# 第十一章 数字信号处理的误差分析

11.1 引言

11.2 数的表示对量化误差的影响

11.3 A/D转换的量化误差分析

11.4 数字滤波器的系数量化误差

11.5 数字滤波器的运算量化误差

# 8.1 引言

- 数字信号处理的实质：一组数值运算。
- 从设计的角度来讨论：认为数字是无限精度的。
- 从实现的角度考虑：数字的精度是有限的
  - 用有限字长的二进制数码表示

- 从设计时的无限精度到实现时的有限精度，会产生相对于原设计系统的误差，严重时会导致系统崩溃。

**数字滤波器的实现方法：**

**a. 利用专用计算机（DSP系统）；**

**b. 直接利用计算机和通用软件编程实现。**

一个数字滤波器的系统函数一般可表示为有理函数形式：

$$H(z)=\frac{\sum_{i=0}^{N} a_i Z^{-i}}{1-\sum_{i=1}^{N} b_i Z^{-i}}$$

为IIR滤波器形式，$\{b_i\}$都为0时就是一个FIR滤波器。

对于这样一个系统，也可用差分方程来表示：

$$y(n)=\sum_{i=0}^{N} a_i x(n-i)+\sum_{i=1}^{N} b_i y(n-i)$$

$$x(n) \longrightarrow \boxed{\text{DF}} \longrightarrow y(n)$$

**IIR、FIR的系统函数**

↓

**网络结构形式（直接、串、并、格型结构）**

↓

**软、硬件实现**

一个输出序列是其过去$N$点输出值的线性组合加上当前输入序列与过去 $N$ 点输入序列的线性组合。

$y(n)$除了与当前的输入 $x(n)$ 有关，同时还与过去的输入和过去的输出有关，系统是带有记忆的。

计算形式　直接计算
分解为多个有理函数相加
分解为多个有理函数相乘

↓

计算结构 ⟶ 不同的效果　实现简单
编程方便
计算精度较高

数字信号_____通过采样和转换得到的，

转换的位数是有限的（一般6、8、10、12、16位）量化误差，

计算机中的数的表示也总是有限的，经此表示的滤波器的系数同样存在量化误差，

在计算过程中因有限字长也会造成误差。

**量化误差主要有三种误差：** 要求记忆

①A/D变换量化效应---信号采集时—抽样定理；

②系数的量化效应--系统函数分子分母系数的量化表示；

③数字运算的有限字长效应--乘法运算--乘积的有效位数比每个因子都增加，须截短或舍入。

# 11.2 数的表示对量化误差的影响

## 11.2.1. 二进制数的定点与浮点表示

任意二进制数可表示成如下形式

$$B = 2^{c} \times M \qquad (11.2.1)$$

如二进制数 $1001.111=2^{100}\times 0.1001111$

### （1）定点制

在整个运算中小数点的位置固定不变

例如：设有一个8位寄存器，按如下方式表示数据：

符号位：0表示正，1表示负

| s | $a_1$ | $a_2$ | $a_3$ | . | $a_4$ | $a_5$ | $a_6$ | $a_7$ |
|---|---|---|---|---|---|---|---|---|

Q4.4格式

| s | . | $a_1$ | $a_2$ | $a_3$ | $a_4$ | $a_5$ | $a_6$ | $a_7$ |
|---|---|---|---|---|---|---|---|---|

Q1.7格式

定点表示 $\beta_0 \bullet \beta_1 \beta_2 \ \cdots \beta_b$

•整个运算中，小数点在数码中的位置固定不变，原则上小数点在数码中的位置是任意的，称为定点制；

•通常定点制总是把数限制在±1之间； 最高位为符号位，0为正，1为负，小数点紧跟在符号位后；数的本身只有小数部分，称为“尾数”；若尾数有$b_m$位，则数B所能表示的数的范围是

$$|B| \le 1 - 2^{-b_m} \qquad (11.2.2)$$

•若数值较大时，可乘上一个衰减因子，保证该数在运算中不超过1；运算后再除以该因子还原。

• 定点数作加减法时结果可能会超出±1，称为“溢出”；

• 乘法运算不溢出，但字长要增加一倍。

为保证字长不变，乘法后，一般要对增加的尾数作截尾或舍入处理，带来误差。

缺点：动态范围小，有溢出；衰减比例系数不好确定。

## （2）浮点制

浮点数的小数点位置是不固定的，它随每个数的大小而变化。

$$B = 2^{c} \times M \qquad (11.2.1)$$

阶码

尾数

尾数、阶码均用带符号的定点数来表示

尾数的第一位表示了浮点数的正负；

设阶码为$b_c$位，尾数为$b_m$位，则浮点数的表示范围是

$$|B| \le 2^{2b_c - 1} \times (1 - 2^{-b_m})$$

**浮点制运算:**

**相加**
- **对阶**
- **相加**
- **归一化,并作尾数处理**

**相乘： 尾数相乘, 阶码相加, 再作截尾或舍入。**

**尾数M的取值在[0.5, 1)，只是要求规格化表示。**

$x = 0.0101 \times 2^{011}$就是非规格化表示。

为充分利用尾数的有效位数，规格化表示为$x = 0.101 \times 2^{010}$

规格化形式：使 $0.5 \leqslant M < 1$。

例如 $x>0$ 时，阶码 $c$ 满足： $2^{c-1} \le |x| < 2^{c}$

举例说明：

$$x_1 = 2^{c_1} M_1, \quad x_2 = 2^{c_2} M_2$$

相乘： $$x_1 \times x_2 = 2^{c_1+c_2}(M_1 \times M_2)$$

相加： 若 $c_1 < c_2$，把 $M_1$ 各位右移变为 $M_1'$，使 $x_1 = 2^{c_2} M_1'$

$$x_1 + x_2 = 2^{c_2}(M_1' + M_2)$$

字长增加，
$M_1'$必须取近似值

优点: 动态范围大,一般不溢出.

缺点: 相乘、相加，都会增加字长，都要对尾数处理作量化处理。

一般，浮点数都用较长的字长，精度较高，所以我们讨论误差影响主要针对定点制。

## 11.2.2 原码、补码和反码

定点数的表示分为三种（原码、反码、补码）：

设有一个（b+1）位码定点数： $\beta_0\beta_1\beta_2---\beta_b$，则

①原码定义为

$$[x]_{原}=\begin{cases} x & 0\le x<1 \\ 1+|x| & -1<x\le 0 \end{cases} \qquad (11.2.3)$$

十进制数值为 $[x]_{10}=(-1)^{\beta_0}\sum_{i=1}^{b}\beta_i 2^{-i}$

例：1.111→-0.875 ，0.010→0.25

原码的优点是乘除法运算方面，而加减法运算要增加时间。

②反码定义：（正数同原码，负数则将原码中的尾数按位求反）

$$[x]_{反} = \begin{cases} x & 0 \le x < 1 \\ (2-2^{-b}) - |x| & -1 < x \le 0 \end{cases} \qquad （11.2.4）$$

十进制数值为 $[x]_{10} = -\beta_0(1-2^{-b}) + \sum_{i=1}^{b} \beta_i 2^{-i}$

例：　$x = -0.625$

正数表示：**0.101**

其反码为：**1.010**

## ③补码表示

（正数同原码，负数则将原码中的尾数求反加1）

$$[x]_{补} = \begin{cases} x & 0 \le x < 1 \\ 2 - |x| & -1 < x \le 0 \end{cases} \qquad (11.2.5)$$

十进制数值为 $$x = -\beta_0 + \sum_{i=1}^{b} \beta_i 2^{-i}$$

例： $x = -0.75$

正数表示：**0.110**

取反：**1.001**

$x$的补码：**1.010**

补码加法运算规律：

正负数可直接相加，符号位同样参加运算，如符号位发生进位，进位的 1 丢掉。

负数以补码形式表示的原因是：

将减法运算变为补码的加法运算。

补码应用最为广泛。

负数（$\beta_0 = 1$）有三种表示方法：原码、反码、补码

〈1〉原码： $\beta_0 = 1$， $\beta_1 \sim \beta_b$ 与 $|x|$ 相同

〈2〉反码：把原码尾数中各位取反（0变1，1变0）

〈3〉补码：反码的末尾+1

例11.1：设 $b$=4，已知 $x_1 = (0.4375)_{10} = (0.0111)_2$,

$$x_2 = (0.625)_{10} = (0.1010)_2,$$

分别将$x_3=(-0.4375)_{10}$，$x_4=(-0.625)_{10}$ 用原码、反码、补码表示。

解：$x_3=(-0.4375)_{10}$，

原码：1.0111；

反码：1.1000；

补码：1.1001；

$x_4 = (-0.625)_{10}$ ，

原码：1.1010；

反码：1.0101；

补码：1.0110；

# 11.2.3 截尾与舍入效应

## 1. 定点制的量化误差

定点制中的乘法，运算完毕后会使字长增加，例如原来是b位字长，运算后增长到b1位，需对尾数作量化处理使b1位字长降低到b位。

量化处理方式：

- 截尾：保留b位，抛弃余下的尾数；
- 舍入：按最接近的值取b位码。

两种处理方式产生的误差不同，另外，码制不同，误差也不同

1、截尾处理：

1）正数（三种码形式相同）

一个$b_1$位的正数 $x$ 为：

$$x = \sum_{i=1}^{b_1} \beta_i 2^{-i}$$

用$[\cdot]_T$表示截尾处理，则

$$[x]_T = \sum_{i=1}^{b} \beta_i 2^{-i}$$

# 截尾误差

$$E_T=[x]_T-x=-\sum_{i=b+1}^{b_1}\beta_i 2^{-i} \qquad (11.2.6)$$

可见，$E_T\leqslant 0$，$\beta_i$全为1时，$E_T$有最大值，

$$E_T=-\sum_{i=b+1}^{b_1}2^{-i}=-(2^{-b}-2^{-b1})$$

“量化宽度”或“量化阶” $q=2^{-b}$：代表b位字长可表示的最小数。

一般 $2^{-b1}\ll 2^{-b}$，因此正数的截尾误差为

$$\mathbf{-q\le E_T\le 0} \qquad (11.2.7)$$

## 2）负数

负数的三种码表示方式不同，所以误差也不同。

原码（β $_0$=1）：

$$x = -\sum_{i=1}^{b_1} \beta_i 2^{-i} \qquad [x]_T = -\sum_{i=1}^{b} \beta_i 2^{-i}$$

