The roots s_{Pi} of the denominator polynomials are poles (∞ points) of the function $G(s)$. The poles and zeros describe the transfer function up to the prefactor K. Over and above this, the nature of the solution of the associated differential equation and therefore the dynamic behaviour of the transfer element or system under consideration depends essentially on the poles s_{Pi}; they stand as solutions λ_i of the characteristic polynomial in the exponent of the exponential functions, from which the solution of the homogenous differential equation will be made up.

By means of the poles and zeros it is possible to create a graphic representation of a transfer function, as shown by the examples in table 3-3. The poles or zeros are usually indicated by crosses or circles in the complex s plane. As poles and zeros are constant complex values, the position of the indicated symbol is not a function of any independent variable. We can recognize (table 3-3) that the transfer function of the proportionally acting element only consists of the transfer factor which cannot be represented by means of poles and zeros and that I-, D- and PT_1-elements will be characterized by means of a single pole or zero. We can also recognize that the multiplication of transfer functions (necessary for serial connections) can be represented by means of superimposing the corresponding pole / zero maps. This is possible because the zeros and poles of a factor are at the same time the zeros or poles of the whole product, providing one or more individual poles and zeros do not have the same value and can therefore be cancelled out in the resulting transfer function.

3.8 Limit theorems

The limit theorems of the Laplace transform represent a useful tool for a variety of applications. Using the information regarding the initial and final values of the function $f(t)$ in table 3-2, we obtain the relationship between the limit values of a transfer function $G(s)$ and the corresponding unit step response $h(t)$

$$\lim_{t\to\infty} h(t) = \lim_{s\to 0} sH(s) = \lim_{s\to 0} sG(s)\frac{1}{s} = \lim_{s\to 0} G(s) \quad ,$$

$$\lim_{t\to 0} h(t) = \lim_{s\to\infty} sH(s) = \lim_{s\to\infty} sG(s)\frac{1}{s} = \lim_{s\to\infty} G(s) \quad , \tag{3.81}$$

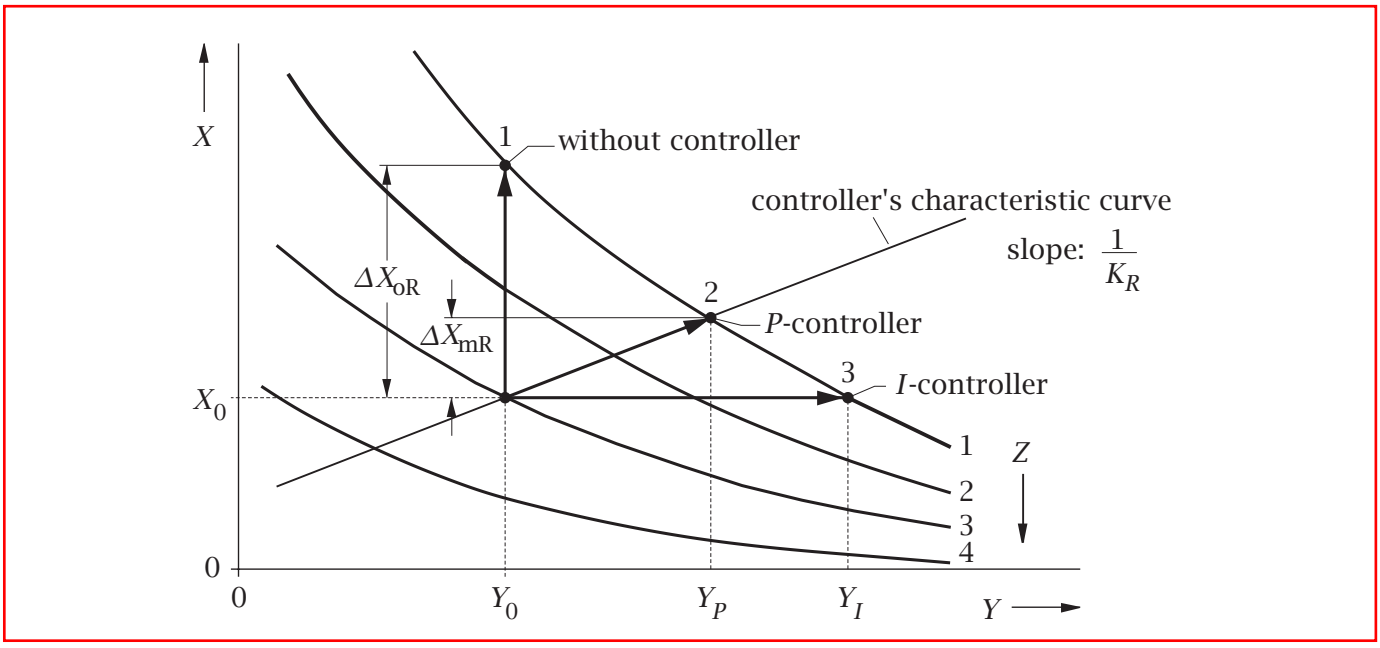


Figure 2-6: Characteristic curves of controlled system and controller

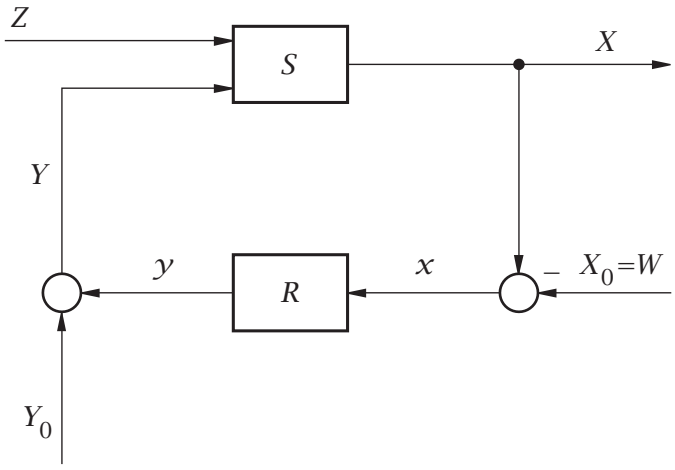


Figure 2-7: Control loop with non-linear controlled system
(here: negative transfer behaviour between the manipulated variable and the controlled variable)

and by rearranging

$$X = X_0 + \frac{1}{K_R}(Y - Y_0) \tag{2.32}$$

the equation of a straight line with slope $1/K_R$, which runs through point X_0, Y_0.

erating point belonging to $Z = 1$ is therefore the intersection 3 of line $X = X_0$ with the characteristic curve of the controlled system for $Z = 1$.

A comparison of the static behaviour of controlled systems with P- and I-controllers will obviously favour the controller with integrating behaviour. There are nevertheless many installations fitted with P-controllers because, apart from the static, also the dynamic characteristics and equipment costs play an important role in the selection of the controller. When generalizing, the behaviour of P- and I-controls can be characterized by the fact that P-controllers react quickly to disturbances and/or deviations, but cannot prevent steady state deviation, while the I-controllers do not permit steady state deviation, but react only slowly.

For characterizing the efficiency of a P-controller in the event of a disturbance, the control factor

$$R = \frac{\Delta X_{\mathrm{mR}}}{\Delta X_{\mathrm{oR}}} \tag{2.33}$$

is used. This factor is smaller than one in the case of control loops that make sense. Its value depends on the gain K_R of the controller and the characteristics of the controlled system. In the case of non-linear controlled systems, which can be described by a set of characteristic curves as shown in figure 2-6, the control factor is not a constant, but depends on the respective changes in the disturbance variable.

In the following we will consider the static behaviour of the closed loop control system described by a set of characteristic curves in figure 2-8 using a P-controller with different gains.

When comparing the sets of characteristic curves in figures 2-8 and 2-6, it is noticeable that the characteristic curves of the controlled system ascend in one case and in the other case descend. It is, for instance, possible to recognize (by entering the respective characteristic curve of the controller) that in figure 2-8 only descending characteristic curves of the controller and in figure 2-6 only ascending curves are suitable to reduce the steady state deviation with a controller to less than the deviation without a controller, i.e. the slope of the characteristic curve of the controller must have a sign different from that of the slope of the controlled system's characteristic curve.

values, the deviation variables are displayed in lower case letters, as illustrated in equation (2.3).

It may also be useful in certain instances to standardize the deviation variables by dividing them by appropriate reference values, e.g. maximum values, operating point values etc. Standardization with respect to operating point values results in

$$\tilde{y} = \frac{Y - Y_0}{Y_0} = \frac{y}{Y_0} \quad , \quad \tilde{u} = \frac{u}{U_0} \quad , \quad \dots \quad . \tag{2.4}$$

Standardization is usefully applied when incorporated dimensions do not provide additional clarity or safety. In most cases, however, standardized variables will not be indicated explicitly, as the standardization can be seen from the context.

Linearization replaces a given non-linear expression

$$Y = f(U, Z_1, Z_2, \dots) \tag{2.1}$$

in the vicinity of an operating point A with

$$Y = Y_0, \quad U = U_0, \quad Z_1 = Z_{10}, \quad Z_2 = Z_{20}, \dots \tag{2.5}$$

by a linear expression

$$y = K_u \cdot u + K_1 \cdot z_1 + K_2 \cdot z_2 + \dots \tag{2.6}$$

with the deviation variables $y, u, z_1, z_2, \dots$ and constants $K_u, K_1, K_2, \dots$.

The coefficients of the linear expression in equation (2.6) are obtained from a Taylor series expansion of the non-linear function equation (2.1). In practice it is necessary to determine the (partial) derivatives of the output variable with respect to the input variables, from which we obtain

$$y = \left[\frac{\partial Y}{\partial U}\right]_A u + \left[\frac{\partial Y}{\partial Z_1}\right]_A z_1 + \left[\frac{\partial Y}{\partial Z_2}\right]_A z_2 + \dots \tag{2.7}$$

and by comparing the coefficients

$$K_u = \left[\frac{\partial Y}{\partial U}\right]_A, \quad K_1 = \left[\frac{\partial Y}{\partial Z_1}\right]_A, \quad K_2 = \left[\frac{\partial Y}{\partial Z_2}\right]_A, \quad \dots \quad . \tag{2.8}$$

line $y = K \cdot u$ in the newly introduced system of coordinates of the deviation variables.

$$y = \left[\frac{\partial Y}{\partial U}\right]_A u = K \cdot u \tag{2.11}$$

with

$$K = \left[\frac{\partial Y}{\partial U}\right]_A \quad , \tag{2.12}$$

i.e. the slope of the replacement line equals the slope of the characteristic curve in the operating point.

Relationships between a dependent and two or more independent variables can be illustrated by a set of characteristic curves. For an approximate determination of the derivative of the dependent output variable with respect to an input variable represented by the parameters of characteristic curves, it is useful to use a suitable differential quotient.

Linearization of the set of characteristic curves illustrated in figure 2-4 provides the coefficients K_u and K_z of the linear equation for deviation variables

$$y = K_u \cdot u + K_z \cdot z \quad . \tag{2.13}$$

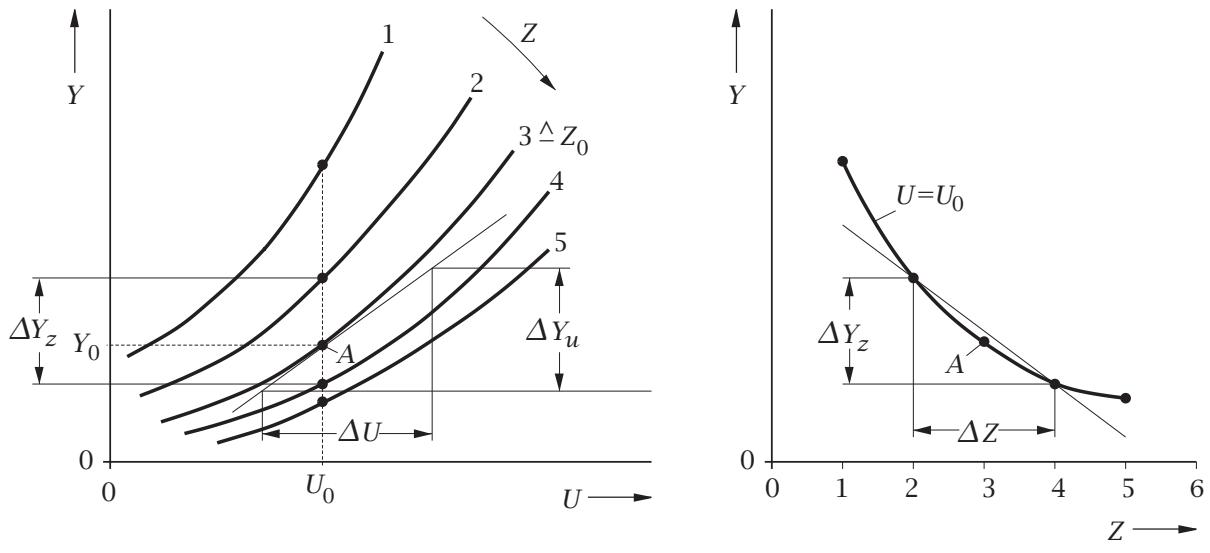


Figure 2-4: Linearization of a set of characteristic curves

the ship, this yaw rate does not change abruptly, even with an abrupt change in the moment, but approximately as displayed in the center block of figure 1-9. An abrupt change of the yaw rate, e.g. from zero to a constant value, results in an angle of rotation α, which can be of any value proportional to time; see the last block in figure 1-9.

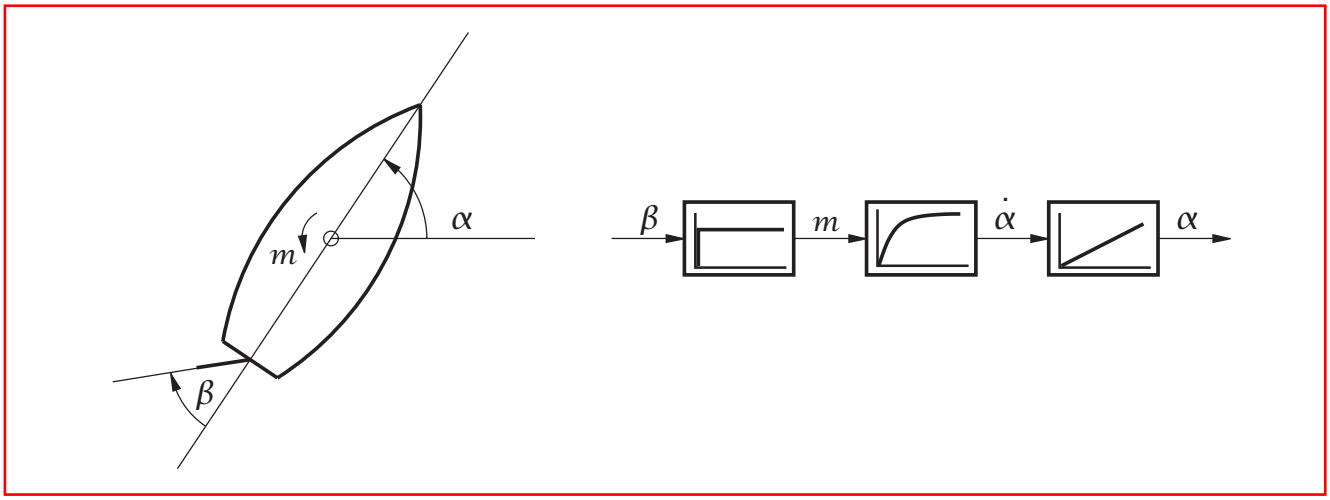


Figure 1-9: Functional diagram of a course steered by a ship

The functional diagram is one of the most important forms of illustration used for technical control tasks and solutions. Only correct and reliable functional diagrams will result in technically viable solutions. When preparing complicated functional diagrams, it is recommended emphatically to proceed, contrary to the method displayed in the example, against the direction of flow of the variables which, beginning with a certain variable, entails looking for the causes of changes in this variable and then to retain these results. This makes it easier to ensure that all influences affecting a variable are in fact detected.

