Quote:

"Floating-point numbers are a lot like sandpiles: every time you move one you lose a little sand and pick up a little dirt."

- Kroughton + Plauger

- IEEE f.p. standard
- Floating point arithmetic
- Backwards stability

IEEE floating pt standard:

$$x = 1.\underbrace{d_1 d_2 \cdots d_t}_{\text{mantissa}} \beta^{e}$$ ← exponent; β ← base

β (base), typically 2 or 16 (2 for our purposes), $t = p-1$.

Single precision: $\beta = 2$, $t = 23$, $-126 \le e \le 127$, total size: 32 bits
($2^7 - 1$)

Double precision: $\beta = 2$, $t = 52$ $-1022 \le e \le 1023$ total size: 64 bits.
($2^{10} - 1$)

$\mathbb{F}$: floating pt numbers.

Machine ε ε_m is the smallest x s.t. $x+1 > 1$ in computer arithmetic.

Algorithm to find ε_m

```
e = 1
y = 2
while y > 1
    e = e/2
    y = 1 + e
end
e_m = 2e
```

{converges in 53 iterations. Why?}

Notice: if $x \in \mathbb{R}$, $x = (1.d_1 d_2 \cdots d_t d_{t+1} \cdots) \times \beta^e$

$\exists\ \hat{x} \in \mathbb{F}_\beta$ $\hat{x} = (1.d_1 d_2 \cdots d_t) \times \beta^e$

$\tilde{x} \in \mathbb{F}_\beta$ $\tilde{x} = (1.d_1 d_2 \cdots d_t') \times \beta^e$

$|x - \hat{x}| < d_{t+1} \times \beta^e$ → chopping

$|x - \tilde{x}| < \frac{1}{\beta} d_t \times \beta^e$ → rounding

Another source of error: operations on floating-pt numbers.

exact operations: $+, -, \times, \div$

floating-pt ops: $\oplus, \ominus, \otimes, \oslash$

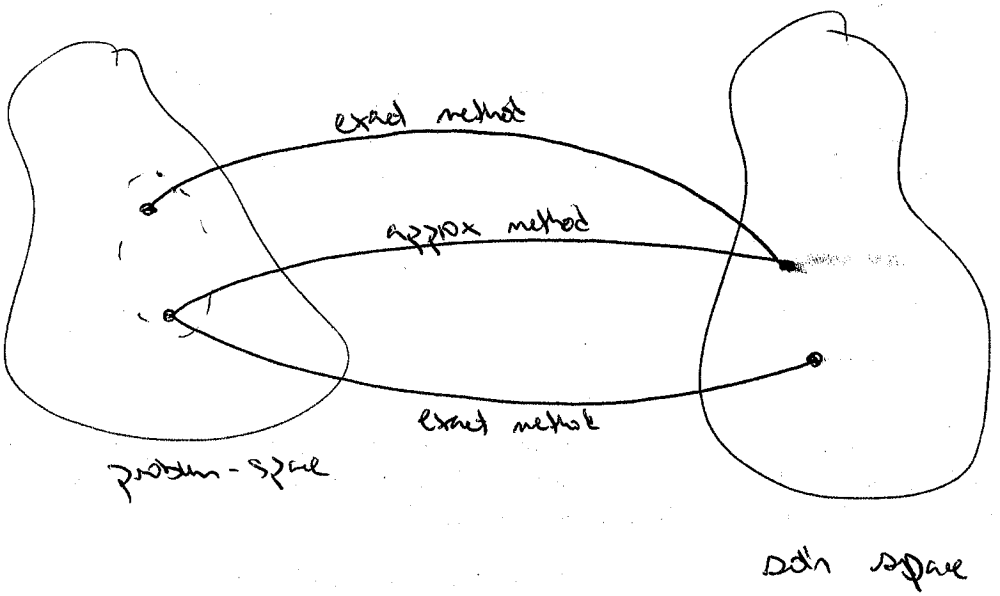
Fundamental axiom of F.P. arithmetic:

For all $x, y \in \mathbb{F}$, $\exists$ δ with $|\delta| \leq \varepsilon_m \leq 2$.

$$x \circledast y = (x * y)(1+\delta), \qquad * = \{+, -, \times, \div\}.$$

$\Rightarrow$ operator perturbations are on the order of ε_m.

Recall: Backwards stability



An algorithm is "backwards stable" if the approximate soln to a given problem is the exact soln of a nearby problem.

~~Next: analysis of complex multiply, back subs~~

Next: rounding error analysis

Stability of Algorithms

$f(x)$ → exact soln of a prob f with data x

$\tilde{f}(x)$ → inexact (numerical soln)

Algorithm is accurate if $\frac{\|f(x) - \tilde{f}(x)\|}{\|f(x)\|} \leq C\varepsilon$ for ε small. (e.g. $\varepsilon = \varepsilon_m$)

Ex: Complex multiply:

EXACT: $x \otimes y = (x \times y)(1 + \sqrt{2}\varepsilon)$ where $|\varepsilon| \leq \varepsilon_m$

$\tilde{f}(x,y)$ $\quad$ $f(x,y)$

$$\frac{|f(x,y) - \tilde{f}(x,y)|}{|f(x,y)|} \leq \underbrace{\sqrt{2}}_{c}\varepsilon_m$$

Backwards stability:

Algorithm is backwards stable if $\tilde{f}(x) = f(\tilde{x})$ for some

inexact soln to prob $\quad$ exact soln to nearby prob

$\tilde{x}$ satisfying

$$\frac{\|x - \tilde{x}\|}{\|x\|} \leq C\varepsilon_m$$

$x \otimes y = x \times (1 + \sqrt{2})\varepsilon_c \, y$, $|\varepsilon_c| < \varepsilon_m$

take $\tilde{x} = x$, $\tilde{y} = (1 + \sqrt{2})\varepsilon_c y$

$\tilde{x} \times \tilde{y} = x \otimes y$

$$\frac{\left\| \begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} \tilde{x} \\ \tilde{y} \end{pmatrix} \right\|}{\left\| \begin{pmatrix} x \\ y \end{pmatrix} \right\|} = \frac{\left\| \begin{pmatrix} 0 \\ (\sqrt{2})\varepsilon_c y \end{pmatrix} \right\|}{\left\| \begin{pmatrix} x \\ y \end{pmatrix} \right\|} = \frac{\sqrt{2}|\varepsilon_c|\|y\|}{\sqrt{x^2 + y^2}} \leq \underbrace{\sqrt{2}}_{c}\varepsilon_m$$