$$E_T = [x]_T - x = \sum_{i=b+1}^{b_1} \beta_i 2^{-i} \qquad (11.2.8)$$

$$\mathbf{0 \le E_T \le q} \qquad (11.2.9)$$

补码（$\beta_0 = 1$）

$$x = -1 + \sum_{i=1}^{b_1} \beta_i 2^{-i}$$

$$[x]_T = -1 + \sum_{i=1}^{b} \beta_i 2^{-i}$$

$$E_T = \sum_{i=1}^{b} \beta_i 2^{-i} - \sum_{i=1}^{b_1} \beta_i 2^{-i}$$

因 $b_1 > b,$ 所以

$$-q < E_T \le 0 \qquad (11.2.10)$$

反码（ $\beta_0 = 1$ ）

$$x = -1 + \sum_{i=1}^{b_1} \beta_i 2^{-i} + 2^{-b_1}$$

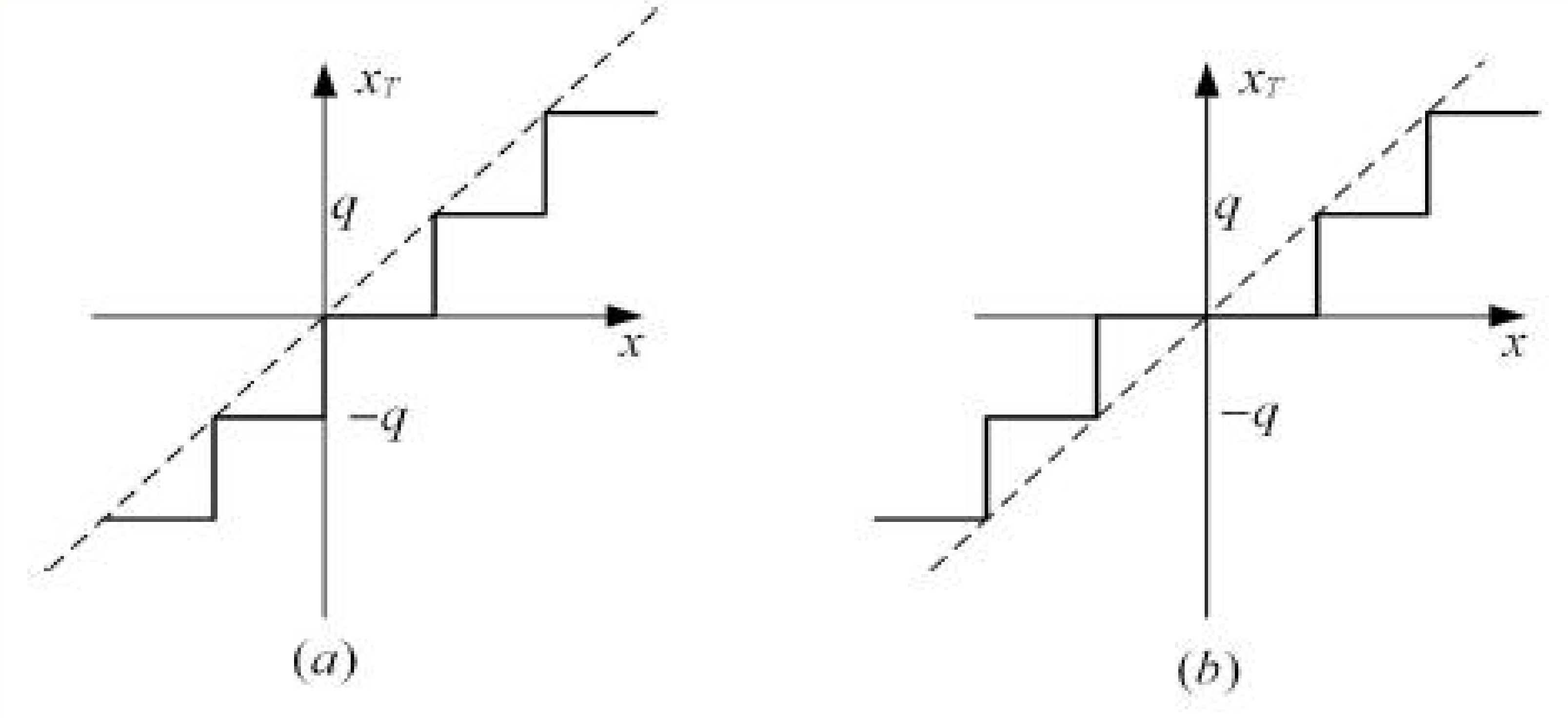
$$[x]_T = -1 + \sum_{i=1}^{b} \beta_i 2^{-i} + 2^{-b}$$

$$E_T = [x]_T - x = -\sum_{i=b+1}^{b_1} \beta_i 2^{-i} + (2^{-b} - 2^{-b_1})$$

$$0 \le E_T < q \qquad \text{（11.2.11）}$$

（ $E_T > 0$ 与原码的相同）

补码的截尾误差均是负值，原码、反码的截尾误差取决于数的正负，正数时为负，负数时为正。




（a）补码表示的截尾过程　（b）原码和反码表示的截尾过程

**图11.2.1　截尾量化处理的非线性特性**

## 2. 舍入处理

通过b+1位上加1后作截尾处理实现。

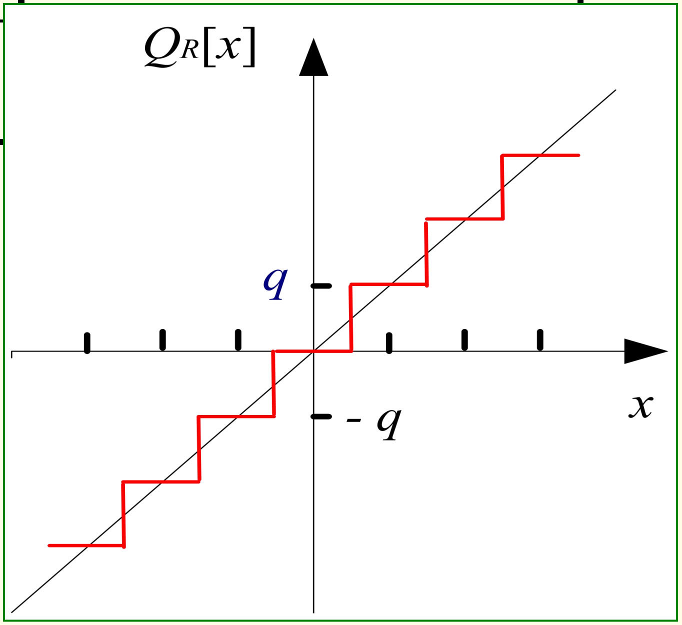
就是通常的四舍五入法，按最接近的数取量化，不论正数、负数，还是原码、补码、反码，误差总是在 $\pm\frac{q}{2}$ 之间，以 $[x]_R$ 表示对x作舍入处理。

舍入处理的误差比截尾处理的误差小，所以对信号进行量化时多用舍入处理。

## （2）定点制量化误差

$$q = 2^{-b}$$

| 截尾误差 | | | 舍入误差 |
|---|---|---|---|
| 正数 | | $-q < E_T \le 0$ | $-\frac{q}{2} < E_R \le \frac{q}{2}$ |
| 负数 | 原码 | $0 \le E_T < q$ | |
| | 反码 | | |
| | 补码 | $-q < E_T \le 0$ | |



## 2. 浮点制量化误差

浮点制中，截尾与舍入只影响尾数的字长，但误差却与阶码的值有关。

例如
$$x_1 = 0.1001 \times 2^{000} (= 0.5625)$$
$$x_{1T} = 0.10 \times 2^{000} (= 0.50)$$

其误差为 $E_1 = x_{1T} - x_1 = -0.0625$

而
$$x_2 = 0.1001 \times 2^{011} (= 4.5)$$
$$x_{2T} = 0.10 \times 2^{011} (= 4.0)$$

其误差为 $E_2 = x_{2T} - x_2 = -0.5$

说明误差与数字本身大小有关。采用相对误差较绝对误差更能反应浮点制的特点。

用$\varepsilon_T$和$\varepsilon_R$表示截尾和舍入的相对误差

$$\varepsilon_T = \frac{x_T - x}{x}; \qquad \varepsilon_R = \frac{x_R - x}{x}; \qquad (11.2.12)$$

相对误差的范围：

当采用舍入处理时，尾数误差在 $\pm q/2$ 之间，设阶码为c，则其绝对误差为：

$$-2^C(q/2) < \varepsilon_R x < 2^C(q/2) \qquad (11.2.13)$$

又由于x是归一化的浮点数，因此

$$2^{c-1} \le |x| < 2^c \qquad (11.2.14)$$

将式（11.2.14）代入式（11.2.13）就可以得到

$$-q < \varepsilon_R \le q \qquad (11.2.15)$$

同理，可以确定浮点数 $x = M \cdot 2^{C}$ 的截尾相对误差的范围：

$$\varepsilon_T = \frac{x_T - x}{x}$$

$$\boxed{q = 2^{-b}}$$

<table>
<tr><th colspan="3">截尾误差</th><th>舍入误差</th></tr>
<tr><td colspan="2">正数</td><td>$-2q < \varepsilon_T \le 0$</td><td rowspan="4">$-q < \varepsilon_R \le q$</td></tr>
<tr><td rowspan="3">负数</td><td>原码</td><td rowspan="2">$-2q < \varepsilon_T \le 0$</td></tr>
<tr><td>反码</td></tr>
<tr><td>补码</td><td>$0 \le \varepsilon_T < 2q$</td></tr>
</table>