The functional diagram is a schematic representation of the fundamental approach to and also the essential objective in automatic control, namely to provide tools to aid in the analysis, target-oriented influencing and design of dynamic technical systems with a complex structure. Such complex structures result, for example, from systems with several variables that interact with each other; they may also result from internal retroactions that frequently – but not exclusively – are due to control loops. In order to obtain a general methodology, i.e. one that can be applied to all specialized disciplines, the relevant analysis and design procedures in automatic control are based on a mathematical

The amplification principle and the superposition principle apply for linear differential equations and linear transfer systems. Both principles assume that the differential equation or the transfer system is provided with an input variable $u(t)$ and gives an output variable $y(t)$. The amplification principle states that an input variable $u(t)$ multiplied by a particular constant factor c gives an output variable $c \cdot y(t)$. The superposition principle deals with the case in which the input variable consists of several components $u(t) = u_1(t) + u_2(t) + \ldots$. It states that the associated output variable can be created in the same way, namely as $y(t) = y_1(t) + y_2(t) + \ldots$. Here $y_i(t)$ is the output variable linked with the input variable $u_i(t)$. Differential equations or transfer systems for which both principles apply are called linear.

3.2 Setting up differential equations

The general form of a linear differential equation with constant coefficients for an element with the input variable u and the output variable y is

$$a_n y^{(n)} + \ldots + a_2 \ddot{y} + a_1 \dot{y} + a_0 y = b_0 u + b_1 \dot{u} + \ldots + b_m u^{(m)}. \quad (3.1)$$

Real, physical-technical transfer elements will be described by means of differential equations, in which the order n of the highest occurring derivative of the output variable is higher than the order m of the highest occurring derivative of the input variable.

In the case of most signal transferring arrangements it is essential that the effects of the storage of material or energy are considered. In the setting up of differential equations for complex relationships a modular procedure is recommended, such as

1. storages are identified and described with suitable basic equations (thus creating subsystems);

2. connections are identified and described (thus describing the combination of subsystems);

3. subsystems are combined by means of connecting equations and elimination of unnecessary internal variables.

3.5 Laplace transform

3.5.1 Transformation of time functions

We can make use of the Laplace transform for the solution of linear differential equations with constant coefficients for given initial conditions and for input variables $u(t)$ which, for the negative value of the argument t, read zero. Many differential equations, which are to be solved in relation to control engineering problems, fulfil the preconditions for a solution using the Laplace transform.

The Laplace transform reversibly and unambiguously allocates another function $F(s)$ in the frequency domain to a function $f(t)$ in the time domain (original domain). The special form of the transform allows the differentiation and integration operations on time functions to be transformed into algebraic operations on the associated transformed functions so that the transformed differential equations read like algebraic equations which can be rearranged and combined with each other more easily than the original equations. As many tasks can be carried out without a relatively extensive knowledge of the theory of the Laplace transform and with the help of so-called correspondence tables, we will present only a short outline of the procedures to be adopted. For more detailed questions please refer to the relevant literature.

The relationship between original and transformed functions will be established by means of the equations

$$F(s) = \int_{-0}^{\infty} f(t) \cdot e^{-st} \mathrm{d}t = \mathscr{L}\{f(t)\} \tag{3.34}$$

$$f(t) = \left\{ \begin{array}{ll} \dfrac{1}{2\pi j} \displaystyle\int_{\alpha - j\infty}^{\alpha + j\infty} F(s) \cdot e^{st} \mathrm{d}s & \text{for} \quad t \geq 0 \\ 0 & \text{for} \quad t < 0 \end{array} \right\} = \mathscr{L}^{-1}\{F(s)\} \tag{3.35}$$

in an unambiguous and reversible manner. Within this $s = \sigma + j\omega$ is a complex variable with a positive real part and α a positive constant, which is to be selected such that the integral in equation (3.35) will converge. The lower integration limit -0 means that a possibly occuring discontinuity in $f(t)$ at $t = 0$ is included into the integration.

$F(s)$	$f(t)$ for $t > 0$	$(f(t) = 0$ for $t \leq 0)$
$\frac{1}{(s-s_p)^n}$	$\frac{1}{(n-1)!} t^{n-1} e^{s_p t}$	$n = 1, 2, 3, \ldots$
1	$\delta(t)$	
$\frac{1}{s}$	$1(t)$	
$\frac{1}{s^2}$	t	
$\frac{1}{1+sT}$	$\frac{1}{T} e^{-t/T}$	
$\frac{\omega_0^2}{s^2 + 2D\omega_0 s + \omega_0^2}$	$\frac{\omega_0}{\sqrt{1-D^2}} e^{-D\omega_0 t} \sin\left(\sqrt{1-D^2}\ \omega_0 t\right)$	$\lvert D\rvert < 1$
	$\omega_0^2 t e^{-D\omega_0 t}$	$\lvert D\rvert = 1$
	$\frac{\omega_0}{\sqrt{D^2-1}} e^{-D\omega_0 t} \sinh\left(\sqrt{D^2-1}\ \omega_0 t\right)$	$\lvert D\rvert > 1$
$\frac{1}{(1+sT_1)(1+sT_2)}$	$\frac{1}{T_1 - T_2}\left(e^{-t/T_1} - e^{-t/T_2}\right)$	$T_1 \neq T_2$
$\frac{s}{1+sT}$	$\frac{1}{T}\left(\delta(t) - \frac{1}{T} e^{-t/T}\right)$	
$\frac{s}{(1+sT_1)(1+sT_2)}$	$\frac{1}{T_1 T_2 (T_1 - T_2)}\left(T_1 e^{-t/T_2} - T_2 e^{-t/T_1}\right)$	$T_1 \neq T_2$
$\frac{s\omega_0^2}{s^2 + 2D\omega_0 s + \omega_0^2}$	$\omega_0^2 e^{-D\omega_0 t}\left(\cos\omega_D t - \frac{D}{\sqrt{1-D^2}} \sin\omega_D t\right)$	$\lvert D\rvert < 1$ $\omega_D = \sqrt{1-D^2}\ \omega_0$
$\frac{1}{s(1+sT)}$	$1 - e^{-t/T}$	
$\frac{1}{s(1+sT_1)(1+sT_2)}$	$1 - \frac{1}{T_1 - T_2}\left(T_1 e^{-t/T_1} - T_2 e^{-t/T_2}\right)$	$T_1 \neq T_2$
$\frac{\omega_0^2}{s(s^2 + 2D\omega_0 s + \omega_0^2)}$	$1 - e^{-D\omega_0 t}\left(\cos\omega_D t + \frac{D}{\sqrt{1-D^2}} \sin\omega_D t\right)$	$\lvert D\rvert < 1$ $\omega_D = \sqrt{1-D^2}\,\omega_0$

Table 3-1: Correspondence tables $F(s) \bullet\!\!-\!\!\circ f(t)$

The inversion

$$g(t) = \frac{\mathrm{d}h(t)}{\mathrm{d}t} \tag{3.73}$$

applies with certain restrictions with regard to the differentiability of the unit step response.

To determine the characteristics of control loop elements and control loops, we will work mainly with the unit step response (see also the presentations in the functional diagrams). The unit impulse response will mainly be used for more theoretical considerations.

3.7 Transfer function

It has already been shown in section 3.5 that the interrelation of the transformed solution of a differential equation with the transformed function of the input variable is multiplicative. In the case of disappearing initial conditions, we always obtain a solution in the frequency domain in the form of

$$Y(s) = G(s) \cdot U(s) \quad . \tag{3.74}$$

Here, $Y(s)$ and $U(s)$ are the Laplace transforms of the corresponding variables. $G(s)$ is a function which is exclusively determined by the differential equation. It is referred to as transfer function, because it describes how the $U(s)$ values are converted into $Y(s)$ values, i.e. how an input value is transferred by the function of the described transfer element into an output. It is very useful that the resulting transfer function of an arbitrary number of transfer elements, arranged in series, is the product of the transfer functions of the individual elements. As it is generally much easier to multiply transformed functions with each other, rather than combine differential equations, we are opening up an easily accessible method of describing the dynamic behaviour of an arrangement of transfer elements in series.

From the differential equation

$$a_n y^{(n)} + \ldots + a_1 \dot{y} + a_0 y = b_0 u + b_1 \dot{u} + \ldots + b_m u^{(m)} \tag{3.75}$$

we obtain, by Laplace transforming both sides and with disappearing initial conditions

$$a_n s^n Y(s) + \ldots + a_1 s Y(s) + a_0 Y(s) = b_0 U(s) + b_1 s U(s) + \ldots + b_m s^m U(s) \tag{3.76}$$

and from that, by rearranging

$$Y(s)(a_n s^n + \ldots + a_1 s + a_0) = U(s)(b_0 + b_1 s + \ldots + b_m s^m). \tag{3.77}$$

This enables us to obtain the transfer function as the quotient

$$\frac{Y(s)}{U(s)} = \frac{b_m s^m + \ldots + b_1 s + b_0}{a_n s^n + \ldots + a_1 s + a_0} = \frac{Z(s)}{N(s)} = G(s) \quad . \tag{3.78}$$

We can now recognize that the transfer function is a rational function of the variable s and that it contains all the coefficients of the differential equation. It therefore describes the relationship between the input and output variables just as well as the differential equation.

While within the following section 3.9 the frequency responses involved can be presented in graphic form without any particular difficulty since they are functions of a single real variable, it is difficult to represent the tranfer functions graphically despite their extensive similarity with frequency responses because they depend on the complex variable $s = \sigma + j\omega$. A frequently used method of representation for rational transfer functions is the use of the roots of the numerator polynomial $Z(s)$ and the denominator polynomial $N(s)$.

According to Viëta's theorem, any polynomial can be expressed through its roots s_i and the coefficient a_n of the highest power of the variable s.

$$a_n s^n + \ldots + a_1 s + a_0 = a_n \cdot (s - s_1) \cdot (s - s_2) \ldots (s - s_n) \quad . \tag{3.79}$$

This permits to rewrite a rational transfer function in accordance with equation (3.75) in the form

$$G(s) = K \cdot \frac{(s - s_{N1}) \cdot (s - s_{N2}) \ldots (s - s_{Nm})}{(s - s_{P1}) \cdot (s - s_{P2}) \ldots (s - s_{Pn})} \tag{3.80}$$

with s_{Ni} as the roots of the numerator polynomial $Z(s)$ and s_{Pi} as the roots of the denominator polynomial $N(s)$.

3.9 Frequency response

3.9.1 General

We can obtain the frequency response from the transfer function $G(s)$ through a relatively simple formal step. We only need to replace the complex variable $s = \sigma + j\omega$ with the imaginary variable $j\omega$, e.g. by letting the real part σ of the variable s become zero. From the previously used

$$Y(s) = G(s) \cdot U(s) \tag{3.74}$$

we obtain

$$Y(j\omega) = G(j\omega) \cdot U(j\omega) \tag{3.82}$$

valid for the frequency response $G(j\omega)$.

Consideration of the fact that the functions of s are defined in the whole s plane will help in the interpretation of this expression; functions of $j\omega$, in particular the frequency response $G(j\omega)$, are only defined on the imaginary axis of the s plane. Although the definition range of the frequency response and the associated transformed functions of the input and output variables in comparison to the transfer function is clearly smaller, the frequency response describes the behaviour of transfer systems just as comprehensively as the transfer function.

The transformed functions of the variables in equation (3.82) can be obtained through the Fourier transform of the corresponding time function. The Fourier transform is, like the Laplace transform, an integral transformation, which creates a one-to-one correspondence between the time domain and the frequency domain.

Because of the major practical significance of the frequency response, a less formal approach to this means of description for dynamic systems will be offered in the next section.

3.9.2 Frequency response and differential equation

We can obtain a plausible relationship between the differential equation and the associated frequency response from the answer to a question

3.9.4 Measurement of frequency responses

Contrary to the transfer function, the frequency response of a transfer system can be measured. Therefore, analogously to the procedures in section 3.9.2 , the system in question will be excited with a harmonic input variable. Then the resulting output variable, having reached stationarity, is compared with the input variable.

It follows from the solution of the linear differential equation derived in section 3.9.2 that a linear transfer system responds to excitation by a harmonic (i.e. sine or cosine shaped) input variable

$$u(t) = U \cdot \cos(\omega t + \varphi_u) \quad , \quad \underline{u} = U e^{j\varphi_u} \tag{3.114}$$

with a harmonic output variable

$$y(t) = Y \cdot \cos(\omega t + \varphi_y) \quad , \quad \underline{y} = Y e^{j\varphi_y} \tag{3.115}$$

of the same frequency. The amplitude and the phase angle of the output variable are generally different from those of the input variable. Both variables can be described by their phasors $\underline{u}$ resp. $\underline{y}$. Figure 3-10 shows an extract from a graphic representation of both variables.

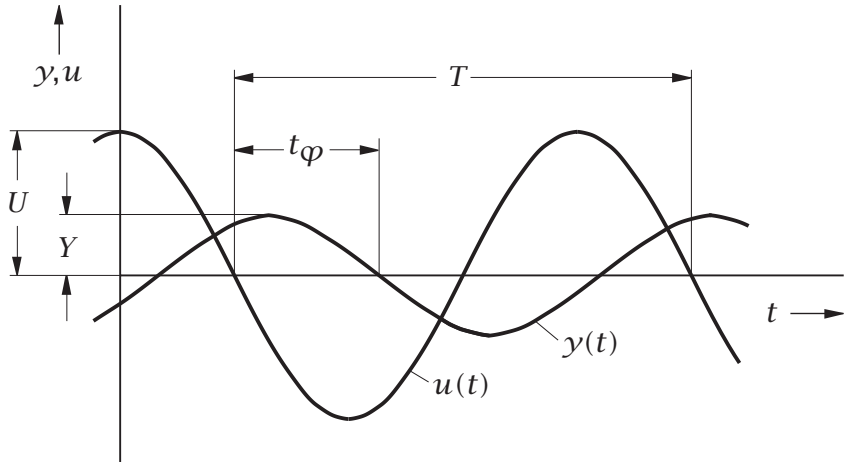


Figure 3-10: Harmonic input and output variables

As we are able to represent a complex value like the frequency response at a particular frequency by the absolute value and the phase angle

$$G(j\omega) = |G(j\omega)| \cdot e^{j\varphi} \tag{3.116}$$

and since the frequency response is defined as

$$G(j\omega) = \frac{\underline{y}}{\underline{u}} = \frac{Y}{U} \cdot e^{j(\varphi_y - \varphi_u)} \quad ,$$

we obtain the absolute value of the frequency response as the quotient of the amplitudes

$$|G(j\omega)| = \frac{Y}{U} \tag{3.117}$$

and the phase angle of the frequency response from the phase shift

$$\varphi(j\omega) = \varphi_y - \varphi_u = -\frac{t_\varphi}{T} \cdot 360° \quad . \tag{3.118}$$

We note that t_φ is the time interval between a zero intersection of the input variable and the correspondent codirectional zero intersection of the output variable. In most technical systems the output variable follows the input variable, so that, based on the situation presented in figure 3-10, the resulting phase angle has a negative value. From this it follows that $t_\varphi > 0$ for $\varphi(j\omega) < 0$ and vice versa.

