MULTIVARIABLE FEEDBACK CONTROL
Analysis and design

Sigurd Skogestad
Norwegian University of Science and Technology

Ian Postlethwaite
University of Leicester

JOHN WILEY & SONS
Chichester · New York · Brisbane · Toronto · Singapore

BORGHEIM, an engineer:

> *Herregud, en kan da ikke gjøre noe bedre enn leke i denne velsignede verden. Jeg synes hele livet er som en lek, jeg!*
>
> Good heavens, one can't do anything better than play in this blessed world. The whole of life seems like playing to me!

Act one, LITTLE EYOLF, Henrik Ibsen.

CONTENTS

PREFACE

This is a book on practical feedback control and not on system theory generally. Feedback is used in control systems to change the dynamics of the system (usually to make the response stable and sufficiently fast), and to reduce the sensitivity of the system to signal uncertainty (disturbances) and model uncertainty. Important topics covered in the book, include

- classical frequency-domain methods
- analysis of directions in multivariable systems using the singular value decomposition
- input-output controllability (inherent control limitations in the plant)
- model uncertainty and robustness
- performance requirements
- methods for controller design and model reduction
- control structure selection and decentralized control

The treatment is for linear systems. The theory is then much simpler and more well developed, and a large amount of practical experience tells us that in many cases linear controllers designed using linear methods provide satisfactory performance when applied to real nonlinear plants.

We have attempted to keep the mathematics at a reasonably simple level, and we emphasize results that enhance *insight* and *intuition*. The design methods currently available for linear systems are well developed, and with associated software it is relatively straightforward to design controllers for most multivariable plants. However, without insight and intuition it is difficult to judge a solution, and to know how to proceed (e.g. how to change weights) in order to improve a design.

The book is appropriate for use as a text for an introductory graduate course in multivariable control or for an advanced undergraduate course. We also think it will be useful for engineers who want to understand multivariable control, its limitations, and how it can be applied in practice. There are numerous worked examples, exercises and case studies which make frequent use of MATLAB™ [1].

[1] MATLAB is a registered trademark of The MathWorks, Inc.

The prerequisites for reading this book are an introductory course in classical single-input single-output (SISO) control and some elementary knowledge of matrices and linear algebra. Parts of the book can be studied alone, and provide an appropriate background for a number of linear control courses at both undergraduate and graduate levels: classical loop-shaping control, an introduction to multivariable control, advanced multivariable control, robust control, controller design, control structure design and controllability analysis.

The book is partly based on a graduate multivariable control course given by the first author in the Cybernetics Department at the Norwegian University of Science and Technology in Trondheim. About 10 students from Electrical, Chemical and Mechanical Engineering have taken the course each year since 1989. The course has usually consisted of 3 lectures a week for 12 weeks. In addition to regular assignments, the students have been required to complete a 50 hour design project using MATLAB. In Appendix B, a project outline is given together with a sample exam.

Examples and internet

Most of the numerical examples have been solved using MATLAB. Some sample files are included in the text to illustrate the steps involved. Most of these files use the μ-toolbox, and some the Robust Control toolbox, but in most cases the problems could have been solved easily using other software packages.

The following are available over the internet from Trondheim[2] and Leicester:

- MATLAB files for examples and figures
- Solutions to selected exercises
- Linear state-space models for plants used in the case studies
- Corrections, comments to chapters, extra exercises

This information can be accessed from the authors' home pages:

- `http://www.kjemi.unit.no/~skoge`
- `http://www.engg.le.ac.uk/staff/Ian.Postlethwaite`

Comments and questions

Please send questions, errors and any comments you may have to the authors. Their email addresses are:

- `Sigurd.Skogestad@kjemi.unit.no`
- `ixp@le.ac.uk`

[2] The internet site name in Trondheim will change from `unit` to `ntnu` during 1996.

Acknowledgements

The contents of the book are strongly influenced by the ideas and courses of Professors John Doyle and Manfred Morari from the first author's time as a graduate student at Caltech during the period 1983-1986, and by the formative years, 1975-1981, the second author spent at Cambridge University with Professor Alistair MacFarlane. We thank the organizers of the 1993 European Control Conference for inviting us to present a short course on applied $\mathcal{H}_\infty$ control, which was the starting point for our collaboration. The final manuscript began to take shape in 1994-95 during a stay the authors had at the University of California at Berkeley – thanks to Andy Packard, Kameshwar Poolla, Masayoshi Tomizuka and others at the BCCI-lab, and to the stimulating coffee at *Brewed Awakening*.

We are grateful for the numerous technical and editorial contributions of Yi Cao, Kjetil Havre, Ghassan Murad and Ying Zhao. The computations for Example 4.5 were performed by Roy S. Smith who shared an office with the authors at Berkeley. Helpful comments and corrections were provided by Richard Braatz, Atle C. Christiansen, Wankyun Chung, Bjørn Glemmestad, John Morten Godhavn, Finn Are Michelsen and Per Johan Nicklasson. A number of people have assisted in editing and typing various versions of the manuscript, including Zi-Qin Wang, Yongjiang Yu, Greg Becker, Fen Wu, Regina Raag and Anneli Laur. We also acknowledge the contributions from our graduate students, notably Neale Foster, Morten Hovd, Elling W. Jacobsen, Petter Lundström, John Morud, Raza Samar and Erik A. Wolff.

The aero-engine model (Chapters 11 and 12) and the helicopter model (Chapter 12) are provided with the kind permission of Rolls-Royce Military Aero Engines Ltd, and the UK Ministry of Defence, DRA Bedford, respectively.

Finally, thanks to colleagues and former colleagues at Trondheim and Caltech from the first author, and at Leicester, Oxford and Cambridge from the second author.

We have made use of material from several books. In particular, we recommend Zhou, Doyle and Glover (1996) as an excellent reference on system theory and $\mathcal{H}_\infty$ control. Of the others we would like to acknowledge, and recommend for further reading, the following: Rosenbrock (1970), Rosenbrock (1974), Kwakernaak and Sivan (1972), Kailath (1980), Chen (1984), Francis (1987), Anderson and Moore (1989), Maciejowski (1989), Morari and Zafiriou (1989), Boyd and Barratt (1991), Doyle et al. (1992), Green and Limebeer (1995), and the MATLAB toolbox manuals of Balas et al. (1993) and Chiang and Safonov (1992).

1

INTRODUCTION

In this chapter, we begin with a brief outline of the design process for control systems. We then discuss linear models and transfer functions which are the basic building blocks for the analysis and design techniques presented in this book. The scaling of variables is critical in applications and so we provide a simple procedure for this. An example is given to show how to derive a linear model in terms of deviation variables for a practical application. Finally, we summarize the most important notation used in the book.

1.1 The process of control system design

The process of designing a control system usually makes many demands of the engineer or engineering team. These demands often emerge in a step by step design procedure as follows:

1. Study the system (plant) to be controlled and obtain initial information about the control objectives.
2. Model the system and simplify the model, if necessary.
3. Analyze the resulting model; determine its properties.
4. Decide which variables are to be controlled (controlled outputs).
5. Decide on the measurements and manipulated variables: what sensors and actuators will be used and where will they be placed?
6. Select the control configuration.
7. Decide on the type of controller to be used.
8. Decide on performance specifications, based on the overall control objectives.
9. Design a controller.
10. Analyze the resulting controlled system to see if the specifications are satisfied; and if they are not satisfied modify the specifications or the type of controller.
11. Simulate the resulting controlled system, either on a computer or a pilot plant.
12. Repeat from step 2, if necessary.
13. Choose hardware and software and implement the controller.
14. Test and validate the control system, and tune the controller on-line, if necessary.

Control courses and text books usually focus on steps 9 and 10 in the above procedure; that is, on methods for controller design and control system analysis. Interestingly, many real control systems are designed without any consideration of these two steps. For example, even for complex systems with many inputs and outputs, it may be possible to design workable control systems, often based on a hierarchy of cascaded control loops, using only on-line tuning (involving steps 1, 4 5, 6, 7, 13 and 14). However, in this case a suitable control structure may not be known at the outset, and there is a need for systematic tools and insights to assist the designer with steps 4, 5 and 6. A special feature of this book is the provision of tools for *input-output controllability analysis* (step 3) and for *control structure design* (steps 4, 5, 6 and 7).

Input-output controllability is the ability to achieve acceptable control performance. It is affected by the location of sensors and actuators, but otherwise it cannot be changed by the control engineer. Simply stated, "even the best control system cannot make a Ferrari out of a Volkswagen". Therefore, the process of control system design should in some cases also include a step 0, involving the design of the process equipment itself. The idea of looking at process equipment design and control system design as an integrated whole is not new, as is clear from the following quote taken from a paper by Ziegler and Nichols (1943):

> In the application of automatic controllers, it is important to realize that controller and process form a unit; credit or discredit for results obtained are attributable to one as much as the other. A poor controller is often able to perform acceptably on a process which is easily controlled. 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 older models, but on these processes, there is a definite end point which can be approached by instrumentation and it falls short of perfection.

Ziegler and Nichols then proceed to observe that there is a factor in equipment design that is neglected, and state that

> …the missing characteristic can be called the "controllability", the ability of the process to achieve and maintain the desired equilibrium value.

To derive simple tools with which to quantify the inherent input-output controllability of a plant is the goal of Chapters 5 and 6.

1.2 The control problem

The objective of a control system is to make the output y behave in a desired way by manipulating the plant input u. The *regulator problem* is to manipulate u to counteract the effect of a disturbance d. The *servo problem* is to manipulate u to keep the output close to a given reference input r. Thus, in both cases we want the *control error* $e = y - r$ to be small. The algorithm for adjusting u based on the available information is the controller K. To arrive at a good design for K we need *a priori* information about the expected disturbances and reference inputs, and of the plant model (G) and disturbance model (G_d). In this book we make use of linear models of the form

$$y = Gu + G_d d \tag{1.1}$$

A major source of difficulty is that the models (G, G_d) may be inaccurate or may change with time. In particular, inaccuracy in G may cause problems because the plant will be part of a feedback loop. To deal with such a problem we will make use of the concept of model uncertainty. For example, instead of a single model G we may study the behaviour of a class of models, $G_p = G + E$, where the "uncertainty" or "perturbation" E is bounded, but otherwise unknown. In most cases weighting functions, $w(s)$, are used to express $E = w\Delta$ in terms of normalized perturbations, Δ, where the magnitude (norm) of Δ is less than or equal to 1. The following terms are useful:

Nominal stability (NS). The system is stable with no model uncertainty.

Nominal Performance (NP). The system satisfies the performance specifications with no model uncertainty.

Robust stability (RS). The system is stable for all perturbed plants about the nominal model up to the worst-case model uncertainty.

Robust performance (RP). The system satisfies the performance specifications for all perturbed plants about the nominal model up to the worst-case model uncertainty.

1.3 Transfer functions

The book makes extensive use of transfer functions, $G(s)$, and of the frequency domain, which are very useful in applications for the following reasons:

- Invaluable insights are obtained from simple frequency-dependent plots.
- Important concepts for feedback such as bandwidth and peaks of closed-loop transfer functions may be defined.
- $G(j\omega)$ gives the response to a sinusoidal input of frequency ω.
- A series interconnection of systems corresponds in the frequency domain to multiplication of the individual system transfer functions, whereas in the time domain the evaluation of complicated convolution integrals is required.
- Poles and zeros appear explicitly in factorized scalar transfer functions.

- Uncertainty is more easily handled in the frequency domain. This is related to the fact that two systems can be described as close (i.e. have similar behaviour) if their frequency responses are similar. On the other hand, a small change in a parameter in a state-space description can result in an entirely different system response.

We consider linear, time-invariant systems whose input-output responses are governed by linear ordinary differential equations with constant coefficients. An example of such a system is

$$\begin{aligned} \dot{x}_1(t) &= -a_1 x_1(t) + x_2(t) + \beta_1 u(t) \\ \dot{x}_2(t) &= -a_0 x_1(t) + \beta_0 u(t) \\ y(t) &= x_1(t) \end{aligned} \tag{1.2}$$

where $\dot{x}(t) \equiv dx/dt$. Here $u(t)$ represents the input signal, $x_1(t)$ and $x_2(t)$ the states, and $y(t)$ the output signal. The system is time-invariant since the coefficients a_1, a_0, β_1 and β_0 are independent of time. If we apply the Laplace transform to (1.2) we obtain

$$\begin{aligned} s\bar{x}_1(s) - x_1(t=0) &= -a_1\bar{x}_1(s) + \bar{x}_2(s) + \beta_1\bar{u}(s) \\ s\bar{x}_2(s) - x_2(t=0) &= -a_0\bar{x}_1(s) + \beta_0\bar{u}(s) \\ \bar{y}(s) &= \bar{x}_1(s) \end{aligned} \tag{1.3}$$

where $\bar{y}(s)$ denotes the Laplace transform of $y(t)$, and so on. To simplify our presentation we will make the usual abuse of notation and replace $\bar{y}(s)$ by $y(s)$, etc.. In addition, we will omit the independent variables s and t when the meaning is clear.

If $u(t), x_1(t), x_2(t)$ and $y(t)$ represent deviation variables away from a nominal operating point or trajectory, then we can assume $x_1(t=0) = x_2(t=0) = 0$. The elimination of $x_1(s)$ and $x_2(s)$ from (1.3) then yields the transfer function

$$\frac{y(s)}{u(s)} = G(s) = \frac{\beta_1 s + \beta_0}{s^2 + a_1 s + a_0} \tag{1.4}$$

Importantly, for linear systems, the transfer function is independent of the input signal (forcing function). Notice that the transfer function in (1.4) may also represent the following system

$$\ddot{y}(t) + a_1\dot{y}(t) + a_0 y(t) = \beta_1\dot{u}(t) + \beta_0 u(t) \tag{1.5}$$

with input $u(t)$ and output $y(t)$.

Transfer functions, such as $G(s)$ in (1.4), will be used throughout the book to model systems and their components. More generally, we consider rational transfer functions of the form

$$G(s) = \frac{\beta_{n_z} s^{n_z} + \cdots + \beta_1 s + \beta_0}{s^n + a_{n-1}s^{n-1} + \cdots + a_1 s + a_0} \tag{1.6}$$

For multivariable systems, $G(s)$ is a matrix of transfer functions. In (1.6) n is the order of the denominator (or pole polynomial) and is also called the *order of the system*, and n_z is the order of the numerator (or zero polynomial). Then $n - n_z$ is referred to as the pole excess or *relative order*.

Definition 1.1

- *A system* $G(s)$ *is* strictly proper *if* $G(s) \to 0$ *as* $s \to \infty$.
- *A system* $G(s)$ *is* semi-proper *or* bi-proper *if* $G(s) \to D \neq 0$ *as* $s \to \infty$.
- *A system* $G(s)$ *which is strictly proper or semi-proper is* proper.
- *A system* $G(s)$ *is* improper *if* $G(s) \to \infty$ *as* $s \to \infty$.

For a proper system, with $n \geq n_z$, we may realize (1.6) by a state-space description, $\dot{x} = Ax + Bu,\ y = Cx + Du$, similar to (1.2). The transfer function may then be written as

$$G(s) = C(sI - A)^{-1}B + D \tag{1.7}$$

Remark. All practical systems will have zero gain at a sufficiently high frequency, and are therefore strictly proper. It is often convenient, however, to model high frequency effects by a non-zero D-term, and hence semi-proper models are frequently used. Furthermore, certain derived transfer functions, such as $S = (I + GK)^{-1}$, are semi-proper.

Usually we use $G(s)$ to represent the effect of the inputs u on the outputs y, whereas $G_d(s)$ represents the effect on y of the disturbances d. We then have the following linear process model in terms of deviation variables

$$y(s) = G(s)u(s) + G_d(s)d(s) \tag{1.8}$$

We have made use of the superposition principle for linear systems, which implies that a change in a dependent variable (here y) can simply be found by adding together the separate effects resulting from changes in the independent variables (here u and d) considered one at a time.

All the signals $u(s)$, $d(s)$ and $y(s)$ are deviation variables. This is sometimes shown explicitly, for example, by use of the notation $\delta u(s)$, but since we always use deviation variables when we consider Laplace transforms, the δ is normally omitted.

1.4 Scaling

Scaling is very important in practical applications as it makes model analysis and controller design (weight selection) much simpler. It requires the engineer to make a judgement at the start of the design process about the required performance of the system. To do this, decisions are made on the expected magnitudes of disturbances and reference changes, on the allowed magnitude of each input signal, and on the allowed deviation of each output.

Let the unscaled (or originally scaled) linear model of the process in deviation variables be

$$\widehat{y} = \widehat{G}\widehat{u} + \widehat{G}_d\widehat{d}; \quad \widehat{e} = \widehat{y} - \widehat{r} \tag{1.9}$$

where a hat ($\widehat{\ }$) is used to show that the variables are in their unscaled units. A useful approach for scaling is to make the variables less than one in magnitude. This is done by *dividing each variable by its maximum expected or allowed change.* For disturbances and manipulated inputs, we use the scaled variables

$$d = \widehat{d}/\widehat{d}_{\max}, \quad u = \widehat{u}/\widehat{u}_{\max} \tag{1.10}$$

where:

- $\widehat{d}_{\max}$ — largest expected change in disturbance
- $\widehat{u}_{\max}$ — largest allowed input change

The maximum deviation from a nominal value should be chosen by thinking of the maximum value one can expect, or allow, as a function of time.

The variables $\widehat{y}$, $\widehat{e}$ and $\widehat{r}$ are in the same units, so the same scaling factor should be applied to each. Two alternatives are possible:

- $\widehat{e}_{\max}$ — largest allowed control error
- $\widehat{r}_{\max}$ — largest expected change in reference value

Since a major objective of control is to minimize the control error $\widehat{e}$, we here usually choose to scale with respect to the maximum control error:

$$y = \widehat{y}/\widehat{e}_{\max}, \quad r = \widehat{r}/\widehat{e}_{\max}, \quad e = \widehat{e}/\widehat{e}_{\max} \tag{1.11}$$

To formalize the scaling procedure, introduce the scaling factors

$$D_e = \widehat{e}_{\max},\ D_u = \widehat{u}_{\max},\ D_d = \widehat{d}_{\max},\ D_r = \widehat{r}_{\max} \tag{1.12}$$

For MIMO systems each variable in the vectors $\widehat{d}$, $\widehat{r}$, $\widehat{u}$ and $\widehat{e}$ may have a different maximum value, in which case D_e, D_u, D_d and D_r become diagonal scaling *matrices*. This ensures, for example, that all errors (outputs) are of about equal importance in terms of their magnitude.

The corresponding scaled variables to use for control purposes are then

$$d = D_d^{-1}\widehat{d},\ u = D_u^{-1}\widehat{u},\ y = D_e^{-1}\widehat{y},\ e = D_e^{-1}\widehat{e},\ r = D_e^{-1}\widehat{r} \tag{1.13}$$

On substituting (1.13) into (1.9) we get

$$D_e y = \widehat{G}D_u u + \widehat{G}_d D_d d; \quad D_e e = D_e y - D_e r$$

and introducing the scaled transfer functions

$$G = D_e^{-1}\widehat{G}D_u, \quad G_d = D_e^{-1}\widehat{G}_d D_d \tag{1.14}$$

then yields the following model in terms of scaled variables

$$y = Gu + G_d d; \quad e = y - r \tag{1.15}$$

Here u and d should be less than 1 in magnitude, and it is useful in some cases to introduce a scaled reference $\widetilde{r}$, which is less than 1 in magnitude. This is done by dividing the reference by the maximum expected reference change

$$\widetilde{r} = \widehat{r}/\widehat{r}_{\max} = D_r^{-1}\widehat{r} \tag{1.16}$$

We then have that

$$r = R\widetilde{r} \quad \text{where} \quad R \triangleq D_e^{-1}D_r = \widehat{r}_{\max}/\widehat{e}_{\max} \tag{1.17}$$

Here R is the largest expected change in reference relative to the allowed control

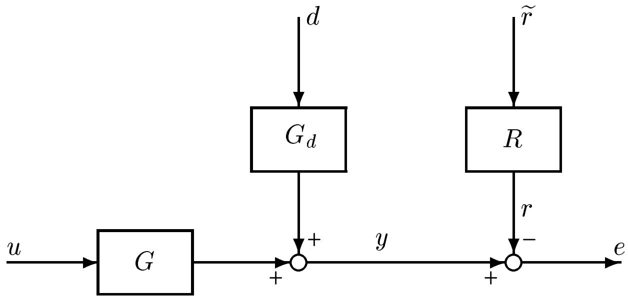


Figure 1.1: Model in terms of scaled variables

error (typically, $R \geq 1$). The block diagram for the system in scaled variables may then be written as in Figure 1.1, for which the following control objective is relevant:

- In terms of scaled variables we have that $|d(t)| \leq 1$ and $|\widetilde{r}(t)| \leq 1$, and our control objective is to manipulate u with $|u(t)| \leq 1$ such that $|e(t)| = |y(t) - r(t)| \leq 1$ (at least most of the time).

Remark 1 A number of the interpretations used in the book depend critically on a correct scaling. In particular, this applies to the input-output controllability analysis presented in Chapters 5 and 6. Furthermore, for a MIMO system one cannot correctly make use of the sensitivity function $S = (I + GK)^{-1}$ unless the output errors are of comparable magnitude.

Remark 2 With the above scalings, the worst-case behaviour of a system is analyzed by considering disturbances d of magnitude 1, and references $\widetilde{r}$ of magnitude 1.

Remark 3 The control error is

$$e = y - r = Gu + G_d d - R\widetilde{r} \tag{1.18}$$

and we see that a normalized reference change $\widetilde{r}$ may be viewed as a special case of a disturbance with $G_d = -R$, where R is usually a constant diagonal matrix. We will sometimes use this to unify our treatment of disturbances and references.

Remark 4 The scaling of the outputs in (1.11) in terms of the control error is used when analyzing a given plant. However, if the issue is to *select* which outputs to control, see Section 10.3, then one may choose to scale the outputs with respect to their expected variation (which is usually similar to $\widehat{r}_{\max}$).

Remark 5 If the expected or allowed variation of a variable about 0 (its nominal value) is not symmetric, then the largest variation should be used for $\widehat{d}_{\max}$ and the smallest variation for $\widehat{u}_{\max}$ and $\widehat{e}_{\max}$. For example, if the disturbance is $-5 \leq \widehat{d} \leq 10$ then $\widehat{d}_{\max} = 10$, and if the manipulated input is $-5 \leq \widehat{u} \leq 10$ then $\widehat{u}_{\max} = 5$. This approach may be conservative (in terms of allowing too large disturbances etc.) when the variations for *several* variables are not symmetric.

A further discussion on scaling and performance is given in Chapter 5 on page 161.

1.5 Deriving linear models

Linear models may be obtained from physical "first-principle" models, from analyzing input-output data, or from a combination of these two approaches. Although modelling and system identification are not covered in this book, it is always important for a control engineer to have a good understanding of a model's origin. The following steps are usually taken when deriving a linear model for controller design based on a first-principle approach:

1. Formulate a nonlinear state-space model based on physical knowledge.
2. Determine the steady-state operating point (or trajectory) about which to linearize.
3. Introduce deviation variables and linearize the model. There are essentially three parts to this step:

   (a) Linearize the equations using a Taylor expansion where second and higher order terms are omitted.

   (b) Introduce the deviation variables, e.g. $\delta x(t)$ defined by

   $$\delta x(t) = x(t) - x^*$$

   where the superscript * denotes the steady-state operating point or trajectory along which we are linearizing.

   (c) Subtract the steady-state to eliminate the terms involving only steady-state quantities.

   These parts are usually accomplished together. For example, for a nonlinear state-space model of the form

   $$\frac{dx}{dt} = f(x, u) \tag{1.19}$$

the linearized model in deviation variables $(\delta x, \delta u)$ is

$$\frac{d\delta x(t)}{dt} = \underbrace{\left(\frac{\partial f}{\partial x}\right)^*}_{A} \delta x(t) + \underbrace{\left(\frac{\partial f}{\partial u}\right)^*}_{B} \delta u(t) \tag{1.20}$$

Here x and u may be vectors, in which case the Jacobians A and B are matrices.

4. Scale the variables to obtain scaled models which are more suitable for control purposes.

In most cases steps 2 and 3 are performed numerically based on the model obtained in step 1. Also, since (1.20) is in terms of deviation variables, its Laplace transform becomes $s\delta x(s) = A\delta x(s) + B\delta u(s)$, or

$$\delta x(s) = (sI - A)^{-1} B\delta u(s) \tag{1.21}$$

Example 1.1 Physical model of a room heating process.

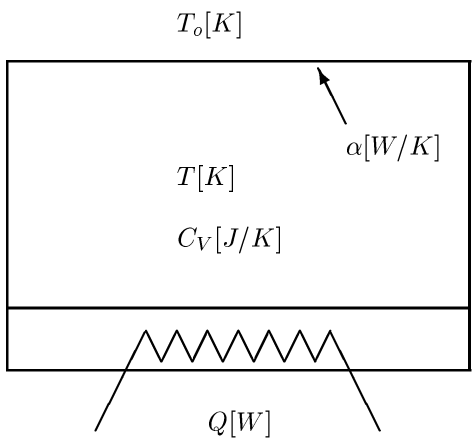


Figure 1.2: Room heating process

The above steps for deriving a linear model will be illustrated on the simple example depicted in Figure 1.2, where the control problem is to adjust the heat input Q to maintain constant room temperature T (within ± 1 K). The outdoor temperature T_o is the main disturbance. Units are shown in square brackets.

1. Physical model. *An energy balance for the room requires that the change in energy in the room must equal the net inflow of energy to the room (per unit of time). This yields the following state-space model*

$$\frac{d}{dt}(C_V T) = Q + \alpha(T_o - T) \tag{1.22}$$

where T [K] is the room temperature, C_V [J/K] is the heat capacity of the room, Q [W] is the heat input (from some heat source), and the term $\alpha(T_o - T)$ [W] represents the net heat loss due to exchange of air and heat conduction through the walls.

2. Operating point. *Consider a case where the heat input Q^* is 2000 W and the difference between indoor and outdoor temperatures $T^* - T_o^*$ is 20 K. Then the steady-state energy balance yields $\alpha^* = 2000/20 = 100$ W/K. We assume the room heat capacity is constant, $C_V = 100$ kJ/K. (This value corresponds approximately to the heat capacity of air in a room of about 100 m^3; thus we neglect heat accumulation in the walls.)*

3. Linear model in deviation variables. *If we assume α is constant the model in (1.22) is already linear. Then introducing deviation variables*

$$\delta T(t) = T(t) - T^*(t),\ \delta Q(t) = Q(t) - Q^*(t),\ \delta T_o(t) = T_o(t) - T_o^*(t)$$

yields

$$C_V \frac{d}{dt}\delta T(t) = \delta Q(t) + \alpha(\delta T_o(t) - \delta T(t)) \tag{1.23}$$

Remark. *If α depended on the state variable (T in this example), or on one of the independent variables of interest (Q or T_o in this example), then one would have to include an extra term $(T^* - T_o^*)\delta\alpha(t)$ on the right hand side of Equation (1.23).*

On taking Laplace transforms in (1.23), assuming $\delta T(t) = 0$ at $t = 0$, and rearranging we get

$$\delta T(s) = \frac{1}{\tau s + 1}\left(\frac{1}{\alpha}\delta Q(s) + \delta T_o(s)\right); \quad \tau = \frac{C_V}{\alpha} \tag{1.24}$$

The time constant for this example is $\tau = 100 \cdot 10^3/100 = 1000$ s ≈ 17 min which is reasonable. It means that for a step increase in heat input it will take about 17min for the temperature to reach 63% of its steady-state increase.

4. Linear model in scaled variables. *Introduce the following scaled variables*

$$y(s) = \frac{\delta T(s)}{\delta T_{\max}}; \quad u(s) = \frac{\delta Q(s)}{\delta Q_{\max}}; \quad d(s) = \frac{\delta T_o(s)}{\delta T_{o,max}} \tag{1.25}$$

In our case the acceptable variations in room temperature T are ± 1 K, i.e. $\delta T_{\max} = \delta e_{\max} = 1$ K. Furthermore, the heat input can vary between 0 W and 6000 W, and since its nominal value is 2000 W we have $\delta Q_{\max} = 2000$ W (see Remark 5 on page 8). Finally, the expected variations in outdoor temperature are ± 10 K, i.e. $\delta T_{o,max} = 10$ K. The model in terms of scaled variables then becomes

$$\begin{aligned} G(s) &= \frac{1}{\tau s + 1}\frac{\delta Q_{\max}}{\delta T_{\max}}\frac{1}{\alpha} = \frac{20}{1000s + 1} \\ G_d(s) &= \frac{1}{\tau s + 1}\frac{\delta T_{o,max}}{\delta T_{\max}} = \frac{10}{1000s + 1} \end{aligned} \tag{1.26}$$

Note that the static gain for the input is $k = 20$, whereas the static gain for the disturbance is $k_d = 10$. The fact that $|k_d| > 1$ means that we need some control (feedback or feedforward) to keep the output within its allowed bound ($|e| \leq 1$) when there is a disturbance of magnitude $|d| = 1$. The fact that $|k| > |k_d|$ means that we have enough "power" in the inputs to reject the disturbance at steady state, that is, we can, using an input of magnitude $|u| \leq 1$, have perfect disturbance rejection ($e = 0$) for the maximum disturbance ($|d| = 1$). We will return with a detailed discussion of this in Section 5.16.2 where we analyze the input-output controllability of the room heating process.

1.6 Notation

There is no standard notation to cover all of the topics covered in this book. We have tried to use the most familiar notation from the literature whenever possible, but an overriding concern has been to be consistent within the book, to ensure that the reader can follow the ideas and techniques through from one chapter to another.

The most important notation is summarized in Figure 1.3, which shows a one degree-of-freedom control configuration with negative feedback, a two degrees-of-freedom control configuration, and a general control configuration. The latter can be used to represent a wide class of controllers, including the one and two degrees-of-freedom configurations, as well as feedforward and estimation schemes and many others; and, as we will see, it can also be used to formulate optimization problems for controller design. The symbols used in Figure 1.3 are defined in Table 1.1. Apart from the use of v to represent the controller inputs for the general configuration, this notation is reasonably standard.

Lower-case letters are used for vectors and signals (e.g. u, y, n), and capital letters for matrices, transfer functions and systems (e.g. G, K). Matrix elements are usually denoted by lower-case letters, so g_{ij} is the ij'th element in the matrix G. However, sometimes we use upper-case letters G_{ij}, for example if G is partitioned so that G_{ij} is itself a matrix, or to avoid conflicts in notation. The Laplace variable s is often omitted for simplicity, so we often write G when we mean $G(s)$.

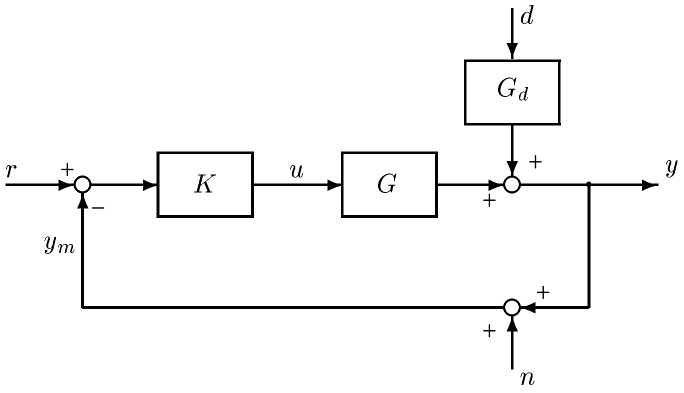
For state-space realizations we use the standard (A, B, C, D)-notation. That is, a system G with a state-space realization (A, B, C, D) has a transfer function $G(s) = C(sI - A)^{-1}B + D$. We sometimes write

$$G(s) \stackrel{s}{=} \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right] \tag{1.27}$$

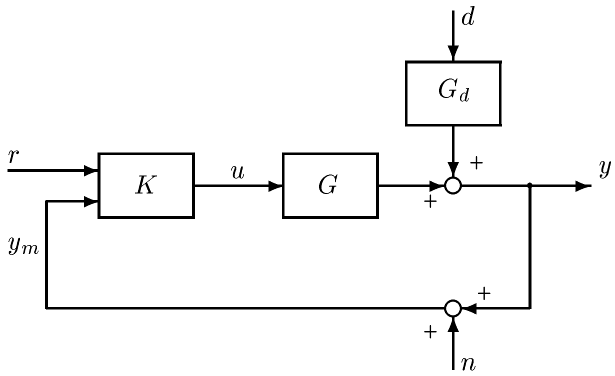
to mean that the transfer function $G(s)$ has a state-space realization given by the quadruple (A, B, C, D).

For closed-loop transfer functions we use S to denote the sensitivity at the plant output, and $T = I - S$ to denote the complementary sensitivity. With negative feedback, $S = (I + L)^{-1}$ and $T = L(I + L)^{-1}$, where L is the transfer function around the loop as seen from the output. In most cases $L = GK$, but if we also include measurement dynamics ($y_m = G_m y + n$) then $L = GKG_m$. The corresponding transfer functions as seen from the input of the plant are $L_I = KG$ (or $L_I = KG_mG$), $S_I = (I + L_I)^{-1}$ and $T_I = L_I(I + L_I)^{-1}$.

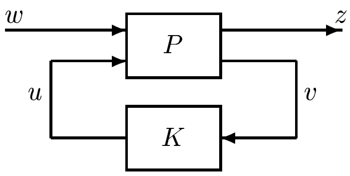
To represent uncertainty we use perturbations E (not normalized) or perturbations Δ (normalized such that their magnitude (norm) is less than or equal to one). The nominal plant model is G, whereas the perturbed model with uncertainty is denoted G_p (usually for a set of possible perturbed plants) or G' (usually for a particular perturbed plant). For example, with additive uncertainty we may have $G_p = G + E_A = G + w_A\Delta_A$, where w_A is a weight representing the magnitude of the uncertainty.



(a) One degree-of-freedom control configuration



(b) Two degrees-of-freedom control configuration



(c) General control configuration

Figure 1.3: Control configurations

Table 1.1: Nomenclature

K	controller, in whatever configuration. Sometimes the controller is broken down into its constituent parts. For example, in the two degrees-of-freedom controller in Figure 1.3(b), $K = \begin{bmatrix} K_r \\ K_y \end{bmatrix}$ where K_r is a prefilter and K_y is the feedback controller.
For the conventional control configurations (Figure 1.3(a) and (b)):	
G	plant model
G_d	disturbance model
r	reference inputs (commands, setpoints)
d	disturbances (process noise)
n	measurement noise
y	plant outputs. These signals include the variables to be controlled ("primary" outputs with reference values r) and possibly some additional "secondary" measurements to improve control. Usually the signals y are measurable.
y_m	measured y
u	control signals (manipulated plant inputs)
For the general control configuration (Figure 1.3(c)):	
P	generalized plant model. It will include G and G_d and the interconnection structure between the plant and the controller. In addition, if P is being used to formulate a design problem, then it will also include weighting functions.
w	exogenous inputs: commands, disturbances and noise
z	exogenous outputs; "error" signals to be minimized, e.g. $y - r$
v	controller inputs for the general configuration, e.g. commands, measured plant outputs, measured disturbances, etc. For the special case of a one degree-of-freedom controller with perfect measurements we have $v = r - y$.
u	control signals

By the right-half plane (RHP) we mean the closed right half of the complex plane, including the imaginary axis ($j\omega$-axis). The left-half plane (LHP) is the open left half of the complex plane, excluding the imaginary axis. A RHP-pole (unstable pole) is a pole located in the right-half plane, and thus includes poles on the imaginary axis. Similarly, a RHP-zero ("unstable" zero) is a zero located in the right-half plane.

We use A^T to denote the transpose of a matrix A, and A^H to represent its complex conjugate transpose.

Mathematical terminology

The symbol $\triangleq$ is used to denote *equal by definition*, $\stackrel{\text{def}}{\Leftrightarrow}$ is used to denote equivalent by definition, and $A \equiv B$ means that A is identically equal to B.

Let A and B be logic statements. Then the following expressions are equivalent:

A $\Leftarrow$ B
A if B, or: If B then A
A is necessary for B
B $\Rightarrow$ A, or: B implies A
B is sufficient for A
B only if A
not A $\Rightarrow$ not B

The remaining notation, special terminology and abbreviations will be defined in the text.

2

CLASSICAL FEEDBACK CONTROL

In this chapter, we review the classical frequency-response techniques for the analysis and design of single-loop (single-input single-output, SISO) feedback control systems. These loop-shaping techniques have been successfully used by industrial control engineers for decades, and have proved to be indispensable when it comes to providing insight into the benefits, limitations and problems of feedback control. During the 1980's the classical methods were extended to a more formal method based on shaping closed-loop transfer functions, for example, by considering the $\mathcal{H}_\infty$ norm of the weighted sensitivity function. We introduce this method at the end of the chapter.

The same underlying ideas and techniques will recur throughout the book as we present practical procedures for the analysis and design of multivariable (multi-input multi-output, MIMO) control systems.

2.1 Frequency response

On replacing s by $j\omega$ in a transfer function model $G(s)$ we get the so-called frequency response description. Frequency responses can be used to describe: 1) a system's response to sinusoids of varying frequency, 2) the frequency content of a deterministic signal via the Fourier transform, and 3) the frequency distribution of a stochastic signal via the power spectral density function.

In this book we use the first interpretation, namely that of frequency-by-frequency sinusoidal response. This interpretation has the advantage of being directly linked to the time domain, and at each frequency ω the complex number $G(j\omega)$ (or complex matrix for a MIMO system) has a clear physical interpretation. It gives the response to an input sinusoid of frequency ω. This will be explained in more detail below. For the other two interpretations we cannot assign a clear physical meaning to $G(j\omega)$ or $y(j\omega)$ at a particular frequency – it is the distribution relative to other frequencies which matters then.

One important advantage of a frequency response analysis of a system is that it provides insight into the benefits and trade-offs of feedback control. Although this insight may be obtained by viewing the frequency response in terms of its relationship between power spectral densities, as is evident from the excellent treatment by Kwakernaak and Sivan (1972), we believe that the frequency-by-frequency sinusoidal response interpretation is the most transparent and useful.

Frequency-by-frequency sinusoids

We now want to give a physical picture of frequency response in terms of a system's response to persistent sinusoids. It is important that the reader has this picture in mind when reading the rest of the book. For example, it is needed to understand the response of a multivariable system in terms of its singular value decomposition. A physical interpretation of the frequency response for a stable linear system $y = G(s)u$ is a follows. Apply a sinusoidal input signal with frequency ω [rad/s] and magnitude u_0, such that

$$u(t) = u_0 \sin(\omega t + \alpha)$$

This input signal is persistent, that is, it has been applied since $t = -\infty$. Then the output signal is also a persistent sinusoid of the same frequency, namely

$$y(t) = y_0 \sin(\omega t + \beta)$$

Here u_0 and y_0 represent magnitudes and are therefore both non-negative. Note that the output sinusoid has a different amplitude y_0 and is also shifted in phase from the input by

$$\phi \triangleq \beta - \alpha$$

Importantly, it can be shown that y_0/u_0 and ϕ can be obtained directly from the Laplace transform $G(s)$ after inserting the imaginary number $s = j\omega$ and evaluating the magnitude and phase of the resulting complex number $G(j\omega)$. We have

$$y_0/u_0 = |G(j\omega)|; \quad \phi = \angle G(j\omega) \text{ [rad]} \tag{2.1}$$

For example, let $G(j\omega) = a + jb$, with real part $a = \text{Re } G(j\omega)$ and imaginary part $b = \text{Im } G(j\omega)$, then

$$|G(j\omega)| = \sqrt{a^2 + b^2}; \quad \angle G(j\omega) = \arctan(b/a) \tag{2.2}$$

In words, (2.1) says that *after sending a sinusoidal signal through a system $G(s)$, the signal's magnitude is amplified by a factor $|G(j\omega)|$ and its phase is shifted by $\angle G(j\omega)$*. In Figure 2.1, this statement is illustrated for the following first-order delay system (time in seconds),

$$G(s) = \frac{ke^{-\theta s}}{\tau s + 1}; \quad k = 5, \theta = 2, \tau = 10 \tag{2.3}$$

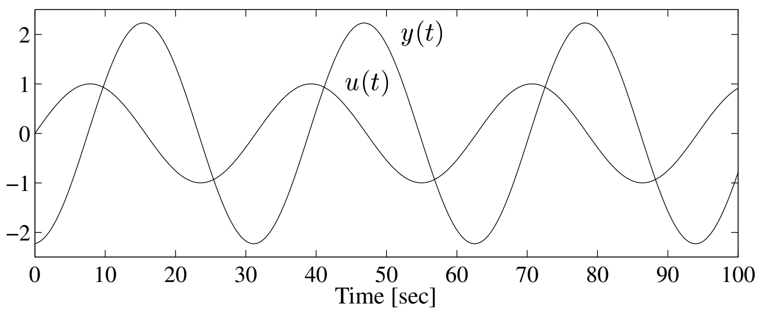


Figure 2.1: Sinusoidal response for system $G(s) = 5e^{-2s}/(10s+1)$ at frequency $\omega = 0.2$ rad/s

At frequency $\omega = 0.2$ rad/s, we see that the output y lags behind the input by about a quarter of a period and that the amplitude of the output is approximately twice that of the input. More accurately, the amplification is

$$|G(j\omega)| = k/\sqrt{(\tau\omega)^2+1} = 5/\sqrt{(10\omega)^2+1} = 2.24$$

and the phase shift is

$$\phi = \angle G(j\omega) = -\arctan(\tau\omega) - \theta\omega = -\arctan(10\omega) - 2\omega = -1.51\,\mathrm{rad} = -86.5^\circ$$

$G(j\omega)$ is called the *frequency response* of the system $G(s)$. It describes how the system responds to persistent sinusoidal inputs of frequency ω. The magnitude of the frequency response, $|G(j\omega)|$, being equal to $|y_0(\omega)|/|u_0(\omega)|$, is also referred to as the *system gain*. Sometimes the gain is given in units of dB (decibel) defined as

$$A\,[\mathrm{dB}] = 20\log_{10} A \tag{2.4}$$

For example, $A = 2$ corresponds to $A = 6.02$ dB, and $A = \sqrt{2}$ corresponds to $A = 3.01$ dB, and $A = 1$ corresponds to $A = 0$ dB.

