term can be justified using facts about term by term differentiation of power series inside their domain of convergence.) The property (11) is the fundamental property of exponentials of matrices. It provides us immediately with this corollary:

The initial value problem $\dfrac{dY}{dt} = AY$*,* $Y(0) = Y_0$ *has the unique solution* $Y(t) = e^{At}Y_0$.

We can, indeed, verify that the formula $Y(t) = e^{At}Y_0$ defines a solution of the IVP:

$$\frac{dY(t)}{dt} = \frac{de^{At}Y_0}{dt} = \frac{de^{At}}{dt}Y_0 = \left(Ae^{At}\right)Y_0 = A\left(e^{At}Y_0\right) = AY(t)\,.$$

(That it is the unique, i.e., the only, solution is proved as follows: if there were another solution $Z(t)$ of the same IVP, then we could let $W(t) = Y(t) - Z(t)$ and notice that $W' = Y' - Z' = A(Y - Z) = AW$, and $W(0) = Y(0) - Z(0) = 0$. Letting $V(t) = e^{-At}W(t)$, and applying the product rule, we have that

$$V' = -Ae^{-At}W + e^{-At}W' = -e^{-At}AW + e^{-At}AW = 0$$

so that V must be constant. Since $V(0) = W(0) = 0$, we have that V must be identically zero. Therefore $W(t) = e^{At}V(t)$ is also identically zero, which because $W = Y - Z$, means that the functions Y and Z are one and the same, which is what we claimed.)

Although we started by declaring Y to be a vector, the equation $Y' = AY$ makes sense as long as Y can be multiplied on the left by A, i.e., whenever Y is a matrix with n rows (and any number of columns). In particular, e^{At} itself satisfies this equation. The result giving uniqueness of solutions of initial value problems applies to matrices since each column satisfies the equation and has the corresponding column of the initial data as its initial value. The value of e^{At} at $t = 0$ is the n by n identity matrix. This initial value problem characterizes e^{At}. Verification of these properties is an excellent check of a calculation of e^{At}.

So we have, in theory, solved the general linear differential equation. A potential problem is, however, that it is not always easy to calculate e^{At}.

Some Examples

We start with this example:

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}\,. \tag{12}$$

We calculate the series by just multiplying A by t:

$$At = \begin{pmatrix} t & 0 \\ 0 & 2t \end{pmatrix}$$

and now calculating the powers of At. Notice that, because At is a diagonal matrix, its powers are very easy to compute: we just take the powers of the diagonal entries *(why? if you don't understand,* **stop** *and think it over right now)*. So, we get

$$e^{At} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} t & 0 \\ 0 & 2t \end{pmatrix} + \frac{1}{2}\begin{pmatrix} t^2 & 0 \\ 0 & (2t)^2 \end{pmatrix} + \frac{1}{6}\begin{pmatrix} t^3 & 0 \\ 0 & (2t)^3 \end{pmatrix} + \cdots + \frac{1}{k!}\begin{pmatrix} t^k & 0 \\ 0 & (2t)^k \end{pmatrix} + \cdots$$

and, just adding coordinate-wise, we obtain:

$$e^{At} = \begin{pmatrix} 1 + t + \frac{1}{2}t^2 + \frac{1}{6}t^3 + \cdots + \frac{1}{k!}t^k + \cdots & 0 \\ 0 & 1 + 2t + \frac{1}{2}(2t)^2 + \frac{1}{6}(2t)^3 + \cdots + \frac{1}{k!}(2t)^k + \cdots \end{pmatrix}$$

which gives us, finally, the conclusion that

$$e^{At} = \begin{pmatrix} e^t & 0 \\ 0 & e^{2t} \end{pmatrix} .$$

So, in this very special case we obtained the exponential by just taking the exponentials of the diagonal elements and leaving the off-diagonal elements zero (observe that we did not end up with exponentials of the non-diagonal entries, since $e^0 = 1$, not 0).

