MA 7007: Numerical Solution of Differential Equations I
Steady States and Boundary Value Problems

Suh-Yuh Yang (楊肅煜)

Department of Mathematics, National Central University
Jhongli District, Taoyuan City 32001, Taiwan

E-mail: syyang@math.ncu.edu.tw
Website: http://www.math.ncu.eud.tw/∼syyang/

The heat equation

The heat equation is derived from Fourier's law and conservation of energy. The basic equation of one-dimensional case is given by
(https://en.wikipedia.org/wiki/Heat_equation)

$$u_t(x,t) = \big(\kappa(x)u_x(x,t)\big)_x + \psi(x,t), \quad a < x < b,\ t > 0.$$

1. $u(x,t)$ is the temperature at point x and time t.
2. $\kappa(x) > 0$ is the coefficient of heat conduction. If the material is homogeneous, then $\kappa(x) \equiv \kappa > 0$ is independent of x.
3. $\psi(x,t)$ is the heat source.

Initial condition: $u(x,0) = u^0(x)$.

Boundary condition:

1. Dirichlet boundary condition (prescribed temperature):
$$u(a,t) = \alpha(t) \quad \text{and} \quad u(b,t) = \beta(t), \quad t \geq 0.$$
2. Neumann boundary condition (prescribed heat flux):
$$u_x(a,t) = 0 \quad \text{and} \quad u_x(b,t) = 0 \quad \text{(insulated)}.$$

The steady-state problem

1. Let $\kappa(x) \equiv \kappa > 0$. If $\psi(x,t)$, $\alpha(t)$, $\beta(t)$ are all time independent, then $u(x,t)$ will converge towards steady state distribution satisfying

$$-\kappa u''(x) = \psi(x), \quad a < x < b \implies u''(x) = f(x), \quad a < x < b,$$

   where $f(x) := -\psi(x)/\kappa$. This is now a second order ODE for $u(x)$.

2. In what follows, we will consider $a = 0, b = 1$ and the Dirichlet boundary condition: $u(0) = \alpha$ and $u(1) = \beta$. (2-point boundary value problem)

Remark: The above steady-state problem can be solved exactly if we integrate f twice and then use the boundary conditions to fix the constants involved.

Example: $u(0) = 20, u(1) = 60, f(x) = -100e^x$
Solution: $u(x) = -100e^x + (100e - 60)x + 120$

A simple finite difference method

Consider the 2-point BVP:

$$u''(x) = f(x), \quad 0 < x < 1, \quad u(0) = \alpha \text{ and } u(1) = \beta.$$

1. Define the grid points $x_j = jh$, $0 \le j \le m+1$, of the interval $[0,1]$, where $h = 1/(m+1)$ is the mesh width (mesh size).
2. Let $U_j \approx u(x_j)$ denote the approximation to $u(x_j)$. From the boundary condition, we know $U_0 = \alpha$ and $U_{m+1} = \beta$, and so we have m unknown values $U_1, U_2, \cdots, U_m$ to compute.

A simple finite difference method (continued)

1. We can approximate the second derivative of u at x_j by

$$u''(x_j) \approx D^2 u(x_j) := \frac{1}{h^2}\Big(u(x_{j-1}) - 2u(x_j) + u(x_{j+1})\Big).$$

2. We then obtain an algebraic system of m linear equations in U_j:

$$\frac{1}{h^2}\Big(U_{j-1} - 2U_j + U_{j+1}\Big) = f(x_j), \quad \text{for } j = 1, 2, \cdots, m,$$

where $U_0 = \alpha$ and $U_{m+1} = \beta$.

Tridiagonal linear system

$$
\begin{aligned}
\frac{1}{h^2}(\alpha - 2U_1 + U_2) &= f(x_1) \qquad (i = 1) \\
\frac{1}{h^2}(U_1 - 2U_2 + U_3) &= f(x_2) \qquad (i = 2) \\
&\cdots \\
\frac{1}{h^2}(U_{m-2} - 2U_{m-1} + U_m) &= f(x_{m-1}) \qquad (i = m-1) \\
\frac{1}{h^2}(U_{m-1} - 2U_m + \beta) &= f(x_m) \qquad (i = m)
\end{aligned}
$$

This linear system can be written in the matrix from as

$$
\frac{1}{h^2}\begin{bmatrix}
-2 & 1 & & & & \\
1 & -2 & 1 & & & \\
 & 1 & -2 & 1 & & \\
 & & \ddots & \ddots & \ddots & \\
 & & & 1 & -2 & 1 \\
 & & & & 1 & -2
\end{bmatrix}
\begin{bmatrix} U_1 \\ U_2 \\ U_3 \\ \vdots \\ U_{m-1} \\ U_m \end{bmatrix}
=
\begin{bmatrix} f(x_1) - \alpha/h^2 \\ f(x_2) \\ f(x_3) \\ \vdots \\ f(x_{m-1}) \\ f(x_m) - \beta/h^2 \end{bmatrix},
$$

or more compactly $AU = F$, where A is weakly diagonally dominant and nonsingular (see Atkinson's book: *An Introduction to Numerical Analysis*, Theorem 8.2).

Measuring error

Let $\widehat{U} = [u(x_1), u(x_2), \cdots, u(x_m)]^\top$ be the vector of true values, then the error vector E defined by

$$E = U - \widehat{U}$$

contains the errors at each grid point.

We define some norms for the grid function E:

1. max-norm (∞-norm): $\|E\|_\infty := \max\limits_{1\le j\le m} |E_j| = \max\limits_{1\le j\le m} |U_j - u(x_j)|.$
2. 1-norm (discrete L^1-norm): $\|E\|_1 := h\sum\limits_{j=1}^{m} |E_j|.$
3. 2-norm (discrete L^2-norm): $\|E\|_2 := \left(h\sum\limits_{j=1}^{m} |E_j|^2\right)^{1/2}.$

Note: Let A be an $m \times m$ real matrix. Then the matrix norm of A associated with a usual vector norm (∞, 1-norm, 2-norm) is equal to the matrix norm of A associated with the corresponding grid function norm (∞, 1-norm, 2-norm).

Local truncation error (LTE)

The LTE is defined by replacing U_j with the true solution $u(x_j)$ in the finite difference formula:

$$\tau_j = \frac{1}{h^2}\Big(u(x_{j-1}) - 2u(x_j) + u(x_{j+1})\Big) - f(x_j), \quad j = 1, 2, \cdots, m.$$

By the Taylor series expansions and $u''(x_j) = f(x_j)$, we know that

$$\begin{aligned} \tau_j &= \Big(u''(x_j) + \frac{1}{12}h^2 u^{(4)}(x_j) + O(h^4)\Big) - f(x_j) \\ &= \frac{1}{12}h^2 u^{(4)}(x_j) + O(h^4) \\ &= O(h^2) \quad \text{for } 0 < h \ll 1. \end{aligned}$$

In other words, the local truncation error is of $O(h^2)$.

Global error

1. Define $\tau := [\tau_1, \tau_2, \cdots, \tau_m]^\top$, then we have $\tau = A\widehat{U} - F$ and

$$AE = A(U - \widehat{U}) = F - (F + \tau) = -\tau.$$

2. Rewrite as the system of equations

$$\frac{1}{h^2}\left(E_{j-1} - 2E_j + E_{j+1}\right) = -\tau(x_j) \quad \text{for } j = 1, 2, \cdots, m$$

with the boundary conditions $E_0 = E_{m+1} = 0$.

3. This can be interpreted as the centered difference discretization of the ODE

$$e''(x) = -\tau(x) \quad \text{for } 0 < x < 1$$

with boundary conditions $e(0) = 0$ and $e(1) = 0$.

4. Roughly speaking, since $\tau(x) \approx \frac{1}{12}h^2 u^{(4)}(x)$, integrating twice shows that

$$e(x) \approx -\frac{1}{12}h^2 u''(x) + \frac{1}{12}h^2\left(u''(0) + \left(u''(1) - u''(0)\right)x\right)$$

and hence the global error should be $O(h^2)$ in ∞-norm, discrete L^1-norm and discrete L^2-norm.

Stability

We rewrite the linear system $AE = -\tau$ as $A^h E^h = -\tau^h$. Then $E^h = -(A^h)^{-1}\tau^h$ and

$$\|E^h\| = \|(A^h)^{-1}\tau^h\| \le \|(A^h)^{-1}\|\|\tau^h\|,$$

where we use a grid function norm (∞, discrete L^1 or discrete L^2). We know that $\|\tau^h\| = O(h^2)$ and we are hoping the same will be true of $\|E^h\|$. So we need

$$\|(A^h)^{-1}\| \le C \qquad \text{for all } 0 < h \ll 1.$$

Definition: Suppose a finite difference method for a linear BVP gives a sequence of matrix equations of the form $A^h U^h = F^h$, where h is the mesh width. We say that the method is *stable* if $(A^h)^{-1}$ exists for all h sufficiently small (say, for $0 < h < h_0$) and if there is a constant C, independent of h, such that

$$\|(A^h)^{-1}\| \le C \quad \text{for all } 0 < h < h_0.$$

Consistency and convergence

1. We say that a finite difference method is *consistent* with the differential equation and boundary conditions if

$$\|\tau^h\| \to 0 \quad \text{as } h \to 0.$$

   Typically the method has $\|\tau^h\| = O(h^p)$ for some integer $p > 0$, and then the method is certainly consistent.

2. A finite difference method is said to be *convergent* if $\|E^h\| \to 0$ as $h \to 0$.

Note that here $\|\cdot\|$ is a grid function norm.

