Digital Filters

Diego Passuello

July 7, 2020

Contents

Chapter 1

The Direct Form

1.1 Introduction

The transfer function of an analog filter normally is represented as a ratio of two real coefficients polynomials in the Laplace variable s, with the denominator's order greater or at least equal to the numerator's one. Every real coefficient polynomial can be decomposed as the product of first or second order polynomials. So an analog filter can be implemented as a series of first or second order filters, where the filter's order is the degree of the denominator's transfer function.

In order to realize a digital filter having approximately the same frequency response of an analog one, a set of transformations is used, which allows the mapping from the Laplace transform to the so-called z-transform. Normally such a transformation consists in the substitution of the s Laplace variable with an appropriate function of z (z^{-1} corresponds to a unit time delay). One of the widely used transformations is the bilinear transformation (See Appendix D on page 59); it consists in the following substitution:

$$s \to \frac{2}{T}\frac{1-z^{-1}}{1+z^{-1}}$$

where T is the sampling period.

Having the z-transform of a digital filter, written as the ratio of two polynomials in the z^{-1} variable, it can be easily implemented with a suitable numerical algorithm. Indeed from the equation:

$$H(z) = A\frac{1+\sum\limits_{k=1}^{M} b_k z^{-k}}{1+\sum\limits_{k=1}^{N} a_k z^{-k}} = \frac{Y(z)}{X(z)} \tag{1.1}$$

remembering the meaning of z^{-1}, it is possible to compute directly the filter's answer with the following recursive procedure:

$$y_n = A\sum_{k=0}^{M} b_k\, x_{n-k} - \sum_{k=1}^{N} a_k\, y_{n-k} \qquad \text{con} \quad b_0 = 1 \tag{1.2}$$

where A is the gain coefficient of the filter and obviously $M \leqslant N$.

This algorithm requires the use of $M+N$ memory words to store previous values of x_n and y_n, as well as $M+N+1$ products and $M+N$ additions. In the case of digital filters the efficiency of its implementation is also limited by the number of memory's read and write cycles. This implementation, known in the literature as *direct form I* (see ref[1] 4.3), requires $2M$ memory accesses for reading x_{n-k} and b_k, $2N$ memory accesses for reading y_{n-k} and a_k plus the A gain reading. In addition, $N+M$ memory accesses are required for writing the new values of x_{n-k} and b_k (one can use cyclic buffers whose management, however, is normally very expensive). In conclusion we need $3N+3M+1$ memory accesses.

We can note though that the equation 1.2 can be interpreted as the series of two filters, the first one realizes the zeros of the transfer function and uses only the input signal and his previous M samples, followed by the second who realizes the poles using the N previous output values. In formulas:

$$\begin{cases} y' = A\displaystyle\sum_{k=0}^{M} b_k\, x_{n-k} \\ y_n = y' - \displaystyle\sum_{k=1}^{N} a_k\, y_{n-k} \end{cases}$$

Since the filters we are considering are linear and time-invariant, the output of the filter does not depend on the order of the factors. We can therefore create the same filter by implementing first the poles and then the zeros. This is equivalent to putting successively (*direct form II*):

$$\frac{Y(z)}{X(z)} = A\frac{Y(z)}{W(z)}\frac{W(z)}{X(z)} \tag{1.3}$$

$$\frac{W(z)}{X(z)} = 1\Big/\left(1+\sum_{k=1}^{N} a_k z^{-k}\right) \tag{1.4}$$

$$\frac{Y(z)}{W(z)} = A\left(1+\sum_{k=1}^{M} b_k z^{-k}\right) \tag{1.5}$$

The pair of equations (1.4) and (1.5) can be immediately translated into a corresponding system of difference equations:

$$\begin{cases} w_n = x_n - \displaystyle\sum_{k=1}^{N} a_k w_{n-k} \\ y_n = A\left(w_n + \displaystyle\sum_{k=1}^{M} b_k w_{n-k}\right) \end{cases} \tag{1.6}$$

The set of N values w_k represents the internal state of the filter. From this state and from the actual value of the input variable x_n, we are able to compute, at each sampling time, the next state and the value of the output variable. This update requires $M+M+1$ multiplications and $N+M$ additions but with only N memory cells to store previous values of w_n. The number of memory accesses is also reduced to $3N+M+1$ with a saving of $2M$ accesses. If the eq. (1.1) is obtained through the use of a bilinear transformation, then $N=M$ necessarily.

It can be shown that the smaller the order of a filter, the less the numerical accuracy required for its implementation; also the parameter's sensitivity of the filter from the coefficient's variations (a_k and b_k) is lower. So it is important to implement the filter as a series of first or second order sections. This doesn't imply, as can be seen easily, a larger computing power for it's implementation.

1.2 Realization: Numerical accuracy

The accuracy that we can obtain in the poles and zeroes placement of a digital filter depends naturally on the arithmetic precision with which the filter's coefficients are represented inside a computer. Let us consider the cases of first and second order filters only: we have seen that they are the fundamental blocks with which we can build an arbitrary filter.

1.2.1 First order filters: pole and zero position

The generic analog transfer function of a first order filter is:

$$H(s) = A\frac{s+\omega_z}{s+\omega_p} \tag{1.7}$$

where ω_z is the angular frequency of the zero, ω_p the angular frequency of the pole and A a gain coefficient. By using a bilinear transformation we obtain:

$$H(z) = A\frac{2+T\omega_z}{2+T\omega_p}\frac{1-\dfrac{2-T\omega_z}{2+T\omega_z}z^{-1}}{1-\dfrac{2-T\omega_p}{2+T\omega_p}z^{-1}} = B\frac{1+bz^{-1}}{1+az^{-1}} \tag{1.8}$$

If in the eq. (1.7) the zero is absent (formally $\omega_z = \infty$) then we have $b = -1$ in eq. 1.8; the same thing happens for the pole's absence and the a coefficient.

Limited arithmetic precision implies a constraint on the possible values of the a and b coefficients and thus to the possible values of the angular frequencies. If we call n the mantissa's number of bits of the floating point number's representation, then the relative accuracy by which we can approximate a number is 2^{-n}. From eq. (1.8), due to the fact that the products $T\omega_z$ and $T\omega_p$ are much lower than 2 (that is very high sampling rate compared to the pole's and zero's frequency), it happens that the a and b coefficients are very close to -1. So, in this case, 2^{-n} is also the absolute precision with which we can represent these coefficients.

Let us examine now the inverse problem: given the a and b coefficients determine the ω_z and ω_p values. We have:

$$a = \frac{T\omega_p - 2}{T\omega_p + 2} \qquad \text{that is} \qquad \omega_p = \frac{2}{T}\frac{1+a}{1-a} \tag{1.9}$$

and an analogous relation between ω_z and b.

The precision with which we can place a low frequency pole (or zero) is given by:

$$\Delta\omega_p = \left|\frac{\partial\omega_p}{\partial a}\right|\Delta a$$

where the partial derivative is evaluated at $a = -1$. Since

$$\frac{\partial\omega_p}{\partial a} = \frac{4}{T}\frac{1}{(1-a)^2}$$

it follows that

$$\Delta\omega_p = \frac{\Delta a}{T}$$

and, remembering that $\Delta a = 2^{-n}$, this implies that the pole's (or zero's) frequency resolution is:

$$\Delta f = \frac{2^{-n}}{2\pi T} = \frac{2^{-n}}{2\pi}f_c \tag{1.10}$$

where f_c is the sampling frequency. We found a linear dependence of the frequency resolution both from the sampling rate and the arithmetic accuracy. In the Virgo experiment f_c is put at 10 KHz, $n = 24$ for single precisione floating point numbers, $n = 32$ for extended precision floating point numbers and finally $n = 52$ for double precision floating point numbers. So the frequency resolution is about equal to 100 μHz for single precision, to 0.4, μHz for extended precision and to 0.4 pHz for double precision arithmetic. In the first realization of the Virgo filters we used extended precision arithmetic and therefore the frequency resolution for first order filters was about 0.4 μHz; this value represent also the lowest frequency that we can implement. In the actual realization we use double precision arithmetic so that the frequency resolution for first order filters is about 0.4 pHz; in the new realization however, the sampling frequency can be increased up to 320 kHz, so that the frequency resolution "drops" to about 12 pHz.

1.2.2 First order filters: numerical noise

We now want to evaluate the arithmetic noise introduced by a first order filter due to the finite numerical precision with which the output sequence is evaluated starting from the input sequence. Let's consider the z-transform of the first order filter, eq. (1.8)

$$H(z) = B\,\frac{1 + b\,z^{-1}}{1 + a\,z^{-1}}$$

Obviously we can write:

$$H(z) = \frac{Y(z)}{X(z)} = \frac{W(z)}{X(z)} \cdot \frac{Y(z)}{W(z)}$$

with:

$$\begin{cases} \dfrac{W(z)}{X(z)} = \dfrac{1}{1 + a\,z^{-1}} \\[2ex] \dfrac{Y(z)}{W(z)} = \dfrac{B \cdot (1 + b\,z^{-1})}{1} \end{cases} \tag{1.11}$$

From the equations (1.11) we can derive the following recursive formulas::

$$\begin{cases} w_n = -a\,w_{n-1} + x_n \\ y_n = B \cdot (w_n + b\,w_{n-1}) \end{cases} \tag{1.12}$$

Now let's compute the filter response for a constant input (zero frequency). For this we must evaluate the transfer function at point $z = 1$, which is equivalent to putting $x_n = x_{dc}$ and, once the transient phase is over, $w_n = w_{n-1} = w_{dc}$. So we get:

$$\begin{cases} w_{dc} = \dfrac{x_{dc}}{1 + a} \\ y_{dc} = B(1 + b)\,w_{dc} \end{cases}$$

The critical part in evaluating the expression (1.12) is given by the sum of $-a\,w_{dc}$ and x_{dc}. Indeed it is the sum of two numbers one of which, $-a\,w_{dc}$, is much larger than the other. In fact, their ratio is given by

$$G_{dc} = \frac{-a\,x_{dc}}{1 + a} \cdot \frac{1}{x_{dc}} = \frac{-a}{1 + a}$$

and using the value of a given by (see 1.9) $a = \dfrac{\omega_p T - 2}{\omega_p T + 2}$ we finally get:

$$G_{dc} = \frac{2 - \omega_p T}{2\,\omega_p T} = \frac{1}{\omega_p T} - \frac{1}{2}.$$

We can therefore write

$$w_{dc} = G_{dc} x_{dc} + x_{dc} = (G_{dc} + 1)\,x_{dc} \tag{1.13}$$

Now suppose that x_{dc} is very close to the maximum allowable value of the input signal. Then the value of w_{dc}, evaluated when x_{dc} has the maximum value , is the maximum value that the variable w can take during the filter operation. It is easy to see that, apart from a multiplicative constant, the sequence w_n is the response of a low-pass filter to the input sequence x_n and, with the same amplitude, the value of the response is maximum for a zero frequency input, that is, for a constant input.

Let m be the number of bits with which the input variable x_n is represented: if x_n is the output of an ADC, m is the number of effective bits of the ADC; if x_n represents an arbitrary signal then m is essentially equal to $ceil(\log_2(dynamical\_range(x_n)))$.

$2^{-m}x_{max}$ is the smallest possible variation that can have an input very close to its maximum value. Also let n be the number of bits of the mantissa of w_n: it essentially coincides with the numerical precision of the processor we use to perform the calculations. Then the smallest variation that w_{max} can have is w_{max} è $2^{-n}w_{max}$.

In order not to lose numerical precision and not to introduce arithmetic noise, w_{max} must be sensitive to the smallest variation of the input signal, that is:

$$2^{-n}w_{max} \leqslant 2^{-m}x_{max}$$

that is:

$$\frac{w_{max}}{x_{max}} \leqslant 2^{n-m}$$

But $\frac{w_{max}}{x_{max}}$ is essentially equal to $G_{dc}+1$ (eq. 1.13). We therefore have:

$$G \simeq \frac{1}{\omega_p T} \leqslant 2^{n-m} \tag{1.14}$$

and finally:

$$\omega_p \geqslant \frac{2^{m-n}}{T} = 2^{m-n} f_c \tag{1.15}$$

It is easy to see, however, that the evaluation of y_n in the eq. (1.12) does not present difficulties since it consists in the algebraic sum of numbers of the same order of magnitude.

We got an estimate of the minimum value that the pole frequency can have in order not to introduce arithmetic noise. So, for example, if f_c is equal to 10 kHz, $n = 24$ (single precision)$m = 16$ (ADC's ENOB), the frequency of the pole must be greater than 6.3 Hz. Instead, using the extended precision ($n = 32$) the frequency must be greater than 24 mHz. Finally, using double precision ($n = 52$) the frequency must be greater than 24 nHz. In current cards we use double precision arithmetic and, as already said, the sampling frequency can reach 320 kHz and the effective resolution of the ADCs is 20 bits ($m = 20$): in this case the minimum pole frequency must be greater than 12 μHz.

1.2.3 Second order filters pole's and zero's position

In this case the generic filter's analog transfer function is:

$$H(s) = A\frac{s^2 + \dfrac{\omega_z s}{Q_z} + \omega_z^2}{s^2 + \dfrac{\omega_p s}{Q_p} + \omega_p^2} \tag{1.16}$$

where ω_z and ω_p are the frequencies of the zeroes and poles, Q_z and Q_p their quality factors and A ,as usual, a gain coefficient. For simplicity we consider only one of the polynomials in the eq.(1.16):

$$P(s) = s^2 + \frac{\omega s}{Q} + \omega^2$$

Applying a bilinear transformation to this last equation, we get:

$$P(z) = \frac{4 + \frac{2\omega T}{Q} + \omega^2T^2}{T^2(1+z^{-1})^2}\cdot$$
$$\cdot\left[1 + \frac{2(\omega^2T^2-4)}{4+\frac{2\omega T}{Q}+\omega^2T^2}z^{-1} + \frac{4-\frac{2\omega T}{Q}+\omega^2T^2}{4+\frac{2\omega T}{Q}+\omega^2T^2}z^{-2}\right] \tag{1.17}$$

So the z-transform of eq. (1.16) is:

$$H(z) = B\frac{1 + cz^{-1} + dz^{-2}}{1 + az^{-1} + z^{-2}} \tag{1.18}$$

where the various coefficients have the following values:

$$\begin{aligned}
B &= A\frac{4 + \dfrac{2\omega_z T}{Q_z} + \omega_z^2 T^2}{4 + \dfrac{2\omega_p T}{Q_p} + \omega_p^2 T^2} && (1.19)\\
a &= \frac{2(\omega_p^2 T^2 - 4)}{4 + \dfrac{2\omega_p T}{Q_p} + \omega_p^2 T^2} && (1.20)\\
b &= \frac{4 - \dfrac{2\omega_p T}{Q_p} + \omega_p^2 T^2}{4 + \dfrac{2\omega_p T}{Q_p} + \omega_p^2 T^2} && (1.21)\\
c &= \frac{2(\omega_z^2 T^2 - 4)}{4 + \dfrac{2\omega_z T}{Q_z} + \omega_z^2 T^2} && (1.22)\\
d &= \frac{4 - \dfrac{2\omega_z T}{Q_z} + \omega_z^2 T^2}{4 + \dfrac{2\omega_z T}{Q_z} + \omega_z^2 T^2} && (1.23)
\end{aligned}$$

Taking into account, as in the case of first order filters, that the product ωT is much smaller than 1, we have: $B \simeq A$, $a \simeq c \simeq -2$ and $b \simeq d \simeq 1$. If in eq. (1.16) the zeroes are absent then we have $c = +2$ and $d = +1$; if only one zero is present then:

$$\begin{aligned}
B &= A\frac{T(\omega_z T + 2)}{4 + \dfrac{2\omega_p T}{Q_p} + \omega_p^2 T^2}\\
c &= \frac{2\omega_z T}{\omega_z T + 2}\\
d &= \frac{\omega_z T - 2}{\omega_z T + 2}
\end{aligned}$$

where in this case ω_z is the zero's angular frequency. Analogous relations holds for the coefficients B, a and b in the case of the absence of one or more poles.

We don't consider the case of the simultaneous absence of poles and zeroes, because this is the case of first order filters. For the same reason we don't consider the case of both real poles and zeroes, because we can implement such a filter as a series of two first order sections.

We will consider thereof only the case for witch at least one polynomial in eq. (1.16) has a couple of complex-conjugate roots. For sake of simplicity we will consider as before only one polynomial, witch z-transform is given by eq.(1.17); let us rewrite it as:

$$P(z) = C(1 + az^{-1} + bz^{-2})$$

where obviously:

$$\begin{cases}
a = \dfrac{2(\omega^2 T^2 - 4)}{4 + \dfrac{2\omega T}{Q} + \omega^2 T^2}\\[2ex]
b = \dfrac{4 - \dfrac{2\omega T}{Q} + \omega^2 T^2}{4 + \dfrac{2\omega T}{Q} + \omega^2 T^2}
\end{cases} \tag{1.24}$$

(for the moment the gain coefficient C is of no interest).

As in the case of first order filters, the a and b coefficients in eq (1.24) have a limited numeric precision. So we assume a and b as independent variables and solve the system of equations (1.24) for ω and Q; once we have ω and Q we are able to compute their accuracy as a function of the a's and b's accuracies.

From eq. (1.24) we get, after some simple algebraic manipulations:

$$\begin{cases} \omega = \dfrac{2}{T}\sqrt{\dfrac{1+b+a}{1+b-a}} \\ \\ Q = \dfrac{\sqrt{(1+b+a)(1+b-a)}}{2(1-b)} \end{cases} \tag{1.25}$$

The accuracy with which we can define ω e Q is given by the following relations:

$$\begin{cases} \Delta\omega = \left|\dfrac{\partial\omega}{\partial a}\right|\Delta a + \left|\dfrac{\partial\omega}{\partial b}\right|\Delta b \\ \\ \Delta Q = \left|\dfrac{\partial Q}{\partial a}\right|\Delta a + \left|\dfrac{\partial Q}{\partial b}\right|\Delta b \end{cases} \tag{1.26}$$

where the partial derivatives are evaluated at $a = -2$ and $b = +1$ (very low frequencies).

For the first partial derivative in (1.26) we have:

$$\begin{aligned} \frac{\partial\omega}{\partial a} &= \frac{2}{T}\frac{1}{2}\frac{1}{\sqrt{\dfrac{1+b+a}{1+b-a}}}\frac{\partial}{\partial a}\frac{1+b+a}{1+b-a} = \\ &= \frac{2}{\omega T^2}\frac{(1+b-a)+(1+b+a)}{(1+b-a)^2} = \frac{4}{\omega T^2}\frac{1+b}{(1+b-a)^2} \end{aligned}$$

so:

$$\left|\frac{\partial\omega}{\partial a}\right|_{\substack{a=-2\\b=+1}} = \frac{1}{2\omega T^2} \tag{1.27}$$

For the second partial derivative we have:

$$\begin{aligned} \frac{\partial\omega}{\partial b} &= \frac{2}{T}\frac{1}{2}\frac{1}{\sqrt{\dfrac{1+b+a}{1+b-a}}}\frac{\partial}{\partial b}\frac{1+b+a}{1+b-a} = \\ &= \frac{2}{\omega T^2}\frac{(1+b-a)-(1+b+a)}{(1+b-a)^2} = \frac{4}{\omega T^2}\frac{-a}{(1+b-a)^2} \end{aligned}$$

so:

$$\left|\frac{\partial\omega}{\partial b}\right|_{\substack{a=-2\\b=+1}} = \frac{1}{2\omega T^2} \tag{1.28}$$

For the third partial derivative we have:

$$\begin{aligned} \frac{\partial Q}{\partial a} &= \frac{1}{2(1-b)}\frac{1}{2\sqrt{(1+b+a)(1+b-a)}}\frac{\partial}{\partial a}[(1+b+a)(1+b-a)] = \\ &= \frac{1}{2(1-b)}\frac{\sqrt{(1+b+a)(1+b-a)}}{2(1+b+a)(1+b-a)}\cdot[(1+b-a)-(1+b+a)] = \\ &= Q\frac{-2a}{2\dfrac{T^2}{4}\dfrac{4}{T^2}\dfrac{1+b+a}{1+b-a}(1+b-a)^2} = \frac{-4Qa}{\omega^2T^2(1+b-a)^2} \end{aligned}$$

so:

$$\left|\frac{\partial Q}{\partial a}\right|_{\substack{a=-2\\b=+1}} = \frac{Q}{2\omega^2T^2} \tag{1.29}$$

Finally, for the last partial derivative we have:

$$\frac{\partial Q}{\partial b}=\frac{\frac{(1+b-a)+(1+b+a)}{2\sqrt{(1+b+a)(1+b-a)}}(1-b)+\sqrt{(1+b+a)(1+b-a)}}{2(1-b)^2}=$$

$$=\frac{(1-b^2)+(1+b+a)(1+b-a)}{2(1-b)^2\sqrt{(1+b+a)(1+b-a)}}=\frac{(2+2b-a^2)\sqrt{(1+b)^2-a^2}}{2(1-b)^2\frac{1+b+a}{1+b-a}(1+b-a)^2}=$$

$$=\frac{(2+2b-a^2)}{1-b}\frac{Q}{\frac{T^2}{4}\omega^2(1+b-a)^2}=\frac{4Q}{\omega^2T^2(1+b-a)^2}\frac{(2+2b-a^2)}{1-b}$$

In this last equation the limit evaluation of the term $\frac{(2+2b-a^2)}{1-b}$ for $a\to-2$ and for $b\to+1$ is problematic. In fact we can easily see that the value of such a limit depends on the order by which we make $a\to-2$ and $b\to+1$: if we allow first $a\to-2$ such a limit evaluates to -2, on the other hand if we allow first $b\to+1$ we get the indeterminate form $0/0$. From the a and b definition (eq. 1.24) we know that they are not independent: In order to evaluate the limit of $\frac{(2+2b-a^2)}{1-b}$ we can discard the quadratic terms in ωT in eq. (1.24); then we replace the approximate values of a and b so obtained into $\frac{(2+2b-a^2)}{1-b}$ getting:

$$\frac{(2+2b-a^2)}{1-b}\simeq\frac{2+2\frac{4-2\omega T/Q}{4+2\omega T/Q}-\left(\frac{-8}{4+2\omega T/Q}\right)^2}{1-\frac{4-2\omega T/Q}{4+2\omega T/Q}}$$

After some simple algebraic manipulations we get:

$$\frac{(2+2b-a^2)}{1-b}\simeq\frac{4}{2+\omega T/Q}$$

so that when we make $\omega T/Q\to0$ we get the value $+2$ for the limit.

So finally:

$$\left|\frac{\partial Q}{\partial b}\right|_{\substack{a=-2\\b=+1}}=\lim_{\substack{a\to-2\\b\to+1}}\left|\frac{4Q}{\omega^2T^2(1+b-a)^2}\cdot2\right|=\frac{Q}{2\omega^2T^2}\tag{1.30}$$

Now we can estimate the ω's and Q's resolution. In fact, from equations (1.26), (1.27), (1.28), (1.29) and (1.30), taking into account that in this case $\Delta a=2\epsilon$ and $\Delta b=\epsilon$, where $\epsilon=2^{-n}$, we have:

$$\begin{cases}\Delta\omega=\dfrac{3\epsilon}{2\omega T^2}\\[2ex]\Delta Q=\dfrac{3\epsilon Q}{2\omega^2T^2}\end{cases}\tag{1.31}$$

We note, first of all, the quadratic dependence of the resolution on the sampling frequency. A third order filter will show a cubic dependence, and so on, for higher order filters. This verifies, as we have stated before, the necessity to implement a generic filter as a series of first and second order sections.

We rewrite eq. (1.31) in order to show the relative precision of ω and Q.

$$\frac{\Delta\omega}{\omega}=\frac{\Delta Q}{Q}=\frac{3\epsilon}{2\omega^2T^2}\tag{1.32}$$

Normally the relative resolution of the frequency should be much lower than the inverse of the quality factor, that is:

$$\frac{\Delta\omega}{\omega} < \frac{1}{Q}$$

and imposing this condition on eq. (1.32), we obtain:

$$\frac{\Delta\omega}{\omega} = \frac{3\epsilon}{2\omega^2 T^2} < \frac{1}{Q}$$

that is:

$$\omega > f_c \sqrt{\frac{3\epsilon Q}{2}} \tag{1.33}$$

We note that the minimum frequency we can implement is proportional to the square root of the arithmetic precision we are using: the number of mantissa bits needed to implement a generic filter grows linearly with the filter's order.

1.2.4 Second order filters: numerical noise

As in the case of the first order filter, we now evaluate the arithmetic noise introduced by the finite numerical precision with which the output sequence is evaluated starting from the input sequence. Let's examine for now the two distinct classic implementations, namely the *direct form I* and the *direct form II.* The z-transform of a generic second order filter, as we have already seen (1.18), is:

$$H(z) = A\frac{1 + cz^{-1} + dz^{-2}}{1 + az^{-1} + bz^{-2}}$$

From this we can derive the following recursive procedure for the *direct form I*:

$$y_n = A(x_n + a\,x_{n-1} + b\,x_{n-2}) - c\,y_{n-1} - d\,y_{n-2} \tag{1.34}$$

or, in the case of the *direct form II* using the state variable w_n:

$$\begin{cases} w_n = -a\,w_{n-1} - b\,w_{n-2} + x_n \\ y_n = A\cdot(w_n + c\,w_{n-1} + d\,w_{n-2}) \end{cases} \tag{1.35}$$

We first deal with the *direct form II* which is the one that is normally used. We proceed as in the case of the first order by evaluating the response of the filter with constant input. For this we have to evaluate the transfer function for $z = 1$, which is equivalent to putting $x_n = x_{dc}$ e $w_n = w_{n-1} = w_{n-2} = w_{dc}$. We thus obtain:

$$\begin{cases} w_{dc} = \dfrac{x_{dc}}{1 + a + b} \\ y_{dc} = A(1 + c + d)\,w_{dc} \end{cases}$$

As in the case of the first order we note that part critical part in evaluating the expression (1.35) is given by the sum of $(-a - b)\,w_{dc}$ and x_{dc}. Again, this is the sum of two numbers one of which, $(-a - b)\,w_{dc}$, is much larger than the other. In fact, their ratio is given by

$$G_{dc} = \frac{(-a - b)\,x_{dc}}{1 + a + b}\cdot\frac{1}{x_{dc}} = \frac{-a - b}{1 + a + b}$$

and using the values of a and b given by (1.20) and (1.21) we obtain:

$$a + b = \frac{2(\omega_p^2 T^2 - 4) + 4 - \dfrac{2\omega_p T}{Q_p} + \omega_p^2 T^2}{4 + \dfrac{2\omega_p T}{Q_p} + \omega_p^2 T^2} = \frac{3\omega_p^2 T^2 - \dfrac{2\omega_p T}{Q_p} - 4}{4 + \dfrac{2\omega_p T}{Q_p} + \omega_p^2 T^2}$$

and

$$1+a+b=\frac{3\omega_p^2T^2-\dfrac{2\omega_pT}{Q_p}-4+4+\dfrac{2\omega_pT}{Q_p}+\omega_p^2T^2}{4+\dfrac{2\omega_pT}{Q_p}+\omega_p^2T^2}=$$

$$=\frac{4\omega_p^2T^2}{4+\dfrac{2\omega_pT}{Q_p}+\omega_p^2T^2} \tag{1.36}$$

from which, finally:

$$G_{dc}=\frac{4+\dfrac{2\omega_pT}{Q_p}-3\omega_p^2T^2}{4\omega_p^2T^2}$$

and, since we consider $\omega_pT \ll 1$, we can put:

$$G_{dc}\simeq\frac{1}{\omega_p^2T^2} \tag{1.37}$$

The maximum value of w_n in the case of a second order filter with a merit factor $Q>\sqrt{2}$, is given by $w_{max}\simeq Q\,w_{dc}$ (if $\omega_pT\ll 1$ three consecutive values of w_n differ very little from each other and, for our purpose, we can essentially assume them equal). We will therefore have:

$$w_{max}\simeq Q_pG_{dc}\simeq\frac{Q_p}{\omega_p^2T^2} \tag{1.38}$$

As with the first order filters, let m be the number of bits of the input variable x_n, with its smallest variation, in case of maximum amplitude, given by $2^{-m}x_{max}$ and let n be the number of bits of the mantissa of w_n, so that the smallest variation that can having w_{max} is given by $2^{-n}w_{max}$. In order not to lose numerical precision and not to introduce arithmetic noise it is necessary, also in this case, that w_{max} is sensitive to the smallest variation of the input, that is:

$$2^{-n}w_{max}\leqslant 2^{-m}x_{max}$$

ossia:

$$\frac{w_{max}}{x_{max}}\leqslant 2^{n-m}$$

But $\dfrac{w_{max}}{x_{max}}$ is essentially equal to Q_pG_{dc} (eq. 1.38)

$$G_{max}\simeq\frac{Q_p}{\omega_p^2T^2}\leqslant 2^{n-m} \tag{1.39}$$

and finally:

$$\omega_p\geqslant\frac{\sqrt{Q_p}}{2^{(n-m)/2}\,T}=\sqrt{Q_p}\,\frac{f_c}{2^{(n-m)/2}} \tag{1.40}$$

As one can see, the situation in the second order case is much more critical since w_{max} is proportional to Q_p and inversely proportional to the textbfsquare of $\boldsymbol{\omega_p T}$.

