Linear Control Systems

course notes

Leonid Mirkin
Faculty of Mechanical Engineering
Technion–IIT

March 4, 2024

Preface

Introductory control courses are traditionally devoted to single-input, single-output (aka SISO) systems. Properties of these systems are well understood and can be characterized in terms of their impulse responses, poles and zeros of their transfer functions, frequency-response gain and phase, etc. Extending these notions to multiple-input, multiple-output (MIMO) systems might not be straightforward though. For example, it is not obvious, perhaps even somewhat counterintuitive, to induce from our SISO insight that the first-order transfer function $\left[\begin{smallmatrix} 1 & 1/s \\ 0 & 1 \end{smallmatrix}\right]$ has not only a pole, but also a zero at the origin. Or that the cascade $P(s)R(s)$ for $P(s) = \left[\begin{smallmatrix} 1/s & 0 \\ 0 & 1/s \end{smallmatrix}\right]$ and $R(s) = \left[\begin{smallmatrix} -1 & 1 \\ 1 & -1 \end{smallmatrix}\right]$, which is a typical interconnection in the diffusively-coupled consensus protocol, has an unstable cancellation.

This text aims primarily at being an introduction to the world of MIMO linear time-invariant systems. The main emphasis is placed on the frequency-domain analysis, which merely reflects my personal preferences. The exposition presumes familiarity with basic classical control notions (such as transfer functions, poles, zeros, time and frequency responses, et cetera) and principles (reviewed in Chapter 1), as well as with some fundamental calculus, complex analysis, and linear algebra. Although linear algebra is reviewed in Chapter 2, the objective of this review is to introduce a number of system-theoretic notions used throughout the text rather than to cover the required background material.

Another goal of these notes is to pave the way for the use of optimization-based analysis and design methods, again, mainly in the frequency domain. To this end, the exposition focuses on problem formulation aspects, such as motivations, the meaning of being optimal from the control engineering viewpoint, and understanding a role played by tuning parameters (weighting functions). The optimization techniques themselves are practically not covered as such issues go well beyond the scope of this text. I am convinced that it is more important to realize, before being acquainted with optimization methods, that "optimal" is not necessarily a synonym of "good," whatever meaning we put on the last term. And that optimization is a *tool* of control engineering rather than its goal. This tool, if used consciously, may indeed be a powerful instrument of producing meaningful results and understanding intrinsic limitations in achieving sought goals. However, its mechanical use could produce senseless optimal control systems.

The text was developed as the lecture notes for the graduate-level course "Linear Control Systems" (036012) taught in the Faculty of Mechanical Engineering at the Technion. The first version was written during the Spring 2000 semester. A major revision was carried out in Spring 2006. The second revision, which is again substantially different from the previous one, was conceived while I was escaping daily routine in a remote location by Lake Issyk-Kul in September 2018. It remains a mystery for myself why the robust stability chapter is unfinished in all these versions, so perhaps it is a feature, rather than a bug, after all…

Lake Issyk-Kul (42.6597,77.5461) Leonid Mirkin
September, 2018

Contents

Nomenclature

$\mathbb{N}$	set of positive integers (natural numbers)
$\mathbb{Z}$	set of integers
$\mathbb{Z}_+$	set of nonnegative integers
$\mathbb{Z}_-$	set of non-positive integers, $\mathbb{Z}_- = \mathbb{Z} \setminus \mathbb{N}$
$\mathbb{Z}_{i_1..i_2}$	integer interval from i_1 to and including i_2, i.e. $\mathbb{Z}_{i_1..i_2} := \{i \in \mathbb{Z} \mid i_1 \leq i \leq i_2\}$
$\mathbb{R}$	set of real numbers, $\mathbb{R} = (-\infty, \infty)$
$\mathbb{R}_+$	set of nonnegative real numbers, $\mathbb{R}_+ = [0, \infty)$
$\mathbb{R}_-$	set of non-positive real numbers, $\mathbb{R}_- = (\infty, 0]$
$\mathrm{j}\mathbb{R}$	set of pure imaginary numbers
$\mathbb{C}$	set of complex numbers
$\operatorname{Re} z$	the real part of $z \in \mathbb{C}$
$\operatorname{Im} z$	the imaginary part of $z \in \mathbb{C}$
$\mathbb{C}_\alpha$	open right half-plain, to the right of $\alpha \in \mathbb{R}$, i.e. $\mathbb{C}_\alpha := \{z \in \mathbb{C} \mid \operatorname{Re} z > \alpha\}$
$\bar{\mathbb{C}}_\alpha$	closed right half-plain, to the right of $\alpha \in \mathbb{R}$, i.e. $\bar{\mathbb{C}}_\alpha := \{z \in \mathbb{C} \mid \operatorname{Re} z \geq \alpha\}$
$\mathbb{T}$	unit circle, $\mathbb{T} := \{z \in \mathbb{C} \mid \lvert z\rvert = 1\}$
$\mathbb{D}$	interior of $\mathbb{T}$ (open unit disk), $\mathbb{D} := \{z \in \mathbb{C} \mid \lvert z\rvert < 1\}$
$\bar{\mathbb{D}}$	closed unit disk, $\bar{\mathbb{D}} := \{z \in \mathbb{C} \mid \lvert z\rvert \leq 1\} = \mathbb{D} \cup \mathbb{T}$
$\mathbb{F}$	general field, frequently used as an alias of either $\mathbb{R}$ or $\mathbb{C}$
e_i	the ith standard basis in $\mathbb{F}^n$; $e_1 := \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \end{bmatrix}'$, $e_2 := \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \end{bmatrix}'$, et cetera
I_n	$n \times n$ identity matrix (just I if the dimension is irrelevant)
$0_{n\times m}$	$n \times m$ zero matrix (just 0 if the dimensions are irrelevant)
a_{ij}	the ijth element of a matrix $A \in \mathbb{F}^{p\times m}$
$a_{i\bullet}$	the ith row of a matrix $A \in \mathbb{F}^{p\times m}$
$a_{\bullet j}$	the jth column of a matrix $A \in \mathbb{F}^{p\times m}$
A'	transpose of a matrix $A \in \mathbb{R}^{n\times m}$ / complex-conjugate transpose of a matrix $A \in \mathbb{C}^{n\times m}$
$\lambda_i(A)$	ith eigenvalue of a matrix $A \in \mathbb{F}^{n\times n}$
$\operatorname{spec}(A)$	spectrum of a matrix $A \in \mathbb{F}^{n\times n}$, i.e. the set of all its eigenvalues
$\rho(A)$	spectral radius of a matrix $A \in \mathbb{F}^{n\times n}$, $\rho(A) := \max\{\lvert\lambda_1(A)\rvert, \ldots, \lvert\lambda_n(A)\rvert\}$
$\bar{\sigma}(A)$	the maximal singular value of a matrix $A \in \mathbb{F}^{p\times m}$, $\bar{\sigma}(A) = \sqrt{\rho(A'A)} = \sqrt{\rho(AA')}$
$\underline{\sigma}(A)$	the minimal singular value of a matrix $A \in \mathbb{F}^{p\times m}$

$\mathrm{tr}(A)$	trace of a matrix $A \in \mathbb{F}^{n\times n}$, $\mathrm{tr}(A) := \sum_{i=1}^{n} a_{ii} = \sum_{i=1}^{n} \lambda_i(A)$
$\|A\|$	spectral norm of a matrix $A \in \mathbb{F}^{n\times m}$, $\|A\|^2 = [\overline{\sigma}(A)]^2 = \rho(A'A) = \rho(AA')$
$\|A\|_{\mathrm{F}}$	Frobenius norm of a matrix $A \in \mathbb{F}^{n\times m}$, $\|A\|_{\mathrm{F}}^2 = \sum_{i=1}^{n}\sum_{j=1}^{m}\|a_{ij}\|^2 = \mathrm{tr}(A'A) = \mathrm{tr}(AA')$
$C^{p\times m}(\mathbb{I})$	class of continuous functions $\mathbb{I} \to \mathbb{F}^{p\times m}$ for $\mathbb{I} \subset \mathbb{R}$ (it is denoted $C^p(\mathbb{I})$ if $m = 1$ and $C(\mathbb{I})$ if the dimensions are irrelevant or clear from the context)
$L_2^{p\times m}(\mathbb{I})$	Lebesgue space of square integrable functions $\mathbb{I} \to \mathbb{F}^{p\times m}$ (or $L_2(\mathbb{I})$)
$L_{2+}^{p\times m}(\mathbb{R})$	space of square integrable functions $\mathbb{R} \to \mathbb{F}^{p\times m}$ with support in $\mathbb{R}_+$ (or $L_{2+}(\mathbb{R})$)
$L_{2-}^{p\times m}(\mathbb{R})$	space of square integrable functions $\mathbb{R} \to \mathbb{F}^{p\times m}$ with support in $\mathbb{R}_-$ (or $L_{2-}(\mathbb{R})$)
$\ell_2^{p\times m}(\mathbb{I})$	space of square summable functions $\mathbb{I} \to \mathbb{F}^{p\times m}$ for $\mathbb{I} \subset \mathbb{Z}$ (or $\ell_2(\mathbb{I})$)
$\ell_{2+}^{p\times m}(\mathbb{Z})$	space of square summable functions $\mathbb{Z} \to \mathbb{F}^{p\times m}$ with support in $\mathbb{Z}_+$ (or $\ell_{2+}(\mathbb{Z})$)
$\ell_{2-}^{p\times m}(\mathbb{Z})$	space of square summable functions $\mathbb{Z} \to \mathbb{F}^{p\times m}$ with support in $\mathbb{Z}_+$ (or $\ell_{2-}(\mathbb{Z})$)
$L_1^{p\times m}(\mathbb{I})$	space of absolute integrable functions $\mathbb{I} \to \mathbb{F}^{p\times m}$ (or $L_1(\mathbb{I})$)
$L_\infty^{p\times m}(\mathbb{I})$	space of essentially bounded functions $\mathbb{I} \to \mathbb{F}^{p\times m}$ (or $L_\infty(\mathbb{I})$)
$H_\infty^{p\times m}(\mathbb{A})$	Hardy space of holomorphic and bounded functions $\mathbb{A} \to \mathbb{F}^{p\times m}$ for some $\mathbb{A} \subset \mathbb{C}$ (or H_∞)
$\mathfrak{F}\{x\}$	Fourier transform of $x : \mathbb{R} \to \mathbb{F}^n$
$\mathfrak{L}\{x\}$	Laplace transform of $x : \mathbb{R} \to \mathbb{F}^n$
$\mathbb{1}_{\mathbb{I}}$	characteristic function of a set $\mathbb{I}$, $\mathbb{1}_{\mathbb{I}}(t) = \begin{cases} 1 & \text{if } t \in \mathbb{I} \subset \mathbb{R} \\ 0 & \text{otherwise} \end{cases}$ or $\mathbb{1}_{\mathbb{I}}[t] = \begin{cases} 1 & \text{if } t \in \mathbb{I} \subset \mathbb{Z} \\ 0 & \text{otherwise} \end{cases}$
$\mathbb{1}$	unit step, $\mathbb{1}(t) = \mathbb{1}_{\mathbb{R}_+}(t)$ or $\mathbb{1}[t] = \mathbb{1}_{\mathbb{Z}_+}[t]$
δ	Dirac delta in continuous time or unit pulse, $\delta[t] = \begin{cases} 1 & \text{if } t = 0 \\ 0 & \text{otherwise} \end{cases}$, in discrete time
$\exp_a$	exponential function, $\exp_a(t) = \mathrm{e}^{at}$
sinc	sine cardinal, $\mathrm{sinc}(t) = \sin(t)/t$
$\mathfrak{D}_G$	domain of an L_2 system G, $\mathfrak{D}_G := \{u \in L_2 \mid Gu \in L_2\}$
g	impulse response of an LTI system G
$G(s)$	transfer function of an LTI system G, $G(s) := (\mathfrak{L}\{g\})(s)$
$\mathrm{pdir_i}(G, p)$	input direction of a pole p of $G(s)$
$\mathrm{pdir_o}(G, p)$	output direction of a pole p of $G(s)$
$\mathrm{zdir_i}(G, z)$	input direction of a zero z of $G(s)$
$\mathrm{zdir_o}(G, z)$	output direction of a pole z of $G(s)$
(A, B, C, D)	state-space realization of a finite-dimensional LTI system G
$R_G(s)$	Rosenbrock system matrix of realization (A, B, C, D) of G, $R_G(s) := \begin{bmatrix} A - sI & B \\ C & D \end{bmatrix}$
lcf	left coprime factorization over H_∞, like $G = \tilde{M}^{-1}\tilde{N}$
rcf	right coprime factorization over H_∞, like $G = NM^{-1}$
$\mathcal{F}_{\mathrm{l}}(G, H)$	lower linear fractional transformation, $\mathcal{F}_{\mathrm{l}}(G, H) = G_{11} + G_{12}H(I - G_{22}H)^{-1}G_{21}$
$\mathcal{F}_{\mathrm{u}}(G, H)$	upper linear fractional transformation, $\mathcal{F}_{\mathrm{u}}(G, H) = G_{22} + G_{21}H(I - G_{11}H)^{-1}G_{12}$
$G \star \tilde{G}$	Redheffer star product

