Lectures on Multivariable Feedback Control

Ali Karimpour

Department of Electrical Engineering, Faculty of Engineering, Ferdowsi University of Mashhad (September 2009)

Chapter 2: Introduction to Multivariable Control

Processes with only one output being controlled by a single manipulated variable are classified as single-input single-output (SISO) systems. Many processes, however, do not confirm to such a simple control configuration. In the industrial process for example, any unit operation capable of manufacturing or refining a product cannot do so with only a single control loop. In fact, each unit operation typically requires control over at least two variables, e.g. product rate and product quality. There are, therefore, usually at least two control loops to content with. Systems with more than one control loop are known as multi-input multi-output (MIMO) or multivariable systems.

This chapter considers some important aspects of multivariable systems. MIMO interconnections, poles and zeros in MIMO system, Smith form for polynomial matrix, Smith McMillan form (SMM) form, matrix fraction description (MFD) and performance specification in MIMO systems are specified in this chapter.

2-1 Multivariable Connections

Figure 2-1 shows cascade (series) interconnection of transfer matrices. The transfer matrix of the overall system is:

$$y(s) = G(s)u(s) = G_2(s)G_1(s)u(s) \qquad 2\text{-}1$$

Note that the transfer matrices must have suitable dimensions.

Parallel interconnection of transfer matrices is shown in Figure 2-2. The transfer matrix of the overall system is:

$$y(s) = G(s)u(s) = (G_1(s) + G_2(s))u(s) \qquad 2\text{-}2$$

Note that the transfer matrices must have suitable dimensions.

Feedback interconnection of transfer matrices is shown in figure 2-3. The transfer matrix of the overall system is:

$$y(s) = G(s)u(s) = (I + G_2(s)G_1(s))^{-1}G_2(s)G_1(s)u(s) \qquad 2\text{-}3$$

Note that the transfer matrices must have suitable dimensions.

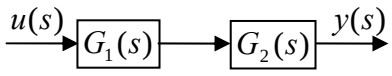


Figure 2-1 Cascade interconnection of transfer matrices

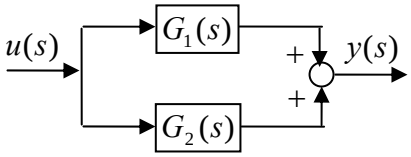


Figure 2-2 Parallel interconnection of transfer matrices

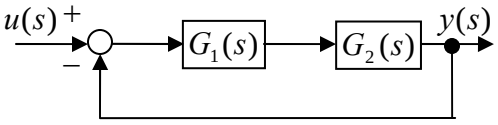


Figure 2-3 Feedback connection of transfer matrices

A useful relation in multivariable is push-through rule. Push-through rule is defined by:

$$(I + G_2(s)G_1(s))^{-1}G_2(s) = G_2(s)(I + G_1(s)G_2(s))^{-1} \qquad 2\text{-}4$$

The cascade and feedback rules can be combined to evaluate closed loop transfer matrix from block diagram.

MIMO rule: To derive the output of a system, start from the output and write down the blocks as you meet them when moving backward (against the signal flow) towards the input. If you exit from a feedback loop then include a term $(I - L)^{-1}$ or $(I + L)^{-1}$ according to the feedback sign where L is the transfer function around that loop (evaluated against the signal flow starting at the point of exit from the loop). Parallel branches should be treated independently and their contributions added together.

Example 2-1

Derive the transfer function of the system shown in figure 2-4.

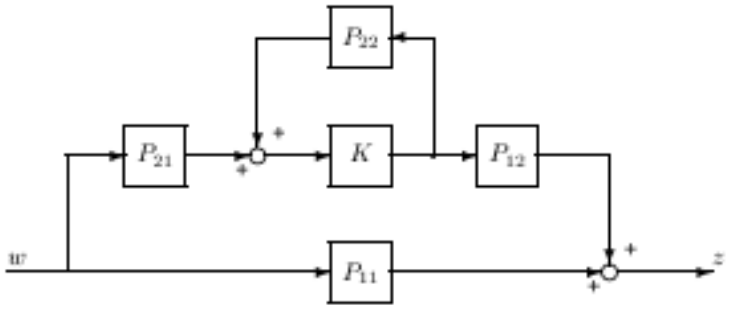


Figure 2-4 System used in Example 2-1

As it has two parallel ways from input to output by MIMO rule the transfer function is:

$$z = \left(P_{11} + P_{12}K(I - P_{22}K)^{-1}P_{21}\right)\omega$$

2-2 Multivariable Poles

Poles of a system can be derived from the state space realizations and the transfer functions.

2-2-1 Poles Derived from State Space Realizations

For simplicity we here define the poles of a system in terms of the eigenvalues of the state space A matrix.

Definition 2-1

The poles p_i of a system with state-space description (A, B, C, D) are eigenvalues $\lambda_i(A), i = 1, 2, ..., n$ of the matrix A. The pole polynomial or characteristic polynomial $\phi(s)$ is defined as $\phi(s) = \det(sI - A)$. Thus the system's poles are the roots of the characteristic polynomial

$$\phi(s) = \det(sI - A) = 0 \qquad 2\text{-}5$$

Note that if A does not correspond to a minimal realization then the poles by this definition will include the poles (eigenvalues) corresponding to uncontrollable and/or unobservable states.

2-2-2 Poles Derived from Transfer Functions

The poles of *G(s)* may be somewhat loosely defined as the finite values *s=p* where *G(p)* has a singularity (is infinite). The following theorem from MacFarlane and Karcanias allows one to

obtain the poles directly from the transfer function matrix *G(s)* and is also useful for hand calculations. It also has the advantage of yielding only the poles corresponding to a minimal realization of the system.

Theorem 2-1

The pole polynomial $\phi(s)$ corresponding to a minimal realization of a system with transfer function *G(s)* is the least common denominator of all non-identically-zero minors of all orders of *G(s)*. A minor of a matrix is the determinant of the square matrix obtained by deleting certain rows and/or columns of the matrix.

Example 2-2

Consider the plant $\frac{(3s+1)^2}{(s+1)}e^{-\vartheta s}$ which has no state-space realization as it contains a delay and is also improper. However from Theorem 2-1 we have that the denominator is s+1 and as expected *G(s)* has a pole at s=-1

Example 2-3

Consider the square transfer function matrix

$$G(s) = \frac{1}{1.25(s+1)(s+2)}\begin{bmatrix} s-1 & s \\ -6 & s-2 \end{bmatrix}$$

The minors of order 1 are the four elements which all have (s+1)(s+2) in the denominator. The minor of order 2 is the determinant

$$\det\big(G(s)\big) = \frac{(s-1)(s-2)+6s}{\big(1.25(s+1)(s+2)\big)^2} = \frac{1}{1.25^2(s+1)(s+2)}$$

Note the pole-zero cancellation when evaluating the determinant. The least common denominator of all the minors is then

$$\varphi(s) = (s+1)(s+2)$$

so a minimal realization of the system has two poles one at $s=-1$ and one at $s=-2$

Example 2-4

Consider the following system with 3 inputs and 2 outputs.

$$G(s)=\frac{1}{(s+1)(s+2)(s-1)}\begin{bmatrix}(s-1)(s+2) & 0 & (s-1)^2\\ -(s+1)(s+2) & (s-1)(s+1) & (s-1)(s+1)\end{bmatrix}$$

The minors of order 1 are the elements of *G(s)*, so they are

$$\frac{1}{s+1},\frac{s-1}{(s+1)(s+2)},\frac{-1}{s-1},\frac{1}{s+2},\frac{1}{s+2}$$

The minor of order 2 corresponding to the deletion of different columns are

$$\frac{2}{(s+1)(s+2)},\frac{1}{(s+1)(s+2)},\frac{-(s-1)}{(s+1)(s+2)^2}$$

By considering all minors we find their least common denominator

$$\varphi(s)=(s+1)(s+2)^2(s-1)$$

The system therefore has four poles one at $s=-1$, one at $s=1$ and two at $s=-2$. From the above examples we see that the MIMO poles are essentially the poles of the elements. However by looking at only the elements it is not possible to determine the multiplicity of the poles.

2-3 Multivariable Zeros

Zeros of a system arises when competing effects internal to the system are such that the output is zero even when the inputs and the states are not themselves identically zero.

2-3-1 Zeros Derived from State Space Realizations

Zeros are usually computed from a state space description of the system. First note that the state space equations of a system may be written as

$$P(s)\begin{bmatrix}x\\ u\end{bmatrix}=\begin{bmatrix}0\\ y\end{bmatrix},\qquad P(s)=\begin{bmatrix}sI-A & -B\\ C & D\end{bmatrix} \tag{2-6}$$

The zeros are then the values of $s=z$ for which the polynomial system matrix P(s) loses rank resulting in zero output for some nonzero input. Numerically the zeros are found as non trivial solutions to the following problem

$$(zI_g - M)\begin{bmatrix} x_z \\ u_z \end{bmatrix} = 0 \qquad \text{2-7}$$

where $M = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$, $I_g = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}$.

This is solved as a generalized eigenvalue problem. (In the conventional eigenvalue problem we have $I_g = I$).

The zeros resulting from a minimal realization are sometimes called the transmission zeros. If one does not have a minimal realization, then numerical computations may yield additional invariant zeros. These invariant zeros plus the transmission zeros are sometimes called the system zeros.

The invariant zeros can be further subdivided into input and output decoupling zeros. These cancel poles associated with uncontrollable or unobservable states and hence have limited practical significance.

If the system outputs contain direct information about each of the states and no direct connection from input, then there are no transmission zeros. This would be the case if $C = I$, $D = 0$, for example.

