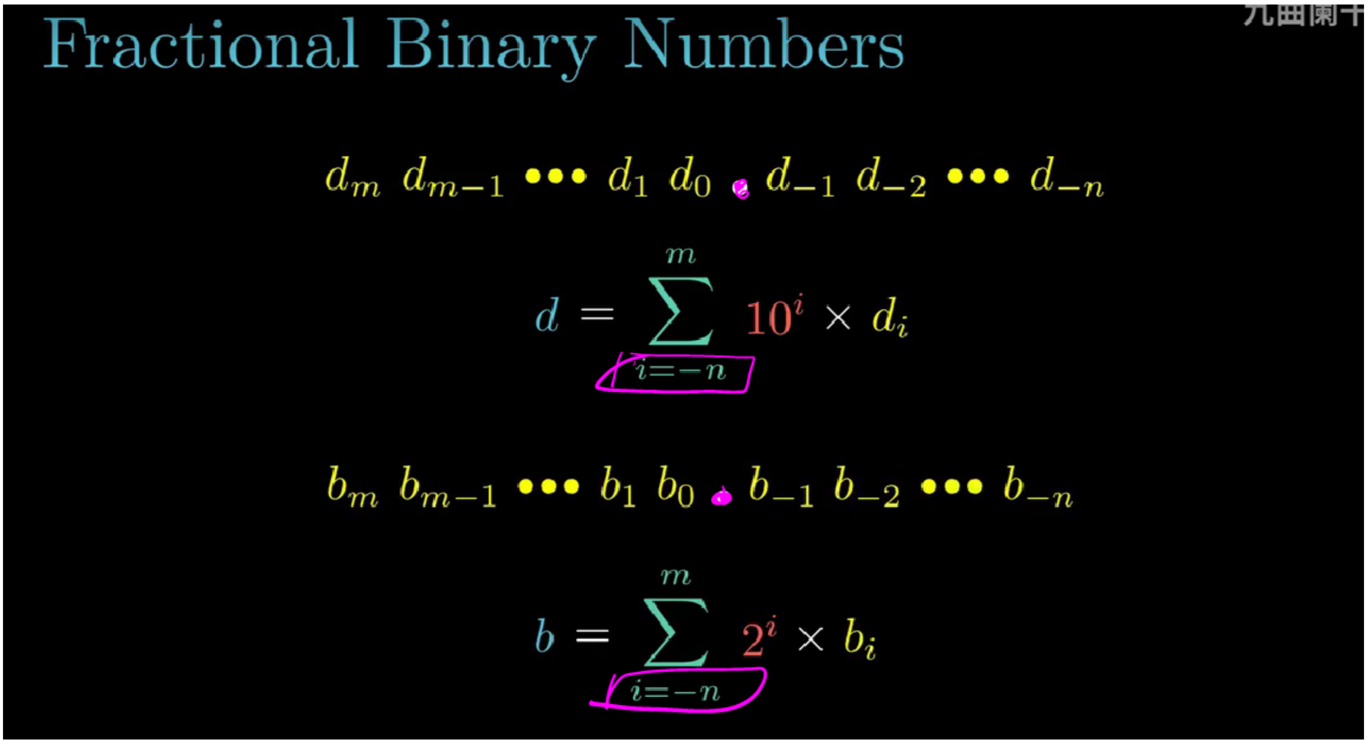


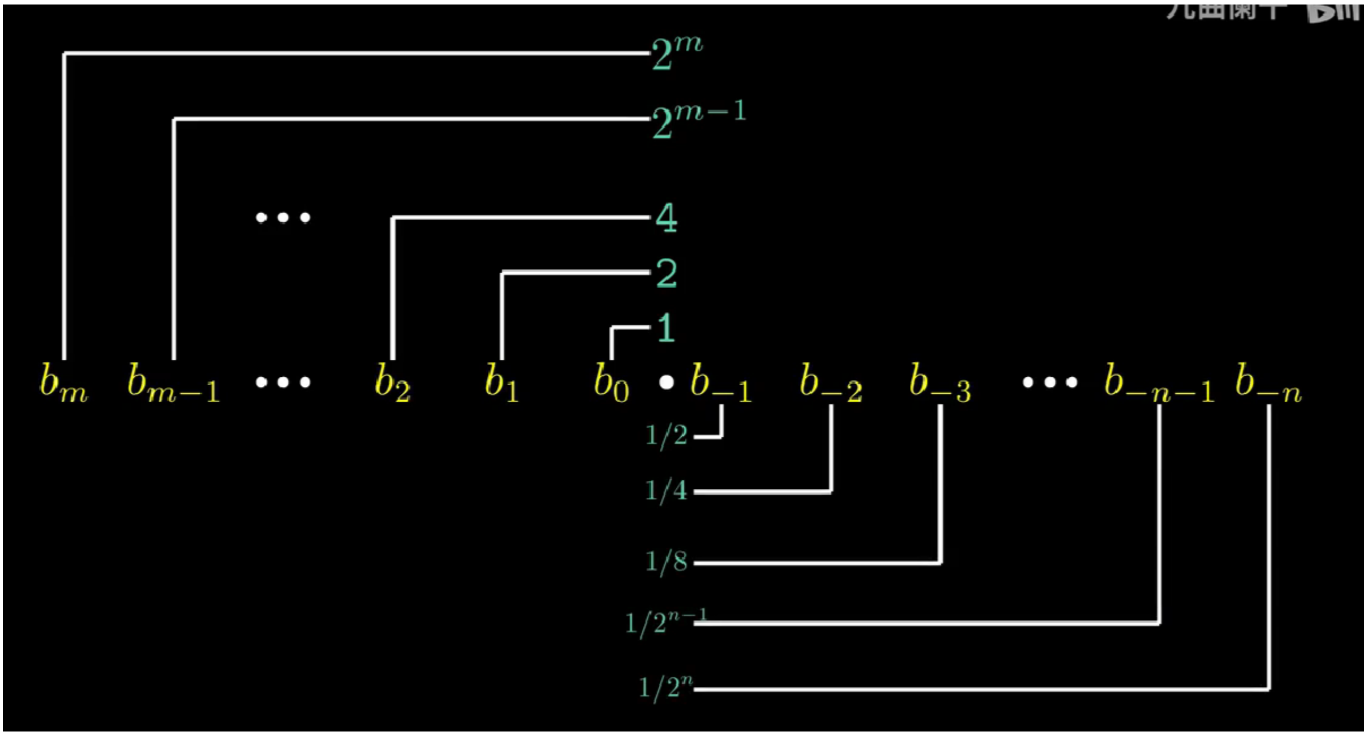
# Fractional Binary Numbers

$$d_m \; d_{m-1} \bullet\bullet\bullet \; d_1 \; d_0 \, . \, d_{-1} \; d_{-2} \bullet\bullet\bullet \; d_{-n}$$

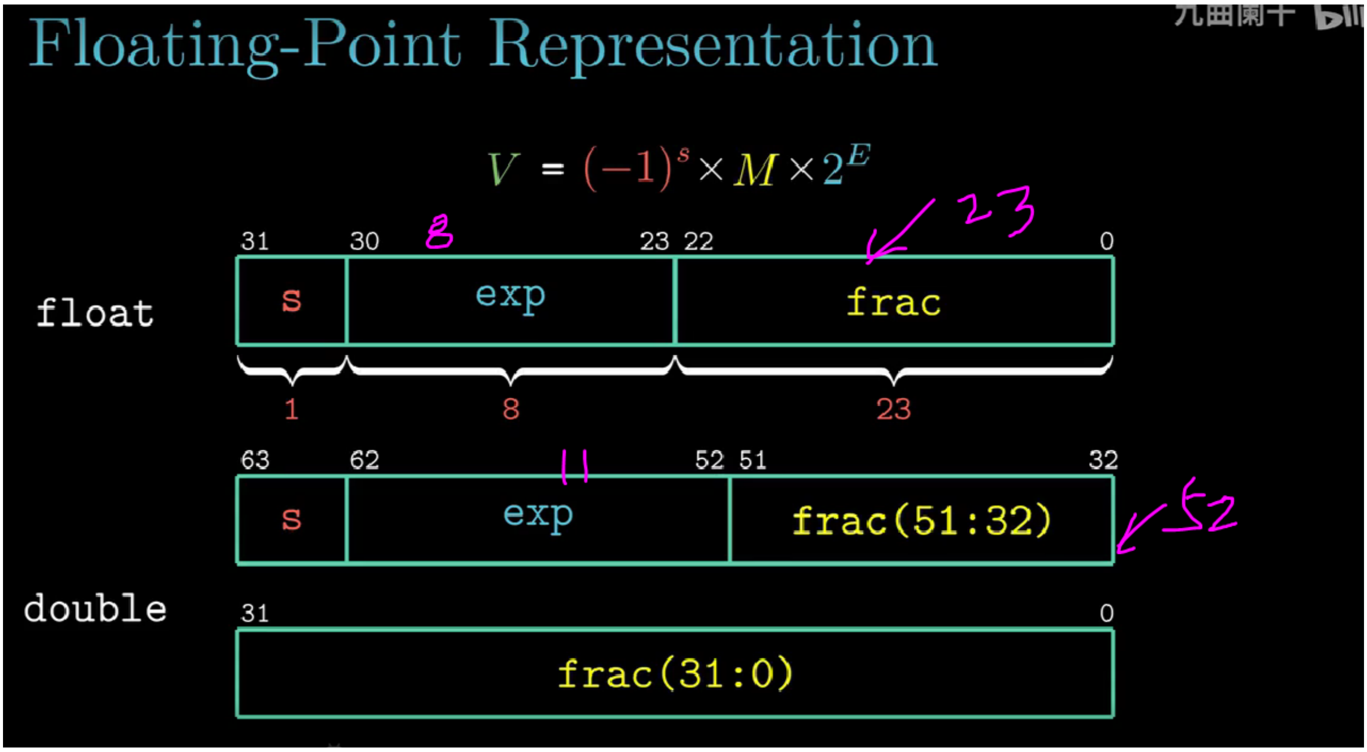
$$d = \sum_{\boxed{i=-n}}^{m} 10^i \times d_i$$