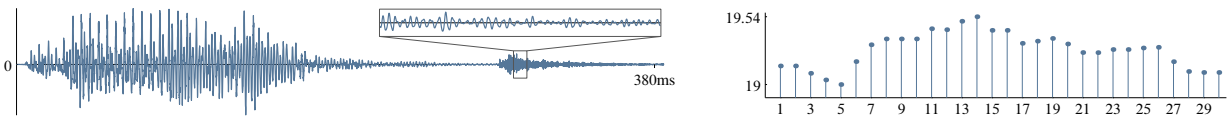
Chapter 1

Preliminary: SISO Control in a Nutshell

THIS CHAPTER revises some basic notions and ideas of classical control methods for SISO (single-input, single-output) LTI (linear time-invariant) systems. It is not aimed at presenting a comprehensive overview, but rather at introducing key notions used throughout these notes and providing a motivation for studying theoretical foundations of MIMO (multiple-input, multiple output) systems. The main emphasis is placed on the plant inversion ideas and frequency-domain analysis and design philosophy.

1.1 Signals and systems

In numerous situations we may be concerned with quantities, whose values change as some *independent variables*, like time, evolve. Two such examples are presented in Fig. 1.1. The audio waveform of a short



(a) Audio waveform of the word "ot" (Hebrew for "signal")

(b) ILS to KGS exchange rates in September 2018

Fig. 1.1: Examples of signals

word is depicted in Fig. 1.1(a). This very waveform makes the voice recognizable (and reproducible, as a matter of fact). The independent variable here is the analog time, a subset of $\mathbb{R}$. Fig. 1.1(b) shows the exchange rates of Israeli new shekel to Kyrgyzstani som during one month. In this case, the independent variable is the discrete time, a subset of $\mathbb{Z}$. These quantities can be described as functions of independent variables, say $f(t)$ if $t \in \mathbb{R}$ or $f[t]$ if $t \in \mathbb{Z}$. When associated with some physical process, such functions are called *signals*. Thus, signals are functions conveying information about the behavior of some phenomenon.

Remark 1.1. The independent variable need be neither time nor scalar. For example, we may be interested in the steady-state temperature profile of a heated body as a function of the distance from the heating element or, in image processing, in the RGB values of pixels as a function of their two spatial coordinates. Nonetheless, time is by far the most prevalent independent variable in control applications, so we frequently interchange these terms in the sequel. ∇

Signals are not "in a vacuum" (although they do exist in a vacuum, thanks Per-Olof). They are related via laws of nature, economics, society, and so on. In some cases these relations, and their applicability scopes, are well understood. For example, we know that applied forces and the mass position are coupled via Newton's law; that the electromotive force around a closed path depends on the magnetic flux enclosed by it (by Faraday's law); that the angles of incidence and refraction of light passing through a boundary

(a) Mass-spring-damper system (b) RLC circuit

Fig. 1.2: Two simple systems

between two different isotropic media are related via Snell's law; et cetera. In other cases, mostly in life sciences, humanities, sociology, economics, and the like, relations between signals may be empirical and vague. For example, it may be accepted that unemployment is linked with losses in a country's production (by Okun's law); the utility of social networks can scale exponentially with the size of the network (by Reed's law); the metabolic rate of animals is related to their mass (by Kleiber's law); passion is inversely proportional to the amount of real information available (by Benford's law of controversy); the bitterness of discussions in academia is in inverse proportion to the importance of issues at stake (by Sayre's law); et cetera. But even then the very fact that signals of interest are interrelated is beyond doubt.

By *systems* we may then understand constraints imposed on interdependent signals. Consider e.g. a mechanical system comprising a mass point with mass m connected to a fixed base via a spring with spring constant k and a damper (dashpot) with damping coefficient c, see Fig. 1.2(a). It can be seen as a constraint on mutual relation between the force f and the mass position x. If the system is in its equilibrium and we change the force, then the position can be uniquely determined via Newton's law. It is worth emphasizing that it also works the other way around. Namely, if somehow we can change the mass position, it applies a unique force (think of a door closer, which works this way). Thus, two involved signals, f and x, are constrained by properties of this system. Another example of a system is the RLC electrical circuit in Fig. 1.2(b), with a resistor having resistance R, a solenoid having inductance L, and a capacitor having capacitance C. Such a system constraints mutual relations between the voltage v and the current i, which can be derived by Kirchhoff's voltage law.

Although all involved signals are "born equal," two groups of signals associated with a given system are often distinguished in control applications. One group is assumed to be an action (input) and another one is then a reaction (output), see the block-diagram in Fig. 1.3, which represents a system P with an input

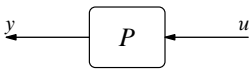


Fig. 1.3: Block-diagram of an I/O system $P : u \mapsto y$

u and an output y. This relation is conventionally written as $y = Pu$, where P should be understood as an operator acting on u. This philosophy is referred to as the *input / output approach*[1] and the corresponding systems as I/O systems. It treats systems as signal processors, in a sense that systems are understood as mappings "$\mapsto$" from input signals to output signals, like $P : u \mapsto y$ in Fig. 1.3, where this relation is reflected in the arrows added to the signal lines. This way of thinking is more restrictive, because the separation between inputs and outputs might impose artificial causality relations. Nevertheless, it is widely acceptable and is sufficiently rich in many situations.

A system $G : u \mapsto y$ is said to be SISO (single-input, single-output) if both its input and output signals are scalar-valued, normally, taking values in $\mathbb{R}$. Otherwise, it is MIMO (multiple-input, multiple-output). It is called *linear* if it satisfies the property of superposition, i.e. if

$$G(\alpha_1 u_1 + \alpha_2 u_2) = \alpha_1 (G u_1) + \alpha_2 (G u_2) \quad \text{for all admissible signals } u_1, u_2 \text{ and all scalars } \alpha_1, \alpha_2.$$

[1] A more general viewpoint, with which we started, is known as the behavioral approach of Jan C. Willems [31].

A system is referred to as *time-invariant* (or shift-invariant) if any constant time shift of its input results in the same time shift of its output. Linear time-invariant systems are often abbreviated to LTI.

1.2 Models

Let us return to the systems in Fig. 1.2 and be more concrete about their behavior. When considered as a mapping $f \mapsto x$, the mechanical system in Fig. 1.2(a) can be described by the differential equation

$$m\ddot{x}(t) + c\dot{x}(t) + kx(t) = f(t). \tag{1.1}$$

This relation follows from Newton's and Hooke's laws and by assuming that the dashpot produces a resistive force proportional to the velocity of its end. The electrical circuit in Fig. 1.2(b) can be viewed as a mapping $v \mapsto q$, where q is the charge across the capacitor, with $i = \dot{q}$. This mapping satisfies

$$L\ddot{q}(t) + R\dot{q}(t) + \frac{1}{C}q(t) = v(t) \tag{1.2}$$

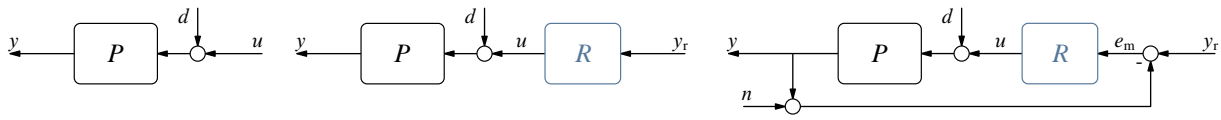
and follows by Kirchhoff's, Faraday's, Ohm's, and Gauss's laws. It is readily seen that (1.2) is essentially the same as (1.1), modulo the replacements $m \to L$, $c \to R$, and $k \to 1/C$. It may then be convenient to present both (1.1) and (1.2) as an abstract mapping $u \mapsto y$ described by the following non-physical second-order differential equation:

$$\ddot{y}(t) + 2\zeta\omega_{\mathrm{n}}\dot{y}(t) + \omega_{\mathrm{n}}^2 y(t) = k_{\mathrm{st}}\omega_{\mathrm{n}}^2 u(t), \tag{1.3}$$

where the parameters ω_{n}, ζ, and k_{st} represent the natural frequency, the damping factor, and the static gain, respectively. Their values have one to one correspondence with physical parameters of the respective systems and determine qualitative and quantitative system properties, like oscillations / decay profiles, speed of transients, steady-state response to various harmonic signals, et cetera. Thus, properties of mechanical and electrical systems in Fig. 1.2 can be studied in a uniform fashion via their abstract models.

