233 Computational Techniques

Solution Sheet for Tutorial 1

Problem 1.

(i) The set of all positive integers is countable. **Solution** $f : \mathbb{N} \to \mathbb{Z}^+$ with $f(n) = n+1$.

(ii) The set of all integers is countable. **Solution** $f : \mathbb{N} \to \mathbb{Z}$ with $f(n) = n/2$ if n is even and $f(n) = -(n+1)/2$ if n is odd.

(iii) We can show by induction on n that the set of ordered lists of natural numbers that have length $n \geq 1$ is countable. **Solution** The base case $k = 1$ is trivial. For the inductive step, regard every list of length $k+1$ as a natural number followed by a list of length k and mimic the proof of countability of rational numbers. as follows.

Inductive step: Assume that the ordered lists of natural numbers of length n is countable and is given by:

$$a_0, a_1, a_2, a_3, \ldots.$$

where each a_i is an ordered list of natural numbers of length n.

Every ordered list of natural numbers of length $n+1$ is of the form $j : a_i$ for some $j \in \mathbb{N}$ and a list a_i for some $i \in \mathbb{N}$ by the inductive step.

Thus, the set of all ordered lists of natural numbers of length $n+1$ is included in the two dimensional array:

$$\begin{array}{cccccccc}
0:a_0 & 0:a_1 & 0:a_2 & 0:a_3 & 0:a_4 & \ldots & 0:a_i & \ldots \\
1:a_0 & 1:a_1 & 1:a_2 & 1:a_3 & 1:a_4 & \ldots & 1:a_i & \ldots \\
2:a_0 & 2:a_1 & 2:a_2 & 2:a_3 & 2:a_4 & \ldots & 2:a_i & \ldots \\
3:a_0 & 3:a_1 & 3:a_2 & 3:a_3 & 3:a_4 & \ldots & 3:a_i & \ldots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ldots
\end{array}$$

The inductive step now follows since these lists are countable by the diagonal method (used in the proof of countability of rational numbers in the notes).

(iv) We can then use (iii) to show that the set of all finite ordered lists of natural numbers is countable. **Solution**

By part (iii), we know that the ordered list of natural numbers of length n is countable and is given, say, by $b_{n0}, b_{n1}, b_{n2}, b_{n3} \ldots$.

Thus, the set of all ordered finite lists of natural numbers is included in the two dimensional array:

$$\begin{array}{ccccc}
a_{10} & a_{11} & a_{12} & a_{13} & \ldots \\
a_{20} & a_{21} & a_{22} & a_{23} & \ldots \\
a_{30} & a_{31} & a_{32} & a_{33} & \ldots \\
\vdots & \vdots & \vdots & \vdots & \ldots
\end{array}$$

By the diagonal method again, this set of countable.

(v) Any non-finite subset of a countable set is countable. **Solution** Assume $A \subset S$ is a non-finite subset of a countable set S. We need to construct an enumeration g of A given an enumeration f of S. We provide two solutions, an inductive solution and a more simple solution. **Inductive solution:** From an enumeration $f : \mathbb{N} \to S$ obtain an enumeration $g : \mathbb{N} \to A$ by an inductive definition. Let $g(0) = f(n)$ where n is the least natural number for which $f(n) \in A$. Now assume $g(k)$ has been defined for $k \geq 0$ and define $g(k+1) = f(m)$ where m is the least integer such that $f(m) \in A$ and $f(m) \neq g(t)$ for $t \leq k$. Note that if f is one to one, then g would also be one to one, which is the advantage of this to the more simple solution explained next. **Simple solution:** Pick any fixed element $c \in A$. For all $n \in \mathbb{N}$, let $g(n) = f(n)$ if $f(n) \in A$, otherwise let $g(n) = c$.

(vi) If S is countable then S^n, i.e., the collection of all n-tuples of elements of S, is countable. **Solution** As in (iii).

(vii) From (vi), we can deduce that the set of integer polynomials (i.e., polynomials with integer co-efficients) is countable. **Solution** Combine (vi) and the analogue of (iv) for a general countable set.

(viii) From (vii) it follows that the set of roots of integer polynomials, the so-called algebraic numbers, is also countable. **Solution** For each integer polynomial of degree n, there are at most n (real or complex) roots (some of which may be repeated roots).

Problem 2.
Accuracy of floating point operations. Assuming rounded binary arithmetic, determine *(a)* the supremum of values $\delta > 0$ such that $fl(1+\delta) = 1$, and *(b)* the number of significant decimal digits, for both single and double precision.
Solution
(a) In rounded SP arithmetic $1+\delta$ is precisely the midpoint between 1, which is represented as $1 = 0.1 \times 2^1$ and the smallest floating point number bigger than 1, namely

$$n = 0.100\ldots\ldots001 \times 2^1,$$

with 21 zeros in the mantissa. Hence $\delta = (n-1)/2 = 2^{-23} \approx 1.2 \times 10^{-7}$. For DP, we get $\delta = 2^{-52} \approx 2.2 \times 10^{-16}$.
Remark: We show below that δ is the same as the machine accuracy or the relative accuracy a_R of a floating point format, which is defined to be the maximum relative error of the numbers in its representable range: $a_R = \sup\{e_R(n) : n \text{ in the representable range}\}$. For rounded arithmetic in SP and DP these are computed as follows. Since we are dealing with relative errors we only need to consider numbers of the form $n = 0.m_1m_2\ldots m_t$ (with $t = 23$ for SP and $t = 52$ for DP), where $m_1 \neq 0$. The supremum of the relative error