$$b_m \; b_{m-1} \bullet\bullet\bullet \; b_1 \; b_0 \, . \, b_{-1} \; b_{-2} \bullet\bullet\bullet \; b_{-n}$$

$$b = \sum_{\boxed{i=-n}}^{m} 2^i \times b_i$$



$2^m$
$2^{m-1}$
4
2
1
$b_m$ $b_{m-1}$ ••• $b_2$ $b_1$ $b_0$ • $b_{-1}$ $b_{-2}$ $b_{-3}$ ••• $b_{-n-1}$ $b_{-n}$
1/2
1/4
1/8
$1/2^{n-1}$
$1/2^n$



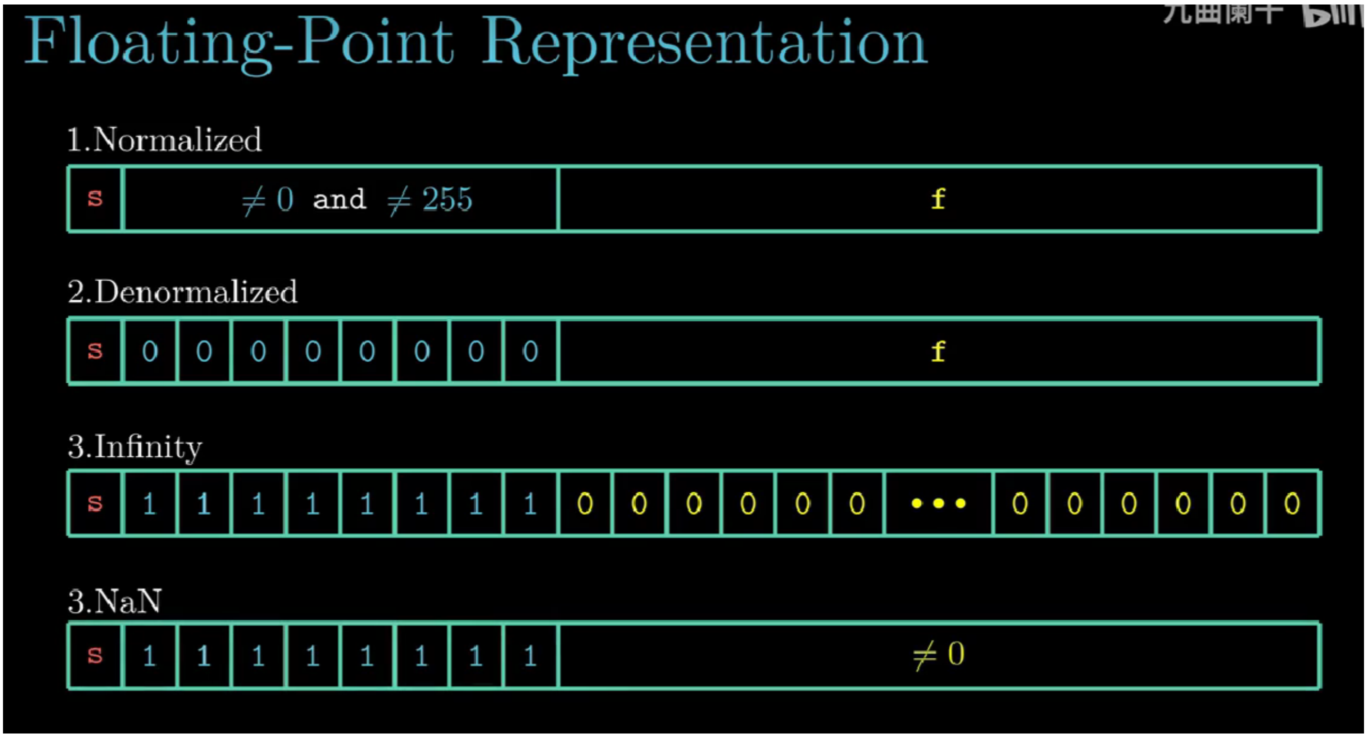
Floating-Point Representation

$$V = (-1)^s \times M \times 2^E$$

floating point types:

- normalized values
- denormalized values
- special values

由阶码字段(exp)来决定是哪一种类型：



Floating-Point Representation

1.Normalized

s | $\neq 0$ and $\neq 255$ | f

2.Denormalized

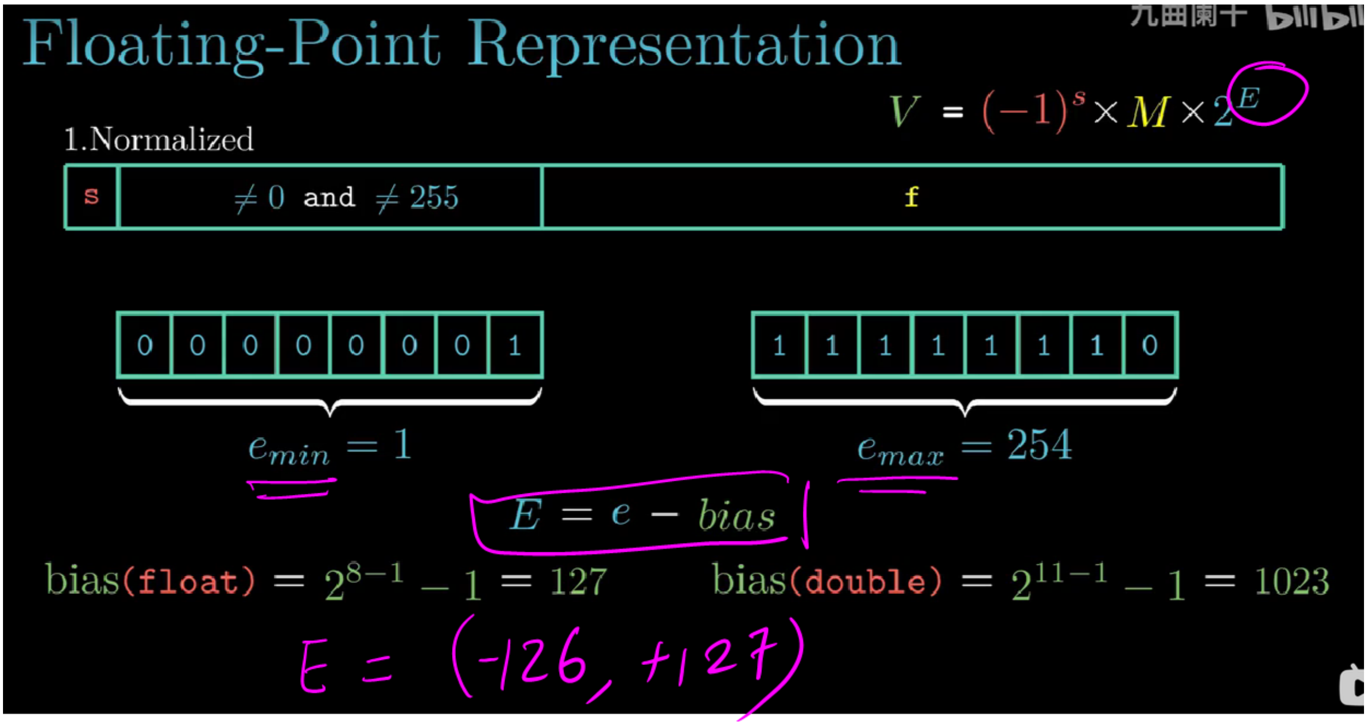
s | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | f

3.Infinity

s | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | ••• | 0 | 0 | 0 | 0 | 0 | 0

3.NaN

s | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | $\neq 0$



# Floating-Point Representation

$V = (-1)^s \times M \times 2^E$

1.Normalized

| s | $\neq 0$ and $\neq 255$ | f |
|---|---|---|

| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
|---|---|---|---|---|---|---|---|