This is a momentous observation. Once we abstract from the physical nature of signals and systems, be they mechanical, electrical, biological, social, etc, we can analyze them from unified points of view by universal theories. This is a key idea behind the notion of *mathematical model*, which is a description of systems in a mathematical language. In control applications we most frequently encounter mathematical language based on differential or difference equations. Yet there are many other ways to describe systems, like logical models used in computer sciences, heuristic models used in psychology, an so on. Model-based way of thinking is ubiquitous in engineering applications, even in methods claiming to be model-free (perhaps the only difference is that some model-based approaches are explicit and others are implicit).

An important aspect of mathematical models, which should always be remembered, is that they are *never perfect*. Any mathematical model is merely a (more or less accurate) approximation of the actual physical process, capturing its properties only to a certain degree. For example, in modeling the mechanical system in Fig. 1.2(a) we assumed that the mass movements are perfectly one-dimensional, the dashpot and the the spring are linear, massless, and do not change their properties when heated up, that there is no friction at the mass base and with the air, and so on and so forth. Likewise, the model (1.2) of the RLC circuit in Fig. 1.2(b) is true only under unrealistic assumptions that the resistor, solenoid, and capacitor are linear and their properties do not depend on working conditions, that the whole circuit is perfectly isolated from its environment, both electrically and thermally, and suchlike, not to mention that the picture becomes substantially more complicated on the atomic level. Still, models (1.1) and (1.2) are useful if deviations from linear regimes and changes in components properties are relatively small. These considerations are true in general as well. It may be safe to claim that there are no linear systems in nature and that no physical system can be exhaustively described by a finite number of ordinary differential / difference equations or



(a) Controlled system (plant) (b) Open-loop control (c) Closed-loop unity-feedback control

Fig. 1.4: Basic control setups

are time invariant. Yet we do use rather simple LTI models in many situations. In general, the required accuracy and model granularity level depend on applications and a model, sufficiently accurate in one application, might not be adequate in another. Occam's razor may be a good principle for deciding on the type of model in every situation. Finally, note that with a certain abuse of terminology, we frequently say "system" meaning its model, this is a common practice.

1.3 A prototype control problem

Control is about affecting systems so that they behave in a desired manner. Systems can be affected in various ways, up to rebuilding their physical structure. Control theory normally addresses situations, when systems are fixed and affected via some of their input signals, naturally called *control inputs*, that can be freely shaped by a controller.

A heavily simplified—yet sufficiently representative as a starting point—control setup can be described in terms of the system presented in Fig. 1.4(a). Here the studied LTI system P, dubbed the *plant*, is affected via its control input u to generate a desired output y, aka the *controlled output*, at the end. The signal d, which also has an effect on y, is called the *disturbance* (sometimes, the load disturbance). Its role is to account for impacts of exogenous phenomena, like impacts of the environment, which are uncertain or too complicated to model. The disturbance signal is often supposed to be unmeasurable (although in some situations it can be) and uncorrelated with control goals. Throughout this and the next section we assume that all involved signals are scalar valued, so that P is SISO.

Requirements to the controlled output may be conveniently expressed in terms of a *reference signal* y_r. The rationale behind its choice varies from task to task. In some problems, called *tracking problems*, this is a physical signal, which should be tracked by the plant output. Its generation might then be an independent task, like recovering y_r from on-line measurements of a target, e.g. in video tracking problems. In other situations, like in *set-point regulation*, the reference signal is calculated analytically from knowing the initial and required steady-state positions of y. Although this is not frequently highlighted, the reference signal in such situations should reflect not only requirements to the behavior of y, but also limitations of the plant, see the discussion in §1.A.2. In any case, given a reference signal y_r, a *basic control problem* can be formulated as selecting the control signal u rendering (at least, approaching)

$$y(t) = y_r(t), \quad \forall t \tag{1.4}$$

despite the presence of unmeasured exogenous signals and modeling uncertainties.

There are essentially two control configurations for meeting this requirement. An *open-loop control* scheme is shown in Fig. 1.4(b). In this case the controller (regulator) R generates the control signal u as a function of the reference signal y_r only (it may also depend on d if the latter is measurable). In *closed-loop control*, whose simplest unity-feedback configuration is depicted in Fig. 1.4(c), the controller R generates the control input using both the reference signal and measurements of the controlled signal y (or signals correlated with y, if it is not directly measurable). Measurements of y require separate sensors, which are never perfect. These imperfections can be accounted for by introducing an additive measurement noise signal, like n in Fig. 1.4(c). In both these configurations, the designed element is the controller R.

1.4 Ultimate control methodology: plant inversion and its limitations

There are many ways to think of designing R. Arguably, they all have the idea of plant inversion under the hood. It is conceptually simple and intuitive, although (practically) never implementable literally, which is why control engineering is such a nontrivial discipline.

The idea can be grasped by considering the idealized version of the system in Fig. 1.4(a) under $d = 0$ and a perfectly known P. Rewriting (1.4) as $y_\mathrm{r} = Pu$ yields the following control signal meeting this requirement:

$$u = P^{-1} y_\mathrm{r}, \tag{1.5}$$

where P^{-1} should be understood as the system producing the unity system when placed in series with P (its transfer function is the reciprocal of the transfer function of the plant, $1/P(s)$). The control law (1.5) is called the plant inversion.

1.4.1 Open-loop plant inversion

The control law (1.5) appears to be a perfect fit for the open-loop system in Fig. 1.4(b). It corresponds to the choice

$$R(s) = \frac{1}{P_0(s)}, \tag{1.6}$$

where P_0 is our model of the plant P. However, the use of this controller has its pitfalls, some of which severely limit its applicability scope.

The first of these pitfalls is associated with the notion of *stability*, which we have not discussed yet. Loosely speaking, an I/O system is said to be stable if its output remains bounded for all bounded inputs. It is then well known that a finite-dimensional LTI system is stable iff its transfer function is proper (i.e. bounded in $s \in \mathbb{C}_\alpha$ for a sufficiently large $\alpha > 0$) and has no poles in the closed right half-plane. Consider now the relation between all exogenous and internal signals in the system in Fig. 1.4(b)

$$\begin{bmatrix} y \\ u \end{bmatrix} = \begin{bmatrix} PR & P \\ R & 0 \end{bmatrix} \begin{bmatrix} y_\mathrm{r} \\ d \end{bmatrix}. \tag{1.7}$$

We obviously need the boundedness of both y and u. We also cannot ignore the effects of either y_r or d (even if d is insignificant, its effect on unstable systems might be destructive). We thus must guarantee that all three nonzero systems in (1.7) are stable. This is known as the *internal stability* of interconnected systems. Internal stability is equivalent to the stability of both P and R. If this is the case, then the third system, PR, is always stable. Consequently, the open-loop *setup in Fig. 1.4(b) cannot be used if the plant is unstable*, no matter what control algorithm is considered. Another outcome of the internal stability requirement is that *controller* (1.6) *cannot be used if $P_0(s)$ is nonminimum-phase or strictly proper*. Indeed, the former property would add unstable poles to the controller and the latter would render $R(s)$ non-proper.

Remark 1.2 (properness). Non-proper controllers may be accepted in some situations, viz. if they always act on analytically known reference signals with sufficiently many bounded derivatives, see the discussion in §1.A.3. Still, properness might be a safe property to impose as a standard requirement. ▽

There is a workaround for problems with nonminimum-phase or strictly proper plant models. Instead of (1.6) we may use

$$R(s) = \frac{T_\mathrm{ref}(s)}{P_0(s)} \tag{1.8}$$

for some stable system T_ref. Technically, the only constraint on T_ref, apart from its own stability, is that the resulting controller is stable. This boils down to the following two conditions:

1. $T_\mathrm{ref}(s)$ must have all nonminimum-phase zeros of $P_0(s)$ as its own zeros (multiplicities counted),

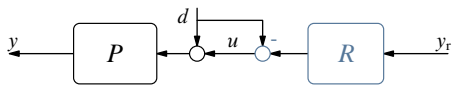


Fig. 1.5: Open-loop control with disturbance feedforward

2. the pole excess of $T_{\text{ref}}(s)$ must be greater than or equal to the pole excess of $P_0(s)$.

While the second of these conditions is effectively non-restrictive, the first one might be, depending on the location of unstable zeros of $P_0(s)$.

Conceptually, the introduction of T_{ref} renders the controlled output (still assuming that $P_0 = P$)

$$y(t) = (T_{\text{ref}} y_{\text{r}})(t), \quad \forall t. \tag{1.9}$$

In other words, we may only expect to match y_{r} processed by T_{ref}. This motivates the term *reference model* for this system, it may be thought of as representing a "best" pragmatic response of the controlled system to y_{r}. Unless $T_{\text{ref}}(s) = 1$, (1.9) is no longer the same as our original goal (1.4). But the latter can be met only if the plant has a bi-proper transfer function (i.e. has infinite bandwidth), which is not realistic. It thus does make sense to switch to a more pragmatic (1.9). This is especially so if we know that y_{r} is not arbitrary, but rather belongs to a more narrow class of signals. For example, we often need to track band-limited signals only. In such situations it may be reasonable to require $T_{\text{ref}}(\mathrm{j}\omega) \approx 1$ only within the frequency range of the spectrum of y_{r}. This normally produces a low-pass $T_{\text{ref}}(s)$ and does not impose any restrictive constraints on the pole excess of $T_{\text{ref}}(s)$.

