MATH 387 ASSIGNMENT 1

DUE WEDNESDAY FEBRUARY 7

1. In this exercise, we will study multiplication algorithms for numbers of the form

$$a = \pm \sum_{k=0}^{\infty} a_k \beta^{k+e},$$

where $0 \le a_k \le \beta - 1$ are the digits (or "big digits") of the mantissa that are stored in a finite (!) array of integers, and $e \in \mathbb{Z}$ is the exponent.

It is clear that multiplication reduces to multiplication of two positive integers. To multiply two positive integers, we first define the *Cauchy product*

$$ab = \Big(\sum_{j=0}^{\infty} a_j \beta^j \Big) \cdot \Big(\sum_{i=0}^{\infty} b_i \beta^i \Big) = \sum_{k=0}^{\infty} \Big(\sum_{j=0}^{k} a_j b_{k-j} \Big) \beta^k = \sum_{k=0}^{\infty} p_k^* \beta^k, \tag{1}$$

where

$$p_k^* = \sum_{j=0}^{k} a_j b_{k-j}, \tag{2}$$

is the k-th *generalized digit* of ab. In general, p_k^* can be larger than $\beta - 1$, and so (1) is not the base-β expansion of the product ab. However, the proper digits $0 \le p_k \le \beta - 1$ of ab can be found by writing p_k^* in base-β and then performing the summation in the right hand side of (1) in base-β arithmetic. One way to find the base-β expansion of p_k^* would be to do the summation in (2) from the beginning in base-β arithmetic.

(a) Formulate a multiplication algorithm that is based on the double summation in (1). In other words, we do not want to compute the generalized digits p_k^* explicitly. Write a convincing argument (i.e., proof) that your algorithm terminates in a finite number of steps, and that it returns the correct answer. Suppose that n is the largest index for which $a_n \neq 0$, and that m is the analogous quantity for b. Basically, n and m measure how much storage each of a and b takes. Then estimate the number of built-in arithmetic operations needed to compute the digits of ab, in terms of n and m, when n and m are large.

(b) With the intent of saving resources, let us ignore the terms with $k < k^*$ in (1), with the truncation parameter k^*, i.e., we replace the product ab by

$$\tilde{p} = \sum_{k=k^*}^{\infty} \Big(\sum_{j=0}^{k} a_j b_{k-j} \Big) \beta^k.$$

Date: Winter 2018.

Show that

$$0 \leq ab - \tilde{p} \leq ab \cdot \beta^{k^*+3-n-m},$$

where n and m are as in (a). What would be a good choice for the value of k^*?

2. (a) Perform a round-off error analysis on the pairwise summation algorithm.
   (b) Perform a round-off error analysis on the "pairwise product" algorithm, which is the analogue of the pairwise summation algorithm for computing products.
3. (a) Let us call the functions $\sin x$, $\arctan x$, e^x, and $\log x$ the *basic functions*. Then reduce the evaluation of $\cos x$, $\tan x$, $\arcsin x$, $\arccos x$, and x^a ($a \in \mathbb{R}$) into basic functions and elementary arithmetic operations. Here by elementary arithmetic operations we understand addition, subtraction, multiplication, division, and n-th root extraction $\sqrt[n]{x}$ (for $n > 0$ integer and $x > 0$ real), and all variables are real (in the sense that they are not complex variables).
   (b) Reduce the argument of $\arctan x$ into $[0, b]$, where $b < 1$ is to be chosen by you (Generally, $b \approx \frac{1}{2}$ would be considered satisfactory). In other words, express $\arctan x$ with $x \in \mathbb{R}$, in terms of $\arctan y$ with $0 \leq y \leq b$.
4. Perform a detailed round-off error analysis on an algorithm that computes $\log y$ by employing the Gregory series

$$\log \frac{1+x}{1-x} = 2\big(x + \frac{x^3}{3} + \frac{x^5}{5} + \dots\big).$$

   You can assume $-\frac{1}{2} \leq x \leq \frac{1}{2}$ in the analysis. Then describe a procedure to reduce the argument into $-\frac{1}{2} \leq x \leq \frac{1}{2}$.
5. In this exercise, we work with *arbitrary precision* floating point numbers, which are numbers of the form $x = m\beta^e$, where $m \in \mathbb{Z}$ and $e \in \mathbb{Z}$. In this setting, floating point operations can be performed with any prescribed accuracy, but the cost of operations grows (linearly to quadratically) with the bit-size of the mantissa. Design an algorithm that computes $\sin x$ ($0 < x \leq \frac{\pi}{4}$) with any given relative precision, by using its Maclaurin series. In other words, you need to prescribe the accuracy of each step in the computation, while keeping the whole computation reasonably efficient. Finally, describe a procedure to reduce the argument of $\sin x$ into $0 < x \leq \frac{\pi}{4}$.

Homework policy

You are welcome to consult each other provided (1) you list all people and sources who aided you, or whom you aided and (2) you write-up the solutions independently, in your own language. If you seek help from other people, you should be seeking general advice, not specific solutions, and must disclose this help. This applies especially to internet fora such as `MathStackExchange`.

Similarly, if you consult books and papers outside your notes, you should be looking for better understanding of or different points of view on the material, not solutions to the problems.