MATH 5587 LECTURE LOG

Lecture 1, 9/5

In the first lecture, the history of the subject was briefly discussed and a few examples of PDEs were mentioned. One of the important examples is the heat equation, which, in one dimension, reads

$$\frac{\partial u}{\partial t} = k \frac{\partial^2 u}{\partial x^2} . \tag{1}$$

We started the derivation of the equation (Section 1.2 of the textbook).

Independently of the derivation of the equation, we did the following calculation, based on an application of the chain rule: Assume a function $u(x,t)$ satisfies the heat equation in the domain $\{(x,t)\,, x \in (a,b)\,,\, t \in (t_1,t_2)\}$, and let λ be a positive number. Then the function $u(\frac{x}{\lambda}, \frac{t}{\lambda^2})\}$ satisfies the heat equation in the domain $\{(x,t)\,, x \in (\lambda a, \lambda b)\,,\, t \in (\lambda^2 t_1, \lambda^2 t_2)\}$. This is an important property of the heat equation, which is very useful to keep in mind for practical applications. I recommend that you go through this calculation in some detail.

Here is a somewhat more difficult calculation you can do as an optional exercise to practice the chain rule. Assume $u(x,t)$ satisfies the heat equation (1) in the domain $(-\infty,\infty)\times(0,\infty)$. We will think of $u(x,t)$ as temperature. We now watch the temperature from another coordinate system $(\tilde{x},\tilde{t})$, which is related to the system (x,t) by

$$\tilde{x} = x - ct\,, \qquad \tilde{t} = t\,, \tag{2}$$

where we can think of c as a velocity. In the new coordinates, the temperature will be described by

$$\tilde{u}(\tilde{x},\tilde{t}) = u(\tilde{x} + c\tilde{t}, \tilde{t})\,. \tag{3}$$

What is the equation satisfied by $\tilde{u}$, in the coordinates $\tilde{x}, \tilde{t}$? If $c \neq 0$, it will not be the original heat equation, the motion of the coordinate system will introduce a new term into the equation. It is instructive to consider also the case $k = 0$.

Lecture 2, 9/7

We finished the derivation of the heat equation (in dimension 1), and discussed the role of the boundary conditions (Section 1.3 in the textbook.) We calculated the steady (= time-independent) solutions of

$$\frac{\partial u}{\partial t} = k\frac{\partial^2 u}{\partial x^2} + f \tag{4}$$

in $(0,L)$ with constant f (i. e. independent of x and t), with the boundary conditions $u(0,t) = u(L,t) = 0$ and also the boundary conditions

$$u(0,t) = 0\,, \quad \frac{\partial u}{\partial x}(L,t) = 0\,,$$

(one insulated end). We also discussed the situation when both ends are insulated, i. e.

$$\frac{\partial u}{\partial x}(0,t) = \frac{\partial u}{\partial x}(L,t) = 0$$

when a steady solution cannot exist for a constant f unless $f = 0$. (Note that when $f = 0$ and the boundary is insulated, then any constant is a steady-state solution, so the steady state is not unique.)

For solutions on the whole real line ($x \in (-\infty, \infty)$) we discussed (as an optional material which may not be in the textbook) the formal formula

$$u(x,t) = e^{tk\partial_x^2} u_0(x) = \left(I + \frac{tk\partial_x^2}{1!} + \frac{(tk\partial_x^2)^2}{2!} + \frac{(tk\partial_x^2)^3}{3!} + \dots \right) u_0(x)\,. \qquad (5)$$

We noticed that the series on the right-hand side is finite when $u_0(x)$ is a polynomial, and hence in this case the formula gives a polynomial in x and t which solves the heat equation (1) and satisfies $u(x,0) = u_0(x)$.

A simple but useful exercise concerning the boundary conditions is the following. If u solves (1) in $(0, L)$ and both ends are insulated, then the quantity

$$U = \frac{1}{L} \int_0^L u(x,t)\, dx$$

is independent of time. As the integral is proportionate to the thermal energy in the rod and both ends are insulated, this is to be expected, but it is important to confirm it directly from the equation.

You can also try to show that in the same situation the quantity

$$\int_0^L (u(x,t) - U)^2 \, dx$$

is decreasing, which is consistent with the intuitive expectation that u should be approaching U as $t \to \infty$ (when the ends are insulated). This calculation is more difficult, one has to use integration by parts.

Lecture 3, 9/12

We essentially went through the material in Sections 2.1 – 2.5. in the textbook.

In addition, we compared the heat equation in the interval $(0, L)$ with the boundary conditions $u(0,t) = u(L,t) = 0$ with the following situation in the theory of ordinary differential equations.[1]

[1] The material below can be at this point considered as optional for this class, but I believe it is good to understand it, as it is a finite-dimensional version of the (infinite-dimensional) PDE situation we are dealing with in connection with the heat equation (and it applies to other equations, too).

Consider a system of n linear differential equation for variables $x_1(t), \ldots, x_n(t)$ given by

$$\begin{array}{rcl} \dot{x}_1 & = & a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \\ \dot{x}_2 & = & a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \\ & \cdots & \\ \dot{x}_n & = & a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n \,, \end{array} \tag{6}$$

where the matrix $A = (a_{ij})_{i,j=1,2,\ldots,n}$ is considered as given. We will write (6) in a compact form as

$$\dot{x} = Ax \,, \tag{7}$$

where x is considered as a function of variable t with values in $\mathbf{R}^n$. We should think of x as a column vector, so that Ax is a compact notation for the product

$$\begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \ldots \\ x_n \end{pmatrix} . \tag{8}$$

Assume now that the matrix A is symmetric, i. e. $a_{ij} = a_{ji}$, $i, j = 1, 2, \ldots, n$. An important theorem in linear algebra says that there is an orthonormal basis of $\mathbf{R}^n$ in which A is diagonal. Let us denote the vectors of this basis as $b^{(1)}, b^{(2)}, \ldots, b^{(n)}$. (We think of these as column vectors.) In this basis the matrix A becomes

$$\begin{pmatrix} \lambda_1 & 0 & 0 & \ldots & 0 \\ 0 & \lambda_2 & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \\ 0 & 0 & 0 & \ldots & \lambda_n \end{pmatrix} , \tag{9}$$

where $\lambda_1, \lambda_2, \ldots, \lambda_n$ are real numbers.[2] This means that that

$$Ab^{(k)} = \lambda_k b^{(k)} \,, \quad k = 1, 2, \ldots, n \,. \tag{10}$$

We can change variables in equation (7) to $y_1, y_2, \ldots, y_n$ by setting

$$x = y_1 b^{(1)} + y_2 b^{(2)} + \cdots + y_n b^{(n)} \,. \tag{11}$$

In the new variables y_j the equation (7) is[3]

$$\begin{array}{rcl} \dot{y}_1 & = & \lambda_1 y_1 \,, \\ \dot{y}_2 & = & \lambda_2 y_2 \,, \\ & \cdots & \\ \dot{y}_n & = & \lambda_n y_n \,, \end{array} \tag{12}$$

[2] In many software packages, such as Matlab or Mathematica you have matrix functions which find both the eigenvectos and the eigenvalues of a given matrix A fairly quickly and with good precision. For having $\lambda_1, \ldots, \lambda_n$ real and $b^{(1)}, \ldots b^{(n)}$ mutually orthogonal, the assumption that A be symmetric is important.

[3] This can be seen for example by replacing by substituting the right-hand side of (11) into equation (7).

for which one can easily write down the general solution:

$$\begin{array}{rcl} y_1 & = & c_1 e^{\lambda_1 t}\,, \\ y_2 & = & c_2 e^{\lambda_2 t}\,, \\ & \dots & \\ y_n & = & c_n e^{\lambda_n t}\,, \end{array} \tag{13}$$

where $c_1, c_2, \dots, c_n$ are constants. Hence from (11) we see that

$$x = c_1 e^{\lambda_1 t} b^{(1)} + c_2 e^{\lambda_2 t} b^{(2)} + \dots + c_n e^{\lambda_n t} b^{(n)}\,. \tag{14}$$

This should be compared with expressions of the form

$$u(x,t) = B_1 e^{\lambda_1 t}\phi^{(1)}(x) + B_2 e^{\lambda_2 t}\phi^{(2)}(x) + \dots \tag{15}$$

for the solutions of the heat equation discussed for example in subsection 2.3.5 of the textbook (and which we also discussed in class).

There is a finite-dimensional approximation of the heat equation which is of the form (7). (This is related to numerical methods discussed in the textbook in Chapter 6.) To consider one of the simplest finite-dimensional approximations of the heat equation in $(0, L)$ with the zero boundary condition on both ends, let us choose a natural number N and set

$$x_0 = 0,\, x_1 = \frac{L}{N},\, x_2 = \frac{2L}{N},\, \dots,\, x_N = L\,, \tag{16}$$

and also

$$U_1(t) = u(x_1, t),\;\; U_2(t) = u(x_2, t),\;\; \dots, U_{N-1}(t) = u(x_{n-1}, t)\,. \tag{17}$$

If we approximate the second derivative $\partial^2 u/\partial x^2$ at x_j by a difference quotient, such as[4]

$$\frac{\partial^2 u(x_j,t)}{\partial x^2} \sim \frac{U_{j+1}(t) - 2U_j(t) + U_{j-1}(t)}{h^2}\,, \quad h = \frac{L}{N}\,, \tag{18}$$

the heat equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} \tag{19}$$

can be approximated as

$$\dot{U} = AU\,, \tag{20}$$

where A is an $(N-1)\times(N-1)$ matrix given by

$$A = h^{-2}\begin{pmatrix} -2 & 1 & 0 & 0 & \dots & 0 \\ 1 & -2 & 1 & 0 & \dots & 0 \\ 0 & 1 & -2 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \\ 0 & 0 & 0 & \dots & -2 & 1 \\ 0 & 0 & 0 & \dots & 1 & -2 \end{pmatrix} \tag{21}$$

[4]The one we use here is perhaps the simplest one which works, but it is not the only one which can be used. Higher precision can be achieved by more sophisticated approximations.

As an optional exercise, you can check with Matlab or Mathematica how the eigenvectors and eigenvalues of this matrix look for some values of N (such as $N = 100$ or $N = 200$). It turns out that in this example one can also compute the eigenvectors explicitly, without a computer, they are still given by $U_j^{(m)} = \sin(\tilde{\lambda}_m x_j)$ for suitable $\tilde{\lambda}_m$. (This is also discussed in Sections 6.3.5 and 6.3.6 of the textbook, with a slightly different notation.)

In Matlab or Mathematica (and other software packages) one can solve (20) by using the matrix function e^{tA}, although this may not be the best way of doing it once N becomes very large. But for $N \sim 100$ or so (which may be already adequate for the heat equation in some situations) one does not have to worry too much about the efficiency, a standard PC is fast enough to allow us to ignore this issue in that case. Of course, this changes when we are in higher dimensions or when N is much larger.

Lecture 4, 9/14

We discussed the orthogonality of functions (loosely along the lines of the *Appendix to 2.3* on pages 54 and 56 in the textbook), and then various form of Fourier series, see the table on page 65 of the textbook.

The textbook uses the convention in which the Fourier series for periodic functions with period $2L$ is written as

$$\begin{aligned} f(x) &= a_0 + a_1 \cos\frac{\pi x}{L} + b_1 \sin\frac{\pi x}{L} + a_2 \cos\frac{2\pi x}{L} + b_2 \sin\frac{2\pi x}{L} + \dots \\ &= a_0 + \sum_{n=1}^{\infty} \left(a_n \cos\frac{\pi n x}{L} + b_n \sin\frac{\pi n x}{L} \right) . \end{aligned} \tag{22}$$

The sine series, which we used to express solutions of the heat equation in $(0, L)$ which vanish at the endpoints, can be thought of as a special case if (22), when f is odd, i. e. $f(-x) = -f(x)$. (In that case, if f is originally defined only on $(0, L)$, we can first extend it as an odd function to $(-L, L)$ and then as a $2L-$periodic function.)
Similarly, the cosine series (used in the textbook to express the solution in a rod with insulated ends), can be thought of as a special case of (22), when f is even, i. e. $f(x) = f(-x)$.

