Limitations on Control System Performance

K. J. Åström

Department of Automatic Control, Lund Institute of Technology
Box 118, S-221 00 Lund, Sweden
Email: kja@control.lth.se
Fax: +46 46 13 81 18

Abstract

Many different factors such as process dynamics, disturbances, process uncertainties and actuator saturation have to be considered when designing a control system. It is important to be aware of factors that limits the achievable performance. In this paper we attempt to formulate some problems that capture the essence of the design problem in a simple way. Results of this type are a complement to computational tools for control system design. They make it possible to quickly make a preliminary assessment before doing massive design computations. They are also useful in process design to determine potential difficulties and to give hints for redesign.

1. Introduction

Much research has been devoted to developing methods for design of control system. A common approach has been to strive for optimality with respect to different criteria, see Newton *et al.* (1957), Horowitz (1963), Maciejowski (1989), Boyd and Barratt (1991), Anderson and Moore (1990), and Green and Limebeer (1995). Optimization methods naturally give the best performance with the specified criteria. Sometimes the methods may indicate that the goals cannot be achieved but they do not always give good insight into the mechanisms that cause the limitations. It is therefore desirable to have complementary techniques that give insight into the factors that fundamentally limits the achievable performance of a control system. This was the spirit of the early work of Bode, see Bode (1945). Bode's tradition was followed in Horowitz (1963), other early works are Frank (1968a), Frank (1968b) and Åström (1968). Many of these results were forgotten when the state space theory emerged and interest in the frequency domain faded, see Horowitz and Shaked (1975). Interest in robustness reemerged in the 1980s with the seminal paper Zames (1981) which started the development of the H_∞ theory. The results also gave good insight into the fundamental limitations. After 1980 there have been a steady stream of results, see Doyle and Stein (1981), Francis and Zames (1984), Boyd and Desoer (1995), Khargonekar and Tannenbaum (1985), Freudenberg and Looze (1985), Zames and Francis (1985), Freudenberg and Looze (1987), Engell (1988), Freudenberg and Looze (1988), Morari and Zafiriou (1989), Stein (1990), Middleton (1991), Doyle *et al.* (1992), Horowitz (1993), Qiu and Davison (1993), Chen (1995), Goodwin *et al.* (1995), Åström (1996), Skogestad and Postlethwaite (1996), Seron *et al.* (1945), Chen (1998), Havre and Skogestad (1999), Yaniv *et al.* (1999). Several of these publications also treat MIMO systems which are outside the scope of this paper.

Insights into fundamental limitations are useful for control system design. They make it possible to quickly check a problem before a design is attempted. For example, the results of this paper will immediately reveal that many of the examples in Keel and Bhattacharyya (1997) give closed-loop systems with very poor robustness. The results are particularly useful for systems with automatic tuning, adaptation and autonomy where performance assessment and design are done automatically, see Åström and Wittenmark (1995) and Åström (1993).

In this tutorial paper we present a unified approach that explicitly gives limitations expressed in terms of the gain crossover frequency. The results are limited to single-input single-output systems. The key result are inequalities for the gain crossover frequency for minimum phase and non-minimum phase systems. Simple design rules can be obtained directly from the inequalities. Related results are found in Engell (1988), Middleton (1991) and Skogestad and Postlethwaite (1996).

The paper is organized as follows. Some preliminary results on feedback systems are summarized in Section 2. minimum phase systems are discussed in Section 3. In this case the fundamental limitations are given by the measurement noise and the highest admissible gain. For non-minimum phase systems the limitations are given by the process dynamics. This is discussed in Section 4 where some simple design rules are given based on requirements on the phase margin. These results are very easy to derive, they are also not very precise because the phase margin gives only partial information about the robustness. Improved results based on the maxima of the sensitivity and the complementary sensitivity are presented in Section 5. The final result is summarized as a collection of design rules. They have the same form as the simple results based on conditions on the phase margin, but the numerical values of some parameters are different. Finally in Section 6 it is shown how the results can be used to resolve a paradox in pole placement design.

2. Preliminaries

Consider a control system with a two degree of freedom configuration as shown in Figure 1. The process to be controlled is assumed to be a single-input single-output system. The process input is u and the output is x. There are load disturbances that drive the system away from its desired behavior and measurement noise that corrupt the information about the system. For simplicity the load disturbances are assumed to act on the process input and the measurement noise on the process output. The controller has two inputs, the command signal c and the measured signal y and one output, the control signal u. There are three inputs r l and n and three interesting output signals u, x and y. If the system is linear the input-output relations are given

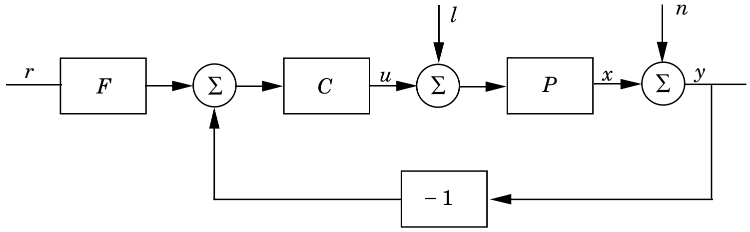

Figure 1 Block diagram of a simple control system.

by the transfer functions

$$
\begin{aligned}
G_{ul} = G_{xn} &= -T & G_{yn} &= S \\
G_{xl} = G_{yl} &= PS & G_{un} &= -CS \\
G_{ur} &= CSF & G_{xr} = G_{yr} &= TF
\end{aligned}
\tag{1}
$$

where $L = PC$ is the loop transfer function, $S = 1/(1+L)$ the sensitivity function and $T = 1 - S$ the complementary sensitivity function.

Equation (1) tells how the closed-loop system reacts to command signals, load disturbances and measurement noise. To reject load disturbances it is desirable to have a large loop transfer function L, because this implies that G_{xl} is small. A large loop transfer function does, however, also imply that G_{xn} is close to one for frequencies lower than the bandwidth of the system. This means that the amount of measurement noise injected into the system via the feedback increases with the bandwidth. The sensitivity function is also small for those frequencies where the loop transfer function is large. This implies that the closed-loop system is insensitive to process uncertainties at those frequencies. Equation (1) implies that the feedback controller C and thus also the loop transfer function can be chosen to satisfy demands on load disturbances, measurement noise and sensitivity to process uncertainties. The feedforward-transfer function F can then be chosen to obtain the desired response to command signals. Control design may thus be considered as a problem of choosing a suitable loop transfer function L and a suitable feedforward transfer function F. A key question is therefore to find features that restrict these choices.

Since $L = PC$ the properties of the process will clearly impose some limitations. In practice, all processes have the property that $P(s)$ goes to zero as s goes to infinity. It is therefore not possible to have a large loop transfer function at all frequencies. The gain cross-over frequency with smallest ω_{gc} such that $|L(i\omega_{gc})| = 1$ is chosen to characterize the gross features of the closed-loop system. The loop transfer function will thus be large for frequencies smaller than ω_{gc} and small for frequencies larger than ω_{gc}. This means that systems with mechanical resonances where the loop transfer function may intersect the unit circle at many points are excluded.

Bode's Relations

Consider a transfer function $G(s)$ with no poles or zeros in the right half plane. Introduce

$$
\log G(i\omega) = A(\omega) + i\Phi(\omega) \tag{2}
$$

a logarithmic frequency scale $u = \log \omega/\omega_0$, $\omega = \omega_0 e^u$, and the functions

$$
a(u) = A(\omega_0 e^u), \qquad \phi(u) = \Phi(\omega_0 e^u).
$$

Assume that $\log G(s)/s$ goes to zero as s goes to infinity, then

$$\begin{aligned} A(\omega_0) - A(\infty) &= -\frac{2}{\pi}\int_0^\infty \frac{v\Phi(v) - \omega_0\Phi(\omega_0)}{v^2 - \omega_0^2}dv \\ &= -\frac{1}{\omega_0\pi}\int_{-\infty}^\infty \frac{d(e^u\phi(u))}{du}\log\coth|\frac{u}{2}|du \\ \Phi(\omega_0) &= \frac{2\omega_0}{\pi}\int_0^\infty \frac{A(v) - A(\omega_0)}{v^2 - \omega_0^2}dv = \frac{1}{\pi}\int_{-\infty}^\infty \frac{da(u)}{du}\log\coth|\frac{u}{2}|du \end{aligned} \tag{3}$$

an approximate version is that

$$\Phi(\omega) \approx \frac{2}{\pi}\frac{da(u)}{du}. \tag{4}$$

This means that if the slope $n = da(u)/du$ of the magnitude curve is constant the phase is $n\pi/2$. This relation appears in practically all elementary courses in feedback control. In this paper we will show that Equation (4) can be used to obtain fundamental limitations on the achievable performance of a control system. Notice also that it is possible to find minimum phase systems that approximate Equation (4) arbitrarily well even for non-integer n, Horowitz (1963).

Bode's Integral

Another limitation on the loop transfer function is expressed by Bode's integral which states that, for systems where $sL(s)$ goes to zero as $s \to \infty$, the sensitivity function S and the complementary sensitivity function T have the properties.

$$\begin{aligned} \int_0^\infty \log|S(i\omega)|d\omega &= \int_0^\infty \log|\frac{1}{1+L(i\omega)}|d\omega = \pi\sum p_i \\ \int_0^\infty \log|T(1/i\omega)|d\omega &= \int_0^\infty \log|\frac{L(1/i\omega)}{1+L(1/i\omega)}|d\omega = \pi\sum\frac{1}{z_i} \end{aligned} \tag{5}$$

where p_i are the right half plane poles of $L(s)$ and z_i its right half plane zeros, a pole or a zero of order n is counted n times. This means that the sensitivity functions cannot be made small for all frequencies, decreasing it at one frequency makes it larger at other frequencies. It also follows from the relation that the sensitivity increases if the process or the controller has poles in the right half plane and that the complementary sensitivity function increases if the process has zeros in the right half plane. See Kwakernak (1995).