For a complete description of a transfer system it is necessary to determine its frequency response for a variety of frequency values. The results of such measurements can be presented as a table of values or in graphic form.

3.9.5 Nyquist plot of frequency responses

As with almost all functions, frequency responses, which are functions of the frequency, can be represented either as analytic expressions, as value tables or in graphic form. From the multitude of possibilities for graphic representation the illustrations most frequently used in control engineering are the Nyquist plot and the logarithmic representation in the so-called Bode diagram.

The Nyquist plot of a frequency response is a line in a complex plane which connects points representing the values of real and imaginary parts of the frequency response for specific values of the frequency (figure 3-11). These points can also be considered to be end points of phasors, which represent the frequency response, or regarded as

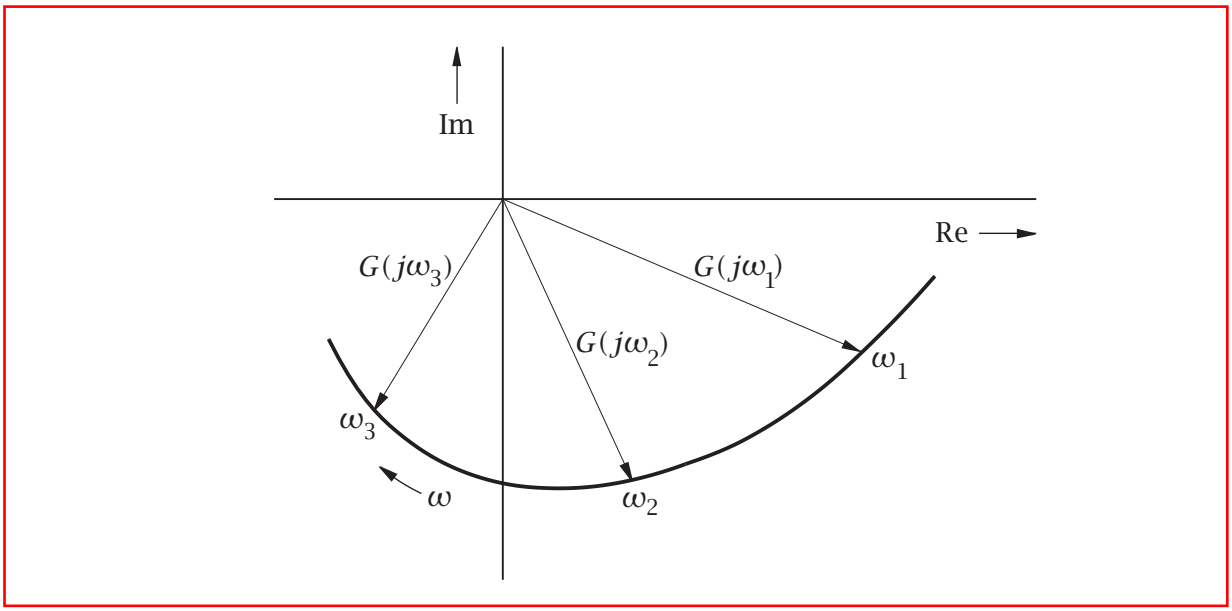


Figure 3-11: Nyquist plot of a frequency response

phasors of the output variable $\underline{y}$ in the case where the input variable is described by means of the unit phasor $\underline{u} = 1$; in this case $\underline{y} = G(j\omega)$.

Nyquist plots of frequency responses are normally drawn as lines with frequency parameterization, but at least with a specification of the direction of the increasing frequency. Some examples are presented in table 3-5.

We can recognize that the Nyquist plot of the frequency response for a P-element degenerates to a point and that the Nyquist plots pertaining to the integrators or differentiators cover the (negative or positive) imaginary axis.

The frequency response of the lag element is not so easily represented graphically. The mathematics show that all complex functions of the form

$$A(\omega) = \frac{j\omega \cdot a + b}{j\omega \cdot c + d} \tag{3.119}$$

for any real values of the coefficients a, b, c, d form circles in the complex plane, whose center points lie on the real axis. By this fact and the limit values of the frequency response for large and small values of the frequency the Nyquist plot of such first order elements is characterized.

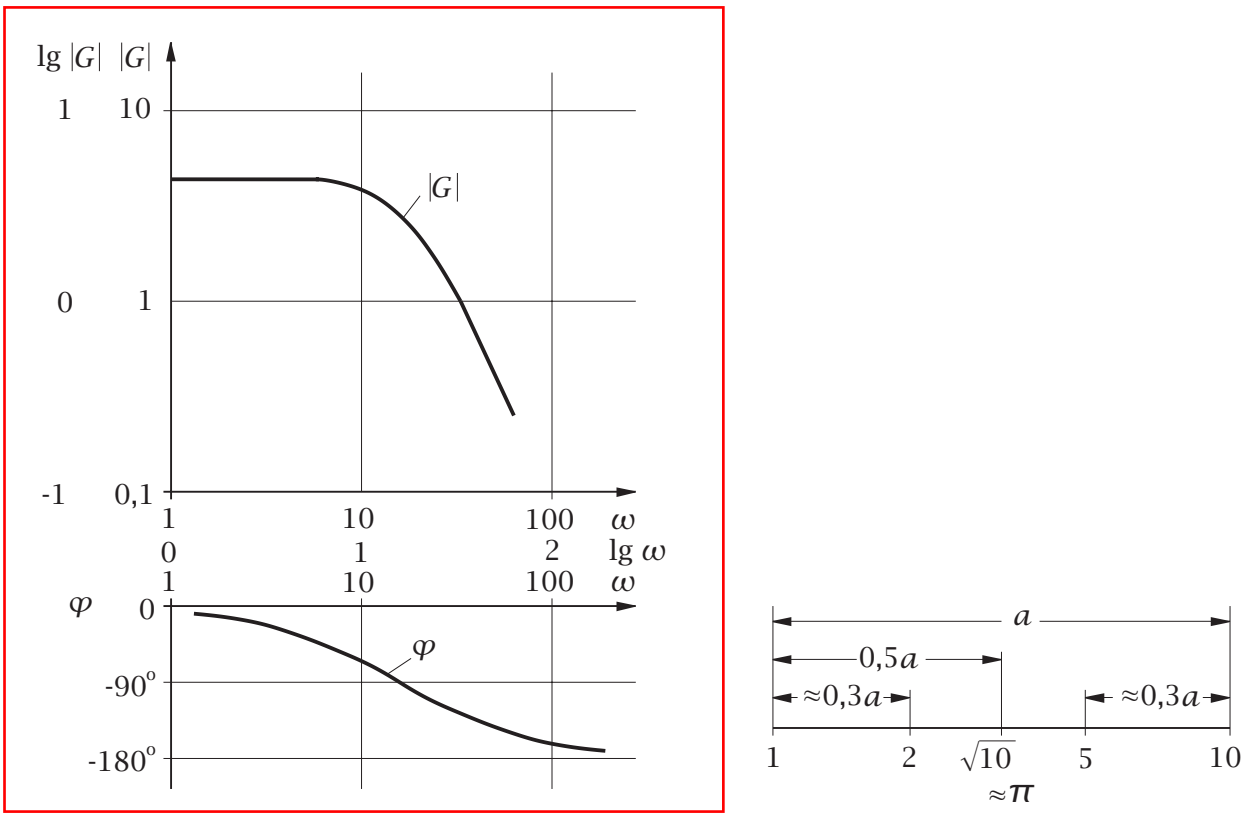


Figure 3-12: Bode diagram and logarithmic scale

The representation of frequency responses in Bode diagrams has substantial advantages over the Nyquist plot representation. For a wide range of frequency responses construction rules can be used, providing sufficiently accurate representations without extensive calculation work. Additionally, the frequently required multiplication of frequency responses can be carried out fairly easily in the Bode diagram.

As the frequency response is represented by the amplitude and the phase angle

$$G(j\omega) = |G| \cdot e^{j\varphi} \tag{3.122}$$

for the product of two frequency responses G_1 and G_2 applies

$$G = G_1 G_2 = |G_1| \cdot e^{j\varphi_1} \cdot |G_2| \cdot e^{j\varphi_2} = |G_1| \cdot |G_2| e^{j(\varphi_1+\varphi_2)} \tag{3.123}$$

and, therefore, the amplitude of G is

$$|G| = |G_1| \cdot |G_2| \tag{3.124}$$

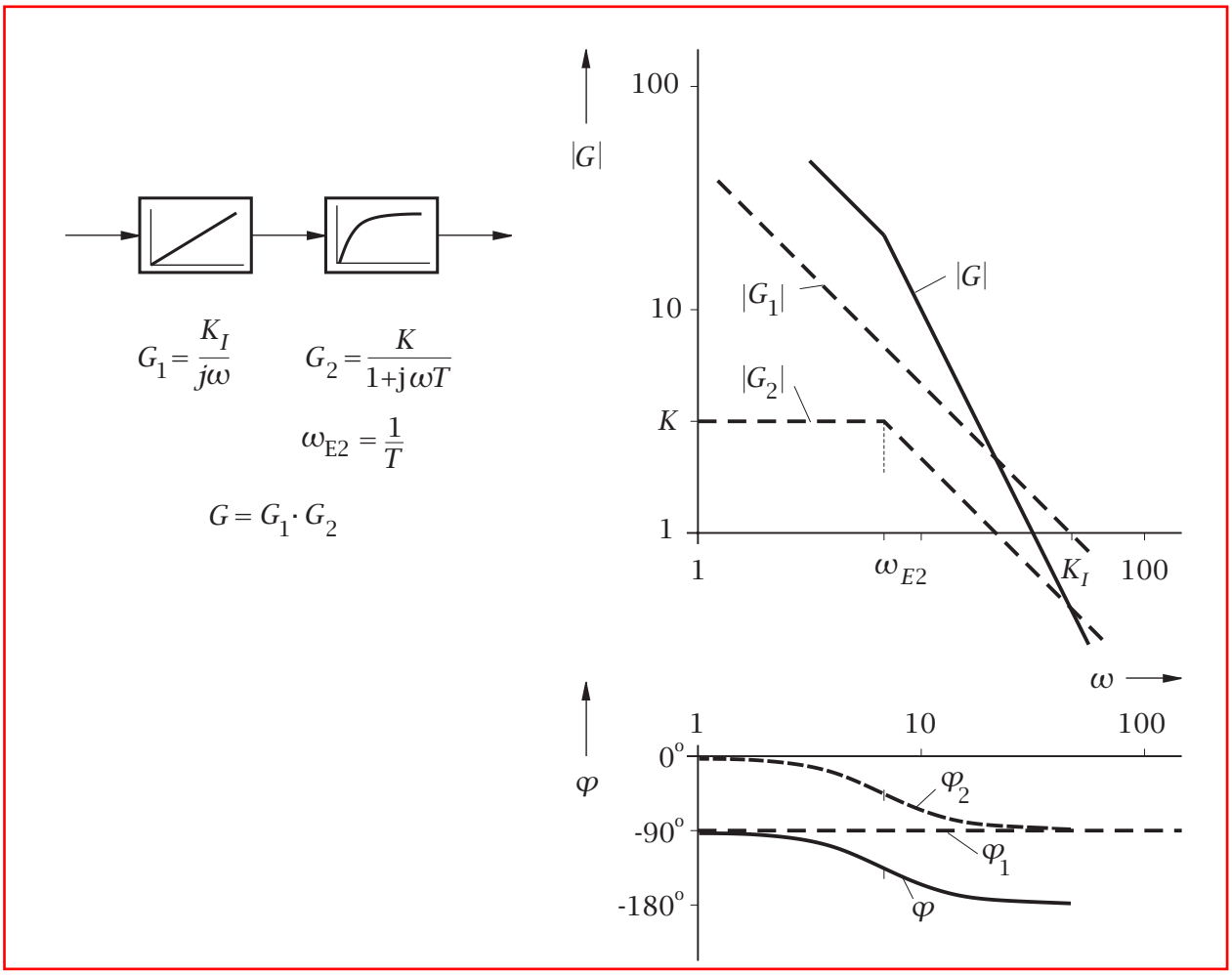
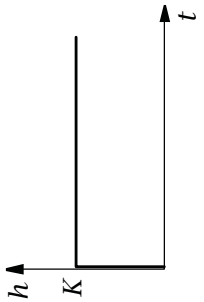
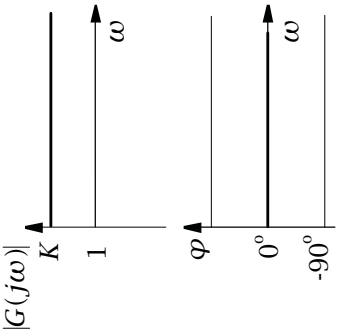
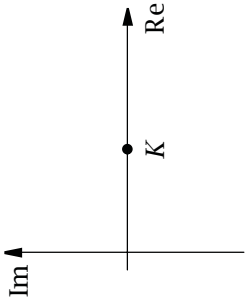
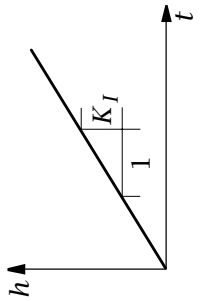
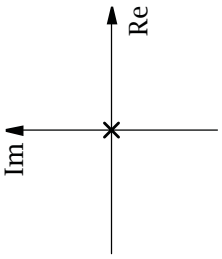
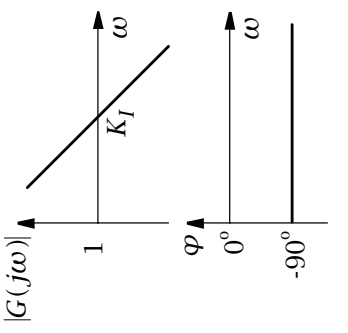
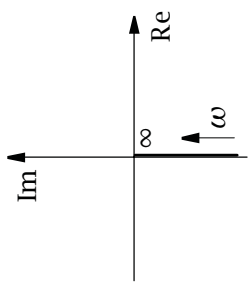
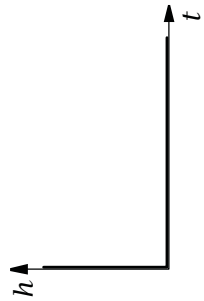
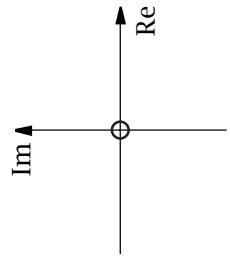
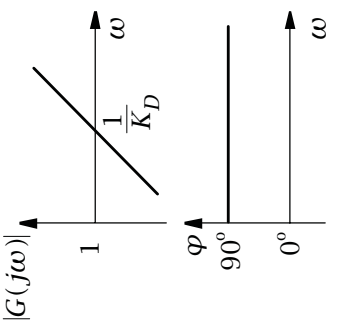
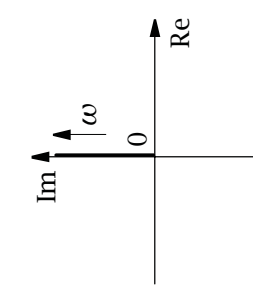


Figure 3-13: Multiplication of frequency responses

3.10 Computing rules for frequency responses and transfer functions

For many control engineering purposes it is necessary to know the frequency response or transfer function of a particular arrangement of signal transfer elements. The connection of several elements is treated as a single transfer element with one input and one output, of which the output variable phasor is associated with the phasor of the input variable by the resulting frequency response of the arrangement (figure 3-14).

