Lecture 19

The "Rayleigh quotient" for approximating eigenvalues and eigenfunctions

`Relevant section of text: 5.6`

The *Rayleigh quotient* is named in honour of John Strutt, Lord Rayleigh (1842-1919), who made a great number of contributions to the study of sound and wave phenomena in general. He was awarded the 1904 Nobel Prize in Physics for the discovery of argon. A number of phenomena are named after him, e.g., Rayleigh scattering, Rayleigh waves. He also developed a number of important approximation methods which have become fundamental tools in applied mathematics and physics. For example, the so-called *Rayleigh-Schrödinger perturbation method*, an adaptation of his perturbation treatment for differential equations, is fundamental in quantum mechanics. The "Rayleigh quotient" is the basis of an important approximation method that is used in solid mechanics as well as in quantum mechanics. In the latter, it is used in the estimation of energy eigenvalues of nonsolvable quantum systems, e.g., many-electron atoms and molecules. A brief description of its use in quantum mechanics will be given at the end of this section.

In order to understand the method, we consider a simple BVP-eigenvalue problem,

$$u'' + \lambda u = 0, \qquad u(0) = u(L) = 0. \tag{1}$$

Of course, we know that solutions to this problem,

$$\lambda_n = \left(\frac{n\pi}{L}\right)^2, \qquad u_n(x) = \sin\left(\frac{n\pi x}{L}\right), \qquad n = 1, 2, \cdots. \tag{2}$$

But we'll pretend that we don't know the solutions and see what we can do. We'll write the above equation for a particular eigenvalue λ_n and eigenfunction u_n:

$$u_n'' + \lambda_n u_n = 0, \qquad u(0) = u(L) = 0. \tag{3}$$

Now multiply by u_n and integrate over $x \in [0, L]$:

$$\int_0^L u_n u_n'' + \lambda_n \int_0^L (u_n)^2 \, dx = 0. \tag{4}$$

(You may recall that this is how we showed that all λ_n were real and positive.) We integrate the first integral by parts to give:

$$\left[u_n(x) u_n'(x)\right]_0^L - \int_0^L (u_n')^2 \, dx. \tag{5}$$

The first terms vanish because of the boundary conditions. We substitute the above result into the equation and rearrange it to yield

$$\lambda_n = \frac{\int_0^L (u_n')^2 \, dx}{\int_0^L (u_n)^2 \, dx}. \tag{6}$$

The goal is to estimate the eigenvalue λ_n without actually solving the differential equation, analytically or numerically. Let's suppose that we wish to estimate the lowest eigenvalue λ_1. We'll consider the special case $L = 1$. The exact answer is $\lambda_1 = \pi^2 \approx 9.8696$, with eigenfunction $u_1(x) = \sin \pi x$.

The idea is to evaluate the ratio on the right-hand side – the "Rayleigh quotient" – for some functions $u(x)$ that satisfy the boundary conditions. Since we are trying to estimate the lowest eigenvalue, we'll use functions that have no zeros in (0,1) – the idea is to use a function that approximates the eigenfunction $u_1(x)$. We'll try several functions, and denote each one as $u_T(x)$ for "trial function":

1. The piecewise linear "hat" function,

$$u_T(x) = \begin{cases} x, & 0 \le x \le 1/2, \\ 1 - x, & 1/2 \le x \le 1. \end{cases} \tag{7}$$

   Note that $u_T'(x) = 1$ for $x \neq 1/2$. As such, the numerator of the Rayleigh quotient is

$$\int_0^1 (u_T'(x))^2 \, dx = 1. \tag{8}$$

   The denominator is

$$\int_0^1 (u_T(x))^2 \, dx = \int_0^{1/2} x^2 \, dx + \int_0^{1/2} (1 - x)^2 \, dx = \frac{1}{24} + \frac{1}{24} = \frac{1}{12}. \tag{9}$$

   Therefore the Rayleigh quotient associated with this trial function, which we'll denote as $RQ(u_T)$, has value 12, which lies above the true value $\lambda_1 = 9.8696$. The percentage relative error associated with this approximation

$$\frac{12 - 9.8696}{9.8696} \times 100 = 21\%. \tag{10}$$

2. The quadratic trial function

$$u_T(x) = x(1 - x) = x - x^2. \tag{11}$$

   After a little computation, we find that

$$RQ(u_T) = \frac{\int_0^1 (u_T'(x))^2 \, dx}{\int_0^1 (u_T(x))^2 \, dx} = 10. \tag{12}$$

   This value is closer to the true value 9.8626. The relative error of this approximation is computed to be 1.3%.

3. The function $u_T(x) = \sin \pi x$. In this case, the Rayleigh quotient is $\pi^2 \approx 9.8626$, which is precisely the value of the eigenvalue. Of course, we're cheating here – $u_T(x)$ is the exact eigenfunction!

Just for fun, let's compute the Rayleigh quotient associated with the function

$$u_T(x) = x(x - \frac{1}{2})(x - 1). \tag{13}$$

It has one node at $x = \dfrac{1}{2}$, and looks something like the second eigenfunction, $u_2 = \sin(2\pi x)$, so we'd expect that the RQ might yield an approximation to the second eigenvalue $\lambda_2 = 4\pi^2 \approx 39.4784$. Numerically (with the help of MAPLE), the Rayleigh quotient produced by this function is

$$RQ[u_T] = 42, \tag{14}$$

which is a reasonable approximation of λ_2. Note that once again it lies above the true value.

Let us now apply the Rayleigh quotient method to the following more general class of eigenvalue problems examined earlier,

$$u'' + \lambda\sigma(x)u = 0, \qquad u(0) = u(L) = 0, \qquad \sigma(x) > 0. \tag{15}$$

Multiplication by u and integration by parts yields the following expression for the Rayleigh quotient:

$$\lambda = \frac{\int_0^L [u'(x)]^2 \, dx}{\int_0^L [u(x)]^2 \sigma(x) \, dx}. \tag{16}$$

From this expression, we see that all eigenvalues are necessarily positive.

We now consider the special case $\sigma(x) = 1 + \dfrac{1}{10}x$, i.e.,

$$u'' + \lambda \left(1 + \frac{1}{10}x\right) u = 0, \qquad u(0) = u(\pi) = 0. \tag{17}$$

This problem is not exactly solvable – a numerical estimate to λ_1 is 0.86346.

