A PROCEDURE FOR CONTROLLABILITY ANALYSIS

Erik A. Wolff, Sigurd Skogestad*, Morten Hovd and Knut W. Mathisen

Chemical Engineering
University of Trondheim - NTH
N-7034 Trondheim, Norway

Abstract

In this paper we give an overview of some of the tools available for linear controllability analysis. We present a procedure which may be described by the following main steps;

1. Generate model
2. Scale the plant
3. Compute controllability measures
4. Analyze controllability

In the paper we raise issues in all of these categories. An FCC reactor is used as an example.

1 Introduction

A common procedure is to design a plant based on steady-state considerations, and then add on a control system at a later stage of the project. This may be acceptable if one at the early design stage can assess whether the plant will be easy to control or not.

Consider for example the benchmark example released by Tennessee Eastman (Downs, 1990). This is an integrated plant with reaction, separation and recycle. An analysis of the plant model reveals that it is unstable, has a high degree of interaction and a complicated dynamic behavior. One may proceed and try to design a control system for this plant. However, before embarking on this it would certainly be useful to know how well this process may be controlled with the best possible controller, that is, what is the controllability of the plant.

A plant with few "inherent control limitations" has good "achievable control performance" and is called "controllable" or "dynamic resilient". Since a plant's dynamic resilience can not be altered by change of the control algorithm, but only by design modifications, it follows that the term dynamic resilience provides a link between process design and process control.

Unfortunately, in standard state-space control the term "controllability" has the rather limited definition in terms of Kalman's state controllability, and this was the reason why Morari (1983) introduced the term "dynamic resilience". However, in engineering practice a plant is called "controllable" if it is possible to achieve the specified control objectives (Rosenbrock, 1970). We will use the term "controllability" in this more general sense, and use the term "state controllability" to avoid confusion.

*Address correspondence to this author. Fax: 47-7-594080, E-mail: skoge@kjemi.unit.no.

Rosenbrock (1970) gives a thorough discussion of the issues of state controllability and state observability, and also defines the term "functional controllability", which for SISO systems is equivalent to requiring $g(s) \not\equiv 0$ and for MIMO system $det(G(s)) \not\equiv 0$. He also introduces the important notion of right half plane (RHP) (transmission) zeros for multivariable systems. Morari (1982) and Stephanopoulos (1982) give good discussions on the issues of control structures and controllability of integrated plants. An initial approach towards quantitative analysis of controllability is given by Morari (1983) who makes use of the important notion of "perfect control", that is, the best achievable control performance. Perkins (1989) gives a good survey of the literature up to 1989.

2 Control objectives and limitations

To examine the controllability of a plant a mathematical model is needed. It is important to stress that for control considerations it is the initial part of the response, corresponding to the closed-loop time constant, that is of main interest. In particular it is important to get a good model of possible RHP-zeros, time delays, and of the interactions in the interesting frequency range. The steady-state behavior is usually of minor interest. One exception is the "sign" of the plant, i.e., $detG(0)$, which must be known. Otherwise it is does not really matter very much what would have happened after several hours if the plant was left uncontrolled. The reason why chemical engineers are often very preoccupied about the steady-state is that this is our natural way of thinking, and because steady-state models are often easily available.

In this paper we consider linear transfer function models on the form

$$y(s) = G(s)u(s) + G_d(s)d(s) \quad (1)$$

where u is the vector of manipulated inputs, d the vector of (physical) disturbances, and y is the vector of outputs (controlled variables). The objective is to keep the error $e = y - r$ small, where r is the vector of reference signals (setpoints). $G(s)$ and $G_d(s)$ are transfer matrices, that need not be square.

The main objective of the control system is to keep the outputs y close to their setpoints r and to reject disturbances (often called "load changes"). The ideal controller will accomplish this by inverting the process such that the manipulated input becomes $u = G^{-1}r - G^{-1}G_d d$. In practice, something close to this may be achieved with with feedback control. With $u = C(s)(r - y)$, the response of the system is

$$y = Tr + SG_d d \quad (2)$$

$$u = G^{-1}Tr - G^{-1}TG_d d \quad (3)$$

Here the sensitivity is $S = (I + GC)^{-1}$ and the complementary sensitivity is $T = GC(I + GC)^{-1}$. It follows from this that at low frequencies where feedback is effective ($\omega < \omega_B$), $S \approx 0$ and $T \approx I$ and the controller corresponds to inversion of the plant. Consequently ideal control (inversion) requires <u>fast</u> feedback (high bandwidth).

