Preserving Significant Digits

1 A silly, but telling, example

Consider three functions $\mathsf{f}(\mathsf{x})$, $\mathsf{g}(\mathsf{x})$, and $\mathsf{h}(\mathsf{x})$ defined as

$$\mathsf{f}(\mathsf{x}) = \mathsf{x} + 1 \tag{1}$$
$$\mathsf{g}(\mathsf{x}) = \mathsf{x} \tag{2}$$
$$\mathsf{h}(\mathsf{x}) = \mathsf{f}(\mathsf{x}) - \mathsf{g}(\mathsf{x}) \tag{3}$$

for $0 \leqslant \mathsf{x} < +\infty$. Using our keens eyes and razor-sharp intellects, we deduce that $\mathsf{h}(\mathsf{x}) = 1$ for all x. Nonetheless, plotting $\mathsf{h}(\mathsf{x})$ over a large range of x produces some disturbing results; see Figure 2.

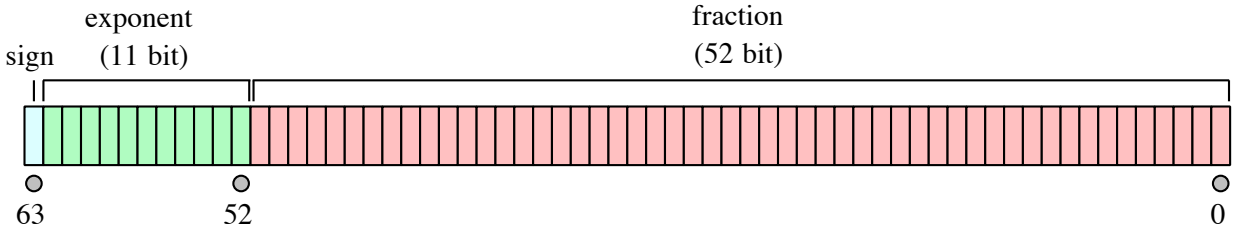
It appears that something unexpected occurs around $\mathsf{x} = 10^{16}$. Let's zoom in on this part of the graph and see what we can see. Figure 3 reveals the schizophrenic behavior of $\mathsf{h}(\mathsf{x})$. On the left it is 1, on the right it is 0, and in between it can not seem to make up its mind.

We know why $\mathsf{h}(\mathsf{x}) = 1$ on the left – that is what it should be, but two questions remain:

- What is going on around $\mathsf{x} = 10^{16}$? It seems to be bouncing between 0 and 2.
- Why is $\mathsf{h}(\mathsf{x}) = 0$ for $\mathsf{x} > 10^{17}$?

Recall, that a double precision floating point number uses 8 consecutive bytes; that is, 64 bits. The representation has a *sign bit*, 11 bits for the *exponent*, and 52 bits for the *fraction*, as shown in Figure 1.

Figure 1: The IEEE 754 double precision binary floating-point format which is used in most civil engineering applications. This figure was extracted from `en.wikipedia.org/wiki/Double_precision`.



Consider the *IEEE 754 double precision floating point number* representation[1] of number 10^{16}:

$$10,000,000,000,000,000 = 0\ 10000110100\ 0001110000110111100100110111111000001000000000000000 \tag{4}$$

[1] I used the online calculator at `http://babbage.cs.qc.edu/IEEE-754/Decimal.html` to generate theses specific numeric examples.

To reconstruct numbers stored in this format we use

$$\boxed{(-1)^{\text{sign}} \times 2^{\text{exponent}-1023} \times 1.\text{fraction}} \tag{5}$$

Lets parse (4) and verify the representation

sign bit	0
exponent	$10000110100_\text{b} = 1076$
fraction	0001110000 1101111001 0011011111 1000001000 0000000000 00_b
place	1234567890 1234567890 1234567890 1234567890 1234567890 12

Substituting these pieces of (4) into (5) we have

$$\begin{aligned}
&(-1)^0 \times 2^{1076-1023} \times 1.0001110000110111100100110111111000001000000000000000 \\
&= 1 \times 2^{53} \times \left(1 + 2^{-4} + 2^{-5} + 2^{-6} + 2^{-11} + 2^{-12} + \ldots + 2^{-37}\right) \\
&= 2^{53} + 2^{53-4} + 2^{53-5} + 2^{53-6} + 2^{53-11} + 2^{53-12} + \ldots + 2^{53-37} \\
&= 2^{53} + 2^{49} + 2^{48} + 2^{47} + 2^{42} + 2^{41} + \ldots + 2^{16} \\
&= 10,000,000,000,000,000
\end{aligned} \tag{6}$$