The Rayleigh coefficient associated with a trial function $u_T(x)$ for this problem is given by

$$RQ(u_T) = \frac{\int_0^\pi [u_T'(x))^2 \, dx}{\int_0^\pi [u_T(x)]^2 (1 + \frac{1}{10}x) \, dx}. \tag{18}$$

We'll try a few test functions again:

1. The piecewise linear "hat" function,

$$u_T(x) = \begin{cases} x, & 0 \le x \le \pi/2, \\ \pi/2 - x, & 1/2 \le x \le \pi. \end{cases} \tag{19}$$

   The numerical value of the Rayleigh quotient corresponding to this function is 1.01630.

2. The quadratic function

$$u_T(x) = x(\pi - x) = \pi x - x^2. \tag{20}$$

   The numerical value of the Rayleigh quotient for this function is 0.875663, which is closer to the true value.

3. The function $u_T(x) = \sin x$.

   The numerical value of the Rayleigh quotient for this function is 0.86424, which is very close to the actual value. In retrospect, this might be expected, since the problem examined, with $\sigma(x) = 1 + x/10$, is a mild perturbation of the problem $u'' + \lambda u = 0$, $u(0) = u(\pi) = 0$, with exact solution $\sin x$ and eigenvalue $\lambda_1 = 1$.

In all of the examples studied above, we notice that the estimated value of the lowest eigenvalue λ_1 furnished by the Rayleigh coefficient is *above* the true value. This is not a coincidence. As we'll state more formally below, the Rayleigh coefficient always yields upper estimates to the lowest eigenvalue. For the general problem just studied we can state that

$$\lambda_1 = 0.863356\ldots = \min\left\{ \frac{\int_0^L (u')^2\, dx}{\int_0^L (u)^2 \sigma\, dx}, \quad u(0) = u(L) = 0, \quad u \text{ piecewise } C^1[0, L]\} \right\} \tag{21}$$

All values of the Rayleigh coefficient will be greater than or equal to the true value.

Rayleigh quotient for general Sturm-Liouville eigenvalue equation

We now compute the Rayleigh quotient for the general Sturm-Liouville equation

$$\frac{d}{dx}\left(p(x)\frac{du}{dx}\right) + q(x)u(x) + \lambda\sigma(x)u = 0, \qquad a < x < b. \tag{22}$$

with homogeneous boundary conditions

$$\begin{aligned} \alpha_1 u(a) + \alpha_2 u'(a) &= 0, \\ \beta_1 u(b) + \beta_2 u'(b) &= 0. \end{aligned} \tag{23}$$

Proceeding as before, we multiply both sides of Eq. (22) by u and integrate:

$$\int_a^b \frac{d}{dx}\left(p(x)\frac{du}{dx}\right) u(x)\, dx + \int_a^b q(x)u(x)^2\, dx + \int_a^b \lambda\sigma(x)u(x)^2\, dx = 0. \tag{24}$$

We integrate the first integral by parts and rearrange to give

$$\lambda = \frac{-pu\frac{du}{dx}|_a^b + \int_a^b [p(u')^2 - qu^2]\, dx}{\int_a^b u^2\sigma\, dx} = RQ(u). \tag{25}$$

In the following special cases of boundary conditions:

$$u(a) = u(b) = 0 \quad \text{(Dirichlet BCs)}, \tag{26}$$

$$u'(a) = u'(b) = 0 \quad \text{(Neumann BCs)}, \tag{27}$$

the first term in the numerator of the Rayleigh quotient vanishes.

The main result is that the Rayleigh quotient yields upper bounds to the true value of the lowest eigenvalue:

$$\lambda_1 = \min\left\{RQ(u) \mid u \text{ is piecewise } C^1[a,b] \text{ and satisfies BCs of problem}\right\} \tag{28}$$

Lecture 20

The Rayleigh Quotient - concluding remarks

Let us consider the eigenvalue problem studied in the previous lecture,

$$u'' + \lambda\sigma(x)u = 0, \qquad u(a) = u(b) = 0, \qquad \sigma(x) > 0 \quad x \in [a, b]. \tag{29}$$

Recall that the Rayleigh quotient associated with this problem is given by

$$RQ(u_T) = \frac{\int_a^b [u_T'(x)]^2 \, dx}{\int_a^b [u_T(x)]\sigma(x) \, dx}. \tag{30}$$

Here, $u_T(x)$ is a suitable "trial function": For continuous functions u_T with no zeros on (a, b), the Rayleigh quotient will yield upper estimates to the lowest eigenvalue λ_1, i.e.,

$$RQ(u_T) \geq \lambda_1. \tag{31}$$

The proof of this statement is a subject of the "Calculus of Variations," where we examine the rate of change of "things" that involve *functions* with respect to variations in the *functions*. Just as a brief aside, the Rayleigh quotient is an example of a *functional*, that is, a real-valued mapping. Here, RQ maps elements of a suitable function space to the positive reals.

Given that the Rayleigh quotient yields upper estimates, or "upper bounds", to the eigenvalue λ_1, one may well be interested in finding better and better approximations. At this point, we may well ask if there is a method that is better than simply "pulling trial functions u_T out of a hat". The answer is "Yes" - there is a systematic method to produce increasingly accurate estimates of eigenvalues, and it is known as the "Rayleigh-Ritz variational method." Once again, the name Rayleigh is involved. A German mathematician, W. Ritz, took the Rayleigh quotient and devised a quite simple, yet powerful, method of approximation.

The idea is to work with a set of basis functions on $[a, b]$ – we'll call them $\phi_k(x)$. (This may be a poor notation – these $\phi_k(x)$ functions are *not* the eigenfunctions of Eq. (29).) The ϕ_k functions need only to be linearly independent. That being said, it is nicer to work with *orthogonal* sets of functions (which we can always normalize to make an *orthonormal* set). But for the moment, we simply assume linear independence. For a given integer, $N > 0$, we then assume that our trial function $u_T(x)$ assumes the form,

$$u_T(x) = \sum_{k=1}^{N} c_k \phi_k(x), \tag{32}$$

where the coefficients c_k are to be determined.

We now substitute this expansion into the Rayleigh quotient in Eq. (30) and now consider it to be a function of the unknowns $c_1, c_2, \cdots, c_N$, i.e.,

$$RQ(u_T) = f(c_1, c_2, \cdots, c_N). \tag{33}$$

The "best" set of coefficients $\{c_1, c_2, \cdots, c_N\}$ are those which minimize the Rayleigh quotient, i.e., the function $f(c_1, c_2, \cdots, c_N)$.

