Prof. W. Gander **Masaryk University Fall 2014**

Finite Arithmetic

1. Consider the following finite decimal arithmetic: 2 digits for the mantissa and one digit for the exponent. So the machine numbers have the form $\pm Z.Z\text{E}\pm Z$ where $Z \in \{0, 1, \dots, 9\}$

   (a) How many normalized machine numbers are available?

   (b) Which is the overflow- and the underflow range?

   (c) What is the machine precision?

   (d) What is the smallest and the largest distance of two consecutive machine numbers.

   Solution:

   - We first count the machine numbers. We can form 19 different exponents: $-9, -8, \dots, 9$. The first digit, before the decimal point, can not be zero, because we consider only normalized numbers, thus we have 9 possibilities for the first digit. Thus in total we have $2 \times 9 \times 10 \times 19 = 3420$ normalized machine numbers plus the number zero. Therefore the grand total is 3421 machine numbers.
   - The largest number is $9.9\text{E}9 = 9'900'000'000$ and the smallest positive number is $1.0\text{E}{-9}$. The overflow range is $|x| > 9.9\text{E}9$ and the underflow range is $0 < |x| < 1.0\text{E}{-9}$.
   - The machine precision is the spacing between the numbers in $(1, 10)$ thus $\varepsilon = 1.1\text{E}0 - 1.0\text{E}0 = 1\text{E}{-1}$.
   - The largest distance between two machine numbers occurs when the exponent is 9: $9.9\text{E}9 - 9.8\text{E}9 = 1\text{E}8$. The smallest distance is $1.1\text{E}{-9} - 1.0\text{E}{-9} = 1\text{E}{-10}$.

2. Compute the condition number when subtracting two real numbers, $\mathcal{P}(x_1, x_2) := x_1 - x_2$.

   Solution: Perturbing the data slightly as before with multiplication, we obtain

$$\frac{|(\hat{x}_1 - \hat{x}_2) - (x_1 - x_2)|}{|x_1 - x_2|} = \frac{|x_1\varepsilon_1 - x_2\varepsilon_2|}{|x_1 - x_2|} \leq \frac{|x_1| + |x_2|}{|x_1 - x_2|}\varepsilon.$$

   We see that if $\operatorname{sign}(x_1) = -\operatorname{sign}(x_2)$, which means the operation is an addition, then the condition number is $\kappa = 1$, meaning that the addition of two numbers is well conditioned. If, however, the signs are the same and $x_1 \approx x_2$, then $\kappa = \frac{|x_1|+|x_2|}{|x_1-x_2|}$ becomes very large, and hence subtraction is ill conditioned in this case.

3. When computing the function

$$f(x) = \frac{e^x - (1+x)}{x^2}$$

on a computer then a large errors occur for $x \approx 0$.

(a) Explain why this happens.

(b) Give a method how to compute $f(x)$ for $|x| < 1$ to machine precision and write a corresponding MATLAB function.

Solution:

(a) for small $x \approx 0$ the expression for $f(x)$ is suffering from cancellation, since $e^x \approx 1 + x$.

(b) a better expression is obtained by expanding e^x in the series:

$$\begin{aligned} f(x) &= \frac{e^x - (1+x)}{x^2} \\ &= \frac{1 + x + x^2/2! + x^3/3! + \cdots - (1+x)}{x^2} \\ &= \frac{1}{2!} + \frac{x}{3!} + \frac{x^2}{4!} + \cdots \end{aligned}$$

Thus

```
function y=f(x)
sn=0.5; s=0; t=0.5; k=2;
while sn~=s
   k=k+1; s=sn;
   t=t*x/k;
   sn=s+t;
end
y=s;

>> x=1e-6; [(exp(x)-(1+x))/x^2  f(x)]
ans =
   0.500044450291171   0.500000166666708
>> x=1e-7; [(exp(x)-(1+x))/x^2  f(x)]
ans =
   0.488498130835069   0.500000016666667
>> x=1e-8; [(exp(x)-(1+x))/x^2  f(x)]
ans =
                   0   0.500000001666667
```

4. Write a MATLAB function to compute the sine function in a machine-independent way using its Taylor series. Since the series is alternating, cancellation will occur for large $|x|$.

To avoid cancellation, reduce the argument x of $\sin(x)$ to the interval $[0, \frac{\pi}{2}]$. Then sum the Taylor series and stop the summation with the machine-independent criterion $s_n + t_n = s_n$, where s_n denotes the partial sum and t_n the next term. Compare the exact values for $[\mathtt{sin}(-10 + k/100)]_{k=0,\ldots,2000}$ with the ones you obtain from your MATLAB function and plot the relative error.

Solution: We reduce the angle in several steps. First we note the sign of the angle and replace x by $|x|$. Then we reduce the angle modulo 2π so that x is in the interval $(0, 2\pi)$. Then if $x > \pi$ we use the relation $\sin(x) = -\sin(\pi - x)$ to reduce $x \in (0, \pi)$. Finally we use $\sin(x) = \sin(\pi - x)$ to reduce $x \in (0, \pi/2)$.

```
function y=sinus(x)
% computing the sin using the Taylor series
% reduce angle x to (0,pi/2)
if x<0, v=-1; else v=1; end   % remember sign of angle
x=abs(x);
x=mod(x,2*pi);                % reduce to (0,2*pi)
if x>pi                       % reduce to (0,pi)
  v=-v; x=x-pi;
end
if x>pi/2                     % reduce to (0,pi/2)
  x=pi-x;
end
s=x; t=x; k=1;                % compute Taylor series
while s+t~=s
  k=k+2; t=-t*x/(k-1)*x/k; s=s+t;
end
y=v*s;                        % add sign
```

With the script

```
Y=[];
for k=0:2000
  x=-10+k/100;
  y=sinus(x);
  ye=sin(x);
  Y = [Y; x abs(y-ye)/abs(ye)];
end
plot(Y(:,1), Y(:,2))
```

we obtain the following figure