*A*dvanced exercises & *S*elf-study Week 1

Numerical Modelling

2020

1 Floating point arithmetic

Question *S*. The following sum (*'The Basel problem'*) was found by Euler:

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} + \cdots = \frac{\pi^2}{6} \approx 1.6449.$$

We test this assertion in MATLAB with a for-loop:

```
bp=0
for i = 1:n
 bp = bp + 1/(single(i)^2);
end
format long; bp % Output of the result with as many digits as possible
```

The function `single` computes the result with single precision floating points (Real4). Of course, we can't sum to infinity, but we can choose `n` to be 'large'. For `n` equal to 10 we find 1.5497677. For `n` equal to 1000 we find 1.6439348, so we're improving. But for `n` equal to 1e4, 1e5, 1e6, etc., we obtain *exactly* 1.6447253. Why doesn't the result improve? *Hint: single precision floating point has how many expected digits of accuracy? What do you think is the result of* `single(1)+single(1e-8)` *in MATLAB?*

Question *S*. There's a simple fix for the question above: change the order of the for-loop.

```
for i = n:-1:1
 bp = bp + 1/(single(i)^2);
end
```

Now, the result improves for increasing `n`. Why? *Hint: roughly 6 digits of accuracy doesn't mean that it can't store very small numbers; the number* `single(1e-39)` *is stored just fine! The problem lies with storing numbers that are 6 orders of magnitude apart.*

Question *S*. What do the above results tell you about associativity when summing numbers with the computer (the rule that $1 + (2 + 3) = (1 + 2) + 3$)? *Hint: what is* $(1 + 10^{-100}) - 1$*? Will your computer agree?.*

2 Numerical integration

Question *S*. Consider the following definition of the rectangle rule:

$$\int_a^b f(x)\,\mathrm{d}x \approx \sum_{i=0}^{N-1} f(a + i\Delta x)\Delta x,$$

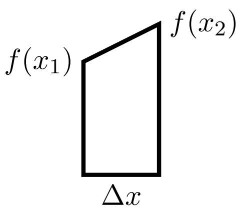
where $\Delta x = \frac{b-a}{N}$. Take the function $f(x) = 2$, and $a = 0$, $b = 4$. Choose N to your own liking. Does the sum return the correct solution? *Hint: draw the rectangles, including the start and end-points, and feel free to use a coarse $\Delta x = 1$.*

Question *S*. Does the rectangle rule also give correct results for linear functions? *Hint: draw the rectangles including the start and end-points for $f(x) = x$ with $a = 0$ and $b = 4$.*

Question *A*. The trapezoidal rule uses the area of a trapezoid rather than a rectangle to integrate. Proof that the area of these trapezoids is given by

$$\frac{f(x_1) + f(x_2)}{2}\Delta x.$$

Hint: draw a trapezoid, and see that it can always be subdivided into two triangles. How do we compute the area of these triangles?



Question *S*. Simplify the following sum,

$$\frac{f(a) + f(a+\Delta x)}{2} + \frac{f(a+\Delta x) + f(a+2\Delta x)}{2} + \cdots + \frac{f(a+(N-1)\Delta x) + f(b)}{2}.$$

Question *S*. The trapezoidal rule can be defined as follows,

$$\int_a^b f(x)\,\mathrm{d}x \approx \left(\frac{f(a) + f(b)}{2} + \sum_{i=1}^{N-1} f(a + i\Delta x)\right)\Delta x,$$

where $\Delta x = \frac{b-a}{N}$. Does the method work for linear functions? *Hint: again, draw the trapezoids for a simple function such as $f(x) = x$, using a coarse Δx. Pay particular attention to the start and end-points.*

3 Gaussian quadrature

Question *A*. 1. Finding clever weights. Say, we want to approximate the following integration by a summation:

$$\int_{-1}^{1} f(x)\,\mathrm{d}x \approx w_1 f(x_1) + w_2 f(x_2) + w_3 f(x_3) + w_4 f(x_4).$$

