MA2501 Numerical Methods

Suggested solutions to exam problems

2nd of June 2006

Problem 1

We are given the function

$$f(x) = \frac{e^{-x}}{1+x}$$

for all $x \geq 0$.

a) We wish to compute the minimum degree polynomial $p(x)$ which interpolates $f(x)$ at the nodes $x_0 = 0$, $x_1 = 2$, $x_2 = 6$, and $x_3 = 8$.

We will use the Newton form of the interpolating polynomial as this is more amenable to hand calculation. We recall briefly that the Newton form is generally given by

$$p(x) = \sum_{i=0}^{n} f[x_0, \ldots, x_i] \prod_{j=0}^{i-1} (x - x_j) \tag{1}$$

in which n is the degree of the resulting polynomial—one less than the number of nodes. In this case, $n = 3$. Moreover, the *divided differences* $f[x_0, \ldots, x_i]$ satisfy the relation

$$f[x_j, \ldots, x_k] = \frac{f[x_{j+1}, \ldots, x_k] - f[x_j, \ldots, x_{k-1}]}{x_k - x_j} \tag{2}$$

for all $0 \leq j < k \leq n$ when we *define* $f[x_j] = f(x_j)$.

The relation (2) gives Table 1 of divided differences, whence the interpolating polynomial

$$\begin{aligned} p(x) &= 1.00000 - 4.77444 \cdot 10^{-1}\, x + 7.77091 \cdot 10^{-2}\, x(x-2) \\ &\quad - 9.48383 \cdot 10^{-3}\, x(x-2)(x-6) \\ &= -9.48383 \cdot 10^{-2}\, x^3 + 0.15358\, x^2 - 0.74667\, x + 1. \end{aligned}$$

0	1.00000			
		$-4.77444 \cdot 10^{-1}$		
2	$4.51118 \cdot 10^{-2}$		$7.77091 \cdot 10^{-2}$	
		$-1.11894 \cdot 10^{-2}$		$-9.48383 \cdot 10^{-3}$
6	$3.54107 \cdot 10^{-4}$		$1.83850 \cdot 10^{-3}$	
		$-1.58417 \cdot 10^{-4}$		
8	$3.72736 \cdot 10^{-5}$			

Table 1: Table of divided differences $f[x_j, \ldots, x_k]$ of Problem 1.

Moreover, $p(1/2) = 0.66388$ which means that

$$f(1/2) - p(1/2) = -0.25952.$$

b) To establish a guaranteed upper bound on the error $|f(x) - p(x)|$, we proceed from a result presented in the lectures. The function $f(x)$ is at least 4 times continuously differentiable, meaning that for any $x \in [0, 8]$ there is a point $\xi_x \in (0, 8)$ for which

$$f(x) - p(x) = \frac{1}{24} f^{(4)}(\xi_x) \cdot x\,(x-2)(x-6)(x-8). \tag{3}$$

Let $w_4(x) = x(x-2)(x-6)(x-8) = x^4 - 16x^3 + 76x^2 - 96x$. Equation (3) means that

$$|f(x) - p(x)| \leq \frac{1}{24} \max_{0 \leq p \leq 8} |f^{(4)}(p)| \cdot \max_{0 \leq p \leq 8} |w_4(p)|. \tag{4}$$

Differentiating gives

$$w_4'(x) = 4x^3 - 48x^2 + 152x - 96 = 4 \cdot (x^3 - 12x^2 + 38x - 24)$$

and we observe that $w_4'(4) = 0$. Polynomial division then gives

$$w_4'(x) = 4 \cdot (x-4) \cdot (x^2 - 8x + 6)$$

from which the extremal points of $w_4(x)$ are

$$(4 - \sqrt{10}, -36), \quad (4, 64), \quad (4 + \sqrt{10}, -36).$$

In other words, $\max_{0 \leq p \leq 8} |w_4(p)| = 64$.