Again the evaluation of y_n in the eq. (1.35) does not present difficulties since it is the algebraic sum of numbers of the same order of magnitude.

We have therefore obtained a limitation on the minimum value that can take the pole frequency so as not to introduce arithmetic noise. So, for example, if f_c is equal to 10 kHz, $n=24$ (single precision arithmetic) $m=16$ (ADC ENOB), the pole frequency must be greater than 100 Hz. With extended precision arithmetic ($n=32$) the frequency must be greater than 6.3 Hz. Finally, using double precision arithmetic ($n=52$) the frequency must be greater than 6.1 mHz. These values must be multiplied by the square root of the factor of merit $Q_p$worsening the situation.

Even in the current implementation, in which we use double precision, a sampling rate of up to 320 kHz and ADC with 20 effective bits ($m = 20$) of resolution, the minimum pole frequency must be greater than $0.78, \mathrm{Hz}$ (multiplied possibly by $\sqrt{Q_p}$).

A detailed analysis of the implementation of the *direct form I* shows that the situation essentially does not change. In fact, the equation (1.34) shows that in this case we sum a very small number given by $A(x_n + a\,x_{n-1} + b\,x_{n-2})$ with a larger number given by $-c\,y_{n-1} - d\,y_{n-2}$. Always taking into account the behavior for constant input the first term is equal to $A\,x_{dc}(1 + a + b)$, while the second one is given by $y_{dc}(-c - d)$. Since the output $y_{dc} = A\,x_{dc}$ and that the sum $(-c - d)$ is of the order of the unit, the ratio of the two terms of the sum is given by:

$$G_{dc} = \frac{1}{1 + a + b} \simeq \frac{1}{\omega_p^2 T^2}$$

.

As one can see, also from the point of view of the arithmetic noise, the two implementations *direct form I* and *direct form II* are perfectly equivalent.

The situation obviously worsens if we examine the direct implementation of higher order filters. Let's take into consideration the equation (1.4) that we rewrite here:

$$\frac{W(z)}{X(z)} = \frac{1}{1 + \sum\limits_{k=1}^{N} a_k z^{-k}} \tag{1.41}$$

Setting $z = 1$ we can immediately compute the value of the ratio between w_{dc} and x_{dc}

$$\frac{w_{dc}}{x_{dc}} = \frac{1}{1 + \sum\limits_{k=1}^{N} a_k}$$

Since the coefficients a_k of z^{-k}, in the equation (1.41), are real, we can decompose the polynomial of z^{-1} in the product of polynomials of at most first and second degree with real coefficients. That is, we can write:

$$1 + \sum_{k=1}^{N} a_k z^{-k} = \prod_{1}^{M} (1 + a_{i_m} z^{-1}) \prod_{1}^{L} (1 + a_{j_l} z^{-1} + b_{j_l} z^{-2})$$

with $N = M + 2L$. Therefore, setting $z = 1$, we have:

$$1 + \sum_{k=1}^{N} a_k = \prod_{1}^{M} (1 + a_{i_m}) \prod_{1}^{L} (1 + a_{j_l} + = b_{j_l}).$$

But, from the above analysis, we have $1 + a_{i_m} \simeq \omega_{i m} T$ and $1 + a_{j_l} + b_{j_l} \simeq \omega_{j}{}_{l}^{2} T^2$ so that ultimately:

$$\frac{w_{dc}}{x_{dc}} \simeq \frac{1}{\prod\limits_{1}^{M} (\omega_{i_m} T) \prod\limits_{1}^{L} (\omega_{j_l}^2 T^2)} = \frac{f_c^N}{\prod\limits_{1}^{M} (\omega_{i_m}) \prod\limits_{1}^{L} (\omega_{j_l}^2)} \tag{1.42}$$

The latter expression clearly shows that arithmetic noise grows exponentially with the filter order. Hence the need to decompose a complex filter into the series of multiple first and second order filters. The implementation of a second-order filter in direct form also introduces an arithmetic noise that can be intolerable. In the next chapter we will study alternative embodiments that solve this problem at the expense of a lower efficiency from the point of view of the number of arithmetic operations and the number of memory access cycles.

Chapter 2

From second to first order

2.1 Cascade filters with complex coefficients

The problem with the direct implementation of second order filters is that they are ... second order filters! The trend of the ratio $\frac{w_{dc}}{x_{dc}}$ as f_c^2 depends on the fact that in the transfer function we are dealing with a second order polynomial.

One might think one can resort to higher numerical precision, e.g. quadruple precision; this solution becomes, however, computationally prohibitive. Instead, without increasing the arithmetic precision needed, we can recast the implementation of a second order filter into a series of two first order filters, for which the relative parameter resolution dependence is linear with the sampling frequency and not quadratic.

The solution is based on the fact that, since we can always decompose a second degree polynomial into the product of two first order polynomials with the use of complex coefficients, then we can decompose a second order filter with a series of two complex first order filters.

First of all we note that if an analog filter has a pair of complex-conjugate poles (or zeroes), this is also true for the corresponding numerical filter obtained with a bilinear transformation. For the moment we will consider filters for which both the poles and the zeroes are complex-conjugate; the case in which the poles or the zeroes, but not both, are real will be considered later. Clearly the case in which both the poles and zeroes are real is of no interest, because in this case the filter is a series of two real first order filters.

So we consider the eq. (1.18), which is the z-transform of a generic second order filter, and we suppose that both the numerator and the denominator have complex-coniugate roots. We can rewrite this equation in the form:

$$H(z) = B\frac{1+cz^{-1}+dz^{-2}}{1+az^{-1}+bz^{-2}} = B\frac{1-uz^{-1}}{1-vz^{-1}}\frac{1-\overline{u}z^{-1}}{1-\overline{v}z^{-1}} \tag{2.1}$$

Such an equation describes a series of two first order filters with complex coefficients. It's clear that if the input to the first filter is a real numerical sequence, it will produce a complex sequence; such a sequence, transformed by the second filter, which is the complex-conjugate of the first one, will produce the real output sequence.

That is, if we neglect the inessential gain term B:

$$\begin{aligned}
\frac{Y(z)}{X(z)} &= \frac{P(z)}{X(z)}\frac{Y(z)}{P(z)} = \frac{1-vz^{-1}}{1-uz^{-1}}\frac{1-\overline{v}z^{-1}}{1-\overline{u}z^{-1}} \\
\frac{P(z)}{X(z)} &= \frac{1-vz^{-1}}{1-uz^{-1}} \\
\frac{Y(z)}{P(z)} &= \frac{1-\overline{v}z^{-1}}{1-\overline{u}z^{-1}}
\end{aligned}$$

All we have said before is explained in the following picture:

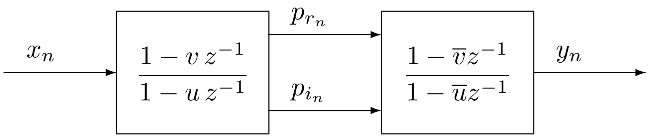
x_n
$\frac{1 - v\,z^{-1}}{1 - u\,z^{-1}}$
p_{r_n}
p_{i_n}
$\frac{1 - \overline{v}z^{-1}}{1 - \overline{u}z^{-1}}$
y_n

Figure 2.1: Complex-conjugate filters

where, obviously, $p(n)_r$ and $p(n)_i$ are the real and imaginary part respectively of the complex sequence $p(n)$.

Now, let $u = G + \mathrm{j}H$ and $v = E + \mathrm{j}F$, we will have necessarily:

$$(1 - (G + \mathrm{j}H)z^{-1})(1 - (G - \mathrm{j}H)z^{-1}) = 1 + az^{-1} + bz^{-2}$$

and so:

$$2G = -a \qquad \text{e} \qquad G^2 + H^2 = b \tag{2.2}$$

analogous equations hold between the coefficients E, F and c, d. We note that, due to the fact that $a \simeq -2$ and $b \simeq 1$, we have $G \simeq -1$ and $H \simeq 0$.

Before we examine the dependency of the ω's and Q's resolution on the new parameters, we take a look on how we could implement such a filter with a set of difference equations.

From the relations:

$$\begin{aligned}
\frac{Y}{X} &= \frac{(1 - (E + \mathrm{j}F)z^{-1})(1 - (E - \mathrm{j}F)z^{-1})}{(1 - (G + \mathrm{j}H)z^{-1})(1 - (G - \mathrm{j}H)z^{-1})} = \frac{Y}{P}\frac{P}{X} \\
\frac{P}{X} &= \frac{(1 - (E + \mathrm{j}F)z^{-1})}{(1 - (G + \mathrm{j}H)z^{-1})} = \frac{P}{V}\frac{V}{X}; \\
\frac{V}{X} &= \frac{1}{(1 - (G + \mathrm{j}H)z^{-1})}; \quad \frac{P}{V} = \frac{(1 - (E + \mathrm{j}F)z^{-1})}{1} \\
\frac{Y}{P} &= \frac{(1 - (E - \mathrm{j}F)z^{-1})}{(1 - (G - \mathrm{j}H)z^{-1})} = \frac{Y}{W}\frac{W}{P}; \\
\frac{W}{P} &= \frac{1}{(1 - (G - \mathrm{j}H)z^{-1})}; \quad \frac{Y}{W} = \frac{(1 - (E - \mathrm{j}F)z^{-1})}{1}
\end{aligned}$$

we obtain the following recursive formulas:

$$\begin{aligned}
v_{r_n} &= x_n + G\,v_{r_{n-1}} - H\,v_{i_{n-1}} \\
v_{i_n} &= + G\,v_{i_{n-1}} + H\,v_{r_{n-1}} \\
p_{r_n} &= v_{r_n} - E\,v_{r_{n-1}} + F\,v_{i_{n-1}} \\
p_{i_n} &= v_{i_n} - E\,v_{i_{n-1}} - F\,v_{r_{n-1}} \\
w_{r_n} &= p_{r_n} + G\,w_{r_{n-1}} + H\,w_{i_{n-1}} \\
w_{i_n} &= p_{i_n} + G\,w_{i_{n-1}} - H\,w_{r_{n-1}} \\
y_n &= w_{r_n} - E\,w_{r_{n-1}} - F\,w_{i_{n-1}}
\end{aligned} \tag{2.3}$$

We note, first of all, that it's necessary to store only the terms $v(n-1)_r$, $v(n-1)_i$, $w(n-1)_r$ and $w(n-1)_i$; from these and from the input variable $x(n)$, we are able to compute the new values of $v(n)$, $w(n)$ and the output variable $y(n)$. So, at every cycle, we need 8 DSP's memory accesses to read and to write the filter's status; 4 more memory accesses are needed to read the filter's coefficients: this amounts to a total of 12 memory accesses. The corresponding second order filter needs, instead, 8 memory accesses: 4 accesses to read and to write the filter's state and 4 accesses to read the filter's coefficients. For what concern the arithmetic, we note that the complex filter needs 13 additions and 14 multiplications instead of 4 additions and 4 multiplications. Due to the fact that, in this case, the computing speed is more limited by the number of memory accesses than by the arithmetic computation, a precise analysis will show that a complex filter implementation requires a computing time which is exactly double the time required by the corresponding second order filter.

2.1.1 Pole's and zero's position

Now we turn back to the estimation of the ω's and Q's accuracy in a complex filter; in order to do that, we will limit ourselves to the analysis of only one polynomial. Using eq. (1.24) we are able to express the new coefficients G and H as a function ofω and Q. So we have:

$$\begin{cases} G = \dfrac{\omega^2 T^2 - 4}{4 + \dfrac{2\omega T}{Q} + \omega^2 T^2} \\ \\ H = \dfrac{2\omega T \sqrt{4 - \dfrac{1}{Q^2}}}{4 + \dfrac{2\omega T}{Q} + \omega^2 T^2} \end{cases} \tag{2.4}$$

and, analogously from eq. (1.25) we obtain ω and Q as a function of G and H:

$$\begin{cases} \omega = \dfrac{2}{T}\sqrt{\dfrac{1 + G^2 + 2G + H^2}{1 + G^2 - 2G + H^2}} \\ \\ Q = \dfrac{\sqrt{(1 + G^2 + 2G + H^2)(1 + G^2 - 2G + H^2)}}{2(1 - G^2 - H^2)} \end{cases} \tag{2.5}$$

To compute the ω's and Q's resolution, we need to compute their partial derivatives with respect to G and H: then we should evaluate these partial derivatives for $G \to -1$ and $H \to 0$. In order to compute these derivatives we can use equations (1.27), (1.28), (1.29) and (1.30).

In fact, making use of the compound derivatives formulas, we get:

$$\frac{\partial \omega}{\partial G} = \frac{\partial \omega}{\partial a}\frac{\partial a}{\partial G} + \frac{\partial \omega}{\partial b}\frac{\partial b}{\partial G}$$

that is:

$$\left(\frac{\partial \omega}{\partial G}\right)_{\substack{G=-1\\H=0}} = \frac{1}{2\omega T^2}\left(\frac{\partial a}{\partial G} + \frac{\partial b}{\partial G}\right) = \frac{1+G}{\omega T^2}$$

We only need to evaluate the term $(1+G)$. From the definition of G (eq. (2.4)) we have:

$$1 + G = 1 + \frac{\omega^2 T^2 - 4}{4 + \dfrac{2\omega T}{Q} + \omega^2 T^2} = \frac{\dfrac{2\omega T}{Q} + 2\omega^2 T^2}{4 + \dfrac{2\omega T}{Q} + \omega^2 T^2}$$

Neglecting the ω^2 term in the numerator of this last equation with respect to the ω term and neglecting both the ω^2 and the ω terms in the denominator compared to the constant term, we obtain the following asymptotic behaviour of the $(1+G)$ term:

$$\lim_{G \to -1}(1 + G) = \frac{\omega T}{2Q}$$

so, finally, we have:

$$\left(\frac{\partial \omega}{\partial G}\right)_{\substack{G=-1\\H=0}} = \frac{1}{\omega T^2}\frac{\omega T}{Q} = \frac{1}{2QT} \tag{2.6}$$

For what concerns the second partial derivative, we have:

$$\frac{\partial\omega}{\partial H} = \frac{\partial\omega}{\partial a}\frac{\partial a}{\partial H} + \frac{\partial\omega}{\partial b}\frac{\partial b}{\partial H}$$

that is:

$$\left(\frac{\partial\omega}{\partial H}\right)_{\substack{G=-1\\H=0}} = \frac{1}{2\omega T^2}\left(\frac{\partial a}{\partial H} + \frac{\partial b}{\partial H}\right) = \frac{0+2H}{2\omega T^2}$$

and from eq (2.4), neglecting both the ω^2 and the ω terms in the denominator compared to the constant term, we obtain:

$$\left(\frac{\partial\omega}{\partial H}\right)_{\substack{G=-1\\H=0}} = \frac{1}{2\omega T^2}\omega T\sqrt{4-\frac{1}{Q^2}} = \frac{1}{2T}\sqrt{4-\frac{1}{Q^2}} \tag{2.7}$$

For what concerns the third partial derivative, we have:

$$\frac{\partial Q}{\partial G} = \frac{\partial Q}{\partial a}\frac{\partial a}{\partial G} + \frac{\partial Q}{\partial b}\frac{\partial b}{\partial G}$$

that is:

$$\left(\frac{\partial Q}{\partial G}\right)_{\substack{G=-1\\H=0}} = \frac{Q}{2\omega^2 T^2}\left(\frac{\partial a}{\partial G} + \frac{\partial b}{\partial G}\right) = \frac{Q(1+G)}{\omega^2 T^2}$$

and making use of what we already done for the first partial derivative, we obtain:

$$\left(\frac{\partial Q}{\partial G}\right)_{\substack{G=-1\\H=0}} = \frac{Q}{\omega^2 T^2}\frac{\omega T}{2Q} = \frac{1}{2\omega T} \tag{2.8}$$

Finally, for the last partial derivative, we have:

$$\frac{\partial Q}{\partial H} = \frac{\partial Q}{\partial a}\frac{\partial a}{\partial H} + \frac{\partial Q}{\partial b}\frac{\partial b}{\partial H}$$

that is:

$$\left(\frac{\partial Q}{\partial H}\right)_{\substack{G=-1\\H=0}} = \frac{Q}{2\omega^2 T^2}\left(\frac{\partial a}{\partial H} + \frac{\partial b}{\partial H}\right) = \frac{QH}{\omega^2 T^2}$$

and making use of what we already done for the second partial derivative, we obtain:

$$\left(\frac{\partial Q}{\partial H}\right)_{\substack{G=-1\\H=0}} = \frac{Q}{\omega^2 T^2}\frac{\omega T}{2}\sqrt{4-\frac{1}{Q^2}} = \frac{Q}{2\omega T}\sqrt{4-\frac{1}{Q^2}} \tag{2.9}$$

Let us write down the equivalent of eq. (1.26):

$$\begin{cases} \Delta\omega = \left|\frac{\partial\omega}{\partial G}\right|\Delta G + \left|\frac{\partial\omega}{\partial H}\right|\Delta H \\ \\ \Delta Q = \left|\frac{\partial Q}{\partial G}\right|\Delta G + \left|\frac{\partial Q}{\partial H}\right|\Delta H \end{cases}$$

where, due to the fact that $G \simeq -1$ and $H \simeq 0$, we have $\Delta G = \epsilon$ and $\Delta H = 0$. So:

$$\begin{cases} \Delta\omega = \frac{\epsilon}{2QT} \\ \\ \Delta Q = \frac{\epsilon}{2\omega T} \end{cases} \tag{2.10}$$

Showing the relative precision of ω and Q we have:

$$\frac{\Delta\omega}{\omega} = \frac{\Delta Q}{Q} = \frac{\epsilon}{2\omega QT} \tag{2.11}$$

We note, first of all, the linear, instead of quadratic, dependance of $\Delta\omega$ and ΔQ as a function of the sampling frequency: so we succeed in our scope. From eq. (2.10) we have that for $Q = 0.5$, that is when the two poles (or zeroes) become real and coincident, we have

$$\Delta\omega = \epsilon/T$$

which is exactly the same expression we obtained for the first order filters.

Imposing, as we have done for the second order filter, that:

$$\frac{\Delta\omega}{\omega} < \frac{1}{Q}$$

we get finally:

$$\omega > \frac{\epsilon f_c}{2} \tag{2.12}$$

and this expression should be compared with eq.(1.33).

What has been developed up to now, considering only the case of both complex poles an zeroes, can be extended easily, with less modifications, to the case where the poles or the zeroes are real. For example, in the case of real zeroes, we get the following relations:

$$\begin{aligned}
\frac{Y}{X} &= \frac{(1+az^{-1})(1+bz^{-1})}{(1+(c'+id')z^{-1})(1+(c'-id')z^{-1})} = \frac{Y}{U}\frac{U}{X} \\
\frac{U}{X} &= \frac{(1+az^{-1})}{(1+(c'+id')z^{-1})} = \frac{U}{V}\frac{V}{X}; \\
\frac{V}{X} &= \frac{1}{(1+(c'+id')z^{-1})}; \quad \frac{U}{V} = \frac{(1+az^{-1})}{1} \\
\frac{Y}{U} &= \frac{(1+bz^{-1})}{(1+(c'-id')z^{-1})} = \frac{Y}{W}\frac{W}{U}; \\
\frac{W}{U} &= \frac{1}{(1+(c'-id')z^{-1})}; \quad \frac{Y}{W} = \frac{(1+bz^{-1})}{1}
\end{aligned}$$

and from these we get the following recursive formulas:

$$\begin{aligned}
v(n)_r &= x(n) - c'v(n-1)_r + d'v(n-1)_i \\
v(n)_i &= - c'v(n-1)_i - d'v(n-1)_r \\
\\
u(n)_r &= v(n)_r + a\,v(n-1)_r \\
u(n)_i &= v(n)_i + a\,v(n-1)_i \\
\\
w(n)_r &= u(n)_r - c'w(n-1)_r - d'w(n-1)_i \\
w(n)_i &= u(n)_i - c'w(n-1)_i + d'w(n-1)_r \\
\\
y(n) &= w(n)_r + b\,w(n-1)_r
\end{aligned}$$

Analogous relations hold in the case of real poles. All that has been said for what concerns the arithmetic precision holds also for the cases where there are real poles or zeroes.

2.1.2 Arithmetic noise

We evaluate, as usual, the response of the filter with constant input. In the equations (2.3) on page. 16 we put $x_n = x_{dc}$. We obtain the following equations:

$$\begin{aligned} v_{dc_r} &= x_{dc} + G\,v_{dc_r} - H\,v_{dc_i} && (2.13)\\ v_{dc_i} &= \phantom{x_{dc}} + G\,v_{dc_i} + H\,v_{dc_r} && (2.14)\end{aligned}$$

$$\begin{aligned} p_{dc_r} &= v_{dc_r} - E\,v_{dc_r} + F\,v_{dc_i} && (2.15)\\ p_{dc_i} &= v_{dc_i} - E\,v_{dc_i} - F\,v_{dc_r} && (2.16)\end{aligned}$$

$$\begin{aligned} w_{dc_r} &= p_{dc_r} + G w_{dc_r} + H\,w_{dc_i} && (2.17)\\ w_{dc_i} &= p_{dc_i} + G\,w_{dc_i} - H\,w_{dc_r} && (2.18)\end{aligned}$$

$$y_{dc} = w_{dc_r} - E\,w_{dc_r} - F\,w_{dc_i} \qquad (2.19)$$

From now on, to simplify writing, we will omit the suffix dc and write the DC values of the signals using capital letters. We first deal with the calculation of the values of V_r, V_i, P_r and P_i. From the second equation (2.14) we can derive the value of V_i that we will replace in the first equation (2.13) obtaining in succession:

$$\begin{cases} V_i = \dfrac{H}{1-G} V_r \\ V_r \left[(1-G) + \dfrac{H^2}{1-G} \right] = X \end{cases}$$

dalle quali:

$$\begin{cases} V_r = \dfrac{1-G}{(1-G)^2 + H^2}\, X \\ V_i = \dfrac{H}{1-G} \cdot \dfrac{1-G}{(1-G)^2 + H^2}\, X = \dfrac{H}{(1-G)^2 + H^2}\, X \end{cases} \qquad (2.20)$$

Also in this case it is important to compute the "gain" of the filter obtained from the equation (2.13):

$$\begin{aligned} G_v = \frac{GV_r - HV_i}{X} &= G\frac{1-G}{(1-G)^2+H^2} - H\frac{H}{(1-G)^2+H^2} \\ &= \frac{G(1-G) - H^2}{(1-G)^2 + H^2} \qquad (2.21)\end{aligned}$$

Since we have (see equations (2.2) on page 16 and (1.36) on page 12):

$$\begin{cases} (1-G)^2 + H^2 = 1 - 2G + G^2 + H^2 = 1 + a + b = \dfrac{4\omega_p^2 T^2}{4 + \dfrac{2\omega_p T}{Q_p} + \omega_p^2 T^2} \\ G(1-G) - H^2 = G - G^2 + H^2 = -a/2 - b = \dfrac{\dfrac{2\omega_p T}{Q_p} - 2\omega_p^2 T^2}{4 + \dfrac{2\omega_p T}{Q_p} + \omega_p^2 T^2} \end{cases}, \qquad (2.22)$$

we ultimately get:

$$G_v = \frac{\dfrac{2\omega_p T}{Q_p} - 2\omega_p^2 T^2}{4\omega_p^2 T^2} = \frac{1}{2Q_p \omega_p T} - \frac{1}{2}.$$

For what concern V_r and V_i, using in the equations (2.20) the values of $1 + a + b$ and H given by (1.36) on page 12 and from (2.4) on page 17, we have:

$$\begin{cases} V_r = G_v\, X + X = \left(\dfrac{1}{2Q_p\omega_p T} + \dfrac{1}{2} \right) X \\ V_i = \dfrac{2\omega_p T \sqrt{4 - \dfrac{1}{Q_p^2}}}{4\omega_p^2 T^2} = \dfrac{1}{2\omega_p T} \sqrt{4 - \dfrac{1}{Q_p^2}} \end{cases} \qquad (2.23)$$

The values of G_v and V_r thus obtained are obviously not the maximum values they can have. We obtain their maximum values at the resonance frequency. For $Q \gg 1$ they are given by (see Appendix E on page 65):

$$V_{max_r} \simeq \left(\frac{2Q_p}{\omega_p T} + \frac{\omega_p T}{2} + \frac{1}{2}\right) X \simeq \left(\frac{2Q_p}{\omega_p T} + \frac{1}{2}\right) X \tag{2.24}$$

$$G_{max} \simeq \frac{2Q_p}{\omega_p T} - \frac{1}{2} \tag{2.25}$$

The value of V_i at the resonance frequency is (always see Appendix E)

$$-\frac{4 - \omega_p^2 T^2}{8\,\omega_p T} X \simeq -\frac{1}{2\,\omega_p T} X$$

The equation (2.25) shows a situation similar to that of the first order filters. The minimum frequency of the complex pole that we can realize without introducing arithmetic noise is $2Q_p$ times greater than the frequency of a corresponding real pole. So for example with $Q_p = 100$, using extended precision, a sampling rate of 10 kHz and 16-bit resolution for the input signal, the minimum pole frequency is about 4.9 Hz. With double precision this limit drops to about 4.9 μHz. Finally with double precision, a sampling frequency of 320 kHz and 20 bit input signal dynamics, the minimum pole frequency is about 2.5 mHz. These values must be compared with those shown on page 7 and on page 13.

The evaluation of P_r and P_i starting from the equations (2.15) and (2.16) on page 20 does not present difficulties from a numerical point of view. In fact it is the sum of numbers whose module ratio is at most equal to about $4Q_z^2 :\ v_{dc_i}$, which is about $2Q$ times greater than v_{dc_r} and, in turn, F which is about $2Q$ times greater than $1 + E$ (see eq. 2.23 and eq. 2.4).

The evaluation of W_r and W_i from the equations (2.17) and (2.18) on page 20 is numerically equivalent to that of V_r e di V_i, just as that of the output Y is analogous to that of P. Therefore the conclusions we have reached regarding the minimum frequency of the achievable complex pole remain the same. One can easily see that the situation does not change if the zeros (or poles) are real (I leave the proof of it as a useful exercise).

One last consideration regarding this realization: we said that 12 memory accesses, 13 sums and 14 products are needed instead of the 8 memory accesses, 4 sums and 4 products of the equivalent direct form realization and therefore this realization turns out to be about 1.5 times less efficient from the memory point of view and about 3.5 times less efficient from the arithmetic point of view. In reality, if there are multiple signals that must be processed by the same filter (this is the case with anti-aliasing or anti-image filtering during down-sampling or up-sampling phases), the difference between the two realizations is much less pronounced. In effect, the filter, as described, processes a real input signal to produce a real output signal. However, nothing prevents you from having a complex input signal and producing a complex output signal as shown in the following figure.