For square systems with *m=p* inputs and outputs and n states, limits on the number of transmission zeros are:

$$\begin{aligned} &D \neq 0 : \textit{At most } n - m + rank(D) \textit{ zeros} \\ &D = 0 : \textit{At most } n - 2m + rank(CB) \textit{ zeros} \\ &D = 0 \textit{ and } rank(CB) = m : \textit{Exactly } n - m \textit{ zeros} \end{aligned} \qquad \text{2-8}$$

Example 2-5

Consider the following state space realization

$$\begin{aligned} \dot{x} &= Ax + Bu \\ y &= Cx + Du \end{aligned}$$

where

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & -6 & -5 \end{bmatrix} \qquad B = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \qquad C = \begin{bmatrix} 4 & 1 & 0 \end{bmatrix} \qquad D = 0$$

Determine the zeros of the system.

Solution: First we derive the number of transmission zeros according to equation 2-8. The product of CB is

$$CB = \begin{bmatrix} 4 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = 0$$

So since $D = 0$ according to equation 2-8 the system has at most $n - 2m + rank(CB) = 3 - 2 = 1$ zero. To find the value of zero we construct M and I_g as follows

$$M = \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & -6 & -5 & 1 \\ 4 & 1 & 0 & 0 \end{bmatrix} \qquad I_g = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Now by use of generalized eigenvalue problem one can find the zeros. The following Matlab m.file can be used to derive zeros.

$$eig(M, I_g)$$

It shows that the system has a zero at $s = -4$.

2-3-2 Zeros Derived from Transfer Functions

For a SISO system the zeros z_i are the solutions to $G(z_i) = 0$. In general it can be argued that zeros are values of s at which $G(s)$ loses rank. This is the basis for the following definition of zeros for a multivariable system (MacFarlane and Karcanias).

Definition 2-2

z_i is a zero of $G(s)$ if the rank of $G(z_i)$ is less than the normal rank of $G(s)$. The zero polynomial is defined as $z(s) = \Pi_{i=1}^{n_z}(s - z_i)$. Where n_z is the number of finite zeros of $G(s)$.

We do not consider zeros at infinity. We require that z_i is finite. Recall that the normal rank of $G(s)$ is the rank of $G(s)$ at all values of s except at a finite number of singularities (which are the zeros) Note that this definition of zeros is based on the transfer function matrix corresponding to a minimal realization of a system. These zeros are sometimes called transmission zeros but we will

simply call them zeros. We may sometimes use the term multivariable zeros to distinguish them from the zeros of the elements of the transfer function matrix.

The following theorem from MacFarlane and Karcanias is useful for hand calculating the zeros of a transfer function matrix $G(s)$.

Theorem 2-2

The zero polynomial $z(s)$ corresponding to a minimal realization of the system is the greatest common divisor of all the numerators of all order-*r* minors of *G(s)* where *r* is the normal rank of $G(s)$ provided that these minors have been adjusted in such a way as to have the pole polynomial $\phi(s)$ as their denominators.

Example 2-6

Consider the transfer function matrix

$$G(s) = \frac{1}{s+2}\begin{bmatrix} s-1 & 4 \\ 4.5 & 2(s-1) \end{bmatrix}$$

The normal rank of G(s) is 2 and the minor of order 2 is the determinant of G(s)

$\det(G(s)) = \dfrac{2(s-1)^2 - 18}{(s+2)^2} = 2\dfrac{s-4}{s+2}$. From Theorem 2-1 the pole polynomial is $\phi(s) = s+2$ and therefore the zero polynomial is $z(s) = s-4$. Thus G(s) has a single RHP-zero at $s = 4$.

This illustrates that in general multivariable zeros have no relationship with the zeros of the transfer function elements. This is also shown by the following example where the system has no zeros.

Example 2-7

Consider the following system

$$G(s) = \frac{1}{1.25(s+1)(s+2)}\begin{bmatrix} s-1 & s \\ -6 & s-2 \end{bmatrix}$$

according to example 2-3 the pole polynomial is:

$$\varphi(s) = (s+1)(s+2)$$

The normal rank of G(s) is 2 and the minor of order 2 is the determinant of G(s), where *det(G(s))* with in $\phi(s)$ as its denominator is

$$\det\big(G(s)\big)=\frac{(s-1)(s-2)+6s}{\big(1.25(s+1)(s+2)\big)^2}=\frac{1}{1.25^2(s+1)(s+2)}$$

Thus the zero polynomial is given by the numerator which is 1, and we find that the system has no multivariable zeros.

Example 2-8

Consider the system

$$G(s)=\begin{bmatrix}\dfrac{s-1}{s+1} & \dfrac{s-2}{s+2}\end{bmatrix}$$

The normal rank of *G(s)* is 1 and since there is no value of s for which both elements become zero, *G(s)* has no zeros.

In general non-square systems are less likely to have zeros than square systems. The following is an example of a non square system which has a zero.

Example 2-9

Consider the following system

$$G(s)=\frac{1}{(s+1)(s+2)(s-1)}\begin{bmatrix}(s-1)(s+2) & 0 & (s-1)^2\\ -(s+1)(s+2) & (s-1)(s+1) & (s-1)(s+1)\end{bmatrix}$$

according to example 2-4 the pole polynomial is:

$$\varphi(s)=(s+1)(s+2)^2(s-1)$$

The minors of order 2 with $\phi(s)$ as their denominators are

$$\frac{2(s-1)(s+2)}{(s+1)(s+2)^2(s-1)},\frac{(s-1)(s+2)}{(s+1)(s+2)^2(s-1)},\frac{-(s-1)^2}{(s+1)(s+2)^2(s-1)}$$

The greatest common divisor of all the numerators of all order-*2* minors is $z(s)=s-1$. Thus, the system has a single RHP-zero located at s = 1.

We also see from the last example that a minimal realization of a MIMO system can have poles and zeros at the same value of *s* provided their directions are different. This is discussed in the next section.

2-4 Directions of Poles and Zeros

Zero directions: Let *G(s)* have a zero at *s = z*, Then *G(s)* loses rank at *s = z* and there will exist nonzero vectors u_z and y_z such that

$$G(z)u_z = 0 \qquad y_z^H G(z) = 0 \tag{2-9}$$

here u_z is defined as the input zero direction and y_z is defined as the output zero direction. We usually normalize the direction vectors to have unit length i.e. $\|u_z\|_2 = 1$ and $\|y_z\|_2 = 1$. From a practical point of view the output zero direction y_z is usually of more important than u_z because y_z gives information about which output _or combination of outputs_ may be difficult to control. In principle we may obtain u_z and y_z from an SVD of $G(z) = Y\Sigma U^H$ and we have that u_z is the last column in *U*, corresponding to the zero singular value of *G(z)* and y_z is the last column of *Y*. A better approach numerically is to obtain u_z from a state space description using the generalized eigenvalue problem in 2-7.

Pole directions: Let *G(s)* have a pole at *s = p*. Then *G(p)* is infinite and we may somewhat crudely write

$$G(p)u_p = \infty \qquad y_p^H G(p) = \infty \tag{2-10}$$

where u_p is the input pole direction and y_p the output pole direction. As for u_z and y_z the vectors u_p and y_p may be obtained from an SVD of $G(p) = Y\Sigma U^H$. Then u_p is the first column in *U* corresponding to the infinite singular value and y_p the first column in *Y*. If the inverse of *G(p)* exists then it follows from the SVD that

$$G^{-1}(p)y_p = 0 \qquad u_p G^{-1}(p) = 0 \tag{2-11}$$

However if we have a state space realization of *G(s)* then it is better to determine the pole directions from the right and left eigenvectors of *A*. Specifically if *p* is pole of *G(s)* then *p* is an eigenvalue of *A*. Let t_p and q_p be the corresponding right and left eigenvectors i.e.

$$At_p = pt_p \qquad q_p^H A = pq_p^H \tag{2-12}$$

then the pole directions are

$$y_p = Ct_p \qquad u_p = B^H q_p \tag{2-13}$$

Example 2-10

Consider the following plant

$$G(s) = \frac{1}{s+2}\begin{bmatrix} s-1 & 4 \\ 4.5 & 2(s-1) \end{bmatrix}$$

It has a RHP zero at $z = 4$ and a LHP pole at $p = -2$. We will use an SVD of *G(z)* and *G(p)* to determine the zero and pole directions. But we stress that this is not a reliable method numerically.

To find the zero direction consider

$$G(z) = G(4) = \frac{1}{6}\begin{bmatrix} 3 & 4 \\ 4.5 & 6 \end{bmatrix} = \frac{1}{6}\begin{bmatrix} 0.55 & -0.83 \\ 0.83 & 0.55 \end{bmatrix}\begin{bmatrix} 9.01 & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} 0.6 & -0.8 \\ 0.8 & 0.6 \end{bmatrix}^H$$

The zero input and output directions are associated with the zero singular value of *G(z)* and we get

$$u_z = \begin{bmatrix} -0.8 \\ 0.6 \end{bmatrix} \text{ and } y_z = \begin{bmatrix} -0.83 \\ 0.55 \end{bmatrix}$$

We see from y_z that the zero has a slightly larger component in the first output. Next, to determine the pole directions consider

$$G(p+\varepsilon) = G(-2+\varepsilon) = \frac{1}{\varepsilon}\begin{bmatrix} -3+\varepsilon & 4 \\ 4.5 & 2(-3+\varepsilon) \end{bmatrix}$$

The SVD as $\varepsilon \to 0$ yields

$$G(-2+\varepsilon) = \frac{1}{\varepsilon}\begin{bmatrix} -0.55 & 0.83 \\ 0.83 & 0.55 \end{bmatrix}\begin{bmatrix} 9.01 & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} 0.6 & 0.8 \\ -0.8 & 0.6 \end{bmatrix}$$

The pole input and outputs directions are associated with the largest singular value, $\frac{9.01}{\varepsilon}$ and we get $u_p = \begin{bmatrix} 0.6 \\ -0.8 \end{bmatrix}$ and $y_p = \begin{bmatrix} -0.55 \\ 0.83 \end{bmatrix}$