Let $f_1(x) = 1/(1+x)$ and $f_2(x) = \mathrm{e}^{-x}$. Then $f_1^{(n)}(x) = (-1)^n n!/(1+x)^{n+1}$ and $f_2^{(n)}(x) = (-1)^n \mathrm{e}^{-x}$. From the given formula for higher derivatives of products we then get

$$\begin{aligned} f^{(4)}(x) &= \frac{(-1)^0 0!}{1+x}\,(-1)^4\mathrm{e}^{-x} + 4\frac{(-1)^1 1!}{(1+x)^2}\,(-1)^3\mathrm{e}^{-x} \\ &\quad + 6\frac{(-1)^2 2!}{(1+x)^3}\,(-1)^2\mathrm{e}^{-x} + 4\frac{(-1)^3 3!}{(1+x)^4}\,(-1)^1\mathrm{e}^{-x} \\ &\quad + \frac{(-1)^4 4!}{(1+x)^5}\,(-1)^0\mathrm{e}^{-x} \\ &= \frac{\mathrm{e}^{-x}}{(1+x)^5}\left(24 + 24(1+x) + 12(1+x)^2 + 4(1+x)^3 + (1+x)^4\right) \end{aligned}$$

and similarly

$$\begin{aligned} f^{(5)}(x) = -\frac{\mathrm{e}^{-x}}{(1+x)^6}\big(&120 + 120(1+x) + 60(1+x)^2 + \\ &20(1+x)^3 + 5(1+x)^4 + (1+x)^5\big). \end{aligned}$$

We notice that $f^{(5)}(x) < 0$ for all $x \geq 0$ and, consequently, that the maximum value of $|f^{(4)}(x)|$ must be attained at either $x = 0$ or at $x = 8$. Moreover, $f^{(4)}(x)$ decays rapidly for increasing values of x yet remains always positive. Thus, the maximum value of $|f^{(4)}(x)|$ is attained at $x = 0$. In summary:

$$\max_{0\leq p\leq 8}|f^{(4)}(p)| = |f^{(4)}(0)| = f^{(4)}(0) = 65.$$

Inserting this and the maximum value of $|w_4(x)|$ on $[0, 8]$ into the error estimate (4) finally yields

$$|f(x) - p(x)| \leq \frac{65\cdot 64}{24} = \frac{520}{3} \approx 173.333.$$

This bound, however, is much too unrefined and inaccurate to be of any practical use. We know that $f(x) \in (0, 1]$ for all $x \geq 0$ and having an error bound that is several orders of magnitude worse than the largest value of the function means we cannot actually control the error. In fact, $\max_{0\leq x\leq 8}|f(x) - p(x)| \approx 0.262$, attained at $x \approx 0.58$.

The main reason for this "bounding failure" is that while the largest value of $|f^{(4)}(x)|$ is certainly big, this largest value does not actually

represent the true nature of $f^{(4)}(x)$ throughout the interval of interest. To establish sharp error bounds, the result (4) implictly assumes that $f^{(4)}(x)$ does not vary too much on $[0, 8]$. This assumption is violated in the present case.

Problem 2

We are given the function

$$f(x) = \frac{\mathrm{e}^{-x}}{1+x}$$

for all $x \geq 0$.

a) We wish to compute the Simpson approximation to $\int_0^8 f(x)\,\mathrm{d}x$ using 8 sub-intervals or, equivalently, a step size of $h = (8-0)/8 = 1$. We get

$$\begin{aligned} S_8(f) &= \tfrac{1}{3}\big(f(0) + 4\big(f(1) + f(3) + f(5) + f(7)\big) \\ &\qquad + 2\big(f(2) + f(4) + f(6)\big) + f(8)\big) \\ &\approx 0.62960. \end{aligned}$$

b) The error committed in computing $\int_0^\infty f(x)\,\mathrm{d}x$ by means of a Simpson method approximation of $\int_0^B f(x)\,\mathrm{d}x$ for some finite $B > 0$ can be divided into two components

- Numerical error in Simpson's method on $\int_0^B f(x)\,\mathrm{d}x$.
- Methodological error (or *truncation error*) incurred by computing $\int_0^B f(x)\,\mathrm{d}x$ rather than $\int_0^\infty f(x)\,\mathrm{d}x$.

Let $S_h(f; 0, B)$ denote the step size h Simpson method approximation to $\int_0^B f(x)\,\mathrm{d}x$. We know that

$$\int_0^B f(x)\,\mathrm{d}x - S_h(f; 0, B) = -\tfrac{1}{180}\,Bh^4 f^{(4)}(\xi)$$

for some $\xi \in (0, B)$. Thus

$$\left|\int_0^B f(x)\,\mathrm{d}x - S_h(f; 0, B)\right| \leq \tfrac{1}{180}\,Bh^4 \max_{0\leq p\leq B}|f^{(4)}(p)| = \tfrac{13}{36}\,Bh^4,$$

the latter equality due to $\max_{0\leq p\leq B}|f^{(4)}(p)| = 65$ for all $B > 0$ as shown in Problem 1**b**).