In table 3-7 the frequency responses of the three most important con-

Term	Differential equation and unit step response	Transfer function and pole / zero plot	Bode diagram amplitude and phase response	Nyquist plot
P	$y=Ku$ 	$G(s)=K$		
I	$y=K_I\int u dt$ 	$G(s)=\frac{K_I}{s}$ 		
D	$y=K_D\dot{u}$ 	$G(s)=K_D s$ 		

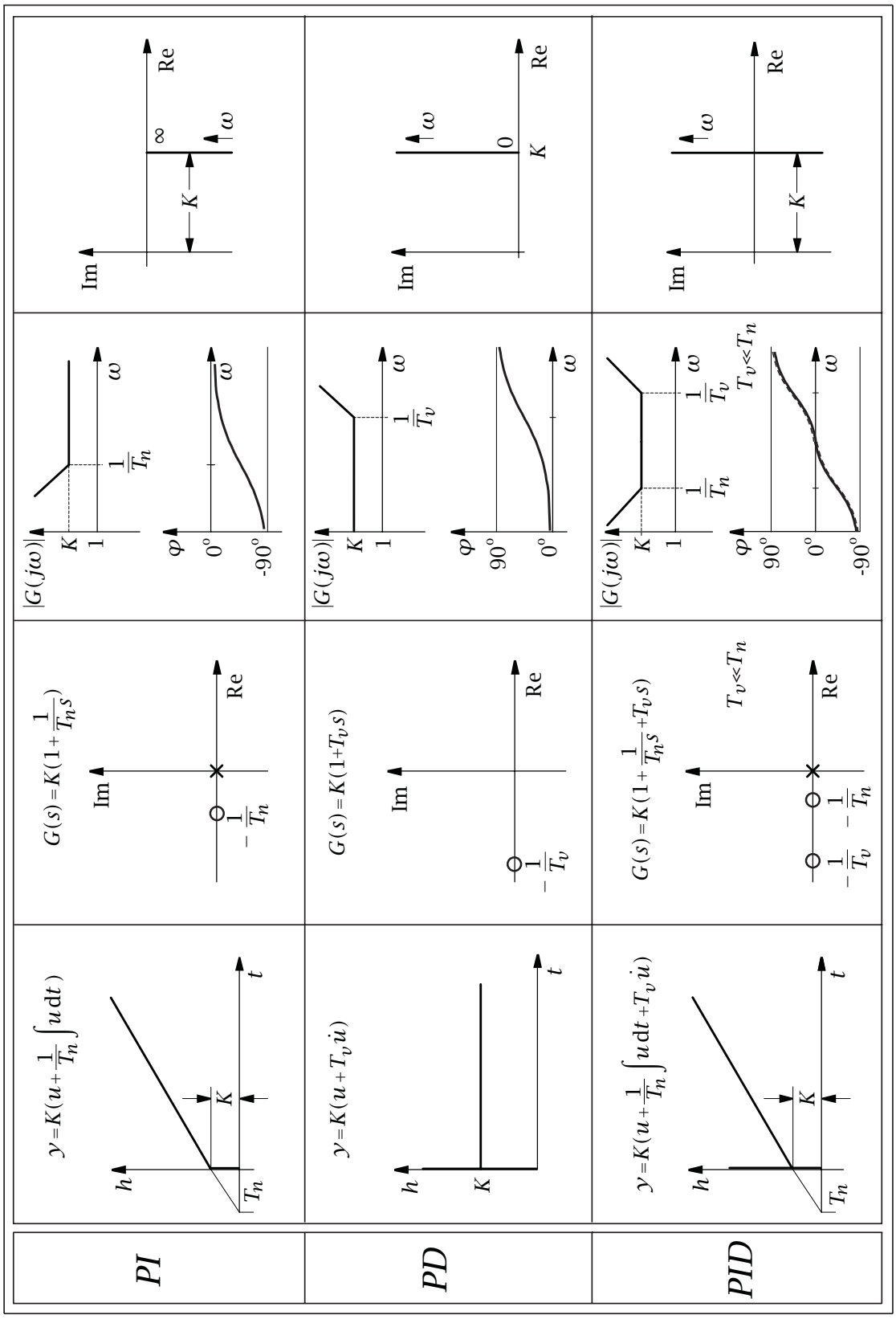
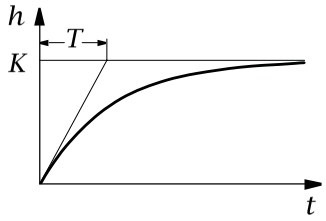
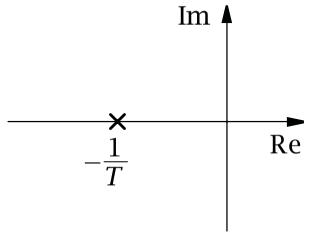
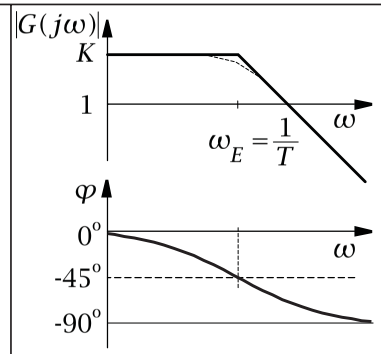
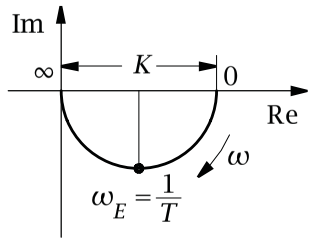
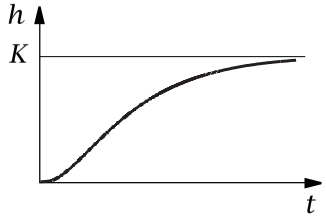
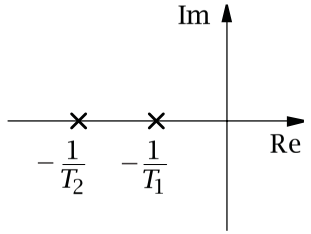
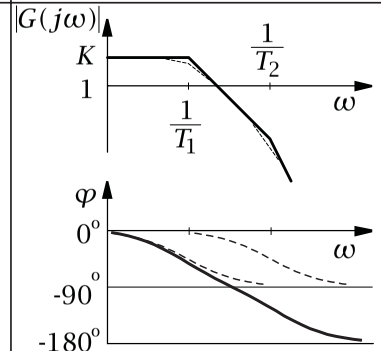
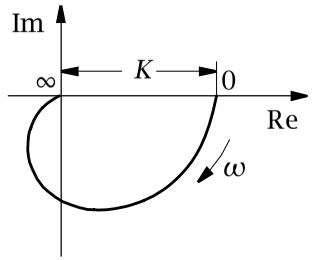
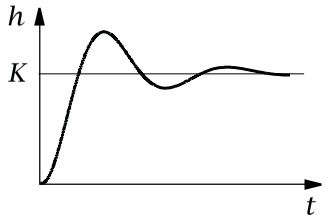
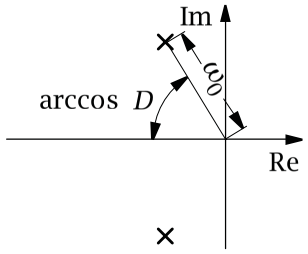
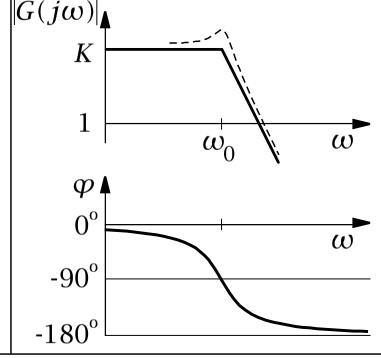
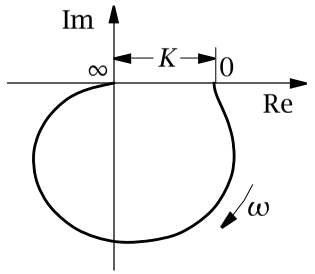


Table 4-2: Control loop elements

Term	Differential equation and unit step response	Transfer function and pole / zero plot	Bode diagram amplitude and phase response	Nyquist plot
PT_1	$T\dot{y} + y = K u$ 	$G(s) = \frac{K}{1+Ts}$ 		
PT_2 ($D \geq 1$)	$T_1 T_2 \ddot{y} + (T_1 + T_2)\dot{y} + y = Ku$ 	$G(s) = \frac{K}{T_1 T_2 s^2 + (T_1+T_2)s + 1}$ 		
PT_2 ($D < 1$)	$\ddot{y} + 2D\omega_0 \dot{y} + \omega_0^2 y = \omega_0^2 Ku$ 	$G(s) = \frac{\omega_0^2 K}{s^2 + 2D\omega_0 s + \omega_0^2}$ 		