Fundamental theorem of finite difference methods

- For a stable method, we have

$$\|E^h\| \le \|(A^h)^{-1}\|\|\tau^h\| \le C\|\tau^h\|,$$

and if the method is consistent, then

$$\|E^h\| \le C\|\tau^h\| \to 0 \quad \text{as } h \to 0.$$

- **Fundamental theorem of finite difference methods:**
For a linear finite difference scheme approximating a linear PDE

$$\text{consistency} + \text{stability} \Longrightarrow \text{convergence}$$

In particular,

$$O(h^p) \text{ local truncation error} + \text{stability} \Longrightarrow O(h^p) \text{ convergence}$$

2-norm and eigenvalues

1. Since the matrix A is symmetric, the 2-norm of A is equal to its spectral radius,

$$\|A\|_2 = \rho(A) = \max_{1\le p\le m} |\lambda_p|.$$

2. The matrix A^{-1} is also symmetric, so

$$\|A^{-1}\|_2 = \rho(A^{-1}) = \max_{1\le p\le m} |\lambda_p^{-1}| = \left(\min_{1\le p\le m} |\lambda_p|\right)^{-1}.$$

3. The m eigenvalues of A are given by

$$\lambda_p = \frac{2}{h^2}\Big(\cos(p\pi h) - 1\Big) \quad \text{for } p = 1, 2, \cdots, m.$$

The eigenvector u^p corresponding to λ_p has components u_j^p given by

$$u_j^p = \sin(p\pi jh) \quad \text{for } j = 1, 2, \cdots, m.$$

Stability in the 2-norm

We see that the smallest eigenvalue of A (in magnitude) is

$$\begin{aligned}\lambda_1 &= \frac{2}{h^2}\Big(\cos(\pi h) - 1\Big) \\ &= \frac{2}{h^2}\Big(-\frac{1}{2}\pi^2 h^2 + \frac{1}{24}\pi^4 h^4 + O(h^6)\Big) \\ &= -\pi^2 + O(h^2),\end{aligned}$$

where we use $\cos(x) = 1 - x^2/2! + x^4/4! - \cdots$. Hence,

$$\|(A^h)^{-1}\|_2 \approx \frac{1}{\pi^2},$$

and the method is stable in the 2-norm. Moreover, we have

$$\|E^h\|_2 \le \|(A^h)^{-1}\|_2 \|\tau^h\|_2 \approx \frac{1}{\pi^2}\|\tau^h\|_2 = O(h^2) \to 0 \quad \text{as } h \to 0.$$

Thus, the method is convergent with the order of accuracy $O(h^2)$ in the discrete L^2-norm.

Eigenvalue and eigenfunction

The jth component of the vector Au^p is

$$\begin{aligned}
(Au^p)_j &= \frac{1}{h^2}\left(u^p_{j-1} - 2u^p_j + u^p_{j+1}\right) \\
&= \frac{1}{h^2}\left(\sin(p\pi(j-1)h) - 2\sin(p\pi jh) + \sin(p\pi(j+1)h)\right) \\
&= \frac{1}{h^2}\left(\sin(p\pi jh)\cos(p\pi h) - 2\sin(p\pi jh) + \sin(p\pi jh)\cos(p\pi h)\right) \\
&= \lambda_p u^p_j \quad \text{for } j = 1, 2, \cdots, m,
\end{aligned}$$

where we define $u^p_0 := 0, u^p_{m+1} := 0$. This is consistent with the fact $Au^p = \lambda_p u^p$.

Note that the eigenvector $u^p = [u^p_1, u^p_2, \cdots, u^p_m]^\top$ with $u^p_j = \sin(p\pi jh)$ is closely related to the eigenfunction of the differential operator $\frac{\partial^2}{\partial x^2} := \prime\prime$.

Eigenvalue and eigenfunction (continued)

The functions $u^p(x) = \sin(p\pi x)$, $p = 1, 2, \cdots$, satisfy

$$\frac{\partial^2}{\partial x^2} u^p(x) = (-p^2\pi^2) u^p(x) := \mu_p u^p(x) \quad \text{and} \quad u^p(0) = u^p(1) = 0.$$

Therefore, $u^p(x) = \sin(p\pi x)$, $p = 1, 2, \cdots$, are eigenfunctions of the differential operator $\frac{\partial^2}{\partial x^2}$ on $[0, 1]$ with homogeneous boundary conditions.

Note that the m eigenvalues of A are given by

$$\begin{aligned} \lambda_p &= \frac{2}{h^2}\Big(\cos(p\pi h) - 1\Big) = \frac{2}{h^2}\Big(-\frac{1}{2}p^2\pi^2h^2 + \frac{1}{24}p^4\pi^4h^4 + \cdots\Big) \\ &= -p^2\pi^2 + O(h^2) \quad \text{as } h \to 0^+ \text{ for } p \text{ fixed.} \end{aligned}$$

Eigenfunctions are used to analyze the differential operator $\frac{\partial^2}{\partial x^2}$.

Norm equivalence

1. Consider the usual vector norms $\|\cdot\|_1$, $\|\cdot\|_2$ and $\|\cdot\|_\infty$ on $\mathbb{R}^N$. Show that for all $\boldsymbol{x} \in \mathbb{R}^N$, we have

$$\begin{aligned} \|\boldsymbol{x}\|_\infty \le \|\boldsymbol{x}\|_1 &\le N\|\boldsymbol{x}\|_\infty, \\ \|\boldsymbol{x}\|_\infty \le \|\boldsymbol{x}\|_2 &\le \sqrt{N}\|\boldsymbol{x}\|_\infty, \\ \|\boldsymbol{x}\|_2 \le \|\boldsymbol{x}\|_1 &\le \sqrt{N}\|\boldsymbol{x}\|_2. \end{aligned}$$

2. The exact error vector $E := U - \widehat{U} \in \mathbb{R}^m$ can be viewed as a grid function. Show that

$$\begin{aligned} h\|E\|_\infty \le \|E\|_1 &\le \|E\|_\infty, \\ \sqrt{h}\|E\|_\infty \le \|E\|_2 &\le \|E\|_\infty, \\ \sqrt{h}\|E\|_2 \le \|E\|_1 &\le \|E\|_2, \end{aligned}$$

where $\|\cdot\|_1$, $\|\cdot\|_2$ and $\|\cdot\|_\infty$ are 1-D grid function norms.

(Note that $2LM \le L^2 + M^2$ for all $L, M \in \mathbb{R}$)

Max-norm stability

1. We have demonstrated that A is stable in the 2-norm and $\|E^h\|_2 = O(h^2)$, which implies $\|E^h\|_1 \le \|E^h\|_2 = O(h^2)$. Suppose that we want a bound on $\|E^h\|_\infty := \max\limits_{1\le j\le m} |E_j^h|$. We can obtain one such bound directly from the bound for the 2-norm in the following way:

$$\|E^h\|_\infty \le \frac{1}{\sqrt{h}}\|E^h\|_2 = O(h^{3/2}) \quad \text{for} \quad h \to 0.$$

   However, this does not show the second order accuracy that we hope to have.

2. In order to show that $\|A^{-1}\|_\infty$ is uniformly bounded in h, i.e., $\|A^{-1}\|_\infty = O(1)$ ($\Longrightarrow \|E^h\|_\infty \le \|A^{-1}\|_\infty \|\tau^h\|_\infty = O(h^2)$ as $h \to 0$),
   in what follows, we will introduce the Green's function solution to the BVP:

$$u''(x) = f(x) \quad 0 < x < 1, \quad u(0) = \alpha, \;\; u(1) = \beta.$$

Green's function solution and Dirac delta function

1. We consider the Green's function solution to the following BVP:

$$u''(x) = f(x) \quad 0 < x < 1, \quad u(0) = \alpha, \ \ u(1) = \beta.$$

For any fixed point $\bar{x} \in (0,1)$, the Green's function $G(x;\bar{x})$ is the function of x that solves the above BVP with the particular source term $f(x) := \delta(x - \bar{x})$ and $\alpha = \beta = 0$, where $\delta(x - \bar{x})$ is the "Dirac delta function (δ-function)" centered at $\bar{x}$.

2. The Dirac delta function $\delta(x)$ can be loosely thought of as a function on the real line which is zero everywhere except at the origin, where it is infinite,

$$\delta(x) = \begin{cases} +\infty & x = 0, \\ 0 & x \neq 0, \end{cases}$$

and which is also constrained to satisfy the identity $\int_{-\infty}^{\infty} \delta(x)dx = 1$. The Dirac delta is not a function in the traditional sense as no function defined on the real numbers has these properties. The Dirac delta function can be rigorously defined either as a *distribution* or as a *measure.*

An approximation to the delta function

Let $\varepsilon > 0$. We consider a sharply peaked function $\varphi_\varepsilon(x)$ that is nonzero only on an interval $(-\varepsilon, \varepsilon)$ near the origin and has the property that

$$\int_{-\infty}^{\infty} \varphi_\varepsilon(x)dx = \int_{-\varepsilon}^{\varepsilon} \varphi_\varepsilon(x)dx = 1.$$

For example, we might take

$$\varphi_\varepsilon(x) = \begin{cases} (\varepsilon + x)/\varepsilon^2 & \text{if } -\varepsilon \le x \le 0, \\ (\varepsilon - x)/\varepsilon^2 & \text{if } 0 \le x \le \varepsilon, \\ 0 & \text{otherweise.} \end{cases}$$

This piecewise linear function is the "hat function" with width ε and height $1/\varepsilon$. Then we can think the δ-function as the limiting case of such functions as $\varepsilon \to 0^+$.