# 8.3 A/D转换的量化误差分析

## 1. AD转换过程的统计模型

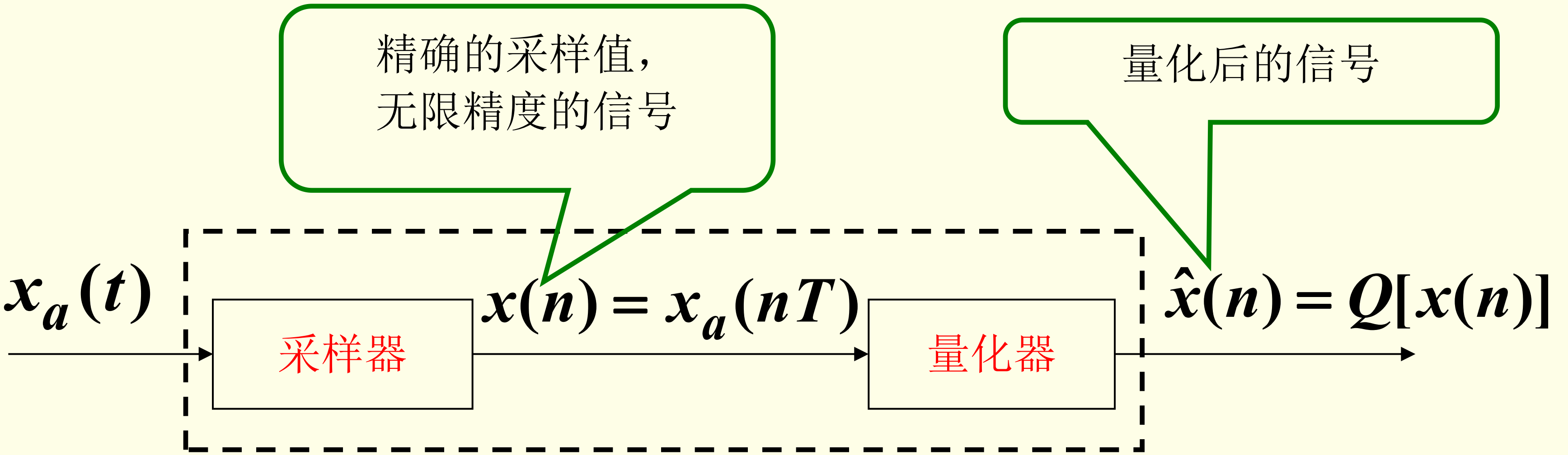



图11.3.1 AD变换的非线性模型

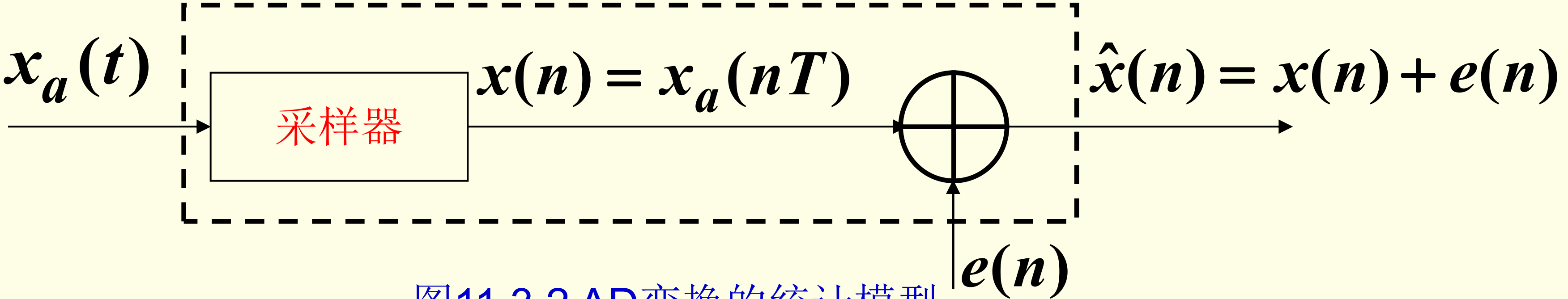



图11.3.2 AD变换的统计模型

A/D变换器分为两部分：

采样：时间离散，幅度连续；

量化：数字编码，对采样序列作舍入或截尾处理，得有限字长数字信号 $\hat{x}(n)$

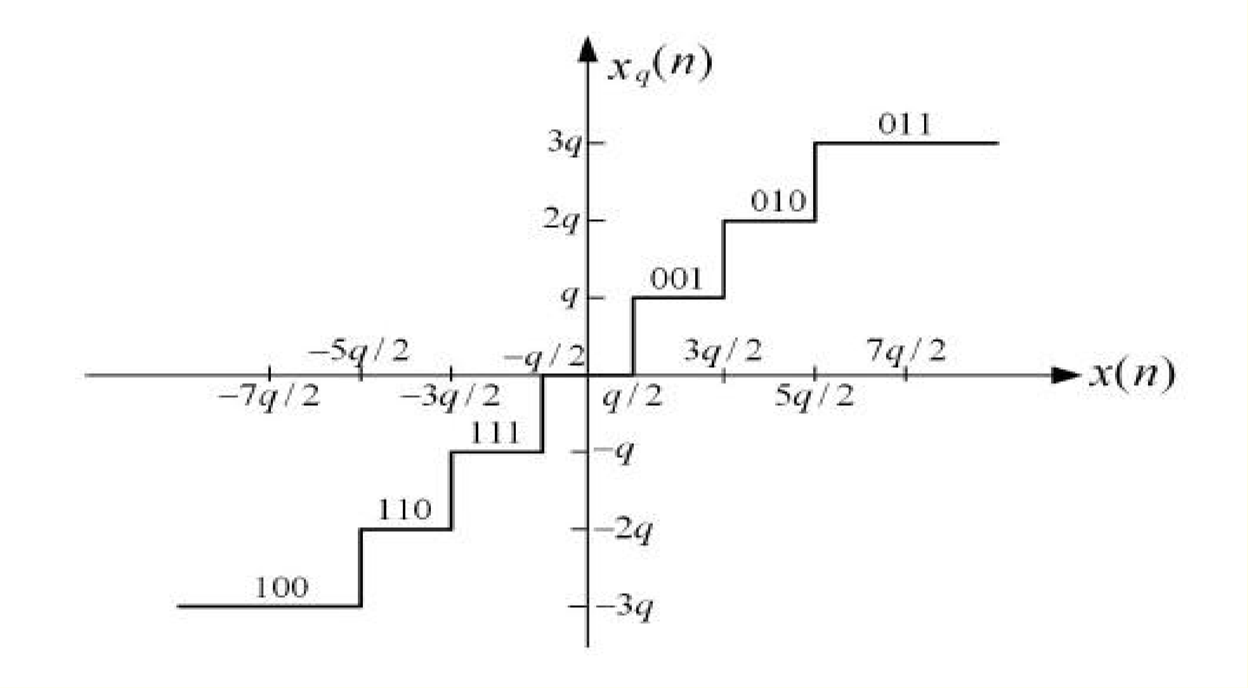



图11.3.3 AD转换器定点补码舍入量化特性

量化误差： $$e(n)=\hat{x}(n)-x(n)=Q[x(n)]-x(n) \qquad (11.3.1)$$

为了讨论A/D的量化效应，先做以下假设：

①. 充分限带：使得采样后不发生混叠失真；

②. 表示成（b+1）位的定点补码小数；

③. 采用舍入量化方式；

④. 模拟信号 $x_a(t)$ 已经归一化了，即$-1<x_a(nT)<1$。

$$\boxed{-\frac{q}{2}<E_R\le\frac{q}{2},\quad q=2^{-b}}$$

## 2. 量化误差的统计分析

假设量化误差 $e(n)$ 具有下列特性：

- $e(n)$ 是一个平稳随机序列；
- $e(n)$ 与采样信号 $x(n)$ 不相关；
- $e(n)$ 本身的任意两个值之间不相关；
- $e(n)$ 在其误差范围内均匀等概率分布。

量化误差 $e(n)$ 是一个与信号序列完全不相关的白噪声序列，称为量化噪声；

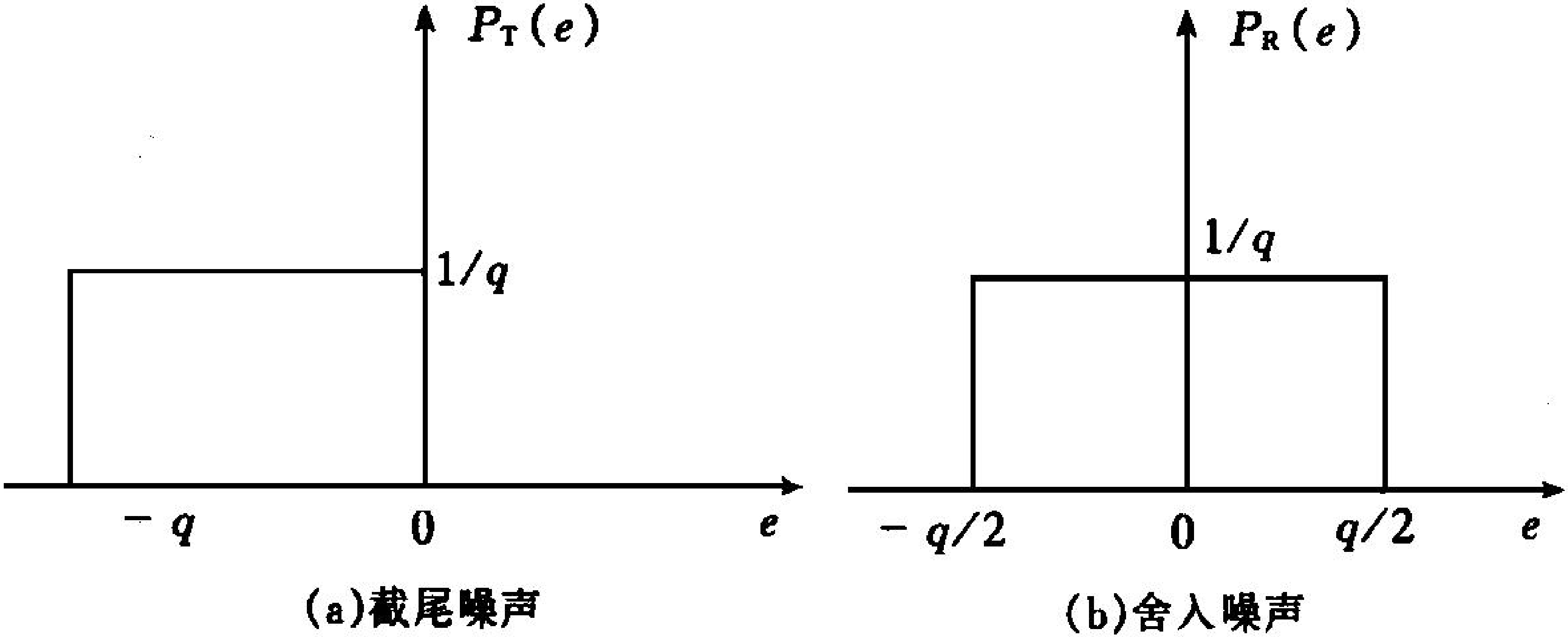
对一个采样数据 $x(n)$ 作截尾和舍入处理，则

截尾量化误差：$$e_T(n) = -\sum_{i=b+1}^{\infty} \beta_i 2^{-i}$$

$$-q < e_T(n) \le 0, \qquad q = 2^{-b}$$

舍入量化误差：$$-\frac{q}{2} < e_R(n) \le \frac{q}{2}$$

上两式给出了量化误差的范围，要精确知道误差的大小很困难。一般，我们总是通过分析量化噪声的统计特性来描述量化误差。可以用一统计模型来表示A/D的量化过程。




量化噪声的概率分布

图11.3.4 e(n)的均匀等概率分布

误差 $e(n)$ 的均值和方差：

截尾量化噪声：

$$m_e = E[e(n)] = \int_{-\infty}^{\infty} ep(e)de = \int_{-q}^{0} \frac{1}{q} ede = -q/2$$

$$\sigma_e^2 = E[(e(n) - m_e)^2] = \int_{-\infty}^{\infty} (e - m_e)^2 p(e)de = q^2/12$$

有直流分量，会影响信号的频谱结构。

舍入量化噪声：

$$m_e = 0$$

$$\sigma_e^2 = q^2 / 12$$

可见，量化噪声的方差与A/D变换的字长直接有关，字长越长，量化噪声越小。

$$q = 2^{-b}$$

对补码舍入和截尾时都假设量化误差的自协方差序列为

$$\gamma_{ee}(m) = \sigma_e^2 \delta(m) \quad （11.3.2）$$

定义量化信噪比：

$$\frac{\sigma_x^2}{\sigma_e^2} = \frac{\sigma_x^2}{q^2/12} = (12 \times 2^{2b})\sigma_x^2 \qquad （11.3.3）$$

用对数表示：

$$SNR = 10\lg\left(\frac{\sigma_x^2}{\sigma_e^2}\right) = 10\lg\left[(12 \times 2^{2b})\sigma_x^2\right]$$

$$= 6.02(b+1) + 10\lg(3\sigma_x^2) = 6.02b + 10.79 + 10\lg(\sigma_x^2)$$

（11.3.4）

字长每增加 1 位，量化信噪比增加6个分贝；

- 信号能量越大，量化信噪比越高。

注：因信号本身有一定的信噪比，单纯提高量化信噪比无意义。

**例11.2：已知x(n)在-1至1之间均匀分布，求8、12位时A/D的SNR。**

因均匀分布，所以有：

均值： $E\left[x(n)\right]=0$

方差： $\sigma_x^2=\int_{-1}^{1}\frac{1}{2}x^2dx=\frac{1}{3}$

当 b=8 位，则SNR=54dB，当 b=12 位，则SNR=78dB.

$$SNR=10\lg(\frac{\sigma_x^2}{\sigma_e^2})=10\lg\left[(12\times2^{2b})\sigma_x^2\right]=6.02(b+1)+10\lg(3\sigma_x^2)$$