Bode's Ideal Loop Transfer Function

In his work on design of feedback amplifiers Bode has suggested an ideal shape of the loop transfer function. He proposed that the loop transfer function should have the form

$$L(s) = (\frac{s}{\omega_{gc}})^n. \tag{6}$$

This transfer function (6) has the properties $d\log|L(i\omega)|/d\log\omega = n$ and $\arg L(i\omega) = n\pi/2$. This means that the Bode diagram is very simple. Notice that with our sign convention the slope n is typically negative. Both the amplitude curve and the phase curve are straight lines. The amplitude curve has constant slope n, and the phase curve is a horizontal line at $n\pi/2$. The Nyquist curve is simply the straight line through the origin, $\arg L(i\omega) = n\pi/2$. In his work on feedback amplifiers Bode called (6) the ideal cut-off characteristic. In the terminology of automatic control we will call it Bode's ideal loop transfer function.

One reason why Bode made the particular choice of $L(s)$ is that it gives a closed-loop system that is insensitive to gain changes. Changes in the process gain will change the crossover frequency but the phase margin is $\varphi_m = \pi(1 + n/2)$ for all values of the gain. The amplitude margin is infinite. The slopes $n = -1.333$, -1.5 and -1.667 correspond to phase margins of $60°$, $45°$ and $30°$. Bode's idea to use loop shaping to design controller that are insensitive to gain variations were later generalized by Horowitz (1963) to systems that are insensitive to other variations of the plant, culminating in the QFT method, see Horowitz (1993).

The transfer function given by Equation (6) is an irrational transfer function for non-integer n. It can be approximated arbitrarily close by rational frequency functions. Bode also suggested that it was sufficient to approximate L over a frequency range around the desired crossover frequency ω_{gc}.

3. Minimum Phase Systems

minimum phase systems are easy to control. Since the process poles and zeros are in the left half plane they can be canceled freely to obtain a desired loop transfer function L. A simple design procedure is to choose a desired complementary sensitivity function T. The corresponding loop transfer function is then $L = T/(1-T)$ and the controller becomes $C = L/P$. Such a controller is said to have high control authority because the properties of the process have no influence on the closed-loop response. It is for example possible to obtain systems with arbitrarily high bandwidth. This is true if the systems are linear without measurement noise. The situation is, however, very different in the presence of measurement noise and actuator saturation, because more measurement noise is injected into the system when the bandwidth is increased. This generates large control signals which will saturate the actuators. Measurement noise and actuator saturations are thus factors that limit the performance for minimum phase systems.

The key problem is to determine how the controller gain is influenced by the specifications. Let the process have the transfer function $P(s)$. Assume that performance is specified by the gain crossover frequency ω_{gc}. To have a phase margin φ_m it is necessary to provide a phase lead of

$$\varphi = -\arg P(i\omega_{gc}) - \pi + \varphi_m. \tag{7}$$

It is also necessary to adjust the gain so that the magnitude of the loop transfer function is unity at ω_{gc}. It is desirable to make estimates that do not depend critically on a particular design method. We will first estimate the gain required to obtain a given phase lead. The following calculation gives a quantitative result.

Gain Required for Given Phase Lead

A nice estimate of the gain associated with a certain phase lead is obtained from the following result by Bode.

THEOREM 1—THE PHASE AREA FORMULA
Consider a transfer function $G(s)$ with no poles and zeros in the right half plane or on the imaginary axis. Assume that i) $\lim_{s\to\infty} G(s) = G_\infty$, ii)$A(-\omega) = A(\omega)$, and

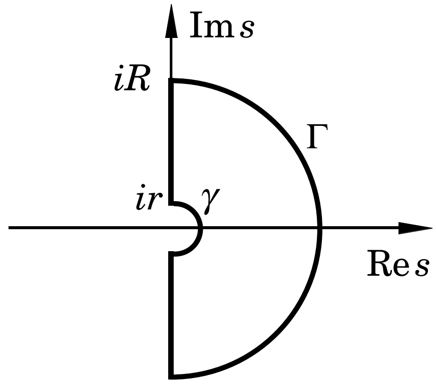

Figure 2 Integration contour used to prove the phase-area formula.

iii) $\Phi(-\omega) = -\Phi(\omega)$. Then

$$A(\infty) - A(0) = \frac{2}{\pi}\int_0^\infty \Phi(\omega)\frac{d\omega}{\omega} = \frac{2}{\pi}\int_{-\infty}^\infty \phi(u)du \tag{8}$$

PROOF 1

The proof which is outlined in Bode's book is a straightforward application of residue calculus. The integral of the function $(G(s) - G_\infty)/s$ along the contour shown in Figure 2 is zero because the function is regular inside the contour. The integral along the large semi circle goes to zero as R goes to infinity. The integral along the small semi circle can be evaluated by residue calculus. Hence

$$\begin{aligned} 0 = \int (G(s) - G_\infty)\frac{ds}{s} &= \int_{-\infty}^0 (A(\omega) - G_\infty + i\Phi(\omega))\frac{d\omega}{\omega} \\ &+ \int_0^\infty (A(\omega) - G_\infty + i\Phi(\omega))\frac{d\omega}{\omega} + i\pi(G(0) - G_\infty) \end{aligned}$$

It follows from ii) and iii) that $G(0) = A(0)$ and $G_\infty = A(\infty)$. Because of iii) the integrals of $A(\omega)$ cancel and the result follows from ii). □

The theorem is called the phase-area theorem because the gain is proportional to the area under the phase curve. It makes it possible to estimate the gain required to obtain a certain phase lead. To keep the gain as low as possible it is useful to have phase lead only over a small frequency range. To have robustness it is however required that the frequency interval is not too short. A representative phase curve is shown in Figure 3. For this curve we have $\int \phi(u)du = 2c\varphi$ and the gain becomes

$$K = e^{\frac{4c}{\pi}\varphi} = e^{2\gamma\varphi} \tag{9}$$

Reasonable values of c are in the range of 1 to 4 which gives γ between 0.6 and 1.2. In the following we will choose $\gamma = 1$.

To get a feel for the orders of magnitude involved consider an analog system where the signal levels are 10V. A measurement noise of 1 mV then saturates the input if the gain is 10^4. If it is only permitted that measurement noise gives control signals of 1V the gain must be less than 10^3. Using Equation (9) with $\gamma = 1$ and requiring a phase margin of $\varphi_m = 45°$ we then find that the maximum crossover frequency is the frequency where the open loop plant has a phase-lag of 332°.

Consider a digital system with 12 bit AD- and DA-converters a change of the input of one bit saturates the DA-converter if the gain is 4096. Assume that we permit one

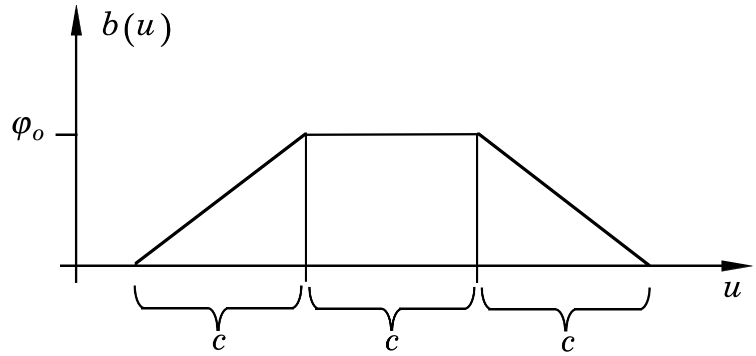

Figure 3 Representative phase curve that gives a phase lead φ.

Table 1 Maximum gain of systems that give a specified phase lead. The table is based on n first order systems with the transfer function (10).

Phase lead	n=*2*	n=*4*	n=*6*	n=*8*	$n=\infty$
90°	34	25	24	24	23
180°	-	1150	730	630	540
225°	-	14000	4800	3300	2600

bit to give a variation of 0.4% of the output range. The gain is then limited by 512. With a phase margin of 60° it means that the maximum-crossover frequency is the frequency where the process has a phase-lag of 299°. High precision analog systems with signal ranges of 1 to 10^4 have been designed. For digital systems the signal ranges are limited by the sensors and the actuators. Special system architectures with sensors and actuators having multiple signal ranges are used in order to obtain systems with a very high signal resolution. In these cases it is possible to have signal ranges up to 1 to 10^6.

The following example illustrates the gain required to obtain a given phase lead for two different systems.

EXAMPLE 1—SIMPLE LEAD NETWORK
Consider a simple lead network with the transfer function

$$G(s) = \frac{s+a}{s/K+a} \tag{10}$$

This system has its largest gain, $\max_\omega |G(i\omega)| = K$, at high frequencies. The largest phase lead

$$\varphi = \max_\omega \arg G(i\omega) = \arctan \frac{K-1}{2\sqrt{K}}$$

is obtained for $\omega = \omega_m = a\sqrt{K}$. The gain at ω_m is $|G(i\omega_m)| = \sqrt{K}$. The achievable phase lead increases with K but it is less than 90°. Several systems can be cascaded to obtain more phase lead. If n identical systems (10) with gain $\sqrt[n]{K}$ are cascaded we get

$$G_n(s) = \Big(\frac{s+a}{s/\sqrt[n]{K}+a}\Big)^n.$$

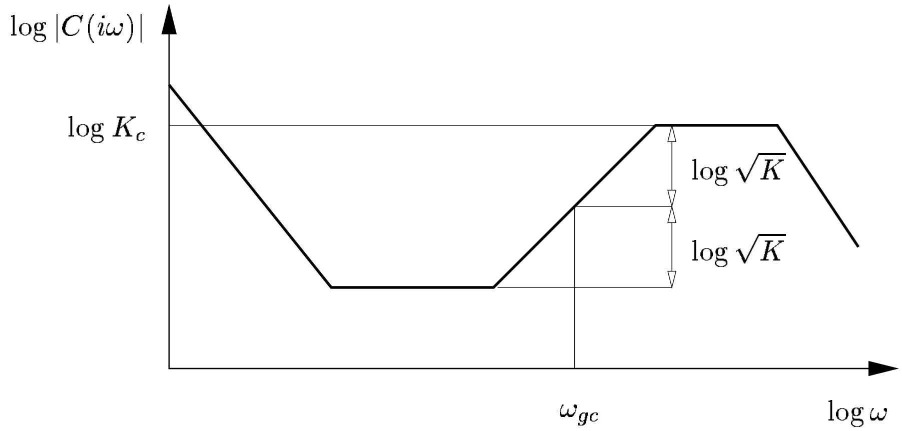

Figure 4 Asymptotes of the magnitude curve of the transfer function of a typical controller that provides phase lead.

