Validated Numerical Proofs

Pre-enrolment Quiz

Calculus and Taylor's theorem

1. Let f be $(n+1)$-times continuously differentiable on an interval containing x_0 and x, and let p_n denote the degree-n Taylor polynomial of f about x_0.

   (a) Take $f(x) = e^x$ and $x_0 = 0$. Find the smallest n for which you can *guarantee*
   $$|e^x - p_n(x)| \le 10^{-10} \qquad \text{for every } x \in [-1, 1].$$

   (b) Now take $f(x) = \dfrac{1}{1+25x^2}$ and $x_0 = 0$. By writing f as a geometric series, show that its Taylor series about 0 converges only for $|x| < 1/5$, and hence that
   $$\sup_{x\in[-1,1]} |f(x) - p_n(x)| \longrightarrow \infty \quad \text{as } n \to \infty.$$

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.

   (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 $\det DF(x, y) \ne 0$ at every point of X.

3. Let $a \ne 0$ and $b \in \mathbb{R}$ 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 $a > 0$.

Linear algebra

4. Write down the formula for $\|A\|_\infty$ in terms of the entries of A, and evaluate it for
   $$B = \begin{pmatrix} 4 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 4 \end{pmatrix}.$$
   Find the eigenvalues of B and hence $\|B\|_2$.

5. Let $\|\cdot\|$ denote a matrix norm induced by a vector norm on $\mathbb{R}^n$ (so it is submultiplicative, that is, $\|AB\| \le \|A\|\|B\|$, and $\|I\| = 1$). Suppose E is an $n \times n$ matrix with $\|E\| < 1$. Show that $I - E$ is invertible and that
   $$\|(I - E)^{-1}\| \le \frac{1}{1 - \|E\|}.$$

Real analysis and metric spaces

6. Let $X = [0, 1]$ with the usual metric and $T(x) = \cos x$. Show that $T(X) \subseteq X$ and that T is a contraction, and give an explicit contraction constant q.

7. Suppose $(a_n)_{n\geq 0}$ satisfies $\sum_{n\geq 0} |a_n| < \infty$. Show that $\sum_{n\geq 0} a_n \cos(n\pi x)$ converges uniformly on $[0, 1]$, and deduce that its sum is continuous.

Numerical methods and programming

8. 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) Implement Newton's method 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). 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) For $f = e^x$, derivatives of f of all orders are e^x, so Taylor's theorem with the Lagrange form of the remainder gives, for $x \in [-1, 1]$ and some $\xi \in (0, x)$,

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

To guarantee the desired bound, we require $(n+1)! \ge e \times 10^{10} \approx 2.72 \times 10^{10}$. 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.)

(b) Substituting $w = -25x^2$ into the geometric series $1/(1-w) = \sum_{k\ge0} w^k$,

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

valid precisely when $|-25x^2| < 1$, i.e. $|x| < 1/5$. Only even powers occur, so the degree-n Taylor polynomial collects exactly the terms with $2k \le n$, giving

$$p_n(x) = \sum_{k=0}^{\lfloor n/2 \rfloor} (-1)^k 25^k x^{2k}, \qquad \text{so in particular } p_{2m} = p_{2m+1}.$$

For $|x| > 1/5$ the terms $25^k x^{2k}$ do not tend to zero, so the series diverges. Taking $x = 1$ for instance,

$$p_n(1) = \sum_{k=0}^{\lfloor n/2 \rfloor} (-25)^k = \frac{1 - (-25)^{\lfloor n/2 \rfloor + 1}}{26},$$

whose modulus tends to infinity, while $f(1) = 1/26$. Hence $|f(1) - p_n(1)| \to \infty$, and the supremum over $[-1, 1]$ diverges.

2. (a)

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

(b) Let $s = x + y$ and $p = xy$. For $F(x, y) = 0$, $p = 1$, and $x^2 + y^2 = s^2 - 2p = s^2 - 2 = 4$, so $s = \pm\sqrt{6}$. Furthermore, $(x - y)^2 = s^2 - 4p = 2$, 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

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

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

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) $\det DF = 2x^2 - 2y^2$. For $(x, y) \in X$, 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) Rearranging and introducing the integrating factor $\mu(t) = e^{-at}$ gives

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

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

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

so that

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

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 $y(t) + \frac{b}{a} = \left(y_0 + \frac{b}{a}\right)e^{at}$.
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. Recall that $\|A\|_\infty = \max_i \sum_j |a_{ij}|$, the maximum absolute row sum. For B, the row sums are 5, 6 and 5, respectively, so $\|B\|_\infty = 6$.