As $\int_0^\infty f(x)\,\mathrm{d}x = \int_0^B f(x)\,\mathrm{d}x + \int_B^\infty f(x)\,\mathrm{d}x$, the methodological error is given by

$$\Big|\int_B^\infty f(x)\,\mathrm{d}x\Big| = \int_B^\infty \frac{\mathrm{e}^{-x}}{1+x}\,\mathrm{d}x \le \frac{1}{1+B}\int_B^\infty \mathrm{e}^{-x}\,\mathrm{d}x \le \int_B^\infty \mathrm{e}^{-x}\,\mathrm{d}x = \mathrm{e}^{-B}.$$

In summary, we find

$$\Big|\int_0^\infty f(x)\,\mathrm{d}x - S_h(f;0,B)\Big| \le \tfrac{13}{36}\,Bh^4 + \mathrm{e}^{-B}$$

as we wanted to prove.

c) We wish to determine the *least* number of sub-intervals n such that the total error incurred in the above method is less than $\varepsilon = \frac{1}{2}\cdot 10^{-3}$. As $h = B/n$, this means finding the least value of n guaranteeing that

$$\frac{13}{36}\frac{B^5}{n^4} + \mathrm{e}^{-B} < \varepsilon$$

which leads to

$$n^4 > \frac{13}{36}\frac{B^5}{\varepsilon - \mathrm{e}^{-B}} = \frac{13}{36}\,g(B) \tag{5}$$

when we define $g(B) = B^5/(\varepsilon - \mathrm{e}^{-B})$. We need in particular $\mathrm{e}^{-B} < \varepsilon$ or $B > -\ln\varepsilon$ lest the methodological error itself be too large. As we want the least possible value of n we thus need to find the minimum value of $g(B)$ when $B > -\ln\varepsilon$. This, then, means that $g(B) > 0$ for all B in the valid domain and as $\lim_{B\downarrow -\ln\varepsilon} g(B) = \lim_{B\to\infty} g(B) = \infty$, the minimum value of $g(B)$ is attained at a point for which $g'(B) = 0$.

Differentiating gives

$$g'(B) = \frac{5B^4(\varepsilon - \mathrm{e}^{-B}) - B^5\mathrm{e}^{-B}}{(\varepsilon - \mathrm{e}^{-B})^2} = -B^4\,\frac{(B+5)\,\mathrm{e}^{-B} - 5\varepsilon}{(\varepsilon - \mathrm{e}^{-B})^2},$$

a zero for which is attained when

$$F(B) = (B+5)\,\mathrm{e}^{-B} - 5\varepsilon = 0 \tag{6}$$

subject to the extra condition that $B > -\ln\varepsilon$.

k	B_k
0	8.600902
1	8.601656
2	8.601656

Table 2: Newton iterates for the minimum point of $g(B)$ in Problem 2.

Formulating Newton's method for the non-linear equation (6) yields the iteration

$$B_{k+1} = B_k - \frac{(B_k+5)\,\mathrm{e}^{-B_k} - 5\varepsilon}{-(B_k+4)\,\mathrm{e}^{-B_k}} = B_k + \frac{(B_k+5)\,\mathrm{e}^{-B_k} - 5\varepsilon}{(B_k+4)\,\mathrm{e}^{-B_k}} \tag{7}$$

for all $k \ge 0$. Using initial value $B_0 = -\ln(\varepsilon) + 1 \approx 8.600902$ yields the Newton iterates of Table 2. In other words $B_{\mathrm{opt}} = 8.601656$ is the best value of the upper limit of the integral when minimising the number of sub-intervals of the Simpson method. Inserting this value into the lower bound (5) yields

$$n > \Big(\frac{13\,B_{\mathrm{opt}}^5}{36\,(\varepsilon - \mathrm{e}^{-B_{\mathrm{opt}}})}\Big)^{1/4} \approx 85.634$$

or, as the number of sub-intervals in Simpson's method must be an even integer, $n = 86$.