$\frac{\overline{n}-n}{n}$ occurs when the absolute error $\overline{n}-n$ is maximum, i.e. $\overline{n}-n = 2^{-t}/2$ and when n is minimum in the denominator i.e., $n = 0.1 = 1/2$. Hence, we have $a_R = 2^{-t}$.
Thus the value of δ computed above is the machine accuracy a_R, which we have just computed. This is because, as our computation for a_R showed, the machine accuracy is precisely the supremum relative error for a number n with $fl(n) = 1$.
(b) There are different definitions for the notion of significant digits. A common and reasonable definition is the following. If a number has relative error e_R, then it has $\lfloor -\log e_R \rfloor$ significant digits in the decimal representation starting with the first non-zero digit. Thus, if a number has relative error $e_R = 10^{-m}$, then it has m significant decimal digits and any further digit can be dispensed with. For example, if 18.774 has relative error 10^{-3} then it has only 3 significant digits and can therefore be replaced by 18.7.
Thus, the number of significant digits in SP or DP will be $\lfloor -\log a_R \rfloor$, since $a_R = 2^{-t}$ is the supremum of the relative error.
For SP ($a_R = 2^{-23}$) we get

$$\lfloor -\log 2^{-23} \rfloor \approx \lfloor 6.923 \rfloor = 6,$$

and for DP ($a_R = 2^{-52}$) we get

$$\lfloor -\log 2^{-52} \rfloor \approx \lfloor 15.653 \rfloor = 15.$$

However, since a_R itself is the supremum of the possible relative error in the representable range, then for most computations in SP we will have $6.923 \approx 7$ significant digits and for DP we will have $15.653 \approx 16$ significant digits.

Problem 3.
Error propagation in the arithmetic operations. Analyse the propagation of the relative error in each of the four arithmetic operations by comparing the relative error of the result with the sum of the relative errors of the operands, assuming that the operations themselves do not introduce additional loss of accuracy. (Treat multiplication first, it's the easiest!)
Solution
Suppose we want to compute $x \circ y$, where $\circ$ is any binary operation. If the machine representations are $x + \Delta x$ and $y + \Delta y$, we end up computing $(x + \Delta x) \circ (y + \Delta y)$. So we make an absolute error of $e_A = |(x + \Delta x) \circ (y + \Delta y) - x \circ y|$, and a relative error of

$$e_R = \frac{e_A}{|x \circ y|} = \frac{|(x + \Delta x) \circ (y + \Delta y) - x \circ y|}{|x \circ y|} .$$

So for *multiplication* ($\circ = \times$), the absolute error $e_A = |x\Delta y + y\Delta x + \Delta x \Delta y|$ is bounded by $|x||\Delta y| + |y||\Delta x| + |\Delta x||\Delta y|$, and so the relative error is

$$e_R \leq \frac{|x||\Delta y| + |y||\Delta x| + |\Delta x||\Delta y|}{|x||y|} = \frac{|\Delta y|}{|y|} + \frac{|\Delta x|}{|x|} + \frac{|\Delta x||\Delta y|}{|x||y|} .$$

So, apart from the last term which is so small that it can be neglected, the relative error of the product is bounded by the sum of the relative errors of x and y.
For *division* (x/y), the absolute error is

$$e_A = \left|\frac{x+\Delta x}{y+\Delta y} - \frac{x}{y}\right| = \frac{|y\Delta x - x\Delta y|}{|y||y+\Delta y|} \leq \frac{|y||\Delta x| + |x||\Delta y|}{|y||y+\Delta y|}$$

and the relative error is

$$\begin{aligned} e_R &= e_A\frac{|y|}{|x|} \leq \frac{|y||\Delta x| + |x||\Delta y|}{|y||y+\Delta y|}\frac{|y|}{|x|} = \frac{|y|}{|y+\Delta y|}\frac{|y||\Delta x| + |x||\Delta y|}{|x||y|} \\ &= \frac{|y|}{|y+\Delta y|}\left(\frac{|\Delta x|}{|x|} + \frac{|\Delta y|}{|y|}\right) . \end{aligned}$$

Here the term in front of the bracket is close to 1 as $|\Delta y| \ll |y|$, and so again the relative error of a quotient is essentially bounded by the sum of the relative errors of numerator and denominator.
For *addition*, the absolute error $e_A = |\Delta x + \Delta y|$ is at most the sum of the absolute errors of x and y, $e_A \leq |\Delta x| + |\Delta y|$. So the relative error of the sum is

$$\begin{aligned} e_R &= \frac{|\Delta x + \Delta y|}{|x+y|} \leq \frac{|\Delta x|}{|x+y|} + \frac{|\Delta y|}{|x+y|} = \frac{|x|}{|x+y|}\frac{|\Delta x|}{|x|} + \frac{|y|}{|x+y|}\frac{|\Delta y|}{|y|} \\ &\leq \frac{\max\{|x|,|y|\}}{|x+y|}\left(\frac{|\Delta x|}{|x|} + \frac{|\Delta y|}{|y|}\right) . \end{aligned}$$

Here the term in the bracket is the sum of the relative errors of x and y. If x and y have the same sign, the term in front is at most 1, and so the relative error of the sum is bounded by the sum of the relative errors of x and y. But if x and y have *opposite* sign, $|x+y|$ can be much smaller than the larger of $|x|$ and $|y|$ and so the factor in front can become very large!
Moral: Subtraction can be dangerous! Subtraction of real numbers with the same sign (or addition of numbers with opposite signs) can lead to heavy loss of accuracy. Try to avoid subtracting real numbers that are almost equal, or if you really can't avoid it, do it in double precision!