

1.4 Floating Point Arithmetic

Today's modern computers, even home personal computers, can usually store enough digits so that roundoff error is not too much of a problem, at least at the storage level. However, computers can literally make millions or billions of calculations in a numerical process, so the original roundoff error can propagate throughout the calculation, rendering the final answer meaningless.

In this section we look at floating point arithmetic and discuss some numerical effects that can have disastrous results on calculations.

Let's begin with an example that demonstrates catastrophic cancellation of digits and loss of precision.

► **Example 1.** Compute $\sqrt{9876} - \sqrt{9875}$ using 5-digit floating point arithmetic.

First, let's examine the relative error in storing $\sqrt{9876}$ in 5-digit floating point form. We're going to use Matlab's *Symbolic Toolbox*, an interface to the computer algebra system (CAS) Maple. Computations in Maple are done symbolically and are exact. Thus, instead of turning a numeric result, an exact expression is returned.

```
>> sym('sqrt(9876)')
ans =
sqrt(9876)
```

The *Symbolic Toolbox* has a command that will return a numerical approximation of a symbolic object, correct to as many places of accuracy that the user needs. The command **vpa** does the work.

```
>> help vpa
VPA    Variable precision arithmetic.
R = VPA(S) numerically evaluates each element of the
double matrix S using variable precision floating point
arithmetic with D decimal digit accuracy, where D is
the current setting of DIGITS. The resulting R is a SYM.

VPA(S,D) uses D digits, instead of the current setting
of DIGITS. D is an integer or the SYM representation of
a number.
```

¹ Copyrighted material. See: <http://msenex.redwoods.edu/Math4Textbook/>

We will use the last paragraph of the help message to approximate $\sqrt{9876}$, correct to 20 digits. Note that the symbolic object must be entered as a string (delimited by single apostrophes (ticks)).

```
>> vpa('sqrt(9876)',20)
ans =
99.378065990438755368
```

Use this result to store $\sqrt{9876}$ in 5-digit floating point format.

$$\sqrt{9876} = 9.9378 \cdot 10^1.$$

We calculate the relative error with

$$\frac{|x^* - x|}{|x|} = \frac{|9.9378 \cdot 10^1 - \sqrt{9876}|}{|\sqrt{9876}|}. \quad (1.1)$$

To find an approximate value of the relative error, we will again turn to the *Symbolic Toolbox*. We'll first store the needed computation as a string in the variable `rel` (use any variable name you like).

```
>> rel='abs(9.9378e1-sqrt(9876))/abs(sqrt(9876))'
rel =
abs(9.9378e1-sqrt(9876))/abs(sqrt(9876))
```

Now we'll use the `vpa` command to approximate to 20 digits of accuracy.

```
>> vpa(rel,20)
ans =
.66403424234193683981e-6
```

Thus, the relative error in storing $x = \sqrt{9876}$ in 5-digit floating point form $x^* = 9.9378 \cdot 10^1$ is approximately $6.6 \cdot 10^{-7}$.

Recall the definition of *significant digits*.

Significant Digits. The number x^* is said to approximate x to n significant digits if n is the largest nonnegative integer for which

$$\frac{|x^* - x|}{|x|} < 5 \cdot 10^{-n}.$$

Because the relative error is less than $5 \cdot 10^{-6}$, we know that $x^* = 9.9378 \cdot 10^1$ approximates $x - \sqrt{9876}$ to 6 significant digits.