$e_{min} = 1$

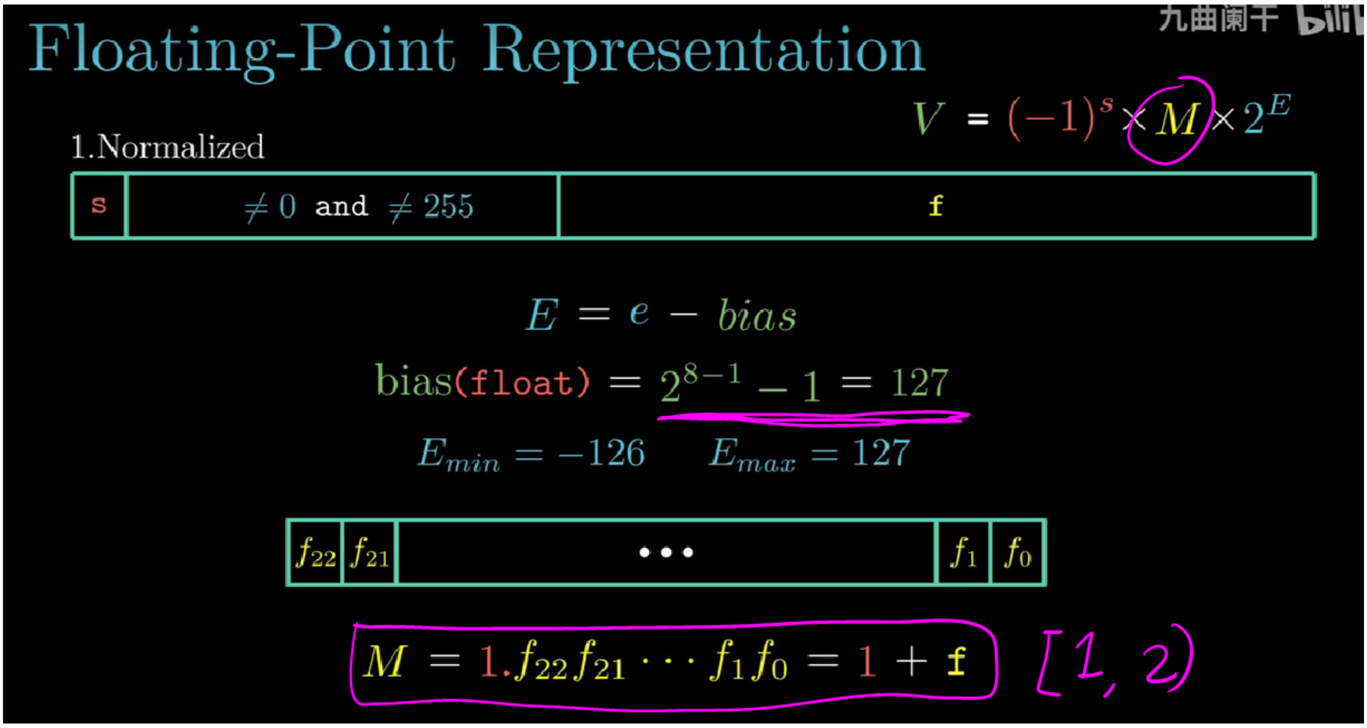
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
|---|---|---|---|---|---|---|---|

