Model Uncertainty and Robust Control

K. J. Åström
Department of Automatic Control
Lund University, Lund, Sweden
Email: kja@control.lth.se, Fax: +46 46 13 81 18

1 Introduction

A key reason for using feedback is to reduce the effects of uncertainty which may appear in different forms as disturbances or as other imperfections in the models used to design the feedback law. Model uncertainty and robustness have been a central theme in the development of the field of automatic control. This paper gives an elementary presentation of the key results.

A central problem in the early development of automatic control was to construct feedback amplifiers whose properties remain constant in spite of variations in supply voltage and component variations. This problem was the key for the telephone industry that emerged in the 1920s. The problem was solved by [9]. We quote from his paper:

> " .. by building an amplifier whose gain is deliberately made say 40 decibels higher than necessary (10 000 fold excess on energy basis) and then feeding the output back on the input in such a way as to throw away excess gain, it has been found possible to effect extraordinarily improvement in constancy of amplification and freedom from nonlinearity."

Blacks invention had a tremendous impact and it inspired much theoretical work. This was required both for understanding and for development of design method. A novel approach to stability was developed in [36], fundamental limitations were explored by [10] who also developed methods for designing feedback amplifiers, see [11]. A systematic approach to design controllers that were robust to gain variations were also developed by Bode.

The work on feedback amplifiers became a central part of the theory of servomechanisms that appeared in the 1940s, see [22], [27]. Systems were then described using

transfer functions or frequency responses. It was very natural to capture uncertainty in terms of deviations of the frequency responses. A number of measures such as amplitude and phase margins and maximum sensitivities were also introduced to describe robustness. Design tools such as the Bode diagram introduced to design feedback amplifiers also found good use in design of servomechanisms. Bode's work on robust design was generalized to deal with arbitrary variations in the process by Horowitz [24]. The design techniques used were largely graphical.

The state-space theory that appeared in the 1960s represented significant paradigm shift. Systems were now described using differential equations. There was a very vigorous development that gave new insight, new concepts [32], [30] and new design methods. Control design problems were formulated as optimization problems which gave effective design methods, see [7], [8], and [40]. Control of linear systems with Gaussian disturbances and quadratic criteria, the LQG problem, was particularly attractive because it admitted analytical solutions, see [8], [29], [28], and [31]. The design computations were also improved because it was possible to capitalize on advances in numerical linear algebra and efficient software. The controller obtained from LQG theory also had a very interesting structure. It was a composition of a Kalman filter and a state feedback.

The state-space theory became the predominant approach, see [5]. Safonov and Athans [42] showed that the phase margin is at least 60° and the amplitude margin is infinite for an LQG problem where all state variables are measured. This result does unfortunately not hold for output feedback as was demonstrated in [15]. There were attempts to recover the robustness of state feedback using special design techniques called loop transfer recovery. The central issue is however that it is not straightforward to capture model uncertainty in a state variable setting. There was also criticism of the state-space theory, see [25].

The paper [49] represented a paradigm shift which brought robustness to the forefront. It started a new development that led to the so called $\mathcal{H}_\infty$ theory. The idea was to develop systematic design methods that were guaranteed to give stable closed loop systems for systems with model uncertainty. The original work was based on frequency responses and interpolation theory which led to compensators of high order. The seminal paper [17] showed how the problem could be solved using state-space methods. Game theory is another approach to $\mathcal{H}_\infty$ theory. The game is to find a controller in the presence of an adversary that changes the process, see [6]. The $\mathcal{H}_\infty$ theory is now well described in books, see [16], [21] and [52].

Major advances in robust design was made in the book [35] where the $\mathcal{H}_\infty$ control problem was regarded as a loop shaping problem. This gave effective design methods and it also reestablished the links with classical control. This line of research has been continued by [48] who has obtained definite results relating modeling errors and robust control. To do this he also had to invent a novel metric for systems, see [46]. This work brings $\mathcal{H}_\infty$ even closer to the classical results.

In this chapter we will try to present the essence of the development in the simple setting of single-input single-output systems. We start with a presentation of some aspects of classical control theory in Section 2. Robustness issues for state-space theory

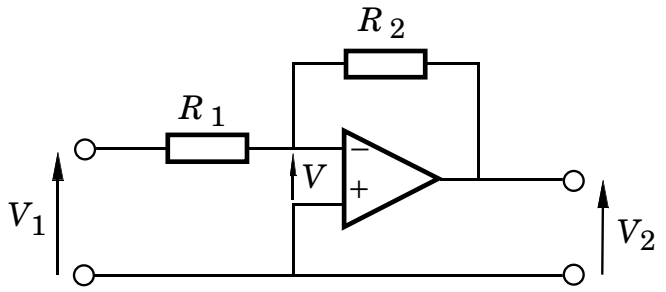

Fig. 1: Schematic diagram of a feedback amplifier.

is reviewed briefly in Section 3 where we also present an example that illustrates that a blind use of state-space theory can lead to closed loop systems with very poor robustness properties. This can be overcome by analyzing the robustness and modifying the design criteria. In Section 4 we discuss fundamental limitations on performance due to time delays and right half plane poles and zeros. This does not appear naturally in state-space theory where the major requirements are observability and controllability.

In Section 5 we present some key results from $\mathcal{H}_\infty$-loop shaping. To do this we have to discuss the important problem of determining if two systems are close from the point of view of feedback. We also have to introduce better stability concepts and ways to characterize model uncertainty. We end the section with a discussion of Vinnicombe's theory which gives much insight, necessary conditions and very nice ties to classical control theory.

2 Classical Control Theory

The fields of automatic control emerged in the mid 1940s when it was realized that it was a common framework for problems associated with feedback control from a wide range of fields such as telecommunication, industrial processes, vehicles, power systems etc. An essential component came from telecommunications where a key problem had been to design accurate reliable amplifiers from components with variable properties.

The Feedback Amplifier

A schematic diagram of the amplifier is shown in Figure 1. Let the raw gain of the amplifier be A. The feedback amplifier has very remarkable properties as can be seen from its input-output relation. It follows from Figure 1 that

$$\frac{V_2}{V_1} = -\frac{R_2}{R_1}\,\frac{1}{1+\frac{1}{A}(1+\frac{R_2}{R_1})}. \tag{1}$$

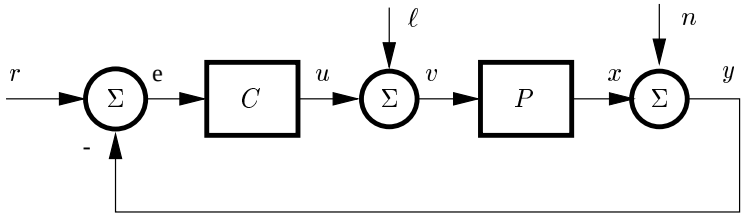

Fig. 2: Block diagram of a feedback system.

Notice that the gain V_2/V_1 is essentially given by the ratio R_2/R_1. If the raw amplifier gain A is large the gain is virtually independent of A. Assume for example that $R_2/R_1 = 100$ and that $A = 10^4$. A 10% change of A the variation gives only a gain variation of 0.1%. Feedback thus has the amazing property of reducing the effects of uncertainty drastically. The linearity is also increased significantly.

The risk for instability is a drawback of feedback. Nyquist developed a theory for analyzing stability of feedback amplifiers, see [36]. Systematic methods to design feedback systems were developed in [10] and further elaborated in [11]. These ideas formed one of the foundations of automatic control.

In today's terminology, we could say that Black used feedback to design an amplifier that was robust to variations in the gain of the process. As a side-effect, he also obtained a closed-loop system that was extremely linear.

Generalization

The ideas of feedback are applicable to a wide range of systems. This is illustrated in Figure 2 which shows a basic feedback loop consisting of a process and a controller. The purpose of the system is to make the process variable x follow the set point r in spite the disturbances ℓ and n that act on the system. The properties of the closed loop system should also be insensitive to variations in the process. There are two types of disturbances. The load disturbance ℓ drives the system away from its desired state and the measurement noise n which corrupts the information about the system obtained from the sensors.

The system in Figure 2 has three inputs r, ℓ and n, and four interesting signals x, y, e and u. There are thus 12 relations that are of potential interest. Assume that the process and the controller are linear time-invariant systems that are characterized by their transfer functions P and C respectively. The relations between the signals are

given by the transfer functions:

$$\begin{aligned}
G_{xr} &= \frac{PC}{1+PC} & G_{x\ell} &= \frac{P}{1+PC} & G_{xn} &= -G_{xr} \\
G_{yr} &= G_{xr} & G_{y\ell} &= G_{x\ell} & G_{yn} &= \frac{1}{1+PC} \\
G_{er} &= 1 - G_{xr} = G_{yn} & G_{e\ell} &= -G_{x\ell} & G_{en} &= -G_{yn} \\
G_{ur} &= \frac{C}{1+PC} & G_{u\ell} &= -G_{xr} & G_{un} &= -G_{ur}
\end{aligned}$$

Here G_{ij} denotes the transfer function from signal j to signal i. Notice that there are only four independent transfer functions.

$$\begin{aligned}
G_{xr} &= \frac{PC}{1+PC} \\
G_{x\ell} &= \frac{P}{1+PC} \\
G_{yn} &= \frac{1}{1+PC} \\
G_{ur} &= \frac{C}{1+PC}
\end{aligned} \tag{2}$$

The transfer functions $G_{xr} = T$ and $G_{yn} = S$ have special names, S is called the sensitivity function, and T is called the complementary sensitivity function. Notice that both S and T depend only on the loop transfer function $L = PC$. The sensitivity functions are are related by

$$S + T = 1. \tag{3}$$

This explains the name complementary sensitivity function. These functions have interesting properties as is discussed in the following.

