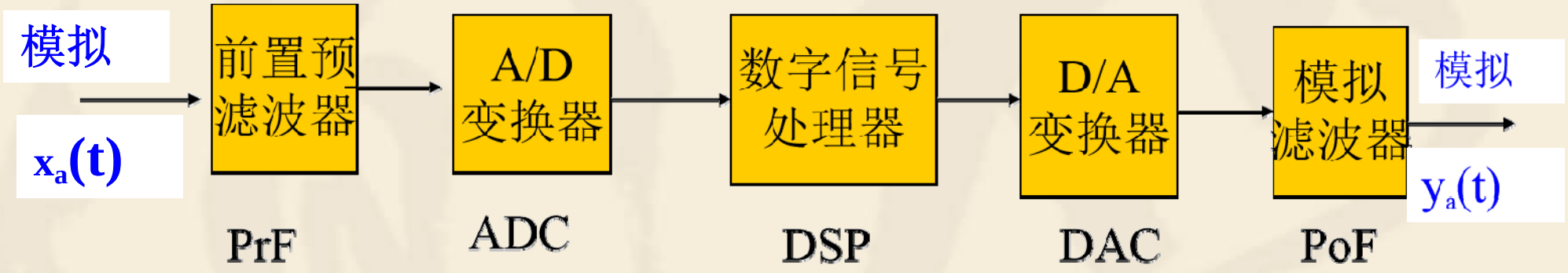
数字信号处理

Digital Signal Processing

第 6 章 数字信号处理中的有限字长效应

数字信号处理中的有限字长效应



- 一个线性、非移变、因果系统的差分方程为：

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

- 在以前的讨论中，认为：
  - 1) 系统的输入序列x(n)
  - 2) 系统的输出序列y(n)
  - 3) 系数都是连续变化的，即具有无限精度。

数字信号处理中的有限字长效应

- 在实际中，无论软件、硬件，都只能用有限字长来表示。这样就会对系统的特性产生一定影响，这就是有限字长效应问题，主要有三个方面的误差：

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

1) A/D 变换中的量化效应**(6.2)**　无限精度模拟信号---有限精度数字信号

2) 系数的量化效应**(6.3)**：零极点位置改变、频率响应变化

3) 运算过程中的尾数处理**(6.4)**：舍入或截尾处理的噪声

数字信号处理中的有限字长效应

- 研究有限字长的目的：

1. 已知字长，可进行误差分析，判断结果可信度，决定是否采取改进措施。 一般情况下，由于计算机字长较长，所以可以不考虑字长的影响。
2. 根据设计要求所需精度确定DSP芯片类型
   - 根据精度确定最小字长。
   - 由最小字长选用DSP芯片类型

由于选用不同DSP芯片，价格差很大。目前TMS320C1X,C2X,C5X,C54X,C62X,C67x等价格差异很大

■ 由于一些问题，还没有系统的办法解决，本章只对一些问题作一些概括的介绍。

■ 本章的目的:

① 认识这三类误差

② 了解这三类误差的大小与采用的数的长度（**ADC**位数、信号字长，系数字长）、与数的表示（数制、码制、量化方式）、与滤波器的结构型式有什么关系

③ 掌握分析误差的实用方法：分别考虑、统计分析

6.1 数的表示及其对量化的影响：数的表示

二进制表示：

任意数　可以表示成 γ 进制数

$$x=\sum b_i\gamma^i,\qquad 0\le b_i\le\gamma-1$$

$$=(b_nb_{n-1}\cdots b_{0\Delta}b_{-1}b_{-2}\cdots b_{-m})_\gamma$$

常数：

十进制　$\gamma=10,\quad x=\sum b_i\cdot 10^i,\ 0\le b_i\le 9$

二进制　$\gamma=2,\quad x=\sum b_i\cdot 2^i,\ b_i=0,1$

十进制数转化为二进制数的步骤：

1）用2反复去除整数部分，并将所得余数以逆序排列

2）用2乘小数部分，并舍去所得的整数部分，这样重复若干次，然后将所得的整数（0或1）用正序排列。

$$x=18.375_{10}\rightarrow 10010_\Delta 011_2$$

$$x=11_\Delta 101_2\rightarrow 1\cdot(2^1)+1\cdot(2^0)+1\cdot(2^{-1})+0\cdot(2^{-2})+1\cdot(2^{-3})=3.625$$

6.1 数的表示及其对量化的影响：数的表示

- 数的表示：（数制）
  - 定点制：动态范围小，精度低，有限字长效应突出
  - 浮点制：动态范围大，精度高。
  - 成组浮点：
- 定点制：（码制）原码、反码、补码
  - 负数
    - 反码：原码逐位取反
    - 补码：反码加“1”
  - 正数：三者一样
- 量化方式：舍入、截尾
- 运算：加法、乘法、延迟（结果不需要量化）

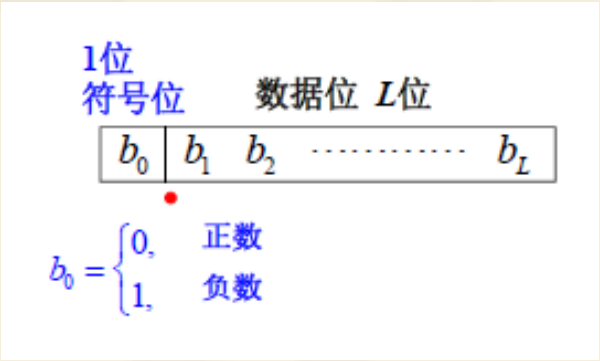
6.1 数的表示及其对量化的影响：定点表示

6.1.2 定点表示：

1. 定义： 用一定字长的二进制位表示，其中

小数点的位置固定（规定在符号位（整数位）与数据位之间）。

纯小数　-1 ＜ M ＜ 1



字长L+1　　$x = b_0 \quad \Delta \quad b_1 b_2 \cdots b_L$

↑　↑　↑

符号位　二进制位　数据位

注：1）运算中数的最大值不超过1（加法）

2）通常采用小数，避免带分数或者整数的原因是：

原因：带分数难于进行乘法运算；

表示一个整数的二进制数的位数不可以用舍入或者截尾的方式进行缩减。

6.1 数的表示及其对量化的影响：

----定点值的原码、补码和反码表示

b_0 小数点 $b_1\ b_2\ \cdots\ b_L$

		正数	例：$\frac{3}{8}$	负数	例：$-\frac{3}{8}$
原码	$x=(-1)^{b_0}\cdot\sum_{i=1}^{L}b_i\cdot 2^{-i}$ （6.3）	0 绝对值	0.011	1 绝对值	1.011
补码	$x=-b_0+\sum_{i=1}^{L}b_i\cdot 2^{-i}$ （6.4）	0 绝对值	0.011	1 绝对值每位求反，末位加1	1.101
反码	$x=-b_0\cdot(1-2^{-L})+\sum_{i=1}^{L}b_i\cdot 2^{-i}$	0 绝对值	0.011	1 绝对值每位求反	1.100

正数　原码表示 = 补码表示 = 反码表示

负数　同一数值（如，十进制 –3/8）要用不同的码字表示

从最右位开始向左，出现‘1’的第一位，从第一位（不含）向左其余取补

反码+1

原码：乘除简单：符号位逻辑加（异或），尾数相乘

加法需判断加数、被加数符号，决定两数次序及和的符号

减法：判断两数绝对值的大小，以便大减小

补码：加法容易，正负直接相加，符号位同样参加运算，符号进位，把进位1去掉，留下其他结果。

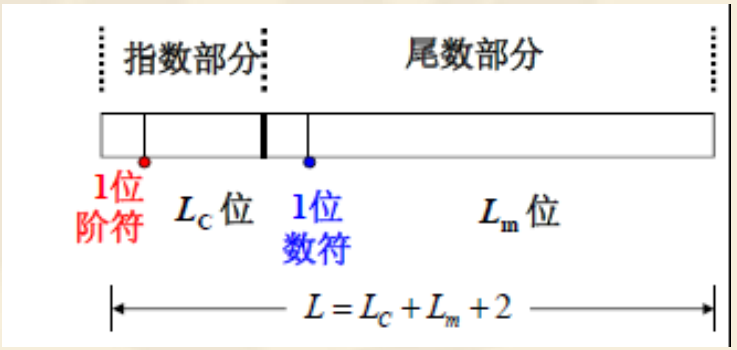
乘法前要先变为原码表示。

反码：应用少

6.1 数的表示及其对量化的影响：浮点表示

$x = \pm M \times 2^{\pm c}$,

M尾数（$\frac{1}{2} \le M < 1$），c为指数（2的幂指数）



尾数采用规格化尾数： 尾数右移一位代表除2
左移一位代表乘2

尾数字长L_m 决定数的精度　　阶码C 的字长L_C 决定数的范围。

阶码和尾数均为补码时的数值范围：　$2^{-2^{L_C}} \frac{1}{2} \le |x| < 2^{(2^{L_C}-1)}(1-2^{-L_m})$

L_C位二进制能表示最大负值　　L_C位二进制能表示最大正值

乘法： 两数阶码相加，尾数相乘；再规格化
加法： 若两数阶码相同，则和的阶码＝两数的阶码

若两数阶码不同，先对阶，再相加

例：求两数F1 和F2之积，

$F_1 = 4,\ F_2 = 5/4, F_1 = 2^{c_1} M_1, F_2 = 2^{c_2} M_2$

$c_1 = 11_\Delta$(二进制的3)

$M_1 =_\Delta 10000$(二进制的1/2)

$c_2 = 01_\Delta$(二进制的1)

$M_2 =_\Delta 10100$(二进制的5/8)

$c_1 + c_2 = 100_\Delta$(二进制的4)

$M_1 M_2 =_\Delta 01010 < \frac{1}{2}$

标准化：$M_1 M_2 =_\Delta 10100$

$c_1 + c_2 = 11_\Delta$

$M_{1\Delta} 10000, M_2 = 1_\Delta 10100$，求$M_1 M_2$

$$
\begin{array}{r}
{}_\Delta 100 \\
+\ {}_\Delta 101 \\
\hline
100 \\
000 \\
100 \\
\hline
010100
\end{array}
$$

$M_1 M_2 = \Delta 01010$

例：求两数F1 和F2之和，

$F_1 = 4,\ F_2 = 5/4, F_1 = 2^{c_1} M_1, F_2 = 2^{c_2} M_2$

$c_1 = 11_\Delta$(二进制的3)

$M_1 =_\Delta 10000$(二进制的1/2)

$c_2 = 01_\Delta$(二进制的1)

$M_2 =_\Delta 10100$(二进制的5/8)

改变c_2使之等于c_1，相应调整M_2

$F_2 = 2^{\hat{c}_2} \hat{M}_2$,其中$\hat{c}_2 = 11,\ \hat{M}_2 =_\Delta 00101$,

尾数相加，得到$c = 11,\ M =_\Delta 10101$

6.1 数的表示及其对量化的影响：浮点定点比较

浮点数的特点：

数值范围大

运算复杂、速度慢

加法、乘法后均需尾数量化

加法前需对阶

定点数的特点：

数值动态范围小 $-1 < M < 1$

运算简单快速

乘法后需进行量化

加法前需防溢出

$x = \pm M \times 2^{\pm c}$,

M尾数（$\frac{1}{2} \leq M < 1$），c为指数（2的幂指数）

6.1 数的表示及其对量化的影响：量化误差

- 定点制中数的位数——寄存器的长度决定，如L+1位，可以表示的最小正数为 2^{-L} ——量化间距。
- 如果要处理的数为 M+1位（含符号位）>L+1，则必须要进行量化。