On the other hand, inherent limitations of the system prevent fast control. These limitations may include a close to singular plant such that the necessary inputs signals become large, non-minimum phase characteristics such as time-delay and RHP-zeros, constraints on the input variables and model uncertainty. These limitations make it desirable to have a <u>low</u> bandwidth. *If these requirements for high and low bandwidth are in conflict then controllability is poor.*

3 Scaling of variables

The relative gain array (RGA) has the advantage of being scaling independent, but for other controllability measures it is crucial that the variables are scaled properly. In general, the variables should be scaled to be within the interval -1 to 1, that is, their desired or expected magnitudes should be normalized to be less than 1 at each frequency. Recommended scalings:

- Inputs (u): Normalize u_j with respect to its allowed range.
- Outputs (y): Normalize e_i with respect to its allowed range.
- Disturbances (d): Normalize d_k with respect to its expected range.

To achieve this we scale the transfer matrices G and G_d. For example, we assume that at each frequency $g_d(j\omega)$ (or the columns in $G_d(j\omega)$) is scaled such that the worst (largest) disturbance corresponds to $|d(j\omega)| = 1$.

Comment: In this paper we scale directly the *transfer matrices* G and G_d and assume that the expected or allowed magnitude of the signals d, u, e and r does not vary with frequency. If their magnitudes vary then we should rather scale the *signals* using frequency-dependent weights. This signal approach is also more general, for example, if the setpoints do not have same size as the allowed errors (as we implicitly have to assume).

4 Tools for controllability analysis

In this section we will briefly discuss a number of methods for evaluating controllability.

All measures are controller independent. We first present general measures, and special measures for decentralized control are given towards the end.

4.1 Functional and state controllability

Probably the first thing that should be checked is that the plant is functional controllable. Essentially, a plant is not functional controllable if the rank of $G(s)$ is for all s less than the number of outputs we want to control. For square plants the requirement is that we should not have $\det G(s) \not\equiv 0$ (Rosenbrock, 1970). A typical example when a plant is not functional controllable is when an entire row of $G(s)$ is zero ("there is no downstream path to a particular output"). Another case is, for example, in a heat exchanger network where we have two control inputs (bypasses) that can only effect the two control outputs (temperatures) by transferring heat through the same stream (the downstream paths coincide) (see Mathisen and Skogestad, 1992).

For unstable plants it should be checked that the unstable states are state controllable and state observable, but otherwise this issue is not of particular interest, as states that we really care about should be included in the output vector y.

4.2 RHP-zeros and time delays

A zero is defined as values for s for which $G(s)$ looses rank, and for square plants that may computed as the solutions to $detG(s) = 0$ (a more careful definition involving the system matrix may be needed in the case the model has internal pole-zero cancellations). A right half plane (RHP) transmission zero of $G(s)$ limits the achievable bandwidth of the plant. This holds regardless of the type of controller used (Holt and Morari, 1985). Plants with RHP transmission zeros within the desired bandwidth should be avoided. If we use a multivariable controller then RHP-zeros in the elements do not imply any particular problem. However, if decentralized controllers are used, then we generally avoid pairing on elements with "significant" RHP-zeros (RHP-zeros close to the origin), because otherwise this loop may go unstable if left by itself (with the other loops open).

the If an RHP transmission zero cannot be avoided, it should preferably be at as high a frequency as possible, and lie in a plant direction (Morari and Zafirou, 1987) such that it affects an output where the performance requirements (as required, e.g., for disturbance rejection) are lax.

Time delays have essentially the same effect as RHP-zeros with $\omega_B < 1/\theta$ where θ is the time delay.

4.3 RHP-poles

Poles of $G(s)$ in the right half plane also put limitations on the control system through stability considerations. The bandwidth of the closed-loop system must be above the frequency of the RHP-pole to ensure a stable system. Freudenberg and Looze have derived some interesting relationships which quantify the effect of RHP-poles and RHP-zeros. These are summarized in Hovd and Skogestad (1992, this Symposium). If there are RHP zeros and RHP-poles in the same direction, it is important that the RHP-pole at p is located at a higher frequency than the RHP-zero at z, i.e., $p > z$.

