Validated Numerical Proofs

Pre-enrolment Quiz

Calculus

1. Throughout this question p_n denotes the degree-n Taylor polynomial of f about the point x_0.

   (a) Take $f(x) = e^x$ and $x_0 = 0$. Write down p_3, and use it to estimate $e^{0.1}$. How large is the error?

   (b) Still with $f(x) = e^x$ and $x_0 = 0$, find the smallest n for which you can *guarantee*

   $$|e^x - p_n(x)| \le 10^{-10} \qquad \text{for every } x \in [-1, 1].$$

   (c) Now take $g(x) = \dfrac{1}{1 + 25x^2}$ and $x_0 = 0$. By writing g as a geometric series, find all x for which its Taylor series about 0 converges.

2. Define $F : \mathbb{R}^2 \to \mathbb{R}^2$ by

   $$F(x, y) = \begin{pmatrix} x^2 + y^2 - 4 \\ xy - 1 \end{pmatrix},$$

   and let $X = [1.8,\, 2.0] \times [0.4,\, 0.6]$.

   (a) Compute the Jacobian matrix $DF(x, y)$.

   (b) Find all real solutions of $F(x, y) = 0$ in closed form. Verify that exactly one of them lies in the rectangle X. (One route is to substitute $y = 1/x$ and solve the resulting equation for x^2.)

   (c) Each entry of DF is a continuous function on the closed bounded set X, so it attains a maximum and a minimum there. Find them.

   (d) Show that $\det DF(x, y) \neq 0$ at every point of X.

3. Let $a \neq 0$ and b be constants.

   (a) Use an integrating factor to solve the initial value problem

   $$y'(t) = a\, y(t) + b, \qquad y(0) = y_0.$$

   (b) Describe the behaviour of the solution as $t \to \infty$, in the two cases $a < 0$ and $a > 0$.

4. Let $T(x) = \cos x$. Show that T maps $[0, 1]$ into $[0, 1]$, and that

   $$|T(x) - T(y)| \le q\,|x - y| \qquad \text{for all } x, y \in [0, 1],$$

   where $q = \sin 1$. Check that $q < 1$.

Linear algebra

5. For a square matrix A with entries a_{ij}, the norm $\|A\|_\infty$ is defined to be the largest of the absolute row sums,
$$\|A\|_\infty = \max_i \sum_j |a_{ij}|.$$

   (a) Evaluate $\|B\|_\infty$ for
$$B = \begin{pmatrix} 4 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 4 \end{pmatrix},$$
   and find the eigenvalues of B.

   (b) It can be shown that if $\|E\|_\infty < 1$ then $I - E$ is invertible, with
$$\|(I-E)^{-1}\|_\infty \le \frac{1}{1-\|E\|_\infty}.$$
   Take $E = \left(\begin{smallmatrix} 0 & 0.3 \\ 0.2 & 0 \end{smallmatrix}\right)$. Check that the hypothesis holds, write down the bound this gives on $\|(I-E)^{-1}\|_\infty$, and compare it against the exact value.

Numerical methods and programming

6. You are not expected to write anything sophisticated in part (c). The point is to be comfortable running a small numerical experiment and interpreting the output. Use whichever language you prefer (Python, Julia, MATLAB, ...).

   (a) Derive Newton's method for solving $f(x) = 0$ from the tangent line approximation to f at x_k. Write down the resulting iteration for $f(x) = x^2 - 2$, and compute x_1 and x_2 by hand starting from $x_0 = 1$.

   (b) The bisection method halves an enclosing interval at each step. Starting from an interval of length 1, how many iterations are needed to reach an interval of width at most 10^{-10}?

   (c) For a system $F(z) = 0$ in several variables, Newton's method reads $z_{k+1} = z_k - DF(z_k)^{-1}F(z_k)$, which in practice means solving $DF(z_k)\,\delta_k = -F(z_k)$ and setting $z_{k+1} = z_k + \delta_k$. Implement this for the system $F(x, y) = 0$ of Question 2, starting from $(x_0, y_0) = (1.8,\ 0.6)$. Tabulate $\|z_k - z^*\|_\infty$ and $\|F(z_k)\|_\infty$ against k, using the exact root z^* from Question 2(b). (For a vector v, $\|v\|_\infty$ means the larger of $|v_1|$ and $|v_2|$.) How does the number of correct digits change from one step to the next? The error should stop decreasing after a handful of steps. At roughly what level does it stall?