2-5 Smith Form of a Polynomial Matrix

Suppose that $\Pi(s)$ is a polynomial matrix. **Smith form** of $\Pi(s)$ is denoted by $\Pi_s(s)$, and it is a pseudo diagonal in the following form

$$\Pi_s(s) = \begin{bmatrix} \Pi_{ds}(s) & 0 \\ 0 & 0 \end{bmatrix} \qquad 2\text{-}14$$

and $\Pi_{ds}(s)$ is a square diagonal matrices in the following form

$$\Pi_{ds}(s) = diag\left\{\bar{\varepsilon}_1(s), \bar{\varepsilon}_2(s), ..., \bar{\varepsilon}_r(s)\right\} \qquad 2\text{-}15$$

Furthermore, $\bar{\varepsilon}_i(s)$ is a factor of $\bar{\varepsilon}_{i+1}(s)$. $\bar{\varepsilon}_i(s)$ is derived from minors of $\Pi(s)$ as $\bar{\varepsilon}_i(s) = \frac{\chi_i}{\chi_{i-1}}$

where χ_i derived by:

$$\begin{aligned} &\chi_0 = 1 \\ &\chi_1 = \text{gcd}\{\text{all monic minors of degree 1}\} \\ &\chi_2 = \text{gcd}\{\text{all monic minors of degree 2}\} \\ &. \\ &. \\ &\chi_r = \text{gcd}\{\text{all monic minors of degree r}\} \end{aligned} \qquad 2\text{-}16$$

gcd stands for greatest common divisor and monic is a polynomial that the coefficient of its greatest degree is one.

The three elementary operations for a polynomial matrix are used to find Smith form.

- Multiplying a row or column by a constant;
- Interchanging two rows or two columns; and
- Adding a polynomial multiple of a row or column to another row or column.

These operations are carried out on a transfer matrix $\Pi(s)$ by either pre-multiplication or post-multiplication by unimodular polynomial matrices known as elementary matrices. A polynomial

matrix is unimodular if its inverse also is a polynomial matrix. Pre-multiplication of $\Pi(s)$ by an elementary matrix produces the corresponding row operation, while post-multiplication produces a column operation. $\Pi_s(s)$ is Smith form of $\Pi(s)$ and they are said to be equivalent, denoted by $\Pi_s(s) \sim \Pi(s)$ if there exists a set of elementary matrices L_i and R_i such that

$$\Pi_s(s) = L_{n2}(s).....L_2(s)L_1(s)\Pi(s)R_1(s)R_2(s).....R_{n1}(s) \qquad 2\text{-}17$$

Example 2-11

Consider the following polynomial matrix

$$\Pi(s) = \begin{bmatrix} 4 & -(s+2) \\ 2(s+2) & -\frac{1}{2} \end{bmatrix}$$

so we have $\chi_0 = 1,\ \chi_1 = \gcd\{1, s+2, s+2, 1\} = 1,\ \chi_2 = \gcd\{s^2 + 4s + 3\} = (s+1)(s+3)$

and now $\bar{\varepsilon}_i(s)$ are:

$$\bar{\varepsilon}_1(s) = \frac{\chi_1}{\chi_0} = 1 \text{ and } \bar{\varepsilon}_2(s) = \frac{\chi_2}{\chi_1} = (s+1)(s+3)$$

the Smith of $\Pi(s)$ is:

$$\Pi_s(s) = \begin{bmatrix} 1 & 0 \\ 0 & (s+1)(s+3) \end{bmatrix}$$

Example 2-12

Consider the following polynomial matrix

$$\Pi(s) = \begin{bmatrix} 1 & -1 \\ s^2 + s - 4 & 2s^2 - s - 8 \\ (s+2)(s-2) & (s+2)(2s-4) \end{bmatrix}$$

so we have

$\chi_0 = 1, \chi_1 = \gcd\{1, s^2 + s - 4, s^2 - 0.5s - 4, s^2 - 4, s^2 - 4\} = 1$ and

$\chi_2 = \gcd\{s^2 - 4, s^2 - 4, s(s^2 - 4)\} = (s^2 - 4)$

and now $\bar{\varepsilon}_i(s)$ are:

$$\bar{\varepsilon}_1(s) = \frac{\chi_1}{\chi_0} = 1 \text{ and } \bar{\varepsilon}_2(s) = \frac{\chi_2}{\chi_1} = (s^2 - 4)$$

the Smith form of $\Pi(s)$ is:

$$\Pi_s(s) = \begin{bmatrix} 1 & 0 \\ 0 & (s^2 - 4) \\ 0 & 0 \end{bmatrix}$$

2-6 Smith-McMillan Forms

The Smith-McMillan form is used to determine the poles and zeros of the transfer matrices of systems with multiple inputs and/or outputs. The transfer matrix is a matrix of transfer functions between the various inputs and outputs of the system. The poles and zeros that are of interest are the poles and zeros of the transfer matrix itself, not the poles and zeros of the individual elements of the matrix. The locations of the poles of the transfer matrix are available by inspection of the individual transfer functions, but the total number of the poles and their multiplicity is not. The location of system zeros, or even their existence, is not available by looking at the individual elements of the transfer matrix.

The transfer matrix will be denoted by *G(s).* The number of rows in *G(s)* is equal to the number of system outputs; that will be denoted by *m*. The number of columns in *G(s)* is equal to the number of system inputs; that will be denoted by *p*. Thus, *G(s)* is an $m \times p$ matrix of transfer functions. The normal rank of *G(s)* is *r*, where $r \leq min\{p, m\}$.

Following theorem gives a diagonal form for a rational transfer-function matrix:

Theorem 2-3 (Smith-McMillan form)

Let $G(s)=[g_{ij}(s)]$ be an $m \times p$ matrix transfer function, where $g_{ij}(s)$ are rational scalar transfer functions, *G(s)* can be represented by:

$$G(s)=\frac{\Pi(s)}{D_G(s)} \qquad 2\text{-}18$$

where $\Pi(s)$ is an $m \times p$ polynomial matrix of rank *r* and $D_G(s)$ is the least common multiple of the denominators of all elements of *G(s)* .

Then, $\tilde{G}(s)$ is Smith McMillan form of *G(s)* and can be derived directly by

$$\tilde{G}(s)=\frac{\Pi_s(s)}{D_G(s)}=\frac{1}{D_G(s)}\begin{bmatrix}\Pi_{ds}(s) & 0\\ 0 & 0\end{bmatrix}=\begin{bmatrix}M(s) & 0\\ 0 & 0\end{bmatrix} \qquad 2\text{-}19$$

where $M(s)$ is:

$$M(s)=diag\left\{\frac{\varepsilon_1(s)}{\delta_1(s)},\frac{\varepsilon_2(s)}{\delta_2(s)},.....,\frac{\varepsilon_r(s)}{\delta_r(s)}\right\} \qquad 2\text{-}20$$

where $\{\varepsilon_i(s),\delta_i(s)\}$ is a pair of monic and coprime polynomials for $i=1,2,...,r$.

Furthermore, $\varepsilon_i(s)$ is a factor of $\varepsilon_{i+1}(s)$ and $\delta_{i+1}(s)$ is a factor of $\delta_i(s)$. Elements of the matrix *M(s)* can be defined by:

$$m_{ii}(s)=\frac{\bar{\varepsilon}_i(s)}{D_G(s)}=\frac{\varepsilon_i(s)}{\delta_i(s)} \qquad 2\text{-}21$$

where $\bar{\varepsilon}_i(s)$ are diagonal elements of $\Pi_s(s)$ (Smith form of $\Pi(s)$) as

$$\Pi_{ds}(s)=diag\left\{\bar{\varepsilon}_1(s),\bar{\varepsilon}_2(s),.....,\bar{\varepsilon}_r(s)\right\} \qquad 2\text{-}22$$

We recall that a matrix *G(s)*, and its Smith-McMillan form $\tilde{G}(s)$ are equivalent matrices. Thus, there exist two unimodular matrices, *L(s)* and *R(s)*, such that

$$\tilde{G}(s)=L(s)G(s)R(s) \qquad 2\text{-}23$$

L(s) and *R(s)* are the unimodular matrices that convert $\Pi(s)$ to its Smith form $\Pi_s(s)$. Then there exist two matrices $\tilde{L}(s)$ and $\tilde{R}(s)$, such as

$$G(s) = \tilde{L}(s)\tilde{G}(s)\tilde{R}(s) \qquad 2\text{-}24$$

where $\tilde{L}(s)$ and $\tilde{R}(s)$ are also unimodular and:

$$\tilde{L}(s) = L(s)^{-1} \ , \ \ \tilde{R}(s) = R(s)^{-1} \qquad 2\text{-}25$$

The poles and zeros of the transfer matrix *G(s)* can be found from the elements of *M(s).* The pole polynomial is defined as

$$\phi(s) = \prod_{i=1}^{r} \delta_i(s) = \delta_1(s)\delta_2(s).....\delta_r(s) \qquad 2\text{-}26$$

Repeated poles can also be identified by inspection of $\phi(s)$. The total number of poles in the system is given by $\deg(\phi(s))$, which is known as the McMillan degree. It is the dimension of a minimal state-space representation of *G(s).*

A state-space representation of *G(s)* may be of higher order than the McMillan degree, indicating pole-zero cancellations in the system.

In similar fashion, the zero polynomial is defined as

$$z(s) = \prod_{i=1}^{r} \varepsilon_i(s) = \varepsilon_1(s)\varepsilon_2(s).....\varepsilon_r(s) \qquad 2\text{-}27$$

The roots of *z(s) = 0* are known as the transmission zeros of *G(s)*. It can be seen that any transmission zero of the system must be a factor in at least one of the $\varepsilon_i(s)$ polynomials. The normal rank of both M(s) and *G(s)* is *r*. It is clear that if any $\varepsilon_i(s)$ be zero, then the rank of M(s) drops below *r*. Therefore, since the ranks of *M(s)* and *G(s)* are always equal, so *G(s)* loses rank. We illustrate the Smith-McMillan form by a simple example.

