MATH661 Homework 1 - Number approximation

Posted: 08/24/22

Due: 09/01/22, 11:55PM

Track 1: 1,2,3,6. Track 2: 1-6.

- Julia preamble

1. Construct a convergence plot in logarithmic coordinates for the following continued fraction approximation of e

$$e = 2 + \cfrac{1}{1 + \cfrac{1}{2 + \cfrac{2}{3 + \cfrac{3}{\ddots}}}} \tag{1}$$

Identify the terms in the general expression of a continued fraction

$$F_n = b_0 + \operatorname*{K}_{k=1}^{n} \frac{a_k}{b_k}.$$

Compare with the additive approximation from a McLaurin series

$$e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \cdots$$

Estimate the rate and order of convergence for both approximations.

Solution. Rewrite (1) as

$$e = 2 + \frac{1}{f},\ f = 1 + \cfrac{1}{2 + \cfrac{2}{3 + \cfrac{3}{\ddots}}},$$

and introduce the continued fraction sequence $\{F_n\}_{n\in\mathbb{N}}$,

$$F_n = b_0 + \operatorname*{K}_{k=1}^{n} \frac{a_k}{b_k} = b_0 + \cfrac{a_1}{b_1 + \cfrac{a_2}{b_2 + \cfrac{a_3}{b_3 + \ddots}}},$$

$$F_0 = b_0,\ F_1 = b_0 + \frac{a_1}{b_1},\ F_2 = b_0 + \cfrac{a_1}{b_1 + \cfrac{a_2}{b_2}},\ F_3 = b_0 + \cfrac{a_1}{b_1 + \cfrac{a_2}{b_2 + \cfrac{a_3}{b_3}}}$$

with assumed limit $\lim_{n\to\infty} = f$. Writing out the first few terms

$$F_0 = 1, F_1 = 1 + \frac{1}{2}, F_2 = 1 + \frac{1}{2 + \frac{2}{3}}, F_3 = 1 + \frac{1}{2 + \frac{2}{3 + \frac{3}{4}}}, \ldots,$$

leads to identification of coefficients as

$$b_k = k + 1, a_k = k.$$

- Define a function f to compute F_n and return $E_n = 2 + 1/F_n \to e$.
- Define a function g to compute

$$M_n = 1 + \frac{1}{1!} + \frac{1}{2!} + \cdots + \frac{1}{n!}.$$

The convergence behavior of the two approximations E_n, M_n is shown in Fig. 1.

-

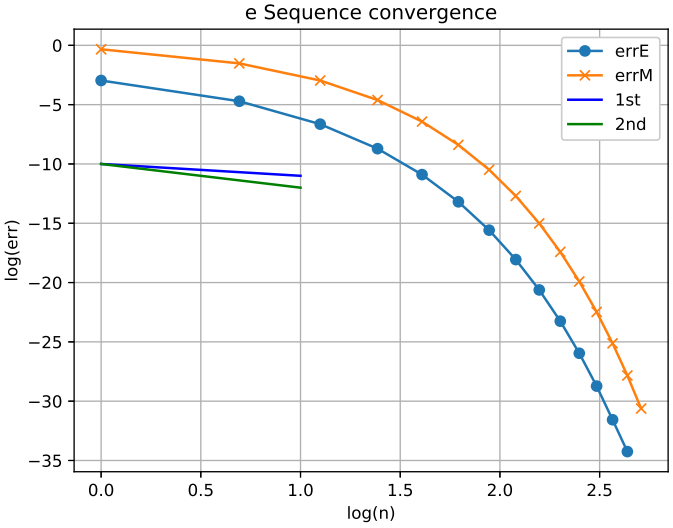

Figure 1. Convergence of continued fraction and additive approximations of e.

The definition of order p and rate r of convergence

$$\lim_{n\to\infty} \frac{|x_{n+1} - x|}{|x_n - x|^p} = r,$$

is based upon the assumption of power-law decrease of the error $e_n = |x_n - x|$,

$$e_{n+1} \sim e_n^p \Leftrightarrow e_{n+1} = r e_n^p .$$

In log-coordinates this assumption leads to a straight line representation

$$\log e_{n+1} = p \log e_n + \log r .$$

The $(\log n, \log e_n)$ representation in Fig. 1 does not allow direct identification of an order of convergence. However a plot of $(\log e_n, \log e_{n+1})$ is easily constructed (Fig. 2), and shows $p \cong 1$, $\ln r \cong -2 \Rightarrow r = e^{-2} \cong 0.135$.