In general, computing an exponential is a little more difficult than this, and it is not enough to just take exponentials of coefficients. Sometimes things that seem surprising (the first time that you see them) may happen. Let us take this example now:

$$A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} . \tag{13}$$

To start the calculation of the series, we multiply A by t:

$$At = \begin{pmatrix} 0 & t \\ -t & 0 \end{pmatrix}$$

and again calculate the powers of At. This is a little harder than in the first example, but not too hard:

$$\begin{aligned}
(At)^2 &= \begin{pmatrix} -t^2 & 0 \\ 0 & -t^2 \end{pmatrix} \\
(At)^3 &= \begin{pmatrix} 0 & -t^3 \\ t^3 & 0 \end{pmatrix} \\
(At)^4 &= \begin{pmatrix} t^4 & 0 \\ 0 & t^4 \end{pmatrix} \\
(At)^5 &= \begin{pmatrix} 0 & t^5 \\ -t^5 & 0 \end{pmatrix} \\
(At)^6 &= \begin{pmatrix} -t^6 & 0 \\ 0 & -t^6 \end{pmatrix}
\end{aligned}$$

and so on. We won't compute more, because by now you surely have recognized the pattern (*right?*). We add these up (not forgetting the factorials, of course):

$$e^{At} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} 0 & t \\ -t & 0 \end{pmatrix} + \frac{1}{2}\begin{pmatrix} -t^2 & 0 \\ 0 & -t^2 \end{pmatrix} + \frac{1}{3!}\begin{pmatrix} 0 & -t^3 \\ t^3 & 0 \end{pmatrix} + \frac{1}{4!}\begin{pmatrix} t^4 & 0 \\ 0 & t^4 \end{pmatrix} + \cdots$$

and, just adding each coordinate, we obtain:

$$e^{At} = \begin{pmatrix} 1 - \frac{t^2}{2} + \frac{t^4}{4!} - \cdots & t - \frac{t^3}{3!} + \frac{t^5}{5!} - \cdots & -t + \frac{t^3}{3!} - \frac{t^5}{5!} + \cdots & 1 - \frac{t^2}{2} + \frac{t^4}{4!} - \cdots \end{pmatrix}$$

which gives us, finally, the conclusion that

$$e^{\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} t} = e^{At} = \begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix} .$$

It is remarkable that trigonometric functions have appeared. Perhaps we made a mistake? How could we make sure? Well, let us *check* that property (11) holds (we'll check only the first equality, you can check the second one). We need to test that

$$\frac{d}{dt}\begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix} = A \begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix} . \tag{14}$$

Since

$$\frac{d}{dt}(\sin t) = \cos t, \quad \text{and} \quad \frac{d}{dt}(\cos t) = -\sin t,$$

we know that
$$\frac{d}{dt}\begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix} = \begin{pmatrix} -\sin t & \cos t \\ -\cos t & -\sin t \end{pmatrix}$$
and, on the other hand, multiplying matrices:
$$\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}\begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix} = \begin{pmatrix} -\sin t & \cos t \\ -\cos t & -\sin t \end{pmatrix}$$
so we have verified the equality (14).

As a last example, let us take this matrix:
$$A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} . \tag{15}$$
Again we start by writing
$$At = \begin{pmatrix} t & t \\ 0 & t \end{pmatrix}$$
and calculating the powers of At. It is easy to see that the powers are:
$$(At)^k = \begin{pmatrix} t^k & kt^k \\ 0 & t^k \end{pmatrix}$$
since this is obviously true for $k = 1$ and, recursively, we have
$$(At)^{k+1} = (At)^k A = \begin{pmatrix} t^k & kt^k \\ 0 & t^k \end{pmatrix}\begin{pmatrix} t & t \\ 0 & t \end{pmatrix} = \begin{pmatrix} t^k t & t^k t + kt^k t \\ 0 & t^k t \end{pmatrix} = \begin{pmatrix} t^{k+1} & (k+1)t^{k+1} \\ 0 & t^{k+1} \end{pmatrix} .$$
Therefore,
$$\begin{aligned} e^{At} &= \sum_{k=0}^{\infty} \begin{pmatrix} t^k/k! & kt^k/k! \\ 0 & t^k/k! \end{pmatrix} \\ &= \begin{pmatrix} \displaystyle\sum_{k=0}^{\infty} \frac{t^k}{k!} & \displaystyle\sum_{k=0}^{\infty} \frac{kt^k}{k!} \\ 0 & \displaystyle\sum_{k=0}^{\infty} \frac{t^k}{k!} \end{pmatrix} \\ &= \begin{pmatrix} e^t & te^t \\ 0 & e^t \end{pmatrix} . \end{aligned}$$
To summarize, we have worked out three examples:

- The first example (12) is a diagonal matrix, and we found that its exponential is obtained by taking exponentials of the diagonal entries.
- The second example (13) gave us an exponential matrix that was expressed in terms of trigonometric functions. Notice that this matrix has imaginary eigenvalues equal to i and $-i$, where $i = \sqrt{-1}$.
- The last example (15) gave us an exponential matrix which had a nonzero function in the $(1, 2)$-position. Notice that this nonzero function was *not* just the exponential of the $(1, 2)$-position in the original matrix. That exponential would give us an e^t term. Instead, we got a more complicated te^t term.

In a sense, these are all the possibilities. Exponentials of all two by two matrices can be obtained using functions of the form e^{at}, te^{at}, and trigonometric functions (possibly multiplied by e^{at}). Indeed, exponentials of any size matrices, not just 2 by 2, can be expressed using just polynomial combinations of t, scalar exponentials, and trigonometric functions. We will not quite prove this fact here; you should be able to find the details in any linear algebra book.

Calculating exponentials using power series is OK for very simple examples, and important to do a few times, so that you understand what this all means. But in practice, one uses very different methods for computing matrix exponentials. (Remember how you first saw the definition of derivative using limits of incremental quotients, and computed some derivatives in this way, but soon learned how to use "the Calculus" to calculate derivatives of complicated expressions using the multiplication rule, chain rule, and so on.)

Computing Matrix Exponentials

We wish to calculate e^{At}. The key concept for simplifying the computation of matrix exponentials is that of *matrix similarity*. Suppose that we have found two matrices, Λ and S, where S is invertible, such that this formula holds:

$$A = S\Lambda S^{-1} \tag{16}$$

(if (16) holds, one says that A and Λ are similar matrices). Then, we claim, it is true that also:

$$e^{At} = S\, e^{\Lambda t}\, S^{-1} \tag{17}$$

for all t. Therefore, if the matrix Λ is one for which $e^{\Lambda t}$ is easy to find (for example, if it is a diagonal matrix), we can then multiply by S and S^{-1} to get e^{At}. To see why (17) is a consequence of (16), we just notice that $At = S(\Lambda t)S^{-1}$ and we have the following "telescopic" property for powers:

$$(At)^k = \Big(S(\Lambda t)S^{-1}\Big)\Big(S(\Lambda t)S^{-1}\Big)\cdots\Big(S(\Lambda t)S^{-1}\Big) = S(\Lambda t)^k S^{-1}$$

since the terms may be regrouped so that all the in-between pairs $S^{-1}S$ cancel out. Therefore,

$$\begin{aligned}
e^{At} &= I + At + \frac{1}{2}(At)^2 + \frac{1}{6}(At)^3 + \cdots + \frac{1}{k!}(At)^k + \cdots \\
&= I + S(\Lambda t)S^{-1} + \frac{1}{2}S(\Lambda t)^2 S^{-1} + \frac{1}{6}S(\Lambda t)^3 S^{-1} + \cdots + \frac{1}{k!}S(\Lambda t)^k S^{-1} + \cdots \\
&= S\left[I + \Lambda t + \frac{1}{2}(\Lambda t)^2 + \frac{1}{6}(\Lambda t)^3 + \cdots + \frac{1}{k!}(\Lambda t)^k + \cdots\right] S^{-1} \\
&= Se^{\Lambda t}S^{-1}
\end{aligned}$$

as we claimed.

The basic theorem is this one:

Theorem. For every n by n matrix A with entries in the complex numbers, one can find an invertible matrix S, and an upper triangular matrix Λ such that (16) holds.

Remember that an upper triangular matrix is one that has the following form:

$$\begin{pmatrix}
\lambda_1 & * & * & \cdots & * & * \\
0 & \lambda_2 & * & \cdots & * & * \\
0 & 0 & \lambda_2 & \cdots & * & * \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & \lambda_{n-1} & * \\
0 & 0 & 0 & \cdots & 0 & \lambda_n
\end{pmatrix}$$

where the stars are any numbers. The numbers $\lambda_1, \ldots, \lambda_n$ turn out to be the eigenvalues of A.