  截尾：低位数截短

  舍入：在数据的L+1位上加1，然后截短
- 量化引入误差e：

  $$e = Q[x] - x$$
- e的范围取决于数的大小、表示形式（字长和码制）及量化方法，Q[x]为x的量化值。

6.1 数的表示及其对量化的影响：量化误差

6.1.4 定点制数的量化

设量化前$M+1$位，量化后$L+1$位，$M>L$

量化误差： 大小与字长、码制、量化方式有关

（1）正数截尾 截尾：保留L位，丢掉后面的$M-L$位
正数的原码、补码、反码表示相同

截尾前 $x=\sum_{i=1}^{M} b_i \cdot 2^{-i},\ b_i=0,1$

截尾后 $Q_t[x]=\sum_{i=1}^{L} b_i \cdot 2^{-i}$

截尾误差 $e_t=Q_t[x]-x=-\sum_{i=L+1}^{M} b_i \cdot 2^{-i}<0, b_i=0, or\ 1$

误差范围 $-(2^{-L}-2^{-M})e_t \le 0, x>0$

近似 $-q \le e_t \le 0, x>0$

量化间隔 $q=2^{-L}$，是保留的尾数中最低码位的位权

负数的截尾误差

（2）原码负数截尾

截尾前 $x=-\sum_{i=1}^{M}b_i2^{-i}$

截尾后 $Q_t[x]=-\sum_{i=1}^{L}b_i2^{-i}$

截尾误差 $e_t=\sum_{i=L+1}^{M}b_i2^{-i}$

误差范围 $0\le e_t\le\left(2^{-L}-2^{-M}\right)$

简记 $0\le e_t<q,\quad x<0$

（3）补码负数截尾

$x=-1+\sum_{i=1}^{M}b_i2^{-i}$

$Q_t[x]=-1+\sum_{i=1}^{L}b_i2^{-i}$

$e_t=-\sum_{i=L+1}^{M}b_i2^{-i}$

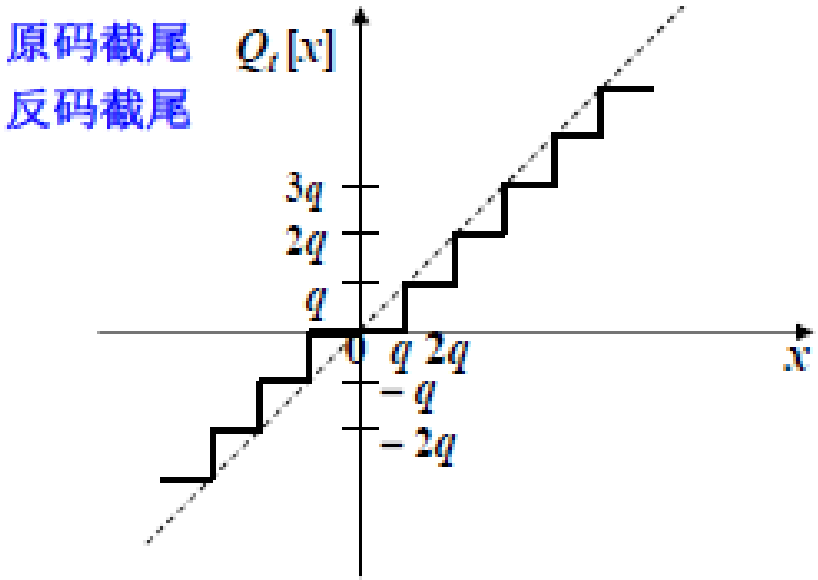
$-(2^{-L}-2^{-M})\le e_t\le 0$

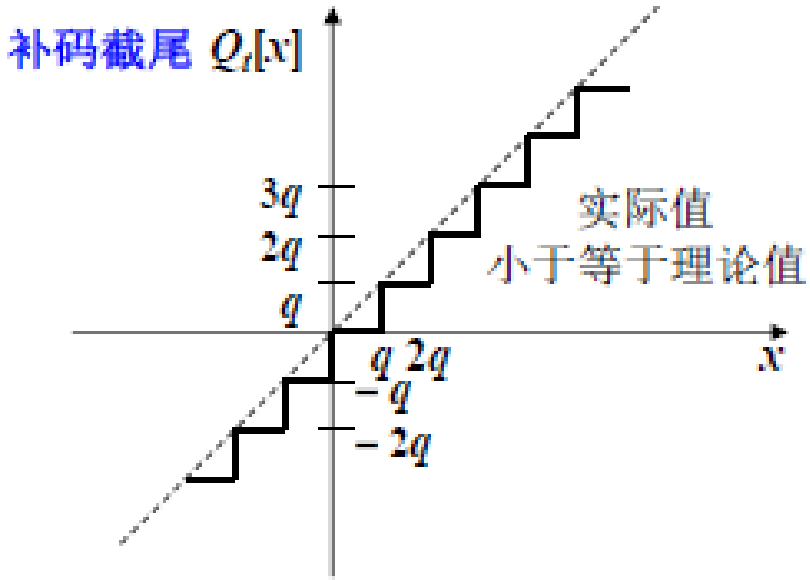
$-q<e_t\le 0,\quad x<0$

（4）反码负数截尾

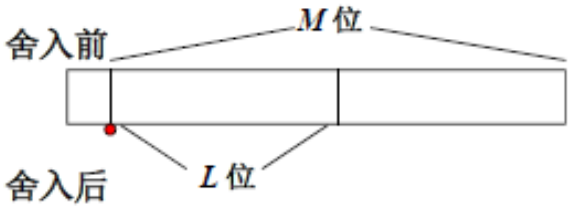
（略）

$0\le e_t<q,\quad x<0$



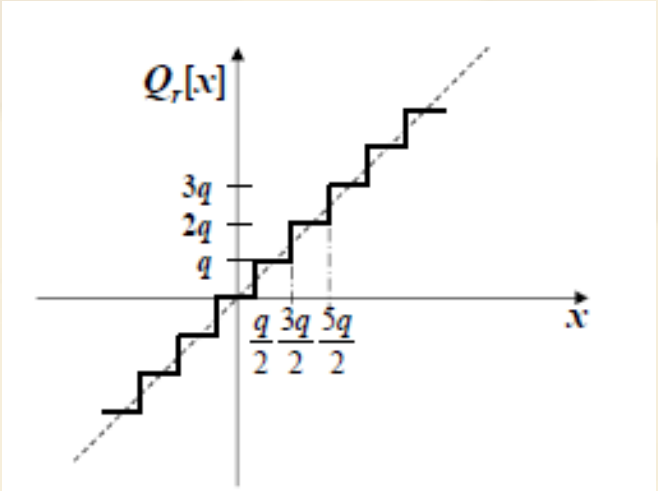


（5）舍入： 不分正负数，不分原码*I* 补码*I* 反码，按最接近的值取*L* 位



舍入误差 $e_r = Q_r[x] - x$ （6.10）

误差范围 $-\frac{q}{2} < e_r \le \frac{q}{2}$ （6.11）



舍入误差的数值较小，
与输入信号的极性无关，
是对称分布的。

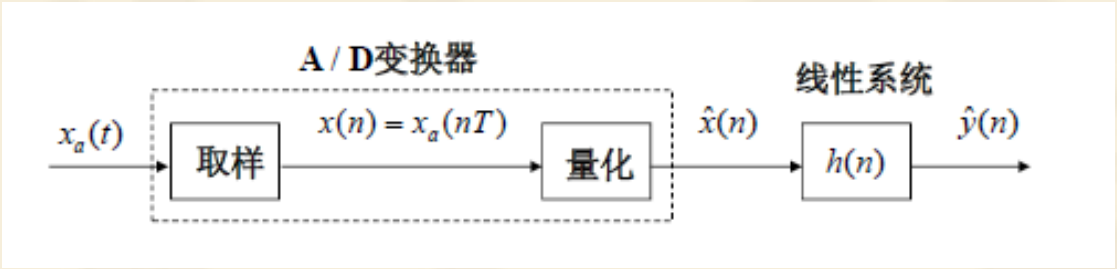
截尾误差的数值较大，
与输入信号的极性相关，
是单极性分布的。

6.1 数的表示及其对量化的影响：数的表示

- 数的表示：（数制）
  - 定点制：动态范围小，精度低，有限字长效应突出
  - 浮点制：动态范围大，精度高。
  - 成组浮点：
- 定点制：（码制）原码、反码、补码
  - 负数
    - 反码：原码逐位取反
    - 补码：反码加“1”
  - 正数：三者一样
- 量化方式：舍入、截尾
- 运算：加法、乘法、延迟（结果不需要量化）

6.2 A/D变换的字长效应

- 单独考虑信号的量化效应，假定系统系数、运算是没有量化的



分两步讨论：1. 无限精度的信号通过**A / D**变换器（ 字长***L* + 1** 位）；
2. 量化信号通过线性系统

讨论的目的：选择码制的依据？
怎样选择**A / D** 变换器字长，以满足信噪比指标？

6.2 A/D变换的字长效应：量化效应的统计分析

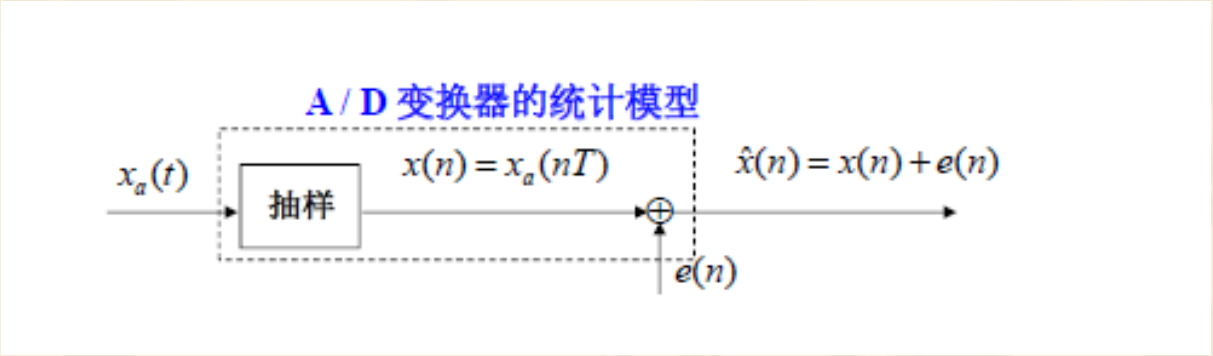
6.2.1 A / D 变换量化误差的统计分析

量化误差（序列） $e(n)=x(n)-\hat{x}(n)$

$e(n)$与输入信号有关，不能确切知道。

量化（舍入或截尾）是非线性运算，不便于分析

将量化误差看作随机变量，运用统计方法来分析



- 为了对此模型进行统计分析，假定：

(1)e(n) 是一个平稳随机序列，均值与时间无关

(2)e(n) 与信号 序列**x (n)** 不相关

(3)e(n) 任意两个抽样间不相关，即为白噪声过程。

(4)e(n) 具有等概率密度分布（在一定的量化间距上）

平稳随机过程

平稳随机过程的概念及数字特征

一、严平稳过程

设$\{X(t), t \in T\}$为一随机过程，若对 T内的任意 n个值 $t_1, t_2, \cdots, t_n$对任意实数 $\tau', t_1+\tau', t_2+\tau', \cdots, t_n+\tau' \in T$可使 n维随机变量 $[X(t_1), X(t_2), \cdots, X(t_n)]$与 $[X(t_1+\tau'), X(t_2+\tau'), \cdots, X(t_n+\tau')]$有相同的分布，即 $\{X(t), t \in T\}$的分布满足

$$F_n(x_1, x_2, \cdots, x_n; t_1, t_2, \cdots, t_n) = F_n(x_1, x_2, \cdots, x_n; t_1+\tau, t_2+\tau, \cdots, t_n+\tau)$$

则称 $\{X(t), t \in T\}$为严平稳过程。

以上定义的是离散型的严平稳过程。对连续型，则用概率密度来描述→

如果一个随机过程m阶矩以下的矩的取值全部与时间无关，则称该过程为m阶平稳过程。特别（见以下二阶平稳过程）：

宽平稳过程　随机过程$\{X_1、X_2、\ldots、X_T\}$的均值函数、方差函数均为常数，自协方差仅是时间间隔$t-s$的函数。即：

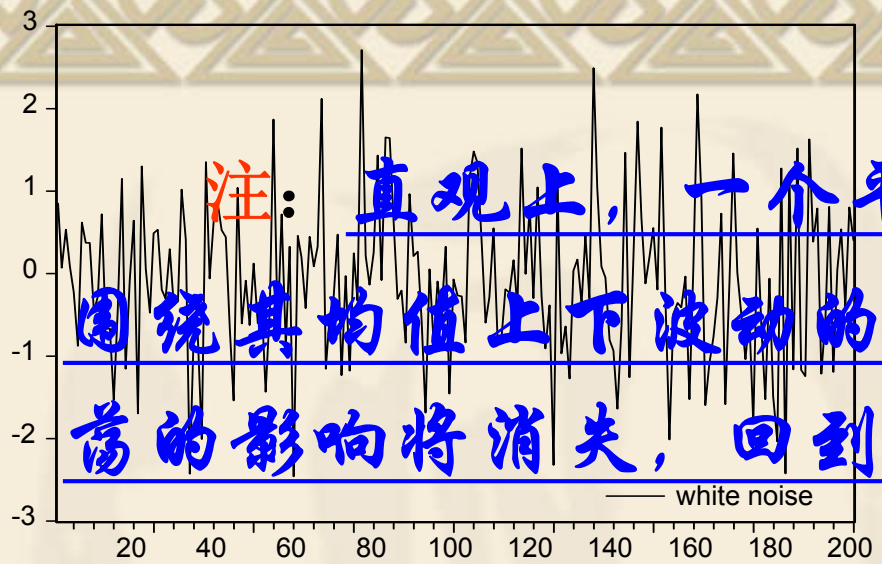
$$E(X_t)=E(X_{t+h})=\mu<\infty \qquad (m\text{为常数；}\ h\text{为任意数})$$

$$Var(X_t)=\sigma^2<\infty \qquad (\sigma^2\text{为常数})$$

$$Cov(X_t,X_s)=Cov(X_{t+h},X_{s+h})=\sigma_{ij}<\infty \qquad t,s\in T$$

则称$X(t)(t\in T)$为二阶宽(广义)平稳过程。

注：平稳随机过程的均值和方差是固定不变的，自协方差只与考察的时间间隔长度有关，而与时间t无关。对平稳过程而言，任何震荡的影响都是暂时的。随着时间t 的推移，影响将消失，回到长期平均水平。