4.4 Singular value analysis

The singular value decomposition of any matrix G is $G = U\Sigma V^H$ with the matrix Σ having the singular values σ_i on the main diagonal. There will be $rank(G)$ singular values.

The singular values are directly related to the vector 2-norm. Specifically, we have for $y = Ax$ that

$$\underline{\sigma}(A) \leq \frac{\|y\|_2}{\|x\|_2} \leq \bar{\sigma}(A) \qquad (4)$$

where $\bar{\sigma}(A)$ is the maximum singular value, and $\underline{\sigma}(A)$ the minimum singular value of A. We may choose the directions for x such that either the lower or upper bounds in (4) is tight.

An SVD on G and G_d is useful for examining which manipulated input combinations have the largest effect and which disturbances give the largest output variations. For example, applied to distillation the singular value analysis shows that much less control action is needed to move top and bottom compositions in the *same* direction (i.e., mole fraction of light component x increases or decreases in distillate and bottoms simultaneously) as the opposite.

Minimum singular value and input magnitudes. $\underline{\sigma}(G)$ was introduced as as a controllability index by Morari (1983). From (3) and (4) we get that the input needed for tracking at a sinusoidally varying reference signal $r(jw)$ is given by (Perkins, 1989)

$$\frac{1}{\bar{\sigma}(G)} \leq \frac{\|u\|_2}{\|r\|_2} \leq \frac{1}{\underline{\sigma}(G)} \qquad (5)$$

Here we have used the fact that $\underline{\sigma}(G^{-1}) = 1/\bar{\sigma}(G)$ and $\bar{\sigma}(G^{-1}) = 1/\underline{\sigma}(G)$. Since r may have any direction, we see that a small value of $\underline{\sigma}(G)$ implies that large input magnitudes may be needed, and such plants are undesirable (Morari, 1983). If the variables have been scaled in accordance with the recommendations above then a requirement for avoiding input constraints for unitary setpoint changes is approximately (since we are looking at the 2-norm and not infinity-norm of u) that $\underline{\sigma}(G(j\omega)) > 1, \forall\omega$. This measure is useful also at steady-state. For SISO plants it simply corresponds to requiring $|g| > 1$, and otherwise preferring designs where the steady-state gain is as large as possible. For decentralized control it is desirable to pair on elements with $|g_{ij}| > 1$.

For perfect control of square plants we need $u = -G^{-1}G_d d$. If we have several disturbances that all are less than 1 in magnitude (i.e., $\|d\|_\infty \leq 1$), then the input magnitude (measured in terms of the infinity-norm) needed for perfect rejection of the worst disturbance is given by

$$\|u\|_\infty = \|G^{-1}G_d\|_\infty \qquad (6)$$

which is equal to the largest row-sum of the matrix $G^{-1}G_d$. A frequency dependent plot of the elements in $G^{-1}G_d$ give useful information about the possibility for reaching input constraints, and which disturbances that cause problems.

4.5 Condition number

The ratio between the largest singular value ($\overline{\sigma}$) and the smallest nonzero singular value ($\underline{\sigma}$), is often denoted the condition number, $\gamma(G) = \frac{\bar{\sigma}(G)}{\underline{\sigma}(G)}$. Plants with a large condition number are called ill-conditioned, and require widely different input magnitudes depending on the direction of the desired output. Note that $\gamma(G)$ depends on the scaling of the inputs and outputs, and it is important that these are scaled properly. There is a close relationship between the optimally scaled condition number, $\gamma^*(G)$, (minimize $\gamma(G)$ with respect to input and output scaling) and the magnitude of the RGA-elements (e.g., Skogestad and Morari, 1987).

If γ is large, then the plant is sensitive to unstructured (uncorrelated) input uncertainty (Skogestad et al., 1988). However, unstructured uncertainty is often unrealistic.

