Normally sat output is:

\[
\begin{align*}
(+) & \quad 011111111 \\
(-) & \quad 100000000
\end{align*}
\]

Examples:

<table>
<thead>
<tr>
<th>11-bit</th>
<th>3-bit</th>
</tr>
</thead>
<tbody>
<tr>
<td>+1023</td>
<td>+3</td>
</tr>
<tr>
<td>-1024</td>
<td>-4</td>
</tr>
</tbody>
</table>

asymmetric creates a bias causes problems:
- RF circuits
- accumulator in path
- effect is more:

1) signal that saturates frequently
2) " " have a narrow word width
If unacceptable, do symmetries saturation

\[ \text{SAT}_{+1} = - \text{SAT}_{-1} \]

<table>
<thead>
<tr>
<th>11-bit</th>
<th>3-bit</th>
</tr>
</thead>
<tbody>
<tr>
<td>+1023</td>
<td>+3</td>
</tr>
<tr>
<td>-1023</td>
<td>-3</td>
</tr>
</tbody>
</table>

(+) $011111\ldots1$

(-) $100000\ldots01$

Now need to detect max neg. case

\[ \text{in} = (\text{SAT}_{-10}+1) \]

Ex: 6-bit input $\Rightarrow$ 5-bit output

Detect $-16$

\[ \text{in} = 110000 \Rightarrow \text{out} = -15 \]

Latter:

\[ \text{in} = 110001 \Rightarrow \text{out} = -15 \]
Saturnator:

Compressor:

+ more gentle magnitude reduction
- the signal is distorted at a smaller magnitude
- now is much more complex
- may require a large memory

Rounding
Format will determine behavior:
- unsigned
- 2's complement
- sign magnitude
- etc.

IEEE 754
- includes many rounding modes

Four main rounding modes functions:

1. round(·)
   - best general purpose
   - symmetric
   - max error of \( \frac{1}{2} \) LSB

2. fix()
   - truncate toward zero
   - max error 1 LSB
(3) floor()
    round towards $-\infty$
    max error 1 LSB

(4) ceil()
    round towards $+\infty$
    max error 1 LSB

**Hardware Rounding**

A) Truncation

XXX.XXXX

XXX.XX--

- don't even calculate truncated bits
- max err $1$ post-rounded LSB

- $\text{Sign Mag}$

same as $\text{fix}() - \text{Sign Mag}$

\[
\begin{align*}
+5.7 &\rightarrow 5 \\
-5.8 &\rightarrow -5 \\
\end{align*}
\]
• 2's Complement
  \[ \begin{align*}
    +5.7 & \rightarrow 5 \\
    -5.8 & \rightarrow -6
  \end{align*} \]

  \[ \text{floor()} \]

  • Signed
    \[ \text{fix()} \]
    \[ \text{floor()} \]

  B) Rounding method #5
  "Add \( \frac{1}{2} \) LSB and Truncate"

  • Not in Matlab
  • Max error of \( \frac{1}{2} \) LSB

  \[ \begin{align*}
    \text{XXX}.XX & \quad \text{in} \\
    YYYY.YXX & \quad \text{intermediate sum} \\
    YYYY... & \quad \text{out}
  \end{align*} \]

  • Often easy to add in "1"

Ex: [Diagram]

\[ \begin{align*}
  \text{4 x 4 bit} & \rightarrow \text{5-bit}
\end{align*} \]
- **Zero case**

- **Better:**

  - Up signed
    - same as round(c)

  - Sign magnitude
    - same as round(c)

  - 2's complement
    - close to round(c)
3 cases:

1) Input `XXXXX.1000` and positive
   
   → `+∞` \( \text{same as round(}) \)

2) Input `XXXXX.0000` and negative
   
   → `+∞` \( \text{not same as round} \)

3) otherwise
   
   `same as round(}

Ex:

\[
(+) \quad XXXX.0
\]

\[
(+) \quad XXXX.1 \quad \{ \text{same as round} \}
\]

\[
(-) \quad XXXX.0
\]

\[
(-1) \quad XXXX.1 \quad \{ \text{same as round(}) + 1 \}
\]

Ex:

\[
(+) \quad XXXXX.000000
\]

\[
32 \quad \{ (+) \quad XXXXX.000000 \quad \{ \text{same at round(}} \}
\]

\[
32 \quad \{ (-) \quad XXXXX.000000 \}
\]

\[
(-) \quad XXXXX.100000 \quad \{ \text{same as round(}} + 1 \}
\]