Both $|G(j\omega)|$ and $\angle G(j\omega)$ depend on the frequency ω. This dependency may be plotted explicitly in Bode plots (with ω as independent variable) or somewhat implicitly in a Nyquist plot (phase plane plot). In Bode plots we usually employ a log-scale for frequency and gain, and a linear scale for the phase.

In Figure 2.2, the Bode plots are shown for the system in (2.3). We note that in this case both the gain and phase fall monotonically with frequency. This is quite common for process control applications. The delay θ only shifts 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 $s = 0$ (or $\omega = 0$). The gain remains relatively constant up to the break frequency $1/\tau$ where it starts falling sharply. Physically, the system responds too slowly to let high-frequency

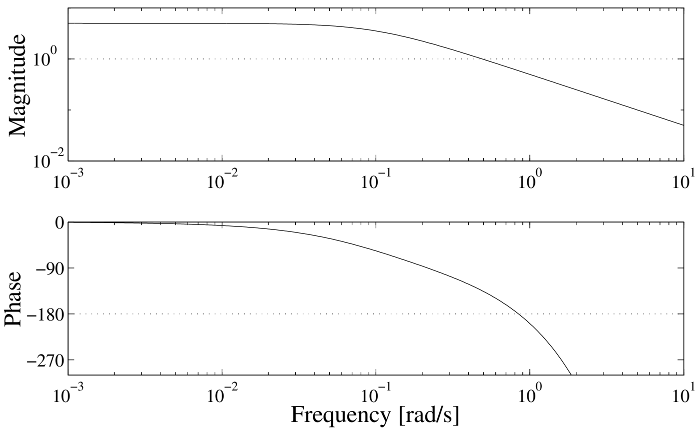


Figure 2.2: Frequency response (Bode plots) of $G(s) = 5e^{-2s}/(10s + 1)$

("fast") inputs have much effect on the outputs, and sinusoidal inputs with $\omega > 1/\tau$ are attenuated by the system dynamics.

The frequency response is also useful for an *unstable plant* $G(s)$, which by itself has no steady-state response. Let $G(s)$ be stabilized by feedback control, and consider applying a sinusoidal forcing signal to the stabilized system. In this case all signals within the system are persistent sinusoids with the same frequency ω, and $G(j\omega)$ yields as before the sinusoidal response from the input to the output of $G(s)$.

Phasor notation. From Euler's formula for complex numbers we have that $e^{jz} = \cos z + j \sin z$. It then follows that $\sin(\omega t)$ is equal to the imaginary part of the complex function $e^{j\omega t}$, and we can write the time domain sinusoidal response in complex form as follows:

$$u(t) = u_0 \text{Im } e^{j(\omega t+\alpha)} \text{ gives as } t \to \infty \quad y(t) = y_0 \text{Im } e^{j(\omega t+\beta)} \tag{2.5}$$

where

$$y_0 = |G(j\omega)|u_0, \quad \beta = \angle G(j\omega) + \alpha \tag{2.6}$$

and $|G(j\omega)|$ and $\angle G(j\omega)$ are defined in (2.2). Now introduce the complex numbers

$$u(\omega) \triangleq u_0 e^{j\alpha}, \quad y(\omega) \triangleq y_0 e^{j\beta} \tag{2.7}$$

where we have used ω as an argument because y_0 and β depend on frequency, and in some cases so may u_0 and α. Note that $u(\omega)$ is *not* equal to $u(s)$ evaluated at $s = \omega$ nor is it equal to $u(t)$ evaluated at $t = \omega$. Since $G(j\omega) = |G(j\omega)| \; e^{j\angle G(j\omega)}$ the sinusoidal response in (2.5) and (2.6) can then be written on complex form as follows

$$y(\omega)e^{j\omega t} = G(j\omega)u(\omega)e^{j\omega t} \tag{2.8}$$

or because the term $e^{j\omega t}$ appears on both sides

$$\boxed{y(\omega) = G(j\omega)u(\omega)} \tag{2.9}$$

which we refer to as the phasor notation. At each frequency, $u(\omega)$, $y(\omega)$ and $G(j\omega)$ are complex numbers, and the usual rules for multiplying complex numbers apply. We will use this phasor notation throughout the book. Thus *whenever we use notation such as $u(\omega)$ (with ω and not $j\omega$ as an argument), the reader should interpret this as a (complex) sinusoidal signal, $u(\omega)e^{j\omega t}$*. (2.9) also applies to MIMO systems where $u(\omega)$ and $y(\omega)$ are complex vectors representing the sinusoidal signal in each channel and $G(j\omega)$ is a complex matrix.

Minimum phase systems. For stable systems which are minimum phase (no time delays or right-half plane (RHP) zeros) there is a unique relationship between the gain and phase of the frequency response. This may be quantified by the Bode gain-phase relationship which gives the phase of G (normalized[1] such that $G(0) > 0$) at a given frequency ω_0 as a function of $|G(j\omega)|$ over the entire frequency range:

$$\angle G(j\omega_0) = \frac{1}{\pi}\int_{-\infty}^{\infty} \underbrace{\frac{d\ln|G(j\omega)|}{d\ln\omega}}_{N(\omega)} \ln\left|\frac{\omega+\omega_0}{\omega-\omega_0}\right| \cdot \frac{d\omega}{\omega} \tag{2.10}$$

The name *minimum phase* refers to the fact that such a system has the minimum possible phase lag for the given magnitude response $|G(j\omega)|$. The term $N(\omega)$ is the slope of the magnitude in log-variables at frequency ω. In particular, the local slope at frequency ω_0 is

$$N(\omega_0) = \left(\frac{d\ln|G(j\omega)|}{d\ln\omega}\right)_{\omega=\omega_0}$$

The term $\ln\left|\frac{\omega+\omega_0}{\omega-\omega_0}\right|$ in (2.10) is infinite at $\omega = \omega_0$, so it follows that $\angle G(j\omega_0)$ is primarily determined by the local slope $N(\omega_0)$. Also $\int_{-\infty}^{\infty} \ln\left|\frac{\omega+\omega_0}{\omega-\omega_0}\right| \cdot \frac{d\omega}{\omega} = \frac{\pi^2}{2}$ which justifies the commonly used approximation for stable minimum phase systems

$$\angle G(j\omega_0) \approx \frac{\pi}{2}N(\omega_0)\ [\mathrm{rad}] = 90^\circ \cdot N(\omega_0) \tag{2.11}$$

The approximation is exact for the system $G(s) = 1/s^n$ (where $N(\omega) = -n$), and it is good for stable minimum phase systems except at frequencies close to those of resonance (complex) poles or zeros.

RHP-zeros and time delays contribute additional phase lag to a system when compared to that of a minimum phase system with the same gain (hence the term *non-minimum phase* system). For example, the system $G(s) = \frac{-s+a}{s+a}$ with a RHP-zero at

[1] The normalization of $G(s)$ is necessary to handle systems such as $\frac{1}{s+2}$ and $\frac{-1}{s+2}$, which have equal gain, are stable and minimum phase, but their phases differ by 180°. Systems with integrators may be treated by replacing $\frac{1}{s}$ by $\frac{1}{s+\epsilon}$ where ϵ is a small positive number.

$s = a$ has a constant gain of 1, but its phase is $-2\arctan(\omega/a)$ [rad] (and not 0 [rad] as it would be for the minimum phase system $G(s) = 1$ of the same gain). Similarly, the time delay system $e^{-\theta s}$ has a constant gain of 1, but its phase is $-\omega\theta$ [rad].

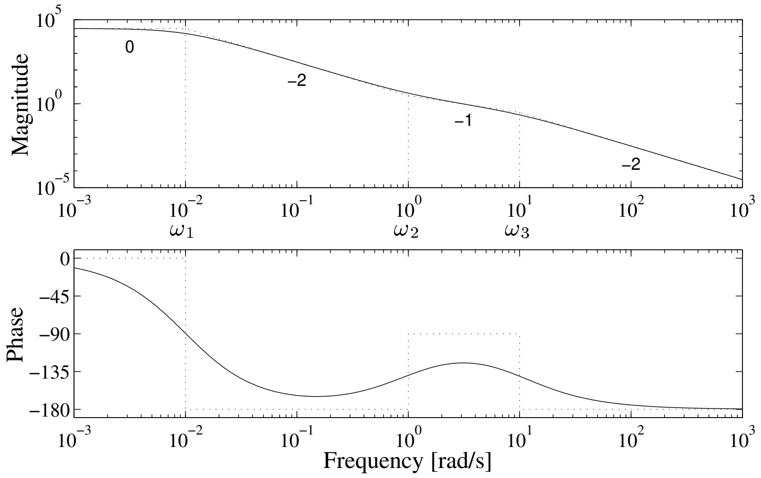


Figure 2.3: Bode plots of transfer function $L_1 = 30\frac{s+1}{(s+0.01)^2(s+10)}$. The asymptotes are given by dotted lines. The vertical dotted lines on the upper plot indicate the break frequencies ω_1, ω_2 and ω_3.

Straight-line approximations (asymptotes). For the design methods used in this book it is useful to be able to sketch Bode plots quickly, and in particular the magnitude (gain) diagram. The reader is therefore advised to become familiar with asymptotic Bode plots (straight-line approximations). For example, for a transfer function

$$G(s) = k\frac{(s+z_1)(s+z_2)\cdots}{(s+p_1)(s+p_2)\cdots} \tag{2.12}$$

the asymptotic Bode plots of $G(j\omega)$ are obtained by using for each term $s + a$ the approximation $j\omega + a \approx a$ for $\omega < a$ and by $j\omega + a \approx j\omega$ for $\omega > a$. These approximations yield straight lines on a log-log plot which meet at the so-called break point frequency $\omega = a$. In (2.12) therefore, the frequencies $z_1, z_2, \ldots, p_1, p_2, \ldots$ are the break points where the asymptotes meet. For complex poles or zeros, the term $s^2 + 2\zeta s\omega_0 + \omega_0^2$ (where $|\zeta| < 1$) is approximated by ω_0^2 for $\omega < \omega_0$ and by $s^2 = (j\omega)^2 = -\omega^2$ for $\omega > \omega_0$. The magnitude of a transfer function is usually close to its asymptotic value, and the only case when there is significant deviation is around the resonance frequency ω_0 for complex poles or zeros with a damping $|\zeta|$ of about 0.3 or less. In Figure 2.3, the Bode plots are shown for

$$L_1(s) = 30\frac{(s+1)}{(s+0.01)^2(s+10)} \tag{2.13}$$

The asymptotes (straight-line approximations) are shown by dotted lines. We note that the magnitude follows the asymptotes closely, whereas the phase does not. In this example the asymptotic slope of L_1 is 0 up to the first break frequency at $\omega_1 = 0.01$ rad/s where we have two poles and then the slope changes to $N = -2$. Then at $\omega_2 = 1$ rad/s there is a zero and the slope changes to $N = -1$. Finally, there is a break frequency corresponding to a pole at $\omega_3 = 10$ rad/s and so the slope is $N = -2$ at this and higher frequencies.

2.2 Feedback control

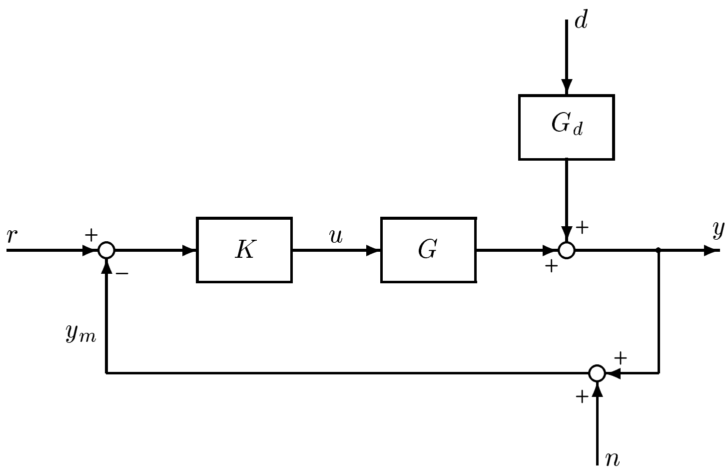


Figure 2.4: Block diagram of one degree-of-freedom feedback control system

2.2.1 One degree-of-freedom controller

In most of this chapter, we examine the simple one degree-of-freedom negative feedback structure shown in Figure 2.4. The input to the controller $K(s)$ is $r - y_m$ where $y_m = y + n$ is the measured output and n is the measurement noise. Thus, the input to the plant is

$$u = K(s)(r - y - n) \tag{2.14}$$

The objective of control is to manipulate u (design K) such that the control error e remains small in spite of disturbances d. The control error e is defined as

$$e = y - r \tag{2.15}$$

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

Remark. In the literature, the control error is frequently defined as $r - y_m$ which is often the controller input. However, this is not a good definition of an error variable. First, the error is normally defined as the actual value (here y) minus the desired value (here r). Second, the error should involve the actual value (y) and not the measured value (y_m).

Note that we do not define e as the controller input $r - y_m$ which is frequently done.

2.2.2 Closed-loop transfer functions

The plant model is written as

$$y = G(s)u + G_d(s)d \tag{2.16}$$

and for a one degree-of-freedom controller the substitution of (2.14) into (2.16) yields

$$y = GK(r - y - n) + G_d d$$

or

$$(I + GK)y = GKr + G_d d - GKn \tag{2.17}$$

and hence the closed-loop response is

$$y = \underbrace{(I + GK)^{-1}GK}_{T}\, r + \underbrace{(I + GK)^{-1}}_{S}\, G_d d - \underbrace{(I + GK)^{-1}GK}_{T}\, n \tag{2.18}$$

The control error is

$$e = y - r = -Sr + SG_d d - Tn \tag{2.19}$$

where we have used the fact $T - I = -S$. The corresponding plant input signal is

$$u = KSr - KSG_d d - KSn \tag{2.20}$$

The following notation and terminology are used

$$\begin{array}{rl} L = GK & \text{loop transfer function} \\ S = (I + GK)^{-1} = (I + L)^{-1} & \text{sensitivity function} \\ T = (I + GK)^{-1}GK = (I + L)^{-1}L & \text{complementary sensitivity function} \end{array}$$

We see that S is the closed-loop transfer function from the output disturbances to the outputs, while T is the closed-loop transfer function from the reference signals to the outputs. The term complementary sensitivity for T follows from the identity:

$$S + T = I \tag{2.21}$$

To derive (2.21), write $S + T = (I + L)^{-1} + (I + L)^{-1}L$ and factor out the term $(I + L)^{-1}$. The term sensitivity function is natural because S gives the sensitivity reduction afforded by feedback. To see this, consider the "open-loop" case i.e. with no feedback. Then

$$y = GKr + G_d d + 0 \cdot n \tag{2.22}$$

and a comparison with (2.18) shows that, with the exception of noise, the response with feedback is obtained by premultiplying the right hand side by S.

Remark 1 Actually, the above is not the original reason for the name "sensitivity". Bode first called S sensitivity because it gives the relative sensitivity of the closed-loop transfer function T to the relative plant model error. In particular, at a given frequency ω we have for a SISO plant, by straightforward differentiation of T, that

$$\frac{dT/T}{dG/G} = S \tag{2.23}$$

Remark 2 Equations (2.14)-(2.22) are written in matrix form because they also apply to MIMO systems. Of course, for SISO systems we may write $S + T = 1$, $S = \frac{1}{1+L}$, $T = \frac{L}{1+L}$ and so on.

Remark 3 In general, closed-loop transfer functions for SISO systems with *negative* feedback may be obtained from the rule

$$\text{OUTPUT} = \frac{\text{“direct”}}{1 + \text{“loop”}} \cdot \text{INPUT} \tag{2.24}$$

where "direct" represents the transfer function for the direct effect of the input on the output (with the feedback path open) and "loop" is the transfer function around the loop (denoted $L(s)$). In the above case $L = GK$. If there is also a measurement device, $G_m(s)$, in the loop, then $L(s) = GKG_m$. The rule in (2.24) is easily derived by generalizing (2.17). In Section 3.2, we present a more general form of this rule which also applies to multivariable systems.

2.2.3 Why feedback?

At this point it is pertinent to ask why we should use feedback control at all — rather than simply using feedforward control. A "perfect" feedforward controller is obtained by removing the feedback signal and using the controller

$$K_r(s) = G^{-1}(s) \tag{2.25}$$

(we assume for now that it is possible to obtain and physically realize such an inverse, although this may of course not be true). We assume that the plant and controller are both stable and that all the disturbances are known, that is, we know $G_d d$, the effect of the disturbances on the outputs. Then with $r - G_d d$ as the controller input, this feedforward controller would yield perfect control:

$$y = Gu + G_d d = GK(r - G_d d) + G_d d = r$$

Unfortunately, G is never an exact model, and the disturbances are never known exactly. *The fundamental reasons for using feedback control are therefore the presence of*

1. Signal uncertainty – Unknown disturbance
2. Model uncertainty
3. An unstable plant

The third reason follows because unstable plants can only be stabilized by feedback (see internal stability in Chapter 4). The ability of feedback to reduce the effect of model uncertainty is of crucial importance in controller design.

2.3 Closed-loop stability

One of the main issues in designing feedback controllers is stability. If the feedback gain is too large, then the controller may "overreact" and the closed-loop system becomes unstable. This is illustrated next by a simple example.

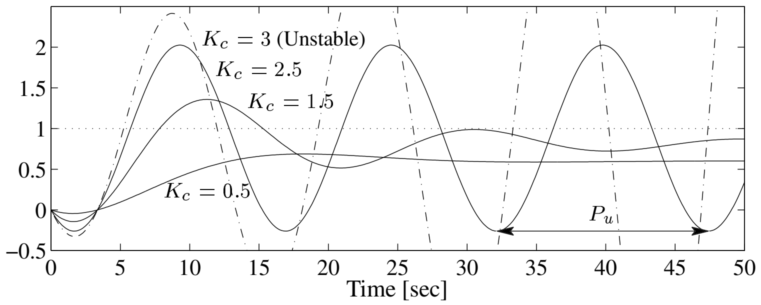


Figure 2.5: Effect of proportional gain K_c on the closed-loop response $y(t)$ of the inverse response process

Example 2.1 **Inverse response process.** *Consider the plant (time in seconds)*

$$\boxed{G(s) = \frac{3(-2s+1)}{(5s+1)(10s+1)}} \tag{2.26}$$

This is one of two main example processes used in this chapter to illustrate the techniques of classical control. The model has a right-half plane (RHP) zero at $s = 0.5$ rad/s. This imposes a fundamental limitation on control, and high controller gains will induce closed-loop instability.

This is illustrated for a proportional (P) controller $K(s) = K_c$ in Figure 2.5, where the response $y = Tr = GK_c(1+GK_c)^{-1}r$ to a step change in the reference ($r(t) = 1$ for $t > 0$) is shown for four different values of K_c. The system is seen to be stable for $K_c < 2.5$, and unstable for $K_c > 2.5$. The controller gain at the limit of instability, $K_u = 2.5$, is sometimes called the ultimate gain and for this value the system is seen to cycle continuously with a period $P_u = 15.2$ s, corresponding to the frequency $\omega_u \triangleq 2\pi/P_u = 0.42$ rad/s.

Two methods are commonly used to determine closed-loop stability:

1. The poles of the closed-loop system are evaluated. That is, the roots of $1+L(s)=0$ are found, where L is the transfer function around the loop. The system is stable *if and only if* all the closed-loop poles are in the open left-half plane (LHP) (that is, poles on the imaginary axis are considered "unstable"). The poles are also equal to the eigenvalues of the state-space A-matrix, and this is usually how the poles are computed numerically.
2. The frequency response (including negative frequencies) of $L(j\omega)$ is plotted in the complex plane and the number of encirclements it makes of the critical point -1 is counted. By Nyquist's stability criterion (for which a detailed statement is given in Theorem 4.7) closed-loop stability is inferred by equating the number of encirclements to the number of open-loop unstable poles (RHP-poles).
   For open-loop stable systems where $\angle L(j\omega)$ falls with frequency such that $\angle L(j\omega)$ crosses -180° only once (from above at frequency ω_{180}), one may equivalently use *Bode's stability condition* which says that the closed-loop system is stable if and only if the loop gain $|L|$ is less than 1 at this frequency, that is
   $$\text{Stability} \quad \Leftrightarrow \quad |L(j\omega_{180})| < 1 \tag{2.27}$$
   where ω_{180} is the phase crossover frequency defined by $\angle L(j\omega_{180}) = -180^\circ$.

Method 1, which involves computing the poles, is best suited for numerical calculations. However, time delays must first be approximated as rational transfer functions, e.g. Padé approximations. Method 2, which is based on the frequency response, has a nice graphical interpretation, and may also be used for systems with time delays. Furthermore, it provides useful measures of relative stability and forms the basis for several of the robustness tests used later in this book.

Example 2.2 **Stability of inverse response process with proportional control.** *Let us determine the condition for closed-loop stability of the plant G in (2.26) with proportional control, that is, with $K(s)=K_c$ and $L(s)=K_cG(s)$.*

1. *The system is stable if and only if all the closed-loop poles are in the LHP. The poles are solutions to $1+L(s)=0$ or equivalently the roots of*
   $$(5s+1)(10s+1) + K_c 3(-2s+1) = 0$$
   $$\Leftrightarrow \quad 50s^2 + (15-6K_c)s + (1+3K_c) = 0 \tag{2.28}$$
   But since we are only interested in the half plane location *of the poles, it is not necessary to solve (2.28). Rather, one may consider the coefficients a_i of the characteristic equation $a_n s^n + \cdots a_1 s + a_0 = 0$ in (2.28), and use the Routh-Hurwitz test to check for stability. For second order systems, this test says that we have stability if and only if all the coefficients have the same sign. This yields the following stability conditions*
   $$(15-6K_c) > 0; \quad (1+3K_c) > 0$$
   or equivalently $-1/3 < K_c < 2.5$. With negative feedback ($K_c \geq 0$) only the upper bound is of practical interest, and we find that the maximum allowed gain ("ultimate gain")

is $K_u = 2.5$ which agrees with the simulation in Figure 2.5. The poles at the onset of instability may be found by substituting $K_c = K_u = 2.5$ into (2.28) to get $50s^2 + 8.5 = 0$, i.e. $s = \pm j\sqrt{8.5/50} = \pm j0.412$. Thus, at the onset of instability we have two poles on the imaginary axis, and the system will be continuously cycling with a frequency $\omega = 0.412$ rad/s corresponding to a period $P_u = 2\pi/\omega = 15.2$ s. This agrees with the simulation results in Figure 2.5.

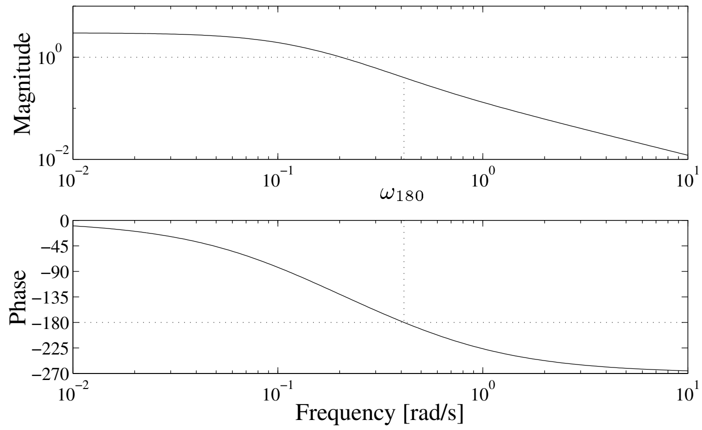


Figure 2.6: Bode plots for $L(s) = K_c \frac{3(-2s+1)}{(10s+1)(5s+1)}$ with $K_c = 1$

2. *Stability may also be evaluated from the frequency response of $L(s)$. A graphical evaluation is most enlightening. The Bode plots of the plant (i.e. $L(s)$ with $K_c = 1$) are shown in Figure 2.6. From these one finds the frequency ω_{180} where $\angle L$ is -180° and then reads off the corresponding gain. This yields $|L(j\omega_{180})| = K_c|G(j\omega_{180})| = 0.4K_c$, and we get from (2.27) that the system is stable if and only if $|L(j\omega_{180})| < 1 \Leftrightarrow K_c < 2.5$ (as found above). Alternatively, the phase crossover frequency may be obtained analytically from:*

$$\angle L(j\omega_{180}) = -\arctan(2\omega_{180}) - \arctan(5\omega_{180}) - \arctan(10\omega_{180}) = -180^\circ$$

which gives $\omega_{180} = 0.412$ rad/s as found in the pole calculation above. The loop gain at this frequency is

$$|L(j\omega_{180})| = K_c \frac{3 \cdot \sqrt{(2\omega_{180})^2 + 1}}{\sqrt{(5\omega_{180})^2 + 1} \cdot \sqrt{(10\omega_{180})^2 + 1}} = 0.4K_c$$

which is the same as found from the graph in Figure 2.6. The stability condition $|L(j\omega_{180})| < 1$ then yields $K_c < 2.5$ as expected.

2.4 Evaluating closed-loop performance

Although closed-loop stability is an important issue, the real objective of control is to improve performance, that is, to make the output $y(t)$ behave in a more desirable manner. Actually, the possibility of inducing instability is one of the disadvantages of feedback control which has to be traded off against performance improvement. The objective of this section is to discuss ways of evaluating closed-loop performance.

2.4.1 Typical closed-loop responses

The following example which considers proportional plus integral (PI) control of the inverse response process in (2.26), illustrates what type of closed-loop performance one might expect.

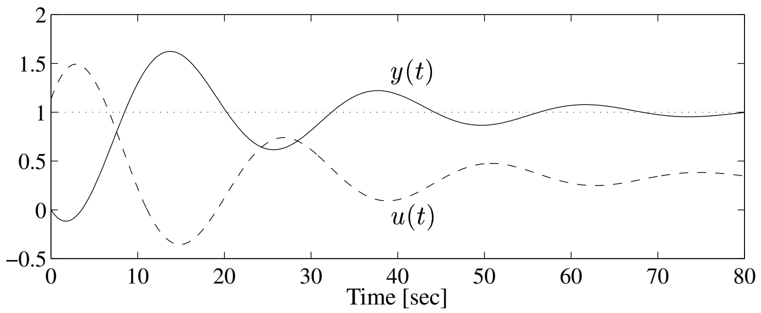


Figure 2.7: Closed-loop response to a step change in reference for the inverse response process with PI-control

Example 2.3 **PI-control of the inverse response process.** *We have already studied the use of a proportional controller for the process in (2.26). We found that a controller gain of* $K_c = 1.5$ *gave a reasonably good response, except for a steady-state offset (see Figure 2.5). The reason for this offset is the non-zero steady-state sensitivity function,* $S(0) = \frac{1}{1+K_cG(0)} = 0.18$ *(where* $G(0) = 3$ *is the steady-state gain of the plant). From* $e = -Sr$ *it follows that for* $r = 1$ *the steady-state control error is* -0.18 *(as is confirmed by the simulation in Figure 2.5). To remove the steady-state offset we add integral action in the form of a PI-controller*

$$K(s) = K_c\left(1 + \frac{1}{\tau_I s}\right) \tag{2.29}$$

The settings for K_c *and* τ_I *can be determined from the classical tuning rules of Ziegler and Nichols (1942):*

$$K_c = K_u/2.2, \quad \tau_I = P_u/1.2 \tag{2.30}$$

where K_u *is the maximum (ultimate) P-controller gain and* P_u *is the corresponding period of oscillations. In our case* $K_u = 2.5$ *and* $P_u = 15.2$ *s (as observed from the simulation in*

Figure 2.5), and we get $K_c = 1.14$ and $\tau_I = 12.7$ s. Alternatively, K_u and P_u can be obtained from the model $G(s)$,

$$K_u = 1/|G(j\omega_u)|, \quad P_u = 2\pi/\omega_u \tag{2.31}$$

where ω_u is defined by $\angle G(j\omega_u) = -180°$.

The closed-loop response, with PI-control, to a step change in reference is shown in Figure 2.7. The output $y(t)$ has an initial inverse response due to the RHP-zero, but it then rises quickly and $y(t) = 0.9$ at $t = 8.0$ s (the rise time). The response is quite oscillatory and it does not settle to within $\pm 5\%$ of the final value until after $t = 65$ s (the settling time). The overshoot (height of peak relative to the final value) is about 62% which is much larger than one would normally like for reference tracking. The decay ratio, which is the ratio between subsequent peaks, is about 0.35 which is also a bit large. However, for disturbance rejection the controller settings may be more reasonable, and one can always add a prefilter to improve the response for reference tracking, resulting in a two degrees-of-freedom controller.

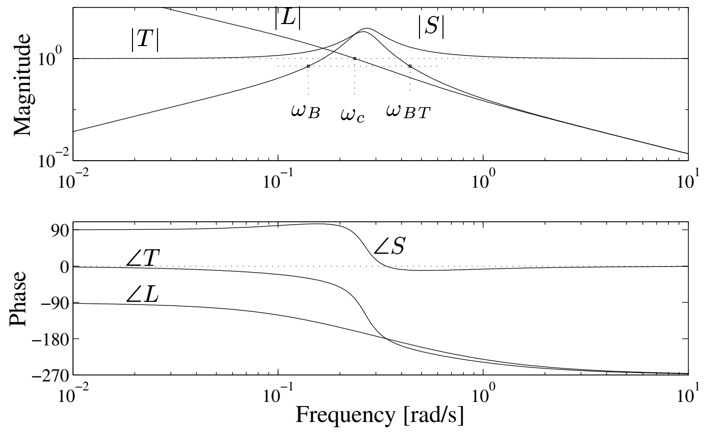


Figure 2.8: Bode magnitude and phase plots of $L = GK$, S and T when $G(s) = \frac{3(-2s+1)}{(5s+1)(10s+1)}$, and $K(s) = 1.136(1 + \frac{1}{12.7s})$ (a Ziegler-Nichols PI controller)

The corresponding Bode plots for L, S and T are shown in Figure 2.8. Later, in Section 2.4.3, we define stability margins, and from the plot of $L(j\omega)$, repeated in Figure 2.11, we find that the phase margin (PM) is 0.34 rad = $19.4°$ and the gain margin (GM) is 1.63. These margins are too small according to common rules of thumb. The peak value of $|S|$ is $M_S = 3.92$, and the peak value of $|T|$ is $M_T = 3.35$ which again are high according to normal design rules.

Exercise 2.1 *Use (2.31) to compute K_u and P_u for the process in (2.26).*

In summary, for this example, the Ziegler-Nichols' PI-tunings are somewhat "aggressive" and give a closed-loop system with smaller stability margins and a more oscillatory response than would normally be regarded as acceptable.

2.4.2 Time domain performance

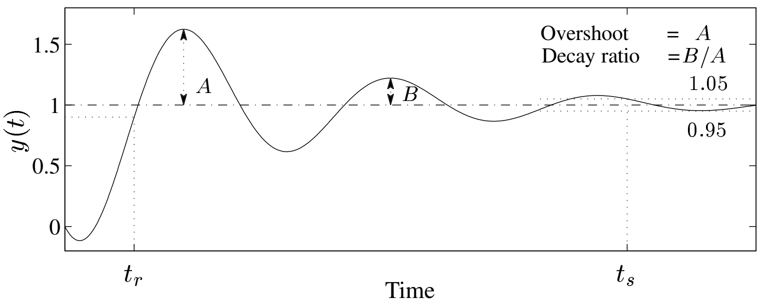


Figure 2.9: Characteristics of closed-loop response to step in reference

Step response analysis. The above example illustrates the approach often taken by engineers when evaluating the performance of a control system. That is, one simulates the response to a step in the reference input, and considers the following characteristics (see Figure 2.9):

- *Rise time:* (t_r) the time it takes for the output to first reach 90% of its final value, which is usually required to be small.
- *Settling time:* (t_s) the time after which the output remains within $\pm 5\%$ of its final value, which is usually required to be small.
- *Overshoot:* the peak value divided by the final value, which should typically be 1.2 (20%) or less.
- *Decay ratio:* the ratio of the second and first peaks, which should typically be 0.3 or less.
- *Steady-state offset:* the difference between the final value and the desired final value, which is usually required to be small.

The rise time and settling time are measures of the *speed of the response*, whereas the overshoot, decay ratio and steady-state offset are related to the *quality of the response*. Another measure of the quality of the response is:

- *Excess variation:* the total variation (TV) divided by the overall change at steady state, which should be as close to 1 as possible.

The total variation is the total movement of the output as illustrated in Figure 2.10. For the cases considered here the overall change is 1, so the excess variation is equal to the total variation.

The above measures address the output response, $y(t)$. In addition, one should consider the magnitude of the manipulated input (control signal, u), which usually should be as small and smooth as possible. If there are important disturbances, then

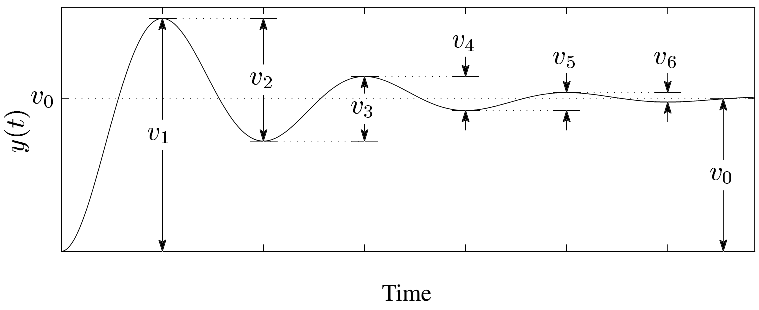


Figure 2.10: Total variation is $\mathrm{TV} = \sum_i v_i$, and Excess variation is TV/v_0

the response to these should also be considered. Finally, one may investigate in simulation how the controller works if the plant model parameters are different from their nominal values.

Remark 1 Another way of quantifying time domain performance is in terms of some norm of the error signal $e(t) = y(t) - r(t)$. For example, one might use the integral squared error (ISE), or its square root which is the 2-norm of the error signal, $\|e(t)\|_2 = \sqrt{\int_0^\infty |e(\tau)|^2 d\tau}$. Note that in this case the various objectives related to both the speed and quality of response are combined into one number. Actually, in most cases minimizing the 2-norm seems to give a reasonable trade-off between the various objectives listed above. Another advantage of the 2-norm is that the resulting optimization problems (such as minimizing ISE) are numerically easy to solve. One can also take input magnitudes into account by considering, for example, $J = \sqrt{\int_0^\infty (Q|e(t)|^2 + R|u(t)|^2) dt}$ where Q and R are positive constants. This is similar to linear quadratic (LQ) optimal control, but in LQ-control one normally considers an impulse rather than a step change in $r(t)$.

Remark 2 The step response is equal to the integral of the corresponding impulse response, e.g. set $u(\tau) = 1$ in (4.11). Some thought then reveals that one can compute the total variation as the integrated absolute area (1-norm) of the corresponding impulse response (Boyd and Barratt, 1991, p. 98). That is, let $y = Tr$, then the total variation in y for a step change in r is

$$\mathrm{TV} = \int_0^\infty |g_T(\tau)| d\tau \triangleq \|g_T(t)\|_1 \tag{2.32}$$

where $g_T(t)$ is the impulse response of T, i.e. $y(t)$ resulting from an impulse change in $r(t)$.

2.4.3 Frequency domain performance

The frequency-response of the loop transfer function, $L(j\omega)$, or of various closed-loop transfer functions, may also be used to characterize closed-loop performance. Typical Bode plots of L, T and S are shown in Figure 2.8. One advantage of the frequency domain compared to a step response analysis, is that it considers a broader class of signals (sinusoids of any frequency). This makes it easier to characterize feedback properties, and in particular system behaviour in the crossover (bandwidth) region. We will now describe some of the important frequency-domain measures used to assess performance e.g. gain and phase margins, the maximum peaks of S and T, and the various definitions of crossover and bandwidth frequencies used to characterize speed of response.

Gain and phase margins

Let $L(s)$ denote the loop transfer function of a system which is closed-loop stable under negative feedback. A typical Bode plot and a typical Nyquist plot of $L(j\omega)$ illustrating the gain margin (GM) and phase margin (PM) are given in Figures 2.11 and 2.12, respectively.

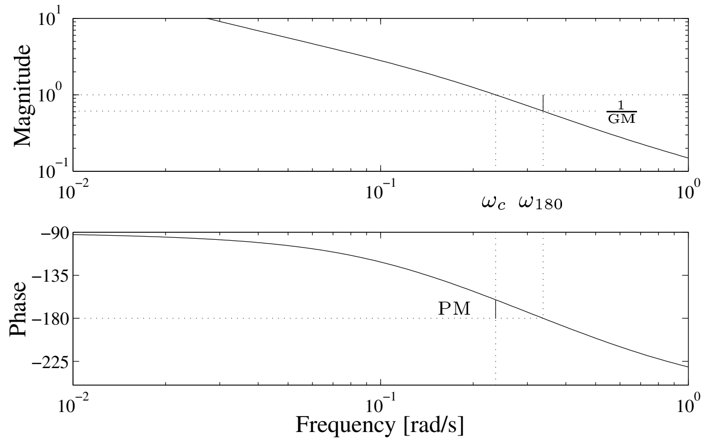


Figure 2.11: Typical Bode plot of $L(j\omega)$ with PM and GM indicated

The *gain margin* is defined as

$$\mathrm{GM} = 1/|L(j\omega_{180})| \tag{2.33}$$

where the *phase crossover frequency* ω_{180} is where the Nyquist curve of $L(j\omega)$ crosses the negative real axis between -1 and 0, that is

$$\angle L(j\omega_{180}) = -180^\circ \tag{2.34}$$

If there is more than one crossing the largest value of $|L(j\omega_{180})|$ is taken. On a Bode plot with a logarithmic axis for $|L|$, we have that GM (in logarithms, e.g. in dB) is the vertical distance from the unit magnitude line down to $|L(j\omega_{180})|$, see Figure 2.11. The GM is the factor by which the loop gain $|L(j\omega)|$ may be increased

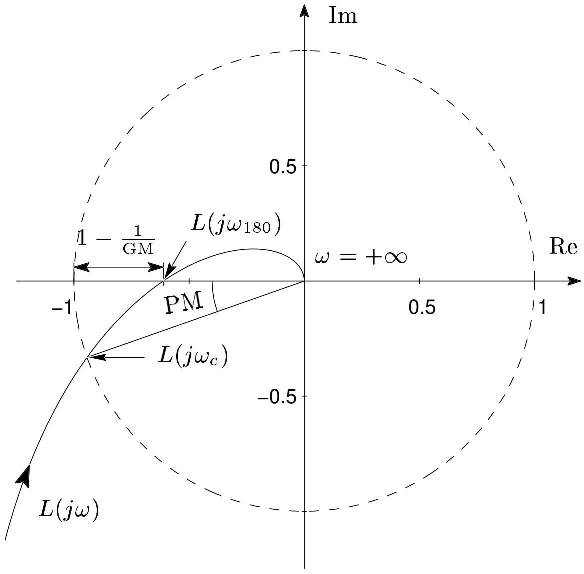


Figure 2.12: Typical Nyquist plot of $L(j\omega)$ for stable plant with PM and GM indicated. Closed-loop instability occurs if $L(j\omega)$ encircles the critical point -1

before the closed-loop system becomes unstable. The GM is thus a direct safeguard against steady-state gain uncertainty (error). Typically we require $\mathrm{GM} > 2$. If the Nyquist plot of L crosses the negative real axis between -1 and $-\infty$ then a *gain reduction margin* can be similarly defined from the smallest value of $|L(j\omega_{180})|$ of such crossings.

The *phase margin* is defined as

$$\mathrm{PM} = \angle L(j\omega_c) + 180^\circ \tag{2.35}$$

where the *gain crossover frequency* ω_c is where $|L(j\omega)|$ first crosses 1 from above, that is

$$|L(j\omega_c)| = 1 \tag{2.36}$$

The phase margin tells how much negative phase (phase lag) we can add to $L(s)$ at frequency ω_c before the phase at this frequency becomes -180° which corresponds to closed-loop instability (see Figure 2.12). Typically, we require PM larger than 30° or more. The PM is a direct safeguard against time delay uncertainty; the system becomes unstable if we add a time delay of

$$\theta_{\max} = \mathrm{PM}/\omega_c \tag{2.37}$$

Note that the units must be consistent, and so if w_c is in [rad/s] then PM must be in radians. It is also important to note that by decreasing the value of ω_c (lowering the closed-loop bandwidth, resulting in a slower response) the system can tolerate larger time delay errors.