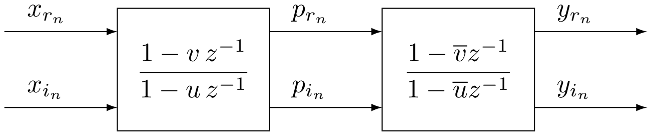
x_{r_n}
x_{i_n}
$\dfrac{1 - v\,z^{-1}}{1 - u\,z^{-1}}$
p_{r_n}
p_{i_n}
$\dfrac{1 - \overline{v} z^{-1}}{1 - \overline{u} z^{-1}}$
y_{r_n}
y_{i_n}

Figure 2.2: Complex filters in series

Due to the linearity of the system and the fact that the overall filter is a real filter, the output, seen as a complex signal, is given by the input multiplied by a real transfer function. Therefore the output sequence y_{r_n} is equal to the filtered input sequence x_{r_n}.

The same relationship exists between the output sequence y_{i_n} and the input sequence x_{i_n}. We can therefore write the following recursive formulas

$$\begin{aligned}
v_{r_n} &= x_{r_n} + G\, v_{r_{n-1}} - H\, v_{i_{n-1}} \\
v_{i_n} &= x_{i_n} + G\, v_{i_{n-1}} + H\, v_{r_{n-1}} \\
p_{r_n} &= v_{r_n} - E\, v_{r_{n-1}} + F\, v_{i_{n-1}} \\
p_{i_n} &= v_{i_n} - E\, v_{i_{n-1}} - F\, v_{r_{n-1}} \\
w_{r_n} &= p_{r_n} + G\, w_{r_{n-1}} + H\, w_{i_{n-1}} \\
w_{i_n} &= p_{i_n} + G\, w_{i_{n-1}} - H\, w_{r_{n-1}} \\
y_{r_n} &= w_{r_n} - E\, w_{r_{n-1}} - F\, w_{i_{n-1}} \\
y_{i_n} &= w_{i_n} - E\, w_{i_{n-1}} + F\, w_{r_{n-1}}
\end{aligned}$$

which require, with respect to the realization (2.3), 3 additional sums and 2 products for a total of 16 sums and 16 products, or 8 sums, 8 products and 6 memory accesses for each signal. The direct form requires, as already said, 4 sums, 4 products and 8 memory accesses. The "stereo" realization would seem even more efficient as regards access to memory and only a factor 2 more penalized as regards arithmetic operations. However, in the case of multiple signals, using the direct form we can, provided we have a sufficient number of available registers, read the filter coefficients once and for all thus saving 4 memory accesses. In the case of two signals, therefore, 8 sums, 8 products and 12 memory accesses are required against 16 sums, 16 products and 12 memory accesses of the "stereo" realization.

2.2 Parallel filters with complex coefficients

Instead of writing the ratio of two second degree polynomials as the product of the ratio of two first degree polynomials, possibly with complex coefficients, we can always write it as the sum of the ratio of two first degree polynomials, always possibly with complex coefficients. We can therefore replace a second order filter with the series of two first order filters with complex coefficients. Starting from the transfer function we can write

$$H(z) = B\frac{1 + cz^{-1} + dz^{-2}}{1 + az^{-1} + bz^{-2}} = B\left[\frac{1/2 - vz^{-1}}{1 - uz^{-1}} + \frac{1/2 - \overline{v}z^{-1}}{1 - \overline{u}z^{-1}}\right]$$

and placing $u = G + \mathrm{j}H$ and $v = E + \mathrm{j}F$ and assuming without loss of generality $B = 1$ we have

$$H(z) = \frac{Y}{X} = \frac{1/2 - (E + \mathrm{j}F)z^{-1}}{1 - (G + \mathrm{j}H)z^{-1}} + \frac{1/2 - (E - \mathrm{j}F)z^{-1}}{1 - (G - \mathrm{j}H)z^{-1}} = \frac{U}{X} + \frac{V}{X}$$

$$\begin{aligned}
\frac{U}{X} &= \frac{U}{S}\frac{S}{X} = \frac{1/2 - (E + \mathrm{j}F)z^{-1}}{1 - (G + \mathrm{j}H)z^{-1}}\,, & \frac{V}{X} &= \frac{V}{T}\frac{T}{X} = \frac{1/2 - (E - \mathrm{j}F)z^{-1}}{1 - (G - \mathrm{j}H)z^{-1}}\;; \\
\frac{S}{X} &= \frac{1/2}{1 - (G + \mathrm{j}H)z^{-1}}\,, & \frac{U}{S} &= \frac{1 - 2(E + \mathrm{j}F)z^{-1}}{1}\;; \\
\frac{T}{X} &= \frac{1/2}{1 - (G - \mathrm{j}H)z^{-1}}\,, & \frac{V}{T} &= \frac{1 - 2(E - \mathrm{j}F)z^{-1}}{1}
\end{aligned} \tag{2.26}$$

Obviously it must be

$$\begin{aligned}
1 + az^{-1} + bz^{-2} &= (1 - (G + \mathrm{j}H)z^{-1})(1 - (G - \mathrm{j}H)z^{-1}) \\
1 + cz^{-1} + dz^{-2} &= (1/2 - (E + \mathrm{j}F)z^{-1})(1 - (G - \mathrm{j}H)z^{-1}) \\
&\quad + (1/2 - (E - \mathrm{j}F)z^{-1})(1 - (G + \mathrm{j}H)z^{-1})
\end{aligned}$$

$$\begin{aligned}
1 + az^{-1} + bz^{-2} &= 1 - 2Gz^{-1} + (G^2 + H^2)z^{-2} \\
1 + cz^{-1} + dz^{-2} &= 1 - (G + 2E)z^{-1} + 2(GE + HF)z^{-2}
\end{aligned}$$

from which

$$\begin{aligned} 2G &= -a & G^2+H^2 &= b \\ G+2E &= -c & 2(GE+HF) &= d \end{aligned} \tag{2.27}$$

From the equations (2.26) we can arrive at the following recursive formulas:

$$\begin{aligned} s_{r_n} &= +\ G\,s_{r_{n-1}} - \ H\,s_{i_{n-1}} + 1/2x_n \\ s_{i_n} &= +\ G\,s_{i_{n-1}} + \ H\,s_{r_{n-1}} \\ u_{r_n} &= -\,2E\,s_{r_{n-1}} + 2\,F\,s_{i_{n-1}} + s_{r_n} \\ t_{r_n} &= +\ G\,t_{r_{n-1}} + \ H\,t_{i_{n-1}} + 1/2x_n \\ t_{i_n} &= +\ G\,t_{i_{n-1}} - \ H\,t_{r_{n-1}} \\ v_{r_n} &= -\,2E\,t_{r_{n-1}} - 2\,F\,t_{i_{n-1}} + t_{r_n} \end{aligned} \tag{2.28}$$

The filter output is $y_n = u_{r_n} + v_{r_n}$. We don't need u_{i_n} and v_{i_n} because with real input the sum $u_{i_n} + v_{i_n}$ is identically zero.

From the relations (2.28) we obtain the following recursive formulas for $(s+t)_{r_n}$, $(s-t)_{i_n}$ e $(u+v)_{r_n}$:

$$\begin{aligned} (s+t)_{r_n} &= \quad G\,(s+t)_{r_{n-1}} - \ H\,(s-t)_{i_{n-1}} + x_n \\ (s-t)_{i_n} &= \quad G\,(s-t)_{i_{n-1}} + \ H\,(s+t)_{r_{n-1}} \\ (u+v)_{r_n} &= -\,2E\,(s+t)_{r_{n-1}} + 2\,F\,(s-t)_{i_{n-1}} + (s+t)_{r_n} \end{aligned} \tag{2.29}$$

replacing in the last of the (2.29) $(s+t)_{r_n}$ with the help of the first, we obtain:

$$y_n = K1(s+t)_{r_{n-1}} + K2(s-t)_{i_{n-1}} + x_n$$

where

$$\begin{aligned} K1 &= G - 2E = c - a \\ K2 &= 2F - H = \frac{a(a-c)+2(d-b)}{\sqrt{4d-c^2}} \end{aligned} \tag{2.30}$$

By placing $w1_n = (s+t)_{r_n}$ and $w2_n = (s-t)_{i_n}$ we finally get the following filter implementation

$$\begin{aligned} w1_n &= \ G\,w1_{n-1} - \ H\,w2_{n-1} + x_n \\ w2_n &= \ H\,w1_{n-1} + \ G\,w2_{n-1} \\ y_n \ &= K1\,w1_{n-1} + K2\,w2_{n-1} + x_n \end{aligned} \tag{2.31}$$

which is an implementation of a State Variable filter.

The filter thus implemented requires 4 memory accesses for reading the filter coefficients (G, H, $K1$ and $K2$), 4 memory accesses for reading/writing the status ($w1_n$ and $w2_n$) for a total of 8 memory accesses, 5 sums and 6 products versus 8 memory accesses, 4 sums and 4 products of direct implementation.

2.2.1 Poles Position

Since in this realization the filter coefficients, which are the coordinates of the poles $v, u = E \pm \mathrm{j}F$, are the same as in the cascade realization, the analysis of the position of the poles leads to exactly the same conclusions as previously discussed (see equations (2.10 and (2.11)). Again, therefore, we have a linear instead of a quadratic dependence of $\Delta\omega$ and ΔQ from the sampling frequency.

2.2.2 Arithmetic noise

Proceeding as with the series filter we calculate the response of the filter with constant input. In the equations (2.31) we put $x_n = x_{dc}$, $w1_n = w1_{dc}$ and $w2_n = w2_{dc}$. We thus obtain the following system:

$$\begin{cases} w1_{dc} = G\,w1_{dc} - H\,w2_{dc} + x_{dc} \\ w2_{dc} = H\,w1_{dc} + G\,w2_{dc} \end{cases} \tag{2.32}$$

namely

$$\begin{cases} (1-G)w1_{dc} + \qquad H\,w2_{dc} = x_{dc} \\ -H \quad w1_{dc} + (1-G)w2_{dc} = 0 \end{cases}$$

which once solved gives

$$\begin{cases} w1_{dc} = \dfrac{(1-G)x_{dc}}{1-2G+G^2+H^2} \\ w2_{dc} = \dfrac{H\,x_{dc}}{1-2G+G^2+H^2} \end{cases} \tag{2.33}$$

and using the values of $((1-G)-H^2$ ((2.22) on page 20), G and H ((2.4) on page 17):

$$\begin{cases} w1_{dc} = \dfrac{\dfrac{2\omega_p T}{Q} + 2\omega_p^2 T^2}{4\omega_p^2 T^2} x_{dc} = \left(\dfrac{1}{2Q_p\omega_p T} + \dfrac{1}{2}\right) x_{dc} \\ w2_{dc} = \dfrac{2\omega_p T\sqrt{4-\dfrac{1}{Q^2}}}{4\omega_p^2 T^2} x_{dc} = \dfrac{1}{2\omega_p T}\sqrt{4-\dfrac{1}{Q^2}}\, x_{dc} \end{cases} \tag{2.34}$$

From the equations (2.32) we can write the "gains" of the filter:

$$\begin{cases} G_{w1} = G\,w1_{dc} - H\,w2_{dc} = \dfrac{G(1-G)-H^2}{1-2G+G^2+H^2} \\ G_{w2} = \dfrac{H\,w1_{dc}}{G\,w2_{dc}} = \dfrac{(1-G)}{G} \end{cases}$$

and using the values of $(G(1-G)-H^2)$ and $(1-2G+G^2+H^2)$ already obtained previously (see eq. (2.22) on page 20)

$$\begin{cases} G_{w1} = \dfrac{\dfrac{2\omega_p T}{Q_p} - 2\omega_p^2 T^2}{4\omega_p^2 T^2} = \dfrac{1}{2Q_p\omega_p T} - \dfrac{1}{2} \\ G_{w2} = \dfrac{\dfrac{2\omega_p T}{Q_p} + 2\omega_p^2 T^2}{4-\omega_p^2 T^2} \simeq \dfrac{\omega_p T}{2Q_p} \end{cases} \tag{2.35}$$

The detailed evaluation of the values of $w1$, $w2$ and of the "gains" at the resonant frequency can be found in the appendix F. Here are the results obtained:

$$\begin{cases} w1_{RIS} \simeq \left(\dfrac{\sqrt{4\,Q_p^2+1}}{2\,\omega_p T} + \dfrac{Q_p}{\sqrt{4\,Q_p^2+1}} \right) x_{RIS} \\ \qquad\;\; \simeq \left(\dfrac{Q_p}{\omega_p T} + \dfrac{1}{2} \right) x_{RIS} \\ w2_{RIS} \simeq \left(\dfrac{\sqrt{4\,Q_p^2+1}}{2\,\omega_p T} + \dfrac{\sqrt{4\,Q_p^2+1}}{8} \right) x_{RIS} \\ \qquad\;\; \simeq \left(\dfrac{Q_p}{\omega_p T} + \dfrac{Q_p}{4} \right) x_{RIS} \end{cases} \tag{2.36}$$

$$\begin{cases} G_{w1_{RIS}} \left(\dfrac{\sqrt{4\,Q_p^2+1}}{2\,\omega_p T} - \sqrt{4\,Q_p^2+1} \right) \simeq \dfrac{Q_p}{\omega_p T} \\ G_{w2_{RIS}} \simeq \dfrac{2\,\omega_p T}{Q_p(4-\Omega'^2)} \sqrt{4\,Q^2+1} \simeq \omega_p T \end{cases} \tag{2.37}$$

The first of the equations (2.37) shows that in the case of parallel filters an improvement of a factor two is obtained compared to the series (see eq. (2.25 on page 21))

In the next chapter we will show that this realization is somewhat optimal.

Chapter 3

Minimum arithmetic noise

3.1 On state variable filters

A generic second order system can be described with the following *state-space* model:

$$\begin{cases} S = z^{-1}\mathbf{A}\,S + \mathbf{B}\,X \\ Y = \mathbf{C}\,S + \mathbf{D}\,X \end{cases} \tag{3.1}$$

where S is the state vector, X is the input vector, Y the output vector, $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, $\mathbf{D}$ real matrices of suitable size. The eigenvalues Of the $\mathbf{A}$ matrix are the poles of the system. Its eigenvectors or eigenstates are the normal modes of the system. The $\mathbf{A}$ matrix is Also known as the transition matrix. It allows the transition from the state at time n to the state at time $n+1$. The dynamics of the system is determined by the properties of the matrix $\mathbf{A}$ which, in the case of a second order system, is a 2x2 square matrix and the state vector S has two components u and v so that we can write $S = \begin{pmatrix} u & v \end{pmatrix}^T$. We will also assume that the input vector X has only one component so that the matrix $\mathbf{B}$ is a 2x1 rectangular matrix whose components we will denote b_1 and b_2. We will now deal with the study of the first of the equations (3.1). The second of the equations (3.1) does not present problems from the numerical point of view and therefore we will not take it into consideration.

We have seen in chapter 1 that a second order filter, whose transfer function, apart from a multiplicative coefficient inessential for our study, is given by

$$H(z) = \frac{1 + cz^{-1} + dz^{-2}}{1 + az^{-1} + bz^{-2}},$$

can be implemented, using an auxiliary variable w, through the following recursive formulas (*direct form II*):

$$\begin{aligned}
&w_n = -a\,w_{n-1} - b\,w_{n-2} + x_n \\
&y_n = w_n + c\,w_{n-1} + d\,w_{n-2} = (c-a)w_{n-1} + (d-b)w_{n-2} + x_n \\
&w_{n-1} = w_n \\
&w_{n-2} = w_{n-1}
\end{aligned}$$

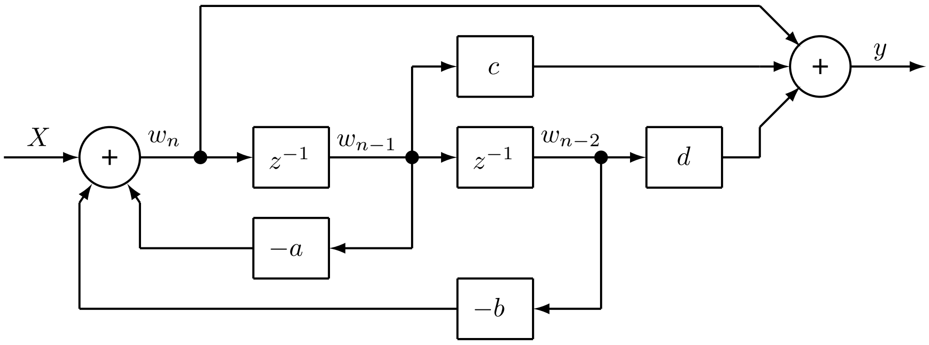

Figure 3.1: Controller canonical form

Identifying the state $S = \begin{pmatrix} u & v \end{pmatrix}^T$ con la coppia $S = \begin{pmatrix} w_{n-1} & w_{n-2} \end{pmatrix}^T$ we can write

$$\begin{aligned} \begin{pmatrix} u \\ v \end{pmatrix} &= z^{-1} \begin{pmatrix} -a & -b \\ 1 & 0 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \end{pmatrix} \begin{pmatrix} x \end{pmatrix} \\ y &= \begin{pmatrix} c-a & d-b \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} + \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} x \\ 0 \end{pmatrix} \end{aligned} \tag{3.2}$$

We obviously have:

$$\begin{aligned} \mathbf{A} &= \begin{pmatrix} -a & -b \\ 1 & 0 \end{pmatrix} & \mathbf{B} &= \begin{pmatrix} 1 \\ 0 \end{pmatrix} \\ \mathbf{C} &= \begin{pmatrix} c-a & d-b \end{pmatrix} & \mathbf{D} &= \begin{pmatrix} 1 & 0 \end{pmatrix} \end{aligned} \tag{3.3}$$

This implementation is represented in figure 3.1 and is known in linear systems theory as *Controller canonical form* (see ref [2] 2.1).

Let us now consider a generic second order system. The evolution of its state is described by the generic transition matrix **A** and by the generic input matrix **B**. The **A** matrix is a 2x2 matrix while the **B** matrix is a 2x1 matrix. We therefore have the following system

$$\begin{cases} u = a_1 z^{-1} u + a_2 z^{-1} v + b_1 x \\ v = a_3 z^{-1} u + a_4 z^{-1} v + b_2 x \end{cases} \tag{3.4}$$

$$\begin{cases} \left(1 - a_1 z^{-1}\right) u - a_2 z^{-1} v = b_1 x \\ -a_3 z^{-1} u + \left(1 - a_4 z^{-1}\right) v = b_2 x \end{cases} \tag{3.5}$$

which solved gives

$$\begin{cases} u = \dfrac{b_1 \left(1 - a_4\, z^{-1}\right) + b_2\, a_2\, z^{-1}}{(1 - a_1\, z^{-1})\,(1 - a_4\, z^{-1}) - a_2\, a_3\, z^{-2}}\, x \\ v = \dfrac{b_2 \left(1 - a_1\, z^{-1}\right) + b_1\, a_3\, z^{-1}}{(1 - a_1\, z^{-1})\,(1 - a_4\, z^{-1}) - a_2\, a_3\, z^{-2}}\, x \end{cases} \tag{3.6}$$

The denominator of the solutions (3.6) is

$$\begin{aligned} \mathbf{Den} &= \left(1 - a_1\, z^{-1}\right) \left(1 - a_4\, z^{-1}\right) - a_2\, a_3\, z^{-2} \\ &= \left(a_1\, a_4 - a_2 a_3\right) z^{-2} - \left(a_1 + a_4\right) z^{-1} + 1 \\ &= \mathrm{Det}(\mathbf{A}) z^{-2} - \mathrm{Tr}(\mathbf{A}) z^{-1} + 1 \end{aligned} \tag{3.7}$$

With the "gains" of the filter

$$\begin{cases} G_u = \dfrac{a_1 z^{-1} u + a_2 z^{-1} v}{b_1 x} \\ G_v = \dfrac{a_3 z^{-1} u + a_4 z^{-1} v}{b_2 x} \end{cases} \tag{3.8}$$

we can write the state's equations (3.4) in the following way:

$$\begin{cases} u = G_u \, b_1 x + b_1 x = (G_u + 1)\, b_1 x \\ v = G_v \, b_2 x + b_2 x = (G_v + 1)\, b_2 x \end{cases} \tag{3.9}$$

From the recursive equations (3.4) we also obtain two other conditions necessary for not introducing arithmetic noise: the ratio of the terms that appear in the sum to the numerators must not have a direct or inverse quadratic dependence on the sampling frequency. Therefore it must be:

$$\begin{cases} G'_u = \dfrac{a_1 u}{a_2 v} = O(\omega_p T) \text{ oppure } O(\omega_p^{-1} T^{-1}) \\ G'_v = \dfrac{a_3 u}{a_4 v} = O(\omega_p T) \text{ oppure } O(\omega_p^{-1} T^{-1}) \end{cases} \tag{3.10}$$

These relations are obviously valid, the first for $a_2 \neq 0$ and the second for $a_4 \neq 0$; otherwise we must take into consideration the reciprocals of the (3.10) in which case either we will have $G'_u = 0$ or $G'_v = 0$. The relationships (3.10) also apply if b_1 or b_2 are null.

By using the (3.6) and remembering the (3.7) the "gains" become

$$\begin{aligned} G_u &= \frac{a_1 b_1 z^{-1} - a_1 a_4 b_1 z^{-2} + a_2 b_2 z^{-1} + a_2 a_3 b_1 z^{-2}}{b_1 \,\text{Den}} \\ &= \frac{(a_1 b_1 + a_2 b_2)\, z^{-1} - (a_1 a_4 - a_2 a_3)\, b_1 z^{-2}}{b_1 \,\text{Den}} \\ &= \frac{(a_1 b_1 + a_2 b_2)\, z^{-1}}{b_1 \,\text{Den}} - \frac{\text{Det}(\mathbf{A}) z^{-2}}{(\text{Det}(\mathbf{A}) z^{-2} - \text{Tr}(\mathbf{A}) z^{-1} + 1)} \end{aligned} \tag{3.11}$$

$$\begin{aligned} G_v &= \frac{a_3 b_1 z^{-1} + a_2 a_3 b_2 z^{-2} + a_4 b_2 z^{-1} - a_1 a_2 b_2 z^{-2}}{b_2 \,\text{Den}} \\ &= \frac{(a_3 b_1 + a_4 b_2)\, z^{-1} - (a_1 a_4 - a_2 a_3)\, b_2 z^{-2}}{b_2 \,\text{Den}} \\ &= \frac{(a_3 b_1 + a_4 b_2)\, z^{-1}}{b_2 \,\text{Den}} - \frac{\text{Det}(\mathbf{A}) z^{-2}}{(\text{Det}(\mathbf{A}) z^{-2} - \text{Tr}(\mathbf{A}) z^{-1} + 1)} \end{aligned} \tag{3.12}$$

Obviously in the *direct form II*, implemented through the (3.2), we have $a_1 = -a$, $a_2 = -b$, $a_3 = 1$, $a_4 = 0$, $b_1 = 1$ and $b_2 = 0$ and therefore $\text{Det}(\mathbf{A}) = b$, $\text{Tr}(\mathbf{A}) = -a$, $\mathbf{Den} = 1 + a\, z^{-1} + b\, z^{-2}$ and the "gain" G_u becomes:

$$\begin{aligned} G_u &= \frac{-a\, z^{-1}}{\text{Den}} - \frac{\text{Det}(\mathbf{A}) z^{-2}}{(\text{Det}(\mathbf{A}) z^{-2} - \text{Tr}(\mathbf{A}) z^{-1} + 1)} \\ &= \frac{-a\, z^{-1} - b\, z^{-2}}{1 + a\, z^{-1} + b\, z^{-2}} = \frac{1}{1 + a\, z^{-1} + b\, z^{-2}} - 1 \end{aligned}$$

Regarding the "gain" G_v we note that, if $b_2 = 0$, G_v loses meaning and eq. (3.4) simply states that the current value of v is the previous value of u, which obviously presents no numerical problem.

We can get the value of G_u with constant input and at the resonance frequency by setting $z = 1$ and $z = \mathrm{e}^{\mathrm{j}\Omega}$ where $\Omega = 2\arctan(\omega_0 T/2)$ (see the properties of the bilinear transformation in the appendix D)

$$\begin{aligned} G_{u_{dc}} &= \frac{1}{1+a+b} - 1 = \frac{1}{\omega_p^2 T^2} + \frac{1}{2\omega_p T Q_p} - \frac{3}{4} \\ G_{u_{ris}} &= \frac{Q_p}{\omega_p^2 T^2}\left(1 + \frac{\omega_p T}{2Q_p} + \frac{1}{2}\right) + O\left(\omega_p T\right) \end{aligned} \tag{3.13}$$

where we made use of the relations (1.36) on page 12.We note the quadratic dependence on the sampling frequency.

The *state-space* representation is not unique. Given any non-singular $\mathbf{T}$ matrix we can obtain a new representation by placing:

$$\widetilde{S} = \mathbf{T}^{-1} S \qquad \widetilde{\mathbf{A}} = \mathbf{T}^{-1}\mathbf{A}\,\mathbf{T} \qquad \widetilde{\mathbf{B}} = \mathbf{T}^{-1}\mathbf{B} \qquad \widetilde{\mathbf{C}} = \mathbf{C}\,\mathbf{T} \qquad \widetilde{\mathbf{D}} = \mathbf{D}$$

The $\widetilde{\mathbf{A}}$ matrix is said to be similar to the $\mathbf{A}$ matrix. The new state becomes

$$\begin{cases} \tilde{u} = \dfrac{\tilde{b}_1\left(1 - \tilde{a}_4\, z^{-1}\right) + \tilde{b}_2\,\tilde{a}_2\, z^{-1}}{\left(1 - \tilde{a}_1 z^{-1}\right)\left(1 - \tilde{a}_4 z^{-1}\right) - \tilde{a}_2\,\tilde{a}_3\, z^{-2}}\, x \\ \tilde{v} = \dfrac{\tilde{b}_2\left(1 - \tilde{a}_1\, z^{-1}\right) + \tilde{b}_1\,\tilde{a}_3\, z^{-1}}{\left(1 - \tilde{a}_1\, z^{-1}\right)\left(1 - \tilde{a}_4\, z^{-1}\right) - \tilde{a}_2\,\tilde{a}_3\, z^{-2}}\, x \end{cases} \tag{3.14}$$

We obviously have that $\mathrm{Det}(\widetilde{\mathbf{A}} - \lambda\mathbf{I}) = \mathrm{Det}(\mathbf{A} - \lambda\mathbf{I})$ and therefore a similitude transformation leaves the determinant and the trace of the matrices unchanged and therefore also leaves the value of **Den**.unchanged.

Let us now consider a generic $\mathbf{T}$ transformation. If we multiply the matrix $\mathbf{T}$ by a constant, this is simply equivalent to a change of scale that absolutely does not change the dynamics of the filter so we can always assume that $\mathrm{Det}(\mathbf{T}) = 1$. Therefore we have for $\mathbf{T}$ and its inverse $\mathbf{T}^{-1}$

$$\mathbf{T} = \begin{pmatrix} t_1 & t_2 \\ t_3 & t_4 \end{pmatrix} \qquad \mathbf{T^{-1}} = \begin{pmatrix} t_4 & -t_2 \\ -t_3 & t_1 \end{pmatrix}$$

and the components of the matrices $\widetilde{\mathbf{A}}$ and $\widetilde{\mathbf{B}}$ become

$$\begin{aligned} \tilde{a}_1 &= a_1\, t_1\, t_4 + a_2\, t_3\, t_4 - a_3\, t_1\, t_2 - a_4\, t_2\, t_3 \\ \tilde{a}_2 &= (a_1 - a_4)\, t_2\, t_4 + a_2\, t_4^2 - a_3\, t_2^2 \\ \tilde{a}_3 &= (a_4 - a_1)\, t_1\, t_3 + a_3\, t_1^2 - a_2\, t_3^2 \\ \tilde{a}_4 &= -a_1\, t_2\, t_3 - a_2\, t_3\, t_4 + a_3\, t_1\, t_2 + a_4\, t_1\, t_4 \\ \\ \tilde{b}_1 &= b_1\, t_4 - b_2\, t_2 \\ \tilde{b}_2 &= -b_1\, t_3 + b_2\, t_1 \end{aligned} \tag{3.15}$$

Two cases now arise depending on whether the value of b_2 is null or not.