An often-used from of the Fourier series (discussed in Section 3.6 of the textbook) is

$$f(x) = \sum_{n=-\infty}^{\infty} c_n e^{\frac{\pi i n x}{L}} . \tag{23}$$

Here we again think of f as a $2L-$periodic function, and we chose the normalization of the coefficients which matches that of (22). In this representation the

coefficients c_n are obtained from f as

$$c_n = \frac{1}{2L}\int_{-L}^{L} f(x)e^{-\frac{\pi i n x}{L}}\,dx\,. \tag{24}$$

The form (23) is in some sense the most natural form of the Fourier series, at least in the context of periodic functions. One of its advantages is that is taking derivatives becomes particularly simple in terms of the Fourier coefficients c_n: The operation

$$f \to \frac{\partial f}{\partial x} \tag{25}$$

becomes just

$$c_n \to \frac{\pi i n}{L}\,c_n \tag{26}$$

on the Fourier side. The relation between (23) and (22) can be seen either by using the expression

$$e^{i\theta} = \cos\theta + i\sin\theta \tag{27}$$

in (23), or the expressions

$$\cos\theta = \frac{e^{i\theta}+e^{-i\theta}}{2}\,, \qquad \sin\theta = \frac{e^{i\theta}-e^{-i\theta}}{2i}\,. \tag{28}$$

in (22). A simple calculation then gives

$$c_0 = a_0\,, \quad c_n = \frac{a_n - ib_n}{2}\,, \quad c_{-n} = \frac{a_n+ib_n}{2}\,, \quad n = 1,2,3,\dots\,, \tag{29}$$

or

$$a_0 = c_0, \quad a_n = c_n + c_{-n}\,, \quad b_n = i(c_n - c_{-n})\,, \quad n = 1,2,3\dots \tag{30}$$

Note that f is real-valued if only if $c_{-n} = \bar{c}_n$ (where the bar denotes the complex conjugation, as usual). The condition that f be odd can be expressed as $c_{-n} = -c_n$ and the condition for f be even can be expressed as $c_{-n} = c_n$.
The form (23) of the Fourier series is also important from the point of view of the Fast Fourier transform, or FFT, which is used in computing for Fourier series. It is one of the most important algorithms in computing, discovered in 1964, see the original paper by Cooley and Tuckey. It appears that Gauss was aware of a similar algorithm already in 1805.

Lecture 5, 9/19

In the beginning of the class we discussed a simple Matlab program for solving the heat equation. We noted that the method used in the program can be applied to quite general equations with constant coefficients, if the problem can be re-formulated in a way suitable for periodic boundary conditions.
As an optional exercise, you can try to apply the program to solve the heat equation backwards, i. e. try to calculate from the data at $t = 0$ the solution

at an earlier time $t = -0.1$, say. You will see that once the number of active modes is not very small, the calculation going backward will "blow-up". This has to do with the fact that solving the heat equation backward in time is an ill-posed problem. The trouble comes from the fact that the exponentials in the formula for the solution can become extremely large if we go to negative times. We also discussed adaptations of the method used for the heat equation to other equations, such as the Schrödinger equation

$$\frac{\partial u}{\partial t} = i\frac{\partial^2 u}{\partial x^2}\,. \tag{31}$$

As an optional (but very useful) exercise, you can think about how one would modify the program so that it would solve the following more general equation for complex-valued functions u:

$$\frac{\partial u}{\partial t} = a\frac{\partial u}{\partial x} + b\frac{\partial^3 u}{\partial x^3} + (k + ci)\frac{\partial^2 u}{\partial x^2}\,, \tag{32}$$

where a, b, c are real numbers and $k > 0$.
The program can also be easily modified to solve the wave equation

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2}\,, \tag{33}$$

which is discussed in Chapter 4.4 of the textbook. The change to the wave equation is somewhat more subtle, as the wave equation is of the second orger in t. It means the the ODE we have to solve for the Fourier coefficients will be of the second order, and we need to specify two quantities to determine it. This reflects the fact that for the wave equation we need to know $u(x, 0)$ *and* $\frac{\partial u}{\partial t}(x, 0)$ to determine the evolution.
The method of separation of variables for the Laplace equation in a rectangle, discussed in Chapter 2.5.1 of the textbook, can also be put in this framework, and we briefly discussed this. The main difference for the Laplace equation is that the corresponding problem we have to solve for the ODE we get for the Fourier modes is a *boundary value problem*[5], rather then the *initial value problem*[6] which appears naturally for the wave equation. See for example the calculation on page 71 in the textbook.
At the end of the lecture we discussed briefly some issues related to the Discrete Fourier Transformation (DFT). This material is optional.

Lecture 6, 9/21, 2017

In this lecture we discussed the material in 2.5.2 and 2.5.3 in the textbook. For the problem discussed in 2.5.3 (the flow outside a circular cylinder) we discussed

[5] A typical ODE boundary value problem is: find a function y on $(0, 1)$ such that $y'' = 4y\,,\ y(0) = 1\,,\ y(1) = 10$.
[6] A typical initial-value problem is: find $y'' = -4y$, with $y(0) = 1, y'(0) = 10\,.$

the special case when the circulation vanishes. In this case the result is that the drag force on a ball moving through an ideal fluid at constant speed vanishes. This results, known as d'Alembert's paradox was derived by d'Alembert around 1751, some years before the equations for the motion of the ideal fluid were derived in a classical paper by Euler published in 1757.
The problem of determining the drag force in real fluids is much more complicated, and there are still many open mathematical problems around it, such as the regularity problem for the Navier-Stokes equations. Classical experiments concerning counter-intuitive behavior of the drag force in certain regimes can be seen in this old video (around 6:55) and also its part 2 (around 3:18). There are many interesting connections to aerodynamics of sports balls, such as the tricky behavior of knuckleballs or free kicks.

Lecture 7, 9/26/2017
(Mean value property of harmonic functions; some topics in Discrete Fourier series)

In the first part of the lecture, we discussed topics from subsection 2.5.4 in the textbook,"Qualitative Properties of the Laplace equation". In particular, we discussed the mean value property of the harmonic functions. In dimension two the mean value property has a simple proof based on the Fourier representation of the solutions, see the short proof on page 79 of the textbook.
The mean value property of harmonic functions holds in any dimension and is closely related to the Shell Theorem in Newtonian gravity (and also electrostatics): If we have a spherical star of a finite radius $R > 0$ centered at the origin and the distribution of the mass inside the star is spherically symmetric, then the gravitational field outside of the star (at distances $> R$ from the origin) is exactly the same as the gravitational field we would get from concentrating all the mass of the star to the origin.[7]
The connection of such statements with the theory of the harmonic functions in the three-dimensional space comes from the observation of P.-S. Laplace (made before 1800) that the function $u(x) = \frac{1}{|x|}$, where $x = (x_1, x_2, x_3)$ and $|x| = \sqrt{x_1^2 + x_2^2 + x_3^2}$ satisfies the Laplace equation

$$\Delta u = 0 \tag{34}$$

in $\mathbf{R}^3 \setminus \{0\}$. In dimension two we have a similar result for the function $u(x) = \log|x|$, where $x = (x_1, x_2)$ and $|x| = \sqrt{x_1^2 + x_2^2}$. These topics are related to Green's functions, which we will study later, and are optional for now.
The connection to the mean value theorem is as follows. Consider two thin shells centered as the origin, with radii $0 < R_1 < R_2$. Assume that the shells have some small finite thickness $\varepsilon > 0$ (small compared to R_1). Let us consider a smooth distribution ρ_1^ε of a positive electric charge in the shell at distance $\sim R_1$,

[7]The statement of the shell theorem is slightly more general, in that it also makes a statement about the field inside the star, but we will not need this.

with total charge q (in some units). We think of ρ_1^ε as a smooth function which depends only on the distance from the origin, and vanishes everywhere outside the thin shell at $\sim R_1$. Similarly, we can think of a spherically symmetric distribution of negative charge at distance $\sim R_2$ from the origin, with total change $-q$, and assume it is described by density $-\rho_2^\varepsilon$, which we think about as a smooth, spherically symmetric functions vanishing outside a small neighborhood the shell of radius R_2.
Let v be the electrostatic potential generated by the distribution of charge given by $\rho_1^\varepsilon - \rho_2^\varepsilon$. We can think of v as $v = v_1 - v_2$, where v_1 is the field generated by the charge ρ_1^ε and v_2 is the field generated by the charge ρ_2^ε. There are now 3 main points from which one can see that the mean-value theorem should be true:

(i) The potential v satisfies

$$-\Delta v = \kappa(\rho_1^\varepsilon - \rho_2^\varepsilon)\,, \tag{35}$$

where $\kappa > 0$ is a suitable constant depending on the choice of units.

(ii) The potential v vanishes outside of the ball $B_{R_2+\varepsilon}$. This is perhaps the most surprising point, and it follows from the discussion above concerning the Shell Theorem.

(iii) When f, g are smooth functions in $\mathbf{R}^3$ and g vanishes outside of some bounded set, then

$$\int_{\mathbf{R}^3} f(x)\Delta g(x)\,dx = \int_{\mathbf{R}^3} (\Delta f(x))g(x)\,dx. \tag{36}$$

This is obtained by integration by parts.

If now u is a harmonic function, i. e. $\Delta u = 0$ and v is the potential above, we obtain from (iii) (which can be applied, because v vanishes outside of $B_{R_2+\varepsilon}$)

$$\int_{\mathbf{R}^3} u(x)\Delta v(x)\,dx = 0\,, \tag{37}$$

which is the same as

$$\int_{\mathbf{R}^3} u(x)\rho_1^\varepsilon(x)\,dx = \int_{\mathbf{R}^3} u(x)\rho_2^\varepsilon(x)\,dx\,. \tag{38}$$

Taking $\varepsilon \to 0_+$ and then $R_1 \to 0_+$, we obtain the mean value theorem.
Our point here is not present a precise proof, but the give some idea why the mean value property of the harmonic functions can be connected to properties of the gravitational or electrostatic potential which people understood early on.

Representing functions in a computer

In the second part of the lecture we started discussing the Fourier series, which are the main topic of Chapter 3 of the textbook. In fact, the topic with which we started off concerns the Discrete Fourier Transformation, which we discussed in connection with the Matlab code for the solution of the heat equation. This is not in the textbook, but it is a topic which is important for Matlab and other practical purposes, so we will discuss it, even though the material can be considered as optional.

We will indentify $2\pi-$periodic functions on $\mathbf{R}$ with functions on the unit circle S^1 in the complex plane, and will use the notation

$$z = e^{i\theta} \,. \tag{39}$$

We consider N points uniformly distributed over the circle as follows. We let $w = e^{\frac{2\pi}{N}}$ and set

$$z_k = w^k \,, \quad k = 0, 1, 2, \dots N-1 \,. \tag{40}$$

When can now represent functions f on the circle $\mathbf{S}^1$ in two different ways:[8]

(i) Represent f by its values its the points $z_0, z_1, \dots z_{N-1}$.
Let us use the notation

$$f_0 = f(z_0) \,,\, f_1 = f(z_1) \,,\, f_2 = f(z_2) \,,\, \dots, f_{N-1} = f(z_{N-1}) \,. \tag{41}$$

This means that we represent f by the vector $(f_0, f_1, f_2, \dots, f_{N-1})$. We will write, with some abuse of notation,

$$f \sim (f_0, f_1, f_2, \dots, f_{N-1}) \,. \tag{42}$$

Of course, from the point of view of Calculus, such a description of a function is not complete, as we did not say what the values of f are at the points which are not in our finite set $z_0, z_1, \dots, z_{N-1}$.
Nevertheless, in practice the values $f_0, f_1, f_2, \dots, f_{N-1}$ may be the only "measurements" we have about the function f, and we often represent function in a computer by such vectors. Note that in this representation we describe f by N numbers.

(ii) Represent f by a (finite) Fourier series, or, equivalently, as a polynomial in z.
Let us start with the representation of f as a polynomial in z and show that it is the same as a finite Fourier series later. Our representation will be of the form

$$f(z) = c_0 + c_1 z + c_2 z^2 + \dots + c_{N-1} z^{N-1} \,. \tag{43}$$

[8] There are of course many more ways to represent functions than the two discussed here. Deciding which representations are best in various situations and what the errors are is in fact a huge topic which has been studied in great depth in Numerical Analysis, Signal Processing and other areas .

In this representation f is again represented by N numbers, this time they are $c_0, c_1, c_2, \ldots, c_{N-1}$. Note that in this representation f looks as a genuine function from the point of view of Calculus: we have a precise rule which apparently for any given z gives us the value of $f(z)$. However, in practice the terms $z^{N-1}, z^{N-2}, \ldots, z^{N-l}$ for l up to $\sim N/2$ should be interpreted as $z^{-1}, z^{-2}, \ldots z^{-l}$, because we really have in mind the truncation $\sum_{-N/2}^{N/2} c_k e^{ik\theta}$ of the Fourier series $\sum_{k=-\infty}^{\infty} c_k e^{ik\theta}$. This does not make a difference on our points $z_0, z_1, z_2, \ldots z_{N-1}$, as $z_j^N = 1$, but does make a difference at most other points. Therefore one has to be somewhat careful with the interpretation of the polynomial (43) for z outside of the set $\{z_0, z_1, \ldots, z_{N-1}\}$.

Transformation between the two representations

How are the two representations of f connected? Given the vector $f_0, f_1, \ldots, f_{N-1}$, there are many different functions f on the circle for which $f(z_k) = f_k$, $k = 0, 1, 2, \ldots N-1$. However, there is precisely one polynomial of degree less or equal to $(N-1)$ which has this property. So if we assume that f is a polynomial of degree at most $N-1$, the vector $f_0, f_1, f_2, \ldots, f_{N-1}$ determines the value $f(z)$ uniquely at any point z. The transformation between the $(f_0, \ldots, f_{N-1})$ representation and the $(c_0, \ldots, c_{N-1})$ representation of such polynomials is precisely the Discrete Fourier Transformation.