Example 2.4 *For the PI-controlled inverse response process example we have* $\mathrm{PM} = 19.4^\circ = 19.4/57.3$ *rad* $= 0.34$ *rad and* $\omega_c = 0.236$ *rad/s. The allowed time delay error is then* $\theta_{\max} = 0.34$ *rad*$/0.236$ *rad/s* $= 1.44$ *s.*

From the above arguments we see that gain and phase margins provide *stability margins* for gain and delay uncertainty. However, as we show below the gain and phase margins are closely related to the peak values of $|S(j\omega)|$ and $|T(j\omega)|$ and are therefore also useful in terms of *performance*. In short, the gain and phase margins are used to provide the appropriate trade-off between performance and stability.

Exercise 2.2 *Prove that the maximum additional delay for which closed-loop stability is maintained is given by (2.37).*

Exercise 2.3 *Derive the approximation for* $K_u = 1/|G(j\omega_u)|$ *given in (5.73) for a first-order delay system.*

Maximum peak criteria

The maximum peaks of the sensitivity and complementary sensitivity functions are defined as

$$M_S = \max_{\omega} |S(j\omega)|; \quad M_T = \max_{\omega} |T(j\omega)| \tag{2.38}$$

(Note that $M_S = \|S\|_\infty$ and $M_T = \|T\|_\infty$ in terms of the $\mathcal{H}_\infty$ norm introduced later.) Typically, it is required that M_S is less than about 2 (6 dB) and M_T is less than about 1.25 (2 dB). A large value of M_S or M_T (larger than about 4) indicates poor performance as well as poor robustness. Since $S + T = 1$ it follows that at any frequency

$$|\,|S| - |T|\,| \leq |S + T| = 1$$

so M_S and M_T differ at most by 1. A large value of M_S therefore occurs if and only if M_T is large. For stable plants we usually have $M_S > M_T$, but this is not a general rule. An upper bound on M_T has been a common design specification in classical control and the reader may be familiar with the use of M-circles on a Nyquist plot or a Nichols chart used to determine M_T from $L(j\omega)$.

We now give some justification for why we may want to bound the value of M_S. Without control ($u = 0$), we have $e = y - r = G_d d - r$, and with feedback control $e = S(G_d d - r)$. Thus, feedback control improves performance in terms of reducing $|e|$ at all frequencies where $|S| < 1$. Usually, $|S|$ is small at low frequencies, for example, $|S(0)| = 0$ for systems with integral action. But because all real systems are strictly proper we must at high frequencies have that $L \to 0$ or equivalently $S \to 1$. At intermediate frequencies one cannot avoid in practice a peak value, M_S, larger than 1 (e.g. see the remark below). Thus, there is an intermediate frequency range where feedback control degrades performance, and the value of M_S is a measure of the worst-case performance degradation. One may also view M_S as a robustness measure, as is now explained. To maintain closed-loop stability the number of encirclements of the critical point -1 by $L(j\omega)$ must not change; so we want L to stay away from this point. The smallest distance between $L(j\omega)$ and the -1 point is M_S^{-1}, and therefore for robustness, the smaller M_S, the better. In summary, both for stability and performance we want M_S close to 1.

There is a close relationship between these maximum peaks and the gain and phase margins. Specifically, for a given M_S we are guaranteed

$$\mathrm{GM} \geq \frac{M_S}{M_S - 1}; \qquad \mathrm{PM} \geq 2 \arcsin\left(\frac{1}{2M_S}\right) \geq \frac{1}{M_S} \text{ [rad]} \tag{2.39}$$

For example, with $M_S = 2$ we are guaranteed $\mathrm{GM} \geq 2$ and $\mathrm{PM} \geq 29.0^\circ$. Similarly, for a given value of M_T we are guaranteed

$$\mathrm{GM} \geq 1 + \frac{1}{M_T}; \qquad \mathrm{PM} \geq 2 \arcsin\left(\frac{1}{2M_T}\right) > \frac{1}{M_T} \text{ [rad]} \tag{2.40}$$

and therefore with $M_T = 2$ we have $\mathrm{GM} \geq 1.5$ and $\mathrm{PM} \geq 29.0^\circ$.

Proof of (2.39) and (2.40): To derive the GM-inequalities notice that $L(j\omega_{180}) = -1/\mathrm{GM}$ (since $\mathrm{GM} = 1/|L(j\omega_{180})|$ and L is real and negative at ω_{180}), from which we get

$$T(j\omega_{180}) = \frac{-1}{\mathrm{GM} - 1}; \quad S(j\omega_{180}) = \frac{1}{1 - \frac{1}{\mathrm{GM}}} \tag{2.41}$$

and the GM-results follow. To derive the PM-inequalities in (2.39) and (2.40) consider Figure 2.13 where we have $|S(j\omega_c)| = 1/|1 + L(j\omega_c)| = 1/|-1 - L(j\omega_c)|$ and we obtain

$$|S(j\omega_c)| = |T(j\omega_c)| = \frac{1}{2\sin(\mathrm{PM}/2)} \tag{2.42}$$

and the inequalities follow. Alternative formulas, which are sometimes used, follow from the identity $2\sin(\mathrm{PM}/2) = \sqrt{2(1 - \cos(\mathrm{PM}))}$. □

Remark. We note with interest that (2.41) requires $|S|$ to be larger than 1 at frequency ω_{180}. This means that provided ω_{180} exists, that is, $L(j\omega)$ has more than -180° phase lag at some frequency (which is the case for any real system), then the peak of $|S(j\omega)|$ must exceed 1.

In conclusion, we see that specifications on the peaks of $|S(j\omega)|$ or $|T(j\omega)|$ (M_S or M_T), can make specifications on the gain and phase margins unnecessary. For instance, requiring $M_S < 2$ implies the common rules of thumb $\mathrm{GM} > 2$ and $\mathrm{PM} > 30^\circ$.

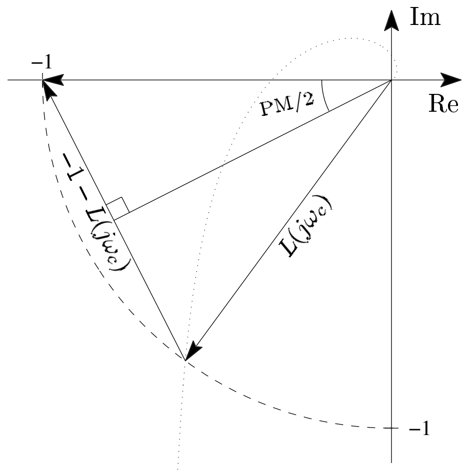


Figure 2.13: At frequency ω_c we see from the figure that $|1 + L(j\omega_c)| = 2\sin(\mathrm{PM}/2)$

2.4.4 Relationship between time and frequency domain peaks

For a change in reference r, the output is $y(s) = T(s)r(s)$. Is there any relationship between the frequency domain peak of $T(j\omega)$, M_T, and any characteristic of the time domain step response, for example the overshoot or the total variation? To answer this consider a prototype second-order system with complementary sensitivity function

$$T(s) = \frac{1}{\tau^2 s^2 + 2\tau\zeta s + 1} \tag{2.43}$$

For underdamped systems with $\zeta < 1$ the poles are complex and yield oscillatory step responses. With $r(t) = 1$ (a unit step change) the values of the overshoot and total variation for $y(t)$ are given, together with M_T and M_S, as a function of ζ in Table 2.1. From Table 2.1, we see that the total variation TV correlates quite well with M_T. This is further confirmed by (A.95) and (2.32) which together yield the following general bounds

$$M_T \ \le \mathrm{TV} \le (2n+1)M_T \tag{2.44}$$

Here n is the order of $T(s)$, which is 2 for our prototype system in (2.43). Given that the response of many systems can be crudely approximated by fairly low-order systems, the bound in (2.44) suggests that M_T may provide a reasonable approximation to the total variation. This provides some justification for the use of M_T in classical control to evaluate the quality of the response.

Table 2.1: Peak values and total variation of prototype second-order system

	Time domain		Frequency domain	
ζ	Overshoot	Total variation	M_T	M_S
2.0	1	1	1	1.05
1.5	1	1	1	1.08
1.0	1	1	1	1.15
0.8	1.02	1.03	1	1.22
0.6	1.09	1.21	1.04	1.35
0.4	1.25	1.68	1.36	1.66
0.2	1.53	3.22	2.55	2.73
0.1	1.73	6.39	5.03	5.12
0.01	1.97	63.7	50.0	50.0

```
% MATLAB code (Mu toolbox) to generate Table:
tau=1;zeta=0.1;t=0:0.01:100;
T = nd2sys(1,[tau*tau 2*tau*zeta 1]); S = msub(1,T);
[A,B,C,D]=unpck(T); y1 = step(A,B,C,D,1,t);
overshoot=max(y1),tv=sum(abs(diff(y1)))
Mt=hinfnorm(T,1.e-4),Ms=hinfnorm(S,1.e-4)
```

2.4.5 Bandwidth and crossover frequency

The concept of bandwidth is very important in understanding the benefits and trade-offs involved when applying feedback control. Above we considered peaks of closed-loop transfer functions, M_S and M_T, which are related to the quality of the response. However, for performance we must also consider the speed of the response, and this leads to considering the bandwidth frequency of the system. In general, a large bandwidth corresponds to a faster rise time, since high frequency signals are more easily passed on to the outputs. A high bandwidth also indicates a system which is sensitive to noise and to parameter variations. Conversely, if the bandwidth is small, the time response will generally be slow, and the system will usually be more robust.

Loosely speaking, *bandwidth* may be defined as the frequency range $[\omega_1, \omega_2]$ over which control is effective. In most cases we require tight control at steady-state so $\omega_1 = 0$, and we then simply call $\omega_2 = \omega_B$ the bandwidth.

The word "effective" may be interpreted in different ways, and this may give rise to different definitions of bandwidth. The interpretation we use is that control is *effective* if we obtain some *benefit* in terms of performance. For tracking performance the error is $e = y - r = -Sr$ and we get that feedback is effective (in terms of improving performance) as long as the relative error $e/r = -S$ is reasonably small, which we may define to be less than 0.707 in magnitude. We then get the following definition:

Definition 2.1 *The (closed-loop) bandwidth, ω_B, is the frequency where $|S(j\omega)|$ first* *crosses* $1/\sqrt{2} = 0.707(\approx -3$ *dB) from below.*

Another interpretation is to say that control is *effective* if it significantly *changes* the output response. For tracking performance, the output is $y = Tr$ and since without control $y = 0$, we may say that control is effective as long as T is reasonably large, which we may define to be larger than 0.707. This leads to an alternative definition which has been traditionally used to define the bandwidth of a control system: *The bandwidth in terms of* T, ω_{BT}, *is the highest frequency at which* $|T(j\omega)|$ *crosses* $1/\sqrt{2} = 0.707(\approx -3$ *dB) from above.*

Remark 1 The definition of bandwidth in terms of ω_{BT} has the advantage of being closer to how the term is used in other fields, for example, in defining the frequency range of an amplifier in an audio system.

Remark 2 In most cases, the two definitions in terms of S and T yield similar values for the bandwidth. In cases where ω_B and ω_{BT} differ, the situation is generally as follows. Up to the frequency ω_B, $|S|$ is less than 0.7, and control is effective in terms of improving performance. In the frequency range $[\omega_B, \omega_{BT}]$ control still affects the response, but does not improve performance — in most cases we find that in this frequency range $|S|$ is larger than 1 and control degrades performance. Finally, at frequencies higher than ω_{BT} we have $S \approx 1$ and control has no significant effect on the response. The situation just described is illustrated in Example 2.5 below (see Figure 2.15).

The *gain crossover frequency*, ω_c, defined as the frequency where $|L(j\omega_c)|$ first crosses 1 from above, is also sometimes used to define closed-loop bandwidth. It has the advantage of being simple to compute and usually gives a value between ω_B and ω_{BT}. Specifically, for systems with PM $< 90^\circ$ we have

$$\omega_B < \omega_c < \omega_{BT} \tag{2.45}$$

Proof of (2.45): Note that $|L(j\omega_c)| = 1$ so $|S(j\omega_c)| = |T(j\omega_c)|$. Thus, when $\mathrm{PM} = 90^\circ$ we get $|S(j\omega_c)| = |T(j\omega_c)| = 0.707$ (see (2.42)), and we have $\omega_B = \omega_c = \omega_{BT}$. For $\mathrm{PM} < 90^\circ$ we get $|S(j\omega_c)| = |T(j\omega_c)| > 0.707$, and since ω_B is the frequency where $|S(j\omega)|$ crosses 0.707 from below we must have $\omega_B < \omega_c$. Similarly, since ω_{BT} is the frequency where $|T(j\omega)|$ crosses 0.707 from above, we must have $\omega_{BT} > \omega_c$. □

Another important frequency is the *phase crossover frequency*, ω_{180}, defined as the first frequency where the Nyquist curve of $L(j\omega)$ crosses the negative real axis between -1 and 0.

Remark. From (2.41) we get that $\omega_{180} > \omega_{BT}$ for $\mathrm{GM} > 2.414$, and $\omega_{180} < \omega_{BT}$ for $\mathrm{GM} < 2.414$, and since in many cases the gain margin is about 2.4 we conclude that ω_{180} is usually close to ω_{BT}. It is also interesting to note from (2.41) that at ω_{180} the phase of T (and of L) is -180°, so from $y = Tr$ we conclude that at frequency ω_{180} the tracking response is completely out of phase. Since as just noted ω_{BT} is often close to ω_{180}, this further illustrates that ω_{BT} may be a poor indicator of closed-loop performance.

Example 2.5 Comparison of ω_B and ω_{BT} as indicators of performance. *An example where ω_{BT} is a poor indicator of performance is the following:*

$$L = \frac{-s+z}{s(\tau s + \tau z + 2)}; \quad T = \frac{-s+z}{s+z}\frac{1}{\tau s + 1}; \quad z = 0.1, \ \tau = 1 \tag{2.46}$$

For this system, both L and T have a RHP-zero at $z = 0.1$, and we have $\mathrm{GM} = 2.1$, $\mathrm{PM} = 60.1°$, $M_S = 1.93$ and $M_T = 1$. We find that $\omega_B = 0.036$ and $\omega_c = 0.054$ are both less than $z = 0.1$ (as one should expect because speed of response is limited by the presence of RHP-zeros), whereas $\omega_{BT} = 1/\tau = 1.0$ is ten times larger than z. The closed-loop response to a unit step change in the reference is shown in Figure 2.14. The rise time is 31.0 s, which is close to $1/\omega_B = 28.0$ s, but very different from $1/\omega_{BT} = 1.0$ s, illustrating that ω_B is a better indicator of closed-loop performance than ω_{BT}.

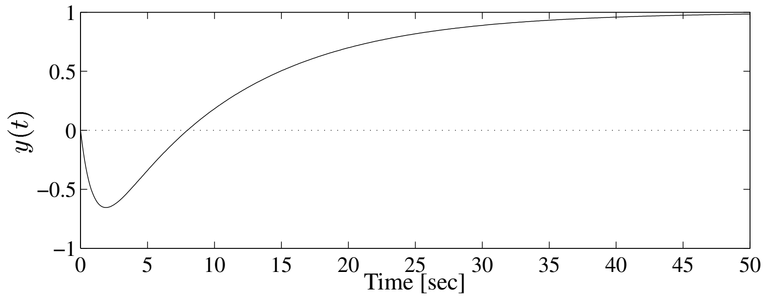


Figure 2.14: Step response for system $T = \frac{-s+0.1}{s+0.1}\frac{1}{s+1}$

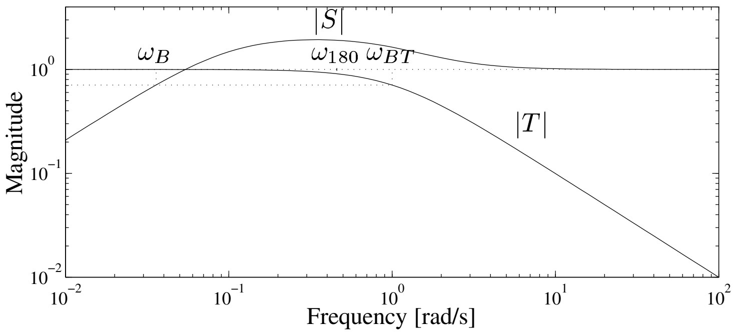


Figure 2.15: Plots of $|S|$ and $|T|$ for system $T = \frac{-s+0.1}{s+0.1}\frac{1}{s+1}$

The magnitude Bode plots of S and T are shown in Figure 2.15. We see that $|T| \approx 1$ up to about ω_{BT}. However, in the frequency range from ω_B to ω_{BT} the phase of T (not shown) drops from about $-40°$ to about $-220°$, so in practice tracking is poor in this frequency range. For *example, at frequency $\omega_{180} = 0.46$ we have $T \approx -0.9$, and the response to a sinusoidally varying reference $r(t) = \sin\omega_{180}t$ is completely out of phase, i.e. $y(t) \approx -0.9r(t)$.*

In conclusion, ω_B (which is defined in terms of $|S|$) and also ω_c (in terms of $|L|$) are good indicators of closed-loop performance, while ω_{BT} (in terms of $|T|$) may be misleading in some cases. The reason is that we want $T \approx 1$ in order to have good performance, and it is not sufficient that $|T| \approx 1$; we must also consider its phase. On the other hand, for for good performance we want S close to 0, and this will be the case if $|S| \approx 0$ irrespective of the phase of S.

2.5 Controller design

We have considered ways of evaluating performance, but one also needs methods for controller design. The Ziegler-Nichols' method used earlier is well suited for on-line tuning, but most other methods involve minimizing some cost function. The overall design process is iterative between controller design and performance (or cost) evaluation. If performance is not satisfactory then one must either adjust the controller parameters directly (for example, by reducing K_c from the value obtained by the Ziegler-Nichols' rules) or adjust some weighting factor in an objective function used to synthesize the controller.

There exist a large number of methods for controller design and some of these will be discussed in Chapter 9. In addition to heuristic rules and on-line tuning we can distinguish between three main approaches to controller design:

1. **Shaping of transfer functions.** In this approach the designer specifies the *magnitude* of some transfer function(s) as a function of frequency, and then finds a controller which gives the desired shape(s).
   (a) **Loop shaping.** This is the classical approach in which the magnitude of the open-loop transfer function, $L(j\omega)$, is shaped. Usually no optimization is involved and the designer aims to obtain $|L(j\omega)|$ with desired bandwidth, slopes etc. We will look at this approach in detail later in this chapter. However, classical loop shaping is difficult to apply for complicated systems, and one may then instead use the Glover-McFarlane $\mathcal{H}_\infty$ loop-shaping design presented in Chapter 9. The method consists of a second step where optimization is used to make an initial loop-shaping design more robust.
   (b) **Shaping of closed-loop transfer functions, such as S, T and KS.** Optimization is usually used, resulting in various $\mathcal{H}_\infty$ optimal control problems such as mixed weighted sensitivity; more on this later.
2. **The signal-based approach.** This involves time domain problem formulations resulting in the minimization of a norm of a transfer function. Here one considers a particular disturbance or reference change and then one tries to optimize the closed-loop response. The "modern" state-space methods from the 1960's, such

as Linear Quadratic Gaussian (LQG) control, are based on this signal-oriented approach. In LQG the input signals are assumed to be stochastic (or alternatively impulses in a deterministic setting) and the expected value of the output variance (or the 2-norm) is minimized. These methods may be generalized to include frequency dependent weights on the signals leading to what is called the Wiener-Hopf (or $\mathcal{H}_2$-norm) design method.

By considering sinusoidal signals, frequency-by-frequency, a signal-based $\mathcal{H}_\infty$ optimal control methodology can be derived in which the $\mathcal{H}_\infty$ norm of a combination of closed-loop transfer functions is minimized. This approach has attracted significant interest, and may be combined with model uncertainty representations, to yield quite complex robust performance problems requiring μ-synthesis; an important topic which will be addressed in later chapters.

3. **Numerical optimization.** This often involves multi-objective optimization where one attempts to optimize directly the true objectives, such as rise times, stability margins, etc. Computationally, such optimization problems may be difficult to solve, especially if one does not have convexity in the controller parameters. Also, by effectively including performance evaluation and controller design in a single step procedure, the problem formulation is far more critical than in iterative two-step approaches. The numerical optimization approach may also be performed on-line, which might be useful when dealing with cases with constraints on the inputs and outputs. On-line optimization approaches such as model predictive control are likely to become more popular as faster computers and more efficient and reliable computational algorithms are developed.

2.6 Loop shaping

In the classical loop-shaping approach to controller design, "loop shape" refers to the magnitude of the loop transfer function $L = GK$ as a function of frequency. An understanding of how K can be selected to shape this loop gain provides invaluable insight into the multivariable techniques and concepts which will be presented later in the book, and so we will discuss loop shaping in some detail in the next two sections.

2.6.1 Trade-offs in terms of L

Recall equation (2.19), which yields the closed-loop response in terms of the control error $e = y - r$:

$$e = -\underbrace{(I+L)^{-1}}_{S} r + \underbrace{(I+L)^{-1}}_{S} G_d d - \underbrace{(I+L)^{-1}L}_{T} n \tag{2.47}$$

For "perfect control" we want $e = y - r = 0$; that is, we would like

$$e \approx 0 \cdot d + 0 \cdot r + 0 \cdot n$$

The first two requirements in this equation, namely disturbance rejection and command tracking, are obtained with $S \approx 0$, or equivalently, $T \approx I$. Since $S = (I + L)^{-1}$, this implies that the loop transfer function L must be large in magnitude. On the other hand, the requirement for zero noise transmission implies that $T \approx 0$, or equivalently, $S \approx I$, which is obtained with $L \approx 0$. This illustrates the fundamental nature of feedback design which always involves a trade-off between conflicting objectives; in this case between large loop gains for disturbance rejection and tracking, and small loop gains to reduce the effect of noise.

It is also important to consider the magnitude of the control action u (which is the input to the plant). We want u small because this causes less wear and saves input energy, and also because u is often a disturbance to other parts of the system (e.g. consider opening a window in your office to adjust your comfort and the undesirable disturbance this will impose on the air conditioning system for the building). In particular, we usually want to avoid fast changes in u. The control action is given by $u = K(r - y_m)$ and we find as expected that a small u corresponds to small controller gains and a small $L = GK$.

The most important design objectives which necessitate trade-offs in feedback control are summarized below:

1. Performance, good disturbance rejection: needs large controller gains, i.e. L large.
2. Performance, good command following: L large.
3. Stabilization of unstable plant: L large.
4. Mitigation of measurement noise on plant outputs: L small.
5. Small magnitude of input signals: K small and L small.
6. Physical controller must be strictly proper: $K \to 0$ and $L \to 0$ at high frequencies.
7. Nominal stability (stable plant): L small (because of RHP-zeros and time delays).
8. Robust stability (stable plant): L small (because of uncertain or neglected dynamics).

Fortunately, the conflicting design objectives mentioned above are generally in different frequency ranges, and we can meet most of the objectives by using a large loop gain ($|L| > 1$) at low frequencies below crossover, and a small gain ($|L| < 1$) at high frequencies above crossover.

2.6.2 Fundamentals of loop-shaping design

By *loop shaping* we mean a design procedure that involves explicitly shaping the magnitude of the loop transfer function, $|L(j\omega)|$. Here $L(s) = G(s)K(s)$ where $K(s)$ is the feedback controller to be designed and $G(s)$ is the product of all other transfer functions around the loop, including the plant, the actuator and the measurement device. Essentially, to get the benefits of feedback control we want the loop gain, $|L(j\omega)|$, to be as large as possible within the bandwidth region. However, due to time delays, RHP-zeros, unmodelled high-frequency dynamics and limitations on the allowed manipulated inputs, the loop gain has to drop below one at and above some frequency which we call the crossover frequency ω_c. Thus, disregarding stability for the moment, it is desirable that $|L(j\omega)|$ falls sharply with frequency. To measure how $|L|$ falls with frequency we consider the logarithmic slope $N = d\ln|L|/d\ln\omega$. For example, a slope $N = -1$ implies that $|L|$ drops by a factor of 10 when ω increases by a factor of 10. If the gain is measured in decibels (dB) then a slope of $N = -1$ corresponds to -20 dB/ decade. The value of $-N$ at high frequencies is often called the *roll-off* rate.

The design of $L(s)$ is most crucial and difficult in the crossover region between ω_c (where $|L| = 1$) and ω_{180} (where $\angle L = -180°$). For stability, we at least need the loop gain to be less than 1 at frequency ω_{180}, i.e. $|L(j\omega_{180})| < 1$. Thus, to get a high bandwidth (fast response) we want ω_c and therefore ω_{180} large, that is, we want the phase lag in L to be small. Unfortunately, this is not consistent with the desire that $|L(j\omega)|$ should fall sharply. For example, the loop transfer function $L = 1/s^n$ (which has a slope $N = -n$ on a log-log plot) has a phase $\angle L = -n \cdot 90°$. Thus, to have a phase margin of $45°$ we need $\angle L > -135°$, and the slope of $|L|$ cannot exceed $N = -1.5$.

In addition, if the slope is made steeper at lower or higher frequencies, then this will add unwanted phase lag at intermediate frequencies. As an example, consider $L_1(s)$ given in (2.13) with the Bode plot shown in Figure 2.3. Here the slope of the asymptote of $|L|$ is -1 at the gain crossover frequency (where $|L_1(j\omega_c)| = 1$), which by itself gives $-90°$ phase lag. However, due to the influence of the steeper slopes of -2 at lower and higher frequencies, there is a "penalty" of about $-35°$ at crossover, so the actual phase of L_1 at ω_c is approximately $-125°$.

The situation becomes even worse for cases with delays or RHP-zeros in $L(s)$ which add undesirable phase lag to L without contributing to a desirable negative slope in L. At the gain crossover frequency ω_c, the additional phase lag from delays and RHP-zeros may in practice be $-30°$ or more.

In summary, a desired loop shape for $|L(j\omega)|$ typically has a slope of about -1 in the crossover region, and a slope of -2 or higher beyond this frequency, that is, the roll-off is 2 or larger. Also, with a proper controller, which is required for any real system, we must have that $L = GK$ rolls off at least as fast as G. At low frequencies, the desired shape of $|L|$ depends on what disturbances and references we are designing for. For example, if we are considering step changes in the references or disturbances which affect the outputs as steps, then a slope for $|L|$ of -1 at low frequencies is acceptable. If the references or disturbances require the outputs to change in a ramp-like fashion then a slope of -2 is required. In practice, integrators are included in the controller to get the desired low-frequency performance, and for offset-free reference tracking the rule is that

- $L(s)$ *must contain at least one integrator for each integrator in* $r(s)$.

Proof: Let $L(s) = \widehat{L}(s)/s^{n_I}$ where $\widehat{L}(0)$ is non-zero and finite and n_I is the number of integrators in $L(s)$ — sometimes n_I is called the *system type*. Consider a reference signal of the form $r(s) = 1/s^{n_r}$. For example, if $r(t)$ is a unit step, then $r(s) = 1/s$ ($n_r = 1$), and if

$r(t)$ is a ramp then $r(s) = 1/s^2$ ($n_r = 2$). The final value theorem for Laplace transforms is

$$\lim_{t\to\infty} e(t) = \lim_{s\to 0} se(s) \tag{2.48}$$

In our case, the control error is

$$e(s) = -\frac{1}{1+L(s)}r(s) = -\frac{s^{n_I-n_r}}{s^{n_I}+\widehat{L}(s)} \tag{2.49}$$

and to get zero offset (i.e. $e(t\to\infty)=0$) we must from (2.48) require $n_I \geq n_r$, and the rule follows. □

In conclusion, one can define the desired loop transfer function in terms of the following specifications:

1. The gain crossover frequency, ω_c, where $|L(j\omega_c)| = 1$.
2. The shape of $L(j\omega)$, e.g. in terms of the slope of $|L(j\omega)|$ in certain frequency ranges. Typically, we desire a slope of about $N = -1$ around crossover, and a larger roll-off at higher frequencies. The desired slope at lower frequencies depends on the nature of the disturbance or reference signal.
3. The system type, defined as the number of pure integrators in $L(s)$.

In Section 2.6.4, we discuss how to specify the loop shape when disturbance rejection is the primary objective of control. Loop-shaping design is typically an iterative procedure where the designer shapes and reshapes $|L(j\omega)|$ after computing the phase and gain margins, the peaks of closed-loop frequency responses (M_T and M_S), selected closed-loop time responses, the magnitude of the input signal, etc. The procedure is illustrated next by an example.

Example 2.6 Loop-shaping design for the inverse response process.
We will now design a loop-shaping controller for the example process in (2.26) which has a RHP-zero at $s = 0.5$. The RHP-zero limits the achievable bandwidth and so the crossover region (defined as the frequencies between ω_c and ω_{180}) will be at about 0.5 rad/s. We require the system to have one integrator (type 1 system), and therefore a reasonable approach is to let the loop transfer function have a slope of -1 at low frequencies, and then to roll off with a higher slope at frequencies beyond 0.5 rad/s. The plant and our choice for the loop-shape is

$$G(s) = \frac{3(-2s+1)}{(5s+1)(10s+1)}; \quad L(s) = 3K_c\frac{(-2s+1)}{s(2s+1)(0.33s+1)} \tag{2.50}$$

The frequency response (Bode plots) of L is shown in Figure 2.16 for $K_c = 0.05$. The controller gain K_c was selected to get a reasonable stability margins (PM and GM). The asymptotic slope of $|L|$ is -1 up to 3 rad/s where it changes to -2. The controller corresponding to the loop-shape in (2.50) is

$$K(s) = K_c\frac{(10s+1)(5s+1)}{s(2s+1)(0.33s+1)}, \quad K_c = 0.05 \tag{2.51}$$

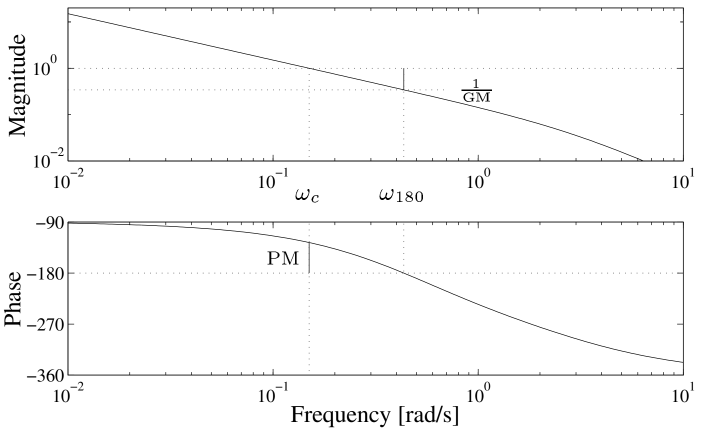


Figure 2.16: Frequency response of $L(s)$ in (2.50) for loop-shaping design with $K_c = 0.05$ ($\mathrm{GM} = 2.92$, $\mathrm{PM} = 54°$, $\omega_c = 0.15$, $\omega_{180} = 0.43$, $M_S = 1.75$, $M_T = 1.11$)

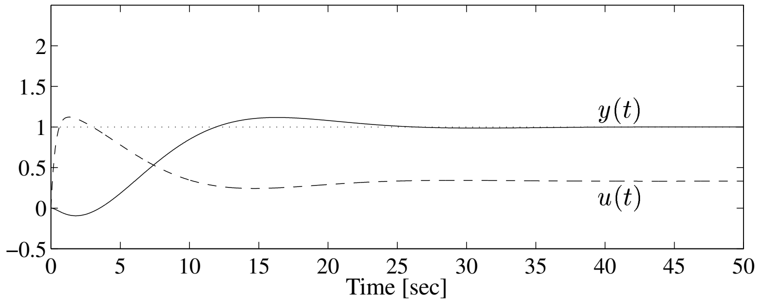


Figure 2.17: Response to step in reference for loop-shaping design

The controller has zeros at the locations of the plant poles. This is desired in this case because we do not want the slope of the loop shape to drop at the break frequencies $1/10 = 0.1$ *rad/s and* $1/5 = 0.2$ *rad/s just before crossover. The phase of* L *is* $-90°$ *at low frequency, and at* $\omega = 0.5$ *rad/s the additional contribution from the term* $\frac{-2s+1}{2s+1}$ *in (2.50) is* $-90°$*, so for stability we need* $\omega_c < 0.5$ *rad/s. The choice* $K_c = 0.05$ *yields* $\omega_c = 0.15$ *rad/s corresponding to* $\mathrm{GM} = 2.92$ *and* PM=$54°$*. The corresponding time response is shown in Figure 2.17. It is seen to be much better than the responses with either the simple PI-controller in Figure 2.7 or with the P-controller in Figure 2.5. Figure 2.17 also shows that the magnitude of the input signal remains less than about* 1 *in magnitude. This means that the controller gain is not too large at high frequencies. The magnitude Bode plot for the controller (2.51) is shown in Figure 2.18. It is interesting to note that in the crossover region around* $\omega = 0.5$ *rad/s the controller gain is quite constant, around* 1 *in magnitude, which is similar to the "best" gain found using a P-controller (see Figure 2.5).*

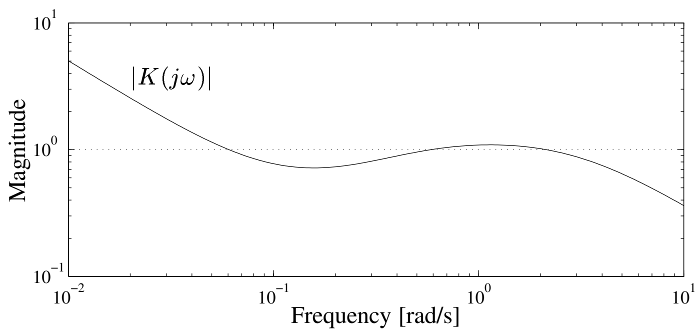


Figure 2.18: Magnitude Bode plot of controller (2.51) for loop-shaping design

Limitations imposed by RHP-zeros and time delays

Based on the above loop-shaping arguments we can now examine how the presence of delays and RHP-zeros limit the achievable control performance. We have already argued that if we want the loop shape to have a slope of -1 around crossover (ω_c), with preferably a steeper slope before and after crossover, then the phase lag of L at ω_c will necessarily be at least $-90°$, even when there are no RHP-zeros or delays. Therefore, if we assume that for performance and robustness we want a phase margin of about $35°$ or more, then the additional phase contribution from any delays and RHP-zeros at frequency ω_c cannot exceed about $-55°$.

First consider a time delay θ. It yields an additional phase contribution of $-\theta\omega$, which at frequency $\omega = 1/\theta$ is -1 rad = $-57°$ (which is more than $-55°$). Thus, for acceptable control performance we need $\omega_c < 1/\theta$, approximately.

Next consider a real RHP-zero at $s = z$. To avoid an increase in slope caused by this zero we place a pole at $s = -z$ such that the loop transfer function contains the term $\frac{-s+z}{s+z}$, the form of which is referred to as all-pass since its magnitude equals 1 at all frequencies. The phase contribution from the all-pass term at $\omega = z/2$ is $-2\arctan(0.5) = -53°$ (which is close to $-55°$), so for acceptable control performance we need $\omega_c < z/2$, approximately.

2.6.3 Inverse-based controller design

In Example 2.6, we made sure that $L(s)$ contained the RHP-zero of $G(s)$, but otherwise the specified $L(s)$ was independent of $G(s)$. This suggests the following possible approach for a minimum-phase plant (i.e. one with no RHP-zeros or time delays). Select a loop shape which has a slope of -1 throughout the frequency range, namely

$$L(s) = \frac{\omega_c}{s} \tag{2.52}$$

where ω_c is the desired gain crossover frequency. This loop shape yields a phase margin of $90°$ and an infinite gain margin since the phase of $L(j\omega)$ never reaches $-180°$. The controller corresponding to (2.52) is

$$K(s) = \frac{\omega_c}{s} G^{-1}(s) \tag{2.53}$$

That is, the controller inverts the plant and adds an integrator ($1/s$). This is an old idea, and is also the essential part of the internal model control (IMC) design procedure (Morari and Zafiriou, 1989) which has proved successful in many applications. However, there are at least two good reasons for why this inverse-based controller may not be a good choice:

1. The controller will not be realizable if $G(s)$ has a pole excess of two or larger, and may in any case yield large input signals. These problems may be partly fixed by adding high-frequency dynamics to the controller.
2. The loop shape resulting from (2.52) and (2.53) is *not* generally desirable, unless the references and disturbances affect the outputs as steps. This is illustrated by the following example.

Example 2.7 **Disturbance process.** *We now introduce our second SISO example control problem in which disturbance rejection is an important objective in addition to command tracking. We assume that the plant has been appropriately scaled as outlined in Section 1.4.*

Problem formulation. *Consider the disturbance process described by*

$$\boxed{G(s) = \frac{200}{10s+1}\frac{1}{(0.05s+1)^2}, \quad G_d(s) = \frac{100}{10s+1}} \tag{2.54}$$

with time in seconds (a block diagram is shown in Figure 2.20). The control objectives are:

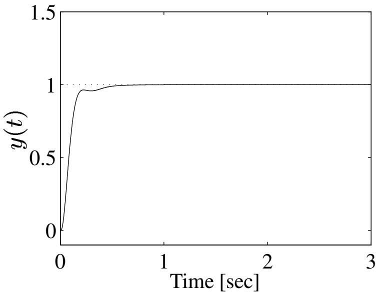
1. *Command tracking: The rise time (to reach 90% of the final value) should be less than 0.3 s and the overshoot should be less than 5%.*
2. *Disturbance rejection: The output in response to a unit step disturbance should remain within the range $[-1, 1]$ at all times, and it should return to 0 as quickly as possible ($|y(t)|$ should at least be less than 0.1 after 3 s).*
3. *Input constraints: $u(t)$ should remain within the range $[-1, 1]$ at all times to avoid input saturation (this is easily satisfied for most designs).*

Analysis. *Since $G_d(0) = 100$ we have that without control the output response to a unit disturbance ($d = 1$) will be 100 times larger than what is deemed to be acceptable. The magnitude $|G_d(j\omega)|$ is lower at higher frequencies, but it remains larger than 1 up to $\omega_d \approx 10$ rad/s (where $|G_d(j\omega_d)| = 1$). Thus, feedback control is needed up to frequency ω_d, so we need ω_c to be approximately equal to 10 rad/s for disturbance rejection. On the other hand, we do not want ω_c to be larger than necessary because of sensitivity to noise and stability problems associated with high gain feedback. We will thus aim at a design with $\underline{\omega_c \approx 10 \text{ rad/s}}$.*

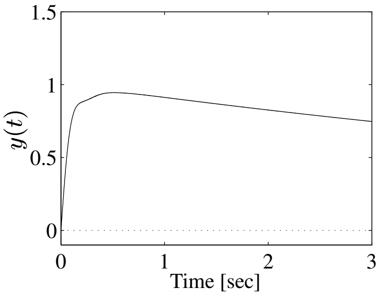
Inverse-based controller design. *We will consider the inverse-based design as given by (2.52) and (2.53) with $\omega_c = 10$. Since $G(s)$ has a pole excess of three this yields an unrealizable controller, and therefore we choose to approximate the plant term $(0.05s+1)^2$*

by $(0.1s+1)$ and then in the controller we let this term be effective over one decade, i.e. we use $(0.1s+1)/(0.01s+1)$ to give the realizable design

$$K_0(s) = \frac{\omega_c}{s}\frac{10s+1}{200}\frac{0.1s+1}{0.01s+1}, \; L_0(s) = \frac{\omega_c}{s}\frac{0.1s+1}{(0.05s+1)^2(0.01s+1)}, \; \omega_c = 10 \quad (2.55)$$



(a) Tracking response



(b) Disturbance response

Figure 2.19: Responses with "inverse-based" controller $K_0(s)$ for the disturbance process

The response to a step reference is excellent as shown in Figure 2.19(a). The rise time is about 0.16 s and there is no overshoot so the specifications are more than satisfied. However, the response to a step disturbance (Figure 2.19(b)) is much too sluggish. Although the output stays within the range $[-1,1]$, it is still 0.75 at $t = 3$ s (whereas it should be less than 0.1). Because of the integral action the output does eventually return to zero, but it does not drop below 0.1 until after 23 s.

The above example illustrates that the simple inverse-based design method where L has a slope of about $N = -1$ at all frequencies, does not always yield satisfactory designs. In the example, reference tracking was excellent, but disturbance rejection was poor. The objective of the next section is to understand why the disturbance response was so poor, and to propose a more desirable loop shape for disturbance rejection.