δ-function arising from differentiating disconti. function

Consider the Heaviside function

$$H(x) = \begin{cases} 0 & x < 0, \\ 1 & x \geq 0. \end{cases}$$

1. For $x \neq 0$, $H(x)$ is constant and so $H'(x) = 0$.
2. At $x = 0$ the derivative is not defined in the classical sense.
3. But it we smooth out the function a little bit, making it continuous and differentiable by changing $H(x)$ only on the interval $(-\varepsilon, \varepsilon)$, then the resulting function $H_\varepsilon(x)$ is differentiable everywhere and has a derivative $H'_\varepsilon(x)$ that looks something like $\varphi_\varepsilon(x)$.
4. The exact shape of $H'_\varepsilon(x)$ depends on how we choose $H_\varepsilon(x)$, but note that regardless of its shape, its integral must be 1, since

$$\int_{-\infty}^{\infty} H'_\varepsilon(x)dx = \int_{-\varepsilon}^{\varepsilon} H'_\varepsilon(x)dx = H_\varepsilon(\varepsilon) - H_\varepsilon(-\varepsilon) = 1 - 0 = 1.$$

By letting $\varepsilon \to 0$, we are led to define $H'(x) = \delta(x)$.

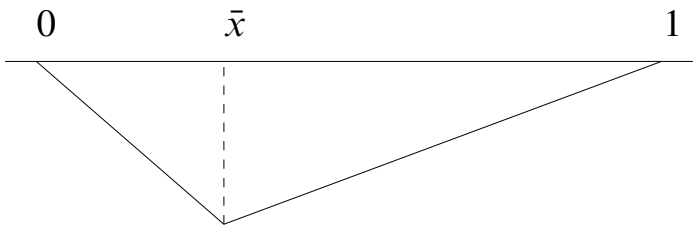
Green's function $G(x;\bar{x})$

Now let us go back to the BVP:

$$u''(x) = f(x) \quad 0 < x < 1, \quad u(0) = \alpha, \;\; u(1) = \beta.$$

If we interpret the problem as a steady-state heat conduction with source $\psi(x) = -\kappa f(x)$ (\$ $\kappa = 1$), then setting $f(x) = \delta(x - \bar{x})$ in the BVP is the mathematical idealization of a heat sink that has a unit magnitude but that is concentrated near a single point $\bar{x}$.

With $f(x) = \delta(x - \bar{x})$, a heat sink at $\bar{x}$, we have the minimum temperature at $\bar{x}$, rising linearly ($\because u''(x) = 0$ away from $\bar{x}$) to each side, as shown in below figure: ($\alpha = 0 = \beta$)



This figure shows a typical Green's function $G(x;\bar{x})$ for one particular choice of $\bar{x}$.

Green's function $G(x;\bar{x})$

To complete the definition of $G(x;\bar{x})$, we need to know the minimum value $G(\bar{x};\bar{x})$. This value is determined by the fact that the jump in slope at this point must be 1, since

$$u'(\bar{x}+\varepsilon) - u'(\bar{x}-\varepsilon) = \int_{\bar{x}-\varepsilon}^{\bar{x}+\varepsilon} u''(x)dx = \int_{\bar{x}-\varepsilon}^{\bar{x}+\varepsilon} \delta(x-\bar{x})dx = 1.$$

Therefore, one can check that the piecewise linear function $G(x;\bar{x})$ is given by

$$G(x;\bar{x}) = \begin{cases} (\bar{x}-1)x & \text{for } 0 \le x \le \bar{x}, \\ \bar{x}(x-1) & \text{for } \bar{x} \le x \le 1. \end{cases}$$

1. If we replaced $f(x)$ with $c\delta(x-\bar{x})$ for any constant c, the solution to the BVP would be $cG(x;\bar{x})$.
2. Any linear combination of Green's functions at different points $\bar{x}$ is a solution to the BVP with the corresponding linear combination of delta functions on the right-hand side.

An example

If we want to solve

$$u''(x) = 3\delta(x - 0.3) - 5\delta(x - 0.7)$$

with $u(0) = u(1) = 0$, the solutions is simply

$$u(x) = 3G(x; 0.3) - 5G(x; 0.7).$$

This is a piecewise linear function with jumps in slope of magnitude 3 at $x = 0.3$ and -5 at $x = 0.7$.

If the right-hand side is a sum of weighted delta functions at any number of points,

$$f(x) = \sum_{k=1}^{n} c_k \delta(x - x_k),$$

then the solution to the BVP is

$$u(x) = \sum_{k=1}^{n} c_k G(x; x_k).$$

General $f(x)$

Suppose $f(x)$ is not a discrete sum of delta functions. We can view this as a continuous distribution of point sources, with $f(\bar{x})$ being a density function for the weight assigned to the delta function at $\bar{x}$, i.e.,

$$f(x) = \int_0^1 f(\bar{x})\delta(x-\bar{x})d\bar{x}. \qquad \text{(treated as a Riemann sum)}$$

This suggests that the solution to $u''(x) = f(x)$, still with $u(0) = u(1) = 0$, is

$$u(x) = \int_0^1 f(\bar{x})G(x;\bar{x})d\bar{x},$$

and indeed it is.

Note: The delta function has the fundamental property that

$$\int_{-\infty}^{\infty} f(x)\delta(x-a)dx = f(a),$$

and, in fact,

$$\int_{a-\varepsilon}^{a+\varepsilon} f(x)\delta(x-a)dx = f(a), \quad \text{for all } \varepsilon > 0.$$

The full solution of the BVP

We now introduce two new function $G_0(x)$ and $G_1(x)$ defined by the BVPs:

$$G_0''(x) = 0, \quad G_0(0) = 1, \quad G_0(1) = 0$$

and

$$G_1''(x) = 0, \quad G_1(0) = 0, \quad G_1(1) = 1.$$

Then the solutions are

$$G_0(x) = 1 - x \quad \text{and} \quad G_1(x) = x.$$

The full solution to

$$u''(x) = f(x) \quad 0 < x < 1, \quad u(0) = \alpha, \ \ u(1) = \beta$$

is thus

$$u(x) = \alpha G_0(x) + \beta G_1(x) + \int_0^1 f(\bar{x}) G(x; \bar{x}) d\bar{x} \qquad (*)$$

or equivalently

$$u(x) = \left(\alpha - \int_0^x \bar{x} f(\bar{x}) d\bar{x} \right) (1 - x) + \left(\beta + \int_x^1 (\bar{x} - 1) f(\bar{x}) d\bar{x} \right) x.$$

The enlarged linear system $AU = F$

Let us return to the study of the max-norm stability of the finite difference method, which will be based on explicitly determining the inverse matrix for the matrix arising in this discretization.

Using the fact $U_0 = \alpha$ and $U_{m+1} = \beta$, we then consider the enlarged linear system $AU = F$, where now

$$A = \frac{1}{h^2}\begin{bmatrix} h^2 & 0 & & & & \\ 1 & -2 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & -2 & 1 \\ & & & & 0 & h^2 \end{bmatrix}, \quad U = \begin{bmatrix} U_0 \\ U_1 \\ U_2 \\ \vdots \\ U_m \\ U_{m+1} \end{bmatrix}, \quad F = \begin{bmatrix} \alpha \\ f(x_1) \\ f(x_2) \\ \vdots \\ f(x_m) \\ \beta \end{bmatrix}.$$

The inverse of matrix A: $B = A^{-1}$

Let B denote the $(m+2) \times (m+2)$ inverse of A, $B = A^{-1}$. We will index the elements of B by B_{00} through $B_{m+1,m+1}$ in the obvious manner. Let B_j denote the jth column of B for $j = 0, 1, \ldots, m+1$. Then

$$AB_j = e_j,$$

where e_j is the jth column of the identity matrix. We can view this as a linear system to be solved for B_j.

The first column B_0 corresponds to the problem with $\alpha = 1, f(x) = 0$, and $\beta = 0$, and so we expect B_0 to be a discrete approximation of the function $G_0(x)$. In fact, the first column of B has elements obtained by evaluating G_0 at the grid points,

$$B_{i0} = G_0(x_i) = 1 - x_i.$$

Similarly, the last ($j = m+1$) column of B has elements

$$B_{i,m+1} = G_1(x_i) = x_i.$$

The inverse of matrix A: $B = A^{-1}$ (continued)

The interior columns ($1 \le j \le m$) correspond to the Green's function for zero boundary conditions and the source concentrated at a single point, since $F_j = 1$ and $F_i = 0$ for $i \neq j$. Note that this is a discrete version of $h\delta(x - x_j)$, namely, $h\varphi_h(x - x_j)$ (see page 20). We expect that the column B_j will be discrete approximation to the function $hG(x; x_j)$.

In fact, it is easy to check that

$$B_{ij} = hG(x_i; x_j) = \begin{cases} h(x_j - 1)x_i, & i = 1, 2, \cdots, j, \\ h(x_i - 1)x_j, & i = j, j+1, \cdots, m. \end{cases}$$

An arbitrary right-hand side F for the linear system can be written as

$$F = \alpha e_0 + \beta e_{m+1} + \sum_{j=1}^{m} f_j e_j,$$

and the solution $U = BF$ is

$$U = \alpha B_0 + \beta B_{m+1} + \sum_{j=1}^{m} f_j B_j,$$

with elements

$$U_i = \alpha(1 - x_i) + \beta x_i + h \sum_{j=1}^{m} f_j G(x_i; x_j),$$

which is the discrete analogue of (*), see page 26.

A modified BVP

Suppose we define a function $v(x)$ by

$$v(x) = \alpha(1-x) + \beta x + h\sum_{j=1}^{m} f_j G(x; x_j).$$

Then $U_i = v(x_i)$ and $v(x)$ is the piecewise linear function that interpolates the numerical solution. This function $v(x)$ is the exact solution to the BVP

$$v''(x) = h\sum_{j=1}^{m} f(x_j)\delta(x-x_j), \quad v(0) = \alpha, \quad v(1) = \beta.$$

Thus we can interpret ths discrete solution as the exact solution to a modified problem in which the right-hand side $f(x)$ has been replaced by a finite sum of delta functions at the grid points x_j, with weights $hf(x_j) \approx \int_{x_{j-1/2}}^{x_{j+1/2}} f(x)dx$.