Solutions

1. (a) Every derivative of e^x is e^x, and all of them equal 1 at $x_0 = 0$, so

$$p_3(x) = 1 + x + \frac{x^2}{2} + \frac{x^3}{6}, \qquad p_3(0.1) = 1.1051\overline{6}.$$

The true value is $e^{0.1} = 1.1051709181\ldots$, so the error is about 4.2×10^{-6}. (Taylor's theorem bounds it by $e^{0.1}(0.1)^4/4! \approx 4.6 \times 10^{-6}$, which is consistent.)

(b) Taylor's theorem with the Lagrange form of the remainder gives, for $x \in [-1, 1]$ and some ξ between 0 and x,

$$|e^x - p_n(x)| = \frac{e^\xi}{(n+1)!}|x|^{n+1} \le \frac{e}{(n+1)!}.$$

We need $(n+1)! \ge e \times 10^{10} \approx 2.72 \times 10^{10}$. Since $13! \approx 6.23 \times 10^9$ and $14! \approx 8.72 \times 10^{10}$, the smallest admissible value is $n+1 = 14$, i.e. $n = 13$, giving the bound $e/14! \approx 3.12 \times 10^{-11}$. (For $n = 12$ the bound is $e/13! \approx 4.37 \times 10^{-10}$, which is not good enough.)

(c) Substituting $w = -25x^2$ into the geometric series $1/(1-w) = \sum_{k \ge 0} w^k$,

$$g(x) = \frac{1}{1 + 25x^2} = \sum_{k=0}^{\infty} (-1)^k 25^k x^{2k}.$$

This geometric series converges precisely when $|-25x^2| < 1$, that is when $|x| < 1/5$, and diverges for $|x| \ge 1/5$ because the terms $25^k |x|^{2k}$ do not tend to zero. Since g is analytic at 0, this series *is* its Taylor series.

So the Taylor series converges only on $(-1/5,\ 1/5)$, a small part of $[-1, 1]$, even though g itself is perfectly well behaved on the whole interval.

2. (a)

$$DF(x, y) = \begin{pmatrix} 2x & 2y \\ y & x \end{pmatrix}.$$

(b) Write $s = x + y$ and $p = xy$. The second equation gives $p = 1$, and the first gives $x^2 + y^2 = s^2 - 2p = s^2 - 2 = 4$, so $s = \pm\sqrt{6}$. Also $(x-y)^2 = s^2 - 4p = 2$, so $x - y = \pm\sqrt{2}$. Hence

$$x = \frac{s \pm \sqrt{2}}{2}, \qquad y = \frac{s \mp \sqrt{2}}{2}, \qquad s = \pm\sqrt{6},$$

giving the four real solutions

$$\pm\left(\tfrac{\sqrt{6}+\sqrt{2}}{2},\ \tfrac{\sqrt{6}-\sqrt{2}}{2}\right) \approx \pm(1.93185,\ 0.51764),$$

$$\pm\left(\tfrac{\sqrt{6}-\sqrt{2}}{2},\ \tfrac{\sqrt{6}+\sqrt{2}}{2}\right) \approx \pm(0.51764,\ 1.93185).$$

Following the hint instead, $y = 1/x$ turns the first equation into $x^2 + x^{-2} = 4$, that is $x^4 - 4x^2 + 1 = 0$, so $x^2 = 2 \pm \sqrt{3}$. These agree with the values above, since $\left(\frac{\sqrt{6}\pm\sqrt{2}}{2}\right)^2 = \frac{8 \pm 4\sqrt{3}}{4} = 2 \pm \sqrt{3}$.

Only $z^* = \left(\frac{\sqrt{6}+\sqrt{2}}{2}, \frac{\sqrt{6}-\sqrt{2}}{2}\right) \approx (1.9318517,\ 0.5176381)$ lies in $X = [1.8, 2.0] \times [0.4, 0.6]$.