2.6.4 Loop shaping for disturbance rejection

At the outset we assume that the disturbance has been scaled such that at each frequency $|d(\omega)| \leq 1$, and the main control objective is to achieve $|e(\omega)| < 1$. With feedback control we have $e = y = SG_d d$, so to achieve $|e(\omega)| \leq 1$ for $|d(\omega)| = 1$ (the worst-case disturbance) we require $|SG_d(j\omega)| < 1, \forall\omega$, or equivalently,

$$|1+L| \geq |G_d| \quad \forall\omega \qquad (2.56)$$

At frequencies where $|G_d| > 1$, this is approximately the same as requiring $|L| > |G_d|$. However, in order to minimize the input signals, thereby reducing the sensitivity to noise and avoiding stability problems, we do not want to use larger loop gains than necessary (at least at frequencies around crossover). A reasonable initial loop shape $L_{\min}(s)$ is then one that just satisfies the condition

$$|L_{\min}| \approx |G_d| \tag{2.57}$$

where the subscript min signifies that $L_{\min}$ is the smallest loop gain to satisfy $|e(\omega)| \leq 1$. Since $L = GK$ the corresponding controller with the minimum gain satisfies

$$|K_{\min}| \approx |G^{-1}G_d| \tag{2.58}$$

In addition, to improve low-frequency performance (e.g. to get zero steady-state offset), we often add integral action at low frequencies, and use

$$|K| = |\frac{s + \omega_I}{s}||G^{-1}G_d| \tag{2.59}$$

This can be summarized as follows:

- For disturbance rejection a good choice for the controller is one which contains the dynamics (G_d) of the disturbance and inverts the dynamics (G) of the inputs (at least at frequencies just before crossover).
- For disturbances entering directly at the plant output, $G_d = 1$, we get $|K_{\min}| = |G^{-1}|$, so an inverse-based design provides the best trade-off between performance (disturbance rejection) and minimum use of feedback.
- For disturbances entering directly at the plant input (which is a common situation in practice – often referred to as a load disturbance), we have $G_d = G$ and we get $|K_{\min}| = 1$, so a simple proportional controller with unit gain yields a good trade-off between output performance and input usage.
- Notice that a reference change may be viewed as a disturbance directly affecting the output. This follows from (1.18), from which we get that a maximum reference change $r = R$ may be viewed as a disturbance $d = 1$ with $G_d(s) = -R$ where R is usually a constant. This explains why selecting K to be like G^{-1} (an inverse-based controller) yields good responses to step changes in the reference.

In addition to satisfying $|L| \approx |G_d|$ (eq. 2.57) at frequencies around crossover, the desired loop-shape $L(s)$ may be modified as follows:

1. Around crossover make the slope N of $|L|$ to be about -1. This is to achieve good transient behaviour with acceptable gain and phase margins.
2. Increase the loop gain at low frequencies as illustrated in (2.59) to improve the settling time and to reduce the steady-state offset. Adding an integrator yields zero steady-state offset to a step disturbance.
3. Let $L(s)$ roll off faster at higher frequencies (beyond the bandwidth) in order to reduce the use of manipulated inputs, to make the controller realizable and to reduce the effects of noise.

The above requirements are concerned with the magnitude, $|L(j\omega)|$. In addition, the dynamics (phase) of $L(s)$ must be selected such that the closed-loop system is stable. When selecting $L(s)$ to satisfy $|L| \approx |G_d|$ one should replace $G_d(s)$ by the corresponding minimum-phase transfer function with the same magnitude, that is, time delays and RHP-zeros in $G_d(s)$ should not be included in $L(s)$ as this will impose undesirable limitations on feedback. On the other hand, any time delays or RHP-zeros in $G(s)$ must be included in $L = GK$ because RHP pole-zero cancellations between $G(s)$ and $K(s)$ yield internal instability; see Chapter 4.

Remark. The idea of including a disturbance model in the controller is well known and is more rigorously presented in, for example, research on the internal model principle (Wonham, 1974), or the internal model control design for disturbances (Morari and Zafiriou, 1989). However, our development is simple, and sufficient for gaining the insight needed for later chapters.

Example 2.8 Loop-shaping design for the disturbance process. *Consider again the plant described by (2.54). The plant can be represented by the block diagram in Figure 2.20, and we see that the disturbance enters at the plant input in the sense that G and G_d share the same dominating dynamics as represented by the term* $200/(10s + 1)$.

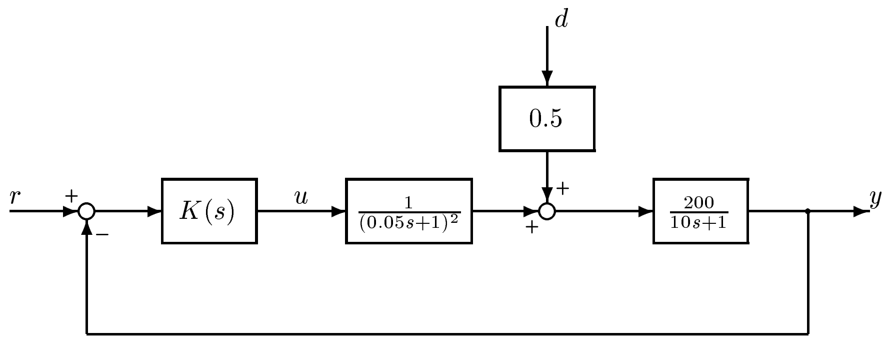
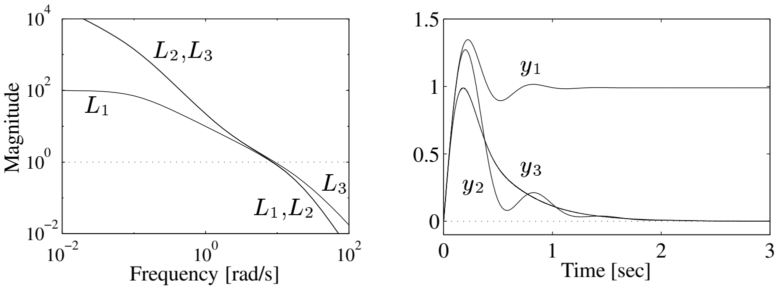


Figure 2.20: Block diagram representation of the disturbance process in (2.54)

Step 1. Initial design. *From (2.57) we know that a good initial loop shape looks like* $|L_{\min}| = |G_d| = \left|\frac{100}{10s+1}\right|$ *at frequencies up to crossover. The corresponding controller is* $K(s) = G^{-1}L_{\min} = 0.5(0.05s + 1)^2$. *This controller is not proper (i.e. it has more zeros than poles), but since the term* $(0.05s+1)^2$ *only comes into effect at* $1/0.05 = 20$ *rad/s, which is beyond the desired gain crossover frequency* $\omega_c = 10$ *rad/s, we may replace it by a constant gain of* 1 *resulting in a proportional controller*

$$K_1(s) = 0.5 \tag{2.60}$$

The magnitude of the corresponding loop transfer function, $|L_1(j\omega)|$, *and the response* ($y_1(t)$) *to a step change in the disturbance are shown in Figure 2.21. This simple controller works* *surprisingly well, and for* $t < 3$ *s the response to a step change in the disturbance is not much different from that with the more complicated inverse-based controller* $K_0(s)$ *of (2.55) as shown earlier in Figure 2.19. However, there is no integral action and* $y_1(t) \to 1$ *as* $t \to \infty$.



(a) Loop gains (b) Disturbance responses

Figure 2.21: Loop shapes and disturbance responses for controllers K_1, K_2 and K_3 for the disturbance process

Step 2. More gain at low frequency. *To get integral action we multiply the controller by the term* $\frac{s+\omega_I}{s}$, *see (2.59), where* ω_I *is the frequency up to which the term is effective (the asymptotic value of the term is 1 for* $\omega > \omega_I$*). For performance we want large gains at low frequencies, so we want* ω_I *to be large, but in order to maintain an acceptable phase margin (which is* 44.7° *for controller* K_1*) the term should not add too much negative phase at frequency* ω_c, *so* ω_I *should not be too large. A reasonable value is* $\omega_I = 0.2\omega_c$ *for which the phase contribution from* $\frac{s+\omega_I}{s}$ *is* $\arctan(1/0.2) - 90^\circ = -11^\circ$ *at* ω_c. *In our case* $\omega_c \approx 10$ *rad/s, so we select the following controller*

$$K_2(s) = 0.5\frac{s+2}{s} \tag{2.61}$$

The resulting disturbance response (y_2*) shown in Figure 2.21(b) satisfies the requirement that* $|y(t)| < 0.1$ *at time* $t = 3$ *s, but* $y(t)$ *exceeds* 1 *for a short time. Also, the response is slightly oscillatory as might be expected since the phase margin is only* 31° *and the peak values for* $|S|$ *and* $|T|$ *are* $M_S = 2.28$ *and* $M_T = 1.89$.

Step 3. High-frequency correction. *To increase the phase margin and improve the transient response we supplement the controller with "derivative action" by multiplying* $K_2(s)$ *by a lead-lag term which is effective over one decade starting at* 20 *rad/s:*

$$K_3(s) = 0.5\frac{s+2}{s}\,\frac{0.05s+1}{0.005s+1} \tag{2.62}$$

This gives a phase margin of 51°, *and peak values* $M_S = 1.43$ *and* $M_T = 1.23$. *From Figure 2.21(b), it is seen that the controller* $K_3(s)$ *reacts quicker than* $K_2(s)$ *and the disturbance response* $y_3(t)$ *stays below* 1.

Table 2.2 summarizes the results for the four loop-shaping designs; the inverse-based design K_0 *for reference tracking and the three designs* K_1, K_2 *and* K_3 *for disturbance rejection. Although controller* K_3 *satisfies the requirements for disturbance rejection, it is* not

Table 2.2: Alternative loop-shaping designs for the disturbance process

	GM	PM	ω_c	M_S	M_T	Reference t_r	Reference $y_{\max}$	Disturbance $y_{\max}$	Disturbance $y(t=3)$
Spec.→			≈ 10			$\leq .3$	≤ 1.05	≤ 1	≤ 0.1
K_0	9.95	72.9°	11.4	1.34	1	0.16	1.00	0.95	0.75
K_1	4.04	44.7°	8.48	1.83	1.33	0.21	1.24	1.35	0.99
K_2	3.24	30.9°	8.65	2.28	1.89	0.19	1.51	1.27	0.001
K_3	19.7	50.9°	9.27	1.43	1.23	0.16	1.24	0.99	0.001

satisfactory for reference tracking; the overshoot is 24% *which is significantly higher than the maximum value of* 5%. *On the other hand, the inverse-based controller* K_0 *inverts the term* $1/(10s+1)$ *which is also in the disturbance model, and therefore yields a very sluggish response to disturbances (the output is still* 0.75 *at* $t=3$ *s whereas it should be less than* 0.1*).*

In summary, for this process none of the controller designs meet all the objectives for both reference tracking and disturbance rejection. The solution is to use a two degrees-of-freedom controller as is discussed next.

2.6.5 Two degrees-of-freedom design

For reference tracking we typically want the controller to look like $\frac{1}{s}G^{-1}$, see (2.53), whereas for disturbance rejection we want the controller to look like $\frac{1}{s}G^{-1}G_d$, see (2.59). We cannot achieve both of these simultaneously with a single (feedback) controller.

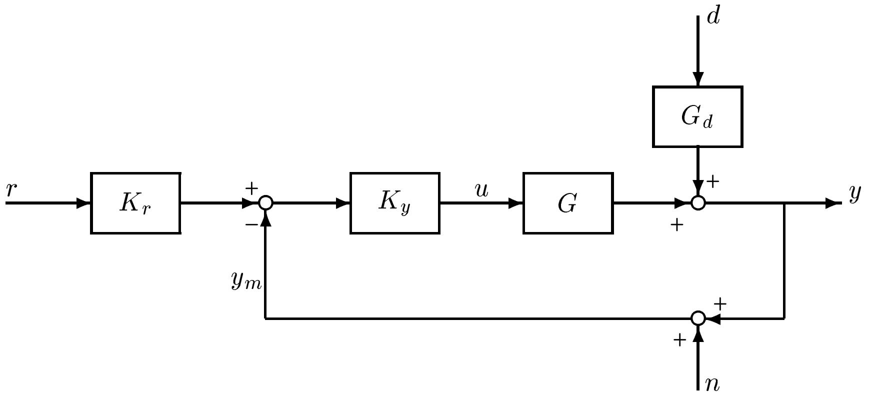


Figure 2.22: Two degrees-of-freedom controller

The solution is to use a two degrees-of-freedom controller where the reference signal r and output measurement y_m are independently treated by the controller, rather than operating on their difference $r - y_m$. There exist several alternative implementations of a two degrees-of-freedom controller. The most general form is shown in Figure 1.3(b) on page 12 where the controller has two inputs (r and y_m) and one output (u). However, the controller is often split into two separate blocks as shown in Figure 2.22 where K_y denotes the feedback part of the controller and K_r a reference prefilter. The feedback controller K_y is used to reduce the effect of uncertainty (disturbances and model error) whereas the prefilter K_r shapes the commands r to improve tracking performance. In general, it is optimal to design the combined two degrees-of-freedom controller K in one step. However, in practice K_y is often designed first for disturbance rejection, and then K_r is designed to improve reference tracking. This is the approach taken here.

Let $T = L(1+L)^{-1}$ (with $L = GK_y$) denote the complementary sensitivity function for the feedback system. Then for a one degree-of-freedom controller $y = Tr$, whereas for a two degrees-of-freedom controller $y = TK_r r$. If the desired transfer function for reference tracking (often denoted the reference model) is $T_{\rm ref}$, then the corresponding ideal reference prefilter K_r satisfies $TK_r = T_{\rm ref}$, or

$$K_r(s) = T^{-1}(s)T_{\rm ref}(s) \tag{2.63}$$

Thus, in theory we may design $K_r(s)$ to get any desired tracking response $T_{\rm ref}(s)$. However, in practice it is not so simple because the resulting $K_r(s)$ may be unstable (if $G(s)$ has RHP-zeros) or unrealizable, and also $TK_r \neq T_{\rm ref}$ if $T(s)$ is not known exactly.

Remark. A convenient practical choice of prefilter is the lead-lag network

$$K_r(s) = \frac{\tau_{\rm lead}s + 1}{\tau_{\rm lag}s + 1} \tag{2.64}$$

Here we select $\tau_{\rm lead} > \tau_{\rm lag}$ if we want to speed up the response, and $\tau_{\rm lead} < \tau_{\rm lag}$ if we want to slow down the response. If one does not require fast reference tracking, which is the case in many process control applications, a simple lag is often used (with $\tau_{\rm lead} = 0$).

Example 2.9 Two degrees-of-freedom design for the disturbance process. *In Example 2.8 we designed a loop-shaping controller $K_3(s)$ for the plant in (2.54) which gave good performance with respect to disturbances. However, the command tracking performance was not quite acceptable as is shown by y_3 in Figure 2.23. The rise time is 0.16 s which is better than the required value of 0.3s, but the overshoot is 24% which is significantly higher than the maximum value of 5%. To improve upon this we can use a two degrees-of-freedom controller with $K_y = K_3$, and we design $K_r(s)$ based on (2.63) with reference model $T_{\rm ref} = 1/(0.1s+1)$ (a first-order response with no overshoot). To get a low-order $K_r(s)$, we may either use the actual $T(s)$ and then use a low-order approximation of $K_r(s)$, or we may start with a low-order approximation of $T(s)$. We will do the latter. From the step response y_3 in Figure 2.23 we approximate the response by two parts; a fast response with time constant 0.1 s and gain 1.5, and a slower response with time constant 0.5 s and gain -0.5 (the sum of the gains is 1). Thus we use $T(s) \approx \frac{1.5}{0.1s+1} - \frac{0.5}{0.5s+1} = \frac{(0.7s+1)}{(0.1s+1)(0.5s+1)}$, from which (2.63) yields*

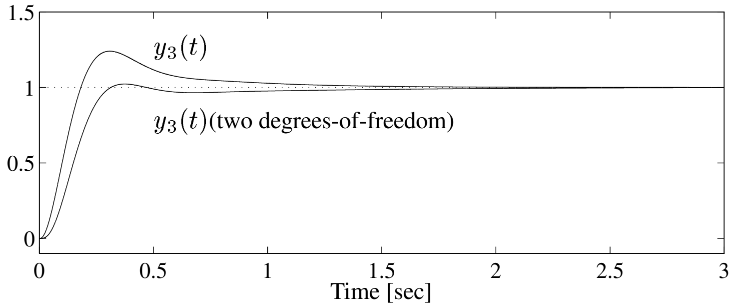


Figure 2.23: Tracking responses with the one degree-of-freedom controller (K_3) and the two degrees-of-freedom controller (K_3, K_{r3}) for the disturbance process

$K_r(s) = \frac{0.5s+1}{0.7s+1}$. *Following closed-loop simulations we modified this slightly to arrive at the design*

$$K_{r3}(s) = \frac{0.5s+1}{0.65s+1} \cdot \frac{1}{0.03s+1} \tag{2.65}$$

where the term $1/(0.03s + 1)$ *was included to avoid the initial peaking of the input signal* $u(t)$ *above* 1. *The tracking response with this two degrees-of-freedom controller is shown in Figure 2.23. The rise time is* 0.25 *s which is better than the requirement of* 0.3 *s, and the overshoot is only* 2.3% *which is better than the requirement of* 5%. *The disturbance response is the same as curve* y_3 *in Figure 2.21. In conclusion, we are able to satisfy all specifications using a two degrees-of-freedom controller.*

Loop shaping applied to a flexible structure

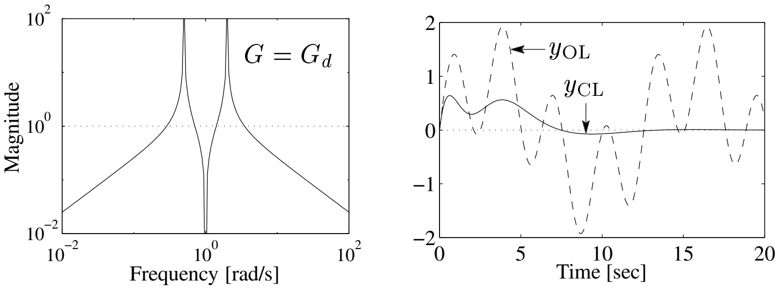
The following example shows how the loop-shaping procedure for disturbance rejection, can be used to design a one degree-of-freedom controller for a very different kind of plant.

Example 2.10 Loop shaping for a flexible structure. *Consider the following model of a flexible structure with a disturbance occurring at the plant input*

$$G(s) = G_d(s) = \frac{2.5s(s^2+1)}{(s^2+0.5^2)(s^2+2^2)} \tag{2.66}$$

From the Bode magnitude plot in Figure 2.24(a) we see that $|G_d(j\omega)| \gg 1$ *around the resonance frequencies of* 0.5 *and* 2 *rad/s, so control is needed at these frequencies. The dashed line in Figure 2.24(b) shows the open-loop response to a unit step disturbance. The output is seen to cycle between* -2 *and* 2 *(outside the allowed range* -1 *to* 1*), which confirms that control is needed. From (2.58) a controller which meets the specification* $|y(\omega)| \leq 1$ *for* $|d(\omega)| = 1$ *is given by* $|K_{\min}(j\omega)| = |G^{-1}G_d| = 1$. *Indeed the controller*

$$K(s) = 1 \tag{2.67}$$



(a) Magnitude plot of $|G| = |G_d|$

(b) Open-loop and closed-loop disturbance responses with $K = 1$

Figure 2.24: Flexible structure in (2.66)

turns out to be a good choice as is verified by the closed-loop disturbance response (solid line) in Figure 2.24(b); the output goes up to about 0.5 *and then returns to zero. The fact that the choice* $L(s) = G(s)$ *gives closed-loop stability is not immediately obvious since* $|G|$ *has* 4 *gain crossover frequencies. However, instability cannot occur because the plant is "passive" with* $\angle G > -180^\circ$ *at all frequencies.*

2.6.6 Conclusions on loop shaping

The loop-shaping procedure outlined and illustrated by the examples above is well suited for relatively simple problems, as might arise for stable plants where $L(s)$ crosses the negative real axis only once. Although the procedure may be extended to more complicated systems the effort required by the engineer is considerably greater. In particular, it may be very difficult to achieve stability.

Fortunately, there exist alternative methods where the burden on the engineer is much less. One such approach is the Glover-McFarlane $\mathcal{H}_\infty$ loop-shaping procedure which is discussed in detail in Chapter 9. It is essentially a two-step procedure, where in the first step the engineer, as outlined in this section, decides on a loop shape, $|L|$ (denoted the "shaped plant" G_s), and in the second step an optimization provides the necessary phase corrections to get a stable and robust design. The method is applied to the disturbance process in Example 9.3 on page 381.

Another design philosophy which deals directly with shaping both the gain and phase of $L(s)$ is the quantitative feedback theory (QFT) of Horowitz (1991).

2.7 Shaping closed-loop transfer functions

In this section, we introduce the reader to the shaping of the magnitudes of closed-loop transfer functions, where we synthesize a controller by minimizing an $\mathcal{H}_\infty$

performance objective. The topic is discussed further in Section 3.4.6 and in more detail in Chapter 9.

Specifications directly on the *open-loop transfer function* $L = GK$, as in the loop-shaping design procedures of the previous section, make the design process transparent as it is clear how changes in $L(s)$ affect the controller $K(s)$ and *vice versa*. An apparent problem with this approach, however, is that it does not consider directly the *closed-loop transfer functions*, such as S and T, which determine the final response. The following approximations apply

$$\begin{aligned} |L(j\omega)| \gg 1 \quad &\Rightarrow \quad S \approx L^{-1}; \quad T \approx 1 \\ |L(j\omega)| \ll 1 \quad &\Rightarrow \quad S \approx 1; \qquad\; T \approx L \end{aligned}$$

but in the crossover region where $|L(j\omega)|$ is close to 1, one cannot infer anything about S and T from the magnitude of the loop shape, $|L(j\omega)|$. For example, $|S|$ and $|T|$ may experience large peaks if $L(j\omega)$ is close to -1, i.e. the phase of $L(j\omega)$ is crucial in this frequency range.

An alternative design strategy is to directly shape the magnitudes of closed-loop transfer functions, such as $S(s)$ and $T(s)$. Such a design strategy can be formulated as an $\mathcal{H}_\infty$ optimal control problem, thus automating the actual controller design and leaving the engineer with the task of selecting reasonable bounds ("weights") on the desired closed-loop transfer functions. Before explaining how this may be done in practice, we discuss the terms $\mathcal{H}_\infty$ and $\mathcal{H}_2$.

2.7.1 The terms $\mathcal{H}_\infty$ and $\mathcal{H}_2$

The $\mathcal{H}_\infty$ norm of a stable scalar transfer function $f(s)$ is simply the peak value of $|f(j\omega)|$ as a function of frequency, that is,

$$\|f(s)\|_\infty \triangleq \max_\omega |f(j\omega)| \tag{2.68}$$

Remark. Strictly speaking, we should here replace "max" (the maximum value) by "sup" (the supremum, the least upper bound). This is because the maximum may only be approached as $w \to \infty$ and may therefore not actually be achieved. However, for engineering purposes there is no difference between "sup" and "max".

The terms $\mathcal{H}_\infty$ norm and $\mathcal{H}_\infty$ control are intimidating at first, and a name conveying the engineering significance of $\mathcal{H}_\infty$ would have been better. After all, we are simply talking about a design method which aims to press down the peak(s) of one or more selected transfer functions. However, the term $\mathcal{H}_\infty$, which is purely mathematical, has now established itself in the control community. To make the term less forbidding, an explanation of its background may help. First, the symbol ∞ comes from the fact that the maximum magnitude over frequency may be written as

$$\max_\omega |f(j\omega)| = \lim_{p\to\infty} \left(\int_{-\infty}^{\infty} |f(j\omega)|^p d\omega \right)^{1/p}$$

Essentially, by raising $|f|$ to an infinite power we pick out its peak value. Next, the symbol $\mathcal{H}$ stands for "Hardy space", and $\mathcal{H}_\infty$ in the context of this book is the set of transfer functions with bounded ∞-norm, which is simply the set of *stable and proper* transfer functions.

Similarly, the symbol $\mathcal{H}_2$ stands for the Hardy space of transfer functions with bounded 2-norm, which is the set of *stable and strictly proper* transfer functions. The $\mathcal{H}_2$ norm of a strictly proper stable scalar transfer function is defined as

$$\|f(s)\|_2 \triangleq \left(\frac{1}{2\pi}\int_{-\infty}^{\infty} |f(j\omega)|^2 d\omega\right)^{1/2} \tag{2.69}$$

The factor $1/\sqrt{2\pi}$ is introduced to get consistency with the 2-norm of the corresponding impulse response; see (4.117). Note that the $\mathcal{H}_2$ norm of a semi-proper (or bi-proper) transfer function (where $\lim_{s\to\infty} f(s)$ is a non-zero constant) is infinite, whereas its $\mathcal{H}_\infty$ norm is finite. An example of a semi-proper transfer function (with an infinite $\mathcal{H}_2$ norm) is the sensitivity function $S = (I + GK)^{-1}$.

2.7.2 Weighted sensitivity

As already discussed, the sensitivity function S is a very good indicator of closed-loop performance, both for SISO and MIMO systems. The main advantage of considering S is that because we ideally want S small, it is sufficient to consider just its magnitude $|S|$; that is, we need not worry about its phase. Typical specifications in terms of S include:

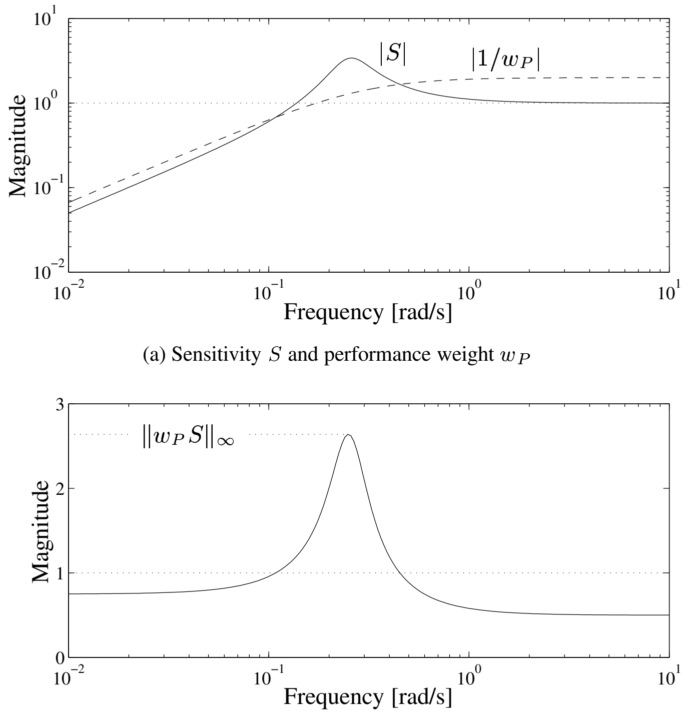
1. Minimum bandwidth frequency ω_B^* (defined as the frequency where $|S(j\omega)|$ crosses 0.707 from below).
2. Maximum tracking error at selected frequencies.
3. System type, or alternatively the maximum steady-state tracking error, A.
4. Shape of S over selected frequency ranges.
5. Maximum peak magnitude of S, $\|S(j\omega)\|_\infty \leq M$.

The peak specification prevents amplification of noise at high frequencies, and also introduces a margin of robustness; typically we select $M = 2$. Mathematically, these specifications may be captured by an upper bound, $1/|w_P(s)|$, on the magnitude of S, where $w_P(s)$ is a weight selected by the designer. The subscript P stands for *performance* since S is mainly used as a performance indicator, and the performance requirement becomes

$$|S(j\omega)| < 1/|w_P(j\omega)|,\ \forall\omega \tag{2.70}$$

$$\Leftrightarrow \quad |w_P S| < 1,\ \forall\omega \quad \Leftrightarrow \quad \boxed{\|w_P S\|_\infty < 1} \tag{2.71}$$

The last equivalence follows from the definition of the $\mathcal{H}_\infty$ norm, and in words the performance requirement is that the $\mathcal{H}_\infty$ norm of the weighted sensitivity, $w_P S$,



(a) Sensitivity S and performance weight w_P

(b) Weighted sensitivity $w_P S$

Figure 2.25: Case where $|S|$ exceeds its bound $1/|w_P|$, resulting in $\|w_P S\|_\infty > 1$

must be less than one. In Figure 2.25(a), an example is shown where the sensitivity, $|S|$, exceeds its upper bound, $1/|w_P|$, at some frequencies. The resulting weighted sensitivity, $|w_P S|$ therefore exceeds 1 at the same frequencies as is illustrated in Figure 2.25(b). Note that we usually do not use a log-scale for the magnitude when plotting weighted transfer functions, such as $|w_P S|$.

Weight selection. An asymptotic plot of a typical upper bound, $1/|w_P|$, is shown in Figure 2.26. The weight illustrated may be represented by

$$\boxed{w_P(s) = \frac{s/M + \omega_B^*}{s + \omega_B^* A}} \tag{2.72}$$

and we see that $1/|w_P(j\omega)|$ (the upper bound on $|S|$) is equal to $A \leq 1$ at low frequencies, is equal to $M \geq 1$ at high frequencies, and the asymptote crosses 1 at the frequency ω_B^*, which is approximately the bandwidth requirement.

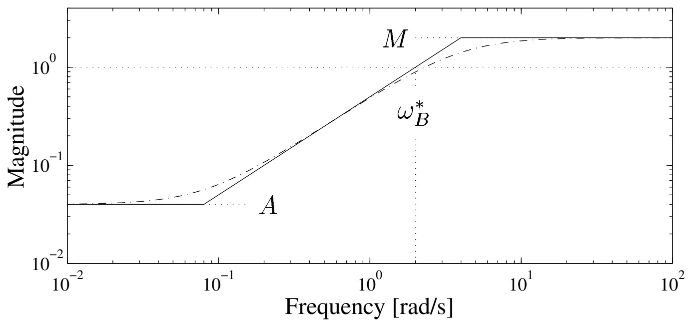


Figure 2.26: Inverse of performance weight. Exact and asymptotic plot of $1/|w_P(j\omega)|$ in (2.72)

Remark. For this weight the loop shape $L = \omega_B^*/s$ yields an S which exactly matches the bound (2.71) at frequencies below the bandwidth and easily satisfies (by a factor M) the bound at higher frequencies. This L has a slope in the frequency range below crossover of $N = -1$.

In some cases, in order to improve performance, we may want a steeper slope for L (and S) below the bandwidth, and then a higher-order weight may be selected. A weight which asks for a slope of -2 for L in a range of frequencies below crossover is

$$\boxed{w_P(s) = \frac{(s/M^{1/2} + \omega_B^*)^2}{(s + \omega_B^* A^{1/2})^2}} \tag{2.73}$$

The insights gained in the previous section on loop-shaping design are very useful for selecting weights. For example, for disturbance rejection we must satisfy $|SG_d(j\omega)| < 1$ at all frequencies (assuming the variables have been scaled to be less than 1 in magnitude). It then follows that a good initial choice for the performance weight is to let $|w_P(j\omega)|$ look like $|G_d(j\omega)|$ at frequencies where $|G_d| > 1$.

Exercise 2.4 *Make an asymptotic plot of* $1/|w_P|$ *in (2.73) and compare with the asymptotic plot of* $1/|w_P|$ *in (2.72).*

2.7.3 Stacked requirements: mixed sensitivity

The specification $\|w_P S\|_\infty < 1$ puts a lower bound on the bandwidth, but not an upper one, and nor does it allow us to specify the roll-off of $L(s)$ above the bandwidth. To do this one can make demands on another closed-loop transfer function, for example, on the complementary sensitivity $T = I - S = GKS$. For instance, one might specify an upper bound $1/|w_T|$ on the magnitude of T to make sure that L rolls off sufficiently fast at high frequencies. Also, to achieve robustness or to restrict the magnitude of the input signals, $u = KS(r - G_d d)$, one may place an upper bound, $1/|w_u|$, on the magnitude of KS. To combine these "mixed sensitivity" specifications, a "stacking approach" is usually used, resulting in the following overall specification:

$$\|N\|_\infty = \max_\omega \bar{\sigma}(N(j\omega)) < 1; \quad N = \begin{bmatrix} w_P S \\ w_T T \\ w_u KS \end{bmatrix} \tag{2.74}$$

We here use the maximum singular value, $\bar{\sigma}(N(j\omega))$, to measure the size of the matrix N at each frequency. For SISO systems, N is a vector and $\bar{\sigma}(N)$ is the usual Euclidean vector norm:

$$\bar{\sigma}(N) = \sqrt{|w_P S|^2 + |w_T T|^2 + |w_u KS|^2} \tag{2.75}$$

After selecting the form of N and the weights, the $\mathcal{H}_\infty$ optimal controller is obtained by solving the problem

$$\min_K \|N(K)\|_\infty \tag{2.76}$$

where K is a stabilizing controller. A good tutorial introduction to $\mathcal{H}_\infty$ control is given by Kwakernaak (1993).

Remark 1 The stacking procedure is selected for mathematical convenience as it does not allow us to exactly specify the bounds on the individual transfer functions as described above. For example, assume that $\phi_1(K)$ and $\phi_2(K)$ are two functions of K (which might represent $\phi_1(K) = w_P S$ and $\phi_2(K) = w_T T$) and that we want to achieve

$$|\phi_1| < 1 \quad \text{and} \quad |\phi_2| < 1 \tag{2.77}$$

This is similar to, but not quite the same as the stacked requirement

$$\bar{\sigma}\begin{bmatrix} \phi_1 \\ \phi_2 \end{bmatrix} = \sqrt{|\phi_1|^2 + |\phi_2|^2} < 1 \tag{2.78}$$

Objectives (2.77) and (2.78) are very similar when either $|\phi_1|$ or $|\phi_2|$ is small, but in the worst case when $|\phi_1| = |\phi_2|$, we get from (2.78) that $|\phi_1| \leq 0.707$ and $|\phi_2| \leq 0.707$. That is, there is a possible "error" in each specification equal to at most a factor $\sqrt{2} \approx 3$ dB. In general, with n stacked requirements the resulting error is at most $\sqrt{n}$. This inaccuracy in the specifications is something we are probably willing to sacrifice in the interests of mathematical convenience. In any case, the specifications are in general rather rough, and are effectively knobs for the engineer to select and adjust until a satisfactory design is reached.

Remark 2 Let $\gamma_0 = \min_K \|N(K)\|_\infty$ denote the optimal $\mathcal{H}_\infty$ norm. An important property of $\mathcal{H}_\infty$ optimal controllers is that they yield a flat frequency response, that is, $\bar{\sigma}(N(j\omega)) = \gamma_0$ at all frequencies. The practical implication is that, except for at most a factor $\sqrt{n}$, the transfer functions resulting from a solution to (2.76) will be close to γ_0 times the bounds selected by the designer. This gives the designer a mechanism for directly shaping the magnitudes of $\bar{\sigma}(S)$, $\bar{\sigma}(T)$, $\bar{\sigma}(KS)$, and so on.

Example 2.11 $\mathcal{H}_\infty$ **mixed sensitivity design for the disturbance process.** *Consider again the plant in (2.54), and consider an $\mathcal{H}_\infty$ mixed sensitivity S/KS design in which*

$$N = \begin{bmatrix} w_P S \\ w_u KS \end{bmatrix} \tag{2.79}$$

Appropriate scaling of the plant has been performed so that the inputs should be about 1 or less in magnitude, and we therefore select a simple input weight $w_u = 1$. The performance weight is chosen, in the form of (2.72), as

$$w_{P1}(s) = \frac{s/M + \omega_B^*}{s + \omega_B^* A}; \quad M = 1.5, \; \omega_B^* = 10, \quad A = 10^{-4} \tag{2.80}$$

A value of $A = 0$ would ask for integral action in the controller, but to get a stable weight and to prevent numerical problems in the algorithm used to synthesize the controller, we have moved the integrator slightly by using a small non-zero value for A. This has no practical significance in terms of control performance. The value $\omega_B^* = 10$ has been selected to achieve approximately the desired crossover frequency ω_c of 10 rad/s. The $\mathcal{H}_\infty$ problem is solved with the μ-toolbox in MATLAB using the commands in Table 2.3.

Table 2.3: **MATLAB program to synthesize an $\mathcal{H}_\infty$ controller**

```
% Uses the Mu-toolbox
G=nd2sys(1,conv([10 1],conv([0.05 1],[0.05 1])),200);            % Plant is G.
M=1.5; wb=10; A=1.e-4; Wp = nd2sys([1/M wb], [1 wb*A]); Wu = 1;   % Weights.
%
% Generalized plant P is found with function sysic:
% (see Section 3.8 for more details)
%
systemnames = 'G Wp Wu';
inputvar = '[ r(1); u(1)]';
outputvar = '[Wp; Wu; r-G]';
input_to_G = '[u]';
input_to_Wp = '[r-G]';
input_to_Wu = '[u]';
sysoutname = 'P';
cleanupsysic = 'yes';
sysic;
%
% Find H-infinity optimal controller:
%
nmeas=1; nu=1; gmn=0.5; gmx=20; tol=0.001;
[khinf,ghinf,gopt] = hinfsyn(P,nmeas,nu,gmn,gmx,tol);
```

For this problem, we achieved an optimal $\mathcal{H}_\infty$ norm of 1.37, so the weighted sensitivity requirements are not quite satisfied (see design 1 in Figure 2.27). Nevertheless, the design seems good with $\|S\|_\infty = M_S = 1.30$, $\|T\|_\infty = M_T = 1.0$, $\mathrm{GM} = 8.04$, $\mathrm{PM} = 71.2^\circ$ and $\omega_c = 7.22$ rad/s, and the tracking response is very good as shown by curve y_1 in Figure 2.28(a). The design is actually very similar to the loop-shaping design for references, K_0, which was an inverse-based controller.

However, we see from curve y_1 in Figure 2.28(b) that the disturbance response is very sluggish. If disturbance rejection is the main concern, then from our earlier discussion in Section 2.6.4 this motivates the need for a performance weight that specifies higher gains at

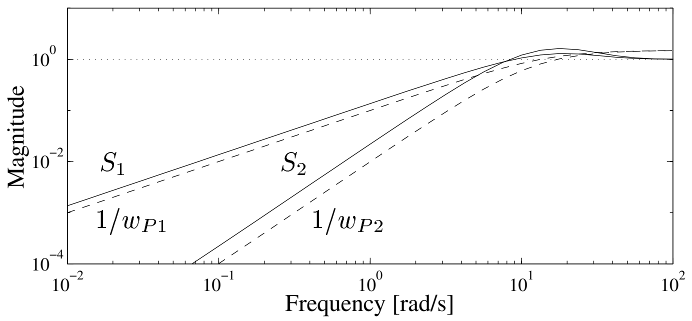
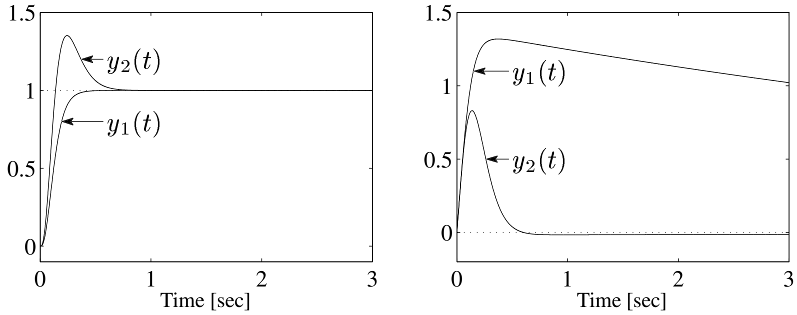


Figure 2.27: Inverse of performance weight (dashed line) and resulting sensitivity function (solid line) for two $\mathcal{H}_\infty$ designs (1 and 2) for the disturbance process



(a) Tracking response (b) Disturbance response

Figure 2.28: Closed-loop step responses for two alternative $\mathcal{H}_\infty$ designs (1 and 2) for the disturbance process

low frequencies. We therefore try

$$w_{P2}(s) = \frac{(s/M^{1/2} + \omega_B^*)^2}{(s + \omega_B^* A^{1/2})^2}, \quad M = 1.5, \omega_B^* = 10, A = 10^{-4} \tag{2.81}$$

The inverse of this weight is shown in Figure 2.27, and is seen from the dashed line to cross 1 *in magnitude at about the same frequency as weight* w_{P1}*, but it specifies tighter control at lower frequencies. With the weight* w_{P2}*, we get a design with an optimal* $\mathcal{H}_\infty$ *norm of* 2.21*, yielding* $M_S = 1.63$*,* $M_T = 1.43$*,* $\mathrm{GM} = 4.76$*,* $\mathrm{PM} = 43.3^\circ$ *and* $\omega_c = 11.34$ *rad/s. The design is actually very similar to the loop-shaping design for disturbances,* K_3*. The disturbance response is very good, whereas the tracking response has a somewhat high overshoot; see curve* y_2 *in Figure 2.28(a).*

In conclusion, design 1 *is best for reference tracking whereas design* 2 *is best for disturbance rejection. To get a design with both good tracking and good disturbance rejection we need a* *two degrees-of-freedom controller, as was discussed in Example 2.9.*

2.8 Conclusion

The main purpose of this chapter has been to present the classical ideas and techniques of feedback control. We have concentrated on SISO systems so that insights into the necessary design trade-offs, and the design approaches available, can be properly developed before MIMO systems are considered. We also introduced the $\mathcal{H}_\infty$ problem based on weighted sensitivity, for which typical performance weights are given in (2.72) and (2.73).