3.2 The $b_2 = 0$ case

Let's first examine the much more important case where the coefficient b_2 is zero and $b_1 = 1$. The state (3.6) simply becomes

$$\begin{cases} u = \dfrac{b_1\left(1 - a_4\, z^{-1}\right)}{\mathbf{Den}}\, x = \dfrac{1 - a_4\, z^{-1}}{\mathbf{Den}}\, x \\ v = \dfrac{b_1\, a_3\, z^{-1}}{\mathbf{Den}}\, x = \dfrac{a_3\, z^{-1}}{\mathbf{Den}}\, x \end{cases} \tag{3.16}$$

The "gain" of the G_u filter is given directly by the first of the (3.8). As we have already noted, if $b_2 = 0$, the second of the (3.8) loses its meaning; but if $a_4 \neq 0$ we can identify the "gain" G_v with G'_v, that is as the ratio of the first to the second term of the sum $a_3z^{-1}u + a_4z^{-1}v$ of the eq. (3.4). We therefore have:

$$\begin{cases} G_u = \dfrac{a_1z^{-1}u + a_2z^{-1}v}{b_1x} = \dfrac{a_1z^{-1}u + a_2z^{-1}v}{x} \\ G_v = \dfrac{a_3z^{-1}u}{a_4z^{-1}v} = \dfrac{a_3}{a_4} \cdot \dfrac{u}{v} \end{cases} \tag{3.17}$$

With the help of the "gains" we can rewrite the state equations in the following way (see eq. (3.4)):

$$\begin{cases} u = G_u\, x + x = (G_u + 1)\, x \\ v = (G_v + 1)\, a_4z^{-1}v \end{cases} \tag{3.18}$$

and using the (3.16)

$$\begin{cases} G_u = \dfrac{a_1\, z^{-1}\left(1 - a_4\, z^{-1}\right) + a_2a_3\, z^{-1}}{\mathbf{Den}}\, z^{-1} = \dfrac{a_1z^{-1}}{\mathbf{Den}} - \dfrac{\mathrm{Det}(\mathbf{A})z^{-2}}{\mathbf{Den}} \\ G_v = \dfrac{1 - a_4z^{-1}}{a_4\, z^{-1}} \end{cases} \tag{3.19}$$

Let's consider the similarity transformations such that even $\tilde{b}_2 = 0$. It must be $t_3 = 0$ (see eq. (3.15)) and since $\mathrm{Det}(\mathbf{T}) = 1$ it must be $t_4 = 1/t_1$. Therefore the matrix $\mathbf{T}$ becomes $\mathbf{T} = \begin{pmatrix} t_1 & t_2 \\ 0 & 1/t_1 \end{pmatrix}$. The matrices $\widetilde{\mathbf{A}}$ and $\widetilde{\mathbf{B}}$, the transforms of the matrices $\mathbf{A}$ and $\mathbf{B}$ of the eq. (3.3), take the following form

$$\widetilde{\mathbf{A}} = \begin{pmatrix} -a - t_1\, t_2 & -a\, t_2/t_1 - b/t_1^2 - t_2^2 \\ t_1^2 & t_1\, t_2 \end{pmatrix} \quad \widetilde{\mathbf{B}} = \begin{pmatrix} 1/t_1 \\ 0 \end{pmatrix} \tag{3.20}$$

It is easy to verify that $\mathrm{Det}(\widetilde{\mathbf{A}}) = b$ and $\mathrm{Tr}(\widetilde{\mathbf{A}}) = -a$ as it necessarily must be. The state (3.14) becomes

$$\begin{cases} \tilde{u} = \dfrac{1 - t_1\, t_2\, z^{-1}}{\mathbf{Den}}\, \dfrac{x}{t_1} \\ \tilde{v} = \dfrac{t_1^2\, z^{-1}}{\mathbf{Den}}\, \dfrac{x}{t_1} \end{cases} \tag{3.21}$$

and the "gains" $\widetilde{G}_u$ and $\widetilde{G}_v$ (with G_u I mean the "gain" of the *direct form II*) become

$$\begin{cases} \widetilde{G}_u = \dfrac{(-a - t_1\, t_2)\, z^{-1}}{\mathbf{Den}} - \dfrac{\mathrm{Det}(\mathbf{A})z^{-2}}{\mathbf{Den}} = G_u - \dfrac{t_1\, t_2\, z^{-1}}{\mathbf{Den}} \\ \widetilde{G}_v = \dfrac{1 - t_1\, t_2\, z^{-1}}{t_1\, t_2\, z^{-1}} \end{cases} \tag{3.22}$$

Let's analyze the behavior with constant input and at the resonance frequency. The asymptotic value of the state as a response to a constant input becomes

$$\begin{cases} \tilde{u} = \dfrac{1 - t_1\, t_2}{\mathbf{Den}}\, \dfrac{x_{dc}}{t_1} \\ \tilde{v} = \dfrac{t_1^2}{\mathbf{Den}}\, \dfrac{x_{dc}}{t_1} \end{cases} \tag{3.23}$$

while the corresponding "gains" value, given by the (3.22) are

$$\begin{cases} \widetilde{G}_{u_{dc}} = G_{u_{dc}} - \dfrac{t_1\, t_2}{1+a+b} \\ \widetilde{G}_{v_{dc}} = \dfrac{1 - t_1\, t_2}{t_1\, t_2} \end{cases} \tag{3.24}$$

Let's first examine the behavior of $\widetilde{G}_u$ in detail. We know from eq (3.13) that the value of $G_{u_{dc}}$ contains a quadratic dependence on the sampling frequency and more precisely it is

$$G_{u_{dc}} = \frac{1}{1+a+b} - 1 = \frac{1}{\omega_p^2 T^2} + \frac{1}{2\omega_p T Q_p} - \frac{3}{4}$$

so that

$$\widetilde{G}_{u_{dc}} = \frac{1 - t_1\, t_2}{\omega_p^2 T^2} + \frac{1 - t_1\, t_2}{2\omega_p T Q_p} + \frac{1 - t_1\, t_2}{4} - 1$$

To remove the quadratic dependence on the sampling frequency we have an infinite number of choices. Just put $t_1\, t_2 = 1 + \omega_p T \cdot R(\omega_p T)$, where $R(\omega_p T)$ is an arbitrary function of the argument with the only obvious limitation that there must be no poles in the origin, to have

$$\widetilde{G}_{u_{dc}} = \frac{R(\omega_p T)}{\omega_p T} + \frac{R(\omega_p T)}{2Q_p} + \frac{\omega_p T \cdot R(\omega_p T)}{4} - 1$$

We must now decide the most appropriate form for $R(\omega_p T)$. The most obvious choice seems to be simply $R(\omega_p T) = 0$, that is $t_1\, t_2 = 1$, in which case the "gain" $\widetilde{G}_{u_{dc}}$ is reduced to the constant -1, the "gain" $\widetilde{G}_{v_{dc}}$ is canceled (eq. (3.24)) and the state becomes $\widetilde{G}_{v_{dc}}$ si annulla (eq. (3.24)) e lo stato diviene

$$\begin{cases} \tilde{u}_{dc} = 0 \\ \tilde{v}_{dc} = \dfrac{t_1 x_{dc}}{1+a+b} = \left(\dfrac{1}{\omega_p^2 T^2} + \dfrac{1}{2\omega_p T Q_p} + \dfrac{1}{4}\right) t_1^2\, \dfrac{x_{dc}}{t_1} \end{cases} \tag{3.25}$$

The fact that the state variable $\tilde{u}_{dc}$ and the "gain" $\widetilde{G}_{v_{dc}}$ cancel each other does not mean that the system is indeterminate. It is only their asymptotic value that cancels out with constant input. The transient phase of the state, with one unit step at the input, is a damped oscillation: the u component of the state is a damped oscillation with zero mean and initial amplitude equal to approximately $1/\omega_p\, T$; the component v , on the other hand, is a damped oscillation around the value of $1/\omega_p^2\, T^2$ and an initial amplitude equal to approximately $1/\omega_p^2\, T^2$. The calculation shows that the "gain" G_u is a damped oscillation with an average value of -1 and an initial amplitude equal to $1/\omega_p\, T$. We have the same behavior for the "gain" G_v with the difference that its asymptotic value is zero. There is therefore a linear dependence of the "gains" with the sampling frequency. For the detailed calculation of the state transient phase and "gains" with $t_1\, t_2 = 1$ see the appendix G.

All this, however, concerns the behavior of the filter with constant input. Let us now analyze its behavior at the resonance frequency. From the equation (3.22) we get the following expression for the "gain" G_u:

$$\widetilde{G}_u = \frac{(-a - t_1\, t_2)\, z^{-1} - \mathrm{Det}(\mathbf{A}) z^{-2}}{\mathbf{Den}} = \frac{(-a - t_1\, t_2)\, z^{-1} - b\, z^{-2}}{\mathbf{Den}}.$$

Its square module at the resonance frequency is

$$\begin{aligned} |\widetilde{G}_u|^2_{RIS} &= \left.\frac{\left[(-a - t_1\, t_2)\, z^{-1} - b\, z^{-2}\right]\left[(-a - t_1\, t_2)\, z - b\, z^2\right]}{|\mathbf{Den}|^2}\right|_{z=e^{j\Omega}} \\ &= \frac{a^2 + 2\, a\, (t_1\, t_2) + (t_1\, t_2)^2 + b^2 + 2a\, b\, \cos\Omega + 2\, (t_1\, t_2)\; b\, \cos\Omega}{|\mathbf{DEN}|^2} \end{aligned}$$

and is a quadratic function of the product $(t_1\, t_2)$. **DEN** is the value of **Den** at the resonance frequency (see (F.6) of the appendix F on page74). The minimum value of $|\widetilde{G}_u|^2_{RIS}$ occurs when

$$2\,a + 2\,(t_1\, t_2) + 2\,b\,\cos\Omega = 0$$

that is when

$$t_1\, t_2 = -a - b\,\cos\Omega = \cos\Omega = \frac{4-\omega_p^2 T^2}{4+\omega_p^2 T^2} \tag{3.26}$$

where we made use of the relationships (E.6) of the appendix Eon page 66. With this choice of the product $(t_1\, t_2)$ the minimum value of $\widetilde{G}_u$ becomes

$$\begin{aligned}
\min(\widetilde{G}_u)_{RIS} &= z^{-1}\,\frac{-a-\cos\Omega - b\cos\Omega + \mathrm{j}\,b\sin\Omega}{\mathbf{DEN}} = \mathrm{e}^{-\mathrm{j}\Omega}\,\frac{-a-\cos\Omega(1+b)+\mathrm{j}\,b\sin\Omega}{\mathbf{DEN}}\\
&= \mathrm{e}^{-\mathrm{j}\Omega}\,\frac{-a+\dfrac{a}{1+b}(1+b)+\,jb\sin\Omega}{\mathbf{DEN}} = \mathrm{e}^{-\mathrm{j}\Omega}\,\frac{\mathrm{j}\,b\sin\Omega}{\mathbf{DEN}}\\
&= \mathrm{e}^{-\mathrm{j}\Omega}\,\frac{\mathrm{j}\,\dfrac{4-2\omega_p T/Q_p+\omega_p^2T^2}{\mathrm{Den}}\,\dfrac{4\omega_p T}{4+\omega_p^2T^2}}{-\mathrm{j}\dfrac{16\omega_p^2T^2/Q_p}{\mathrm{Den}(4+\omega_p^2T^2)}\,\mathrm{e}^{-\mathrm{j}\Omega}}\\
&= -\frac{4-2\omega_p T/Q_p+\omega_p^2T^2}{4\omega_p T/Q_p} = -\frac{Q_p}{\omega_p T}+\frac{1}{2}-\frac{\omega_p T\,Q_p}{4}
\end{aligned}$$

Again from the equation (3.22) we obtain for the square module of the "gain" G_v at the resonance frequency the following expression:

$$|\widetilde{G}_v|^2_{RIS} = \frac{(1-(t_1\,t_2)\,\mathrm{e}^{-\mathrm{j}\Omega})(1-(t_1\,t_2)\,\mathrm{e}^{\mathrm{j}\Omega})}{(t_1\,t_2)^2} = \frac{1+(t_1\,t_2)^2-2(t_1\,t_2)\cos\Omega}{(t_1\,t_2)^2}$$

also a quadratic function of the product $(t_1\, t_2)$.We get the minimum value of $|\widetilde{G}_v|^2_{RIS}$ when

$$\begin{aligned}
&\frac{[2(t_1\,t_2)-2\cos\Omega]\,(t_1\,t_2)^2-\left[1+(t_1\,t_2)^2-2(t_1\,t_2)\cos\Omega\right]2(t_1\,t_2)}{(t_1\,t_2)^4}=0\\
&(t_1\,t_2)^2-(t_1\,t_2)\cos\Omega-1-(t_1\,t_2)^2+2(t_1\,t_2)\cos\Omega=(t_1\,t_2)\cos\Omega-1=0
\end{aligned}$$

that is when

$$t_1\,t_2 = \frac{1}{\cos\Omega} = \frac{4+\omega_p^2T^2}{4-\omega_p^2T^2} \tag{3.27}$$

With this choice of the product $(t_1\, t_2)$ the minimum value of G_v becomes

$$\begin{aligned}
\min(\widetilde{G}_v)_{RIS} &= \cos\Omega(\cos\Omega+\mathrm{j}\,\sin\Omega)-1=\sin^2\Omega+\mathrm{j}\,\sin\Omega\cos\Omega\\
&= \sin\Omega(\sin\Omega\mathrm{j}\,\cos\Omega) = \frac{4\,\omega_p\,T}{4+\omega_p^2T^2}\,\mathrm{e}^{\mathrm{j}\Omega}
\end{aligned}$$

whose module is obviously $\dfrac{4\,\omega_p\,T}{4+\omega_p^2T^2}$

From all this analysis it turns out that the optimal value of the product $t_1\,t_2$ is of the order of unity. The most obvious choice still seems to be to put $t_1\,t_2 = 1$. However, given that all the values of $t_1\,t_2$ close to 1 give a linear dependence of the filter parameters with the sampling frequency and that moreover the values of the parameters themselves do not change significantly from their optimal values, the choice more opportune to adopt is to minimize the number of filter coefficients in order to minimize the number of processor memory accesses. This is achieved, for example, by making $\tilde{a}_1 = \tilde{a}_4$ i.e. for

$-a - t_1\, t_2 = t_1\, t_2$ i.e. forr $t_1\, t_2 = -a/2 = g$. With this choice and bearing in mind that $\mathrm{Det}(\widetilde{\mathbf{A}}) = b = g^2 + h^2$ the matrix $\mathbf{A}$ becomes

$$\widetilde{\mathbf{A}} = \begin{pmatrix} g & -h^2/t_1^2 \\ t_1^2 & g \end{pmatrix}$$

The value of t_1^2 does not affect the parameter values so we can make $\tilde{a}_2$ and $\tilde{a}_3$ take opposite values equal to h. In this way, only 2 coefficients are needed for the $\widetilde{\mathbf{A}}$ matrix. Finally, we can multiply the matrix $\mathbf{T}$ per t_1 without, as we have seen, altering the matrix $\widetilde{\mathbf{A}}$. In this way, the input matrix takes the form $\widetilde{\mathbf{b}} = (1 \quad 0)^T$. Note that this choice satisfies all the constraints that we had established regarding the ratios of the state matrix coefficients. The state variable filter thus obtained is the one we developed in the previous chapter and is the one currently implemented in Virgo.

Appendix A

Digital Signal Processing

The content of this appendix is a brief introduction to Digital Signal Processing. It is based mostly on the first chapters of [1] and is essentially a brief summary of it. For further details, see the original work.

A.1 Numerical Sequences

In numerical signal processing, we are dealing with signals that are defined only for discrete values of t and are therefore represented as sequences of numbers. Discrete-time signals can be generated by sampling continuous-time signals or directly generated by some discrete-time processes (synthesizers, etc.). In the case of signals generated by the sampling of continuous-time signals, the sampling time is normally a multiple of a fundamental time T, the reciprocal of which is called the sampling frequency.

In analogy to the treatment of continuous time signals carried out by analog systems, we can define numerical signal processing systems in which both the input and the output are represented by numerical sequences. If the numbers of the numerical sequences take only discrete values, then we speak of digital processing systems.

The theory of discrete-time systems deals with the processing of signals which are represented by numerical sequences. The expression

$$x = \{x(n)\} \quad -\infty < n < +\infty$$

represents a numerical sequence x, in which the n-th number of the sequence is denoted by $x(n)$.

Just as in the analog world there are significant signals such as the Dirac's Delta Function, the unit step etc., also in the digital world there are significant numerical sequences of fundamental importance.

- The sequence *unit-impulse* also called simply impulse

$$\delta(n) = \begin{cases} 0 & n \neq 0 \\ 1 & n = 0 \end{cases}$$

- The sequence *unit-step*

$$u(n) = \begin{cases} 0 & n < 0 \\ 1 & n \geqslant 0 \end{cases}$$

We have the following obvious relationships

$$u(n) = \sum_{k=-\infty}^{+\infty} \delta(k)$$

$$\delta(n) = u(n) - u(n-1)$$

A sequence $x(n)$ is said to be periodic with period N if

$$x(n) = x(n+N) \quad \forall n$$

and furthermore it is said that y is a shifted version of the sequence x if

$$y(n) = \sum k = -\infty^{+\infty} x(n-n_0)$$

Various operations can be defined on the sequences and in particular

$$\begin{array}{lll} x \cdot y &= \{x(n)y(n)\} & \text{Product} \\ x+y &= \{x(n)+y(n)\} & \text{Sum} \\ \alpha \cdot x &= \{\alpha x(n)\} & \text{Product with a constant} \end{array}$$

An arbitrary sequence can be represented as the sum of delayed and suitably scaled unit samples

$$x(n) = \sum_{k=-\infty}^{+\infty} x(n)\delta(n-k) \tag{A.1}$$

A.2 Linear Time-Invariant Systems

A system is defined by an operator $T[\cdot]$ which maps an input sequence $x(n)$ into an output sequence $y(n)$

$$y(n) = T[x(n)]$$

.

A system is linear if the superposition principle holds: if $y_1(n)$ and $y_2(n)$ are the system's responses to inputs $x_1(n)$ and $x_2(n)$ then we have:

$$T[a\, x_1(n) + b\, x_2(n)] = a\, T[x_1(n)] + b\, T[x_2(n)] = a\, y_1(n) + b\, y_2(n)$$

with a and b arbitrary constants.

Representing a generic sequence $x(n)$ as the sum of scaled and delayed unit pulses, see (A.1), assuming the linearity of the system and denoting with $h_k(n)$ the response of the system to the impulse $\delta(n-k)$, we have:

$$y(n) = \sum_{k=-\infty}^{+\infty} x(k)T[\delta(n-k)] = \sum_{k=-\infty}^{+\infty} x(k)h_k(n). \tag{A.2}$$

Therefore the system is completely characterized by the response $h_k(n)$ to the impulse $\delta(n-k)$. If only linearity is imposed on the system then $h_k(n)$ will generally depend on both n and k.

The class of time-invariant (shift-invariant) systems is characterized by the property that if $y(n)$ is the system's response to the input $x(n)$ then $y(n-k)$ is the system's response to the input $x(n-k)$. When the index n is associated with time, it is called invariance by time translation. The property of the time-invariance implies that if $h(n)$ is the answer to $\delta(n)$ then the answer to $\delta(n-k)$ is simply $h(n-k)$. Therefore we have:

$$y(n) = \sum_{k=-\infty}^{+\infty} x(k)h(n-k). \tag{A.3}$$

Each linear and time-invariant system is completely characterized by its response to the unit impulse $h(n)$.

The expression (A.3) is commonly called convolution sum and $y(n)$ is said to be the convolution of $x(n)$ with $h(n)$. It is indicated by:

$$y(n) = x(n)\ starh(n)$$

. With a change in the sum index we have

$$y(n) = \sum_{k=-\infty}^{+\infty} h(k)x(n-k)$$

so that the convolution sum has the commutative property. The order in which two sequences are convolved is not important: the system response is the same if we exchange the role of the input and the impulse response.

Two linear and time-invariant systems in cascade correspond to a linear and time-invariant system whose response to the unit impulse is the convolution sum of the two responses. Since the order in which two sequences are convolved is not important, the answer does not depend on the order in which the two systems are cascaded

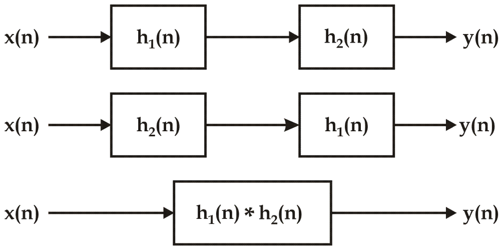

Figure A.1:

Since the convolution sum is a linear operation, it follows that a linear and time-invariant system consisting of the parallel of 2 systems is equivalent to a single system whose response to the unit impulse is the sum of the two individual responses

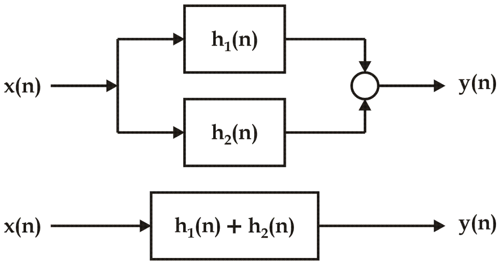

Figure A.2:

A.3 Stability e Causality

A stable system is a system for which each limited input sequence outputs a sequence that is also limited. The linear and time-invariant systems are stable if and only if:

$$S = \sum_{k=-\infty}^{+\infty} |h(k)| < \infty$$

In fact, if the above relationship holds and i x is a limited sequence, that is if $|x(n)| > M$ for every n, then:

$$|y(n)| = \left| \sum_{k=-\infty}^{+\infty} h(k)x(n-k) \right| \leqslant M \sum_{k=-\infty}^{+\infty} |h(k)| < \infty$$

that is, the output sequence y is limited. On the other hand, if S is not limited then a limited input sequence can be found which produces an unlimited output sequence. One such sequence is for example the following:

$$h(n) = \begin{cases} \dfrac{h^*(-n)}{|h(n)|} & h(n) \neq 0 \\ 0 & h(n) = 0 \end{cases}$$

$x(n)$ is obviously limited while the output value for $n = 0$ is:

$$y(0) = \sum_{k=-\infty}^{+\infty} x(-k)h(k)) = \sum_{k=-\infty}^{+\infty} \frac{|h(k)|^2}{h(k)} = S = \infty$$

A system is said to be causal if the value of the output sequence for $n = n_0$ depends only on the values of the input sequence for $n \leqslant n_0$. Therefore:

$$\begin{array}{ll} \text{se} \quad\quad x_1(n) = x_2(n) & \text{per } n < n_0 \\ \text{allora: } y_1(n) = y_2(n) & \text{per } n < n_0 \end{array}$$

A linear and time-invariant system is causal if and only if the response to the unit impulse is null for $n < 0$.

So, for example, if the response to the unit impulse is given by the sequence $h(n) = a^n u(n)$, since the answer is null for $n < 0$, the system is causal. With regards to stability, we have:

$$S = \sum_{k=-\infty}^{+\infty} |h(k)| = \sum_{k=0}^{+\infty} |a|^k$$

If $|a| < 1$ the geometric series converges to the value $S = 1/(1 - |a|)$ and the system is stable, while for $|a| \geqslant 1$ the series diverges and the system is unstable.

A.4 Difference Equations

An important class of linear and time-invariant systems is that consisting of systems for which the input $x(n)$ and the output $y(n)$ satisfy a linear, constant coefficients difference equation of the form:

$$\sum_{k=0}^{N} a_k y(n-k) = \sum_{r=0}^{M} b_r x(n-r) \tag{A.4}$$

In general, such a system is not necessarily causal. For example the difference equation $y(n) - ay(n-1) = x(n)$ for $x(n) = \delta(n)$ is satisfied both by the sequence $y(n) = a^n u(n)$ than from the sequence $y(n) = -a^n u(-n-1)$. The first solution corresponds to a causal system that is stable for $|a| < 1$ and the second solution to a stable non-causal system only if $|a| > 1$.Normally it is assumed that a difference equation represents a causal system.

As with ordinary differential equations, a difference equation has a family of solutions: a solution of the associated homogeneous equation can be added to each particular solution of the equation. The solution becomes unique if the so-called initial conditions are specified. If the system is causal, we must specify the initial conditions so that if $x(n) = 0$ for $n < n_0$ then $y(n) = 0$ for $n < n_0$. If we assume that the system is causal, we can explicitly write the relationship that links the output to the input:

$$y(n) = -\sum_{k=1}^{N} \frac{a_k}{a_0} y(n-k) + \sum_{r=0}^{M} \frac{b_r}{a_0} x(n-r)$$

A.5 FIR and IIR Systems

Generally in a linear and time-invariant system, the response sequence to the unit impulse can be of finite or infinite duration. If the response to the unit impulse is of finite duration, we speak of **FIR** (*finite impulse response*); vice versa, if the response to the unit impulse is of infinite duration we speak of textbf IIR systems (*infinite impulse response*).

If in the difference equation (A.4) that describes the system it is $N = 0$ so that:

$$y(n) = \frac{1}{a_0}\left[\sum_{r=0}^{M} b_r x(n-r)\right]$$

then the system is a FIR system: in fact the expression is identical to the sum of convolution with:

$$h(n) = \begin{cases} \dfrac{b_n}{a_0} & n = 0, 1, \cdots, M \\ 0 & \text{altrimenti} \end{cases}$$

A FIR type system can always be described by a difference equation with $N = 0$; in a type IIR system, on the other hand, you must necessarily have $N > 0$.

A.6 The numerical signals in the frequency domain

A fundamental property of linear and time-invariant systems is that their stationary response to a sinusoidal input signal is a sinusoidal output signal of the same input frequency and with amplitude and phase determined by the system itself. It is this property that makes the representation of signals by sinusoids or complex exponentials (Fourier representation) so useful in the theory of linear systems.

So for discrete-time systems we have:

$$\begin{aligned} x(n) &= e^{\mathrm{j}n\Omega} \qquad\qquad\qquad \text{per } -\infty < n < +\infty \\ y(n) &= \sum_{k=-\infty}^{+\infty} h(k) e^{\mathrm{j}(n-k)\Omega} = e^{\mathrm{j}n\Omega} \sum_{k=-\infty}^{+\infty} h(k) e^{-jk\Omega} \end{aligned}$$

If we define

$$H(e^{\mathrm{j}\Omega}) = \sum_{k=-\infty}^{+\infty} h(k) e^{-jk\Omega} \tag{A.5}$$

we can write:

$$y(n) = H(e^{\mathrm{j}\Omega}) e^{\mathrm{j}n\Omega}$$

Therefore $H(e^{\mathrm{j}\Omega})$ describes the change in the complex amplitude of a complex exponential with angular frequency Ω and is called the system frequency response at the unit impulse $h(n)$. In terms of amplitude and phase we have:

$$H(e^{\mathrm{j}\Omega}) = |H(e^{\mathrm{j}\Omega})| e^{-\mathrm{j}\arg\left[H(e^{\mathrm{j}\Omega})\right]}$$

Since a sinusoidal signal can be expressed as the sum of two complex exponentials, the frequency response to a sinusoidal signal is:

$$\begin{aligned} x(n) &= A\cos(n\Omega_0 + \varphi) = \frac{A}{2} e^{\mathrm{j}\varphi} e^{\mathrm{j}n\Omega_0} + \frac{A}{2} e^{-\mathrm{j}\varphi} e^{-\mathrm{j}n\Omega_0} \\ y(n) &= \frac{A}{2}\left[H(e^{\mathrm{j}\Omega_0} e^{\mathrm{j}\varphi} e^{\mathrm{j}n\Omega_0} + H(e^{-\mathrm{j}\Omega_0} e^{-\mathrm{j}\varphi} e^{-\mathrm{j}n\Omega_0}\right] = \\ &= A|H(e^{\mathrm{j}\Omega_0})| \cos(n\Omega_0 + \varphi + \theta) \end{aligned}$$

where

$$\theta = \arg\left[H(e^{\mathrm{j}\Omega_0})\right]$$

is the system phase response for the angular frequency Ω_0.