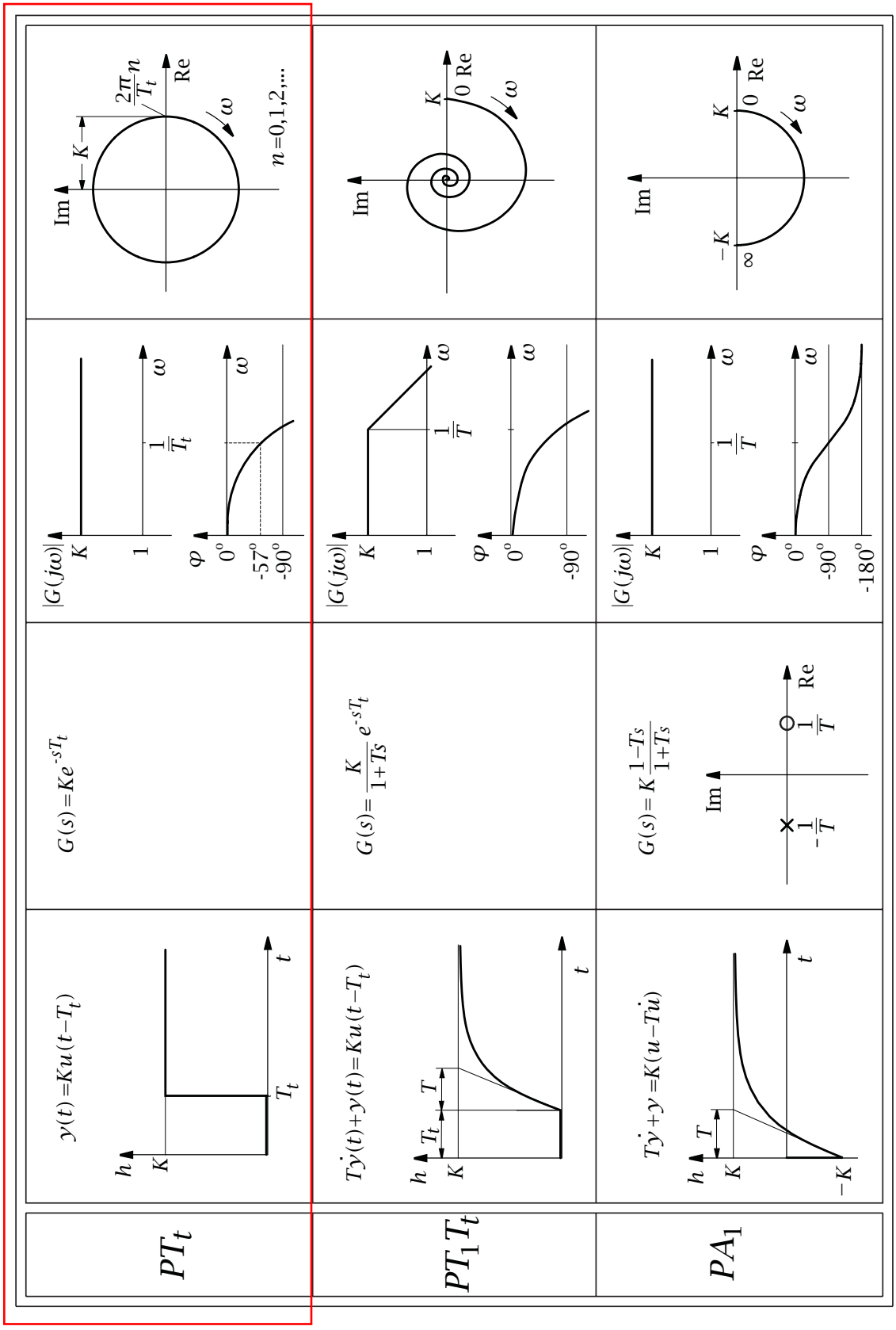


Table 4-4: Control loop elements

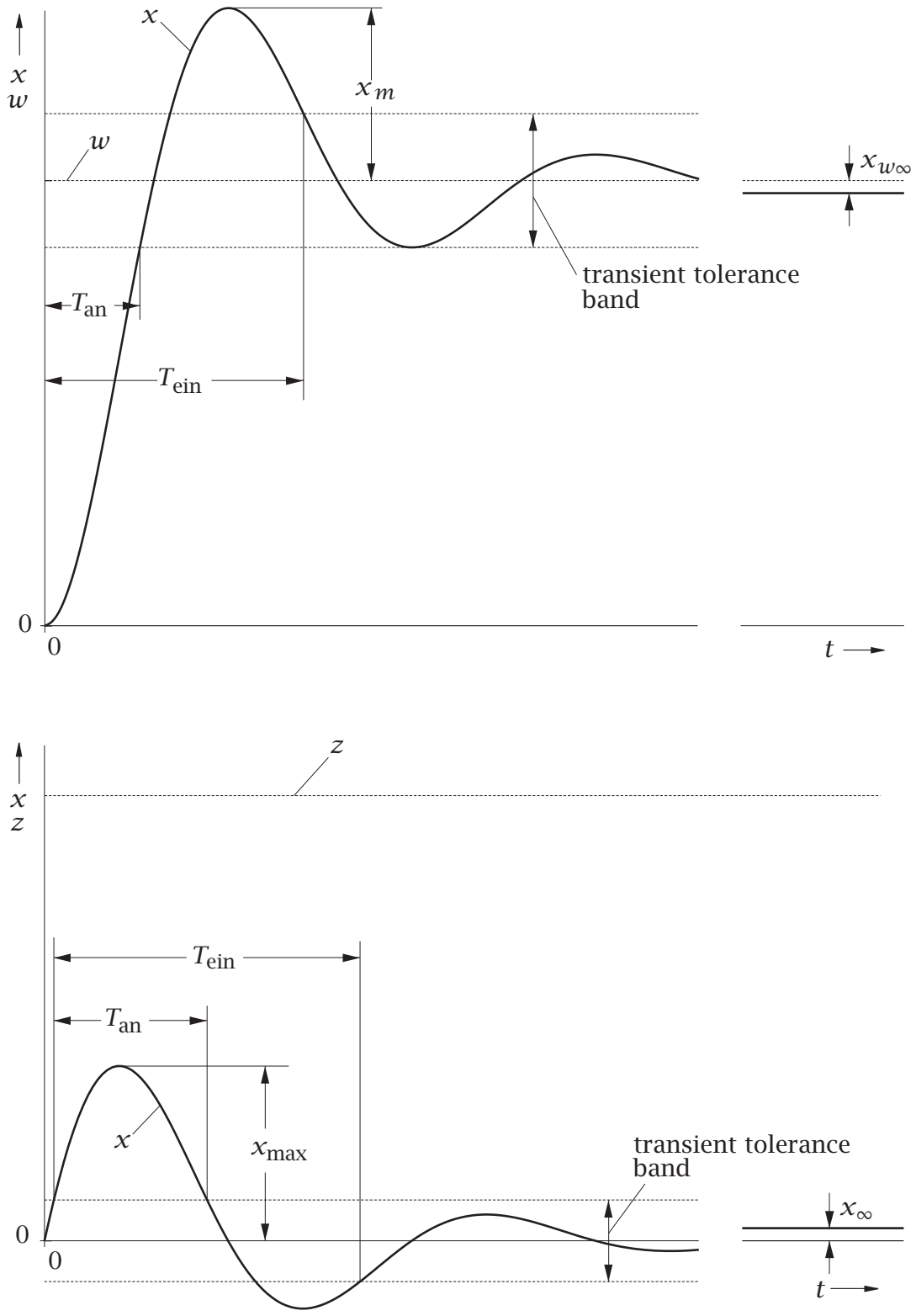


Figure 5-14: Step response to changes in the reference (above) or disturbance variable (below)

5-2 , we obtain recommendable controller settings from the steady state gain $K_{R\,\mathrm{krit}}$ so obtained and the resulting period of oscillation T_{krit}. This procedure is particularly suited for controls with unknown measuring and actuator devices because their characteristics are also included in the oscillation test. We can, however, also obtain $K_{R\,\mathrm{krit}}$ and T_{krit} from an analysis of the frequency response without actual operational trials.

For the control loop according to figure 5-4 and using a P-controller, we obtain from figure 5-5 a $K_{R\,\mathrm{krit}} = 8$ and from table 5-2 the recommendation $K_R = 4$.

Controller	K_R	T_n	T_v
P	$0{,}5 \cdot K_{R\,\mathrm{krit}}$	-	-
PI	$0{,}45 \cdot K_{R\,\mathrm{krit}}$	$0{,}85 \cdot T_{\mathrm{krit}}$	-
PID	$0{,}6 \cdot K_{R\,\mathrm{krit}}$	$0{,}5 \cdot T_{\mathrm{krit}}$	$0{,}12 \cdot T_{\mathrm{krit}}$

Table 5-2: Controller settings according to an oscillation test

Stability

5.5 Algebraic stability criteria

In order to be technically useful a control system must be stable. For this reason methods and tools to test the stability of dynamic systems are an important part of control engineering. Control theory defines different types of stability, of which only the so-called transfer stability will be dealt with here and regarded quite simply as the stability. Following the Anglo-American terminology, this type of stability will also be referred to as BIBO stability (Bounded Input-Bounded Output). It is based on the requirement that a stable system must respond with a bounded output variable to any bounded input variable.

Control theory has developed tools (stability criteria) with which, from the description of the dynamic behaviour of the system, conclusions can be drawn regarding its stability. Such criteria evaluate differential equations, transfer functions and frequency responses describing the system and avoid the determination of special time functions, e.g. those of the controlled variable.

For all linear systems, and therefore also for all control loops containing exclusively linear elements, the superposition principle applies and it follows from this that the stability of such a system is a property which is not dependent on the input variables of the system. For a test on stability it is sufficient, therefore, to investigate the solution of the homogeneous differential equation of the system.

It has been shown in section 3.4 that the solution of a linear differential equation with constant coefficients

$$a_n x^{(n)} + \ldots + a_1 \dot{x} + a_0 x = x_e \tag{5.19}$$

consists of the solution of the homogeneous differential equation and a particular solution. The solution of the homogeneous differential equation is thereby in the form of

$$x_h(t) = C_1 e^{\lambda_1 t} + C_2 e^{\lambda_2 t} + \ldots + C_n e^{\lambda_n t} \tag{5.20}$$

with λ_i, $(i = 1, \ldots, n)$ as the roots of the characteristic polynomial

$$a_n \lambda^n + \ldots + a_1 \lambda + a_0 = 0 \quad . \tag{5.21}$$

Among the previously considered control loop elements only the delay elements result in control loops with differential equations that do not fit into this pattern. Such control loops can therefore not be checked for stability using the algebraic criteria described in the following text. It will be shown that stability criteria which evaluate the frequency response of the open control loop, e.g. the Nyquist criterion, are applicable also in this case.

We can, with the help of the convolution integral, show that the solution (5.20) of the homogeneous differential equation for $t \to \infty$ must approach zero, in order that the associated system shows transfer stability. In view of the particular characteristics of the exponential function, this means that each of the summands of $x_h(t)$ must tend to zero for large time values, i.e.

$$\lim_{t \to \infty} C_i e^{\lambda_i t} = 0 \quad , \quad i = 1, \ldots, n \tag{5.22}$$

must be fulfilled if the system is stable. From this follows the requirement

$$\mathrm{Re}\, \lambda_i < 0 \quad , \quad i = 1, \ldots, n \quad . \tag{5.23}$$

If this requirement is violated by even one real root of the characteristic polynomial being positive, then the solution of the differential equation contains a part which monotonously increases with time above all limits. Such a system is described as monotonously unstable. A conjugate complex pair of roots with a positive real part among the λ_is results in an oscillating part in the solution with a constant frequency and an amplitude monotonously increasing beyond all bounds; the system is therefore oscillatory unstable. In the case that one or more roots differing from each other exhibit a vanishing real part, the solution contains parts which neither increase nor decay. If all other roots then have negative real parts, the system is at the limit of stability. Conjugate complex pairs of roots with real parts tending to zero lead to sustained oscillations in the solution. Figure 5-16 shows roots of the characteristic polynomial in the complex plane as well as the associated parts in the solution of the homogeneous differential equation belonging to the real roots and conjugate complex pairs of roots. Systems at the limit of stability are not stable in the sense of the definition of the transfer stability (BIBO) because at least one bounded input may be found which leads to an unbounded output.

Based on these considerations, we arrive at a first stability criterion, namely

> a transfer system is stable only if all the roots of the characteristic polynomial belonging to its differential equation have negative real parts.

As the poles of the transfer function and the roots of the characteristic polynomial are identical, the above statement also applies with regard to the real part of the poles of the transfer function.

For control loops described by means of higher-order differential equations, application of the above-mentioned criterion means that the roots of a higher-degree polynomial must be determined. As this often can only be done numerically, algebraic criteria have been developed which not only circumvent the explicit solution of the differential equation but also that of the characteristic polynomial.

With the aid of algebraic stability criteria we can check whether all the roots of a polynomial, in this case those of the characteristic polyno-

Based on the homogeneous differential equation of the system considered (in this case that of the closed control loop)

$$a_n x^{(n)} + \ldots + a_1 \dot{x} + a_0 x = 0 \tag{5.24}$$

a necessary condition common to both criteria states:

1st condition: The system is only stable when all the coefficients $a_n \ldots a_0$ are present and positive.

In the case that all coefficients are negative, the condition can be satisfied by multiplying the differential equation with -1.

Necessary and sufficient, according to the Hurwitz criterion, is the

2nd condition: The system is only stable when the Hurwitz determinant and its subdeterminants (formed in line with the scheme of equation (5.25)) are all greater than zero.

$$\begin{vmatrix} a_{n-1} & a_{n-3} & a_{n-5} & \cdots & \cdots & \cdots & 0 \\ a_n & a_{n-2} & a_{n-4} & \cdots & \cdots & \cdots & 0 \\ 0 & a_{n-1} & a_{n-3} & \cdots & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & & & \\ \vdots & \vdots & \vdots & & a_3 & a_1 & 0 \\ \vdots & \vdots & \vdots & & a_4 & a_2 & a_0 \\ 0 & 0 & 0 & & a_5 & a_3 & a_1 \end{vmatrix} = H \tag{5.25}$$

The rule for creating the Hurwitz determinant in equation (5.25) can be described in such a way that the coefficients $a_{n-1},\ldots,a_1$ are put in their natural sequence on the main diagonal and the columns are filled in such a way that the coefficients are arranged in increasing order of their indices from top to bottom. Missing coefficients are represented by zeros. The definition of the subdeterminants results from the diagram.

Necessary and sufficient for the Routh criterion we have the

2nd condition: The system is only stable, if the Routh test functions R_i are all greater than zero.

The test functions will be determined by equation (5.26) (or similar schemes). To this end, the coefficients of the differential equation are arranged in two rows. The next row is then always computed from the preceding two rows. This procedure is to be continued until we have two rows with, in each case, only one element. The elements of the first column of this scheme are the Routh test functions $R_n, \ldots, R_0$ (equation (5.27)).

$$
\begin{array}{lll}
a_n & a_{n-2} & a_{n-4} \\
a_{n-1} & a_{n-3} & a_{n-5} \\
\underbrace{a_{n-2} - \dfrac{a_n}{a_{n-1}} a_{n-3}}_{a'_{n-2}} & \underbrace{a_{n-4} - \dfrac{a_n}{a_{n-1}} a_{n-5}}_{a'_{n-4}} & \underbrace{a_{n-6} - \dfrac{a_n}{a_{n-1}} a_{n-7}}_{a'_{n-6}} \\
a_{n-3} - \dfrac{a_{n-1}}{a'_{n-2}}\, a'_{n-4} & a_{n-5} - \dfrac{a_{n-1}}{a'_{n-2}}\, a'_{n-6} &
\end{array}
\tag{5.26}
$$

$$
R_n = a_n \quad , \quad R_{n-1} = a_{n-1} \quad , \quad R_{n-2} = a'_{n-2} \quad , \quad \ldots \tag{5.27}
$$

Apart from the statement on stability, we can determine from the sequence of test functions the number of roots of the characteristic polynomial with positive real parts because it equals the number of sign changes in the sequence of the test functions.

Comparison of the two criteria shows that the Hurwitz criterion appears to be more elegant, but, because of the numerous determinant calculations, it is more difficult to use than that of Routh. The Routh criterion is generally preferred for a numerical calculation, particularly with systems of higher than third order.

When using both criteria on lower order systems, special cases are encountered which can be dealt with quite easily. When checking differential equations of first and second order, stability is already assured as soon as the first of the already mentioned conditions has been satisfied. From the schemes for the second condition we can see for the second order differential equation that the Hurwitz determinant is $H = a_1$ and

to be considered is the part of the curve C identified by a dot-dash line, the semicircle with an infinite radius. This part is mapped to the origin of the $G_0(s)$ plane by any transfer function describing real systems because the denominator of these transfer functions is of higher degree than the numerator. Finally, the part of curve C identified by a dashed line designates negative imaginary argument values $s = -j\omega$. Because the transfer functions consist of polynomials with only real coefficients,

$$G_0(-j\omega) = G_0^*(j\omega) \tag{5.48}$$

applies here, i.e. the function values are the same as the conjugate complex of the values associated with the corresponding positive arguments. Therefore the corresponding part of curve C'' (illustrated with a dashed line) is created from the previously obtained partial curve by reflection at the real axis.

	$m = n - p$ (5.46)	
m	number of revolutions of C' in $N(s)$ plane (opposite mathematically positive direction)	number of revolutions of C'' around -1 (≡ Nyquist plot of $G_0(j\omega)$ for $-\infty < \omega < +\infty$) "open loop"
n	number of zeros of $N(s)$ inside C (right s halfplane)	number of poles of $G_z(s)$ in the right s halfplane (cf. eq.(5.35)) [e.g.: $G_z(s)$ stable → n = 0] "closed loop"
p	number of poles of $N(s)$ inside C (right s halfplane)	number of poles of $G_0(s)$ in the right s halfplane (cf. eq.(5.45)) [e.g.: $G_0(s)$ stable → $p = 0$] "open loop"

Table 5-3: Summary of the considerations on the Nyquist criterion

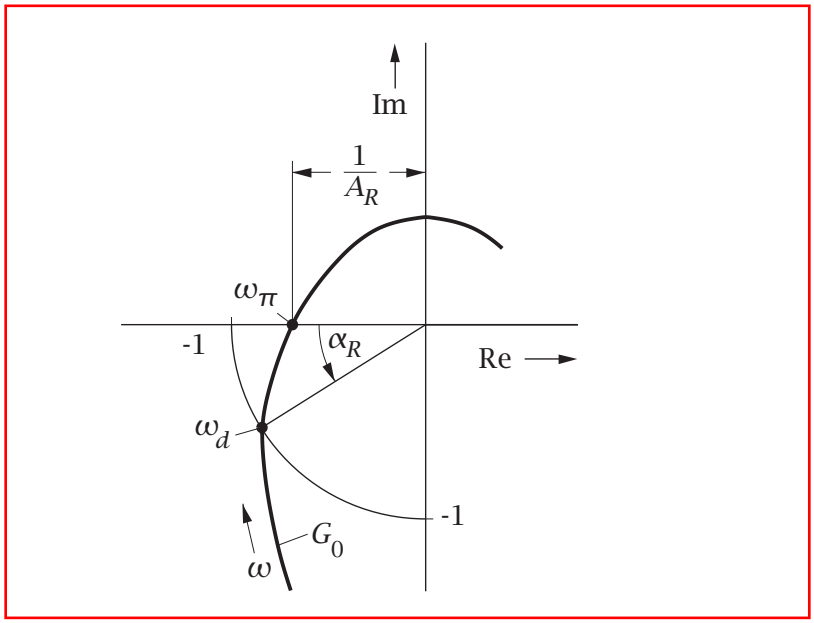


Figure 5-24: Gain and phase margins

The gain margin is the factor by which the static gain of the open control loop must be multiplied so that the associated closed control loop is at the limit of stability.

$$A_R = \frac{1}{|G_0(j\omega_\pi)|} \tag{5.63}$$

with ω_π defined by

$$\varphi_0(\omega_\pi) = -(2n+1)\pi \quad . \tag{5.64}$$

The gain margin is also called amplitude margin or amplitude distance.