## 3. 量化噪声通过线性系统

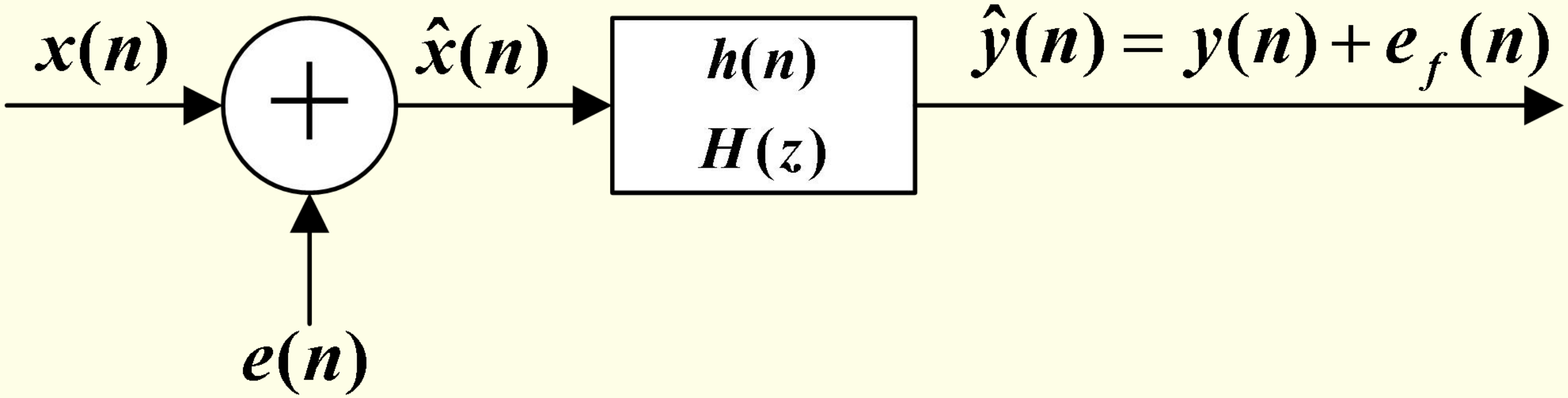



图11. 3. 5 量化噪声通过线性系统的线性模型

系统的输出：

$$\hat{y}(n)=\hat{x}(n)*h(n)=\left[x(n)+e(n)\right]*h(n)$$

$$=x(n)*h(n)+e(n)*h(n)=y(n)+e_f(n) \qquad (11.3.5)$$

设e(n)是定点补码舍入误差，e(n)的均值为$m_e$、方差为 $\sigma_e^2$

则系统量化噪声的输出$e_f(n)$的均值 $m_f$和方差 $\sigma_f^2$ 计算如下：

$$m_f = E[e_f(n)] = E[e(n) * h(n)] = m_e \sum_{m=0}^{\infty} h(m) = 0 \quad （11.3.6）$$

$$\sigma_f^2 = E\left[e_f^2(n)\right] = E\left[\sum_{m=0}^{\infty} h(m)e(n-m) \sum_{l=0}^{\infty} h(l)e(n-l)\right]$$

$$= \sum_{m=0}^{\infty}\sum_{l=0}^{\infty} h(m)h(l)E\left[e(n-m)e(n-l)\right] \quad （11.3.7）$$

根据Parseval定理， $\sigma_f^2$ 也可以用下式表示：

$$\sigma_f^2 = \sigma_e^2 \sum_{m=0}^{\infty} \left| h(m) \right|^2 = \frac{\sigma_e^2}{2\pi j} \oint_c H(z) H(z^{-1}) \frac{dz}{z} \qquad (11.3.8)$$

或者

$$\sigma_f^2 = \sigma_e^2 \sum_{m=0}^{\infty} \left| h(m) \right|^2 = \frac{\sigma_e^2}{2\pi} \int_{-\pi}^{\pi} \left| H(e^{j\omega}) \right|^2 d\omega \qquad (11.3.9)$$

H(z)全部极点在单位圆内，$\oint_c$ 表示沿单位圆逆时针方向的圆周积分。由留数定理：

$$\sigma_f^2 = \sigma_e^2 \sum_k \mathrm{Res}\left[\frac{H(z)H(z^{-1})}{z}, z_k\right]$$

*该式为求解量化噪声输出的方法之二。*

如 $e(n)$ 为截尾噪声，则输出噪声中还有一直流分量

$$m_f = E\left[\sum_{m=0}^{\infty} h(m)e(n-m)\right] = m_e \cdot \sum_{m=0}^{\infty} h(m) = m_e \cdot H(e^{j0})$$

例10.3：一个8位A/D变换器（$b=7$，最高位为符号位），其输出$\hat{x}(n)$作为IIR滤波器的输入，求滤波器输出端的量化噪声功率，已知IIR滤波器的系统函数为：

$$H(z)=\frac{z}{z-0.999}$$

复习求解方法：$\sigma_f^2=\sigma_e^2\sum_{m=0}^{\infty}h^2(m)=\frac{\sigma_e^2}{2\pi j}\oint_c H(z)H(z^{-1})\frac{dz}{z}$

解：由于A/D的量化效应，滤波器输入端的噪声功率：

$$\sigma_e^2=\frac{q^2}{12}=\frac{2^{-14}}{12}=\frac{2^{-16}}{3}$$

滤波器的输出噪声功率为：

$$\sigma_f^2 = \frac{\sigma_e^2}{2\pi j}\oint_c \frac{1}{(z-0.999)(z^{-1}-0.999)}\frac{dz}{z}$$

其积分值等于单位圆内所有极点留数的和。单位圆内有一个极点 z=0.999，所以

$$\sigma_f^2 = \sigma_e^2 \frac{1}{\frac{1}{0.999}-0.999}\cdot\frac{1}{0.999}$$

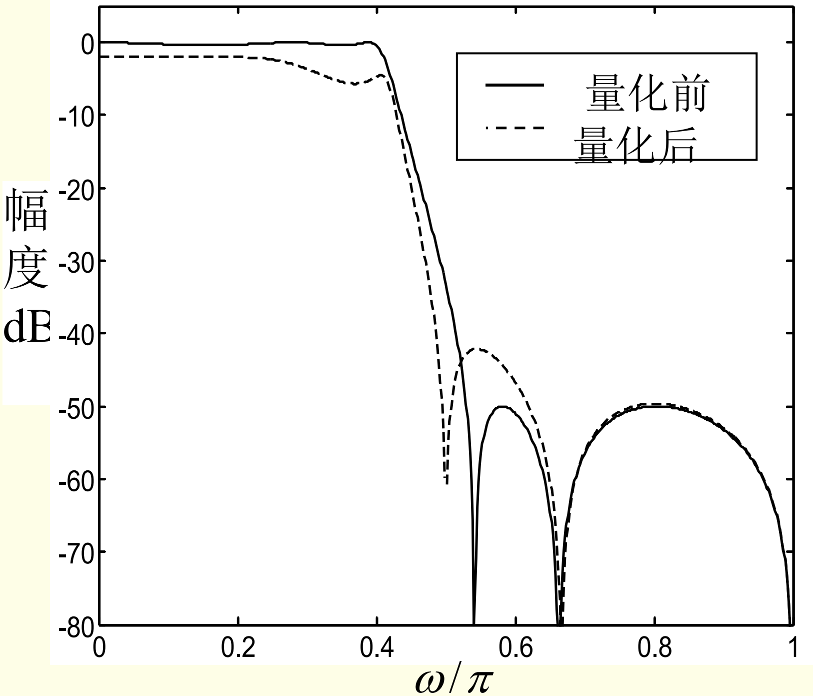
$$= \frac{2^{-16}}{3}\frac{1}{1-0.999^2} = 2.5444\times 10^{-3}$$

## 10.4 数字滤波器的系数量化误差

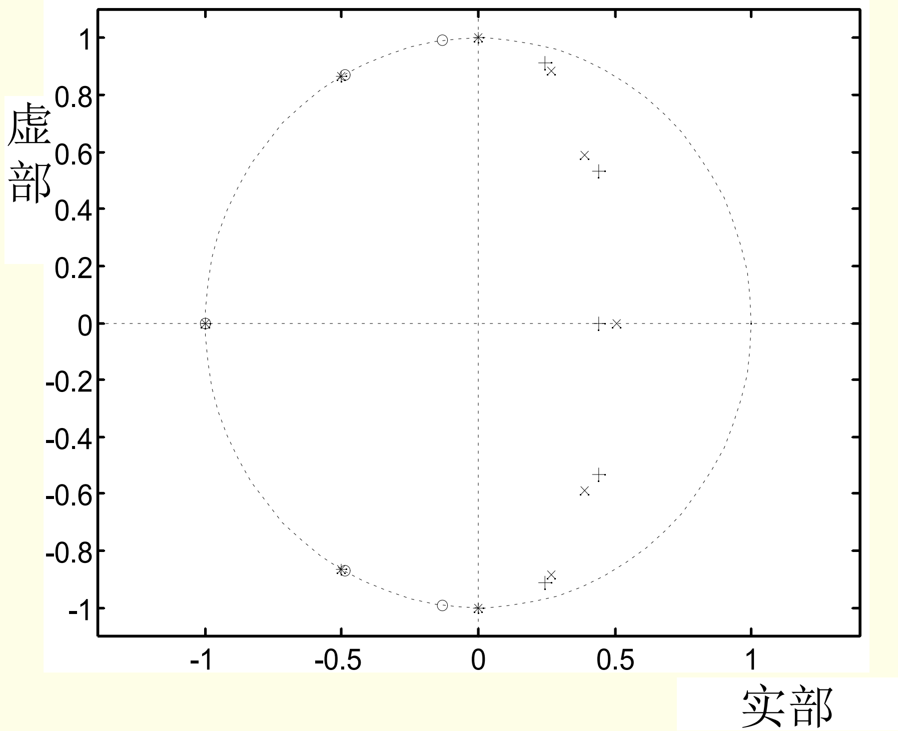
由于滤波器的所有系数必须以有限长度的二进码形式存放在存储器中，所以必然对理想系数值取量化，造成实际系数存在误差，使零、极点位置发生偏离，影响滤波器性能。

一个设计正确的滤波器，在实现时，由于系数量化，可能会导致实际滤波器的特性不符合要求，严重时甚至使单位圆内的极点偏离到单位圆外，从而系统失去稳定性。

系数量化对滤波器的影响与字长有关，也与滤波器的结构有关，选择合适的结构可改善系数量化的影响。




(a) 系数量化前后的频率响应




(b) 系数量化前后的零极点分布
'o'量化前的零点, '*'量化后的零点,
'x'量化前的极点, '+'量化后的极点

## 五阶椭圆低通滤波器的量化效应

## 极点位置灵敏度

指每个极点位置对各系数偏差的敏感程度。极点位置的变化将直接影响系统的稳定性。所以极点位置灵敏度可以反映系数量化对滤波器稳定性的影响。

设系数量化后的系统函数为：

$$\hat{H}(z)=\frac{\sum_{i=1}^{N}\hat{a}_i z^{-i}}{1-\sum_{i=1}^{N}\hat{b}_i z^{-i}}=\frac{A(z)}{B(z)}$$

量化后的系数

$$\hat{a}_i=a_i+\Delta a_i$$

$$\hat{b}_i=b_i+\Delta b_i$$

分析量化偏差 $\Delta a_i, \Delta b_i$ 造成的极点位置偏差。

设理想极点为 $z_i$， $i = 1, 2, \cdots, N$ ，则

$$B(z) = 1 - \sum_{i=1}^{N} b_i z^{-i} = \prod_{i=1}^{N} (1 - z_i z^{-1})$$

系数量化后，极点变为 $z_i + \Delta z_i$ 位置偏差 $\Delta z_i$ 是由 $\Delta b_i$ 引起的。

$\Delta b_i$ 对 $\Delta z_i$ 的影响：

因每个极点都与 $N$ 个$b_i$系数有关，

$$z_i = z_i(b_1, b_2, \cdots, b_N), \qquad i = 1, \cdots, N$$

$$\therefore \quad \Delta z_i = \frac{\partial z_i}{\partial b_1}\Delta b_1 + \frac{\partial z_i}{\partial b_2}\Delta b_2 + \cdots + \frac{\partial z_i}{\partial b_N}\Delta b_N = \sum_{k=1}^{N} \frac{\partial z_i}{\partial b_k}\Delta b_k \qquad i = 1, \cdots, N$$