3

INTRODUCTION TO MULTIVARIABLE CONTROL

In this chapter, we introduce the reader to multi-input multi-output (MIMO) systems. We discuss the singular value decomposition (SVD), multivariable control, and multivariable right-half plane (RHP) zeros. The need for a careful analysis of the effect of uncertainty in MIMO systems is motivated by two examples. Finally we describe a general control configuration that can be used to formulate control problems. Many of these important topics are considered again in greater detail later in the book. The chapter should be accessible to readers who have attended a classical SISO control course.

3.1 Introduction

We consider a multi-input multi-output (MIMO) plant with m inputs and l outputs. Thus, the basic transfer function model is $y(s) = G(s)u(s)$, where y is an $l \times 1$ vector, u is an $m \times 1$ vector and $G(s)$ is an $l \times m$ transfer function matrix.

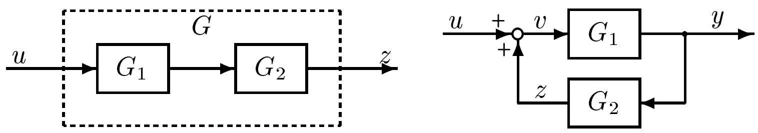
If we make a change in the first input, u_1, then this will generally affect all the outputs, $y_1, y_2, \ldots, y_l$, that is, there is *interaction* between the inputs and outputs. A non-interacting plant would result if u_1 only affects y_1, u_2 only affects y_2, and so on.

The main difference between a scalar (SISO) system and a MIMO system is the presence of *directions* in the latter. Directions are relevant for vectors and matrices, but not for scalars. However, despite the complicating factor of directions, most of the ideas and techniques presented in the previous chapter on SISO systems may be extended to MIMO systems. The singular value decomposition (SVD) provides a useful way of quantifying multivariable directionality, and we will see that most SISO results involving the absolute value (magnitude) may be generalized to multivariable systems by considering the maximum singular value. An exception to this is Bode's stability condition which has no generalization in terms of singular values. This is related to the fact that it is difficult to find a good measure of phase for MIMO transfer functions.

The chapter is organized as follows. We start by presenting some rules for determining multivariable transfer functions from block diagrams. Although most of the formulas for scalar systems apply, we must exercise some care since matrix multiplication is not commutative, that is, in general $GK \neq KG$. Then we introduce the singular value decomposition and show how it may be used to study directions in multivariable systems. We also give a brief introduction to multivariable control and decoupling. We then consider a simple plant with a multivariable RHP-zero and show how the effect of this zero may be shifted from one output channel to another. After this we discuss robustness, and study two example plants, each 2×2, which demonstrate that the simple gain and phase margins used for SISO systems do not generalize easily to MIMO systems. Finally, we consider a general control problem formulation.

At this point, you may find it useful to browse through Appendix A where some important mathematical tools are described. Exercises to test your understanding of this mathematics are given at the end of this chapter.

3.2 Transfer functions for MIMO systems



(a) Cascade system (b) Positive feedback system

Figure 3.1: Block diagrams for the cascade rule and the feedback rule

The following three rules are useful when evaluating transfer functions for MIMO systems.

1. **Cascade rule.** *For the cascade (series) interconnection of G_1 and G_2 in Figure 3.1(a), the overall transfer function matrix is $G = G_2G_1$.*

Remark. The order of the transfer function matrices in $G = G_2G_1$ (from left to right) is the reverse of the order in which they appear in the block diagram of Figure 3.1(a) (from left to right). This has led some authors to use block diagrams in which the inputs enter at the right hand side. However, in this case the order of the transfer function blocks in a feedback path will be reversed compared with their order in the formula, so no fundamental benefit is obtained.

2. **Feedback rule.** *With reference to the positive feedback system in Figure 3.1(b), we have $v = (I - L)^{-1}u$ where $L = G_2G_1$ is the transfer function around the loop.*

3. **Push-through rule.** *For matrices of appropriate dimensions*

$$G_1(I - G_2G_1)^{-1} = (I - G_1G_2)^{-1}G_1 \tag{3.1}$$

Proof: Equation (3.1) is verified by pre-multiplying both sides by $(I - G_1G_2)$ and post-multiplying both sides by $(I - G_2G_1)$. □

Exercise 3.1 *Derive the cascade and feedback rules.*

The cascade and feedback rules can be combined into the following MIMO rule for evaluating closed-loop transfer functions from block diagrams.

> **MIMO Rule:** *Start from the output and write down the blocks as you meet them when moving backwards (against the signal flow), taking the most direct path towards the input. If you exit from a feedback loop then include a term $(I-L)^{-1}$ for positive feedback (or $(I + L)^{-1}$ for negative feedback) where L is the transfer function around that loop (evaluated against the signal flow starting at the point of exit from the loop).*

Care should be taken when applying this rule to systems with nested loops. For such systems it is probably safer to write down the signal equations and eliminate internal variables to get the transfer function of interest. The rule is best understood by considering an example.

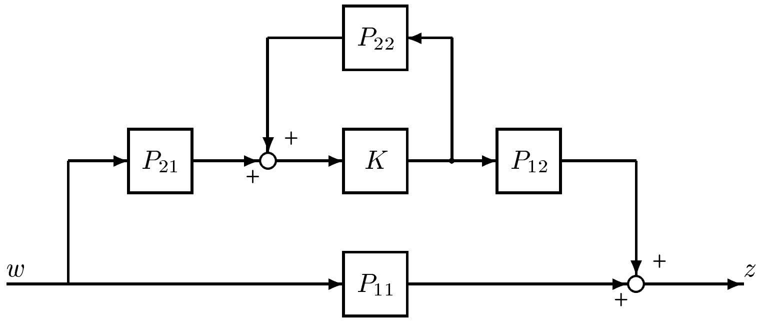


Figure 3.2: Block diagram corresponding to (3.2)

Example 3.1 *The transfer function for the block diagram in Figure 3.2 is given by*

$$z = (P_{11} + P_{12}K(I - P_{22}K)^{-1}P_{21})w \tag{3.2}$$

To derive this from the MIMO rule above we start at the output z and move backwards towards w. There are two branches, one of which gives the term P_{11} directly. In the other branch we move backwards and meet P_{12} and then K. We then exit from a feedback loop and get a term $(I - L)^{-1}$ (positive feedback) with $L = P_{22}K$, and finally we meet P_{21}.

Exercise 3.2 *Use the MIMO rule to derive the transfer functions from u to y and from u to z in Figure 3.1(b). Use the push-through rule to rewrite the two transfer functions.*

Exercise 3.3 *Use the MIMO rule to show that (2.18) corresponds to the negative feedback system in Figure 2.4.*

Negative feedback control systems

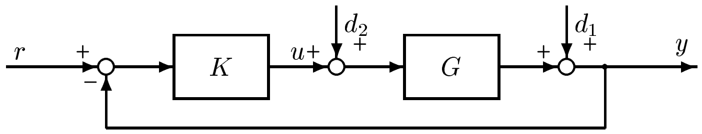


Figure 3.3: Conventional negative feedback control system

For the negative feedback system in Figure 3.3, we define L to be the loop transfer function as seen when breaking the loop at the *output* of the plant. Thus, for the case where the loop consists of a plant G and a feedback controller K we have

$$L = GK \tag{3.3}$$

The sensitivity and complementary sensitivity are then defined as

$$S \triangleq (I + L)^{-1}; \quad T \triangleq I - S = L(I + L)^{-1} \tag{3.4}$$

In Figure 3.3, T is the transfer function from r to y, and S is the transfer function from d_1 to y; also see equations (2.16) to (2.20) which apply to MIMO systems.

S and T are sometimes called the *output sensitivity* and *output complementary sensitivity*, respectively, and to make this explicit one may use the notation $L_O \equiv L$, $S_O \equiv S$ and $T_O \equiv T$. This is to distinguish them from the corresponding transfer functions evaluated at the *input* to the plant.

We define L_I to be the loop transfer function as seen when breaking the loop at the *input* to the plant with negative feedback assumed. In Figure 3.3

$$L_I = KG \tag{3.5}$$

The *input* sensitivity and *input* complementary sensitivity functions are then defined as

$$S_I \triangleq (I + L_I)^{-1}; \quad T_I \triangleq I - S_I = L_I(I + L_I)^{-1} \tag{3.6}$$

In Figure 3.3, $-T_I$ is the transfer function from d_2 to u. Of course, for SISO systems $L_I = L$, $S_I = S$, and $T_I = T$.

Exercise 3.4 *In Figure 3.3, what transfer function does S_I represent? Evaluate the transfer functions from d_1 and d_2 to $r - y$.*

The following relationships are useful:

$$(I+L)^{-1} + L(I+L)^{-1} = S + T = I \tag{3.7}$$

$$G(I+KG)^{-1} = (I+GK)^{-1}G \tag{3.8}$$

$$GK(I+GK)^{-1} = G(I+KG)^{-1}K = (I+GK)^{-1}GK \tag{3.9}$$

$$T = L(I+L)^{-1} = (I+(L)^{-1})^{-1} \tag{3.10}$$

Note that the matrices G and K in (3.7)-(3.10) need not be square whereas $L = GK$ is square. (3.7) follows trivially by factorizing out the term $(I+L)^{-1}$ from the right. (3.8) says that $GS_I = SG$ and follows from the push-through rule. (3.9) also follows from the push-through rule. (3.10) can be derived from the identity $M_1^{-1}M_2^{-1} = (M_2M_1)^{-1}$.

Similar relationships, but with G and K interchanged, apply for the transfer functions evaluated at the plant input. To assist in remembering (3.7)-(3.10) note that G comes first (because the transfer function is evaluated at the output) and then G and K alternate in sequence. A given transfer matrix never occurs twice in sequence. For example, the closed-loop transfer function $G(I+GK)^{-1}$ does *not* exist (unless G is repeated in the block diagram, but then these G's would actually represent two different physical entities).

Remark 1 The above identities are clearly useful when deriving transfer functions analytically, but they are also useful for numerical calculations involving state-space realizations, e.g. $L(s) = C(sI-A)^{-1}B + D$. For example, assume we have been given a state-space realization for $L = GK$ with n states (so A is a $n \times n$ matrix) and we want to find the state space realization of T. Then we can first form $S = (I+L)^{-1}$ with n states, and then multiply it by L to obtain $T = SL$ with $2n$ states. However, a minimal realization of T has only n states. This may be obtained numerically using model reduction, but it is preferable to find it directly using $T = I - S$, see (3.7).

Remark 2 Note also that the right identity in (3.10) can only be used to compute the state-space realization of T if that of L^{-1} exists, so L must be semi-proper with $D \neq 0$ (which is rarely the case in practice). On the other hand, since L is square, we can always compute the frequency response of $L(j\omega)^{-1}$ (except at frequencies where $L(s)$ has $j\omega$-axis poles), and then obtain $T(j\omega)$ from (3.10).

Remark 3 In Appendix A.6 we present some factorizations of the sensitivity function which will be useful in later applications. For example, (A.139) relates the sensitivity of a perturbed plant, $S' = (I+G'K)^{-1}$, to that of the nominal plant, $S = (I+GK)^{-1}$. We have

$$S' = S(I+E_OT)^{-1}, \quad E_O \triangleq (G'-G)G^{-1} \tag{3.11}$$

where E_O is an output multiplicative perturbation representing the difference between G and G', and T is the nominal complementary sensitivity function.

3.3 Multivariable frequency response analysis

The transfer function $G(s)$ is a function of the Laplace variable s and can be used to represent a dynamic system. However, if we fix $s = s_0$ then we may view $G(s_0)$ simply as a complex matrix, which can be analyzed using standard tools in matrix algebra. In particular, the choice $s_0 = j\omega$ is of interest since $G(j\omega)$ represents the response to a sinusoidal signal of frequency ω.

3.3.1 Obtaining the frequency response from $G(s)$

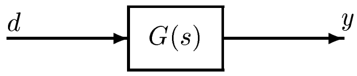


Figure 3.4: System $G(s)$ with input d and output y

The frequency domain is ideal for studying directions in multivariable systems at any given frequency. Consider the system $G(s)$ in Figure 3.4 with input $d(s)$ and output $y(s)$:

$$y(s) = G(s)d(s) \tag{3.12}$$

(We here denote the input by d rather than by u to avoid confusion with the matrix U used below in the singular value decomposition). In Section 2.1 we considered the sinusoidal response of scalar systems. These results may be directly generalized to multivariable systems by considering the elements g_{ij} of the matrix G. We have

- $g_{ij}(j\omega)$ represents the sinusoidal response from input j to output i.

To be more specific, apply to input channel j a scalar sinusoidal signal given by

$$d_j(t) = d_{j0}\sin(\omega t + \alpha_j) \tag{3.13}$$

This input signal is persistent, that is, it has been applied since $t = -\infty$. Then the corresponding persistent output signal in channel i is also a sinusoid with the same frequency

$$y_i(t) = y_{i0}\sin(\omega t + \beta_i) \tag{3.14}$$

where the amplification (gain) and phase shift may be obtained from the complex number $g_{ij}(j\omega)$ as follows

$$\frac{y_{io}}{d_{jo}} = |g_{ij}(j\omega)|, \quad \beta_i - \alpha_j = \angle g_{ij}(j\omega) \tag{3.15}$$

In phasor notation, see (2.7) and (2.9), we may compactly represent the sinusoidal time response described in (3.13)-(3.15) by

$$y_i(\omega) = g_{ij}(j\omega)d_j(\omega) \tag{3.16}$$

where

$$d_j(\omega) = d_{jo}e^{j\alpha_j}, \quad y_i(\omega) = y_{io}e^{j\beta_i} \tag{3.17}$$

Here the use of ω (and not $j\omega$) as the argument of $d_j(\omega)$ and $y_i(\omega)$ implies that these are complex numbers, representing at each frequency ω the magnitude and phase of the sinusoidal signals in (3.13) and (3.14).

The overall response to simultaneous input signals of the same frequency in several input channels is, by the superposition principle for linear systems, equal to the sum of the individual responses, and we have from (3.16)

$$y_i(\omega) = g_{i1}(j\omega)d_1(\omega) + g_{i2}(j\omega)d_2(\omega) + \cdots = \sum_j g_{ij}(j\omega)d_j(\omega) \tag{3.18}$$

or in matrix form

$$\boxed{y(\omega) = G(j\omega)d(\omega)} \tag{3.19}$$

where

$$d(\omega) = \begin{bmatrix} d_1(\omega) \\ d_2(\omega) \\ \vdots \\ d_m(\omega) \end{bmatrix} \quad \text{and} \quad y(\omega) = \begin{bmatrix} y_1(\omega) \\ y_2(\omega) \\ \vdots \\ y_l(\omega) \end{bmatrix} \tag{3.20}$$

represent the vectors of sinusoidal input and output signals.

Example 3.2 *Consider a 2×2 multivariable system where we simultaneously apply sinusoidal signals of the same frequency ω to the two input channels:*

$$d(t) = \begin{bmatrix} d_1(t) \\ d_2(t) \end{bmatrix} = \begin{bmatrix} d_{10}\sin(\omega t + \alpha_1) \\ d_{20}\sin(\omega t + \alpha_2) \end{bmatrix} \tag{3.21}$$

The corresponding output signal is

$$y(t) = \begin{bmatrix} y_1(t) \\ y_2(t) \end{bmatrix} = \begin{bmatrix} y_{10}\sin(\omega t + \beta_1) \\ y_{20}\sin(\omega t + \beta_2) \end{bmatrix} \tag{3.22}$$

which can be computed by multiplying the complex matrix $G(j\omega)$ by the complex vector $d(\omega)$:

$$y(\omega) = G(j\omega)d(\omega); \quad y(\omega) = \begin{bmatrix} y_{10}e^{j\beta_1} \\ y_{20}e^{j\beta_2} \end{bmatrix}, \ d(\omega) = \begin{bmatrix} d_{10}e^{j\alpha_1} \\ d_{20}e^{j\alpha_2} \end{bmatrix} \tag{3.23}$$

3.3.2 Directions in multivariable systems

For a SISO system, $y = Gd$, the gain at a given frequency is simply

$$\frac{|y(\omega)|}{|d(\omega)|} = \frac{|G(j\omega)d(\omega)|}{|d(\omega)|} = |G(j\omega)|$$

The gain depends on the frequency ω, but since the system is linear it is independent of the input magnitude $|d(\omega)|$.

Things are not quite as simple for MIMO systems where the input and output signals are both vectors, and we need to "sum up" the magnitudes of the elements in each vector by use of some norm, as discussed in Appendix A.5.1. If we select the vector 2-norm, the usual measure of length, then at a given frequency ω the magnitude of the vector input signal is

$$\|d(\omega)\|_2 = \sqrt{\sum_j |d_j(\omega)|^2} = \sqrt{d_{10}^2 + d_{20}^2 + \cdots} \tag{3.24}$$

and the magnitude of the vector output signal is

$$\|y(\omega)\|_2 = \sqrt{\sum_i |y_i(\omega)|^2} = \sqrt{y_{10}^2 + y_{20}^2 + \cdots} \tag{3.25}$$

The *gain* of the system $G(s)$ for a particular input signal $d(\omega)$ is then given by the ratio

$$\frac{\|y(\omega)\|_2}{\|d(\omega)\|_2} = \frac{\|G(j\omega)d(\omega)\|_2}{\|d(\omega)\|_2} = \frac{\sqrt{y_{10}^2 + y_{20}^2 + \cdots}}{\sqrt{d_{10}^2 + d_{20}^2 + \cdots}} \tag{3.26}$$

Again the gain depends on the frequency ω, and again it is independent of the input magnitude $\|d(\omega)\|_2$. However, for a MIMO system there are additional degrees of freedom and the gain depends also on the *direction* of the input d.

Example 3.3 *For a system with two inputs, $d = \begin{bmatrix} d_{10} \\ d_{20} \end{bmatrix}$, the gain is in general different for the following five inputs:*

$$d_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix},\ d_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix},\ d_3 = \begin{bmatrix} 0.707 \\ 0.707 \end{bmatrix},\ d_4 = \begin{bmatrix} 0.707 \\ -0.707 \end{bmatrix},\ d_5 = \begin{bmatrix} 0.6 \\ -0.8 \end{bmatrix}$$

(which all have the same magnitude $\|d\|_2 = 1$ but are in different directions). For example, for the 2×2 system

$$G_1 = \begin{bmatrix} 5 & 4 \\ 3 & 2 \end{bmatrix} \tag{3.27}$$

(a constant matrix) we compute for the five inputs d_j the following output vectors

$$y_1 = \begin{bmatrix} 5 \\ 3 \end{bmatrix},\ y_2 = \begin{bmatrix} 4 \\ 2 \end{bmatrix},\ y_3 = \begin{bmatrix} 6.36 \\ 3.54 \end{bmatrix},\ y_4 = \begin{bmatrix} 0.707 \\ 0.707 \end{bmatrix},\ y_5 = \begin{bmatrix} -0.2 \\ 0.2 \end{bmatrix}$$

and the 2-norms of these five outputs (i.e. the gains for the five inputs) are

$$\|y_1\|_2 = 5.83,\ \|y_2\|_2 = 4.47,\ \|y_3\|_2 = 7.30,\ \|y_4\|_2 = 1.00,\ \|y_5\|_2 = 0.28$$

This dependency of the gain on the input direction is illustrated graphically in Figure 3.5 where we have used the ratio d_{20}/d_{10} as an independent variable to represent the input direction. We see that, depending on the ratio d_{20}/d_{10}, the gain varies between 0.27 and 7.34.

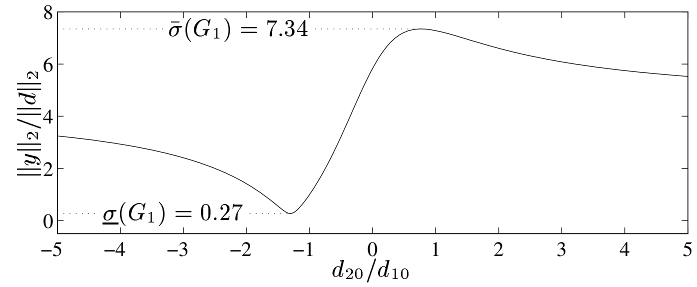


Figure 3.5: Gain $\|G_1 d\|_2/\|d\|_2$ as a function of d_{20}/d_{10} for G_1 in (3.27)

The maximum value of the gain in (3.26) as the direction of the input is varied is the maximum singular value of G,

$$\max_{d\neq 0} \frac{\|Gd\|_2}{\|d\|_2} = \max_{\|d\|_2=1} \|Gd\|_2 = \bar{\sigma}(G) \tag{3.28}$$

whereas the minimum gain is the minimum singular value of G,

$$\min_{d\neq 0} \frac{\|Gd\|_2}{\|d\|_2} = \min_{\|d\|_2=1} \|Gd\|_2 = \underline{\sigma}(G) \tag{3.29}$$

We will discuss this in detail below. The first identities in (3.28) and (3.29) follow because the gain is independent of the input magnitude for a linear system.

3.3.3 Eigenvalues are a poor measure of gain

Before discussing the singular values we want to demonstrate that the magnitudes of the eigenvalues of a transfer function matrix, e.g. $|\lambda_i(G(j\omega)|$, do *not* provide a useful means of generalizing the SISO gain, $|G(j\omega)|$. First of all, eigenvalues can only be computed for square systems, and even then they can be very misleading. To see this, consider the system $y = Gd$ with

$$G = \begin{bmatrix} 0 & 100 \\ 0 & 0 \end{bmatrix} \tag{3.30}$$

which has both eigenvalues λ_i equal to zero. However, to conclude from the eigenvalues that the system gain is zero is clearly misleading. For example, with an input vector $d = [0 \;\; 1]^T$ we get an output vector $y = [100 \;\; 0]^T$.

The "problem" is that the eigenvalues measure the gain for the special case when the inputs and the outputs are in the same direction, namely in the direction of the eigenvectors. To see this let t_i be an eigenvector of G and consider an input $d = t_i$. Then the output is $y = Gt_i = \lambda_i t_i$ where λ_i is the corresponding eigenvalue. We get

$$\|y\|/\|d\| = \|\lambda_i t_i\|/\|t_i\| = |\lambda_i|$$

so $|\lambda_i|$ measures the gain in the direction t_i. This may be useful for stability analysis, but not for performance.

To find useful generalizations of $|G|$ for the case when G is a matrix, we need the concept of a *matrix norm*, denoted $\|G\|$. Two important properties which must be satisfied for a matrix norm are the *triangle inequality*

$$\|G_1 + G_2\| \leq \|G_1\| + \|G_2\| \tag{3.31}$$

and the multiplicative property

$$\|G_1 G_2\| \leq \|G_1\| \cdot \|G_2\| \tag{3.32}$$

(see Appendix A.5 for more details). As we may expect, the magnitude of the largest eigenvalue, $\rho(G) \triangleq |\lambda_{max}(G)|$ (the spectral radius), does *not* satisfy the properties of a matrix norm; also see (A.115).

In Appendix A.5.2 we introduce several matrix norms, such as the Frobenius norm $\|G\|_F$, the sum norm $\|G\|_{\rm sum}$, the maximum column sum $\|G\|_{i1}$, the maximum row sum $\|G\|_{i\infty}$, and the maximum singular value $\|G\|_{i2} = \bar{\sigma}(G)$ (the latter three norms are induced by a vector norm, e.g. see (3.28); this is the reason for the subscript i). We will use all of these norms in this book, each depending on the situation. However, in this chapter we will mainly use the induced 2-norm, $\bar{\sigma}(G)$. Notice that $\bar{\sigma}(G) = 100$ for the matrix in (3.30).

Exercise 3.5 *Compute the spectral radius and the five matrix norms mentioned above for the matrices in (3.27) and (3.30).*

3.3.4 Singular value decomposition

The singular value decomposition (SVD) is defined in Appendix A.3. Here we are interested in its physical interpretation when applied to the frequency response of a MIMO system $G(s)$ with m inputs and l outputs.

Consider a fixed frequency ω where $G(j\omega)$ is a constant $l \times m$ complex matrix, and denote $G(j\omega)$ by G for simplicity. Any matrix G may be decomposed into its singular value decomposition, and we write

$$G = U \Sigma V^H \tag{3.33}$$

where

Σ is an $l \times m$ matrix with $k = \min\{l, m\}$ non-negative singular values, σ_i, arranged in descending order along its main diagonal; the other entries are zero. The singular values are the positive square roots of the eigenvalues of $G^H G$, where G^H is the complex conjugate transpose of G.

$$\sigma_i(G) = \sqrt{\lambda_i(G^H G)} \tag{3.34}$$

U is an $l \times l$ unitary matrix of output singular vectors, u_i,

V is an $m \times m$ unitary matrix of input singular vectors, v_i,

This is illustrated by the SVD of a real 2×2 matrix which can always be written in the form

$$G = \underbrace{\begin{bmatrix} \cos\theta_1 & -\sin\theta_1 \\ \sin\theta_1 & \cos\theta_1 \end{bmatrix}}_{U} \underbrace{\begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix}}_{\Sigma} \underbrace{\begin{bmatrix} \cos\theta_2 & \pm\sin\theta_2 \\ -\sin\theta_2 & \pm\cos\theta_2 \end{bmatrix}^T}_{V^T} \tag{3.35}$$

where the angles θ_1 and θ_2 depend on the given matrix. From (3.35) we see that the matrices U and V involve rotations and that their columns are orthonormal.

The singular values are sometimes called the principal values or principal gains, and the associated directions are called principal directions. In general, the singular values must be computed numerically. For 2×2 matrices however, analytic expressions for the singular values are given in (A.36).

Caution. It is standard notation to use the symbol U to denote the matrix of *output* singular vectors. This is unfortunate as it is also standard notation to use u (lower case) to represent the *input* signal. The reader should be careful not to confuse these two.

Input and output directions. The column vectors of U, denoted u_i, represent the *output directions* of the plant. They are orthogonal and of unit length (orthonormal), that is

$$\|u_i\|_2 = \sqrt{|u_{i1}|^2 + |u_{i2}|^2 + \ldots + |u_{il}|^2} = 1 \tag{3.36}$$

$$u_i^H u_i = 1, \quad u_i^H u_j = 0, \quad i \neq j \tag{3.37}$$

Likewise, the column vectors of V, denoted v_i, are orthogonal and of unit length, and represent the *input directions*. These input and output directions are related through the singular values. To see this, note that since V is unitary we have $V^H V = I$, so (3.33) may be written as $GV = U\Sigma$, which for column i becomes

$$Gv_i = \sigma_i u_i \tag{3.38}$$

where v_i and u_i are vectors, whereas σ_i is a scalar. That is, if we consider an *input* in the direction v_i, then the *output* is in the direction u_i. Furthermore, since $\|v_i\|_2 = 1$ and $\|u_i\|_2 = 1$ we see that the i'th singular value σ_i gives directly the gain of the matrix G in this direction. In other words

$$\sigma_i(G) = \|Gv_i\|_2 = \frac{\|Gv_i\|_2}{\|v_i\|_2} \tag{3.39}$$

Some advantages of the SVD over the eigenvalue decomposition for analyzing gains and directionality of multivariable plants are:

1. The singular values give better information about the gains of the plant.

2. The plant directions obtained from the SVD are orthogonal.
3. The SVD also applies directly to non-square plants.

Maximum and minimum singular values. As already stated, it can be shown that the largest gain for *any* input direction is equal to the maximum singular value

$$\bar{\sigma}(G) \equiv \sigma_1(G) = \max_{d\neq 0} \frac{\|Gd\|_2}{\|d\|_2} = \frac{\|Gv_1\|_2}{\|v_1\|_2} \tag{3.40}$$

and that the smallest gain for any input direction is equal to the minimum singular value

$$\underline{\sigma}(G) \equiv \sigma_k(G) = \min_{d\neq 0} \frac{\|Gd\|_2}{\|d\|_2} = \frac{\|Gv_k\|_2}{\|v_k\|_2} \tag{3.41}$$

where $k = \min\{l, m\}$. Thus, for any vector d we have that

$$\underline{\sigma}(G) \leq \frac{\|Gd\|_2}{\|d\|_2} \leq \bar{\sigma}(G) \tag{3.42}$$

Define $u_1 = \bar{u}, v_1 = \bar{v}, u_k = \underline{u}$ and $v_k = \underline{v}$. Then it follows that

$$G\bar{v} = \bar{\sigma}\bar{u}, \qquad G\underline{v} = \underline{\sigma}\;\underline{u} \tag{3.43}$$

The vector $\bar{v}$ corresponds to the input direction with largest amplification, and $\bar{u}$ is the corresponding output direction in which the inputs are most effective. The directions involving $\bar{v}$ and $\bar{u}$ are sometimes referred to as the "strongest", "high-gain" or "most important" directions. The next most important directions are associated with v_2 and u_2, and so on (see Appendix A.3.5) until the "least important", "weak" or "low-gain" directions which are associated with $\underline{v}$ and $\underline{u}$.

Example 3.4 *Consider again the system (3.27) in Example 3.3,*

$$G_1 = \begin{bmatrix} 5 & 4 \\ 3 & 2 \end{bmatrix} \tag{3.44}$$

The singular value decomposition of G_1 is

$$G_1 = \underbrace{\begin{bmatrix} 0.872 & 0.490 \\ 0.490 & -0.872 \end{bmatrix}}_{U} \underbrace{\begin{bmatrix} 7.343 & 0 \\ 0 & 0.272 \end{bmatrix}}_{\Sigma} \underbrace{\begin{bmatrix} 0.794 & -0.608 \\ 0.608 & 0.794 \end{bmatrix}^H}_{V^H}$$

The largest gain of 7.343 is for an input in the direction $\bar{v} = \begin{bmatrix} 0.794 \\ 0.608 \end{bmatrix}$, and the smallest gain of 0.272 is for an input in the direction $\underline{v} = \begin{bmatrix} -0.608 \\ 0.794 \end{bmatrix}$. This confirms the findings in Example 3.3.

Since in (3.44) both inputs affect both outputs, we say that the system is *interactive*. This follows from the relatively large off-diagonal elements in G_1. Furthermore, the system is *ill-conditioned*, that is, some combinations of the inputs have a strong effect on the outputs, whereas other combinations have a weak effect on the outputs. This may be quantified by the *condition number*; the ratio between the gains in the strong and weak directions; which for the system in (3.44) is $\bar{\sigma}/\underline{\sigma} = 7.343/0.272 = 27.0$.

Example 3.5 **Shopping cart**. *Consider a shopping cart (supermarket trolley) with fixed wheels which we may want to move in three directions; forwards, sideways and upwards. This is a simple illustrative example where we can easily figure out the principal directions from experience. The strongest direction, corresponding to the largest singular value, will clearly be in the forwards direction. The next direction, corresponding to the second singular value, will be sideways. Finally, the most "difficult" or "weak" direction, corresponding to the smallest singular value, will be upwards (lifting up the cart).*

For the shopping cart the gain depends strongly on the input direction, i.e. the plant is ill-conditioned. Control of ill-conditioned plants is sometimes difficult, and the control problem associated with the shopping cart can be described as follows: Assume we want to push the shopping cart sideways (maybe we are blocking someone). This is rather difficult (the plant has low gain in this direction) so a strong force is needed. However, if there is any uncertainty in our knowledge about the direction the cart is pointing, then some of our applied force will be directed forwards (where the plant gain is large) and the cart will suddenly move forward with an undesired large speed. We thus see that the control of an ill-conditioned plant may be especially difficult if there is input uncertainty which can cause the input signal to "spread" from one input direction to another. We will discuss this in more detail later.

Example 3.6 **Distillation process.** *Consider the following steady-state model of a distillation column*

$$G = \begin{bmatrix} 87.8 & -86.4 \\ 108.2 & -109.6 \end{bmatrix} \tag{3.45}$$

The variables have been scaled as discussed in Section 1.4. Thus, since the elements are much larger than 1 *in magnitude this suggests that there will be no problems with input constraints. However, this is somewhat misleading as the gain in the low-gain direction (corresponding to the smallest singular value) is actually only just above* 1*. To see this consider the SVD of G:*

$$G = \underbrace{\begin{bmatrix} 0.625 & -0.781 \\ 0.781 & 0.625 \end{bmatrix}}_{U} \underbrace{\begin{bmatrix} 197.2 & 0 \\ 0 & 1.39 \end{bmatrix}}_{\Sigma} \underbrace{\begin{bmatrix} 0.707 & -0.708 \\ -0.708 & -0.707 \end{bmatrix}^H}_{V^H} \tag{3.46}$$

From the first input singular vector, $\bar{v} = [\,0.707 \quad -0.708\,]^T$*, we see that the gain is* 197.2 *when we increase one input and decrease the other input by a similar amount. On the other hand, from the second input singular vector,* $\underline{v} = [\,-0.708 \quad -0.707\,]^T$*, we see that if we increase both inputs by the same amount then the gain is only* 1.39*. The reason for this is that the plant is such that the two inputs counteract each other. Thus, the distillation process is ill-conditioned, at least at steady-state, and the condition number is* $197.2/1.39 = 141.7$*. The physics of this example is discussed in more detail below, and later in this chapter we will consider a simple controller design (see Motivating robustness example No. 2 in Section 3.7.2).*

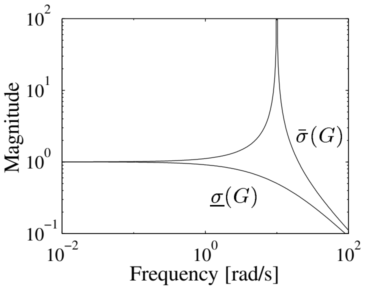
Example 3.7 **Physics of the distillation process.** *The model in (3.45) represents two-point (dual) composition control of a distillation column, where the top composition is to be controlled at* $y_D = 0.99$ *(output* y_1*) and the bottom composition at* $x_B = 0.01$ *(output* y_2*), using reflux L (input* u_1*) and boilup V (input* u_2*) as manipulated inputs (see Figure 10.6 on page 426). Note that we have here returned to the convention of using* u_1 *and* u_2 *to denote the manipulated inputs; the output singular vectors will be denoted by* $\bar{u}$ *and* $\underline{u}$*.*

The $1,1$*-element of the gain matrix G is* 87.8*. Thus an increase in* u_1 *by* 1 *(with* u_2 *constant) yields a large steady-state change in* y_1 *of* 87.8*, that is, the outputs are very sensitive to changes* *in u_1. Similarly, an increase in u_2 by 1 (with u_1 constant) yields $y_1 = -86.4$. Again, this is a very large change, but in the opposite direction of that for the increase in u_1. We therefore see that changes in u_1 and u_2 counteract each other, and if we increase u_1 and u_2 simultaneously by 1, then the overall steady-state change in y_1 is only $87.8 - 86.4 = 1.4$.*

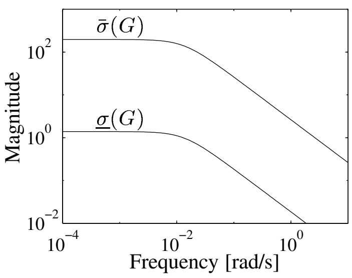
Physically, the reason for this small change is that the compositions in the distillation column are only weakly dependent on changes in the internal flows *(i.e. simultaneous changes in the internal flows L and V). This can also be seen from the smallest singular value, $\underline{\sigma}(G) = 1.39$, which is obtained for inputs in the direction $\underline{v} = \begin{bmatrix} -0.708 \\ -0.707 \end{bmatrix}$. From the output singular vector $\underline{u} = \begin{bmatrix} -0.781 \\ 0.625 \end{bmatrix}$ we see that the effect is to move the outputs in different directions, that is, to change $y_1 - y_2$. Therefore, it takes a large control action to move the compositions in different directions, that is, to make both products purer simultaneously. This makes sense from a physical point of view.*

On the other hand, the distillation column is very sensitive to changes in external flows *(i.e. increase $u_1 - u_2 = L - V$). This can be seen from the input singular vector $\bar{v} = \begin{bmatrix} 0.707 \\ -0.708 \end{bmatrix}$ associated with the largest singular value, and is a general property of distillation columns where both products are of high purity. The reason for this is that the external distillate flow (which varies as $V - L$) has to be about equal to the amount of light component in the feed, and even a small imbalance leads to large changes in the product compositions.*

For dynamic systems the singular values and their associated directions vary with frequency, and for control purposes it is usually the frequency range corresponding to the closed-loop bandwidth which is of main interest. The singular values are usually plotted as a function of frequency in a Bode magnitude plot with a log-scale for frequency and magnitude. Typical plots are shown in Figure 3.6.



(a) Spinning satellite in (3.77)



(b) Distillation process in (3.82)

Figure 3.6: Typical plots of singular values

Non-Square plants

The SVD is also useful for non-square plants. For example, consider a plant with 2 inputs and 3 outputs. In this case the third output singular vector, u_3, tells us in which output direction the plant cannot be controlled. Similarly, for a plant with more inputs than outputs, the additional input singular vectors tell us in which directions the input will have no effect.

Exercise 3.6 *For a system with m inputs and 1 output, what is the interpretation of the singular values and the associated input directions (V)? What is U in this case?*

Use of the minimum singular value of the plant

The minimum singular value of the plant, $\underline{\sigma}(G(j\omega))$, evaluated as a function of frequency, is a useful measure for evaluating the feasibility of achieving acceptable control. If the inputs and outputs have been scaled as outlined in Section 1.4, then with a manipulated input of unit magnitude (measured by the 2-norm), we can achieve an output magnitude of at least $\underline{\sigma}(G)$ in *any* output direction. We generally want $\underline{\sigma}(G)$ as large as possible.

Remark. The requirement $\underline{\sigma}(G) > 1$, to avoid input saturation, is discussed in Section 6.9. In Section 10.3, it is shown that it may be desirable to have $\underline{\sigma}(G(j\omega))$ large even when input saturation is not a concern. The minimum singular value of the plant and its use is also discussed by Morari (1983), and Yu and Luyben (1986) call $\underline{\sigma}(G(j\omega))$ the "Morari resilience index".

3.3.5 Singular values for performance

So far we have used the SVD primarily to gain insight into the directionality of MIMO systems. But the maximum singular value is also very useful in terms of frequency-domain performance and robustness. We here consider performance.

For SISO systems we earlier found that $|S(j\omega)|$ evaluated as a function of frequency gives useful information about the effectiveness of feedback control. For example, it is the gain from a sinusoidal reference input (or output disturbance) to the control error, $|e(\omega)|/|r(\omega)| = |S(j\omega)|$.

For MIMO systems a useful generalization results if we consider the ratio $\|e(\omega)\|_2/\|r(\omega)\|_2$, where r is the vector of reference inputs, e is the vector of control errors, and $\|\cdot\|_2$ is the vector 2-norm. As explained above, this gain depends on the *direction* of $r(\omega)$ and we have from (3.42) that it is bounded by the maximum and minimum singular value of S,

$$\underline{\sigma}(S(j\omega)) \leq \frac{\|e(\omega)\|_2}{\|r(\omega)\|_2} \leq \bar{\sigma}(S(j\omega)) \tag{3.47}$$

In terms of *performance*, it is reasonable to require that the gain $\|e(\omega)\|_2/\|r(\omega)\|_2$ remains small for any direction of $r(\omega)$, including the "worst-case" direction which gives a gain of $\bar{\sigma}(S(j\omega))$. Let $1/|w_P(j\omega)|$ (the inverse of the performance weight) represent the maximum allowed magnitude of $\|e\|_2/\|r\|_2$ at each frequency. This results in the following performance requirement:

$$\begin{aligned} \bar{\sigma}(S(j\omega)) < 1/|w_P(j\omega)|,\ \forall\omega \quad &\Leftrightarrow \quad \bar{\sigma}(w_P S) < 1,\ \forall\omega \\ &\Leftrightarrow \quad \|w_P S\|_\infty < 1 \end{aligned} \tag{3.48}$$

where the $\mathcal{H}_\infty$ norm (see also page 55) is defined as the peak of the maximum singular value of the frequency response

$$\|M(s)\|_\infty \triangleq \max_\omega \bar{\sigma}(M(j\omega)) \tag{3.49}$$

Typical performance weights $w_P(s)$ are given in Section 2.7.2, which should be studied carefully.

The singular values of $S(j\omega)$ may be plotted as functions of frequency, as illustrated later in Figure 3.10(a). Typically, they are small at low frequencies where feedback is effective, and they approach 1 at high frequencies because any real system is strictly proper:

$$\omega \to \infty : \quad L(j\omega) \to 0 \quad \Rightarrow \quad S(j\omega) \to I \tag{3.50}$$

The maximum singular value, $\bar{\sigma}(S(j\omega))$, usually has a peak larger than 1 around the crossover frequencies. This peak is undesirable, but it is unavoidable for real systems.

As for SISO systems we define the bandwidth as the frequency up to which feedback is effective. For MIMO systems the bandwidth will depend on directions, and we have a *bandwidth region* between a lower frequency where the maximum singular value, $\bar{\sigma}(S)$, reaches 0.7 (the low-gain or worst-case direction), and a higher frequency where the minimum singular value, $\underline{\sigma}(S)$, reaches 0.7 (the high-gain or best direction). If we want to associate a single bandwidth frequency for a multivariable system, then we consider the worst-case (low-gain) direction, and define

- *Bandwidth*, ω_B: Frequency where $\bar{\sigma}(S)$ crosses $\frac{1}{\sqrt{2}} = 0.7$ from below.