This system has the phase lead

$$\varphi = n \arctan \frac{\sqrt[n]{K} - 1}{2\sqrt[2n]{K}}$$

. The gain K_n required to obtain a phase lead φ with n cascaded systems is

$$K_n = \Big(1 + 2\tan^2 \frac{\varphi}{n} + 2\tan \frac{\varphi}{n}\sqrt{1 + \tan^2 \frac{\varphi}{n}}\;\Big)^n .$$

Numerical values of the gain K_n is given in Table 1. Notice that it is advantageous to use systems of high order because the desired lead is obtained with less gain. As n goes to infinity we have $K_\infty = e^{2\varphi}$, which is the same gain as given by Equation (9) for $\gamma = 1$. Furthermore

$$\lim_{n\to\infty} G_n(s) = K^{\frac{s}{s+a}}.$$

Achievable Crossover Frequency

The achievable crossover frequency for minimum phase systems is essentially determined by the largest gain of the controller at high frequencies. Consider a minimum phase system with transfer function $P(s)$. Assume that it is desired to have a loop transfer function with a gain crossover frequency ω_{gc} and a phase margin φ_m. Introduce

$$K_c = \max_{\omega \ge \omega_{gc}} |C(i\omega)|$$

i.e., the largest controller gain above the crossover frequency. Figure 4 shows the magnitude curve of a typical controller transfer function. The controller gain at the crossover frequency is $1/|P(i\omega_{gc})|$. The phase of the process transfer function at ω_{gc} is $\arg P(i\omega_{gc})$. The controller must have a phase lead of $\varphi_l = -\pi + \varphi_m - \arg P(i\omega_{gc})$ at the crossover frequency. It follows from Equation (9) that this phase lead corresponds to a gain of $K = e^{2\gamma\varphi_l}$. Assuming that the controller transfer function is symmetric around the neighborhood of the crossover frequency, as shown in Figure 4, the maximum high frequency gain of the controller becomes

$$K_c = \frac{\sqrt{K}}{|P(i\omega_{gc})|} = \frac{e^{\gamma\varphi_l}}{|P(i\omega_{gc})|} = \frac{1}{|P(i\omega_{gc})|} e^{\gamma(-\pi+\varphi_m - \arg P(i\omega_{gc}))}. \tag{11}$$

Table 2 Maximum-high frequency controller gain, K_c, and phase advance, φ_l, required for a specified crossover frequency for the process $P(s) = 1/(s+1)^2$

ω_{gc}	10	20	50	100	200	500
K_c	181.5	796	5.3 10^3	2.2 10^4	8.7 10^4	5.5 10^5
φ_l	33.6	39.3	42.7	43.8	44.4	44.8
$K_c\lvert P(i\omega_{gc})\rvert$	1.8	2.0	2.1	2.1	2.2	2.2
$-\arg P(i\omega_{gc})$	168	174	178	179	179	179

Notice that K_c is a product of two factors, one is inversely proportional to the process gain at the crossover frequency, the other is the gain required to obtain the required phase-lead. The particular value of γ depends on the details of the design method, a reasonable value is $\gamma = 1$. It follows from Equation (11) that

$$\arg P(i\omega_{gc}) = -\pi + \varphi_m - \frac{\log K^*}{\gamma} \tag{12}$$

where

$$K^* = \frac{K_c}{|P(i\omega_{gc})|} = K_c|P(i\omega_{gc})|.$$

The value of K_c depends strongly on the type of controller used, the measurement noise and the saturation levels.

For proportional or PI control we have $K^* \approx 1$ and for a PID controller where the derivative term is implemented as $sT_d/(1+sT_d/N)$ we have $K^* \approx \sqrt{N}$. Choosing $\varphi_m = 45°$ and $\gamma = 1$ it follows from (12) that ω_{gc} for a P or PI controller is the frequency ω_{135} where the process has a phase lag of 135°. For a PID controller with $N = 10$ the crossover frequency can be increased to ω_{200}.

We will illustrate the results with a few examples. Consider a system with the transfer function $P(s) = 1/(s+1)^n$. We have $\arg P(i\omega) = -n \arctan \omega$, furthermore it follows from Equation (11) that

$$K_c = \left(1 + \omega_{gc}^2\right)^{n/2} e^{\gamma(n \arctan \omega_{gc} - \pi + \varphi_m)}.$$

We choose the numerical values $\gamma = 1$ and $\varphi_m = \pi/4$.

For $n = 2$ the crossover frequency with P or PI control is $\omega_{gc} = \arctan(3\pi/8) = 0.87$. Values of K_c and $K_c|P(i\omega_{gc})|$ for different values of ω_{gc} are given in Table 2. With a K_c restricted to 10000 we find that the maximum-crossover frequency is about 68.5 rad/s, almost 80 times faster than with PI control. The main contribution of the gain is to compensate for the decrease in process gain. The phase advance required is not more than 43°.

For $n = 8$ the crossover frequency with P or PI control is $\omega_{gc} = \arctan(3\pi/32) = 0.29$. Values of K_c and $K_c|P(i\omega_{gc})|$ for different values of ω_{gc} are given in Table 3. With a K_c restricted to 10000 we find that the maximum-crossover frequency is about 1.34 rad/s, which is about 5 times faster than with PI control. A significant portion of the gain is due to the phase lead required to give a phase advance of 292°. □

Table 3 Maximum-high frequency controller gain, K_c, and phase advance, φ_l, required for a specified crossover frequency for the process $P(s) = 1/(s+1)^8$

ω_{gc}	0.5	1.0	1.2	1.4	1.5	2
K_c	9.4	812	$3.7\ 10^3$	$1.5\ 10^4$	$2.7\ 10^4$	$4.2\ 10^5$
φ_l	78	225	266	300	315	372
$K_c\lvert P(i\omega_{gc})\rvert$	3.9	74	104	190	246	665
$-\arg P(i\omega_{gc})$	212	360	401	435	450	507

4. Non-Minimum Phase Systems

In the previous section we found that for minimum phase systems the achievable crossover frequency is determined by the admissible high frequency gain of the controller. For systems with poles and zeros in the right half plane there are additional restrictions that are due to the process dynamics. Consider a system with the transfer function $P(s)$. Factor the transfer function as

$$P(s) = P_{mp}(s)P_{nmp}(s) \tag{13}$$

where P_{mp} is the minimum phase part and P_{nmp} is the non-minimum phase part. Let $\Delta P(s)$ denote the uncertainty in the process transfer function. It is assumed that the factorization is normalized so that $|P_{nmp}(i\omega)| = 1$ and the sign is chosen so that P_{nmp} has negative phase. The achievable bandwidth is characterized by the gain crossover frequency ω_{gc}.

The Crossover Frequency Inequality

We will now derive an inequality for the gain crossover frequency. The loop transfer function is $L(s) = P(s)C(s)$. Requiring that the phase margin is φ_m we get.

$$\arg L(i\omega_{gc}) = \arg P_{nmp}(i\omega_{gc}) + \arg P_{mp}(i\omega_{gc}) + \arg C(i\omega_{gc}) \geq -\pi + \varphi_m. \tag{14}$$

To proceed it is necessary to make some assumptions about the controller-transfer function. We will assume that the controller is chosen so that the loop transfer function $P_{mp}C$ is equal to Bode's ideal loop transfer function given by Equation (6). Equation (14) then becomes

$$\arg P_{mp}(i\omega) + \arg C(i\omega) = n\frac{\pi}{2} \tag{15}$$

where n is the slope of the loop transfer function at the crossover frequency. If replace n by $_{gc}$ the Equation (15) is a good approximation for other controllers, with no RHP poles and zeros, because the amplitude curve is typically close to a straight line at the crossover frequency. It follows from Bode's relations (3) that the phase is $n_{gc}\pi/2$. If the controller has poles or zeros in the right half plane its non-minimum phase parts have to be included in P_{nmp}.

It follows from Equations (14) and (15) that the crossover frequency satisfies the inequality

$$\arg P_{nmp}(i\omega_{gc}) \geq -\pi + \varphi_m - n_{gc}\frac{\pi}{2}. \tag{16}$$

This condition which we call the *crossover frequency inequality* gives the limitations imposed by non-minimum phase factors. A straightforward method to assess the crossover frequencies that can be obtained for a given system is simply to plot the left hand side of Equation (16) and determine when the inequality holds. The following example gives a simple rule of thumb.

EXAMPLE 2—A SIMPLE RULE OF THUMB

To see the implications of (16) we will make some reasonable design choices. The phase margin is chosen to be 45° ($\varphi_m = \pi/4$), and the slope of the compensated minimum phase part is chosen to $n_{gc} = -1/2$. The crossover frequency inequality (16) then becomes

$$\arg P_{nmp}(i\omega_{gc}) \geq -\frac{\pi}{2}. \tag{17}$$

This gives the simple rule that the phase lag of the minimum phase components should be less than 90° at the gain crossover frequency. If the slope is instead chosen as $n = -1$ with a phase margin of 45° the phase lag of the non-minimum phase components can be at most $-45°$. With reasonable design choices we thus find that the phase lag of the non-minimum phase components should be in the range of $-45°$ to $-90°$. □

The crossover frequency inequality implies that non-minimum phase components impose restrictions on possible crossover frequencies. It also means that there are systems that cannot be controlled with sufficient stability margins. If the process has an uncertainty $\Delta P(i\omega_{gc})$ the crossover frequency inequality (16) becomes more stringent

$$\arg P_{nmp}(i\omega_{gc}) \geq -\pi + \varphi_m - n_{gc}\frac{\pi}{2} - \arg \Delta P(i\omega_{gc}).$$

A Zero in the Right Half Plane

We will now discuss limitations imposed by right half plane zeros. We will first consider systems with only one zero in the right half plane. The non-minimum phase part of the plant transfer function then becomes

$$P_{nmp}(s) = \frac{z-s}{z+s}. \tag{18}$$

Notice that P_{nmp} should be chosen to have unit gain and negative phase. We have

$$\arg P_{nmp}(i\omega) = -2\arctan\frac{\omega}{z}$$

. Since the phase of P_{nmp} decreases with frequency the inequality (16) gives an upper bound on the crossover frequency. Introducing

$$\alpha = \frac{\pi}{2} - \frac{\varphi_m}{2} + n_{gc}\frac{\pi}{4} \tag{19}$$

the crossover frequency inequality, Equation (16), can be written as

$$\frac{\omega_{gc}}{z} \leq \tan\alpha. \tag{20}$$

With $\varphi_m = \pi/4$ and $n_{gc} = -1/2$ which corresponds to the simple rule of thumb (17) we get $\omega_{gc} < z$. With $\varphi_m = \pi/4$ and $n_{gc} = -1$ we get $\omega_{gc} \leq 0.4z$.