$\frac{\partial z_i}{\partial b_k}$ 决定量化影响大小，反映极点 $z_i$ 对系数 $b_k$变化的敏感程度。

$\frac{\partial z_i}{\partial b_k}$ 大，$\Delta b_k$ 对 $\Delta z_i$ 影响大；$\frac{\partial z_i}{\partial b_k}$ 小，$\Delta b_k$ 对 $\Delta z_i$ 的影响小。

$\frac{\partial z_i}{\partial b_k}$ --称之为极点位置灵敏度

下面由B(Z)求灵敏度 $\frac{\partial z_i}{\partial b_k}$ ：

利用偏微分关系：

$$\left.\frac{\partial B(z)}{\partial b_k}\right|_{z=z_i} = \left.\frac{\partial B(z)}{\partial z_i}\right|_{z=z_i}\left(\frac{\partial z_i}{\partial b_k}\right)$$

故 $$\frac{\partial z_i}{\partial b_k} = \left.\frac{\partial\ B(z)/\partial b_k}{\partial\ B(z)/\partial z_i}\right|_{z=z_i}$$

$$\because \frac{\partial B(z)}{\partial b_k} = -z^{-k}$$

又 $$\frac{\partial B(z)}{\partial z_i} = -z^{-N}\prod_{\substack{k=1\\k\neq i}}^{N}(z-z_k)$$

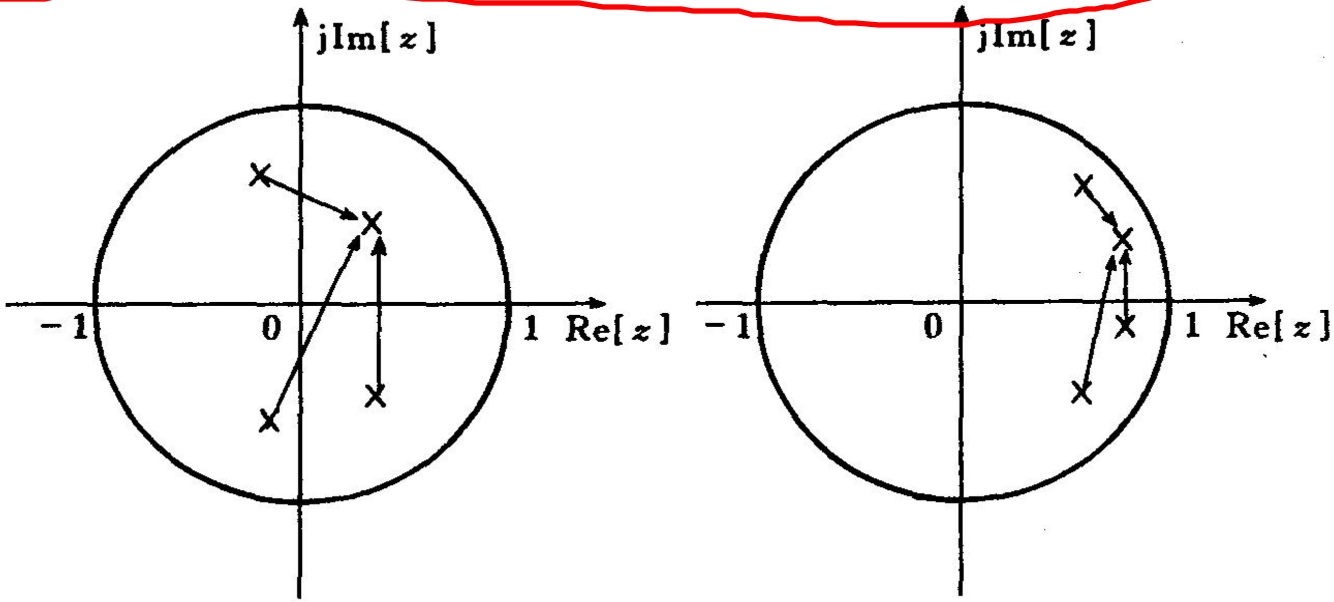
故

$$\frac{\partial z_i}{\partial b_k} = \frac{z_i^{N-k}}{\prod_{\substack{k=1\\k\neq i}}^{N}(z_i-z_k)}$$

上式分母中每个因子$(z_i-z_k)$是一个由极点$z_k$指向当前极点$z_i$的矢量，整个分母是所有极点指向极点$z_i$的矢量积，这些矢量越长，极点彼此间的距离越远，极点位置灵敏度越低；矢量越短，极点位置灵敏度越高。即极点位置灵敏度与极点间距离成反比。

例1，一个共轭极点在虚轴附近的滤波器如图（a）一个共轭极点在实轴附近的滤波器如图（b）

两者比较，前者极点位置灵敏度比后者小，即系数量化程度相同时，前者造成的误差比后者小。




(a)带通滤波器　　(b)低通滤波器

极点位置灵敏度与极点间距离成反比

例2 一个三对共轭极点的滤波器 H(z)，用三种结构实现。

1）用直接型结构实现，极点分布如图a ，

2）用三个二阶网络级联的形式实现，极点分布如图b ，

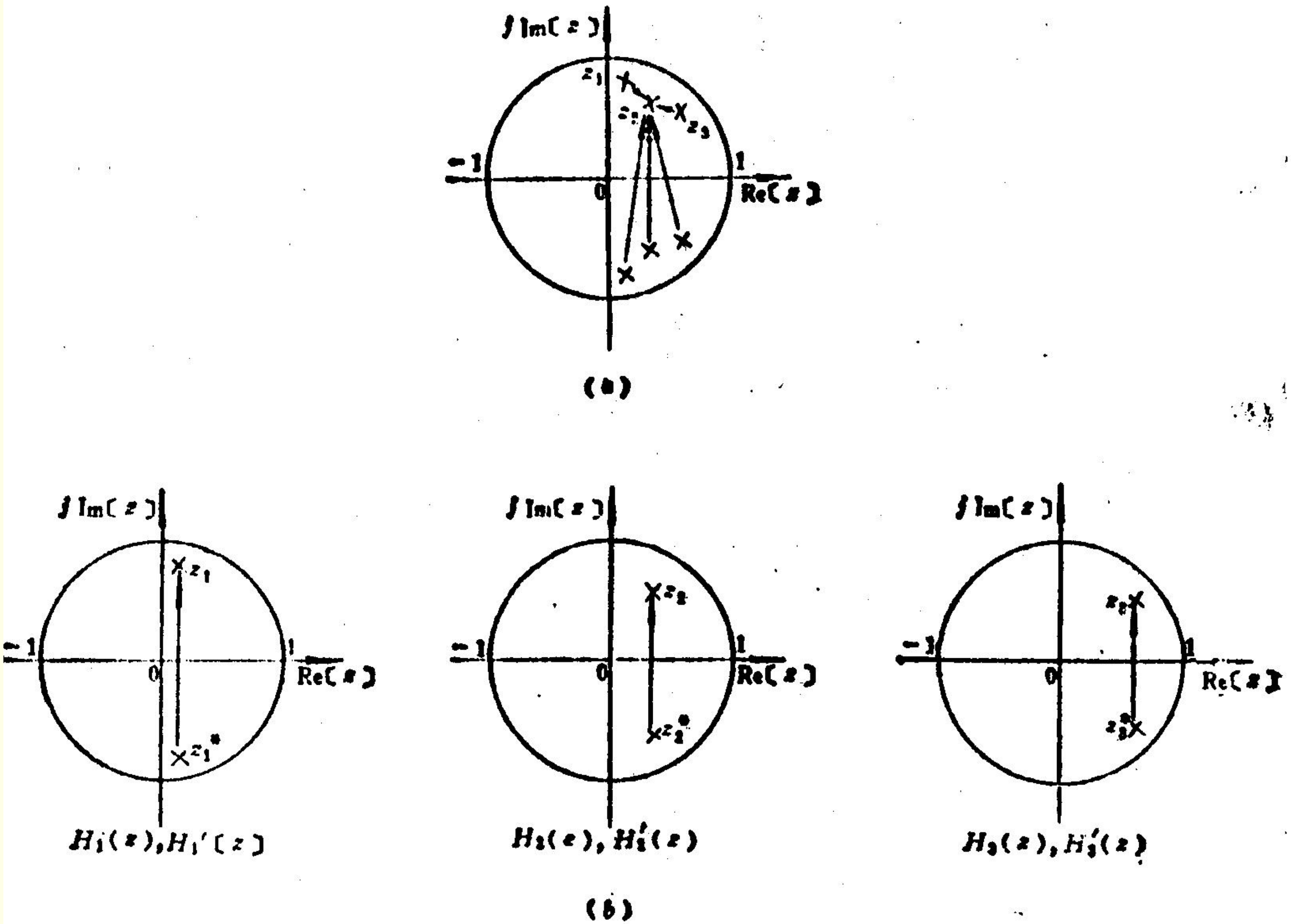
$$H(z)=H_1(z)\bullet H_2(z)\bullet H_3(z)$$

3）用三个并联二阶网络实现，极点分布如图b 。

$$H(z)=H'_1(z)+H'_2(z)+H'_3(z)$$

直接型极点分布密，极点位置灵敏度高。

级联和并联型，极点分布稀，极点位置灵敏度下降。




直接型与级联型、并联型极点密度的比较

(a) 直接型 $H(z)$ 的极点密度；(b) 级联型、并联型极点密度。

**影响极点位置灵敏度的几个因素：**

- 与零极点的分布状态有关；极点位置灵敏度大小与极点间距离成反比；
- 与滤波器结构有关。高阶直接型极点位置灵敏度高；并联或级联型，系数量化误差的影响小；
- 高阶滤波器避免用直接型，尽量分解为低阶网络的级联或并联。

