NATO-ASI on Methods of Model Based Process Control, Antalya, Turkey, Aug. 7-17, 1994.

Frequency Domain Methods for Analysis and Design
II. CONTROLLABILITY ANALYSIS OF SISO SYSTEMS

Sigurd Skogestad *
Chemical Engineering
University of Trondheim, NTH
N-7034 Trondheim, Norway

These notes consists of two parts. Part I gives an overview of modern frequency-domain methods, including H-infinity methods and robust control using the structured singular velue, mu. Part II gives a tutorial introduction to controllability analysis for scalar systems using the frequency domain. Some readers may prefer to read Part II first. This part II is an extended version of some lecture notes used both for a graduate and undergraduate control course. Parts of the notes have previously been presented at the IFAC conferences ADCHEM'94 in Kyoto (mainly Section 3 on the background for the controllability analysis) and IPDC'94 in Baltimore (mainly Section 4 on the pH-application).

Abstract

The objective of this paper is to derive some fundamental results for controllability analysis of single-input single-output (SISO) systems. The effects of disturbances, delays, constraints and RHP-zeros are quantified. These results are applied to a neutralization process where it is shown that the process must be modified to get acceptable controllability.

1 INTRODUCTION

In process control courses the issues of controller design and stability analysis are often emphasized. However, in practice the following three issues are usually more important.

I. How well can the plant be controlled? Before attempting to start any controller design one should have some idea of how easy the plant actually is to control. Is it a difficult control problem? Indeed, does there even exist a controller which meets the required performance objectives?

II. What control strategy should be used? What to measure, what to manipulate, how to pair? In textbooks one finds qualitative rules to address this issue. For example in Seborg et al. (1989) one finds in a chapter called "The art of process control" the rules:

1. Control outputs that are not self-regulating
2. Control outputs that have favorable dynamic and static characteristics, i.e., there should exist an input with a significant, direct and rapid effect.
3. Select inputs that have large effects on the outputs.
4. Select inputs that rapidly effect the controlled variables

These rules are reasonable, but what is "self-regulating", "large", "rapid" and "direct". One objective of this paper is to quantify this.

III. How should the process be changed to improve control ? For example, one may want to design a buffer tank for damping a disturbance, or one may want to know how fast a measurement should be to get acceptable control.

Controllability analysis. All the above three questions are related to the inherent control characteristics of the process itself, that is, to what is denoted the *controllability* of the process. We shall use the following definition:

(Input-output) controllability is the ability to achieve acceptable control performance, that is, to keep the outputs (y) within specified bounds or displacements from their setpoints (r), in spite of unknown variations such as disturbances (d) and plant changes, using available inputs (u) and available measurements (e.g.,y_m or d_m).

In summary, a plant is controllable if there *exists* a controller (connecting measurements and inputs) that yields acceptable performance for all expected plant variations. Thus, controllability is independent of the controller, and is solely a property of the plant (process) only. It can only be affected by changing the plant itself, that is, by *design modifications.* Surprisingly, in spite of the fact that mathematical methods are used extensively for control system design, the methods available when it comes to controllability analysis are usually qualitative. In most cases the "simulation approach" is used. However, this requires a specific controller design and specific values of

*E-mail: skoge@kjemi.unit.no; phone: +47-73-594154; fax: +47-73-594080

disturbances and setpoint changes. In the end one never really knows if the assessment is a fundamental property of the plant or if it depends on the specific choices made.

The objective of this paper is to present quantitative controllability measures which can replace this *ad hoc* procedure. The paper deals with scalar (SISO) systems, but all the tools presented may be generalized to multivariable (MIMO) systems. Disturbances are considered in detail, but model uncertainty, which also necessitates the use of feedback control, is not included in this paper. Linear control theory is used, and most of the tools make use of the frequency response. One reason for this is the very useful idea of "bandwidth" which is a purely frequency-domain concept.

One shortcoming with the controllability analysis presented in this paper is that all the measures are linear. This may seem to be very restrictive, but in most cases it is not. In fact, one of the most important nonlinearities, namely input constraints, can be handled with the linear approach. To deal with slowly varying changes one may perform a controllability analysis at several selected operating points. As a last step of the controllability analysis one should perform some nonlinear simulations to confirm the results of the linear controllability analysis. The experience from a large number of case studies has been that the agreement is generally very good.

Remarks on the definition of controllability. The above definition is in agreement with one's intuitive feeling about the term, and is also how the term was used originally in the control literature. For example, Ziegler and Nichols (1943) define controllability as *"the ability of the process to achieve and maintain the desired equilibrium value"*. Rosenbrock (1970, p. 161) notes that *"in engineering practice, a system is called controllable if it possible to achieve the specified aims of control, whatever these may be"*. Unfortunately, in the 60's the term "controllability" became synonymous with the rather narrow concept of "state controllability" introduced by Kalman, and the term is still used in this restrictive manner by the system theory community. "State controllability" is the ability to bring a system from a given initial state to any final state (but with no regard to the dynamic repsonse between and after these two states). This concept is of interest for realizations and numerical calculations, but as long as we know that all the unstable modes are both controllable and observable, it has little practical significance. For example, Rosenbrock (1970, p. 177) notes that "most industrial plants are controlled quite satisfactorily though they are not [state] controllable". He also remarks that "the chief point to be stressed is that controllability is an engineering term with a wide connotation. To restict its meaning to one particular type of controllability seems wrong, and leads to confusion."To avoid confusion with Kalman's state controllability, Morari (1983) introduced the term "dynamic resilience". However, this term does not capture the fact that "controllability" is related to control, and so instead we propose to use the term "input-output controllability" to make the distinction with "state controllability".