Recalling (40), we can write

$$\begin{array}{rcccccccccc} f_0 & = & c_0 & + & c_1 & + & c_2 & + & \ldots & + c_{N-1} \\ f_1 & = & c_0 & + & c_1 w & + & c_2 w^2 & + & \ldots & + c_{N-1} w^{N-1} \\ f_2 & = & c_0 & + & c_1 w^2 & + & c_2 w^4 & + & \ldots & + c_{N-1} w^{2(N-1)} \\ \ldots & \ldots & \ldots & & & & & & & \\ f_{N-1} & = & c_0 & + & c_1 w^{N-1} & + & c_2 w^{2(N-1)} & + & \ldots & + c_{N-1} w^{(N-1)(N-1)} \end{array} . \tag{44}$$

This is the same as

$$\begin{pmatrix} f_0 \\ f_1 \\ f_2 \\ \ldots \\ f_{N-1} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & \ldots & 1 \\ 1 & w & w^2 & \ldots & w^{N-1} \\ 1 & w^2 & w^4 & \ldots & w^{2(N-1)} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & w^{N-1} & w^{2(N-1)} & \ldots & w^{(N-1)(N-1)} \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \\ c_2 \\ \ldots \\ c_{N-1} \end{pmatrix} \tag{45}$$

Denoting the $N \times N$ matrix on the right-hand side by $A = A(w, N)$ and slightly abusing notation by using f for the vector on the left-hand-side and c for the vector on the right-hand side, we can abbreviate the last identity as

$$f = A\, c\,. \tag{46}$$

This can be thought of as the "discrete Fourier representation" of the vector f, and gives a rule for calculating f from c. The matrix A is invertible, with

$$A^{-1} = \frac{1}{N}\, \overline{A}\,, \tag{47}$$

where the bar denotes the complex conjugation.[9] (Each entry of $\overline{A}$ is the complex conjugate of the corresponding entry of A.) Hence c can be calculated from f by

$$c = A^{-1} f = \frac{1}{N}\, \overline{A}\, f\,. \tag{48}$$

As already mentioned in Lecture 4, for $N = 2^m$ the matrix multiplications in (47) and (48) can be done in about $N \log N$ steps with the Fast Fourier Transform algorithm. This has far-reaching consequences in applications. In Matlab these operations are performed by the `fft` and `ifft` commands, although the normalization is a bit different from the one used above: `fft(u)` calculates $\overline{A}u$ and `ifft(u)` calculates $\frac{1}{N}Au$.

Comparison with the classical Fourier series in the textbook

To compare the representation (43) with the representation

$$f(\theta) = \sum_{k=-\infty}^{k=\infty} c_k e^{ik\theta} \tag{49}$$

we discussed in Lecture 4, will write $z = e^{i\theta}$ and recall that $w = e^{\frac{2\pi i}{N}}$. This means that $w^N = 1$. Hence if we evaluate f only on $\theta_0 = 0\,,\, \theta_1 = \frac{2\pi}{N}\,,\, \theta_2 = \frac{2\cdot 2\pi}{N}\,,\, \theta_3 = \frac{3\cdot 2\pi}{N}\,, \ldots,$ it is natural to replace the infinite series (49) by a finite sum of N terms. Moreover, still assuming we only evaluate f at $\theta_0, \theta_1, \ldots, \theta_{N-1}$, in the expression

$$f(\theta) = c_0 + c_1 e^{i\theta} + c_2 e^{2i\theta} + \cdots + c_{N-1} e^{(N-1)i\theta} \tag{50}$$

we can do the following replacements

$$\begin{array}{lll} c_{N-1}e^{(N-1)i\theta} \;\sim\; c_{-1}e^{-i\theta}\,, & & c_{-1} = c_{N-1} \\ c_{N-2}e^{(N-2)i\theta} \;\sim\; c_{-2}e^{-2i\theta}\,, & & c_{-2} = c_{N-2} \\ \cdots & & \cdots \\ c_{N-l}e^{(N-l)i\theta} \;\sim\; c_{-l}e^{-li\theta}\,, & & c_{-l} = c_{N-l}\,. \end{array} \tag{51}$$

It clearly makes sense to do it up only up to $l \sim \frac{N}{2}$. One should keep this "conversion table" in mind when using the `fft` and `ifft` functions in Matlab.

Lecture 8, 9/28/2017
(More on Fourier series)

We continued to discuss the Fourier series (Section 3 of the textbook). We focused on the complex form on $(-\pi, \pi)$ (or on the unit circle)

$$f(\theta) = \sum_{k=-\infty}^{\infty} c_k e^{ik\theta} \tag{52}$$

[9] This is in fact not difficult to prove, and for the mathematically inclined students this may be a good exercise.

discussed in section 3.6, see also Lecture 4. Recall that the coefficient c_k can be calculated from

$$c_k = \frac{1}{2\pi} \int_{\pi}^{\pi} f(\theta)\, e^{-ik\theta}\, d\theta\,. \tag{53}$$

One of the main issues in the study of the Fourier series is their convergence. The main theorem discussed in the textbook in this direction is the theorem on page 89 concerning point-wise convergence of the Fourier series of a piece-wise smooth function.

A simple rule of thumb for the convergence of the series that it converges fast for smooth functions and not so fast for non-smooth function. For example, if a function is not continuous *as a periodic function*, the Fourier series cannot converge absolutely, in the sense that any discontinuity of the function implies for its Fourier series that

$$\sum_k |c_k e^{ik\theta}| = \sum_k |c_k| = +\infty\,, \tag{54}$$

and therefore the point-wise convergence has to rely on suitable cancellations in the sum.

An important property of the Fourier series (52) is the identity

$$\int_{-\pi}^{\pi} |f(\theta)|^2 = 2\pi \sum |c_k|^2\,, \tag{55}$$

which follows from the othogonality of the functions $e^{ik\theta}$:

$$\int_{-\pi}^{\pi} e^{ik\theta}\overline{e^{il\theta}}\, d\theta = \int_{-\pi}^{\pi} e^{ik\theta} e^{-il\theta}\, d\theta = \begin{cases} 0 & \text{when } k \neq l\,, \\ 2\pi & \text{when } k = l\,. \end{cases} \tag{56}$$

Based on this one can prove (and may not be in the textbook) that for any function f for which the integral $\int_{-\pi}^{\pi} |f(\theta)|^2\, d\theta$ is well-defined and finite, the Fourier series satisfies the following:

If we denote

$$S_n f(\theta) = \sum_{k=-n}^{k=n} c_k e^{ik\theta} \tag{57}$$

the partial sum of the series, then

$$\int_{-\pi}^{\pi} |f(x) - S_n f(x)|^2\, d\theta \to 0 \quad \text{for } n \to \infty\,. \tag{58}$$

This is another type of convergence which is different from the point-wise convergence (although still related), and plays an important role in the theory of Fourier series, and PDEs in general.

We calculated the Fourier coefficients c_k of the 2π-periodic function defined by

$$f(x) = x\,, \quad x \in (-\pi, \pi)\,. \tag{59}$$

and the periodicity condition $f(x+2\pi) = f(x)$ for $x \in (-\infty, \infty)$. We did not specify $f(\pi)$, as this is not important for finding the Fourier series. Note that the function is not continuous as a periodic function (no matter how $f(\pi)$ would be defined), as it has a discontinuity at $x = \pm\pi$. We evaluated the integrals

$$c_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) e^{-ikx}\, dx \tag{60}$$

by integrating by parts and obtained

$$c_k = (-1)^k \, \frac{i}{k}\,, \quad k \neq 0\,, \qquad c_0 = 0. \tag{61}$$

Note that we have $\sum_k |c_k| = +\infty$, as expected (given that the functions is discontinuous as a periodic function). On the other hand, the sum $\sum_k |c_k|^2$ is finite, as expected due to the finiteness of $\int_{-\pi}^{\pi} |f(x)|^2\, dx$. In addition, from (55) we get the classical identity

$$\sum_{k=1}^{\infty} \frac{1}{k^2} \;=\; \frac{\pi^2}{6}\,. \tag{62}$$

Lecture 9, 10/3/2017
(More on Fourier series; the wave equation)

In the first part of the lecture, we continued to discuss some of the topics on Fourier series in Chapter 3.
In addition to the notion of the pointwise convergence, considered in the textbook, there are other important notions of convergence, which we will not discuss in detail, but it is still good to known about them. Although a Fourier series may not converge point-wise, usually it converges if we choose the notion of convergence appropriately.
Let us illustrate this by a few examples. These are optional.
(i) It was already known to Euler that for a finite sum

$$f(x) = \sum_{k=-n}^{n} c_k e^{ikx}\,, \tag{63}$$

we have

$$\int_{-\pi}^{\pi} |f(x)|^2\, dx \;= 2\pi \sum_{k=-n}^{n} |c_k|^2\,. \tag{64}$$

This raises the following question: if we have a an infinite sequence $\{c_k\}_{k=-\infty}^{\infty}$ with

$$\sum_{k=-\infty}^{\infty} |c_k|^2 < +\infty \tag{65}$$

does the infinite series

$$f(x) = \sum_{k=-\infty}^{\infty} c_k e^{ikx} \tag{66}$$

define a "genuine function", the Fourier series of which is given by c_k?
This question was only resolved in 1907 by what is now called the Riesz-Fischer Theorem . It states that the series converges if we use a different kind of convergence, which we mentioned in the last lecture, see (58). The proof was made possible by the then new tool of Lebesgue integration, introduced in 1904, which is at the basis in many advances of modern analysis.
Even after the Riesz-Fisher theorem, it was not clear what can be said about the point-wise convergence of the Fourier series satisfying (65). This remained open until 1966, when the question was settled by L. Carleson.

(ii) Let us consider the Fourier series (61) of the function (60) from the last lecture:

$$f(x) = \sum_{k\neq 0} (-1)^k \frac{i}{k}\, e^{ikx}\,. \tag{67}$$

Can this sum be differentiated term-by term? This is discussed in the textbook in section 3.4, and it is shown there that if we differentiate f "naively", only as a function on $(-\pi, \pi)$, a term-by-term differentiation would lead to an incorrect result.
However, there is a way to interpret the differentiation in which the term-by-term differentiation becomes correct. This is again a topic beyond the scope of this course, but it is good to mention it briefly. (This is of course optional). The first important point to realize for the correct interpretation is that although f is smooth in $(-\pi, \pi)$, where we have $f(x) = x$, the function is not smooth (or even continuous) as a periodic function in $(-\infty, \infty)$. There is a discontinuity at the points $x = (2l + 1)\pi$, where l is an integer.
How do we define $f'(x)$ at those points? This is a non-trivial question, which was successfully resolved only the 1950 by the theory of distributions.[10] In that theory, the correct formula for the differentiation of f is

$$f'(x) = 1 - 2\pi\delta(x - \pi)\,, \tag{68}$$

where $\delta(x)$ is the so-called Dirac delta function. With these definitions the series (67) can be differentiated term-by-term, although one still faces another difficulty: the series which we obtain by differentiation

$$\sum_{k\neq 0} (-1)^{k+1} e^{ikx} \tag{69}$$

is obviously not convergent in a classical sense, so one must re-interpret the notion of convergence, which is again done by using the theory of distributions mentioned above.

[10] The theory formalized in a elegant way many observations which were known for some time, starting with the work of Riemann, and later Heaviside, Dirac and others.

So, overall, you see that it is a lot of work to come up with good notions of differentiation and convergence in which the Fourier series could be differentiated term-by-term, but it is possible, and in many respects the new notions of convergence and differentiation are the right ones for the theory of PDEs.

In the second part of the lecture we starting discussing the wave equation, along the lines of the material in Section 4. We emphasized that the behavior of the solutions of the wave equation is quite different from the behavior of the heat equation and the Laplace equation. While the solutions of the Laplace equation and the heat equation in some sense "try to become constant", the solutions of the wave equation have a tendency to oscillate, especially in bounded domains. The solutions do not satisfy the maximum principle, and we do not see the "smoothing effects" which we observed for the heat and Laplace equations.
Still, at the level of calculation, one can again solve many interesting problems by separation of variables, and the calculations are in fact quite similar to the other equations, in some sense we just replace the functions $e^{-\lambda t}$ by $e^{i\lambda t}$.

Lecture 10, 10/5/2017
(More on the wave equation)

We continued to discuss the wave equation

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}\,, \tag{70}$$

which we derived as an equation for oscillations of a string. Recall that $c = \sqrt{\frac{T}{\rho}}$, where T is the tension of the string (the force which pulls on it), and ρ is its density (mass per unit length).
Let us first assume that the string is parameterized by $x \in (0, L)$ and that $u(0,t) = u(L,0) = 0$. The general solution of the equation in that case can be written for example as

$$u(x,t) = \sum_{k=1}^{\infty} A_k \sin\left(\frac{\pi k x}{L}\right) \sin\left(\frac{\pi k c (t - t_k)}{L}\right)\,, \tag{71}$$

where $t_1, t_2, \ldots$ are any numbers[11] and $A_1, A_2, A_3, \ldots$ are any numbers with for which the series is convergent in a suitable sense. This condition is of course somewhat vague, but at this point we do not worry about it.
We note that the time frequencies contained in the solution are

$$\omega_k = \frac{\pi c}{L} k\,, \qquad k = 1, 2, 3, \ldots \tag{72}$$

These are the frequencies which we will hear, when listening to the vibrations of the string. On a guitar or violin the most audible frequency typically is

[11] Note that due to periodicity of sin the numbers t_k can be taken from a bounded interval, without loss of generality.