What is the next larger number, above 10^{16}, that can be represented in the IEEE 754 format? To answer this we change the right-most bit of the fraction-part from a 0 to a 1:

sign bit	0
exponent	$10000110100_\text{b} = 1076$
fraction	0001110000 1101111001 0011011111 1000001000 0000000000 01_b
place	1234567890 1234567890 1234567890 1234567890 1234567890 12

Substituting these pieces into (5) we have

$$\begin{aligned}
&(-1)^0 \times 2^{1076-1023} \times 1.0001110000110111100100110111111000001000000000000001 \\
&= 1 \times 2^{53} \times \left(1 + 2^{-4} + 2^{-5} + 2^{-6} + 2^{-11} + 2^{-12} + \ldots + 2^{-37} \ldots + 2^{-52}\right) \\
&= 2^{53} + 2^{53-4} + 2^{53-5} + 2^{53-6} + 2^{53-11} + 2^{53-12} + \ldots + 2^{53-37} + \ldots + 2^{53-52} \\
&= 2^{53} + 2^{49} + 2^{48} + 2^{47} + 2^{42} + 2^{41} + \ldots + 2^{16} + \ldots + 2 \\
&= 10,000,000,000,000,002
\end{aligned} \tag{7}$$

There is a gap of 2 in the IEEE 754 format between $(10,000,000,000,000,000)$ and $(10,000,000,000,000,00$
Thus, the representation for $(10,000,000,000,000,000)$ and $(10,000,000,000,000,001)$ are identical in IEEE 754 format, and

$$\begin{aligned}
&(10,000,000,000,000,001) - (10,000,000,000,000,000) \\
= \;& 0
\end{aligned} \tag{8}$$

while

$$\begin{aligned}
&(10,000,000,000,000,002) - (10,000,000,000,000,000) \\
= \;& (10,000,000,000,000,002) - (10,000,000,000,000,001) \\
= \;& 2
\end{aligned} \tag{9}$$

2 A more challenging example

Consider the function

$$f(x) = \sqrt{x+1} - \sqrt{x} \tag{10}$$

Figure 4 shows what appears to be a smooth function that asymptotically approaches 0, which is what we would expect. Zooming in on the range between $x = 10^{13}$ and 10^{17} tells a different story; see Figure 5. Are we again suffering from the *slings and arrows of outrageous fortune*? No, it is just the consequences of the truncation error in the IEEE 754 format.

Do we have to simply live with the noise? No. We have to outsmart the noise.

$$\begin{aligned}
f(x) &= \sqrt{x+1} - \sqrt{x} && (11)\\
&= \left(\sqrt{x+1} - \sqrt{x}\right)\left(\frac{\sqrt{x+1} + \sqrt{x}}{\sqrt{x+1} + \sqrt{x}}\right) && (12)\\
&= \frac{\left(\sqrt{x+1} - \sqrt{x}\right)\left(\sqrt{x+1} + \sqrt{x}\right)}{\sqrt{x+1} + \sqrt{x}} && (13)\\
&= \frac{\left(\sqrt{x+1}\right)^2 - \left(\sqrt{x}\right)^2}{\sqrt{x+1} + \sqrt{x}} && (14)\\
&= \frac{(x+1) - (x)}{\sqrt{x+1} + \sqrt{x}} && (15)\\
&= \frac{1}{\sqrt{x+1} + \sqrt{x}} && (16)
\end{aligned}$$

Thus, we let

$$g(x) = \frac{1}{\sqrt{x+1} + \sqrt{x}} \tag{17}$$

replace $f(x)$, and we slay the dragon.

3 The "take home" message

Why are $h(x)$ in (3), and $f(x)$ in (10), so badly behaved for large values of x? Because both are attempting to evaluate a small difference between very large values of almost identical magnitude.

> Be afraid of subtractions between very large values of similar magnitude if you need to maintain significant digits.

On the bright side, in almost all cases, once you see the problem, a mathematical solution is possible.