(c) Each entry is a monotone function of a single variable, so the extreme values occur at the endpoints of the corresponding edge of the rectangle, giving

$$2x \in [3.6,\ 4.0], \qquad 2y \in [0.8,\ 1.2], \qquad y \in [0.4,\ 0.6], \qquad x \in [1.8,\ 2.0].$$

(d) $\det DF = 2x^2 - 2y^2$. On X we have $x^2 \geq 1.8^2 = 3.24$ and $y^2 \leq 0.6^2 = 0.36$, so

$$\det DF(x,y) \geq 2(3.24 - 0.36) = 5.76 > 0,$$

and likewise $\det DF \leq 2(4.0 - 0.16) = 7.68$. (The true value at z^* is $4\sqrt{3} \approx 6.928$.)

3. (a) Write the equation as $y' - ay = b$ and multiply by the integrating factor $\mu(t) = e^{-at}$, chosen so that the left-hand side becomes an exact derivative,

$$\big(e^{-at}y\big)' = e^{-at}\big(y' - ay\big) = b\,e^{-at}.$$

Integrating from 0 to t and using $y(0) = y_0$,

$$e^{-at}y(t) - y_0 = b\int_0^t e^{-as}\,\mathrm{d}s = \frac{b}{a}\big(1 - e^{-at}\big),$$

so that

$$y(t) = y_0 e^{at} + \frac{b}{a}\big(e^{at} - 1\big) = \left(y_0 + \frac{b}{a}\right)e^{at} - \frac{b}{a}.$$

As a check, setting $t = 0$ returns y_0, and differentiating gives $y' = a\left(y_0 + \frac{b}{a}\right)e^{at} = ay + b$.

(b) The constant $-b/a$ is the equilibrium, being the unique solution of $ay + b = 0$, and the formula above says the deviation from it is $y(t) + \frac{b}{a} = \big(y_0 + \frac{b}{a}\big)e^{at}$.
For $a < 0$ this deviation decays, so every solution tends to $-b/a$ exponentially at rate $|a|$, whatever the initial value. For $a > 0$ it grows, so unless y_0 is exactly $-b/a$ the solution leaves every bounded set, again exponentially, at rate a.

4. cos is decreasing on $[0,1]$, so $T([0,1]) = [\cos 1, 1] \approx [0.5403, 1] \subseteq [0,1]$. By the mean value theorem, for $x, y \in [0,1]$ there is ξ between them with

$$|\cos x - \cos y| = |\sin \xi|\,|x - y| \leq \sin(1)\,|x - y|,$$

using that sin is increasing on $[0,1]$. So the stated inequality holds with $q = \sin 1 \approx 0.8414710$, and $q < 1$ because $1 < \pi/2$, so that $\sin 1$ is below its maximum value of 1.

5. (a) The absolute row sums of B are 5, 6 and 5, so $\|B\|_\infty = 6$.
The characteristic polynomial factorises as $\det(B - \lambda I) = (4 - \lambda)\big((4 - \lambda)^2 - 2\big)$, giving eigenvalues

$$\lambda = 4 - \sqrt{2} \approx 2.5858, \qquad 4, \qquad 4 + \sqrt{2} \approx 5.4142.$$

(b) The row sums of E are 0.3 and 0.2, so $\|E\|_\infty = 0.3 < 1$ and the hypothesis holds. The bound is therefore

$$\|(I - E)^{-1}\|_\infty \leq \frac{1}{1 - 0.3} = \frac{10}{7} \approx 1.4286.$$

For the exact value, $I - E = \left(\begin{smallmatrix} 1 & -0.3 \\ -0.2 & 1 \end{smallmatrix}\right)$ has determinant $1 - 0.06 = 0.94$, so

$$(I - E)^{-1} = \frac{1}{0.94}\begin{pmatrix} 1 & 0.3 \\ 0.2 & 1 \end{pmatrix}, \qquad \|(I - E)^{-1}\|_\infty = \frac{1.3}{0.94} \approx 1.3830.$$

The bound holds, and is a mild overestimate.