That all sounds fine, but there is one complication: The Rayleigh quotient in Eq. (30) is – guess what? – a *quotient*. Quotients are cumbersome to work with – just go back to the quotient rule of differentiation. As such, it is convenient to reformulate the minimization problem as follows:

Minimize the numerator function

$$g(c_1, c_2, \cdots, c_N) = \int_a^b [u_T'(x)]^2 \, dx, \tag{34}$$

subject to the constraint that the denominator function be unity, i.e.,

$$h(c_1, c_2, \cdots, c_N) = \int_a^b [u_T(x)]^2 \sigma(x) \, dx = 1. \tag{35}$$

Just to get a better idea of what is involved, the numerator function g will be given by

$$\begin{aligned} g(c_1, \cdots, c_N) &= \int_a^b \left[\sum_{k=1}^N c_k \phi_k'(x) \right]^2 dx \\ &= \sum_{k=1}^N \sum_{l=1}^N c_k c_l \int_a^b \phi_k'(x) \phi_l'(x) \, dx. \end{aligned} \tag{36}$$

In other words, g is a *quadratic form* in the coefficients c_k. In some very "nice" cases, e.g., sine and cosine functions, not only do the ϕ_k form an orthogonal set, but also the ϕ_k' functions. And in the special case that the ϕ_k' are orthonormal, the g function reduces to a simple sum of squares of the c_k.

The denominator function h will be given by

$$\begin{aligned} h(c_1, \cdots, c_N) &= \int_a^b \left[\sum_{k=1}^N c_k \phi_k(x) \right]^2 \sigma(x) \, dx \\ &= \sum_{k=1}^N \sum_{l=1}^N c_k c_l \int_a^b \phi_k(x) \phi_l(x) \sigma(x) \, dx, \end{aligned} \tag{37}$$

which is also a quadratic form in the c_k.

As you may recall from an earlier Calculus course on multivariable functions, the minimization of an objective function with constraints may often be treated with the method of Lagrange multipliers. The above problem lends itself quite nicely to this approach – the result is a linear system of equations in the unknowns c_k, which will be left as an exercise.

In the next section, we describe briefly the use of the Rayleigh quotient in quantum mechanics, which may be of interest to students in the Mathematical Physics programme.

Application of Rayleigh quotient in quantum mechanics (Optional)

The Rayleigh quotient is an extremely important approximation method in quantum mechanics. We illustrate it very briefly below.

We start with the time-independent Schrödinger equation, an eigenvalue/BVP of the form

$$H\psi = E\psi \qquad \psi(a) = \psi(b) = 0. \tag{38}$$

(Here, a could be 0 or $-\infty$ and b could be ∞.) The components of this equation are as follows:

1. H is a second-order linear differential operator – the *hamiltonian* - specific to the quantum mechanical system of concern.

2. Ψ is the *wavefunction* of the system – in principle, it contains all of the information of the system. $|\Psi(x)|^2 dx$ represents the probability of finding the quantum particle in the interval $[x, x + dx]$.

3. E is the *energy eigenvalue* of the system.

In general, these eigenvalue equations are not exactly solvable. There is an additional condition on the wavefunction – that it be normalized, i.e.,

$$\langle \Psi, \Psi \rangle = \int_a^b \Psi^2 \, dx = 1. \tag{39}$$

This corresponds to the statement that the probability of finding the particle in the entire space $[a, b]$ is unity.

A typical goal in applications is to estimate the lowest eigenvalue, or *ground-state energy* of the quantum mechanical system. Using the same method as above, we multiply both sides by Ψ (in

general, Ψ^*, since Ψ can be complex-valued, but we ignore this technicality here) and integrate over $[a, b]$ to give

$$\langle \Psi, H\Psi \rangle = E\langle \Psi, \Psi \rangle, \tag{40}$$

where $\langle u, v \rangle$ denotes the inner product over $[a, b]$. The Rayleigh quotient then becomes

$$E = \frac{\langle \Psi, H\Psi \rangle}{\langle \Psi, \Psi \rangle}. \tag{41}$$

One may now employ various trial functions Ψ_T in an effort to obtain suitable estimates of the ground-state energy. All of these estimates are upper estimates.

A further step in this approximation method is known as the *Rayleigh-Ritz* or *variational method*, a systematic method to produce good trial functions Ψ_T which, in turn, will furnish good estimates of the eigenvalue E. Although it was discussed in the previous section, we describe it briefly below, as it is usually performed in quantum mechanics. In this case, one normally relies on the approximations furnished by a complete orthonormal basis set for $L^2[a, b]$.

Let $\{\phi_n\}_1^\infty$ denote such an orthonormal basis. We then consider approximations to Ψ of the form

$$\Psi_N = \sum_{n=1}^{N} c_n \phi_n. \tag{42}$$

Note that, from Parseval's identity,

$$\langle \Psi, \Psi \rangle = \sum_{n=1}^{N} (c_n)^2. \tag{43}$$

Let $\mathbf{c}_N = (c_1, c_2, \cdots, c_N)$. We then have

$$E_1 \leq E_N(\mathbf{c}) = \frac{\langle \Psi_N, H\Psi_N \rangle}{\langle \Psi_N, \Psi_N \rangle}. \tag{44}$$

For each value of N, we could try to minimize the above ratio. However, the ratio will produce a ratio of functions involving the coefficients c_n.

A more tractable problem is the following:

$$\text{minimize } \langle \Psi_N, H\Psi_N \rangle, \quad \text{subject to the condition } \sum_{n=1}^{N} c_i^2 = 1. \tag{45}$$

If we use Lagrangian multipliers to accomodate the condition on the c_i, then the above problem reduces to finding the eigenvalues of the matrix $\mathbf{H}$ that is the representation of the hamiltonian operator H in the orthonormal basis $\{\phi_n\}$ (Exercise).

This is the basis of most approximations methods in quantum chemistry that seek to estimate the energies of nontrivial quantum mechanical systems, e.g., many-electron atoms, molecules. In fact, the diagonalization of finite-dimensional matrix representations of the hamiltonian operator is performed virtually automatically in most applications. That being said, it is always a good idea to return to the source of this approximation method.

A brief look at numerical methods for PDEs

`Relevant section of text: Chapter 6`

Here we provide some idea of the basic methods behind numerical methods of solving – or, more precisely, providing approximations to – PDEs. In no way is this description to be considered complete. One could spend an entire term on numerical methods for PDEs.

Most, if not all, numerical methods for solving ODEs and PDEs are based on some kind of "finite differences" where

1. functions are defined on a set of discrete points, often referred to as a "grid" or "mesh," and
2. their derivatives are approximated by appropriate "finite differences" involving these values.