Finally, one should note that the term "controllable" does not quite mean the same as "easy to control". The latter usually means that one can "easily design a simple controller" and get acceptable performance. On the other hand, "controllable" means that there *exists* a controller which yields acceptable performance, although this controller may be very complex and require a detailed model of the plant. It is possible to restict the definition of controllability to make it closer to the term "easy to control". For example, one may require the controller to be linear (as is done throughout this paper), or to be decentralized, or to be of a certain order or form (e.g., PID controller).

One may also consider controllability using feedback control, which is the main topic in this paper, although we do also have some discussion on the use of feedforward control.

The link between process design and control. The terms controllability provides the link between process design and control. This is explained very nicely by Ziegler and Nichols (1943):

"The finest controller made, when applied to a miserably designed process, may not deliver the desired performance. True, on badly designed processes, advanced controllers are able to eke out better results than olde models, but on these processes, there is a definite end point which can be approached by instrumentation and it falls short of perfection. The chronology in process design is evidently wrong. Nowadays an engineer first designs his equipment so that it will be capable of performing its intended function at the normal throughput rate plus a safety factor. The control engineer or instrumentman is then told to put on a controller capable of maintaining the static equilibrium for which the apparatus was designed. .... When the plant is started, however, it may be belatedly discovered that, in spite of the correct equipment design for steady-state condition and the correct instrument selection, control results are not within the desired tolerance. A long expensive process of "cut and try" is then begun in order to make the equipment work. Both engineers realize that some factor in equipment design was neglected but generally can neither identify the missing ingredient nor correct it in future design.

The missing characteristic can be called "controllability", the ability of the process to achieve and maintain the desired equilibrium value. Design for steady-state conditions is not enough if exact maintenance of variables is necessary. "

Ziegler and Nichols then point out that although "a great many factors affecting controllability have been identified" the problem is complex, and "as it now stands the plant designer is almost justified in disregarding the entire matter. .... Sooner or later, however, these factors affecting process controllability will have to be smoked out and reduced to definite "good-practice" rules which will be as much a part of equipement design as safety factors".

It it probably fair to say that progress has been slow, and now, more than 50 years later, such good-practice rules are still not in common use. It is hoped, however, that this tutorial paper will contribute to the "smoking-out" process.

Design modifications. As pointed out above, controllability can only be affected by design modifications. These may include:

1. Change the apparatus itself (type, size, etc.)
2. Relocate sensor and actuators
3. Add new equipment to dampen disturbances, for example, buffer tanks.
4. Add extra sensors for measurement (cascade control)
5. Add extra actuators (parallell control)
6. Change the control objectives
7. Change the control structure of the lower levels

In most cases controllability is improved by bringing the actuator and measurement device closer together in order to improve the speed of response, for example, by reducing the process delay. This applies to the first items above, which usually are quite problem specific and are not treated in this paper.

It it arguable whether or not the last two items are design modifications, but at least they address issues which come before the actual controller design. The last issue is important because control systems are usually designed in a hierarchical manner, and the lower-level loops are assumed closed when designing the control system at a given level. Thus, a change in the lower-level control structure may drastically change the achievable control performance of the levels above, and therefore may be viewed as a design modification as seen from the level above.

Previous work on controllability analysis. The topic has been addressed in many application papers, but mostly on an *ad hoc* basis since the theoretical basis for a controllability analysis has been relatively poor (one reason for this is probably the unfortunate use of the term in the meaning of state controllability, which led to the belief that there was nothing more to).

Except for the initial work by Ziegler and Nichols (1943), there does not seem to have been much progress on input-output controllability analysis until Rosenbrock (1966, 1970) presented a thorough discussion on the various definitions of state controllability and observability, and introduced similar concepts in terms of the *outputs*. This led to the introduction of the term "functional controllability" (which for scalar systems is equivalent to requiring that the transfer function $g(s)$ is not identically equal to zero) and to the important notion of right half plane (RHP) zeros (which for scalar systems is directly related to inverse responses). Another important step towards a quantitative analysis was made by Morari (1983) who made use of the notion of "perfect control" in an attempt to quantify the term controllability. Balchen and Mumme (1988, pp. 16-21, pp.47-48) present some nice controllability guidelines which are more specific than the rules from Seborg er al. (1989) given above, but most of them lack a theoretical justification.

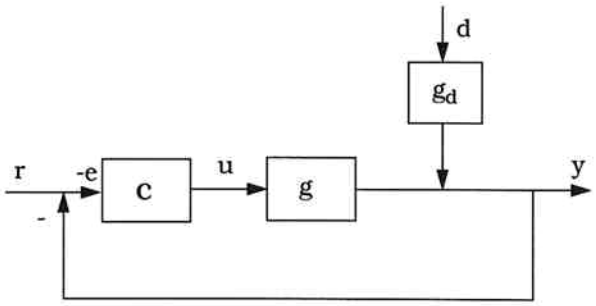

Figure 1: Block diagram of feedback control system.

One important issue which was missing from most of Morari's and Rosenbrock's analyses was an explicit consideration of disturbances. Disturbances have of course been discussed in many application papers, but only recently have their relationship to controllability been treated in a systematic manner (e.g., Skogestad and Wolff, 1992).