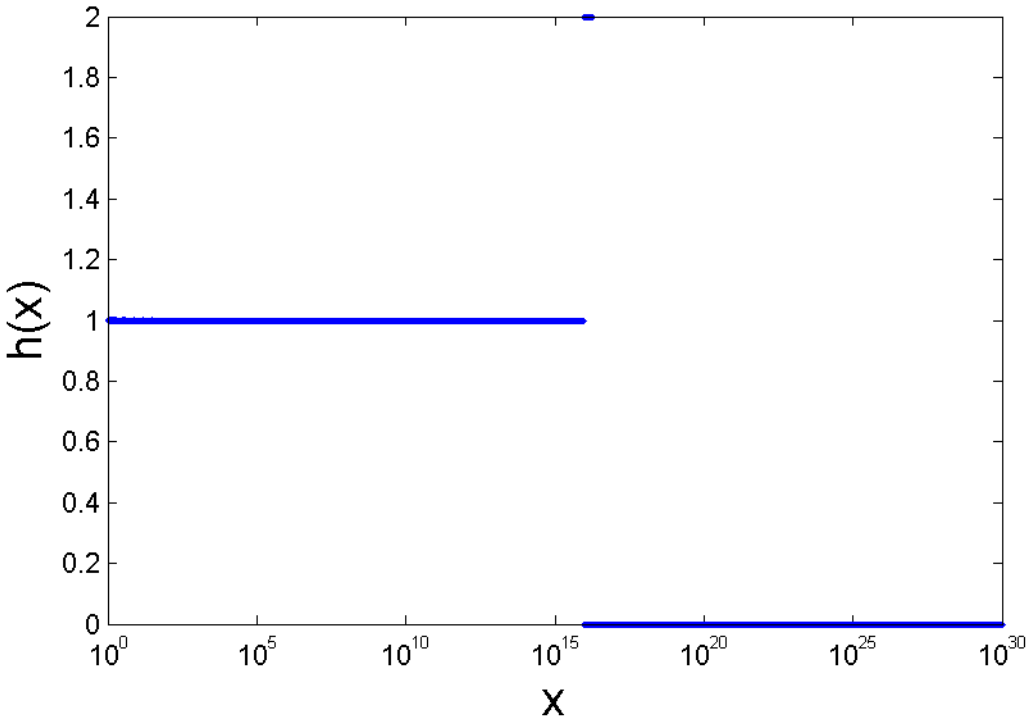


Figure 2: Plotting $h(x) = (x + 1) - (x))$ over a large range of x produces disturbing results.

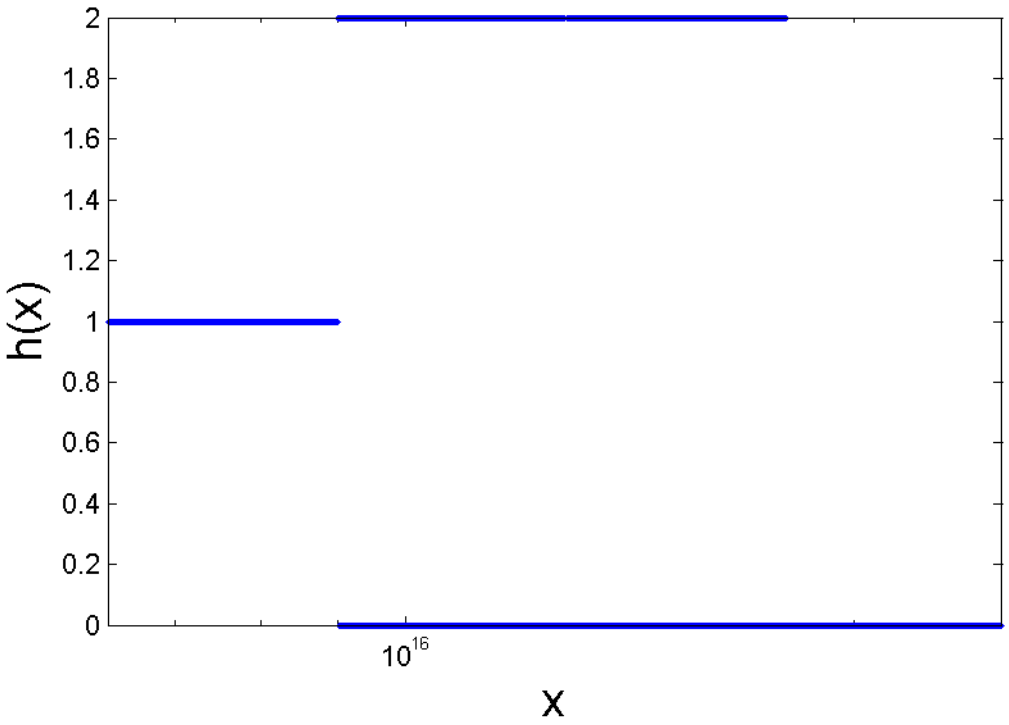


Figure 3: Zooming in around $x = 10^{16}$ reveals the schizophrenic behavior of $h(x)$.