In similar fashion, we can use the *Symbolic Toolbox* to approximate $\sqrt{9875}$ and find the relative error in storing this number in 5-digit floating point form. First, an approximation of $\sqrt{9875}$.

```
>> vpa('sqrt(9875)',20)
ans =
99.373034571758952600
```

Thus, the 5-digit floating point storage of $x = \sqrt{9875}$ is $x^* = 9.9373 \cdot 10^1$. The relative error in storing $\sqrt{9875}$ is next.

```
>> rel='abs(9.9373e1-sqrt(9875))/abs(sqrt(9875))'
rel =
abs(9.9373e1-sqrt(9875))/abs(sqrt(9875))
>> vpa(rel,20)
ans =
.34789879469399867106e-6
```

Thus, the relative error is approximately $3.4 \cdot 10^{-7}$, which is about the same as the relative error in storing $\sqrt{9876}$ in 5-digit floating point form. Because the relative error is less than $5 \cdot 10^{-7}$, we know that $x^* = 9.9373 \cdot 10^1$ approximates $x - \sqrt{9875}$ to 7 significant digits.

Now, let's subtract the stored 5-digit floating point form numbers and see what happens.

$$9.9378 \cdot 10^1 - 9.9373 \cdot 10^1 = 0.0005 \cdot 10^1$$

Note that all of the digits of the mantissa are gone save one. Further, note that there are no more digits to the right of the 5 in $0.00005 \cdot 10^1$, so when we adjust the exponent to place this result in 5-digit floating point form, as in

$$5.0000 \cdot 10^{-3},$$

the digits to the right of the decimal point in $5.0000 \cdot 10^{-3}$ are meaningless. Indeed, the computer could quite possibly shove in some digits from memory that are not zeros, just to pad the number.

Now, let's look at the relative error in approximating $x = \sqrt{9876} - \sqrt{9875}$ with this result.

```

>> rel=
'abs(5.0e-3-(sqrt(9876)-sqrt(9875)))/abs(sqrt(9876)-sqrt(9875))'
rel =
abs(5.0e-3-(sqrt(9876)-sqrt(9875)))/abs(sqrt(9876)-sqrt(9875))
>> vpa(rel,20)
ans =
.62444971890114329880e-2

```

The relative error is approximately $6.2 \cdot 10^{-3}$. Recall that the relative error in approximating $\sqrt{9876}$ and $\sqrt{9875}$ with 5-digit floating point numbers was of the order of 10^{-7} , so roughly speaking, the relative error made by subtracting, which is of the order 10^{-3} , is roughly $10^{-3}/10^{-7} = 10^4$, or 10 000 times larger! Further, because the relative error of the subtraction is less than $5 \cdot 10^{-2}$, only 2 significant digits remain!

This phenomenon is called *catastrophic cancellation* of digits and occurs whenever you attempt to subtract two numbers that are very close to one another. The programmer needs to be aware of this phenomenon and avoid subtracting two nearly equal numbers. In this case, we can change the subtraction into addition by rationalizing the numerator with this calculation.

$$\sqrt{9876} - \sqrt{9875} = \frac{9876 - 9875}{\sqrt{9876} + \sqrt{9875}} = \frac{1}{\sqrt{9876} + \sqrt{9875}}$$

If we add the floating point representations of $\sqrt{9876}$ and $\sqrt{9875}$, we get

$$9.9378 \cdot 10^1 + 9.9373 \cdot 10^1 = 19.8751 \cdot 10^2,$$

which when stored in 5-digit floating point form is $1.9875 \cdot 10^2$. Next, we perform the division,

$$\frac{1.0000 \cdot 10^0}{1.9875 \cdot 10^2} = 0.50314465408805 \cdot 10^{-2},$$

which when stored in 5-digit floating form is $x^* = 5.0314 \cdot 10^{-3}$. The relative error when approximating $x = \sqrt{9876} - \sqrt{9875}$ with $x^* = 5.0314 \cdot 10^{-3}$ is

$$\frac{|x^* - x|}{|x|} = \frac{|5.0314 \cdot 10^{-3} - (\sqrt{9876} - \sqrt{9875})|}{|\sqrt{9876} - \sqrt{9875}|}.$$