There are two reasons that this theorem is interesting. First, it provides a way to compute exponentials, because it is not difficult to find exponentials of upper triangular matrices (the example (15) is actually quite typical) and second because it has important theoretical consequences.

Although we don't need more than the theorem stated above, there are two stronger theorems that you may meet elsewhere. One is the "Jordan canonical form" theorem, which provides a matrix Λ that is not only upper triangular but which has an even more special structure. Jordan canonical forms are theoretically important because they are essentially unique (that is what "canonical" means in this context). Hence, the Jordan form allows you to determine whether or not two matrices are similar. However, it is not very useful from a computational point of view, because they are what is known in numerical analysis as "numerically unstable", meaning that small perturbations of A can give one totally different Jordan forms. A second strengthening is the "Schur unitary triangularization theorem" which says that one can pick the matrix S to be *unitary*. (A unitary matrix is a matrix with entries in the complex numbers whose inverse is the complex conjugate of its transpose. For matrices S with real entries, then we recognize it as an *orthogonal* matrix. For matrices with complex entries, unitary matrices turn out to be more useful than other generalization of orthogonal matrices that one may propose.) Schur's theorem is extremely useful in practice, and is implemented in many numerical algorithms.

We do not prove the theorem here in general, but only show it for $n = 2$; the general case can be proved in much the same way, by means of a recursive process.

We start the proof by remembering that every matrix has at least one eigenvalue, let us call it λ, and an associate eigenvector, v. That is to say, v is a vector **different from zero**, and

$$Av = \lambda v \,. \tag{18}$$

If you stumble on a number λ and a vector v that you believe to an eigenvalue and its eigenvector, you should *immediately* see if (18) is satisfied, since that is an easy calculation. Numerical methods for finding eigenvalues and eigenvectors take this approach.

For theoretical purposes, it is useful to note that the the eigenvalues λ can be characterized as the roots of the characteristic equation

$$\det(\lambda I - A) = 0.$$

For two-dimensional systems, this is the same as the equation

$$\lambda^2 - \text{trace}\,(A)\lambda + \det(A) = 0$$

with

$$\begin{aligned} \text{trace}\begin{pmatrix} a & b \\ c & d \end{pmatrix} &= a + d \\ \det\begin{pmatrix} a & b \\ c & d \end{pmatrix} &= ad - bc \,. \end{aligned}$$

Now, quadratic equations are easy to solve, so this approach is also computationally useful for 2 by 2 matrices.

There are, for 2 by 2 matrices with *real* entries, either two real eigenvalues, one real eigenvalue with multiplicity two, or two complex eigenvalues. In the last case, the two complex eigenvalues must be conjugates of each other.

If you have λ, an eigenvector associated to an eigenvalue λ is then found by solving the linear system

$$(A - \lambda I)v = 0$$

(*since λ is a root of the characteristic equation*, there are an infinite number of solutions; we pick any nonzero one).

With an eigenvalue λ and eigenvector v found, we next pick *any* vector w with the property that the two vectors v and w are linearly independent. For example, if

$$v = \begin{pmatrix} a \\ b \end{pmatrix}$$

and a is not zero, we can take

$$w = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

(what would you pick for w is a were zero?). Now, since the set $\{v, w\}$ forms a basis (this is the key idea for all n: once you know v, you need to find $n-1$ other vectors to fill out a basis containing v) of two-dimensional space, we can find coefficients c and d so that

$$Aw = cv + dw\,. \tag{19}$$

We can summarize both (18) and (19) in one matrix equation:

$$A\,(v\,w) = (v\,w)\begin{pmatrix} \lambda & c \\ 0 & d \end{pmatrix}.$$

Here $(v\,w)$ denotes the 2 by2 matrix whose columns are the vectors v and w. To complete the construction, we let $S = (v\,w)$ and

$$\Lambda = \begin{pmatrix} \lambda & c \\ 0 & d \end{pmatrix}.$$

Then,

$$AS = S\Lambda$$

which is the same as what we wanted to prove, namely $A = S\Lambda S^{-1}$. Actually, we can even say more. It is a fundamental fact in linear algebra that, if two matrices are similar, then their eigenvalues must be the same. Now, the eigenvalues of Λ are λ and d, because the eigenvalues of any triangular matrix are its diagonal elements. Therefore, since A and Λ are similar, d must be also an eigenvalue of A.