A right half plane zero gives an upper bound to the achievable bandwidth. The bandwidth decreases with decreasing frequency of the zero. It is thus more difficult to control systems with slow zeros.

Time Delays

The transfer function for such systems has an essential singularity at infinity. The non-minimum phase part of the transfer function of the process is

$$P_{nmp}(s) = e^{-sL}. \tag{21}$$

We have

$$\arg P_{nmp}(i\omega) = -\omega L$$

. It follows from the crossover frequency inequality, Equation (16), that

$$\omega_{gc} L \leq \pi - \varphi_m + n_{gc}\frac{\pi}{2} = 2\alpha. \tag{22}$$

The simple rule of thumb (17) gives $\omega_{gc} L \leq \frac{\pi}{2} = 1.57$. Approximating the time delay by a first order Pade' approximation and using the results for systems with a right half plane zero gives $\omega_{gc} L \leq 2$.

Time delays thus give an upper bound on the achievable bandwidth.

A Pole in the Right Half Plane

Consider a system with one pole in the right half plane. The non-minimum phase part of the transfer function is thus

$$P_{nmp}(s) = \frac{s+p}{s-p} \tag{23}$$

where $p > 0$. Notice that the transfer function is normalized so that P_{nmp} has unit gain and negative phase. We have

$$\arg P_{nmp}(i\omega) = -2 \arctan \frac{p}{\omega}$$

. It follows from the crossover frequency inequality, Equation (16), that

$$-2 \arctan \frac{p}{\omega_{gc}} \geq -\pi + \varphi_m + n_{gc}\frac{\pi}{2} = -2\alpha$$

where α is given by (19). Hence

$$\omega_{gc} \geq \frac{p}{\tan \alpha}. \tag{24}$$

The simple rule of thumb (17), which corresponds to $\varphi_m = \pi/4$ and $n_{gc} = -1/2$ gives $\omega_{gc} \geq p$. Requiring that $\varphi_m = \pi/4$ and $n_{gc} = -1$ we get instead $\omega_{gc} \geq 2.4p$. □

Unstable poles thus give a lower bound on the crossover frequency. For systems with right half plane poles the bandwidth must thus be sufficiently large. By computing the phase lag of the minimum phase part of the system at ω_{gc} we can determine the phase lead required to compensate the minimum phase part. Equation (9) then gives the gain required to achieve the phase lead. Knowledge of the measurement noise and the saturation levels of the control signal then indicates the feasibility of stabilizing the system.

Table 4 Achievable phase margin for for $\varphi_m = \pi/4$ and $n_{gc} = -1/2$ and different zero-pole ratios z/p.

z/p	2	5	10	20
φ_m	-6.0	38.6	64.8	84.6

Table 5 Ratio z/p required to give a specified phase margin for $\varphi_m = \pi/4$ and $n_{gc} = -1/2$.

φ_m	0	30	45	60
z/p	2.24	3.86	5.83	8.68

Poles and Zeros in the Right Half Plane

The calculations can be extended to cases with both poles and zeros in the right half plane. We will give the results for a pole-zero pair. The non-minimum phase part of the transfer function is then

$$P_{nmp}(s) = \frac{(z-s)(s+p)}{(z+s)(s-p)}. \tag{25}$$

For $z > p$ we have

$$\arg P_{nmp}(i\omega) = -2\arctan\frac{\omega}{z} - 2\arctan\frac{p}{\omega} = -2\arctan\frac{\omega_{gc}/z + p/\omega_{gc}}{1 - p/z}.$$

The right hand side has its maximum for $\omega = \sqrt{pz}$, hence

$$\arg P_{nmp}(i\omega) \leq -4\arctan\sqrt{\frac{p}{z}}$$

and the inequality (16) becomes

$$\frac{z}{p} \geq \tan^2\frac{\alpha}{2} = \tan^2\Bigl(\frac{\pi}{4} - \frac{\varphi_m}{4} + n_{gc}\frac{\pi}{8}\Bigr). \tag{26}$$

Introducing the values $\varphi_m = \pi/4$ and $n_{gc} = -1/2$ into (26) gives $z \geq 5.83p$. With $\varphi_m = \pi/4$ and $n_{gc} = -1$ equation (26) gives $z \geq 25.3p$. Table 4 gives the phase margin as a function of the ratio z/p for $\varphi_m = \pi/4$ and $n_{gc} = -1/2$. The phase-margin that can be achieved for a given ratio p/z is

$$\varphi_m < \pi + n_{gc}\frac{\pi}{2} - 4\arctan\sqrt{\frac{p}{z}}. \tag{27}$$

Table 5 gives the zero-pole ratio as a function of the phase margin.

When the unstable zero is faster than the unstable pole, i.e. $z > p$, the ratio z/p thus must be sufficiently large in order to have the desired phase margin. The best gain crossover frequency is the geometric mean of the unstable pole and zero.

Example 3
Control of a system with the transfer function

$$P(s) = \frac{s-1}{s^2+0.5s-0.5}$$

is discussed in Doyle *et al.* (1992) and Keel and Bhattacharyya (1997). The system has a zero at $s = 1$ and a pole at $s = 0.78$. The ratio $z/p = 1.28$ is so small so there is no controller that will give a reasonably robust closed-loop system. □

Example 4—The X-29
Considerable design effort has been devoted to the design of the flight control system for the X-29 aircraft, see Clarke *et al.* (1994) and Huang and Knowles (1990). One of the design criteria was that the phase margin should be greater than 45° for all flight conditions. At one flight condition the model has the following non-minimum phase component

$$P_{nmp}(s) = \frac{s-26}{s-6}$$

Since $z = 4.33p$, it follows from Equation 27 that the achievable phase margins for $n_{gc} = -0.5$ and $n_{gc} = -1$ are $\varphi_m = 32.3°$ and $\varphi_m = -12.6°$. It is interesting to note that many design methods were used in a futile attempt to reach the design goal. A simple calculation of the type given in this paper would have given much insight. We illustrate the results with a few examples. □

Example 5—Klein's Unridable Bicycle
An interesting bicycle with rear wheel steering which is impossible to ride was designed by Professor Klein in Illinois, see Klein (1986). The theory presented in this paper is well suited to explain why it is impossible to ride this bicycle. The transfer function from steering angle to tilt angle is given by

$$\frac{\theta(s)}{\delta(s)} = \frac{m\ell V}{Jc}\frac{V-as}{s^2-mg\ell/J}$$

where m is the total mass of the bicycle and the rider, J the moment of inertia for tilt with respect to the contact line of the wheels and the ground, h the height of the center of mass from the ground, a the vertical distance from the center of mass to the contact point of the front wheel, V the forward velocity, and g the acceleration of gravity. The system has a RHP pole at $s = p = \sqrt{mg\ell/J}$, caused by the pendulum effect. Because of the rear wheel steering the system also has a RHP zero at $s = z = V/l$. Typical values $m = 70$ kg, $\ell = 1.2$ m, $a = 0.7$, $J = 120$ kgm^2 and $V = 5$ m/s, give $z = V/a = 7.14$ rad/s and $p = \omega_0 = 2.6$ rad/s. The ratio of the zero and the pole is thus $p/z = 2.74$, with $n_{gc} = -0.5$ the inequality (16) shows that the phase margin can be at most $\varphi_m = 10.4$.

The reason why the bicycle is impossible to ride is thus that the system has a right half plane pole and a right half plane zero that are too close together. Klein has verified this experimentally by making a bicycle where the ratio z/p is larger. This bicycle is indeed possible to ride. □

So far we have only discussed the case $z > p$. When the unstable zero is slower than the unstable pole the crossover frequency inequality (16) cannot be satisfied unless $\varphi_m < 0$ and $n_{gc} > 0$. The closed-loop systems have low gain and poor control at low frequencies, as will be discussed in Section 5.

A Pole in the Right Half Plane and Time Delay

Consider a system with one pole in the right half plane and a time delay T. The non-minimum phase part of the transfer function is thus

$$P_{nmp}(s) = \frac{s+p}{s-p} e^{-sT}. \tag{28}$$

It follows from the the crossover condition, Equation (16), that

$$2 \arctan \frac{\omega_{gc}}{p} - \omega_{gc} T \geq \varphi_m - n_{gc} \frac{\pi}{2}. \tag{29}$$

The system cannot be stabilized if $pT > 2$. If $pT < 2$ the left hand side has its smallest value for $\omega_{gc}/p = \sqrt{2/(pT) - 1}$. Introducing this value of ω_{gc} into (29) we get

$$2 \arctan \sqrt{\frac{2}{pT} - 1} - pT \sqrt{\frac{2}{pT} - 1} > \varphi_m - n_{gc} \frac{\pi}{2}$$

. The simple rule of thumb with to $\varphi_m = \pi/4$ and $n_{gc} = -0.5$ gives

$$pT \leq 0.326 \tag{30}$$

and $\varphi_m = \pi/4$ and $n_{gc} = -1$ gives $pT \leq 0.0781$.

EXAMPLE 6—POLE BALANCING

To illustrate the results we can consider balancing of an inverted pendulum. A pendulum of length ℓ has a right half plane pole $\sqrt{g/\ell}$. Assuming that the neural lag of a human is 0.07 s. The inequality (30) gives $\sqrt{g/l}\ 0.07 < 0.326$, hence $\ell > 0.45$. The calculation thus indicate that a human with a lag of 0.07 s should be able to balance a pendulum whose length is 0.5 m. To balance a pendulum whose length is 0.1 m the time delay must be less than 0.03s. Pendulum balancing has also been done using video cameras as angle sensors. The limited video rate imposes strong limitations on what can be achieved. With a video rate of 20 Hz it follows from (30) that the shortest pendulum that can be balanced with $\varphi_m = 45°$ and $n_{gc} = -0.5$ is $\ell = 0.23$m. □

The calculations can be extended to complex poles and zeros and to multiple singularities.