It is then understood that the bandwidth is at least ω_B for any direction of the input (reference or disturbance) signal. Since $S = (I + L)^{-1}$, (A.52) yields

$$\underline{\sigma}(L) - 1 \leq \frac{1}{\bar{\sigma}(S)} \leq \underline{\sigma}(L) + 1 \tag{3.51}$$

Thus at frequencies where feedback is effective (namely where $\underline{\sigma}(L) \gg 1$) we have $\bar{\sigma}(S) \approx 1/\underline{\sigma}(L)$, and at the bandwidth frequency (where $1/\bar{\sigma}(S(j\omega_B)) = \sqrt{2} = 1.41$) we have that $\underline{\sigma}(L(j\omega_B))$ is between 0.41 and 2.41. Thus, the bandwidth is approximately where $\underline{\sigma}(L)$ crosses 1. Finally, at higher frequencies where for any real system $\underline{\sigma}(L)$ (and $\bar{\sigma}(L)$) is small we have that $\bar{\sigma}(S) \approx 1$.

3.4 Control of multivariable plants

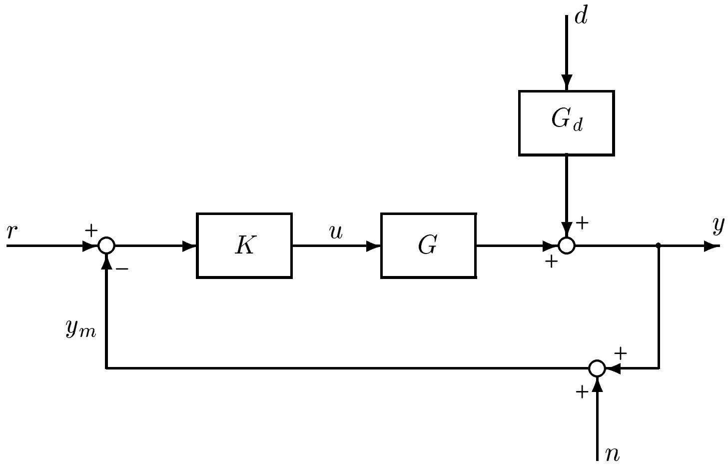


Figure 3.7: One degree-of-freedom feedback control configuration

Consider the simple feedback system in Figure 3.7. A conceptually simple approach to multivariable control is given by a two-step procedure in which we first design a "compensator" to deal with the interactions in G, and then design a *diagonal* controller using methods similar to those for SISO systems. This approach is discussed below.

The most common approach is to use a pre-compensator, $W_1(s)$, which counteracts the interactions in the plant and results in a "new" shaped plant:

$$G_s(s) = G(s)W_1(s) \tag{3.52}$$

which is more diagonal and easier to control than the original plant $G(s)$. After finding a suitable $W_1(s)$ we can design a *diagonal* controller $K_s(s)$ for the shaped plant $G_s(s)$. The overall controller is then

$$K(s) = W_1(s)K_s(s) \tag{3.53}$$

In many cases effective compensators may be derived on physical grounds and may include nonlinear elements such as ratios.

Remark 1 Some design approaches in this spirit are the Nyquist Array technique of Rosenbrock (1974) and the characteristic loci technique of MacFarlane and Kouvaritakis (1977).

Remark 2 The $\mathcal{H}_\infty$ loop-shaping design procedure, described in detail in Section 9.4, is similar in that a pre-compensator is first chosen to yield a shaped plant, $G_s = GW_1$, with desirable properties, and then a controller $K_s(s)$ is designed. The main difference is that in $\mathcal{H}_\infty$ loop shaping, $K_s(s)$ is a full multivariable controller, designed based on optimization (to optimize $\mathcal{H}_\infty$ robust stability).

3.4.1 Decoupling

Decoupling control results when the compensator is chosen such that G_s in (3.52) is diagonal at a selected frequency. The following different cases are possible:

1. *Dynamic decoupling:* $G_s(s)$ is diagonal (at all frequencies). For example, with $G_s(s) = I$ and a square plant, we get $W_1 = G^{-1}(s)$ (disregarding the possible problems involved in realizing $G^{-1}(s)$). If we then select $K_s(s) = l(s)I$ (e.g. with $l(s) = k/s$), the overall controller is

$$K(s) = K_{\mathrm{inv}}(s) \triangleq l(s)G^{-1}(s) \tag{3.54}$$

   We will later refer to (3.54) as an *inverse-based* controller. It results in a decoupled nominal system with identical loops, i.e. $L(s) = l(s)I$, $S(s) = \frac{1}{1+l(s)}I$ and $T(s) = \frac{l(s)}{1+l(s)}I$.

   Remark. In some cases we may want to keep the diagonal elements in the shaped plant unchanged by selecting $W_1 = G^{-1}G_{diag}$. In other cases we may want the diagonal elements in W_1 to be 1. This may be obtained by selecting $W_1 = G^{-1}((G^{-1})_{diag})^{-1}$, and the off-diagonal elements of W_1 are then called "decoupling elements".

2. *Steady-state decoupling:* $G_s(0)$ is diagonal. This may be obtained by selecting a constant pre-compensator $W_1 = G^{-1}(0)$ (and for a non-square plant we may use the pseudo-inverse provided $G(0)$ has full row (output) rank).
3. *Approximate decoupling at frequency w_o:* $G_s(j\omega_o)$ is as diagonal as possible. This is usually obtained by choosing a constant pre-compensator $W_1 = G_o^{-1}$ where G_o is a real approximation of $G(j\omega_o)$. G_o may be obtained, for example, using the align algorithm of Kouvaritakis (1974). The bandwidth frequency is a good selection for ω_o because the effect on performance of reducing interaction is normally greatest at this frequency.

The idea of decoupling control is appealing, but there are several difficulties:

1. As one might expect, decoupling may be very sensitive to modelling errors and uncertainties. This is illustrated below in Section 3.7.2.
2. The requirement of decoupling and the use of an inverse-based controller may not be desirable for disturbance rejection. The reasons are similar to those given for SISO systems in Section 2.6.4, and are discussed further below; see (3.58).
3. If the plant has RHP-zeros then the requirement of decoupling generally introduces extra RHP-zeros into the closed-loop system (see Section 6.5.1).

Even though decoupling controllers may not always be desirable in practice, they are of interest from a theoretical point of view. They also yield insights into the limitations imposed by the multivariable interactions on achievable performance. One popular design method, which essentially yields a decoupling controller is the internal model control (IMC) approach (Morari and Zafiriou, 1989).

Another common strategy, which avoids most of the problems just mentioned, is to use *partial (one-way) decoupling* where $G_s(s)$ in (3.52) is upper or lower triangular.

3.4.2 Pre- and post-compensators and the SVD-controller

The above pre-compensator approach may be extended by introducing a post-compensator $W_2(s)$, as shown in Figure 3.8. One then designs a *diagonal* controller

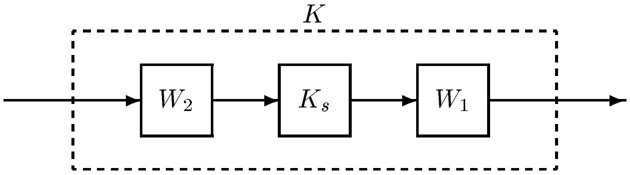


Figure 3.8: Pre- and post-compensators, W_1 and W_2. K_s is diagonal

K_s for the shaped plant $W_2 G W_1$. The overall controller is then

$$K(s) = W_1 K_s W_2 \tag{3.55}$$

The *SVD-controller* is a special case of a pre- and post-compensator design. Here

$$W_1 = V_o \quad \text{and} \quad W_2 = U_o^T \tag{3.56}$$

where V_o and U_o are obtained from a singular value decomposition of $G_o = U_o \Sigma_o V_o^T$, where G_o is a real approximation of $G(j\omega_o)$ at a given frequency w_o (often around the bandwidth). SVD-controllers are studied by Hung and MacFarlane (1982), and by Hovd et al. (1994) who found that the SVD controller structure is optimal in some cases, e.g. for plants consisting of symmetrically interconnected subsystems.

In summary, the SVD-controller provides a useful class of controllers. By selecting $K_s = l(s)\Sigma_o^{-1}$ a decoupling design is achieved, and by selecting a diagonal K_s with a low condition number ($\gamma(K_s)$ small) generally results in a robust controller (see Section 6.10).

3.4.3 Diagonal controller (decentralized control)

Another simple approach to multivariable controller design is to use a diagonal or block-diagonal controller $K(s)$. This is often referred to as decentralized control. Clearly, this works well if $G(s)$ is close to diagonal, because then the plant to be controlled is essentially a collection of independent sub-plants, and each element in $K(s)$ may be designed independently. However, if off-diagonal elements in $G(s)$ are large, then the performance with decentralized diagonal control may be poor because no attempt is made to counteract the interactions.

3.4.4 What is the shape of the "best" feedback controller?

Consider the problem of disturbance rejection. The closed-loop disturbance response is $y = SG_d d$. Suppose we have scaled the system (see Section 1.4) such that at each frequency the disturbances are of magnitude 1, $||d||_2 \leq 1$, and our performance requirement is that $||y||_2 \leq 1$. This is equivalent to requiring $\bar{\sigma}(SG_d) \leq 1$. In many cases there is a trade-off between input usage and performance, such that the controller that minimizes the input magnitude is one that yields all singular values of SG_d equal to 1, i.e. $\sigma_i(SG_d) = 1, \forall\omega$. This corresponds to

$$S_{\min} G_d = U_1 \tag{3.57}$$

where $U_1(s)$ is some all-pass transfer function (which at each frequency has all its singular values equal to 1). The subscript min refers to the use of the smallest loop gain that satisfies the performance objective. For simplicity, we assume that G_d is square so $U_1(j\omega)$ is a unitary matrix. At frequencies where feedback is effective we have $S = (I + L)^{-1} \approx L^{-1}$, and (3.57) yields $L_{\min} = GK_{\min} \approx G_d U_1^{-1}$. In conclusion, the controller and loop shape with the minimum gain will often look like

$$K_{\min} \approx G^{-1} G_d U_2, \quad L_{\min} \approx G_d U_2 \tag{3.58}$$

where $U_2 = U_1^{-1}$ is some all-pass transfer function matrix. This provides a generalization of $|K_{\min}| \approx |G^{-1}G_d|$ which was derived in (2.58) for SISO systems, and the summary following (2.58) on page 48 therefore also applies to MIMO systems. For example, we see that for disturbances entering at the plant inputs, $G_d = G$, we get $K_{\min} = U_2$, so a simple constant unit gain controller yields a good trade-off between output performance and input usage. We also note with interest that it is generally not possible to select a unitary matrix U_2 such that $L_{\min} = G_d U_2$ is diagonal, so a decoupling design is generally not optimal for disturbance rejection. These insights can be used as a basis for a loop-shaping design; see more on $\mathcal{H}_\infty$ loop-shaping in Chapter 9.

3.4.5 Multivariable controller synthesis

The above design methods are based on a two-step procedure in which we first design a pre-compensator (for decoupling control) or we make an input-output pairing selection (for decentralized control) and then we design a diagonal controller $K_s(s)$. Invariably this two-step procedure results in a suboptimal design.

The alternative is to synthesize directly a multivariable controller $K(s)$ based on minimizing some objective function (norm). We here use the word *synthesize* rather than *design* to stress that this is a more formalized approach. Optimization in controller design became prominent in the 1960's with "optimal control theory" based on minimizing the expected value of the output variance in the face of stochastic disturbances. Later, other approaches and norms were introduced, such as $\mathcal{H}_\infty$ optimal control.

3.4.6 Summary of mixed-sensitivity $\mathcal{H}_\infty$ design (S/KS)

We here provide a brief summary of the S/KS and other mixed-sensitivity $\mathcal{H}_\infty$ design methods which are used in later examples. In the S/KS problem, the objective is to minimize the $\mathcal{H}_\infty$ norm of

$$N = \begin{bmatrix} W_P S \\ W_u KS \end{bmatrix} \tag{3.59}$$

This problem was discussed earlier for SISO systems, and another look at Section 2.7.3 would be useful now. A sample MATLAB file is provided in Example 2.11, page 60.

The following issues and guidelines are relevant when selecting the weights W_P and W_u:

1. KS is the transfer function from r to u in Figure 3.7, so for a system which has been scaled as in Section 1.4, a reasonable initial choice for the input weight is $W_u = I$.
2. S is the transfer function from r to $-e = r - y$. A common choice for the performance weight is $W_P = \mathrm{diag}\{w_{Pi}\}$ with

$$w_{Pi} = \frac{s/M_i + \omega_{Bi}^*}{s + \omega_{Bi}^* A_i}, \quad A_i \ll 1 \tag{3.60}$$

   (see also Figure 2.26 on page 58). Selecting $A_i \ll 1$ ensures approximate integral action with $S(0) \approx 0$. Often we select M_i about 2 for all outputs, whereas ω_{Bi}^* may be different for each output. A large value of ω_{Bi}^* yields a faster response for output i.
3. To find a reasonable initial choice for the weight W_P, one can first obtain a controller with some other design method, plot the magnitude of the resulting diagonal elements of S as a function of frequency, and select $w_{Pi}(s)$ as a rational approximation of $1/|S_{ii}|$.
4. For disturbance rejection, we may in some cases want a steeper slope for $w_{Pi}(s)$ at low frequencies than that given in (3.60), e.g. as see the weight in (2.73). However, it may be better to consider the disturbances explicitly by considering the $\mathcal{H}_\infty$ norm of

$$N = \begin{bmatrix} W_P S & W_P S G_d \\ W_u KS & W_u KS G_d \end{bmatrix} \tag{3.61}$$

   or equivalently

$$N = \begin{bmatrix} W_P S W_d \\ W_u KS W_d \end{bmatrix} \text{ with } W_d = [\, I \quad G_d \,] \tag{3.62}$$

   where N represents the transfer function from $\begin{bmatrix} r \\ d \end{bmatrix}$ to the weighted outputs $\begin{bmatrix} W_P e \\ W_u u \end{bmatrix}$. In some situations we may want to adjust W_P or G_d in order to satisfy better our original objectives. The helicopter case study in Section 12.2 illustrates this by introducing a scalar parameter α to adjust the magnitude of G_d.

5. T is the transfer function from $-n$ to y. To reduce sensitivity to noise and uncertainty, we want T small at high frequencies, and so we may want additional roll-off in L. This can be achieved in several ways. One approach is to add $W_T T$ to the stack for N in (3.59), where $W_T = \mathrm{diag}\{w_{Ti}\}$ and $|w_{Ti}|$ is smaller than 1 at low frequencies and large at high frequencies. A more direct approach is to add high-frequency dynamics, $W_1(s)$, to the plant model to ensure that the resulting shaped plant, $G_s = GW_1$, rolls off with the desired slope. We then obtain an $\mathcal{H}_\infty$ optimal controller K_s for this shaped plant, and finally include $W_1(s)$ in the controller, $K = W_1 K_s$.

More details about $\mathcal{H}_\infty$ design are given in Chapter 9.

3.5 Introduction to multivariable RHP-zeros

By means of an example, we now give the reader an appreciation of the fact that MIMO systems have zeros even though their presence may not be obvious from the elements of $G(s)$. As for SISO systems, we find that RHP-zeros impose fundamental limitations on control.

The zeros z of MIMO systems are defined as the values $s = z$ where $G(s)$ loses rank, and we can find the *direction* of a zero by looking at the direction in which the matrix $G(z)$ has zero gain. For square systems we essentially have that the poles and zeros of $G(s)$ are the poles and zeros of $\det G(s)$. However, this crude method may fail in some cases, as it may incorrectly cancel poles and zeros with the same location but different directions (see Sections 4.5 and 4.6.1 for more details).

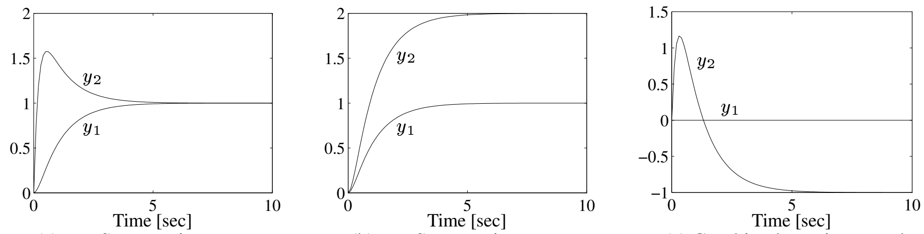
Example 3.8 *Consider the following plant*

$$G(s) = \frac{1}{(0.2s+1)(s+1)} \begin{bmatrix} 1 & 1 \\ 1+2s & 2 \end{bmatrix} \tag{3.63}$$

The responses to a step in each individual input are shown in Figure 3.9(a) and (b). We see that the plant is interactive, but for these two inputs there is no inverse response to indicate the presence of a RHP-zero. Nevertheless, the plant does have a multivariable RHP-zero at $z = 0.5$*, that is,* $G(s)$ *loses rank at* $s = 0.5$*, and* $\det G(0.5) = 0$*. The singular value decomposition of* $G(0.5)$ *is*

$$G(0.5) = \frac{1}{1.65} \begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix} = \underbrace{\begin{bmatrix} 0.45 & 0.89 \\ 0.89 & -0.45 \end{bmatrix}}_{U} \underbrace{\begin{bmatrix} 1.92 & 0 \\ 0 & 0 \end{bmatrix}}_{\Sigma} \underbrace{\begin{bmatrix} 0.71 & -0.71 \\ 0.71 & 0.71 \end{bmatrix}^H}_{V^H} \tag{3.64}$$

and we have as expected $\underline{\sigma}(G(0.5)) = 0$*. The input and output directions corresponding to the RHP-zero are* $\underline{v} = \begin{bmatrix} -0.71 \\ 0.71 \end{bmatrix}$ *and* $\underline{u} = \begin{bmatrix} 0.89 \\ -0.45 \end{bmatrix}$*. Thus, the RHP-zero is associated with*



(a) Step in u_1, $u = [\,1 \quad 0\,]^T$ (b) Step in u_2, $u = [\,0 \quad 1\,]^T$ (c) Combined step in u_1 and u_2, $u = [\,1 \quad -1\,]^T$

Figure 3.9: Open-loop response for $G(s)$ in (3.63)

both inputs and with both outputs. The presence of the multivariable RHP-zero is also observed from the time response in Figure 3.9(c), which is for a simultaneous input change in opposite directions, $u = \begin{bmatrix} 1 \\ -1 \end{bmatrix}^T$*. We see that y_2 displays an inverse response whereas y_1 happens to remain at zero for this particular input change.*

To see how the RHP-zero affects the closed-loop response, we design a controller which minimizes the $\mathcal{H}_\infty$ norm of the weighted S/KS matrix

$$N = \begin{bmatrix} W_P S \\ W_u K S \end{bmatrix} \tag{3.65}$$

with weights

$$W_u = I,\ W_P = \begin{bmatrix} w_{P1} & 0 \\ 0 & w_{P2} \end{bmatrix}, w_{Pi} = \frac{s/M_i + \omega^*_{Bi}}{s + w^*_{Bi} A_i}, \quad A_i = 10^{-4} \tag{3.66}$$

The MATLAB file for the design is the same as in Table 2.3 on page 60, except that we now have a 2×2 system. Since there is a RHP-zero at $z = 0.5$ we expect that this will somehow limit the bandwidth of the closed-loop system.

Design 1. *We weight the two outputs equally and select*

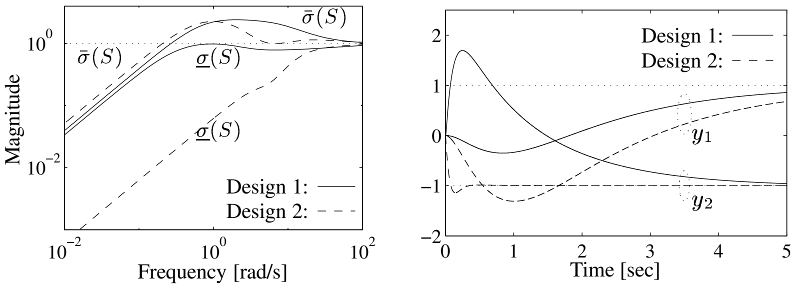
$$\text{Design 1}: \quad M_1 = M_2 = 1.5; \quad \omega^*_{B1} = \omega^*_{B2} = z/2 = 0.25$$

This yields an $\mathcal{H}_\infty$ norm for N of 2.80 and the resulting singular values of S are shown by the solid lines in Figure 3.10(a). The closed-loop response to a reference change $r = [\,1 \quad -1\,]^T$ is shown by the solid lines in Figure 3.10(b). We note that both outputs behave rather poorly and both display an inverse response.

Design 2. *For MIMO plants, one can often move most of the deteriorating effect (e.g. inverse response) of a RHP-zero to a particular output channel. To illustrate this, we change the weight w_{P2} so that more emphasis is placed on output 2. We do this by increasing the bandwidth requirement in output channel 2 by a factor of 100:*

$$\text{Design 2}: \quad M_1 = M_2 = 1.5; \quad \omega^*_{B1} = 0.25,\ \omega^*_{B2} = 25$$

This yields an $\mathcal{H}_\infty$ norm for N of 2.92. In this case we see from the dashed line in Figure 3.10(b) that the response for output 2 (y_2) is excellent with no inverse response.



(a) Singular values of S

(b) Response to change in reference, $r = [\,1 \quad -1\,]^T$

Figure 3.10: Alternative designs for 2×2 plant (3.63) with RHP-zero

However, this comes at the expense of output 1 *(*y_1*) where the response is somewhat poorer than for Design* 1.

Design 3. *We can also interchange the weights* w_{P1} *and* w_{P2} *to stress output* 1 *rather than output* 2. *In this case (not shown) we get an excellent response in output* 1 *with no inverse response, but output* 2 *responds very poorly (much poorer than output* 1 *for Design* 2*). Furthermore, the* $\mathcal{H}_\infty$ *norm for* N *is* 6.73, *whereas it was only* 2.92 *for Design* 2.

Thus, we see that it is easier, for this example, to get tight control of output 2 *than of output* 1. *This may be expected from the output direction of the RHP-zero,* $\underline{u} = \begin{bmatrix} 0.89 \\ -0.45 \end{bmatrix}$, *which is mostly in the direction of output* 1. *We will discuss this in more detail in Section 6.5.1.*

Remark 1 We find from this example that we can direct the effect of the RHP-zero to either of the two outputs. This is typical of multivariable RHP-zeros, but there are cases where the RHP-zero is associated with a particular output channel and it is *not* possible to move its effect to another channel. The zero is then called a “pinned zero” (see Section 4.6.2).

Remark 2 It is observed from the plot of the singular values in Figure 3.10(a), that we were able to obtain by Design 2 a very large improvement in the “good” direction (corresponding to $\underline{\sigma}(S)$) at the expense of only a minor deterioration in the “bad” direction (corresponding to $\bar{\sigma}(S)$). Thus Design 1 demonstrates a shortcoming of the $\mathcal{H}_\infty$ norm: only the worst direction (maximum singular value) contributes to the $\mathcal{H}_\infty$ norm and it may not always be easy to get a good trade-off between the various directions.

3.6 Condition number and RGA

Two measures which are used to quantify the degree of directionality and The level of (two-way) interactions in MIMO systems, are the condition number and the relative gain array (RGA), respectively. We here define the two measures and present an overview of their *practical use*. We do not give detailed proofs, but refer to other places in the book for further details.

3.6.1 Condition number

We define the *condition number* of a matrix as the ratio between the maximum and minimum singular values,

$$\gamma(G) \stackrel{\triangle}{=} \bar{\sigma}(G)/\underline{\sigma}(G) \tag{3.67}$$

A matrix with a large condition number is said to be *ill-conditioned*. For a non-singular (square) matrix $\underline{\sigma}(G) = 1/\bar{\sigma}(G^{-1})$, so $\gamma(G) = \bar{\sigma}(G)\bar{\sigma}(G^{-1})$. It then follows from (A.119) that the condition number is large if both G and G^{-1} have large elements.

The condition number depends strongly on the scaling of the inputs and outputs. To be more specific, if D_1 and D_2 are diagonal scaling matrices, then the condition numbers of the matrices G and $D_1 G D_2$ may be arbitrarily far apart. In general, the matrix G should be scaled on physical grounds, for example, by dividing each input and output by its largest expected or desired value as discussed in Section 1.4.

One might also consider minimizing the condition number over all possible scalings. This results in the *minimized or optimal condition number* which is defined by

$$\gamma^*(G) = \min_{D_1, D_2} \gamma(D_1 G D_2) \tag{3.68}$$

and can be computed using (A.73).

The condition number has been used as an input-output controllability measure, and in particular it has been postulated that a large condition number indicates sensitivity to uncertainty. This is not true in general, but the reverse holds; if the condition number is small, then the multivariable effects of uncertainty are not likely to be serious (see (6.72)).

If the condition number is large (say, larger than 10), then this may *indicate* control problems:

1. A large condition number $\gamma(G) = \bar{\sigma}(G)/\underline{\sigma}(G)$ may be caused by a small value of $\underline{\sigma}(G)$, which is generally undesirable (on the other hand, a large value of $\bar{\sigma}(G)$ need not necessarily be a problem).
2. A large condition number may mean that the plant has a large minimized condition number, or equivalently, it has large RGA-elements which indicate fundamental control problems; see below.
3. A large condition number *does* imply that the system is sensitive to "unstructured" (full-block) input uncertainty (e.g. with an inverse-based controller, see (8.135)), but this kind of uncertainty often does not occur in practice. We therefore *cannot* generally conclude that a plant with a large condition number is sensitive to uncertainty, e.g. see the diagonal plant in Example 3.9.

3.6.2 Relative Gain Array (RGA)

The relative gain array (RGA) of a non-singular square matrix G is a square matrix defined as

$$\mathrm{RGA}(G) = \Lambda(G) \triangleq G \times (G^{-1})^T \tag{3.69}$$

where $\times$ denotes element-by-element multiplication (the Hadamard or Schur product). For a 2×2 matrix with elements g_{ij} the RGA is

$$\Lambda(G) = \begin{bmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{bmatrix} = \begin{bmatrix} \lambda_{11} & 1-\lambda_{11} \\ 1-\lambda_{11} & \lambda_{11} \end{bmatrix}; \quad \lambda_{11} = \frac{1}{1 - \frac{g_{12}g_{21}}{g_{11}g_{22}}} \tag{3.70}$$

Bristol (1966) originally introduced the RGA as a steady-state measure of interactions for decentralized control. Unfortunately, based on the original definition, many people have dismissed the RGA as being "only meaningful at $\omega = 0$". To the contrary, in most cases it is the value of the RGA at frequencies close to crossover which is most important.

The RGA has a number of interesting *algebraic properties*, of which the most important are (see Appendix A.4 for more details):

1. It is independent of input and output scaling.
2. Its rows and columns sum to one.
3. The sum-norm of the RGA, $\|\Lambda\|_{\mathrm{sum}}$, is very close to the minimized condition number γ^*; see (A.78). This means that plants with large RGA-elements are always ill-conditioned (with a large value of $\gamma(G)$), but the reverse may not hold (i.e. a plant with a large $\gamma(G)$ may have small RGA-elements).
4. A relative change in an element of G equal to the negative inverse of its corresponding RGA-element yields singularity.
5. The RGA is the identity matrix if G is upper or lower triangular.

From the last property it follows that the RGA (or more precisely $\Lambda - I$) provides a measure of *two-way interaction*. The definition of the RGA may be generalized to non-square matrices by using the pseudo inverse; see Appendix A.4.2.

In addition to the algebraic properties listed above, the RGA has a surprising number of useful *control properties*:

1. The RGA is a good indicator of sensitivity to uncertainty:
   (a) *Uncertainty in the input channels (diagonal input uncertainty).* Plants with large RGA-elements around the crossover frequency are fundamentally difficult to control because of sensitivity to input uncertainty (e.g. caused by uncertain or neglected actuator dynamics). In particular, decouplers or other inverse-based controllers should not be used for plants with large RGA-elements (see page 244).
   (b) *Element uncertainty.* As implied by algebraic property no. 4 above, large RGA-elements imply sensitivity to element-by-element uncertainty. However, this

kind of uncertainty may not occur in practice due to physical couplings between the transfer function elements. Therefore, diagonal input uncertainty (which is always present) is usually of more concern for plants with large RGA-elements.

2. *RGA and RHP-zeros.* If the sign of an RGA-element changes from $s = 0$ to $s = \infty$, then there is a RHP-zero in G or in some subsystem of G (see Theorem 10.5).
3. *Non-square plants*. Extra inputs: If the sum of the elements in a column of RGA is small ($\ll 1$), then one may consider deleting the corresponding input. Extra outputs: If all elements in a row of RGA are small ($\ll 1$), then the corresponding output cannot be controlled (see Section 10.4).
4. *Diagonal dominance.* The RGA can be used to measure diagonal dominance, by the simple quantity

$$\text{RGA-number} = \|\Lambda(G) - I\|_{\text{sum}} \tag{3.71}$$

For decentralized control we prefer pairings for which the RGA-number at crossover frequencies is close to 1 (see pairing rule 1 on page 435). Similarly, for certain multivariable design methods, shaping, it is simpler to choose the weights and shape the plant if we first rearrange the inputs and outputs to make the plant diagonally dominant with a small RGA-number.
5. *RGA and decentralized control.*

   (a) *Integrity:* For stable plants avoid input-output pairing on negative steady-state RGA-elements. Otherwise, if the sub-controllers are designed independently each with integral action, then the interactions will cause instability either when all of the loops are closed, or when the loop corresponding to the negative relative gain becomes inactive (e.g. because of saturation) (see Theorem 10.4 page 439). Interestingly, this is the only use of the RGA directly related to Bristol's original definition.

   (b) *Stability:* Prefer pairings corresponding to an RGA-number close to 0 at crossover frequencies (see page 435).

Remark. An iterative evaluation of the RGA, $\Lambda^2(G) = \Lambda(\Lambda(G))$ etc., has in applications proved to be useful for choosing pairings for large systems. Wolff (1994) found numerically that

$$\Lambda^\infty \triangleq \lim_{k\to\infty} \Lambda^k(G) \tag{3.72}$$

is a permuted identity matrix (with the exception of "borderline" cases, the result is proved for a positive definite Hermitian matrix G by Johnson and Shapiro (1986)). Typically, Λ^k approaches Λ^∞ for k between 4 and 8. This permuted identity matrix may then be used as a candidate pairing choice. For example, for $G = \begin{bmatrix} 1 & 2 \\ -1 & 1 \end{bmatrix}$ we get $\Lambda = \begin{bmatrix} 0.33 & 0.67 \\ 0.67 & 0.33 \end{bmatrix}$, $\Lambda^2 = \begin{bmatrix} -0.33 & 1.33 \\ 1.33 & -0.33 \end{bmatrix}$, $\Lambda^3 = \begin{bmatrix} -0.07 & 1.07 \\ 1.07 & -0.07 \end{bmatrix}$ and $\Lambda^4 = \begin{bmatrix} 0.00 & 1.00 \\ 1.00 & 0.00 \end{bmatrix}$, which indicates that the off-diagonal pairing should be considered. Note that Λ^∞ may sometimes "recommend" a pairing on negative RGA-elements, even if a positive pairing is possible.

Example 3.9 *Consider a diagonal plant and compute the RGA and condition number,*

$$G = \begin{bmatrix} 100 & 0 \\ 0 & 1 \end{bmatrix}, \; \Lambda(G) = I, \; \gamma(G) = \frac{\bar{\sigma}(G)}{\underline{\sigma}(G)} = \frac{100}{1} = 100, \; \gamma^*(G) = 1 \tag{3.73}$$

Here the condition number is large which means that the plant gain depends strongly on the input direction. However, since the plant is diagonal there are no interactions so $\Lambda(G) = I$ *and* $\gamma^*(G) = 1$*, and no sensitivity to uncertainty (or other control problems) is normally expected.*

Remark. *An exception would be if there was uncertainty caused by unmodelled or neglected off-diagonal elements in G. This would couple the high-gain and low-gain directions, and the large condition number implies sensitivity to this off-diagonal ("unstructured") uncertainty.*

Example 3.10 *Consider a triangular plant G for which we get*

$$G = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}, \; G^{-1} = \begin{bmatrix} 1 & -2 \\ 0 & 1 \end{bmatrix}, \; \Lambda(G) = I, \; \gamma(G) = \frac{2.41}{0.41} = 5.83, \; \gamma^*(G) = 1 \tag{3.74}$$

Note that for a triangular matrix, the RGA is always the identity matrix and $\gamma^*(G)$ *is always* 1.

Example 3.11 *Consider again the distillation process for which we have at steady-state*

$$G = \begin{bmatrix} 87.8 & -86.4 \\ 108.2 & -109.6 \end{bmatrix}, \; G^{-1} = \begin{bmatrix} 0.399 & -0.315 \\ 0.394 & -0.320 \end{bmatrix}, \; \Lambda(G) = \begin{bmatrix} 35.1 & -34.1 \\ -34.1 & 35.1 \end{bmatrix} \tag{3.75}$$

In this case $\gamma(G) = 197.2/1.391 = 141.7$ *is only slightly larger than* $\gamma^*(G) = 138.268$. *The magnitude sum of the elements in the RGA-matrix is* $\|\Lambda\|_{\mathrm{sum}} = 138.275$. *This confirms (A.79) which states that, for* 2×2 *systems,* $\|\Lambda(G)\|_{\mathrm{sum}} \approx \gamma^*(G)$ *when* $\gamma^*(G)$ *is large. The condition number is large, but since the minimum singular value* $\underline{\sigma}(G) = 1.391$ *is larger than* 1 *this does not by itself imply a control problem. However, the large RGA-elements indicate control problems, and fundamental control problems are expected if analysis shows that* $G(j\omega)$ *has large RGA-elements also in the crossover frequency range. (Indeed, the idealized dynamic model (3.82) used below has large RGA-elements at all frequencies, and we will confirm in simulations that there is a strong sensitivity to input channel uncertainty with an inverse-based controller).*

Example 3.12 *Consider a* 3×3 *plant for which we have*

$$G = \begin{bmatrix} 16.8 & 30.5 & 4.30 \\ -16.7 & 31.0 & -1.41 \\ 1.27 & 54.1 & 5.40 \end{bmatrix}, \; \Lambda(G) = \begin{bmatrix} 1.50 & 0.99 & -1.48 \\ -0.41 & 0.97 & 0.45 \\ -0.08 & -0.95 & 2.03 \end{bmatrix} \tag{3.76}$$

and $\gamma = 69.6/1.63 = 42.6$ *and* $\gamma^* = 7.80$. *The magnitude sum of the elements in the RGA is* $\|\Lambda\|_{\mathrm{sum}} = 8.86$ *which is close to* γ^* *as expected from (A.78). Note that the rows and the columns of* Λ *sum to* 1. *Since* $\underline{\sigma}(G)$ *is larger than 1 and the RGA-elements are relatively small, this steady-state analysis does not indicate any particular control problems for the plant.*

Remark. *The plant in (3.76) represents the steady-state model of a fluid catalytic cracking (FCC) process. A dynamic model of the FCC process in (3.76) is given in Exercise 6.16.*

For a detailed analysis of achievable performance of the plant (input-output controllability analysis), one must also consider the singular values, RGA and condition number as functions of frequency. In particular, the crossover frequency range is important. In addition, disturbances and the presence of unstable (RHP) plant poles and zeros must be considered. All these issues are discussed in much more detail in Chapters 5 and 6 where we discuss achievable performance and input-output controllability analysis for SISO and MIMO plants, respectively.

3.7 Introduction to MIMO robustness

To motivate the need for a deeper understanding of robustness, we present two examples which illustrate that MIMO systems can display a sensitivity to uncertainty not found in SISO systems. We focus our attention on diagonal input uncertainty, which is present in any real system and often limits achievable performance because it enters between the controller and the plant.

3.7.1 Motivating robustness example no. 1: Spinning Satellite

Consider the following plant (Doyle, 1986; Packard et al., 1993) which can itself be motivated by considering the angular velocity control of a satellite spinning about one of its principal axes:

$$G(s) = \frac{1}{s^2+a^2}\begin{bmatrix} s-a^2 & a(s+1) \\ -a(s+1) & s-a^2 \end{bmatrix}; \quad a = 10 \tag{3.77}$$

A minimal, state-space realization, $G = C(sI-A)^{-1}B + D$, is

$$\left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right] = \left[\begin{array}{cc|cc} 0 & a & 1 & 0 \\ -a & 0 & 0 & 1 \\ \hline 1 & a & 0 & 0 \\ -a & 1 & 0 & 0 \end{array}\right] \tag{3.78}$$

The plant has a pair of $j\omega$-axis poles at $s = \pm ja$ so it needs to be stabilized. Let us apply negative feedback and try the simple diagonal constant controller

$$K = I$$

The complementary sensitivity function is

$$T(s) = GK(I+GK)^{-1} = \frac{1}{s+1}\begin{bmatrix} 1 & a \\ -a & 1 \end{bmatrix} \tag{3.79}$$

Nominal stability (NS). The closed-loop system has two poles at $s = -1$ and so it is stable. This can be verified by evaluating the closed-loop state matrix

$$A_{cl} = A - BKC = \begin{bmatrix} 0 & a \\ -a & 0 \end{bmatrix} - \begin{bmatrix} 1 & a \\ -a & 1 \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$$

(To derive A_{cl} use $\dot{x} = Ax + Bu$, $y = Cx$ and $u = -Ky$).

Nominal performance (NP). The singular values of $L = GK = G$ are shown in Figure 3.6(a), page 76. We see that $\underline{\sigma}(L) = 1$ at low frequencies and starts dropping off at about $\omega = 10$. Since $\underline{\sigma}(L)$ never exceeds 1, we do not have tight control in the low-gain direction for this plant (recall the discussion following (3.51)), so we expect poor closed-loop performance. This is confirmed by considering S and T. For example, at steady-state $\bar{\sigma}(T) = 10.05$ and $\bar{\sigma}(S) = 10$. Furthermore, the large off-diagonal elements in $T(s)$ in (3.79) show that we have strong interactions in the closed-loop system. (For reference tracking, however, this may be counteracted by use of a two degrees-of-freedom controller).

Robust stability (RS). Now let us consider stability robustness. In order to determine stability margins with respect to perturbations in each input channel, one may consider Figure 3.11 where we have broken the loop at the first input. The loop transfer function at this point (the transfer function from w_1 to z_1) is $L_1(s) = 1/s$ (which can be derived from $t_{11}(s) = \frac{1}{1+s} = \frac{L_1(s)}{1+L_1(s)}$). This corresponds to an infinite gain margin and a phase margin of 90°. On breaking the loop at the second input we get the same result. This suggests good robustness properties irrespective of the value of a. However, the design is far from robust as a further analysis shows. Consider

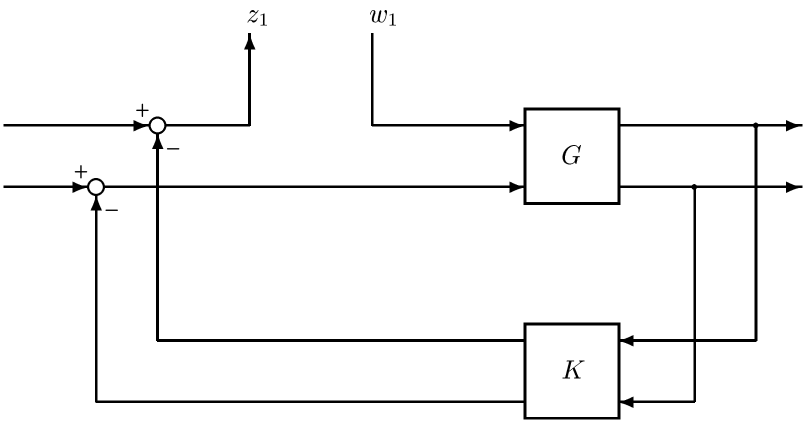


Figure 3.11: Checking stability margins "one-loop-at-a-time"

input gain uncertainty, and let ϵ_1 and ϵ_2 denote the relative error in the gain in each input channel. Then

$$u_1' = (1 + \epsilon_1)u_1, \quad u_2' = (1 + \epsilon_2)u_2 \tag{3.80}$$

where u_1' and u_2' are the actual changes in the manipulated inputs, while u_1 and u_2 are the desired changes as computed by the controller. It is important to stress that this diagonal input uncertainty, which stems from our inability to know the exact values of the manipulated inputs, is *always* present. In terms of a state-space description, (3.80) may be represented by replacing B by

$$B' = \begin{bmatrix} 1+\epsilon_1 & 0 \\ 0 & 1+\epsilon_2 \end{bmatrix}$$

The corresponding closed-loop state matrix is

$$A'_{cl} = A - B'KC = \begin{bmatrix} 0 & a \\ -a & 0 \end{bmatrix} - \begin{bmatrix} 1+\epsilon_1 & 0 \\ 0 & 1+\epsilon_2 \end{bmatrix} \begin{bmatrix} 1 & a \\ -a & 1 \end{bmatrix}$$

which has a characteristic polynomial given by

$$\det(sI - A'_{cl}) = s^2 + \underbrace{(2+\epsilon_1+\epsilon_2)}_{a_1} s + \underbrace{1+\epsilon_1+\epsilon_2+(a^2+1)\epsilon_1\epsilon_2}_{a_0} \tag{3.81}$$

The perturbed system is stable if and only if both the coefficients a_0 and a_1 are positive. We therefore see that *the system is always stable if we consider uncertainty in only one channel at a time* (at least as long as the channel gain is positive). More precisely, we have stability for $(-1 < \epsilon_1 < \infty, \epsilon_2 = 0)$ and $(\epsilon_1 = 0, -1 < \epsilon_2 < \infty)$. This confirms the infinite gain margin seen earlier. However, the system can only tolerate *small simultaneous changes* in the two channels. For example, let $\epsilon_1 = -\epsilon_2$, then the system is unstable ($a_0 < 0$) for

$$|\epsilon_1| > \frac{1}{\sqrt{a^2+1}} \approx 0.1$$

In summary, we have found that checking single-loop margins is inadequate for MIMO problems. We have also observed that large values of $\bar{\sigma}(T)$ or $\bar{\sigma}(S)$ indicate robustness problems. We will return to this in Chapter 8, where we show that with input uncertainty of magnitude $|\epsilon_i| < 1/\bar{\sigma}(T)$, we are guaranteed robust stability (even for “full-block complex perturbations”).