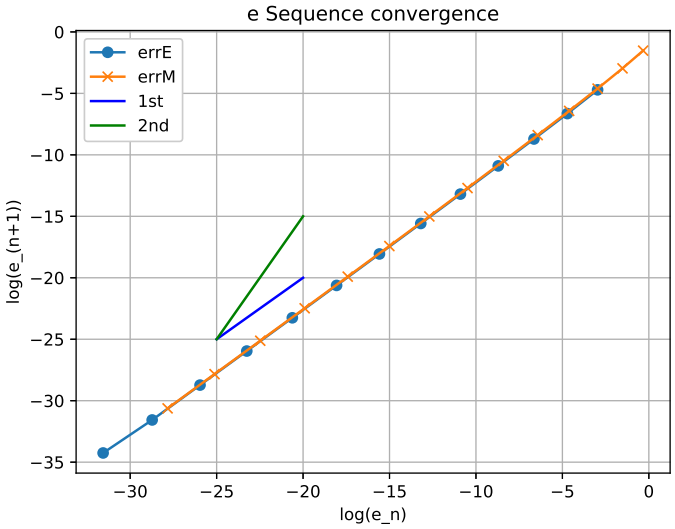

Figure 2. Order-of-convergence plot for approximations of e. Visual estimation indicates $p = 1$, linear sequence convergence.

2. Apply convergence acceleration to both the above approximations of e. Construct the convergence plot of the accelerated sequences, and estimate the new rate and order of convergence.

   Solution. Since both sequences exhibit linear convergence the Aitken formula

$$a_n = x_n - \frac{(x_n - x_{n-1})^2}{x_n - 2x_{n-1} + x_{n-2}}$$

   is applicable. The resulting accelerated convergence plot is shown in Fig. 3. Convergence acceleration to second order is observed for a small range of errors, $\ln e \in [-7, -4]$. For smaller errors, the floating point system cannot separate the small differences appearing in the Aitken correction.

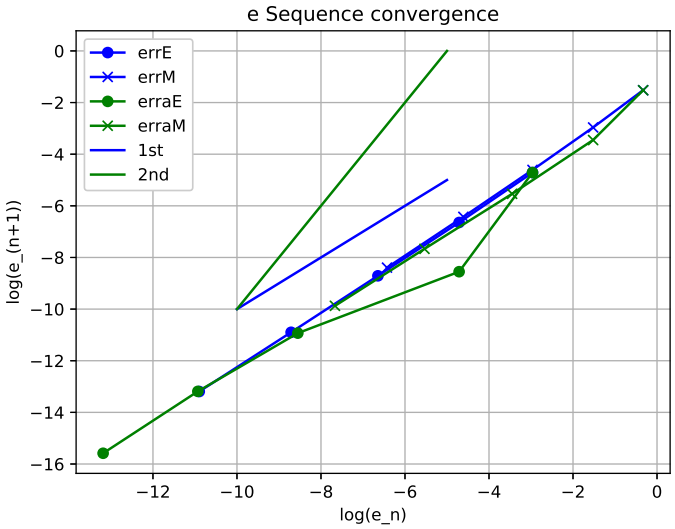

Figure 3. Convergence of Aitken acceleration of e approximation sequences

3. Completely state the mathematical problem of taking the n^{th} root of a positive real, $n \in \mathbb{N}$. Find the absolute and relative condition numbers.

   Solution. First, assume $n > 0$ fixed leading to the problem $f\colon \mathbb{R}_+ \to \mathbb{R},\ f(x) = x^{1/n}$. The absolute condition number is

   $$\hat{\kappa} = \lim_{\varepsilon \to 0} \sup_{|\delta x| \leqslant \varepsilon} \frac{\|f(x+\delta x) - f(x)\|}{\|\delta x\|}.$$

   The condition number furnishes a bound for the change in the solution upon a change in the input

   $$\|f(x+\delta x) - f(x)\| \leqslant \hat{\kappa}\, \|\delta x\|.$$

   Using the absolute value norm $\|x\| = |x|$ for $x \in \mathbb{R}_+$ gives

   $$\hat{\kappa} = \lim_{\varepsilon \to 0} \sup_{|\delta x| \leqslant \varepsilon} \frac{|f(x+\delta x) - f(x)|}{|\delta x|} = \left| \frac{\mathrm{d}f(x)}{\mathrm{d}x} \right| = \frac{1}{n} x^{\frac{1}{n}-1} = \frac{1}{n} \frac{1}{x^{(n-1)/n}}, \text{for } x > 0.$$

   Consider some simple cases:

   - $n = 1,\ f(x) = x,\ \hat{\kappa} = 1$, hence perturbations in the input are not amplified

     $$f(x) = x,\ f(x+\delta x) = x + \delta x,\ f(x+\delta x) - f(x) = \delta x.$$

     - The problem is well-conditioned.