Stability and Stability Margins

Many properties of the system in Figure 2 can be derived from the loop transfer function $L = PC$.

The stability of the system can be investigated by Nyquist's stability criterion which says that the closed loop system is stable if

$$\frac{1}{2\pi}\Delta \arg_{\Gamma}(1 + L(s)) = -\mathcal{P}_{rhp}(L) \tag{4}$$

where $\Delta \arg$ is the argument variation when s traverses a contour Γ that encloses the right half plane (RHP) and $\mathcal{P}_{rhp}(L)$ is the number of poles of L in the right half plane.

Stability is normally investigated by analyzing the Nyquist curve, see Figure 3.

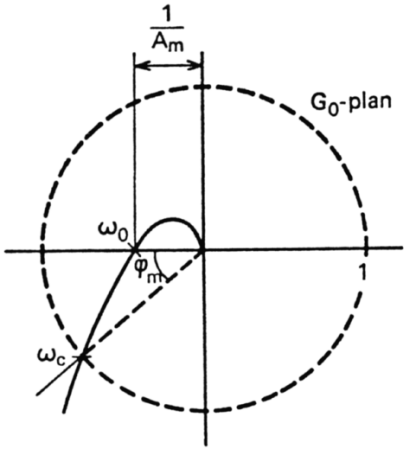

Fig. 3: Nyquist curve with phase and amplitude margins.

To achieve stability the Nyquist curve must be sufficiently far away from the critical point -1. The distance from the critical point can also be used as a measure of the degree of stability. In this way the notions of amplitude margin A_m and phase margin φ_m, see Figure 3. An amplitude margin A_m implies that the gain can be increased with a factor less than A_m without making the system unstable. Similarly for a system with a phase margin φ_m it is possible to increase the phase shift in the loop by a quantity less than φ_m without making the system unstable.

2.1 Small Process Variations

The effects of small variations in the process will now be investigated. The signal transmission of the closed loop system from set point r to output y is given by the complementary sensitivity function

$$T = \frac{PC}{1+PC}$$

We have

$$\frac{dT}{T} = \frac{dP}{P} - \frac{CdP}{1+PC} = \frac{1}{1+PC}\frac{dP}{P} = S\frac{dP}{P} \tag{5}$$

The function S tells how the closed loop properties are influenced by small variations in the process. The maximum sensitivities

$$M_s = \max|S(i\omega)|, \qquad M_t = \max|T(i\omega)| \tag{6}$$

are also used as robustness measures. The variable $1/M_s$ can be interpreted as the shortest distance between the Nyquist curve and the critical point -1, see Figure 3. classically the maximum complementary sensitivity is denoted by M_p.

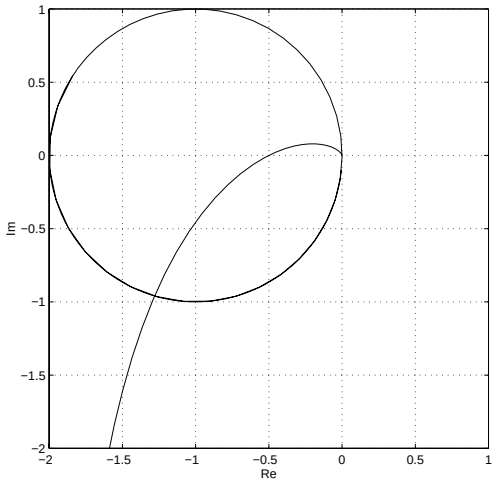

Fig. 4: Nyquist curve with circle for constant sensitivity $S(i\omega) = 1$. Disturbances with frequencies outside the circle are attenuated and frequencies inside the circle are amplified by the feedback.

2.2 A Property of the Sensitivity Function

The sensitivity function also has other physical interpretations. Consider the system in Figure 2. If there is no feedback the Laplace transform of the output is Y_{ol}. The output under closed loop control is given by

$$Y_{cl} = \frac{1}{1+PC} Y_{ol}$$

It thus follows that

$$\frac{Y_{cl}}{Y_{ol}} = \frac{1}{1+PC} = S \tag{7}$$

The sensitivity function thus tells how the disturbances are influenced by feedback. Disturbances with frequencies such that $|S(i\omega)|$ is less than one are reduced by an amount equal to the distance to the critical point and disturbances with frequencies such that $|S(i\omega)|$ is larger than one are amplified by the feedback. This is illustrated in Figure 4.

2.3 Large Process Variations

Equation 5 gives the sensitivity for small perturbations of the process. It is also possible to get expressions for large variations in the process. To see how much the process can change without making the closed loop system unstable we will use the Nyquist

diagram. Consider a point on the Nyquist curve in Figure 4. The distance to the critical point is $|1+L|$. If the process changes by ΔP, the point changes by $C\Delta P$. The system will remain stable as long as

$$|C\Delta P| < |1+PC|$$

and the number of right hand poles of PC does not change. This implies that the perturbations must have the property that ΔP does not have any poles in the right half plane.

The admissible variation is process dynamics is thus given by

$$|\frac{\Delta P}{P}| < \left|\frac{1+PC}{PC}\right| = \left|\frac{1}{T}\right| \leq \frac{1}{M_t} \tag{8}$$

which can also be expressed as

$$|\Delta P| < \left|\frac{P}{T}\right| \leq \frac{|P|}{M_t} \tag{9}$$

A crude estimate of the largest admissible variation in the process is thus given by the largest value M_t of the complementary sensitivity. It follows from the Figure 4 that large variations in P are permitted for the frequencies where P either is large or small. The smallest admissible variations are for frequencies where $|T|$ is large.

A similar estimate based on the maximum sensitivity is that

$$|\Delta P| < \left|\frac{1}{SC}\right| \leq \frac{1}{M_s|C|} \tag{10}$$

Bode's Integrals

It follows from Equations (5) and (8) that it would be highly desirable to make the sensitivity functions S and T as small as possible. This is unfortunately not possible because it follows from Equation (3) that $S+T=1$. There are also other constraints on the sensitivities. It was shown in [11] that

$$\begin{aligned}\int_0^\infty \log|S(i\omega)|d\omega &= \int_0^\infty \log|\frac{1}{1+L(i\omega)}|d\omega = \pi\sum p_i \\ \int_0^\infty \log|T(1/i\omega)|d\omega &= \int_0^\infty \log|\frac{L(1/i\omega)}{1+L(1/i\omega)}|d\omega = \pi\sum\frac{1}{z_i}\end{aligned} \tag{11}$$

where p_i are the right half plane poles of L and z_i are the right half plane zeros of L. These equations imply that the sensitivities can be made small at one frequency only at the expense of increasing the sensitivity at other frequencies. This phenomena is sometimes called the water bed effect. It also follows from the equations that the presence of poles in the right half plane increase the sensitivity and that zeros in the right half plane increase the complementary sensitivity. A fast RHP pole gives higher sensitivity than a slow pole, and a slow RHP zero gives higher sensitivity than a fast zero.

2.4 Bode's Relations

The amplitude and the phase curves are also related. It is not possible to achieve high phase advance without using high gains and it is not possible to obtain transfer functions that decrease rapidly without having large phase lags. These facts are expressed analytically by some relations derived in [10].

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

$$\log G(i\omega) = A(\omega) + i\Phi(\omega) \tag{12}$$

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

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

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

$$\begin{aligned} A(\omega_0) - A(\infty) &= -\frac{2}{\pi}\int_0^\infty \frac{v\Phi(v) - \omega_0\Phi(\omega_0)}{v^2 - \omega_0^2} dv \\ &= -\frac{1}{\omega_0 \pi}\int_{-\infty}^\infty \frac{d(e^u\phi(u))}{du} \log \coth |\frac{u}{2}| du \\ \Phi(\omega_0) &= \frac{2\omega_0}{\pi}\int_0^\infty \frac{A(v) - A(\omega_0)}{v^2 - \omega_0^2} dv = \frac{1}{\pi}\int_{-\infty}^\infty \frac{da(u)}{du} \log \coth |\frac{u}{2}| du \end{aligned} \tag{13}$$

an approximate version is that

$$\Phi(\omega) \approx \frac{2}{\pi}\frac{da(u)}{du}. \tag{14}$$

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.

Bode's relations imposes fundamental limitations on the performance that can be achieved. A simple observation is that even if it is desirable that the loop gain decreases rapidly at the crossover frequency, it is not possible to have a steeper slope than -2 without violating stability constraints.