The phase margin is the angle which a phasor drawn to the point where the Nyquist plot intersects a circle with a radius of one forms with the negative real axis

$$\alpha_R = \varphi_0(\omega_d) - \varphi_0(\omega_\pi) \tag{5.65}$$

with ω_d defined by

$$|G_0(j\omega_d)| = 1 \quad . \tag{5.66}$$

If the Nyquist plot intersects the real axis several times, the least gain margin or the least phase margin is the crucial factor.

number of controlled (sub)systems in a predefined time sequence, the treatment of only one control loop results in the structure illustrated in figure 6-1. This structure shall form the basis for the following considerations.

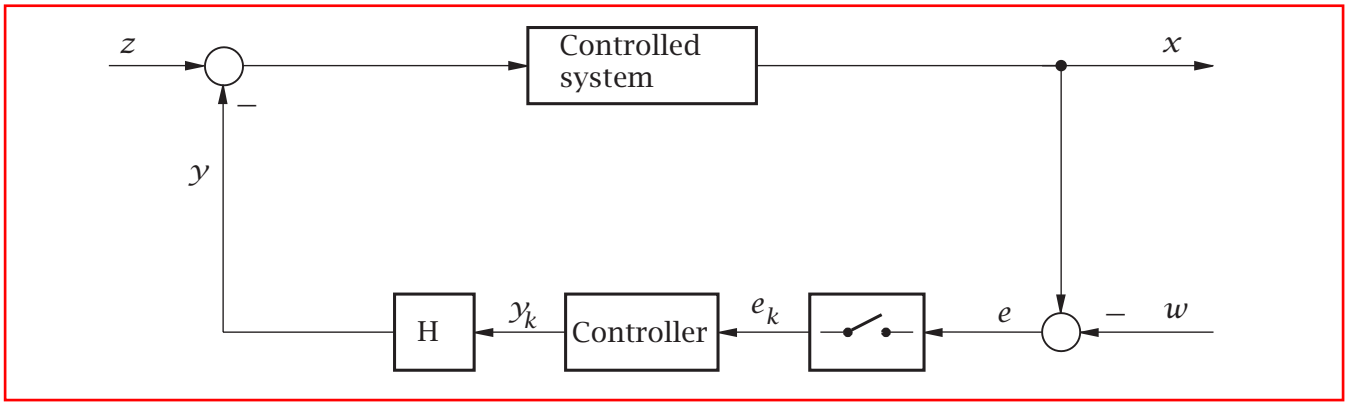


Figure 6-1: Simple sampling control

Time discretization of continuous variables, henceforth called sampling, can be interpreted in such a way that the continuous variables $e(t)$ is either associated with a series of equidistant impulses $e^*(t)$ or a time-series of values e_k (figure 6-2). The areas of the impulses correspond to the associated values of the continuous variable. Their time spacing is the sampling interval T.

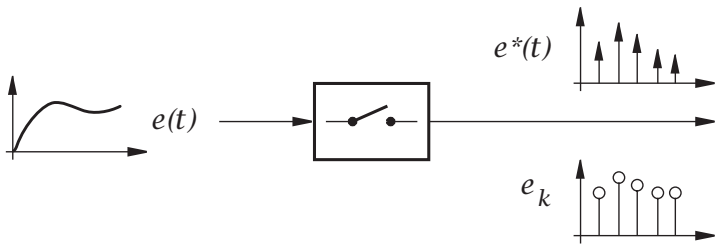


Figure 6-2: Sampler

The time-discrete variable $e^*(t)$ or the time-series of values e_k can be processed by time-discrete transfer elements. These transfer elements can be treated as impulse transferring systems which convert an input impulse sequence $e^*(t)$ into an output impulse sequence $y^*(t)$ or as systems which produce an output time-series y_k from an input time-series e_k (figure 6-3). Input and output time-series have the same sampling interval T. In most cases the time-discrete transfer systems are

ence equation and replace the continuous reference and disturbance variables with time-discrete variables.

Without going into detail, we can state that the representation of the overall system as a continuous system is useful if the sampling interval T is so small in relation to the dynamics of the system that the sampling procedure does not substantially influence the overall behaviour. In such cases it is sufficient to replace the sampler and hold element with an element with a delay time of $T/2$ which is then considered as part of the controlled system, to specify a continuous controller suitable for this controlled system and then to allocate a time-discrete control algorithm to this controller. The details are dealt with in the following section 6.3.

If the sampling interval in relation to the dynamics of the system is no longer negligibly small, it is generally prudent to aim at describing the overall system as a time-discrete system. For this we need the time-discrete representation of the subsystem consisting of hold element, controlled system and sampler, which is then combined with the representation of the time-discrete controller. The necessary procedures will be dealt with in a different lecture (HRT).

6.2 Linear time-discrete transfer systems

The controller in figure 6-1 is a linear, time-discrete transfer element. In general, such transfer elements convert a time-series of input values u_k or input impulses $u^*(t)$ into a time-series of output values y_k or output impulses $y^*(t)$.

In analogy to the description of linear continuous transfer systems by linear differential equations, it is possible to describe time-discrete transfer systems by difference equations. Such difference equations for time-series are of the following form

$$a_0 y_k + a_1 y_{k-1} + \ldots + a_n y_{k-n} = b_0 u_k + b_1 u_{k-1} + \ldots + b_m u_{k-m} \tag{6.2}$$

with (u_k) as the time-series of the input values and (y_k) as the time-series of the output values. The equation (6.2) can be shortened to

the Shannon theorem must be shorter than half of the shortest cycle duration in the function to be sampled

$$T < \frac{T_{\min}}{2} \quad . \tag{6.34}$$

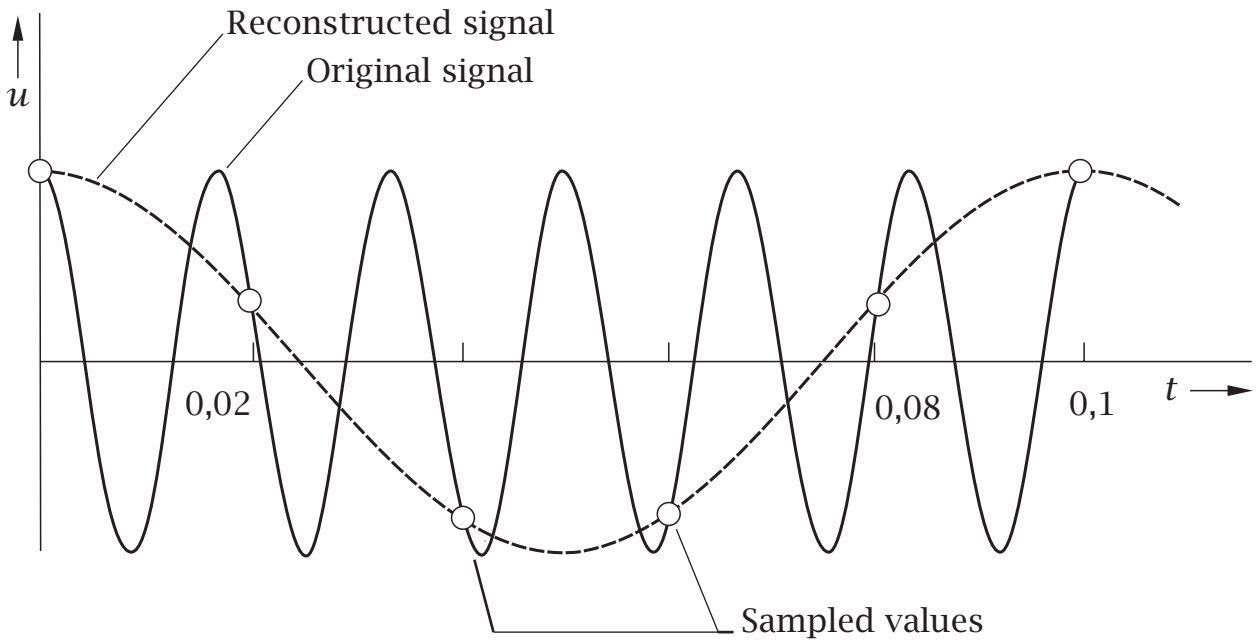


Figure 6-8: Falsification due to undersampling

Figure 6-8 illustrates in a particularly drastic manner the consequences of a violation of the Shannon theorem. A sinusoidal original signal with a frequency of 60 Hz is sampled with a frequency of 50 Hz, although the sampling frequency should be higher than 120 Hz. The consequence of this so-called undersampling is that, with careful reconstruction of the sampled signal, a sinusoidal signal with a frequency of 10 Hz – the difference between the sampling frequency and the frequency of the original signal – is created, a result which has practically nothing in common with the original. The example also illustrates that undersampling not only causes a loss of higher frequency signal parts, which might be acceptable in some circumstances, but that there is also a falsification of the signal in the – often more interesting – area of lower frequencies. This aliasing can generally not be eliminated by means of some form of reworking of the samples.

In applied measuring and control engineering, the undesirable higher

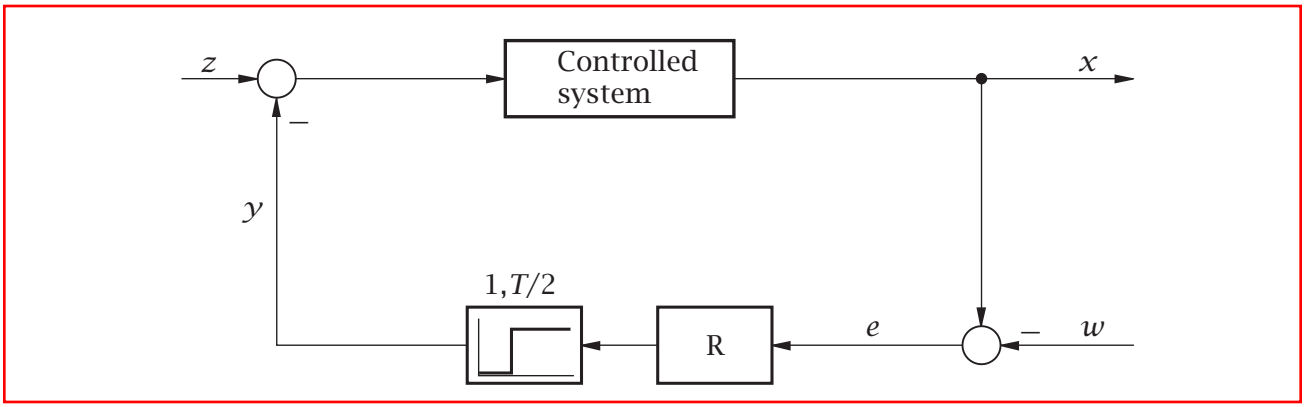


Figure 6-12: Equivalent control loop

The disadvantage of the time-discrete method of operation of the control algorithm is often compensated by the fact that various difficulties related to analog pneumatic or electronic devices are avoided and that, for instance, there is no restriction in the range of realizable parameter values, as in the case of *PID*-controllers with *PD* input network and lag-differential feedback.

Using the approximations introduced in section 6.2 for differentiation and integration, it is relatively simple to obtain the difference equation which corresponds to a time-continuous *PID*-controller. From the differential equation

$$y = K_R \left(u + \frac{1}{T_n} \int_0^t u \, d\tau + T_v \dot{u} \right) \tag{6.35}$$

one obtains through differentiation with respect to time

$$\dot{y} = K_R \left(\dot{u} + \frac{1}{T_n} u + T_v \ddot{u} \right) \quad . \tag{6.36}$$

With the help of the corresponding difference quotient – for the second derivative, the backward difference has to be established twice – follows

$$\frac{y_k - y_{k-1}}{T} = K_R \left(\frac{u_k - u_{k-1}}{T} + \frac{1}{T_n} u_k + T_v \frac{u_k - 2u_{k-1} + u_{k-2}}{T^2} \right) \quad . \tag{6.37}$$

Correctly arranged, we obtain

there are numerous others, which partially represent mixed forms of the above-mentioned methods. The technical conditions are so multifarious that the following descriptions can only be seen as a guide. Later on in this chapter we will deal with some of the peculiarities of multivariable control.

7.2 Precontrol

Precontrols have the task of reducing as far as possible the disturbances to the process to be controlled, so that influencing variables with disturbing changes are kept fully or practically constant by means of additional, mostly very simply constructed controls.

Figure 7-2 shows the functional diagram of a control loop with a closed precontrol loop, consisting of the controlled system S_v and the controller R_v. The precontrol is intended to reduce the disturbance variable z, so that only the reduced disturbance variable z' affects the main control loop. It is obvious that disturbance variables, to be reduced by the precontrol, must be capable of being measured and influenced.

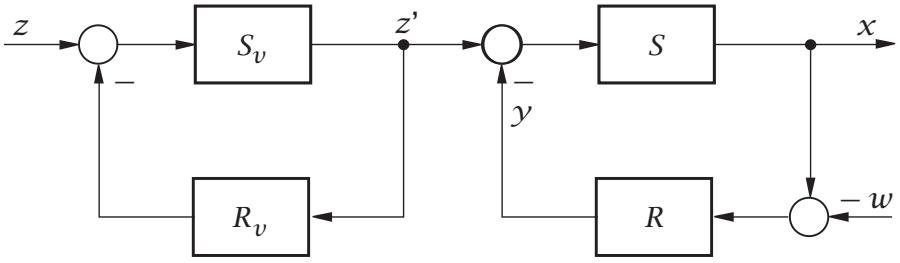


Figure 7-2: Precontrol

The precontrol of the gas pressure of gas heated ovens can be used in this case as an example. The disturbance variable z in figure 7-2 corresponds to the pressure fluctuations in the supply network, which will be reduced by a pressure regulator S_v, R_v so that their influence on the oven temperature x remains minimal. A simple pressure regulator without auxiliary power is often used as the precontroller. However, because the temperature can still be changed by other influencing factors which cannot be handled by a precontrol, we cannot avoid using the main controller R.

A precontrol does not only improve the disturbance response of a con-

trol with regard to the disturbance variables handled by the precontrol, in many cases it also permits a reduction of the range of the manipulated variable (y in figure 7-2) influenced by the main controller, and this often improves the economics of the whole process.

7.3 Feedforward control with a disturbance variable

By feedforward of disturbance variables useful changes in the manipulated variable of the main control loop are derived from measurements of the change in variables which disturb the process to be controlled. In an ideal situation, the manipulated variable can accurately compensate for the influence of the disturbance, so that the controlled variable is not influenced by this disturbance.

As illustrated in figure 7-3, the disturbance variable z is measured and a variable generated by the compensating unit A is superimposed on the manipulated variable generated by the controller. We can recognize that for

$$G_A = 1 \tag{7.1}$$

the resulting change in the manipulated variable would fully compensate the disturbance variable z.