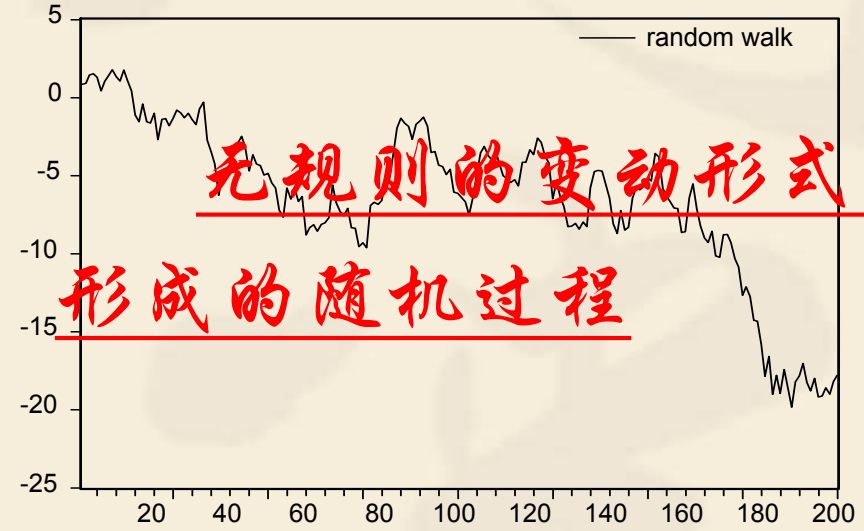


由白噪声过程产生的时间序列

日元对美元汇率的收益率序列

注：直观上，一个平稳的时间序列可以看作一条围绕其均值上下波动的曲线。随着时间t的推移，震荡的影响将消失，回到长期平均水平。

白噪声是（宽）平稳随机过程，均值为零，方差不随时间变化



由随机游走过程产生时间序列

日元对美元汇率（300 天，1995 年）

无规则的变动形式，如同一个人酒后乱步，所形成的随机过程

$Y_t = Y_{t-1} + u_t$，$u_t \sim N(0, \sigma^2)$ 为白噪声序列

随机游走过程是非平稳随机过程，

均值为零、方差随时间变化

科学研究方法：复杂问题简单化

“把复杂的东西简单化，可以发现新定律。”——牛顿

“万物之始，大道至简，衍化至繁。”——老子《道德经》

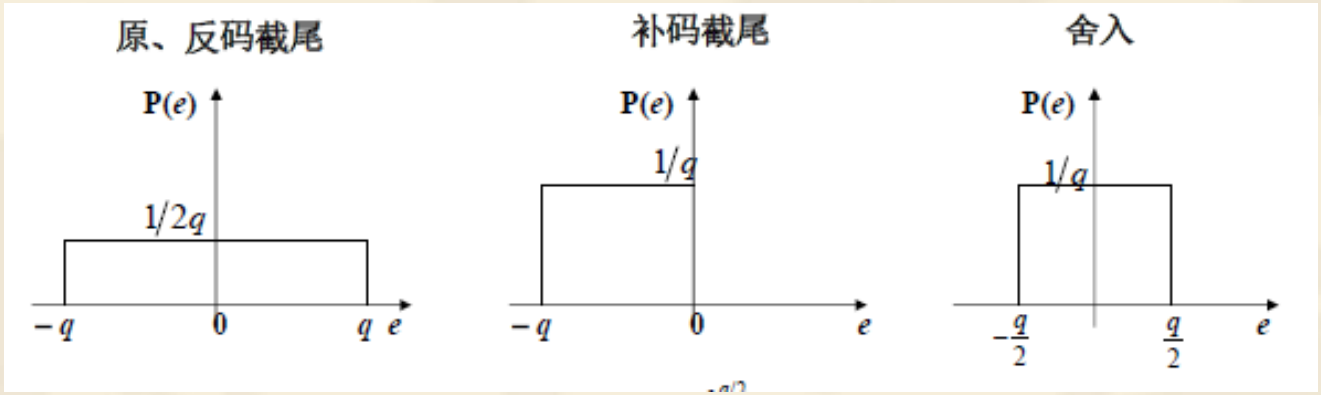
牛顿第二定律：F=ma，揭示了在远低于光速条件下，质量、加速度与力之间的本质联系。

胡克定律：s=Ee，对线弹性材料，材料中的应力和应变满足线性关系。

非线性系统线性化：**Kalman filter，Extended Kalman filter**

6.2 A/D变换的字长效应：量化效应的统计分析

量化噪声的概率密度函数



（1）定点舍入量化

均值 $m_e = E[e(n)] = \int_{-q/2}^{q/2} e \cdot p(e) \cdot \mathrm{d}e = 0$

方差 $\sigma_e^2 = E\{[e(n) - m_e]^2\} = \int_{-q/2}^{q/2} [e - m_e]^2 \cdot p(e) \cdot \mathrm{d}e$

$$= \int_{-q/2}^{q/2} e^2 \cdot \frac{1}{q} \cdot \mathrm{d}e = \frac{q^2}{12} = \frac{2^{-2L}}{12} \qquad (6.13)$$

（2）定点补码截尾量化

均值 $m_e = \int_{-q}^{0} e \cdot \frac{1}{q} \cdot \mathrm{d}e = -\frac{q}{2} = -\frac{2^{-L}}{2}$

方差 $\sigma_e^2 = \int_{-q}^{0} \left[e + \frac{q}{2}\right]^2 \cdot \frac{1}{q} \cdot \mathrm{d}e = \frac{q^2}{12} = \frac{2^{-2L}}{12}$ （6.14）

6.2 A/D变换的字长效应：量化效应的统计分析

由给定的信噪比指标来确定必需的字长

$$\hat{x}(n) = x(n) + e(n)$$

舍入量化噪声序列：均值：$m_e = 0$

均方差（能量）：$E[e^2(n)] = \sigma_e^{\ 2}$

信噪比（**signal to noise ratio, SNR**）：$\dfrac{\text{信号能量}}{\text{噪声能量}}$或$\dfrac{\text{信号功率}}{\text{噪声功率}}$

$$\frac{\sigma_x^{\ 2}}{\sigma_e^{\ 2}} = \frac{\sigma_x^{\ 2}}{2^{-2L}/12} = 12 \times 2^{2L} \sigma_x^{\ 2}$$

对数形式

$$SNR = 10\log_{10}\left(\frac{\sigma_x^{\ 2}}{\sigma_e^{\ 2}}\right) = 10\log_{10}12 + 2L \times 10\log_{10}2 + 10\log_{10}\left(\sigma_x^{\ 2}\right)$$

$$= 10.79 + 6.02L + 10\log_{10}\left(\sigma_x^{\ 2}\right)$$

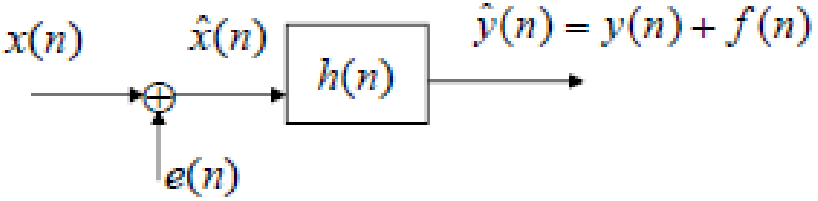
字长增加**1** 位，**SNR** 增加**6 dB**

6.2 A/D变换的字长效应：量化噪声通过线性系统

- 当已量化的信号通过一线性系统，实际的输入信号为：

$$\hat{x}(n) = x(n) + e(n)$$

$$\begin{aligned}\hat{y}(n) &= \hat{x}(n) * h(n) \\ &= [x(n) * h(n)] + [e(n) * h(n)] \\ &= y(n) + f(n)\end{aligned}$$



舍入误差序列

$$f(n) = e(n) * h(n) = \sum_{m=0}^{\infty} h(m)e(n-m)$$

$$m_f = E[f(n)] = E[\sum_{m=0}^{\infty} h(m)e(n-m)]$$

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

6.2 A/D变换的字长效应：量化噪声通过线性系统

- 因为序列***e(n)***本身任意两个值之间是不相关的，则

$$E[e(n-m)e(n-l)]=\begin{cases}q^2/12=\sigma_e^{\ 2}, l=m, q=2^{-L}\\ 0, l\neq m\end{cases}$$

- 因此，当冲激响应为实序列时，有

$$\sigma_f^2=E[f^2(n)]=\sigma_e^2\sum_{m=0}^{\infty}h^2(m)=\sigma_e^2\frac{1}{2\pi j}\oint_c z^{-1}H(z)H(z^{-1})dz$$

$$=\frac{\sigma_e^2}{2\pi}\int_{-\pi}^{\pi}|H(e^{jw})|^2dw$$

与信号的量化字长***L*** 有关，
与系统特性***h*(*n*)**、***H*(z)**有关

6.2 A/D变换的字长效应：量化噪声通过线性系统

- 由于数字系统中 $\sum_{m=0}^{\infty} h^2(m)$ 代表系统的能量，是个定值，所以：

$$\sigma_f^{\,2} = k'\sigma_e^{\,2} = k \cdot 2^{-2L}$$ （其中k’, k为常数）

- 所以量化误差在输出处的方差仍与 2^{-2L} 成正比，仍直接与字长L相联系。量化噪声的平均值 m_e 随量化的方法不同而不同。

6.2 A/D变换的字长效应：量化噪声通过线性系统

舍入：$(m_e)_K = 0$ 经线性系统 $(m_f)_K = 0$

截尾：$(m_e)_T = q/2$ 经线性系统

$$(m_f)_T = (m_e)_T \sum_{m=-\infty}^{\infty} h(n) = (m_e)_T H(e^{j0})$$

- 所以，截尾处理后线性系统的输出中有直流分量，将对信号的频谱结构产生影响，这是应避免的。

例：设有一个8*bit*（L=8）的A/D转换器，它的输出$\hat{x}(n)$经过一个*IIR*滤波器，

其中$H(z)=\dfrac{z}{z-0.999}$，

求滤波器输出端的量化噪声功率。

解： 输入信号功率：$\sigma_e^2=\dfrac{1}{12}\times q^2=\dfrac{2^{-16}}{3}$(数据位7位)

$\sigma_f^2=\sigma_e^2 I$，其中$I=\dfrac{1}{2\pi j}\oint_c z^{-1}H(z)H(z^{-1})dz$，而$z^{-1}H(z)H(z^{-1})=\dfrac{z^{-1}}{(z-a)(z^{-1}-a)}$，$a=0.999$

$R_{z=a}=\mathrm{Re}s[z^{-1}H(z)H(z^{-1}),z=a]=\dfrac{z^{-1}}{(z^{-1}-a)}|_{z=a}=\dfrac{1}{1-a^2}$($z=a$单位圆内)

$R_{z=a}=\mathrm{Re}s[z^{-1}H(z)H(z^{-1}),z=a]$

$=\mathrm{Re}s[\dfrac{z^{-1}}{(z-a)(z^{-1}-a)},z=a]=\mathrm{Re}s[\dfrac{1}{(z-a)(1-az)},z=a]$

$=\mathrm{Re}s[\dfrac{z^{-1}}{(z^{-1}-a)}|_{z=a}]$

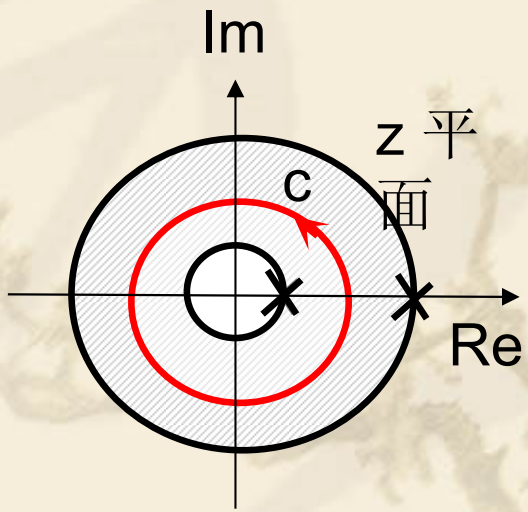
$=\dfrac{z^{-1}}{(z^{-1}-a)}|_{z=a}$

$=\dfrac{1}{1-a^2}$($z=a$单位圆内)

$I=\dfrac{1}{2\pi j}\oint_c z^{-1}H(z)H(z^{-1})dz=\dfrac{1}{1-a^2}$，

当$a=0.999$时，I=500.25

$\sigma_f^2=\sigma_e^2 I=\dfrac{2^{-16}}{3}\times 500.25=0.00254$



6.3 系数量化误差

设计出的系统，其传输函数的系数具有无限精度。
软件、硬件实现时，系数字长有限。

系数量化导致系统的的零点、极点位置偏差、频率响应偏差

系数量化的影响大小与下列因素有关：

- 量化字长
- 滤波器结构
- 极点（零点）个数和密集程度

讨论目的： 了解不同结构及特性的滤波器对系数量化的灵敏程度，

从而为滤波器系数字长的选择、滤波器结构的选择提供依据。

6.3.1 系数量化误差

量化前：

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

量化后：

$$\hat{H}(z)=\frac{\sum_{i=0}^{N}\hat{a}_i z^{-i}}{1-\sum_{i=1}^{N}\hat{b}_i z^{-i}}$$

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

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

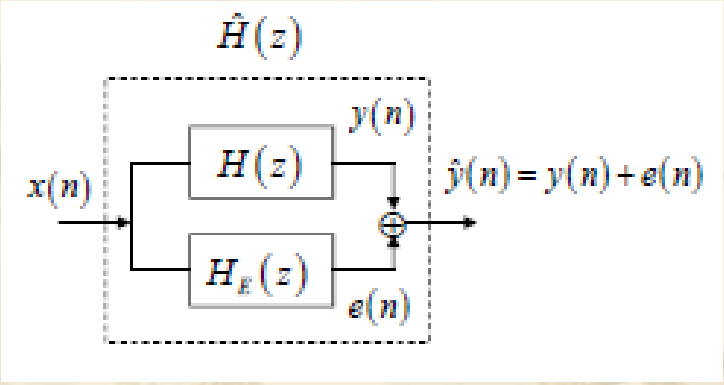
有精度系统

$$\hat{H}(z)=\frac{\sum_{i=0}^{N}\hat{a}_i z^{-i}}{1-\sum_{i=1}^{N}\hat{b}_i z^{-i}}=\frac{\sum_{i=0}^{N}a_i z^{-i}+\sum_{i=0}^{N}\Delta a_i z^{-i}}{1-\sum_{i=1}^{N}b_i z^{-i}-\sum_{i=1}^{N}\Delta b_i z^{-i}}$$

$$=\frac{A(z)+\Delta A(z)}{B(z)+\Delta B(z)}=\frac{A(z)}{B(z)}+\left[\frac{A(z)+\Delta A(z)}{B(z)+\Delta B(z)}-\frac{A(z)}{B(z)}\right]$$

$$=H(z)+H_e(z)$$

无精度系统　　误差系统



6.3.1 系数量化误差

IIR 滤波器（比**FIR**更一般）

$$H(z)=\frac{\sum_{i=0}^{N}a_i z^{-i}}{1-\sum_{i=1}^{N}b_i z^{-i}}=G\frac{\prod_{k=1}^{M}(1-z_k z^{-1})}{\prod_{k=1}^{N}(1-z_i z^{-1})}$$

零点、极点位置决定系统的频响 $H(e^{iw})$

极点位置影响系统的稳定性

系数量化

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

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

实际系统函数

$$\hat{H}(z)=\frac{\sum_{i=0}^{N}\hat{a}_i z^{-i}}{1-\sum_{i=1}^{N}\hat{b}_i z^{-i}}=\hat{G}\frac{\prod_{k=1}^{M}(1-\hat{z}_k z^{-1})}{\prod_{k=1}^{N}(1-\hat{z}_i z^{-1})},\hat{G}=\hat{G}+\Delta G$$