From the definition of $H(e^{\mathrm{j}\Omega})$ we note that it is a continuous function of Ω and also that it is periodic function of period 2π. This means that the frequency response of a discrete-time system to a sinusoidal signal with angular frequency Ω is strictly the same response to a signal with angular frequency $\Omega_0 + 2\pi$.

Since $H(e^{\mathrm{j}\Omega})$ is a periodic function of Ω, it can certainly be represented by a Fourier series. Indeed its definition:

$$H(e^{\mathrm{j}\Omega}) = \sum_{k=-\infty}^{+\infty} h(k)e^{\text{-}jk\Omega}$$

represents it in the form of a Fourier series whose coefficients are the sequence $h(n)$, i.e. the response to the unit impulse. We can therefore obtain the sequence $h(n)$ from the frequency response using the relationship that gives us the Fourier coefficients of a periodic function:

$$h(n) = \frac{1}{2\pi}\int_{-\pi}^{+\pi} H(e^{\mathrm{j}\Omega})e^{\mathrm{j}n\Omega}\mathrm{d}\Omega$$

This obviously can be generalized to an arbitrary sequence: thus given a sequence $x(n)$ we define its Fourier transform and its anti-transform through the relations:

$$X(e^{\mathrm{j}\Omega}) = \sum_{n=-\infty}^{+\infty} x(n)e^{\text{-}jn\Omega} \tag{A.6}$$

$$x(n) = \frac{1}{2\pi}\int_{-\pi}^{+\pi} X(e^{\mathrm{j}\Omega})e^{\mathrm{j}n\Omega}\mathrm{d}\Omega \tag{A.7}$$

all this obviously if the series converges.

It can easily be shown that if the sequence $y(n)$ is the system response to the sequence $x(n)$, i.e.

$$y(n) = \sum_{k=-\infty}^{+\infty} x(k)h(n-k)$$

then

$$X(e^{\mathrm{j}\Omega}) = H(e^{\mathrm{j}\Omega})X(e^{\mathrm{j}\Omega})$$

A.6.1 Some properties of the Fourier transform

Here is a summary of the most important properties of the Fourier transform for numerical sequences.

- Definizioni:
  - $x_e(n)$ è **conjugate-symmetric** if $x_e(n) = x_e^*(-n)$
  - $x_o(n)$ è **coniugate-antisymmetric** if $x_o(n) = -x_e^*(-n)$
- One can always write $x(n) = x_e(n) + x_o(n)$ with
  - $x_e(n) = \frac{1}{2}[x(n) + x^*(-n)]$
  - $x_o(n) = \frac{1}{2}[x(n) - x^*(-n)]$
- One can always write $X(e^{\mathrm{j}\Omega}) = X_e(e^{\mathrm{j}\Omega}) + X_o(e^{\mathrm{j}\Omega})$ with
  - $X_e(e^{\mathrm{j}\Omega}) = \frac{1}{2}[X(e^{\mathrm{j}\Omega}) + X^*(e^{-\mathrm{j}\Omega})]$
  - $X_o(e^{\mathrm{j}\Omega}) = \frac{1}{2}[X(e^{\mathrm{j}\Omega}) - X^*(e^{-\mathrm{j}\Omega})]$

- In general:

$$\begin{cases} x(n) & \rightarrow X(e^{\mathrm{j}\Omega}) \\ x^*(n) & \rightarrow X^*(e^{-\mathrm{j}\Omega}) \\ x^*(-n) & \rightarrow X^*(e^{\mathrm{j}\Omega}) \\ \mathrm{Re}[x(n)] & \rightarrow X_e(e^{\mathrm{j}\Omega}) \qquad \text{coniugate-symmetric part of} \quad X(e^{\mathrm{j}\Omega}) \\ \mathrm{j}\ \mathrm{Im}[x(n)] & \rightarrow X_o(e^{\mathrm{j}\Omega}) \qquad \text{coniugate-antisymmetric part of} \quad X(e^{\mathrm{j}\Omega}) \\ x_e(n) & \rightarrow \mathrm{Re}[X(e^{\mathrm{j}\Omega})] \\ x_o(n) & \rightarrow \mathrm{j}\ \mathrm{Im}[X(e^{\mathrm{j}\Omega})] \end{cases}$$

- If $x(n)$ is real:

$$\begin{cases} X(e^{\mathrm{j}\Omega}) & = X^*(e^{-\mathrm{j}\Omega}) \qquad \text{The Fourier transform is coniugate-symmetric} \\ \mathrm{Re}[X(e^{\mathrm{j}\Omega})] & = \mathrm{Re}[X(e^{-\mathrm{j}\Omega})] \qquad \text{the real part is even} \\ \mathrm{Im}[X(e^{\mathrm{j}\Omega})] & = -\mathrm{Im}[X(e^{-\mathrm{j}\Omega})] \qquad \text{the imaginary part is odd} \\ |X(e^{\mathrm{j}\Omega})| & = |X^*(e^{-\mathrm{j}\Omega})| \qquad \text{the amplitude is even} \\ \arg[X(e^{\mathrm{j}\Omega})] & = -\arg[X(e^{-\mathrm{j}\Omega})] \qquad \text{the phase is odd} \end{cases}$$

Appendix B

The Sampling Theorem

B.1 Sampling of continuous time signals

Discrete-time signals are often obtained by sampling continuous-time signals. We intend to derive the relationship between the spectra $X_a(\mathrm{j}\omega)$ of the analog signal and $X(e^{\mathrm{j}\Omega})$ of the sampled signal. If $x_a(t)$ is an analog signal whose Fourier representation is given by:

$$\frac{1}{2\pi}\int_{-\infty}^{+\infty} X_a(\mathrm{j}\omega)e^{\mathrm{j}\omega t}\mathrm{d}\omega$$

with

$$X_a(\mathrm{j}\omega) = \int_{-\infty}^{+\infty} x_a(t)e^{-\mathrm{j}\omega t}\mathrm{d}t$$

indicating with $x(n) = x_a(nT_s)$ the sequence obtained by sampling with period T_s of the continuous signal we have:

$$\begin{aligned} x(n) = x_a(nT_s) &= \frac{1}{2\pi}\int_{-\infty}^{+\infty} X_a(\mathrm{j}\omega)e^{\mathrm{j}\omega nT_s}\mathrm{d}\omega \\ &= \frac{1}{2\pi}\sum_{r=-\infty}^{+\infty}\int_{(2r-1)\pi/T_s}^{(2r+1)\pi/T_s} X_a(\mathrm{j}\omega)e^{\mathrm{j}\omega nT_s}\mathrm{d}\omega. \end{aligned}$$

With the change of variable $\omega = \Omega/T_s + 2\pi r/T_s$ and changing the order of addition and integration we have:

$$\begin{aligned} x(n) &= \frac{1}{2\pi}\sum_{r=-\infty}^{+\infty}\int_{(2r-1)\pi/T_s}^{(2r+1)\pi/T_s} X_a(\mathrm{j}\omega)e^{\mathrm{j}\omega nT_s}\mathrm{d}\omega \\ &= \frac{1}{2\pi}\int_{(2r-1)\pi/T_s}^{(2r+1)\pi/T_s}\sum_{r=-\infty}^{+\infty} X_a(\mathrm{j}\omega)e^{\mathrm{j}\omega nT_s}\mathrm{d}\omega \\ &= \frac{1}{2\pi}\int_{-\pi}^{+\pi}\sum_{r=-\infty}^{+\infty} X_a(\mathrm{j}\Omega/T_s + \mathrm{j}2\pi r/T_s)e^{\mathrm{j}n\Omega}\mathrm{d}\Omega \end{aligned}$$

to be compared with the definition:

$$x(n) = \frac{1}{2\pi}\int_{-\pi}^{+\pi} X(e^{\mathrm{j}\Omega})e^{\mathrm{j}n\Omega}\mathrm{d}\Omega$$

Ultimately we have:

$$X(e^{\mathrm{j}\Omega}) = \frac{1}{T_s}\sum_{r=-\infty}^{+\infty} X_a(\mathrm{j}\Omega/T_s + \mathrm{j}2\pi r/T_s) \tag{B.1}$$

i.e. in terms of the analog frequency ω:

$$X(\mathrm{j}\omega) = \frac{1}{T_s} \sum_{r=-\infty}^{+\infty} X_a(\mathrm{j}\omega + \mathrm{j}2\pi r/T_s) \tag{B.2}$$

This result constitutes the so-called Sampling Theorem. It determines the Fourier transform of a numerical sequence obtained by sampling an analog signal as a function of the Fourier transform of the analog signal itself. As can be seen, the spectrum of the sampled signal is given by the superposition of the analog spectrum and its shifted copies of multiples of the sampling frequency.

If the sampling frequency is too low, the shifted versions of the analog spectrum overlap and information is lost. In this case the high frequency components of $X_a(\mathrm{j}\omega)$ are reflected in the low frequencies of $X(e^{\mathrm{j}\Omega})$ and this effect, for which a high component frequency is converted to a low frequency, it is called *aliasing*.

Vice versa, if the sampling frequency is at least double the maximum frequency present in the analog signal, there is no such overlap of the spectra and therefore there is no loss of information. The least frequency for which this occurs is known as Nyquist frequency. In reality, this statement is a bit too restrictive: it is easy to see that in order not to have information loss it is sufficient that the sampling frequency is at least equal to twice the analog band for which the signal spectrum is significantly different from zero. In fact, even in this case the shifted copies of the analog spectrum do not overlap and from the spectrum of the sampled signal it is always possible to go back to the starting analog spectrum. The reconstruction of the analog signal starting from the sampled one, provided that the Nyquist criterion has been satisfied, is a procedure known as interpolation and is the topic of the next paragraph.

Before continuing we will give a further demonstration of the sampling theorem which, although not very rigorous from a mathematical point of view, has the advantage of being more direct and more intuitive.

We define sampler a linear system whose output, for an input signal $x(t)$, is an analog signal $\tilde{x}(t)$ consisting of a sequence of Dirac deltas, spaced from each other by the sampling period T_s, and whose amplitude is equal to the value that the input signal assumes at the time nT_s. We can formally write the sampler output as follows:

$$\tilde{x}(t) = \sum_{l=-\infty}^{+\infty} x(lT_s)\delta(t - lT_s). \tag{B.3}$$

It is obvious that such a linear system **is not time-invariant**. In fact, if the input signal is a shifted version of $x(t)$ the sampler output is

$$\sum_{l=-\infty}^{+\infty} x(lT_s - \tau)\delta(t - lnT_s) \neq \tilde{x}(t - \tau)$$

especially since $\tilde{x}(t-\tau) \equiv 0$ unless it is $\tau = kT_s$ with k integer. The fact that the sampler is not invariant for time translations also implies that in this case the commutative property does not apply. For example, a system consisting of a sampler followed by a forming filter produces a completely different output from a filter followed by a sampler.

Keeping in mind the properties of the Dirac delta, we can rewrite the sum in eq. (B.3) as follows

$$\tilde{x}(t) = \sum_{l=-\infty}^{+\infty} x(lT_s)\delta(t - lT_s) = \sum_{l=-\infty}^{+\infty} x(t)\delta(t - lT_s) = x(t) \sum_{l=-\infty}^{+\infty} \delta(t - lT_s) \tag{B.4}$$

We can therefore characterize the process of sampling a signal as the product of the signal itself with a periodic Dirac delta sequence of period T_s. T_s.

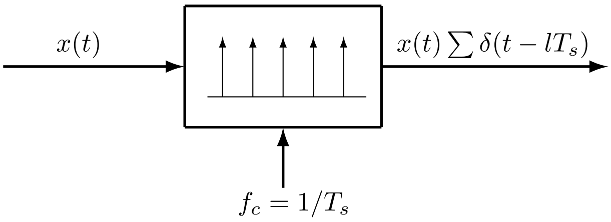

Figure B.1: Campionatore ideale

Since the Dirac delta sequence is periodic, we can represent it with a Fourier series

$$\sum_{l=-\infty}^{+\infty}\delta(t-lT_s)=\sum_{r=-\infty}^{+\infty}C_r e^{\mathrm{j}r\omega_s t}$$

where

$$\omega_s = 2\pi f_s = \frac{2\pi}{T_s}$$

and the development coefficients C_r are given by

$$\begin{aligned}C_r &= \frac{1}{T_s}\int_{-T_s/2}^{+T_s/2}\left[\sum_{l=-\infty}^{+\infty}\delta(t-lT_s)\right]e^{\mathrm{j}r\omega_s t}\mathrm{d}t\\ &= \frac{1}{T_s}\sum_{l=-\infty}^{+\infty}\int_{-T_s/2}^{+T_s/2}\delta(t-lT_s)e^{\mathrm{j}r\omega_s t}\mathrm{d}t = \frac{1}{T_s}\end{aligned}$$

Therefore

$$\sum_{l=-\infty}^{+\infty}\delta(t-lT_s)=\frac{1}{T_s}\sum_{r=-\infty}^{+\infty}e^{\mathrm{j}r\omega_s t} \tag{B.5}$$

Using the (B.5) in the (B.5) we have

$$\tilde{x}(t)=x(t)\frac{1}{T_s}\sum_{r=-\infty}^{+\infty}e^{\mathrm{j}r\omega_s t} \tag{B.6}$$

whose Fourier transform is

$$\begin{aligned}X(\mathrm{j}\omega) &= \frac{1}{T_s}\int_{-\infty}^{+\infty}\left[x(t)\sum_{r=-\infty}^{+\infty}e^{\mathrm{j}r\omega_s t}\right]e^{-\mathrm{j}\omega t}\mathrm{d}t\\ &= \frac{1}{T_s}\sum_{r=-\infty}^{+\infty}\int_{-\infty}^{+\infty}x(t)e^{-\mathrm{j}(\omega-r\omega_s)}t\mathrm{d}t\end{aligned} \tag{B.7}$$

The integral following the sum symbol is the Fourier transform of the input signal evaluated at the angular frequency $\omega - r\omega_s$

$$\int_{-\infty}^{+\infty}x(t)e^{-\mathrm{j}(\omega-r\omega_s)}t\mathrm{d}t = X_a[\mathrm{j}(\omega-r\omega_s)]$$

Therefore

$$X(\mathrm{j}\omega)=\frac{1}{T_s}\sum_{r=-\infty}^{+\infty}X_a[\mathrm{j}(\omega-r\omega_s)] \tag{B.8}$$

that is, the spectrum of the sampled signal is given by the spectrum of the original signal superimposed on all its shifted copies of $r\omega_s$.

A further proof derives from the observation that, as seen from (B.4), the sampled signal is given by the product of the analog signal with a sequence of Dirac deltas. It is known from the theory that the Fourier transform of the product two functions of time is given by the product of convolution in the frequency domain of the transforms of the functions themselves. It is therefore a matter of carrying out the convolution product of the Fourier transform of $x_a(t)$ with the Fourier transform of a series of Dirac deltas. To compute the Fourier transform of a series of $\delta(t-lT_s)$ we use its Fourier series representation given by eq. (B.5). Let $\boldsymbol{\Delta}(\mathrm{j}\omega)$ denote the Fourier transform of the Dirac delta series

$$\begin{aligned}\boldsymbol{\Delta}(\mathrm{j}\omega) &= \int_{-\infty}^{+\infty}\left[\frac{1}{T_s}\sum_{r=-\infty}^{+\infty} e^{\mathrm{j}r\omega_s t}\right] e^{-\mathrm{j}\omega t}\mathrm{d}t = \frac{1}{T_s}\sum_{r=-\infty}^{+\infty}\int_{-\infty}^{+\infty} e^{\mathrm{j}(r\omega_s-\omega)t}\mathrm{d}t \\ &= \frac{2\pi}{T_s}\sum_{r=-\infty}^{+\infty}\delta(\omega-r\omega_s) = \omega_s\sum_{r=-\infty}^{+\infty}\delta(\omega-r\omega_s)\end{aligned} \tag{B.9}$$

where we used the well-known relationship $\delta(x) = \frac{1}{2\pi}\int e^{\mathrm{j}xt}\mathrm{d}t$ and the fact that the Dirac delta is an even function.

Therefore the spectrum of a series of Dirac deltas of period T is also a series of Dirac deltas of period ω_s multiplied by ω_s.

We now evaluate the spectrum of the sampled signal with the convolution product of $X_a(\mathrm{j}\omega)$ with $\mathrm{j}\boldsymbol{\Delta}(\omega)$

$$\begin{aligned}X(\mathrm{j}\omega) &= \frac{\omega_s}{2\pi}\int_{-\infty}^{+\infty} X_a(\mathrm{j}\omega')\sum_{r=-\infty}^{+\infty}\delta(r\omega_s-\omega-\omega')\mathrm{d}\omega' \\ &= \frac{1}{T_s}\sum_{r=-\infty}^{+\infty} X_a[\mathrm{j}(\omega-r\omega_s)]\end{aligned}$$

result that coincides with eq. (B.8)

From the proof of the sampling theorem we obtain a further important important result: **the spectrum of a sampled signal (or a numerical sequence) is periodic of period $\boldsymbol{\omega_s}$.**

B.2 Reconstruction of a band limited signal

If during the sampling of an analog signal the Nyquist criterion is satisfied, having no loss of information, the inverse procedure is possible, i.e. the recovery of the analog signal starting from its samples. Such a procedure is called interpolation. A possible and widely used implementation of the interpolation is the following: we start from the result obtained from the sampling theorem and represent the spectrum of the sampled signal as superposition of the spectrum of the continuous signal with its translated copies of integer multiples of ω_s (see eq. (B.2))

$$X(\mathrm{j}\omega) = \frac{1}{T_s}\sum_{r=-\infty}^{+\infty} X_a(\mathrm{j}\omega + \mathrm{j}2\pi r/T_s)$$

The signal $x(t)$ is obviously equal to the Fourier anti-transform of its spectrum. Compliance with Nyquist's criterion requires that $X_a(\mathrm{j}\omega)$ be different from zero for $-\pi/T < \omega < \pi/T$. We can therefore write

$$x(t) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} X_a(\mathrm{j}\omega)e^{\mathrm{j}\omega t}\mathrm{d}\omega = \frac{1}{2\pi}\int_{-\pi/T}^{+\pi/T} X_a(\mathrm{j}\omega)e^{\mathrm{j}\omega t}\mathrm{d}\omega$$

But, again due to the Nyquist criterion, within the range $-\pi/T < \omega < \pi/T$ the spectrum of the analog signal $X_a(\mathrm{j}\omega)$ coincides, apart the factor T, with that of the sampled signal $X(\mathrm{j}\omega)$ so

$$x(t) = \frac{1}{2\pi}\int_{-\pi/T}^{+\pi/T} T\,X(\mathrm{j}\omega)e^{\mathrm{j}\omega t}\mathrm{d}\omega$$

We also know that the numerical sequence $x(kT)$, that is the samples of the analog signal, are the coefficients of the Fourier series with which the spectrum of the sequence itself is represented (see eq (A.6)). Substituting therefore in the previous expression $X(\mathrm{j}\omega)$ with its Fourier series we obtain

$$x(t) = \frac{T}{2\pi}\int_{-\pi/T}^{+\pi/T}\left[\sum_{k=-\infty}^{+\infty} x(kT)e^{-\mathrm{j}k\omega t}\right]e^{\mathrm{j}\omega t}\mathrm{d}\omega = \sum_{k=-\infty}^{+\infty} x(kT)\left[\frac{T}{2\pi}\int_{-\pi/T}^{+\pi/T} e^{\mathrm{j}\omega(t-kT)}\mathrm{d}\omega\right]$$

The evaluation of the integral in square brackets is elementary

$$\int_{-\pi/T}^{+\pi/T} e^{\mathrm{j}\omega(t-kT)}\mathrm{d}\omega = \frac{1}{\mathrm{j}(t-kT)}e^{\mathrm{j}\omega(t-kT)}\Bigg|_{-\pi/T}^{+\pi/T} = \frac{2\pi}{T}\frac{\sin[(\pi/T)(t-kT)]}{(\pi/T)(t-kT)}$$

for which we ultimately have

$$x(t) = \sum_{k=-\infty}^{+\infty} x(kT)\frac{\sin[(\pi/T)(t-kT)]}{(\pi/T)(t-kT)} \tag{B.10}$$

To conclude an observation: the relationship (B.10) that we have obtained allows the reconstruction of the analog signal starting from its samples. However, it has a peculiarity: to give the value $x(t)$ at time t requires knowledge not only of the samples prior to time t but also of future values. It is therefore a non-causal process and cannot be used in real-time applications but proves very useful for off-line processing.

Appendix C

The z-Transform

One of the characteristics of the Laplace transform, which facilitates the study and design of time-invariant and continuous-time linear systems, consists in the fact that it transforms the differential equation that describes the system itself into an algebraic equation that is much simpler. to be treated.

The z-transform is the equivalent procedure applicable to time-invariant and discrete-time linear systems, transforming the difference equation that describes the system into an algebraic equation whose study is much simpler.

C.1 Definition

Given a numerical sequence $x(n)$ its z-transform $X(z)$ is a function of the complex variable z given by

$$X(z) = \sum_{n=-\infty}^{+\infty} x(n)z^{-n} \tag{C.1}$$

The transform thus defined is also called the bilateral z-transform; sometimes it is useful to also consider the so-called unilateral z-transform

$$X(z) = \sum_{n=0}^{+\infty} x(n)z^{-n} \tag{C.2}$$

Obviously the bilateral and unilateral z-transform coincide for causal numerical sequences for which $x(n) = 0$ for $n < 0$.

Expressing the complex variable z in polar form, $z = \rho\, e^{\mathrm{j}\Omega}$, we have:

$$X(r\, e^{\mathrm{j}\Omega}) = \sum_{n=-\infty}^{+\infty} x(n)(\rho\, e^{\mathrm{j}\Omega})^{-n} = \sum_{n=-\infty}^{+\infty} x(n)\rho^{-n}e^{\mathrm{j}\Omega n} \tag{C.3}$$

Therefore we can interpret the z-transform of the sequence $x(n)$ as the Fourier transform of $x(n)$ multiplied by an exponential sequence. For $r = 1$, that is for $|z| = 1$, the z-transform coincides with the Fourier transform of the sequence. There is a perfect analogy with the Laplace transform: its value along the imaginary axis, that is, for $s = \mathrm{j}\omega$, coincides with the Fourier transform.

C.2 Region of Convergence

In general, the z-transform does not necessarily converge for all numerical sequences or for all the values of the complex variable z. For each given sequence, the set of values of

z for which the z-transform converges, that is:

$$\left|\sum_{n=-\infty}^{+\infty} x(n)z^{-n}\right| < \infty$$

it is called the convergence region.

It may happen that the z-transform converges even if the corresponding Fourier transform does not converge and vice versa. For example. the sequence $x(n) = u(n)$ is not absolutely convergent and therefore its Fourier transform does not converge; however its z-transform is absolutely convergent for $|z| > 1$.

Generally, the convergence region of the bilateral transform is an annular region of the z plane

$$R_- < |z| < R_+$$

where it can be $R_- = 0$ and $R_+ = \infty$.

The power series that defines the bilateral z-transform is a Laurent series. A Laurent series represents an analytic function within the convergence region and therefore, within the convergence region, the z-transform and all its derivatives must be continuous functions of z.

An important class of z-transforms are those for which $X(z)$ is a rational function, that is, it is the ratio of two polynomials in z. The roots of the numerator are the zeros of $X(z)$ and the roots of the denominator are its poles.

For example the sequence $x(n) = a^n u(n)$ whose transform Z is

$$X(z) = \sum_{n=-\infty}^{+\infty} a^n u(n) z^{-n} = \sum_{n=0}^{+\infty} (a\,z^{-1})^n = \frac{1}{1 - a\,z^{-1}} = \frac{z}{z-a} \quad \text{per} \quad |z| < |a|, \tag{C.4}$$

it has a zero for $z = 0$ and a pole for $z = a$. The convergence region is the region outside the circle of radius a.

In general:

- The convergence region of a sequence of finite length is the entire z plane with at most the exclusion of the point $z = 0$ and the point $z = \infty$.

- The convergence region of a right one-sided sequence is external to a circle of radius R_- and can include the point $z = \infty$ (causal sequence)

- The convergence region of a left one-sided sequence is inside a circle of radius R_+ and can include the point $z = 0$

- The convergence region of a bilateral sequence, seen as the sum of a left unilateral sequence converging for $|z| < R_+$ and a right unilateral sequence converging for $|z| > R_-$, is the common convergence region given by $R_- < |z| < R_+$ if $R_- < R_+$; if instead $R_- > R_+$ this region does not exist and the series does not converge.

For example in the case of a bilateral sequence whose z-transform is

$$X(z) = \sum_{n=n_1}^{+\infty} x(n)z^{-n},$$

if we assume that the series is absolutely convergent for $z = z_1$, that is

$$\sum_{n=n_1}^{+\infty} |x(n)z_1^{-n}| < \infty,$$

then the series obviously converges also for every $|z| > |z_1|$. If $n_1 \geqslant 0$ (causal sequence) the series also converges for $z = \infty$.

We also observe that if a right one-sided series converges for $z = z_1$, then each term of the series is limited and therefore there is a finite constant A such that

$$|x(n)z_1^{-n}| < A \quad \text{per } n \geqslant n_1,$$

placing $|z_1| = \rho$ with $\rho > R_-$ we have

$$|x(n| < A\rho^n$$

that is, the sequence, for $n \to +\infty$, cannot grow faster than an exponential. If the convergence region of $x(n)$ includes the unit radius circumference, so that $\rho < 1$,can be chosen, then $|x(n)|$ must tend to zero at least exponentially.

Similar considerations can be made for left unilateral sequences, i.e. the sequence, for $n \to -\infty$, cannot grow faster than an exponential and, if the convergence region includes the unit circle, $x(n)$ tends to zero for $n \to -\infty$.

The z-transform of the left unilateral sequence $x(n) = -b^n u(-n-1)$ is:

$$X(z) = \sum_{n=-\infty}^{-1} -b^n z^{-n} = \sum_{n=1}^{\infty} -b^{-n} z^n = 1 - \sum_{n=0}^{\infty} b^{-n} z^n$$

which is convergent if $|z| < |b|$ nel in which case

$$X(z) = 1 - \frac{1}{1 - b^{-1}z} = \frac{z}{z - b}$$

A comparison with the analogous right one-sided sequence $x(n) = a^n u(n)$ shows that to uniquely define the z-transform, in addition to the function $X(z)$, the convergence region must also be specified.

C.3 Inversion

The inverse of the z-transform can be easily evaluated using Cauchy's integral theorem which states:

$$\frac{1}{2\pi j} \oint_C z^{k-1} \mathrm{d}z = \begin{cases} 1, & \text{se } k = 0 \\ 0, & \text{altrimenti.} \end{cases}$$

where C is a closed path traveled in a counterclockwise direction that includes the origin.

By applying this theorem to the z-transform of a sequence $x(n)$ multiplied by z^{k-1} and integrating along a path that includes the origin and which lies entirely within the convergence region we have:

$$\begin{aligned} \frac{1}{2\pi j} \oint_C X(z) z^{k-1} \mathrm{d}z &= \frac{1}{2\pi j} \oint_C \sum_{n=-\infty}^{+\infty} x(n) z^{-n+k-1} \mathrm{d}z \\ &= \sum_{n=-\infty}^{+\infty} x(n) \frac{1}{2\pi j} \oint_C z^{-n+k-1} \mathrm{d}z = x(k) \end{aligned}$$

The request that the integration path lies entirely within the convergence region allows us to exchange the integration order with the addition order.

Definitely:

$$x(n) = \frac{1}{2\pi j} \oint_C X(z) z^{n-1} \mathrm{d}z$$

The above expression is normally useful for $n \geqslant 0$; for $n < 0$ we can use the following

$$x(n) = \frac{1}{2\pi j} \oint_C X(p^{-1}) p^{-n-1} \mathrm{d}p \qquad z = p^{-1}$$

If the z-transform is a rational function, the evaluation of its anti-transform can be considerably simplified by using the residue theorem:

$$x(n) = \frac{1}{2\pi j}\oint_C X(z)z^{n-1}\mathrm{d}z = \sum[\text{sum of residues of } X(z)z^{n-1} \text{ in the poles inside } C]$$

If $X(z)$ has a pole of order k for $z = z_0$, then we can write

$$X(z)z^{n-1} = \frac{\Psi(z)}{(z-z_0)^k} \quad \text{where } \Psi(z) \text{ is analytic in } z = z_0$$

and

$$\text{Res}\left[X(z)z^{n-1} \text{ in } z = z_0\right] = \frac{1}{(k-1)!}\left[\frac{\mathrm{d}^{k-1}\Psi(z)}{\mathrm{d}z^{k-1}}\right]_{z=z_0}$$

In particular if the pole is of the first order then

$$\text{Res}\left[X(z)z^{n-1} \text{ per } z = z_0\right] = \Psi(z_0)$$

Another widely used method for evaluating the z-anti-transform of a rational function is based on the partial fraction expansion of the function itself. This method is especially applied when dealing with causal sequences that are equal to zero for $n < 0$.