We shall be making good use of Taylor's theorem for the approximation of a function. In the following equation, x_0 is a reference point, and $\Delta x > 0$. Assuming that the required derivatives exist, we have, for $n \geq 1$,

$$f(x + \Delta x) = f(x_0) + f'(x_0)\Delta x + \frac{1}{2}f''(x_0)(\Delta x)^2 + \cdots + \frac{1}{n!}f^{(n)}(x_0)(\Delta x)^n + R_n, \tag{46}$$

where the remainder term R_n is given by

$$R_n = \frac{1}{(n+1)!}f^{(n+1)}(c_n)(\Delta x)^{n+1}, \tag{47}$$

and c_n lies between x_0 and $x_0 + \Delta x$.

In the case $n = 1$, we may use Eq. (46) to approximate the derivative $f'(x_0)$ by means of the well-known formula,

$$f'(x_0) \approx \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x}. \tag{48}$$

This is an example of a *finite difference* approximation. Since it involves the values of $f(x)$ at x_0 and $x_0 + \Delta x > x_0$, it is known as the *forward difference* formula.

If we replace Δx by $-\Delta x$ in Eq. (46), i.e.,

$$f(x - \Delta x) = f(x_0) - f'(x_0)\Delta x + \frac{1}{2}f''(x_0)(\Delta x)^2 + \cdots, \tag{49}$$

then we obtain another finite difference approximation for $f'(x_0)$,

$$f'(x_0) \approx \frac{f(x_0) - f(x_0 - \Delta x)}{\Delta x}. \tag{50}$$

Because this approximation involves the values of $f(x)$ at x_0 and $x_0 - \Delta x < x_0$, it is known as the *backward difference* formula.

If we add both sides of Eqs. (48) and (50) and divide by 2, we obtain another approximation for $f'(x_0)$,

$$f'(x_0) \approx \frac{f(x_0 + \Delta x) - f(x_0 - \Delta x)}{2\Delta x}. \tag{51}$$

Since this approximation involves points on either side of x_0, it is known as the *centered difference* approximation. In practice (see textbook), it is found that this approximation yields better estimates of $f'(x_0)$: We are essentially using information from both sides of x_0, as opposed to "one-sided" information in the case of the forward and backward difference schemes.

With an eye to the heat equation, which involves second derivatives, we look for a finite difference approximation to the second derivative, $f''(x_0)$. We set $n = 2$ in Taylor's formula, Eq. (46), and consider both cases Δx and $-\Delta x$,

$$\begin{aligned} f(x + \Delta x) &= f(x_0) + f'(x_0)\Delta x + \frac{1}{2}f''(x_0)(\Delta x)^2 + R_2^+, \\ f(x - \Delta x) &= f(x_0) - f'(x_0)\Delta x + \frac{1}{2}f''(x_0)(\Delta x)^2 + R_2^-, \end{aligned} \tag{52}$$

where $R_2^{\pm}$ denote appropriate remainder terms. If we add both equations, divide by $(\Delta x)^2$ and ignore the remainder terms, we obtain the following finite difference approximation,

$$f''(x_0) \approx \frac{f(x + \Delta x) - 2f(x_0) + f(x - \Delta x)}{(\Delta x)^2}. \tag{53}$$

Because it involves x_0 along with the points $x_0 \pm \Delta x$ on either side of it, this approximation is a *centered difference* approximation.

Application to simple ODE/IVP problems

Before looking at PDEs, it is instructive to consider the application of finite differences to a simple class of initial value problems for ODEs. Suppose that $u(x)$ is the unique solution to the IVP,

$$\frac{du}{dx} = g(u, x), \qquad u(x_0) = u_0. \tag{54}$$

(Of course, we are assuming that g satisfies the conditions for the existence of a unique solution to the IVP.) A simple example is $g(u, x) = Au$, where A is constant, i.e.,

$$\frac{du}{dx} = Au, \qquad u(x_0) = u_0. \tag{55}$$

The solution to this IVP is (Exercise),

$$u(x) = u_0 e^{A(x-x_0)}. \tag{56}$$

Our finite difference scheme will be defined on the *uniform grid* of points x_k on the real line defined by

$$x_k = x_0 + k\Delta x. \tag{57}$$

On this set of points, we denote the values of the solution $u(x)$ as follows,

$$u_k = u(x_k). \tag{58}$$

At the grid points $x = x_k$, the differential equation in Eq. (54) dictates that

$$u'(x_k) = g(u_k, x_k). \tag{59}$$

This is an *exact result*. But the finite difference method works only with values of $u(x)$ on the grid points, i.e., $u(x_k)$. Therefore, we must approximate the derivative on the LHS by a finite difference. If we use the forward difference approximation, then

$$u'(x_k) \approx \frac{u_{k+1} - u_k}{\Delta x}. \tag{60}$$

We now employ this approximation in Eq. (59) to give the approximation,

$$\frac{u_{k+1} - u_k}{\Delta x} \approx g(u_k, x_k). \tag{61}$$

The finite difference scheme involves replacing the "$\approx$" by an "$=$", i.e.,

$$\frac{u_{k+1} - u_k}{\Delta x} = g(u_k, x_k). \tag{62}$$

This is the *finite difference scheme* associated with the ODE in Eq. (54). Note that we can rearrange it as follows,

$$u_{k+1} = u_k + g(u_k, x_k)\Delta x, \qquad k = 0, 1, 2, \cdots. \tag{63}$$

We see that the approximation u_{k+1} is obtained from u_k, i.e., the approximation to the left of it. *This is basically using the linear approximation to estimate $u_{k+1} = u(x_{k+1})$ from a knowledge of $u_k = u(x_k)$.*

Eq. (63) defines an iterative procedure to compute the approximations u_k: Starting with $k = 0$ and the prescribed initial value

$$u_0 = u(x_0), \tag{64}$$

we compute successive estimates, i.e.,

$$\begin{aligned} u_1 &= u_0 + g(u_0, x_0)\Delta x, \\ u_2 &= u_1 + g(u_1, x_1)\Delta x, \\ &\vdots \\ u_k &= u_{k-1} + g(u_{k-1}, x_{k-1})\Delta x. \end{aligned} \tag{65}$$

At each step, we must compute the RHS of the ODE, i.e., the function $g(u, x)$ at $u = u_k$ and $x = x_k$.

