COL726 Assignment 1

13–27 January, 2022

Note: All answers should be accompanied by a rigorous justification, unless the question explicitly states that a justification is not necessary.

1. Suppose a function $\mathbf{f} : X \to Y$ is given as a black box, and let $\mathbf{x} \in X$ be an input vector. I choose several random vectors $\mathbf{h}_1, \ldots, \mathbf{h}_k \in X$ with small norm $\|\mathbf{h}_i\| \leq d$, and evaluate $\mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{x}+\mathbf{h}_1), \ldots, \mathbf{f}(\mathbf{x}+\mathbf{h}_k)$.

   (a) From these values (and their linear combinations, inner products, norms, etc.), what can I conclude about the relative condition number $\kappa(\mathbf{x})$? Be as specific as possible. Assume d is small enough that nonlinearities in f are not significant.

   (b) What problems do you expect if this procedure is carried out on a machine with floating-point arithmetic?

2. Consider the function $f(x) = 1 - \cos x$.

   (a) What is its relative condition number at $x = 0$?

   (b) Assume that floating-point functions $\widetilde{\sin}, \widetilde{\cos}, \widetilde{\tan} : \mathbf{F} \to \mathbf{F}$ are provided which compute the corresponding trigonometric functions in a stable way. Is the naive algorithm $\tilde{f}(x) = 1 \ominus \widetilde{\cos}(\mathrm{fl}(x))$ stable?

   (c) Suggest an algorithm to compute $f(x)$ with improved stability. Plot the results of both algorithms for small positive x on a log-log scale.

3. Recall that I can represent a quadratic polynomial $p(z) = az^2 + bz + c$ by a vector $\begin{bmatrix} p(0) \\ p(1) \\ p(2) \end{bmatrix} \in \mathbb{C}^3$.

   (a) Find the corresponding basis polynomials $e_1(z), e_2(z), e_3(z)$, such that $p(z)$ is represented by a vector $\mathbf{x}$ if and only if $p(z) = x_1 e_1(z) + x_2 e_2(z) + x_3 e_3(z)$.

   (b) Show that differentiation is a linear transformation of polynomials, and find the matrix $\mathbf{D}$ that represents it in this basis.

   (c) Without performing any arithmetic calculations, prove that $\mathbf{D}$ does not have full rank. Your proof should not depend on the values of the entries of $\mathbf{D}$ found in part (b).

4. Suppose I have an orthonormal set of n vectors $\{\mathbf{q}_1, \ldots \mathbf{q}_n\} \subset \mathbb{C}^m$, with $n < m$. Give a stable algorithm to find a new vector $\mathbf{q}_{n+1}$, so that $\{\mathbf{q}_1, \ldots \mathbf{q}_n, \mathbf{q}_{n+1}\}$ is still an orthonormal set. Here stability means that $\mathbf{q}_i^* \mathbf{q}_{n+1} = O(\epsilon_{\mathrm{m}})$ for all $i = 1, \ldots, n$, and $\|\mathbf{q}_{n+1}\|_2 = 1 + O(\epsilon_{\mathrm{m}})$.

Hint: Consider all m standard basis vectors $\mathbf{e}_1, \ldots, \mathbf{e}_m$. Which one will give the best stability when used to construct $\mathbf{q}_{n+1}$?

5. Consider the function $f : \mathbb{C}^m \to \mathbb{R}$ given by $f(\mathbf{x}) = \frac{1}{2}(\min\{|x_1|, \ldots, |x_m|\} + \max\{|x_1|, \ldots, |x_m|\})$, the average of the minimum and maximum absolute entries of $\mathbf{x}$.

   (a) Show from first principles that f is a norm when $m = 2$, i.e. that it satisfies all three norm conditions. You may assume that $|x + y| \leq |x| + |y|$ for $x, y \in \mathbb{C}$.

   (b) Show that f is not a norm when $m \geq 3$.

6. The infinite series $\sum_{i=1}^{\infty} i^{-2} = 1 + \frac{1}{4} + \frac{1}{9} + \cdots$ sums to $s_\infty = \pi^2/6$. For any finite n, it can be shown that the partial sum $s_n = \sum_{i=1}^{n} i^{-2}$ has truncation error $|s_n - s_\infty| = O(n^{-1})$.

   (a) Suppose the sum is computed by initializing $s = 0$ and then iteratively accumulating s += n^{-2} in a loop. For what n will the floating-point error in this addition be comparable to the term n^{-2} being added? What will be the truncation error at that point? Give both answers in big O notation in terms of ϵ_{m}.

   (b) It is observed that the sequence $a_n = ns_n - (n-1)s_{n-1}$ has the same limit and converges faster to s_∞. What will be the additional floating-point error if this expression is evaluated as written, i.e. $a_n = (n \otimes s_n) \ominus ((n-1) \otimes s_{n-1})$? For what n will this exceed the truncation error in s_n itself? Again, state your answers in terms of big O of ϵ_{m}.

   (c) Write a program `baselSeries(n)` to compute both sequences in single precision (see below), and return the pair $([s_1, \ldots, s_n], [a_1, \ldots, a_n])$. Evaluate the relative errors of both sequences (using $s_\infty \approx$ `(math.pi**2)/6`), and plot them for $n = 2^0, 2^1, \ldots, 2^{20}$ on a log-log scale. Explain whether the plots match what you predicted in parts (a) and (b).

   We are using single precision only to make the effects of rounding visible sooner. To use single-precision arithmetic in Python, make sure all your integers are converted to single precision using `np.single`, e.g. `s = np.single(0)` and `s += np.single(1)/np.single(n**2)`.

Collaboration policy: Refer to the policy on the course webpage.

If you collaborated with others to solve any question(s) of this assignment, give their names in your submission. If you found part of a solution using some online resource, give its URL.

Submission: This assignment has two submission forms on Gradescope. In one form, you have to submit a PDF of your answers for all questions, and in the other you have to submit your code for Question 6. Both submissions must be uploaded before the assignment deadline.

Code submissions should contain a single file `a1.py` which contains the requested function and any helper functions. You are permitted but not required to include the code for producing the plot. The plot itself should be included in your PDF.