The tools for controllability analysis are now reaching a more mature state, but still the fundamental ideas are not well known. The objective of this paper is to present the ideas for scalar systems in a tutorial manner. For decentralized control of multivariable processes the results may be generalized directly by introducing the Closed Loop Disturbance Gain (CLDG) and the Performance Relative Gain Array (PRGA) (Hovd and Skogestad, 1992).

2 LINEAR CONTROL THEORY

Notation. Consider a linear process model in terms of deviation variables

$$y = gu + g_d d \tag{1}$$

Here y denotes the output, u the manipulated input and d the disturbance (including what is of-

ten referred to as "load changes"). $g(s)$ and $g_d(s)$ are transfer function models for the effect on the output of the input and disturbance, and all controllability results in this paper are based on this information. The Laplace variable s is often omitted to simplify notation. The control error e is defined as

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

where r denotes the reference value (setpoint) for the output.

Feedback control. Consider a simple feedback scheme

$$u = c(s)(r - y) \tag{3}$$

where $c(s)$ is the controller. Eliminating u from equations (1) and (3) yields the closed-loop response

$$y = Tr + Sg_d d \tag{4}$$

Here the sensitivity is $S = (I + gc)^{-1}$ and the complementary sensitivity is $T = gc(I + gc)^{-1} = 1 - S$. The transfer function around the feedback loop is denoted L. In this case $L = gc$. The corresponding input signal is

$$u = -ce = cSr - cSg_d d \tag{5}$$

The frequency domain. Most of the results in this paper are based on the frequency domain. Unfortunately, few process engineers feel comfortable with this domain, so a simple introduction is given first. Consider the effect of a small change in the input (input signal) u on the output (output signal) y. In the Laplace domain this may be represented as

$$\Delta y(s) = g(s)\Delta u(s)$$

where Δu represents a small change in the input (independent variable), and $\Delta y(s)$ is the resulting change in the output. $g(s)$ is the transfer function of the system. The Δ is included to show explicitly that we are dealing with deviation variables, but since we will only deal with deviation variables in this paper the Δ will be omitted to simplify notation.

Let us now consider the time domain where most engineer feel more comfortable. The problem with the time domain is that we have to consider specific input signals $u(t)$ and have to recompute $y(t)$ for each signal. The favorite input test signal for engineers is a step. However, in general a step response does not provide sufficient information for a controllability analysis. Therefore the frequency domain shoul be used.

The physical interpretation of the frequency domain for a system $y = g(s)u$ is a follows: A persistent sinusoidal input with frequency ω, $u(t) = u_0 sin(\omega t)$, yields a persistent sinusoidal output with the same frequency, $y(t) = y_0 sin(\omega t + \phi)$, but shifted in phase by ϕ. This is shown graphically in Figure 2 for a first-order system with time delay,

$$g(s) = \frac{ke^{-\theta s}}{1 + \tau s}; \quad k = 5, \theta = 2, \tau = 10 \tag{6}$$

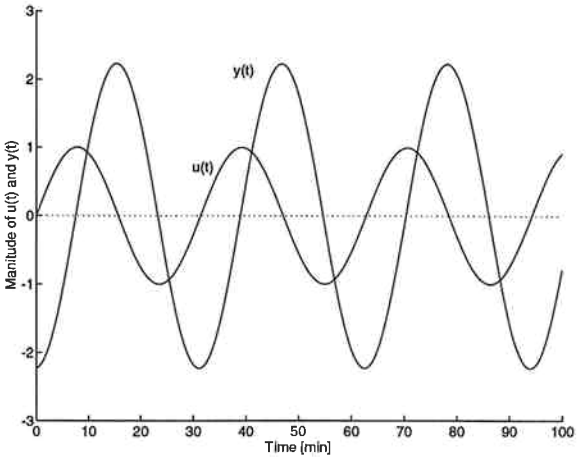

Figure 2: Sinusoidal response for system $g(s) = 5e^{-2s}/(1 + 10s)$ at frequency $\omega = 0.2$ [rad/min]. Period $P = 2\pi/\omega = 31.4$ min. Gain $|g(j\omega)| = 5/\sqrt{1 + (10\omega)^2} = 2.24$. Phase shift $\phi = -\arctan(10\omega) - 2\omega = -1.51$ rad $= -86.3^o$ corresponding to time shift $\Delta t = -\phi/\omega = 7.6$ min.

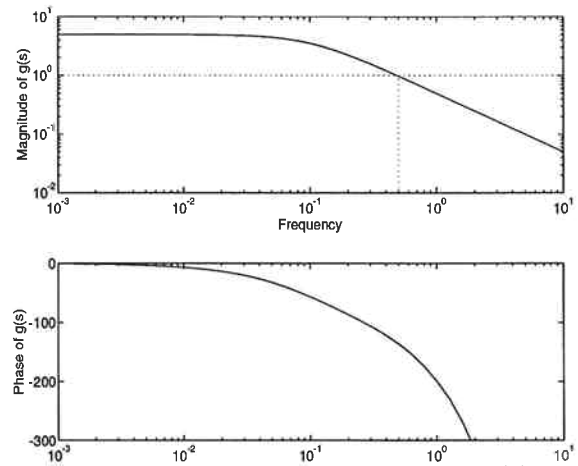

Figure 3: Frequency response of $g(s) = 5e^{-2s}/(1 + 10s)$.