一个a_k的量化对全部零点z_k 的位置都有影响　　一个零点位置的偏差 Δz_k取决于所有 a_k的量化误差

一个b_k 的量化对全部极点 z_i 的位置都有影响　　一个极点位置的偏差 Δz取决于所有 b_k的量化误差

6.3.2 极点位置灵敏度

例：二阶IIR 滤波器

直接型：

$$H(z)=\frac{1}{1-b_1z^{-1}-b_2z^{-2}}=\frac{1}{1-0.9z^{-1}+0.2z^{-2}},\quad z_1=0.5,z_2=0.4$$

系数量化：符号位1位，数据位为3位

$$\hat{H}_1(z)=\frac{1}{1-0.875z^{-1}+0.125z^{-2}}$$

$$\hat{z}_1=0.695,\quad \hat{z}_2=0.18$$ 极点位置偏差大

级联型：$$=\frac{1}{1-a_1z^{-1}}\frac{1}{1-a_2z^{-1}}=\frac{1}{1-0.5z^{-1}}\frac{1}{1-0.4z^{-1}},\quad z_1=0.5,z_2=0.4$$

$$\hat{z}_1=0.5,\quad \hat{z}_2=0.375$$ 极点位置偏差小

系数量化后的传输函数 $$H(z)=\frac{1}{1-0.5z^{-1}}\frac{1}{1-0.375z^{-1}}$$

用低阶级联型结构实现，对系数量化较不敏感

6.3.2 极点位置灵敏度

- 系数量化使系统函数的零极点偏离准确的位置，这就会产生：极点在Z平面的单位圆内，系统是稳定的，经系数量化，使零极点发生移动，DF的性能偏离技术要求，甚至使极点移到单位圆外，破坏了系统的稳定性。
- IIR DF的传递函数为：

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

6.3.2 极点位置灵敏度

1、系统极点（零点）位置对系数量化的灵敏度

灵敏度高：系数的微小变化会导致极点（零点）位置较大变化

什么结构的系统，其零、极点位置灵敏度低？

如果IIR DF的网络结构为直接型或正准型，当系数量化后，*H(z)*变为 $\hat{H}(z)$：

$$\hat{H}(z)=\frac{\sum_{i=0}^{N}\hat{a}_i z^{-i}}{1-\sum_{i=1}^{N}\hat{b}_i z^{-i}}$$

其中 $\hat{a}_i,\hat{b}_i$ 表示量化后的系数 $\hat{a}_i=a_i+\Delta a_i \quad \hat{b}_i=b_i+\Delta b_i$

6.3.2 极点位置灵敏度

分析极点的情况 $B(z)=1-\sum_{i=1}^{N}b_i z^{-i}=\prod_{k=1}^{N}\left(1-z_k z^{-1}\right)$

- 其中，z_k 是系数为无限精度时的极点，

  由于$\hat{b}_i=b_i+\Delta b_i$的量化误差引起的极点偏离为 $z_k+\Delta z_k$

注：一个极点(零点)位置的偏差 $\Delta z_k(\Delta z_i)$ 取决对于所有的 $a_k(b_k)$ 的量化误差

6.3.2 极点位置灵敏度

- 极点灵敏度：系数 b_i 的变化所引起的 z_k 位置的变化率，

  用偏导数 $\dfrac{\partial z_k}{\partial b_i}$

$$\Delta z_k = \sum_{i=1}^{N} \frac{\partial z_k}{\partial b_i} \Delta b_i \qquad k = 1,2,\ldots\ldots, N$$

- 所以极点灵敏度 $\dfrac{\partial z_k}{\partial b_i}$ 的大小决定了系数偏差对极点位置的影响程度，它越大，Δb_i 对 Δz_k 的影响越大。

6.3.2 极点位置灵敏度

因为 $$\frac{\partial B(z)}{\partial b_i}=\frac{\partial B(z)}{\partial z_k}\cdot\frac{\partial z_k}{\partial b_i}$$

而由 $$B(z)=1-\sum_{i=1}^{N}b_i z^{-i}=\prod_{k=1}^{N}\left(1-z_k z^{-1}\right)$$

可得 $$\frac{\partial B(z)}{\partial b_i}=-z^{-i}$$

及 $$\frac{\partial B(z)}{\partial z_k}=-z^{-1}\prod_{\substack{l=1\\ l\neq k}}^{N}\left(1-z_l z^{-1}\right)$$

6.3.2 极点位置灵敏度

所以 $$\frac{\partial z_k}{\partial b_i}=\frac{\partial B(z)}{\partial b_i}\Big/\frac{\partial B(z)}{\partial z_k}=\frac{-z^{-i}}{-z^{-1}\prod\limits_{\substack{l=1\\l\neq k}}^{N}\left(1-z_l z^{-1}\right)}=\frac{z^{-i}}{z^{-N}\prod\limits_{\substack{l=1\\l\neq k}}^{N}\left(z-z_l\right)}=\cdots\cdots$$

将上式分子分母同乘以 z^N ， 得

$$\frac{\partial z_k}{\partial b_i}=\frac{z^{N-i}}{\prod\limits_{\substack{l=1\\l\neq k}}^{N}\left(z-z_l\right)}$$

表示 z_k 外的一极点指向 z_k 的矢量

$$\left.\frac{\partial z_k}{\partial b_i}\right|_{z=z_k}=\frac{z_k^{N-i}}{\prod\limits_{\substack{l=1\\l\neq k}}^{N}\left(z_k-z_l\right)}$$

6.3.2 极点位置灵敏度

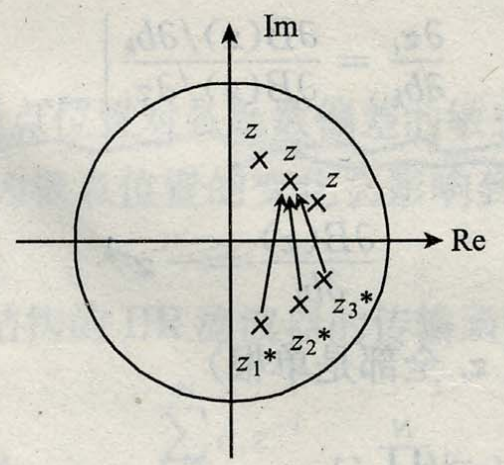
$$\left.\frac{\partial z_k}{\partial b_i}\right|_{z=z_k} = \frac{z_k^{N-i}}{\prod_{\substack{l=1 \\ l\neq k}}^{N}(z_k - z_l)}$$

- 当$H(z)$的极点靠拢越近，$|z_i - z_k|$越小，灵敏度越高；
- N越大（极点数越多），极点越密集，$|z_i - z_k|$越小，越敏感。
- $|z_i| \to 1$，则灵敏度越高

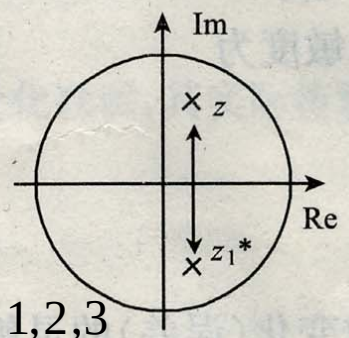
零点位置灵敏度：类似

高阶直接型结构　极（零）点多而密集，极（零）点位置灵敏度高，难以调整。应避免采用

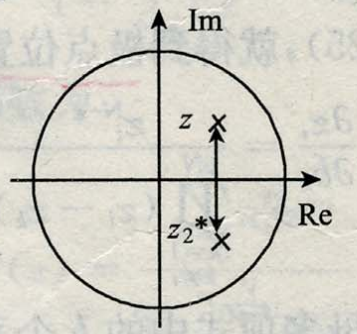
低阶基本节　极（零）点少而稀疏，极（零）点位置灵敏度低。
建议采用低阶基本节的级联/ 并联



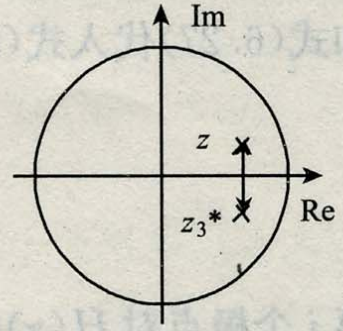
(a) 直接型$H(z)$的极点



$H_1(z), H'_1(z)$



$H_2(z), H'_2(z)$



$H_3(z), H'_3(z)$

(b) 级联型、并联型的极点

$$H(z)=\prod_{i=1}^{3}H_i(z),$$

其中 $H_i(z)=\dfrac{A_i(z)}{(1-z_i z^{-1})(1-z^*_i z^{-1})}, i=1,2,3$

式中分母对每一极点 z_i 都只有一个因子 $\left|\dfrac{\partial z_i}{\partial b_{1i}}=\dfrac{z_i}{(z_i-z_i^*)}\right.$

$$\frac{\partial z_i}{\partial b_{2i}}=\frac{z_i}{(z_i-z_i^*)},$$

级联型每个网络的极点 密集度比直接型网络稀 疏

6.3.3 系数量化对FIR滤波器的影响

一、零点位置灵敏度

---其分析类似于对IIR 滤波器的极点位置灵敏度的分析。

---应该用低阶节的级联，使得零点位置对系数量化的灵敏度低。

二、线性相位滤波器，长度为*M*

1）用直接型结构实现，系数量化后单位抽样响应仍然对称，系统仍是线性相位的。

若系数被舍入到***L*+1**位，则滤波器单位取样响应的误差

$$e_h(n)=\hat{h}(n)-h(n),\quad -\frac{1}{2}2^{-(L+1)}<e_h(n)<\frac{1}{2}2^{-(L+1)}$$

滤波器频率响应的误差 $H_e(e^{jw})=\hat{H}(e^{jw})-H(e^{jw})=\sum_{n=0}^{M-1}e_h(n)e^{-jwn}$

$e_h(n)$ 是零均值的， 所以 $H_e(e^{jw})$ 是零均值的

假设系数误差序列中的***M*** 个取样互不相关，则

$$\sigma_e^2=\frac{2^{-2(L+1)}}{12}M=\frac{4^{-L}}{48}M$$

给定***M***和允许的频率响应误差σ_e^2， 可确定所需要的数据位量化字长***L***。

2）用级联型结构实现，系数量化后，零点仍然互为镜像，系统仍是线性相位的。

其基本节的误差及其统计分析同直接型结构

6.3.3 IIR DF中系数量化对FIR滤波器的影响

上式说明，在FIR DF 的系数量化时，会使传递函数产生误差，此误差不会超过

$$|e(n)| \le \frac{q}{2}, q = 2^{-L}$$

所以

$$\left|E(e^{j\omega})\right| \le \sum_{n=0}^{M} \left|e(n)\right| \le \frac{(M+1)q}{2}$$

例

- 一个FIR DF其阶数（最大延迟数）M = 20，如果要求由于系数量化产生的误差小于1/100，问字长需要几位？

解：如果 $\frac{(M+1)q}{2} \le \frac{1}{100}$

则满足：$\left|E(e^{j\omega})\right| \le \frac{1}{100}$

已知 $q = 2^{-L}$

所以 $\frac{M+1}{2} \cdot 2^{-L} \le \frac{1}{100}$

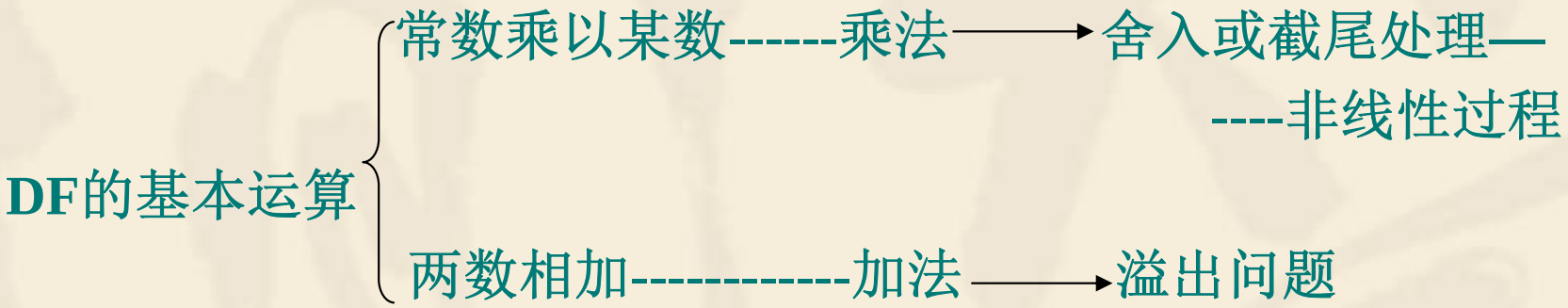
即 $$2^{-L} \le \frac{1}{1050}$$

所以 $$L \ge 10.04$$

即需要**11**比特的字长才能满足需要。

- 强调：不论是**IIR DF**或**FIR DF**，系数量化后的频率响应都必须用计算机校核，以确保其性能符合要求。

6.4 运算过程中的有限字长效应



舍入或截尾处理都是非线性过程，非线性问题的分析很复杂，而且有许多问题并不清楚，所以下面只讨论一些简单的情况。

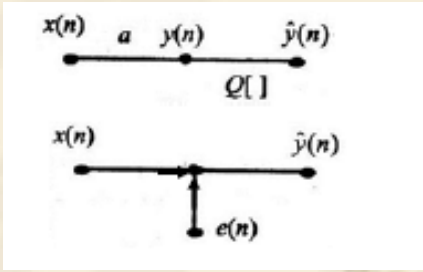
6.4 运算过程中的有限字长效应

- 典型的相乘：$y(n) = a\,x(n)$

(B+C)位　B位　C位长

对(B+C)位的$y(n)$进行舍入或截尾，产生量化误差（或称量化噪声）。

- 统计模型：$$\hat{y}(n) = Q[y(n)] = y(n) + e(n) = ax(n) + e(n)$$



进行分析之前，假定：

(1)$e(n)$ 是白噪声序列

(2)在一个量化区间内$e(n)$是等概率密度分布

(3)$e(n)$ 与输入信号、输出信号及中间计算结果不相关。

步骤：

- 写出数学模型$H(z)$、$h(n)$
- 画流图，每一乘法后加上等效的噪声序列
- 假设各噪声：白色；均匀等概；与信号（输入、中间、输出）不相关

  则 $m_e = 0, \sigma_e^2 = \frac{2^{-2L}}{12} = \frac{q^2}{12}, q = 2^{-L}$ （舍入量化）

- 输出端总的噪声（由运算量化引起）

  $f(n) = \ldots + \ldots + \ldots$ （项数=乘法次数）

- 求$f(n)$的方差 σ_f^2
- 计算输出端的信噪比 $\frac{\sigma_y^2}{\sigma_f^2}$

一阶*IIR*滤波器：$y(n)=ay(n-1)+x(n),n\geq 0,|a|<1$

$\hat{y}(n)=y(n)+f(n)$, $f(n)$是噪声源$e(n)$造成的输出误差

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

$$I=\frac{1}{2\pi j}\oint_c z^{-1}H(z)H(z^{-1})dz,\qquad \text{其中}z^{-1}H(z)H(z^{-1})=\frac{z^{-1}}{(z-a)(z^{-1}-a)}$$