$e_{max} = 254$

$E = e - bias$

bias(float) $= 2^{8-1} - 1 = 127$ bias(double) $= 2^{11-1} - 1 = 1023$

$E = (-126, +127)$



# Floating-Point Representation

$V = (-1)^s \times M \times 2^E$

1.Normalized

| s | $\neq 0$ and $\neq 255$ | f |
|---|---|---|

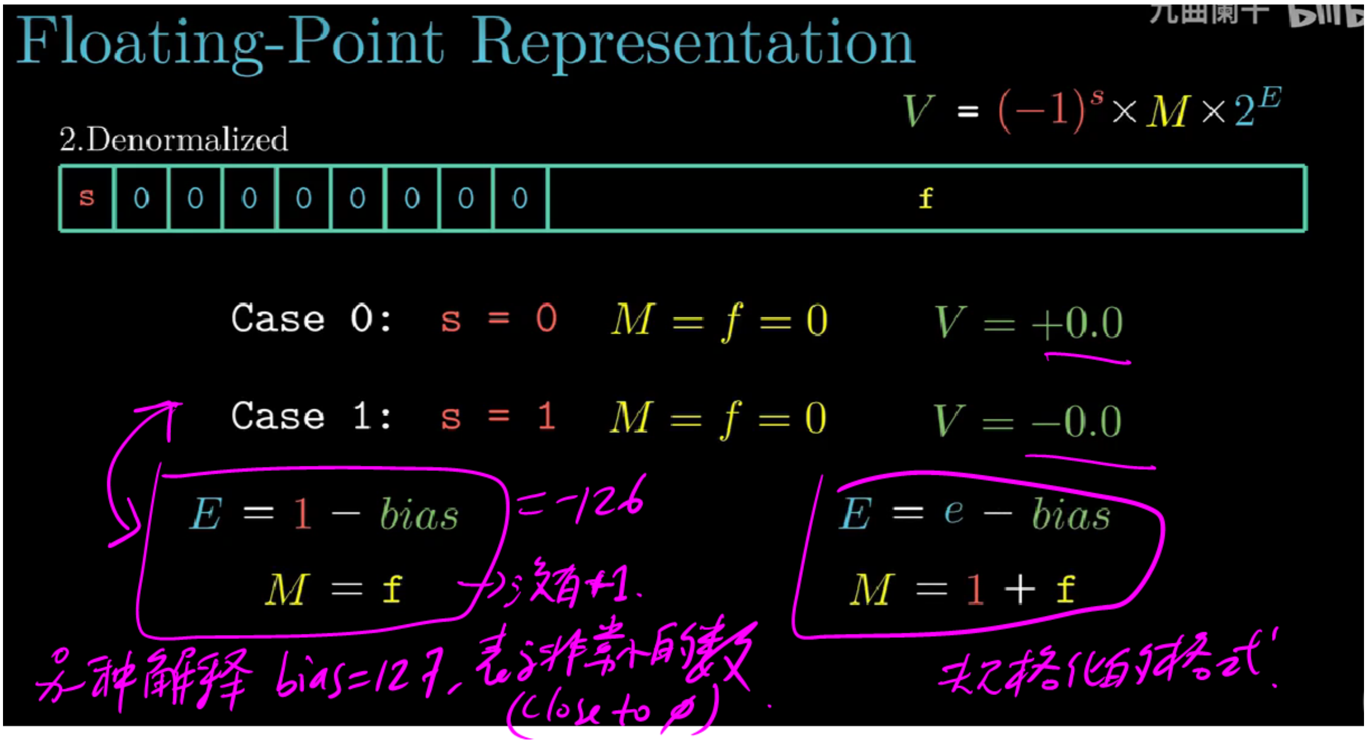
$E = e - bias$

bias(float) $= 2^{8-1} - 1 = 127$

$E_{min} = -126$ $E_{max} = 127$

| $f_{22}$ | $f_{21}$ | • • • | $f_1$ | $f_0$ |
|---|---|---|---|---|

$M = 1.f_{22}f_{21}\cdots f_1 f_0 = 1 + \mathtt{f}$ $[1, 2)$



## Floating-Point Representation

$$V = (-1)^s \times M \times 2^E$$

2.Denormalized

| s | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | f |
|---|---|---|---|---|---|---|---|---|---|

Case 0: s = 0 $M = f = 0$ $V = \underline{+0.0}$

Case 1: s = 1 $M = f = 0$ $V = \underline{-0.0}$

$E = 1 - bias$ =-126

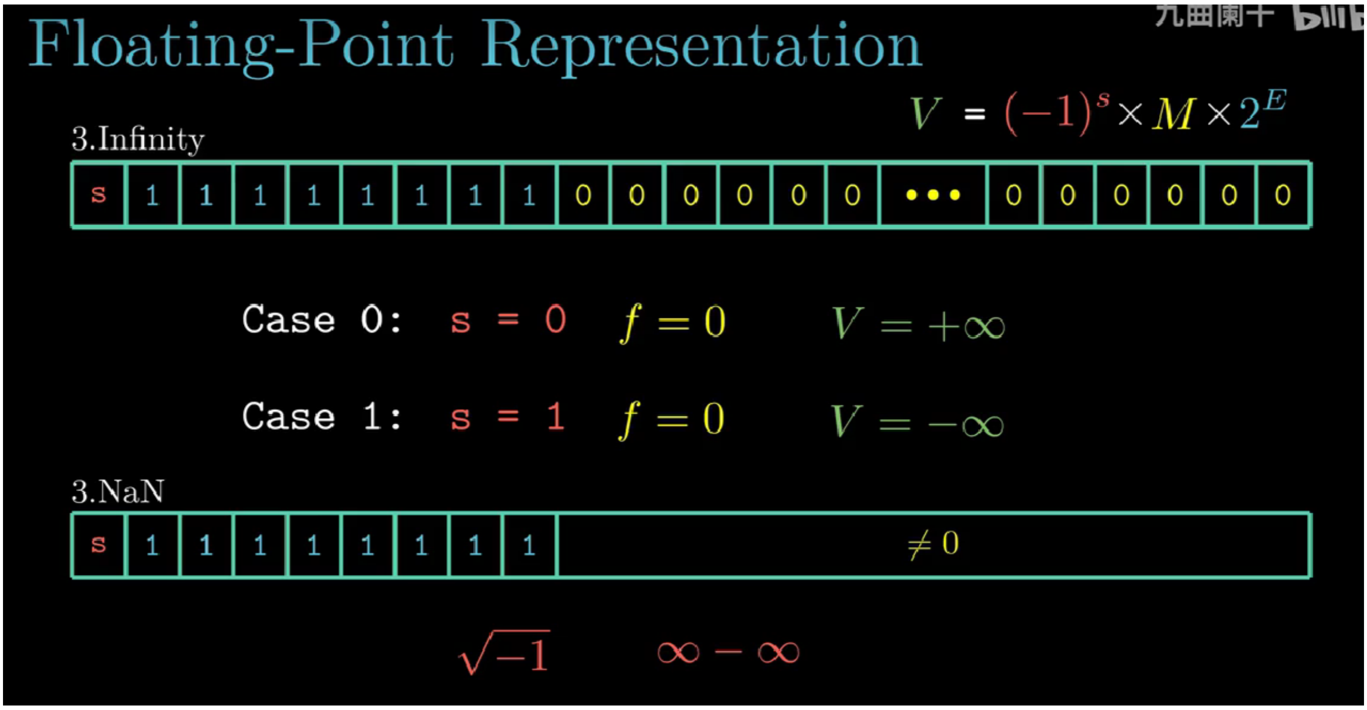
$M = $ f →没有+1.

另种解释 bias=127, 表示非常小的数 (close to 0).

$E = e - bias$

$M = 1 + $ f

规格化的格式'.