The proof of Schur's theorem follows the same pattern, except for having fewer choices for v and w.

The Three Cases for $n = 2$

The following special cases are worth discussing in detail:

1. A has two different real eigenvalues.
2. A has two complex conjugate eigenvalues.
3. A has a repeated real eigenvalue.

In cases 1 and 2, one can always find a *diagonal* matrix Λ. To see why this is true, let us go back to the proof, but now, instead of taking just any linearly independent vector w, let us pick a special one, namely an eigenvector corresponding to the other eigenvalue of A:

$$Aw = \mu w\,.$$

This vector is always linearly independent of v, so the proof can be completed as before. Notice that Λ is now diagonal, because $d = \mu$ and $c = 0$.

To prove that v and w are linearly independent if they are eigenvectors for different eigenvalues, assume the contrary and show that it leads to a contradiction. Thus, suppose that $\alpha v + \beta w = 0$. Apply A to get

$$\alpha\lambda v + \beta\mu w = A(\alpha v + \beta w) = A(0) = 0.$$

On the other hand, multiplying $\alpha v + \beta w = 0$ by λ we would have $\alpha\lambda v + \beta\lambda w = 0$. Subtracting gives $\beta(\lambda - \mu)w = 0$, and as $\lambda - \mu \neq 0$ we would arrive at the conclusion that $\beta w = 0$. But w, being an eigenvector, is required to be nonzero, so we would have to have $\beta = 0$. Plugging this back into our linear dependence would give $\alpha v = 0$, which would require $\alpha = 0$ as well. This shows us that there are no nonzero coefficients α and β for which $\alpha v + \beta w = 0$, which means that the eigenvectors v and w are linearly independent.

Notice that in cases 1 and 3, the matrices Λ and S are both real. In case 1, we will interpret the solutions with initial conditions on the lines that contain v and w as "straight line solutions".

In case 2, the matrices Λ and S are, in general, not real. Note that, in case 2, if $Av = \lambda v$, taking complex conjugates gives

$$A\bar{v} = \bar{\lambda}\bar{v}$$

and we note that

$$\bar{\lambda} \neq \lambda$$

because λ is not real. So, we can always pick w to be the conjugate of v. It will turn out that solutions can be re-expressed in terms of trigonometric functions — remember example (13) — as we'll see in the next section.

Now let's consider Case 3 (the repeated real eigenvalue). We have that

$$\Lambda = \begin{pmatrix} \lambda & c \\ 0 & \lambda \end{pmatrix}$$

so we can also write $\Lambda = \lambda I + cN$, where N is the following matrix:

$$N = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

Observe that:

$$(\lambda I + cN)^2 = (\lambda I)^2 + c^2N^2 + 2\lambda cN = \lambda^2 I + 2\lambda cN$$

(because $N^2 = 0$) and, for the general power k, recursively:

$$\begin{aligned}(\lambda I + cN)^k &= \left(\lambda^{k-1}I + (k-1)\lambda^{k-2}cN\right)(\lambda I + cN) \\ &= \lambda^k I + (k-1)\lambda^{k-1}cN + \lambda^{k-1}cN + (k-1)\lambda^{k-2}a^2N^2 \\ &= \lambda^k I + k\lambda^{k-1}cN\end{aligned}$$

so

$$(\lambda I + cN)^k t^k = \left(\lambda^k I + k\lambda^{k-1}cN\right)t^k = \begin{pmatrix} \lambda^k t^k & k\lambda^{k-1}ct^k \\ 0 & \lambda^k t^k \end{pmatrix}$$

and therefore

$$e^{\Lambda t} = \begin{pmatrix} e^{\lambda t} & cte^{\lambda t} \\ 0 & e^{\lambda t} \end{pmatrix} \tag{20}$$

because $0 + ct + (2\lambda c)t^2/2 + (3\lambda^2 c)t^3/6! + \cdots = cte^{\lambda t}$. (This generalizes the special case in example (15).)

A Shortcut

If we just want to find the form of the general solution of $Y' = AY$, we do not need to actually calculate the exponential of A and the inverse of the matrix S.