It is useful to have in mind this physical picture of the frequency domain when one interprets the controllability results presented later. The magnitude y_0 and phase shift ϕ is easily computed from the Laplace transform $g(s)$ by inserting the imaginary number $s = j\omega$ and evaluating the magnitude and phase of the resulting complex number:

$$y_0/u_0 = |g(j\omega)|; \quad \phi = \angle g(j\omega) \text{ [rad]} \tag{7}$$

In this paper we use a "frequency-by-frequency" approach and at each frequency consider the response to a sinusoidal input of unit magnitude ($u_0(\omega) = 1$) as illustrated in Fig 2. This results in the "frequency response" of the system where we consider the system gain $y_o(\omega)/u_o(\omega) = |g(j\omega)|$ (and possibly the phase shift $\angle g(j\omega)$) as a function of ω. Graphically, this frequency response is usually represented in a Bode plot with a log-scale for frequency and gain.

In Fig. 3 the frequency response (Bode-plot) is

shown for the example in (6). We note that in this case both the gain (amplitude) and phase falls monotonically with frequency. This is quite common for open-loop (no feedback control) chemical engineering systems. The delay θ will only shift the sinusoid in time, and thus affects the phase but not the gain. The system gain $|g(j\omega)|$ is equal to k at low frequencies (this is the steady-state gain and is obtained by setting $s = 0$). The gain remains relatively constant up to a frequency of about $1/\tau$ where it start falling sharply. Physically, the system responds too slowly to let high-frequency ("fast") inputs have much effect on the outputs, that is, high-frequency sinusoidal inputs are smoothened ("dampened", "attenuated") by the system dynamics.

Assume that $k > 1$ and note for later reference the frequency ω_d where the gain is 1, that is, $|g(j\omega_d)| = 1$ (this frequency is of particular interest when $g(s)$ is the disturbance model, and this is the reason for the subscript d). The exact value is given by $k/\sqrt{1+(\omega_d\tau)^2} = 1$, but we often use the asymptotic approximation $k/(\omega_d\tau) \approx 1$, and obtain

$$\omega_d \approx k/\tau \tag{8}$$

Thus, we see that ω_d is large if the steady-state gain k is large (the input has a large effect on the output) or if the time constant τ is small (the input has a fast effect on the output).

Frequency responses may be obtained for any transfer function. In this paper we consider frequency responses of three transfer functions: $g(j\omega)$ (effect of manipulated inputs u on outputs y), $g_d(j\omega)$ (effect of disturbances d on outputs y), and $L = gc(j\omega)$ (frequency response of loop transfer function). The frequency responses of g and g_d are often similar to the response shown in Fig. 3, whereas the magnitude of $L = gc$ is often infinite at low frequency because the controller $c(s)$ usually contains an integrator.

Bandwidth. Here bandwidth is defined as *the frequency ω_B where the loop gain is one in magnitude*, i.e. $|L(j\omega_B)| = 1$ (or more precicely where the low-frequency asymptote of $|L|$ first crosses 1 from above). This frequency is often called the "crossover frequency".

At frequencies lower than the bandwidth ($\omega < \omega_B$) feedback is effective and will affect the frequency response. However, for sinusoidal signals (for example, a disturbance) with frequencies higher than ω_B the response will not be much affected by the feedback.

Other definitions of bandwidth are also in use, for example, as the frequency where $|S(j\omega)| = 0.7$ or the frequency where $|T(j\omega)| = 0.7$. The above definition in terms of the loop transfer function is preferred because it is simple. It usually yields a value between the two alternative definitions in terms of $|S|$ and $|T|$.

A frequency domain analysis, in particular in the frequency-region corresponding to the bandwidth, is very useful for systems under feedback control. This is the case even when the disturbances and setpoints entering the system are *not* sinusoids. One reason for this is that the feedback control system will usually amplify frequencies corresponding to the closed-loop bandwidth, ω_B. [1]. For example, the effect of disturbances is usually largest around the bandwidth frequency; slower disturbances are attenuated by the feedback control, and faster disturbances are usually attenuated by the process itself. Thus, the magnitude of g_d at the bandwidth frequency, $|g_d(j\omega_B)|$, is usually a very good approximation of the worst-case amplification of a disturbance when using feedback control. This means that if we can somehow estimate the best achievable ω_B, we can say a lot about how sensitive the system is to disturbances under feedback control. The implication for design is to look for plant modifications which makes the plant more "self-regulating" in terms of reducing the magnitude of $|g_d(j\omega_B)|$.

For pure feedforward control the frequency domain may not be quite as relevant. For example, if the disturbances are always steps then a step response analysis may be more relevant. However, in many cases the disturbances are sinusoidal since they are generated from feedback loops in other parts of the system.

3 CONTROLLABILITY ANALYSIS

Scaling. The interpretation of most measures presented in this paper assumes that the transfer functions g and g_d are in terms of scaled variables. The first step in a controllability analysis is therefore to scale (normalize) all variables (input, disturbance, output) to be less than 1 in magnitude (i.e., within the interval -1 to 1).

Thus, in the following *we assume that the signals are persistent sinusoids, and that g and g_d have been scaled, such that at each frequency the allowed input $|u(j\omega)| < 1$, the expected disturbance $|d(j\omega)| < 1$, the allowed control error $|e(j\omega)| < 1$, and the expected reference signal $|r(j\omega)| < R_{max}$*. Note that e and r are measured in the same units so R_{max} is the magnitude of the expected setpoint change relative to the allowed control error. The detailed scaling procedure is outlined in the Appendix.