In the next example we find that there can be sensitivity to diagonal input uncertainty even in cases where $\bar{\sigma}(T)$ and $\bar{\sigma}(S)$ have no large peaks. This can not happen for a diagonal controller, see (6.77), but it will happen if we use an inverse-based controller for a plant with large RGA-elements, see (6.78).

3.7.2 Motivating robustness example no. 2: Distillation Process

The following is an idealized dynamic model of a distillation column,

$$G(s) = \frac{1}{75s+1} \begin{bmatrix} 87.8 & -86.4 \\ 108.2 & -109.6 \end{bmatrix} \tag{3.82}$$

(time is in minutes). The physics of this example was discussed in Example 3.7. The plant is ill-conditioned with condition number $\gamma(G) = 141.7$ at all frequencies. The plant is also strongly two-way interactive and the RGA-matrix at all frequencies is

$$\mathrm{RGA}(G) = \begin{bmatrix} 35.1 & -34.1 \\ -34.1 & 35.1 \end{bmatrix} \tag{3.83}$$

The large elements in this matrix indicate that this process is fundamentally difficult to control.

Remark. (3.82) is admittedly a very crude model of a real distillation column; there should be a high-order lag in the transfer function from input 1 to output 2 to represent the liquid flow down to the column, and higher-order composition dynamics should also be included. Nevertheless, the model is simple and displays important features of distillation column behaviour. It should be noted that with a more detailed model, the RGA-elements would approach 1 at frequencies around 1 rad/min, indicating less of a control problem.

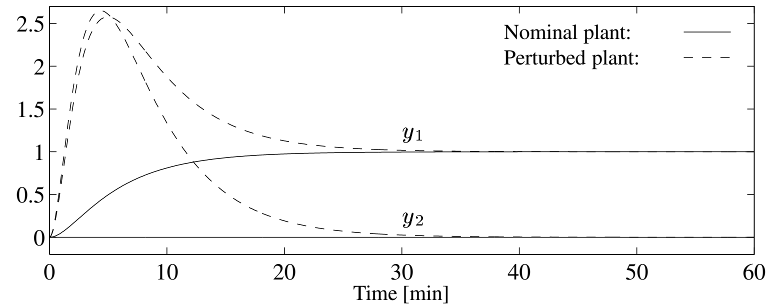


Figure 3.12: Response with decoupling controller to filtered reference input $r_1 = 1/(5s+1)$. The perturbed plant has 20% gain uncertainty as given by (3.86).

We consider the following inverse-based controller, which may also be looked upon as a steady-state decoupler with a PI controller:

$$K_{\mathrm{inv}}(s) = \frac{k_1}{s} G^{-1}(s) = \frac{k_1(1+75s)}{s} \begin{bmatrix} 0.3994 & -0.3149 \\ 0.3943 & -0.3200 \end{bmatrix}, \quad k_1 = 0.7 \tag{3.84}$$

Nominal performance (NP). We have $GK_{\mathrm{inv}} = K_{\mathrm{inv}}G = \frac{0.7}{s}I$. With no model error this controller should counteract all the interactions in the plant and give rise to two decoupled first-order responses each with a time constant of $1/0.7 = 1.43$ min. This is confirmed by the solid line in Figure 3.12 which shows the simulated response to a reference change in y_1. The responses are clearly acceptable, and we conclude that *nominal performance (NP) is achieved with the decoupling controller*.

Robust stability (RS). The resulting sensitivity and complementary sensitivity functions with this controller are

$$S = S_I = \frac{s}{s+0.7}I; \quad T = T_I = \frac{1}{1.43s+1}I \tag{3.85}$$

Thus, $\bar{\sigma}(S)$ and $\bar{\sigma}(T)$ are both less than 1 at all frequencies, so there are no peaks which would indicate robustness problems. We also find that this controller gives an infinite gain margin (GM) and a phase margin (PM) of $90°$ in each channel. Thus, use of the traditional margins and the peak values of S and T indicate no robustness problems. However, from the large RGA-elements there is cause for concern, and this is confirmed in the following.

We consider again the input gain uncertainty (3.80) as in the previous example, and we select $\epsilon_1 = 0.2$ and $\epsilon_2 = -0.2$. We then have

$$u_1' = 1.2u_1, \quad u_2' = 0.8u_2 \tag{3.86}$$

Note that the uncertainty is on the *change* in the inputs (flow rates), and not on their absolute values. A 20% error is typical for process control applications (see Remark 2 on page 300). The uncertainty in (3.86) does not by itself yield instability. This is verified by computing the closed-loop poles, which, assuming no cancellations, are solutions to $\det(I + L(s)) = \det(I + L_I(s)) = 0$ (see (4.102) and (A.12)). In our case

$$L_I'(s) = K_{\text{inv}} G' = K_{\text{inv}} G \begin{bmatrix} 1+\epsilon_1 & 0 \\ 0 & 1+\epsilon_2 \end{bmatrix} = \frac{0.7}{s} \begin{bmatrix} 1+\epsilon_1 & 0 \\ 0 & 1+\epsilon_2 \end{bmatrix}$$

so the perturbed closed-loop poles are

$$s_1 = -0.7(1+\epsilon_1), \quad s_2 = -0.7(1+\epsilon_2) \tag{3.87}$$

and we have closed-loop stability as long as the input gains $1+\epsilon_1$ and $1+\epsilon_2$ remain positive, so we can have up to 100% error in each input channel. We thus conclude that *we have robust stability (RS) with respect to input gain errors for the decoupling controller*.

Robust performance (RP). For SISO systems we generally have that nominal performance (NP) and robust stability (RS) imply robust performance (RP), but this is not the case for MIMO systems. This is clearly seen from the dotted lines in Figure 3.12 which show the closed-loop response of the perturbed system. It differs drastically from the nominal response represented by the solid line, and even though it is stable, the response is clearly not acceptable; it is no longer decoupled, and $y_1(t)$ and $y_2(t)$ reach a value of about 2.5 before settling at their desired values of 1 and 0. *Thus RP is not achieved by the decoupling controller.*

Remark 1 There is a simple reason for the observed poor response to the reference change in y_1. To accomplish this change, which occurs mostly in the direction corresponding to the low plant gain, the inverse-based controller generates relatively *large* inputs u_1 and u_2, while trying to keep $u_1 - u_2$ very *small*. However, the input uncertainty makes this impossible – the result is an undesired *large* change in the actual value of $u_1' - u_2'$, which subsequently results in large changes in y_1 and y_2 because of the large plant gain ($\bar{\sigma}(G) = 197.2$) in this direction, as seen from (3.46).

Remark 2 The system remains stable for gain uncertainty up to 100% because the uncertainty occurs only at one side of the plant (at the input). If we also consider uncertainty at the output then we find that the decoupling controller yields instability for relatively small errors in the input and output gains. This is illustrated in Exercise 3.8 below.

Remark 3 It is also difficult to get a robust controller with other standard design techniques for this model. For example, an S/KS-design as in (3.59) with $W_P = w_P I$ (using $M = 2$ and $\omega_B = 0.05$ in the performance weight (3.60)) and $W_u = I$, yields a good nominal response (although not decoupled), but the system is very sensitive to input uncertainty, and the outputs go up to about 3.4 and settle very slowly when there is 20% input gain error.

Remark 4 Attempts to make the inverse-based controller robust using the second step of the Glover-McFarlane $\mathcal{H}_\infty$ loop-shaping procedure are also unhelpful; see Exercise 3.9. This shows that robustness with respect to coprime factor uncertainty does not necessarily imply robustness with respect to input uncertainty. In any case, the solution is to avoid inverse-based controllers for a plant with large RGA-elements.

Exercise 3.7 *Design a SVD-controller $K = W_1 K_s W_2$ for the distillation process in (3.82), i.e. select $W_1 = V$ and $W_2 = U^T$ where U and V are given in (3.46). Select K_s in the form*

$$K_s = \begin{bmatrix} c_1 \frac{75s+1}{s} & 0 \\ 0 & c_2 \frac{75s+1}{s} \end{bmatrix}$$

and try the following values:

(a) $c_1 = c_2 = 0.005$;
(b) $c_1 = 0.005$, $c_2 = 0.05$;
(c) $c_1 = 0.7/197 = 0.0036$, $c_2 = 0.7/1.39 = 0.504$.

Simulate the closed-loop reference response with and without uncertainty. Designs (a) and (b) should be robust. Which has the best performance? Design (c) should give the response in Figure 3.12. In the simulations, include high-order plant dynamics by replacing $G(s)$ by $\frac{1}{(0.02s+1)^5}G(s)$. What is the condition number of the controller in the three cases? Discuss the results. (See also the conclusion on page 244).

Exercise 3.8 *Consider again the distillation process (3.82) with the decoupling controller, but also include output gain uncertainty $\widehat{\epsilon}_i$. That is, let the perturbed loop transfer function be*

$$L'(s) = G' K_{\rm inv} = \frac{0.7}{s} \underbrace{\begin{bmatrix} 1+\widehat{\epsilon}_1 & 0 \\ 0 & 1+\widehat{\epsilon}_2 \end{bmatrix} G \begin{bmatrix} 1+\epsilon_1 & 0 \\ 0 & 1+\epsilon_2 \end{bmatrix} G^{-1}}_{L_0} \tag{3.88}$$

where L_0 is a constant matrix for the distillation model (3.82), since all elements in G share the same dynamics, $G(s) = g(s)G_0$. The closed-loop poles of the perturbed system are solutions to $\det(I + L'(s)) = \det(I + (k_1/s)L_0) = 0$, *or equivalently*

$$\det(\frac{s}{k_1} I + L_0) = (s/k_1)^2 + \mathrm{tr}(L_0)(s/k_1) + \det(L_0) = 0 \tag{3.89}$$

For $k_1 > 0$ we have from the Routh-Hurwitz stability condition indexRouth-Hurwitz stability test that instability occurs if and only if the trace and/or the determinant of L_0 are negative.

Since $\det(L_0) > 0$ *for any gain error less than* 100%, *instability can only occur if* $\mathrm{tr}(L_0) < 0$. *Evaluate* $\mathrm{tr}(L_0)$ *and show that with gain errors of equal magnitude the combination of errors which most easily yields instability is with* $\widehat{\epsilon}_1 = -\widehat{\epsilon}_2 = -\epsilon_1 = \epsilon_2 = \epsilon$. *Use this to show that the perturbed system is unstable if*

$$|\epsilon| > \sqrt{\frac{1}{2\lambda_{11} - 1}} \tag{3.90}$$

where $\lambda_{11} = g_{11}g_{22}/\det G$ *is the* $1,1$*-element of the RGA of* G. *In our case* $\lambda_{11} = 35.1$ *and we get instability for* $|\epsilon| > 0.120$. *Check this numerically, e.g. using MATLAB.*

Remark. The instability condition in (3.90) for simultaneous input and output gain uncertainty, applies to the very special case of a 2×2 plant, in which all elements share the same dynamics, $G(s) = g(s)G_0$, and an inverse-based controller, $K(s) = (k_1/s)G^{-1}(s)$.

Exercise 3.9 *Consider again the distillation process* $G(s)$ *in (3.82). The response using the inverse-based controller* K_{inv} *in (3.84) was found to be sensitive to input gain errors. We want to see if the controller can be modified to yield a more robust system by using the Glover-McFarlane* $\mathcal{H}_\infty$ *loop-shaping procedure. To this effect, let the shaped plant be* $G_s = GK_{\mathrm{inv}}$, *i.e.* $W_1 = K_{\mathrm{inv}}$, *and design an* $\mathcal{H}_\infty$ *controller* K_s *for the shaped plant (see page 382 and Chapter 9), such that the overall controller becomes* $K = K_{\mathrm{inv}}K_s$. *(You will find that* $\gamma_{min} = 1.414$ *which indicates good robustness with respect to coprime factor uncertainty, but the loop shape is almost unchanged and the system remains sensitive to input uncertainty.)*

3.7.3 Robustness conclusions

From the two motivating examples above we found that multivariable plants can display a sensitivity to uncertainty (in this case input uncertainty) which is fundamentally different from what is possible in SISO systems.

In the first example (spinning satellite), we had excellent stability margins (PM and GM) when considering one loop at a time, but small simultaneous input gain errors gave instability. This might have been expected from the peak values ($\mathcal{H}_\infty$ norms) of S and T, defined as

$$\|T\|_\infty = \max_\omega \bar{\sigma}(T(j\omega)), \quad \|S\|_\infty = \max_\omega \bar{\sigma}(S(j\omega)) \tag{3.91}$$

which were both large (about 10) for this example.

In the second example (distillation process), we again had excellent stability margins (PM and GM), and the system was also robustly stable to errors (even simultaneous) of up to 100% in the input gains. However, in this case small input gain errors gave very poor output performance, so robust performance was not satisfied, and adding simultaneous output gain uncertainty resulted in instability (see Exercise 3.8). These problems with the decoupling controller might have been expected because the plant has large RGA-elements. For this second example the $\mathcal{H}_\infty$ norms of S and T were both about 1, so the absence of peaks in S and T does not guarantee robustness.

Although sensitivity peaks, RGA-elements, etc. are useful indicators of robustness problems, they provide no exact answer to whether a given source of uncertainty will yield instability or poor performance. This motivates the need for better tools for analyzing the effects of model uncertainty. We want to avoid a trial-and-error procedure based on checking stability and performance for a large number of candidate plants. This is very time consuming, and in the end one does not know whether those plants are the limiting ones. What is desired, is a simple tool which is able to identify the worst-case plant. This will be the focus of Chapters 7 and 8 where we show how to represent model uncertainty in the $\mathcal{H}_\infty$ framework, and introduce the structured singular value μ as our tool. The two motivating examples are studied in more detail in Example 8.10 and Section 8.11.3 where a μ-analysis predicts the robustness problems found above.

3.8 General control problem formulation

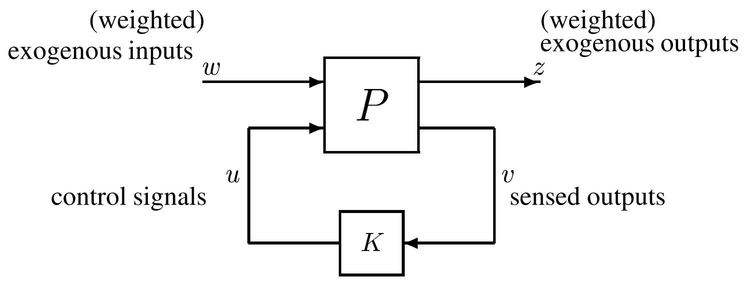


Figure 3.13: General control configuration for the case with no model uncertainty

In this section we consider a general method of formulating control problems introduced by Doyle (1983; 1984). The formulation makes use of the general control configuration in Figure 3.13, where P is the generalized plant and K is the generalized controller as explained in Table 1.1 on page 13. Note that positive feedback is used.

The overall control objective is to minimize some norm of the transfer function from w to z, for example, the $\mathcal{H}_\infty$ norm. The controller design problem is then:

- Find a controller K which based on the information in v, generates a control signal u which counteracts the influence of w on z, thereby minimizing the closed-loop norm from w to z.

The most important point of this section is to appreciate that almost any linear control problem can be formulated using the block diagram in Figure 3.13 (for the nominal case) or in Figure 3.21 (with model uncertainty).

Remark 1 The configuration in Figure 3.13 may at first glance seem restrictive. However, this is not the case, and we will demonstrate the generality of the setup with a few examples, including the design of observers (the estimation problem) and feedforward controllers.

Remark 2 We may generalize the control configuration still further by including diagnostics as additional outputs from the controller giving the *4-parameter controller* introduced by Nett (1986), but this is not considered in this book.

3.8.1 Obtaining the generalized plant P

The routines in MATLAB for synthesizing $\mathcal{H}_\infty$ and $\mathcal{H}_2$ optimal controllers assume that the problem is in the general form of Figure 3.13, that is, they assume that P is given. To derive P (and K) for a specific case we must first find a block diagram representation and identify the signals w, z, u and v. To construct P one should note that it is an *open-loop* system and remember to break all "loops" entering and exiting the controller K. Some examples are given below and further examples are given in Section 9.3 (Figures 9.9, 9.10, 9.11 and 9.12).

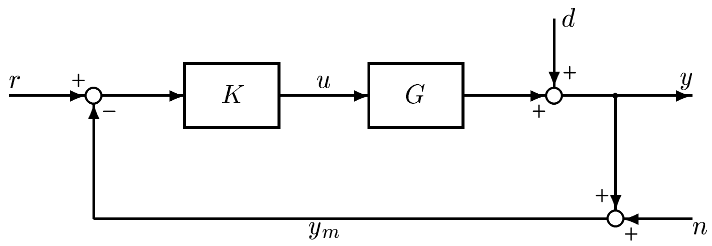


Figure 3.14: One degree-of-freedom control configuration

Example 3.13 **One degree-of-freedom feedback control configuration.** *We want to find P for the conventional one degree-of-freedom control configuration in Figure 3.14. The first step is to identify the signals for the generalized plant:*

$$w = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} = \begin{bmatrix} d \\ r \\ n \end{bmatrix}; \quad z = e = y - r; \quad v = r - y_m = r - y - n \tag{3.92}$$

With this choice of v, the controller only has information about the deviation $r - y_m$. Also note that $z = y - r$, which means that performance is specified in terms of the actual output y and not *in terms of the measured output y_m. The block diagram in Figure 3.14 then yields*

$$\begin{aligned} z &= y - r = Gu + d - r = Iw_1 - Iw_2 + 0w_3 + Gu \\ v &= r - y_m = r - Gu - d - n = -Iw_1 + Iw_2 - Iw_3 - Gu \end{aligned}$$

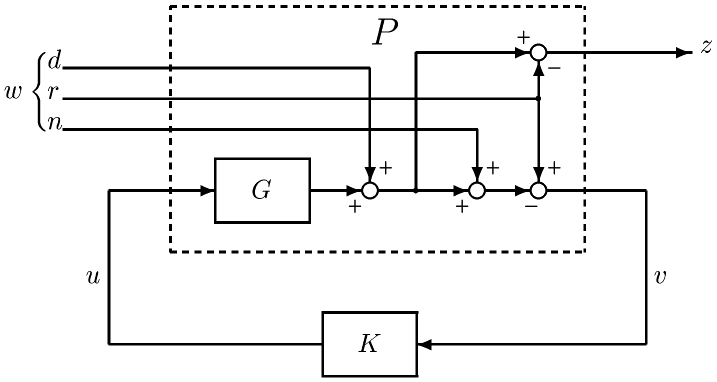


Figure 3.15: Equivalent representation of Figure 3.14 where the error signal to be minimized is $z = y - r$ and the input to the controller is $v = r - y_m$

and P which represents the transfer function matrix from $[\, w \quad u \,]^T$ to $[\, z \quad v \,]^T$ is

$$P = \begin{bmatrix} I & -I & 0 & G \\ -I & I & -I & -G \end{bmatrix} \tag{3.93}$$

Note that P does not *contain the controller. Alternatively, P can be obtained by inspection from the representation in Figure 3.15.*

Remark. Obtaining the generalized plant P may seem tedious. However, when performing numerical calculations P can be generated using software. For example, in MATLAB we may use the `simulink` program, or we may use the `sysic` program in the μ-toolbox. The code in Table 3.1 generates the generalized plant P in (3.93) for Figure 3.14.

Table 3.1: **MATLAB program to generate P in (3.93)**

```
% Uses the Mu-toolbox
systemnames = 'G';                                   % G is the SISO plant.
inputvar = '[d(1);r(1);n(1);u(1)]';                  % Consists of vectors w and u.
input_to_G = '[u]';
outputvar = '[G+d-r; r-G-d-n]';                      % Consists of vectors z and v.
sysoutname = 'P';
sysic;
```

3.8.2 Controller design: Including weights in P

To get a meaningful controller synthesis problem, for example, in terms of the $\mathcal{H}_\infty$ or $\mathcal{H}_2$ norms, we generally have to include weights W_z and W_w in the generalized plant P, see Figure 3.16. That is, we consider the weighted or normalized exogenous inputs w (where $\widetilde{w} = W_w w$ consists of the "physical" signals entering the system;

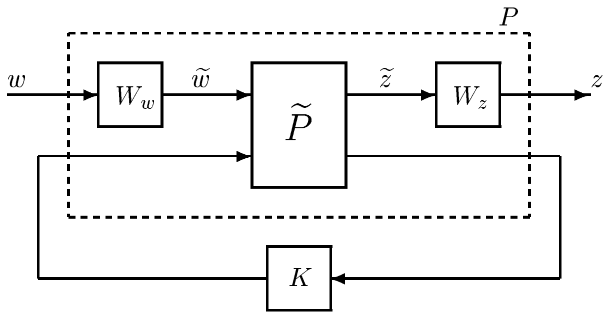


Figure 3.16: General control configuration for the case with no model uncertainty

disturbances, references and noise), and the weighted or normalized controlled outputs $z = W_z\widetilde{z}$ (where $\widetilde{z}$ often consists of the control error $y - r$ and the manipulated input u). The weighting matrices are usually frequency dependent and typically selected such that weighted signals w and z are of magnitude 1, that is, the norm from w to z should be less than 1. Thus, in most cases only the magnitude of the weights matter, and we may without loss of generality assume that $W_w(s)$ and $W_z(s)$ are stable and minimum phase (they need not even be rational transfer functions but if not they will be unsuitable for controller synthesis using current software).

Example 3.14 **Stacked $S/T/KS$ problem.** *Consider an $\mathcal{H}_\infty$ problem where we want to bound $\bar{\sigma}(S)$ (for performance), $\bar{\sigma}(T)$ (for robustness and to avoid sensitivity to noise) and $\bar{\sigma}(KS)$ (to penalize large inputs). These requirements may be combined into a stacked $\mathcal{H}_\infty$ problem*

$$\min_K \|N(K)\|_\infty, \quad N = \begin{bmatrix} W_uKS \\ W_TT \\ W_PS \end{bmatrix} \tag{3.94}$$

where K is a stabilizing controller. In other words, we have $z = Nw$ and the objective is to minimize the $\mathcal{H}_\infty$ norm from w to z. Except for some negative signs which have no effect when evaluating $\|N\|_\infty$, the N in (3.94) may be represented by the block diagram in Figure 3.17 (convince yourself that this is true). Here w represents a reference command ($w = -r$, where the negative sign does not really matter) or a disturbance entering at the output ($w = d_y$), and z consists of the weighted input $z_1 = W_uu$, the weighted output $z_2 = W_Ty$, and the weighted control error $z_3 = W_P(y - r)$. We get from Figure 3.17 the following set of equations

$$\begin{aligned} z_1 &= W_uu \\ z_2 &= W_TGu \\ z_3 &= W_Pw + W_PGu \\ v &= -w - Gu \end{aligned}$$

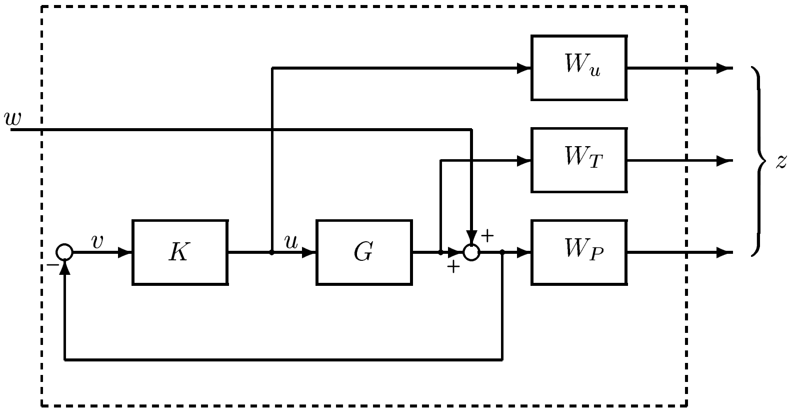


Figure 3.17: Block diagram corresponding to $z = Nw$ in (3.94)

so the generalized plant P from $[\,w \quad u\,]^T$ *to* $[\,z \quad v\,]^T$ *is*

$$P = \begin{bmatrix} 0 & W_u I \\ 0 & W_T G \\ W_P I & W_P G \\ -I & -G \end{bmatrix} \tag{3.95}$$

3.8.3 Partitioning the generalized plant P

We often partition P as

$$P = \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} \tag{3.96}$$

such that its parts are compatible with the signals w, z, u and v in the generalized control configuration,

$$z = P_{11} w + P_{12} u \tag{3.97}$$

$$v = P_{21} w + P_{22} u \tag{3.98}$$

The reader should become familiar with this notation. In Example 3.14 we get

$$P_{11} = \begin{bmatrix} 0 \\ 0 \\ W_P I \end{bmatrix}, \quad P_{12} = \begin{bmatrix} W_u I \\ W_T G \\ W_P G \end{bmatrix} \tag{3.99}$$

$$P_{21} = -I, \quad P_{22} = -G \tag{3.100}$$

Note that P_{22} has dimensions compatible with the controller, i.e. if K is an $n_u \times n_v$ matrix, then P_{22} is an $n_v \times n_u$ matrix. For cases with one degree-of-freedom negative feedback control we have $P_{22} = -G$.

3.8.4 Analysis: Closing the loop to get N

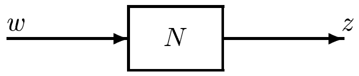


Figure 3.18: General block diagram for analysis with no uncertainty

The general feedback configurations in Figures 3.13 and 3.16 have the controller K as a separate block. This is useful when synthesizing the controller. However, for *analysis* of closed-loop performance the controller is given, and we may absorb K into the interconnection structure and obtain the system N as shown in Figure 3.18 where

$$z = Nw \tag{3.101}$$

where N is a function of K. To find N, first partition the generalized plant P as given in (3.96)-(3.98), combine this with the controller equation

$$u = Kv \tag{3.102}$$

and eliminate u and v from equations (3.97), (3.98) and (3.102) to yield $z = Nw$ where N is given by

$$N = P_{11} + P_{12}K(I - P_{22}K)^{-1}P_{21} \triangleq F_l(P, K) \tag{3.103}$$

Here $F_l(P, K)$ denotes a lower *linear fractional transformation (LFT)* of P with K as the parameter. Some properties of LFTs are given in Appendix A.7. In words, N is obtained from Figure 3.13 by using K to close a lower feedback loop around P. Since positive feedback is used in the general configuration in Figure 3.13 the term $(I - P_{22}K)^{-1}$ has a negative sign.

Remark. To assist in remembering the sequence of P_{12} and P_{21} in (3.103), notice that the first (last) index in P_{11} is the same as the first (last) index in $P_{12}K(I - P_{22}K)^{-1}P_{21}$. The lower LFT in (3.103) is also represented by the block diagram in Figure 3.2.

The reader is advised to become comfortable with the above manipulations before progressing much further.

Example 3.15 *We want to derive N for the partitioned P in (3.99) and (3.100) using the LFT-formula in (3.103). We get*

$$N = \begin{bmatrix} 0 \\ 0 \\ W_P I \end{bmatrix} + \begin{bmatrix} W_u I \\ W_T G \\ W_P G \end{bmatrix} K(I + GK)^{-1}(-I) = \begin{bmatrix} -W_u KS \\ -W_T T \\ W_P S \end{bmatrix}$$

where we have made use of the identities $S = (I + GK)^{-1}$, $T = GKS$ and $I - T = S$. With the exception of the two negative signs, this is identical to N given in (3.94). Of course, the negative signs have no effect on the norm of N.

Again, it should be noted that deriving N from P is much simpler using available software. For example in the MATLAB μ-Toolbox we can evaluate $N = F_l(P, K)$ using the command `N=starp(P,K)`. Here `starp` denotes the matrix star product which generalizes the use of LFTs (see Appendix A.7.5).

Exercise 3.10 *Consider the two degrees-of-freedom feedback configuration in Figure 1.3(b). (i) Find P when*

$$w = \begin{bmatrix} d \\ r \\ n \end{bmatrix}; \quad z = \begin{bmatrix} y - r \\ u \end{bmatrix}; \quad v = \begin{bmatrix} r \\ y_m \end{bmatrix} \tag{3.104}$$

(ii) Let $z = Nw$ and derive N in two different ways; directly from the block diagram and using $N = F_l(P, K)$.

3.8.5 Generalized plant P: Further examples

To illustrate the generality of the configuration in Figure 3.13, we now present two further examples: one in which we derive P for a problem involving feedforward control, and one for a problem involving estimation.

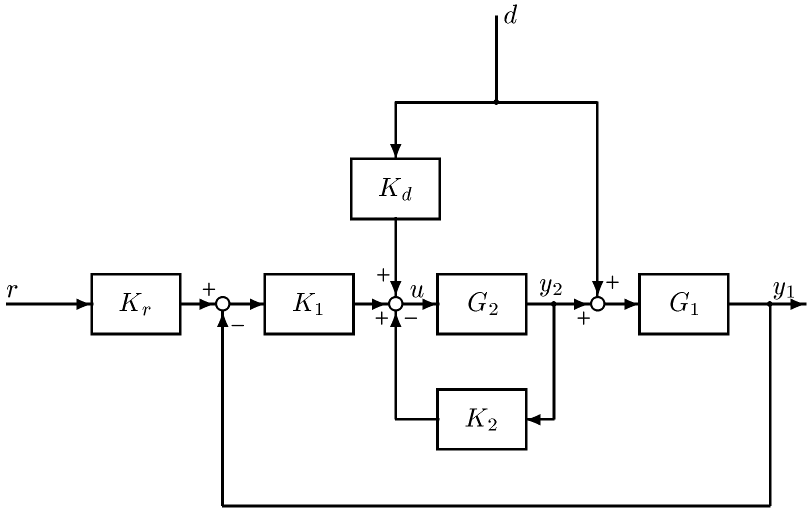


Figure 3.19: System with feedforward, local feedback and two degrees-of-freedom control

Example 3.16 *Consider the control system in Figure 3.19, where y_1 is the output we want to control, y_2 is a secondary output (extra measurement), and we also measure the disturbance d. By secondary we mean that y_2 is of secondary importance for control, that is, there is no control objective associated with it. The control configuration includes a two degrees-of-freedom controller, a feedforward controller and a local feedback controller based on the extra* *measurement y_2. To recast this into our standard configuration of Figure 3.13 we define*

$$w = \begin{bmatrix} d \\ r \end{bmatrix}; \quad z = y_1 - r; \quad v = \begin{bmatrix} r \\ y_1 \\ y_2 \\ d \end{bmatrix} \tag{3.105}$$

Note that d and r are both inputs and outputs to P and we have assumed a perfect measurement of the disturbance d. Since the controller has explicit information about r we have a two degrees-of-freedom controller. The generalized controller K may be written in terms of the individual controller blocks in Figure 3.19 as follows:

$$K = [\, K_1 K_r \quad -K_1 \quad -K_2 \quad K_d \,] \tag{3.106}$$

By writing down the equations or by inspection from Figure 3.19 we get

$$P = \begin{bmatrix} G_1 & -I & G_1 G_2 \\ 0 & I & 0 \\ G_1 & 0 & G_1 G_2 \\ 0 & 0 & G_2 \\ I & 0 & 0 \end{bmatrix} \tag{3.107}$$

Then partitioning P as in (3.97) and (3.98) yields $P_{22} = [\, 0^T \quad (G_1 G_2)^T \quad G_2^T \quad 0^T \,]^T$.

Exercise 3.11 Cascade implementation. *Consider further Example 3.16. The local feedback based on y_2 is often implemented in a cascade manner; see also Figure 10.4. In this case the output from K_1 enters into K_2 and it may be viewed as a reference signal for y_2. Derive the generalized controller K and the generalized plant P in this case.*

Remark. From Example 3.16 and Exercise 3.11, we see that a cascade *implementation* does not usually limit the achievable performance since, unless the optimal K_2 or K_1 have RHP-zeros, we can obtain from the optimal overall K the subcontrollers K_2 and K_1 (although we may have to add a small D-term to K to make the controllers proper). However, if we impose restrictions on the *design* such that, for example K_2 or K_1 are designed "locally" (without considering the whole problem), then this will limit the achievable performance. For example, for a *two degrees-of-freedom controller* a common approach is to first design the feedback controller K_y for disturbance rejection (without considering reference tracking) and then design K_r for reference tracking. This will generally give some performance loss compared to a simultaneous design of K_y and K_r.

Example 3.17 Output estimator. *Consider a situation where we have no measurement of the output y which we want to control. However, we do have a measurement of another output variable y_2. Let d denote the unknown external inputs (including noise and disturbances) and u_G the known plant inputs (a subscript $_G$ is used because in this case the output u from K is not the plant input). Let the model be*

$$y = G u_G + G_d d; \quad y_2 = F u_G + F_d d$$

The objective is to design an estimator, K_{est}, such that the estimated output $\widehat{y} = K_{\text{est}} \begin{bmatrix} y_2 \\ u_G \end{bmatrix}$ is as close as possible in some sense to the true output y; see Figure 3.20. This problem may be written in the general framework of Figure 3.13 with

$$w = \begin{bmatrix} d \\ u_G \end{bmatrix},\ u = \widehat{y},\ z = y - \widehat{y},\ v = \begin{bmatrix} y_2 \\ u_G \end{bmatrix}$$

Note that $u = \widehat{y}$, that is, the output u from the generalized controller is the estimate of the plant output. Furthermore, $K = K_{\text{est}}$ and

$$P = \begin{bmatrix} G_d & G & -I \\ F_d & F & 0 \\ 0 & I & 0 \end{bmatrix} \tag{3.108}$$

We see that $P_{22} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ since the estimator problem does not involve feedback.

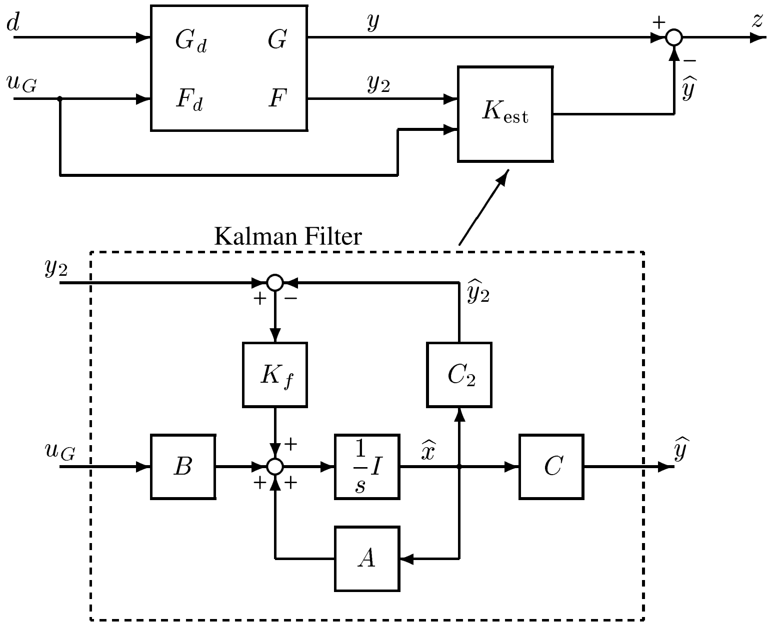


Figure 3.20: Output estimation problem. One particular estimator K_{est} is a Kalman Filter

Exercise 3.12 State estimator (observer). *In the Kalman filter problem studied in Section 9.2 the objective is to minimize $x - \widehat{x}$ (whereas in Example 3.17 the objective was to minimize $y - \widehat{y}$). Show how the Kalman filter problem can be represented in the general configuration of Figure 3.13 and find P.*

3.8.6 Deriving P from N

For cases where N is given and we wish to find a P such that

$$N = F_l(P, K) = P_{11} + P_{12}K(I - P_{22}K)^{-1}P_{21}$$

it is usually best to work from a block diagram representation. This was illustrated above for the stacked N in (3.94). Alternatively, the following procedure may be useful:

1. Set $K = 0$ in N to obtain P_{11}.
2. Define $Q = N - P_{11}$ and rewrite Q such that each term has a common factor $R = K(I - P_{22}K)^{-1}$ (this gives P_{22}).
3. Since $Q = P_{12}RP_{21}$, we can now usually obtain P_{12} and P_{21} by inspection.

Example 3.18 Weighted sensitivity. *We will use the above procedure to derive P when $N = w_P S = w_P(I + GK)^{-1}$, where w_P is a scalar weight.*

1. $P_{11} = N(K = 0) = w_P I$.
2. $Q = N - w_P I = w_P(S - I) = -w_P T = -w_P GK(I + GK)^{-1}$, and we have $R = K(I + GK)^{-1}$ so $P_{22} = -G$.
3. $Q = -w_P GR$ so we have $P_{12} = -w_P G$ and $P_{21} = I$, and we get

$$P = \begin{bmatrix} w_P I & -w_P G \\ I & -G \end{bmatrix} \tag{3.109}$$

Remark. When obtaining P from a given N, we have that P_{11} and P_{22} are unique, whereas from Step 3 in the above procedure we see that P_{12} and P_{21} are not unique. For instance, let α be a real scalar then we may instead choose $\widetilde{P}_{12} = \alpha P_{12}$ and $\widetilde{P}_{21} = (1/\alpha)P_{21}$. For P in (3.109) this means that we may move the negative sign of the scalar w_P from P_{12} to P_{21}.

Exercise 3.13 Mixed sensitivity. *Use the above procedure to derive the generalized plant P for the stacked N in (3.94).*

3.8.7 Problems not covered by the general formulation

The above examples have demonstrated the generality of the control configuration in Figure 3.13. Nevertheless, there are some controller design problems which are not covered. Let N be some closed-loop transfer function whose norm we want to minimize. To use the general form we must first obtain a P such that $N = F_l(P, K)$. However, this is not always possible, since there may not exist a block diagram representation for N. As a simple example, consider the stacked transfer function

$$N = \begin{bmatrix} (I + GK)^{-1} \\ (I + KG)^{-1} \end{bmatrix} \tag{3.110}$$

The transfer function $(I + GK)^{-1}$ may be represented on a block diagram with the input and output signals *after* the plant, whereas $(I + KG)^{-1}$ may be represented by another block diagram with input and output signals *before* the plant. However, in N there are no cross coupling terms between an input before the plant and an output after the plant (corresponding to $G(I+KG)^{-1}$), or between an input after the plant and an output before the plant (corresponding to $-K(I+GK)^{-1}$) so N cannot be represented in block diagram form. Equivalently, if we apply the procedure in Section 3.8.6 to N in (3.110), we are not able to find solutions to P_{12} and P_{21} in Step 3.

Another stacked transfer function which *cannot* in general be represented in block diagram form is

$$N = \begin{bmatrix} W_P S \\ S G_d \end{bmatrix} \tag{3.111}$$

Remark. The case where N cannot be written as an LFT of K, is a special case of the Hadamard weighted $\mathcal{H}_\infty$ problem studied by van Diggelen and Glover (1994a). Although the solution to this $\mathcal{H}_\infty$ problem remains intractable, van Diggelen and Glover (1994b) present a solution for a similar problem where the Frobenius norm is used instead of the singular value to "sum up the channels".

Exercise 3.14 *Show that N in (3.111) can be represented in block diagram form if $W_P = w_P I$ where w_P is a scalar.*

3.8.8 A general control configuration including model uncertainty

The general control configuration in Figure 3.13 may be extended to include model uncertainty as shown by the block diagram in Figure 3.21. Here the matrix Δ is a *block-diagonal* matrix that includes all possible perturbations (representing uncertainty) to the system. It is usually normalized in such a way that $\|\Delta\|_\infty \leq 1$.