Problem 3

We are given the function

$$R(z) = \frac{1+z/2}{1-z/2}$$

for all $-2 < z < 2$.

a) We wish to show that $\mathrm{e}^z - R(z) = -\frac{z^3}{12} + \mathcal{O}(z^4)$. Using either knowledge of the geometric series or explicit Taylor series expansion, we find that

$$\frac{1}{1-x} = \sum_{k=0}^\infty x^k = 1 + x + x^2 + x^3 + \cdots$$

whenever $-1 < x < 1$. Consequently

$$\begin{aligned} R(z) &= (1+\frac{z}{2})\cdot(1+\frac{z}{2}+\frac{z^2}{4}+\frac{z^3}{8}+\frac{z^4}{16}+\cdots) \\ &= 1+\frac{z}{2}+\frac{z^2}{4}+\frac{z^3}{8}+\frac{z^4}{16}+\cdots+\frac{z}{2}+\frac{z^2}{4}+\frac{z^3}{8}+\frac{z^4}{16}+\cdots \\ &= 1+z+\frac{z^2}{2}+\frac{z^3}{4}+\frac{z^4}{8}+\cdots \end{aligned}$$

for all $-2 < z < 2$. We know in addition that

$$e^z = \sum_{k=0}^{\infty} \frac{z^k}{k!} = 1+z+\frac{z^2}{2}+\frac{z^3}{6}+\frac{z^4}{24}+\cdots$$

so

$$e^z - R(z) = (\frac{1}{6}-\frac{1}{4})\,z^3 + (\frac{1}{24}-\frac{1}{8})\,z^4+\cdots = -\frac{z^3}{12}+\mathcal{O}(z^4)$$

as we wanted to prove.

b) We know that LU factorisation amounts to Gaussian elimination and additionally storing the *multipliers* in the lower triangular matrix L. From the initial matrix

$$A = \begin{bmatrix} 1.10 & -0.05 & 0.00 \\ -0.05 & 1.10 & -0.05 \\ 0.00 & -0.05 & 1.10 \end{bmatrix}$$

we calculate the multipliers

$$\ell_{21} = \tfrac{a_{21}}{a_{11}} = -\tfrac{0.05}{1.10} \approx -0.0455, \quad \ell_{31} = \tfrac{a_{31}}{a_{11}} = 0.0000$$

and obtain the reduced matrix

$$\tilde{A} = \begin{bmatrix} 1.1000 & -0.0500 & 0.0000 \\ 0.0000 & 1.0977 & -0.0500 \\ 0.0000 & -0.0500 & 1.1000 \end{bmatrix}.$$

Repeating the elimination step we obtain the multiplier

$$\ell_{32} = \tfrac{\tilde{a}_{32}}{\tilde{a}_{22}} \approx -0.0455$$

and the final reduced matrix

$$\tilde{\tilde{A}} = \begin{bmatrix} 1.1000 & -0.0500 & 0.0000 \\ 0.0000 & 1.0977 & -0.0500 \\ 0.0000 & 0.0000 & 1.0977 \end{bmatrix}.$$

Thus, the unit diagonal lower triangular matrix L and the upper triangular matrix U such that $LU = A$ are, respectively,

$$L = \begin{bmatrix} 1.0000 & 0.0000 & 0.0000 \\ -0.0455 & 1.0000 & 0.0000 \\ 0.0000 & -0.0455 & 1.0000 \end{bmatrix},$$

and

$$U = \begin{bmatrix} 1.1000 & -0.0500 & 0.0000 \\ 0.0000 & 1.0977 & -0.0500 \\ 0.0000 & 0.0000 & 1.0977 \end{bmatrix}.$$

c) Let $P(z) = 1 + z/2$ and $Q(z) = 1 - z/2$. Thus $R(z) = P(z)/Q(z)$ which means that $Q(z)R(z) = P(z)$. Substituting $z = hX$ we find

$$P(hX) = I + \tfrac{hX}{2} = \begin{bmatrix} 0.90 & 0.05 & 0.00 \\ 0.05 & 0.90 & 0.05 \\ 0.00 & 0.05 & 0.90 \end{bmatrix}$$

$$Q(hX) = I - \tfrac{hX}{2} = \begin{bmatrix} 1.10 & -0.05 & 0.00 \\ -0.05 & 1.10 & -0.05 \\ 0.00 & -0.05 & 1.10 \end{bmatrix}.$$