It is known that a rational function $F(z) = N(z)/Q(z)$, where $N(z)$ e $Q(z)$ are polynomials of the complex variable z, we can decompose it into partial fraction. The procedure to be followed is based on the fact that if $F(z)$ has at the point $z = p_i$ a pole of orde l, the function $G(z) = (z-p_i)^l F(z)$ is regular at point $z = p_i$ and can therefore be developed in Taylor series. Its serial development is

$$G(z) = A_{il} + A_{i(l-1)}(z-p_i) + A_{i(l-2)}(z-p_i)^2 + \cdots + Aa_{i(l-k)}(z-p_i)^k + \cdots$$

where

$$A_{ij} = \frac{1}{(l-j)!}\left[\frac{\mathrm{d}^{l-j}(z-p_i)^l}{\mathrm{d}z^{l-j}}\right]_{z=p_i} \tag{C.5}$$

We can therefore write

$$F(z) = \frac{G(z)}{(z-p_i)^l} = \sum_{j=1}^{l}\frac{A_{ij}}{(z-p_i)^j} + H(z) \tag{C.6}$$

where $H(z)$ is in turn a rational function that contains the remaining poles of $F(z)$. The part under the summation symbol in the (C.6) is known in complex variable function theory as the main part of Laurent's series development relative to the pole p_i. Continuing with the same method we can finally write

$$F(z) = \sum_{i=1}^{k}\sum_{j=1}^{l}\frac{A_{ij}}{(z-p_i)^j} + P(z) \tag{C.7}$$

where k is the number of distinct poles of $F(z)$, the coefficients A_{ij} are obtained by the (C.5), and $P(z)$ is a polynomial z. Normally, however, since the degree $N(z)$ is less than or equal to that of $Q(z)$, $P(z)$ is reduced to a possibly zero constant.

Since the z-transform is a linear operation, the z-anti-transform of a rational function is given by the sum of the z-anti-transforms of the elements of its decomposition into simple fractions. Let's first examine the case of $j = 1$. From the relationship (C.4) we know that the z-anti-transform of $z/(z-a)$ the sequence $x(n) = a^n u(n)$. For the z-anti-transform of $A_{i1}/(z-p_i)$ we can proceed in this way

$$\frac{A_{i1}}{z-p_i} = \frac{A_{i1}}{p_i}\left(\frac{z}{z-p_i} - 1\right)$$

whose anti-transform is

$$x(n) = \frac{A_{i1}}{p_i}\,(p_i^n u(n) - \delta(n)) = A_{i1} p_i^{n-1} u(n-1)$$

that is, the exponential sequence $A_{i1} p_i^n u(n)$ shifted to the right by a sample and not divergent $|p_i| < 1$.

As for the generic term $A_{ij}/(z - p_i)^j$ we observe that it is, apart from a numerical multiplicative coefficient, the derivative of order $j - 1$ of $1/(z - p_i)$ and therefore it is easy to see that its anti-transform is given by the usual delayed exponential sequence of j samples multiplied by a certain polynomial of the index n.

In short, the anti-transform of the main part relative to the pole p_i is given by a suitable polynomial of n of degree $l - 1$ multiplied by an exponential sequence of the type p_i^k. It does not diverge if $|p_i| < 1$. We have thus shown that a numerical system is stable if all its poles lie within the unit circle of the z plane. It is the analogue of the stablity condition of continuous time systems which are stable if all its poles lie in the left half plane of the plane of the Laplace variable s.

Keeping in mind what has already been said about the eq.C.3 on page 49, we can see how the z-transform plays exactly the same role in the numeric field as the Laplace transform in a continuous field. The left half plane of the s plane corresponds to the inside of the unit circle of the z plane; the circumference of the unit circle corresponds to the imaginary axis $textj\omega$ of the plane s; the spectrum of a continuous signal is obtained by evaluating the Laplace transform along the imaginary axis while the spectrum of a numeric or sampled signal is obtained by evaluating its z-transform along the unit circumference. In this regard, we observe that in the numerical case we can travel the unit circumference several times and this implies the periodicity of the spectrum of the numerical signals.

There are other methods used for the evaluation of the z-anti-transform such as the Laurent series development around the origin, the long division between polynomials etc. For their description and use, refer to the reference [1] of the bibliography.

C.4 Proprieties

Here we list the main properties of the z-transform. The properties described here obviously apply within the convergence region of the individual z-transforms.

- **Linearity**

  It derives directly from the definition of the z-transform. Given two numerical sequences $x_1(n)$ and $x_2(n)$ the z-transform of their linear combination $a\,x_1(n) + b\,x_2(n)$ is

$$\begin{aligned}\mathcal{Z}\{a\,x_1(n) + b\,x_2(n)\} &= \sum_{n=-\infty}^{+\infty} [a\,x_1(n) + b\,x_2(n)] z^{-n} \\ &= a \sum_{n=-\infty}^{+\infty} x_1(n) z^{-n} + b \sum_{n=-\infty}^{+\infty} x_2(n) z^{-n} = aX_1(z) + bX_2(z)\end{aligned}$$

  i.e. the linear combination of the respective Z transforms.

- **Time delay**

  Given a numerical sequence $x(n)$ whose z-transform is $X(z)$, the z-transform of a delayed version of k samples is given by $z^{-k}X(z)$. Indeed

$$\mathcal{Z}\{x(n-k)\} = \sum_{n=-\infty}^{+\infty} x(n-k) z^{-n} = \sum_{n=-\infty}^{+\infty} x(m) z^{-m-k}$$

$$z^{-k} \sum_{n=-\infty}^{+\infty} x(n) z^{-n} = z^{-k} X(z)$$

where the change of variable $m = n - k$ was made followed by the substitution $n = m$

- **Time advance**

  Given a numerical sequence $x(n)$ whose z-transform is $X(z)$, the z-transform of an early version of k samples is given by $z^k X(z)$. Obviously the proof is the same as that given for the time delay with the only replacement of $-k$ with k. Things differ in the case of causal sequences: a delayed causal numerical sequence remains a causal sequence, not so for a time-advanced causal sequence. To obtain the z-transform of a time-advanced causal sequence it is necessary to delete the part of the series with positive powers of z. Starting from the causal sequence $x(n)$ and indicating the bilateral transform with $\mathcal{Z}$ and $\mathcal{Z}_+$ the one-sided right one respectively we have

$$\begin{aligned}\mathcal{Z}\{x(n+k)\} = z^k X(z) = \sum_{n=-\infty}^{+\infty} x(n) z^{k-n} = \sum_{n=0}^{+\infty} x(n) z^{k-n} \\ = \sum_{n=0}^{k-1} x(n) z^{k-n} + \sum_{n=k}^{+\infty} x(n) z^{k-n} = \sum_{n=0}^{k-1} x(n) z^{k-n} + \mathcal{Z}_+\{x(n+k)\}\end{aligned}$$

  so the right unilateral z-transform of $x(n+k)$ is

$$\mathcal{Z}_+\{x(n+k)\} = z^k X(z) - \sum_{n=0}^{k-1} x(n) z^{k-n}$$

  This property will be used later to find the solution of difference-equations of order N with constant coefficients when initial conditions are specified.

- **Multiplication by an Exponential Sequence**

  Given a numerical sequence $x(n)$ whose z-transform is $X(z)$, the z-transform of the sequence $a^n x(n)$ is given by $X(z/a)$. The proof of this property is trivial. It can be used, for example, to evaluate the z-transform of a damped sinusoidal sequence.

- **Differentiation of $X(z)$**

  Given a numerical sequence $x(n)$ whose z-transform is $X(z)$, the z-transform of the $nx(n)$ sequence is given by $-z\mathrm{d}X(z)/textdz$. Indeed

$$-z\frac{\mathrm{d}X(z)}{\mathrm{d}z} = -z \sum_{n=-\infty}^{+\infty} (-n) x(n) z^{-n-1} = \sum_{n=-\infty}^{+\infty} n x(n) z^{-n} = \mathcal{Z}\{nx(n)\}$$

- **Convolution product**

  Let $x(n)$ be the numerical input sequence of a linear, time-invariant system with $h(n)$ as its unit impulse response, the output sequence $y(n)$ is given by (see eq. (A.3) on page 36)

$$y(n) = \sum_{k=-\infty}^{+\infty} x(k) h(n-k).$$

  then the z-transform of the output sequence $y(n)$ is given by $Y(z) = H(z)X(z)$. Indeed

$$\begin{aligned}H(z)X(z) = \sum_{k=-\infty}^{+\infty} x(k) z^{-k} \sum_{m=-\infty}^{+\infty} h(m) z^{-m} = \sum_{k=-\infty}^{+\infty} \sum_{m=-\infty}^{+\infty} x(k) h(m) z^{-(k+m)} \\ = \sum_{n=-\infty}^{+\infty} \sum_{k=-\infty}^{+\infty} x(k) h(n-k) z^{-n} = \sum_{n=-\infty}^{+\infty} y(n) z^{-n} = Y(z)\end{aligned}$$

Therefore the z-transform of the convolution sum of two numerical sequences is the product of the respective z-transforms. It is exactly the same property of the Laplace transform: the Laplace transform of the convolution product of two functions of time is given by the product of the respective Laplace transforms. Once again we notice the perfect parallel between the Laplace transform in the continuous time domain and the z-transform in the discrete time domain.

- **Initial value theorem**

  If $x(n)$ is a causal sequence, i.e. $x(n) = 0$ for $n < 0$ evidently

  $$x(0) = \lim_{z \to \infty} X(z)$$

- **Final value theorem**

  If $x(n)$ is a causal sequence, i.e. $x(n) = 0$ for $n < 0$ we have

  $$\lim_{n \to \infty} x(n) = \lim_{z \to 1}(1 - z^{-1})X(z)$$

C.5 z-Transform of notable causal signals

We give here a list of z-transforms of remarkable causal sequences. They can be easily obtained from the definition of the z-transform using one or more properties listed in the previous paragraph.

$f(nT)$	$F(z)$
$\delta(nT)$	1
$u(nT)$	$\frac{1}{1 - z^{-1}}$
nT	$\frac{Tz^{-1}}{(1 - z^{-1})^2}$
n^2T^2	$\frac{T^2z^{-1}(1 + z^{-1})}{(1 - z^{-1})^3}$
n^3T^3	$\frac{T^3z^{-1}(1 + 4z^{-1} + z^{-2})}{(1 - z^{-1})^4}$
e^{-anT}	$\frac{1}{1 - e^{-aT}z^{-1}}$
nTe^{-anT}	$\frac{Te^{-aT}z^{-1}}{(1 - e^{-aT}z^{-1})^2}$
$n^2T^2e^{-anT}$	$\frac{T^2e^{-aT}z^{-1}(1 + e^{-aT}z^{-1})}{(1 - e^{-aT}z^{-1})^3}$
$cos(anT)$	$\frac{1 - z^{-1}\cos(aT)}{1 - 2z^{-1}\cos(aT) + z^{-2}}$
$\sin(aT)$	$\frac{z^{-1}\sin(aT)}{1 - 2z^{-1}\sin(aT) + z^{-2}}$
$e^{-bnT}cos(anT)$	$\frac{1 - e^{-bT}z^{-1}\cos(aT)}{1 - 2e^{-bT}z^{-1}\cos(aT) + e^{-2bT}z^{-2}}$
$e^{-bnT}\sin(aT)$	$\frac{e^{-bT}z^{-1}\sin(aT)}{1 - 2e^{-bT}z^{-1}\cos(aT) + e^{-2bT}z^{-2}}$

C.6 Applications and Examples

As the Laplace transform finds application in the analysis and in the evaluation of the response of linear time-invariant systems that are governed by linear differential equations with constant coefficients, so the z-transform, which is the analogue of the Laplace transform for numerical systems, allows the study and solution of linear and time-invariant systems, governed by linear difference equations. It allows to transform a system of difference equations into an equivalent system of algebraic equations, much simpler to handle.

Give a system that is described by a linear and constant coefficient difference equation with input $x(n)$ and output $y(n)$

$$\sum_{k=0}^{N} a_k y(n-k) = \sum_{r=0}^{M} b_r x(n-r) \tag{C.8}$$

the output $y(n)$ for $n > N$ can be obtained, starting from the sequence $x(n)$, by the following recursive formula

$$y(n) = \frac{1}{a_0}\left[\sum_{r=0}^{M} b_r x(n-r) - \sum_{k=1}^{N} a_k y(n-k)\right]$$

once we have assigned the N initial values $y(0), y(1), \cdots, y(N-1)$.

However, if we are interested in the frequency response of the system described by eEq. (C.8) or obtain an expression of $y(n)$ in closed form, we must resort to the z-transform. By applying the z-transform to the first and second members of the eq. (C.8) we have

$$\mathcal{Z}\left\{\sum_{k=0}^{N} a_k y(n-k)\right\} = \mathcal{Z}\left\{\sum_{r=0}^{M} b_r x(n-r)\right\}$$
$$\sum_{k=0}^{N} a_k \mathcal{Z}\left\{y(n-k)\right\} = \sum_{r=0}^{M} b_r \mathcal{Z}\left\{x(n-r)\right\}$$

where we made use of the linearity of the z-transform. Using the time delay property, for which we have $\mathcal{Z}\{x(n-r)\} = z^{-r}X(z)$, and an analogous relation for the variable y we can write

$$\sum_{k=0}^{N} a_k z^{-k} Y(z) = \sum_{r=0}^{M} b_r z^{-r} X(z)$$

that resolved gives

$$Y(z) = \frac{\displaystyle\sum_{r=0}^{M} b_r z^{-r}}{\displaystyle\sum_{k=0}^{N} a_k z^{-k}} X(z)$$

The ratio $H(z) = Y(z)/X(z)$ is the transfer function of the system and is obviously the z-transform of the response to the impulse sequence $\delta(n)$. Its value evaluated along the unit circumference gives us the frequency response of our system

$$H(e^{\mathrm{j}\Omega}) = \frac{\displaystyle\sum_{r=0}^{M} b_r e^{-\mathrm{j}r\Omega}}{\displaystyle\sum_{k=0}^{N} a_k e^{-\mathrm{j}k\Omega}}$$

where the angular frequency Ω, in the case of a sampled system with period T, is equal to $\Omega = \omega T$, with ω analog angular frequency.

Let us now give an example of a solution of the sequence $y(n)$ in closed form. We aim to obtain a closed formula that provides the n-th term of the Fibonacci sequence. The Fibonacci series is defined recursively by the following equation

$$F_{n+2} = F_{n+1} + F_n \quad \text{con } n > 1 \text{ e condizioni iniziali } F_0 = 0, F_1 = 1$$

By indicating with $\mathcal{F}(z)$ the z-transform of the sequence and making use of the time advance property we can write

$$\begin{aligned}
\mathcal{F}\{F_{n+2}\} &= z^2\mathcal{F}(z) - F_0 z^2 - F_1 z^1 = z^2\mathcal{F}(z) - z \\
\mathcal{F}\{F_{n+1}\} &= z^1\mathcal{F}(z) - F_0 z^1 = z\mathcal{F}(z) \\
\mathcal{F}\{F_n\} \quad &= \mathcal{F}(z)
\end{aligned}$$

so that

$$\begin{aligned}
&\mathcal{F}\{F_{n+2}\} = \mathcal{F}\{F_{n+1}\} + \mathcal{F}\{F_n\} \\
&z^2\mathcal{F}(z) - z = z\mathcal{F}(z) + \mathcal{F}(z) \\
&\mathcal{F}(z) = \frac{z}{z^2 - z - 1}
\end{aligned}$$

The poles of $\mathcal{F}(z)$ are the roots of the equation $z^2 - z - 1$ that is

$$z_+ = (1 + \sqrt{5})/2 \quad \text{e} \quad z_- = (1 - \sqrt{5})/2$$

che sono poli semplici. which are simple poles. Let's decompose $\mathcal{F}(z)$ into simple fractions

$$\mathcal{F}(z) = \frac{z}{z^2 - z - 1} = \frac{z}{z_+ - z_-}\left(\frac{1}{z - z_+} - \frac{1}{z - z_-}\right) = \frac{1}{\sqrt{5}}\left(\frac{z}{z - z_+} - \frac{z}{z - z_-}\right)$$

and anti-transforming is finally achieved

$$F_n = \frac{1}{\sqrt{5}}\left(\frac{1 + \sqrt{5}}{2}\right)^n - \frac{1}{\sqrt{5}}\left(\frac{1 - \sqrt{5}}{2}\right)^n$$

We observe that the pole at the point z_- lies inside the unit circle and produces the sequence that tends to zero exponentially while the pole z_+ is outside the unit circle and generates the divergent sequence. Furthermore, the module of $\left(\frac{1+\sqrt{5}}{2}\right)^n$ is less than 0.5 since from $n = 2$, and, bearing in mind also the $1/\sqrt{5}$ factor, we conclude that the n-th term of the Fibonacci series is given by the integer closest to $\frac{1}{\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^n$.

Appendix D

The bilinear transformation

Suppose we have the following first order differential equation:

$$a_1 y'(t) + a_0 y(t) = b_0 x(t) \tag{D.1}$$

where, obviously, $y'(t)$ is the time derivative of $y(t)$.

Using a Laplace transform and assuming as initial condition $y(0^+) = 0$ we get:

$$a_1 sY(s) + a_0 Y(S) = b_0 X(s)$$

from which we obtain the transfer function:

$$H_a(s) = \frac{Y(s)}{X(s)} = \frac{b_0}{a_1 s + a_0}. \tag{D.2}$$

Obviously we have:

$$y(t) = \int_{t_0}^{t} y'(\tau) d\tau + y(t_0).$$

Putting $t = nT$ and $t_0 = (n-1)T$ in this last equation ($1/T$ is the sampling frequency) we get:

$$y(nT) = \int_{(n-1)T}^{nT} y'(\tau) d\tau + y((n-1)T)$$

Approximating the integral by the trapezoidal rule we obtain:

$$y(nT) \simeq y((n-1)T) + \frac{T}{2}[y'(nT) + y'((n-1)T)]. \tag{D.3}$$

On the other hand we have from eq. D.1:

$$y'(nT) = -\frac{a_0}{a_1} y(nT) + \frac{b_0}{a_1} x(nT)$$

putting this last equation into eq. D.3 and writing y_n instead of $y(nT)$ we get:

$$y_n - y_{n-1} \simeq \frac{T}{2}\left[-\frac{a_0}{a_1}(y_n + y_{n-1}) + \frac{b_0}{a_1}(x_n + x_{n-1})\right].$$

Now, using the Z transform and remembering that if $x_n \to X(z)$ then $x_{n-1} \to z^{-1}X(z)$, we obtain:

$$H(z) = \frac{Y(z)}{X(z)} = \frac{b_0}{a_1 \frac{2}{T} \frac{1 - z^{-1}}{1 + z^{-1}} s + a_0}. \tag{D.4}$$

Comparing the analog transfer function expression (eq. D.2) with the discrete one (eq. D.4), we see that we can obtain these last from the first one simply by substituting the s Laplace's variable with the following expression:

$$s = \frac{2}{T}\frac{1-z^{-1}}{1+z^{-1}}$$

Obviously we can obtain an analog transfer function from a discrete one using the inverse transformation:

$$z = \frac{1+T/2\ s}{1-T/2\ s}.$$

The transformation $s \rightarrow \frac{2}{T}\frac{1-z^{-1}}{1+z^{-1}}$ is called bilinear transformation. We obtain it in the case of a first order linear differential equation. Nevertheless its validity is more general, because an Nth-order linear differential equation can be written as a set of N first order linear differential equations.

The bilinear transformation $s = \frac{2}{T}\frac{1-z^{-1}}{1+z^{-1}}$ and its inverse $z = \frac{1+T/2\ s}{1-T/2\ s}$ are conformal transformations from the complex plane s to the complex plane z and vice versa. They are a special case of flat linear conformal transformations, known in the theory of complex functions also as homographic transformations.

D.1 Conformal homographic transformations

The generic conformal homographic transformation is of the form:

$$w = \frac{az+b}{cz+d}, \qquad ad-bc \neq 0. \tag{D.5}$$

The condition $ad - bc \neq 0$ is necessary because otherwise the function (D.5) is reduced to a constant. It is shown that conformal homographic transformations enjoy the following remarkable properties:

1) they are the only conformal transformations that at each point of the complex z plane, including the point at infinity, correspond one and only one point of the w plane;

2) transforms each circumference of the z plane into a circumference of the w plane;

3) transforms each pair of points symmetrical with respect to a circumference C into a pair of points symmetrical with respect to the image of C.

The straight lines are particular circumferences that pass through the point at infinity.

Proof of the **1)** property is based on the following considerations:

- It cannot have essential singularities because a function that has an essential singularity assumes in any neighborhood of this singularity any complex value, excluding at most only one, an infinite number of times (Picard's theorem): there would no longer be biunivocal correspondence between the z plane and the w plane.

- It must have only one pole (possibly at infinity) because by definition of homography the infinity point of the w plane must correspond to one and only one point of the z plane. The only pole must also be a first order pole because in the vicinity of a pole of multiplicity greater than one the function is no longer injective.

- If the pole P is not at infinity then the main part of $w(z)$ near the pole P has the form $\frac{B}{z-P}$; subtracting the principal part from $w(z)$ we will get the function $\varphi(z) = w(z) - \frac{B}{z-P}$ which is free of singularities in the whole plane and therefore $\varphi(z)$ is a constant function. Consequently it must be $w(z) = Az + B$, or a homographic transformation.

- If the pole P is at infinity then the principal part of $w(z)$ is of the form Az and proceeding as in the previous point we conclude that in this case it must be $w(z) = Az + B$, a particular form of homographic transformation.

The **2)** property is easily proven. If $c = 0$ the (D.5) is reduced to a linear transformation, that is to a rotation of the plane around the origin with dilation followed by a translation, and therefore transforms straight lines into straight lines and circumferences into circumferences.

So let's assume $c \neq 0$. The (D.5) transformation can be written in the following way

$$w = \frac{az+b}{cz+d} = a/c + \frac{(bc-ad)/c^2}{z+d/c} \tag{D.6}$$

It can therefore be broken down into the sequence of three elementary transformations: (a) a translation $w_1 = z + d/c$, (b) a transformation of the form $w_2 = k/w_1$ with $k = (bc - ad)/c^2$ and finally (c) of a further linear transformation $w = w_2 + a/c$. Clearly both the transformation (a), which refers to a simple translation, and the transformation (c), which is a rotation of the plane around the origin with dilation followed by a translation, transform the circumferences into circumferences. It is therefore sufficient to consider only the elementary transformation (b) of the form

$$w = \frac{k}{z} \tag{D.7}$$

and prove that it too transforms circumferences into circumferences. The equation of a generic circumference l in the Cartesian plane is

$$A(x^2 + y^2) + 2Bx + 2Cy + D = 0, \tag{D.8}$$

where $A = 0$ n the case of a straight line. We can rewrite it in form

$$Az\overline{z} + \Delta z + \overline{\Delta z} + D = 0 \quad \text{dove} \quad \Delta = B - iC. \tag{D.9}$$

The D.7 transformation transforms the circumference l into the curve of equation

$$Dw\overline{w} + \overline{\Delta k}z + \Delta k\overline{z} + Ak\overline{k} = 0 \tag{D.10}$$

which is the equation of a circumference (or a straight line).

Let's now consider the property **3)**. Given a circumference C with center O and of radius R, two points A_1 e A_2 are symmetrical with respect to C if they lie on the same radius, one on radius itself and the other on its extension, and are such that the product of the distance $\overline{OA_1}$ by the distance $\overline{OA_2}$ is equal to the square of the radius R. Let's now take a generic circumference passing through the points A_1 e A_2.

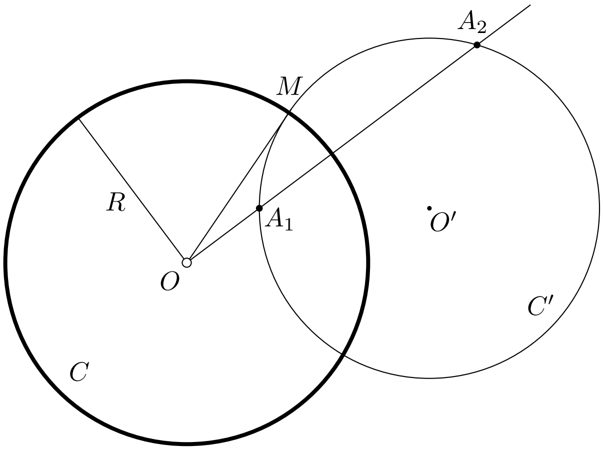

Figure D.1: Symmetrical points

Since, by a well-known geometry theorem, the product of the length of the secant $\overline{OA_2}$ by the length of its external part $\overline{OA_1}$ must be equal to the square of the length of the tangent conducted by the point O at the circumference C' and on the other hand, since the points A_1 and A_2 are symmetrical, this product must be equal to the square of the radius R, we deduce that the radius $\overline{OM}$ is tangent to the circumference C', i.e. the circumference C' is orthogonal to the circumference C. But the circumference C' is a generic circumference passing through the points A_1 and A_2 and therefore each circumference belonging to the bundle of circumferences passing through two symmetrical points with respect to the circumference C is orthogonal to the circumference C itself. Since a homographic transformation $F(\cdot)$, transforms circumferences into circumferences and that one of the most remarkable properties of conformal transformations is the conservation of angles, we can conclude that the bundle of circumferences passing through the points $F(A_1)$ and $F(A_2)$ are orthogonal to the circumference $F(C)$. But this is only possible if the points $F(A_1)$ and $F(A_2)$ are symmetrical with respect to the circumference $F(C)$. This proves the property **3)**.

D.2 Proprieties if the bilinear transformation

The inverse of the bilinear transformation $s = \frac{2}{T}\frac{1-z^{-1}}{1+z^{-1}}$, that is $z = \frac{1+sT/2}{1-sT/2}$, maps the entire s plane into the z plane. This mapping is one-to-one: to a point of the s plane corresponds one and only one point of the z plane and vice versa. The image of the point $s = \sigma + \mathrm{j}\omega$ becomes

$$z = \frac{1+\sigma T/2 + \mathrm{j}\omega T/2}{1-\sigma T/2 + \mathrm{j}\omega T/2} \tag{D.11}$$

From the eq. (D.11) we see that if $\sigma < 0$ then $|z| < 1$ for any value of ω, while for $\sigma > 0$ we have $|z| > 1$ and e for $\sigma = 0$ we have $|z| = 1$. Therefore the bilinear transformation maps points of the left half-plane of s plane, inside the unit circle in the s plane and vice versa. A pole located in the left half plane of the s plane will have as image a pole inside the unit circle in the z plane. **The bilinear transformation transforms stable analog filters into corresponding stable numeric filters and vice versa.** The frequency axis $\mathrm{j}\omega$ of the s plane is placed in one-to-one correspondence with the unit circumference of the z plane.

We now find which value of the "angular frequency" Ω in the z plane corresponds to a particular value of the analog frequency ω. We have

$$|z| = e^{\mathrm{j}\Omega} = \frac{1+\mathrm{j}\omega T/2}{1-\mathrm{j}\omega T/2}$$

from which

$$\Omega = \arg\left(1+\frac{\mathrm{j}\omega T}{2}\right) - \arg\left(1-\frac{\mathrm{j}\omega T}{2}\right) = 2\arctan\left(\frac{\omega T}{2}\right) \tag{D.12}$$

This means that a generic analog function $H_a(s)$ assumes at the point $\mathrm{j}\omega$ the same value as the corresponding $H(z)$, obtained from $H_a(s)$ by means of the bilinear transformation, at the point $e^{\mathrm{j}\Omega}$ with Ω given by eq. (D.12). In other words, the spectrum of di $H(z)$ is obtained from that of $H_a(s)$ by simply "compressing" the frequency axis $-\infty \leqslant \omega \leqslant +\infty$ in the range $-\pi \leqslant \Omega \leqslant +\pi$ using the equation (D.12). There is therefore a distortion of the frequency axis, known as *frequency warping* which is small for $\omega \ll 1/T$.