We can use the *Symbolic Toolbox* to help with this calculation.

```
>> rel=
'abs(5.0313e-3-(sqrt(9876)-sqrt(9875)))/abs(sqrt(9876)-sqrt(9875))',
rel =
abs(5.0313e-3-(sqrt(9876)-sqrt(9875)))/abs(sqrt(9876)-sqrt(9875))
>> vpa(rel,20)
ans =
.23587741414644558543e-4
```

Thus, the relative error when approximating $x = \sqrt{9876} - \sqrt{9875}$ with $x^* = 5.0314 \cdot 10^{-3}$ is approximately $2.3 \cdot 10^{-5}$, which is less than $5 \cdot 10^{-5}$, so this time we've kept 5 significant digits, which is much better than the 2 significant digits of the previous computation.

Let's look at another example.

► **Example 2.** Solve the quadratic equation $x^2 - 1634x + 2 = 0$ using 10-digit floating point arithmetic.

Using the quadratic formula to solve $x^2 - 1634x + 2 = 0$, we get

$$x = \frac{1634 \pm \sqrt{1634^2 - 4(1)(2)}}{2(1)} = 817 \pm \sqrt{667487}.$$

We can use the *Symbolic Toolbox* to approximate $\sqrt{667487}$.

```
>> vpa('sqrt(667487)',20)
ans =
816.99877600887505858
```

Thus, in 10-digit floating point form, $\sqrt{667487} \approx 8.169987760 \cdot 10^2$. Thus, one solution of the quadratic is

$$x_1^* = 8.170000000 \cdot 10^2 + 8.169987760 \cdot 10^2 = 16.339987760 \cdot 10^2,$$

or $x_1^* = 1.633998776 \cdot 10^3$ in 10-digit floating point form. We can use the *Symbolic Toolbox* to calculate the relative error in approximating $x_1 = 817 + \sqrt{667487}$ with this 10-digit floating point number.

```

>> rel=
'abs(1.633998776e3-(817+\sqrt(667487)))/abs(817+sqrt(667487))'
rel =
abs(1.633998776e3-(817+\sqrt(667487)))/abs(817+sqrt(667487))
>> vpa(rel,20)
ans =
.54314964676275835537e-11

```

Thus, the relative error is approximately $5.4 \cdot 10^{-12}$, which is less than $5 \cdot 10^{-11}$, so we are approximating $817 + \sqrt{667487}$ to 11 significant digits.

The second root is

$$x_2^* = 8.170000000 \cdot 10^2 - 8.169987760 \cdot 10^2 = 0.000012240 \cdot 10^2,$$

or $x_2 = 1.224000000 \cdot 10^{-3}$, in 10-digit floating point form. Note the catastrophic cancellation of digits. The last 5 zeros of 1.224000000 are completely meaningless and in some cases the computer will pad these places with nonzero digits that happen to be lying around in memory.

We can use the *Symbolic Toolbox* to calculate the relative error made when approximating $x_2 = 817 - \sqrt{667487}$ with the 10-digit floating point number $x_2^* = 1.224000000 \cdot 10^{-3}$.

```

>> rel=
'abs(1.224000000e-3-(817-\sqrt(667487)))/abs(817-sqrt(667487))'
rel =
abs(1.224000000e-3-(817-\sqrt(667487)))/abs(817-sqrt(667487))
>> vpa(rel,20)
ans =
.72509174283635093702e-5

```

Thus, the relative error is approximately $7.3 \cdot 10^{-6}$, which is less than $5 \cdot 10^{-5}$, so $x_2^* = 1.224000000 \cdot 10^{-3}$ is approximating $x_2 = 817 - \sqrt{667487}$ to only 5 significant digits!

Programmers have to be aware of catastrophic cancellation of digits any time to numbers are subtracted that are very nearly equal in value. The programmer has to look for an algorithm that avoids subtraction.