Example: We consider again the special case $g(u) = Au$, with exact solution $u(x) = u_0 e^{A(x-x_0)}$. The finite difference equation (63) becomes

$$u_{k+1} = u_k + Au_k\Delta x. \tag{66}$$

In this special case, however, we may rewrite the above equation as follows,

$$u_{k+1} = (1 + A\Delta x)u_k. \tag{67}$$

Starting with u_0, note that

$$\begin{aligned} u_1 &= (1 + A\Delta x)u_0 \\ u_2 &= (1 + A\Delta x)u_1 = (1 + A\Delta x)^2 u_2 \\ &\vdots \\ u_k &= (1 + A\Delta x)^k u_0. \end{aligned} \tag{68}$$

We expect that as Δx gets smaller and smaller, the approximations u_k will be better. However, as Δx gets smaller, the points x_k get closer together and we move away from x_0 at a lesser rate. Suppose that we would like to get an approximation to the solution $u(x)$ at a fixed point $x > x_0$. And suppose that we choose Δx so that

$$x_0 + n\Delta x = x, \tag{69}$$

for some integer $n > 0$. In other words, we arrive at x after n steps of the above procedure. From the above equation, it follows that

$$n = \frac{x - x_0}{\Delta x}. \tag{70}$$

As expected, as Δx decreases, the number of steps n increases. In fact, as $\Delta x \to 0$, $n \to \infty$. If we rewrite the above equation as follows,

$$\Delta x = \frac{x - x_0}{n}, \tag{71}$$

and insert this result into the expression for u_k in Eq. (68), with $k = n$, we obtain

$$u_n = u_0 \left(1 + \frac{A(x - x_o)}{n}\right)^n. \tag{72}$$

Recalling the definition of e in terms of limits,

$$\begin{aligned} \lim_{n\to\infty} u_n &= u_0 \lim_{n\to\infty} \left(1 + \frac{A(x - x_o)}{n}\right)^n \\ &= u_0 e^{A(x-x_0)}, \end{aligned} \tag{73}$$

which is indeed the exact solution $u(x)$ at x. In this special case, we have shown that the finite difference scheme yields the exact result in the limit $\Delta x \to 0$.

We now return to the general ODE problem in Eq. (54) and consider the *centered finite difference scheme* associated with it. Once again, we consider approximations to this problem evaluated at the grid points x_k. Over these grid points, recall that the differential equation takes the form of Eq. (59). The centered finite difference approximation for the derivative on the left has the form

$$u'(x_k) = \frac{u_{k+1} - u_{k-1}}{2\Delta x}. \tag{74}$$

The resulting finite difference scheme for this problem will be

$$\frac{u_{k+1} - u_{k-1}}{2\Delta x} = g(u_k, x_k). \tag{75}$$

Rearranging, we obtain the difference equation,

$$u_{k+1} = u_{k-1} + 2g(u_k, x_k)\Delta x, \tag{76}$$

which should be compared with the finite difference scheme in Eq. (63) obtained from the forward difference method.

One quickly notes that the estimate at u_{k+1} is not determined from the estimate u_k immediately to its left but rather from u_{k-1}, which is two spaces to its left. Once again, we must start the procedure with the prescribed initial value u_0. But we now encounter a problem: If we set $k = 0$ in the above scheme in order to compute the next estimate, u_1, we need to know u_{-1}, which we don't! If we simply choose to avoid this problem and go on to compute u_2, we'll face the same problem when trying to compute u_3: we need to know u_1, which we don't!

In order to bypass this difficulty, we use the forward difference scheme of Eq. (65) to compute u_1 from u_0, and then proceed with the centered difference scheme of Eq. (76), i.e.,

$$\begin{aligned} u_1 &= u_0 + g(u_0, x_0)\Delta x, \\ u_k &= u_{k-2} + 2g(u_{k-1}, x_{k-1})\Delta x, \quad k \geq 2, \end{aligned} \tag{77}$$

where we have altered the indices in the second formula.

At first glance, it might appear that since all $u_{k,odd}$ are determined from u_1 and $u_{k,even}$ are determined from u_0, the two sequences are independent. Even worse, they could be viewed as separate *forward difference schemes*. In fact, they are coupled – note that for $k \geq 2$, u_k depends on both u_{k-2} as well as the function g evaluated over the other sequence, i.e., $g(u_{k-1}, x_{k-1})$. Therefore, the two sequences are not independent.

Example: We compare the results of the forward and centered difference schemes as applied to the simple ODE/IVP

$$u' = u, \qquad u(0) = 1, \tag{78}$$

with exact solution $u(x) = e^x$. In the table on the next page are presented the results corresponding to $\Delta x = 0.1$ for $0 \leq k \leq 20$. The second column presents the estimates u_k obtained from the forward difference scheme, Eq. (65) and the third column presents the estimates u_k obtained from the centered difference scheme, Eq. (76). The final column lists the exact values $u_k = u(x_k) = e^{x_k}$. Of course, both schemes start with the prescribed initial condition $u_0 = 1$. And, as discussed above, they share the same value u_1. But for larger k, we see that the forward difference scheme provides increasingly poorer estimates of the exact solution, as compared to the centered scheme.

x_k	Forward	Centered	Exact
0.0	1.00000	1.00000	1.00000
0.1	1.10000	1.10000	1.10517
0.2	1.21000	1.22000	1.22140
0.3	1.33100	1.34400	1.34986
0.4	1.46410	1.48880	1.49182
0.5	1.61051	1.64176	1.64872
0.6	1.77156	1.81715	1.82212
0.7	1.94872	2.00519	2.01375
0.8	2.14359	2.21819	2.22554
0.9	2.35795	2.44883	2.45960
1.0	2.59374	2.70796	2.71828
1.1	2.85312	2.99042	3.00417
1.2	3.13843	3.30604	3.32012
1.3	3.45227	3.65163	3.66930
1.4	3.79750	4.03637	4.05520
1.5	4.17725	4.45890	4.48169
1.6	4.59497	4.92815	4.95303
1.7	5.05447	5.44453	5.47395
1.8	5.55992	6.01705	6.04965
1.9	6.11591	6.64794	6.68589
2.0	6.72750	7.34664	7.38906

Lecture 21

Numerical methods (cont'd)

Application to PDEs

We now show some simple examples of how finite difference methods can be applied to PDEs. We start with the standard (homogeneous) heat equation with no sources,

$$\frac{\partial u}{\partial t} = k\frac{\partial^2 u}{\partial x^2}, \quad 0 \le x \le L. \tag{79}$$

For the moment, the boundary conditions and initial conditions will be ignored.