Now it is time to address the effect of other idealizations. Specifically, consider the behavior of the controlled output under a nonzero disturbance d and not perfectly known plant, with $P_0 \neq P$. Define $e := T_{\text{ref}} y_{\text{r}} - y$ as the error signal representing the deviation from required behavior. It is readily seen that (1.8) renders

$$e = \left(1 - PP_0^{-1}\right) T_{\text{ref}} y_{\text{r}} - Pd =: e_{y_{\text{r}}} - e_d. \tag{1.10}$$

Because y_{r} and d are independent, we cannot cause $e_{y_{\text{r}}}$ and e_d to cancel each other or alleviate their effect. Thus, these two terms should be considered individually.

An important observation is that we are left with no tools to affect the mismatch terms in (1.10). This would also be true for any other choice of R in Fig. 1.4(b). Equation (1.10) would then remain unchanged modulo the replacement of T_{ref} with the model $P_0 R$ of the attained system $y_{\text{r}} \mapsto y$. The effect of modeling uncertainty, $e_{y_{\text{r}}}$, is entirely determined by the relative modeling error system, $1 - PP_0^{-1}$. Hence, open-loop control can only be effective in frequency ranges where this error is small. The effect of the exogenous disturbance, e_d, is also independent of the controller choice. In fact, a controller acting in open loop can affect the disturbance response only if it can measure d or at least a signal correlated with it.

Remark 1.3 (disturbance feedforward). If d is measurable, the control configuration in Fig. 1.5 can be used instead, with the very same rationale behind the design of R, just now with $y_{\text{r}} = Pd + Pu$ to satisfy (1.4). For this scheme we obviously have $e_d = 0$. Even if d is not measured perfectly, such measurements can still be useful to reduce the impact of d on the controlled variables, at least to some extent. ▽

Finally, it should be mentioned that the two conditions on p. 5 are not the only limiting factors in the choice of T_{ref}. Another limitation, a soft one, is the control effort. Actuating resources are always limited, so we would be interested in attaining control goals with "affordable" u. It may be convenient to normalize control signals, so that $|u| = 1$ is the borderline between "high" and "low" control efforts. The condition $|R(\mathrm{j}\omega)| \not\gg 1$ may be then viewed as an indication of "affordable" control. In addition, we may be interested in attaining $|R(\mathrm{j}\omega)| \ll 1$ at high frequencies, to avoid amplifying parasitic components in y_{r}. These considerations, viewed in the context of (1.8), can be translated to the rule of thumb that *the bandwidth of T_{ref} should not considerably exceed that of P*. Control problems tend to become more complicated if this suggestion is not followed.

1.4.2 Closed-loop plant inversion

Relationships between (1.4) and the setup in Fig. 1.4(c) are less evident than those between (1.4) and the setup in Fig. 1.4(b). Yet they exist in an unexpected and fascinating way.

To start with, consider relations between all exogenous and internal signals for the system in Fig. 1.4(c). It is not hard to see that

$$\begin{bmatrix} y \\ u \\ e \end{bmatrix} = \begin{bmatrix} T & T_{\mathrm{d}} & -T \\ T_{\mathrm{c}} & -T & -T_{\mathrm{c}} \\ S & -T_{\mathrm{d}} & T \end{bmatrix} \begin{bmatrix} y_{\mathrm{r}} \\ d \\ n \end{bmatrix}, \tag{1.11}$$

where $e := r - y = e_{\mathrm{m}} + n$ and

$$\begin{bmatrix} S(s) & T_{\mathrm{c}}(s) \\ T_{\mathrm{d}}(s) & T(s) \end{bmatrix} := \frac{1}{1 + P(s)R(s)} \begin{bmatrix} 1 & R(s) \\ P(s) & P(s)R(s) \end{bmatrix} = \frac{1}{1 + P(s)R(s)} \begin{bmatrix} 1 \\ P(s) \end{bmatrix} \begin{bmatrix} 1 & R(s) \end{bmatrix} \tag{1.12}$$

are four fundamental closed-loop transfer functions, termed the *sensitivity* (S), *complementary sensitivity* (as $T = 1 - S$), *control sensitivity* ($T_{\mathrm{c}} = TP^{-1}$), and *disturbance sensitivity* ($T_{\mathrm{d}} = PS$) transfer functions. They are also known as the *Gang of Four* (the term coined by Karl Johan Åström back in 2000).

In line with its definition for open-loop systems, the *internal stability* of the system in Fig. 1.4(c) is defined as the stability of all possible closed-loop systems, i.e. the four systems whose transfer functions are defined in (1.12). The internal stability requirement rules out unstable pole / zero cancellations between $P(s)$ and $R(s)$. Indeed, if unstable poles of $P(s)$ are canceled by the controller, then they are still poles of $T_{\mathrm{d}}(s)$. Likewise, if unstable zeros of $P(s)$ are canceled by $R(s)$, these zeros show up as poles of $T_{\mathrm{c}}(s)$. Internal stability also rules out non-proper $R(s)$'s, because then $T_{\mathrm{c}}(s)$ is not proper either. But if there are no unstable pole / zero cancellations between $P(s)$ and $R(s)$ and the latter transfer function is proper, internal stability is equivalent to the stability of either one of the four transfer functions in (1.12).

Consider now the control signal. Assuming for the moment that measurements of y are perfect, i.e. that $n = 0$, we have:

$$u = T_{\mathrm{c}} r - Td = \frac{R}{1 + PR} y_{\mathrm{r}} - \frac{PR}{1 + PR} d = \frac{1}{1/R + P} y_{\mathrm{r}} - \frac{P}{1/R + P} d. \tag{1.13}$$

It is worth emphasizing that the mechanical inversion of the plant by the controller, as in (1.6), is futile here. Indeed, even if admissible, the choice $R = P^{-1}$ would result in $u = 0.5P^{-1}y_{\mathrm{r}} - 0.5d$ and then in $y = 0.5y_{\mathrm{r}} + 0.5P^{-1}d$, which makes little sense. A proper "plant inversion" choice can be seen by noticing that the terms $1/R$ on the right-hand side of (1.13) become less significant as the controller gain grows. In the limit (ignore mathematical meaning of this for now), as $R \to \infty$, the control signal

$$u \to u_{\infty} := P^{-1} y_{\mathrm{r}} - d \tag{1.14}$$

and the controlled output meets (1.4) despite the presence of d. This u_{∞} is the very control signal generated by the open-loop control system with disturbance feedforward in Fig. 1.5 under the plant-inversion choice of the controller, as in (1.5). But unlike its open-loop version, the closed-loop plant inversion requires neither the knowledge of the plant model nor direct measurements of d (although measuring y can be thought of as an indirect measurement of the disturbance). This sounds like a miracle and should naturally arouse suspicions of being too simple to be true.

These suspicions would be well founded. The situation in the closed-loop case is even more complicated than that in the open-loop case. Stability is again a primary concern. It can be shown, e.g. by root-locus asymptotes arguments, that closed-loop systems can be stable under controllers with uniformly high gains only for a handful of plants. Namely, these are minimum-phase systems with a pole excess of

at most two. And even for these plants the use of such controllers is not practical. First, no measurements are perfect. If the measurement noise is taken into account, the control signal in (1.14) turns

$$u_\infty = P^{-1}(y_\mathrm{r} - n) - d, \tag{1.15}$$

resulting in $y = y_\mathrm{r} - n$. Tracking measurement noise, especially its high-frequency components, is not what we need. Moreover, P^{-1} is typically a high-pass system, so fast components of n would be amplified by (1.15), producing large high-frequency oscillations in the control signal, which is not healthy either. Second, uniformly high-gain feedback frequently produces lightly damped modes in closed-loop systems. This, in turn, yields poor transients, with fast but slowly decaying oscillations. Third, feedback systems with high loop gains at high frequencies are normally quite sensitive to modeling uncertainty, loop delays, digital implementation with sampling and roundoff errors, et cetera, i.e. they are not *robust*. This lack of robustness is deemed unacceptable in applications.

Still, the underlying idea behind (1.14)—that a high loop gain effectively inverts the plant and measures disturbances—is insightful and can be recognized as the guiding principle behind many control design methods. For example, adding an integral action to the controller renders the loop gain infinite at the zero frequency. This is why controllers with integral action can guarantee zero steady-state errors under reference signals converging to a constant value despite the presence of constant disturbances and without the need to know the static gain of the plant. The loop-shaping method, which manipulates the loop gain ideas explicitly, is discussed in more details below.

1.4.3 Loop shaping

Consider again the relation (1.11). Our goal is to render e "small" and u "not too large," in whatever senses. Similarly to treating the open-loop problem in §1.4.1, we assume that the exogenous signals are not correlated and cannot be used to cancel the effect of each other. The requirements above might then be loosely expressed as requirements to reduce gains of S, T_d, and T, while keeping gains of T_c and T not too large. The problem is that these requirements are intrinsically conflicting, because $S + T = 1$, so S and T cannot be made small simultaneously.

An elegant way to circumvent this obstacle can be found via treating the system in Fig. 1.4(c) in the frequency domain. The underlying assumptions are that *(i)* spectra of y_r and n are well separated and *(ii)* only low-frequency components of d are harmful and should be compensated by the controller. Indeed, it is the case in many applications that y_r is relatively slow, sensors are less reliable at high frequencies, and P is low-pass, so high-frequency components of d do not affect y anyway. Adopting these assumptions, we may reformulate the requirements as rendering $|S(\mathrm{j}\omega)|$ and $|T_\mathrm{d}(\mathrm{j}\omega)|$ small at low frequencies, $|T(\mathrm{j}\omega)|$ small at high frequencies, and keeping $|T_\mathrm{c}(\mathrm{j}\omega)|$ and $|T(\mathrm{j}\omega)|$ not too large at all frequencies. From the transient performance point of view, it is beneficial to avoid high resonance peaks in $|T(\mathrm{j}\omega)|$ (smoother transients) and to have a sufficiently high bandwidth of $T(s)$ (faster transients). At the same time, aiming at too high bandwidth of $T(s)$, especially that exceeding the bandwidth of $P(s)$, would result in high peaks in $|T_\mathrm{c}(\mathrm{j}\omega)|$ (cf. the discussion at the end of §1.4.1) and poor robustness. This set of requirements is rather simplified, but still provides a flavor of the ideas behind frequency-domain design methods and complexity of goals involved. The complexity is manifested in both the number of transfer functions involved and the fact that they all are nonlinear functions of the design parameter, $R(s)$.