The ideal controller and plant inversion. The objective of the control system is to manipulate u such that the control error e remains small in spite of disturbances and changes in the setpoint. The ideal controller will accomplish this by inverting the process (Morari, 1983) such that

[1] The bandwidth frequency will often show up as oscillations in the time response and we usually have $\omega_B \approx 2\pi/P$ where P is the period of the oscillations.

the manipulated input becomes (set $y = r$ in (1) and solve for u):

$$u = g^{-1}r - g^{-1}g_d d \tag{9}$$

For example, an ideal feedforward controller operates in this manner. Usually, the disturbance is not measured and feedback control is used instead. As may be expected, the input signal generated under feedback is also given by Eq.(9) at frequencies where feedback is effective. To see this, consider Eq. (5) and use the fact $cS = g^{-1}T$ to derive the following expression for the input signal under feedback control

$$u = g^{-1}Tr - g^{-1}Tg_d d \tag{10}$$

At low frequencies, $\omega < \omega_B$, where $|gc(j\omega)| > 1$ and feedback is effective we have $S \approx 0$ and $T \approx 1$, and we rederive (9). Consequently ideal control (inversion) requires *fast* feedback control (high bandwidth).

On the other hand, inherent limitations of the system may prevent fast control. The limitations may include constraints on the allowed input signal u and non-minimum phase elements in $g(s)$ such as time delay and right half plane zeros. *If these requirements for high and low bandwidth are in conflict then controllability is poor.* The objective of the remaining part of this section is to quantify these statements. The results are derived for feedback control, although some of them also apply to feedforward control.

3.1 Disturbances and bandwidth

The effect of a disturbance on the output at a frequency ω in the absence of control is

$$y(j\omega) = g_d(j\omega)d(j\omega) \tag{11}$$

(we are here assuming that $r = 0$ such that the control error $e = y$). The worst-case disturbance at this frequency has magnitude 1, i.e., $|d(j\omega)| = 1$. Furthermore, at each frequency the output should be less than 1 in magnitude, i.e., we need control if $|y(j\omega)| > 1$. *Consequently, at frequencies where $|g_d(j\omega)| > 1$ we need control (feedforward or feedback) in order to prevent the output exceeding its allowed bound.* Typically, $|g_d(j\omega)|$ is larger than 1 at low frequencies and drops to zero at high frequencies. *In this case the frequency, ω_d, where $|g_d(j\omega_d)| = 1$ is a useful controllability measure:* At frequencies lower than ω_d we need control to reject the disturbance, and thus ω_d provides a minimum bandwidth requirement for control, and we have the approximate requirement

$$\omega_B > \omega_d \tag{12}$$

Example. Consider the disturbance model (recall Fig.3)

$$g_d(s) = k_d e^{-\theta_d s}/(1 + \tau_d s) \tag{13}$$

where $k_d = 5$ and $\tau_d = 10$ [min]. Scaling has been applied to g_d, so this means that with no control, the effect of disturbances on the outputs at low frequencies is $k_d = 5$ times larger than what we allow. Thus control is required, and since g_d crosses 1 at a frequency $\omega_d \approx k_d/\tau_d = 0.5$ rad/min, the minimum bandwidth requirement for disturbance rejection using feedback control is $\omega_B > 0.5$ rad/min.

Remarks.

1. Scaling is critical for any controllability measure involving disturbance rejection.
2. Recall the following rule from the introduction:
   - Control outputs that are not self-regulating

   This rule can be quantified as follows: Control outputs y for which $|g_d(j\omega)| > 1$ at some frequency.
3. In words we have proved that "large disturbances with a fast effect" require fast control. Specifically, if the disturbance is increased, then to get acceptable performance the bandwidth (speed of response) of the control system has to be increased.
4. To be more specific assume that the disturbance is increased by a factor f, and assume that at frequency ω_d the slope of $|g_d(j\omega)|$ on the log-log Bode-plot is $-\beta$, that is, $g_d \sim 1/s^\beta$ at the frequency ω_d (in the example above $\beta = 1$). Then the bandwidth has to be increased by a factor $f^{1/\beta}$ to counteract the increased disturbance.
5. Note that a delay in the disturbance model has no effect on the required bandwidth.
6. On the other hand, with feedforward control where the disturbance is measured, a delay in the disturbance model makes control easier.

3.2 Input constraints

Consider the response to a "worst-case" sinusoidal disturbance of magnitude 1 ($|d(j\omega)| = 1$) and assume $r = 0$. From Eq.(9) the input magnitude needed for perfect control ($e = 0$) is

$$|u| = |g^{-1}g_d d| = |g_d|/|g| \tag{14}$$

(Strictly speaking, perfect control is not required, and the input needed for "acceptable" control ($|e| < 1$) is $|u| = (|g_d| - 1)/|g|$. The difference is small at frequencies where $|g_d|$ is larger than 1, and the input needed for perfect control will be used in the following[2]).

Consider frequencies $\omega < \omega_d$ where control is needed to reject disturbances. The requirement is that $|u(j\omega)| \leq 1$ at each frequency. To fulfill this one must require

$$|g(j\omega)| > |g_d(j\omega)|, \quad \forall \omega < \omega_d \tag{15}$$

Similarly, to perfectly track a setpoint $r(j\omega) = R_{max}$ at each frequency with $|u| < 1$ one must from Eq.(9) require

[2] For multivariable systems the differences between perfect and acceptable control may be large if the plant is ill-conditioned.

$$|g(j\omega)| > R_{max}, \quad \forall \omega < \omega_r \tag{16}$$

where ω_r is the frequency up to which setpoint tracking is desired.

