MA5233 Computational Mathematics

Assignment 1

Simon Etter
Deadline: 10 September 2020, 7pm
Total marks: 20

1 Floating-point numbers [4 marks]

Answer the following questions using pen-and-paper-style analysis [2 marks each].

1. What is the smallest positive integer which cannot be represented exactly in `Float64`?
2. What is the largest positive and finite integer which can be represented exactly in `Float64`?

Hint. You can check your answers using Julia.

2 Big O notation [4 marks]

Consider the statement

$$\big(1 + O(\varepsilon)\big)^{-1} = 1 + O(\varepsilon) \quad \text{for} \quad \varepsilon \to 0.$$

1. [2 marks] Verify this statement using the definition

$$f(x) = O\big(g(x)\big) \text{ for } x \to x_0 \qquad \Longleftrightarrow \qquad \lim_{x \to x_0} \frac{|f(x)|}{|g(x)|} < \infty.$$

2. [2 marks] Verify this statement using the definition

$$f(x) = O\big(g(x)\big) \text{ for } x \to x_0 \qquad \Longleftrightarrow \qquad \exists\, \delta, C > 0\ \forall x \in x_0 + [-\delta, \delta] : |f(x)| \le C\, |g(x)|.$$

3 The `expm1` function [8 marks]

Consider the function $f(x) = \exp(x) - 1$ and its numerical implementation $\tilde f(x) = \mathtt{T}(\exp(x)) \ominus 1$ (we ignore the rounding of the input of $\tilde f(x)$ for simplicity).

1. [2 marks] Show that $\kappa(f, x)$ is bounded for all $x \in [-1, 1]$.
2. [2 marks] Complete `expm1_errors()` such that it produces a plot demonstrating that

$$\frac{|\tilde f(x) - f(x)|}{|f(x)|} = O\big(x^{-1}\big) \qquad \text{for } x \to 0.$$

3. [2 marks] Determine $x \in [-1, 1]$ such that $\frac{|\tilde f(x) - f(x)|}{|f(x)|} = 1$. Demonstrate analytically that your choice of x indeed leads to a relative error of 1.
4. [1 mark] Determine functions $g(y)$, $h(x)$ such that $f(x) = g(h(x))$ and at least one of the two functions is ill conditioned for $x \approx 0$. Compute the condition number of the ill-conditioned function.

5. [1 mark] Explain in your own words why Julia provides a special function `expm1(x)` for evaluating $f(x) = \exp(x) - 1$.
6. [unmarked] Extend `expm1_errors()` such that it also shows the relative errors for $\tilde{f}(x) = \texttt{expm1}(x)$.

4 Rounding errors when solving linear systems [4 marks]

Assume $\tilde{x} \in \mathbb{R}^n$ is a numerically computed approximation to the solution of the linear system $Ax = b$ where $A \in \mathbb{R}^{n \times n}$ and $x, b \in \mathbb{R}^n$.

1. [3 marks] Determine a constant $C \in \mathbb{R}$ depending only on the singular values $\sigma_1 \geq \ldots \geq \sigma_n > 0$ of A and $\|x\|_2$, $\|b\|_2$ such that

$$\frac{\|\tilde{x} - x\|_2}{\|x\|_2} = C\, O(\texttt{eps()}).$$

   You may assume that the numerical solution $\tilde{x}$ is computed using a backward stable algorithm, i.e. $\tilde{x}$ is the solution to $\tilde{A}\tilde{x} = b$ where $\tilde{A} \in \mathbb{R}^{n \times n}$ satisfies

$$\frac{\|\tilde{A} - A\|_2}{\|A\|_2} = O(\texttt{eps()}).$$

2. [1 mark] Complete the function `linear_system_error()` to check your answer to Task 1.

   Hint. You can compute the singular values of `A` using `svdvals(A)`, and you can compute the 2-norm of a vector `x` using `norm(x)`. The estimated relative error may be one or two order of magnitudes larger than the exact relative error (order of magnitude = power of 10).