An interesting problem is if the limitations imposed by Bode's relations can be avoided by using nonlinear systems. The Clegg integrator [13] is a nonlinear system where the magnitude curve has the slope -1 and the phase lag is only 38°.

2.5 Bode's Ideal Loop Transfer Function

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

$$L(s) = \Big(\frac{s}{\omega_{gc}}\Big)^n. \tag{15}$$

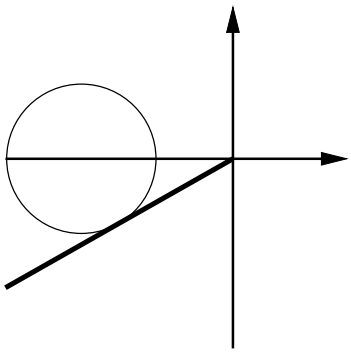

Fig. 5: Nyquist curve for Bode's ideal loop transfer function.

The Nyquist curve for this loop transfer function is simply a straight line through the origin with $\arg L(i\omega) = n\pi/2$, see Figure 5. Bode called (15) 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 $L(s)$ given by Equation (15) 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 [24] to systems that are insensitive to other variations of the plant, culminating in the QFT method, see [26].

The transfer function given by Equation (15) 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}. Assume for example that the gain of the process varies between k_{min} and k_{max} and that it is desired to have a loop transfer function that is close to (15) in the frequency range $(\omega_{min}, \omega_{max})$. It follows from (15) that

$$\frac{\omega_{max}}{\omega_{min}} = \Big(\frac{k_{max}}{k_{min}}\Big)^{1/n}$$

With $n = -5/3$ and a gain ratio of 100 we get a frequency ratio of about 16 and with $n = -4/3$ we get a frequency ratio of 32. To avoid having too large a frequency range it is thus useful to have n as small as possible. There is, however, a compromise because the phase margin decreases with decreasing n and the system becomes unstable for $n = -2$.

2.6 Fractional Systems

It follows from Equation (15) that the loop transfer function is not a rational function. We illustrate this with an example.

Example.

Consider a process with the transfer function

$$P(s) = \frac{k}{s(s+1)} \tag{16}$$

Assume that we would like to have a closed loop system that is insensitive to gain variations with a phase margin of 45°. Bode's ideal loop transfer function that gives this phase margin is

$$L(s) = \frac{1}{s\sqrt{s}} \tag{17}$$

Since $L = PC$ we find that the controller transfer function is

$$C(s) = \frac{s+1}{\sqrt{s}} = \sqrt{s} + \frac{1}{\sqrt{s}} \tag{18}$$

To implement a controller the transfer function is approximated with a rational function. This can be done in many ways. One possibility is the following

$$\hat{C}(s) = k\frac{(s+1/64)(s+1/16)(s+1/4)(s+1)^2(s+4)(s+16)(s+64)}{(s+1/128)(s+1/32)(s+1/8)(s+1/2)(s+2)(s+8)(s+32)(s+128)} \tag{19}$$

where the gain k is chosen to equal the gain of $\sqrt{s} + 1/\sqrt{s}$ for $s = i$. Notice that the controller is composed of sections of equal length having slopes 0, +1 and -1 in the Bode diagram. Figure 6 shows the Bode diagram for the loop transfer function. The figure shows that the phase margin will be close to 45° with a tolerance of less than 2° for a gain variation of 3 orders of magnitude. With a tolerance of 5° we can even allow a gain variation of 4 orders of magnitude. The range of gains can be extended by making the controller more complex. Even if the closed loop system has the same phase margin when the gain changes the response speed will change with the gain.

The example shows that robustness is obtained by increasing controller complexity. The range of gain variation that the system can tolerate can be increased by increasing the complexity of the controller.

Fractional systems did not receive much attention after Bodes work. In the 1990s there was however a resurgence in the interest of fractional systems, see e.g. [53, 54, 55, 56, 57]. Oustaloup coined the acronyme CRONE from the french Commande Robuste d'Ordre Non Entier (Robust Control of Fractional Order) for his controller.

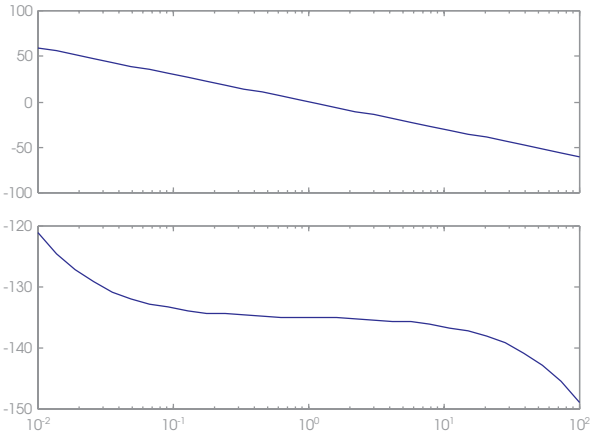

Fig. 6: Bode diagram of the loop transfer function obtained by approximating the fractional controller with a rational transfer function.

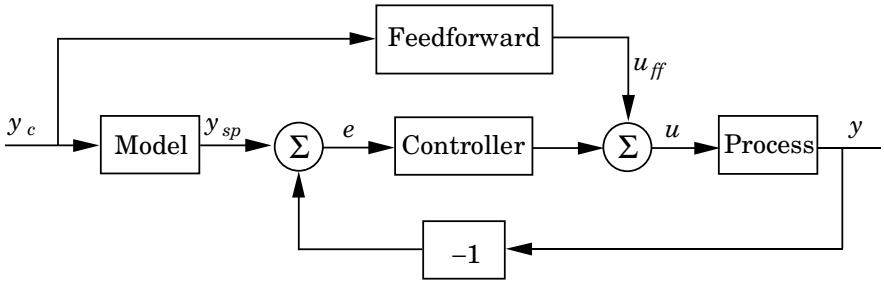

Fig. 7: Block diagram of a system with two degrees of freedom.

Two Degree of Freedom Systems (2DOF)

The system in Figure 2 is a system with error feedback because the controller acts on the error $e = r - y$ which is the difference between the set point and the output. There are significant advantages in having control systems with other configurations. An example of such a system is shown in Figure 7. In this system the set point is fed through a model before it is compared with the process output. There is also a feedforward link which essentially is a combination of the model and the inverse process model or an approximate inverse. Ideally the feedforward signal u_{ff} generates a signal which when applied to the process produces the ideal output in response to set point changes. The feedback controller which acts on the error will only make some corrections if there are deviations from the ideal behavior.

A system with error feedback only is called a system with one degree of freedom. The system in Figure 7 is called a system with two degrees of freedom (2DOF) because the signal paths from the set point to the control signal is different from the signal path

from the output to the control signal. This terminology was introduced by [24] who analyzed these systems carefully.

A very nice property of systems with two degrees of freedom is that the problem of set point response can be separated from the problems of robustness and disturbance rejection. Referring to Figure 2 we will first design a feedback by compromising between disturbance attenuation and robustness. When this is done we will then design a model and a feedforward which gives the desired response to the setpoint.

There are many variations of systems with two degrees of freedom, the following quote from [24] is still valid

> "Some structures have been presented as fundamentally different from the others. It has been suggested that they have virtues not possessed by others, and have been given special names ... all 2DOF configurations have basically the same properties and potentials ...."

Quantitative Feedback Theory (QFT)

Bode's technique of dealing with gain variations were both elegant and effective. A limitation of Bode's work was that it was limited to gain variations only. A very nice generalization of Bode's work was done by Horowitz who extended it to arbitrary variations of a process transfer function. He characterized model uncertainty by sets of amplitudes and phase for each frequency called templates. Horowitz also developed a graphical design techniques to design feedback systems that were robust to these types of disturbances. He used a system configuration with two degrees of freedom to deal with set point responses. Horowitz design technique called quantitative feedback theory (QFT) is described in several books, see [24] and [26]. It has been applied successfully to a wide range of problems.

Summary

In this section we have reviewed classical control theory with a focus on model uncertainty and robustness. It is worthwhile to note that model uncertainty was a key motivation for introducing feedback and that classical control theory had very effective ways of dealing with uncertainty both qualitatively and quantitatively. Process uncertainty could be described very easily as a variation in the process transfer function with the caveat that the disturbances do not change the number of right half plane poles of the system. The theory has given important concepts and tools such as the transfer function, Nyquist's stability theory, the Nyquist curve, Bode diagrams, Bode's integrals and Bode's ideal loop transfer function. Robustness measures such as amplitude and phase margins and the maximum sensitivities were also introduced. Bode's ideal loop transfer function is probably the first design method that addressed robustness explicitly. Horowitz quantitative feedback theory is a continuation of this idea.

3 State-Space Theory

The state-space theory represented a paradigm shift which led to many useful system concepts and new methods for analysis and design. The systems was represented by differential equations instead of transfer functions. For linear systems the standard model used was

$$\begin{aligned} \frac{dx}{dt} &= Ax + Bu + v \\ y &= Cx + e \end{aligned} \tag{20}$$

where u is the input, y the output and x is the state. The uncertainty is represented by the disturbances v and e and by variations in the elements of the matrices A, B and C. The disturbances e and v were typically described as stochastic processes, see [20] and [4].

