Floating point

Recap: \[1 \leq \xi \leq 2\]

\[217.3 = \left( 1. \underbrace{1 \ 0 \ 1 \ 1}_{\xi = 2} \right) \cdot 2^{\frac{3}{\xi}}\]

\[\left( \underbrace{1 \ \cdots \ \cdots}_{\xi = 2} \right)\]
Unrepresentable numbers?

- Can you think of a somewhat central number that we cannot represent as
  \[ x = (1.\_\_\_\_\_\_\_\_)_2 \cdot 2^{-p}? \]

<table>
<thead>
<tr>
<th>0</th>
<th>→ Pos.0: significant, all zeros + special exp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exp. range: -1023, -1024</td>
<td></td>
</tr>
<tr>
<td>-1024</td>
<td>special meaning: &quot;turn off the learning!&quot;</td>
</tr>
</tbody>
</table>
Demo: Picking apart a floating point number

\[
\hat{\text{exponent}} = 1023 - \text{stored} \\
\text{implicit} \text{ (not stored)}
\]

\[
\begin{align*}
(3)_{10} & = (11)_2 = (1.1) \cdot 2^1 = (1.1) \cdot 2 \\
& = 1023 \cdot 1024
\end{align*}
\]

\[
\begin{align*}
2^{-1022} & = (1) \cdot 2^{-1022} \\
2^{-1023} & = (0.100000 \ldots) \cdot 2^{-1023} \\
0 & = (0.000000 \ldots) \cdot 2^{-1023}
\end{align*}
\]

\text{normal} \quad \text{subnormal}
Subnormal Numbers

- What is the smallest representable number in an FP system with 4 stored bits in the significand and an exponent range of \([-7, 7]\)?

\[
\begin{align*}
(0) \left[ \begin{array}{cccc}
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0
\end{array} \right]_2 & \cdot 2^{-7} = 2^{-7} \\
(0) \left[ \begin{array}{cccc}
1 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 \\
1 & 0 & 0 & 0
\end{array} \right]_2 & \cdot 2^{-8} = 2^{-8} \\
(1) \left[ \begin{array}{cccc}
0 & 0 & 0 & 1 \\
0 & 0 & 0 & 1 \\
0 & 0 & 0 & 1 \\
0 & 0 & 0 & 1
\end{array} \right]_2 & \cdot 2^{-8} = 2^{-11}
\end{align*}
\]

FP assist \(\rightarrow\) work slow w/ subnormal numbers

\(\rightarrow\) super slow
Demo: Density of Floating Point Numbers
Demo: Floating Point vs. Program Logic

"Underflow"

"Subnormals" $\rightarrow$ "Gradual underflow"
Floating Point and Rounding Error

What is the relative error produced by working with floating point numbers?

○ What is smallest floating point number > 1? Assume 4 stored bits in the significand.

\[
\left(1.\underbrace{-\frac{1}{2}}_{\text{ulp: unit in the last place}}\right) \approx 1 + 2^{-y}
\]

machine eps

○ What’s the smallest FP number > 1024 in that same system?

○ Can we give that number a name?

○ What does this say about the relative error incurred in floating point calculations?

rel. error introduced in every fp op. is

\sim \text{machine eps}

○ What’s that same number for double-precision floating point? (52 bits in the significand)

\[2^{-52}\]
Implementing Arithmetic

- How is floating point addition implemented?
  Consider adding \( a = (1.101)_2 \cdot 2^1 \) and \( b = (1.001)_2 \cdot 2^{-1} \) in a system with three bits in the significand.

\[
\begin{align*}
\text{Shift onto same exponent} \\
\hline
(1.101)_2 \cdot 2^1 \\
(\underline{1.001})_2 \cdot 2^{-1}
\end{align*}
\]

\[
(1.101) \quad (1.010) \\
(1.1 \underline{1} \ 1 \ 1) \cdot 2^1 \\
\]
Demo: Floating point and the harmonic series
Problems with FP Addition

- What happens if you subtract two numbers of very similar magnitude? As an example, consider $a = (1.1011)_2 \cdot 2^0$ and $b = (1.1010)_2 \cdot 2^0$. 
Demo: Catastrophic Cancellation
In-class activity: Floating Point 2