4.6 Relative gain array (RGA)

The most widespread controllability measure is probably the RGA which was introduced by Bristol (1966). Skogestad and Hovd (1990) give a thorough survey of the frequency-dependent RGA and its properties. For a square plant $G(s)$ the relative gain is defined as the ratio of the "open-loop" and "closed-loop" gains between input j and output i. It is defined at each frequency as

$$\lambda_{ij}(s) = \frac{(\partial y_i/\partial u_j)_{u_{l\neq j}}}{(\partial y_i/\partial u_j)_{y_{l\neq i}}} = g_{ij}(s)[G^{-1}(s)]_{ji} \qquad (7)$$

and a RGA-matrix is computed from

$$\Lambda(j\omega) = G(j\omega) \times (G^{-1}(j\omega))^T \qquad (8)$$

where $\times$ denotes element-by-element multiplication. It is worth noting that the RGA is independent of scaling, and must only be rearranged (*not* recomputed) when considering different control pairings.

Plants with large RGA-values are ill-conditioned ($\gamma(G)$ is large) irrespective of input and output scaling. Triangular plants yield $\Lambda = I$, and plants where Λ is different from I are called interactive (G has significant offdiagonal elements). It is established that plants with large RGA-values, in particular at high frequencies, are fundamentally difficult to control. In particular, it is known that one should never use decouplers in such cases because of a strong sensitivity to (structured) input uncertainty in each channel, i.e., one should never use a controller with large RGA-values (Skogestad and Morari, 1987).

The relative gains λ_{ij} give a direct measure of the sensitivity of the plant to independent element-by-element uncertainty (which actually occurs relatively rarely): $G(j\omega)$ becomes singular (and the plant impossible to control at this frequency) if any element $g_{ij}(j\omega)$ changes by $-1/\lambda_{ij}(j\omega)$, thus large RGA-elements imply that $G(j\omega)$ is close to singularity.

For interactive plants which do not have large RGA-elements, a decoupler may be useful. In particular, this applies to the case where the RGA-elements vary in magnitude with frequency (e.g., between 0 and 2), and it may be difficult to find a good pairing for decentralized control (see below). A steady-state decoupler may be used if the directions do not change too much with frequency.

4.7 Disturbance sensitivity

We will only give a short presentation of measures for evaluating disturbance sensitivity, as they are treated in more detail in the paper "Controllability measures for disturbance rejection" by Skogestad and Wolff also at this meeting.

Open-loop disturbance sensitivity. For one disturbance i the open-loop disturbance sensitivity is directly given by the i'th element of the vector g_d, that is $\left(\frac{\partial y_i}{\partial d}\right)_{u_j} = g_{di}$. If appropriately scaling has been applied and any of the elements in G_d are larger than 1 then control is needed to get acceptable performance.

Consider a SISO plant. Typically, $|g_d|$ is larger than 1 at low frequencies and drops to zero at high frequencies. The frequency ω_d where $|g_d(j\omega)|$ crosses 1 is then of particular interest, since it yields the minimum bandwidth requirement for feedback control, i.e., $\omega_B > \omega_d$. ω_d is thus a measure of the controllability that one needs to impose on the system. If the plant has a RHP-zero at $s = z$ then we must require $z > \omega_d$ (It is stressed that all these relationships involving RHP-zeros and RHP-poles are approximate).

For MIMO plants we have a bandwidth region ranging from ω_B (worst direction, $\underline{\sigma}(GC)$) to ω'_B (best direction, $\bar{\sigma}(GC)$). For a single disturbance consider the frequency ω_d where $\|g_d\|_2$ crosses 1. Then we *must* require that $\omega'_B > \omega_d$ and we *may* have to require $\omega_b > \omega_d$ (depending of the direction of the disturbance).

Disturbance condition number. To study specifically the *direction* of a disturbance, Skogestad and Morari (1987) introduced the disturbance condition number of the matrix A

$$\gamma_d(A) = \frac{\|A^{-1}g_d\|_2}{\|g_d\|_2}\bar{\sigma}(A) \tag{9}$$

where A may be G or $L = GC$. The disturbance condition number of G, $\gamma_d(G)$, tells us for a particular disturbance how much larger the input magnitude needs to be to reject a unit disturbance, compared to if the disturbance was in the best possible direction of the plant (corresponding to the direction of $\bar{\sigma}(G)$).