The control problem was formulated as to minimize the criterion

$$J = E \lim_{T\to\infty} \frac{1}{T} \int_0^T (x^T Q_1 x + u^T Q_2 u) dt \tag{21}$$

Since the equations are linear with stochastic disturbances and the criterion is quadratic the problem was called the linear quadratic Gaussian control problem (LQG). The solution to the control problem is given by

$$\begin{aligned} u &= L(x_m - \hat{x}) + u_{ff} \\ \frac{d\hat{x}}{dt} &= A\hat{x} + Bu + K(y - C\hat{x}) \end{aligned} \tag{22}$$

This control law has a very nice interpretation as feedback from the error $x_m - \hat{x}$ which is the difference between the ideal states x_m and the estimated states $\hat{x}$. The estimated states are given by the Kalman filter. Controllability and observability are key conditions for solving the problem.

There are many other design methods based on the state-space formulation which gives controllers with the structure (22) for example pole placement. They differ from the LQG method in the sense that other techniques are used to obtain the matrices K and L.

In Figure 8 we show a block diagram of the controller obtained from LQG theory. In the figure we have also used a system configuration with two degrees of freedom. The system has a very attractive structure. The observer or the Kalman filter delivers an estimate of the state based on a model of the system and the input and output signals of the system. Notice that the state may also have components that represent the disturbances. There is a feedback from the deviations of the estimated state from its desired value x_m. Set point following is obtained by the usual two degree of freedom configuration.

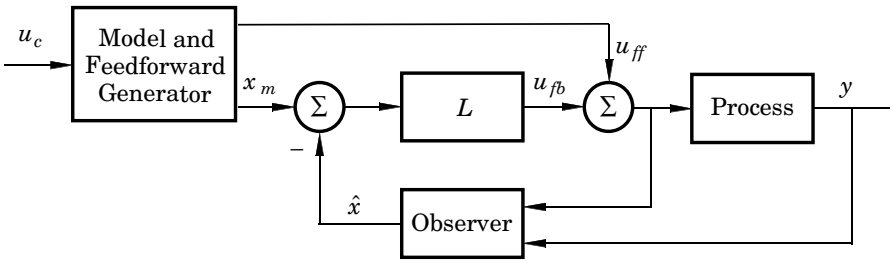

Fig. 8: Block diagram of a system state feedback having two degrees of freedom.

Robustness

In the model (20) it is natural to describe model uncertainties as variations in the elements of the matrices A, B and C. This is, however, a very restricted class of perturbations which does not cover neglected dynamics or small time delays such uncertainties are easier to describe in the frequency domain. The LQG theory was also criticized heavily by classic control theorist because it did not take robustness into account, see [25].

Very strong robustness properties could be established when all states were measured. In this case the system has a phase margin of 60° and infinite amplitude margin, see [42], which indicates a very good robustness.

The nice robustness properties of systems with state feedback do not hold for systems with output feedback. Nice counterexamples were given in the paper [15]. For systems with output feedback it was attempted to recover the robustness of full state feedback making very fast observers. This approach led to a design technique called loop transfer recovery, see [18].

The only formal requirements on the system to be controlled in state-space theory is that the system is observable and controllable. There are no consideration of right half plane poles and zeros or time delays. Because of this it is necessary to investigate the robustness of the design and to make appropriate modifications to achieve good robustness. We use an example to illustrate what happens if this is not done.

Example: A fast system with a low bandwidth

Consider a system that is described by

$$\frac{dx}{dt} = \begin{pmatrix} -1 & 1 \\ 0 & 0 \end{pmatrix} x + \begin{pmatrix} a \\ 1 \end{pmatrix} u$$
$$y = \begin{pmatrix} 1 & 0 \end{pmatrix} x$$

The system is controllable if $a \neq 1$. We will assume that $a = -10$. The system is of second order and one state variable is measured directly. The system can be controlled

with an observer of first degree. The closed loop system is then of order 3. We assume that a state feedback and an observe is designed so that the closed loop system is

$$(s + \alpha\omega_0)(s^2 + \omega_0 s + \omega_0^2). \tag{23}$$

The transfer function of the system is

$$P(s) = \frac{as + 1}{s(s + 1)}. \tag{24}$$

To obtain a fast closed-loop system we choose $\omega_0 = 10$ and $\alpha = 1$. straightforward calculations show that the controller has the transfer function

$$C(s) = \frac{s_0 s + s_1}{s + r} \tag{25}$$

with $r = 9274.5$, $s_0 = 925.5$ 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 9.

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 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 $\varphi_m = 0.15^\circ$. 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 and observability. For this particular problem there are severe limitations caused by the right half plane zero. 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.

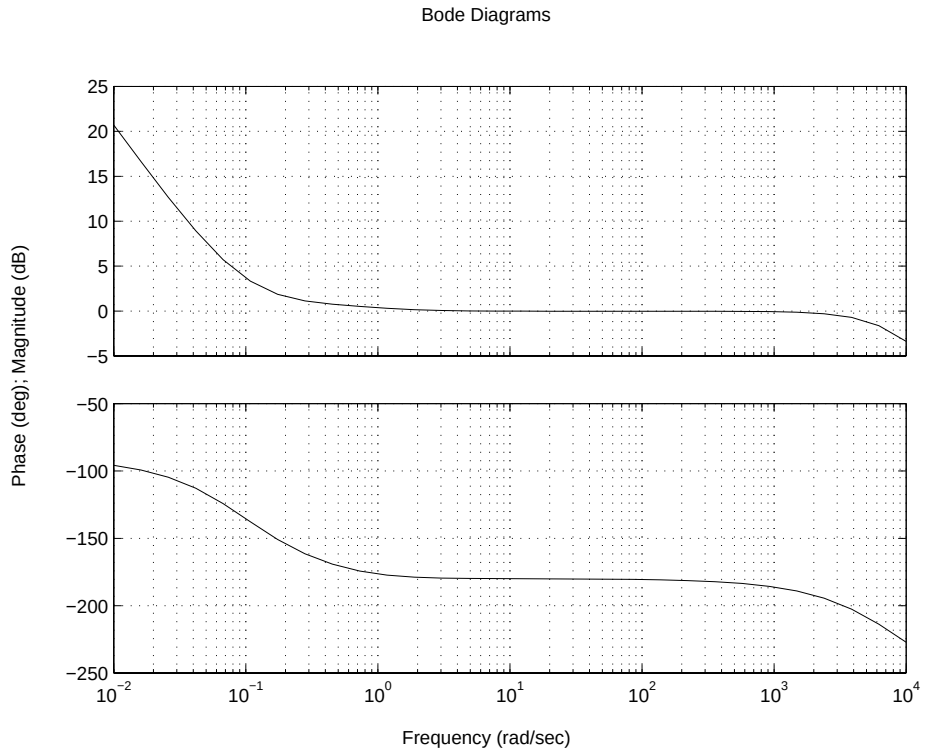

Fig. 9: Bode diagram for the loop transfer function of a system with a slow unstable zero at $s = 0.1$, where the specifications are $\omega_0 = 10$ and $\alpha = 1$.

Summary

state-space theory is an elegant way to approach a control problem. A nice aspect is that it naturally deals with multi-variable systems. The theory has given important concepts such as observability and reachability, It has also given several design methods, such as linear quadratic control, loop transfer recovery and pole placement. It has also introduced powerful computational methods based on numerical linear algebra. A severe drawback is that robustness is not dealt with properly. This means that it is possible to formulate control problems which give solutions that have very poor robustness properties. It is easy to avoid the difficulties when we are aware of them, simply by evaluating the robustness of a design and to reduce the requirements until a suitable compromise is reached.

4 Fundamental Limitations

It is very useful to determine the performance that can be achieved without sacrificing robustness. Such estimates will be provided in this section. In particular we will consider limitations that arise from poles and zeros in the right half plane and time delays. The results are based on [2] and [1].

Consider a system with the transfer function $P(s)$. Factor the transfer function as

$$P(s) = P_{mp}(s)P_{nmp}(s) \tag{26}$$

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}.

4.1 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.

$$\arg L(i\omega_{gc}) = \arg P_{nmp}(i\omega_{gc}) + \arg P_{mp}(i\omega_{gc}) + \arg C(i\omega_{gc}) \geq -\pi + \varphi_m. \tag{27}$$

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 (15), then

$$\arg P_{mp}(i\omega) + \arg C(i\omega) = n\frac{\pi}{2} \tag{28}$$

where n is the slope of the loop transfer function at the crossover frequency. Equation (28) is also a good approximation for other controllers because the amplitude curve is typically close to a straight line at the crossover frequency. The parameter n in (28) is then the slope n_{gc} at the crossover frequency. It follows from Bode's relations (13) that the phase is $n_{gc}\pi/2$. It follows from Equations (27) and (28) that the crossover frequency satisfies the inequality

$$\arg P_{nmp}(i\omega_{gc}) \geq -\alpha, \tag{29}$$

where

$$\alpha = \pi - \varphi_m + n_{gc}\frac{\pi}{2}. \tag{30}$$

This *crossover frequency inequality* gives the limitations imposed by non-minimum phase factors. A straightforward method to determine the crossover frequencies that can be obtained is to plot the left hand side of Equation (29) and determine when the inequality holds. The following example gives a simple rule of thumb.