Example 2-13

Consider the following transfer-function matrix

$$G(s)=\begin{bmatrix} \dfrac{4}{(s+1)(s+2)} & \dfrac{-1}{(s+1)} \\ \dfrac{2}{(s+1)} & \dfrac{-1}{2(s+1)(s+2)} \end{bmatrix}$$

We can then express *G(s)* in the form:

$$G(s)=\frac{\Pi(s)}{D_G(s)},\qquad \Pi(s)=\begin{bmatrix} 4 & -(s+2) \\ 2(s+2) & -0.5 \end{bmatrix},\qquad D_G(s)=(s+1)(s+2)$$

According example 2-11 the Smith form of $\Pi(s)$ is:

$$\Pi_s(s)=\begin{bmatrix} 1 & 0 \\ 0 & (s+1)(s+3) \end{bmatrix}$$

So the Smith McMillan form of *G(s)* is:

$$\tilde{G}(s)=\frac{\Pi_s(s)}{D_G(s)}=\begin{bmatrix} \dfrac{1}{(s+1)(s+2)} & 0 \\ 0 & \dfrac{s+3}{s+2} \end{bmatrix}$$

Clearly the pole polynomial and the zero polynomial are:

$$\phi(s)=(s+1)(s+2)^2\ ,\ z(s)=s+3$$

Example 2-14

Consider the following example of a system with *m* = 3 outputs and *p* = 2 inputs. The transfer matrix is shown below;

$$G(s)=\begin{bmatrix} \dfrac{1}{(s+1)(s+2)} & \dfrac{-1}{(s+1)(s+2)} \\ \dfrac{s^2+s-4}{(s+1)(s+2)} & \dfrac{2s^2-s-8}{(s+1)(s+2)} \\ \dfrac{s-2}{(s+1)} & \dfrac{2s-4}{(s+1)} \end{bmatrix}$$

We can then express *G(s)* in the form:

$$G(s)=\frac{\Pi(s)}{D_G(s)},\quad \Pi(s)=\begin{bmatrix} 1 & -1 \\ s^2+s-4 & 2s^2-s-8 \\ (s-2)(s+2) & (2s-4)(s+2) \end{bmatrix},\quad D_G(s)=(s+1)(s+2)$$

according example 2-12 the Smith form of $\Pi(s)$ is:

$$\Pi_s(s)=\begin{bmatrix} 1 & 0 \\ 0 & (s^2-4) \\ 0 & 0 \end{bmatrix}$$

So the Smith McMillan form of *G(s)* is:

$$\tilde{G}(s)=\frac{\Pi_s(s)}{D_G(s)}=\begin{bmatrix} \dfrac{1}{(s+1)(s+2)} & 0 \\ 0 & \dfrac{s-2}{s+1} \\ 0 & 0 \end{bmatrix}$$

Clearly pole polynomial and zero polynomial are:

$$\phi(s)=(s+2)(s+1)^2\ ,\ z(s)=s-2$$

2-7 Matrix Fraction Description (MFD)

A model structure that is related to the Smith-McMillan form is matrix fraction description (MFD). There are two types, namely a right matrix fraction description (RMFD) and a left matrix fraction description (LMFD).

First of all suppose $\tilde{G}(s)$ is a $m \times m$ matrix and is the Smith McMillan form of G(s), define the following two matrices:

$$N(s) \stackrel{\Delta}{=} diag\left(\varepsilon_1(s), ..., \varepsilon_r(s), 0, ..., 0\right) \quad \text{2-28}$$

$$D(s) \stackrel{\Delta}{=} diag\left(\delta_1(s), ..., \delta_r(s), 1, ..., 1\right) \quad \text{2-29}$$

where $N(s)$ and $D(s)$ are $m \times m$ matrices. Hence $\tilde{G}(s)$, can be written as

$$\tilde{G}(s) = N(s)D(s)^{-1} \quad \text{2-30}$$

Combining 2-24 and 2-30, we can write

$$G(s) = \tilde{L}(s)\tilde{G}(s)\tilde{R}(s) = \tilde{L}(s)N(s)D(s)^{-1}\tilde{R}(s) = \tilde{L}(s)N(s)\left(R(s)D(s)\right)^{-1} = G_N(s)G_D(s)^{-1} \quad \text{2-31}$$

This is known as a **right matrix fraction description** (RMFD) where:

$$G_N(s) = \tilde{L}(s)N(s) \quad , \quad G_D(s) = R(s)D(s) \quad \text{2-32}$$

If one start with $\tilde{G}(s) = D(s)^{-1}N(s)$ then combining with 2-24

$$G(s) = \tilde{L}(s)\tilde{G}(s)\tilde{R}(s) = \tilde{L}(s)D(s)^{-1}N(s)\tilde{R}(s) = \left(D(s)L(s)\right)^{-1}N(s)\tilde{R}(s) = \overline{G}_D(s)^{-1}\overline{G}_N(s) \quad \text{2-33}$$

This is known as a **left matrix fraction description** (LMFD) where:

$$\overline{G}_D(s) = D(s)L(s) \quad , \quad \overline{G}_N(s) = N(s)\tilde{R}(s) \quad \text{2-34}$$

The left and right matrix descriptions have been initially derived starting from the Smith-McMillan form. Hence, the factors are polynomial matrices. However, it is immediate to see that they provide a more general description. In particular, $G_N(s), G_D(s), \overline{G}_D(s)$ and $\overline{G}_N(s)$ are

generally matrices with rational entries. One possible way to obtain this type of representation is to divide the two polynomial matrices forming the original MFD by the same (stable) polynomial.

We also observe that the RMFD (LMFD) is not unique, because, for any nonsingular $m \times m$ matrix $\Omega(s)$ we can write *G(s)* as

$$G(s) = G_N(s)\left(\Omega(s)\Omega(s)^{-1}\right)G_D(s)^{-1} = \left(G_N(s)\Omega(s)\right)\left(G_D(s)\Omega(s)\right)^{-1} \qquad 2\text{-}35$$

where $\Omega(s)$ is said to be a right common factor. When the only right common factors of $G_N(s)$ and $G_D(s)$ is unimodular matrix, then, we say that $G_N(s)$ and $G_D(s)$ are right coprime. In this case, we say that the RMFD $\left(G_N(s), G_D(s)\right)$ is **irreducible**.

It is easy to see that when a RMFD is irreducible, then

- $s = z$ is a zero of *G(s)* if and only if $G_N(s)$ loses rank at $s = z$; and
- $s = p$ is a pole of *G(s)* if and only if $G_D(s)$ is singular at $s = p$. This means that the pole polynomial of *G(s)* is $\phi(s) = \det\left(G_D(s)\right)$.

An example showing the above concepts is considered next.

Example 2-15

Consider a 2×2 MIMO system having the transfer function

$$G(s) = \begin{bmatrix} \dfrac{4}{(s+1)(s+2)} & \dfrac{-0.5}{s+1} \\ \dfrac{1}{s+2} & \dfrac{2}{(s+1)(s+2)} \end{bmatrix}$$

a) Find the Smith-McMillan form by performing elementary row and column operations.

b) Find the poles and zeros.

c) Build a RMFD for the model.

Solution

a) We first compute its Smith-McMillan form by performing elementary row and column operations.

$$\tilde{G}(s) = L(s)G(s)R(s) = \begin{bmatrix} \dfrac{1}{(s+1)(s+2)} & 0 \\ 0 & \dfrac{s^2+3s+18}{(s+1)(s+2)} \end{bmatrix}$$

where

$$L(s) = \begin{bmatrix} 0.25 & 0 \\ -2(s+1) & 8 \end{bmatrix}, \qquad R(s) = \begin{bmatrix} 1 & \dfrac{s+2}{8} \\ 0 & 1 \end{bmatrix}$$

b) We see that the observable and controllable part of the system has zero and pole polynomials given by

$$z(s) = s^2 + 3s + 18 \text{ , } \phi(s) = (s+1)^2(s+2)^2$$

So the poles are -1, -1, -2 and -2 and zeros are $-1.5 \pm j3.97$

c) To derive RMFD we need

$$\tilde{L}(s) = L(s)^{-1} = \begin{bmatrix} 4 & 0 \\ s+1 & 0.125 \end{bmatrix}, \qquad \tilde{R}(s) = R(s)^{-1} = \begin{bmatrix} 1 & -\dfrac{s+2}{8} \\ 0 & 0 \end{bmatrix}$$

$$N(s) = \begin{bmatrix} 1 & 0 \\ 0 & s^2+3s+18 \end{bmatrix}, \qquad D(s) = \begin{bmatrix} (s+1)(s+2) & 0 \\ 0 & (s+1)(s+2) \end{bmatrix}$$

So

$$G_N(s)=\begin{bmatrix}4 & 0\\ s+1 & 0.125\end{bmatrix}\begin{bmatrix}1 & 0\\ 0 & s^2+3s+18\end{bmatrix}=\begin{bmatrix}4 & 0\\ s+1 & \dfrac{s^2+3s+18}{8}\end{bmatrix}$$

$$G_D(s)=\begin{bmatrix}1 & \dfrac{s+2}{8}\\ 0 & 1\end{bmatrix}\begin{bmatrix}(s+1)(s+2) & 0\\ 0 & (s+1)(s+2)\end{bmatrix}=\begin{bmatrix}(s+1)(s+2) & \dfrac{(s+1)(s+2)^2}{8}\\ 0 & (s+1)(s+2)\end{bmatrix}$$

Matrix fraction description (MFD) can be extended to $n\times m$ non square matrix *G(s)*. In RMFD, $G_N(s)$ is $n\times m$ and $G_D(s)$ is $m\times m$ and in LMFD, $\overline{G}_D(s)$ is $n\times n$ and $\overline{G}_N(s)$ is $n\times m$. Following example shows the procedure of finding RMFD and LMFD for a non square matrix *G(s)*.