In this case, consider the idea that if x_1 and x_2 are roots of the quadratic equation $x^2 - 1634x + 2 = 0$, then the quadratic $x^2 - 1634x + 2$ can be

factored as $(x - x_1)(x - x_2)$. If we expand and compare to the original quadratic, then

$$\begin{aligned} x^2 - 1634x + 2 &= (x - x_1)(x - x_2) \\ &= x^2 - (x_1 + x_2)x + x_1x_2. \end{aligned}$$

Comparing the constant terms, $x_1x_2 = 2$, or equivalently,

$$x_2 = \frac{1}{x_1}.$$

This last result will allow avoid subtraction in calculation the second root x_2 .

$$\begin{aligned} x_2^* &= \frac{2}{x_1^*} \\ &= \frac{2.000000000 \cdot 10^0}{1.633998776 \cdot 10^3} \\ &= 1.2239911369332985296 \cdot 10^{-3} \\ &= 1.223991137 \cdot 10^{-3} \end{aligned}$$

We can use Matlab to calculate the relative error in approximating $x_2 = 817 - \sqrt{667487}$ with the 10-digit floating point number $x_2^* = 1.223991137 \cdot 10^{-3}$.

```
>> rel=
'abs(1.223991137e-3-(817-\sqrt{667487}))/abs(817-\sqrt{667487})'
rel =
abs(1.223991137e-3-(817-\sqrt{667487}))/abs(817-\sqrt{667487})
>> vpa(rel,20)
ans =
.98518524802025190487e-8
```

Thus, the relative error is approximately $9.9 \cdot 10^{-9}$, which is less than $5 \cdot 10^{-8}$, so this time we've managed to hang on to 8 significant digits.

1.4 Exercises

In **Exercises 1-4**, perform each of the following tasks for the given number.

- i. Use Matlab's **vpa** command to approximate the given number to 20 digits.
- ii. Place the result of the **vpa** command into n -digit floating point format for the given value of n .
- iii. Use the **vpa** command to determine the relative error.
- iv. Use the definition of significant digits to determine the number of significant digits in your n -digit floating point approximation.

1. $\sqrt{14\,385}$, $n = 5$

2. $\sqrt{23\,888}$, $n = 5$

3. $\ln 888\,475$, $n = 10$

4. $\ln 1\,234\,555$, $n = 10$

In **Exercises 5-8**, perform each of the following tasks for the given expression.

- i. Use the **vpa** command of the *Symbolic Toolbox* to assist in finding n -digit floating point approximations of each root for the given value of n . Calculate the relative error for each and state the number of significant digits for each.
- ii. Use the results of the previous part to hand calculate an n -digit floating point approximation for the given expression.
- iii. Use Matlab's **vpa** command to find

the relative error in approximating the given expression with the n -digit floating point number of part (ii).

- iv. Use the definition of significant digits to determine the number of significant digits in the approximation of part (ii).

5. $\sqrt{8355} - \sqrt{8354}$, $n = 5$

6. $\sqrt{7565} - \sqrt{7564}$, $n = 5$

7. $\sqrt{45619} - \sqrt{45617}$, $n = 10$

8. $\sqrt{387159} - \sqrt{387156}$, $n = 10$

In **Exercises 9-12**, perform each of the following tasks for the given quadratic equation.

- i. Solve the given quadratic by hand. Place your solution in simple radical form.
- ii. Use Matlab's **vpa** command to approximate the radical in your solution to 20 digits, then place the result in 10-digit floating point format.
- iii. Use hand calculations to determine the solution that doesn't cause catastrophic cancellation of digits in 10-digit floating point form. Use the **vpa** command to determine the relative error and the number of significant digits of this result.
- iv. Follow the lead of **Example 2** in the narrative to obtain a 10-digit floating point approximation of the second solution. Then use the **vpa** command to determine the rela-

tive error and the number of significant digits.

9. $x^2 - 4744x + 2 = 0$

10. $x^2 - 5666x + 4 = 0$

11. $x^2 - 388x + 2 = 0$

12. $x^2 - 644x + 1 = 0$