5. Other Criteria

The phase margin and the slope at crossover are quite crude measures. The advantage in using them is that it is very easy to obtain the crossover frequency inequality, (16). To get a deeper understanding we will introduce other criteria and carry out a more detailed design for some of the cases. The criteria we will consider are the maximal values of the sensitivity function $S = 1/(1+L)$ and the complementary sensitivity function $T = L/(1+L)$, i.e.

$$M_s = \max_\omega |S(i\omega)| \qquad M_t = \max_\omega |T(i\omega)|.$$

Right Half Plane Zeros

A system with a right half plane zero at $s = 1$ is first considered. We will carry out a design where the minimum phase part is compensated to give Bodes ideal loop transfer function. The loop transfer function is

$$L(s) = \Big(\frac{s}{\omega_{gc}}\Big)^n \frac{1-s}{1+s}$$

. The simple rule (16) with $\varphi_m = \pi/4$ and $n_{gc} = -0.5$ gives $\omega_{gc} = 1$. This results in a loop transfer function with $M_s = 3.2$ and $M_t = 2.3$. Reducing the crossover frequency to $\omega_{gc} = 0.5014$ gives a loop transfer function with $M_s = 2$ and $M_t = 1.1257$. The crossover frequency must be reduced further to get lower sensitivities, with $n_{gc} = -1$, and $\omega_{gc} = 0.2044$ we get $M_s = 1.4$ and $M_t = 1.0$. The sensitivity constraint is more critical than the constraint on the complementary sensitivity for systems having a right half plane zero.

Time Delay

A system with a time delay will now be considered. We will carry out a design where the minimum phase part is compensated to give Bodes ideal loop transfer function. The loop transfer function is

$$L(s) = \Big(\frac{s}{\omega_{gc}}\Big)^n e^{-s}$$

. The simple design rule (16), with $\varphi_m = \pi/4$ and $n_{gc} = -0.5$ gives $\omega_{gc} = 0.78$, $M_t = 1.45$ and $M_s = 2.21$. By reducing the value of ω_{gc} to 0.7 we get $M_t = 1.26$ and $M_s = 2.0$. To bring M_s down to 1.4 ω_0 has to be reduced to 0.37.

Right Half Plane Pole

Consider a system with a right half plane pole at $s = 1$. We will make a design where the minimum phase part is compensated to give Bodes ideal loop transfer function. The loop transfer function is then

$$L(s) = \Big(\frac{s}{\omega_{gc}}\Big)^n \frac{s+1}{s-1}$$

The simple design rule (16), with $\varphi_m = \pi/4$ and $n_{gc} = -0.5$ gives $\omega_{gc} = 1.99$. This gives a loop transfer function with $M_s = 1.4$ and $M_t = 2.06$. Choosing $\varphi_m = \pi/4$ and $n_{gc} = -1$ gives $\omega_{gc} = 4.9$ and $M_t = 1.40$ and $M_s = 1.0$. The constraint on the complementary sensitivity is more critical than the sensitivity constraint for systems having a right half plane pole.

Right Half Plane Pole-Zero Pair

For systems with one right half plane pole-zero pair the $\mathcal{H}_\infty$ theory gives the following bounds on the sensitivity functions, see Khargonekar and Tannenbaum (1985)

$$M_s = M_t = \Big|\frac{z+p}{z-p}\Big|..$$

These inequalities imply that

$$\frac{z}{p} \geq \frac{M+1}{M-1}, \quad \text{or} \quad \frac{z}{p} \leq \frac{M-1}{M+1}$$

where M is M_s or M_t. It follows from these inequalities that to have $M \leq 2$ we must have either $z/p \geq 3$ or $p/z \geq 3$. Notice that the controllers that minimize M_s and M_t are not the same. If we want to guarantee that both M_s and M_t are less than M the inequalities become

$$\frac{z}{p} \geq \Big(\frac{M}{M-1}\Big)^2, \quad \text{or} \quad \frac{z}{p} \leq \Big(\frac{M-1}{M}\Big)^2. \tag{31}$$

These equations imply that to have $M_s \leq 2$ and $M_t \leq 2$ we must have either $z/p \geq 4$ or $p/z \geq 4$. To have sensitivities less than 1.4 we must have $z/p \geq 12.25$ or $p/z \geq 12.25$. The simple estimates based on (16) give more conservative values, one reason being that these estimates are based on the assumption that the loop transfer function has infinite gain at zero frequency. To see this we will make a design.

Consider a system with a right half plane pole-zero pair where the zero is faster than the pole. Assume $c > 1$ and let $z = 1/\sqrt{c}$ and $p = \sqrt{c}$. The ratio of the RHP zero to the RHP pole is thus equal to $z/p = c > 1$. A design where the minimum phase part is compensated to give Bodes ideal loop transfer function the loop transfer function

$$L(s) = \Big(\frac{s}{\omega_{gc}}\Big)^n \frac{(\sqrt{c}-s)(s+1/\sqrt{c})}{(\sqrt{c}+s)(s-1/\sqrt{c})}$$

. A design with $n_{gc} = -0.5$ and a phase margin of 45° has $c = 5.83$ and $M_s = M_t = 2.1$. The zero-pole ratio has to be increased to 6.5 to obtain $M_s = M_t = 2$. The zero-pole ratio has to be increased to $|z/p| \geq 14.4$ to obtain $M_s = M_t = 1.4$.

To illustrate what happens when a system has a right half plane pole-zero pair where the pole is faster than the zero we will consider an example.

Example 7—Slow RHP zero and fast RHP pole

Consider a system with the transfer function

$$P(s) = \frac{z-s}{s-p} \tag{32}$$

where $z < p$. The characteristic equation of the closed loop is

$$k(s-z) + p - s = (k-1)s + p - kz = 0.$$

To have a stable closed-loop system the gain must be in the range $1 < k < p/z$. With k in this range the Nyquist curve is a semi circle whose diameter is the interval $-k \leq L \leq -kz/p$. The complete Nyquist contour encloses the critical point which is necessary in order to have a stable closed-loop system. Adjusting the gain so that the maxima of the sensitivity and the complementary sensitivity are both equal to M and as small as possible we get

$$k = \sqrt{p/z}, \qquad M = \frac{p+\sqrt{pz}}{p-z}, \qquad \frac{p}{z} = \Big(\frac{M}{M-1}\Big)^2.$$

Compare with (31).

Notice that the gain of the loop transfer function is less than one at low frequencies and both the phase margin and slope are negative at the crossover frequency. □

The example shows that it is possible to stabilize a process with a right half plane pole-zero pair with $z < p$ with a proportional controller. The process in the example had finite gain for high frequencies. In Youla and Jabr (1974) it is shown that if the process is proper it cannot be stabilized with a stable controller unless $z > p$.

Right Half Plane Pole and Time Delay

Consider a system with a right half plane pole at $s = 1$ and a time delay T. A design where the minimum phase part is compensated to give Bodes ideal loop transfer function has the loop transfer function

$$L(s) = \left(\frac{s}{\omega_{gc}}\right)^n \frac{s+1}{s-1} e^{-sT}$$

. According to the inequality (30) the condition $T < 0.326$ must be fulfilled in order to have a phase margin of $\pi/4$. The time delay has to be decreased in order to obtain a more robust design. For $T = 0.156$ we get $M_s = 2.00$ and $M_t = 1.69$. With $n = -1$ and $T = 0.0545$ we get $M_s = 1.40$ and $M_t = 1.38$.

Summary

The phase margin is a crude indicator of the stability margin. By carrying out detailed designs the results obtained can be refined and other criteria can be considered. In this section we have made designs that give $M_s = M_t = 2$ and $M_s = M_t = 1.4$. The results give the following design inequalities.

- A RHP zero z

$$\frac{\omega_{gc}}{z} \le \begin{cases} 0.5 & \text{for } M_s,\ M_t < 2 \\ 0.2 & \text{for } M_s,\ M_t < 1.4. \end{cases}$$

- A time delay T

$$\omega_{gc} T \le \begin{cases} 0.7 & \text{for } M_s,\ M_t < 2 \\ 0.37 & \text{for } M_s,\ M_t < 1.4. \end{cases}$$

- A RHP pole p

$$\frac{p}{\omega_{gc}} \ge \begin{cases} 2 & \text{for } M_s,\ M_t < 2 \\ 5 & \text{for } M_s,\ M_t < 1.4. \end{cases}$$

- A RHP pole-zero pair with $z > p$

$$\frac{z}{p} \ge \begin{cases} 6.5 & \text{for } M_s,\ M_t < 2 \\ 14.4 & \text{for } M_s,\ M_t < 1.4. \end{cases}$$

- A RHP pole p and a time delay T

$$pT \le \begin{cases} 0.16 & \text{for } M_s,\ M_t < 2 \\ 0.05 & \text{for } M_s,\ M_t < 1.4. \end{cases}$$

A time delay or a zero in the right half plane gives an upper bound of the bandwidth that can be achieved. The bound decreases when the zero z decreases and the time delay increases. A pole in the right half plane gives a lower bound on the bandwidth. The bandwidth increases with increasing p. For a pole zero pair there is a upper bound on the pole-zero ratio.

The inequalities obtained agree qualitatively with the investigation of Middleton (1991), who proposes $\omega_B \le 2z$ for a right half plane zero and $\omega_B \ge 6p$ for a right half plane pole, $z/p \ge 6$ for a right half plane pole-zero pair, and $\omega_B T < 0.2$ for a time delay. Middleton bases his results on the bandwidth which is larger than the crossover frequency. Bounds based on the bandwidth are also given in Skogestad and Postlethwaite (1996). They propose $\omega_B \le z M_s/(M_s - 1)$ for a right half plane zero and $\omega_B \ge p M_t/(M_t - t)$ for a right half plane pole, $z/p \ge (M+1)/(M-1)$ for a right half plane pole-zero pair, and $\omega_B L < 1$ for a time delay.