Say we choose $x_{1,2,3,4}$ arbitrarily on the domain $[-1,1]$, then how do we cleverly choose our weights $w_{1,2,3,4}$? Without knowing anything about $f(x)$, we may say that we would at least demand that if $f(x) = 1$ (with $\int_{-1}^{1} \mathrm{d}x = 2$), we would want our weights to accurately compute that. Plugging that into the equation above, we find that we want $w_1+w_2+w_3+w_4 = 2$. We can also demand that $f(x) = x$ (with $\int_{-1}^{1} x\,\mathrm{d}x = 0$) is accurately integrated. Plugging that into the equation above, that gives $w_1x_1+w_2x_2+w_3x_3+w_4x_4 = 0$. Confirm that we can repeat this procedure also for x^2 and x^3 to find

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ x_1 & x_2 & x_3 & x_4 \\ x_1^2 & x_2^2 & x_3^2 & x_4^2 \\ x_1^3 & x_2^3 & x_3^3 & x_4^3 \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \end{pmatrix} = \begin{pmatrix} 2 \\ 0 \\ \frac{2}{3} \\ 0 \end{pmatrix}. \tag{1}$$

We can solve this system with, e.g., MATLAB. We can then integrate polynomials up to degree 3 using 4 arbitrarily chosen points. This is of higher formal accuracy than the rectangular and trapezoidal rule! If you choose N points, you can accurately integrate polynomials up to degree $N-1$.

Question *A*. Confirm that we can approximate an integral using two points $x_1 = -1$ and $x_2 = 1$, as

$$\int_{-1}^{1} f(x)\,\mathrm{d}x \approx f(-1) + f(1), \tag{2}$$

which is accurate for polynomials up to degree 1. *Hint: use a system like in eq.* (1).

Question *A*. 2. Legendre polynomials. Legendre polynomials of order n (called P_n) have a special property regarding integration from -1 to 1:

$$\int_{-1}^{1} P_n(x)\ \left(a_0 + a_1x + \cdots + a_{n-1}x^{n-1}\right)\mathrm{d}x = 0.$$

That is, the n-th Legendre polynomial multiplied with an $(n-1)$-th degree polynomial integrates to zero on the domain $(-1,1)$. Test this property by integrating $P_3(x) = \frac{1}{2}(5x^3 - 3x)$ multiplied with $a_0 + a_1x + a_2x^2$ on the domain $(-1,1)$.

Question *A*. 3. Polynomial division. We can always divide two polynomials and write the result in terms of a 'quotient' $Q(x)$ and a 'remainder' $R(x)$: $\frac{f(x)}{g(x)} = Q(x) + \frac{R(x)}{g(x)}$. An easy way to check the result is to multiply both sides with $g(x)$, finding $f(x) = Q(x)g(x)+R(x)$. Use that check to confirm the following polynomial division:

$$\underbrace{\frac{1 + 2x + x^2 + 5x^3}{\frac{1}{2}(3x^2 - 1)}}_{f(x)/g(x)} = \underbrace{\frac{2}{3} + \frac{10x}{3}}_{Q(x)} + \underbrace{\frac{\frac{4}{3} + \frac{11x}{3}}{\frac{1}{2}(3x^2 - 1)}}_{R(x)/g(x)}.$$

An important observation is that when we divide an M-th degree polynomial by an N-th degree polynomial, we make $Q(x)$ and $R(x)$ of maximal degree $(M - N)$. Indeed, in this example we divided a 3rd degree polynomial by a 2nd degree polynomial, and both $Q(x)$ and $R(x)$ are of 1st degree only.

Question *A*. 4. Gauss-Legendre quadrature: putting it all together. Confirm the following sketch of a proof. Say $f(x)$ is a $(2n-1)$th degree polynomial. Do polynomial division with the n-th order Legendre polynomial to write $f(x) = Q(x)P_n(x) + R(x)$. We know that $Q(x)$ and $R(x)$ are of maximum degree $n - 1$. Integrate on both sides,

$$f(x) = Q(x)P_n(x) + R(x) \Longleftrightarrow \int_{-1}^{1} f(x)\,\mathrm{d}x = \underbrace{\int_{-1}^{1} Q(x)P_n(x)\,\mathrm{d}x}_{=0} + \int_{-1}^{1} R(x)\,\mathrm{d}x. \qquad (3)$$