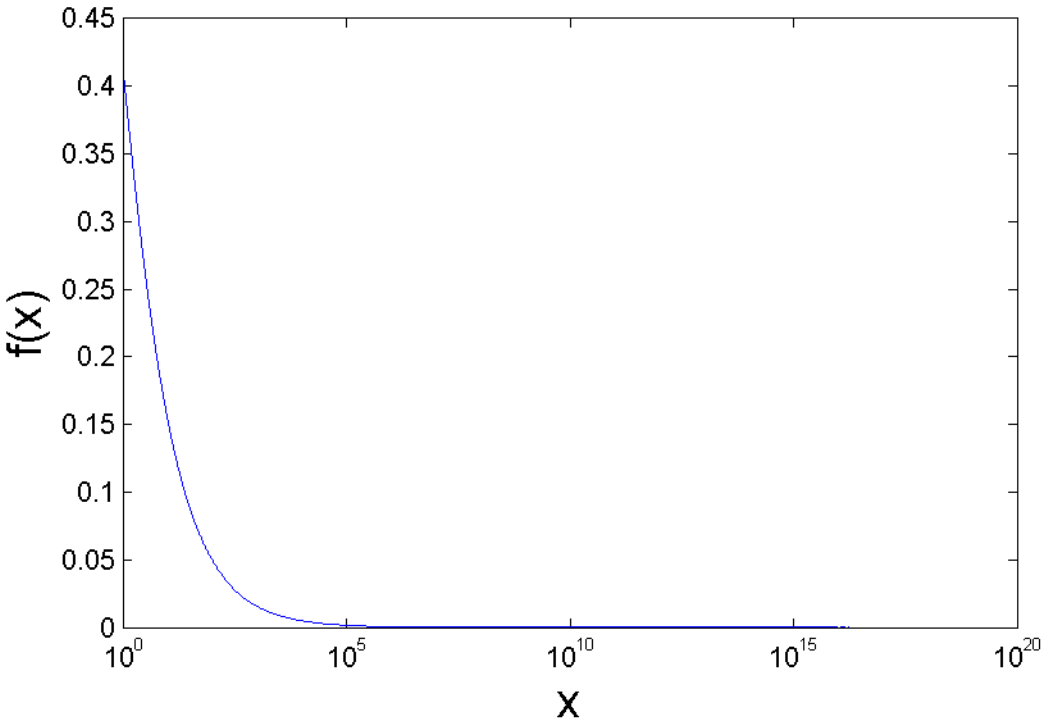


Figure 4: Plotting $f(x) = \sqrt{x+1} - \sqrt{x}$ over a large range of x .

