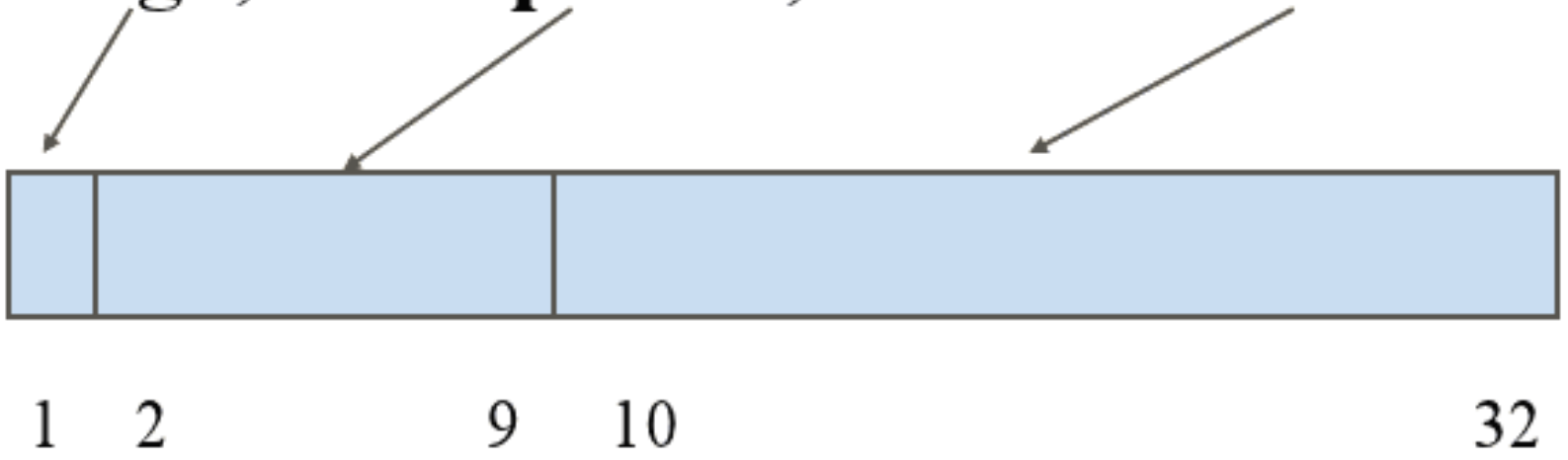
Recitation 4

IEEE Floating Point Representation

Floating point numbers can be represented by binary codes by dividing them into three parts:

the **sign**, the **exponent**, and the **mantissa.**



The second field of the floating point number will be the exponent.

Since we must be able to represent both positive and negative exponents, we will use a convention which uses a value known as a **bias of 127** to determine the representation of the exponent.

- An exponent of 5 is therefore stored as 127 + 5 or 132;
- an exponent of -5 is stored as 127 + (-5) OR 122.

The **biased exponent**, the value actually stored, will range from 0 through 255. This is the range of values that can be represented by 8-bit, unsigned binary numbers.

The mantissa is the set of 0's and 1's to the left of the radix point of the **normalized** (when the digit to the left of the radix point is 1) binary number.

- ex:1.**00101** X 2^3

The mantissa is stored in a 23 bit field,

- Example: find the IEEE FP representation of 40.15625
- 40: 101000
- .15625: .00101
- So 40.15625 in binary is : 101000.00101
- Normalize: 1.0100000101 * 2^5
- Convert the exp to biased: 127 + 5 = 132, in binary : 10000100
- Result : 0 1000100 01000001010...0

Description	Bit representation
Zero	0 0000 000
Smallest pos.	0 0000 001
	0 0000 010
	0 0000 011
	⋮
Largest denorm.	0 0000 111
Smallest norm.	0 0001 000
	0 0001 001
	⋮
	0 0110 110
	0 0110 111
One	0 0111 000
	0 0111 001
	0 0111 010
	⋮
	0 1110 110
Largest norm.	0 1110 111
Infinity	0 1111 000