The integral with the Legendre polynomial goes to zero because $P_n(x)$ is of higher degree than the polynomial $Q(x)$. Apparently, $f(x)$ (of degree $2n - 1$) integrates to the same value as $R(x)$ (of degree $n - 1$). And we can exactly integrate the latter $(n - 1)$-th order polynomial using n sampled points,

$$\int_{-1}^{1} f(x)\,\mathrm{d}x = \int_{-1}^{1} R(x)\,\mathrm{d}x = w_1 R(x_1) + w_2 R(x_2) + \cdots + w_n R(x_n);$$

we obtain the weights using, e.g., eq. (1). Gauss realized that we don't even *need* to know $R(x)$, because $R(x) = f(x)$ where $P_n(x) = 0$, (see eq. (3), on the left). So if we choose the points $x_{1,2,\ldots,n}$ as those where $P_n(x_i) = 0$, we get $R(x)$ straight from $f(x)$, and write

$$\int_{-1}^{1} f(x)\,\mathrm{d}x = \int_{-1}^{1} R(x)\,\mathrm{d}x = w_1 f(x_1) + w_2 f(x_2) + \cdots + w_n f(x_n).$$

This is the Gaussian quadrature. We choose our n sampling points as the n zeroes of $P_n(x)$. Then using eq. (1), we obtain the n associated weights. The weighted sum then correctly integrates polynomials of order $2n - 1$. With 10 weights, for example, we can integrate a 19th order polynomial correctly!

Question *A*. Assume the function $f(x)$ to integrate is of degree $2N - 1 = 3$, so $N = 2$. We will thus use 2 points to approximate the integration. Confirm that the zeroes of the 2nd degree Legendre polynomial, $P_2(x) = \frac{1}{2}(3x^2 - 1)$, are $x_{1,2} = \pm\sqrt{1/3}$. Solve a system like eq. (1) at the given x_i to obtain the weights $w_1 = w_2 = 1$. So, with just two points,

$$\int_{-1}^{1} f(x)\,\mathrm{d}x \approx f\left(-\sqrt{\frac{1}{3}}\right) + f\left(\sqrt{\frac{1}{3}}\right),$$

the integration is fully accurate for any polynomial of type $a + bx + cx^2 + dx^3$! Compare this to eq. (2) which only holds for polynomials of type $a + bx$, while using the same number of samples! The clever choice of the sampling *locations* improves the accuracy.

Question *A*. The 3rd degree Legendre polynomial is $P_3 = \frac{1}{2}(5x^3 - 3x)$. Go through the steps outlined above to find

$$\int_{-1}^{1} f(x)\,\mathrm{d}x \approx \frac{5}{9} f\left(-\sqrt{\frac{3}{5}}\right) + \frac{8}{9} f(0) + \frac{5}{9} f\left(\sqrt{\frac{3}{5}}\right), \tag{4}$$

which is accurate for polynomials up to 5th order. *Hint: obtain the zeroes of the Legendre polynomial and solve eq.* (1).

Question *A*. 5. Gauss-Legendre-Lobatto quadrature. Later on, in the lecture on the spectral element method, we will use the Gauss-Legendre-Lobatto (GLL) quadrature to integrate functions on our domain. What sets this method apart is that we also want the points at $x = -1$ and $x = 1$ to be part of the integration limits. It turns out that this method is identical to the Gauss-Legendre quadrature, except that we use the zeroes of $P_n'(x)$, alongside the points $x = -1$ and $x = 1$. This quadrature is only accurate up to $2n - 3$. Confirm that for $P_2 = \frac{1}{2}(3x^2 - 1)$, the GLL needs a sample at $x = 0$; confirm that for $P_3 = \frac{1}{2}(5x^3 - 3x)$ the GLL needs samples from $x = \pm \sqrt{1/5}$. Then we simply solve the system of eq. (1) to obtain the associated weights.

$$\int_{-1}^{1} f(x)\,\mathrm{d}x \overset{\mathrm{GLL}_2}{\approx} \frac{1}{3} f(-1) + \frac{4}{3} f(0) + \frac{1}{3} f(1),$$
$$\int_{-1}^{1} f(x)\,\mathrm{d}x \overset{\mathrm{GLL}_3}{\approx} \frac{1}{6} f(-1) + \frac{5}{6} f(-\sqrt{1/5}) + \frac{5}{6} f(\sqrt{1/5}) + \frac{1}{6} f(1).$$

Hint: take the derivative of the Legendre polynomials, and find its zeroes.