Extreme Cases in Floating Point Numbers

Floating point numbers are represented in IEEE formats.
Consider IEEE 754 32-bit format also called Single Precision format or Short real format.

S	Biased Exponent	Mantissa
(1)	(8) Bias value = 127	(23 bits)

- **32 bits** are used to store the **number**.
- **23 bits** are used for the **Mantissa**.
- **8 bits** are used for the Biased **Exponent**.
- **1 bit** used for the **Sign** of the number.
- The **Bias** value is $(\mathbf{127})_{10}$.
- The range is $\pm\mathbf{1 \times 10^{-38}}$ **to** $\pm\mathbf{3 \times 10^{38}}$ approximately.
- It is called as the **Single Precision Format** for Floating-Point Numbers.

For a value, 1.0 the normalized form will be

$\mathbf{(-1)^0 \times 1.0 \times 2^0}$

Here, the True Exponent is 0.

```
If:  TE = 0,     BE = 127    Representation = 0111 1111
If:  TE = 1,     BE = 128    Representation = 1000 0000
If:  TE = 2,     BE = 129    Representation = 1000 0000
…
If:  TE = 127,   BE = 254    Representation = 1111 1110
If:  TE = 128,   BE = 255    Representation = 1111 1111
If:  TE = 129,   BE = 255    Representation = 1111 1111
If:  TE = 130,   BE = 255    Representation = 1111 1111
```

This is because the 8-bit biased exponent cannot hold a value more than 255.

Hence, all cases where the TE = 128 or more, the BE will be represented as 1111 1111.
This indicates as exception (error) called OVERFLOW.
The number is called **NaN** (Not a Number).
It is identified by Exponent being all 1s (1111 1111).
Here, the Mantissa can be anything!

The Extreme case of NaN is Infinity.
It is also an OVERFLOW and hence the Exponent will be 1111 1111.
To differentiate Infinity from NaN, the Mantissa in infinity is 0000 0000.
Hence Infinity is identified as Exponent all 1s and Mantissa all 0s.

Lets look at the opposite picture, where the magnitude of the number is diminishing and reaching 0.

Suppose the number is 0.1
It will be normalized as

$(-1)^0$ x 1.0 x 2^{-1}

Here, the True Exponent is -1.

```
If:  TE = -1,    BE = 126    Representation = 0111 1110
If:  TE = -2,    BE = 125    Representation = 0111 1101
…
If:  TE = -126,  BE = 1      Representation = 0000 0001
If:  TE = -127,  BE = 0      Representation = 0000 0000
If:  TE = -128,  BE = 0      Representation = 0000 0000
If:  TE = -129,  BE = 0      Representation = 0000 0000
```

This is because the 8-bit biased exponent cannot hold a value more than 255.

Hence, all cases where the TE = -127 or less, the BE will be represented as 0000 0000.
This indicates as exception (error) called UNDERFLOW.
The number is called **De-Normal** Number.
It is identified by Exponent being all 0s (0000 0000).
Here, the Mantissa can be anything!

The Extreme case of De-Normal Number is Zero.
It is also an UNDERFLOW and hence the Exponent will be 0000 0000.
To differentiate Zero from De-Normal Number, the Mantissa in Zero is 0000 0000.
Hence Zero is identified as Exponent all 0s and Mantissa all 0s.
This means Zero is represented as all 0s.

Summary

Number	Exception	Exponent	Mantissa
Normal Number	No Error	0 < E < 2555	Anything
NaN	Overflow	1111 1111	Anything
Infinity	Overflow	1111 1111	0000 0000
De-Normal Number	Underflow	0000 0000	Anything
Zero	Underflow	0000 0000	0000 0000