Example: A Simple Rule of Thumb

To see the implications of (29) we will make some reasonable design choices. With a phase margin of 45° ($\varphi_m = \pi/4$), and a slope of $n_{gc} = -1/2$ we get $\alpha = \pi/2$ and Equation (29) becomes

$$\arg P_{nmp}(i\omega_{gc}) \geq -\frac{\pi}{2}. \tag{31}$$

This gives the simple rule that the phase lag of the minimum phase components should be less than 90° at the gain crossover frequency.

4.2 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

$$P_{nmp}(s) = \frac{z-s}{z+s}. \tag{32}$$

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

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

and (29) gives the following upper bound on the crossover frequency.

$$\frac{\omega_{gc}}{z} \leq \tan\alpha/2. \tag{33}$$

and the simple rule of thumb (31) we get $\omega_{gc} < z$.

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.

4.3 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

$$P_{nmp}(s) = e^{-sT}. \tag{34}$$

We have $\arg P_{nmp}(i\omega) = -\omega T$ and the crossover frequency inequality, Equation (29) becomes

$$\omega_{gc}T \leq \pi - \varphi_m + n_{gc}\frac{\pi}{2} = \alpha. \tag{35}$$

The simple rule of thumb (31) gives $\omega_{gc}T \leq \frac{\pi}{2} = 1.57$.

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

4.4 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

$$P_{nmp}(s) = \frac{s+p}{s-p} \tag{36}$$

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

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

and the crossover frequency inequality] (29) gives

$$\omega_{gc} \geq \frac{p}{\tan\alpha/2}. \tag{37}$$

The simple rule of thumb (31) gives $\omega_{gc} \geq p$.

Unstable poles give a lower bound on the crossover frequency. For systems with right half plane poles the bandwidth must thus be sufficiently large.

4.5 Poles and Zeros in the Right Half Plane

Consider a system with

$$P_{nmp}(s) = \frac{(z-s)(s+p)}{(z+s)(s-p)}. \tag{38}$$

For $z > p$ we have

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

The right hand side has its maximum for $\omega = \sqrt{pz}$ and the inequality (29) becomes

$$\frac{z}{p} \geq \tan^2\alpha/2 = \tan^2\Big(\frac{\pi}{4} - \frac{\varphi_m}{4} + n_{gc}\frac{\pi}{8}\Big). \tag{39}$$

The simple rule of thumb (31) gives $z \geq 25.3p$. Table 1 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

$$\varphi_m < \pi + n_{gc}\frac{\pi}{2} - 4\arctan\sqrt{\frac{p}{z}}. \tag{40}$$

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 largest gain crossover frequency is the geometric mean of the unstable pole and zero.

z/p	2	2.24	3.86	5	5.83	8.68	10	20
φ_m	-6.0	0	30	38.6	45	60	64.8	84.6

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

Example: The X-29

Considerable design effort has been devoted to the design of the flight control system for the X-29 aircraft, see [12] and [41]. 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 (40) that the achievable phase margins for $n_{gc} = -0.5$ and $n_{gc} = -1$ are $\varphi_m = 32.3^\circ$ and $\varphi_m = -12.6^\circ$. 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 section would have given much insight.

Example: Klein's Unridable Bicycle

An interesting bicycle with rear wheel steering which is impossible to ride was designed by Professor Klein in Illinois, see [33]. 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 (29) 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 (29) cannot be satisfied unless $\varphi_m < 0$ and $n_{gc} > 0$.

4.6 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

$$P_{nmp}(s) = \frac{s+p}{s-p}e^{-sT}. \tag{41}$$

The crossover frequency condition (29) gives

$$2\arctan\frac{\omega_{gc}}{p} - \omega_{gc}T \geq \varphi_m - n_{gc}\frac{\pi}{2}. \tag{42}$$

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 (42) 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

$$pT \leq 0.326 \tag{43}$$

Example: 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 (43) gives $\sqrt{g/\ell}\ 0.07 < 0.326$, 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 (43) that the shortest pendulum that can be balanced with $\varphi_m = 45^\circ$ and $n_{gc} = -0.5$ is $\ell = 0.23$m.

4.7 Other Criteria

The phase margin is a crude indicator of the stability margin. By carrying out detailed designs the results can be refined. This is done in [1] which gives results for designs with $M_s = M_t = 2$ and $M_s = M_t = 1.4$.

- A RHP zero z

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

- A RHP pole p

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

- A time delay T

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

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

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

- A RHP pole p and a time delay T

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

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.

5 $\mathcal{H}_\infty$ Loop Shaping

A consequence of the introduction of state-space theory was that interest shifted from robustness to optimization. New developments that started by the development of $\mathcal{H}_\infty$ control by George Zames in the 1980s gave a strong revival of robustness, see [50]. This led to a very vigorous development that has given new insight and new design methods. These results will be discussed in this Section. To keep the presentation simple we will only deal with systems having one input and one output, but techniques as well as results can be generalized to systems with many inputs and many outputs. For more extensive treatments we refer to [16], [21], [52] [23] and [48].

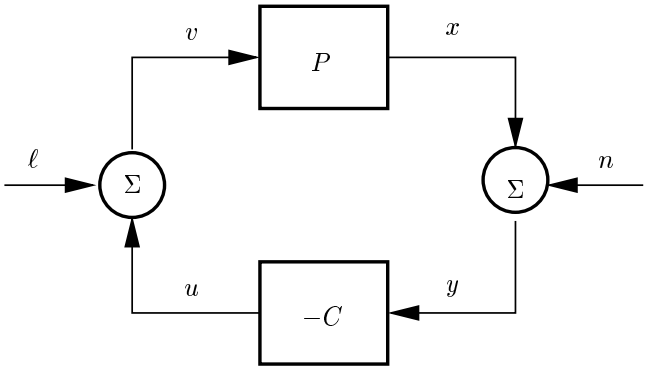

Fig. 10: Block diagram of a simple feedback system.

Problem Formulation

If a system structure with two degrees of freedom is used the problems of setpoint response can can be dealt with separately and we can therefore focus on robustness and attenuation of disturbances. The set point will not be considered and Figure 2 can be simplified to the system shown in Figure 10. The system has two inputs the measurement noise n and the load disturbance d. The problem we will consider is to design a controller with the properties.

- Insensitive to changes in the process properties.
- Ability to reduce the effects of the load disturbance d.
- Does not inject too much measurement noise into the system.

Stability

Stability is a primary robustness requirement. There is one problem with the stability concept introduced in Section 2. Because the stability is based on the loop transfer function only there may be cancellations of poles and zeros in the process and the controller. This does not pose any problems if the cancelled factors are stable. The results will however be strongly misleading if the canceled factors are unstable because there will be internal signals in the system that will diverge. We illustrate this with an example.

Example: Pole Zero Cancellation

Consider the system in Figure 10 with

$$C(s) = \frac{s-1}{s}$$
$$P(s) = \frac{1}{s-1}$$

The loop transfer function is $L = 1/s$ and the system thus appears stable. Notice however that the transfer function from disturbance d to output is.

$$G_{y\ell} = \frac{s}{(s+1)(s-1)}$$

A load disturbance will thus make the output diverge and it does not make sense to call the system stable.

The problem illustrated in Example 5 is well known. Classically it is resolved by formally introducing the rule that cancellation of unstable poles are not permissible. This can also be encapsulated in an algebra for manipulating systems which does not permit division by factors having roots in the right half plane, see [38] and [39].

Another way is to introduce a stability concept that takes care of the problem directly. It follows from the analysis in Section 2 that the closed loop system is completely characterized by the transfer functions given by Equation (2). Based on this we can say that a system is stable if all these transfer functions (2) are stable. This is sometimes called internal stability. The transfer functions (2) can be conveniently combined in the matrix

$$G(s) = \begin{pmatrix} \dfrac{1}{1+PC} & -\dfrac{C}{1+PC} \\ \dfrac{P}{1+PC} & -\dfrac{PC}{1+PC} \end{pmatrix} \tag{44}$$

Notice that this transfer function (44) represents the signal transmission from the disturbances d and n to the signals v and x in the block diagram in Figure 2. Let the transfer functions of the process and the controller be represented as

$$C(s) = \frac{B_c}{A_c}$$
$$P(s) = \frac{B_c}{A_c}$$

The matrix (44) can then be represented as

$$G(s) = \begin{pmatrix} \dfrac{A_c A_p}{A_c A_p + B_c B_p} & -\dfrac{A_p B_c}{A_c A_p + B_c B_p} \\ \dfrac{A_c B_p}{A_c A_p + B_c B_p} & -\dfrac{B_p B_c}{A_c A_p + B_c B_p} \end{pmatrix} \tag{45}$$

and the stability criterion is that the equation

$$C_{pol} = A_c A_p + B_c B_p \tag{46}$$

should have all its roots in the left half plane. This is also called internal stability. We will simply say that the system (P, C) is stable.