We first form a grid on the x-axis by setting $\Delta x = \dfrac{L}{N}$ for an $N > 0$, and defining the grid points,

$$x_m = m\Delta x, \qquad m = 0, 1, \cdots, N. \tag{80}$$

We'll also need to discretize the time variable: Let $\Delta t > 0$ and define

$$t_n = n\Delta t, \quad n = 0, 1, 2, \cdots. \tag{81}$$

The finite difference scheme will be concerned with the values of $u(x,t)$ on these grid points, i.e., the values $u(x_m, t_n)$: For each $n \ge 0$, the $N+1$ set of values $u(x_m, t_n)$, $0 \le m \le N$, form a "snapshot" of the temperature function u on the spatial grid points x_m.

For notational convenience, we let

$$u_m^{(n)} = u(x_m, t_n). \tag{82}$$

(This is the notation used by the textbook. The most important point is that the superscript "(n)" denotes to the nth time step.)

For the time derivative $\dfrac{\partial u}{\partial t}$, we'll employ the forward difference:

$$\frac{\partial u}{\partial t}(x_m, t_n) \approx \frac{u(x_m, t_{n+1}) - u(x_m, t_n)}{\Delta t} = \frac{u_m^{(n+1)} - u_m^{(n)}}{\Delta t}. \tag{83}$$

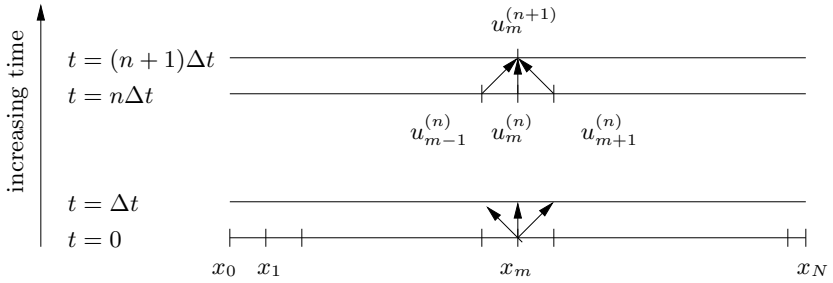
The spatial derivative $\dfrac{\partial^2 u}{\partial x^2}$ will be approximated by the following centered difference,

$$\frac{\partial^2 u}{\partial x^2}(x_m, t_n) \approx \frac{u(x_{m+1}, t_n) - 2u(x_m, t_n) + u(x_{m-1}, t_n)}{(\Delta x)^2} = \frac{u_{m+1}^{(n)} - 2u_m^{(n)} + u_{m-1}^{(n)}}{(\Delta x)^2}. \tag{84}$$

Substitution of these differences into the heat equation, Eq. (79), and rearrangement yields the following difference scheme,

$$u_m^{(n+1)} = u_m^{(n)} + \frac{k\Delta t}{(\Delta x)^2}\left[u_{m+1}^{(n)} - 2u_m^{(n)} + u_{m-1}^{(n)}\right]. \tag{85}$$

This scheme shows that the temperature at $x = x_m$ and time $t = t_{n+1}$ is computed from the temperatures at the three points $x = x_{m-1}, x_m, x_{m+1}$ at time $t = t_n$. Graphically, we can depict this relationship as follows:



This relationship also implies that each value $u_k^{(n)}$ is used to compute the three values $u_{k-1}^{(n+1)}$, $u_k^{(n)}$ and $u_{k+1}^{(n+1)}$ in the next time step, i.e., the temperature at the same spot along with the temperatures at the neighbouring spots. In other words, the temperature – or concentration of chemical in the case of the diffusion equation – x_k influences the temperatures at its neighbours $x_{k\pm1}$. But we should have expected this from our model of heat flow/diffusion. We'll return to this idea a little later.

Matrix representation of the finite difference scheme

If, for each $n \geq 0$ we place the components $u_m^{(n)}$ into an $N+1$-dimensional column vector, denoted as $\mathbf{u}^{(n)}$, then the finite difference scheme in Eq. (85) may be written vectorially as follows,

$$\begin{aligned} \mathbf{u}^{(n+1)} &= \mathbf{T}\mathbf{u}^{(n)} \\ &= (\mathbf{I} + s\mathbf{A})\mathbf{u}^{(n)}, \end{aligned} \tag{86}$$

where $\mathbf{I}$ and $\mathbf{A}$, hence $\mathbf{T}$, are $(N+1)\times(N+1)$ matrices: $\mathbf{I}$ is the identity matrix and $\mathbf{A}$ is a symmetric tridiagonal matrix with values -2 along the diagonal and 1 on each off-diagonal. The scalar s is given by

$$s = \frac{k\Delta t}{(\Delta x)^2}. \tag{87}$$

Note that the matrix $\mathbf{T}$ is constant, which implies that for any $n \geq 0$, $\mathbf{u}^{(n)}$ may be expressed in terms of the initial condition vector $u^{(0)}$ as follows,

$$\mathbf{u}^{(n)} = (\mathbf{I} + s\mathbf{A})^n \mathbf{u}^{(0)}. \tag{88}$$

The factor on the right has the appearance of something that could approach an exponential as $n \to \infty$ – in this case, however, there are matrices inside the brackets. This would then suggest a matrix exponential. The appearance of Δt and Δx in the s factor also complicates matters somewhat. We'll see later (in the section on Fourier transforms) that, indeed, exponentials are involved in the solution of the heat equation. At that point, we shall return to the above result.

Incorporation of initial values/boundary conditions

We now have a scheme to propagate the approximations, $u_m^{(n)} = u(x_m, t_n)$, in time. It now remains to incorporate the initial conditions and the boundary conditions for this problem.

1. **Initial values:** Recall that the initial condition was given by

   $$u(x, 0) = f(x). \tag{89}$$

   How do we incorporate this initial condition into our finite difference scheme. A "natural" choice would be to employ the values of $f(x)$ at the spatial grid points, i.e.,

   $$u_m^{(0)} = u(x_m, 0) = f(x_m), \quad 0 \leq m \leq N, \tag{90}$$

   which is traditionally done in computations.

   However, as a student, who is familiar with the concept of "sampling" in electrical engineering and signal processing, pointed out, it may (or may not!) be desirable to perform other kinds of "sampling" of the function values $f(x)$. Indeed, selecting the values $f(x_m)$ is a form of sampling. But it is a rather "crude" method. Another way to sample is to "convolve" the function $f(x)$ with a suitable "kernel," for example a Gaussian function of small width – the result is a kind of "local averaging" of $f(x)$ about each sample point x_m. This is a standard procedure in signal processing. That being said, we won't pursue this idea any further in our very simple numerical treatment of the heat equation.