Example 2-16

Consider the following transfer matrix;

$$G(s)=\begin{bmatrix}\dfrac{1}{(s+1)(s+2)} & \dfrac{-1}{(s+1)(s+2)}\\ \dfrac{s^2+s-4}{(s+1)(s+2)} & \dfrac{2s^2-s-8}{(s+1)(s+2)}\\ \dfrac{s-2}{(s+1)} & \dfrac{2s-4}{(s+1)}\end{bmatrix}$$

Find RMFD and LMFD of the system.

Solution

According example 2-14 the SMM form of *G(s)* is:

$$\tilde{G}(s)=\frac{\Pi_s(s)}{D_G(s)}=\begin{bmatrix}\dfrac{1}{(s+1)(s+2)} & 0\\ 0 & \dfrac{s-2}{s+1}\\ 0 & 0\end{bmatrix}$$

To derive the RMFD and LMFD we must find the unimodular matrices that convert $\Pi(s)$ to $\Pi_s(s)$. So

$$\Pi_s(s) = L(s)\Pi(s)R(s) = \begin{bmatrix} 1 & 0 & 0 \\ -(s^2+s-4) & 1 & 0 \\ s & -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ s^2+s-4 & 2s^2-s-8 \\ (s-2)(s+2) & (2s-4)(s+2) \end{bmatrix} \begin{bmatrix} 1 & \frac{1}{3} \\ 0 & \frac{1}{3} \end{bmatrix}$$

Now we write *G(s)* according to $\tilde{G}(s)$:

$$G(s) = \tilde{L}(s)\tilde{G}(s)\tilde{R}(s) = \begin{bmatrix} 1 & 0 & 0 \\ s^2+s-4 & 1 & 0 \\ s^2-4 & 1 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{(s+1)(s+2)} & 0 \\ 0 & \frac{s-2}{s+1} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 0 & 3 \end{bmatrix} =$$

$$\begin{bmatrix} 1 & 0 & 0 \\ s^2+s-4 & 1 & 0 \\ s^2-4 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & s-2 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} (s+2)(s+1) & 0 \\ 0 & s+1 \end{bmatrix}^{-1} \begin{bmatrix} 1 & -1 \\ 0 & 3 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ s^2+s-4 & s-2 \\ s^2-4 & s-2 \end{bmatrix} \begin{bmatrix} (s+2)(s+1) & \frac{s+1}{3} \\ 0 & \frac{s+1}{3} \end{bmatrix}^{-1}$$

Above equation leads to RMFD and

$$G_N(s) = \begin{bmatrix} 1 & 0 \\ s^2+s-4 & s-2 \\ s^2-4 & s-2 \end{bmatrix}, \; G_D(s) = \begin{bmatrix} (s+2)(s+1) & \frac{s+1}{3} \\ 0 & \frac{s+1}{3} \end{bmatrix}$$

To derive LMFD the $\tilde{G}(s)$ must partitioned as LMFD so:

$$G(s)=\tilde{L}(s)\tilde{G}(s)\tilde{R}(s)=\begin{bmatrix}1 & 0 & 0\\ s^2+s-4 & 1 & 0\\ s^2-4 & 1 & 1\end{bmatrix}\begin{bmatrix}\dfrac{1}{(s+1)(s+2)} & 0\\ 0 & \dfrac{s-2}{s+1}\\ 0 & 0\end{bmatrix}\begin{bmatrix}1 & -1\\ 0 & 3\end{bmatrix}=$$

$$\begin{bmatrix}1 & 0 & 0\\ s^2+s-4 & 1 & 0\\ s^2-4 & 1 & 1\end{bmatrix}\begin{bmatrix}(s+2)(s+1) & 0 & 0\\ 0 & s+1 & 0\\ 0 & 0 & 1\end{bmatrix}^{-1}\begin{bmatrix}1 & 0\\ 0 & s-2\\ 0 & 0\end{bmatrix}\begin{bmatrix}1 & -1\\ 0 & 3\end{bmatrix}=\begin{bmatrix}(s+2)(s+1) & 0 & 0\\ (s+1)(4-s^2-s) & s+1 & 0\\ s & -1 & 1\end{bmatrix}^{-1}\begin{bmatrix}1 & -1\\ 0 & 3(s-2)\\ 0 & 0\end{bmatrix}$$

so we find

$$\overline{G}_D(s)=\begin{bmatrix}(s+2)(s+1) & 0 & 0\\ (s+1)(4-s^2-s) & s+1 & 0\\ s & -1 & 1\end{bmatrix},\quad \overline{G}_N(s)=\begin{bmatrix}1 & -1\\ 0 & 3(s-2)\\ 0 & 0\end{bmatrix}$$

2-8 Scaling

Scaling is very important in practical applications as it makes model analysis and controller design (weight selection) much simple. It requires the engineer to make a judgment at the start of the design process about the required performance of the system. To do this, decisions are made on the expected magnitudes of disturbances and reference changes, on the allowed magnitude of each input signal, and on the allowed deviation of each output. Let the unscaled (or original system) linear model of the process in Figure 2-5a be

$$\hat{y}=\hat{G}\hat{u}+\hat{G}_d\hat{d};\qquad \hat{e}=\hat{y}-\hat{r} \tag{2-36}$$

where a hat (^) is used to show that the variables are in their unscaled (or originally system) units. A useful approach for scaling is to make the variables less than one in magnitude. This is done by dividing each variable by its maximum expected or allowed change. For disturbances and manipulated inputs, we use the scaled variables

$$d = \frac{\hat{d}}{\hat{d}_{\max}}, \quad u = \frac{\hat{u}}{\hat{u}_{\max}} \qquad \text{2-37}$$

where

$\_\hat{d}_{\max}$: largest expected change in disturbance

$\_\hat{u}_{\max}$: largest allowed input change

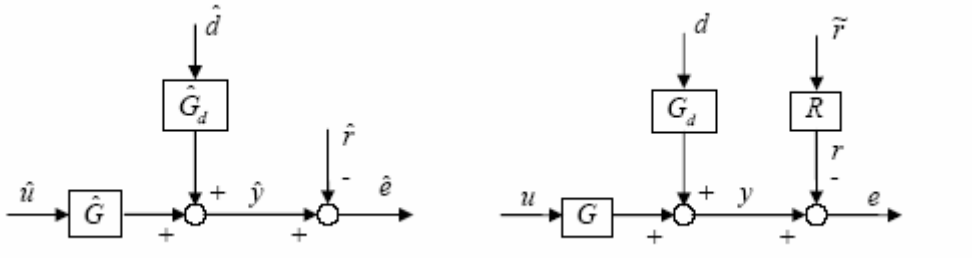
The maximum deviation from the nominal value should be chosen by thinking of the maximum value one can expect (or allow) as a function of time. The variables $\hat{y}, \hat{e}$ and $\hat{r}$ are in the same units, so the same scaling factor should be applied to each. Two alternatives are possible:

$\_\hat{e}_{\max}$: largest allowed control error

$\_\hat{r}_{\max}$: largest expected change in reference value

Since a major objective of control is to minimize the control error, we here usually choose to scale with respect to the maximum control error:

$$e = \frac{\hat{e}}{\hat{e}_{\max}}, \quad r = \frac{\hat{r}}{\hat{e}_{\max}}, \quad y = \frac{\hat{y}}{\hat{e}_{\max}} \qquad \text{2-38}$$



(a) **(b)**

Figure 2-5 Model in terms of (a) original variable and (b) scaled variable

To formalize the scaling procedure, introduce the scaling factors

$$D_e = \hat{e}_{\max} \ , \ D_u = \hat{u}_{\max} \ , \ D_d = \hat{d}_{\max} \ , \ D_r = \hat{r}_{\max} \qquad 2\text{-}39$$

For MIMO systems each variable in the vectors $\hat{d}$, $\hat{r}$, $\hat{u}$ and $\hat{e}$ may have a different maximum value, in which case D_e , D_u , D_d and D_r become diagonal scaling matrices. This ensures, for example, that all errors (outputs) are of about equal importance in terms of their magnitude.

The corresponding scaled variables to use for control purposes are then

$$d = D_d^{-1}\hat{d} \ , \ u = D_u^{-1}\hat{u} \ , \ y = D_e^{-1}\hat{y} \ , \ e = D_e^{-1}\hat{e} \ , \ r = D_e^{-1}\hat{r} \qquad 2\text{-}40$$

On substituting 2-40 into 2-36 we get

$$D_e y = \bar{G} D_e u + \bar{G}_d D_d d \ ; \qquad D_e e = D_e y - D_e r \qquad 2\text{-}41$$

and introducing the scaled transfer functions

$$G = D_e^{-1}\hat{G}D_u \ , \ G_d = D_e^{-1}\hat{G}_d D_d \qquad 2\text{-}42$$

then yields the following model in terms of scaled variables

$$y = Gu + G_d d \ ; \qquad e = y - r \qquad 2\text{-}43$$

Here *u* and *d* should be less than 1 in magnitude, and it is useful in some cases to introduce a scaled reference $\tilde{r}$ which is less than 1 in magnitude. This is done by dividing the reference by the maximum expected reference change

$$\tilde{r} = \frac{\hat{r}}{\hat{r}_{\max}} = D_r^{-1}\hat{r} \qquad 2\text{-}44$$

We then have that

$$r = R\tilde{r} \quad \text{where} \quad R \cong D_e^{-1} D_r = \frac{\hat{r}_{\max}}{\hat{e}_{\max}} \qquad \text{2-45}$$

Here R is the largest expected change in reference relative to the allowed control error, typically, $R \geq 1$. The block diagram for the system in scaled variables may then be written as in Figure 2-5b for which the following control objective is relevant:

In terms of scaled variables we have that $|d(t)| \leq 1$ and $|\tilde{r}(t)| \leq 1$, and our control objective is to design *u* with $|u(t)| \leq 1$ such that $|e(t)| = |y(t) - r(t)| \leq 1$ (at least most of the time).