Example: Pole Zero Cancellation

Applying the result to the problem in Example 5 we find that

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

The characteristic polynomial is

$$C_{pol} = (s-1)s + s - 1 = (s-1)(s+1)$$

which clearly has a root in the right half plane.

How to Compare two Systems

A fundamental problem when discussing robustness is to determine when two systems are close. This seemingly innocent problem is not as simple as it may appear. For feedback control it would be natural to claim that two systems are close if they have similar behavior under a given feedback, see [45] and [43]. The fact that two systems have similar open loop characteristics does not mean that they will behave similarly under feedback.

Example: Similar Open Loop Different Closed Loop

Systems with the transfer functions

$$G_1(s) = \frac{1000}{s+1}, \qquad G_2(s) = \frac{1000a^2}{(s+1)(s+a)^2}$$

have very similar open loop responses for large values of a. This is illustrated in Figure 11 which shows the step responses of for $a = 100$. The differences between the step responses is barely noticeable in the figure. The closed loop systems obtained with unit feedback have the transfer functions

$$G_{1cl} = \frac{1000}{s+1001}, \qquad G_{2cl} = \frac{10^7}{(s-287)(s^2+86s+34879)}$$

The closed loop systems are very different because the system G_{2cl} is unstable.

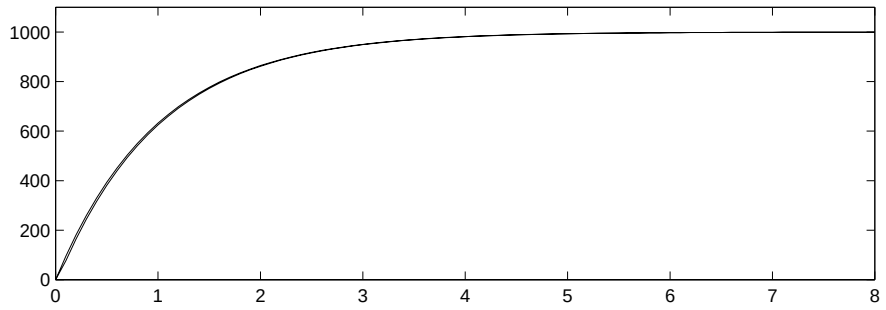

Fig. 11: Step responses for systems with the transfer functions $G_1(s) = 1000/(s+1)$ and $G_2(s) = 10^7/((s+1)(s+100)^2)$.

Example: Different Open Loop Similar Closed Loop

Systems with the transfer functions

$$G_1(s) = \frac{1000}{s+1}, \qquad G_2(s) = \frac{1000}{s-1}$$

have very different open loop properties because one system is unstable and the other is stable. The closed loop systems obtained with unit feedback are however

$$G_{1cl}(s) = \frac{1000}{s+1001} \qquad G_{2cl}(s) = \frac{1000}{s+999}$$

which are very close.

There are many examples of this in the literature of adaptive control where the importance of considering the closed loop properties of a model has been recognized for a long time, see e.g. [3]. The examples given above show that the naive way of comparing two systems by analyzing their responses to a given input signal is not appropriate for feedback control. The difficulty is that it does not work when one or both systems are unstable as in Example 5 and 5.

One approach is to compare the outputs when the inputs are restricted to the class of inputs that give bounded outputs. This approach was introduced in [51] and [19] using the notion of gap metric. Another approach was introduced in [44] and [45]. To describe this approach we assume that the process is described by the rational transfer function

$$P(s) = \frac{B(s)}{A(s)}$$

where $A(s)$ and $B(s)$ are polynomials. Introduce a stable polynomial $C(s)$ whose degree is not smaller than the degrees of $A(s)$ and $B(s)$. The transfer function $P(s)$

can then be written as

$$P(s) = \frac{B(s)/C(s)}{A(s)/C(s)} = \frac{D(s)}{N(s)} \tag{47}$$

Vidyasagar proposed to compare two systems by comparing the stable rational transfer functions D and N. This is called the graph metric. A difficulty was that the graph metric was difficult to compute.

Coprime Factorization

The polynomial C in (47) can be chosen in many different ways. We will now discuss a convenient choice.

We start by introducing a suitable concept. Two rational functions D and N are called coprime if there exist rational functions X and Y which satisfy the equation

$$XD + YN = 1$$

The condition for coprimeness is essentially that $D(s)$ and $N(s)$ do not have any common factors. The functions $D(s)$ and $N(s)$ will now be chosen so that

$$DD^* + NN^* = 1 \tag{48}$$

where we have used the notation $D^*(s) = D(-s)$. A factorization (47) of P where N and D satisfy (48) is called a normalized coprime factorization of P. Such a factorization the polynomials A and B in (47) do not have common factors.

5.1 Vinnicombe's Metric

A very nice solution to the problem of comparing two systems that is appropriate for feedback was given by Vinnicombe, see [46] and [48]. Consider two systems with the normalized coprime factorizations

$$P_1 = \frac{D_1}{N_1}$$
$$P_2 = \frac{D_2}{N_2}$$

To compare the systems it must be required that

$$\frac{1}{2\pi}\Delta \arg_\Gamma (N_1 N_2^* + D_1 D_2^*) = 0 \tag{49}$$

where Γ is the Nyquist contour. In the polynomial representation this condition implies that

$$\frac{1}{2\pi}\Delta \arg_\Gamma (B_1 B_2^* + A_1 A_2^*) = \deg A_2 \tag{50}$$

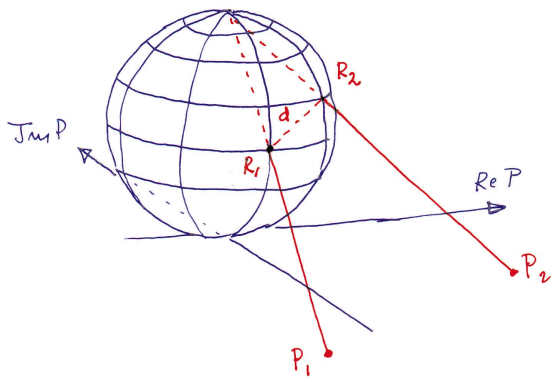

Fig. 12: Geometric interpretation of the Vinnicombe metric.

If the winding number constraint is satisfied the distance between the systems is defined as

$$\delta_\nu(P_1, P_2) = \sup_\omega \frac{|P_1(i\omega) - P_2(i\omega)|}{\sqrt{(1+|P_1(i\omega)|^2)(1+|P_2(i\omega)|^2)}} \tag{51}$$

We have $|\delta_\nu(P_1, P_2)| \leq 1$. If the winding number condition is not satisfied the distance is defined as $\delta_\nu = 1$. Vinnicombe showed that δ_ν is a metric and he called it the ν-gap metric.

Geometric Interpretation

Vinnicombe's metric is easy to compute and it also has a very nice geometric interpretation. The expression

$$d = \frac{|P_1 - P_2|}{\sqrt{(1+|P_1|^2)(1+|P_2|^2)}}$$

can be interpreted graphically as follows. Let P_1 and P_2 be two complex numbers. The Riemann sphere is located above the complex plane. It has diameter 1 and its south pole is at the origin of the complex plane. Points in the complex plane are projected onto the sphere by a line through the point and the north pole, see Figure 12. Let R_1 and R_2 be the projections of P_1 and P_2 on the Riemann sphere. The number d is then the shortest chordal distance between the points R_1 and R_2, see Figure 12.

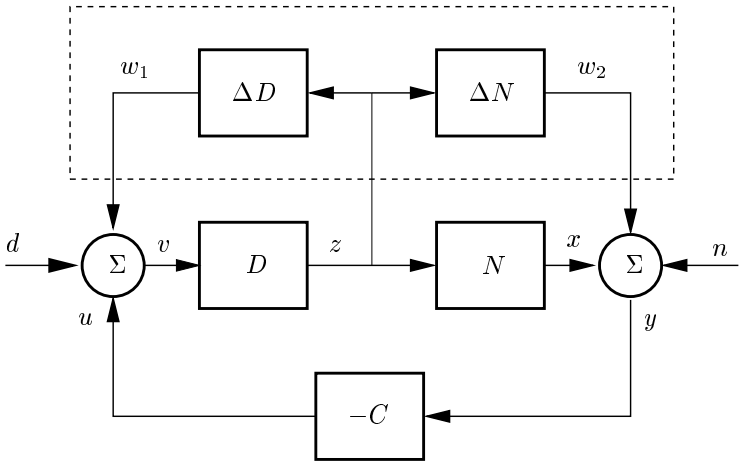

Fig. 13: Block diagram of a process with coprime factor uncertainty and a controller.