In other words, the matrix $R(hX)$ must satisfy the simultaneous equations

$$\begin{bmatrix} 1.10 & -0.05 & 0.00 \\ -0.05 & 1.10 & -0.05 \\ 0.00 & -0.05 & 1.10 \end{bmatrix} R(hX) = \begin{bmatrix} 0.90 & 0.05 & 0.00 \\ 0.05 & 0.90 & 0.05 \\ 0.00 & 0.05 & 0.90 \end{bmatrix}.$$

In particular, the second column of $R(hX)$, here denoted by the symbol $\mathbf{r}^{(2)}$, must satisfy the linear system

$$\begin{bmatrix} 1.10 & -0.05 & 0.00 \\ -0.05 & 1.10 & -0.05 \\ 0.00 & -0.05 & 1.10 \end{bmatrix} \mathbf{r}^{(2)} = \begin{bmatrix} 0.05 \\ 0.90 \\ 0.05 \end{bmatrix},$$

the coefficient matrix of which is the matrix A of Problem **b**).

Using the LU decomposition of Problem **b**) and *defining* $\mathbf{y} = U\mathbf{r}^{(2)}$, we first solve the linear system

$$\begin{bmatrix} 1.0000 & 0.0000 & 0.0000 \\ -0.0455 & 1.0000 & 0.0000 \\ 0.0000 & -0.0455 & 1.0000 \end{bmatrix} \mathbf{y} = \begin{bmatrix} 0.05 \\ 0.90 \\ 0.05 \end{bmatrix},$$

to get

$$y_1 = 0.05,$$
$$y_2 = 0.90 - (-0.0455) \cdot 0.05 \approx 0.9023,$$
$$y_3 = 0.05 - (-0.0455) \cdot 0.9023 \approx 0.0911.$$

Then, to compute the final result $\mathbf{r}^{(2)}$, we must solve the linear system

$$\begin{bmatrix} 1.1000 & -0.0500 & 0.0000 \\ 0.0000 & 1.0977 & -0.0500 \\ 0.0000 & 0.0000 & 1.0977 \end{bmatrix} \mathbf{r}^{(2)} = \begin{bmatrix} 0.0500 \\ 0.9023 \\ 0.0911 \end{bmatrix}$$

to obtain

$$(\mathbf{r}^{(2)})_3 = \frac{0.0911}{1.0977} \approx 0.0830$$
$$(\mathbf{r}^{(2)})_2 = \frac{1}{1.0977}(0.9023 - (-0.05) \cdot 0.0830) \approx 0.8257,$$
$$(\mathbf{r}^{(2)})_1 = \frac{1}{1.1000}(0.0500 - (-0.05) \cdot 0.8257) \approx 0.0830$$

or in vector form, $\mathbf{r}^{(2)} = [0.0830, 0.8257, 0.0830]^\mathsf{T}$.

Problem 4

We are given the (non-linear) partial differential equation with initial and boundary conditions

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \frac{1}{1+u^2}, \quad (x,t) \in [0,1] \times [0,1]$$
$$u(0,t) = 0, \quad u(1,t) = 0$$
$$u(x,0) = x \cdot (x-1), \quad 0 \le x \le 1.$$

a) We have seen in the lectures that an arbitrary, sufficiently differentiable function $v(x,y)$ satisfies the relation

$$\frac{v(x+h,y) - 2v(x,y) + v(x-h,y)}{h^2} = \frac{\partial^2 v}{\partial x^2}(x,y) + \mathcal{O}(h^2).$$

Thus at the point (x_i, t) we find

$$\frac{\partial^2 u}{\partial x^2}(x_i,t) = \frac{u(x_{i+1},t) - 2u(x_i,t) + u(x_{i-1},t)}{h^2} + \mathcal{O}(h^2)$$
$$= \frac{U_{i+1}(t) - 2U_i(t) + U_{i-1}(t)}{h^2} + \mathcal{O}(h^2)$$
$$\approx \frac{U_{i+1}(t) - 2U_i(t) + U_{i-1}(t)}{h^2}$$

Moreover,

$$\frac{\partial u}{\partial t}(x_i,t) = U_i'(t), \quad \text{and} \quad \frac{1}{1+u(x_i,t)^2} = \frac{1}{1+\big(U_i(t)\big)^2}.$$