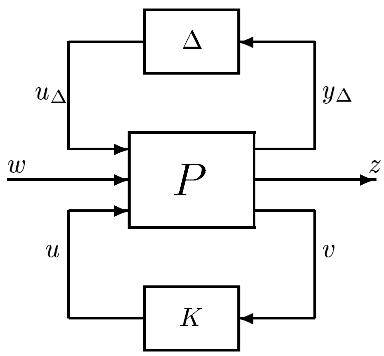


Figure 3.21: General control configuration for the case with model uncertainty

The block diagram in Figure 3.21 in terms of P (for synthesis) may be transformed into the block diagram in Figure 3.22 in terms of N (for analysis) by using K to close

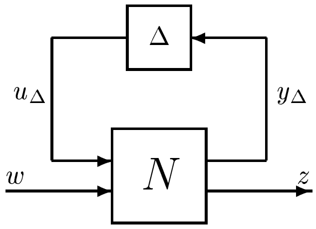


Figure 3.22: General block diagram for analysis with uncertainty included

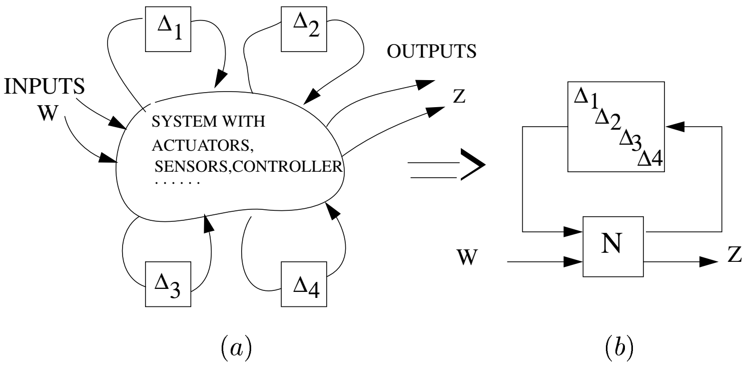


Figure 3.23: Rearranging a system with multiple perturbations into the $N\Delta$-structure

a lower loop around P. If we partition P to be compatible with the controller K, then the same *lower LFT* as found in (3.103) applies, and

$$N = F_l(P, K) = P_{11} + P_{12}K(I - P_{22}K)^{-1}P_{21} \tag{3.112}$$

To evaluate the perturbed (uncertain) transfer function from external inputs w to external outputs z, we use Δ to close the upper loop around N (see Figure 3.22), resulting in an *upper LFT* (see Appendix A.7):

$$z = F_u(N, \Delta)w; \quad F_u(N, \Delta) \triangleq N_{22} + N_{21}\Delta(I - N_{11}\Delta)^{-1}N_{12} \tag{3.113}$$

Remark 1 Controller synthesis based on Figure 3.21 is still an unsolved problem, although good practical approaches like DK-iteration to find the "μ-optimal" controller are in use (see Section 8.12). For analysis (with a given controller), the situation is better and with the $\mathcal{H}_\infty$ norm an assessment of robust performance involves computing the structured singular value, μ. This is discussed in more detail in Chapter 8.

Remark 2 In (3.113) N has been partitioned to be compatible with Δ, that is N_{11} has dimensions compatible with Δ. Usually, Δ is square in which case N_{11} is a square matrix of the same dimension as Δ. For the nominal case with no uncertainty we have $F_u(N, \Delta) = F_u(N, 0) = N_{22}$, so N_{22} is the nominal transfer function from w to z.

Remark 3 Note that P and N here also include information about how the uncertainty affects the system, so they are *not* the same P and N as used earlier, for example in (3.103). Actually, the parts P_{22} and N_{22} of P and N in (3.112) (with uncertainty) are equal to the P and N in (3.103) (without uncertainty). Strictly speaking, we should have used another symbol for N and P in (3.112), but for notational simplicity we did not.

Remark 4 The fact that almost any control problem with uncertainty can be represented by Figure 3.21 may seem surprising, so some explanation is in order. First represent each source of uncertainty by a perturbation block, Δ_i, which is normalized such that $\|\Delta_i\| \leq 1$. These perturbations may result from parametric uncertainty, neglected dynamics, etc. as will be discussed in more detail in Chapters 7 and 8. Then "pull out" each of these blocks from the system so that an input and an output can be associated with each Δ_i as shown in Figure 3.23(a). Finally, collect these perturbation blocks into a large block-diagonal matrix having perturbation inputs and outputs as shown in Figure 3.23(b). In Chapter 8 we discuss in detail how to obtain N and Δ.

3.9 Additional exercises

Most of these exercises are based on material presented in Appendix A. The exercises illustrate material which the reader should know before reading the subsequent chapters.

Exercise 3.15 *Consider the performance specification $\|w_P S\|_\infty < 1$. Suggest a rational transfer function weight $w_P(s)$ and sketch it as a function of frequency for the following two cases:*

1. *We desire no steady-state offset, a bandwidth better than 1 rad/s and a resonance peak (worst amplification caused by feedback) lower than 1.5.*
2. *We desire less than 1% steady-state offset, less than 10% error up to frequency 3 rad/s, a bandwidth better than 10 rad/s, and a resonance peak lower than 2. Hint: See (2.72) and (2.73).*

Exercise 3.16 *By $\|M\|_\infty$ one can mean either a spatial or temporal norm. Explain the difference between the two and illustrate by computing the appropriate infinity norm for*

$$M_1 = \begin{bmatrix} 3 & 4 \\ -2 & 6 \end{bmatrix}, \quad M_2(s) = \frac{s-1}{s+1}\frac{3}{s+2}$$

Exercise 3.17 *What is the relationship between the RGA-matrix and uncertainty in the individual elements? Illustrate this for perturbations in the $1,1$-element of the matrix*

$$A = \begin{bmatrix} 10 & 9 \\ 9 & 8 \end{bmatrix} \tag{3.114}$$

Exercise 3.18 *Assume that A is non-singular. (i) Formulate a condition in terms of the maximum singular value of E for the matrix $A+E$ to remain non-singular. Apply this to A in (3.114) and (ii) find an E of minimum magnitude which makes $A+E$ singular.*

Exercise 3.19 *Compute* $\|A\|_{i1}$, $\bar{\sigma}(A) = \|A\|_{i2}$, $\|A\|_{i\infty}$, $\|A\|_F$, $\|A\|_{\max}$ *and* $\|A\|_{\text{sum}}$ *for the following matrices and tabulate your results:*

$$A_1 = I; \quad A_2 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}; A_3 = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}; A_4 = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}, A_5 = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}$$

Show using the above matrices that the following bounds are tight (i.e. we may have equality) for 2×2 *matrices (*$m = 2$*):*

$$\bar{\sigma}(A) \le \|A\|_F \le \sqrt{m}\,\bar{\sigma}(A)$$

$$\|A\|_{\max} \le \bar{\sigma}(A) \le m\|A\|_{\max}$$
$$\|A\|_{i1}/\sqrt{m} \le \bar{\sigma}(A) \le \sqrt{m}\|A\|_{i1}$$
$$\|A\|_{i\infty}/\sqrt{m} \le \bar{\sigma}(A) \le \sqrt{m}\|A\|_{i\infty}$$
$$\|A\|_F \le \|A\|_{\text{sum}}$$

Exercise 3.20 *Find example matrices to illustrate that the above bounds are also tight when* A *is a square* $m \times m$ *matrix with* $m > 2$.

Exercise 3.21 *Do the extreme singular values bound the magnitudes of the elements of a matrix? That is, is* $\bar{\sigma}(A)$ *greater than the largest element (in magnitude), and is* $\underline{\sigma}(A)$ *smaller than the smallest element? For a non-singular matrix, how is* $\underline{\sigma}(A)$ *related to the largest element in* A^{-1}*?*

Exercise 3.22 *Consider a lower triangular* $m \times m$ *matrix* A *with* $a_{ii} = -1$, $a_{ij} = 1$ *for all* $i > j$, *and* $a_{ij} = 0$ *for all* $i < j$.

a) What is $\det A$ *?*
b) What are the eigenvalues of A *?*
c) Show that the smallest singular value is less than or equal to 2^{-m}.
d) What is the RGA of A*?*
e) Let $m = 4$ *and find an* E *with the smallest value of* $\bar{\sigma}(E)$ *such that* $A + E$ *is singular.*

Exercise 3.23 *Find two matrices* A *and* B *such that* $\rho(A+B) > \rho(A)+\rho(B)$ *which proves that the spectral radius does not satisfy the triangle inequality and is thus not a norm.*

Exercise 3.24 *Write* $T = GK(I + GK)^{-1}$ *as an LFT of* K, *i.e. find* P *such that* $T = F_l(P, K)$.

Exercise 3.25 *Write* K *as an LFT of* $T = GK(I + GK)^{-1}$, *i.e. find* J *such that* $K = F_l(J, T)$.

Exercise 3.26 *State-space descriptions may be represented as LFTs. To demonstrate this find* H *for*

$$F_l(H, 1/s) = C(sI - A)^{-1}B + D$$

Exercise 3.27 *Show that the set of all stabilizing controllers in (4.91) can be written as* $K = F_l(J, Q)$ *and find* J.

Exercise 3.28 *In (3.11) we stated that the sensitivity of a perturbed plant, $S' = (I + G'K)^{-1}$, is related to that of the nominal plant, $S = (I + GK)^{-1}$ by*

$$S' = S(I + E_O T)^{-1}$$

where $E_O = (G' - G)G^{-1}$. This exercise deals with how the above result may be derived in a systematic (though cumbersome) manner using LFTs (see also (Skogestad and Morari, 1988a)).

a) First find F such that $S' = (I + G'K)^{-1} = F_l(F, K)$, and find J such that $K = F_l(J, T)$ (see Exercise 3.25).

b) Combine these LFTs to find $S' = F_l(N, T)$. What is N in terms of G and G'?. Note that since $J_{11} = 0$ we have from (A.156)

$$N = \begin{bmatrix} F_{11} & F_{12}J_{12} \\ J_{21}F_{21} & J_{22} + J_{21}F_{22}J_{12} \end{bmatrix}$$

c) Evaluate $S' = F_l(N, T)$ and show that

$$S' = I - G'G^{-1}T(I - (I - G'G^{-1})T)^{-1}$$

d) Finally, show that this may be rewritten as $S' = S(I + E_O T)^{-1}$.

3.10 Conclusion

The main purpose of this chapter has been to give an overview of methods for analysis and design of multivariable control systems.

In terms of analysis, we have shown how to evaluate MIMO transfer functions and how to use the singular value decomposition of the frequency-dependent plant transfer function matrix to provide insight into multivariable directionality. Other useful tools for analyzing directionality and interactions are the condition number and the RGA. Closed-loop performance may be analyzed in the frequency domain by evaluating the maximum singular value of the sensitivity function as a function of frequency. Multivariable RHP-zeros impose fundamental limitations on closed-loop performance, but for MIMO systems we can often direct the undesired effect of a RHP-zero to a subset of the outputs. MIMO systems are often more sensitive to uncertainty than SISO systems, and we demonstrated in two examples the possible sensitivity to input gain uncertainty.

In terms of controller design, we discusssed some simple approaches such as decoupling and decentralized control. We also introduced a general control configuration in terms of the generalized plant P, which can be used as a basis for synthesizing multivariable controllers using a number of methods, including LQG, $\mathcal{H}_2$, $\mathcal{H}_\infty$ and μ-optimal control. These methods are discussed in much more detail in Chapters 8 and 9. In this chapter we have only discussed the $\mathcal{H}_\infty$ weighted sensitivity method.

BIBLIOGRAPHY

Anderson, B. D. O. (1986). Weighted Hankel-norm approximation: Calculation of bounds, *Systems & Control Letters* **7**(4): 247–255.

Anderson, B. D. O. and Liu, Y. (1989). Controller reduction: Concepts and approaches, *IEEE Transactions on Automatic Control* **34**(8): 802–812.

Anderson, B. D. O. and Moore, J. B. (1989). *Optimal Control: Linear Quadratic Methods*, Prentice-Hall, Englewood Cliffs, New Jersey.

Balas, G. J., Doyle, J. C., Glover, K., Packard, A. and Smith, R. (1993). *μ-Analysis and Synthesis Toolbox User's Guide*, MathWorks, Natick, Mass.

Balchen, J. G. and Mumme, K. (1988). *Process Control. Structures and Applications*, Van Nostrand Reinhold, New York.

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

Boyd, S. and Barratt, C. (1991). *Linear Controller Design — Limits of Performance*, Prentice-Hall, Englewood Cliffs.

Boyd, S. and Desoer, C. A. (1985). Subharmonic functions and performance bounds in linear time-invariant feedback systems, *IMA J. Math. Contr. and Info.* **2**: 153–170.

Boyd, S., Ghaoui, L. E., Feron, E. and Balakrishnan, V. (1994). *Linear Matrix Inequalities in System and Control Theory*, SIAM, Philadelphia.

Braatz, R. D. (1993). *Robust Loopshaping for Process Control*, PhD thesis, California Institute of Technology, Pasadena.

Braatz, R. D. and Morari, M. (1994). Minimizing the Euclidean condition number, *SIAM Journal on Control and Optimization* **32**(6): 1763–1768.

Braatz, R. D., Morari, M. and Skogestad, S. (1996). Loopshaping for robust performance, *International Journal of Robust and Nonlinear Control* **6**.

Bristol, E. H. (1966). On a new measure of interactions for multivariable process control, *IEEE Transactions on Automatic Control* **AC-11**: 133–134.

Campo, P. J. and Morari, M. (1994). Achievable closed-loop properties of systems under decentralised control: Conditions involving the steady-state gain, *IEEE Transactions on Automatic Control* **AC-39**: 932–942.

Cao, Y. (1995). *Control Structure Selection for Chemical Processes Using Input-output Controllability Analysis*, PhD thesis, University of Exeter.

Chang, J. W. and Yu, C. C. (1990). The relative gain for non-square multivariable systems, *Chem. Eng. Sci.* **45**: 1309–1323.

Chen, C. T. (1984). *Linear System Theory and Design*, Holt, Rinehart and Winston, Inc., New York.

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

Chiang, R. Y. and Safonov, M. G. (1992). *Robust Control Toolbox User's Guide*, MathWorks, South Natick.

Churchill, R. V., Brown, J. W. and Verhey, R. F. (1974). *Complex Variables and Applications*, McGraw-Hill, New York.

Dahl, H. J. and Faulkner, A. J. (1979). Helicopter simulation in atmospheric turbulence, *Vertica* pp. 65–78.

Daoutidis, P. and Kravaris, C. (1992). Structural evaluation of control configurations for multivariable nonlinear processes, *Chemical Engineering Science* **47**: 1091–1107.

Davison, E. J. (ed.) (1990). *Benchmark Problems for Control System Design*, Report of the IFAC Theory Committee, International Federation of Automatic Control.

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

Doyle, J. C. (1978). Guaranteed margins for LQG regulators, *IEEE Transactions on Automatic Control* **AC-23**(4): 756–757.

Doyle, J. C. (1982). Analysis of feedback systems with structured uncertainties, *IEE Proceedings, Part D* **129**(6): 242–250.

Doyle, J. C. (1983). Synthesis of robust controllers and filters, *Proc. IEEE Conf. on Decision and Control*, San Antonio, Texas, pp. 109–114.

Doyle, J. C. (1984). *Lecture Notes on Advances in Multivariable Control*, ONR/Honeywell Workshop, Minneapolis, October.

Doyle, J. C. (1986). Redondo Beach lecture notes, Internal Report, Caltech, Pasadena.

Doyle, J. C., Francis, B. and Tannenbaum, A. (1992). *Feedback Control Theory*, Macmillan Publishing Company.

Doyle, J. C., Glover, K., Khargonekar, P. P. and Francis, B. A. (1989). State-space solutions to standard $\mathcal{H}_2$ and $\mathcal{H}_\infty$ control problems, *IEEE Transactions on Automatic Control* **AC-34**(8): 831–847.

Doyle, J. C. and Stein, G. (1981). Multivariable feedback design: Concepts for a classical/modern synthesis, *IEEE Transactions on Automatic Control* **AC-26**(1): 4–16.

Doyle, J. and Stein, G. (1979). Robustness with observers, *IEEE Transactions on Automatic Control* **24**(4): 607–611.

Eaton, J. W. and Rawlings, J. B. (1992). Model-predictive control of chemical processes, *Chemical Engineering Science* **69**(1): 3–9.

Engell, S. (1988). *Optimale Lineare Regelung*, Vol. 18 of *Fachberichte Messen, Steuern, Regeln*, Springer-Verlag, Berlin.

Enns, D. (1984). Model reduction with balanced realizations: An error bound and a frequency weighted generalization, *Proceedings of the 23rd IEEE Conference on Decision and Control*, Las Vegas, NV, USA, pp. 127–32.

Fernando, K. V. and Nicholson, H. (1982). Singular perturbational model reduction of balanced systems, *IEEE Transactions on Automatic Control* **AC-27**(2): 466–468.

Foss, A. S. (1973). Critique of chemical process control theory, *AIChE Journal* **19**: 209–214.

Foster, N. P., Spurgeon, S. K. and Postlethwaite, I. (1993). Robust model-reference tracking control with a sliding mode applied to an act rotorcraft, *19th European Rotorcraft Forum*, Italy.

Francis, B. (1987). *A course in $\mathcal{H}_\infty$ control theory*, Lecture Notes in Control and Information Sciences, Springer-Verlag, Berlin.

Francis, B. A. and Zames, G. (1984). On $\mathcal{H}_\infty$ optimal sensitivity theory for SISO feedback systems, *IEEE Transactions on Automatic Control* **AC-29**(1): 9–16.

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

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

Freudenberg, J. S. and Looze, D. P. (1988). *Frequency Domain Properties of Scalar and*

Multivariable Feedback Systems, Vol. 104 of *Lecture Notes in Control and Information Sciences*, Springer-Verlag, Berlin.

Freudenberg, J.S. Looze, D. (1985). Right half planes poles and zeros and design tradeoffs in feedback systems, *IEEE Transactions on Automatic Control* pp. 555–565.

Gjøsæter, O. B. (1995). *Structures for Multivariable Robust Process Control*, PhD thesis, Norwegian University of Science and Technology, Trondheim.

Glover, K. (1984). All optimal Hankel-norm approximations of linear multivariable systems and their L^∞-error bounds, *International Journal of Control* **39**(6): 1115–93.

Glover, K. and Doyle, J. C. (1988). State-space formulae for all stabilizing controller that satisfy an $\mathcal{H}_\infty$ norm bound and relations to risk sensitivity, *Systems and Control Letters* **11**: 167–172.

Glover, K. and McFarlane, D. (1989). Robust stabilization of normalized coprime factor plant descriptions with $\mathcal{H}_\infty$ bounded uncertainty, *IEEE Transactions on Automatic Control* **AC-34**(8): 821–830.

Golub, G. H. and van Loan, C. F. (1989). *Matrix Computations*, John Hopkins University Press, Baltimore.

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

Grosdidier, P. and Morari, M. (1986). Interaction measures for systems under decentralized control, *Automatica* **22**: 309–319.

Grosdidier, P., Morari, M. and Holt, B. R. (1985). Closed-loop properties from steady-state gain information, *Industrial and Engineering Chemistry Process Design and Development* **24**: 221–235.

Haggblom, K. E. and Waller, K. (1988). Transformations and consistency relations of distillation control structures, *AIChE Journal* **34**: 1634–1648.

Hanus, R., Kinnaert, M. and Henrotte, J. (1987). Conditioning technique, a general anti-windup and bumpless transfer method, *Automatica* **23**(6): 729–739.

Havre, K. (1995). Personal communication. Norwegian University of Science and Technology, Trondheim.

Helton, J. (1976). Operator theory and broadband matching, *In Proceedings of the 11th Annual Allerton Conference on Communications, Control and Computing* .

Holt, B. R. and Morari, M. (1985a). Design of resilient processing plants V — The effect of deadtime on dynamic resilience, *Chemical Engineering Science* **40**: 1229–1237.

Holt, B. R. and Morari, M. (1985b). Design of resilient processing plants VI — The effect of right plane zeros on dynamic resilience, *Chemical Engineering Science* **40**: 59–74.

Horn, R. A. and Johnson, C. R. (1985). *Matrix Analysis*, Cambridge University Press.

Horn, R. A. and Johnson, C. R. (1991). *Topics in Matrix Analysis*, Cambridge University Press.

Horowitz, I. M. (1963). *Synthesis of Feedback Systems*, Academic Press, London.

Horowitz, I. M. (1991). Survey of quantitative feedback theory (QFT), *International Journal of Control* **53**(2): 255–291.

Horowitz, I. M. and Shaked, U. (1975). Superiority of transfer function over state-variable methods in linear time-invariant feedback system design, *IEEE Transactions on Automatic Control* **AC-20**(1): 84–97.

Hovd, M. (1992). *Studies on Control Structure Selection and Design of Robust Decentralized and SVD Controllers*, PhD thesis, Norwegian University of Science and Technology, Trondheim.

Hovd, M., Braatz, R. D. and Skogestad, S. (1994). SVD controllers for $\mathcal{H}_2$-, $\mathcal{H}_\infty$-, and μ-optimal control, *Proc. 1994 American Control Conference*, Baltimore, pp. 1233–1237.

Hovd, M. and Skogestad, S. (1992). Simple frequency-dependent tools for control system analysis, structure selection and design, *Automatica* **28**(5): 989–996.

Hovd, M. and Skogestad, S. (1993). Procedure for regulatory control structure selection with application to the FCC process, *AIChE Journal* **39**(12): 1938–1953.

Hovd, M. and Skogestad, S. (1994a). Pairing criteria for decentralised control of unstable plants, *Industrial and Engineering Chemistry Research* **33**: 2134–2139.

Hovd, M. and Skogestad, S. (1994b). Sequential design of decentralized controllers, *Automatica* **30**(10): 1601–1607.

Hoyle, D., Hyde, R. A. and Limebeer, D. J. N. (1991). An $\mathcal{H}_\infty$ approach to two degree of freedom design, *Proceedings of the 30th IEEE Conference on Decision and Control*, Brighton, UK, pp. 1581–1585.

Hung, Y. S. and MacFarlane, A. G. J. (1982). *Multivariable Feedback: A Quasi-Classical Approach*, Vol. 40 of *Lecture Notes in Control and Information Sciences*, Springer-Verlag, Berlin.

Hyde, R. A. (1991). *The Application of Robust Control to VSTOL Aircraft*, PhD thesis, University of Cambridge.

Hyde, R. A. and Glover, K. (1993). The application of scheduled $\mathcal{H}_\infty$ controllers to a VSTOL aircraft, *IEEE Transactions on Automatic Control* **AC-38**(7): 1021–1039.

Jaimoukha, I. M., Kasenally, E. M. and Limebeer, D. J. N. (1992). Numerical solution of large scale Lyapunov equations using Krylov subspace methods, *Proceedings of the 31st IEEE Conference on Decision and Control*.

Johnson, C. R. and Shapiro, H. M. (1986). Mathematical aspects of the relative gain array ($A \circ A^{-T}$), *SIAM Journal on Algebraic and Discrete Methods* **7**(4): 627–644.

Kailath, T. (1980). *Linear Systems*, Prentice-Hall, Englewood Cliffs.

Kalman, R. (1964). When is a linear control system optimal?, *Journal of Basic Engineering — Transaction on ASME — Series D* **86**: 51–60.

Kouvaritakis, B. (1974). *Characteristic Locus Methods for Multivariable Feedback Systems Design*, PhD thesis, University of Manchester Institute of Science and Technology, UK.

Kwakernaak, H. (1969). Optimal low-sensitivity linear feedback systems, *Automatica* **5**(3): 279–286.

Kwakernaak, H. (1985). Minimax frequency-domain performance and robustness optimization of linear feedback systems, *IEEE Transactions on Automatic Control* **AC-30**(10): 994–1004.

Kwakernaak, H. (1993). Robust control and $\mathcal{H}_\infty$-optimization — Tutorial paper, *Automatica* **29**: 255–273.

Kwakernaak, H. and Sivan, R. (1972). *Linear Optimal Control Systems*, Wiley Interscience, New York.

Laub, A. J., Heath, M. T., Page, C. C. and Ward, R. C. (1987). Computation of balancing transformations and other applications of simultaneous diagonalization algorithms, *IEEE Transactions on Automatic Control* **AC-32**(2): 115–122.

Laughlin, D. L., Jordan, K. G. and Morari, M. (1986). Internal model control and process uncertainty – mapping uncertainty regions for SISO controller-design, *International Journal of Control* **44**(6): 1675–1698.

Laughlin, D. L., Rivera, D. E. and Morari, M. (1987). Smith predictor design for robust performance, *International Journal of Control* **46**(2): 477–504.

Lee, J. H., Braatz, R. D., Morari, M. and Packard, A. (1995). Screening tools for robust control structure selection, *Automatica* **31**(2): 229–235.

Limebeer, D. J. N. (1991). The specification and purpose of a controller design case study, *Proc. IEEE Conf. Decision Contr.*, Brighton, UK, pp. 1579–1580.

Limebeer, D. J. N., Kasenally, E. M. and Perkins, J. D. (1993). On the design of robust two degree of freedom controllers, *Automatica* **29**(1): 157–168.

Liu, Y. and Anderson, B. D. O. (1989). Singular perturbation approximation of balanced systems, *International Journal of Control* **50**(4): 1379–1405.

Lundström, P. (1994). *Studies on Robust Multivariable Distillation Control*, PhD thesis, Norwegian University of Science and Technology, Trondheim.

Lundström, P., Skogestad, S. and Doyle, J. C. (1996). Two degrees of freedom controller design for an ill-conditioned plant using μ-synthesis, *To appear in IEEE Transactions on Control System Technology* .

Lunze, J. (1992). *Feedback Control of Large-Scale Systems*, Prentice-Hall, Englewood Cliffs.

MacFarlane, A. G. J. and Karcanias, N. (1976). Poles and zeros of linear multivariable systems: A survey of algebraic, geometric and complex variable theory, *International Journal of Control* **24**: 33–74.

MacFarlane, A. G. J. and Kouvaritakis, B. (1977). A design technique for linear multivariable feedback systems, *International Journal of Control* **25**: 837–874.

Maciejowski, J. M. (1989). *Multivariable Feedback Design*, Addison-Wesley, Wokingham U.K.

Manness, M. A. and Murray-Smith, D. J. (1992). Aspects of multivariable flight control law design for helicopters using eigenstructure assignment, *Journal of American Helicopter Society* pp. 18–32.

Manousiouthakis, V., Savage, R. and Arkun, Y. (1986). Synthesis of decentralized process control structures using the concept of block relative gain, *AIChE Journal* **32**: 991–1003.

Marlin, T. (1995). *Process Control*, Mc-Graw Hill.

McFarlane, D. and Glover, K. (1990). *Robust Controller Design Using Normalized Coprime Factor Plant Descriptions*, Vol. 138 of *Lecture Notes in Control and Information Sciences*, Springer-Verlag, Berlin.

McMillan, G. K. (1984). *pH Control*, Instrument Society of America, Research Triangle Park, North Carolina.

Meinsma, G. (1995). Unstable and nonproper weights in $\mathcal{H}_\infty$ control, *Automatica* **31**(1): 1655–1658.

Mesarovic, M. (1970). Multilevel systems and concepts in process control, *Proc. of the IEEE*, Vol. 58, pp. 111–125.

Meyer, D. G. (1987). *Model Reduction via Factorial Representation*, PhD thesis, Stanford University.

Moore, B. C. (1981). Principal component analysis in linear systems: controllability, observability and model reduction, *IEEE Transactions on Automatic Control* **AC-26**(1): 17–32.

Morari, M. (1982). Integrated plant control: A solution at hand or a research topic for the next decade?, *in* T. F. Edgar and D. E. Seborg (eds), *Proc. Chemical Process Control 2*, AIChE, New York, pp. 467–495.

Morari, M. (1983). Design of resilient processing plants III – a general framework for the assessment of dynamic resilience, *Chemical Engineering Science* **38**: 1881–1891.

Morari, M., Arkun, Y. and Stephanopoulos, G. (1980). Studies in the synthesis of control structures for chemical process, Part I: Formulation of the problem. Process decomposition and the classification of the control tasks. Analysis of the optimizing control structures., *AIChE Journal* **26**(2): 220–232.

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

Nett, C. N. (1986). Algebraic aspects of linear-control system stability, *IEEE Transactions on Automatic Control* **AC-31**(10): 941–949.

Nett, C. N. (1989). A quantitative approach to the selection and partitioning of measurements and manipulations for the control of complex systems, *Presentation at Caltech Control Workshop*, Pasadena, USA, Jan. 1989.

Nett, C. N. and Manousiouthakis, V. (1987). Euclidean condition and block relative gain - connections conjectures, and clarifications, *IEEE Transactions on Automatic Control* **AC-32**(5): 405–407.

Nett, C. N. and Minto, K. D. (1989). A quantitative approach to the selection and partitioning of measurements and manipulations for the control of complex systems, Copy of transparencies from talk at *American Control Conference*, Pittsburgh, Pennsylvania, June 1989.

Niemann, H. and Stoustrup, J. (1995). Special Issue on Loop Transfer Recovery, *International Journal of Robust and Nonlinear Control* **7**(7): November.

Nwokah, O. D. I. and Perez, R. (1991). On multivariable stability in the gain space, *Automatica* **27**(6): 975–983.

Owen, J. G. and Zames, G. (1992). Robust $\mathcal{H}_\infty$ disturbance minimization by duality, *Systems & Control Letters* **19**(4): 255–263.

Packard, A. (1988). *What's New with μ*, PhD thesis, University of California, Berkeley.

Packard, A. and Doyle, J. C. (1993). The complex structured singular value, *Automatica* **29**(1): 71–109.

Packard, A., Doyle, J. C. and Balas, G. (1993). Linear, multivariable robust-control with a μ-perspective, *Journal of Dynamic Systems Measurement and Control — Transactions of the ASME* **115**(2B): 426–438.

Padfield, G. D. (1981). Theoretical model of helicopter flight mechanics for application to piloted simulation, *Technical Report 81048*, Defence Research Agency (formerly Royal Aircraft Establishment), UK.

Perkins, J. D. (ed.) (1992). *IFAC Workshop on Interactions Between Process Design and Process Control*, (London, Sept. 1992), Pergamon Press, Oxford.

Pernebo, L. and Silverman, L. M. (1982). Model reduction by balanced state space representation, *IEEE Transactions on Automatic Control* **AC-27**(2): 382–387.

Poolla, K. and Tikku, A. (1995). Robust performance against time-varying structured perturbations, *IEEE Transactions on Automatic Control* **40**(9): 1589–1602.

Postlethwaite, I., Foster, N. P. and Walker, D. J. (1994). Rotorcraft control law design for rejection of atmospheric turbulence, *Proceedings of IEE Conference, Control 94*, Warwick, pp. 1284–1289.

Postlethwaite, I. and MacFarlane, A. G. J. (1979). *A Complex Variable Approach to the Analysis of Linear Multivariable Feedback Systems*, Vol. 12 of *Lecture Notes in Control and Information Sciences*, Springer-Verlag, Berlin.

Postlethwaite, I., Samar, R., Choi, B.-W. and Gu, D.-W. (1995). A digital multi-mode $\mathcal{H}_\infty$ controller for the spey turbofan engine, *3rd European Control Conference*, Rome, Italy, pp. 147–152.

Postlethwaite, I. and Walker, D. J. (1992). Advanced control of high performance rotorcraft, *Institute of Mathematics and Its Applications Conference on Aerospace Vehicle Dynamics and Control, Cranfield Institute of Technology* pp. 615–619.

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

Rosenbrock, H. H. (1966). On the design of linear multivariable systems, *Third IFAC World Congress*. Paper 1a.

Rosenbrock, H. H. (1970). *State-space and Multivariable Theory*, Nelson, London.

Rosenbrock, H. H. (1974). *Computer-Aided Control System Design*, Academic Press, New York.

Safonov, M. G. (1982). Stability margins of diagonally perturbed multivariable feedback systems, *IEE Proceedings, Part D* **129**(6): 251–256.

Safonov, M. G. and Athans, M. (1977). Gain and phase margin for multiloop LQG regulators, *IEEE Transactions on Automatic Control* **AC-22**(2): 173–179.

Safonov, M. G. and Chiang, R. Y. (1989). A Schur method for balanced-truncation model reduction, *IEEE Transactions on Automatic Control* **AC-34**: 729–733.

Safonov, M. G., Limebeer, D. J. N. and Chiang, R. Y. (1989). Simplifying the $\mathcal{H}_\infty$ theory via loop-shifting, matrix-pencil and descriptor concepts, *International Journal of Control* **50**(6): 2467–2488.

Samar, R. (1995). *Robust Multi-Mode Control of High Performance Aero-Engines*, PhD thesis, University of Leicester.

Samar, R. and Postlethwaite, I. (1994). Multivariable controller design for a high performance areo engine, *Proceedings IEE Conference Control 94*, Warwick, pp. 1312–1317.

Samar, R., Postlethwaite, I. and Gu, D.-W. (1995). Model reduction with balanced realizations, *International Journal of Control* **62**(1): 33–64.

Samblancatt, C., Apkarian, P. and Patton, R. J. (1990). Improvement of helicopter robustness and performance control law using eigenstructure techniques and $\mathcal{H}_\infty$ synthesis, *16th European Rotorcraft Forum*, Scotland. Paper No. 2.3.1.

Seborg, D. E., Edgar, T. F. and Mellichamp, D. A. (1989). *Process Dynamics and Control*, Wiley, New York.

Sefton, J. and Glover, K. (1990). Pole-zero cancellations in the general $\mathcal{H}_\infty$ problem with reference to a two block design, *Systems & Control Letters* **14**: 295–306.

Shamma, J. S. (1994). Robust stability with time-varying structured uncertainty, *IEEE Transactions on Automatic Control* **AC-39**: 714–724.

Shinskey, F. G. (1984). *Distillation Control*, 2nd edn, McGraw Hill, New York.

Shinskey, F. G. (1988). *Process Control Systems*, 3rd edn, McGraw Hill, New York.

Skogestad, S. (1992). Dynamics and control of distillation columns — a critical survey, *IFAC symposium DYCORD+'92*, Maryland, pp. 1–25.

Skogestad, S. and Havre, K. (1996). The use of RGA and condition number as robustness measures, In *Proc. European Symposium on Computer-Aided Process Engineering (ESCAPE'96)*, Rhodes, Greece, May 1996.

Skogestad, S., Lundström, P. and Jacobsen, E. (1990). Selecting the best distillation control configuration, *AIChE Journal* **36**(5): 753–764.

Skogestad, S. and Morari, M. (1987a). Control configuration selection for distillation columns, *AIChE Journal* **33**(10): 1620–1635.

Skogestad, S. and Morari, M. (1987b). Effect of disturbance directions on closed-loop performance, *Industrial and Engineering Chemistry Research* **26**: 2029–2035.

Skogestad, S. and Morari, M. (1987c). Implications of large RGA elements on control performance, *Industrial and Engineering Chemistry Research* **26**: 2323–2330.

Skogestad, S. and Morari, M. (1988a). Some new properties of the structured singular value, *IEEE Transactions on Automatic Control* **AC-33**(12): 1151–1154.

Skogestad, S. and Morari, M. (1988b). Variable selection for decentralized control, *AIChE Annual Meeting*, Washington DC. Paper 126f. Reprinted in *Modeling, Identification and Control*, 1992, Vol. **13**, No. 2, 113–125.

Skogestad, S. and Morari, M. (1989). Robust performance of decentralized control systems by independent designs, *Automatica* **25**(1): 119–125.

Skogestad, S., Morari, M. and Doyle, J. C. (1988). Robust control of ill-conditioned plants: High-purity distillation, *IEEE Transactions on Automatic Control* **AC-33**(12): 1092–1105.

Skogestad, S. and Wolff, E. A. (1992). Controllability measures for disturbance rejection, *IFAC Workshop on Interactions between Process Design and Process Control*, London, UK, pp. 23–29.

Sourlas, D. D. and Manousiouthakis, V. (1995). Best achievable decentralized performance, *IEEE Transactions on Automatic Control* **AC-40**(11): 1858–1871.

Stanley, G., Marino-Galarraga, M. and McAvoy, T. J. (1985). Shortcut operability analysis. 1. The relative disturbance gain, *Industrial and Engineering Chemistry Process Design and Development* **24**: 1181–1188.

Stein, G. and Athans, M. (1987). The LQG/LTR procedure for multivariable feedback design, *IEEE Transactions on Automatic Control* **32**(2): 105–114.

Stein, G. and Doyle, J. C. (1991). Beyond singular values and loopshapes, *AIAA J. of Guidance and Control* **14**: 5–16.

Stephanopoulos, G. (1989). *Chemical Process Control*, Prentice-Hall, Englewood Cliffs, New Jersey.

Strang, G. (1976). *Linear Algebra and Its Applications*, Academic Press, New York.

Takahashi, M. D. (1993). Synthesis and evaluation of an $\mathcal{H}_2$ control law for a hovering helicopter, *Journal of Guidance, Control and Dynamics* **16**: 579–584.

Tøffner-Clausen, S., Andersen, P., Stoustrup, J. and Niemann, H. H. (1995). A new approach to μ-synthesis for mixed perturbation sets, *Proc. of 3rd European Control Conference*, Rome, Italy, pp. 147–152.

Toker, O. and Ozbay, H. (1995). On np-hardness of the purely complex μ computation, *Proceedings of the American Control Conference*, pp. 447–451.

Tombs, M. S. and Postlethwaite, I. (1987). Truncated balanced realization of a stable non-minimal state-space system, *International Journal of Control* **46**: 1319–1330.

Tsai, M., Geddes, E. and Postlethwaite, I. (1992). Pole-zero cancellations and closed-loop properties of an $\mathcal{H}_\infty$ mixed sensitivity design problem, *Automatica* **3**: 519–530.

van de Wal, M. and de Jager, B. (1995). Control structure design: A survey, *Proc. of American Control Conference*, Seattle, pp. 225–229.

van Diggelen, F. and Glover, K. (1994a). A Hadamard weighted loop shaping design procedure for robust decoupling, *Automatica* **30**(5): 831–845.

van Diggelen, F. and Glover, K. (1994b). State-space solutions to Hadamard weighted $\mathcal{H}_\infty$ and $\mathcal{H}_2$ control-problems, *International Journal of Control* **59**(2): 357–394.

Vidyasagar, M. (1985). *Control System Synthesis: A Factorization Approach*, MIT Press, Cambridge, MA.

Walker, D. J. (1996). On the structure of a two degrees-of-freedom controller, *To appear in International Journal of Control* .

Walker, D. J. and Postlethwaite, I. (1996). Advanced helicopter flight control using two degrees-of-freedom $\mathcal{H}_\infty$ optimization, *Journal of Guidance, Control and Dynamics* **19**(2): March–April.

Walker, D. J., Postlethwaite, I., Howitt, J. and Foster, N. P. (1993). Rotorcraft flying qualities improvement using advanced control, *American Helicopter Society/NASA Conference*, San Francisco.

Waller, K. V., Häggblom, K. E., Sandelin, P. M. and Finnerman, D. H. (1988). Disturbance sensitivity of distillation control structures, *AIChE Journal* **34**: 853–858.

Wang, Z. Q., Lundström, P. and Skogestad, S. (1994). Representation of uncertain time delays in the $\mathcal{H}_\infty$ framework, *International Journal of Control* **59**(3): 627–638.

Weinmann, A. (1991). *Uncertain Models and Robust Control*, Springer-Verlag, Berlin.

Whidborne, J. F., Postlethwaite, I. and Gu, D. W. (1994). Robust controller design using $\mathcal{H}_\infty$ loop shaping and the method of inequalities, *IEEE Transactions on Control Systems technology* **2**(4): 455–461.

Willems (1970). *Stability Theory of Dynamical Systems*, Nelson.

Wolff, E. (1994). *Studies on Control of Integrated Plants*, PhD thesis, Norwegian University of Science and Technology, Trondheim.

Wonham, M. (1974). *Linear Multivariable Systems*, Springer-Verlag, Berlin.

Youla, D. C., Jabr, H. A. and Bongiorno, J. J. (1976). Modern Wiener-Hopf design of optimal controllers, part II: The multivariable case., *IEEE Transactions on Automatic Control* **AC-21**: 319–38.

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

Young, P. M. (1993). *Robustness with Parametric and Dynamic Uncertainties*, PhD thesis, California Institute of Technology, Pasadena.

Young, P. M. (1994). Controller design with mixed uncertainties, *Proceedings of the American Control Conference*, Baltimore, Maryland, USA, pp. 2333–2337.

Young, P. M., Newlin, M. and Doyle, J. C. (1992). Practical computation of the mixed μ problem, *Proceedings of the American Control Conference*, Chicago, pp. 2190–2194.

Yu, C. C. and Fan, M. K. H. (1990). Decentralised integral controllability and D-stability, *Chemical Engineering Science* **45**: 3299–3309.

Yu, C. C. and Luyben, W. (1986). Design of multiloop SISO controllers in multivariable processes, *Industrial and Engineering Chemistry Process Design and Development* **25**: 498–503.

Yu, C. C. and Luyben, W. L. (1987). Robustness with respect to integral controllability, *Industrial and Engineering Chemistry Research* **26**(5): 1043–1045.

Yue, A. and Postlethwaite, I. (1990). Improvement of helicopter handling qualities using $\mathcal{H}_\infty$ optimization, *IEE Proceedings-D Control Theory and Applications* **137**: 115–129.

Zafiriou, E. (ed.) (1994). *IFAC Workshop on Integration of Process Design and Control*, (Baltimore, June 1994), Pergamon Press, Oxford. See also special issue of *Computers & Chemical Engineering*, Vol. **20**, No. 4, 1996.

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

Zhou, K., Doyle, J. C. and Glover, K. (1996). *Robust and Optimal Control*, Prentice-Hall, Upper Saddle River.

Ziegler, J. G. and Nichols, N. B. (1942). Optimum settings for automatic controllers, *Transactions of the A.S.M.E.* **64**: 759–768.

Ziegler, J. G. and Nichols, N. B. (1943). Process lags in automatic-control circuits, *Transactions of the A.S.M.E.* **65**: 433–444.

INDEX

[1] Page numbers in *italic* refer to definitions.