# 11.5 数字滤波器的运算量化误差

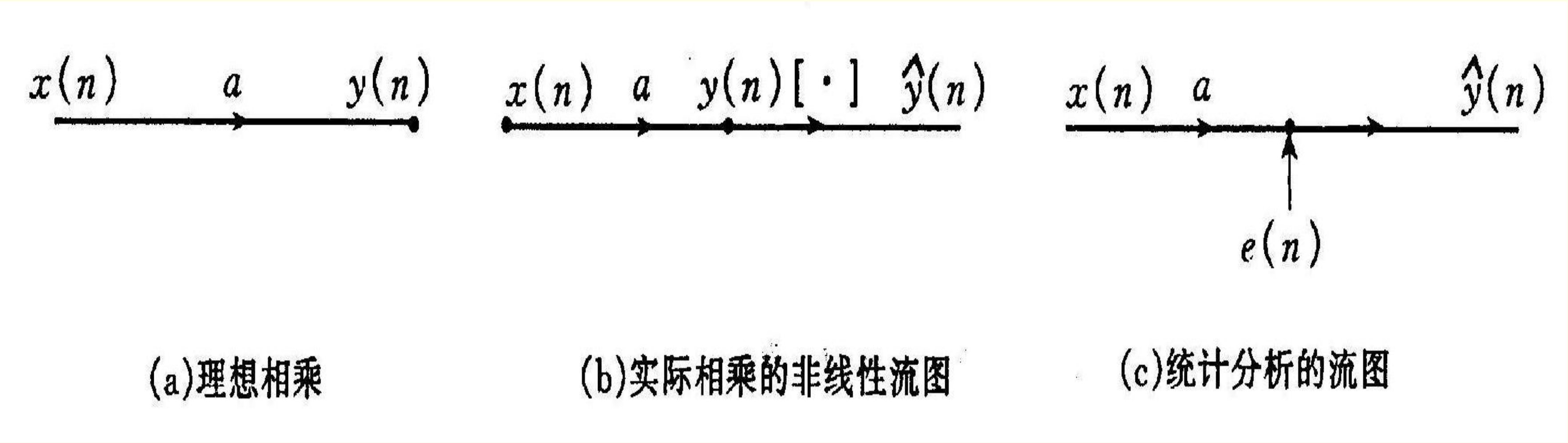
数字滤波器的实现，涉及到两种运算：相乘、求和。

定点制运算中，每一次乘法运算之后都要作一次舍入（截尾）处理，因此引入了非线性，采用统计分析的方法，将舍入误差作为独立噪声e(n)迭加在信号上，因而仍可用线性流图表示定点相乘。

本节主要分析：

(1) 乘积的舍入误差在滤波器系统中形成的噪声输出的求解方法。

(2) 同一系统而不同结构时，乘积的舍入误差在输出端的噪声输出的变化。




定点相乘运算统计分析的流图表示

# 对舍入噪声e(n)作如下的假设：

1. e(n) 为平稳随机噪声序列；
2. e(n) 与输入序列 x(n) 不相关，各噪声之间也互不相关。
3. ~~e(n) 为白色噪声；~~
4. 在量化间隔内均匀分布（即每个噪声都是均匀等概率分布）。

有了这些条件，整个系统就可作为线性系统处理。每一个噪声可用第十章所讲的线性离散系统的理论求出其输出噪声，所有输出噪声经线性迭加得到总的噪声输出。

# 1. 递归数字滤波器IIR 的定点运算量化误差

（1）以一阶IIR滤波器为例，论述乘积误差的影响

输入与输出关系可用差分方程表示为：

$$y(n) = ay(n-1) + x(n) \qquad n \geq 0, |a| < 1$$

乘积项将引入一个舍入噪声，如图

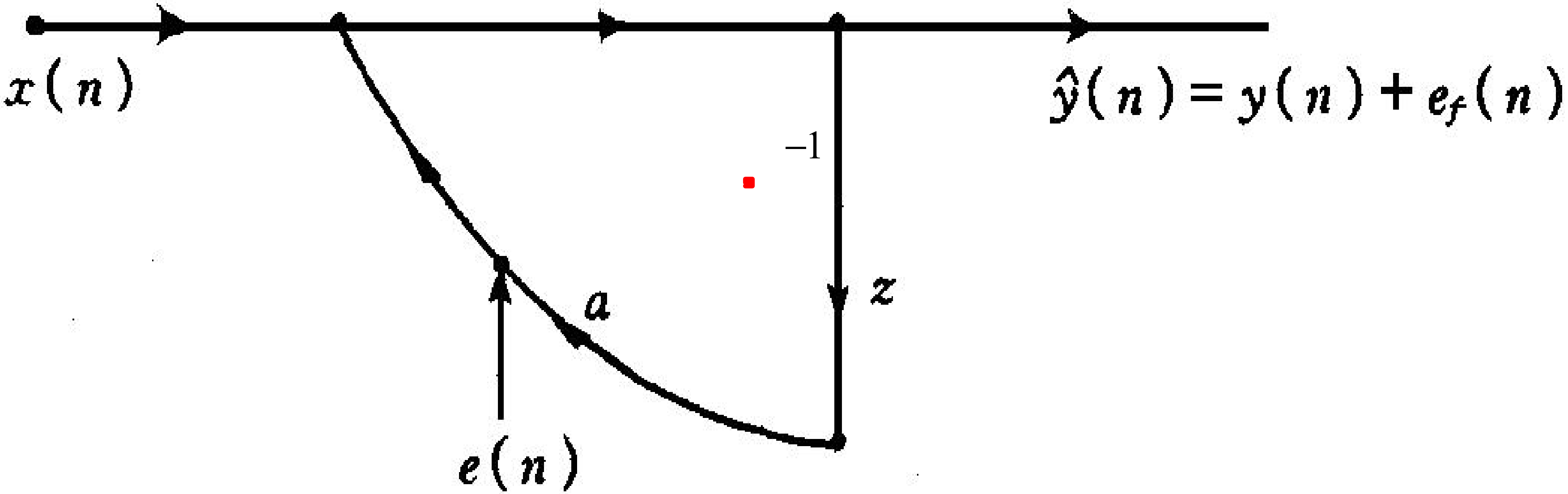
上述一阶系统的单位脉冲响应为

$$h(n) = a^n u(n)$$

系统函数为 $$H(z) = \frac{z}{z-a}$$

由于 $e(n)$ 是迭加在输入端的，故由 $e(n)$ 造成的输出误差为：$e_f = e(n) * h(n) = e(n) * a^n u(n)$

$$y(n) = ay(n-1) + x(n)$$




一阶 IIR 滤波器的舍入噪声分析

输出噪声方差

$$\sigma_f^2 = \sigma_e^2 \sum_{m=0}^{\infty} h^2(m) = \sigma_e^2 \sum_{m=0}^{\infty} a^{2m}$$

或

$$\sigma_f^2 = \frac{\sigma_e^2}{2\pi j} \oint_c H(z) H(z^{-1}) \frac{dz}{z}$$

由上两式均可求得

$$\sigma_f^2 = \frac{\sigma_e^2}{1-a^2} = \frac{q^2}{12(1-a^2)} = \frac{2^{-2b}}{12(1-a^2)}$$

可见字长 $b$越大，输出噪声越小，同样的方法可分析其它高阶DF的输出噪声。

(2)举例说明乘积的量化误差与IIR滤波器结构的关系

例：一个二阶IIR低通数字滤波器，系统函数为

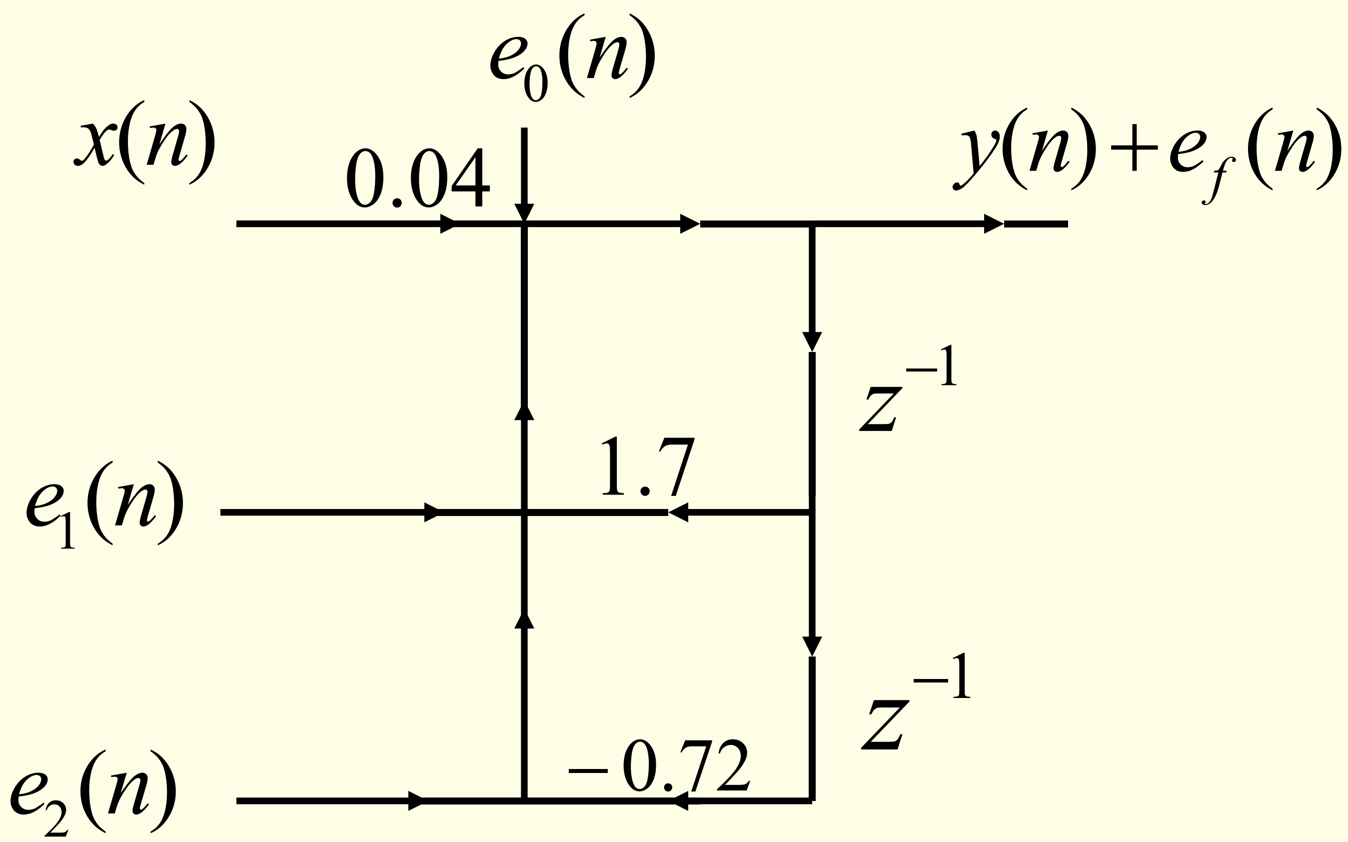
$$H(z)=\frac{0.04}{(1-0.9z^{-1})(1-0.8z^{-1})}$$

采用定点制算法，尾数作舍入处理，分别计算其直接型、级联型、并联型三种结构的舍入误差。

解：①直接型

$$H(z)=\frac{0.04}{1-1.7z^{-1}+0.72z^{-2}}=\frac{0.04}{B(z)}$$

直接型结构流图如图



图中 $e_0(n)$、$e_1(n)$、$e_2(n)$ 分别为系数0.04、1.7 、-0.72相乘后引入的舍入噪声。采用线性迭加的方法，从图上可看出输出噪声 $e_f(n)$ 是这三个舍入噪声通过网络形成的，如图b，因此

$$H_0(z)=\frac{1}{B(z)}$$

$h_0(n)$ 是$H_0(z)$的单位脉冲响应

$$e_f(n)=[e_0(n)+e_1(n)+e_2(n)]*h_0(n)$$

输出噪声的方差为：

$$\sigma_f^2 = 3\sigma_e^2 \cdot \frac{1}{2\pi j}\oint_c \frac{1}{B(z)B(z^{-1})}\frac{dz}{z}$$

将 $\sigma_e^2 = q^2/12$ 和B(z)代入，利用留数定理得：

$$\sigma_f^2 = 22.4q^2$$

②级联型

将H(z)分解

$$H(z)=\frac{0.04}{1-0.9z^{-1}}\cdot\frac{1}{1-0.8z^{-1}}=\frac{0.04}{B_1(z)}\cdot\frac{1}{B_2(z)}$$