   - $n = 2,\ f(x) = \sqrt{x}$,

     $$\hat{\kappa} = \frac{1}{2\sqrt{x}}.$$

- at $x = 1/4$, $\hat{\kappa} = 1$, indicating no amplification of input perturbation
- as $x \to \infty$, $\hat{\kappa} \to 0$, indicating input perturbations have negligible effect upon output. This indicates incorrect identification of the variables in a problem.
- as $x \to 0_+$, $\hat{\kappa} \to \infty$, indicating small input perturbations have arbitrarily large effects upon output. This indicates an ill-posed problem

In general, the conditioning of the n^{th} root operation

$$\hat{\kappa} = \frac{1}{n} \frac{1}{x^{(n-1)/n}}, \text{for } x > 0$$

indicates ill-conditioning for $x \to 0$, well conditioning for $x \sim 1$, incorrect model for $x \to \infty$.

Consider now what happens when n is also allowed to vary. The definition of the condition number cannot be directly applied since the limit process is not defined for $n \in \mathbb{N}$. One can however consider the problem $g: \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}$

$$g(x, y) = x^{1/y},$$

and notice that $h: \mathbb{R}_+ \times \mathbb{N} \to \mathbb{R}$, $h(x, n) = x^{1/n}$ is a restriction of g. The condition number of h can be inferred from that of g

$$\hat{\kappa} = \lim_{\varepsilon \to 0} \sup_{\|\delta z\| \leqslant \varepsilon} \frac{|g(x + \delta x, y + \delta y) - g(x, y)|}{\|\delta z\|}, z = \begin{bmatrix} \delta x & \delta y \end{bmatrix}^T.$$

Use the ∞-norm for $\delta z \in \mathbb{R}^2_+$, $\|\delta z\|_\infty = \max(\delta x, \delta y)$. When approaching zero perturbation, $\|\delta z\| \to 0$ above the first bisector, the inequality

$$\delta x < \delta y \leqslant \varepsilon$$

implies

$$\hat{\kappa} = \left| \frac{\partial g}{\partial y} \right|.$$

Conversely, for $\|\delta z\| \to 0$ below the first bisector

$$\hat{\kappa} = \left| \frac{\partial g}{\partial x} \right|.$$

In general, for some arbitrary path to approach zero,

$$\hat{\kappa} = \max\left(\left| \frac{\partial g}{\partial x} \right|, \left| \frac{\partial g}{\partial y} \right| \right) = \max\left(\frac{1}{y} \frac{1}{x^{(y-1)/y}}, \frac{x^{1/y} \ln x}{y^2} \right).$$

When restricted to $y = n \in \mathbb{N}$,

$$\hat{\kappa} = \max\left(\frac{1}{n}\frac{1}{x^{(n-1)/n}}, \frac{x^{1/n}\ln x}{n^2}\right).$$

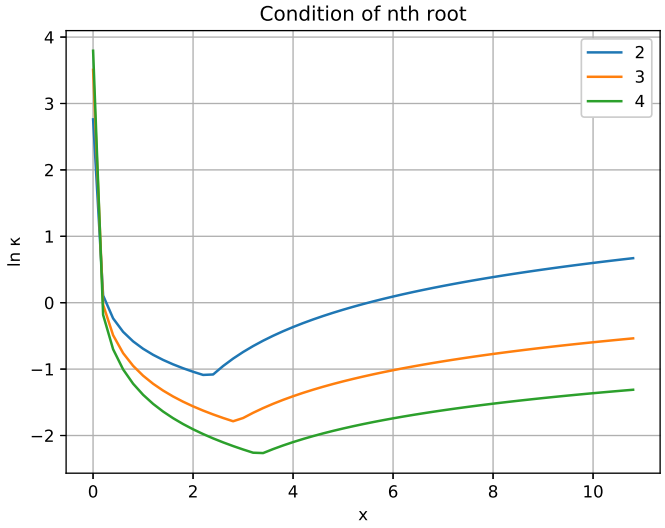

Figure 4. Conditioning of n^{th}-root $x^{1/n}$ with perturbations allowed in n.