The max-norm stability

To verify max-norm stability of the numerical method, we must show that $\|B\|_\infty$ is uniformly bounded as $h \to 0$. The infinity norm of the matrix is given by

$$\|B\|_\infty = \max_{0 \le i \le m+1} \sum_{j=0}^{m+1} |B_{ij}|,$$

the maximum row sum of elements in the matrix. The intermediate rows are dense and the first and last elements are bounded by 1. The other m elements of each of these rows are all bounded by h, and hence

$$\sum_{j=0}^{m+1} |B_{ij}| \le 1 + 1 + mh < 3$$

Every row sum is bounded by 3 at most, and so $\|A^{-1}\|_\infty < 3$ for all h, and stability is proved.

The max-norm stability (continued)

Note that we have also worked out the inverse of the original matrix A,

$$A = \frac{1}{h^2}\begin{bmatrix} -2 & 1 & & & & \\ 1 & -2 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & -2 & 1 \\ & & & & 1 & -2 \end{bmatrix}.$$

Because the first row of B consists of zeros beyond the first element, and the last row consists of zeros, except for the last element, it is easy to check that the inverse of the $m \times m$ matrix is the $m \times m$ central block of B consisting of B_{11} through B_{mm}. The max-norm of this matrix is bounded by 1 for all h, so our original formulation is stable as well.

Neumann boundary conditions

Let us consider the second-order differential equation

$$u''(x) = f(x) \quad \text{for } 0 < x < 1,$$

with the Neumann BC at left-endpoint, $u'(0) = \sigma$, and the Dirichlet BC at right-endpoint, $u(1) = \beta$. Therefore, the approximation U_0 is one of the unknowns. Then the first row of matrix A in the enlarged linear system at page 27 must be modified to model $u'(0) = \sigma$, $u(1) = \beta$.

First approach: first order accuracy

1. Using the one-sided expression, $u'(0) \approx D_+u(x_0) := \dfrac{u(x_1) - u(x_0)}{h}$, we have

$$(U_1 - U_0)/h = \sigma.$$

This is only first order accurate, since the local truncation error is

$$\begin{aligned}
\tau_0 &= \frac{1}{h}\Big(u(x_1) - u(x_0)\Big) - \sigma \\
&= \frac{1}{h}\Big(u(x_0) + hu'(x_0) + \frac{1}{2}h^2u''(x_0) + O(h^3) - u(x_0)\Big) - \sigma \\
&= u'(x_0) + \frac{1}{2}hu''(x_0) + O(h^2) - \sigma = \frac{1}{2}hu''(x_0) + O(h^2) = O(h).
\end{aligned}$$

2. We obtain the system of linear equations for the unknowns $U_0, \cdots, U_{m+1}$:

$$\frac{1}{h^2}\begin{bmatrix} -h & h & & & & \\ 1 & -2 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & -2 & 1 \\ & & & & 0 & h^2 \end{bmatrix}\begin{bmatrix} U_0 \\ U_1 \\ U_2 \\ \vdots \\ U_m \\ U_{m+1} \end{bmatrix} = \begin{bmatrix} \sigma \\ f(x_1) \\ f(x_2) \\ \vdots \\ f(x_m) \\ \beta \end{bmatrix}.$$

3. The global error is only $O(h)$. In this case, it is mainly due to $\tau_0 = O(h)$.

Second approach: second order accuracy

1. Using a centered approximation to $u'(0) = \sigma$ (second order accuracy), we introduce another unknown U_{-1} and use the following two equations:

$$\begin{aligned} \frac{1}{h^2}(U_{-1} - 2U_0 + U_1) &= f(x_0), \\ \frac{1}{2h}(U_1 - U_{-1}) &= \sigma. \end{aligned}$$

   This results in a linear system of $m+3$ equations in $m+3$ unknowns.

2. Eliminating U_{-1} from above equations, we have:

$$\frac{1}{h}(-U_0 + U_1) = \sigma + \frac{h}{2}f(x_0),$$

   which reduces the linear system to one with only $m+2$ equations for $m+2$ unknowns $U_0, U_1, \cdots, U_{m+1}$.

Second approach: second order accuracy (continued)

1. The resulting linear system can be written in the matrix form:

$$\frac{1}{h^2}\begin{bmatrix} -h & h & & & & \\ 1 & -2 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & -2 & 1 \\ & & & & 0 & h^2 \end{bmatrix}\begin{bmatrix} U_0 \\ U_1 \\ U_2 \\ \vdots \\ U_m \\ U_{m+1} \end{bmatrix} = \begin{bmatrix} \sigma + \frac{h}{2}f(x_0) \\ f(x_1) \\ f(x_2) \\ \vdots \\ f(x_m) \\ \beta \end{bmatrix}.$$

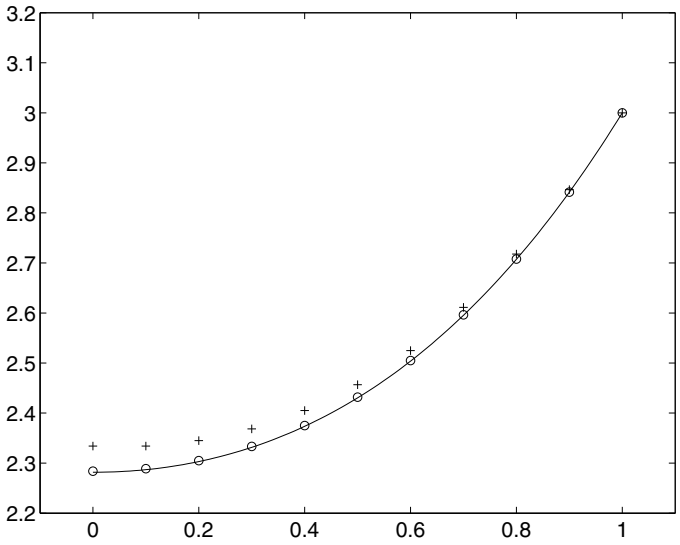
   Note that the matrix is exactly the same as the previous matrix in the first approach, but with different right-hand side, the first component is changed from σ to $\sigma + \frac{h}{2}f(x_0)$.

2. Alternative idea: By the Taylor series expansion, we have

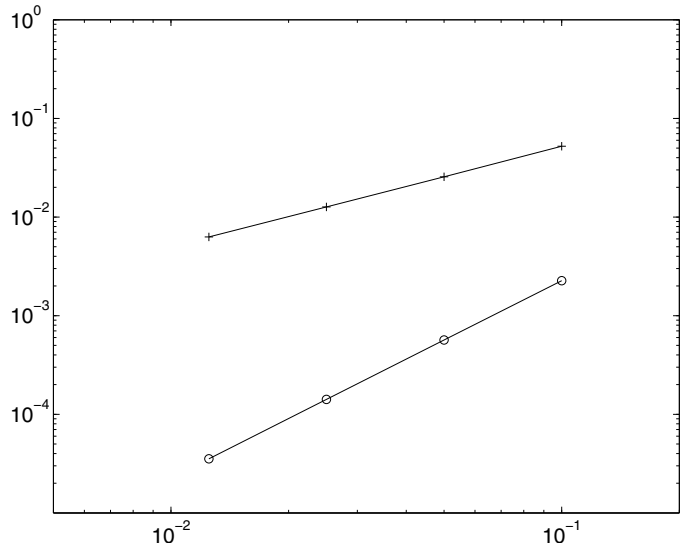
$$\frac{u(x_1) - u(x_0)}{h} \approx u'(x_0 + \frac{h}{2}) = u'(x_0) + \frac{h}{2}u''(x_0) + O(h^2) = \sigma + \frac{h}{2}f(x_0) + O(h^2).$$

Neumann boundary condition

(a)



(b)



(a) Exact and finite difference solutions to the steady-state heat equation $u''(x) = e^x$, $u'(0) = 0$ and $u(1) = 3$. The solid line is the true solution $u(x) = e^x - x + 4 - e$. The plus sign shows a solution on a grid with 20 points using first approach. The circle shows the solution on the same grid using second approach.

(b) A log-log plot of the max-norm error as the grid is refined is also shown for each case.

Third approach: second order accuracy

1. We could use a second order accurate one-sided approximation based on the three unknowns U_0, U_1, and U_2: (cf. Chapter 1)

$$\frac{1}{2h}(-3U_0 + 4U_1 - U_2) = \sigma.$$

That is,

$$\frac{1}{h^2}\left(-\frac{3h}{2}U_0 + 2hU_1 - \frac{h}{2}U_2\right) = \sigma.$$

2. The resulting linear system is given by

$$\frac{1}{h^2}\begin{bmatrix} -\frac{3h}{2} & 2h & -\frac{h}{2} & & & & \\ 1 & -2 & 1 & & & & \\ & 1 & -2 & 1 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & 1 & -2 & 1 & \\ & & & & 0 & h^2 \end{bmatrix}\begin{bmatrix} U_0 \\ U_1 \\ U_2 \\ \vdots \\ U_m \\ U_{m+1} \end{bmatrix} = \begin{bmatrix} \sigma \\ f(x_1) \\ f(x_2) \\ \vdots \\ f(x_m) \\ \beta \end{bmatrix}.$$

The use of this equation slightly disturbs the tridiagonal structure but adds little to the cost of solving the system of equations.