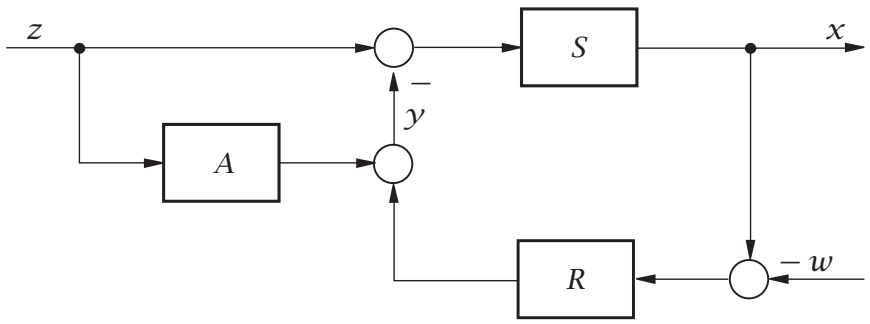


Figure 7-3: Feedforward of disturbance variable

When controlling with individual analog units, the structure illustrated in figure 7-4 is frequently used, where the disturbance variable is added to the deviation. The effect of both compensation structures is the same if in figure 7-4 for full compensation

$$G_A \cdot G_R = 1 \tag{7.2}$$

problems. On the other hand, the success of such a method depends strongly on the correct tuning of this open loop control.

A widely used application example for feedforward of a disturbance variable is the control of the heating fluid entry temperature in central heating systems as a function of changes of the outside temperature. The outside temperature is one of the most important disturbance variables of room temperature control. If, by suitable control of the radiator inlet fluid temperature, the effect of outside temperature fluctuations on the room temperature is partially or wholly compensated for, the room temperature controller can generally be simpler and therefore cheaper while it can carry out its intended task better than a controller in a single control loop.

7.4 Auxiliary manipulated variable

If the process to be controlled can be represented in the functional diagram in the form of a serial connection of several lag elements (figure 7-5), it may well be sensible to use an additional manipulated variable, the auxiliary manipulated variable y_h. The most important prerequisite for this is, of course, that such a manipulated variable can be applied to the system to be controlled.

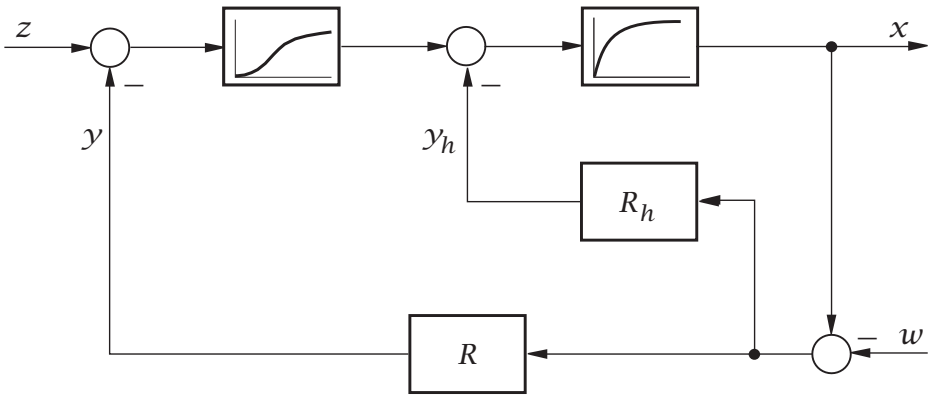


Figure 7-5: Application of an auxiliary manipulated variable

The auxiliary manipulated variable and the controller R_h generating it build a sub-control loop with a part of the controlled system. This loop can generally have more favourable dynamic characteristics than the main control loop, because the associated controlled subsystem is of

of an auxiliary manipulated variable. Due to more favourable conditions, this additional control loop can be designed for good dynamics and will also improve the stability characteristics of the main control loop.

However, as the auxiliary controller does not process the actually interesting controlled variable x and the associated reference variable w, it must be designed so as to ensure that it does not obstruct the main controller. This is generally implemented by equipping the auxiliary controller with P, PD or differentiating behaviour. In the case of controls with analog equipment, the solution illustrated in figure 7-7b is frequently used for the auxiliary controller R'_h to minimize the costs. If

$$G_{R'_h} \cdot G_R = G_{R_h} \tag{7.4}$$

applies, then both solutions are equivalent.

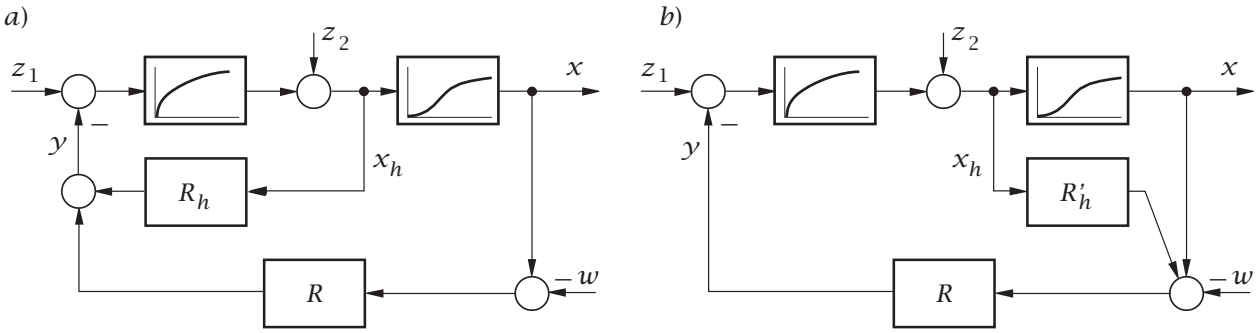


Figure 7-7: Application of an auxiliary controlled variable

In the event that the improvement in the dynamics of the main control loop, created by the auxiliary controlled variable, is used to increase the steady state gain of the main controller, it must be noted that, as in the case of the auxiliary manipulated variable, failure of the auxiliary control loop can also endanger the stability of the whole system. If this possibility is not used, the auxiliary controller reduces only the effects of the disturbances (z_1, z_2 in figure 7-7) covered by it.

The previously mentioned steam temperature control shall again be used as an example. According to figure 7-8, it is generally not only the temperature of the steam leaving the superheater which will be measured but also the steam temperature between the spray cooler and the

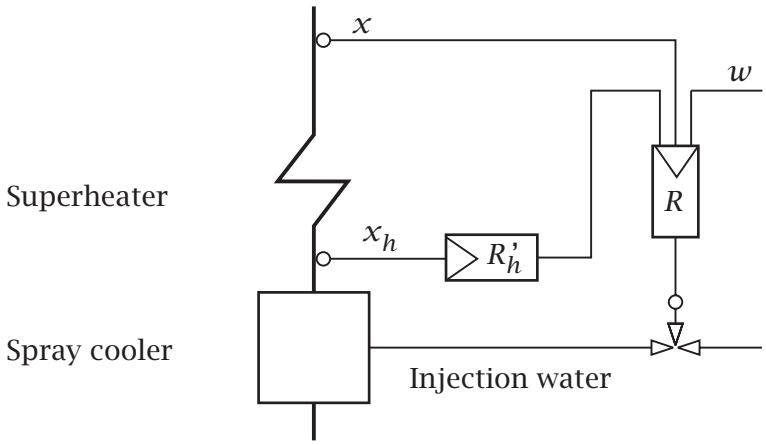


Figure 7-8: Steam temperature control with the temperature upstream of the superheater as auxiliary controlled variable

superheater. This measured quantity x_h is also used to influence the injection water flow.

7.6 Cascade control

If the prerequisites for the use of an auxiliary controlled variable exist, it is possible to arrange the main and the auxiliary controller in such a way that the main controller R produces the reference variable of the auxiliary controller R_h (figure 7-9). This creates a secondary control loop in which all the disturbances acting on the front part of the controlled system (z_1, z_2 in figure 7-9) are balanced out by the auxiliary controller.

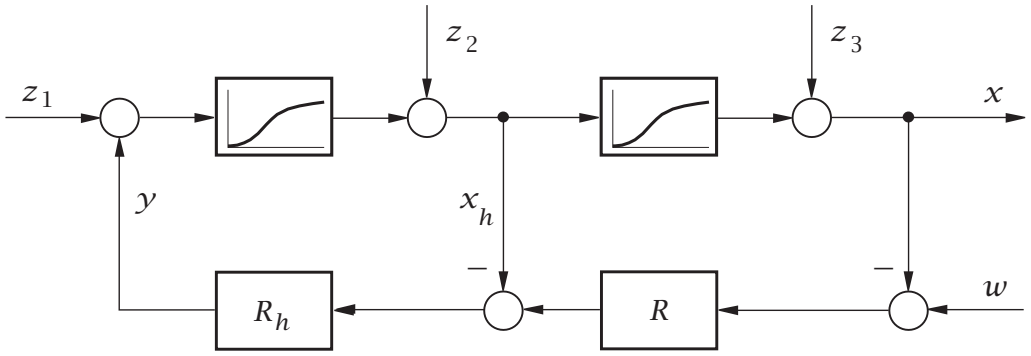


Figure 7-9: Cascade control

Cascade controls are a very frequently used form of multi-loop controls.

7.7 Feedforward control and reference variable filter

A feedforward of the reference variable according to figure 7-11 is frequently used with follow-up controls. This is somewhat related to the feedforward of a disturbance variable dealt with in section 7.3. The compensation unit A can be used to reduce the dynamic errors of the follow-up control without unfavourable effects on the stability of the system.

The following applies for the structure in figure 7-11

$$G(s) = \frac{X(s)}{W(s)} = \frac{(G_A + G_R)G_S}{1 + G_R G_S} = \frac{G_A G_S + G_R G_S}{1 + G_R G_S} \tag{7.5}$$

and we can recognize that for

$$G_A = \frac{1}{G_S} \tag{7.6}$$

the follow-up system with feedforward control does not have any dynamic errors because its transfer function is $G = 1$.

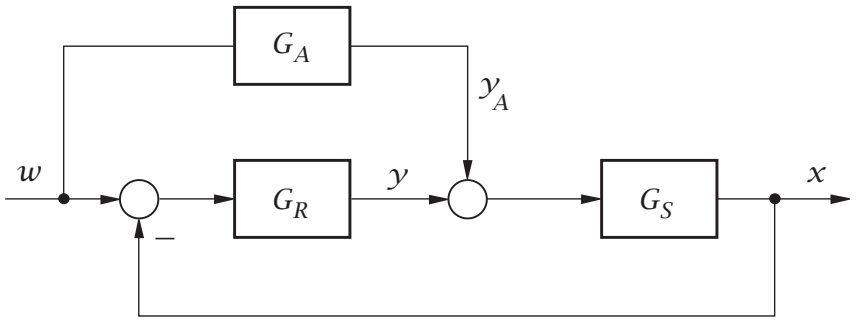


Figure 7-11: Follow-up control with feedforward control

In the majority of cases the controlled systems in such follow-up controls have integrating behaviour with a lag (IT_1 or IT_n behaviour). The compensation unit must therefore calculate multiple derivatives and produce a manipulated variable which consists of a weighted sum of the derivatives of the reference variable with respect to time. Because of difficulties related to the technical equipment and because higher frequency signal components are considerably increased by differentiation, we must restrict ourselves to only a few derivatives and abstain from a complete compensation for the dynamic errors.

8 State space

8.1 General

The relationship between the input and output variables of dynamic transfer systems may be described not just in terms of various differential equations, generally of a higher order, but also in terms of systems of first order differential equations. The variables that appear in addition to the input and output variables in such differential equation systems must conform to certain definite conditions, and are then generally characterised by the letter x as state variables.

The system of differential equations is then constructed in such a way that the n derivatives $\dot{x}_i$ of the state variables x_i are expressed as functions of these state variables and the p input variables u_i

$$\begin{aligned} \dot{x}_1 &= f_1(x_1, \ldots, x_n, u_1, \ldots, u_p, t) \\ &\vdots \\ \dot{x}_n &= f_n(x_1, \ldots, x_n, u_1, \ldots, u_p, t) \quad . \end{aligned} \tag{8.1}$$

The q output variables y_i are represented as functions of the state variables and input variables:

$$\begin{aligned} y_1 &= g_1(x_1, \ldots, x_n, u_1, \ldots, u_p, t) \\ &\vdots \\ y_q &= g_q(x_1, \ldots, x_n, u_1, \ldots, u_p, t) \quad . \end{aligned} \tag{8.2}$$

In abbreviated form, the input, output and state variables are combined as vectors, and one obtains

$$\begin{aligned} \dot{\boldsymbol{x}} &= \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{u}, t) \\ \boldsymbol{y} &= \boldsymbol{g}(\boldsymbol{x}, \boldsymbol{u}, t) \quad . \end{aligned} \tag{8.3}$$

In case of a linear time-invariant system, equation (8.3) simplifies to:

$$\begin{aligned} \dot{\boldsymbol{x}} &= \boldsymbol{A} \cdot \boldsymbol{x} + \boldsymbol{B} \cdot \boldsymbol{u} \\ \boldsymbol{y} &= \boldsymbol{C} \cdot \boldsymbol{x} + \boldsymbol{D} \cdot \boldsymbol{u} \end{aligned} \tag{8.4}$$

where $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, $\boldsymbol{D}$ are matrices with time-independent coefficients.

can be obtained in an analogous manner, if the expression e^{At} is first defined in an appropriate way. The scalar expression e^{at} can be represented in the form of an infinite series

$$e^{at} = \sum_{k=0}^{\infty} \frac{(a \cdot t)^k}{k!} = 1 + \frac{a \cdot t}{1!} + \frac{(a \cdot t)^2}{2!} + \ldots \quad . \tag{8.46}$$

From this the definition

$$e^{\boldsymbol{A}t} = \sum_{k=0}^{\infty} \frac{(\boldsymbol{A} \cdot t)^k}{k!} = \boldsymbol{I} + \frac{t}{1!}\boldsymbol{A} + \frac{t^2}{2!}\boldsymbol{A}^2 + \ldots \tag{8.47}$$

which will subsequently be used, can be derived. It can be seen that the $(n \times n)$ matrix $e^{\boldsymbol{A}t}$ arises from the $(n \times n)$ matrix $\boldsymbol{A}$, because equation (8.47) represents a sum of $(n \times n)$ matrices. It can be shown that the series (8.47) is convergent and that

$$\frac{d}{dt} e^{\boldsymbol{A}t} = \boldsymbol{A} \cdot e^{\boldsymbol{A}t} = e^{\boldsymbol{A}t} \cdot \boldsymbol{A} \quad . \tag{8.48}$$

On the basis of these determinations, the solution of the state equation (8.45) corresponding to equation (8.44) is

$$\boldsymbol{x}(t) = e^{\boldsymbol{A}t}\boldsymbol{x}(0) + \int_0^t e^{\boldsymbol{A}(t-\tau)}\boldsymbol{B}\,\boldsymbol{u}(\tau)\,d\tau \quad . \tag{8.49}$$

In the case of a vanishing input variable $u = 0$, equation (8.49) becomes

$$\boldsymbol{x}(t) = e^{\boldsymbol{A}t}\boldsymbol{x}(0) = \boldsymbol{\Phi}(t)\boldsymbol{x}(0) \tag{8.50}$$

on the basis of what is known as the state transition matrix

$$\boldsymbol{\Phi}(t) = e^{\boldsymbol{A}t} \quad , \tag{8.51}$$

which is to be determined according to equation (8.47).