Remarks.

1. Recall the following rule from the introduction:
   - Select inputs that have large effects on the outputs.

   This rule may be quantified as follows: In terms of scaled variables we should have $|g| > |g_d|$ at frequencies where $|g_d| > 1$, and additionally we should have $|g| > R_{max}$ at frequencies where setpoint tracking is desired.
2. The following remark applies also to the previous subsection on disturbances and bandwidth. If there are several disturbances then they should be analyzed individually to identify the most difficult ones. This could be the starting point for proposing design modifications. (The worst-case combined effect of several disturbances may be obtained by simply adding together their individual effects. For example, let the effect of disturbance d_k on y be g_{dk}. Then to consider the worst-case combination one may simply replace $|g_d|$ by $\sum_k |g_{dk}|$ in the above expressions.)
3. For unstable plants we need a minimum bandwidth p to stabilize the system (see below). In this case we need $|g| > |g_d|$ up to the frequency p. Otherwise, the input will saturate, and the plant can not be stabilized.
4. The bounds (15) and (16) are strictly speaking only *necessary* conditions for controllability. This follows since we have used a frequency-by-frequency analysis and have not considered whether there actually exist a causal controller that can achieve the performance required by perfect control. In other words, we must always satisfy the bounds (15) and (16) (or at least the modified bound for "acceptable" control), but this may not be sufficient to avoid input constraints in the presence of delays or RHP-zeros.
5. Since the input needed for perfect control is independent of the control implementation, the bounds (15) and (16) also apply to feedforward control.

3.3 Time delay and right half plane zeros

It is well-known that time delays and right half plane (RHP) zeros limit the achievable speed of response. We shall here quantify this statement in terms of upper bounds on the allowed bandwidth. The derivation makes use of the complementary sensitivity function T which for a controller without a prefilter on r is the transfer function from setpoint to output, i.e., $y = Tr$.

Consider an "ideal" controller which is integral square error (ISE)-optimal for the case with step changes in the setpoint (this controller is "ideal" in the sense that it may not be realizable in practice because the required inputs may be infinite). That is, the objective is to minimize $\int_o^\infty |e(t)|^2 dt$ for the case where $r(t)$ is a step, and with no penalty on the input u. In this case the corresponding "ideal" complementary sensitivity for a plant with RHP-zeros at z_i and a time delay θ is (see Morari and Zafiriou, 1989, p. 58)

$$T = \prod_i \frac{-s + z_i}{s + \bar{z}_i} e^{-\theta s} \tag{17}$$

where $\bar{z}_i$ is the complex conjugate of z_i. Note that T is "all pass" since $|T(j\omega)| = 1$ at all frequencies. Given T we can compute the loop transfer function $L = T/(1-T)$, and then obtain the bandwidth as the frequency where $|L(j\omega)|$ crosses 1.

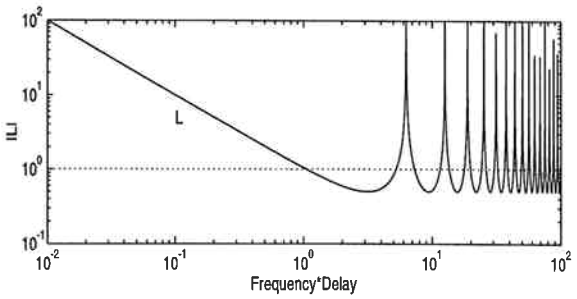

Figure 4: "Ideal" loop transfer function for plant with delay

Time delay. Consider a plant with a time delay, that is, $g(s)$ contains the term $e^{-\theta s}$. The "ideal" controller can "invert away" most of the dynamics in $g(s)$, but it cannot remove the delay. Thus, even the "ideal" complementary sensitivity function will contain the delay,

$$T = e^{-\theta s} \tag{18}$$

The loop transfer function corresponding to this ideal response is $L = T/(1-T) = e^{-\theta s}/(1-e^{-\theta s})$. The magnitude $|L|$ is plotted in Figure 4. At low frequencies, $\omega\theta < 1$, we have $e^{-\theta s} \approx 1 - \theta s$ (by a Taylor series expansion of the exponential) and $L \approx \frac{1}{\theta s}$, and thus the low frequency asymptote of $|L(j\omega)|$ crosses 1 at frequency $1/\theta$ (the exact frequency where $|L(j\omega)|$ crosses 1 in Fig. 4 is $\frac{\pi}{3}\frac{1}{\theta} = 1.05/\theta$). This is the bandwidth frequency. In practice, the "ideal" controller cannot be realized, and so this analysis provides an upper bound on the bandwidth of approximately

$$\omega_B < 1/\theta \tag{19}$$

Real RHP zero. Consider a plant with an inverse response, that is, $g(s)$ contains a term $(-s + z)$ corresponding to a real RHP zero at z. Again, the "ideal" controller cannot remove the effect of this RHP zero. Thus, even the "ideal" complementary sensitivity function will contain the RHP-zero

$$T = \frac{-s + z}{s + z} \tag{20}$$

The loop transfer function corresponding to this ideal response is $L = (-s + z)/2s$. The magnitude $|L|$ is plotted in Figure 5. The low frequency