ω_1, which is usually described in term of the specific note we perceive when listening to the vibrations. Note that when we know L and T, we can derive the density of the string ρ from ω_1. In a similar way, we can measure the tension T by measuring ω_1, L and ρ. The "overtones", given by $\omega_2, \omega_3, \dots$ and their amplitudes $A_2, A_3, \dots$ determine the "color" of the tone (distinguishing the guitar from the piano, say). Our ear is not sensitive to the parameters t_k, which determine the phase shifts of the various modes.
We note that such a decompositions of small vibrations near an equilibrium is quite general, and applies to systems described by

$$\ddot{x} = Ax\,, \tag{73}$$

where, in case of finitely many degrees of freedom, n, say, A is an $n\times n$ symmetric matrix. If we diagonalize this matrix, the system will simplify to a set n non-interacting oscillators, with the $n-$th oscillator being governed by

$$\ddot{y}_k = -\omega_k^2\, y_k\,, \tag{74}$$

for suitable $\omega_1, \dots, \omega_n$, with $y_1, \dots, y_n$ denoting the variables in which A is diagonal. The general solution of (74) can be written for example as $y_k(t) = A_k \sin(\omega_k(t - t_k))$.

Periodic string (in x)

The wave equation can also be considered with the periodic boundary conditions. In that case it is sometimes useful to use the complex form of the Fourier representation. Let us consider the case of $2\pi-$periodic functions (in x), for example. In that case the general solution of (70) can be written as for example as

$$u(x,t) = \sum_{k=-\infty}^{\infty} e^{ikx}\left(c_k^- e^{-ikct} + c_k^+ e^{ikct}\right)\,. \tag{75}$$

where c_k^-, c_k^+ are (complex) numbers for which the series converges in a suitable sense. (We are again a little vague on the type of convergence.) The fact that the solution is possibly complex valued is not a problem - when we are dealing with the classical periodic string, the physical meaning can be attached to the real part of the solution.
Note that (75) can be written as

$$u(x,t) = \sum_k c_k^- e^{ik(x-ct)} + \sum_k c_k^+ e^{ik(x+ct)} = f(x-ct) + g(x+ct)\,. \tag{76}$$

It is easy to verify, independently of the above considerations, that for any (sufficiently regular) functions f, g defined on the real line, the functions $f(x-ct)$ and $g(x+ct)$ satisfy the wave equation. Therefore, in the periodic case, the general solution of (70) can be also written as

$$u(x,t) = f(x-ct) + g(x+ct)\,, \tag{77}$$

where f, g are any "reasonable" functions. This is an important fact about the solutions of the one-dimensional wave equation (70) in the periodic case, and it also extends to the case of the general solution on the real line $(-\infty, \infty)$.
Solutions of the form $f(x - ct)$ can be thought of as "travelling waves", and represent a disturbance which travels from the left to the right at speed c, without changing its shape.
The existence of such solution has to do with the fact that the coefficient c is constant. When $c = c(x)$ (as is the case when the density of the string changes with x), such "pure" traveling waves no longer exist, as the excitations can be reflected back at the inhomogeneities. This phenomenon can be used for measuring many things, and also for imaging methods in medicine (X-rays, ultrasound, ...), geology (inverse imaging via seismic waves), and other areas.

Lecture 11, 10/12/2017
(Equations with variable coefficients, Sturm-Liouville problems, finite-dimensional analogues)

We started discussing the material in Section 5. The equation for the vibrations of a string of variable density $\rho(x)$ is

$$\rho(x)\frac{\partial^2 u}{\partial t^2} = T\frac{\partial^2 u}{\partial x^2}\,. \tag{78}$$

The heat equation in a rod with variable heat conduction coefficient $K_0 = K_0(x)$, variable density $\rho = \rho(x)$, and variable heat capacity $c = c(x)$ is

$$c(x)\rho(x)\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(K_0(x)\frac{\partial u}{\partial x}\right)\,. \tag{79}$$

The consideration concerning the boundary conditions are the same as in the case with constant coefficients. Assuming the equations are considered in $(0, L)$, we will consider the boundary conditions $u(0, t) = 0$ and $u(L, t) = 0$ as an example.
We can still apply the method of separation of variables - the reason why it still works in the two above examples (and many other examples) is that the coefficients of the equations do not depend on time.
Let us look at the wave equation (78). Setting

$$u(x, t) = \phi(x)h(t) \tag{80}$$

we obtain

$$\rho(x)\phi(x)h''(t) = T\phi''(x)h(t)\,, \tag{81}$$

and "separating the variables", we arrive at

$$\frac{h''(t)}{h(t)} = \frac{T\phi''(x)}{\rho(x)\phi(x)} = -\lambda\,, \tag{82}$$

where we put the minus sign in front of the λ so that the values of λ for which we have solutions are positive. (Of course, if we did these problems for the first time, we would not know that and would write (82) with λ, and then probably change the sign later.)
This leads to an equation

$$h''(t) = -\lambda h(t) \tag{83}$$

for h which we already know how to solve. The equation for ϕ is more complicated:

$$\phi''(x) = -\lambda \frac{\rho(x)}{T} \phi(x)\,, \tag{84}$$

with the boundry conditions $\phi(0) = 0$ and $\phi(L) = 0$.
This is an example of a *Sturm-Liouville problem*, and these problems are discussed in some detail in the textbook, in section 5.3. The main theorem is on page 157, and we discussed some parts of it in the lecture.
The main takeaway for these problems is that the behavior of solutions is qualitatively similar to the case with constant coefficients, but in most cases it is not possible to express the solutions explicitly in terms of elementary functions. There are various special functions often introduced for the purpose of expressing solutions of such problems.
In the case of a homogeneous string we saw (see (72)) that the frequencies which we hear in the oscillations are given by the multiples of a certain "base frequency". This may no longer be the case in the case of a string with variable density, and the overtones which we hear in that case may be dissonant with the base frequency.
There is a finite-dimensional analogy of some of the mathematical issues which arise in the context of strings with variable density.[12] Let A be a $n \times n$ symmetric matrix, let M be a positive definite $n \times n$ symmetric matrix, and consider the equation

$$M\ddot{x} = Ax\,. \tag{85}$$

We know how to approach

$$\ddot{x} = Ax\,. \tag{86}$$

In that case there exists another orthogonal basis in which A is diagonal and the system splits into n independent simple equations of the form

$$\ddot{y}_j = \lambda_j y_j\,, \qquad j = 1, 2, \ldots, n\,. \tag{87}$$

Can we do the same for (85)? The answer is yes, and there are several ways of doing it. For example, let us start by writing

$$\ddot{x} = M^{-1}Ax = Bx\,. \tag{88}$$

The problem of comparing this equation with (86) is that the matrix B is no longer symmetric. One of the tricks to deal with this is to use a different scalar product. Instead of working with the usual scalar product

$$(x, y) = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n\,, \tag{89}$$

[12] This material is optional.

we work with

$$(x, y)_M = (Mx, y) = \sum_{ij} M_{ij} x_j y_i \,. \tag{90}$$

It is easy to check that

$$(Bx, y)_M = (x, By)_M \tag{91}$$

and hence we are back to the familiar situation of (85) (but with a new scalar product). The basis in which B is diagonal will in general not be orthogonal for the old scalar product, but it will be orthogonal for the new scalar product.
Another (mathematically equivalent) way to deal with (85) is the following: First, we can write M as a square of another positive definite symmetric matrix, which we will denote $M^{\frac{1}{2}}$,

$$M = M^{\frac{1}{2}} M^{\frac{1}{2}} \,. \tag{92}$$

This is easy to see when M is diagonal, and the general case can be reduced to the case of a diagonal matrix, if we work in the basis in which our matrix is diagonal.
Let us now introduce a new variable y in (85) by

$$M^{\frac{1}{2}} x = y \,, \tag{93}$$

which is the same as

$$x = M^{-\frac{1}{2}} y \,. \tag{94}$$

We can then write

$$M^{\frac{1}{2}} \ddot{y} = A M^{-\frac{1}{2}} y \,, \tag{95}$$

or

$$\ddot{y} = M^{-\frac{1}{2}} A M^{-\frac{1}{2}} y = \tilde{A} y \,, \tag{96}$$

where $\tilde{A} = M^{-\frac{1}{2}} A M^{-\frac{1}{2}}$. The main point now is that $\tilde{A}$ is a symmetric matrix (with respect to the canonical scalar product (89)).
In some sense, the change of coordinates (94) enabled us to replace the expression $M^{-1}A$ in (88) (which may not give a symmetric matrix even when A, M are symmetric) by the more symmetric expression $M^{-\frac{1}{2}} A M^{-\frac{1}{2}}$, which does give a symmetric matrix (under our assumptions).

Lecture 12, 10/17/2017
(Sturm-Liouville problems, finite-dimensional analogues, Rayleigh quotient)

We continued to discuss the Sturm-Liouville problems in Sections 5, and started discussing the Rayleigh quotient, see also Section 5.6 of the textbook. Our discussion so far mostly concerned the finite-dimensional situation.

Lecture 13, 10/19/2017
(Sturm-Liouville problems, finite-dimensional analogues, Rayleigh quotient, introduction to numerical methods)

In the first part of the lecture we continued to discuss the Rayleigh quotient and constrained minimization of quadratic forms, first in finite dimension. The basic points are as follows.

To each $n \times n$ symmetric matrix $A = \{a_{ij}\}$ we can associate a quadratic form $q_A(x) = \frac{1}{2}(Ax, x) = \frac{1}{2}\sum_{i,j} a_{ij}x_j x_i$. We note that $q'_A(x)y = \sum_{i,j} a_{ij}x_j y_i$ and hence the vector of the partial derivatives $\{\frac{\partial q_A}{\partial x_i}(x)\}_{i=1}^n$can be identified with the vector Ax.

Let M be a strictly positive definite symmetric matrix, with the corresponding quadratic form $q_M(x)$. The sets $\{x\,,\, q_M(x) = \text{const.}\}$ are ellipsoids (when the constant is strictly positive).

It is clear that the function q_A restricted to the ellipsoid $\{q_M(x) = \frac{1}{2}\}$ attains its minimum, assume that this at a point $\overline{x}$. At the point $\overline{x}$ we must have

$$q'_A(\overline{x}) - \lambda q'_M(\overline{x}) = 0 \tag{97}$$

for some $\lambda \in \mathbf{R}^n$, due to basic rules for constrained minimization using the Lagrange multipliers. This is the same as

$$A\overline{x} = \lambda M\overline{x}\,. \tag{98}$$

Minimizing q_A over ellipsoids as above is the same as minimizing the function

$$\frac{q_A(x)}{q_M(x)}\,, \tag{99}$$

over $\mathbf{R}^n \setminus \{0\}$. Expression (99) is sometimes called the Rayleigh quotient, and we see from (98) (by taking the scalar product with $\overline{x}$) that

$$\lambda = \frac{q_A(\overline{x})}{q_M(\overline{x})} \tag{100}$$

The minimal eigenvalue of A with respect to M is given by the minimal value of this quotient (and similarly for the maximal eigenvalue and the maximum of the quotient). This has a number of applications, some of which we discussed. There are also analogous considerations in the Sturm-Liouville problems, and these are discussed in the textbook.

We started discissing the numerical methods in Chapter 6 of the textbook.

Lecture 14, 10/24/2017
(Numerical methods - a first look at their accuracy and stability)

Consider the simple (ordinary) differential equation

$$\dot{y} = ay\,, \tag{101}$$

where $y = y(t)$ is a function of one variable and a is a parameter. Eventually we might want to allow complex a, but for now we can assume that a is real. (We use he usual notation $\dot{y}$ for the time derivative $\frac{dy}{dt}$.)

We can solve this equation explicitly, the solution is

$$y_{\text{exact}}(t) = e^{at} y(0)\,, \tag{102}$$

so there is no need to use numerical methods in this case. However, we can use the explicit solution for analyzing numerical methods and checking how they perform in this simple case. The information which we get from from such an analysis turns out to be very useful. Some of the main points one has to keep in mind in connection with numerical calculations already transpire in this elementary example.

Let us choose a small $\tau > 0$ and approximate (101) by

$$\frac{y(t+\tau) - y(t)}{\tau} = a y(t)\,. \tag{103}$$

This is the same as

$$y(t+\tau) = (1 + a\tau) y(t)\,, \tag{104}$$

and by iterating this formula we have

$$y(n\tau) = (1 + a\tau)^n y(0)\,. \tag{105}$$

We can also write it in a different way: for $t > 0$ we choose τ so that $n\tau = t$ and write

$$y_{\text{approx},n}(t) = \left(1 + \frac{at}{n}\right)^n y(0)\,. \tag{106}$$

We now wish to compare a precise solution (102) with the approximate solution (106). Let us calculate

$$\varepsilon(t,n) = \log y_{\text{exact}}(t) - \log y_{\text{approx},n}(t) \tag{107}$$

for large n. Note that

$$y_{\text{exact}}(t) = e^{\varepsilon(t,n)} y_{\text{approx},n}(t)\,, \tag{108}$$

so knowing $\varepsilon(t,n)$ is enough to see what the error is. We recall that for $|\xi| < 1$ we have

$$\log(1+\xi) = 1 + \xi - \frac{1}{2}\xi^2 + O(|\xi^3|)\,, \tag{109}$$

where we use the "O-notation": $O(\xi^3)$ means that the error is below $C|\xi^3|$ when ξ is small, where the exact value of C is not important for our argument. Using (102)together with (106) and (109), we obtain

$$\varepsilon(t,n) = \frac{1}{2}\frac{(at)^2}{n} + O\left(\frac{|at|^3}{n^2}\right)\,. \tag{110}$$

We see that the error approaches zero linearly in $\frac{1}{n}$. We say that the method is of the first order. This is only good enough when we do not need a lot of precision.