A workaround is to translate requirements on closed-loop transfer functions to those on the loop transfer function

$$L(s) = P(s)R(s),$$

which is a linear function of $R(s)$. It is readily seen that $S = (1+L)^{-1}$ and $T = L(1+L)^{-1}$. Hence, if the loop gain is high, i.e. if $|L(\mathrm{j}\omega)| \gg 1$, we have small $|S(\mathrm{j}\omega)|$ and, in many cases, $|T_\mathrm{d}(\mathrm{j}\omega)|$. If the loop gain is

low, i.e. if $|L(\mathrm{j}\omega)| \ll 1$, we have a small $|T(\mathrm{j}\omega)|$. These relations are intuitive. Indeed, at frequencies where the loop gain is high, we effectively invert the plant and have good tracking and disturbance attenuation. But this comes at a price of an increased sensitivity to measurement noise. At frequencies where the loop gain is low we effectively disconnect the sensor with its noise, thus rendering measurement noise irrelevant. Less intuitive is that the other requirements can also be expressed in terms of $L(\mathrm{j}\omega)$. This is true for the peaks of $|T(\mathrm{j}\omega)|$ ($L(\mathrm{j}\omega)$ should be sufficiently far from the critical point $(-1, 0)$ on the Nyquist complex plane to avoid them), its bandwidth (connected with the crossover frequency ω_c, at which $|L(\mathrm{j}\omega_\mathrm{c})| = 1$), and even the closed-loop stability (the Nyquist criterion).

A representative loop-shaping cascade design involves the following steps.

1. Pick the required crossover frequency ω_c around which the whole procedure is carried out. Too high ω_c would complicate the design and might render the closed-loop bandwidth too high, while an overly low ω_c would produce a too slow closed-loop system. Arguably, this is the most important choice in the whole procedure. The right choice is normally reached via trial-and-error iterations.
2. Add a low-pass filter, whose bandwidth exceeds ω_c, to provide a required high-frequency roll-off.
3. Add a static gain to render the actual crossover frequency equal to the chosen ω_c.
4. If the closed-loop system is unstable in this stage (check via the Nyquist criterion), or if stability margins are insufficient, add phase lead element(s). A properly parametrized lead does not alter ω_c.
5. If the low-frequency gain is not sufficiently high, add a phase lag element (includes a PI as a special case). A properly parametrized lag almost does not alter ω_c.
6. Simulate the resulted closed-loop system. If some of its properties are unsatisfactory, return to Step 1, change ω_c, and repeat all steps. Sometimes, it might be possible to return to some later steps, keeping ω_c intact and tuning loop gains, roll-off, stability margins, etc.

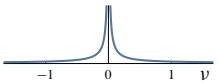
In some situations one may use skewed notch filters as a subtler lead element. It is more localized, yet comes at a lower price. Loop-shaping procedure for systems with lightly damped modes might require different tools. For instance, a phase lag might be required there to increase the distance from the critical point, which might not appear intuitive at first sight. A (not overly detailed) example of the use of loop-shaping ideas can be found in §1.A.4.

Although loop-shaping design guidelines are to a large extent ad hoc, there are analytic results shedding light on their potential limitations. One of them is Bode's gain-phase relation, which quantifies the fact that the magnitude and the phase of the loop frequency response $L(\mathrm{j}\omega)$ cannot be manipulated independent of each other. For example, if $L(s)$ is stable, its phase at every frequency ω_0 must satisfy

$$\arg L(\mathrm{j}\omega_0) = \frac{1}{\pi}\int_{\mathbb{R}} \frac{\mathrm{d}\ln|L(\mathrm{j}\omega_0 \mathrm{e}^{\nu})|}{\mathrm{d}\nu} \ln\Big(\coth\frac{|\nu|}{2}\Big)\mathrm{d}\nu - 2\sum_{i=1}^{n_\mathrm{nmpz}} \arctan\frac{\omega_0 + \operatorname{Im} z_i}{\operatorname{Re} z_i}, \tag{1.16}$$

where $\nu := \ln(\omega/\omega_0)$, $n_\mathrm{nmpz} \in \mathbb{Z}_+$ is the number of nonminimum-phase zeros of $L(s)$, and z_i, with $\operatorname{Re} z_i \geq 0$, are these zeros. The function

$$\ln\Big(\coth\frac{|\nu|}{2}\Big) = \quad = \ln\Big|\frac{\omega + \omega_0}{\omega - \omega_0}\Big|$$



may be thought of as an approximate Dirac delta. The first term on the right-hand side of (1.16) indicates then that the phase depends on the slope of the Bode magnitude plot,

$$\frac{\mathrm{d}\ln|L(\mathrm{j}\omega_0\mathrm{e}^{\nu})|}{\mathrm{d}\nu} = \frac{\mathrm{d}\ln|L(\mathrm{j}\omega)|}{\mathrm{d}\ln\omega} = \frac{\mathrm{d}\log|L(\mathrm{j}\omega)|}{\mathrm{d}\log\omega},$$

around $\omega = \omega_0$. The faster we need to reduce $|L(\mathrm{j}\omega)|$, the more phase lag we have to tolerate. Because the Nyquist stability criterion normally requires the phase lag of $L(\mathrm{j}\omega)$ to be limited around the crossover

frequency, the first term on the right-hand side of (1.16) effectively necessitates the crossover region to be sufficiently wide, so that the negative slope of the gain is not too steep. The second terms on the right-hand side of (1.16) imply that every nonminimum-phase zero aggravates this situation, especially at frequencies exceeding $|z_i|$. This effectively imposes limitations on the attainable crossover frequency for nonminimum-phase systems.

Another classical quantitative relation is Bode's sensitivity integral. If the loop transfer function $L(s)$ is real-rational and strictly proper, then the following relation holds whenever $S(s)$ is stable:

$$\int_0^\infty \ln|S(\mathrm{j}\omega)|\mathrm{d}\omega = \pi\left(\sum_{i=1}^{n_{\mathrm{rhpp}}} \mathrm{Re}\, p_i - \frac{1}{2}\lim_{s\to\infty} sL(s)\right), \tag{1.17}$$

where $n_{\mathrm{rhpp}} \in \mathbb{Z}_+$ is the number of poles of $L(s)$ in the open right half-plane $\mathbb{C}_0$ and p_i, with $\mathrm{Re}\, p_i > 0$, are these poles. Relation (1.17) is known [32] as *generalized*[2] *Bode's sensitivity integral*. The integrand on the left-hand side of (1.17) is negative if $|S(\mathrm{j}\omega)| < 1$ and nonnegative otherwise. Thus, (1.17) effectively says that any feedback design with a strictly proper $sL(s)$ is an *art of tradeoffs*, where a reduction of $|S(\mathrm{j}\omega)|$ at some frequencies inevitably leads to its increase at others. This phenomenon is known as the *waterbed effect*. Bode's sensitivity integral itself does not indicate that $|S(\mathrm{j}\omega)|$ necessarily grows outside the frequency range where the $|S(\mathrm{j}\omega)|$ is pushed down as this range widens. But this kind of results can be derived if some additional information on the decay of $|L(\mathrm{j}\omega)|$ at high frequencies is available.

1.5 Naïve MIMO extensions

Classical controller design methods in the frequency domain are rather well understood for SISO systems. However, their extensibility to MIMO processes might not be natural. The goal of this section is to demonstrate that via a simple academic example.

Consider the closed-loop system in Fig. 1.4(c). To avoid potential complications in the stability analysis, the plant and controller are both static. The plant in all examples is taken as the 2×2 family of constant matrices

$$P(s) = \begin{bmatrix} 1+\alpha & 1-\alpha \\ -1+\alpha & -1-\alpha \end{bmatrix}, \tag{1.18}$$

parametrized by $\alpha \in [0, 1]$. Also, assume that measurements are perfect, i.e. that $n = 0$.

In general, the relation between the signals of interest in the MIMO case is slightly different from that defined by (1.11), mainly because MIMO systems do not necessarily commute. Specifically, we have that

$$\begin{bmatrix} u \\ e \end{bmatrix} = \begin{bmatrix} T_{\mathrm{c}} & -T_{\mathrm{i}} \\ S_{\mathrm{o}} & -T_{\mathrm{d}} \end{bmatrix} \begin{bmatrix} y_{\mathrm{r}} \\ d \end{bmatrix},$$

where

$$\begin{bmatrix} T_{\mathrm{c}}(s) & T_{\mathrm{i}}(s) \\ S_{\mathrm{o}}(s) & T_{\mathrm{d}}(s) \end{bmatrix} := \begin{bmatrix} R(s) \\ I \end{bmatrix} (I + P(s)R(s))^{-1} \begin{bmatrix} I & P(s) \end{bmatrix}. \tag{1.19}$$

Note that $S_{\mathrm{o}} \neq I - T_{\mathrm{i}}$ in general, which explains adding indices to them, "i" for the "input" complementary sensitivity and "o" for the "output" sensitivity.