4.8 Partial disturbance sensitivity

The following measure is useful when considering if one may let one of the outputs be uncontrolled, for example, if the original control problem is difficult. Recall, that the open-loop disturbance sensitivity for an output i and disturbance k is $\left(\frac{\partial y_i}{\partial d_k}\right)_{u_j} = g_{dik}$. The corresponding disturbance sensitivity with all the *other* outputs $l \neq i$ perfectly controlled can be expressed as

$$\left(\frac{\partial y_i}{\partial d_k}\right)_{u_j, y_{j\neq i}} = [G^{-1}G_d]_{ik}/[G^{-1}]_{ji} \tag{10}$$

We denote this measure the partial disturbance gain (PDG). The term partial is used since the system is only partially controlled. For simultaneous disturbances we should evaluate the worst overall effect of them by taking the 1-norm (sum of element magnitudes). This gives rise to a combined PDG-matrix, denoted G_{PDG} with elements

$$[G_{PDG}]_{ij} = \sum_k |[G^{-1}G_d]_{ik}|/|[G^{-1}]_{ji}| \tag{11}$$

It it desirable to find an "uncontrolled pairing" $u_j - y_i$ for which the G_{PDG}-element is less than 1. Note that also the steady-state values of G_{PDG} are important.

For the case $j = i$ (that is, we have paired up the uncontrolled output with the output we want in manual), the PDG is equal to the ratio between the CLDG (see below) and the corresponding RGA-element:

$$\left(\frac{\partial y_i}{\partial d_k}\right)_{u_i, y_{l\neq i}} = \delta_{ik}/\lambda_{ii} \tag{12}$$

4.9 Relative order and phase lag

The relative order is sometimes used a controllability measure (e.g., Daoutidis and Kravaris, 1992). The relative order may be defined also for nonlinear plants, and for linear plants it corresponds to the high-frequency rolloff, that is, the pole excess of the transfer function. Of course, we want the inputs to directly affect the outputs, and the relative order should be small. However, the usefulness of the concept of relative order is rather limited since it depends on the modeling detail. In fact, a more useful measure to consider is the phase lag of the model at the bandwidth frequencies, for decentralized control we want to pair on variables where the phase lag is as small as possible, and it should be less than -180^o (see Balchen, 1988).

4.10 Special measures for decentralized control.

Pairing and use of RGA. For decentralized control of stable plants one should always try to pair on positive steady-state RGA-elements. Otherwise one will with integral control get instability of either 1) the overall system, 2) the individual loop, or 3) the remaining system when the loop in question is removed. Hovd and Skogestad (1992) have extended the use of the steady-state RGA to unstable plants.

However, also for decentralized control the most important frequency regions is around the closed-loop bandwidth, and we usually prefer pairings corresponding to relative gains close to 1 (with the other elements close to zero) in the this frequency region.

PRGA. One inadequacy of the RGA (eg., McAvoy, 1983, p. 166) is that it may indicate that interactions is no problem, but significant one-way coupling may exist. This follows since the RGA is equal to the identity matrix (I) when $G(s)$ is triangular. To overcome this problem Hovd and Skogestad (1992) introduced the performance relative gain array (PRGA). The PRGA-matrix is defined as

$$\Gamma(s) = G_{diag}(s)G(s)^{-1} \tag{13}$$

where $G_{diag}(s)$ is the matrix consisting of only the diagonal elements of $G(s)$, i.e., $G_{diag} = diag\{g_{ii}\}$. Note that the diagonal elements of RGA and PRGA are identical, but otherwise PRGA does not have all the nice algebraic properties of the RGA. For example, PRGA is independent of *input* scaling, but it depends on output scaling. This is reasonable since performance is defined in terms of the magnitude of the outputs. Note that PRGA $= G_s^{-1}$ where G_s is obtained by input scaling of G such that all the diagonal elements are 1 (at all frequencies).

As is clear from 14 below, we prefer the PRGA-elements to be small at low frequencies, but at high frequency we want the PRGA-matrix to be triangular (i.e., $G(j\omega)$ triangular) with the diagonal elements (corresponding to the chosen pairings) close to 1.

CLDG and RDG. Consider decentralized control. Then at low frequencies the closed-loop response for loop i when also *all* the other loops are closed is (Skogestad and Hovd, 1990):

$$e_i \approx -\frac{\gamma_{ij}}{L_i}r_i + \frac{\delta_{dik}}{L_i}d_k; \quad \omega \leq \omega_B \tag{14}$$