4. Completely state the mathematical problem of solving the initial value problem for an ordinary differential equation of first order. Find the absolute and relative condition numbers.

   Solution. The IVP

   $$y' = f(y), y(0) = y_0$$

   is formulated as the mathematical problem

   $$F: C^{0,1}(\mathbb{R}) \times \mathbb{R} \to C(\mathbb{R}),$$

   that when evaluated for some slope function f and initial condition y_0 gives the integral curve $y: [0, a) \to \mathbb{R}$, $a > 0$. $F(f, y_0) = y$. In the above $C(\mathbb{R})$ is the space of continuous functions and $C^{0,1}$ is the space of Lipschitz continuous functions.

   As in the n^{th}-root problem there are two inputs to F, and it is useful to start with the case in which the slope function f is fixed and only y_0 varies. The Lyapunov exponent L is defined as

   $$\delta y(t) \cong e^{Lt}\,\delta y_0,$$

   to characterize this case and the condition number is simply

   $$\hat{\kappa}(t) = e^{Lt},$$

i.e., the condition number and the Lyapunov exponent express the same concept.

The condition number for the case of variation of the slope function requires a concept of taking a derivative with respect to a function f, a generalization of the calculus concept of taking a derivative with respect to a variable. This is known as a functional derivative and can be defined in the Fréchet sense for normed spaces and in the Gateaux sense for Banach spaces.

5. Completely state the mathematical problem of finding the roots of a cubic polynomial. Find the absolute and relative condition numbers.

   Solution. Consider the cubic $x^3 + a_2x^2 + a_1x + a_0 = 0$ with roots x_1, x_2, x_3 related to the polynomial coefficients by the Vieta relations

$$x_1 + x_2 + x_3 = -a_2, x_1x_2 + x_1x_3 + x_2x_3 = a_1, x_1x_2x_3 = -a_0 \,.$$

   The mathematical problem of finding the roots of the cubic is $f: \mathbb{R}^3 \to \mathbb{C}^3$

$$f(\boldsymbol{a}) = \boldsymbol{x}, \boldsymbol{a} = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix}, \boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

   Consider the effect of small changes $\delta\boldsymbol{a}$ upon the roots by taking differentials

$$\begin{aligned} \delta x_1 + \delta x_2 + \delta x_3 &= -\delta a_2 \\ (x_2 + x_3)\delta x_1 + (x_3 + x_1)\delta x_2 + (x_1 + x_2)\delta x_3 &= \delta a_1 \\ x_2x_3\,\delta x_1 + x_3x_1\,\delta x_2 + x_1x_2\,\delta x_3 &= -\delta a_0 \end{aligned}$$

   This is a linear system for $\delta\boldsymbol{x}$ with matrix

$$\boldsymbol{B} = \begin{bmatrix} 1 & 1 & 1 \\ x_2 + x_3 & x_3 + x_1 & x_1 + x_2 \\ x_2x_3 & x_3x_1 & x_1x_2 \end{bmatrix}.$$

   The condition number for f is the maximal amplification by the matrix $\boldsymbol{B}$ or

$$\hat{\kappa} = \|\boldsymbol{B}\|.$$

   Consider some specific cases:

   - For $x_1 = x_2 = x_3 = \xi$,

$$\boldsymbol{B} = [\ \boldsymbol{b}\ \ \boldsymbol{b}\ \ \boldsymbol{b}\ ], \boldsymbol{b} = \begin{bmatrix} 1 \\ 2\xi \\ \xi^2 \end{bmatrix}$$

6. Numerically compare the approximation of $b(t) = t(\pi - t)(2\pi - t)$ by linear combination of $\mathcal{T} = \{\sin t, \sin 2t, \sin 3t, ...\}$ with that of linear combinations of $\mathcal{E} = \{1, e^t, e^{2t}, e^{3t}, ...\}$. Present a study of the aproximation error as the number of terms in the linear combination increases. Estimate the order of convergence in both cases.