Example 1.1. First, consider the P (proportional) controller of the form

$$R(s) = k \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$$

[2] Bode's original formula is for a stable $L(s)$ with a pole excess of at least 2, for which the right-hand side of (1.17) is zero.

for some $k \in \mathbb{R}$. With this choice,

$$T_{\mathrm{i}}(s) = \frac{k}{2k+1}\begin{bmatrix} \frac{4\alpha k-1-\alpha}{2\alpha k+1} & \frac{1-\alpha}{2\alpha k+1} \\ \frac{1-\alpha}{2\alpha k+1} & \frac{4\alpha k+1+\alpha}{2\alpha k+1} \end{bmatrix}, \qquad T_{\mathrm{c}}(s) = \frac{k}{2k+1}\begin{bmatrix} \frac{(1+\alpha)k+1}{2\alpha k+1} & \frac{(1-\alpha)k}{2\alpha k+1} \\ -\frac{(1-\alpha)k}{2\alpha k+1} & -\frac{(1+\alpha)k+1}{2\alpha k+1} \end{bmatrix},$$

$$T_{\mathrm{d}}(s) = \frac{1}{2k+1}\begin{bmatrix} \frac{4\alpha k-1-\alpha}{2\alpha k+1} & \frac{1-\alpha}{2\alpha k+1} \\ -\frac{1-\alpha}{2\alpha k+1} & -\frac{4\alpha k+1+\alpha}{2\alpha k+1} \end{bmatrix}, \qquad S_{\mathrm{o}}(s) = \frac{1}{2k+1}\begin{bmatrix} \frac{(1+\alpha)k+1}{2\alpha k+1} & \frac{(1-\alpha)k}{2\alpha k+1} \\ \frac{(1-\alpha)k}{2\alpha k+1} & \frac{(1+\alpha)k+1}{2\alpha k+1} \end{bmatrix}.$$

If $\alpha \neq 0$, the limit as $k \to \infty$ yields the following closed-loop transfer functions:

$$T_{\mathrm{i}}(s) \to I, \quad T_{\mathrm{c}}(s) \to \frac{1}{4\alpha}\begin{bmatrix} 1+\alpha & 1-\alpha \\ -1+\alpha & -1-\alpha \end{bmatrix}, \quad S_{\mathrm{o}}(s) = T_{\mathrm{d}}(s) \to 0,$$

from which

$$u \to u_\infty = \frac{1}{4}\left(\frac{1}{\alpha}\begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix} + \begin{bmatrix} 1 & -1 \\ 1 & -1 \end{bmatrix}\right) y_{\mathrm{r}} - d = P^{-1} y_{\mathrm{r}} - d \quad \text{and} \quad e \to e_\infty = 0, \tag{1.20}$$

similarly to what happens in the SISO case (cf. (1.14)). If $\alpha = 1$, the plant is diagonal, $P = \left[\begin{smallmatrix} 2 & 0 \\ 0 & -2 \end{smallmatrix}\right]$, and the control signal decouples, in a sense that each component of u depends exclusively on the corresponding components of y_{r} and d. Otherwise, there is a cross-coupling in the control signal. The output $y = y_{\mathrm{r}}$ is decoupled for all $\alpha > 0$.

If $\alpha = 0$, the situation is different. In that case the plant $P = \left[\begin{smallmatrix} 1 & 1 \\ -1 & -1 \end{smallmatrix}\right]$ and

$$T_{\mathrm{i}}(s) = \frac{k}{2k+1}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \qquad T_{\mathrm{c}}(s) = \frac{k}{2k+1}\begin{bmatrix} k+1 & k \\ -k & -k-1 \end{bmatrix},$$

$$T_{\mathrm{d}}(s) = \frac{1}{2k+1}\begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix}, \qquad S_{\mathrm{o}}(s) = \frac{k}{2k+1}\begin{bmatrix} 1+1/k & 1 \\ 1 & 1+1/k \end{bmatrix}$$

for all k. In the limit, as $k \to \infty$, we have:

$$T_{\mathrm{i}}(s) = S_{\mathrm{o}}(s) \to \frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \quad T_{\mathrm{c}}(s) \to \frac{1}{2}\begin{bmatrix} k+1 & k \\ -k & -k-1 \end{bmatrix}\Bigg|_{k\to\infty}, \quad T_{\mathrm{d}}(s) \to 0,$$

and then

$$u \to u_\infty = \frac{1}{2}\begin{bmatrix} k+1 & k \\ -k & -k-1 \end{bmatrix}\Bigg|_{k\to\infty} y_{\mathrm{r}} - \frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} d \quad \text{and} \quad e \to e_\infty = \frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} y_{\mathrm{r}}.$$

This is not what we could have expected. The error signal

$$\begin{bmatrix} e_1 \\ e_2 \end{bmatrix} = \frac{1}{2}\begin{bmatrix} y_{\mathrm{r}1} + y_{\mathrm{r}2} \\ y_{\mathrm{r}1} + y_{\mathrm{r}2} \end{bmatrix}$$

vanishes only if $y_{\mathrm{r}1} = -y_{\mathrm{r}2}$. If it happens that $y_{\mathrm{r}1} = y_{\mathrm{r}2}$, then the errors $e = y_{\mathrm{r}}$. This signal actually equals the error signal attained with the choice $R(s) = 0$, i.e. using no controller at all, if $d = 0$. The control signal

$$\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \frac{1}{2}\begin{bmatrix} k(y_{\mathrm{r}1} + y_{\mathrm{r}2}) + y_{\mathrm{r}1} \\ -k(y_{\mathrm{r}1} + y_{\mathrm{r}2}) - y_{\mathrm{r}2} \end{bmatrix} - \frac{1}{2}\begin{bmatrix} d_1 + d_2 \\ d_1 + d_2 \end{bmatrix}$$

is bounded also only if $y_{\mathrm{r}1} = -y_{\mathrm{r}2}$. Moreover, if this condition does not hold, two components of u diverge synchronously, approaching $u_1 = -u_2$ no matter what y_{r} we have (except $y_{\mathrm{r}1} = -y_{\mathrm{r}2}$). ◊

Example 1.2. Now, let

$$R(s) = kI = k\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

In this case we end up with

$$T_{\mathrm{i}}(s) = \frac{k}{4\alpha k^2 - 1}\begin{bmatrix} 4\alpha k - 1 - \alpha & -1 + \alpha \\ 1 - \alpha & 4\alpha k + 1 + \alpha \end{bmatrix}, \quad T_{\mathrm{c}}(s) = \frac{k}{4\alpha k^2 - 1}\begin{bmatrix} (1+\alpha)k - 1 & (1-\alpha)k \\ -(1-\alpha)k & -(1+\alpha)k - 1 \end{bmatrix},$$
$$T_{\mathrm{d}}(s) = \frac{1}{4\alpha k^2 - 1}\begin{bmatrix} 4\alpha k - 1 - \alpha & -1 + \alpha \\ 1 - \alpha & 4\alpha k + 1 + \alpha \end{bmatrix}, \quad S_{\mathrm{o}}(s) = \frac{1}{4\alpha k^2 - 1}\begin{bmatrix} (1+\alpha)k - 1 & (1-\alpha)k \\ -(1-\alpha)k & -(1+\alpha)k - 1 \end{bmatrix}.$$

If $\alpha \neq 0$, we end up with the very same closed-loop properties as in Example 1.1. So consider only the case of $\alpha = 0$, for which

$$T_{\mathrm{d}}(s) = \begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix}, \quad S_{\mathrm{o}}(s) = \begin{bmatrix} -k+1 & -k \\ k & k+1 \end{bmatrix}, \quad T_{\mathrm{i}}(s) = kT_{\mathrm{d}}(s), \quad \text{and} \quad T_{\mathrm{c}}(s) = kS_{\mathrm{o}}(s).$$

Because $R(s)$ is a scaled identity matrix, it commutes with the plant and hence $S_{\mathrm{o}} = I - T_{\mathrm{i}}$. Thus,

$$u \to u_\infty = \left(\begin{bmatrix} -k(k-1) & -k^2 \\ k^2 & k(k+1) \end{bmatrix} y_{\mathrm{r}} - k\begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix} d\right)\Bigg|_{k\to\infty} \quad \text{and} \quad e \to e_\infty = \frac{1}{k}u_\infty.$$

Now the error signal

$$\begin{bmatrix} e_1 \\ e_2 \end{bmatrix} = \begin{bmatrix} -k^2(y_{\mathrm{r}1} + y_{\mathrm{r}2}) + y_{\mathrm{r}1} \\ k(y_{\mathrm{r}1} + y_{\mathrm{r}2}) + y_{\mathrm{r}2} \end{bmatrix} + \begin{bmatrix} -(d_1 + d_2) \\ d_1 + d_2 \end{bmatrix}$$

can be zero only if $y_{\mathrm{r}} = 0$ and $d_1 = -d_2$, as $\det(S_{\mathrm{o}}(s)) = 1$. Moreover, it actually blows up as $k \to \infty$ unless, again, $y_{\mathrm{r}1} = -y_{\mathrm{r}2}$. This looks more related to instability, although the system is stable. The situation with u is similar, as $u = ke$. The only difference is that u grows even faster as k increases. ◊

The moral of the examples above is that MIMO systems are qualitatively different from and way richer than their "poor SISO cousins." Intuitively, discrepancies between SISO and MIMO systems may be connected to differences in processing vectors with various mutual relations between their components. This is indeed the case and the notion is known as the (spatial) direction of MIMO systems. We shall study it later on and explain what happened in Examples 1.1 and 1.2, see Examples 2.4 and 2.5.

1.A Case study: set-point control of a DC motor

To illustrate the ideas discussed in this chapter, consider the problem of controlling the shaft angle θ of a armature-controlled DC motor connected to a rigid mechanical load.

1.A.1 Model

A basic property of armature-controller DC motors is that the generated torque, τ, is proportional to the armature current, i, i.e.

$$\tau(t) = K_{\mathrm{m}} i(t),$$

where K_{m} is called the torque constant of the motor. The torque is then applied to the load (which includes the rotor itself), whose angular velocity $\omega = \dot{\theta}$ is assumed to satisfy

$$J\tfrac{\mathrm{d}}{\mathrm{d}t}\omega(t) + f\omega(t) = \tau(t) + \tau_{\mathrm{e}}(t),$$

Parameter	K_a	K_m [N m/A]	R_a [Ω]	L_a [H]	J [kg m^2]	f [N m s/rad]	τ_{max} [N m]
Value	12	0.126	2.08	0.000264	0.008	0.005	0.235

Table 1.1: Numerical values of motor and load parameters

where J and f are the moment of inertia and the viscous friction coefficient of the load, respectively, and τ_e is an external torque representing possible interactions with the environment. The armature electric circuit satisfies

$$L_a \frac{d}{dt} i(t) + R_a i(t) = v(t) - v_b(t),$$

where L_a and R_a are the inductance and the resistance of the motor armature, respectively, v is the external voltage supplied to the armature and v_b is the back emf (electromotive force), which by Lenz's law is proportional to the rotor angular velocity, i.e.

$$v_b(t) = K_b \omega(t),$$

where K_b is the back emf constant (or speed constant) of the motor. Normally, it satisfies $K_b = K_m$, if expressed in compatible units. Finally, we assume that the armature voltage is generated by an amplifier as

$$v(t) = K_a u(t),$$

where K_a is its gain and u is a normalized control signal, which is assumed to be limited as $|u(t)| \leq 1$ for all t. Numerical values of all parameters above, which will be used throughout this example, are presented in Table 1.1.