结构流图为

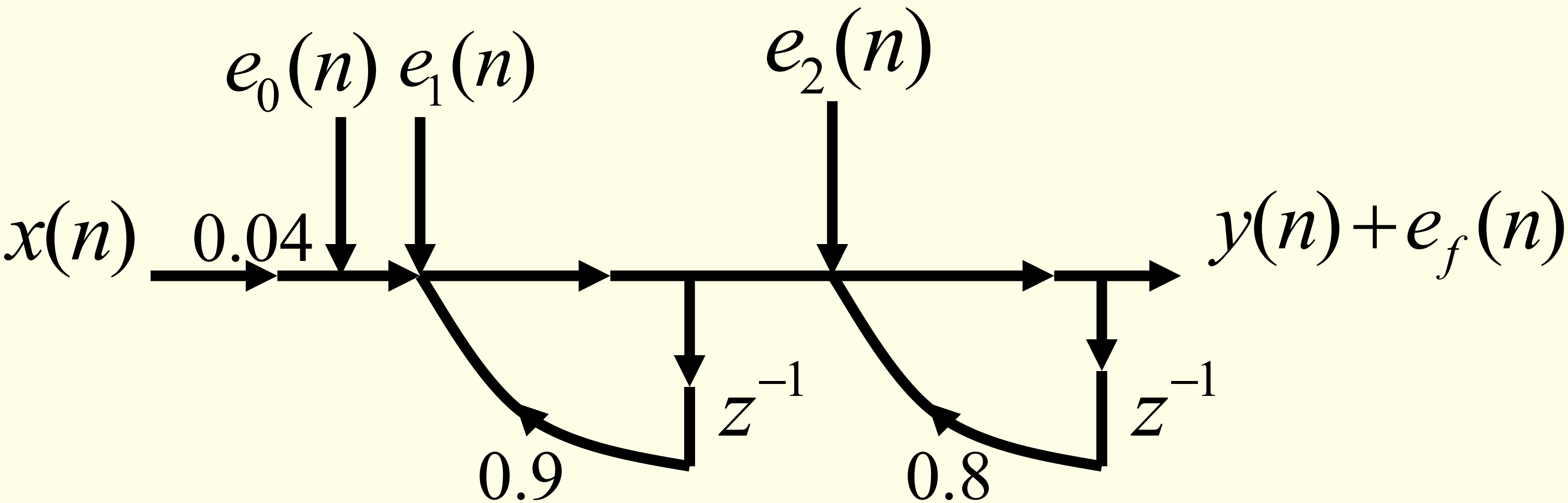



图 IIR级联型的舍入噪声分析

由图中可见，噪声$e_0(n)$、$e_1(n)$通过 $H_1(z)$ 网络，

$$H_1(z)=\frac{1}{B_1(z)B_2(z)}$$

噪声$e_2(n)$只通过网络 $H_2(z)$，

$$H_2(z)=\frac{1}{B_2(z)}$$

即

$$e_f(n)=\{e_0(n)+e_1(n)\}*h_1(n)+e_2(n)*h_2(n)$$

$h_1(n)$和$h_2(n)$分别是$\mathbf{H_1(z)}$和$\mathbf{H_2(z)}$的单位脉冲响应，

因此：

$$\sigma_f^2=\frac{2\sigma_e^2}{2\pi j}\oint_c\frac{1}{B_1(z)B_2(z)B_1(z^{-1})B_2(z^{-1})}\frac{dz}{z}$$

$$+\frac{\sigma_e^2}{2\pi j}\oint_c\frac{1}{B_2(z)B_2(z^{-1})}\frac{dz}{z}$$

将 $B_1(z)=1-0.9z^{-1}, B_2(z)=1-0.8z^{-1}, \sigma_e^2={q^2}/{12}$

代入，得：

$$\sigma_f^2=15.2q^2$$

（思考：如果将$H_1(z)$和$H_2(z)$次序颠倒，结果会怎样）

③并联型

将H(z)分解为部分分式

$$H(z)=\frac{0.36}{1-0.9z^{-1}}+\frac{-0.32}{1-0.8z^{-1}}=\frac{0.36}{B_1(z)}+\frac{-0.32}{B_2(z)}$$

其结构如图：

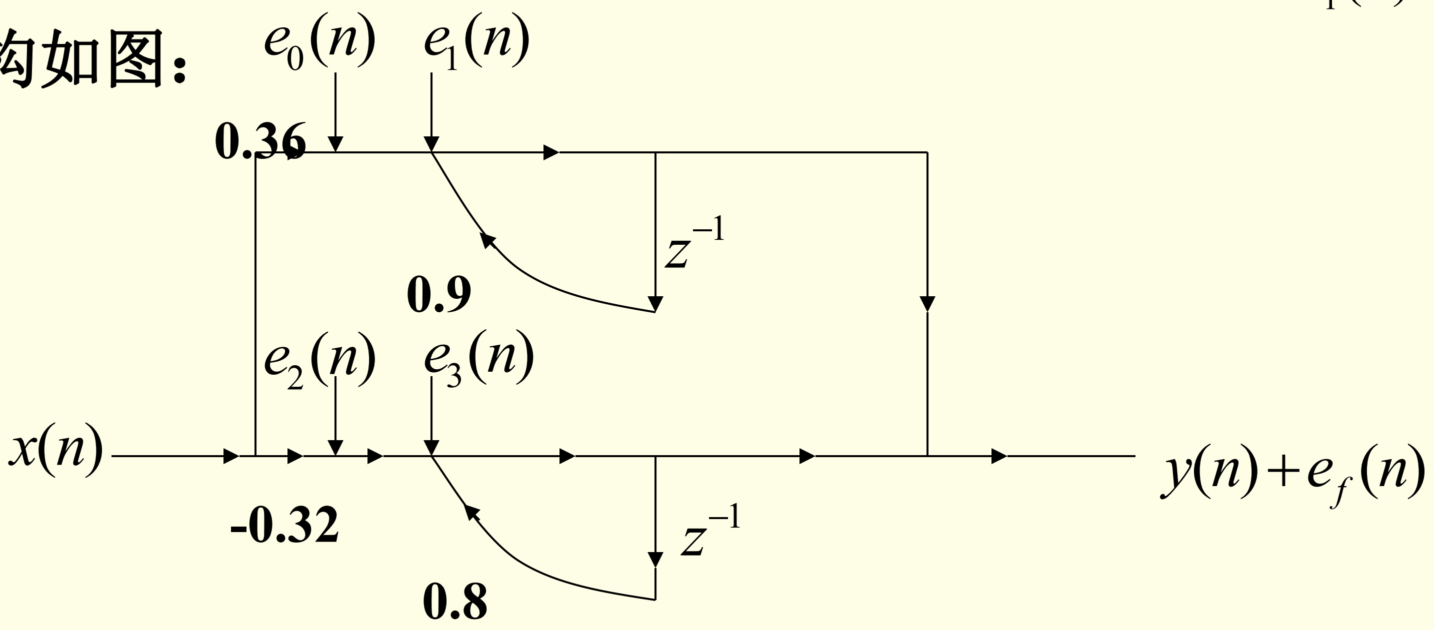



图 IIR并联型的舍入噪声分析

并联型结构有4个系数，有4个舍入噪声，其中

$[e_0(n)+e_1(n)]$ 只通过 $1/B_1(z)$ 网络，

$[e_2(n)+e_3(n)]$ 通过 $1/B_2(Z)$ 网络。

输出噪声方差为：

$$\sigma_f^2=\frac{2\sigma_e^2}{2\pi j}\oint_c\frac{1}{B_1(z)B_1(z^{-1})}\frac{dz}{z}+\frac{2\sigma_e^2}{2\pi j}\oint_c\frac{1}{B_2(z)B_2(z^{-1})}\frac{dz}{z}$$

代入$B_1(z)$和$B_2(z)$及 $\sigma_e^2$ 的值，得：

$$\sigma_f^2=1.34q^2$$

比较三种结构的误差大小，可知

直接型 > 级联型 > 并联型

原因：

- 直接型结构的所有舍入误差都经过全部网络的反馈环节，反馈过程中误差积累，输出误差很大。

- 级联型结构，每个舍入误差只通过其后面的反馈环节，而不通过它前面的反馈环节，误差小于直接型。

- 并联型结构 ：每个并联网络的舍入误差只通过本身的反馈环节，与其它并联网络无关，积累作用最小，误差最小。

该结论对IIR DF有普遍意义。

因此，从有限字长效应看，直接型（Ⅰ、Ⅱ型）结构最差，运算误差最大，高阶时避免采用。级联型结构较好。并联型结构最好，运算误差最小。

结论：IIR滤波器的有限字长效应与它的结构有关。

## 2. 非递归数字滤波器FIR的运算量化误差

IIR的分析方法同样适用于FIR滤波器，FIR滤波器无反馈环节（频率采样型结构除外），不会造成舍入误差的积累，舍入误差的影响比同阶IIR滤波器小，不会产生非线性振荡。

以横截型结构为例分析FIR的有限字长效应。

## ① 舍入噪声

N-1 阶FIR的系统函数为:

$$H(z)=\sum_{m=0}^{N-1}h(m)z^{-m}$$

无限精度下，直接型结构的差分方程为:

$$y(n)=\sum_{m=0}^{N-1}h(m)x(n-m)$$

有限精度运算时，

$$\hat{y}(n)=y(n)+e_f(n)=\sum_{m=0}^{N-1}\left[h(m)x(n-m)\right]_R$$

每一次相乘后产生一个舍入噪声

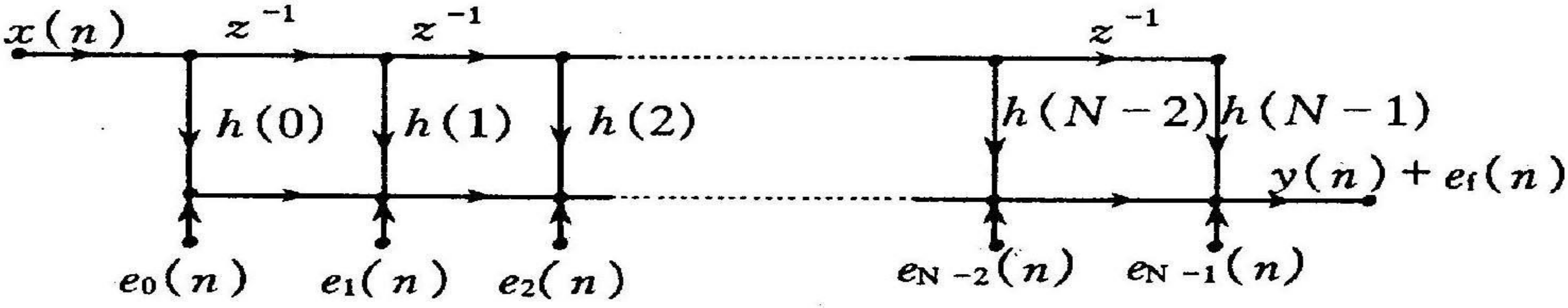
$$\left[h(m)x(n-m)\right]_R = h(m)x(n-m)+e_m(n)$$