计算单位圆内极点的留数之和

$$R_{z=a}=\mathrm{Re}s[z^{-1}H(z)H(z^{-1}),z=a]=\frac{z^{-1}}{(z^{-1}-a)}|_{z=a}=\frac{1}{1-a^2}(z=a\text{单位圆内})$$

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

$$\begin{aligned}R_{z=a}&=\mathrm{Re}s[z^{-1}H(z)H(z^{-1}),z=a]\\&=\mathrm{Re}s[\frac{z^{-1}}{(z-a)(z^{-1}-a)},z=a]=\mathrm{Re}s[\frac{1}{(z-a)(1-az)},z=a]\\&=\mathrm{Re}s[\frac{z^{-1}}{(z^{-1}-a)}|_{z=a}]\\&=\frac{z^{-1}}{(z^{-1}-a)}|_{z=a}\\&=\frac{1}{1-a^2}(z=a\text{单位圆内})\end{aligned}$$

$$\sigma_f^2=\sigma_e^2\frac{1}{1-a^2}=\frac{1}{12}\times 2^{-2L}\cdot\frac{1}{1-a^2}=\frac{2^{-2(L+1)}}{3(1-a^2)}$$

有一*IIR*滤波器：$H(z)=H_1(z)H_2(z)$，其中

$$H_1(z)=\frac{1}{(1-a_1z^{-1})},\quad |a_1|<1 \qquad H_2(z)=\frac{1}{(1-a_2z^{-1})},\quad |a_2|<1$$

如图6.12所示，研究两者乘积的舍入误差。

$$\sigma_f^2=\sigma_e^2(I_1+I_2)$$

$$I_2=\frac{1}{2\pi j}\oint_c[\frac{1}{1-a_2z^{-1}}][\frac{1}{1-a_2z}]z^{-1}dz=\frac{1}{1-a_2^{\ 2}}$$

$$I_1=\frac{1}{2\pi j}\oint_c[\frac{1}{1-a_1z^{-1}}][\frac{1}{1-a_1z}][\frac{1}{1-a_2z^{-1}}][\frac{1}{1-a_2z}]z^{-1}dz$$

$$=\frac{a_1}{(1-a_1^{\ 2})(a_1-a_2)(1-a_1a_2)}+\frac{a_2}{(1-a_2^{\ 2})(a_2-a_1)(1-a_1a_2)}$$

$$\sigma_f^2=\sigma_e^2(I_1+I_2)=\frac{1}{12}\times 2^{-2L}\cdot(I_1+I_2)$$

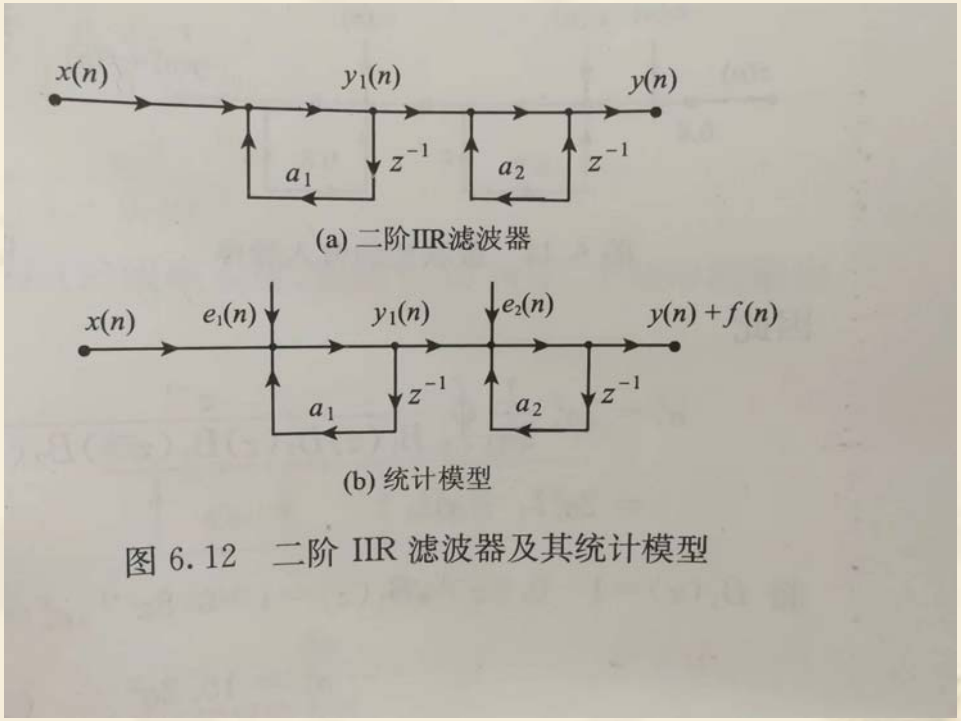


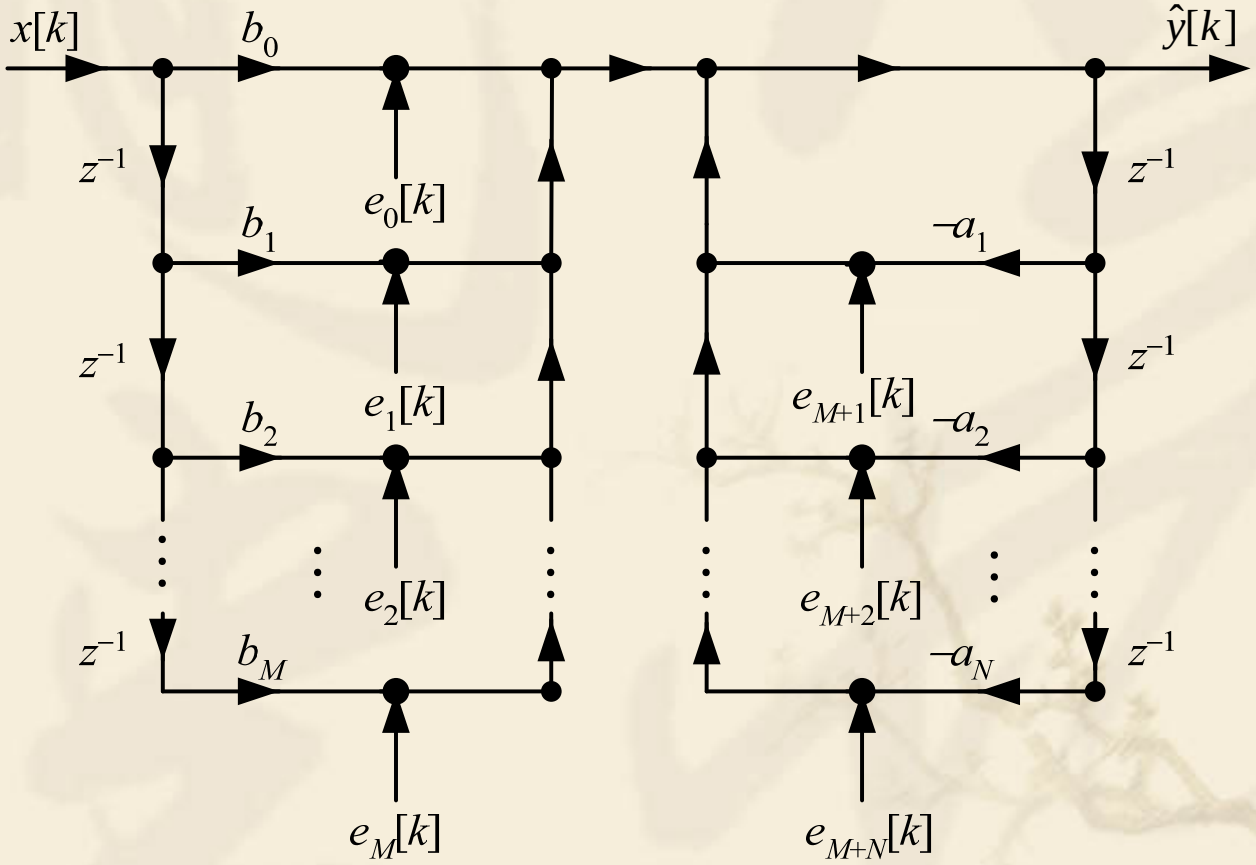
图 6.12　二阶 IIR 滤波器及其统计模型

6.4.1 乘积的舍入误差: IIR滤波器

直接I型结构乘积量化误差分析

单个噪声源方差

$$\sigma_0^2 = E\left\{e_i^2[k]\right\} = \frac{q^2}{12}$$



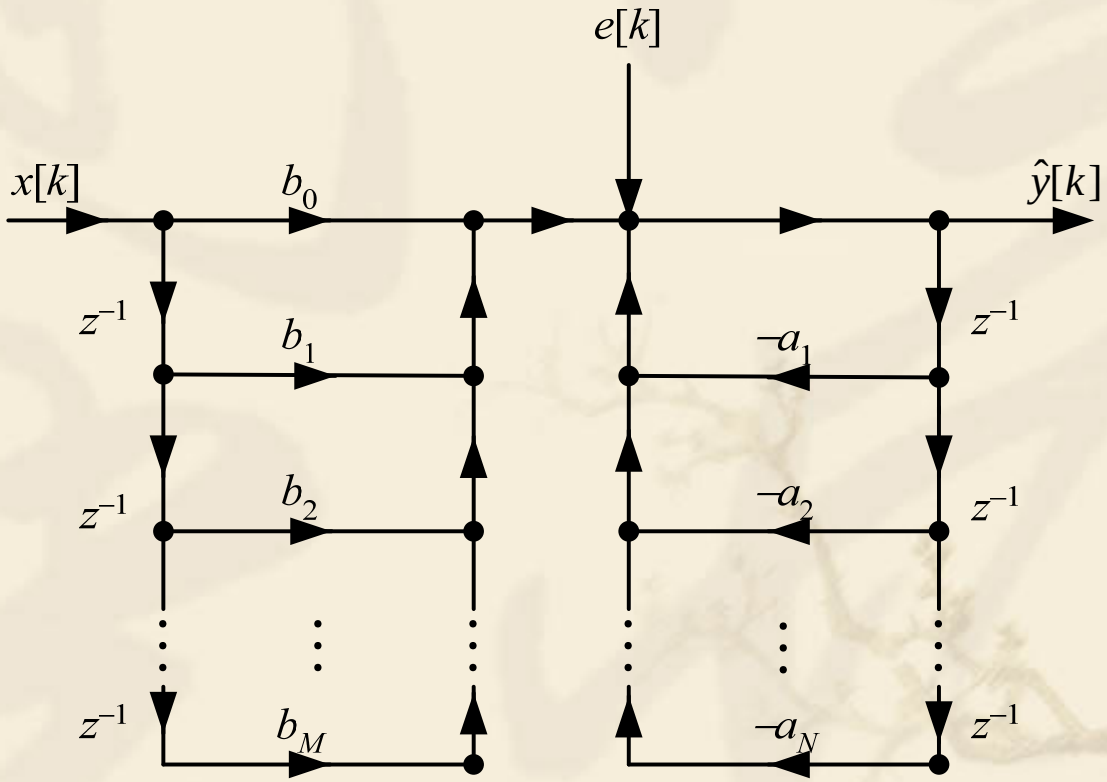
直接I型结构乘积量化误差单个噪声源模型

6.4.1 乘积的舍入误差: IIR滤波器

直接I型结构乘积量化误差分析

联合噪声方差

$$\sigma_e^2 = E\left\{e^2[k]\right\} = (M+N+1)\frac{q^2}{12}$$



直接I型结构乘积量化误差联合噪声源模型

6.4.1 乘积的舍入误差: IIR滤波器

直接I型结构乘积量化误差分析

$e[k]$通过系统的平均噪声功率

$$\sigma_v^2 = (M+N+1)\frac{q^2}{12}\frac{1}{2\pi j}\oint_C H_e(z)H_e(z^{-1})z^{-1}dz$$