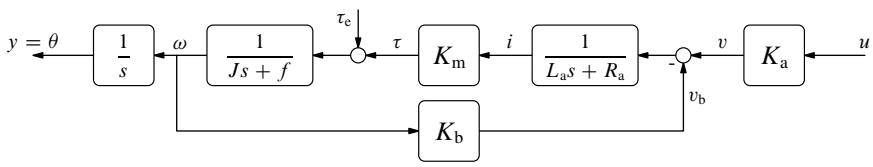


Fig. 1.6: Armature-controlled DC motor connected to a rigid mechanical load

The equations above can be combined in the block-diagram in Fig. 1.6, with the control input u and the controlled output $y = \theta$. This corresponds to the system in Fig. 1.4(a), with $P = P_L$ under

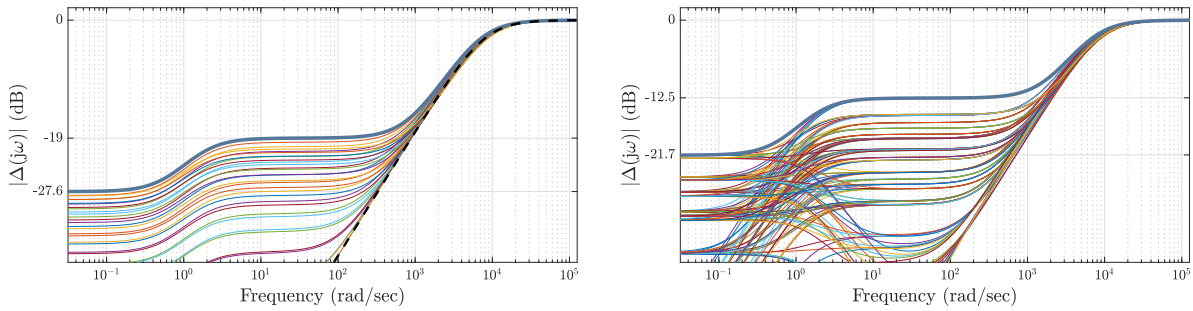
$$P_L(s) := \frac{K_m K_a}{s((L_a s + R_a)(Js + f) + K_m K_b)} \quad \text{and} \quad D(s) = \frac{L_a s + R_a}{K_m K_a} T_e(s). \tag{1.21}$$

This plant P_L is unstable, because of a pole at the origin. The disturbance might be unbounded in its high frequency range because τ_e passes through a system with a non-proper transfer function. Yet this should not be a problem as those high-frequency components are filtered out by the plant, which has a pole excess of two and natural low-pass properties. In many situations the time constant of the electrical part is substantially smaller than that of the mechanical load, i.e. $L_a/R_a \ll J/f$. In this case we may take $L = 0$ without substantial loss of accuracy. This yields the simplified transfer function of the plant

$$P_0(s) = \frac{K_m K_a}{s(R_a J s + (R_a f + K_b K_m))}, \tag{1.22}$$

which is still unstable. In what follows, this P_0 will be used as the plant model in controller design because of its simplicity.

In any case, both P_L and P_0 above are approximations of a real, possibly nonlinear, DC motor P. Moreover, physical parameters of the motor might change in course of its operation. For example, the



(a) $\tilde{R}_\mathrm{a} \in [1.87, 2.29]$, dashed line corresponds to $\tilde{R}_\mathrm{a} = R_\mathrm{a}$ (b) $\tilde{R}_\mathrm{a} \in [1.87, 2.29]$, $\tilde{J} \in [7.2, 8.8] \cdot 10^{-3}$, $\tilde{f} \in [4.5, 5.5] \cdot 10^{-3}$

Fig. 1.7: Modeling errors Δ for various uncertainty sources

resistance increases as the armature circuit heats up and the load changes frequently. On top of this, there are complex high-frequency (like flexible structural modes of the load) and nonlinear (like dry friction) phenomena, which are not taken into account in deriving (1.21). All this implies that we have to account for modeling errors. Such errors are convenient to express in terms of *multiplicative modeling uncertainty*

$$\Delta := 1 - \frac{P}{P_0}, \tag{1.23}$$

where P_0 is a *nominal plant*, that in (1.22) for some chosen values of its parameters, cf. (1.10). If $P = P_L$ from (1.21) under $L_\mathrm{a} \neq 0$ and known other parameters, then the relative modeling error is LTI with the transfer function

$$\Delta(s) = \frac{L_\mathrm{a} s(Js + f)}{(L_\mathrm{a} s + R_\mathrm{a})(Js + f) + K_\mathrm{m} K_\mathrm{b}}. \tag{1.24}$$

This is a stable high-pass system with the zero DC gain, $\Delta(0) = 0$, and a monotonically increasing magnitude of it frequency-response, approaching 1 at high frequencies (the magnitude of its frequency response is shown by the black dashed line in Fig. 1.7(a)). If $P = P_L$ under a mismatched resistance, say $\tilde{R}_\mathrm{a} \neq R_\mathrm{a}$, then

$$\Delta(s) = \frac{(L_\mathrm{a} s + \tilde{R}_\mathrm{a} - R_\mathrm{a})(Js + f)}{(L_\mathrm{a} s + \tilde{R}_\mathrm{a})(Js + f) + K_\mathrm{m} K_\mathrm{b}}, \tag{1.25}$$

which is also stable, but now $\Delta(0) = (\tilde{R}_\mathrm{a} - R_\mathrm{a}) f / (\tilde{R}_\mathrm{a} f + K_\mathrm{m} K_\mathrm{b}) \neq 0$, indicating that modeling error exists also at low frequencies. Thin lines in Fig. 1.7(a) represent $|\Delta(\mathrm{j}\omega)|$ for various values of $\tilde{R}_\mathrm{a} \in [1.87, 2.29]$, which is about 10% deviation from the nominal $\tilde{R}_\mathrm{a} = R_\mathrm{a}$.

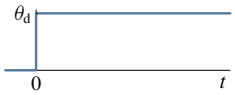
If the value of $\tilde{R}_\mathrm{a}$ is not known a priori, or drifts during the motor operation, then Δ is not a fixed given LTI system, but rather a family of systems. There are many approaches to handle such situations, some of which are discussed in Chapter 8. Conceptually, it may be the simplest to treat modeling mismatch as an arbitrary stable system, whose norm (size) is bounded by corresponding bounds on the actual Δ. Such a treatment significantly simplifies the analysis, rendering it suitable for small-gain arguments [6, Sec. III.2]. However, it introduces *conservatism*, in the sense that the class of modeling errors is enlarged and a negative result for the whole norm-bounded class does not necessarily imply that for the original class of modeling errors. Still, simplicity is a great asset and below we analyze the system via such an upper bound. Obtaining it is relatively simple for constant $\tilde{R}_\mathrm{a}$'s, in which case it is merely the frequency-wise upper bound on $|\Delta(\mathrm{j}\omega)|$ over all admissible $\tilde{R}_\mathrm{a}$, see the thick line in Fig. 1.7(a). This approach readily extends to modeling errors arising from uncertainty in other constant parameters, see for example Fig. 1.7(b), which represents combined uncertainty in R_a, J, and f, each is again about 10% deviation from nominal values. Obtaining an upper-bound Δ in the time-varying or nonlinear case might be more involved.

1.A.2 Reference signal

The control goal here is to rotate the motor shaft to a specific destination angle θ_d as fast as possible and stay there. To simplify the formulae and without loss of generality we may assume that the initial steady state is at $y(0) = 0$ and that $\theta_\mathrm{d} > 0$. This kind of problems arise in numerous applications and are dubbed the *set-point regulation*.

In a constraint-free world, this kind of problems would prompt a step reference signal, like

$$y_\mathrm{r}(t) = \theta_\mathrm{d}\mathbb{1}(t) = \quad ,$$



which is obviously the fastest move to any θ_d. Yet this signal would not be realistic. No motor can generate an infinite torque or admit an infinite input voltage required for a shaft jump. Of course, we may still use the step reference, or its filtered version, essentially as a declaration of intent. This will not render y matching y_r closely in its initial stage anyway. But this will put the control system into saturation modes if θ_d is "too large" and might cause unnecessary, and definitely unwanted, spikes / oscillations in y and u.

A far better approach is to incorporate practical constraints into the choice of y_r. The rationale behind this is twofold. First, it is always healthier to demand from a system something that it can supply than something unachievable anyway. This is a good recipe to avoid complications, like saturation and overloads, which might then be way harder to correct by a controller. Second, the fastest trajectory in a constrained world is almost certainly a nonlinear function of the destination point and system parameters. It is then always beneficial, analysis-wise, to pull nonlinear parts of a control algorithm outside feedback loops, i.e. to y_r.