   Solution. The approximation is stated as

$$b(t) \cong \hat{b}_n(t) = \sum_{k=1}^{n} c_k \, a_k(t)$$

   with the basis functions chosen either as $a_k(t) = \sin(kt)$ or $a_k(t) = e^{(k-1)t}$. The approximation error $e(t) = b(t) - \hat{b}_n(t)$ can be measured in various norms, e.g., the 2-norm

$$\varepsilon^2 = \|e(t)\|_2^2 = \int_0^{2\pi} (b(t) - \hat{b}_n(t))^2 \, \mathrm{d}t \cong \frac{2\pi}{m} \sum_{i=1}^{m} (b(t_i) - \hat{b}_n(t_i))^2.$$

   - Based upon the code from Fig.1 in L04, define a function that returns the approximation error for given basis set $a_k(t)$, number of terms n, and evaluation points m.

```
∴ function err(m,n,a,dbg=false)
    h=2.0*π/m; j=1:m; t=h*j;
    A = a.(1,t)
    for k=2:n
      A = [A a.(k,t)]
    end
    if dbg
      return A
    end
    bt=t.*(π.-t).*(2*π.-t)
    x=A\bt; b=A*x;
    return norm(b-bt)*(2*pi)/m
  end
```

```
err
```

```
∴
```

   - Define the two basis sets of interest

```
∴ s(k,t) = sin(k*t); e(k,t)=exp((k-1)*t);
```

```
∴
```

   - Verify the err function constructs the expected matrix. Note that the basis function is passed as an argument.

The numerical values within the matrix $\boldsymbol{A}$ are hard to interpret for large m, hence plot the columns in Fig. 5. The basis fubnctino plots already indicate that the exponential family is likely to lead to bad approximations since with respect to $e^{(n-1)t}$, all e^{kt} for $k < n-1$ are negligibly small, hence $\boldsymbol{A}$ is likely to have only one independent column vector.

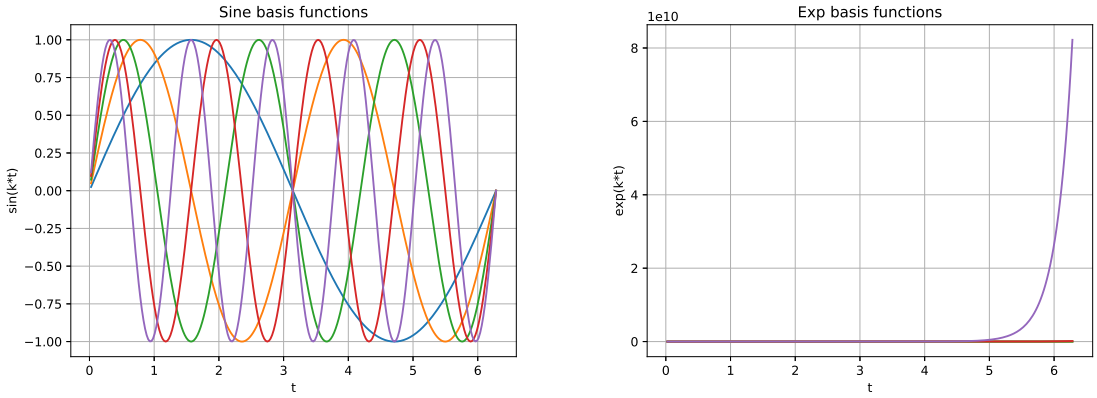

Figure 5. Left: Sine basis functions. Right: Exponential basis functions.

- Test the err function for larger m values (more samples).

The convergence behavior is shown in Fig. 6. As expected as the number of terms in the linear combination increases the sine approximation converges, while that for the exponential basis has a constant error (numerically, the rank of the matrix remains 1).

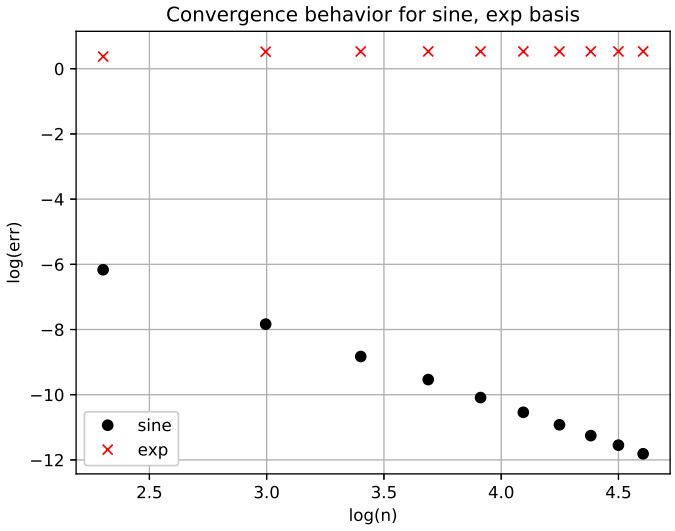

Figure 6. Convergence of sine, exp basis approximation of $b(t)$ with increasing number of terms.