Rounding Error after Using Guard Digits I

Theorem

Using $p+1$ *digits for* $x-y \Rightarrow$ *relative rounding error* $< 2\epsilon$ *(*ϵ*: machine epsilon)*

Proof:

- Assume $x > y$
- Assume $x = x_0.x_1 \cdots x_{p-1} \times \beta^0$
  The proof is similar if it's not β^0
- If $y = y_0.y_1 \cdots y_{p-1}$ no error
- If $y = 0.y_1 \cdots y_p \Rightarrow$ 1 guard digit, exact $x-y$

Rounding Error after Using Guard Digits II

rounded to a closest number $\Rightarrow$ relative error $\leq \epsilon$

- In general $y = 0.0 \cdots 0 y_{k+1} \cdots y_{k+p}$

  $\bar{y}$: y truncated to $p+1$ digits

  $$|y - \bar{y}| < (\beta - 1)(\beta^{-p-1} + \beta^{-p-2} + \cdots + \beta^{-p-k}) \quad (1)$$

  $-p-1$: we have $p+1$ digits now

  (Think about $p = 3, \beta = 10$, first digit truncated $\leq 9 \times 0.0001 = 9 \times 10^{-4}$)

Rounding Error after Using Guard Digits III

After y is truncated, we need to calculate

$$x - \bar{y}$$

Now we round a number of $p+1$ digits to p:

$$x - \bar{y} + \delta$$

Thus

$$\text{error} \leq 0.\underbrace{0\ldots0}_{p-1 \text{ digits}}(\beta/2)$$

Rounding Error after Using Guard Digits IV

Therefore,

$$|\delta| \leq (\beta/2)\beta^{-p} \tag{2}$$

- The error is

$$(x - y) - (x - \bar{y} + \delta) = \bar{y} - y - \delta$$

Rounding Error after Using Guard Digits V

- **case** 1: if $x - y \geq 1$, from (1) and (2),

$$
\begin{aligned}
& \text{relative error} \\
= \; & \frac{|\bar{y} - y - \delta|}{x - y} \leq \frac{|\bar{y} - y - \delta|}{1} \\
\leq \; & \beta^{-p}[(\beta - 1)(\beta^{-1} + \cdots + \beta^{-k}) + \beta/2] \\
= \; & \beta^{-p}[(\beta - 1)\beta^{-k}(1 + \cdots + \beta^{k-1}) + \beta/2] \\
= \; & \beta^{-p}[(\beta - 1)\beta^{-k}\frac{1 - \beta^{k}}{1 - \beta} + \beta/2] \\
= \; & \beta^{-p}[(1 - \beta^{-k}) + \beta/2] \\
< \; & \beta^{-p}(1 + \beta/2) \leq 2\epsilon
\end{aligned}
$$

Rounding Error after Using Guard Digits VI

- **case** 2: $x - \bar{y} \leq 1$: enough digits to store $x - \bar{y}$ so $\delta = 0$

  The relative error is now

  $$\frac{|\bar{y} - y|}{x - y}$$

  The smallest $x - y$: (smallest x - largest y) is

  $$1.0 - 0.0 \ldots 0\rho \ldots \rho > (\beta - 1)(\beta^{-1} + \cdots + \beta^{-k})$$

Rounding Error after Using Guard Digits VII

k zeros, p ρ's, $\rho = \beta - 1$, from (1),

$$
\begin{aligned}
&\text{relative error} \\
\le\ & \frac{|\bar{y} - y|}{(\beta - 1)(\beta^{-1} + \cdots + \beta^{-k})} \\
<\ & \frac{(\beta - 1)\beta^{-p}(\beta^{-1} + \cdots + \beta^{-k})}{(\beta - 1)(\beta^{-1} + \cdots + \beta^{-k})} = \beta^{-p} < 2\epsilon
\end{aligned}
$$

- **case** 3: $x - y < 1$ but $x - \bar{y} > 1$

  We show that this situation is impossible

Rounding Error after Using Guard Digits VIII

If $x - \bar{y} = 1.\underbrace{0 \cdots 1}_{p} \Rightarrow x - y \geq 1$: a contradiction

Why $x - y$ must be ≥ 1:

$$|y - \bar{y}| < \beta^{-p} = 0.\underbrace{0 \cdots 1}_{p}$$

- Conclusion: adding some guard digits can reduce the error

  Especially when subtracting two nearby numbers

Rounding Error after Using Guard Digits IX

- Cost: the adder is one bit wider (cheap)
  Most modern computers have guard digits