## Floating-Point Representation

$$V = (-1)^s \times M \times 2^E$$

3.Infinity

| s | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | ••• | 0 | 0 | 0 | 0 | 0 | 0 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|

Case 0: s = 0 $f = 0$ $V = +\infty$

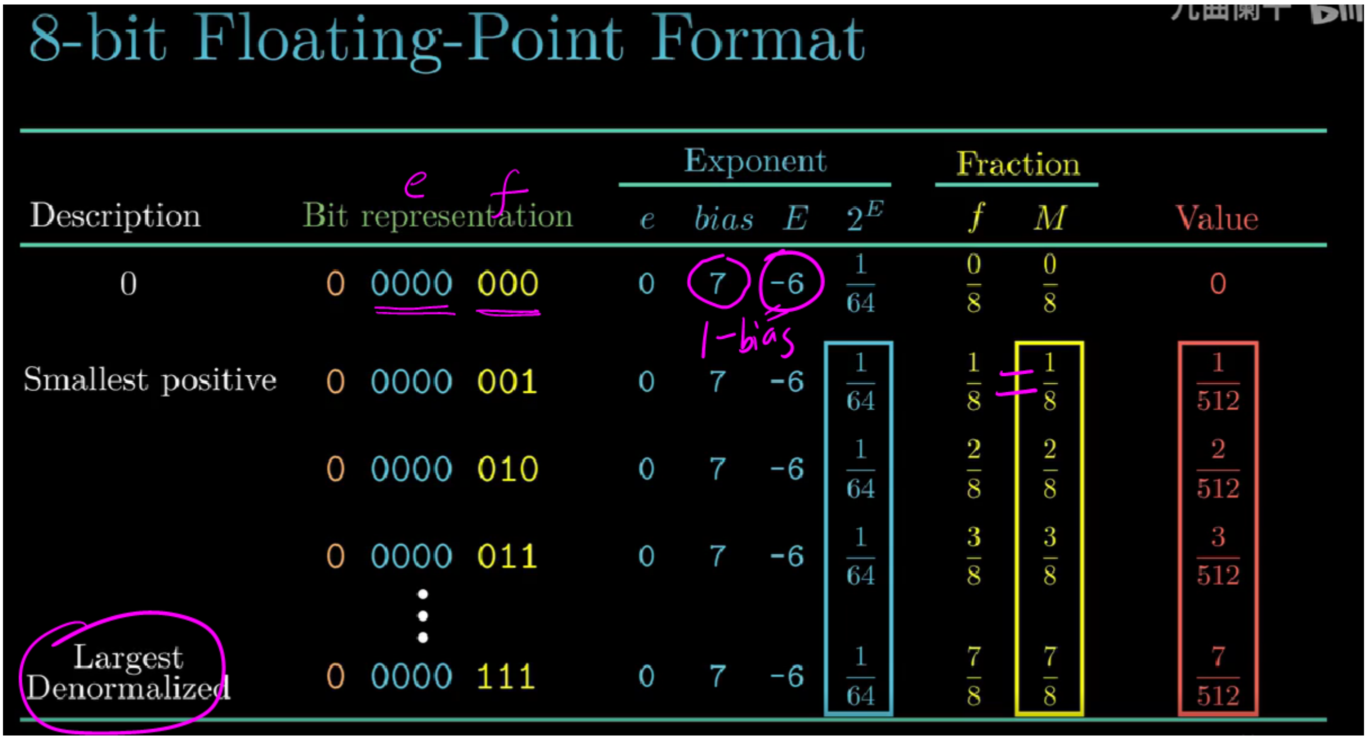
Case 1: s = 1 $f = 0$ $V = -\infty$

3.NaN

| s | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | $\neq 0$ |
|---|---|---|---|---|---|---|---|---|---|