Let us first take the cases of different eigenvalues (real or complex, that is, cases 1 or 2, it doesn't matter which one). As we saw, Λ can be taken to be the diagonal matrix consisting of these eigenvalues (which we call here

λ and μ instead of λ_1 and λ_2), and $S = (v\,w)$ just lists the two eigenvectors as its columns. We then know that the solution of every initial value problem $Y' = AY$, $Y(0) = Y_0$ will be of the following form:

$$Y(t) = e^{At}Y_0 = S\,e^{\Lambda t}\,S^{-1}Y_0 = (v\,w)\begin{pmatrix} e^{\lambda t} & 0 \\ 0 & e^{\mu t} \end{pmatrix}\begin{pmatrix} a \\ b \end{pmatrix} = a\,e^{\lambda t}v + b\,e^{\mu t}w$$

where we just wrote $S^{-1}Y_0$ as a column vector of general coefficients a and b. In conclusion: *The general solution of $Y' = AY$, when A has two eigenvalues λ and μ with respective eigenvectors v and w, is of the form*

$$a\,e^{\lambda t}v + b\,e^{\mu t}w \tag{21}$$

for some constants a and b. So, one approach to solving IVP's is to first find eigenvalues and eigenvectors, write the solution in the above general form, and then plug-in the initial condition in order to figure out what are the right constants.

In the case of non-real eigenvalues, recall that we showed that the two eigenvalues must be conjugates of each other, and the two eigenvectors may be picked to be conjugates of each other. Let us show now that we can write (21) in a form which does not involve any complex numbers. In order to do so, we start by decomposing the first vector function which appears in (21) into its real and imaginary parts:

$$e^{\lambda t}v = Y_1(t) + iY_2(t) \tag{22}$$

(let us not worry for now about what the two functions Y_1 and Y_2 look like). Since μ is the conjugate of λ and w is the conjugate of v, the second term is:

$$e^{\mu t}w = Y_1(t) - iY_2(t)\,. \tag{23}$$

So we can write the general solution shown in (21) also like this:

$$a(Y_1 + iY_2) + b(Y_1 - iY_2) = (a+b)Y_1 + i(a-b)Y_2\,. \tag{24}$$

Now, it is easy to see that a and b must be conjugates of each other. (Do this as an optional homework problem. Use the fact that these two coefficients are the components of $S^{-1}Y_0$, and the fact that Y_0 is real and that the two columns of S are conjugates of each other.) This means that *both coefficients $a+b$ and $i(a-b)$ are real numbers.* Calling these coefficients "k_1" and "k_2", we can summarize the complex case like this: *The general solution of $Y' = AY$, when A has a non-real eigenvalue λ with respective eigenvector v, is of the form*

$$k_1\,Y_1(t) \;+\; k_2\,Y_2(t) \tag{25}$$

for some real constants k_1 and k_2. The functions Y_1 and Y_2 are found by the following procedure: calculate the product $e^{\lambda t}v$ and separate it into real and imaginary parts as in Equation (22). What do Y_1 and Y_2 really look like? This is easy to answer using Euler's formula, which gives

$$e^{\lambda t} = e^{\alpha t + i\beta t} = e^{\alpha t}(\cos\beta t + i\sin\beta t) = e^{\alpha t}\cos\beta t + ie^{\alpha t}\sin\beta t$$

where α and β are the real and imaginary parts of λ respectively.

Finally, in case 3 (repeated eigenvalues) we can write, instead:

$$\begin{aligned} Y(t) = e^{At}Y_0 = S\,e^{\Lambda t}\,S^{-1}Y_0 &= (v\,w)\begin{pmatrix} e^{\lambda t} & cte^{\lambda t} \\ 0 & e^{\lambda t} \end{pmatrix}\begin{pmatrix} a \\ b \end{pmatrix} \\ &= a\,e^{\lambda t}v + b\,e^{\lambda t}(ctv + w)\,. \end{aligned}$$

When $c = 0$ we have from $A = S\Lambda S^{-1}$ that A must have been the diagonal matrix

$$\begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix}$$

to start with (because S and Λ commute). When $c \neq 0$, we can write $k_2 = bc$ and redefine w as $\frac{1}{c}w$. Note that then (19) becomes $Aw = v + \lambda w$, that is, $(A - \lambda I)w = v$. Any vector w with this property is linearly independent from v (why?).