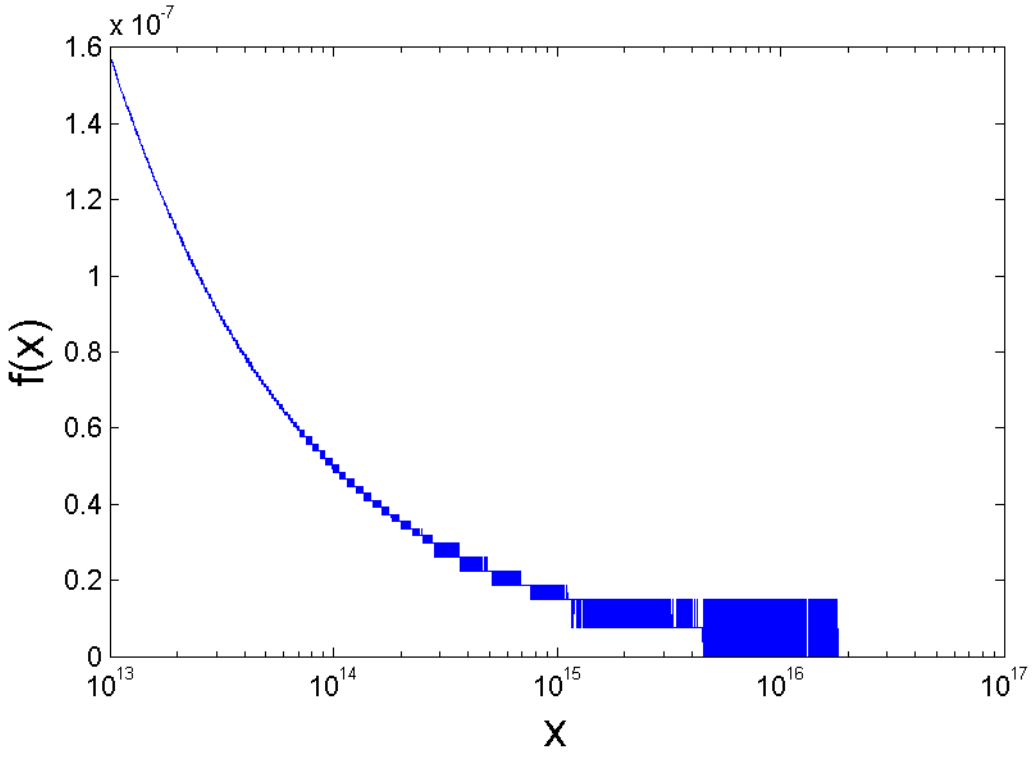


Figure 5: Zooming in between $x = 10^{13}$ and 10^{17} reveals unexpected misbehavior for $f(x) = \sqrt{x+1} - \sqrt{x}$.

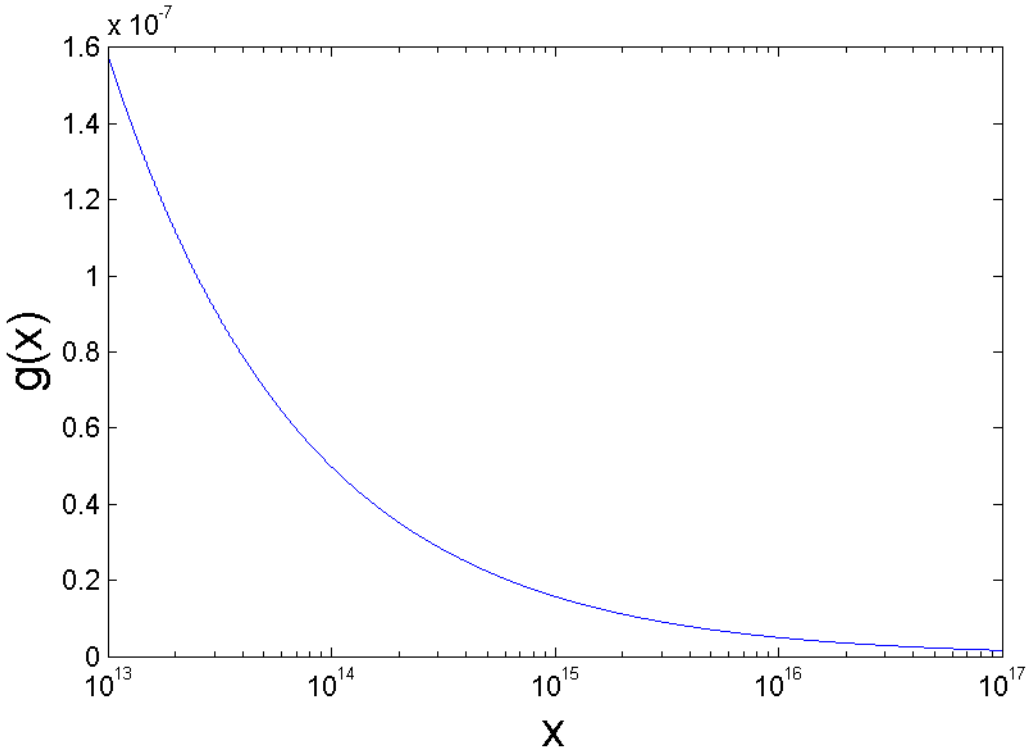


Figure 6: Plotting $g(x) = \frac{1}{\sqrt{x+1}+\sqrt{x}}$ over the previously troublesome range of x .