$\sqrt{-1}$ $\infty - \infty$

**bias=7 =>> 2 ^ (n-1) - 1 = 7 (n=4)**



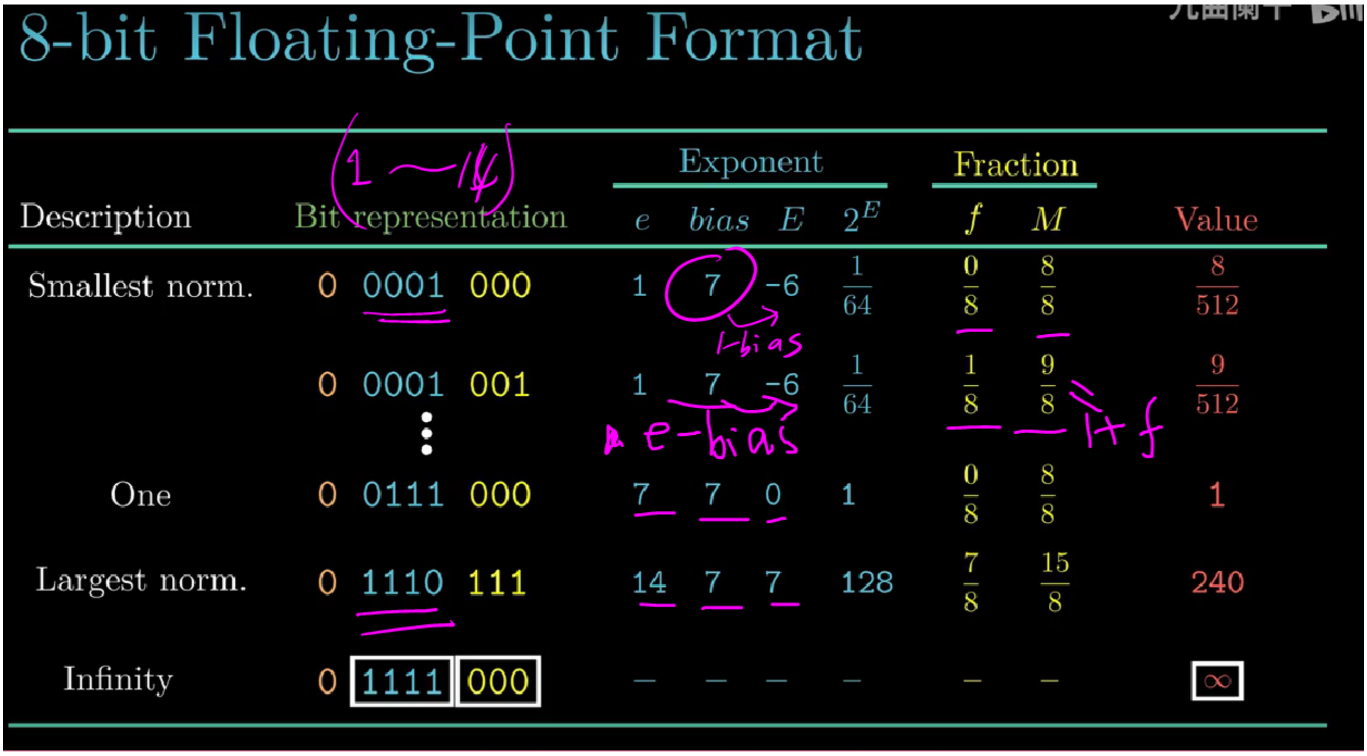
## 8-bit Floating-Point Format

|  |  | Exponent |  |  |  | Fraction |  |  |
|---|---|---|---|---|---|---|---|---|
| Description | Bit representation | $e$ | $bias$ | $E$ | $2^E$ | $f$ | $M$ | Value |
| 0 | 0 0000 000 | 0 | 7 | -6 | $\frac{1}{64}$ | $\frac{0}{8}$ | $\frac{0}{8}$ | 0 |
| Smallest positive | 0 0000 001 | 0 | 7 | -6 | $\frac{1}{64}$ | $\frac{1}{8}$ | $\frac{1}{8}$ | $\frac{1}{512}$ |
|  | 0 0000 010 | 0 | 7 | -6 | $\frac{1}{64}$ | $\frac{2}{8}$ | $\frac{2}{8}$ | $\frac{2}{512}$ |
|  | 0 0000 011 | 0 | 7 | -6 | $\frac{1}{64}$ | $\frac{3}{8}$ | $\frac{3}{8}$ | $\frac{3}{512}$ |
|  | ⋮ |  |  |  |  |  |  |  |
| Largest Denormalized | 0 0000 111 | 0 | 7 | -6 | $\frac{1}{64}$ | $\frac{7}{8}$ | $\frac{7}{8}$ | $\frac{7}{512}$ |

e f

1-bias

=



## 8-bit Floating-Point Format

|  |  | Exponent |  |  |  | Fraction |  |  |
|---|---|---|---|---|---|---|---|---|
| Description | Bit representation | $e$ | $bias$ | $E$ | $2^E$ | $f$ | $M$ | Value |
| Smallest norm. | 0 0001 000 | 1 | 7 | -6 | $\frac{1}{64}$ | $\frac{0}{8}$ | $\frac{8}{8}$ | $\frac{8}{512}$ |
|  | 0 0001 001 | 1 | 7 | -6 | $\frac{1}{64}$ | $\frac{1}{8}$ | $\frac{9}{8}$ | $\frac{9}{512}$ |
|  | ⋮ |  |  |  |  |  |  |  |
| One | 0 0111 000 | 7 | 7 | 0 | 1 | $\frac{0}{8}$ | $\frac{8}{8}$ | 1 |
| Largest norm. | 0 1110 111 | 14 | 7 | 7 | 128 | $\frac{7}{8}$ | $\frac{15}{8}$ | 240 |
| Infinity | 0 1111 000 | – | – | – | – | – | – | ∞ |

(1 ~ 14)

1-bias

e-bias

= 1+f