The general solution for the output variable $\boldsymbol{y}$ may be obtained with

$$\boldsymbol{y} = \boldsymbol{C} \cdot \boldsymbol{x} + \boldsymbol{D} \cdot \boldsymbol{u} \tag{8.52}$$

from equation (8.49) as

$$\boldsymbol{y}(t) = \boldsymbol{C} \cdot e^{\boldsymbol{A}t} \cdot \boldsymbol{x}(0) + \int_0^t \boldsymbol{C} \cdot e^{\boldsymbol{A}(t-\tau)} \cdot \boldsymbol{B} \cdot \boldsymbol{u}(\tau) d\tau + \boldsymbol{D} \cdot \boldsymbol{u}(t) \quad . \tag{8.53}$$

With conversions that shall not be explained in detail,

$$\boldsymbol{y}(t) = \boldsymbol{C} \cdot e^{\boldsymbol{A}t} \cdot \boldsymbol{x}(0) + \int_0^t \boldsymbol{G}(t-\tau) \cdot \boldsymbol{u}(\tau) d\tau \tag{8.54}$$

can be derived from this, with $\boldsymbol{G}(t)$ as the so-called weighting matrix of the transfer system.

The Laplace transform has proven to be a useful tool for solving differential equations. It is used in such a way that the time functions arising in the differential equation are to be replaced by their transformed functions in the frequency domain of the Laplace transform and the relationships between the time functions are to be transferred into the corresponding ones between the transformed functions. The most advantageous part of this procedure is that the mathematical integration and differentiation operations turn into simple algebraic operations. Since the Laplace transform is applicable in the same way to both, systems of differential equations and single differential equations, one can also use this tool with the state equations, if one thereby considers the rules of matrix calculus.

The state equations in the frequency domain of the Laplace transform are as follows

$$\begin{aligned} s \cdot \boldsymbol{X}(s) - \boldsymbol{x}(0) &= \boldsymbol{A} \cdot \boldsymbol{X}(s) + \boldsymbol{B} \cdot \boldsymbol{U}(s) \\ \boldsymbol{Y}(s) &= \boldsymbol{C} \cdot \boldsymbol{X}(s) + \boldsymbol{D} \cdot \boldsymbol{U}(s) \quad . \end{aligned} \tag{8.55}$$

In equation (8.55) the vectors of the transformed functions of the state, input and output variables were designated with bold capital letters $\boldsymbol{X}(s)$, $\boldsymbol{U}(s)$, $\boldsymbol{Y}(s)$, although they are not matrices. $\boldsymbol{x}(0)$ is the vector of initial conditions of the state equation. The coefficient matrices $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, $\boldsymbol{D}$ are changed by the transformation just as little as the coefficients of a single differential equation.

In the case of a system with a single input variable u and a single output variable y, the weighting matrix $\boldsymbol{G}(t)$ consists only of a single element — the weighting function $g(t)$, and the so-called transfer matrix $\boldsymbol{G}(s)$ consists only of a single element — the transfer function $G(s)$.

In the case of multiple input and output variables which has to be considered here, the weighting matrix $\boldsymbol{G}(t)$ consists of a corresponding number of weighting functions that connect the input and output variables with one another; in an equivalent way, the elements of the transfer matrix $\boldsymbol{G}(s)$ are transfer functions that connect the transformed functions of the input and output variables with one another.

8.4 Controllability and observability

From the general solutions of the state space equations (8.49) and (8.54), some important statements about the described system can be derived. Among these characteristics are the controllability and the observability of the system - terms that were introduced by Kalman in 1960.

A system

$$\begin{aligned} \dot{\boldsymbol{x}} &= \boldsymbol{A} \cdot \boldsymbol{x} + \boldsymbol{B} \cdot \boldsymbol{u} \\ \boldsymbol{y} &= \boldsymbol{C} \cdot \boldsymbol{x} + \boldsymbol{D} \cdot \boldsymbol{u} \end{aligned} \tag{8.61}$$

is said to be controllable if its state $\boldsymbol{x}$ can be transferred from any arbitrary initial state $\boldsymbol{x}(t_0)$ to the final state $\boldsymbol{0}$ in finite time by means of an appropriate input value, the control vector $\boldsymbol{u}(t)$.

Correspondingly, the system (8.61) is said to be observable if from the known input vector $\boldsymbol{u}(t)$ and from the measurement of $\boldsymbol{y}(t)$ over a finite time interval, the initial state $\boldsymbol{x}(t_0)$ can be determined uniquely. For observable systems, one can design so-called state observers which generate estimates of the state variables from the input and output variables.

One can demonstrate, that a system with a single input variable u and a single output variable y is controllable , if the vectors

$$\boldsymbol{b},\ \boldsymbol{A} \cdot \boldsymbol{b},\ \boldsymbol{A}^2 \cdot \boldsymbol{b},\ \ldots,\ \boldsymbol{A}^{n-1} \cdot \boldsymbol{b} \tag{8.62}$$

are linearly independent . Thus, the (n,n)-controllability matrix

$$\boldsymbol{Q}_S = \left[\boldsymbol{b},\ \boldsymbol{A} \cdot \boldsymbol{b},\ \boldsymbol{A}^2 \cdot \boldsymbol{b},\ \ldots,\ \boldsymbol{A}^{n-1} \cdot \boldsymbol{b}\right] \tag{8.63}$$

is nonsingular if and only if the system is controllable. In other words, controllability is given when

$$\det \boldsymbol{Q}_S \neq 0 \quad . \tag{8.64}$$

A system with a single input variable u, n state variables and a single output variable y is said to be observable, if the vectors

$$\boldsymbol{c}^T,\ \boldsymbol{c}^T \cdot \boldsymbol{A},\ \ldots,\ \boldsymbol{c}^T \cdot \boldsymbol{A}^{n-1} \tag{8.65}$$

are linearly independent . In other words, observability is given if the (n,n)-observability matrix

$$\boldsymbol{Q}_B = \begin{bmatrix} \boldsymbol{c}^T \\ \boldsymbol{c}^T \cdot \boldsymbol{A} \\ \vdots \\ \boldsymbol{c}^T \cdot \boldsymbol{A}^{n-1} \end{bmatrix} \tag{8.66}$$

is nonsingular.

For systems with multiple input variables and multiple output variables, correspondingly extended conditions apply.

8.5 Stability and closed-loop control in state space

8.5.1 Stability and state feedback

The poles of the transfer function, which are the zeros of its denominator polynomial, determine the dynamic characteristics of the system, in particular its stability and its damping characteristics. Transferring this statement to equation (8.60), it follows that the roots of the equation

$$\det(s \cdot \boldsymbol{I} - \boldsymbol{A}) = 0 \tag{8.67}$$

are essential for the behaviour of the system. The determinant in equation (8.67) is a n-th order polynomial in s and corresponds to the characteristic polynomial. The roots of the determinant in equation (8.67) are also designated as the eigenvalues of the matrix $\boldsymbol{A}$. All of them must exhibit negative real parts, if the system described by the matrix $\boldsymbol{A}$ is supposed to be stable.

which will be combined to yield

$$\dot{\boldsymbol{x}} = (\boldsymbol{A} - \boldsymbol{B} \cdot \boldsymbol{K}) \cdot \boldsymbol{x} \tag{8.71}$$

Equation (8.71) describes a system without any input variables with the system matrix

$$\boldsymbol{A}_K = \boldsymbol{A} - \boldsymbol{B} \cdot \boldsymbol{K} \quad . \tag{8.72}$$

8.5.2 Pole placement

One possibility for the controller design is to select desirable eigenvalues of the matrix $\boldsymbol{A}_K$ and to determine from this and the known matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ the controller or feedback matrix $\boldsymbol{K}$.

As an example a state feedback is to be determined according to the mentioned procedure of pole placement for a transfer system with a single input and a single ouput variable. Figure 8-6 shows the functional diagram of the system with feedback.

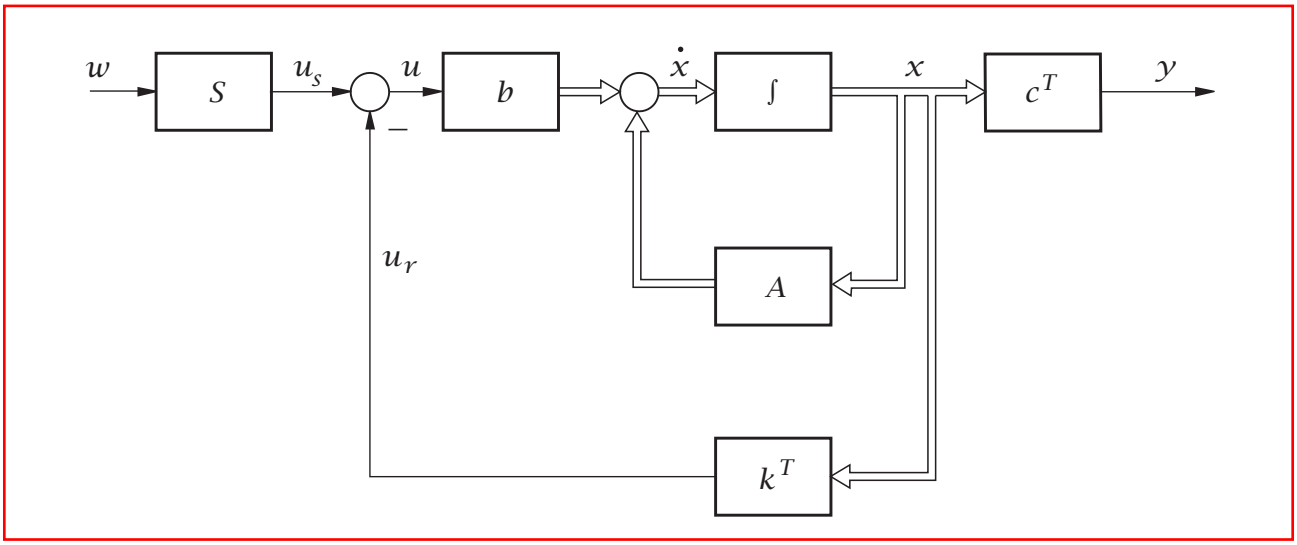


Figure 8-6: SISO-system with state feedback

The transfer system may be be stated in controller canonical form according to equation (8.23). The state variables of the controller canonical form can be obtained for this purpose by transformation of the original state variables in the way described in chapter 8.2. According

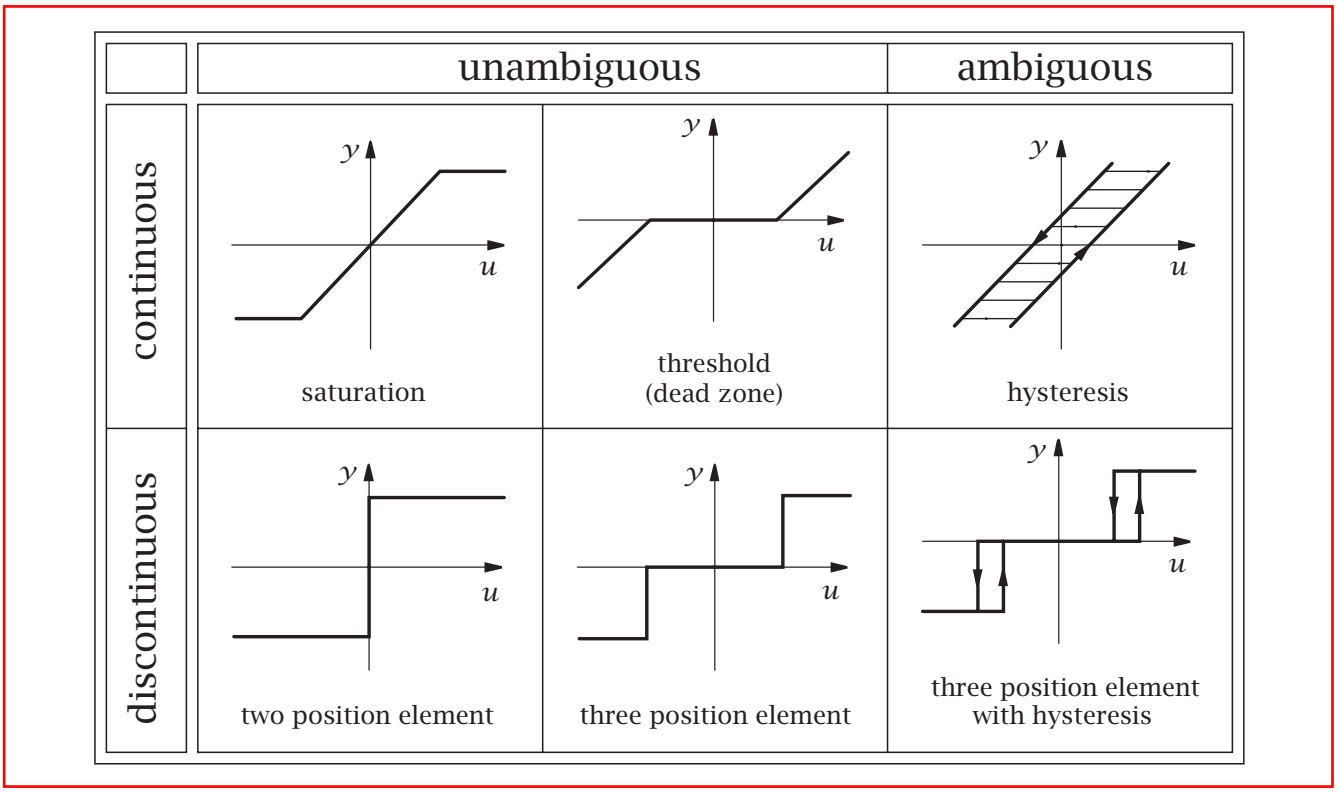


Figure 9-1: Types of characteristic curves

The most frequent nonlinearities in simple follow-up systems are the saturation of the manipulating speed and the threshold (dead zone). With the functional diagram in figure 9-2 both will be considered by inserting the appropriate characteristic curve into the nonlinear element. Some information regarding the effects of such nonlinear elements can already be provided by elementary considerations relating to the behaviour of follow-up controls in the time domain with simple input variables.

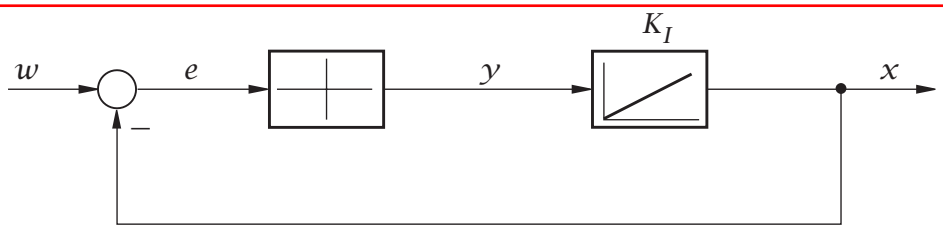


Figure 9-2: Follow-up system with a nonlinear transfer element

The saturation of the manipulating speed y_{max} (figure 9-3) will become effective when the operating limits of the integrating element have been