6. A Paradox in Pole-Placement Design

It is well-known that if a system is controllable and observable it is possible to find a feedback which gives arbitrary closed-loop poles. In the previous section it has been shown that time delays and right half plane zero restrict the crossover frequency that can be obtained. We then have the paradox: How can the crossover frequency be limited when the closed-loop poles can be made arbitrarily fast? When resolving this paradox we also obtain rules for choosing reasonable closed-loop poles in a pole-placement design. We start with a simple example.

EXAMPLE 8—A FAST SYSTEM WITH A LOW BANDWIDTH
Consider a process with the transfer function

$$P(s) = \frac{as+1}{s(s+1)}. \tag{33}$$

The zero is unstable if $a < 0$. It will be assumed that $a = -10$. It then follows from the crossover frequency-inequality (16) that it is difficult to obtain crossover frequencies higher than 0.1 rad/s.

To carry out a pole-placement design assume that it is desirable to have a closed-loop system with the characteristic polynomial

$$(s + \alpha\omega_0)(s^2 + \omega_0 s + \omega_0^2). \tag{34}$$

This can be accomplished with a controller with the transfer function

$$C(s) = \frac{s_0 s + s_1}{s + r} \tag{35}$$

where

$$s(s+1)(s+r) + (as+1)(s_0 s + s_1) = (s + \alpha\omega_0)(s^2 + \omega_0 s + \omega_0^2).$$

Equating equal powers of s gives a linear equation which has the solution

$$\begin{aligned} r &= \frac{-\alpha a^2\omega_0^3 + (1+\alpha)a\omega_0^2 - (1+\alpha)\omega_0 + 1}{a-1} \\ s_0 &= \frac{\alpha a\omega_0^3 - (1+\alpha)\omega_0^2 + (1+\alpha)\omega_0 - 1}{a-1} \\ s_1 &= \alpha\omega_0^3. \end{aligned} \tag{36}$$

To obtain a fast closed-loop system we choose $\omega_0 = 10$ and $\alpha = 1$. Equation (36) then gives the controller parameters $r = 9274.5$, $s_0 = 925.5$ and $s_1 = 1000$. The loop transfer function is

$$L(s) = \frac{(as+1)(s_0 s + s_1)}{s(s+1)(s+r)} = -9255\frac{(s-0.1)(s+1.0805)}{s(s+1)(s+9274)}.$$

The process pole at $s = -1$ is almost canceled by the controller zero at $s = -s_1/s_0 = -1.0805$. The Bode diagram of the loop transfer function is shown in Figure 5 and the Nyquist curve is shown in Figure 6.

The loop transfer function has a low frequency asymptote that intersects the line $\log|L(i\omega)| = 0$ at $\omega = -1/a$, i.e. at the slow unstable zero. The magnitude then

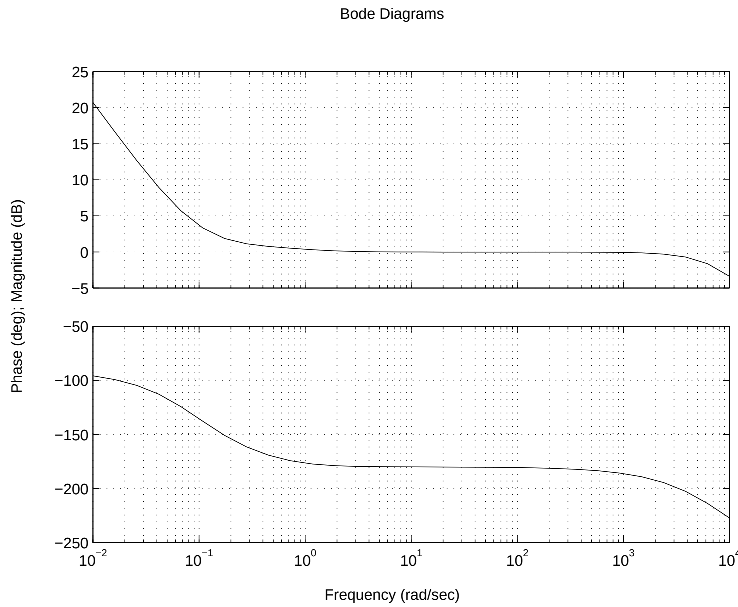

Figure 5 Bode diagram for the loop transfer function of a system with a slow unstable zero at $s = 0.1$, where the specifications are $\omega_0 = 10$ and $\alpha = 1$.

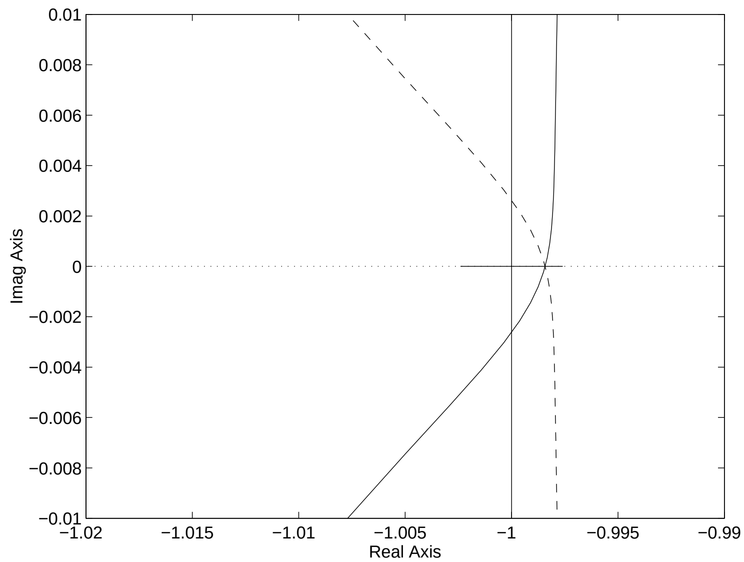

Figure 6 Nyquist curve for the loop transfer function of a system with a slow unstable zero at $s = 0.1$. The specifications for the controller are, $\omega_0 = 10$ and $\alpha = 1$.

becomes close to one and it remains so until the the break point at $\omega = r \approx \alpha a \omega_0^3$, i.e. the controller pole. The phase is also close to $-180°$ over that frequency range which means that the stability margin is very poor. The crossover frequency is 6.58 and the phase margin is $\varphi_m = 0.15°$. The maximum sensitivities are $M_s = 678$ and $M_t = 677$ which also shows that the system is extremely sensitive. The slope of the magnitude curve at crossover is also very small which is another indication of the poor robustness of the system. □

The example illustrates clearly the danger of using a design method in a routine manner. It also shows that it is not sufficient to check controllability. For this particular problem there are severe limitations caused by the right half plane zero.

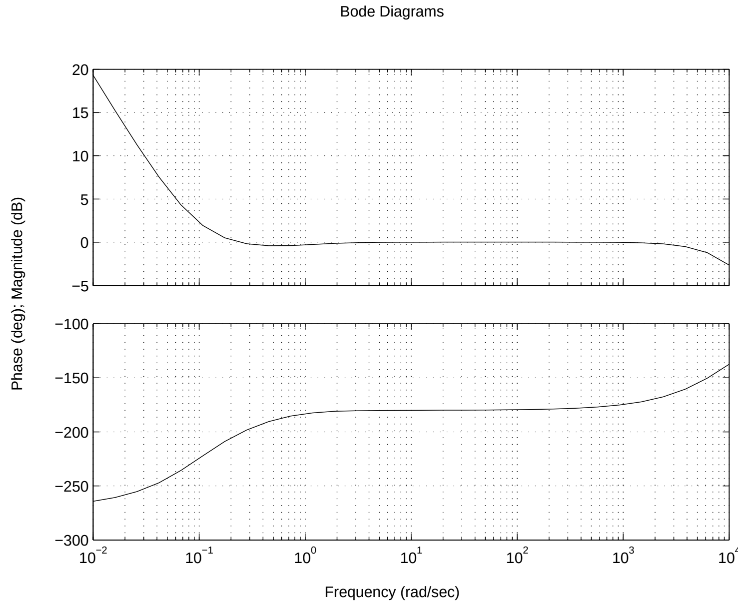

Figure 7 Bode diagram for the loop transfer function of a system with a slow stable zero, $a = 10$. The specifications are $\omega_0 = 10$ and $\alpha = 1$.

Trying to make designs which violate these limitations by making a closed-loop system that is too fast we obtain a closed-loop system that has very poor stability margins even if the closed-loop poles are quite well damped. Also notice that even if the gain crossover frequency is 6.58 rad/s the sensitivity becomes larger than one for $\omega = 0.107$, which is close to the right half plane zero. Feedback is thus not effective for disturbances having higher frequencies than 0.107, because disturbances will be amplified by the feedback. The example shows that it is important to be aware of the limitations when designing control systems. It is of course possible to obtain sensible control designs, for example by choosing smaller values of ω_0.

Next we will investigate what happens if a system has slow but stable zeros. We will consider a system similar to the one in Example 8 but we will assume that the system has a slow zero in the left half plane.

EXAMPLE 9—A SLOW STABLE ZERO

Consider the system given by Equation (33) with $a = 10$, which means that the system has a zero at $s = -0.1$. Let the desired closed-loop characteristic polynomial be given by (34). Equation (36) then gives the controller parameters $r = -10891$, $s_0 = 1091$ and $s_1 = 1000$. The controller thus has an unstable pole $p = 10891$ in the right half plane. Since the magnitude of this pole is very large it follows from Bode's integral that the closed-loop system is very sensitive. The loop transfer function is

$$L(s) = \frac{(as+1)(s_0 s + s_1)}{s(s+1)(s+r)} = 1091 \frac{(s+0.1)(s+0.9166)}{s(s+1)(s-10891)}.$$

The process pole at $s = -1$ is almost canceled by the controller zero at $s = -s_1/s_0 = -0.9166$. The Bode diagram of the loop transfer function is shown in Figure 7. The amplitude curve is similar to the one in Figure 5. The system has a very small phase margin and it is extremely sensitive. The design is thus very poor. One indication of this is the presence of the unstable mode in the controller.