Well-posedness (適定性) of BVPs

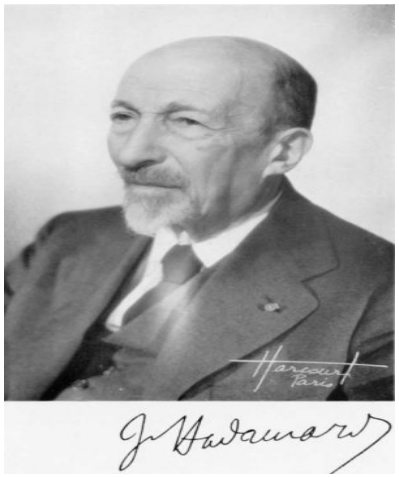
Well-posed problem (defined by Jacques Hadamard): mathematical models of physical phenomena should have the properties that

(1) A solution exists;

(2) The solution is unique;

(3) The solution's behavior changes continuously with the data.

But, we will show that even seemingly simple BVPs may fail to be well posed.



Jacques Salomon Hadamard (French mathematician, 1865-1963)

Example: Consider the following BVP with Neumann BCs at both ends,

$$\begin{cases} u''(x) = f(x) \quad \text{for } 0 < x < 1, \\ u'(0) = \sigma_0 \text{ and } u'(1) = \sigma_1. \end{cases}$$

Not well posed

Using the centered difference to discretize the differential equation and the second approach to deal with the Neumann BCs, we obtain the discrete linear system:

$$\frac{1}{h^2}\begin{bmatrix} -h & h & & & & & \\ 1 & -2 & 1 & & & & \\ & 1 & -2 & 1 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & 1 & -2 & 1 & \\ & & & & h & -h & \end{bmatrix}\begin{bmatrix} U_0 \\ U_1 \\ U_2 \\ \vdots \\ U_m \\ U_{m+1} \end{bmatrix} = \begin{bmatrix} \sigma_0 + \frac{h}{2}f(x_0) \\ f(x_1) \\ f(x_2) \\ \vdots \\ f(x_m) \\ -\sigma_1 + \frac{h}{2}f(x_{m+1}) \end{bmatrix}.$$

The matrix is singular, and in general the system has no solution or it has infinitely many solutions. This isn't a failure in our numerical model. It reflects that the problem is *not well posed*, and the differential equation will also have either no solution or infinitely many solutions. Note that $A[1, 1, \cdots, 1]^\top = \mathbf{0}$

Fredholm alternative: Let $\boldsymbol{Ax} = \boldsymbol{b}$ describe p equations in p unknowns. Then either (1) or (2) holds, *but not both*:

1. There is exactly one solution for each arbitrary $\boldsymbol{b}$. ($\boldsymbol{A}$ is nonsingular)
2. There is a nonzero solution $\boldsymbol{z} \neq \mathbf{0}$ to $\boldsymbol{Az} = \mathbf{0}$. ($\boldsymbol{A}$ is singular)

$\sigma_0 = \sigma_1 = 0$ **and** $f(x) \equiv 0$

We consider the case of $\sigma_0 = \sigma_1 = 0$ and $f(x) \equiv 0$. In other words, both ends of the rod are insulated, there is no heat flux through the ends, and there is no heat source within the rod.

Recall the BVP is a simplified equation for finding the steady-state solution of $u_t(x,t) = \kappa u_{xx}(x,t) + \psi(x,t)$ with some initial data $u^0(x)$. How does $u(x,t)$ behave with time?

1. Total heat energy must be conserved in t: $\int_0^1 u(x,t)dx = \int_0^1 u^0(x)dx, t \geq 0$.
2. Diffusion of the heat tends to redistribute it until it is uniformly distributed throughout the rod, so we expect the steady state solution $u(x) = c$. By conservation of energy, $c = \int_0^1 u^0(x)dx$.

But any $u(x) = c$ is a solution of the steady-state BVP. It has infinitely many solutions. The physical problem has only one solution, but in attempting to simplify it by solving for the steady state alone, we have thrown away a crucial piece of data, which is the heat content of the initial data for the heat equation.

$\sigma_0 = \sigma_1 = 0$

1. Suppose we have a source term $f(x)$ and $f(x) < 0$ everywhere, then we are constantly adding heat to the rod. (Note that $f(x) = -\psi(x)/\kappa \Rightarrow \psi(x) > 0$). Since no heat can escape through the insulated ends, we expect the temperature to keep rising without bound.

   In this case we never reach a steady state, and the BVP has no solution.

2. If f is positive over part of the interval and negative elsewhere, and the net effect of the heat sources and sinks exactly cancels out, then we expect that a steady state might exist.

   In fact, solving the BVP exactly by integrating twice and trying to determine the constants of integration from the boundary conditions show that a solution exists only if $\int_0^1 f(x)dx = 0$, in which case there are infinitely many solutions.

$$\int_0^t u''(s)ds = \int_0^t f(s)ds \Rightarrow u'(t) - u'(0) = \int_0^t f(s)ds$$

$$\Longrightarrow \begin{cases} 0 = \sigma_1 - \sigma_0 = u'(1) - u'(0) = \int_0^1 f(s)ds. \\ \int_0^x u'(t)dt = \int_0^x \int_0^t f(s)dsdt \Rightarrow u(x) = u(0) + \int_0^x \int_0^t f(s)dsdt, \end{cases}$$

where $u(0)$ can be arbitrary.

$\sigma_0 \neq 0$ and/or $\sigma_1 \neq 0$

If σ_0 and/or σ_1 are nonzero, then there is heat flow at the boundaries and the net heat source must cancel the boundary fluxes. Since

$$u'(1) - u'(0) = \int_0^1 u''(x)dx = \int_0^1 f(x)dx,$$

this requires

$$\int_0^1 f(x)dx = \sigma_1 - \sigma_0.$$

With this condition, we have infinitely many solutions:

$$\int_0^t u''(s)ds = \int_0^t f(s)ds \Rightarrow u'(t) - u'(0) = \int_0^t f(s)ds \Rightarrow u'(t) = \sigma_0 + \int_0^t f(s)ds$$

$$\Rightarrow \int_0^x u'(t)dt = \sigma_0 x + \int_0^x \int_0^t f(s)dsdt \Rightarrow u(x) = \sigma_0 x + u(0) + \int_0^x \int_0^t f(s)dsdt,$$

where $u(0)$ can be arbitrary.

$\sigma_0 \neq 0$ and/or $\sigma_1 \neq 0$

Similarly, the singular linear system $AU = F$,

$$\frac{1}{h^2}\begin{bmatrix} -h & h & & & & & \\ 1 & -2 & 1 & & & & \\ & 1 & -2 & 1 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & 1 & -2 & 1 & \\ & & & & h & -h & \end{bmatrix}\begin{bmatrix} U_0 \\ U_1 \\ U_2 \\ \vdots \\ U_m \\ U_{m+1} \end{bmatrix} = \begin{bmatrix} \sigma_0 + \frac{h}{2}f(x_0) \\ f(x_1) \\ f(x_2) \\ \vdots \\ f(x_m) \\ -\sigma_1 + \frac{h}{2}f(x_{m+1}) \end{bmatrix},$$

has a solution (in fact infinitely many solutions) only if F is orthogonal to the null space of $A^\top$. *Proof:* Assume that we have a solution U, i.e., $AU = F$, and $A^\top V = 0$. Then $\langle F, V\rangle = \langle AU, V\rangle = \langle U, A^\top V\rangle = \langle U, 0\rangle = 0$.

Note that we have $A^\top [1, h, \cdots, h, 1]^\top = 0$. This gives the condition

$$\frac{h}{2}\left(f(x_0) + 2\sum_{i=1}^{m} f(x_i) + f(x_{m+1})\right) = \sigma_1 - \sigma_0,$$

which is the composite trapezoidal rule approximation to $\int_0^1 f(x)dx = \sigma_1 - \sigma_0$.

A general linear second order equation

We now consider the more general linear differential equation

$$a(x)u''(x) + b(x)u'(x) + c(x)u(x) = f(x) \quad \text{for } a < x < b$$

with two Dirichlet boundary conditions, $u(a) = \alpha$ and $u(b) = \beta$. This equation can be discretized by using the centered difference approximations,

$$a_i\left(\frac{U_{i-1} - 2U_i + U_{i+1}}{h^2}\right) + b_i\left(\frac{U_{i+1} - U_{i-1}}{2h}\right) + c_iU_i = f_i, \qquad (*)$$

where $a_i := a(x_i)$, $b_i := b(x_i)$ and $f_i := f(x_i)$, with $U_0 = \alpha$ and $U_{m+1} = \beta$. This gives $AU = F$, where

$$A = \frac{1}{h^2}\begin{bmatrix} (h^2c_1 - 2a_1) & (a_1 + hb_1/2) & & & \\ (a_2 - hb_2/2) & (h^2c_2 - 2a_2) & (a_2 + hb_2/2) & & \\ \ddots & \ddots & \ddots & & \\ & (a_{m-1} - hb_{m-1}/2) & (h^2c_{m-1} - 2a_{m-1}) & (a_{m-1} + hb_{m-1}/2) & \\ & & (a_m - hb_m/2) & (h^2c_m - 2a_m) \end{bmatrix} \qquad (2.67)$$

and

$$U = \begin{bmatrix} U_1 \\ U_2 \\ \vdots \\ U_{m-1} \\ U_m \end{bmatrix}, \qquad F = \begin{bmatrix} f_1 - (a_1/h^2 - b_1/2h)\alpha \\ f_2 \\ \vdots \\ f_{m-1} \\ f_m - (a_m/h^2 + b_m/2h)\beta \end{bmatrix}. \qquad (2.68)$$

The better discretization

1. The above discretization may not be the best discretization to use for certain problems of this type even if it has second-order accuracy. Often the physical problem has certain properties that we would like to preserve with our discretization, and it is important to understand the underlying problem and be aware of its mathematical properties before blindly applying a numerical method.
2. Consider heat conduction in a rod with varying heat conduction properties, where $\kappa(x)$ varies with x and $\kappa(x) > 0$ for all $a < x < b$,

$$(\kappa(x)u'(x))' = f(x) \quad a < x < b,$$

with two boundary conditions, $u(a) = \alpha$ and $u(b) = \beta$. Applying the product rule to the above differential equation, we obtain

$$\kappa(x)u''(x) + \kappa'(x)u'(x) = f(x) \quad a < x < b,$$

and then apply the discretization (*). However, this is not the best approach because the resulting linear system may not be symmetric.