Coprime Factor Uncertainty

Classical sensitivity results such as (8) were obtained based on additive perturbations. The system P was perturbed to $P + \Delta P$ where ΔP is a stable transfer function. These types of perturbations are not well suited to deal with feedback systems as is illustrated by Example 5. A more sophisticated way to describe perturbations are required for this. The development of the metrics for systems gave good insight into what should be done. Uncertainty will be described in terms of the normalized coprime factorization of a system. Consider a system described by

$$P + \Delta P = \frac{N + \Delta N}{D + \Delta D} = ND^{-1} = D^{-1}N \tag{52}$$

where N and D is a normalized coprime factorization of P and the perturbations ΔN and $\Delta D$are stable proper transfer functions. Figure 13 shows a block diagram of the closed loop system with a the perturbed plant.

We will now investigate how large the perturbations can be without violating the stability condition. For the system in Figure 13 we have

$$z = \frac{D^{-1}}{1+PC}w_1 - \frac{D^{-1}C}{1+PC}w_2 = D^{-1}\begin{pmatrix} \frac{1}{1+PC} & -\frac{C}{1+PC} \end{pmatrix}\begin{pmatrix} w_1 \\ w_2 \end{pmatrix}$$

and

$$\begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \begin{pmatrix} \Delta D \\ \Delta N \end{pmatrix} z$$

The system can thus be represented with the block diagram in Figure 14. We can then

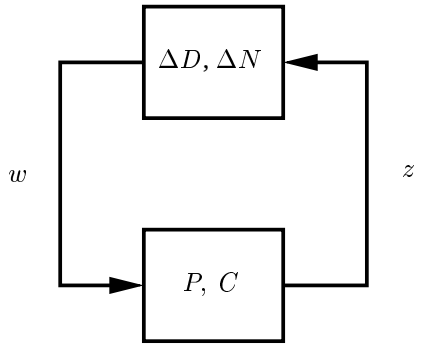

Fig. 14: Simplification of the block diagram in Figure 13.

invoke the small gain theorem and conclude that the perturbed system will be stable if the loop gain is less than one, see [14]. Hence

$$\left\| \begin{pmatrix} \Delta N \\ \Delta D \end{pmatrix} \right\|_\infty \left\| D^{-1} \begin{pmatrix} \frac{1}{1+PC} & -\frac{C}{1+PC} \end{pmatrix} \right\|_\infty < 1 \tag{53}$$

This condition can be simplified if we use the fact that N and D is a normalized coprime factorization. This gives

$$\left\| D^{-1} \begin{pmatrix} \frac{1}{1+PC} & -\frac{C}{1+PC} \end{pmatrix} \right\|_\infty = \left\| \begin{pmatrix} D \\ N \end{pmatrix} D^{-1} \begin{pmatrix} \frac{1}{1+PC} & -\frac{C}{1+PC} \end{pmatrix} \right\|_\infty$$

$$\left\| \begin{pmatrix} I \\ P \end{pmatrix} \begin{pmatrix} \frac{1}{1+PC} & -\frac{C}{1+PC} \end{pmatrix} \right\|_\infty = \|G(P,C)\|_\infty$$

where $G(P,C)$ denotes the system matrix

$$G(P,C) = \begin{pmatrix} \frac{1}{1+PC} & -\frac{C}{1+PC} \\ \frac{C}{1+PC} & -\frac{PC}{1+PC} \end{pmatrix} \tag{54}$$

Introducing

$$\gamma(P,C) = \sup_\omega \|G(P(i\omega), C(i\omega))\|_\infty \tag{55}$$

we thus find that the closed loop system is stable for all normalized coprime perturbations ΔD and ΔN such that

$$\left\| \begin{pmatrix} \Delta N \\ \Delta D \end{pmatrix} \right\|_\infty < \frac{1}{\gamma} \tag{56}$$

This equation is a natural generalization of Equation (8) in classical control theory. Notice that since the systems have one input and one output we have

$$|G(P,C)| = \bar{\sigma}G(P,C) = \frac{\sqrt{(1+|C|^2)(1+|P|^2)}}{|1+PC|} \tag{57}$$

and Equation (55) can thus be written

$$\gamma(P,C) = \sup_{\omega} \frac{\sqrt{(1+|C(i\omega)|^2)(1+|P(i\omega)|^2)}}{|1+P(i\omega)C(i\omega)|} \tag{58}$$

$\mathcal{H}_\infty$-Loop Shaping

The goal of $\mathcal{H}_\infty$ is to design control systems that are insensitive to model uncertainty. It follows from Equations (55) and (56) that this can be accomplished by finding a controller C that gives a stable closed loop system and minimizes the $\mathcal{H}_\infty$ norm of the transfer function $G(P,C)$ given by Equation eq:gmatrix. It follows from Equation (56) that such a design permits the largest deviation of the normalized coprime deviations.

It is interesting to observe that the transfer function G also describes the signal transmission from the disturbances d and n to v and x in Figure 2. A robust controller obtained in this way will also attenuate the disturbances very well.

A state-space solution to the $\mathcal{H}_\infty$ control problem was given in [17]. A loop shaping design procedure was developed in [34] and [35].

Frequency Weighting

In the design procedure presented in [35] it is also possible to introduce a frequency weighting W as a design parameter. The $\mathcal{H}_\infty$ control problem for the process $P' = PW$ is then solve giving the controller C'. The controller for the process P is then C. In this way it is possible to obtain controller that have high gain at specified frequency ranges and high frequency roll off.

Generalized Stability Margin

A generalization of the classical stability margin was also introduced in [34]. For a closed loop system consisting of the process P and the controller C such that the closed loop system we define the generalized stability margin as

$$b(P,C) = \begin{cases} \dfrac{1}{\gamma} & \text{if } (P,C) \text{ is stable,} \\ 0 & \text{otherwise} \end{cases} \tag{59}$$

Notice that the generalized stability margin takes values between 0 and 1. The margin is 0 if the system is unstable. A value close to one indicates a good margin of stability. Reasonable practical values of the margin are in the range of $1/3$ to $1/5$.

The H_∞-loop shaping in [34] gives a controller that maximizes the stability margin giving

$$b_{opt} = \sup_C b(P, C) \tag{60}$$

Vinnicombe's Theorems

A number of interesting theorems that relate model uncertainty to robustness have been derived in [47] and [48]. These results, which can be seen as the natural conclusion of the work that began in [50], give very nice relations between robust control and model uncertainty. Vinnicombe has proven the following results.

Proposition 1

Consider a nominal processes P and a controller C and a parameter β. Then the controller C stabilizes all plants P_1 such that $\delta_v(P, P_1) \leq \beta$, if and only if $b(P, C) > \beta$.

Proposition 2

Given a nominal process P, a perturbed process P_1 and a number $\beta < b_{opt}(P, C)$. Then (P_1, C) is stable for all compensators C, such that $b(P_1, C) > \beta$ if and only if $\delta(P, P_1) \leq \beta$.

The first proposition tells that a controller C designed for process P with a generalized stability margin greater than β will stabilize all processes P_1 in a δ_v environment of P provided that $\delta_v(P, P_1) < \beta$.

Proofs of these theorems are given in [48]. Vinnicombe has actually given sharper results which only requires the inequalities to hold pointwise for each frequency.

Connections to the Classical Control Theory

The $\mathcal{H}_\infty$-loop shaping cannot be directly related to the classical robustness criteria. The classical robustness criteria such as M_t and M_s depend only on the loop transfer function $L = PC$. But the generalized stability margin $b(P, C)$ and the generalized sensitivity $\gamma(P, C)$ depend on both P and C. The generalized stability margin will therefore change if the process transfer function is multiplied by a constant and the controller transfer function is divided by the same number. One reason for this is that the criterion (57) implicitly assumes that the disturbances ℓ and n have equal weight. This is a reasonable assumption if sufficient information about the disturbances are

available but very often we do not have this information. One possibility to formulate the design problem in this case is to choose the most favorable disturbance relation. This can be done by introducing a weighting of the disturbances. Let $P' = PW$ be the weighted process let CW^{-1} be the weighted controller CW^{-1}. We have

$$L' = P'C' = PC = L$$

The loop transfer function is invariant to W. The generalized sensitivity γ becomes

$$\gamma(P', C') = \sup_{\omega} \frac{\sqrt{(1 + |C(i\omega)W^{-1}(i\omega)|^2)(1 + |P(i\omega)W(i\omega)|^2)}}{|1 + P(i\omega)C(i\omega)|}$$

A straightforward calculation shows that γ' is minimized for the weight

$$W = \sqrt{|C|/|P|}$$

see [37]. This weight we get the following expression for the weighted sensitivity

$$\gamma^* = \sup_{\omega}(|S(i\omega)| + |T(i\omega)|) \tag{61}$$

Notice that the weighted γ^* only depends on the loop transfer function $L = PC$.