故

$$y(n)+e_f(n)=\sum_{m=0}^{N-1}h(m)x(n-m)+\sum_{m=0}^{N-1}e_m(n)$$

输出噪声为：$$e_f(n)=\sum_{m=0}^{N-1}e_m(n)$$

如图。




横截型 FIR 滤波器的舍入噪声分析

图中可见，所有舍入噪声都直接加在输出端，因此输出噪声是这些噪声的简单和。

于是，
$$\sigma_f^2 = N\sigma_e^2 = Nq^2/12$$

输出噪声方差与字长有关，与阶数有关，N越高，运算误差越大，或者，在运算精度相同的情况下，阶数越高的滤波器需要的字长越长。

**例：FIR滤波器，N=10，b=17时**

$$\sigma_f^2=\frac{Nq^2}{12}=10\times 2^{-34}/_{12}=4.85\times 10^{-11}\qquad(-103db)$$

**N=1024时，**

$$\sigma_f^2={Nq^2}/_{12}=1024\times 2^{-34}/_{12}=4.97\times 10^{-9}\qquad(-83db)$$

$$\sigma_f=0.705*10^{-4}$$

**因此，滤波器输出中，小数点后只有4位数字是有效的。**

## 3. 极限环振荡

理解概念

在IIR滤波器中由于存在反馈环，舍入处理在一定条件下引起非线性振荡，如零输入极限环振荡。

掌握：概念、 产生的原因、克服方法。

一 、 IIR DF零输入极限环振荡

条件：
1。存在反馈
2.舍入处理

量化处理是非线性的，在DF中由于运算过程中的尾数处理，使系统引入了非线性环节，数字滤波器变成了非线性系统。对于非线性系统，当系统存在反馈时，在一定条件下会产生振荡，数字滤波器也一样。

IIR滤波器是一个反馈系统，在无限精度情况下，如果它的所有极点都在单位圆内，这个系统总是稳定的，当输入信号为零后，IIR 数字滤波器的响应将逐步变为零。

但同一滤波器，以有限精度进行运算时，当输入信号为零时，由于舍入引入的非线性作用，输出不会趋于零，而是停留在某一数值上，或在一定数值间振荡，这种现象为“零输入极限环振荡”。

例：设一阶IIR DF的系统函数为：

$$H(z)=\frac{1}{1-az^{-1}}$$

无限精度运算时，差分方程为：

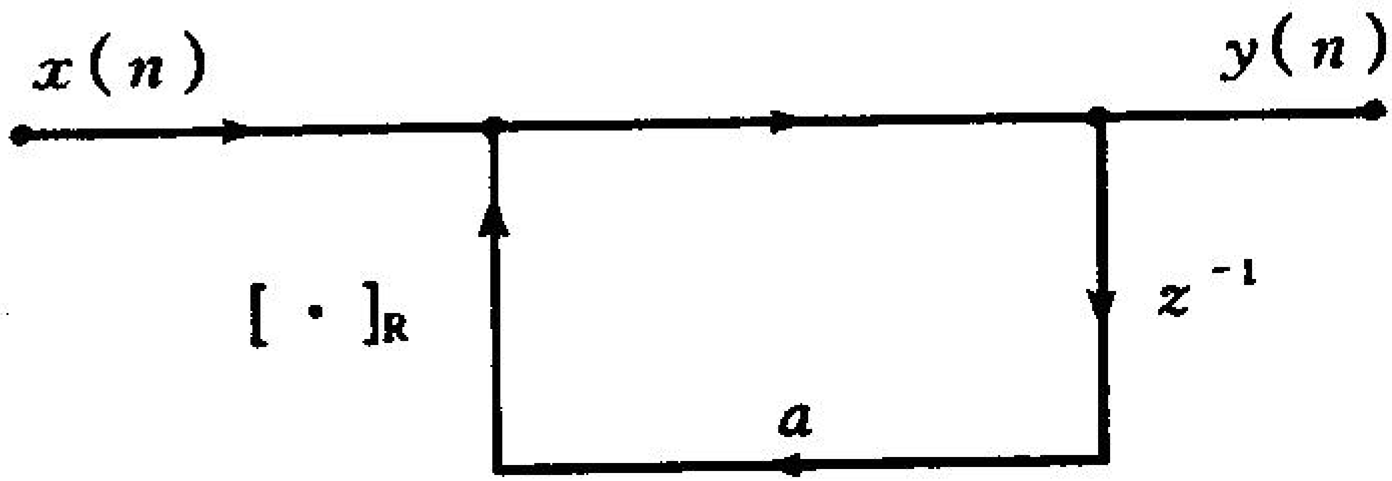
$$y(n)=ay(n-1)+x(n)$$

在定点制中，每次乘法运算后都必须对尾数作舍入处理，这时的非线性差分方程为：

$$\hat{y}(n)=\left[a\,\hat{y}(n-1)\right]_R+x(n)$$

（有限精度）

$[\cdot]_R$表示舍入运算，上述运算过程的非线性流图 如图。




一阶 IIR 网络的非线性流图

若输入为

$$x(n)=\begin{cases}7/8 & n=0\\ 0 & n>0\end{cases}$$

字长 b=3，系数 a=0.100。

无限精度时，系统的极点为 z=a=0.5<1，在单位圆内，系统稳定。

若输入变为零，输出也逐渐衰减到零，

$$y(n)=\frac{7}{8}\times 0.5^{n}\qquad n>0$$

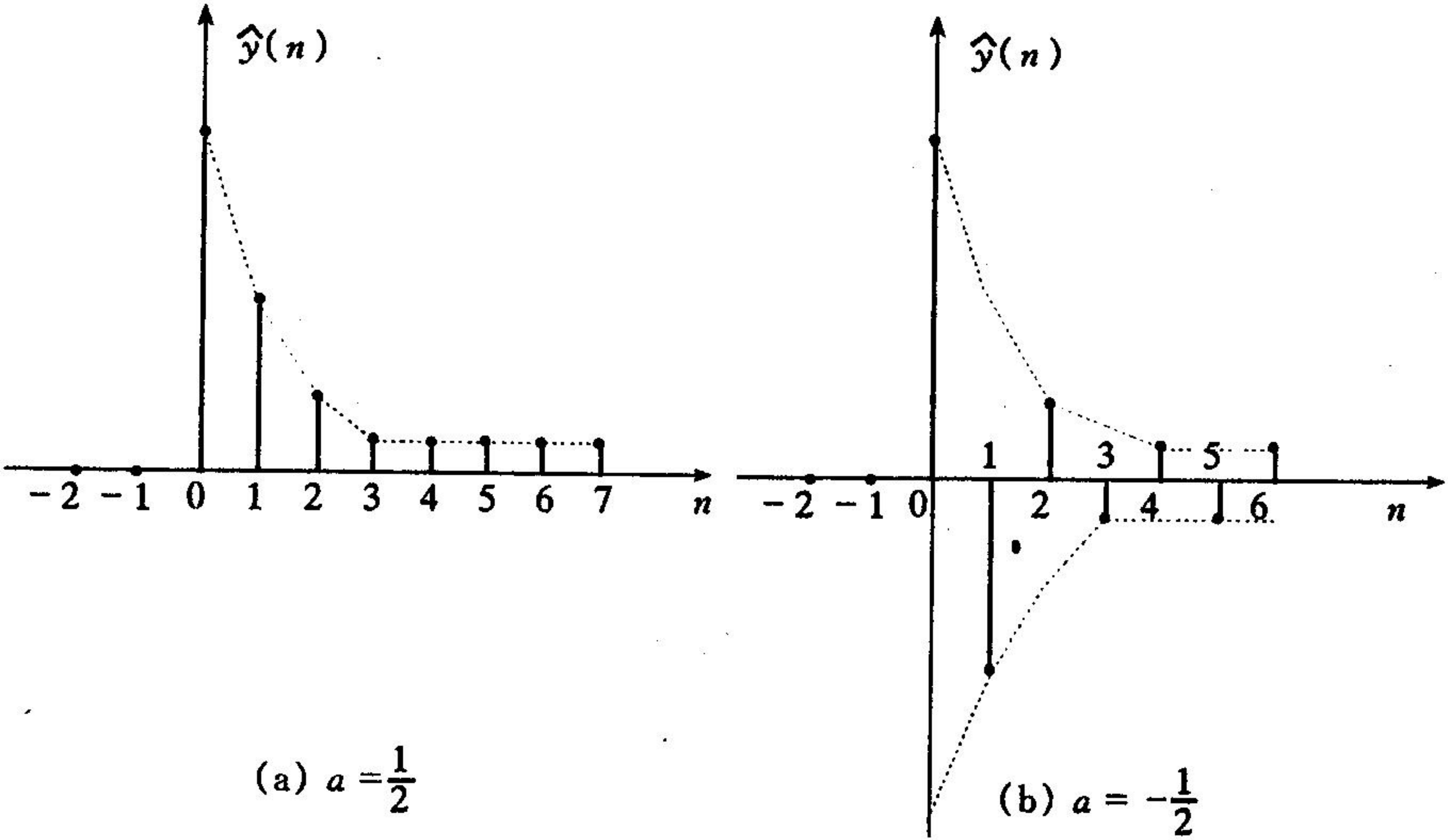
但有限精度时，由于舍入处理，系统可能会进入死区。

下面是非线性差分方程的运算结果，

| n | x（n） | $\hat{y}(n-1)$ | $a\hat{y}(n-1)$ | $[a\hat{y}(n-1)]_R$ | $\hat{y}(n)$ |
|---|---|---|---|---|---|
| 0 | 0.111 | 0.000 | 0.0000 | 0.000 | 0.111(7/8) |
| 1 | 0.000 | 0.111 | 0.0111 | 0.100 | 0.100(1/2) |
| 2 | 0.000 | 0.100 | 0.0100 | 0.010 | 0.010(1/4) |
| 3 | 0.000 | 0.010 | 0.0010 | 0.001 | 0.001(1/8) |
| 4 | 0.000 | 0.001 | 0.0001 | 0.001 | 0.001(1/8) |

……

可见，输出停留在y（n）=0.001上再也衰减不下去了，如图（a），y（n）=0.001以下也称为“死带”区域，如果系数a=-0.5，为负数，则每乘一次a就改变一次符号，因此输出将是正负相间的,如图（b），这时y（n）在±0.125之间作不衰减的振荡，这种振荡现象就是“零输入极限环振荡”。




**零输入极限环振荡**

振荡产生的原因：

考察上述非线性差分方程的运算结果，在最后一行，当 $\hat{y}(n-1)$ =0.001时，$a\hat{y}(n-1)$ =0.0001，经舍入处理后又进位为 $[a\hat{y}(n-1)]_R$=0.001 ,仍与 $\hat{y}(n-1)$ 的值相同， 因此输出保持不变。

这可解释为，只要满足 $\left|[a\hat{y}(n-1)]_R\right| = \left|\hat{y}(n-1)\right|$ 时，舍入处理使系数 a 失效，或者说相当于将 a 换成了一个绝对值为1的等效系数 $a'$，$a' = a/|a|$，这时 $H(z) = \dfrac{1}{1 \pm z^{-1}}$ 极点等效迁移到单位圆上，系统失去稳定，出现振荡。

极限振荡幅度与字长的关系：

$$\because \left| \left[ a\hat{y}(n-1) \right]_R - a\hat{y}(n-1) \right| \le \frac{q}{2}$$

$$\therefore \left| \hat{y}(n-1) \right| \le \frac{q/2}{1-|a|}$$

- 极限环振荡的幅度与量化阶成正比；与极点位置和滤波器阶数有关；
- 增加字长，可减小极限环振荡。

高阶IIR网络中，同样有这种极限环振荡现象，但振荡的形式更复杂。不一一讨论。

# 本章总结

- 数的表示对量化误差的影响
- A/D转换的量化误差分析
- 数字滤波器的系数量化误差
- 数字滤波器的运算量化误差