Assume now that we wish to apply the same method to a system

$$\dot{u} = Au\,, \tag{111}$$

where u is a vector with r components and A is an $r \times r$ matrix. We can do exactly the same calculation and arrive at

$$u(n\tau) = (I + n\tau A)^n u(0)\,, \tag{112}$$

where I is the $r \times r$ identity matrix.

There is a hidden danger in this formula. Assume for example that A is obtained from discretising the Laplace operator. Then it is symmetric, and by a suitable change of coordinates it can be diagonalized, so that (111) is equivalent to

$$\dot{y}_k = a_k y_k\,, \quad k = 1, 2, \ldots, r\,. \tag{113}$$

Now some of the eigenvalues a_k may be large negative (as is the case for the discrete Laplacian when r is large). We have

$$y_k(n\tau) = (1 + \tau a_k)^n y_k(0)\,, \tag{114}$$

and if $(1 + \tau a_k) < -1$, the iteration (114) will catastrophically diverge, whereas the real solution $y_k(t)$ very quickly decays to zero. We have encountered a numerical instability (in a severe form). In the scalar equation (101) with a negative a one would probably never even think about choosing τ with $1 + a\tau < -1$, but for the system (111) something similar can happen "by accident", as things may no longer be as explicit as in the obvious scalar case. Clearly, for the calculation to be reasonable, the step τ has to be chosen so that $\tau a_k > -1$ for all k. This is a typical *stability condition*. Stability conditions of one form or another are necessary for many numerical schemes.

One can try to improve the simple scheme we discussed for example as follows. The difference quotient in (103) is a more precise expression for the derivative $\dot{y}$ at the point $t + \tau/2$, rather then t. So one could try to consider

$$\frac{y(t+\tau) - y(t)}{\tau} = a y(t + \frac{1}{2}\tau)\,. \tag{115}$$

However, of our we evalue the solution on at times $k\tau$ with $k = 0, 1, 2, 3, \ldots$, we cannot use $\tau/2$. We can try instead

$$\frac{y(t+\tau) - y(t)}{\tau} = a\frac{1}{2}(y(t) + y(t+\tau))\,, \tag{116}$$

which is the same as

$$y(t+\tau) = \frac{1 + \frac{1}{2}\tau a}{1 - \frac{1}{2}\tau a}\, y(t)\,. \tag{117}$$

This is the (special case of the) Crank-Nicolson scheme. In case of matrices it amounts to

$$u(t+\tau) = (I - \frac{1}{2}\tau A)^{-1}(I + \frac{1}{2}\tau A)u(t)\,. \tag{118}$$

As an optional exercise, you can check by a similar calculation as above that the scheme has a higher precision than (104), the error will be of order $O(n^{-2})$ for large n. Also, for symmetric matrices with negative eigenvalues it never becomes unstable, because

$$\left|\frac{1-\xi}{1+\xi}\right| < 1 \tag{119}$$

when $\xi \geq 0$.

Lecture 15, 10/26/2017

We continued to discuss stability issues for numerical schemes, and also the probabilistic interpretation of a simple finite difference scheme for the heat equation in terms of random walk, see Section 6.3.4 of the textbook.

Lecture 16, 10/31/2017
(stability of numerical methods - continuation)

Let us consider a simple transport equation

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0\,, \tag{120}$$

where $a \in \mathbf{R}$ is a parameter. The solutions of this equation can be easily characterized: they are functions of the form

$$u(x,t) = f(x-at)\,, \tag{121}$$

where f is any continuously differentiable function. (Larger classes of solutions may be considered, but this is not our focus here.) The equation can be considered on $(-\infty,\infty)$, or on a circle (which is the same as on $(-\infty,\infty)$ with an extra condition $u(x+2L,t) = u(x,t)$, where $2L$ is the length of the circle).
Equation (120) is closely related to the wave equation, due to the identity

$$\left(\frac{\partial}{\partial t} + a\frac{\partial}{\partial x}\right)\left(\frac{\partial}{\partial t} - a\frac{\partial}{\partial x}\right) = \frac{\partial^2}{\partial t^2} - a^2\frac{\partial^2}{\partial x^2}\,. \tag{122}$$

Why would we do numerical analysis for a simple equation which we can solve explicitly? The reason is that the equation provides a good test for numerical methods - we can compare the exact solution (which in this case is available) and the solution we get from computer. Lessons we learn from this comparison can be then used for more complicated equations, such as the Navier-Stokes equation, where more complicated versions of the transport term $a\frac{\partial u}{\partial x}$ come up. Some of the issues which need to be addressed in the more complicated case are already present in the simple model (120).
Assume that in the computer the function $u(x,t)$ is represented by its values as $x_0 = 0, x_1 = h, x_2 = 2h, \ldots, x_{n-1} = (n-1)h$, with the understanding that $u(x_n,t) = u(x_0,t)$. This represents the situation when the point

$x_0, x_1, x_3, \ldots x_{n-1}$ are uniformly distributed an a circle of length $2L = nh$. In this situation, our information at each time consists of the values of the function u at the point $x_0, x_1, \ldots x_{n-1}$. If we are given $u(x,0)$ only as the vector $u(x_0,0) = f(x_0), u(x_1,0) = f(x_1), \ldots, u(x_{n-1},0) = f(x_{n-1})$, it is not completely clear what value for $u(x_1,\tau)$ we should take for $\tau = \frac{h}{2a}$. The precise solution is $u(x_1,\tau) = f(x_1 - a\tau) = f(x_1 - \frac{h}{2})$, but we do not have access to the value of f at $f(x_1 - \frac{h}{2})$ as we know f only at our "grid points" x_k. We see that evaluating the solution at time τ at the point x_k necessarily involves some kind of guessing what the value of f should be outside of the grid points.
As a side remark, it is worth noting that when $a > 0$ and $\tau = \frac{h}{a}$, the exact solution (121) satisfies

$$u(x, t+\tau) = u(x-h, t)\,, \tag{123}$$

which can also be written as

$$\frac{u(x,t+\tau) - u(x,t)}{\tau} + a\frac{u(x-h,t) - u(x,t)}{-h} = 0\,, \quad \tau = \frac{h}{a}\,. \tag{124}$$

We note that this is slightly different than the "naive" approximation

$$\frac{u(x,t+\tau) - u(x,t)}{\tau} + a\frac{u(x+h,t) - u(x,t)}{h} = 0\,. \tag{125}$$

In fact, this last approximation is dangerous, as can be seen from the following. Evaluating $u(x,t+\tau)$ from (125), we obtain

$$u(x,t+\tau) = (1 + a\frac{\tau}{h})u(x,t) - a\frac{\tau}{h}u(x+h,t)\,. \tag{126}$$

Let us see what this formula gives when $u(x,0) = e^{ikx}$. Then

$$u(x,\tau) = \lambda_k u(x,0)\,, \qquad \lambda_k = 1 + a\frac{\tau}{h}\left(1 - e^{ikh}\right)\,. \tag{127}$$

We note that the real part of λ_k is greater than 1, and when $kh \sim \pi$, then in fact $\lambda_k \sim 1 + 2a\frac{\tau}{h}$. This means that when we iterate the formula (126), obtaining

$$u(x, m\tau) = \lambda_k^m u(x,0)\,, \tag{128}$$

the solution of (126) will quickly grow. We see that the approximation (125) leads to serious numerical instability, and this formula cannot be used in a real computation. On the other hand, formula (124) would give reasonable results. Note, however, that that formula has its own problems, e. g. when a is negative. We see that the situation is quite subtle, and designing a good numerical scheme for this problem is non-trivial. A simple stable method which works is for example a version of the Crank-Nicolson scheme:

$$\frac{u(x,t+\tau) - u(x,t)}{\tau} + a\frac{u(x+h,t+\tau) + u(x+h,t) - u(x-h,t+\tau) - u(x-h,t)}{4h} = 0\,. \tag{129}$$

A slight disadvantage of this method from the computational point of view is that in it "implicit", i. e. computing the vector $u(x,t+\tau)$ involves a solution of a system of equations. Also, the scheme has some other disadvantages, but they are not catastrophic. In general, when dealing with the simple equation (120) in a numerical calculation, one should be careful. Often a reasonable method to use, at least when we are dealing with periodic boundary conditions and smooth functions, is related to discrete Fourier transformation (the so called pseudospectral method).
Taking $\tau \to 0_+$ in (129) gives

$$\frac{\partial u(x,t)}{\partial t} + a\frac{u(x+h,t)-u(x-h,t)}{2h} = 0\,. \tag{130}$$

This is an example of a *semi-distrete scheme*, where we discretize only the spatial variable: x is considered to be in our discrete set $0, h, 2h, \ldots, (n-1)h$. The last equation can be written as

$$\dot{u} = Au\,, \tag{131}$$

where u is an time-dependent $n-$vector $u_0, u_1, \ldots, u_{n-1}$ (with $u_k(t) = u(x_k,t)$ and A is the following $n \times n$ matrix

$$A = -\frac{a}{h}\begin{pmatrix} 0 & 1 & 0 & \ldots & 0 & 0 & -1 \\ -1 & 0 & 1 & \ldots & 0 & 0 & 0 \\ 0 & -1 & 0 & \ldots & 0 & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & -1 & 0 & 1 \\ 1 & 0 & 0 & \ldots & 0 & -1 & 0 \end{pmatrix} \tag{132}$$

One can check that the eigenvalue of this matrix are purely imaginary, so our semi-discrete scheme is not unstable. It introduces an interesting error related to dispersion. One way to get an idea of what behavior the semi-discrete scheme (130) will lead to is to use the following approximation: write

$$u(x \pm h, t) = u(x) \pm h\frac{\partial u(x,t)}{\partial x} + \frac{h^2}{2}\frac{\partial^2 u(x,t)}{\partial x^2} \pm \frac{h^3}{6}\frac{\partial^3 u(x,t)}{\partial x^3} + O(h^4)\,, \tag{133}$$

and substitute this into (130), neglecting the terms coming from $O(h^4)$. This gives

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} + \frac{ah^2}{6}\frac{\partial^3 u}{\partial x^3} = 0\,. \tag{134}$$

We see that the equation which we would like to model gains in this semi-discrete approximation an additional (unwanted) term with the third derivative. The term has h^2 in front of it, so this influence vanishes as $h \to 0$.
As an optional exercise, you can analyze stability of the semi-discrete scheme

$$\frac{\partial u(x,t)}{\partial t} + a\frac{u(x+h,t)-u(x,t)}{h} \tag{135}$$

by looking at solutions of the form $u(x,t) = c(t)e^{ikx}$. Equation (135) then gives

$$\dot{c} + a\lambda_k c = 0\,, \tag{136}$$

with

$$\lambda_k = \frac{e^{ikh} - 1}{h}\,, \tag{137}$$

and we see that for k with e^{ikh} some distance away from 1 and $a > 0$ we will have instability. This can be also seen by other ways. For example, a calculation similar to (123) shows that if we neglect the terms of order h^3, the scheme should be have approximately as

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} + \frac{ah^2}{2}\frac{\partial^2 u}{\partial x^2} = 0\,, \tag{138}$$

which is not a good equation to solve for in the direction of positive time when $a > 0$. Note that in that case the second derivative comes into the equation with a sign opposite to what it would be for the heat equation.