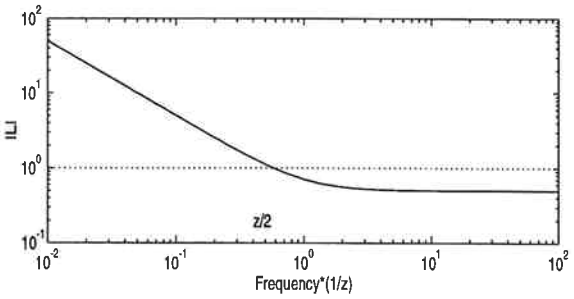

Figure 5: "Ideal" loop transfer function for plant with RHP zero.

asymptote of $|L(j\omega)|$ crosses 1 at frequency $z/2$. In practice, the "ideal" controller cannot be realized, and we obtain an upper bound on the bandwidth of approximately

$$\omega_B < \frac{z}{2} \tag{21}$$

Remarks on bounds (19) and (21).

1. The bounds are independent of scaling.
2. The bounds provide a quantification of the rules
   - Control outputs that have favorable dynamic and static characteristics, i.e., there should exist an input with a significant, direct and rapid effect.
   - Select inputs that rapidly effect the controlled variables
3. To reject a disturbance we obtained the requirement $\omega_B > \omega_d$. Combining this with (19) yields an upper limit on the allowed delay, $\theta < 1/\omega_d$. Similarly, we get $\omega_d < z/2$.
4. It will be possible to have a slightly higher bandwidth than given by these two bounds, but only at the expense of a very oscillatory response (corresponding to a large peak in T and S).
5. The above derivation applies when the delay or RHP zero is in the plant itself (between the input u and the output y). However, with feedback control a delay or RHP zero in the measurement of y yields similar limitations, and the above bounds still apply.
6. The bound (21) for RHP-zeros assumes that we want to use u for "slow control" of y for frequencies lower than $z/2$. However, if this is not the case, then one may instead use u for fast (transient) control of y for frequencies higher than z (with the sign of the controller gain reversed compared to the "normal" case[3]). This is further discussed below. This assumes that we are not concerned with the long-term behavior of the output[4], or that we have a "parallell" control system where another input may be used for long-term control of the output.
7. Zeros in the left half plane, corresponding to "overshoots" in the time response, do not present a *fundamental* limitation on control, but *in practice* a LHP-zero located close to the origin may cause problems. First, one may encounter problems with input constraints at low frequency (because the steady-state gain is often low). Second, a simple controller can probably not be used. Specifically, a simple PID controller contains no poles that can be used to counteract the effect of a LHP zero.
8. Similar restrictions to those given by the bounds above also apply to feedforward control. This follows since the ideal T in (17) corresponds to the input u which minimizes the ISE of the output irrespective of the control implementation.

[3]To see that the controller gain must be reversed one may consider the formulas in Morari and Zafiriou (1989, p. 63) where we see that the sign of $\tilde{q}$ and thus of the feedback controller c is zero if the desired response time τ is such that $\tau = 1/z$.

[4]In process control we are usually concerned with the long-time behavior and often require perfect control at steady-state, but there are cases where the control objective is to reject transient disturbances and the steady-state does not matter. One example is the use of a buffer tank to eliminate high-frequency flowrate disturbances.

Further remarks on the limitation of RHP-zeros and the use of positive feedback.

In remark 6 it was claimed that one may essentially choose whether a RHP-zero should pose control limitations at low or high frequencies. This may need some further discussion, as it was certainly not clear to me when I first looked at it.

Let us start with a simple time domain interpretation. A RHP-zero corresponds to an inverse response, that is, to a gain reversal. Usually, one wants good steady-state control and applies negative feedback. In this case one cannot get good transient response as one has to wait until the effect of the apllied input goes in the right direction (after the gain reversal). On the other hand, if one wants good transient response then one can react immediately, but since the gain is in the opposite direction one must positive feedback, and due to the gain reversal one gets poor steady-state control.

Of course, one can use a controller which itself has a RHP-zero and thus a gain reversal, but as shown below this does not really help. Another, even more tempting approach is to use an unstable controller with a pole at the location of the RHP-zero of the plant. However, it is well known that this does not work as it results in an internally unstable system where something eventually will blow up.

Let us now consider an example in more detail. The problem is to design a feedback controller for the plant

$$g(s) = \frac{-s+z}{s+z} \tag{22}$$

which has a a RHP-zero at z. We shall first consider negative feedback, then positive feedback, and finally the combination of the two.

I. Negative feedback. Let us first consider the conventional case where we want good steady-state control and where the bandwidth is approximately limited to $\omega_B < z/2$. The "ideal" controller in terms of minimizing the integral square error to step setpoints has loop transfer function $L = (-s+z)/2s$ corresponding to a PI-controller

$$c(s) = K\frac{s+z}{s} \tag{23}$$

with gain $K = 0.5$. With the controller (23) the sensitivity function is

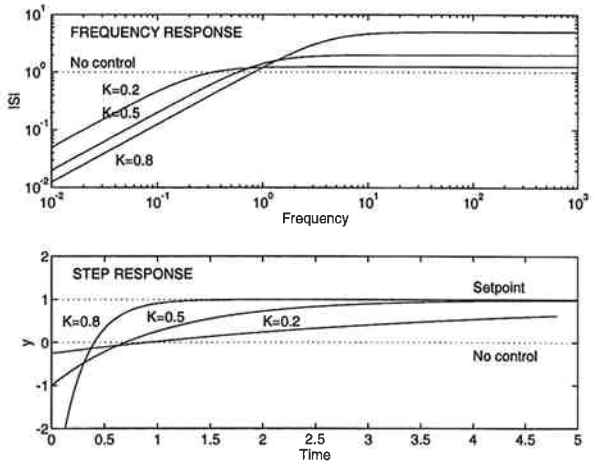

Figure 6: Plant with RHP-zero at $z = 1$ using negative feedback. $c(s) = K\frac{s+1}{s}$