2-9 Performance Specification

In the application of automatic controllers, it is important to realize that controller and process form a unit, credit or discredit for results obtained are attributable to one as much as the other. A poor controller is often able to perform acceptably on a process which is easily controllable. The finest controller made, when applied to a miserably designed process, may not deliver the desired performance. There are some important definitions in this matter.

- **Nominal stability NS**: The system is stable with no model uncertainty.
- **Nominal Performance NP**: The system satisfies the performance specifications with no model uncertainty.
- **Robust stability RS**: The system is stable for all perturbed plants about the nominal model up to the worst case model uncertainty.
- **Robust performance RP**: The system satisfies the performance specifications for all perturbed plants about the nominal model up to the worst case model uncertainty.

Performance specification can be considered in time and frequency domain.

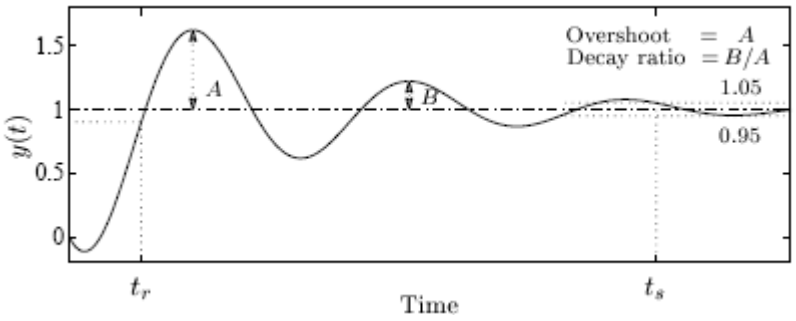


Figure 2-6 Step response of a system

2-9-1 Time Domain Performance

Although closed loop stability is an important issue, the real objective of control is to improve performance, that is, to make the output y(t) behave in a more desirable manner. Actually, the possibility of inducing instability is one of the disadvantages of feedback control which has to be traded off against performance improvement. The objective of this section is to discuss ways of evaluating closed loop performance.

Step response analysis approach, often taken by engineers when evaluating the performance of a control system. That is, one simulates the response to a step in the reference input, and considers characteristics shown in Figure 2-6.

- **Rise time, t_r** , the time it takes for the output to first reach 90% of its final value, which is usually required to be small.
- **Settling time, t_s** , the time after which the output remains within $\pm 5\%$ (or $\pm 2\%$) of its final value, which is usually required to be small.
- **Overshoot, P.O**, the peak value divided by the final value, which should typically be less than 20% or less.
- **Decay ratio**, the ratio of the second and first peaks, which should typically be 0.3 or less.
- **Steady state offset, e_{ss},** the difference between the final value and the desired final value, which is usually required to be small.
- **Excess variation**, the total variation (TV) divided by the overall change at steady state, which should be as close to 1 as possible. The total variation is the total movement of the

output as illustrated in Figure 2-7. For the cases considered here the overall change is 1, so the excess variation is equal to the total variation.

$$TV = \sum_i v_i \quad \text{Excess variation} = TV / v_0 \tag{2-46}$$

Note that the step response is equal to the integral of the corresponding impulse response, e.g. set u=1 in the following convolution integral.

$$y(t) = \int_0^t g(\tau)u(t-\tau)d\tau \tag{2-47}$$

where $g(\tau)$ is the impulse response. One can compute the total variation as the integrated absolute area (1-norm), of the corresponding impulse response

$$TV = \int_0^\infty g(\tau)d\tau = \|g(t)\|_1 \tag{2-48}$$

ISE, IAE, ITSE, ITAE: These measures are integral squared error, integral absolute error, integral time weighted squared error and integral time weighted absolute error respectively. For example IAE is defined as

$$IAE = \int_0^\infty |e(\tau)|d\tau \tag{2-49}$$

The rise time and settling time are measures of the speed of the response, whereas the overshoot, decay ratio, *TV*, ISE, IAE, ITSE, ITAE and steady state offset are related to the quality of the response.

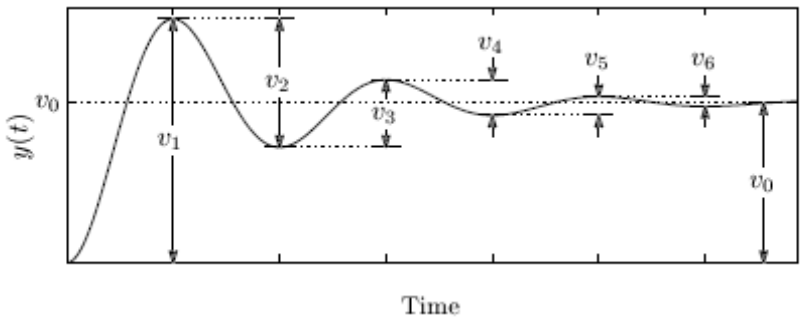


Figure 2-7 Total variation in the step response of a system

2-9-2 Frequency Domain Performance

The frequency response of the loop transfer function, $L(j\omega)$, or of various closed-loop transfer functions, may also be used to characterize closed-loop performance. Typical Bode plot of *L* is shown in Figure 2-8. One advantage of the frequency domain compared to a step response analysis is that it considers a broader class of signals (sinusoids of any frequency). This makes it easier to characterize feedback properties, and in particular system behaviors in the crossover (bandwidth) region. We will now describe some of the important frequency domain measures used to assess performance, e.g. gain and phase margins, the maximum peaks of *T* and *S,* and the various definitions of crossover and bandwidth frequencies used to characterize speed of response.

Let *L(s)* denote the loop transfer function of a system which is closed-loop stable under negative feedback. A typical Bode plot and a typical Nyquist plot of $L(j\omega)$ illustrating the gain margin (GM) and phase margin (PM) are given in Figures 2-8 and 2-9, respectively.

From Nyquist's stability condition, the closeness of the curve $L(j\omega)$ to the point -1 in the complex plane is a good measure of how close a stable closed-loop system is to instability.

We see from Figure 2-8 that GM measures the closeness of $L(j\omega)$ to -1 along the real axis, whereas PM is a measure along the unit circle.

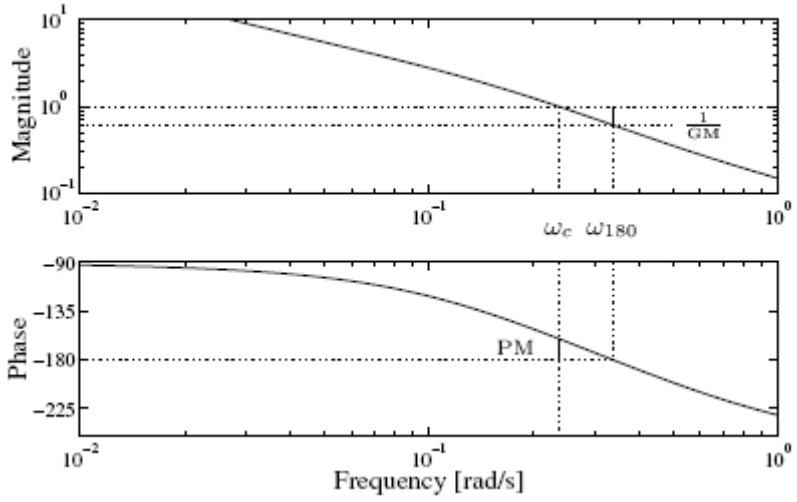


Figure 2-8 Bode plot of $L(j\omega)$.

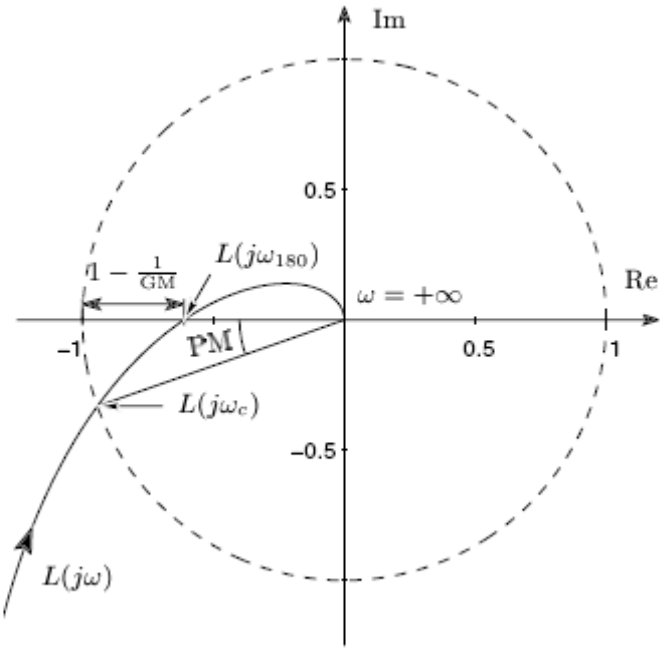


Figure 2-9 Nyquist plot of $L(j\omega)$.