We could remove this distortion in the following way: we build a new analog function $H'_a(s)$ which assumes at the point $s = 2\mathrm{j}/T \arctan(\omega T/2)$ the same value that $H_a(s)$ assumes at the point $\mathrm{j}\omega$. Let's say that

$$H'_a\left(\mathrm{j}\frac{2}{T}\arctan\frac{\omega T}{2}\right) = H_a(\mathrm{j}\omega)$$

that is

$$H'_a(\mathrm{j}\omega) = H_a\left(\mathrm{j}\frac{2}{T}\tan\frac{\omega T}{2}\right). \tag{D.13}$$

We obviously want that if $H_a(s)$ is a rational function of s, so is $H'_a(s)$. Therefore we cannot simply set $H'_a(s) = H_a(2/T\tanh(sT/2))$ which, even if it satisfies the condition (D.13), is a transcendent function. However, we can obtain a rational function that satisfies the condition (D.13) at least for the critical frequencies of the filter that we want to implement.

This pre-distortion of frequencies is called *frequency pre-warping.*

For example, if we want to build a Butterworth filter of a certain order k and that has a cutoff frequency of $\omega = \omega_0$, we first design an analog filter of order k that has a cutoff frequency of $\omega = 2/T\tan(\omega_0 T/2)$. We then apply the bilinear transformation to the latter filter and starting from th $H(z)$ thus obtained we get a corresponding difference equation that we can solve with a simple recursive algorithm.

We can apply this procedure for the realization of more complex Chebyshev, elliptical etc. filters. Butterworth filters have the characteristic of being most flat in the passband, those of Chebyshev of the first type of being equi-ripple in the pass-band, those of Chebyshev of the second type of being equi-ripple in the stop band and the elliptical ones of being equi-ripple in both pass-band and stop-band. These properties are preserved by applying the procedure described above, which shows a remarkable utility of the use of bilinear transformation.

Summarizing the procedure to be used is the following:

1. Specify the set of critical frequencies $\{\omega_k\}$ for the filter you want to realize.
2. Pre-warping of critical frequencies in $\{\omega'_k = 2/T\tan(\omega_k T/2)\}$
3. Design an analog filter with transfer function $H'_a(s)$ using the critical frequencies $\{\omega'_k\}$ obtained with pre-warping
4. Obtain the corresponding numeric filter $H(z)$ using the bilinear transformation
5. Implement $H(z)$ starting from the corresponding difference equation preferably using low order cascade sections.

$$\begin{tikzcd}
\sum_n \frac{\mathrm{d}^n y(t)}{dt^n} = \sum_k \frac{\mathrm{d}^n x(t)}{dt^n} \arrow[r, "d/dt \rightarrow s"] \arrow[dd, "\frac{\mathrm{d}^k y(t)}{dt^k} \rightarrow y(n-k)"] & H_a(s) \arrow[r, "\omega_k \rightarrow 2/T\tan(\omega_k T/2)", "\text{pre-warp}"'] & H'_a(s) \arrow[dd, "s \rightarrow \frac{2}{T}\frac{1-z^{-1}}{1+z^{-1}}", "\text{bilinear transform}"'] \\
 & & \\
\sum_n \frac{\mathrm{d}^n y(t)}{dt^n} = \sum_k \frac{\mathrm{d}^n x(t)}{dt^n} & & H(z) \arrow[ll, "y(n-k) \leftarrow z^{-k}"']
\end{tikzcd}$$

Figure D.2: Use of the bilinear transformation

Appendix E

Cascade filters with complex coefficients

Consideriamo un filtro passa basso analogico del secondo ordine. Sia Let's consider a second order analog low pass filter. Let

$$H(s) = \frac{{\omega_0}^2}{s^2 + \dfrac{\omega_0 s}{Q} + {\omega_0}^2}$$

be its analog transfer function. With the help of the bilinear transformation we can obtain its z-transform:

$$H(z) = A\frac{(1+z^{-1})^2}{1+az^{-1}+bz^{-2}}$$

where

$$\begin{aligned} A &= \frac{{\omega_0}^2 T^2}{4 + \dfrac{2\omega_0 T}{Q} + {\omega_0}^2 T^2} \\ a &= \frac{2({\omega_0}^2 T^2 - 4)}{4 + \dfrac{2\omega_0 T}{Q} + {\omega_0}^2 T^2} \\ b &= \frac{4 - \dfrac{2\omega_0 T}{Q} + {\omega_0}^2 T^2}{4 + \dfrac{2\omega_0 T}{Q} + {\omega_0}^2 T^2} \end{aligned} \tag{E.1}$$

If we indicate, using polar coordinates, with $\rho\,\mathrm{e}^{\mathrm{j}\alpha}$ and $\rho\,\mathrm{e}^{-\mathrm{j}\alpha}$ the position of the complex conjugated poles in the z plane, we must force:

$$\left(1 - \rho\,\mathrm{e}^{\mathrm{j}\alpha} z^{-1}\right)\left(1 - \rho\,\mathrm{e}^{-\mathrm{j}\alpha} z^{-1}\right) = 1 + az^{-1} + bz^{-2} \tag{E.2}$$

from witch

$$\begin{cases} \rho^2 & = b \\ 2\rho\cos\alpha & = -a \end{cases} \tag{E.3}$$

We can implement the second order filter through the series of two first order sections with complex conjugated coefficients.

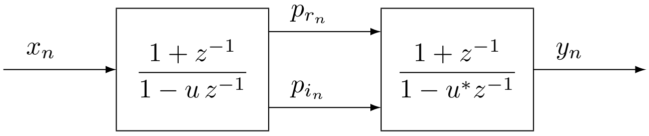
x_n
$\dfrac{1+z^{-1}}{1-u\,z^{-1}}$
p_{r_n}
p_{i_n}
$\dfrac{1+z^{-1}}{1-u^* z^{-1}}$
y_n

Figure E.1: Cascade filters with complex coefficients

where p_{r_n} and p_{i_n} represent the real and the imaginary part of the complex sequence p_n and $u = g + \mathrm{j}h = \rho\,\mathrm{e}^{\mathrm{j}\alpha}$ is the pole of cartesian coordinates (g, h) or polar coordinates (ρ, α). Let's write some realations that will come in handy later. To simplify the writing we put $\Omega = 2\arctan(\omega_0 T/2)$, which is the image of ω_0 on the unit circumference.

$$\begin{cases}
\Omega & = 2\arctan(\omega_0 T/2) \\
\Omega' & = \omega_0 T \\
\mathrm{Den} & = 4 + 2\Omega'/Q + \Omega'^2 \\
a & = \dfrac{2\Omega'^2 - 8}{\mathrm{Den}} \\
b & = \dfrac{4 - 2\Omega'/Q + \Omega'^2}{\mathrm{Den}} \\
g = \rho\cos\alpha & = -\dfrac{a}{2} = \dfrac{4 - \Omega'^2}{\mathrm{Den}} \\
h = \rho\sin\alpha & = \sqrt{b - \dfrac{a^2}{4}} = \dfrac{4\Omega'\sqrt{1 - \dfrac{1}{4Q^2}}}{\mathrm{Den}} \\
1 + b & = \dfrac{2\Omega'^2 + 8}{\mathrm{Den}} \\
1 - b & = \dfrac{4\Omega'/Q}{\mathrm{Den}} \\
1 + b + a & = \dfrac{4\Omega'^2}{\mathrm{Den}} \\
1 + b - a & = \dfrac{16}{\mathrm{Den}} \\
1 + g = 1 - a/2 & = \dfrac{2\Omega'/Q + 8}{\mathrm{Den}} \\
1 - g = 1 + a/2 & = \dfrac{2\Omega'^2 + 2\Omega'/Q}{\mathrm{Den}}
\end{cases} \tag{E.4}$$

Moreover, since it is $\Omega'/2 = \tan(\Omega/2)$ we have

$$\begin{cases}
\sin^2\dfrac{\Omega}{2} = \dfrac{\Omega'^2}{4 + \Omega'^2} \\
\cos^2\dfrac{\Omega}{2} = \dfrac{4}{4 + \Omega'^2}
\end{cases} \tag{E.5}$$

$$\begin{cases}
\sin\Omega = \dfrac{4\Omega'}{4 + \Omega'^2} = 2\,Q\dfrac{1 - b}{1 + b} \\
\cos\Omega = \dfrac{4 - \Omega'^2}{4 + \Omega'^2} = \dfrac{-a}{1 + b}
\end{cases} \tag{E.6}$$

Let's just consider the first section of the filter:

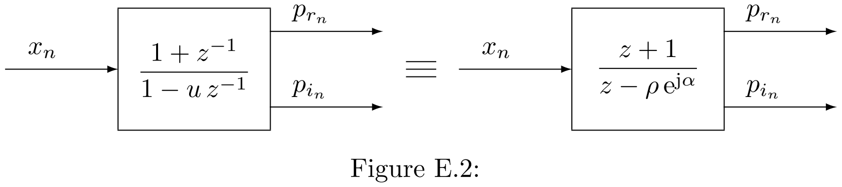
x_n
$\dfrac{1 + z^{-1}}{1 - u\,z^{-1}}$
p_{r_n}
p_{i_n}
$\equiv$
x_n
$\dfrac{z + 1}{z - \rho\,\mathrm{e}^{\mathrm{j}\alpha}}$
p_{r_n}
p_{i_n}

Figure E.2:

We aim to evaluate the response of the filter with constant input and at the resonance frequency. We break down the filter into the cascade of two stages: the first for the realization of the pole and the second for the realization of the zero. That is, we write:

$$\begin{aligned}
&\frac{P}{X} = \frac{1+z^{-1}}{1-\rho\,\mathrm{e}^{\mathrm{j}\alpha}z^{-1}} = \frac{V}{X}\frac{P}{V};\\
&\frac{V}{X} = \frac{1}{1-\rho\,\mathrm{e}^{\mathrm{j}\alpha}z^{-1}}; \quad \frac{P}{V} = \frac{1+z^{-1}}{1}
\end{aligned}$$

For the response with constant input we have to compute the value of the transfer function for $z = 1$, while for the response to the resonance frequency we have to compute the value of the transfer function for $z = \mathrm{e}^{\mathrm{j}\Omega}$ where $\Omega = 2\arctan(\omega_0 T/2)$ (see properties of the bilinear transformation).

E.1 Constant input response

Let's put $z = 1$ and note that in this case it is $V = P/2$, so we will only deal with P. We have: P. Abbiamo:

$$p_{dc} = x_{dc}\frac{2}{1-\rho\,e^{\mathrm{j}\alpha}} = x_{dc}\frac{2(1-\rho\,e^{-\mathrm{j}\alpha})}{(1-\rho\,e^{\mathrm{j}\alpha})(1-\rho\,e^{-\mathrm{j}\alpha})} = x_{dc}\frac{2(1-\rho\,e^{-\mathrm{j}\alpha})}{1+a+b},$$

separating the real and the imaginary part

$$\begin{cases}
p_{dc_r} &= x_{dc}\dfrac{2-2\rho\cos\alpha}{1+a+b} = \dfrac{2+a}{1+a+b}x_{dc}\\
p_{dc_i} &= x_{dc}\dfrac{-2\rho\sin\alpha}{1+a+b} = \dfrac{-\sqrt{4\rho^2-a^2}}{1+a+b}x_{dc}
\end{cases} \tag{E.7}$$

and using some of the (E.4) we finally get:

$$\begin{cases}
p_{dc_r} = x_{dc}\dfrac{{\omega_0}^2T^2 + \dfrac{\omega_0 T}{Q}}{{\omega_0}^2T^2} = \left(\dfrac{1}{Q\omega_0 T}+1\right)x_{dc}\\
p_{dc_i} = x_{dc}\dfrac{-4\omega_0 T\sqrt{4-\dfrac{1}{Q^2}}}{4{\omega_0}^2T^2} = -\dfrac{1}{\omega_0 T}\sqrt{4-\dfrac{1}{Q^2}}x_{dc}.
\end{cases} \tag{E.8}$$

As regards the "gain" for constant input, remember that its value is $G_v = (V_r/x_{dc}-1)$ (see eq. (2.23) on page 20

$$G_v = \frac{1}{2\,Q\omega_0 T} - \frac{1}{2}$$

In the case of the response with constant input we could reach the same result by solving the difference equation of the filter. In fact from the transfer function:

$$\frac{P}{X} = \frac{1+z^{-1}}{(1+(g+\mathrm{j}h)z^{-1})} \tag{E.9}$$

we get the following recursive formulas:

$$\begin{cases}
p_{r_n} = x_n + x(n-1) + g\,p_r(n-1) - h\,p_i(n-1)\\
p_{i_n} = \qquad\qquad\qquad + g\,p_i(n-1) + h\,p_r(n-1)
\end{cases}$$

from which, placing $x_n = x_{dc}$, $p_{r_n} = p_{dc_r}$ and $p_{i_n} = p_{dc_i}$, we have sequentially

$$\begin{cases}
p_{dc_i} = -\dfrac{h}{1-g}p_{dc_r}\\
p_{dc_r}(1-g) = 2x_{dc} + \dfrac{h^2}{1-g}p_{dc_r}
\end{cases}, \tag{E.10}$$

$$\begin{cases} p_{dc_r} = \dfrac{2}{1-g+\dfrac{h^2}{1-g}} x_{dc} = \dfrac{2(1-g)}{1+g^2+h^2-2g} x_{dc} \\ \quad = \dfrac{2(1+a/2)}{1+\rho^2+a} x_{dc} = \dfrac{2+a}{1+a+b} x_{dc} \\ p_{dc_i} = -\dfrac{h}{1-g}\dfrac{2+a}{1+a+b} x_{dc} = -\dfrac{2h}{1+a+b} x_{dc} = -\dfrac{\sqrt{4\rho^2-a^2}}{1+a+b} x_{dc} \end{cases} \tag{E.11}$$

which are exactly the same relations obtained previously (see (E.7)).

E.2 Resonance frequency response

Let's now compute the response of the first section of the filter to the resonance frequency.

$$\begin{cases} \dfrac{V}{X} = \dfrac{1}{1-\rho\,\mathrm{e}^{\mathrm{j}\alpha}z^{-1}} \\ \dfrac{P}{V} = 1+z^{-1} \end{cases}$$

placing $z = \mathrm{e}^{\mathrm{j}\Omega}$

$$\frac{V}{X} = \frac{1}{1-\rho\,\mathrm{e}^{\mathrm{j}\alpha}\mathrm{e}^{-\mathrm{j}\Omega}} = \frac{1-\rho\,\mathrm{e}^{-\mathrm{j}\alpha}\mathrm{e}^{\mathrm{j}\Omega}}{1+\rho^2-2\rho\cos(\Omega-\alpha)}$$

and separating the real and the imaginary part

$$\begin{cases} V_r = \dfrac{1-\rho\cos(\Omega-\alpha)}{1+\rho^2-2\rho\cos(\Omega-\alpha)} X \\ V_i = \dfrac{\rho\sin(\alpha-\Omega)}{1+\rho^2-2\rho\cos(\Omega-\alpha)} X \end{cases} \tag{E.12}$$

The numerator of V_r becomes

$$\begin{aligned} 1-\rho\cos(\Omega-\alpha) &= 1-\rho\cos\alpha\cos\Omega-\rho\sin\alpha\sin\Omega \\ &= 1-\frac{4-\Omega'^2}{\text{Den}}\cdot\frac{4-\Omega'^2}{4+\Omega'^2}-\frac{4\Omega'\sqrt{1-\dfrac{1}{4Q^2}}}{\text{Den}}\cdot\frac{4\Omega'}{4+\Omega'^2} \\ &= \frac{\dfrac{8\,\Omega'}{Q}+\dfrac{2\,\Omega'^3}{Q}+16\,\Omega'^2-16\Omega'^2\sqrt{1-\dfrac{1}{4Q^2}}}{(4+\Omega'^2)\,\text{Den}} \\ &= \frac{\dfrac{2\,\Omega'}{Q}(4+\Omega'^2)+16\,\Omega'^2\left(1-\sqrt{1-\dfrac{1}{4Q^2}}\right)}{(4+\Omega'^2)\,\text{Den}}, \end{aligned}$$

the numerator of V_i

$$\begin{aligned} \rho\sin(\alpha-\Omega) &= \rho\sin\alpha\cos\Omega-\rho\cos\alpha\sin\Omega = \\ &= \frac{4\Omega'\sqrt{1-\dfrac{1}{4Q^2}}}{\text{Den}}\cdot\frac{4-\Omega'^2}{4+\Omega'^2}-\frac{4-\Omega'^2}{\text{Den}}\cdot\frac{4\Omega'}{4+\Omega'^2} \\ &= \frac{-4\Omega'(4-\Omega'^2)\left(1-\sqrt{1-\dfrac{1}{4Q^2}}\right)}{(4+\Omega'^2)\,\text{Den}}, \end{aligned}$$

and finally the denominator

$$\begin{aligned}1+\rho^2-2\rho\cos(\Omega-\alpha) &= 1+b-2\rho\cos\alpha\cos\Omega-2\rho\sin\alpha\sin\Omega\\ &= 1+b+a\cos\Omega-2h\sin\Omega\\ &= \frac{2\Omega'^2+8}{\text{Den}}+\frac{2\Omega'^2-8}{\text{Den}}\cdot\frac{4-\Omega'^2}{4+\Omega'^2}-\frac{8\Omega'\sqrt{1-\dfrac{1}{4Q^2}}}{\text{Den}}\cdot\frac{4\Omega'}{4+\Omega'^2}\\ &= \frac{32\Omega'^2\left(1-\sqrt{1-\dfrac{1}{4Q^2}}\right)}{(4+\Omega'^2)\,\text{Den}},\end{aligned}$$

We have therefore finally

$$\begin{aligned}V_r &= \frac{\dfrac{2\,\Omega'}{Q}(4+\Omega'^2)+16\,\Omega'^2\left(1-\sqrt{1-\dfrac{1}{4Q^2}}\right)}{(4+\Omega'^2)\,\text{Den}}\cdot\frac{(4+\Omega'^2)\,\text{Den}}{32\Omega'^2\left(1-\sqrt{1-\dfrac{1}{4Q^2}}\right)}\\ &= \left(\frac{4+\Omega'^2}{16\Omega'\,Q}\cdot\frac{1}{1-\sqrt{1-\dfrac{1}{4Q^2}}}+\frac{1}{2}\right)X=\left(\frac{(4+\Omega'^2)(2\,Q+\sqrt{4\,Q^2-1})}{8\Omega'}+\frac{1}{2}\right)X\end{aligned}$$

$$\begin{aligned}V_i &= \frac{-4\Omega'(4-\Omega'^2)\left(1-\sqrt{1-\dfrac{1}{4Q^2}}\right)}{(4+\Omega'^2)\,\text{Den}}\cdot\frac{(4+\Omega'^2)\,\text{Den}}{32\Omega'^2\left(1-\sqrt{1-\dfrac{1}{4Q^2}}\right)}\\ &= -\frac{4-\Omega'^2}{8\,\Omega'}X\end{aligned} \tag{E.13}$$

For what concern P, since it is $P=(1+z^{-1})V$, placing $z=\cos\Omega+\mathrm{j}\sin\Omega$ we have

$$\begin{aligned}P_r &= V_r(1+\cos\Omega)+V_i\sin\Omega=\\ &= \left(\frac{(4+\Omega'^2)(2\,Q+\sqrt{4\,Q^2-1})}{8\Omega'}+\frac{1}{2}\right)\frac{8}{4+\Omega'^2}X-\frac{4-\Omega'^2}{8\,\Omega'}\frac{4\,\Omega'}{4+\Omega'^2}X\\ &= \left(\frac{2\,Q+\sqrt{4\,Q^2-1}}{\Omega'}+\frac{1}{2}\right)X=\left(\frac{2\,Q+\sqrt{4\,Q^2-1}}{\omega_0T}+\frac{1}{2}\right)X\end{aligned} \tag{E.14}$$

$$\begin{aligned}P_i &= -V_r\sin\Omega-V_i(1+\cos\Omega)=\\ &= -\left(\frac{(4+\Omega'^2)(2\,Q+\sqrt{4\,Q^2-1})}{8\Omega'}+\frac{1}{2}\right)\frac{4\,\Omega'}{4+\Omega'^2}X-\frac{4-\Omega'^2}{8\,\Omega'}\frac{8}{4+\Omega'^2}X\\ &= -\left(\frac{1}{\Omega'}+\frac{2\,Q+\sqrt{4\,Q^2-1}}{2}\right)X=-\left(\frac{1}{\omega_0T}+\frac{2\,Q+\sqrt{4\,Q^2-1}}{2}\right)X\end{aligned}$$

At the limit for $Q\gg 1$ we have

$$\begin{cases} P_r\simeq\left(\dfrac{4\,Q}{\omega_0T}+\dfrac{1}{2}\right)X\simeq\dfrac{4\,Q}{\omega_0T}X\\ P_i\simeq-\left(\dfrac{1}{\omega_0T}+2Q\right)X\end{cases}$$

while for $Q = 1/2$, which is the limit case of two coincident poles, we have

$$\begin{cases} P_r = +\left(\dfrac{1}{\omega_0 T} + \dfrac{1}{2}\right) X \\ P_i = -\left(\dfrac{1}{\omega_0 T} + \dfrac{1}{2}\right) X \end{cases}$$

Appendix F

State Space Filters

Let's consider a system described by the following *state-space* model:

$$S = z^{-1}\mathbf{A}\,S + \mathbf{B}\,X, \qquad Y = \mathbf{C}\,S + \mathbf{D}\,X$$

where S is the state vector, X is the input vector, Y the output vector, $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, $\mathbf{D}$ real matrices of suitable size.

We want to examine the case of a second order filter with complex conjugated poles. Its transfer function is

$$H(z) = \frac{Y(z)}{X(z)} = \frac{1 + c\,z^{-1} + d\,z^{-2}}{1 + a\,z^{-1} + b\,z^{-2}} = \frac{W(z)}{X(z)}\frac{Y(z)}{W(z)}$$

where

$$\begin{aligned} \frac{W(z)}{X(z)} &= \frac{1}{1 + a\,z^{-1} + b\,z^{-2}} \\ \frac{Y(z)}{W(z)} &= \frac{1 + c\,z^{-1} + d\,z^{-2}}{1} \end{aligned}$$

so we get the following recursive formulas

$$\begin{aligned} &w_n = -a\,w_{n-1} - b\,w_{n-2} + x_n \\ &y_n = w_n + c\,w_{n-1} + d\,w_{n-2} = (c-a)w_{n-1} + (d-b)w_{n-2} + x_n \\ &w_{n-1} = w_n \\ &w_{n-2} = w_{n-1} \end{aligned}$$

By identifying the state $(u\ v)^T$ with the pair $(w_{n-1}\ w_{n-2})^T$ we can write

$$\begin{pmatrix} u \\ v \end{pmatrix} = z^{-1}\begin{pmatrix} -a & -b \\ 1 & 0 \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} x \\ 0 \end{pmatrix}$$

$$y = \begin{pmatrix} c-a & d-b \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix} + \begin{pmatrix} 1 & 0 \end{pmatrix}\begin{pmatrix} x \\ 0 \end{pmatrix}$$

We obviously have:

$$\mathbf{A} = \begin{pmatrix} -a & -b \\ 1 & 0 \end{pmatrix} \quad \mathbf{B} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \mathbf{C} = \begin{pmatrix} c-a & d-b \end{pmatrix} \quad \mathbf{D} = \begin{pmatrix} 1 & 0 \end{pmatrix}$$

As you know, the state-space representation is not unique. Given any non-singular matrix $\mathbf{T}$, a new representation can be obtained by placing:

$$\widetilde{S} = \mathbf{T}^{-1}S \qquad \widetilde{\mathbf{A}} = \mathbf{T}^{-1}\mathbf{A}\,\mathbf{T} \qquad \widetilde{\mathbf{B}} = \mathbf{T}^{-1}\mathbf{B} \qquad \widetilde{\mathbf{C}} = \mathbf{C}\,\mathbf{T} \qquad \widetilde{\mathbf{D}} = \mathbf{D}$$

The filter's poles are given by the solutions of the equation $z^2 + a\,z + b = 0$, that is

$$z_{1,2} = -\frac{a}{2} \pm \mathrm{j}\sqrt{b - \frac{a^2}{4}} = \rho\,\mathrm{e}^{\pm \mathrm{j}\alpha}. \tag{F.1}$$

To simplify the writing we put

$$g = -\frac{a}{2} \quad \text{e} \quad h = \sqrt{b - \frac{a^2}{4}}$$

and build the following transformation matrix and its inverse:

$$\mathbf{T} = \begin{pmatrix} 1 & g/h \\ 0 & 1/h \end{pmatrix} \qquad \mathbf{T^{-1}} = \begin{pmatrix} 1 & -g \\ 0 & h \end{pmatrix}$$

so the new matrices become:

$$\begin{aligned}
\widetilde{\mathbf{A}} &= \begin{pmatrix} 1 & -g \\ 0 & h \end{pmatrix} \begin{pmatrix} -a & -b \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & g/h \\ 0 & 1/h \end{pmatrix} \\
&= \begin{pmatrix} -a-g & -b \\ h & 0 \end{pmatrix} \begin{pmatrix} 1 & g/h \\ 0 & 1/h \end{pmatrix} = \begin{pmatrix} g & -h \\ h & g \end{pmatrix} \\
\widetilde{\mathbf{B}} &= \begin{pmatrix} 1 & -g \\ 0 & h \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & -g \\ 0 & h \end{pmatrix} \\
\widetilde{\mathbf{C}} &= \begin{pmatrix} c-a & d-b \end{pmatrix} \begin{pmatrix} 1 & g/h \\ 0 & 1/h \end{pmatrix} = \begin{pmatrix} c-a & (cg-ag+d-b)/h \end{pmatrix} \\
&= \begin{pmatrix} c-a & \dfrac{a(a-c)+2(d-b)}{\sqrt{4b-a^2}} \end{pmatrix} \\
\widetilde{\mathbf{D}} &= \begin{pmatrix} 1 & 0 \end{pmatrix}
\end{aligned}$$

and the equations of state of the new representation in turn become

$$\begin{aligned}
\begin{pmatrix} u \\ v \end{pmatrix} &= z^{-1} \begin{pmatrix} g & -h \\ h & g \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} + \begin{pmatrix} 1 & -g \\ 0 & h \end{pmatrix} \begin{pmatrix} x \\ 0 \end{pmatrix} \\
y &= \begin{pmatrix} c-a & \dfrac{a(a-c)+2(d-b)}{\sqrt{4b-a^2}} \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} + \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} x \\ 0 \end{pmatrix}
\end{aligned} \tag{F.2}$$

and from these we can write the following recursive formulas for the evolution of the state of the filter

$$\begin{aligned}
u_n &= g\,u_{n-1} - h\,v_{n-1} + x_n \\
v_n &= h\,u_{n-1} + g\,v_{n-1}
\end{aligned} \tag{F.3}$$

Since in this case $\widetilde{\mathbf{B}}\,X = X$ we have for the evolution of the state S the following equation $S = z^{-1}\widetilde{\mathbf{A}}\,S + X$ whose solution is $S = (I - z^{-1}\widetilde{\mathbf{A}})^{-1}X$. The determinant of $(I - z^{-1}\widetilde{\mathbf{A}})$ is $(1-z^{-1}g)^2 + z^{-2}h^2 = 1 + a\,z^{-1} + b\,z^{-2}$. The second component of the input vector is zero so we only need the 11 and 21 components of the matrix $(I - z^{-1}\widetilde{\mathbf{A}})^{-1}$

$$(1 - z^{-1}g)/(1 + a\,z^{-1} + b\,z^{-2})$$

and

$$z^{-1}h/(1 + a\,z^{-1} + b\,z^{-2}).$$

Therefore, the u, v components of the S state are

$$\begin{aligned} u &= \frac{1 - g\,z^{-1}}{1 + a\,z^{-1} + b\,z^{-2}}\,x \\ v &= \frac{h\,z^{-1}}{1 + a\,z^{-1} + b\,z^{-2}}\,x \end{aligned} \tag{F.4}$$

Using the (F.3) and the (F.4) we can copute the ”gains” of the filter

$$\begin{aligned} G_u &= \frac{g\,u - h\,v}{x} = \frac{g - (g^2 + h^2)z^{-1}}{1 + a\,z^{-1} + b\,z^{-2}} = \frac{-a/2 - b\,z^{-1}}{1 + a\,z^{-1} + b\,z^{-2}} \\ G_v &= \frac{h\,u}{g\,v} = \frac{h(1 - g\,z^{-1})}{g\,h\,z^{-1}} = \frac{1 - g\,z^{-1}}{g\,z^{-1}} = \frac{2 + a\,z^{-1}}{-a\,z^{-1}} \end{aligned} \tag{F.5}$$

Now putting $z = 1$ and $z = \mathrm{e}^{\mathrm{j}\Omega}$ where $\Omega = 2\arctan(\omega_0 T/2)$ (see properties of the bilinear transformation), we can compute the value of the state and of the ”gains” with constant input and at the resonance frequency.