Lecture 17, 11/2/2017
(normal matrices and their diagonaization)

When we solve PDEs on a computer, one natural way to represent functions is as vectors: a functions $f(x)$ is represented by its values on a a set of grid points, such as $x_0 = 0$, $x_1 = h$, $x_2 = 2h$, $x_3 = 3h$, ... A natural vector representing f is has coordinates $f_0, f_1, f_2, \ldots, f_{n-1}$ when we think of a periodic function f and assume that $f(x + nh) = f(x)$. (For $2L-$periodic functions we would take $h = \frac{2L}{n}$ so that $nh = 2L$.) Linear operations on vectors are usually best described in terms of matrices.
We will consider $n \times n$ matrices, in general with entries which are complex numbers. Let $\mathbf{C}$ denote the set of complex numbers and $\mathbf{C}^n$ the set of complex $n-$vectors. By abuse of notation, such vectors will be written for example as as $z = (z_1, \ldots, z_n)$, even though they should be written, strictly speaking, as column vectors. Notice that we also changed our numbering, and our indices for the vectors now go from 1 to n.
We define the *Hermitian product* of two vectors z and w in $\mathbf{C}^n$ as

$$\langle z, w\rangle = z_1\overline{w}_1 + z_2\overline{w}_2 + \ldots z_n\overline{w}_n\,, \tag{139}$$

where $\overline{w}_j$ means, as usual, the complex conjugate of w_j. The real part of the Hermitian product can be thought of as the usual real scalar product in $\mathbf{R}^{2n}$, and has the usual meaning that $\mathrm{Re}\,\langle\, z\,,\, w\,\rangle = |z||w|\cos\theta$, where θ is the angle between the two vectors and $|z|, |w|$ are respectively their lengths. Given a $n \times n$ matrix $A = A_k l$ and a vector $z \in \mathbf{C}^n$, the vector Az is defined as usual by $(Az)_k = \sum_{l=1}^n A_{kl} z_l$. The adjoint matrix A^* is defined by

$$\langle\, Az\,,\, w\,\rangle = \langle\, z\,,\, A^* w\,\rangle\,. \tag{140}$$

This is the same as

$$(A^*)_{kl} = \overline{A}_{lk}\,. \tag{141}$$

Definition

A matrix A is called *normal* if $AA^* = A^*A$.

It is easy to see that any diagonal matrix is normal. The converse of the statement is of course not correct if we formulate it naively – not every normal matrix is diagonal. However, it become true if we include a change of coordinates:

Theorem

If a matrix A is normal, then there exists a basis of vectors mutually orthogonal with respect to the Hermitian product, such that A becomes diagonal when represented in this basis.

This can be thought of as an extension of the statement which we emphasized many times - namely that each real symmetric matrix can be diagonalized in a suitable orthogonal basis. (It is a good exercise to derive this last statement from the above theorem.)

The theorem is in fact not hard to prove, the interested reader can find proofs of the statement online (or in any number of textbooks).

A natural class of normal matrices is the set of *unitary matrices*, which are matrices U for which $U^* = U^{-1}$. An important example of such a matrix is the matrix of a "shift" $(z_1, z_2, z_3 \dots, z_n) \to (z_2, z_3, z_4, \dots z_n, z_1)$ which can be identified with the matrix

$$S = \begin{pmatrix} 0 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & 1 \\ 1 & 0 & 0 & 0 & \dots & 0 \end{pmatrix} . \tag{142}$$

This matrix is easily seen to be unitary, and its eigenvectors and eigenvalues can be computed explicitly: the eigenvectors are column vectors with coordinates $1, \zeta, \zeta^2, \dots, \zeta^{n-1}$, where ζ is a complex number with $\zeta^n = 1$. The matrix S and and its eigenvectors are important in the analysis of difference quotient approximations of PDEs with constant coefficients, as such approximation can typically be expresses in terms of S. For example, the matrix $S + S^* - 2I$ is the discrete version of the second derivative.
The transformation between the standard canonical basis of $\mathbf{C}^n$ and the basis consisting of eigenvectors of S is essentially the discrete Fourier transform.

Lecture 18, 11/7/2017
(finite elements - an elementary example)

Let us start with a simple observation concerning the elementary equation

$$ax = b\,. \tag{143}$$

If $a > 0$, then solving the equation is the same as finding the minimum of the function

$$q(x) = \frac{1}{2}ax^2 - bx\,. \tag{144}$$

Note that q is a convex function (recall that we assume $a > 0$), its graph is a parabola, $q(x) \to +\infty$ when $x \to \pm\infty$, and hence q attains its minimum at exactly one point. That point is given by the equation $q'(x) = 0$, which is the same as (143).

The same applies to a more complicated equation (or, more precisely, system of equations)

$$Ax = b\,, \tag{145}$$

where A is a $n \times n$ symmetric matrix satisfying $(Ax, x) > 0$ for each vector $x = (x_1, \ldots, x_n) \neq 0$, and $b = (b_1, \ldots, b_n)$ is a given vector. (We again slightly abuse notation by writing the vectors above as a row vectors, even though they should really be thought about as column vectors.)

In connection with (145) we define

$$Q(x) = Q(x_1, \ldots, x_n) = \frac{1}{2}(Ax, x) - (b, x)\,, \tag{146}$$

and note that under our assumptions $Q(x) \to \infty$ when $|x| \to \infty$ and Q is (strictly) convex. Hence it attains its minimum at exactly one point. At the minimum all the partial derivatives of Q have to vanish:

$$\frac{\partial Q}{\partial x_i}(x) = 0\,, \qquad i = 1, 2, \ldots, n\,, \tag{147}$$

and one can check that these n equations are exactly the same as the n equations symbolized by the compact notation (145).

The takeaway from the above is that sometimes solving a system of equations can be equivalent to finding a minimum of a function. The simplest examples of the method of finite elements are probably best understood in this context (when the space over which we minimize a function is itself a space of functions), although the applicability of the method goes beyond such situations.

We will ilustrate the idea of the method on the boundary value problem

$$-\frac{d^2}{dx^2}u(x) = f(x) \qquad x \in (a, b)\,, \qquad u(a) = 0\,, \quad u(b) = 0\,. \tag{148}$$

The analogy with the previous example is the following:

- The unknown $x = (x_1, \ldots, x_n)$ in (146) corresponds to the unknown function u in (148) .
- The right-hand side b in (146) corresponds to the function f in (148)
- The matrix A in (146) corresponds to the differential operator $\frac{d^2}{dx^2}$ in (148).

- The "space" $\mathbf{R}^n$ from which we take the vectors x in (146) corresponds to the "space of functions u on (a,b) vanishing at the end-points" in (148).

The class of functions described in the last point is not very precisely defined, we did not say how regular these functions should be, but at this point we can consider this as a technicality which we can neglect.

The key point now is that there is a function on the space of functions, which we will denote by J, which is analogous to Q. It can be written as

$$J(u) = \int_a^b \left(\frac{1}{2}(-u''(x))u(x) - f(x)u(x) \right) dx = \int_a^b \left(\frac{1}{2}(u'(x))^2 - f(x)u(x) \right) dx \,. \tag{149}$$

The equality between the two integrals in (149) is valid only for functions which are "sufficiently regular" and vanish at the endpoint of the interval, and for this class of function one can obtain the equality by integration by parts.

The problem of solving the equation (148) is now equivalent (modulo technicalities) to the problem of finding the function u at which J attains its minimum, among all sufficiently regular functions vanishing at the endpoints of the interval (a,b).

One can see this by taking a smooth function φ which vanishes at the endpoint of the interval, and calculating

$$\frac{d}{dt}|_{t=0} J(u+t\varphi) = \int_a^b \left(u'\varphi' - \varphi f \right) dx = \int_a^b \left(-u'' - f \right) \varphi \, dx \,, \tag{150}$$

where we have again used integration by parts. If we are at a "point" (a function) u where J attains its minimum, the derivative (150) has to vanish for each φ as above, and this means that $-u'' - f = 0$, which is the same as (148).

A "function on functions", such as our J, is usually called a *functional*. To solve the full problem (148), we need to minimize J over an infinite-dimensional space of functions. The idea of the finite element method is to approximate this problem by minimizing J over some finite-dimensional subspace of functions, such as functions which are continuous and "piece-wise affine" with respect to some particular partition of (a,b) into small intervals. For example, for each positive integer $n \geq 1$ we can consider the space of function X_n defined as follows. Take $n+1$ points $x_0 = a < x_1 < \cdots < x_{n-1} < x_n = b$ in (a,b). A function u is in X_n if it satisfies the following conditions

- u is continuous in $[a,b]$.
- $u(a) = u(b) = 0$.
- u is affine (i. e. of the form $u(x) = a_i x + b_i$) on each interval $[x_i, x_{i+1}]$, $i = 0, 1, \ldots n-1$

It is easy to see that a function u in X_n is uniquely described by its values at $x_1, x_2, \ldots, x_{n-1}$, and hence the space X_n has dimension $n-1$. The points x_i are often chosen so that $x_{i+1} - x_i = (b-a)/n$, but other choices are possible.

Note that J is well-defined on X_n and is in fact of the form (146) on that space, no matter how we choose the partition $x_0, x_1, \ldots, x_n$. (The exact form of the matrix A will of course depend on how exactly we choose these points, but it will always be a positive definite symmetric matrix.) Therefore the problem of minimizing J on X_n is equivalent to solving (145) for suitable A, b. The exact calculation of A and b may need some work, but an attractive feature of the method is that once we choose X_n, everything lese is uniquely determined, we do not have to make any guesses. (There are many other choices of the finite-dimensional space we can use.)

Lecture 19, 11/14/2017
(finite elements - continuation)

Let us return to the space X_n defined at the end of the last lecture. We note that a function $u \in X_n$ is uniquely determined by its values at the points $x_1, \ldots, x_{n-1}$, let us denote them $u_1, \ldots, u_{n-1}$. The functional J can therefore be considered as a function on the space of vectors in $\mathbf{R}^{n-1}$ with coordinates $(u_1, \ldots, u_{n-1})$. Let us calculate the equation we get from the condition that the differential of the functional J attains its minimum on X_n at a vector $u \in X_n$ with coordinates $(u_1, \ldots, u_{n-1})$.
Note that our notation is somewhat loose: we use u for both the vector $(u_1, \ldots, u_{n-1})$ and the function in X_n associated to it. Also, we use J for the original functional, as well we for the function on $\mathbf{R}^{n-1}$ defined by restricting J to X_n and expressing in in the coordinates $(u_1, \ldots, u_{n-1})$ on X_n, so that for $u \in X_n$ we may write $J(u) = J(u_1, \ldots, u_{n-1})$. Such a notation has an advantage in its flexibility, but in some situations it may have a disadvantage of being a little ambiguous, and it works only if the reader has the same objects in mind as the writer. Hopefully this will be the case in our situation here.
Let us fix an integer $k \in \{1, 2, \ldots, n-1\}$. We will calculate the equation one gets from

$$\frac{d}{dt}|_{t=0} J(u + t\varphi) = 0 \tag{151}$$

when we choose $\varphi = \varphi^{(k)} \in X_n$ which has values 0 at all points x_i with the exception of x_k, where we will assume $\varphi^{(k)}(x_k) = 1$. Then (151) is the same as

$$\frac{\partial J(u_1, \ldots, u_{n-1})}{\partial u_k} = 0\,. \tag{152}$$

We have

$$u'(x) = \frac{u_{i+1} - u_i}{h}\,, \qquad x \in [x_i, x_{i+1}]\,, \tag{153}$$

and

$$\frac{d}{dx}\varphi^{(k)}(x) = \begin{cases} 0 & x \neq (x_{k-1}, x_{k+1})\,, \\ \frac{1}{h} & x \in (x_{k-1}, x_k)\,, \\ -\frac{1}{h} & x \in (x_k, x_{k+1})\,. \end{cases} \tag{154}$$

We note that $\int_a^b \varphi^{(k)}(x)\,dx = h$. Let us define

$$\tilde{f}_k = \frac{1}{h}\int_a^b f(x)\varphi^{(k)}(x)\,dx\,. \tag{155}$$

Note that $\tilde{f}_k$ gives a certain average of f at the point x_k which approaches $f(x_k)$ when f is continuous and $h \to 0$.
Using this notation, together with (153) and (154), the equation (151) (which is the same as (152)) becomes

$$-\frac{u_{k-1} - 2u_k + u_{k+1}}{h^2} = \tilde{f}_k\,. \tag{156}$$

We recognize the expression on the left as the approximation of the operator $-\frac{\partial^2}{\partial x^2}$ which we have seen before.
Other choices of the finite-dimensional space which we use to approximate the function space lead to different approximations of the operator. Note again that once we have the functional J, our approximation of the problem is uniquely determined by the choice of X, all the other details are determined by the method.

Other boundary conditions

The method is also quite flexible as far as the boundary conditions are concerned. Let us for example consider the problem of minimizing the function J above over the space $X^{\mathrm{reg}}_{a,0}$ of sufficiently regular functions which vanish only at the point a, and can attain non-zero values at the point b.
We again use the condition (151), this time for all functions $\varphi \in X^{\mathrm{reg}}_{a,0}$. First we use functions φ which vanish at both endpoints to get the equation

$$-u''(x) = f(x) \quad x \in (a,b)\,, \tag{157}$$

exactly in the same way as in the last lecture. Next, we take φ which vanishes at a but not necessarily at b, and integrate by parts:

$$0 = \int_a^b (u'\varphi' - \varphi f)\,dx = u'(x)\varphi(x)|_{x=a}^{x=b} - \int_a^b (-u''\varphi - f\varphi)\,dx\,. \tag{158}$$

As we already know that $-u'' = f$ in (a,b) and we also assume $\varphi(a) = 0$, the last identity amounts to

$$0 = u'(b)\varphi(b) \qquad \varphi \in X^{\mathrm{reg}}_{a,0}\,, \tag{159}$$

which is the same as $u'(b) = 0$. We see that the choice of $X^{\mathrm{reg}}_{a,0}$ as the space over which we minimize J leads to the boundary condition $u'(b) = 0$.