More precisely, if the Nyquist plot of $L(j\omega)$ crosses the negative real axis between -1 and 0, then the (upper) gain margin is defined as

$$GM = \frac{1}{|L(j\omega_{180})|} \tag{2-50}$$

where the phase crossover frequency ω_{180} is where the Nyquist curve of $L(j\omega)$ crosses the negative real axis between -1 and 0, i.e.

$$\angle L(j\omega_{180}) = 180 \tag{2-51}$$

The phase margin is defined as

$$PM = \angle L(j\omega_c) + 180 \tag{2-52}$$

where the gain crossover frequency ω_c is the frequency where $L(j\omega)$ crosses 1, i.e.

$$|L(j\omega_c)| = 1 \tag{2-53}$$

The PM is a direct safeguard against time delay uncertainty; the system becomes unstable if we add a time delay of

$$\theta_{\max} = PM / \omega_c \tag{2-54}$$

Note that the units must be consistent, and so if ω_c is in [rad/s] then PM must be in radians. It is also important to note that by decreasing the value of ω_c (lowering the closed-loop bandwidth, resulting in a slower response) the system can tolerate larger time delay errors.

Stability margins are measures of how close a stable closed-loop system is to instability. From the above arguments we see that the GM and PM provide stability margins for gain and delay uncertainty. More generally, to maintain closed-loop stability, the Nyquist stability condition tells us that the number of encirclements of the critical point -1 by $L(j\omega)$ must not change. As discussed next, the actual closest distance is equal to $1/M_s$ where M_s is the peak value of the sensitivity $S(j\omega)$. As expected, the GM and PM are closely related to M_s, and since $|S|$ is also a measure of performance; they are therefore also useful in terms of performance. In summary, specifications on the GM and PM (e.g. GM > 2 and PM $>30^\circ$) are used to provide the appropriate trade-off between performance and stability robustness.

The maximum peaks of the sensitivity and complementary sensitivity functions are defined as

$$M_s = \max_\omega |S(j\omega)| \qquad M_T = \max_\omega |T(j\omega)| \tag{2-55}$$

Since *S+T=1* so *S* and *T* differ at most by 1. A large value of M_s therefore occurs if and only if M_T is large.

We now give some justification for why we may want to reduce the value of M_s. Consider the one degree-of-freedom configuration in Figure 2-10. Let we define error signal as $e = y - r$, then without control and noise ($u = n = 0$), we have $e = y - r = G_d d - r$, and with feedback control

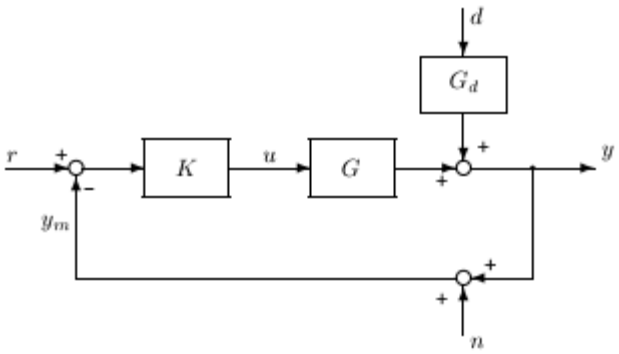


Figure 2-10 One degree-of-freedom configuration

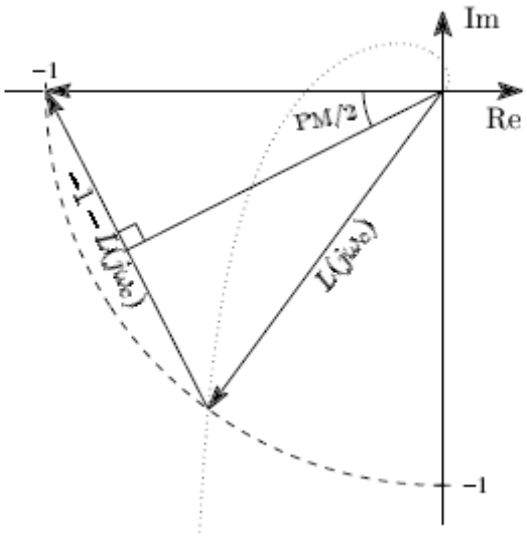


Figure 2-11 Nyquist plot of $L(j\omega)$.

$e = y - r = SG_d d - Sr = S(G_d d - r)$. Thus, feedback control improves performance in terms of reducing $|e|$ at all frequencies where $|S| < 1$.

One may also view M_s as a robustness measure. To maintain closed-loop stability, we want $L(j\omega)$ to stay away from the critical point -1. According to Figure 2-11 the smallest distance between $L(j\omega)$ and -1 is M_s^{-1}, and therefore for robustness, the smaller M_s, is better. In summary, both for stability and performance we want M_s close to 1.

There is a close relationship between these maximum peaks and the GM and PM. Specifically, for a given M_s we are guaranteed

$$GM \geq \frac{M_s}{M_s - 1}; \qquad PM \geq 2\sin^{-1}\left(\frac{1}{2M_s}\right) \geq \frac{1}{M_s}[rad] \tag{2-56}$$

For example, with $M_s = 2$ we are guaranteed GM > 2 and PM > 29°. Similarly, for a given value of M_T we are guaranteed

$$GM \geq 1 + \frac{1}{M_T}; \qquad PM \geq 2\sin^{-1}\left(\frac{1}{2M_T}\right) \geq \frac{1}{M_T}[rad] \tag{2-57}$$

and specifically with $M_T = 2$ we have GM > 1.5 and PM > 29°.

2-10 Trade-offs in Frequency Domain

Consider the simple one degree-of-freedom configuration in Figure 2-10. The input to the controller $K(s)$ is $r - y_m$ and the measured output is $y_m = y + n$ where *n* is the measurement noise. Thus, the input to the plant is

$$u = K(s)(r - y - n) \tag{2-58}$$

The objective of control is to manipulate *u* (design *K*) such that the control error *e* remains small in spite of disturbances *d* and noises *n*. The control error *e* is defined as

$$e = y - r \tag{2-59}$$

where *r* denotes the reference value (set point) for the output. Note that we do not define *e* as the controller input $r - y_m$ which is frequently done.

The plant model is written as

$$y = G(s)u + G_d(s)d \tag{2-60}$$

and for a one degree-of-freedom controller the substitution of 2-58 and 2-59 into 2-60 yields

$$y = G(s)K(s)(r - y - n) + G_d(s)d \tag{2-61}$$

or

$$(I + G(s)K(s))y = G(s)K(s)(r - n) + G_d(s)d \tag{2-62}$$

and hence the closed-loop response is

$$y = \underbrace{(I + G(s)K(s))^{-1}G(s)K(s)}_{T}r + \underbrace{(I + G(s)K(s))^{-1}}_{S}G_d(s)d - \underbrace{(I + G(s)K(s))^{-1}G(s)K(s)}_{T}n \tag{2-63}$$

The control error is

$$e = y - r = -Sr + SG_d d - Tn \tag{2-64}$$

where we have used the fact $T + S = I$. The corresponding plant input signal is

$$u = KSr - KSG_d d - KSn \tag{2-65}$$

The following notation and terminology are used

$L = GK$ loop transfer function

$S = (I + GK)^{-1} = (I + L)^{-1}$ sensitivity function

$T = (I + GK)^{-1}GK = (I + L)^{-1}L$ complementary sensitivity function

We see that S is the closed-loop transfer function from the output disturbances to the outputs, while T is the closed-loop transfer function from the reference signals to the outputs. The term complementary sensitivity for T follows from the identity

$$T + S = I \tag{2-66}$$

The term sensitivity function is natural because S gives the sensitivity reduction afforded by feedback. To see this, consider the "open-loop" case i.e. with no control (K=0). Then the error is

$$e = y - r = -r + G_d d + 0n \tag{2-67}$$

and a comparison with 2-64 shows that, with the exception of noise, the response with feedback is obtained by pre multiplying the right hand side by S.

Remark: Actually, the above is not the original reason for the name "sensitivity" Bode first called S as sensitivity because it gives the relative sensitivity of the closed-loop transfer function T to the relative plant model error. In particular, at a given frequency ω we have for a SISO plant, by straightforward differentiation of T, that

$$S = \frac{dT/T}{dG/G} \tag{2-68}$$

Recall equation 2-64 which yields the closed-loop response in terms of the control error e,

$$e = y - r = -Sr + SG_d d - Tn$$

For "perfect control" we want $e = y - r = 0$ that is, we would like

$$S = 0, \quad T = 0 \tag{2-69}$$

The first requirements in 2-69 is namely disturbance rejection and command tracking, and is obtained with $S \approx 0$ or equivalently $T \approx I$. Since $S = (I + L)^{-1}$ this implies that the loop transfer function L must be large in magnitude. On the other hand, the requirement for zero noise transmission implies that $T \approx 0$ or equivalently $S \approx I$, which is obtained with $L \approx 0$. This illustrates the fundamental nature of feedback design which always involves a trade-off between conflicting objectives, in this case between large loop gains for disturbance rejection and tracking, and small loop gains to reduce the effect of noise.

It is also important to consider the magnitude of the control action u, (which is the input to the plant). We want u small because this causes less wear and saves input energy, and also because u

is often a disturbance to other parts of the system (e.g. consider opening a window in your office to adjust your body temperature and the undesirable disturbance this will impose on the air conditioning system for the building. In particular, we usually want to avoid fast changes in *u.* The control action is given by $u = K(r - y_m)$ and we find as expected that a small *u* corresponds to small controller gains and a small $L = GK$.

The most important design objectives which necessitate trade-offs in feedback control are summarized below.

1- Performance, good disturbance rejection: needs large controller gains, i.e. *L* large or $T \approx I$.

2- Performance, good command following: *L* large or $T \approx I$.

3- Stabilization of unstable plant: *L* large or $T \approx I$.

4- Mitigation of measurement noise on plant outputs: *L* small or $T \approx 0$.

5- Small magnitude of input signals: *K* small and *L* small or $T \approx 0$.

6- Physical controller must be strictly proper: *L* has approach to 0 at high frequencies or $T \approx 0$.