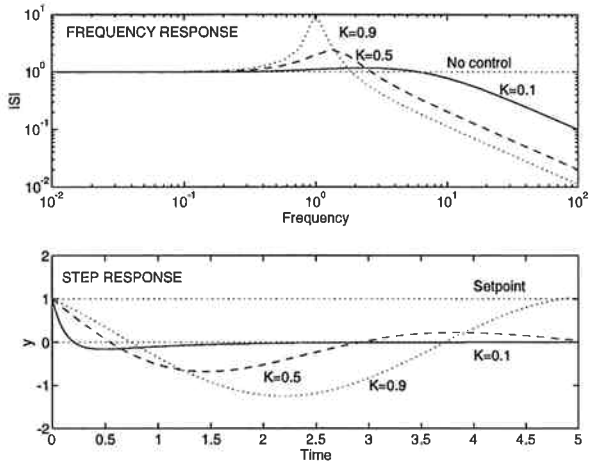

Figure 7: Plant with RHP-zero at $z = 1$ using positive feedback, $c(s) = -Ks$.

$$S = \frac{1}{1+gc} = \frac{s}{(1-K)s + Kz} \tag{24}$$

and we see that the system is stable for $0 < K < 1$. For the numerical calculations let $z = 1$. In Fig. 6 are shown the frequency response of the sensitivity function S and the response to a step setpoint change for various values of K. We note that with higher gains the controller is able to reduce the sensitivity at low frequencies, but only at the expense of a higher peak at some frequency above $z = 1$. Similarly, we see from the time response that an increased gain yields faster settling towards the steady-state, but a poorer initial response (in all cases the output goes in the wrong direction initially). The value $K = 0.5$ is seen to yield a reasonable trade-off between the two situations.

Positive feedback. If we do not care about the steady-state behavior then we may "reverse" the gain of the controller and instead achieve good control at frequencies higher than z. To this effect consider postive feedback using a derivative controller

$$c(s) = -\frac{K}{z}s \tag{25}$$

This yields $S = \frac{s+z}{(K/z)s^2+(1-K)s+z}$. The system is stable for $0 < K < 1$. In Fig. 7 the frequency and step responses for various values of K are shown. We note

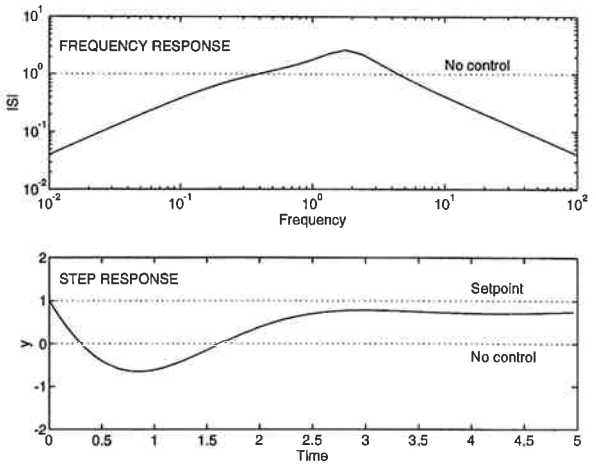

Figure 8: Plant with RHP-zero at $z = 1$ using combined negative and positive feedback, $c(s) = K(-s + \frac{s+1}{s}), K = 0.25$

that with higher gains the controller is able to reduce the sensitivity at high frequencies, but only at the expense of a peak in $|S(j\omega)|$ at intermediate frequencies around $z = 1$. Similarly, we see from the time response that an increased gain yields somewhat better setpoint tracking initially (in all cases the output "jumps" directly up to the desired value of $y = 1$ at $t = 0$), but at the expense of much larger oscillations (in all cases there is no tracking at steady-state). The value $K = 0.5$ yields a reasonable trade-off between the two situations.

III. Combined negative and positive feedback. One may finally consider combining the two control actions at low and high frequency, for example, with the controller

$$c(s) = K(-\frac{s}{z} + \frac{s+z}{s}) = \frac{K}{z}\frac{(-s+1.62z)(s+0.62z)}{s} \tag{26}$$

It is interesting to note that the controller contains a RHP-zero which gives the desired gain reversal. To have stability we must require $0 < K < 0.5$. In Fig. 8 the frequency and step responses for $K = 0.25$ are shown. We are able to reduce the sensitivity below one at all frequencies except around the frequency z corresponding to the RHP-zero. Similarly, we note from the step response that the error is quite large around time $t = 1/z = 1$.

In summary, this combined approach with both negative and positive feedback does not yield much (if any) improvement compared to negative feedback alone. In particular, the settling towards the steady-state is poor. In a practical situation one must, in order to improve the controllability, add another manipulated input to the process to take care of the control at either high or low frequencies. This is commonly done in cases when there is an input with a fast (direct) effect, which has no steady-state effect (i.e., a zero at the origin), and thus can only be used for transient control (Balchen and Mumme, 1988, p.47). In this case a second input must be used for the steady state control.