2. **Boundary conditions:** Here we are considering the simple fixed-temperature BCs,

   $$u(0, t) = A, \quad u(L, t) = B, \tag{91}$$

which amounts to setting

$$u_0^{(n)} = A, \quad u_N^{(n)} = B, \quad n = 1, 2, \cdots. \tag{92}$$

Note: In general, the imposition of these endpoint conditions at each step may require that some extra attention be paid to points close to the endpoints, in particular, the points next to them. For the simple scheme constructed above, however, this is not the case. (It would, however, be the case if the thermal conductivity coefficient K_0 were a function of x.)

The roles of Δt and Δx

We expect that the approximations afforded by the finite difference scheme constructed above will improve as we make Δt and Δx smaller, since each partial derivative in the heat equation will be better approximated. These space and time steps, however, are not independent, as may have been suggested by the appearance of the factor

$$s = \frac{k\Delta t}{(\Delta x)^2} \tag{93}$$

in Eq. (85).

A "stability analysis" (see textbook) shows that the numerical scheme in Eq. (85) is guaranteed to be stable if $s \leq 1/2$, that is,

$$\frac{k\Delta t}{(\Delta x)^2} < \frac{1}{2}. \tag{94}$$

Otherwise, it is possible that errors are amplified exponentially - a kind of "resonance" phenomenon. Because of time constraints, we have simply stated the above result without derivation and mention that it is derived in the textbook by Haberman, Chapter 6, Section 6.3.4, "Fourier-von Neumann Stability Analysis."

The result implies that for a given spatial refinement $\Delta x > 0$, the time step Δt should be chosen as follows,

$$\Delta t < \frac{1}{2k}(\Delta x)^2. \tag{95}$$

The obvious consequence is that if the refinement of the spatial grid is increased, i.e., Δx is decreased, then the time step Δt must be decreased. This in turn implies that more iterations will be required to propagate the solution to a fixed time $T > 0$.

The underlying reason for this requirement on the time step Δt is to ensure that the changes in the temperature function are computed sufficiently accurately with respect to time before being propagated "outward" from points to neighbouring points.

Indeed, the relationship between Δx and Δt goes even deeper, and is related to the physics behind the diffusion process. To see this, a rearrangment of the s factor yields

$$\Delta x = \sqrt{\frac{s}{k}}\sqrt{\Delta t}. \tag{96}$$

Now recall Eq. (85), the finite difference approximation for the heat equation. In the limit $\Delta t \to 0$, the factor s should be unity, implying that

$$dx = \sqrt{k}\sqrt{dt}. \tag{97}$$

In other words, an infinitesimal change dx in space is connected with an infinitesimal change $\sqrt{dt}$ in time. This accounts for the phenomenon that one-dimensional diffusion proceeds with a speed proportional to $\sqrt{t}$. (Note also that the proportionality constant is given by $\sqrt{k}$ – recall that k is the thermal conductivity.) We shall verify this result in our solution of the heat equation via Fourier transforms.

Some computational examples

The finite difference scheme formulated above allows us to compute solutions to the heat equation in cases that cannot be solved in closed form. This, in reality, means *most cases* – since homogeneous rods (implying constant coefficients) represent a very small subset of the more general set of inhomogeneous, i.e., nonconstant, problems.

As a check, however, the scheme was applied to a case that could be solved in closed form – the results are shown in the top plot in Figure 1. The solutions are observed to approach the equilibrium solution which can easily be determined in closed form.

The bottom plot in Figure 1 shows the evolution of solutions for the case in which the thermal conductivity was assumed to be piecewise constant: $k = 2$ for $0 \leq k \leq 1/2$ and $k = 1/2$ for $1/2 < x \leq 1$. This models a rod that is formed by welding two rods of different composition at $x = 1/2$, with perfect thermal contact there. As discussed in Problem Set 4, the discontinuity in k produces a discontinuity in the first derivative of the temperature function $u(x, t)$ at $x = 1/2$ – this is necessary to ensure that there is no loss/gain of heat at the junction.

In Figure 2, we present numerical solutions for the case when heat sources are present, i.e., the PDE

$$\frac{\partial u}{\partial t} = k\frac{\partial^2 u}{\partial x^2} + a. \tag{98}$$

The top set of solutions correspond to a constant a value over the rod. In this case, the equilibrium solution can be computed in closed form. (It is given in the figure caption.) The bottom set of solutions are produced when there is a constant heat source over a subinterval, in this case, $[1/4, 1/2]$. The solutions are seen to converge to an equilibrium distribution – this distribution can also be computed in closed form but we omit the calculation. (Exercise: It will be piecewise linear over $[0, 1/4]$ and $[1/2, 1]$, and piecewise quadratic over $[1/4, 1/2]$.)

The finite difference scheme used in this section can be modified to accomodate other problems. First of all, the forward difference for the time derivative (known as "Euler method") could be replaced by a central difference – the results would be more accurate. And for more general problems that involve nonuniformities, i.e.,

$$c(x)\rho(x)\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(K_0(x)\frac{\partial u}{\partial x}\right), \tag{100}$$

one would have to revise the finite difference scheme for the spatial derivatives on the right. One method would be to expand the derivative by means of the product rule for differentiation and then apply appropriate finite differences to terms being differentiated. Another method would be to apply a finite difference to the term in brackets first and then apply a finite difference to that finite difference.

