PURDUE UNIVERSITY · CS 31400
NUMERICAL ANALYSIS

HOMEWORK
David F. Gleich
January 29, 2021

Please answer the following questions in complete sentences in a typed manuscript and submit the solution on Gradescope by Feb 7 / Feb 8th (5am).

Problem 0: Homework checklist

- Please identify anyone, whether or not they are in the class, with whom you discussed your homework. This problem is worth 1 point, but on a multiplicative scale.
- Make sure you have included your source-code and prepared your solution according to the most recent Piazza note on homework submissions.

Problem 1: Gautschi Exercise 1.1

1. Enumerate all positive elements of $\mathbb{R}(3,2)$ that correspond to normalized floating point values. (Recall that $\mathbb{R}(3,2)$ is a floating point system that uses 2-bits for the exponent and 3-bits for the significand or mantissa.) List each number on one line as a decimal.
2. What is the distance $d(x)$ of a positive normalized floating point number to its next larger floating-point number? (Do this for all values in $\mathbb{R}(3,2)$.)
3. Determine the relative distance $r(x) = d(x)/x$ for all numbers and give upper and lower bounds.

Problem 2: Gautschi Exercise 1.11

Let $f(x) = \sqrt{1+x^2} - 1$.

1. Explain the difficulty of computing $f(x)$ for a small value of $|x|$ and show how it can be circumvented.
2. Determine an expression for the condition number of $f(x)$ and discuss conditioning for small $|x|$.
3. How can the answers to 1 and 2 be reconciled?

Problem 3: Gautschi Machine Exercise 1.4

Euler's constant $\gamma = 0.57721566490153286\ldots$ is defined as the limit

$$\gamma = \lim_{n\to\infty} \gamma_n \text{ where } \gamma_n = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} - \log(n).$$

Assuming that $\gamma - \gamma_n \sim cn^{-d}$, try to determine c and d on the computer.

Problem 4: Gautschi Machine Exercise 1.6a.

Write a program to compute

$$S_N = \sum_{n=1}^{N} (\frac{1}{n} - \frac{1}{n+1}) = \sum_{n=1}^{N} \frac{1}{n(n+1)}$$

once using the first formula and once using the second formula. For $N = 10^k, k = 1, 2, \ldots, 7$, print the respective absolute errors. Comment on the results.

Problem 5: The Hurwitz zeta function

This is a *real* problem that I ran into! I was using a program written by someone else to compute a particular statistical estimate. The details don't really matter, but at one step, the code needed to compute the Hurwitz zeta function:

$$H(s,q) = \sum_{n=0}^{\infty} \frac{1}{(q+n)^s}$$

where s ranged from 1 to 7 and where q ranged from 1 to 500. They were using Matlab, which doesn't provide a function for the Hurtwitz zeta function, but does provide a function for the Riemann zeta function:

$$R(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}.$$

Note that the only difference between these two functions is that the first term in the Riemann zeta function is $1/(1^s)$ and the first term in the Hurwitz zeta function is $1/(q^s)$. To account for this difference, the code I was using did the following:

```
function h=hzeta(s,q)
z = zeta(s)
h = z - sum((1:(q-1)).^(-s));
```

Was I happy when I realized this? You should use what you've learned about floating point computations to answer this question. You can evaluate the Hurwitz zeta function to arbitrary precision via Wolfram Alpha.

Problem 6

Suppose you want to evaluate $a+b+c$ on a machine. Suppose that $a > b > c > 0$ and that a, b, c are all exact on the machine. Empirically illustrate that

$$\mathrm{fl}(\mathrm{fl}(a+b)+c) \neq \mathrm{fl}(a+\mathrm{fl}(b+c)).$$

Now prove which of the three orders of evaluation result in the smallest error. (The three are (a+b)+c, (a+c)+b, a+(b+c).)

Problem 7

The Intel Penitum FDIV bug was illustrated with the procedure:

$$A - \frac{A}{B}B.$$

Determine the accuracy of this computation assuming that A and B can be exactly stored in floating point on the computer.

Problem 8

1. We want to understand computer implementations of methods to compute the roots of the quadratic equation given the coefficients a, b, c in

   $$ax^2 + bx + c = 0$$

   using the high school formula

   $$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

   Show empirically that this formula can result in catastrophic errors in the computed roots and identify the cause.

2. Consider Gautschi's suggestion in problem 9. Suppose we take (without loss of generality) the quadratic

$$x^2 + px + q = 0$$

with roots x_1, x_2. Suppose we compute the larger root first, and then use $x_1 x_2 = q$ to determine the additional root. This results in the Julia program:

```
x1=abs(p/2) + sqrt(p*p/4 - q)
if p > 0
    x1 = -x1
end
x2 = q/x1
```

Find two serious flaws with this program as a "general-purpose" quadratic equation solver. Support your arguments with specific examples, if necessary.

3. Consider reading Kahan's note on quadratics for more on how to accurately evaluate them. https://www.cs.berkeley.edu/~wkahan/Qdrtcs.pdf