F.1 Constant input response

Let’s put in (F.4) and (F.5) $z = 1$. As for the state we have

$$\begin{aligned} \frac{u_{dc}}{x_{dc}} &= \frac{1 - g}{1 + a + b} = \frac{2\left({\omega_0}^2 T^2 + \dfrac{\omega_0 T}{Q}\right)}{4{\omega_0}^2 T^2} = \left(\frac{1}{2\omega_0 T\,Q} + \frac{1}{2}\right) \\ \frac{v_{dc}}{x_{dc}} &= \frac{h}{1 + a + b} = \frac{2\omega_0 T\sqrt{4 - \dfrac{1}{Q^2}}}{4{\omega_0}^2 T^2} = \frac{1}{\omega_0 T}\sqrt{1 - \frac{1}{4\,Q^2}} \end{aligned}$$

where we made use of (E.4) and (F.1). As for the ”gains” we have (F.1). Per quanto riguarda i ”guadagni” abbiamo

$$\begin{aligned} G_{u_{dc}} &= \frac{-a/2 - b}{1 + a + b} = \frac{\dfrac{2\omega_0 T}{Q} - 2{\omega_0}^2 T^2}{4{\omega_0}^2 T^2} = \frac{1}{2\omega_0 T\,Q} - \frac{1}{2} \\ G_{v_{dc}} &= \frac{2 + a}{-a} = \frac{\dfrac{2\omega_0 T}{Q} + 2{\omega_0}^2 T^2}{4 - {\omega_0}^2 T^2} \simeq \frac{\omega_0 T}{2Q} \end{aligned}$$

The results obtained so far coincide with (2.34) and (2.35).

F.2 Resonance frequency response

If we denote by **DEN** the denominator of the (F.4) at the resonant frequency, i.e. for $z = \mathrm{e}^{\mathrm{j}\Omega}$ we have

$$\begin{aligned} \mathbf{DEN} &= (1 + a\,z^{-1} + b\,z^{-2})|_{z=\mathrm{e}^{\mathrm{j}\Omega}} \\ &= 1 + a\cos\Omega + b\cos 2\Omega + \mathrm{j}(a\sin\Omega + b\sin 2\Omega) \end{aligned}$$

The real part is

$$\begin{aligned} 1 + a\cos\Omega + b\cos 2\Omega &= 1 + a\cos\Omega + b\,(2\cos^2\Omega - 1) \\ &= 1 - b + \cos\Omega\,(a + 2b\cos\Omega) = 1 - b - \frac{a}{1 + b}\left(a - \frac{2ab}{1 + b}\right) \\ &= 1 - b - \frac{a}{(1 + b)}\frac{a - ab}{(1 + b)} = (1 - b)\left[1 - \frac{a^2}{(1 + b)^2}\right] = \frac{(1 + b)\sin^2\Omega}{2Q}\sin\Omega \end{aligned}$$

and the imaginary part is

$$\begin{aligned}
a\sin\Omega + b\sin 2\Omega &= a\sin\Omega + 2b\sin\Omega\cos\Omega = \sin\Omega\,(a + 2b\cos\Omega)\\
&= \sin\Omega\left(a - \frac{2ab}{1+b}\right) = \sin\Omega\,\frac{a - ab}{(1+b)} = a\frac{\sin^2\Omega}{2Q}\\
&= -\frac{(1+b)\sin^2\Omega}{2Q}\cos\Omega
\end{aligned}$$

where we made use of the relations $\cos\Omega = \dfrac{-a}{1+b}$ and $\sin\Omega = 2Q\,\dfrac{1-b}{1+b}$ (see eq. (E.6)).

The denominator therefore becomes

$$\begin{aligned}
\mathbf{DEN} &= \frac{(1+b)\sin^2\Omega}{2Q}\sin\Omega - \mathrm{j}\frac{(1+b)\sin^2\Omega}{2Q}\cos\Omega\\
&= \frac{(1+b)\sin^2\Omega}{2Q}(\sin\Omega - \mathrm{j}\cos\Omega) = -\mathrm{j}\frac{(1+b)\sin^2\Omega}{2Q}\,\mathrm{e}^{-\mathrm{j}\Omega} \qquad \text{(F.6)}\\
&= -\mathrm{j}\frac{(8+2\Omega'^2)16\Omega'^2}{2Q\mathrm{Den}(4+\Omega'^2)^2}\,\mathrm{e}^{-\mathrm{j}\Omega} = -\mathrm{j}\frac{16\Omega'^2/Q}{\mathrm{Den}(4+\Omega'^2)}\,\mathrm{e}^{-\mathrm{j}\Omega}
\end{aligned}$$

and its square modulus

$$|\mathbf{DEN}|^2 = \frac{256\,\Omega'^4/Q^2}{\mathrm{Den}^2(4+\Omega'^2)^2}$$

The u component of the state vector at the resonance frequency divided by the input is (eq. (F.4))

$$\begin{aligned}
\left(\frac{u}{x}\right)_{RIS} &= \left.\frac{1 - g\,z^{-1}}{1 + a\,z^{-1} + b\,z^{-2}}\right|_{z=\mathrm{e}^{\mathrm{j}\Omega}} = \mathrm{j}\,(1 - g\,\mathrm{e}^{-\mathrm{j}\Omega})\frac{\mathrm{Den}(4+\Omega'^2)}{16\Omega'^2/Q}\,\mathrm{e}^{\mathrm{j}\Omega}\\
&= \mathrm{j}\,\frac{\left(\cos\Omega + \mathrm{j}\sin\Omega - \dfrac{4-\Omega'^2}{\mathrm{Den}}\right)\mathrm{Den}(4+\Omega'^2)}{16\Omega'^2/Q}\\
&= -\frac{4\Omega'\mathrm{Den}}{16\Omega'^2/Q} + \mathrm{j}\,\frac{(4-\Omega'^2)\mathrm{Den} - (4-\Omega'^2)(4+\Omega'^2)}{16\Omega'^2/Q}\\
&= -\frac{Q\,\mathrm{Den}}{4\Omega'} + \mathrm{j}\,\frac{(4-\Omega'^2)2\Omega'/Q}{16\Omega'^2/Q} = -\frac{4Q + 2\Omega' + \Omega'^2\,Q}{4\Omega'} + \mathrm{j}\,\frac{(4-\Omega'^2)}{8\Omega'}
\end{aligned}$$

and its modulus

$$\begin{aligned}
\left|\frac{u}{x}\right|_{RIS} &= \frac{\sqrt{64Q^2 + 16\Omega'^2 + 4\Omega'^4\,Q^2 + 64\Omega'\,Q + 32\Omega'^2\,Q^2 + 16\Omega'^3\,Q + 16 - 8\Omega'^2 + \Omega'^4}}{8\Omega'}\\
&= \frac{4+\Omega'^2}{8\Omega'}\sqrt{\frac{16\Omega'\,Q}{4+\Omega'^2} + 4Q^2 + 1}\\
&= \frac{(4+\Omega'^2)\sqrt{4Q^2+1}}{8\Omega'}\left[1 + \frac{8\Omega'\,Q}{(4+\Omega'^2)(4Q^2+1)} + O(\Omega'^2)\right]\\
&= \frac{\sqrt{4Q^2+1}}{2\Omega'} + \frac{Q}{\sqrt{4Q^2+1}} + O(\Omega')
\end{aligned}$$

As regards the component v of the state vector at the resonance frequency we have

$$\begin{aligned}
\left(\frac{v}{x}\right)_{RIS} &= \left.\frac{h\,z^{-1}}{1 + a\,z^{-1} + b\,z^{-2}}\right|_{z=\mathrm{e}^{\mathrm{j}\Omega}} = \mathrm{j}\,h\,\mathrm{e}^{-\mathrm{j}\Omega}\,\frac{\mathrm{Den}(4+\Omega'^2)}{16\Omega'^2/Q}\,\mathrm{e}^{\mathrm{j}\Omega}\\
&= \mathrm{j}\,\frac{2\Omega'\sqrt{4Q^2-1}}{Q\,\mathrm{Den}}\,\frac{\mathrm{Den}(4+\Omega'^2)}{16\Omega'^2/Q} = \mathrm{j}\,\frac{(4+\Omega'^2)\sqrt{4Q^2-1}}{8\Omega'}
\end{aligned}$$

and of course its modulus

$$\left|\frac{v}{x}\right|_{RIS} = \frac{(4+\Omega'^2)\sqrt{4Q^2-1}}{8\Omega'}$$

Now let's compute the values of the "gains" at the resonance frequency. For the "gain" G_u we have (see eq. F.5)

$$\begin{aligned}
G_{u_{RIS}} &= \left.\frac{-a/2-bz^{-1}}{1+az^{-1}+bz^{-2}}\right|_{z=\mathrm{e}^{\mathrm{j}\Omega}} = \mathrm{j}(-a/2-b\mathrm{e}^{-\mathrm{j}\Omega})\frac{\mathrm{Den}(4+\Omega'^2)}{16\Omega'^2/Q}\mathrm{e}^{\mathrm{j}\Omega} \\
&= \mathrm{j}\left[(4-\Omega'^2)\cos\Omega+\mathrm{j}(4-\Omega'^2)\sin\Omega-(4+\Omega'^2-2\,\Omega'/Q)\right]\frac{4+\Omega'^2}{16\Omega'^2/Q} \\
&= -\frac{Q(4-\Omega'^2)}{4\,\Omega'}+\mathrm{j}\,\frac{(4-\Omega'^2)^2-(4+\Omega'^2)^2+2(4+\Omega'^2)\,\Omega'/Q}{16\Omega'^2/Q} \\
&= -\frac{Q(4-\Omega'^2)}{4\,\Omega'}+\mathrm{j}\left(\frac{4+\Omega'^2}{8\Omega'}-Q\right)
\end{aligned} \tag{F.7}$$

whose modulus is

$$\begin{aligned}
|G_u|_{RIS} &= \sqrt{\frac{Q^2}{16\,\Omega'^2}\left[(4-\Omega'^2)^2+\frac{(4+\Omega'^2)^2}{4Q^2}+16\,\Omega'^2-\frac{4\,\Omega'(4+\Omega'^2)}{Q}\right]} \\
&= \frac{Q}{4\,\Omega'}\sqrt{(4+\Omega'^2)^2+\frac{(4+\Omega'^2)^2}{4Q^2}-\frac{4\,\Omega'(4+\Omega'^2)}{Q}} \\
&= \frac{Q(4+\Omega'^2)}{4\,\Omega'}\sqrt{1+\frac{1}{4Q^2}-\frac{4\,\Omega'}{Q(4+\Omega'^2)}} \\
&= \frac{\sqrt{4Q^2+1}}{2\,\Omega'}-\frac{Q}{\sqrt{4Q^2+1}}+O(\Omega')
\end{aligned} \tag{F.8}$$

while for the "gain" G_v we have (always see eq. F.5)

$$\begin{aligned}
G_{v_{RIS}} &= \left.\frac{2+az^{-1}}{-az^{-1}}\right|_{z=\mathrm{e}^{\mathrm{j}\Omega}} = \frac{2\mathrm{e}^{\mathrm{j}\Omega}+a}{-a} \\
&= \frac{4+2\,\Omega'/Q+\Omega'^2}{4-\Omega'^2}\cos\Omega-1+\mathrm{j}\,\frac{4+2\,\Omega'/Q+\Omega'^2}{4-\Omega'^2}\sin\Omega \\
&= \frac{2\,\Omega'/Q}{4+\Omega'^2}+\mathrm{j}\,\frac{4+2\,\Omega'/Q+\Omega'^2}{4-\Omega'^2}\,\frac{4\,\Omega'}{4+\Omega'^2}
\end{aligned} \tag{F.9}$$

whose modulus is

$$\begin{aligned}
|G_v|_{RIS} &= \frac{4\,\Omega'}{4-\Omega'^2}\sqrt{\frac{(4-\Omega'^2)^2+4Q^2(4+2\,\Omega'/Q+\Omega'^2)^2}{4Q^2\,(4+\Omega'^2)^2}} \\
&= \frac{4\,\Omega'}{4-\Omega'^2}\sqrt{\frac{(4-\Omega'^2)^2+4Q^2(4+\Omega'^2)^2+16\,\Omega'^2+16Q\,\Omega'(4+\Omega'^2)}{4Q^2\,(4+\Omega'^2)^2}} \\
&= \frac{4\,\Omega'}{4-\Omega'^2}\sqrt{1+\frac{1}{4Q^2}+\frac{4\,\Omega'}{Q(4+\Omega'^2)}} \\
&= \frac{\Omega'\sqrt{4Q^2+1}}{2Q}+\frac{\Omega'^2}{\sqrt{4Q^2+1}}+O(\Omega'^3)
\end{aligned} \tag{F.10}$$

We summarize the results obtained in the following table

	Valore	$\omega_0 T \to 0$	$Q \gg 1$ $\omega_0 T \to 0$	$Q = 1/2$ $\omega_0 T \to 0$
$\dfrac{u_{dc}}{x_{dc}}$	$\dfrac{1}{2\omega_0 T Q} + \dfrac{1}{2}$	$\dfrac{1}{2\omega_0 T Q}$	$\dfrac{1}{2\omega_0 T Q}$	$\dfrac{1}{\omega_0 T}$
$\dfrac{v_{dc}}{x_{dc}}$	$\dfrac{1}{\omega_0 T}\sqrt{1 - \dfrac{1}{4 Q^2}}$	$\dfrac{\sqrt{4 Q^2 + 1}}{2 \omega_0 T Q}$	$\dfrac{1}{\omega_0 T}$	0
$G_{u_{dc}}$	$\dfrac{1}{2\omega_0 T Q} - \dfrac{1}{2}$	$\dfrac{1}{2\omega_0 T Q}$	$\dfrac{1}{2\omega_0 T Q}$	$\dfrac{1}{\omega_0 T}$
$G_{v_{dc}}$	$\dfrac{\dfrac{2\omega_0 T}{Q} + 2\omega_0^2 T^2}{4 - \omega_0^2 T^2}$	$\dfrac{\omega_0 T}{2Q}$	$\dfrac{\omega_0 T}{2Q}$	$\omega_0 T$
$\left\lvert\dfrac{u}{x}\right\rvert_{RIS}$	$\dfrac{\sqrt{4 Q^2 + 1}}{2 \omega_0 T} + \dfrac{Q}{\sqrt{4 Q^2 + 1}} + O(\omega_0 T)$	$\dfrac{\sqrt{4 Q^2 + 1}}{2 \omega_0 T}$	$\dfrac{Q}{\omega_0 T}$	$\dfrac{\sqrt{2}}{2\omega_0 T}$
$\left\lvert\dfrac{v}{x}\right\rvert_{RIS}$	$\dfrac{(4 + \omega_0^2 T^2)\sqrt{4Q^2 - 1}}{8 \omega_0 T}$	$\dfrac{\sqrt{4 Q^2 + 1}}{2 \omega_0 T}$	$\dfrac{Q}{\omega_0 T}$	$\dfrac{\sqrt{2}}{2\omega_0 T}$
$\lvert G_u \rvert_{RIS}$	$\dfrac{\sqrt{4 Q^2 + 1}}{2 \omega_0 T} - \dfrac{Q}{\sqrt{4 Q^2 + 1}} + O(\omega_0 T)$	$\dfrac{\sqrt{4 Q^2 + 1}}{2 \omega_0 T}$	$\dfrac{Q}{\omega_0 T}$	$\dfrac{\sqrt{2}}{2\omega_0 T}$
$\lvert G_v \rvert_{RIS}$	$\dfrac{\omega_0 T \sqrt{4 Q^2 + 1}}{2 Q} + O(\omega_0^2 T^2)$	$\dfrac{\omega_0 T \sqrt{4Q^2 + 1}}{2 Q}$	$\omega_0 T$	$\omega_0 T \sqrt{2}$

Appendix G

State Space Filters: Unit step response

In this appendix we intend to explicitly compute the evolution of the state S when the matrices $\mathbf{A}$ e $\mathbf{B}$ have the following form and the input$x(n)$ is the sequence "unit step" $u(n)$

$$\mathbf{A} = \begin{pmatrix} -a-1 & -a-b-1 \\ 1 & 1 \end{pmatrix} \qquad \mathbf{B} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

with the usual meaning of the coefficients a and b. The state $(u\ v)^T$ is given by

$$\begin{cases} u = \dfrac{1 - z^{-1}}{1 + az^{-1} + bz^{-2}}\, x \\ v = \dfrac{z^{-1}}{1 + az^{-1} + bz^{-2}}\, x \end{cases}$$

If we indicate, using polar coordinates, with $\rho\, e^{j\alpha}$ and $\rho\, e^{-j\alpha}$ the position of the complex conjugated poles in the z plane, then

$$1 + az^{-1} + bz^{-2} = \left(1 - \rho\, e^{j\alpha} z^{-1}\right)\left(1 - \rho\, e^{-j\alpha} z^{-1}\right)$$

The z-transform of the input is $x(z) = \mathcal{Z}[u(n)] = \dfrac{1}{1 - z^{-1}}$. The z-transform of the u component of the state is

$$\begin{aligned} u(z) &= \frac{1 - z^{-1}}{\left(1 - \rho\, e^{j\alpha} z^{-1}\right)\left(1 - \rho\, e^{-j\alpha} z^{-1}\right)} \frac{1}{1 - z^{-1}} = z\, \frac{z}{\left(z - \rho\, e^{j\alpha}\right)\left(z - \rho\, e^{-j\alpha}\right)} \\ &= z \left(\frac{A}{z - \rho\, e^{j\alpha}} + \frac{B}{z - \rho\, e^{-j\alpha}} \right) \end{aligned}$$

where A and B are the residues of the function $F(z) = \dfrac{z}{\left(z - \rho\, e^{j\alpha}\right)\left(z - \rho\, e^{-j\alpha}\right)}$ respectively in the polesi $z = \rho\, e^{j\alpha}$ and $z = \rho\, e^{-j\alpha}$. Since the poles are of the first order, their values are

$$\begin{aligned} A &= \lim_{z \to \rho\, e^{j\alpha}} F(z)\left(z - \rho\, e^{j\alpha}\right) = \frac{\rho\, e^{j\alpha}}{\rho\, e^{j\alpha} - \rho\, e^{-j\alpha}} = \frac{e^{j\alpha}}{2j \sin\alpha} \\ B &= \lim_{z \to \rho\, e^{-j\alpha}} F(z)\left(z - \rho\, e^{-j\alpha}\right) = \frac{\rho\, e^{-j\alpha}}{\rho\, e^{-j\alpha} - \rho\, e^{j\alpha}} = \frac{-e^{-j\alpha}}{2j \sin\alpha} \end{aligned}$$

Therefore, the z-transform of the u component of the state becomes

$$\begin{aligned} u(z) &= \frac{1}{2j \sin\alpha} \left(\frac{z\, e^{j\alpha}}{z - \rho\, e^{j\alpha}} - \frac{z\, e^{-j\alpha}}{z - \rho\, e^{-j\alpha}} \right) \\ &= \frac{1}{2j \sin\alpha} \left(\frac{e^{j\alpha}}{1 - \rho\, e^{j\alpha} z^{-1}} - \frac{e^{-j\alpha}}{1 - \rho\, e^{-j\alpha} z^{-1}} \right) \end{aligned}$$

and since $\rho < 1$

$$
\begin{aligned}
u(z) &= \frac{1}{2\mathrm{j}\sin\alpha}\left(\mathrm{e}^{\mathrm{j}\alpha}\sum_{n=0}^{\infty}\rho^n \mathrm{e}^{\mathrm{j}\,n\alpha} z^{-n} - \mathrm{e}^{-\mathrm{j}\alpha}\sum_{n=0}^{\infty}\rho^n \mathrm{e}^{-\mathrm{j}\,n\alpha} z^{-n}\right) \\
&= \frac{1}{2\mathrm{j}\sin\alpha}\sum_{n=0}^{\infty}\rho^n\left(\mathrm{e}^{\mathrm{j}\,(n+1)\alpha} - \mathrm{e}^{-\mathrm{j}\,(n+1)\alpha}\right) z^{-n} = \frac{1}{\sin\alpha}\sum_{n=0}^{\infty}\rho^n \sin[(n+1)\alpha] z^{-n}
\end{aligned}
$$

and the response of the u component of the state to the unit step is

$$
u(n) = \frac{\rho^n \sin[(n+1)\alpha]}{\sin\alpha} = \frac{4 + 2\Omega_0\,T/Q + \Omega_0^2\,T^2}{2\,\Omega_0\,T}\,\frac{Q}{\sqrt{4Q^2-1}}\rho^{n+1}\sin[(n+1)\alpha]
$$

Proceeding analogously for the component v of the state we have for its z-transform

$$
\begin{aligned}
v(z) &= \frac{z^{-1}}{(1-\rho\,\mathrm{e}^{\mathrm{j}\alpha}z^{-1})\,(1-\rho\,\mathrm{e}^{-\mathrm{j}\alpha}z^{-1})}\,\frac{1}{1-z^{-1}} = z\,\frac{z}{(z-\rho\,\mathrm{e}^{\mathrm{j}\alpha})\,(z-\rho\,\mathrm{e}^{-\mathrm{j}\alpha})\,(z-1)} \\
&= z\left(\frac{A}{z-\rho\,\mathrm{e}^{\mathrm{j}\alpha}} + \frac{B}{z-\rho\,\mathrm{e}^{-\mathrm{j}\alpha}} + \frac{C}{z-1}\right),
\end{aligned}
$$

A, B and C are the residues of $F(z) = \dfrac{z}{(z-\rho\,\mathrm{e}^{\mathrm{j}\alpha})\,(z-\rho\,\mathrm{e}^{-\mathrm{j}\alpha})\,(z-1)}$. Also in this case the poles are of the first order and therefore the residue values are

$$
\begin{aligned}
A &= \lim_{z\to\rho\,e^{\mathrm{j}\alpha}} F(z)\left(z-\rho\,\mathrm{e}^{\mathrm{j}\alpha}\right) &&= \frac{\rho\,\mathrm{e}^{\mathrm{j}\alpha}}{(\rho\,\mathrm{e}^{\mathrm{j}\alpha}-\rho\,\mathrm{e}^{-\mathrm{j}\alpha})\,(\rho\mathrm{e}^{\mathrm{j}\alpha}-1)} \\
&= \frac{\mathrm{e}^{\mathrm{j}\alpha}\left(\rho\mathrm{e}^{-\mathrm{j}\alpha}-1\right)}{2\mathrm{j}\,\sin\alpha\,(1+\rho^2-2\rho\cos\alpha)} &&= \frac{\rho-\mathrm{e}^{\mathrm{j}\alpha}}{2\mathrm{j}\,\sin\alpha\,(1+\rho^2-2\rho\cos\alpha)} \\
B &= \lim_{z\to\rho\,e^{-\mathrm{j}\alpha}} F(z)\left(z-\rho\,\mathrm{e}^{-\mathrm{j}\alpha}\right) &&= \frac{\rho\,\mathrm{e}^{-\mathrm{j}\alpha}}{(\rho\,\mathrm{e}^{-\mathrm{j}\alpha}-\rho\,\mathrm{e}^{\mathrm{j}\alpha})\,(\rho\mathrm{e}^{-\mathrm{j}\alpha}-1)} \\
&= \frac{\mathrm{e}^{-\mathrm{j}\alpha}\left(\rho\mathrm{e}^{\mathrm{j}\alpha}-1\right)}{-2\mathrm{j}\,\sin\alpha\,(1+\rho^2-2\rho\cos\alpha)} &&= \frac{\rho-\mathrm{e}^{-\mathrm{j}\alpha}}{-2\mathrm{j}\,\sin\alpha\,(1+\rho^2-2\rho\cos\alpha)} \\
C &= \lim_{z\to 1} F(z)\,(z-1) = \frac{1}{(1-\rho\,\mathrm{e}^{\mathrm{j}\alpha})\,(1-\rho\,\mathrm{e}^{-\mathrm{j}\alpha})} &&= \frac{1}{1+\rho^2-2\rho\cos\alpha}
\end{aligned}
$$

Therefore, the z-transform of the v component of the state becomes

$$
\begin{aligned}
v(z) &= \frac{1}{1+\rho^2-2\rho\cos\alpha}\left[\frac{1}{2\mathrm{j}\sin\alpha}\left(\frac{\rho-\mathrm{e}^{\mathrm{j}\alpha}}{1-\rho\,\mathrm{e}^{\mathrm{j}\alpha}z^{-1}} - \frac{\rho-\mathrm{e}^{-\mathrm{j}\alpha}}{1-\rho\,\mathrm{e}^{-\mathrm{j}\alpha}z^{-1}}\right) + \frac{1}{1-z^{-1}}\right] \\
&= \frac{1}{1+a+b}\left\{\frac{1}{2\mathrm{j}\sin\alpha}\sum_{n=0}^{\infty}\left[\rho^{n+1}\left(\mathrm{e}^{\mathrm{j}\,n\alpha}-\mathrm{e}^{-\mathrm{j}\,n\alpha}\right)-\right.\right. \\
&\qquad\qquad \left.\left. -\rho^n\left(\mathrm{e}^{\mathrm{j}\,(n+1)\alpha}-\mathrm{e}^{-\mathrm{j}\,(n+1)\alpha}\right)\right] + 1\right\} z^{-n} \\
&= \frac{1}{1+a+b}\left\{\frac{1}{\sin\alpha}\sum_{n=0}^{\infty}\left[\rho^n\left(\rho\sin n\alpha - \sin(n+1)\alpha\right)\right] + 1\right\} z^{-n}
\end{aligned}
$$

and the response of the v component of the state to the unit step is

$$
\begin{aligned}
v(n) &= \frac{1}{1+a+b}\left\{\frac{1}{\sin\alpha}\left[\rho^n\left(\rho\sin n\alpha - \sin(n+1)\alpha\right)\right] + 1\right\} \\
&= \frac{1}{1+a+b}\left\{\frac{\rho^n\sin n\alpha(\rho^2-\rho\cos\alpha)}{\rho\sin\alpha} - \frac{\rho^n\cos n\alpha(\rho\sin\alpha)}{\rho\sin\alpha} + 1\right\} \\
&= \frac{4+2\Omega_0\,T/Q+\Omega_0^2\,T^2}{4\,\Omega_0^2\,T^2}\left\{\frac{\Omega_0\,T\,Q-1}{2\sqrt{4Q^2-1}}\rho^n\sin n\alpha - \rho^n\cos n\alpha + 1\right\}
\end{aligned}
$$

Bibliography

[1] Alan V. Oppenheim and Ronald W. Schafer, *Digital Signal Processing.* Englewood Cliffs, NJ: Prentice Hall, 1978.

[2] Thomas Kailath, *Linear Systems.*
Englewood Cliffs, NJ: Prentice Hall, 1980.

[3] Vladimir Ivanovič Smirnov, *Corso di matematica superiore, Vol. 3 Parte seconda.*
Roma: Editori Riuniti, 1978.