Let us now investigate how the design should be changed to give a stable controller. To be specific we will investigate the consequences of changing the parameter

α which gives the location of the real closed-loop pole. With $a = 10$ and $\omega_0 = 10$ it follows from Equation (36) that

$$
\begin{aligned}
r &= \frac{991 - 99010\alpha}{9} \\
s_0 &= \frac{-91 + 9910\alpha}{9} \\
s_1 &= 100\alpha.
\end{aligned}
$$

These equations show that the value of α is critical. To have a stable controller α must be less than 0.01001, and to avoid that the controller has a zero in the right half plane α has to be larger than 0.00918. The parameter α must thus be in the narrow range $0.00918 < \alpha < 0.01001$ in order to give good control. Let us choose $\alpha = 0.01$, ($\alpha = 1/a\omega_0$ in general) which means that the slow process zero is canceled by a slow controller pole. The controller parameters then become

$$
\begin{aligned}
r &= \frac{1}{a} \\
s_0 &= \omega_0 - 1 \\
s_1 &= \omega_0^2
\end{aligned}
$$

and the loop transfer function becomes

$$
L(s) = \frac{(\omega_0 - 1)s + \omega_0^2}{s(s+1)}.
$$

For low frequencies we have $L(s) \approx \omega_0^2/s$. The loop transfer function thus has a low frequency asymptote then intersects at $s = \omega_0$. There is a pole at $s = -1$ and a zero at $s = -\omega_0^2/(\omega_0 - 1)$. With $\omega_0 = 10$ and $\alpha = 0.01$ the controller parameters become $r = 0.1$, $s_0 = 0.9$ and $s_1 = 10$. The Bode diagram of the system in Figure 8 indicates that the design is quite robust. The phase margin is $\varphi_m = 52°$. □

The example shows that the selection of closed-loop poles has a drastic influence on the robustness of the closed-loop system obtained. In the particular example where the system has a slow mode it was advantageous to have a process zero close to the slow process pole. A major difference between the non-minimum phase case is that it is not possible to cancel right half plane zeros. The problem of pole cancellation is also discussed in Graebe and Middleton (1995).

Properties of the Sensitivity Functions

A very good insight into the phenomena observed in Examples 8 and 9 can be obtained by investigating the sensitivity function and the complementary sensitivity function. Let the process-transfer function be $P = N_p/D_p$, and let the controller-transfer function be $C = N_c/D_c$. The sensitivity functions then becomes

$$
\begin{aligned}
S &= \frac{D_pD_c}{N_pN_c + D_pD_c} = \frac{D_pD_c}{D_{cl}} \\
T &= \frac{N_pN_c}{N_pN_c + D_pD_c} = \frac{N_pN_c}{D_{cl}}
\end{aligned}
$$

where D_{cl} denotes the characteristic polynomial of the closed-loop system.

Let us first investigate the case when the process has a slow zero and the closed-loop poles are considerably faster. Consider the amplitude curve of the Bode diagram

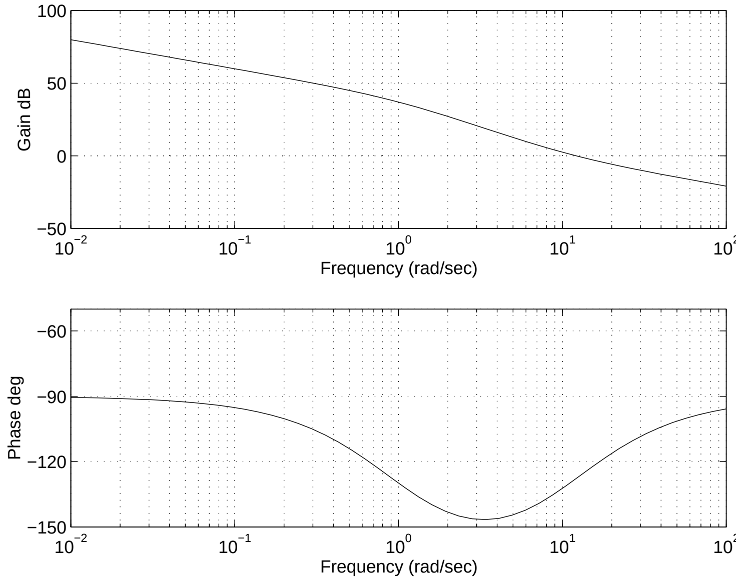

Figure 8 Bode diagram for the loop transfer function of a system with a slow stable zero, $e = 10$. The specifications are, $\omega_0 = 10$ and $\alpha = 0.01$.

of the complementary sensitivity function T. For low frequencies the function has a horizontal asymptote. The first break point occurs at the slow process zero. At that zero the sensitivity function has an asymptote with slope 1 and the complementary sensitivity function increases. If there are more slow zeros the function increases even faster. The magnitude curve starts to decrease for frequencies corresponding to the closed-loop poles. It thus follows that the complementary sensitivity can be very large if there is a wide gap between the slowest process zero and the slowest closed-loop pole.

We will now investigate the case when the process or the controller has a pole that is faster than the closed-loop poles. The sensitivity function has the property $S(\infty) = 1$. At high frequencies the Bode diagram has a horizontal asymptote. The sensitivity function will increase with decreasing frequency if there is a fast process or controller pole until a closed-loop pole is encountered. The sensitivity function will be larger if the fast process pole is smaller than the fastest closed-loop pole. We illustrate with an example.

EXAMPLE 10—EFFECT OF A SLOW ZERO

The sensitivity functions of the system in Example 8 are

$$
\begin{aligned}
S(s) &= \frac{s(s+1)(s+9274)}{(s+10)(s^2+10s+100)} \\
T(s) &= \frac{9255(0.1-s)(s+1.0805)}{(s+10)(s^2+10s+100)}
\end{aligned}
$$

The Bode diagrams of these transfer functions are shown in Figure 9. The magnitude curve of the complementary sensitivity is unity at low frequencies. It starts to increase with slope 1 at the first break point which corresponds to the slow zero at $\omega = 0.1$. The next break point is close to $\omega = 1$, the sensitivity function then increases with slope 1 until the next break point approximately at $\omega = 1$ where the slope increases to 2. The function continues to increase until the break points that corresponds to the poles which are located approximately at $\omega = 10$ where the function has its

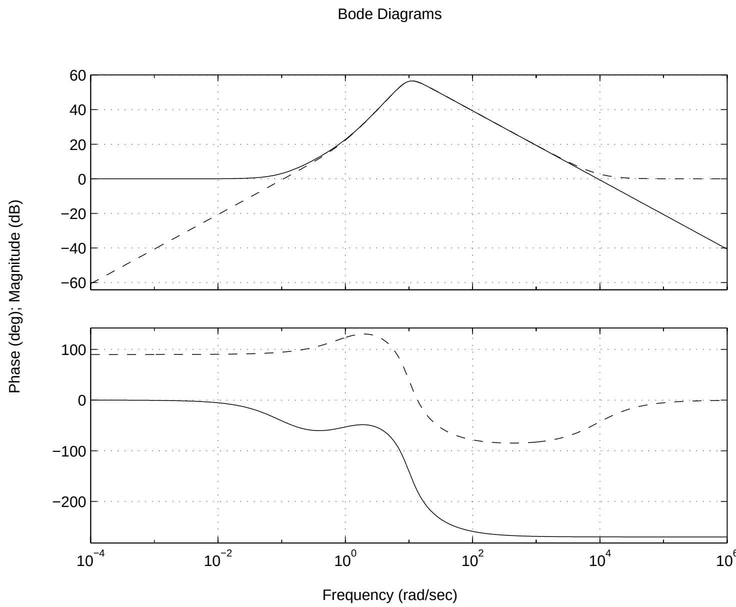

Figure 9 Bode diagrams of the sensitivity function (full line) and the complementary sensitivity function (dashed line).

largest value. A crude estimate based on the asymptotes gives a peak of about 1000. The actual maximum of the complementary sensitivity is $M_t = 677$.

Next we will consider the sensitivity function. We will explore the magnitude curve starting at high frequencies. The first break point occurs at $\omega = 9274$, the sensitivity function then increases until the next break point is reached at $\omega = 10$. A crude estimate based on the asymptotes gives a peak of about 927. The actual maximum of the complementary sensitivity is $M_s = 678$. □

Relations to Cheap Control and Peaking

The phenomena observed in this section are closely related to results on cheap control, see Bongiorno and Youla (1968), Kwakernak (1969), Bongiorno and Youla (1970), Kwakernak and Sivan (1972), and peaking Mita (1977), Francis and Glover (1978), Sussman and Kokotovic (1991).

Let $A(s)$ and $B(s)$ be polynomials of degree n and m, consider a single-input single-output system with transfer function $P(s) = B(s)/A(s)$. Assume that there are no zeros on the imaginary axis and let $B(s)$ be factored as $B^+(s)$ and $B^-(s)$ where B^+ has all its zeros in the left half plane and $B^-(s)$ has all its zeros in the right half plane. The problem of minimizing the criterion

$$J = \int_0^\infty \big(y^2(t) + \rho u^2(t)\big)dt$$

was investigated in Kwakernak (1969), see also Kwakernak and Sivan (1972). The closed-loop poles of the optimal system were determined as the parameter ρ goes to zero. It was shown that m closed-loop poles converge to the zeros of $B^+(s)B^-(-s)$ and that $n - m$ zeros goes to infinity in a Butterworth pattern. This result indicate that stable system zeros are cancelled in an optimal controller with fast response.

The phenomena of peaking implies that some state variables may have very large excursions in the output even for small disturbances. This occurs if it is attempted to assign stable poles far from the origin. In Mita (1977) it was shown that for single-input single-output systems there are initial values in an m dimensional subspace that will give arbitrary large deviations in the state variables for feedbacks that give stable but fast closed-loop poles.

The fact that a fast closed-loop system with a slow stable zero should have a slow closed-loop pole close to the slow zero was also well-known in the typical dipole pattern that was recommended in classical servo systems, see Truxal (1955).