7- Nominal stability (stable plant): *L* small

Fortunately, the conflicting design objectives mentioned above are generally in different frequency ranges, and we can meet most of the objectives by using a large loop gain $(|L| > 1)$ at low frequencies below crossover, and a small gain $(|L| < 1)$ at high frequencies above crossover.

2-11 Bandwidth and Crossover Frequency

The concept of bandwidth is very important in understanding the benefits and trade-offs involved when applying feedback control. Above we considered peaks of closed-loop transfer functions, M_T and M_s which are related to the quality of the response. However, for performance we must also consider the speed of the response, and this leads to considering the bandwidth frequency of the system. In general, a large bandwidth corresponds to a smaller rise time, since high-frequency signals are more easily passed on to the outputs. A high bandwidth also indicates a system which

is sensitive to noise. Conversely, if the bandwidth is small, the time response will generally be slow, and the system will usually be more robust.

Loosely speaking, bandwidth may be defined as the frequency range $[\omega_1, \omega_2]$ over which control is effective. In most cases we require tight control at steady-state so $\omega_1 = 0$, and we then simply call $\omega_2 = \omega_B$.

The word "effective" may be interpreted in different ways, and this may give rise to different definitions of bandwidth. The interpretation we use is that control is effective if we obtain some benefit in terms of performance. For tracking performance the error is $e = y - r = Sr$ (see Figure 2-10 and let $n = d = 0$) and we get that feedback is effective (in terms of improving performance) as long as the relative error $|e|/|r| = |S|$ is reasonably small, which we may define to be $|S| \le 0.707$.

We then get the following definition:

Definition 2-3

The (closed-loop) bandwidth, ω_B, is the frequency where $|S(j\omega)|$ first crosses $\frac{1}{\sqrt{2}} = -3db$ from below.

Remark. Another interpretation is to say that control is effective if it significantly changes the output response. For tracking performance, the output is $y = Tr$ (see Figure 2-10 and let $n = d = 0$), we may say that control is effective as long as T is reasonably large, which we may define to be larger than 0.707.

This leads to an alternative definition which has been traditionally used to define the bandwidth of a control system.

Definition 2-4

The (closed-loop) bandwidth, ω_{BT}, is the highest frequency at which $|T(j\omega)|$ crosses $\frac{1}{\sqrt{2}} = -3db$ from above.

However, we would argue that this alternative definition, although being closer to how the term is used in some other fields, is less useful for feedback control.

Another important definition for bandwidth is as follows:

Definition 2-5

The gain crossover frequency, ω_c, defined as the frequency where $|L(j\omega)|$ first crosses 1from above, is also sometimes used to define closed-loop bandwidth.

It has the advantage of being simple to compute and usually gives a value between ω_B and ω_{BT}. Specifically, for systems with PM < 90° (most practical systems) we have

$$\omega_B \le \omega_c \le \omega_{BT} \qquad 2\text{-}70$$

In conclusion ω_B (which is defined in terms of *S*) and also ω_c (in terms of *L*) are good indicators of closed-loop performance, while ω_{BT} (in terms of *T*) may be misleading in some cases.

Example 2-17 Comparison of ω_B and ω_{BT} as indicators of performance.

Following is an example where ω_{BT} is a poor indicator of performance.

$$L = \frac{-s+z}{s(\tau s+\tau z+2)}; \qquad T = \frac{-s+z}{s+z}\frac{1}{\tau s+1}; \qquad z = 0.1,\ \tau = 1$$

For this system, both *L* and *T* have a RHP-zero at $z = 0.1$ and we have GM=2.1, PM=60.1°, $M_s = 1.93$, and $M_T = 1$. We find that $\omega_B = 0.036$ and $\omega_c = 0.054$ are both less than $z = 0.1$ (as one should expect because speed of response is limited by the presence of RHP-zeros), whereas $\omega_{BT} = 1/\tau = 1$ is ten times larger than $z = 0.1$. The closed-loop response to a unit step change in the reference is shown in Figure 2-12. The rise time is 31.0 sec which is close to $1/\omega_B = 28.0\,\text{sec}$ but very different from $1/\omega_{BT} = 1.0\,\text{sec}$ illustrating that ω_B is a better indicator of closed-loop performance than ω_{BT}.

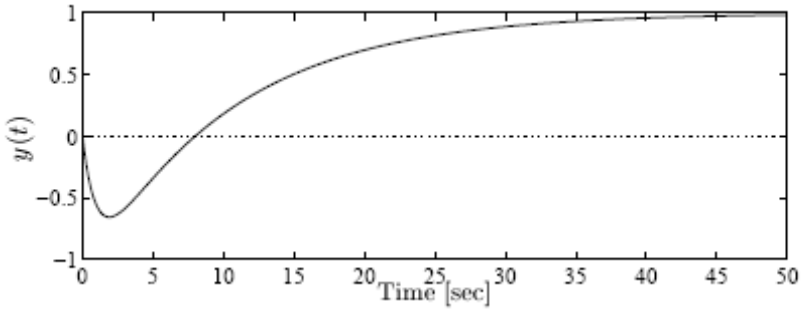


Figure 2-12 Step response for system $T = \frac{-s+0.1}{s+0.1}\frac{1}{s+1}$

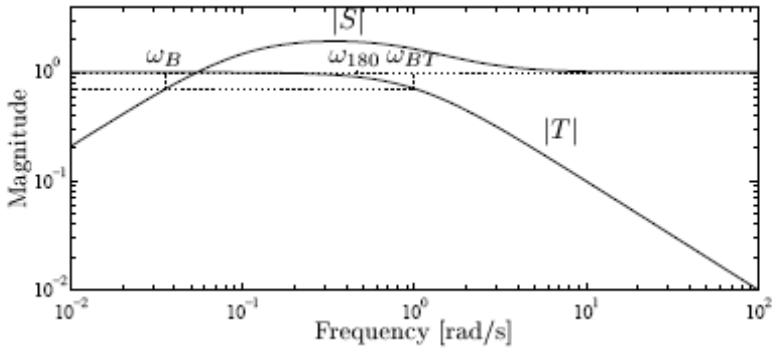


Figure 2-13 Plot of $|S|$ and $|T|$ for system $T = \dfrac{-s+0.1}{s+0.1}\dfrac{1}{s+1}$

The Bode plots of *S* and *T* are shown in Figure 2-13. We see that $|T| \approx 1$ up to about ω_{BT}. However, in the frequency range from ω_B to ω_{BT} the phase of *T* (not shown) drops from about -40° to about -220°, so in practice tracking is poor in this frequency range. For example, at frequency $\omega_{180} = 0.46$ we have $T = -0.9$ and the response to a sinusoidally varying reference $r(t) = \sin \omega_{180} t$ is completely out of phase, i.e. $y(t) \approx -0.9 r(t)$.

We thus conclude that $|T|$ by itself is not a good indicator of performance, we must also consider its phase. The reason is that we want $T \approx 1$ in order to have good performance, and it is not sufficient that $|T| \approx 1$. On the other hand, $|S|$ by itself is a reasonable indicator of performance, it is not necessary to consider its phase. The reason for this is that for good performance we want *S* close to 0 and this will be the case if $|S| \approx 0$ irrespective of the phase of *S*.

Exercises

2-1 Proof the equation 2-4.

2-2 Derive the pre and post-multiplication matrices change $\Pi(s)$ to $\Pi_s(s)$ in example 2-11.

2-3 Derive the pre and post-multiplication matrices change $\Pi(s)$ to $\Pi_s(s)$ in example 2-12.

2-4 Derive the LMFD of the system in example 2-15.

2-5 Consider following system.

$$\dot{x} = \begin{bmatrix} 1 & 3 & -1 \\ -2 & 0 & 1 \\ 3 & 1 & 2 \end{bmatrix} x \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} u$$

$$y = [1 \quad 0 \quad 0]x$$

a) Find the SMM form of the system.

b) Find the pole and zero polynomials of the system.

c) Find the RMFD and LMFD of the system.

2-6 Consider following transfer matrix:

$$G(s) = \begin{bmatrix} \dfrac{1}{s+2} \\ \dfrac{s-2}{s+4} \\ \\ 3 \end{bmatrix}$$

d) Find the SMM form of the system.

e) Find the pole and zero polynomials of the system.

f) Find the RMFD and LMFD of the system.

2-7 Consider following transfer matrix:

$$G(s) = \begin{bmatrix} \dfrac{1}{s+2} & \dfrac{s+2}{s+4} \\ \dfrac{s-2}{s+4} & \dfrac{s+1}{s-3} \\ \dfrac{3}{s+2} & \dfrac{4}{2s-1} \end{bmatrix}$$

g) Find the SMM form of the system.

h) Find the pole and zero polynomials of the system.

i) Find the RMFD and LMFD of the system.

2-8 Consider $g(s) = \frac{100}{s(s+2)}$ with a unity negative feedback.

a) Draw the step response of the system.

b) From the figure derived in part "a" determine: rise time, settling time, overshoot, decay ratio, steady state offset and excess variation.

c) Find excess variation and IAE by eq.2-48 and 2-49 respectively.

2-9 By use of figure 2-6 derive equations 2-56 and 2-57.

2-10 In the example 1.1 of the main reference (Skogestd,2005) suppose the acceptable variations in room temperature are 0.5° K, furthermore, the heat input can vary between 0 W and 1000 W, finally, the expected variations in outdoor temperature are -10° K, and $+20^\circ$ K. Find the scaled transfer function.

2-11 By use of Figure 2-11 derive equations 2-70.

2-12 By use of the main reference (Skogestd,2005) show by an example that ω_B is a better index for performance than ω_{BT}.

References

Skogestad Sigurd and Postlethwaite Ian. (2005) *Multivariable Feedback Control*: England, John Wiley & Sons, Ltd.

Maciejowski J.M. (1989). *Multivariable Feedback Design*: Adison-Wesley.