So we conclude, for the case of repeated eigenvalues: *The general solution of $Y' = AY$, when A has a repeated (real) eigenvalue λ is either of the form $e^{\lambda t}Y_0$ (if A is a diagonal matrix) or, otherwise, is of the form*

$$k_1 \, e^{\lambda t} v \; + \; k_2 \, e^{\lambda t}(tv + w) \tag{26}$$

for some real constants k_1 and k_2, where v is an eigenvector corresponding to λ and w is any vector which satisfies $(A - \lambda I)w = v$. Observe that $(A - \lambda I)^2 w = (A - \lambda I)v = 0$. general, one calls any nonzero vector such that $(A - \lambda I)^k w = 0$ a *generalized eigenvector* (of order k) of the matrix A (since, when $k = 1$, we have eigenvectors).

Forcing Terms

The use of matrix exponentials also helps explain much of what is done in chapter 4 (forced systems), and renders Laplace transforms unnecessary. Let us consider non-homogeneous linear differential equations of this type:

$$\frac{dY}{dt}(t) = AY(t) \, + \, u(t) \, . \tag{27}$$

We wrote the arguments "t" just this one time, to emphasize that everything is a function of t, but from now on we will drop the t's when they are clear from the context.

Let us write, *just as we did when discussing scalar linear equations*, $Y' - AY = u$. We consider the "integrating factor" $M(t) = e^{-At}$. Multiplying both sides of the equation by M, we have, since $(e^{-At}Y)' = e^{-At}Y' - e^{-At}AY$ (*right?*):

$$\frac{de^{-At}Y}{dt} = e^{-At}u \, .$$

Taking antiderivatives:

$$e^{-At}Y = \int_0^t e^{-As}u(s)\,ds \, + \, Y_0$$

for some constant vector Y_0. Finally, multiplying by e^{-At} and remembering that $e^{-At}e^{At} = I$, we conclude:

$$Y(t) = e^{At}Y_0 \, + \, e^{At}\int_0^t e^{-As}u(s)\,ds \, . \tag{28}$$

This is sometimes called the "variation of parameters" form of the general solution of the forced equation (27). Of course, $Y_0 = Y(0)$ (just plug-in $t = 0$ on both sides).

Notice that, if the vector function $u(t)$ is a polynomial in t, then the integral in (28) will be a combination of exponentials and powers of t (integrate by parts). Similarly, if $u(t)$ is a combination of trigonometric functions, the integral will also combine trigonometric functions and polynomials. This observation justifies the "guesses" made for forced systems in chapter 4 (they are, of course, not guesses, but consequences of integration by parts).

Problems

1. In each of the following, factor the matrix A into a product $S\Lambda S^{-1}$, with Λ diagonal:

a. $\quad A = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}$

$$\text{b.} \qquad A = \begin{pmatrix} 5 & 6 \\ -1 & -2 \end{pmatrix}$$

$$\text{c.} \qquad A = \begin{pmatrix} 2 & -8 \\ 1 & -4 \end{pmatrix}$$

$$\text{d.} \qquad A = \begin{pmatrix} 2 & 2 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & -1 \end{pmatrix}$$

2. For each of the matrices in Exercise 1, use the $S\Lambda S^{-1}$ factorization to calculate A^6 (do *not* just multiply A by itself).

3. For each of the matrices in Exercise 1, use the $S\Lambda S^{-1}$ factorization to calculate e^{At}.

4. Calculate e^{At} for this matrix:

$$\begin{pmatrix} 0 & 1 & 2 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

   using the power series definition.

5. Consider these matrices:

$$A = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} \quad B = \begin{pmatrix} 0 & -1 \\ 0 & 0 \end{pmatrix}$$

   and calculate e^{At}, e^{Bt}, and $e^{(A+B)t}$.

   Answer, true or false: is $e^{At}e^{Bt} = e^{(A+B)t}$?

6. (Challenge problem) Show that, for any two matrices A and B, it is true that

$$e^{At}e^{Bt} = e^{(A+B)t} \text{ for all } t$$

   if and only if $AB - BA = 0$. (The expression "$AB - BA$" is called the "Lie bracket" of the two matrices A and B, and it plays a central role in the advanced theory of differential equations.)