The characteristic polynomial factorises as $\det(B - \lambda I) = (4 - \lambda)\big((4 - \lambda)^2 - 2\big)$, giving eigenvalues

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

and since B is symmetric, its 2-norm is the largest absolute eigenvalue, so $\|B\|_2 = 4 + \sqrt{2} \approx 5.4142$.

5. Suppose $(I - E)v = 0$ for some v. Then $v = Ev$, so $\|v\| = \|Ev\| \leq \|E\|\,\|v\|$. As $\|E\| < 1$ this ensures $\|v\| = 0$ and so $v = 0$. Hence, $(I - E)$ is invertible, so let $C = (I - E)^{-1}$. From $C(I - E) = I$ we get $C = I + CE$, hence

$$\|C\| \leq \|I\| + \|C\|\,\|E\| = 1 + \|C\|\,\|E\| \quad \Longrightarrow \quad \|C\|\,(1 - \|E\|) \leq 1,$$

and dividing by $1 - \|E\| > 0$ gives $\|C\| \leq 1/(1 - \|E\|)$.

(Equivalently, $\sum_{k\geq 0} E^k$ converges absolutely because $\sum_k \|E^k\| \leq \sum_k \|E\|^k < \infty$, and telescoping $(I - E)\sum_{k=0}^m E^k = I - E^{m+1}$ identifies the sum as $(I - E)^{-1}$. This is the Neumann series.)

6. $\cos(x)$ is decreasing on $x \in [0, 1]$, 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 $\xi \in (x, y)$ with

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

since $\sin x$ is increasing on $x \in [0, 1]$. So T is a contraction with $q = \sin(1) \approx 0.8414710 < 1$.

7. Letting $s_m(x) = \sum_{n=0}^{m} a_n \cos(n\pi x)$, since $|a_n \cos(n\pi x)| \le |a_n|$ for all x, and $\sum_n |a_n|$ converges, the Weierstrass M-test (with $M_n = |a_n|$) gives uniform convergence of (s_m) on $[0,1]$.

   In particular, since the tail of a convergent series tends to zero, for any $\epsilon > 0$ there exists N such that for $m, m' \ge N$ with $m' > m$, $\|s_{m'} - s_m\|_\infty \le \sum_{n=m+1}^{m'} |a_n| < \epsilon$. So (s_m) is uniformly Cauchy and therefore uniformly convergent. Each s_m is a finite sum of continuous functions, and is hence continuous. Since a uniform limit of continuous functions is itself continuous, the sum is continuous.

8. (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.

   (c) Here is one implementation in Python.

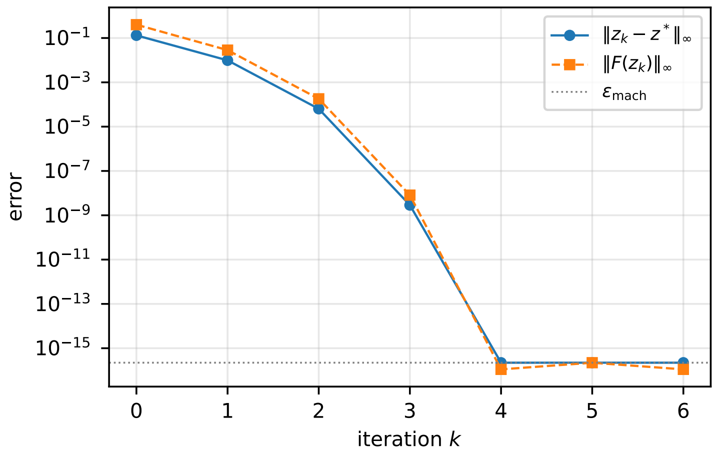
```
import numpy as np

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, $\|z_{k+1} - z^*\| \leq C\|z_k - z^*\|^2$. 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.