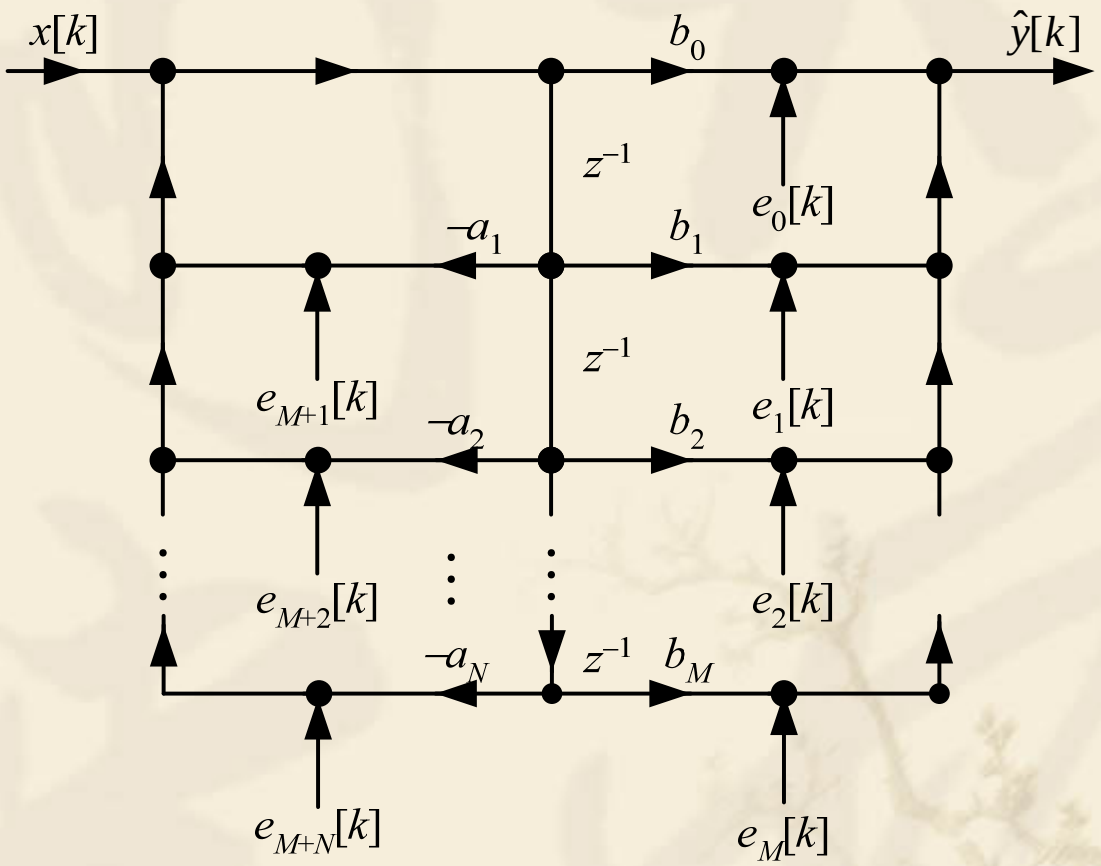
$H_e(z)$：$e[k]$所通过系统的系统函数

$H_e(z)=1/A(z)$

就是分母项

6.4.1 乘积的舍入误差: IIR滤波器

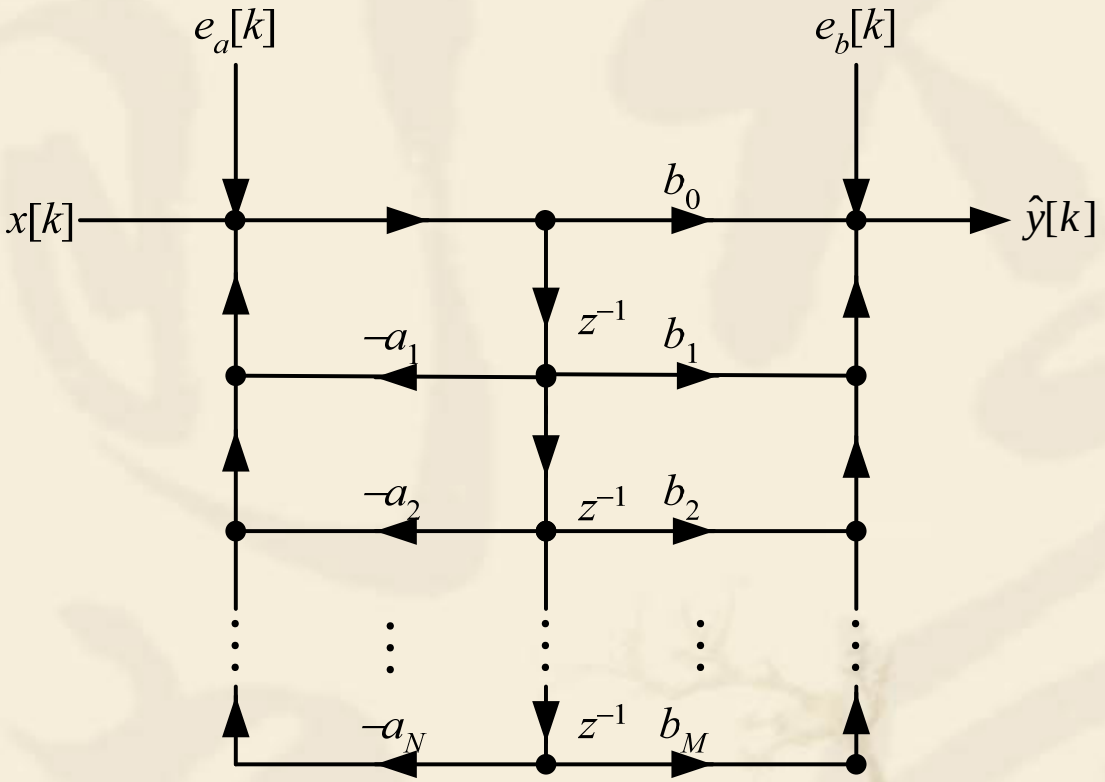
正准I型结构乘积量化误差分析



直接II型结构乘积量化误差单个噪声源模型

6.4.1 乘积的舍入误差: IIR滤波器

正准I型结构乘积量化误差分析



直接II型结构乘积量化误差联合噪声源模型

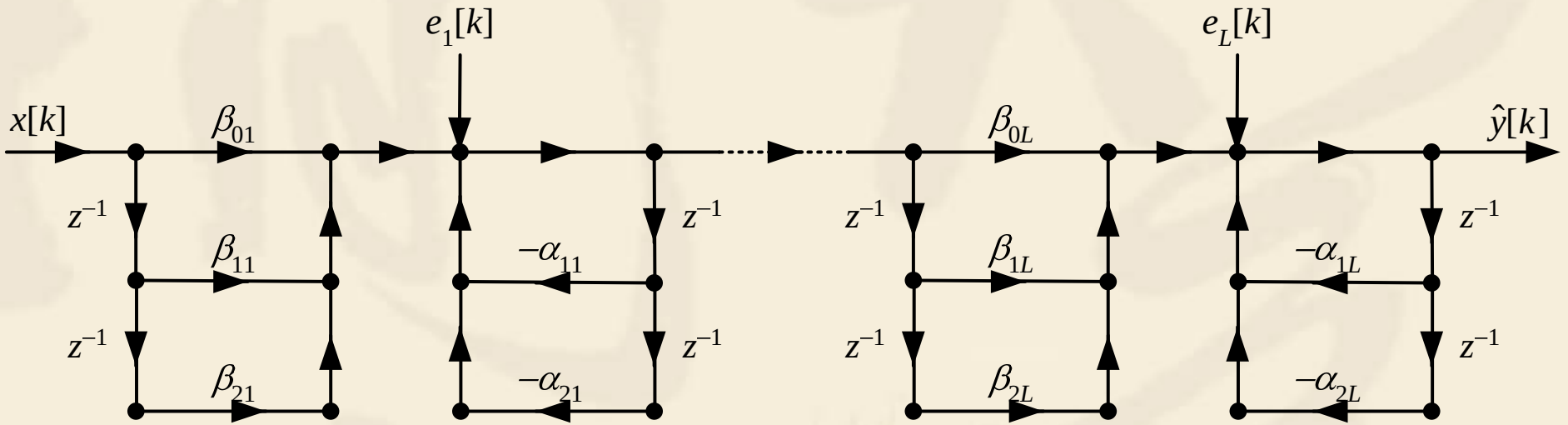
极点网络N经过传输函数
零点网络直接输出

$e_a[k]$和$e_b[k]$通过系统的输出噪声方差

$$\sigma_v^2 = N\frac{2^{-2L}}{12}\sum_{k=0}^{\infty}\left|h[k]\right|^2 + (M+1)\frac{2^{-2L}}{12}$$

6.4.1 乘积的舍入误差: IIR滤波器

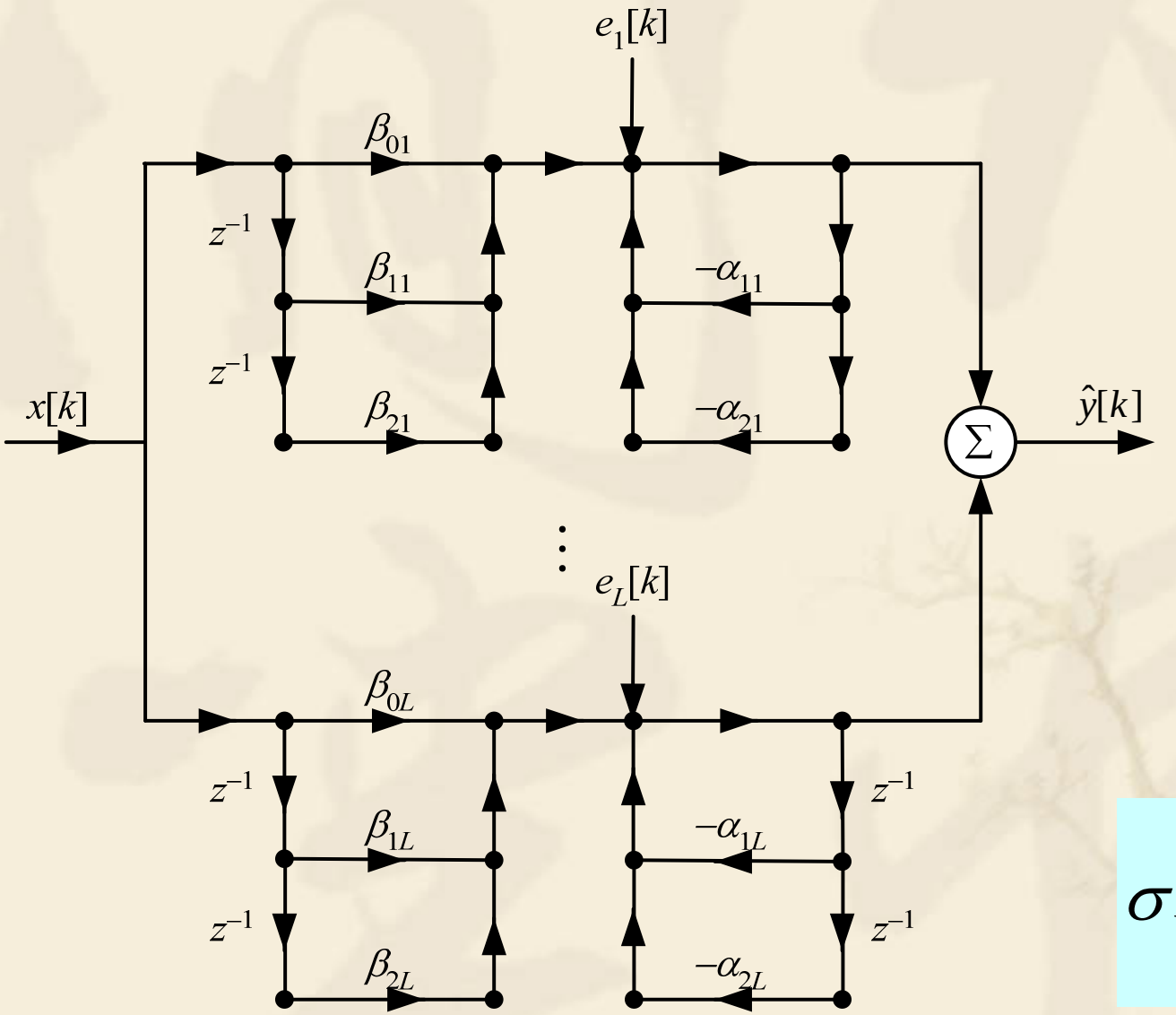
IIR DF级联结构乘积量化误差分析



$$\sigma_v^2 = \sum_{i=1}^{L} \sigma_i^2 \frac{1}{2\pi j} \oint_C \frac{\prod_{l=i+1}^{L} H_l(z) H_l(z^{-1})}{A_i(z) A_i(z^{-1})} z^{-1} dz$$

6.4.1 乘积的舍入误差: IIR滤波器

IIR DF并联结构乘积量化误差分析



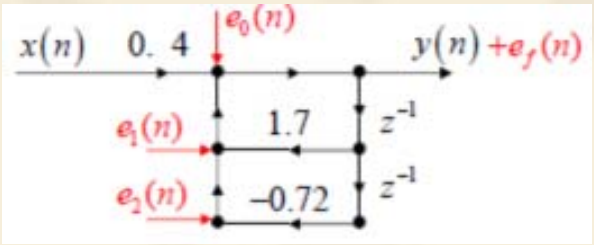
$$\sigma_v^2 = \sum_{i=1}^{L} \sigma_{e_i}^2 \left| h_i[k] \right|^2$$

6.4.1 乘积的舍入误差: IIR滤波器

例：二阶低通 $H(z)=\frac{0.4}{1-0.9z^{-1}}\cdot\frac{1}{1-0.8z^{-1}},|z|<0.8$ 因果，稳定，低通

计算直接型、级联型、并联型结构时定点运算的舍入误差

直接型结构



（1）数学模型 $H(z)=\frac{0.4}{1-1.7z^{-1}+0.72z^{-2}}$

（2）画流图，每一乘法加等效噪声

（3）假设各噪声的统计特性

（4）输出端总噪声

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

$$H_0(z)=\frac{1}{1-1.7z^{-1}+0.72z^{-2}}$$

（5）总噪声的方差

$$\sigma_f^2=E[(e_f(n)-m_f)^2]=E[(e_f(n)]^2$$

$$=\{E[e_0^2(n)]+E[e_1^2(n)]+E[e_2^2(n)]\}\cdot\sum_{n=0}^{\infty}h_0^2(n)$$

$$\sigma_f^2=3\sigma_e^2\frac{1}{2\pi j}\oint_c z^{-1}H_0(z)H_0(z^{-1})dz$$

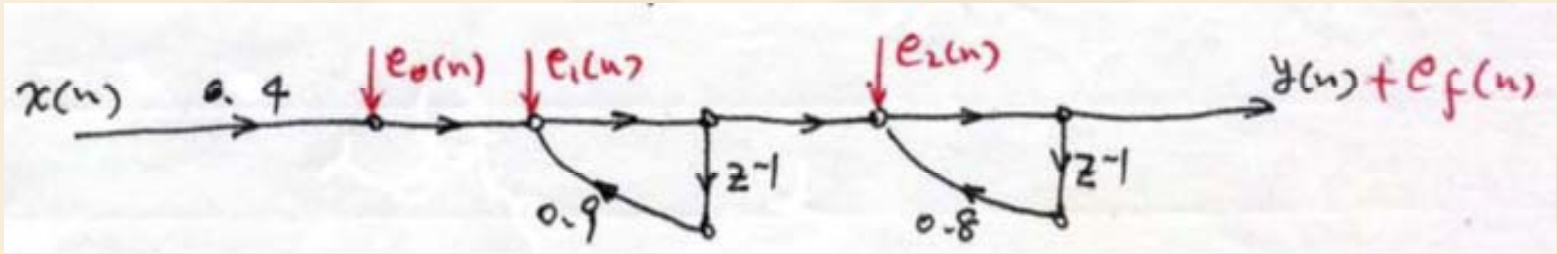
$$=3\frac{q^2}{12}\frac{1}{2\pi j}\oint_c\frac{z^{-1}}{(1-1.7z^{-1}+0.72z^{-2})(1-1.7z^{-1}+0.72z^{-2})}dz=22.4q$$