The better discretization (continued)

It is better to discretize the physical problem directly,

$$\kappa(x_{i+1/2})u'(x_{i+1/2}) = \kappa_{i+1/2}\Big(\frac{U_{i+1}-U_i}{h}\Big),$$

and the analogous approximation at $x_{i-1/2}$. Differencing these then gives a centered approximation to $(\kappa u')'$ at the grid x_i,

$$\begin{aligned}(\kappa u')'(x_i) &\approx \frac{1}{h}\left\{\kappa_{i+1/2}\Big(\frac{U_{i+1}-U_i}{h}\Big) - \kappa_{i-1/2}\Big(\frac{U_i-U_{i-1}}{h}\Big)\right\}\\ &= \frac{1}{h^2}\Big(\kappa_{i-1/2}U_{i-1} - (\kappa_{i-1/2}+\kappa_{i+1/2})U_i + \kappa_{i+1/2}U_{i+1}\Big).\end{aligned}$$

This leads to a symmetric matrix,

$$A = \frac{1}{h^2}\begin{bmatrix} -(\kappa_{1/2}+\kappa_{3/2}) & \kappa_{3/2} & & & \\ \kappa_{3/2} & -(\kappa_{3/2}+\kappa_{5/2}) & \kappa_{5/2} & & \\ & \ddots & \ddots & \ddots & \\ & & \kappa_{m-3/2} & -(\kappa_{m-3/2}+\kappa_{m-1/2}) & \kappa_{m-1/2} \\ & & & \kappa_{m-1/2} & -(\kappa_{m-1/2}+\kappa_{m+1/2}) \end{bmatrix}.$$

Some good properties

1. The matrix yielded by above method has the advantage of being symmetric, as we would hope since the original differential equation is self-adjoint.
2. Moreover since $\kappa > 0$, the matrix can be shown to be nonsigular and negative definite.
3. When solving the resulting linear system by iterative methods it is also often desirable that the matrix have properties such as negative definiteness, since some iterative methods (e.g., the conjugate-gradient method) depend on such properties.

Nonlinear equations: motion of a pendulum

Consider the motion of a pendulum with mass m at the end of a rigid (massless) bar of length L, and let $\theta(t)$ be the angle of the pendulum from vertical at time t. Ignoring the mass of the bar and forces of friction and air resistance, the pendulum motion can be modeled as

$$\theta''(t) = -\frac{g}{L}\sin(\theta(t)),$$

where g is the gravitational constant.

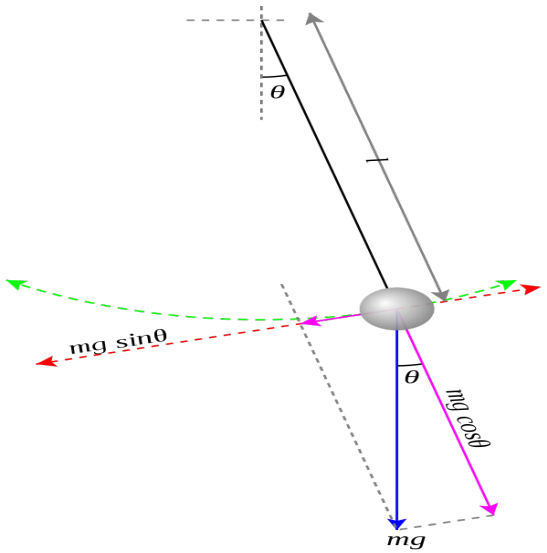
$$F = ma = -mg\sin\theta \implies a = -g\sin\theta$$

$$\text{arc length } s = L\theta \implies v = \frac{ds}{dt} = L\frac{d\theta}{dt}$$

$$\text{and } a = \frac{d^2s}{dt^2} = L\frac{d^2\theta}{dt^2}$$

$$\text{Therefore, } L\frac{d^2\theta}{dt^2} = -g\sin\theta$$

$$\implies \frac{d^2\theta}{dt^2} = -\frac{g}{L}\sin\theta$$



Nonlinear equations

Taking $g/L = 1$ for simplicity, we have

$$\begin{aligned} &\theta''(t) = -\sin(\theta(t)) \quad \text{for } 0 < t < T, \\ &\theta(0) = \alpha \quad \text{and} \quad \theta(T) = \beta \qquad (\text{given } \theta'(0) \text{ is more natural } \Rightarrow \text{ IVP}). \end{aligned}$$

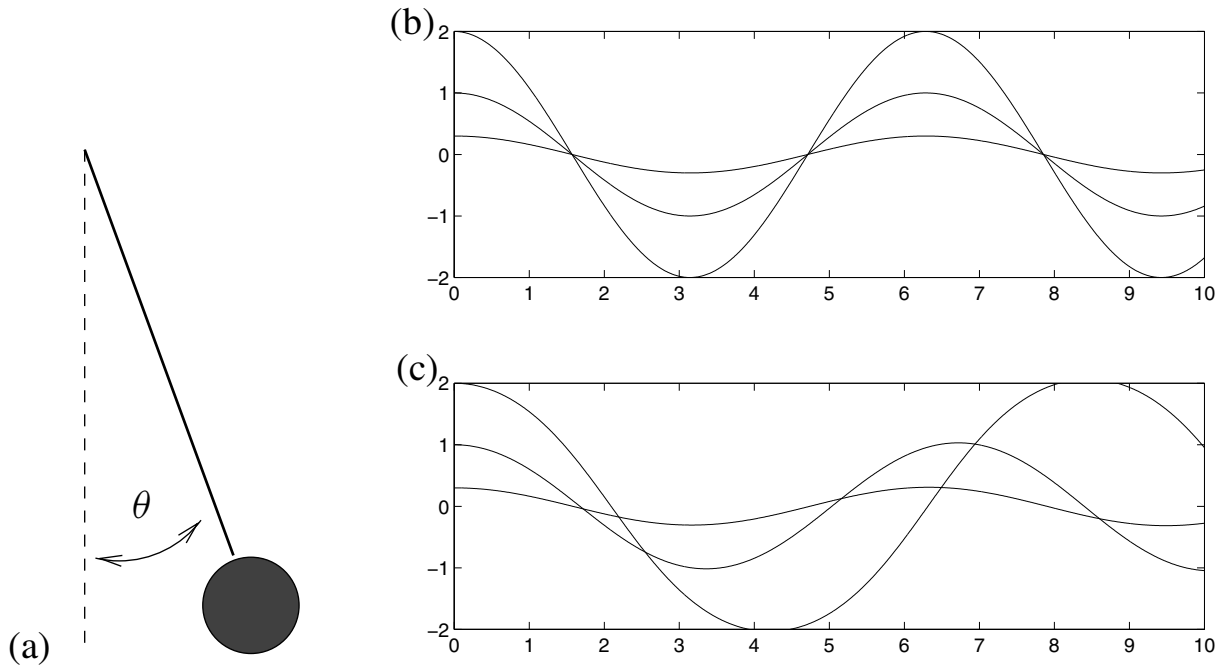
For small amplitudes of the angle θ, we have $\sin(\theta(t)) \approx \theta(t)$ and

$$\begin{aligned} &\theta''(t) = -\theta(t) \quad \text{for } 0 < t < T, \\ &\theta(0) = \alpha \quad \text{and} \quad \theta(T) = \beta. \end{aligned}$$

The general solutions are of the form:

$$\theta(t) = A\cos(t) + B\sin(t). \qquad (\text{has period } 2\pi)$$

Solutions for various $\theta(0)$ and $\theta'(0) = 0$ (IVPs)



(a) pendulum; (b) solutions to the linear equation and (c) solutions to the nonlinear equation for various initial θ and zero initial velocity.

Discretization of the nonlinear BVP

1. Following the approach for linear problems, we obtain the system of equations

$$\frac{1}{h^2}\left(\theta_{i-1} - 2\theta_i + \theta_{i+1}\right) + \sin(\theta_i) = 0,$$

for $i = 1, 2, \cdots, m$, where $h := T/(m+1)$ and $\theta_0 := \alpha$, $\theta_{m+1} := \beta$.

2. This is now a nonlinear system of equations of the form

$$G(\theta) = 0,$$

where $G : \mathbb{R}^m \to \mathbb{R}^m$ and $\theta := (\theta_1, \theta_2, \cdots, \theta_m)^\top$.

3. This cannot be solved as easily as the tridiagonal linear systems. Instead of a direct method, we must generally use some iterative method, such as Newton's method.