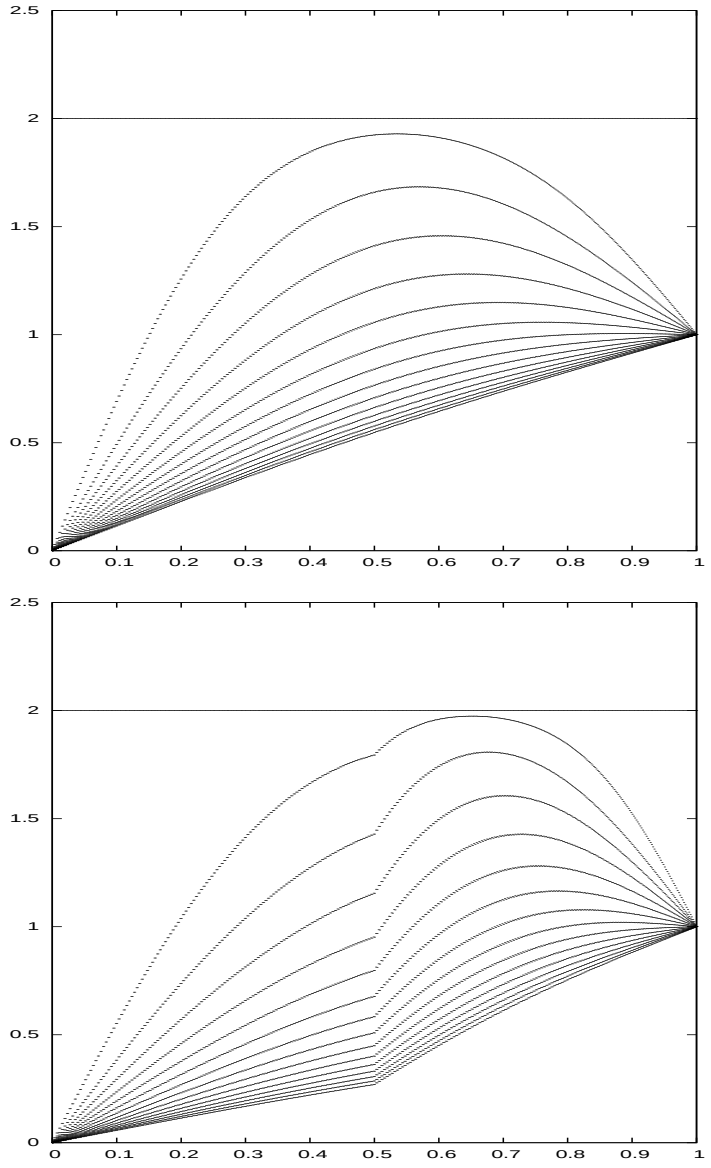


Figure 1. Solutions $u(x_m, t_n)$ to heat equation $u_t = ku_{xx}$ for $x \in [0, 1]$ with boundary conditions $u(0, t) = 0$, $u(1, t) = 1$ and initial condition $u(x, 0) = 2$. Solutions were computed using simple finite-difference method: $\Delta x = 1/500$, $\Delta t = (\Delta x)^2/4$.

Top: $k = 1$ for $0 \le x \le 1$.

Bottom: $k = 2$ for $0 \le x \le 1/2$ and $k = 1/2$ for $1/2 < x \le 1$.

In both cases, $u(x, t) \to u_{eq}(x)$. The discontinuity in the derivative at $x = 1/2$ in the bottom case is due to the discontinuity of the thermal conductivity k there. Because k is smaller for $x > 1/2$, the spatial derivatives of u must be larger in order to produce the fluxes necessary to transfer heat from the rod into the environment.

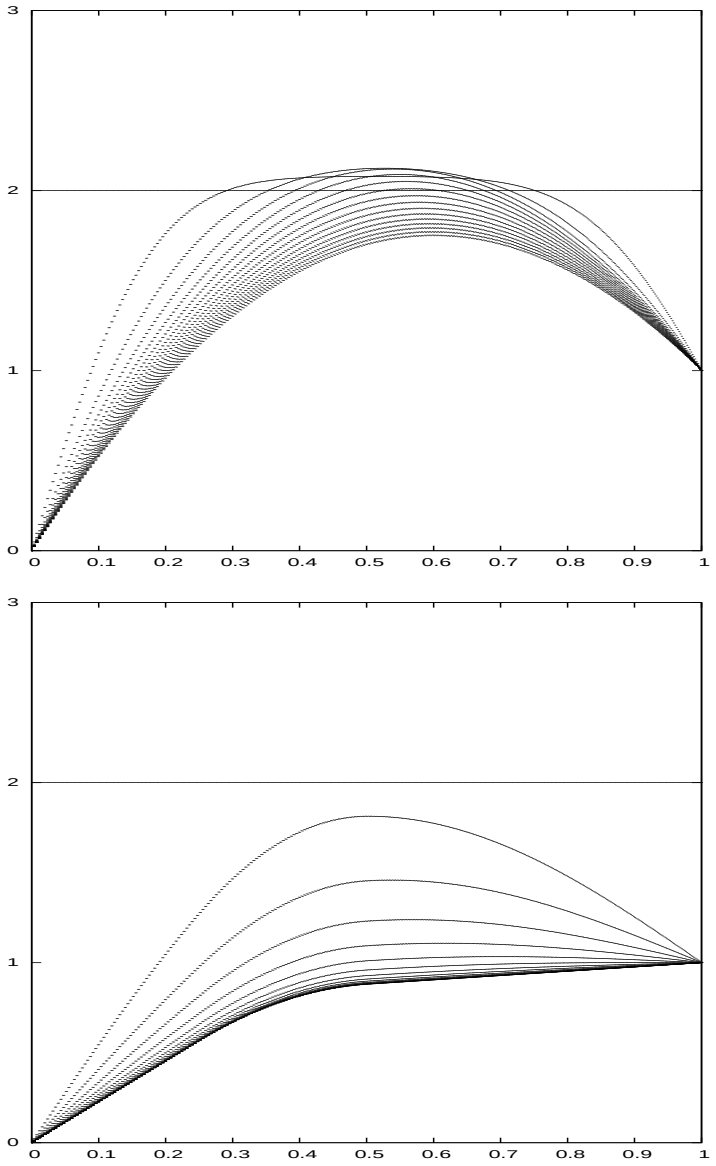


Figure 2. Solutions $u(x_m, t_n)$ to heat equation $u_t = ku_{xx} + a$ for $x \in [0, 1]$ with boundary conditions $u(0, t) = 0$, $u(1, t) = 1$ and initial condition $u(x, 0) = 2$. Solutions were computed using simple finite-difference method: $\Delta x = 1/500$, $\Delta t = (\Delta x)^2/4$.

Top: $k = 1$ and $a = 8$ for $0 \le x \le 1$. Here, the equilibrium solution is given by

$$u_{eq}(x) = -\frac{1}{2}ax^2 + \left(\frac{1}{2}a + 1\right)x. \tag{99}$$

Bottom: $k = 1$ for $0 \le x \le 1$; $a = 8$ for $\frac{1}{4} \le x \le \frac{1}{2}$ and $a = 0$ elsewhere. Here $u_{eq}(x)$ is linear on $[0, \frac{1}{4}]$ and $[\frac{1}{2}, 1]$ (no sources on these intervals) and quadratic on $[\frac{1}{4}, \frac{1}{2}]$ (constant source term a).

In both cases, $u(x, t) \to u_{eq}(x)$.