Using the weighted sensitivity function we thus obtain an interesting connection between $\mathcal{H}_\infty$-loop shaping and classical robustness theory. The number γ^* defined by Equation (55) and Equation (58) is a natural generalization of the maxima M_s, M_t of the sensitivity function and the complementary sensitivity. It is useful to introduce a combined sensitivity M by requiring that both M_s and M_t should at most be equal to M. The combined sensitivity implies that the Nyquist curve of the loop transfer function is outside a circle with diameter on

$$\Big(-\frac{M+1}{M-1}, -\frac{M-1}{M+1}\Big)$$

The following inequalities are shown

$$2M - 1 < \gamma < 2M$$
$$\frac{\gamma}{2} < M < \frac{\gamma + 1}{2}$$

in [37] where also sharper inequalities are presented.

The generalized stability margin b is also a natural generalization of the classical stability margin A_m. There are however some scale changes. The normal stability margin takes values between 1 and ∞ while the generalized stability margin takes values between zero and one. To get compatibility the classical stability margin should be redefined as the distance between the critical point and the intersection of the Nyquist curve with the negative real axis. Hence

$$A_m^* = 1 - \frac{1}{A_m}$$

References

[1] K. J. Åström. Limitations on control system performance. *European Journal of Control*, 6:1–19, 2000.

[2] Karl Johan Åström. Limitations on control system performance. In *European Control Conference*, Brussels, Belgium, July 1997.

[3] Karl Johan Åström and Björn Wittenmark. *Adaptive Control*. Addison-Wesley, Reading, Massachusetts, second edition, 1995.

[4] Karl Johan Åström and Björn Wittenmark. *Computer-Controlled Systems*. Prentice Hall, third edition, 1997.

[5] M. Athans and P. L. Falb. *Optimal Control*. McGraw-Hill, New York, 1966.

[6] T. Basar and P. Bernhard. *$\mathcal{H}^\infty$-Optimal control and related minimax design problems - A Dynamic game approach*. Birkhauser, Boston, 1991.

[7] R. Bellman. *Dynamic Programming*. Princeton University Press, New Jersey, 1957.

[8] R. Bellman, I. Glicksberg, and O. A. Gross. Some aspects of the mathematical theory of control processes. Technical Report R-313, The RAND Corporation, Santa Monica, Calif., 1958.

[9] H. S. Black. Stabilized feedback amplifiers. *Bell System Technical Journal*, 13:1–18, 1934.

[10] H. W. Bode. Relations between attenuation and phase in feedback amplifier design. *Bell System Technical Journal*, 19:421–454, 1940.

[11] H. W. Bode. *Network Analysis and Feedback Amplifier Design*. Van Nostrand, New York, 1945.

[12] R. Clarke, J. J. Burken, J. T. Bosworth, and Bauer J.E. X-29 flight control system - Lessons learned. *International Journal of Control*, 59(1):199–219, 1994.

[13] J. C. Clegg. A nonlinear integrator for servomechanis. *Trans. AIEE Part II*, 77:41–42, 1958.

[14] C. A. Desoer and M. Vidyasagar. *Feedback Systems: Input-Output Properties*. Academic Press, New York, 1975.

[15] J. C. Doyle. Guaranteed margins for LQG regulators. AC-23:756–757, 1978.

[16] J. C. Doyle, B. A. Francis, and A. R. Tannenbaum. *Feedback control theory*. Macmillan, New York, 1992.

[17] J. C. Doyle, K. Glover, P.P Khargonekar, and B. A. Francis. State-space solutions to standard h_2 and $\mathcal{H}_\infty$ control problems. AC-34:831–847, 1989.

[18] J. C. Doyle and G. Stein. Multivariable feedback design: Concepts for a classical/modern synthesis. AC-26:4–16, 1981.

[19] A. K. El-Sakkary. The gap metric: Robustness of stabilization of feedback systems. AC-26:240–247, 1985.

[20] Gene F. Franklin, J. David Powell, and Abbas Emami-Naeini. *Feedback Control of Dynamic Systems*. Addison-Wesley, third edition, 1994.

[21] Michael Green and D. J. N. Limebeer. *Linear Robust Control*. Prentice Hall, Englewood Cliffs, N.J., 1995.

[22] A. C. Hall. Application of circuit theory to the design of servomechanisms. *Journal of the Franklin Institute*, 242:279–307, 1946.

[23] J. W. Helton and O. Merino. *Classical control using H^∞ Methods*. SIAM, Philadelphia, 1999.

[24] I. M. Horowitz. *Synthesis of Feedback Systems*. Academic Press, New York, 1963.

[25] I. M. Horowitz and U. Shaked. Superiority of transfer function over state-variable methods in linear time-invariant feedback system design. AC-20:84–97, 1975.

[26] Isac M. Horowitz. *Quantitative Feedbac Design Theory (QFT)*. QFT Publications, Boulder, Colorado, 1993.

[27] H. M. James, N. B. Nichols, and R. S. Phillips. *Theory of Servomechanisms*. Mc Graw-Hill, New York, 1947.

[28] R. E. Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME*, 82D:35–45, march.

[29] R. E. Kalman. Contributions to the theory of optimal control. *Boletin de la Sociedad Matématica Mexicana*, 5:102–119, 1960.

[30] R. E. Kalman. When is a linear control system optimal? *Trans. ASME Ser. D: J. Basic Eng.*, 86:1–10, March 1964.

[31] R. E. Kalman and R. S. Bucy. New results in linear filtering and prediction theory. *Trans ASME (J. Basic Engineering)*, 83 D:95–108, 1961.

[32] R. E. Kalman, Y. Ho, and K. S. Narendra. *Controllability of Linear Dynamical Systems*, volume 1 of *Contributions to Differential Equations*. John Wiley & Sons, Inc., New York, 1963.

[33] Richard E. Klein. Using bicycles to teach system dynamics. *IEEE Control Systems Magazine*, CSM:4–9, April 1986.

[34] D.C. MacFarlane and K. Glover. *Robust controller design using normalized coprime factor plant descriptions*. Springer, New York, 1990.

[35] D.C. MacFarlane and K. Glover. A loop shaping design procedure using $\mathcal{H}_\infty$-synthesis. AC-37:759–769, 1992.

[36] H. Nyquist. Regeneration theory. *Bell System Technical Journal*, 11:126–147, 1932.

[37] H. Panagopoulos and K. J. Åström. PID control design and H_∞ loop shaping design of PI controllers based on non-convex optimization. In *Proceedings 1999 IEEE Int. Conf. Control Applications and the Symp. Computer Aided Control Systems Design (CCA'99&CACSD'99)*, Kohala Coast, Hawaii, August 1999.

[38] L. Pernebo. An algebraic theory for the design of controllers for linear multivariable system - Part I: Structure matrices and feedforward design. AC-26:171–182, 1981.

[39] L. Pernebo. An algebraic theory for the design of controllers for linear multivariable system - Part II: Feeback realizations and feedback design. AC-26:173–194, 1981.

[40] L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mischenko. *The Mathematical Theory of Optimal Processes*. John Wiley, New York, 1962.

[41] W. L. Rogers and D. J. Collins. X-29 $\mathcal{H}_\infty$ controller synthesis. *Journal of Guidance Control and Dynamics*, 15(4):962–967, 1992.

[42] M. G. Safonov and M. Athans. Gain and phase margins for multiloop lqg regulators. AC-22:173–179, 1977.

[43] R. Skelton. Model error concepts in control design. *International Journal of Control.*, 49:1725–1753, 1989.

[44] M Vidyasagar. The graph metric for unstable plants and robustness estimates for feedback stability. AC-29:403–417, 1984.

[45] Mathukumalli Vidyasagar. *Control System Synthesis: A Factorization Approach*. MIT Press, Cambridge, Massachusetts, 1985.

[46] G. Vinnicombe. Frequency domain uncertainty and the graph topology. AC-38:1371–1383, 1993.

[47] G. Vinnicombe. The robustness of feedback systems with bounded complexity controllers. AC-41:795–803, 1996.

[48] G. Vinnicombe. *Uncertainty and Feedback: $\mathcal{H}_\infty$ loop-shaping and the ν-gap metric*. Imperial College Press, London, 1999.

[49] G. Zames. Feedback and optimal sensitivity: Model reference transformations, multiplicative seminorms, and approximative inverse. AC-26(2):301–320, 1981.

[50] G. Zames. Feedback and optimal sensitivity: Model reference transformations, multiplicative seminorms, and approximative inverses. AC-26:301–320, 1981.

[51] G. Zames and A. K. El-Sakkary. Unstable systems and feedback: The gap metric. In *Proc. Allerton Conference*, pages 380–385, 1980.

[52] J. C. Zhou, J. C. Doyle, and K. Glover. *Robust and optimal control*. Prentice Hall, New Jersey, 1996.

[53] R. L. Bagley and R. A. Calico. Fractional-order state equations for the control of viscoelastic damped structures. J. Guidance, Control and Dynamics (14) 304-311, 1991.

[54] A. Oustaloup. La Commande CRONE. Hermes, Edition CNRS, Paris, France, 1991.

[55] A. Oustaloup. La Derivation Non Entiere: Theorie, Synthese et Applications. Hermes, Paris, France, 1995.

[56] I. Podlubny. Fractional Differential Equations. Academic Press, NY, 1999a.

[57] I. Podlubny. Fractional-order systems and PID controllers. IEEE Trans AC-44, 208–214, 1999.