Newton's method for nonlinear system of equations

If $\theta^{[k]}$ is our approximation to θ in step k, then Newton's method is derived via the Taylor expansion

$$G(\theta^{[k+1]}) = G(\theta^{[k]}) + G'(\theta^{[k]})\left(\theta^{[k+1]} - \theta^{[k]}\right) + \cdots .$$

Setting $G(\theta^{[k+1]}) = 0$ as desired, and dropping the higher order terms, results in

$$0 \approx G(\theta^{[k]}) + G'(\theta^{[k]})\left(\theta^{[k+1]} - \theta^{[k]}\right).$$

This gives the Newton update

$$\theta^{[k+1]} := \theta^{[k]} + \delta^{[k]},$$

where $\delta^{[k]}$ solves the linear system

$$J(\theta^{[k]})\delta^{[k]} = -G(\theta^{[k]}).$$

where $J(\theta) := G'(\theta) \in \mathbb{R}^{m\times m}$ is the Jacobian matrix with elements

$$J_{ij}(\theta) = \frac{\partial}{\partial\theta_j} G_i(\theta) \quad \text{for } i, j = 1, 2, \cdots, m.$$

Jacobian matrix $J(\theta)$

In our case,

$$G_i(\theta) = G_i(\theta_1, \theta_2, \cdots, \theta_m) = \frac{1}{h^2}\left(\theta_{i-1} - 2\theta_i + \theta_{i+1}\right) + \sin(\theta_i),$$

and hence

$$J_{ij}(\theta) = \begin{cases} 1/h^2 & \text{if } j = i-1 \text{ or } j = i+1, \\ -2/h^2 + \cos(\theta_i) & \text{if } j = i, \\ 0 & \text{otherwise,} \end{cases}$$

so that

$$J(\theta) = \frac{1}{h^2}\begin{bmatrix} -2 + h^2\cos(\theta_1) & 1 & & & \\ 1 & -2 + h^2\cos(\theta_2) & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & & 1 & \\ & & & 1 & -2 + h^2\cos(\theta_m) \end{bmatrix}.$$

Some remarks

1. With Newton's method, we need an initial guess that has to be close enough to an exact solution.
2. Newton's method can be shown to converge quadratically if we start with an initial guess that is sufficiently close to an exact solution.
3. The solution of the nonlinear problem found above is an *isolated solution* in the sense that there are no other solutions very nearby (it is also said to be locally unique). It does not follow that this is the unique solution to the nonlinear BVP.

Accuracy on nonlinear equations

1. Keep clear the distinction between the convergence of Newton's method to a solution of the finite difference equations and the convergence of this finite difference approximation to the solution of the differential equation.
2. **Local truncation error:** inserting the true solution into the finite difference equations:

$$\begin{aligned}
\tau_i &:= \frac{1}{h^2}\Big(\theta(t_{i-1}) - 2\theta(t_i) + \theta(t_{i+1})\Big) + \sin(\theta(t_i)) - 0 \\
&= (\theta''(t_i) + \sin(\theta(t_i))) + \frac{1}{12}h^2\theta^{(4)}(t_i) + O(h^4) \\
&= \frac{1}{12}h^2\theta^{(4)}(t_i) + O(h^4), \quad i = 1, 2, \cdots, m.
\end{aligned}$$

   Hence, the LTE is $O(h^2)$.

3. **Global error**: Let $\hat{\theta}$ be the vector of true values at the grid points. Let $\tau = (\tau_1, \tau_2, \cdots, \tau_m)^\top$. Then $G(\hat{\theta}) = \tau$. Let $E := \theta - \hat{\theta}$ be the global error, then we have

$$G(\theta) - G(\hat{\theta}) = 0 - \tau = -\tau$$

Global error

1. Recall the linear case (i.e., $G(\theta) = A\theta - F$):

$$G(\theta) - G(\hat{\theta}) = A\theta - A\hat{\theta} = AE = -\tau \implies \cdots \implies E^h = (A^h)^{-1}(-\tau^h).$$

2. Use Taylor series expansions to write

$$G(\theta) = G(\hat{\theta}) + J(\hat{\theta})E + O(\|E\|^2) \implies J(\hat{\theta})E = -\tau + O(\|E\|^2).$$

   If we ignore the higher order terms, then we again have a linear relation between the local and global errors.

3. Let $\hat{J}^h := J(\hat{\theta})$ on a grid with grid spacing h.

   Definition: The nonlinear difference method $G(\theta) = 0$ is *stable* in some norm $\|\cdot\|$ if the matrices $(\hat{J}^h)^{-1}$ are uniformly bounded in this norm as $h \to 0$, i.e., there exist $C > 0$ and $h_0 > 0$ such that

$$\|(\hat{J}^h)^{-1}\| \le C \quad \text{for all } 0 < h < h_0.$$

4. It can be shown that if the method is stable in this sense and consistent ($\|\tau^h\| \to 0$ as $h \to 0^+$), then the method converges ($\|E^h\| \to 0$ as $h \to 0^+$). (This is not obvious in the nonlinear case)

Singular perturbation problems

1. Singular perturbation problem $\Longrightarrow$ boundary and/or interior layers $\Longrightarrow$ solution varies rapidly $\Longrightarrow$ difficult to solve numerically.
2. Consider the time-dependent problem which models the temperature $u(x,t)$ of a fluid flowing through a pipe with constant velocity a and the fluid has constant heat diffusion coefficient $\kappa > 0$ and ψ is the source term:

$$u_t + au_x = \kappa u_{xx} + \psi(x) \quad 0 < x < 1 \quad \oplus \quad \text{IC} \quad \oplus \quad \text{BC}.$$

Suppose that $a > 0$. Then we naturally have a boundary condition at the left boundary $x = 0$, specifying the temperature of the incoming fluid:

$$u(0,t) = \alpha(t).$$

Since $\kappa > 0$, the heat can diffuse upstream, we need to specify

$$u(1,t) = \beta(t)$$

to determine a unique solution. If $\kappa = 0$ (no diffusion), we only need BC at $x = 0$ since $a > 0$.

Steady-state convection-diffusion (對流-擴散) problem

If α, β and ψ are all independent of time t, then we expect to having a steady-state solution by solving the following two-point BVP, called the convection/advection-diffusion problem ($\kappa, a > 0$):

$$\begin{cases} -\kappa u''(x) + a u'(x) = \psi(x), \quad 0 < x < 1, \\ u(0) = \alpha \quad \text{and} \quad u(1) = \beta, \end{cases}$$

1. a is small relative to $\kappa \Rightarrow$ smooth solution, the problem is easy to solve.
2. a is large relative to $\kappa \Rightarrow$ convection-dominated (對流佔優) $\Rightarrow$ mainly hyperbolic nature, non-smooth solution.

   Define the Péclet number by $Pe := a/\kappa$, which is the ratio of advection velocity to transport speed due to diffusion.

   To reduce to one-parameter case, let $\varepsilon = \kappa/a > 0$ and rewrite equation in the form

   $$\varepsilon u''(x) - u'(x) = f(x) \left(:= -\frac{1}{a}\psi(x) \right).$$

   Then a large relative to κ (large Péclet number) $\Longleftrightarrow 0 < \varepsilon \ll 1$.

The singularly perturbed problem

As $\varepsilon \to 0^+$, the equation

$$\varepsilon u''(x) - u'(x) = f(x)$$

reduces to a first order equation

$$-u'(x) = f(x),$$

which allows only one boundary condition ($u(0) = \alpha$), rather than two.

However, for $\varepsilon > 0$, no matter how small, we have a second order equation that needs two conditions. Thus, we expect to see strange behavior at the outflow boundary ($x = 1$) as $\varepsilon \to 0^+$.

Indeed, the solution u may exhibit the behavior of boundary layer of width $O(\varepsilon)$ at the outflow boundary, namely, a narrow region where the solution u changes rapidly. In this case, we call

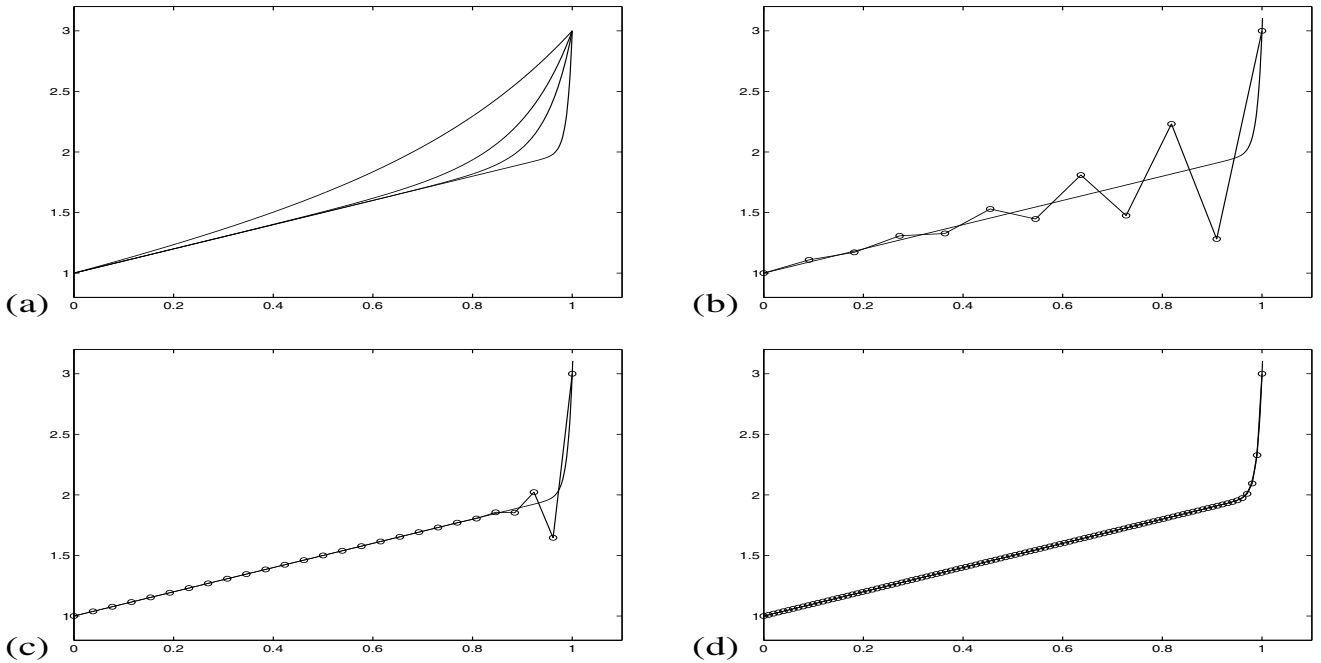
$$\begin{cases} -\kappa u''(x) + a u'(x) = \psi(x), \quad 0 < x < 1, \\ u(0) = \alpha \quad \text{and} \quad u(1) = \beta, \end{cases}$$

with $0 < \varepsilon \ll 1$ a *singularly perturbed problem.*

An example of boundary layer

Let $\alpha = 1$, $\beta = 3$, and $f(x) = -1$. Then the exact solution is given by

$$u(x) = \alpha + x + (\beta - \alpha - 1)\left(\frac{e^{x/\varepsilon} - 1}{e^{1/\varepsilon} - 1}\right).$$



(a) Exact solutions: $\varepsilon = 0.3, 0.1, 0.05$, and 0.01 from top to bottom.
(b) Numerical solution for $\varepsilon = 0.01$ with $h = 1/10$.
(c) $h = 1/25$. (d) $h = 1/100$.