The heat equation and finite elements

The finite element method can also be applied to the heat equations. For that purpose it is instructive to interpret the finite dimensional equation

$$\dot{x} = -Ax\,, \tag{160}$$

for $n-$dimensional vector $(x_1, x_2, \ldots, x_n)$ with a positive-definite $n \times n$ matrix A as a "gradient flow". We let

$$f(x) = \frac{1}{2}(Ax, x) \tag{161}$$

and note that the vector ∇f consiting of the partial derivatives $\frac{\partial f}{\partial x_i}$ of f coincides with the vector Ax. The equation (160) can then be written as

$$\dot{x} = -\nabla f(x)\,. \tag{162}$$

One can think of it as the "steepest descent" in the "landscape" defined by the function f. The vector $\nabla f(x)$ is perpendicular to the surfaces defined by $\{f = \text{const.}\}$. Note that to draw the picture of ∇f as a vector in the same space where x "lives", we need to know what "perpendicular" means, or, in other words, we are using the scalar product in $\mathbf{R}^n$. (The partial derivatives $\frac{\partial f}{\partial x_i}$ are, of course, defined regardless of the scalar product, but without the scalar product structure they are coordinates of a linear functional on $\mathbf{R}^n$ (which is often denoted $f'(x)$), rather then a vector in $\mathbf{R}^n$).

The heat equation can be viewed in the same way, except we work with a suitable space of functions X, rather than a finite-dimensional space $\mathbf{R}^n$. For example, when X is the space of all sufficiently regular functions on the interval (a, b) which vanish at the endpoints, then the heat equation $\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + f(x)$ for $x \in (a, b)$ and $t \in (0, \infty)$ can be interpreted as the steepest descent defined by the functional J above (on the space X), if we take the scalar product on X to be defined as $(u, v) = \int_a^b u(x)v(x)\,dx$.

Here we do not wish to go into the subtleties of the exact interpretation of this heuristics. For practical purposes it is enough to know that we can consider the time derivative as a right-hand side of the corresponding steady-state problem, and get the right equation from the minimization of $\int_a^b \frac{1}{2}(u'(x))^2 - F(x)u(x))\,dx$ where we set $F(x) = f(x) - \frac{\partial u}{\partial t}$ after the minimization.

A finite dimensional approximation of the heat equation can now be obtained by choosing a suitable finite-dimensional subspace X_n, and apply the same consideration with X replaced by X_n.

Lecture 20, 11/16/2017
(equations with a right-hand side, inhomogeneos equations, Green's functions - introduction)

Duhamel's principle for simple ODEs

We start discussing the material in Chapter 8 in the textbook, although from a slightly different angle. We start with a simple ODE problem. Assume $a > 0$ and consider the a simple ODE for functions $x = x(t)$ on time interval $(0, \infty)$.

$$\dot{x} = -ax\,, \qquad x(0) = x_0\,. \tag{163}$$

We know how to to solve this problem: the solution $x(t)$ is defined in $(0, \infty)$ as

$$x(t) = x_0 e^{-at}\,. \tag{164}$$

From a certain point of view there is not much we can add to this, but one can also interpret this solution differently. Let us define a function $X(t)$ for $t \in (-\infty, \infty)$ as follows:

$$X(t) = \begin{cases} 0 & t < 0\,, \\ e^{-at} & t \geq 0\,. \end{cases} \tag{165}$$

The function has a discontinuity at $t = 0$, its value jumps by 1 as we cross $t = 0$ when moving on the $t-$axis from the left to the right.
Let us approximate $X(t)$ by a continuous function X_ε defined as

$$X_\varepsilon(t) = \begin{cases} 0 & t < -\varepsilon\,, \\ \frac{1}{\varepsilon}(t+\varepsilon) & t \in (-\varepsilon, 0)\,, \\ e^{-at} & t \geq 0\,. \end{cases} \tag{166}$$

The function X_ε is quite similar to X, except the "jump" at $t = 0$ does not happen at once, but it happens gradually (although still quickly, when ε is small) over a short interval $(-\varepsilon, 0)$. Note that X_ε increases linearly from 0 to 1 as t moves through the small time interval $(-\varepsilon, 0)$. Let us set

$$f_\varepsilon(t) = X_\varepsilon'(t) + aX_\varepsilon(t)\,, \qquad t \in (-\infty, \infty)\,. \tag{167}$$

Note that f_ε vanishes outside the interval $(-\varepsilon, 0)$, and for $t \in (-\varepsilon, 0)$ we have

$$f_\varepsilon(t) = \frac{1}{\varepsilon} + g_\varepsilon(t)\,, \qquad g_\varepsilon(t) = \frac{a}{\varepsilon}(t+\varepsilon)\,, \quad t \in (-\varepsilon, 0)\,. \tag{168}$$

The dominant part of f_ε in (168) is $\frac{1}{\varepsilon}$, the function $g_\varepsilon(t)$ is of order a, and its integral over $(-\varepsilon, 1)$ is of order εa, becoming negligible in the limit $\varepsilon \to 0_+$. On the other hand, the integral of $\frac{1}{\varepsilon}$ over $(-\varepsilon, 0)$ is 1. Hence

$$\lim_{\varepsilon \to 0_+} \int_{-\infty}^{\infty} f_\varepsilon(t)\, dt = 1 \tag{169}$$

and $f_\varepsilon(t)$ vanishes for t outside of the interval $(-\varepsilon, 0)$. The limit of functions f_ε as $\varepsilon \to 0_+$ is an object which is not really a function in the traditional sense. It is called the *Dirac function*, and usually denoted by $\delta(x)$. Formally, $\delta(x)$ vanishes everywhere except at $x = 0$ and $\int_{-\infty}^{\infty} \delta(x)\, dx = 1$. An important

property of the Dirac function is that for any continuous function $\phi(t)$ one has $\int_{-\infty}^{\infty} \varphi(s)\delta(s)\,ds = \varphi(0)$ and, more generally,

$$\int_{-\infty}^{\infty} \varphi(s)\delta(t-s),ds = \varphi(t)\,. \tag{170}$$

In the textbook the Dirac function is discussed in Section 9.3.4 starting at page 384. Additional useful material can be found on the Wikipedia page linked above.

As the function X given by (166) is the limit of the functions X_ε as $\varepsilon \to 0_+$, it is natural to expect that

$$\dot{X} + aX = \delta \qquad \text{in } (-\infty, \infty)\,. \tag{171}$$

One can think about the situation as follows: We consider a system which, in the absence of any outside disturbances, is described by the equation $\dot{x} + ax = 0$. (For example, we can think of $x(t)$ as a mass of some radioactive substance in a sample, with the parameter a being related to the half-life time of the substance.) The function X describes the situation when up to time $t = 0$ "nothing is going on", at the time $t = 0$ the system gets a "kick" (or "impulse") normalized to a unit strength, and then is left alone for the rest of the time. The Dirac function δ symbolizes the impulse. (In the example with the radioactive material the impulse could represent an injection of a unit amount of a fresh radioactive material.)

Mathematically there is nothing new in his picture in comparison with our original viewpoint (163). However, the new interpretation has some advantages, especially when considering an inhomogeneous equation

$$\dot{x} + ax = f(t)\,. \tag{172}$$

Let us first assume

$$f(t) = f_1\delta(t-t_1) + f_2\delta(t-t_2) + \cdots + f_n\delta(t-t_n)\,. \tag{173}$$

We can think of this f as giving the system a series of impulses at times $t_1 < t_2 < \cdots < t_n$, with the strength of the impulse of time t_k being f_k. (In the context of the model with radioactive decay, we can think of injecting f_j units of the radioactive material at time t_j.) The solution of (172) with f given by (173) and x vanishing for $t < t_1$ is given by

$$x(t) = f_1X(t-t_1) + f_2X(t-t_2) + \cdots + f_nX(t-t_n)\,. \tag{174}$$

This formula is is a consequence of the linearity of the equation. For example, the term $f_1X(t-t_1)$ represents the contribution to $x(t)$ from the "kick" at time $t = t_1$. The influence of this kick is proportional to its strength f_1, and is not influenced by the other kicks. The solution is then just a sum of the contributions from the individual kicks. Such behavior of the solutions is a consequence of the linearity of the equation $\dot{x} + ax = 0$. If x_1, x_2 are two

solutions and α_1, α_2 are two real numbers, then $x = \alpha_1 x_2 + \alpha_2 x_2$ is again a solution.
Now a general function $f(t)$ can be thought of as composed of a continuous family of kicks, in the sense

$$f(t) = \int_{-\infty}^{\infty} f(s)\delta(t-s)\, ds\,. \tag{175}$$

This is the same as (170), but the iterpretation is a bit different. We think of the numbers $f(s)$ as coefficients in he decomposition of the function f into "elementary impulses" $t \to \delta(t-s)$.
The analogy of (174) now is

$$X(t) = \int_{-\infty}^{\infty} f(s)X(t-s)\, ds\,. \tag{176}$$

This describes the solution of (172) which vanishes in the limit $t \to -\infty$. (Some assumptions about f are needed in order for the integral be convergent, but at this point we are neglecting such technicalities.) The same formula can be obtain in various other ways. In the context of this formula, the function $X(t)$ might be called the Green's function of the differential operator $x \to \dot{x} + ax$.
Using the above considerations, we can write the solution of the problem

$$\dot{x} + ax = f(t)\,, \quad t > 0\,, \qquad x(0) = x_0 \tag{177}$$

as

$$x(t) = x_0 e^{-at} + \int_0^t f(s)e^{-a(t-s)}\, ds\,. \tag{178}$$

The same formula can again be arrived in many other ways, including the standard "variation of parameters" discussed in the textbook for second order equations on page 361. The interpretation above is closely related to the Duhamel's integral, see also the Duhamel's formula.

Green's function for $-u''(x) = f(x)$ in (a,b) , $u(a) = u(b) = 0$.

The idea of decomposing a function $f(x)$ into the Dirac functions, calculating the solution for a single Dirac functions, and then using linearity and superposition work also for our next example:

$$-u''(x) = f(x)\,,\ x \in (a,b)\ , \quad u(a) = u(b) = 0\,. \tag{179}$$

We first calculate u when $f(x) = \delta_y(x) = \delta(x-y)$, the Dirac function located at the point $y \in (a,b)$. Note that the solutions of $u''(x) = 0$ a given interval are only linear functions (of the form $px + q$) on that interval. Hence, given that $u(a) = 0$, when $f = \delta_y$, we must have

$$u(x) = u(y)\frac{x-a}{y-a}\ , \ x \in (a,y)\,, \tag{180}$$

and

$$u(x) = u(y)\frac{b-x}{b-y} \quad , \quad x \in (y,b)\,. \tag{181}$$

It remains to calculate $u(y)$. Note that when $a < x < y$ the derivative $u'(x)$ is given by $\frac{u(y)}{y-a}$, and when $y < x < b$ the derivative $u'(x)$ is given by $-\frac{u(y)}{b-y}$ The jump in $-u'(x)$ when x crosses y from left to right must be 1, which gives

$$u(y)\left(\frac{1}{y-a}+\frac{1}{b-y}\right) = 1\,. \tag{182}$$

and hence

$$u(y) = \frac{(y-a)(b-y)}{b-a}\,. \tag{183}$$

Recall that $u(x)$ is the solution of (179) when $f(x) = \delta_y(x) = \delta(x-y)$. Let us denote this solution $G(x,y)$, to indicate also the dependence on y. From (180) and (181) we obtain

$$G(x,y) = \begin{cases} \frac{(x-a)(b-y)}{b-a}\,, & a \le x \le y\,, \\ \frac{(y-a)(b-x)}{b-a}\,, & y \le x \le b\,. \end{cases} \tag{184}$$

Note that $G(x,y) = G(y,x)$.

The solution of (179) for a general $f(x)$ is now given by

$$u(x) = \int_a^b G(x,y)f(y)\,dy\,. \tag{185}$$

The function G is called the Green's function of the problem (179). Note that formula (185) resembles the formula

$$x_i = \sum_j (A^{-1})_{ij} b_j \tag{186}$$

for the solution of the system $Ax = b$, where A is an $n \times n$ matrix and A^{-1} is its inverse. The only difference is that the indices i, j in (186) run through a finite set $\{1, 2, \ldots, n\}$, whereas the "indices" x, y in (185) run through an interval (a, b). Also, the summation in (186) is replaced by the integration in (185).

The above example illustrates an important point in PDE theory. Namely, the inverse operators to differential operators (such as the function $G(x,y)$ above) are sometimes more transparent than the differential operators themselves (such as the operaor $\frac{d^2}{dx^2}$ with the zero boundary conditions at the endpoints a, b in the example above). In the numerical approximations we have discussed, our (linear) differential operators were always represented by matrices, although the operators themselves (such as $-\frac{\partial^2}{\partial x^2}$) at the fist glance may not look like matrices. On the other hand the inverses of the differential operators, such as the one given by (185), do look quite similar to matrices (with indices which are

real numbers, rather than integers, and with summation replaced by integration, when compared with the usual formulae from linear algebra).