若0.4在后？ $e_f(n)=e_0(n)+[e_1(n)+e_2(n)]*h_0(n)$

6.4.1 乘积的舍入误差: IIR滤波器

级联型结构

$$H(z)=0.4\cdot\frac{1}{1-0.9z^{-1}}\cdot\frac{1}{1-0.8z^{-1}}=\frac{0.4}{B_1(z)}\cdot\frac{1}{B_2(z)}$$



$$e_0(n),e_1(n)\rightarrow H_1(z)=\frac{1}{B_1(z)B_2(z)}\qquad e_2(n)\rightarrow H_2(z)=\frac{1}{B_2(z)}$$

$$e_f(n)=[e_0(n)+e_1(n)]*h_1(n)+e_2(n)*h_2(n)$$

其中：$h_1(n)\leftrightarrow H_1(z)=\frac{1}{1-0.9z^{-1}}\cdot\frac{1}{1-0.8z^{-1}}$

$$\begin{aligned}\sigma_f^2&=2\sigma_e^2\frac{1}{2\pi j}\oint_c\frac{z^{-1}}{B_1(z)B_2(z)B_1(z^{-1})B_2(z^{-1})}dz+\sigma_e^2\frac{1}{2\pi j}\oint_c\frac{z^{-1}}{B_2(z)B_2(z^{-1})}dz\\&=2\sigma_e^2I_1+\sigma_e^2I_2\\&=15.2q^2\end{aligned}$$

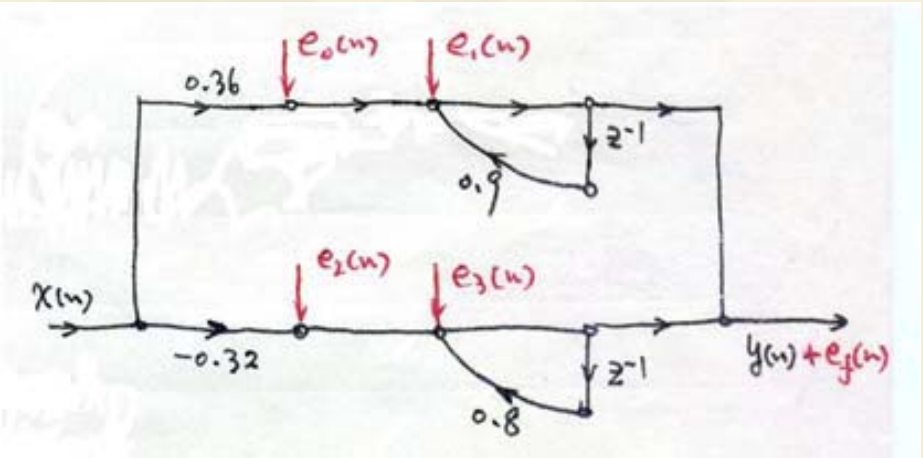
若改变三个子系统的级联次序，$e_f(n)$不同

级联次序对运算量化误差有影响

6.4.1 乘积的舍入误差: IIR滤波器

并联型结构

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



$$e_f(n)=[e_0(n)+e_1(n)]*h_0(n)+(e_2(n)+e_3(n))*h_1(n)$$

其中：$h_0(n)\leftrightarrow H_0(z)=\frac{1}{1-0.9z^{-1}}=\frac{1}{B_1(z)}$

$$h_1(n)\leftrightarrow H_1(z)=\frac{1}{1-0.8z^{-1}}=\frac{1}{B_2(z)}$$

$$\sigma_f^2=2\sigma_e^2\frac{1}{2\pi j}\oint_c\frac{z^{-1}}{B_1(z)B_1(z^{-1})}dz+2\sigma_e^2\frac{1}{2\pi j}\oint_c\frac{z^{-1}}{B_2(z)B_2(z^{-1})}dz$$
$$=2\sigma_e^2I_1+2\sigma_e^2I_2$$
$$=1.34q^2$$

6.4.1 乘积的舍入误差: IIR滤波器

$$H(z)=H_1(z)H_2(z)$$

$$H_1(z)=\frac{0.4z^{-1}}{1-0.8z^{-1}} \qquad H_2(z)=\frac{z^{-1}}{1-0.9z^{-1}}$$

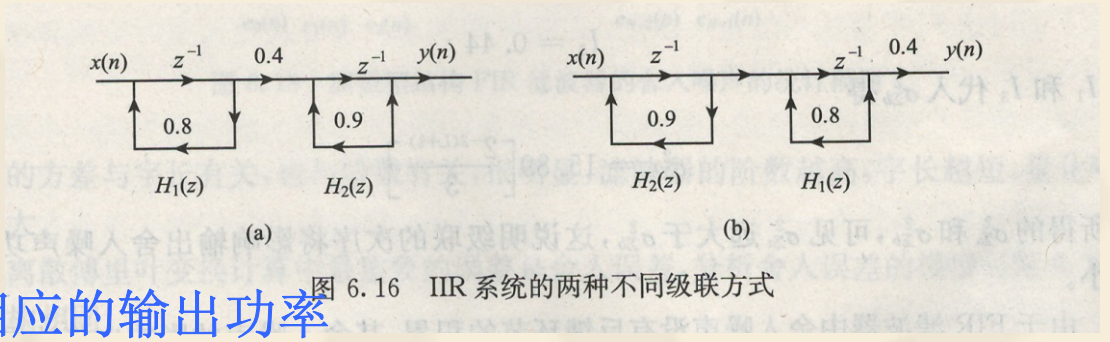


图 6.16 IIR 系统的两种不同级联方式

用 $H(z)=H_1(z)H_2(z)$

和 $H(z)=H_2(z)H_1(z)$ 级联实现，求相应的输出功率

$$\sigma_a^2=\sigma_e^2 I_1+2\sigma_e^2 I_2$$

$$I_1=\frac{1}{2\pi j}\oint_c z^{-1}H_1(z)H_2(z)H_1(z^{-1})H_2(z^{-1})dz=14.45$$

$$I_2=\frac{1}{2\pi j}\oint_c z^{-1}H_2(z)H_2(z^{-1})dz=5.28$$

$$\sigma_a^2=\sigma_e^2 I_1+2\sigma_e^2 I_2=25.01[\frac{2^{-2(L+1)}}{3}]$$

$$\sigma_b^2=\sigma_e^2 I_1+\sigma_e^2 I_3+\sigma_e^2$$

$$I_3=\frac{1}{2\pi j}\oint_c z^{-1}H_1(z)H_1(z^{-1})dz=0.44$$

$$\sigma_b^2=15.89[\frac{2^{-2(L+1)}}{3}]$$

总结：

- 用不同的结构实现，由运算量化引起的误差不同
- 运算次序不同，运算量化引起的误差也不同
- 从减少运算量化引起的误差角度看

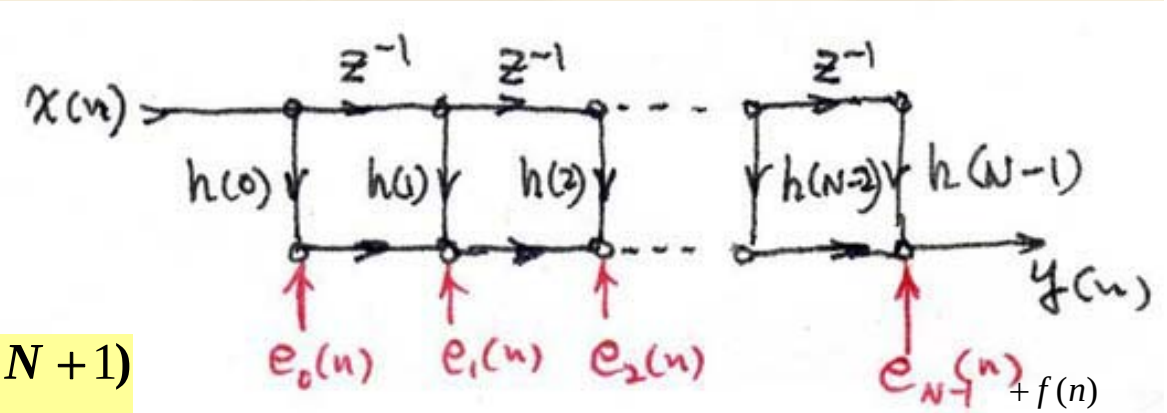
直接型结构	误差积累起来，输出误差最大	差
级联型结构	误差只通过其后面的反馈环节	中
并联型结构	误差仅通过本子系统的反馈环节	优

直接型误差>级联型误差>并联型误差

应避免在高阶时采用直接型。

6.4.1 乘积的舍入误差:

FIR滤波器中的乘积舍入误差



直接型结构

$$y(n)=b_0x(n)+b_1x(n-1)+\cdots+b_{N-1}x(n-N+1)$$
$$=\sum_{k=0}^{N-1}b_kx(n-k)=\sum_{i=0}^{N-1}h(k)x(n-k)$$

所有误差在输出端相加

$$f(n)=e_0(n)+e_1(n)+\cdots+e_{N-1}(n)$$

$$\sigma_f^2=N\sigma_e^2=N\frac{q^2}{12}$$

知道 σ_f^2和N， 可以求出所需字长L

为达到相同的运算精度，阶次N 越高，需字长越长

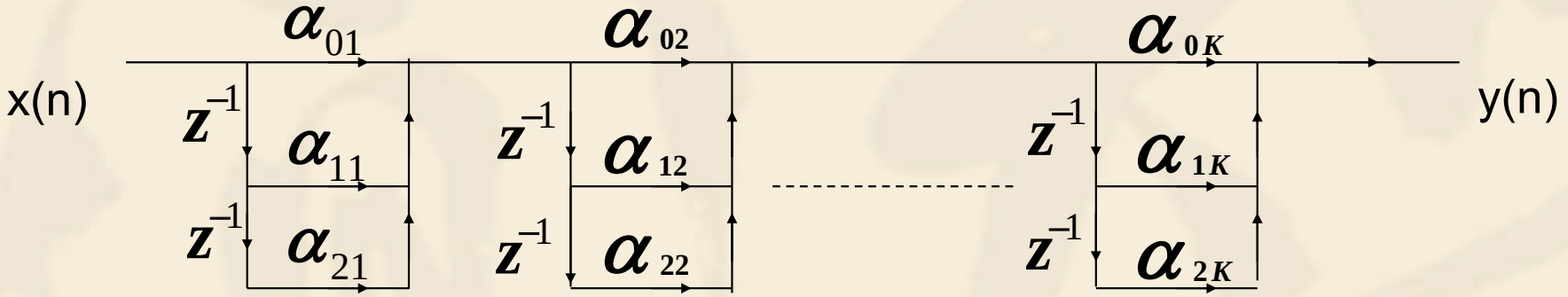
线性相位FIR 滤波器中的乘积舍入误差

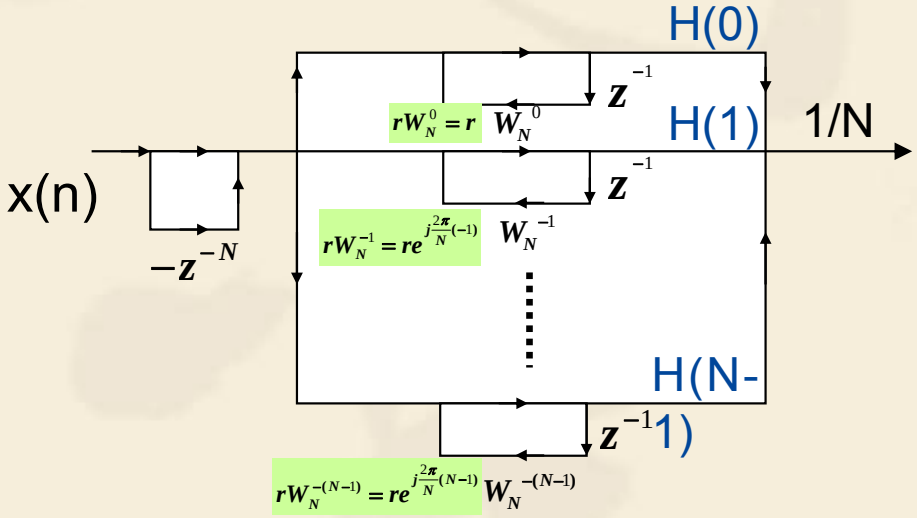
DFT计算中的有限字长效应

FFT计算中的有限字长效应

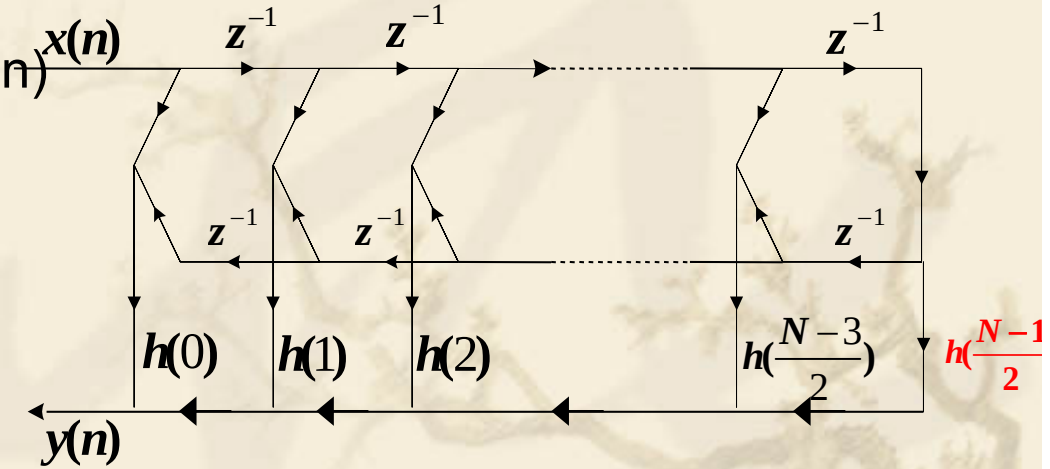
6.4.1 乘积的舍入误差:

级联型结构





频率取样型结构



线性相位结构

一个长度为*N*的复序列的*N*的DFT

$$X(k)=\sum_{n=0}^{N-1}x(n)W_N^{nk},k=0,1,\cdots,N-1$$

一次复乘-----4次实数乘法----4次量化误差，假设共K次复乘

假设：

（1）所有的4K个误差彼此不相关且输入序列无关

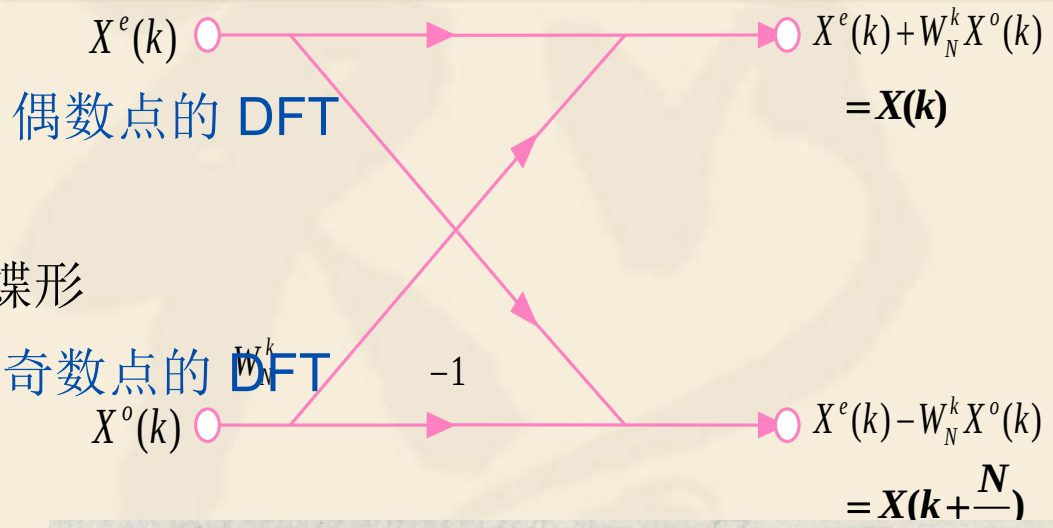
（2）量化误差是方差为 $\sigma_0^2=2^{-2b}/12$ 均匀分布的随机变量

一个单离散样本的N点DFT：

N次复乘-----4N次实数乘法----4N次量化误差

$$\sigma_e^2=4N\sigma_0^2=\frac{2^{-2b}}{3}N$$

1个蝶形运算需要1次复乘和2次复加



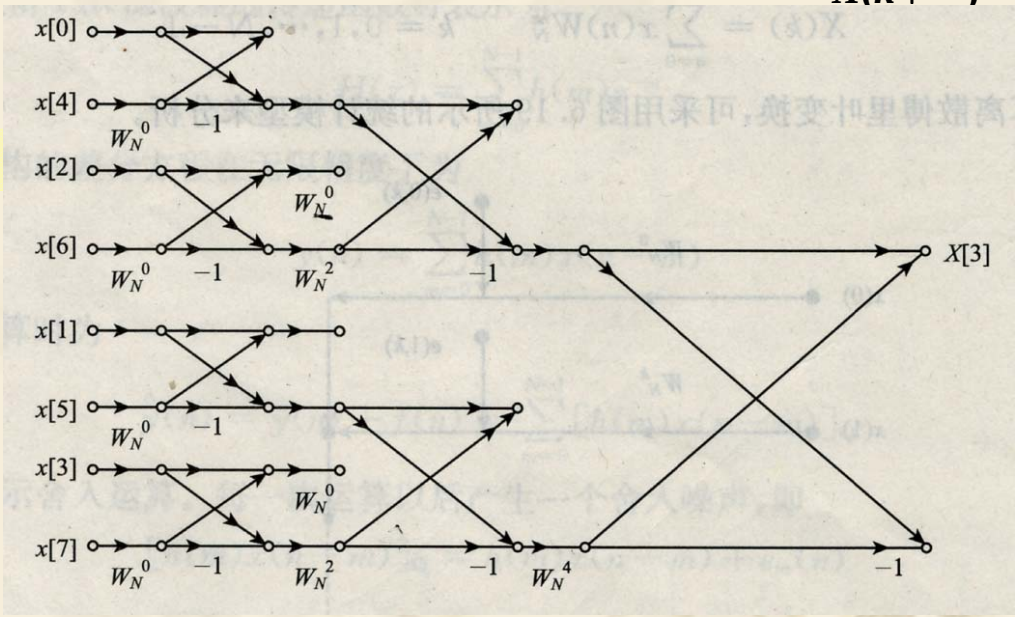
一个长度为N的复序列的N的FFT

共 $\log_2 N$ 级， 每一级 $N/2^r = 2^{m-r}, r = 1,2,\cdots,m$ 蝶形

共:$1+2+2^2+\cdots 2^{m-1} = 2^m - 1 = N-1$

舍入误差来源： $4(N-1)$（一次复乘 $\rightarrow$ 4次实乘运算）

$$\sigma_e^2 = 4(N-1)\sigma_0^2 \approx \frac{2^{-2b}}{3}N$$



说明：快速傅里叶变换并不改变计算一个单离散傅里叶样本的复数乘法的总数；是一种有组织的计算，利用旋转因子的周期性，相应减少了所有N个离散傅里叶变换样本的总的相乘数目。

6.4.2 极限环振荡

设计出的滤波器是个线性系统。
运算量化的滤波器是个非线性系统。可能不稳定，产生两种极限环振荡。

1. 颗粒型极限环振荡
—— **IIR**滤波器定点运算中有限寄存器长度引起的零输入极限环振荡
输入消失后，还有输出且不衰减

表**6.2** 一阶**IIR**滤波器的有限精度运算过程 $(a=1/2=0_{\Delta}100)$

n	$x(n)$	$\hat{y}(n-1)$	$a\,\hat{y}(n-1)$	$Q[a\,\hat{y}(n-1)]$	$\hat{y}(n)=Q[a\,\hat{y}(n-1)]+x(n)$	若 $a=-1/2$
0	$0_{\Delta}111$	$0_{\Delta}000$	$0_{\Delta}000000$	$0_{\Delta}000$	$0_{\Delta}111=7/8$	$7/8$
1	$0_{\Delta}000$	$0_{\Delta}111$	$0_{\Delta}011100$	$0_{\Delta}100$	$0_{\Delta}100=1/2$	$-1/2$
2	$0_{\Delta}000$	$0_{\Delta}100$	$0_{\Delta}010000$	$0_{\Delta}010$	$0_{\Delta}010=1/4$	$+1/4$
3	$0_{\Delta}000$	$0_{\Delta}010$	$0_{\Delta}001000$	$0_{\Delta}001$	$0_{\Delta}001=1/8$	$-1/8$
4	$0_{\Delta}000$	$0_{\Delta}001$	$0_{\Delta}000100$	$0_{\Delta}001$	$0_{\Delta}001=1/8$	$+1/8$
..	...	...	...	...	...	...

不再衰减，等幅振荡

6.4.2 极限环振荡

零输入极限环振荡

任意一阶系统 $H(z)=\dfrac{1}{1-az^{-1}}$，$|z|>|a|$，$|a|<1$，$a$起衰减作用，系统稳定

有限字长运算 $\hat{y}(n)=Q[a\hat{y}(n-1)]+x(n)$

$n\geq n_0$后，$|Q[a\hat{y}(n-1)]|=|\hat{y}(n-1)|$

即：舍入处理使系数**a** 的衰减作用失效

相当于将系数**a** 换成**a’**，而**| a’| =1**

这时的等效系统函数 $H'(z)=\dfrac{1}{1\pm z^{-1}}$，

舍入处理使系数**a** 的衰减作用失效

其极点在单位圆上，输出是等幅振荡

$|Q[a\hat{y}(n-1)]|=|\hat{y}(n-1)|$

$|\hat{y}(n-1)|-|a\hat{y}(n-1)|\leq\dfrac{1}{2}2^{-L}$

$|\hat{y}(n-1)|\leq\dfrac{\frac{1}{2}2^{-L}}{1-|a|}$

振荡幅度 $\dfrac{q/2}{1-|a|}$ 与字长**L** 和系数**| a |** 有关

如何使极限环振荡减弱？

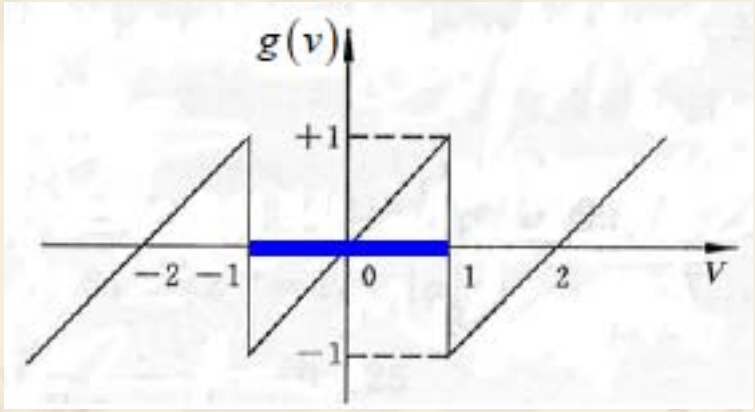
1. 提高L
2. 减小**| a |**

6.4.2 极限环振荡

2. 溢出振荡　　定点数的表示范围：绝对值**< 1**

定点加法引起，和的数值超出了定点数的表示范围，产生溢出。

补码加法器的输入输出特性曲线



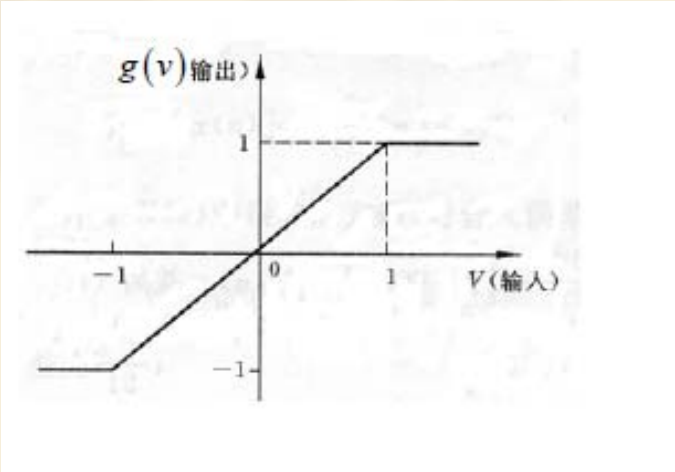
$v=\sum(\cdot)$ 是加法的结果，是和。

当 $-1\le v<1$，补码加法器的输出 $g(v)=v$
超过此范围，非线性，

为避免产生溢出振荡，要保证**|*v*| < 1**

二阶**IIR**（基本节） 系统**|*v*| < 1** 的充要条件：分母多项式系数**|*b*1| + |*b*2| < 1**

2. 溢出振荡



消除或减弱溢出振荡的办法：

用带有饱和特性的加法器
截顶使得溢出振荡消除，
但带来限幅失真（非线性失真）

减小非线性失真的方法：

对输入序列做尺度变换（减小其幅度），使得在系统中任何加法节点处都不会溢出。

6.4.2 极限环振荡

减小非线性失真的方法：

对输入序列做尺度变换（减小其幅度），使得在系统中任何加法节点处都不会溢出。

对输入信号**x(n)**，系统的第**k** 个节点的响应

$$|y_k(n)|=|\sum_{m=-\infty}^{\infty}h_k(m)x(n-m)|\le\sum_{m=-\infty}^{\infty}|h_k(m)|\cdot|x(n-m)|$$

输入信号**x(n)**的上界 $\le A_x\sum_{m=-\infty}^{\infty}|h_k(m)|$

≤ 1

允许的输入信号上界 $A_x<\dfrac{1}{\sum_{m=-\infty}^{\infty}|h_k(m)|},k=1,2$ 事先对输入信号进行尺度变换

但是，求出的A_x 可能很小，造成输入序列被过度缩小，输出精度降低。

6.4.2 极限环振荡

2. 溢出振荡

结论：

① 对高阶的系统，发生溢出的可能性大

∴ 一般不采用高阶直接型结构，而采用低阶节级联、并联

② 极点配对，使得基本节满足：分母多项式系数$|b1| + |b2| < 1$

小结

误差的三个来源：

1、输入信号的量化误差
2、系统系数的量化误差
3、运算误差

本章的目的:

• 了解误差的大小与数的长度（**ADC**位数、信号字长，系数字长）、数的表示（数制、码制、量化）、滤波器的结构有何关系• 掌握分析误差的实用方法：分别考虑、统计分析

数的表示和运算对量化的影响　　定点/浮点　　加法/乘法
原码/补码/反码　　截尾/舍入

信号量化效应

A / D 变换量化误差的统计分析
量化噪声通过线性系统

数字滤波器的**系数量化**效应
系数量化对系统极点（零点）位置的影响
系统极点（零点）位置对系数量化的灵敏度
系数量化对**FIR** 滤波器的影响

运算量化效应

乘积的舍入误差　IIR滤波器结构的影响；FIR滤波器的运算量化分析；

DFT、FFT的运算量化分析

IIR滤波器定点运算中的零输入极限环振荡（乘积尾数处理造成），

溢出振荡（加法结果超限）