Difficulties in numerical computation

1. Most numerical methods exhibit spurious oscillations or low accuracy for singular perturbation problems (see Figure (b) and (c)).
2. Since the solution changes rapidly over a very small interval in space, derivatives of $u(x)$ are large. For example,

$$u''(x) = D^2u(x) - \frac{1}{12}h^2u''''(x) + O(h^4).$$

   If h is not small enough, then the local truncation error will be very large in the boundary layer. Moreover, even if the truncation error is large only in the boundary layer, the resulting global error $E = -A^{-1}\tau$ may be large everywhere, since A^{-1} is a dense matrix.
3. On finer grids the solution looks better (Figure (c) and (d)), and as $h \to 0$ the method does exhibit second order accurate convergence.

   $\Longrightarrow$ a huge number of linear equations.

Interior layers (ILs)

Consider the nonlinear 2-point boundary value problem

$$\begin{cases} \varepsilon u''(x) + u(x)(u'(x) - 1) = 0 \quad \text{for } a < x < b, \\ u(a) = \alpha \quad \text{and} \quad u(b) = \beta. \end{cases}$$

Setting $\varepsilon = 0$ gives a reduced equation

$$\begin{aligned} & u(x)(u'(x) - 1) = 0 \quad \text{for } a < x < b \\ \Longrightarrow \quad & u(x) = 0 \text{ or } u(x) = x + C \quad \text{for some } C \in \mathbb{R}, \end{aligned}$$

for which we generally can enforce only one boundary condition:

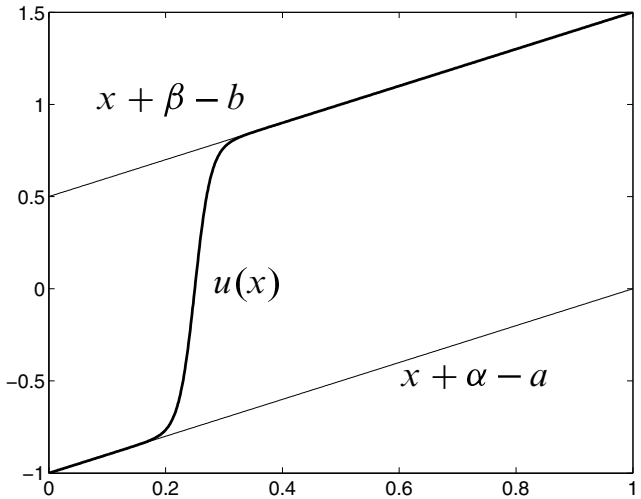
1. $u(x) = x + \alpha - a$ if $u(a) = \alpha$ is imposed.
2. $u(x) = x + \beta - b$ if $u(b) = \beta$ is imposed.

Example of IL

For $0 < \varepsilon \ll 1$, the solution of the full equation

$$\begin{cases} \varepsilon u''(x) + u(x)(u'(x) - 1) = 0 \quad \text{for } a < x < b, \\ u(a) = \alpha \quad \text{and} \quad u(b) = \beta \end{cases}$$

must satisfy both boundary conditions. The figure below shows a solution.



Outer solutions and full solution to the singular perturbation problem with $a = 0$, $b = 1$, $\alpha = -1$, and $\beta = 1.5$. The solution has an interior layer centered about $\bar{x} = 0.25$

How to determine the location and width of the IL? – Perturbation Theory

Assume the interior layer is centered at some location $\bar{x} \in (a, b)$, and we zoom in on the solution by assuming that $u(x)$ has the approximate form

$$u(x) = W((x - \bar{x})/\varepsilon^k)$$

for some power k to be determined. Then we have

$$\begin{aligned} u'(x) &= \varepsilon^{-k} W'((x - \bar{x})/\varepsilon^k), \\ u''(x) &= \varepsilon^{-k} W''((x - \bar{x})/\varepsilon^k). \end{aligned}$$

Substituting these into the equation

$$\varepsilon u''(x) + u(x)(u'(x) - 1) = 0 \quad \text{for } a < x < b$$

gives

$$\varepsilon \cdot \varepsilon^{-2k} W''(\xi) + W(\xi)(\varepsilon^{-k} W'(\xi) - 1) = 0$$

$$\text{multiply by } \varepsilon^{2k-1} \Longrightarrow \; W''(\xi) + W(\xi)(\varepsilon^{k-1} W'(\xi) - \varepsilon^{2k-1}) = 0,$$

where $\xi = (x - \bar{x})/\varepsilon^k$.

Determine the layer width $O(\varepsilon^k)$

By rescaling the independent variable by a factor ε^k, we have converted the singular perturbation problem into a problem where the highest order derivative W'' has coefficient 1 and the small parameter appears only in the lower order term:

$$W''(\xi) + W(\xi)(\varepsilon^{k-1}W'(\xi) - \varepsilon^{2k-1}) = 0, \quad \text{where } \xi := \frac{x - \bar{x}}{\varepsilon^k}.$$

1. For $k < 1$, the lower order term blows up as $\varepsilon \to 0^+$, or dividing by ε^{k-1} shows that we still have a singular perturbation problem.
2. The lower order term behaves well in the limit $\varepsilon \to 0^+$ only if we take $k \geq 1$.

Boundary conditions

Boundary conditions: Fix x at any value away from $\bar{x}$, then

$$\xi := \frac{x - \bar{x}}{\varepsilon^k} \to \pm\infty \quad \text{as } \varepsilon \to 0^+.$$

So we define boundary conditions at $\pm\infty$,

$$W(\xi) \to \bar{x} + \alpha - a \quad \text{as } \xi \to -\infty,$$
$$W(\xi) \to \bar{x} + \beta - b \quad \text{as } \xi \to +\infty.$$

We also require

$$W'(\xi) = \varepsilon^k u'(x) \to 0 \quad \text{as } \xi \to \pm\infty,$$

since outside the layer the linear functions have the desired slope.

Observe that if $k > 1$, the lower order term vanishes as $\varepsilon \to 0^+$ and the equation reduces to $W''(\xi) = 0$. This implies the solution simply appears linear, while it does not allow us to capture the full behavior in the interior layer.

Approximate solution

Taking $k = 1$ gives the proper interior problem

$$W''(\xi) + W(\xi)(W'(\xi) - \varepsilon) = 0.$$

Now letting $\varepsilon \to 0$, we have

$$W''(\xi) + W(\xi)W'(\xi) = 0,$$

which has the solutions

$$W(\xi) = w_0 \tanh(w_0\xi/2), \qquad \left(\tanh(x) = \frac{\sinh(x)}{\cosh(x)} = \frac{\frac{e^x - e^{-x}}{2}}{\frac{e^x + e^{-x}}{2}} = \frac{1 - e^{-2x}}{1 + e^{-2x}} \right)$$

for arbitrary constants w_0. The boundary conditions lead to

$$w_0 = \frac{1}{2}(a - b + \beta - \alpha) \quad \text{and} \quad \bar{x} = \frac{1}{2}(a + b - \alpha - \beta).$$

Combining the inner and outer solutions, we obtain an approximate solution

$$u(x) \approx \tilde{u}(x) := x - \bar{x} + w_0 \tanh(w_0(x - \bar{x})/2\varepsilon).$$

Finite difference approximation

In summary, the solution has an interior layer of width $O(\varepsilon)$ at $x = \bar{x}$ with roughly linear solution outside the layer. This type of information may be all we need to know about the solution for some applications.

This nonlinear problem

$$\begin{cases} \varepsilon u''(x) + u(x)(u'(x) - 1) = 0 \quad \text{for } a < x < b, \\ u(a) = \alpha \quad \text{and} \quad u(b) = \beta \end{cases}$$

can be solved numerically on a uniform grid using the finite difference equations

$$G_i(U) := \varepsilon \left(\frac{U_{i-1} - 2U_i + U_{i+1}}{h^2} \right) + U_i \left(\frac{U_{i+1} - U_{i-1}}{2h} - 1 \right) = 0$$

for $i = 1, 2, \cdots, m$ with $U_0 = \alpha$ and $U_{m+1} = \beta$. This gives a nonlinear system of equations $G(U) = 0$ that can be solved using Newton's method. The initial guess for Newton's method can be chosen as $U_i = \tilde{u}(x_i)$, $1 \le i \le m$, which is already very accurate at nearly all grid points.

Nonuniform grids

On the other hand, when ε is very small, highly accurate numerical results can be obtained with less computation by using a nonuniform grid, with grid points clustered in the layer (see figure below). The width of the layer is $O(\varepsilon)$ and, moreover, from

$$u(x) \approx \tilde{u}(x) := x - \bar{x} + w_0 \tanh(w_0(x - \bar{x})/2\varepsilon),$$

we expect that most of the transition occurs for, say, $|\frac{1}{2}w_0\xi| < 2$. This translates into $|x - \bar{x}| < 4\varepsilon/w_0$.