Lecture 21, 11/21/2017
(non-homogeneous problems - continuation)

the non-homogeneous ODE $\ddot{x} + \omega^2 x = f(t)$

Let us apply considerations from the last lecture to the equation

$$\ddot{x} + \omega^2 x = f(t)\,. \tag{187}$$

We have seen in previous lectures how to solve the initial-value (ODE) problem

$$\ddot{x} + \omega^2 x = 0\,, \quad x(0) = x_0\,, \quad \dot{x}(0) = x_1 \tag{188}$$

We can write down the solution explicitly:

$$x(t) = x_0 \cos \omega t + \frac{x_1}{\omega} \sin \omega t\,. \tag{189}$$

Going back to (187), let us assume that f vanishes for $t \in (-\infty, t_0)$ for some t_0, and let us search for a solution which also vanishes in $(-\infty, t_0)$. We can think of an oscillator which is at rest from the time $-\infty$ to time t_0, and after time t_0 forcing $f(t)$ is applied, and the oscillator may be "excited" by f into a non-trivial motion (once f becomes non-zero). Motivated by the discussion in the previous lecture, let us imagine that $f(t)$ is a superposition of infinitesimal "kicks" $f(s)\delta(t-s)\,ds$, in the sense that

$$f(t) = \int_{-\infty}^{\infty} f(s)\delta(t-s)\,ds\,, \tag{190}$$

where $\delta(t)$ is the Dirac function (discussed in the last lecture). Let X be the solution of (187) with $f(t) = \delta(t)$ satisfying $X(t) = 0$ for $t < 0$. The interpretation quite similar to what we discussed last time: we think of (187) as describing a physical system, such as an oscillator. Up to time $t = 0$ nothing is going on, and the system is at rest, corresponding to $X(t) = 0$ for $t < 0$. At time $t = 0$ the system receives a "kick", or an impulse of force. The kick is normalized so that the jump in the first derivative $\dot{X}(t)$ at $t = 0$ is 1. In other words, $\dot{X}(t) \to 1$ as $t \to 0_+$. The function X itself will be continuous at $t = 0$ (and at any other point, of course).
So we wish to solve

$$\ddot{X} + \omega^2 X = \delta(t)\,, \quad X(t) = 0 \text{ when } t < 0\,. \tag{191}$$

Based on our considerations above, it is not hard to see that the solution is

$$X(t) = \begin{cases} 0 & \text{when } t < 0\,, \\ \frac{1}{\omega} \sin \omega t & \text{when } t \geq 0\,. \end{cases} \tag{192}$$

Note that for $t \geq 0$ this coincides with (188) when $x_0 = 0$ and $x_1 = 1$.
The solution of (187) which vanishes as $t \to -\infty$ will then be given (assuming $f(t)$ also vanishes sufficiently quickly as $t \to -\infty$) by

$$X(t) = \int_{\infty}^{\infty} X(t-s) f(s)\, ds\,, \tag{193}$$

by considerations mirroring those in the last lecture concerning the equation $\dot{x} + ax = f(t)$. If we wish to solve

$$\ddot{x} + \omega^2 x = f(t) \qquad t \in (0,\infty)\,, \quad x(0) = x_0\,, \quad \dot{x}(0) = x_1\,, \tag{194}$$

we can use the following version of the formula (combined also with formula (189))

$$x(t) = x_0 \cos \omega t + \frac{x_1}{\omega} \sin \omega t + \int_0^t f(s) \frac{1}{\omega} \sin \omega (t-s)\, ds\ . \tag{195}$$

By now we already know how to interpret this: the first two terms represent the solution we would have for $f = 0$, which is given by (189), and the integral represents the superposition of the contributions from the "infinitesimal kicks" $\delta(t-s) f(s)\, ds$ from which we can imagine f being composed.

In the rest of the lecture we discussed applications of these formulae to non-homogeneous PDEs for which we can decompose the solution into Fourier modes (or more general eigenvalue modes), as explained in Chapter 8 of the textbook. For an example with the heat equation, see Section 8.3. Section 8.5 deals with the 2d wave equations. We have only done the wave equation in 1d so far, but the method works in any dimension, and is in fact essentially the same as for finite-dimensional systems

$$\ddot{x} + Ax = f(t) \tag{196}$$

where x is an $n-$dimensional vector with components $x_1, \ldots x_n$, the $n \times n$ matrix A is symmetric, positive definite, and the forcing terms $f(t)$ is again an $n-$vector with components $f_1, \ldots, f_n$ (which are functions of t).
The main idea is the same as already discussed previously in some of the past lectures: we can introduce new coordinates $y_1, \ldots, y_n$ in which the matrix A becomes diagonal, with the entries on the diagonal being strictly positive. Let us denote them $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$. In the new coordinates the equation becomes

$$\ddot{y}_j + \lambda_j y_j = g_j(t)\,, \qquad j = 1, 2, \ldots, n\,, \tag{197}$$

where $g_1(t), \ldots, g_n(t)$ are the coordinates of the forcing term in the new coordinate system. These equations do not interact with each other, and hence each of them can be solved independently, using formula (195).
The vibrations of a string or membrane can be approached in the same way. Let us consider for example and inhomogeneous string of length L (parametrized by the interval $(0, L)$), with density $\rho(x)$, tension T and forcing $f(x,t)$, which is fixed at the endpoints. The equation is

$$\rho(x) \frac{\partial^2 u}{\partial t^2} = T \frac{\partial^2 u}{\partial x^2} + f(x,t)\,, \qquad u(0,t) = 0\,, \quad u(L,t) = 0\,. \tag{198}$$

Let $\phi_j(x), \lambda_j$ be respectively the eigenfunctions and eigenvalues of the Sturm-Liouville problem

$$-T\phi'' = \lambda\rho(x)\phi\,, \quad \phi(0) = 0\,, \quad \phi(L) = 0\,. \tag{199}$$

We write $f(x,t) = \sum_j f_j(t)\rho(x)\phi_j(x)$ and search the solution $u(x,t)$ as

$$u(x,t) = \sum_j c_j(t)\phi_j(x)\,. \tag{200}$$

Substituting these expressions into (198), we obtain

$$\ddot{c}_k + \lambda_k c_k = f_k(t)\,, \quad k = 1, 2, \ldots \tag{201}$$

Each of these equations can now be solved using (195).

Lecture 22, 11/21/2017
resonance; fundamental solution of the heat equation

Resonance

In practice one often encounters situations when the forcing is periodic. An important case is described by the following equation:

$$\ddot{x} + \omega^2 x = be^{i\kappa t} \tag{202}$$

Here b can be a complex number, and the solution x can also be complex. The the physical quantity is then the real part of the solution. Note that when $x(t)$ is a complex solution, then its real part satisfies the same equation with $be^{i\kappa t}$ replaced by its real part. One can calculate the solution of (202) by the method we discussed last time, but one can try to use a shortcut which – as we will see – works when $\kappa^2 \neq \omega^2$.
We search a particular solution in the form

$$x(t) = Ae^{i\kappa t}\,. \tag{203}$$

With this Ansatz, the equation gives

$$-\kappa^2 A + \omega^2 A = b\,, \tag{204}$$

which can be solved when $\kappa^2 \neq \omega^2$:

$$A = \frac{b}{-\kappa^2 + \omega^2}\,. \tag{205}$$

The general solution of the ODE (202) then is

$$x(t) = C_1 e^{i\omega t} + C_2 e^{-i\omega t} + \frac{b}{-\kappa^2 + \omega^2} e^{i\kappa t}\,, \tag{206}$$

where C_1, C_2 are any complex numbers.

This formula obvious works only when $\kappa^2 \neq \omega^2$. It is an interesting excercise to work with the formula to calculate the limit $\kappa \to \pm\omega$. It will of course give the same result as we obtain from the calculations we discussed last time.

The main conclusion for us is that when $\kappa^2 \neq \omega^2$, the solution will stay bounded.

As the equation is linear, there is no difficulty in passing from (202) to the more general case

$$\ddot{x} + \omega^2 x = b_1 e^{i\kappa_1 t} + b_2 e^{i\kappa_2 t} + \cdots + b_m e^{i\kappa_m t}\,. \tag{207}$$

The general solution will be

$$x(t) = C_1 e^{i\omega t} + C_2 e^{-i\omega t} + \frac{b_1}{-\kappa_1^2 + \omega^2} e^{i\kappa_1 t} + \cdots + \frac{b_m}{-\kappa_m^2 + \omega^2} e^{i\kappa_m t}\,, \tag{208}$$

assuming, of course, that $\omega^2 \neq \kappa_j^2\,,\ \ j = 1, 2, \ldots, m$.

The above can be applied to the wave equation with a right-hand side of the from $f(x,t) = g(x)e^{i\kappa t}$.

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2} + g(x) e^{i\kappa t} \tag{209}$$

The solutions can again be complex, and for the physical interpretation we take the real part. Let us consider the equation on $(0, L)$ with the boundary conditions $u(0,t) = 0\,,\ u(L,t) = 0$. We write $g(x) = \sum_n g_n \sin\frac{\pi x}{L}$ and search the solution as

$$u(x,t) = \sum_n B_n(t) \sin\frac{\pi x}{L}\,. \tag{210}$$

For each $B_n(t)$ we get an equation of the form (202), with $\omega_n^2 = n^2 \frac{\pi^2 c^2}{L^2}$. The formula (206) will work if $\kappa^2 \neq \omega_n^2$ for all n for which $g_n \neq 0$. Of $\omega_n^2 = \kappa^2$ for some n then the mode given by this n is at resonance, will grow unboundedly if $g_n \neq 0$, and one has to work with a different formula.

In practice the resonance effect are important, and one has to take them very seriously in various engineering constructions. The computation of the possible resonant frequencies can be complicated for the real-world systems, and often has to realy on numerical simulation.

In the second part of the lecture we started discussing the fundamental solution of the 1d heat equation. The key formulae in the textbook in this context are 10.4.6 and 10.4.7 on page 451 see also the example on page 453.

In class we discussed some important properties of the *heat kernel*

$$\Gamma(x,t) = \frac{1}{\sqrt{4\pi t}}\ e^{-\frac{x^2}{4t}}\,. \tag{211}$$

The formula defines Γ for $x \in (-\infty, \infty)$ and $t > 0$, but one can consider it as a functions defined also for $t \in (-\infty, \infty)$ by setting $\Gamma(x,t) = 0$ for $t \leq 0$. Such a function is defined for all (x,t) with the exception of $(x,t) = (0,0)$, where it has a singularity. However the singularity is "under control". Note that $\Gamma \geq 0$ and in the class we saw in class that $\int_{-\infty}^{\infty} \Gamma(x,t)\, dx = 1$ when $t \geq 0$. It is also worth noting that the function $\Gamma(x,t)$ extended in this way is smooth everywhere except at the origin. [13]

The function Γ can be thought of as a response of out system (infinite heat conducting rod) to the following situation:
Up to time $t = 0$ the system is at rest (at zero temperature) and nothing is going on. At time $t = 0$ we inject into the system a unit amount of the heat energy, exactly at the origin. (This is of course an idealization, in practice heat energy cannot be concentrated at a point.) This idea can be mathematically captured by the equation

$$\frac{\partial \Gamma}{\partial t} = \frac{\partial^2 \Gamma}{\partial x^2} + \delta(x,t)\,, \qquad x \in (-\infty, \infty)\,,\ t \in (-\infty, \infty)\,. \tag{212}$$

where $\delta(x,t)$ is a two dimensional Dirac function (which can be thought of as $\delta(x,t) = \delta(x)\delta(t)$, where the functions on the right-hand-side are the one-dimensional Dirac functions). If instead injecting the unit amount of the hear energy at time $t = 0$ and location $x = 0$ we do it at time $t = s$ and location $x = y$, the equation will be

$$\frac{\partial \tilde{\Gamma}}{\partial t} = \frac{\partial^2 \tilde{\Gamma}}{\partial x^2} + \delta(x - y, t - s)\,, \qquad x \in (-\infty, \infty)\,,\ t \in (-\infty, \infty)\,. \tag{213}$$

and the solution will be $\tilde{\Gamma}(x,t) = \Gamma(x - y, t - s)\,.$

If one now considers the equation

$$\frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2} = f(x,t) \tag{214}$$

in all space and for $t \in (-\infty, \infty)$ assuming the system is "undisturbed" at $t \sim -\infty$ and f vanishes at $t \sim -\infty$ one can imagine f as a superposition "infinitesimal injections" of the heat energy $f(y,s)\, dy\, ds$ and the solution is a superposition of the solutions corresponding to these "infinitesim injections"

$$u(x,t) = \int_{-\infty}^{t} \int_{-\infty}^{\infty} \Gamma(x - y, t - s) f(y,s)\, ds\,. \tag{215}$$

If instead we solve the initial-value problem

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2} + f(x,t)\,, (x,t) \in \mathbf{R} \times (0, \infty)\,, \qquad u(x,0) = u_0(x)\,, \tag{216}$$

[13] However, it cannot be analytic across $t = 0$, as its Taylor expansion at any point $(x,0)$, $x \neq 0$ is trivial.

where u_0 is a given function (describing the temperature at time $t = 0$), then the solutions will be

$$u(x,t) = \int_{-\infty}^{\infty} \Gamma(x-y,t)u_0(y)\,dy + \int_0^t \int_{-\infty}^{\infty} \Gamma(x-y,t-s)f(y,s)\,dy\,ds\,. \quad (217)$$