6. (a) Replacing f by its tangent line at x_k gives $f(x) \approx f(x_k) + f'(x_k)(x - x_k)$. Setting the right-hand side to zero and calling the solution x_{k+1} gives

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}.$$

For $f(x) = x^2 - 2$ we have $f'(x) = 2x$, so

$$x_{k+1} = x_k - \frac{x_k^2 - 2}{2x_k} = \frac{1}{2}\left(x_k + \frac{2}{x_k}\right),$$

the classical iteration for $\sqrt{2}$. From $x_0 = 1$,

$$x_1 = \frac{1}{2}(1 + 2) = \frac{3}{2}, \qquad x_2 = \frac{1}{2}\left(\frac{3}{2} + \frac{4}{3}\right) = \frac{17}{12} = 1.41\overline{6},$$

against $\sqrt{2} = 1.4142135624\ldots$, and the next iterate is $577/408 = 1.4142156863\ldots$, in error by about 2×10^{-6}.

(b) After n steps the width is 2^{-n}, so we need $2^{-n} \le 10^{-10}$, that is $n \ge \log_2 10^{10} = 33.22\ldots$, giving 34 iterations. The width halves deterministically, so this count is fixed in advance and does not depend on f at all.

(c) Here is one implementation in Python.

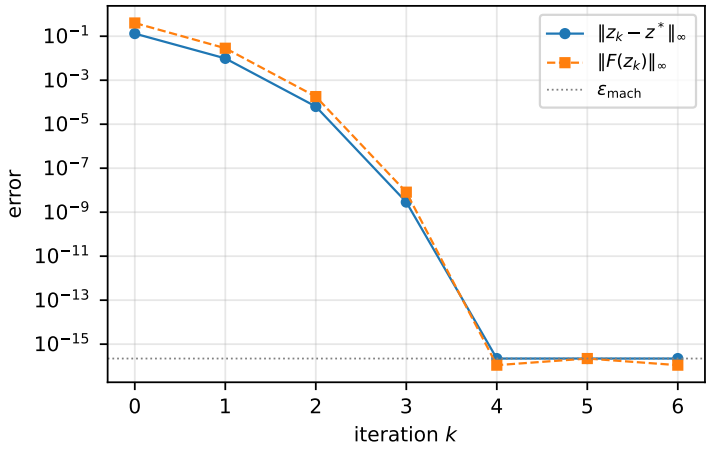
```
import numpy as np

F = lambda z: np.array([z[0]**2 + z[1]**2 - 4, z[0]*z[1] - 1])
DF = lambda z: np.array([[2*z[0], 2*z[1]], [z[1], z[0]]])
zstar = np.array([(np.sqrt(6)+np.sqrt(2))/2, (np.sqrt(6)-np.sqrt(2))/2])

z = np.array([1.8, 0.6])
for k in range(6):
    print(k, np.linalg.norm(z-zstar, np.inf), np.linalg.norm(F(z), np.inf))
    z = z - np.linalg.solve(DF(z), F(z))
```

The iterates converge to $z^* \approx (1.9318516526,\ 0.5176380902)$, with the errors shown below.

k	$\|z_k - z^*\|_\infty$	$\|F(z_k)\|_\infty$
0	1.32×10^{-1}	4.00×10^{-1}
1	9.81×10^{-3}	2.85×10^{-2}
2	6.38×10^{-5}	1.81×10^{-4}
3	2.87×10^{-9}	8.13×10^{-9}
4	2.22×10^{-16}	1.11×10^{-16}
5	2.22×10^{-16}	2.22×10^{-16}



The errors run $10^{-1}, 10^{-2}, 10^{-4}, 10^{-9}, 10^{-16}$, so the number of correct digits roughly doubles from one step to the next, and each step squares the previous error up to a

constant factor, $\|z_{k+1} - z^*\| \le C\|z_k - z^*\|^2$. This is the characteristic behaviour of Newton's method near a simple root, and here it relies on $DF(z^*)$ being nonsingular, which is what Question 2(d) established.

The error stalls at about 2×10^{-16}, the level of the machine epsilon of IEEE double precision ($2^{-52} \approx 2.2 \times 10^{-16}$). Once z_k is within one unit in the last place of z^*, no further improvement is representable, and the computed iterates simply jitter at that level.