Inserting these relations into the partial differential equation and using $h = 1/N$ we find

$$U_i'(t) = \frac{1}{h^2}\big(U_{i+1}(t) - 2U_i(t) + U_{i-1}(t)\big) + \frac{1}{1+\big(U_i(t)\big)^2}$$
$$= N^2\big(U_{i-1}(t) - 2U_i(t) + U_{i+1}(t)\big) + \frac{1}{1+\big(U_i(t)\big)^2}$$

which must hold for all $i = 1, \dots, N-1$. Additionally, $U_0(t) = U_N(t) \equiv 0$ for all $t \in [0,1]$ due to the boundary conditions and $U_i(0) = x_i \cdot (1 - x_i)$ for all $i = 1, \dots, N-1$ due to the PDE initial condition.

Using this knowledge we arrive at the final system of $N-1$ ordinary differential equations given by

$$U_1' = N^2 \cdot (-2U_1 + U_2) + \frac{1}{1+U_1^2}$$
$$U_i' = N^2 \cdot (U_{i-1} - 2U_i + U_{i+1}) + \frac{1}{1+U_i^2}, \quad i = 2, \dots, N-2$$
$$U_{N-1}' = N^2 \cdot (U_{N-2} - 2U_{N-1}) + \frac{1}{1+U_{N-1}^2}.$$

b) There are several ways, mostly differing in computational efficiency, of implementing the MATLAB function `ode_rhs`. The only two rules that *must* be obeyed are that

- the function signature be

  ```
  function dy = ode_rhs(t, y)
  ```

  in which `t` and `y` are the current values of the independent variable t and the dependent variable $\mathbf{y}$, respectively
- the return value `dy` be a *column* vector of the same size as `y`

We can implement the function using a MATLAB `for`-loop as follows

```
function dy = ode_rhs(t, y)
N = numel(y) + 1;       % system dimension: N - 1
N2 = N * N;             % N^2, convenience
dy = zeros(size(y));

dy(1) = N2 * (-2*y(1) + y(2)) + 1./(1 + y(1).^2);
for i = 2 : N - 2,
   dy(i) = N2 * (y(i-1) - 2*y(i) + y(i+1)) + ...
                1./(1 + y(i).^2);
end
dy(N-1) = N2 * (y(N-2) - 2*y(N-1)) + ...
                1./(1 + y(N-1).^2);
```

Another possibility is to use MATLAB's powerful indexing and array operations as follows

```
function dy = ode_rhs(t, y)
N = numel(y) + 1;
N2 = N * N;

dy = N2 .* [         - 2.*y(1)     + y(2);      ...
            y(1:N-3) - 2.*y(2:N-2) + y(3:N-1); ...
            y(N-2)   - 2.*y(N-1)]           + ...
     1 ./ (1 + y.^2);
```

However the function is implemented, though, the final PDE resolution process is effectuated through the statements

```
>> N = 200;
>> x = linspace(0, 1, N + 1); % N intervals: N+1 points
>> y0 = x .* (1 - x);         % initial condition
>> y0 = y0(2 : end-1);        % ‘internal’ points
>> [t, y] = ode15s('ode_rhs', [0, 1], y0);
```

The final result is shown in Figure 1.

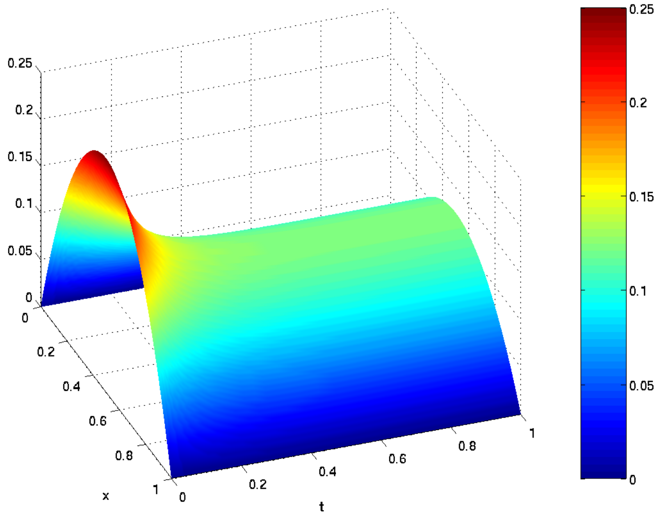

Figure 1: Solution to non-linear PDE of Problem 4.