Design Rules

We thus find that the closed-loop poles should be chosen carefully in a pole placement design. Based on the insight derived from the examples the following rules can be suggested:

- For minimum phase systems which have zeros that are slower than the dominant closed-loop poles we should position additional closed-loop poles which are slightly faster than the process zeros. If there are open loop poles that are significantly faster than the dominant poles we should position additional closed-loop poles that are slightly slower than the fast poles. This means that there are controller poles close to the zeros. In such a case a two-degree of freedom configuration should be chosen so that the slow controller poles are not excited by set point changes.
- For non-minimum phase systems the design inequalities should be used to select the dominant closed-loop poles. Slow unstable zeros and fast unstable poles impose severe limitations on what can be achieved.

Even if we have focused on the pole placement design method similar phenomena occur for other design techniques. There is no way to avoid the fundamental limitations and their consequences for design.

7. Conclusions

In this paper we have formulated and solved several problems that make it possible to estimate the limitations on control-system performance that are imposed by measurement noise, actuator saturation, and singularities of the transfer function in the right half plane. For minimum phase systems the crossover frequency is given by Equation (12). For non-minimum phase systems the crossover frequency is instead determined by Equation (16) which implies that the minimum phase part of the open loop transfer function at ω_{gc} has a phase lag in the range of $\pi/4$ to $\pi/2$. From these conditions it is possible to derive simple design rules. The results can be taught in elementary courses in feedback control. The consequences of making designs without considering the limitations have been illustrated.

8. References

Anderson, B. D. O. and J. B. Moore (1990): *Optimal Control Linear Quadratic Methods*. Prentice Hall, Englewood Cliffs, N.J.

Åström, K. J. (1968): *Reglerteori (Control Theory)*. (In Swedish). Almqvist & Wiksell, Uppsala, Sweden.

Åström, K. J. (1993): "Autonomous controllers." *Control Eng. Practice*, **1:2**, pp. 227–232.

Åström, K. J. (1996): "Fundamental limitations of control system performance." In Paulraj *et al.*, Eds., *Mathematical Engineering: A Kailath Festschrift*. Kluwer.

Åström, K. J. and B. Wittenmark (1995): *Adaptive Control*, second edition. Addison-Wesley, Reading, Massachusetts.

Bode, H. W. (1945): *Network Analysis and Feedback Amplifier Design*. Van Nostrand, New York.

Bongiorno, Jr., J. J. and D. C. Youla (1968): "On observers in multi-variable control systems." *International Journal of Control*, **8:3**, pp. 221–243.

Bongiorno, Jr., J. J. and D. C. Youla (1970): "Discussion of 'on observers in multi-variable control systems'. correspondence." *International Journal of Control*, **12:1**, pp. 183–190.

Boyd, S. and C. A. Desoer (1995): "Subharmonic function and performance bounds in linear timeinvariant feedback systems." *IMA J. Maht. Contr. and Info.*, **2**, pp. 153–170.

Boyd, S. P. and C. H. Barratt (1991): *Linear Controller Design – Limits of Performance*. Prentice Hall Inc., Englewood Cliffs, New Jersey.

Chen, J. (1995): "Sensitivity integral relations and design tradeoffs in linear multivariable feedback systems." *IEEE Transactions on Automatic Control*, **AC-40:10**, pp. 1700–1716.

Chen, J. (1998): "Multivariable gain-phase and sensitivity integral relations and design tradeoffs." *IEEE Transactions on Automatic Control*, **AC-43:3**, pp. 373–385.

Clarke, R., J. J. Burken, J. T. Bosworth, and J. E. Bauer (1994): "X-29 flight control system: lessons learned." *Int. J. Control*, **59:1**, pp. 199–219.

Doyle, J. C., B. A. Francis, and A. R. Tannenbaum (1992): *Feedback control theory*. Macmillan, New York.

Doyle, J. C. and G. Stein (1981): "Multivariable feedback design: Concepts for a classical/modern synthesis." *IEEE Transactions on Automatic Control*, **AC-26**, pp. 4–16.

Engell, S. (1988): *Optimale Lineare Regelung*. Springer, Berlin.

Francis, B. A. and K. Glover (1978): "Bounded peaking in the optimal linear regulator with cheap control." *IEEE Transactions on Automatic Control*, **23:4**, pp. 608–617.

Francis, B. A. and G. Zames (1984): "On H_∞ optimal sensitivity theory for SISO feedback systems." *IEEE Transactions on Automatic Control*, **AC-29:1**, pp. 9–16.

Frank, P. M. (1968a): "Vollständige Vorhersage im stetigen Regelkreis mit Totzeit, Teil I." *Regelungstechnik*, **16:3**, pp. 111–116.

Frank, P. M. (1968b): "Vollständige Vorhersage im stetigen Regelkreis mit Totzeit, Teil II." *Regelungstechnik*, **16:5**, pp. 214–218.

Freudenberg, J. S. and D. P. Looze (1985): "Right half plane pole and zeros and design tradeoffs in feedback systems." *IEEE Transactions on Automatic Control*, **AC-30**, pp. 555–565.

Freudenberg, J. S. and D. P. Looze (1987): "A sensitivity trade-off for plants with time delay." In *IEEE Trans. Aut. Control*, vol. AC-32, pp. 99–104.

Freudenberg, J. S. and D. P. Looze (1988): *Frequency Domain Properties fo Scalar and Multivariable Feedback Systems*. Springer Verlag, Berlin.

Goodwin, G., D. Q. Mayne, and J. Shim (1995): "Tradeoffs in linear filter design." *Automatica*, **31:10**, pp. 1367–1376.

Graebe, S. F. and R. H. Middleton (1995): "Stable open loop poles: to cancel or not to cancel?" In *Proc. 34th IEEE Conference on Decision and Control*. New Orleans, LA.

Green, M. and D. J. N. Limebeer (1995): *Linear Robust Control*. Prentice Hall, Englewood Cliffs, N.J.

Havre, K. and S. Skogestad (1999): "Achievable H_∞-performance of multivariable systems with unstable zeros and poles." In *Preprints 14th Triennial IFAC World Congress*, vol. D, pp. 391–396. Pergamon, Beijing.

Horowitz, I. M. (1963): *Synthesis of Feedback Systems*. Academic Press, New York.

Horowitz, I. M. (1993): *Quantitative Feedbac Design Theory (QFT)*. QFT Publications, Boulder, Colorado.

Horowitz, I. M. and U. Shaked (1975): "Superiority of transfer function over state-variable methods in linear time-invariant feedback system design." *IEEE Transactions on Automatic Control*, **AC-20**, pp. 84–97.

Huang, C. Y. and G. A. Knowles (1990): "Application of nonlinear control strategies to aircraft at high angle of attack." In *29th IEEE Conference on Decision and Control*, pp. 188–193. Honolulu, Hawai.

Keel, L. H. and S. P. Bhattacharyya (1997): "Robust, fragile, or optimal?" *IEEE Transactions on Automatic Control*, **AC-42:8**, pp. 1098–1105.

Khargonekar, P. and A. Tannenbaum (1985): "Non-euclidean metrics and the robust stabilization of systems with parameter uncertainty." *IEEE Transactions on Automatic Control*, **AC-30:10**, pp. 1005–1013.

Klein, R. E. (1986): "Using bicycles to teach system dynamics." *IEEE Control Systems Magazine*, **CSM**, April, pp. 4–9.

Kwakernak, H. (1969): "Optimal low-sensitivity linear feedback systems." *Automatica*, **5**, pp. 279–285.

Kwakernak, H. (1995): "Symmetries in control system design." In Isidori, Ed., *Trends in Control - A European Perspective*. Springer, Berlin.

Kwakernak, H. and R. Sivan (1972): "The maximally achievable accuracy of linear optimal regulators and linear optimal filters." *IEEE Transactions on Automatic Control*, **AC-17:1**, pp. 79–85.

Maciejowski, J. M. (1989): *Multivariable Feedback Design*. Addison-Wesley, Reading, Massachusetts.

Middleton, R. H. (1991): "Trade-offs in linear control system design." *Automatica*, **27:2**, pp. 281–292.

Mita, T. (1977): "On zeros and responses of linear regulators and linear observers." *IEEE Transactions on Automatic Control*, **AC-31:10**, pp. 423–439.

Morari, M. and E. Zafiriou (1989): *Robust Process Control*. Prentice-Hall, Englewood Cliffs, New Jersey.

Newton, Jr, G. C., L. A. Gould, and J. F. Kaiser (1957): *Analytical Design of Linear Feedback Controls*. John Wiley & Sons.

Qiu, L. and E. J. Davison (1993): "Performance limitations of non-minimum phase systems in the servomechanism problem." *Automatica*, **29:2**, pp. 337–349.

Seron, M. M., J. H. Braslavsky, and G. C. Goodwin (1945): *Fundamental Limitations in Filtering and Control*. Springer, Berlin.

Skogestad, S. and I. Postlethwaite (1996): *Multivariable Feedback Control: Analysis and Design*. Wiley, Chichester, UK.

Stein, G. (1990): "Respect the unstable - the 1995 Bode Lecture." In *30th IEEE Conference on Decision and Control*. Honolulu, Hawai.

Sussman, H. J. and P. V. Kokotovic (1991): "The peaking phenomenon and the global stabilization of nonlinear systems." *IEEE Transactions on Automatic Control*, **AC-36:4**, pp. 424–439.

Truxal, J. (1955): *Automatic Feedback Control System Synthesis*. McGraw-Hill, New York.

Yaniv, O., P.-O. Gutman, and I.Chepovetsky (1999): "Sensitivity reduction limitations in mimo non-minimum phase feedback systems." In *Preprints 14th Triennial IFAC World Congress*, vol. D, pp. 417–424. Pergamon, Beijing.

Youla, D. C. and C. N. Jabr, H. A. amd Lu (1974): "Single-loop feedback stabilization of linear multivariable dynamical systems." *Automatica*, **10:2**, pp. 159–173.

Zames, G. (1981): "Feedback and optimal sensitivity: model reference transformations, multiplicative seminorms, and approximative inverse." *IEEE Transactions on Automatic Control*, **AC-26:2**, pp. 301–320.

Zames, G. and B. A. Francis (1985): "Feedback, minimax sensitivity and optimal robustness." *IEEE Transactions on Automatic Control*, **AC-28:5**, pp. 585–600.