Apparently, the most restrictive constraint for our problem is that on the torque generated by the motor (equivalently, on the current in its armature circuit). In many cases, at least if f is not too small, the voltage constraint is inactive while the current constraint is satisfied. A possible choice of the reference trajectory then is to rotate the shaft to its target position in minimum time t_f under given load dynamics and the constraint that $|\tau(t)| \le \tau_\mathrm{max}$ for all t and a given $\tau_\mathrm{max} > 0$ dependent on the concrete motor. This trajectory can be derived by minimum-time optimization techniques [4, Ch. 7] for the second-order $1/(Js^2 + fs)$. The optimal torque should be of a bang-bang type, with exactly one switch point t_s in $(0, t_\mathrm{f})$, because both poles of the plant are real. The Laplace transform of the resulting y_r, which is a nonlinear function of J, f, θ_d, and τ_max, is

$$Y_\mathrm{r}(s) = \frac{1}{s(Js+f)}\frac{1-2\mathrm{e}^{-t_\mathrm{s}s}+\mathrm{e}^{-t_\mathrm{f}s}}{s}\tau_\mathrm{max}, \tag{1.26}$$

where the time instances

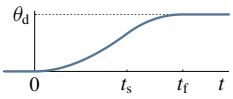
$$t_\mathrm{s} = \frac{f\theta_\mathrm{d}}{\tau_\mathrm{max}} + \frac{J}{f}\ln\Big(1+\sqrt{1-\mathrm{e}^{-f^2\theta_\mathrm{d}/(J\tau_\mathrm{max})}}\Big) \quad \text{and} \quad t_\mathrm{f} = \frac{f\theta_\mathrm{d}}{\tau_\mathrm{max}} + 2\frac{J}{f}\ln\Big(1+\sqrt{1-\mathrm{e}^{-f^2\theta_\mathrm{d}/(J\tau_\mathrm{max})}}\Big)$$

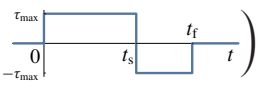
are derived from the conditions

$$\lim_{s\to 0} sY_\mathrm{r}(s) = \theta_\mathrm{d} \quad \text{and} \quad \lim_{Js\to -f}(Js+f)Y_\mathrm{r}(s) = 0 \tag{1.27}$$

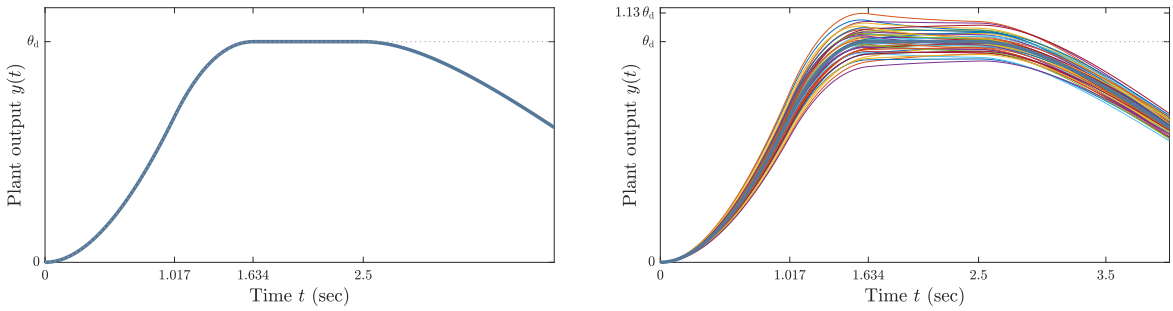
(require $f > 0$, which is naturally assumed to hold true). This reference signal has the form

$$y_\mathrm{r}(t) = \quad \Big(\text{attained by the bang-bang } \tau(t) = \quad \Big). \tag{1.28}$$





The numerical values here and in what follows are taken from Table 1.1 and with $\theta_\mathrm{d} = 6\pi$ [rad], resulting in $t_\mathrm{s} = 1.017$ [s] and $t_\mathrm{f} = 1.634$ [s].



(a) Response of P_L from (1.21)

(b) Responses of P's in Fig. 1.7(b)

Fig. 1.8: Open-loop controlled command responses y, the disturbance τ_e satisfies $\tau_e(t) = 0.1\mathbb{1}(t-2.5)$

It should be mentioned that in some applications there might be additional constraints, like limited load velocity, acceleration, or jerk (e.g. if the load is the cabin of an elevator), accounting for which would result in more complicated reference profiles. Still, handling such constraints on the level of y_r is way easier than on the level of a controller.

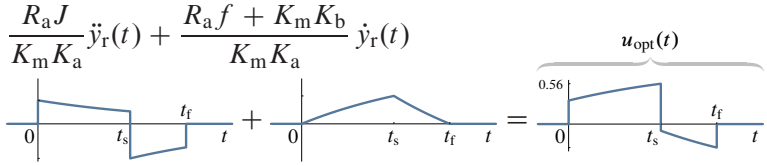
1.A.3 Open-loop control

First, consider the design of an open-loop controller in the configuration presented in Fig. 1.4(b). Let us use the plain plant inversion strategy (1.6) here, namely select

$$R(s) = \frac{1}{P_0(s)} = \frac{R_a J}{K_m K_a} s^2 + \frac{R_a f + K_m K_b}{K_m K_a} s$$

Although this transfer function is non-proper, and hence unstable, it acts only on y_r known analytically. We just need to calculate two first derivatives of y_r above and verify that they are bounded. This is indeed the case and the control trajectory corresponding to the optimal y_r is

$$u(t) = \frac{R_a J}{K_m K_a}\ddot{y}_r(t) + \frac{R_a f + K_m K_b}{K_m K_a}\dot{y}_r(t)$$



(1.29)

Note that the peak value of the input signal, 0.56, is still far from its maximal allowable value of 1. In fact, even when θ_d grows, the peak voltage never exceeds $(K_b/f + R_a/K_m)\tau_{max} = 0.818K_a < K_a$, which justifies ignoring the voltage constraint in the choice of y_r in our case.

As discussed in §1.4.1, the controller above attains (1.4) only under the condition that $P = P_0$, which is not the case. If we consider P as in (1.21), still with $d = 0$, the deviation from the ideal response in the Laplace domain, according to (1.10) and (1.24), is

$$E(s) = Y_r(s) - Y(s) = \Delta(s)Y_r(s) = \frac{L_a}{(L_a s + R_a)(Js + f) + K_m K_b}\frac{1 - 2e^{-t_s s} + e^{-t_f s}}{s}\tau_{max}.$$

This corresponds to an exponentially decaying e, which is proportional to L. If L is very small, as assumed in justifying the use of P_0, the error is also insignificant, as can be seen from the thick line in Fig. 1.8(a) for $t \in [0, 2.5]$, which is virtually indistinguishable from the ideal response shown by the dotted gray line. Moreover, the steady-state value is still θ_d for every L, as $\lim_{t\to\infty} e(t) = \lim_{s\to 0} sE(s) = 0$. This is because L does not affect properties of $P(s)$ at low frequencies.

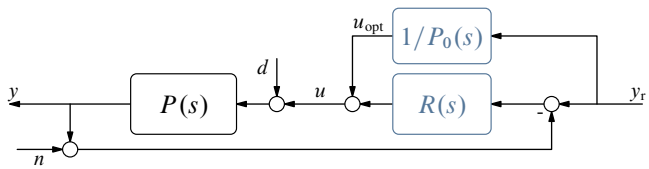


Fig. 1.9: 2DOF closed-loop control aiming at $y = y_{\text{r}}$ under $P = P_0$ and $d = n = 0$

The resistance and load friction do affect low-frequency properties of P, so the open-loop control is less successful if either of them changes. Thin lines in Fig. 1.8(b) for $t \in [0, 2.5]$ represent responses of the system to the control signal (1.29) for plants from the family depicted in Fig. 1.7(b). It is readily seen that those up to 10% deviations from nominal values of R_{a}, J, and f produce rather diverse responses, with steady-state errors up to 8% of the value of θ_{d}. The responses are also visually different from the expected response represented by the thick line in transient phases, demonstrating that open-loop control has no ability to cope with modeling mismatches.

Expectably, the shaft angle θ starts to diverge if a constant external torque τ_{e}, even a very small one, is applied to the system as shown in Fig. 1.6. After all, the plant is unstable. This is seen from the plots in Fig. 1.8, which represent, after $t = 2.5$ [sec], responses to a step torque of -0.1 [N m] applied at that time instance. This external torque causes the angle to diverge for every P, decaying linearly after short transients. This is a typical problem with controlling unstable processes in open loop, where no mechanism for compensating the effect of unmeasurable disturbances exist.

1.A.4 Closed-loop control

Because the transfer function of the plant in (1.21) has a pole excess of 3, no proper $R(s)$ with uniformly high gain will stabilize it. We thus cannot use the arguments of the beginning of §1.4.2. Rather, consider designing R via loop-shaping techniques. The design is simplified by the use of the 2-degrees-of-freedom (2DOF) controller architecture, a simplified version of which is depicted in Fig. 1.9. This scheme assumes that the signal $(P_0^{-1})y_{\text{r}}$ can be implemented, which is indeed the case for our choice of the reference signal in (1.28) as is evident from (1.29).

It can be verified that the tracking error signal $e = y_{\text{r}} - y$ in this case is of the form

$$e = S(\Delta y_{\text{r}} - Pd) + Tn =: e_{y_{\text{r}}} - e_d + e_n, \tag{1.30}$$

where $S = (1 + PR)^{-1}$ is the *actual* sensitivity function, $T = 1 - S$ is the complementary sensitivity function, the modeling error Δ is as defined in (1.23), and the components in the right-hand side are counterparts of what we had in the open-loop case in (1.10). An important property of this closed-loop relation is that in the nominal case, with $P = P_0$ and $\Delta = 0$, the effect of the reference signal on the tracking error $e_{y_{\text{r}}} = 0$, exactly as in the open-loop control case, regardless the choice of R (as long as it stabilizes the system, of course). This is an advantage of the 2DOF architecture over the unity-feedback one in Fig. 1.4(c). The latter, by (1.11), has $e_{y_{\text{r}}} = Sy_{\text{r}}$, which can be zero only in the infeasible uniformly high-gain case. Moreover, the choice of R does not need to take nominal tracking considerations into account, which simplifies the design procedure.

By the logic of the discussion on p. 6, consider each of the components in the right-hand side of (1.30) separately.

$e_{y_{\text{r}}}$: The only source of problems here is the modeling error. Unlike the open-loop case, where $e_{y_{\text{r}}} = \Delta y_{\text{r}}$, the closed-loop sensitivity S is also a factor now. We thus can affect $e_{y_{\text{r}}}$ via a choice of the feedback controller R. A natural approach to reduce $e_{y_{\text{r}}}$ is to decrease $|S(\mathrm{j}\omega)|$ at dominant frequencies of

possible Δy_{r}. To estimate this dominant part, observe that by (1.26)

$$|Y_{\mathrm{r}}(\mathrm{j}\omega)| = \frac{|1 - 2\mathrm{e}^{-\mathrm{j}\omega t_{\mathrm{s}}} + \mathrm{e}^{-\mathrm{j}\omega t_{\mathrm{f}}}|\tau_{\max}}{\omega^2\sqrt{f^2 + J^2\omega^2}} \leq \frac{4\tau_{\max}}{\omega^2\sqrt{f^2 + J^2\omega^2}}, \quad \forall\omega \in \mathbb{R} \tag{1.31}$$

which follows by the triangle inequality (see §A.1.2) and can be arbitrarily tight at every ω under some θ_{d}. Remarkably, the bound above is independent of θ_{d}, so that requirements on S can be formulated independently of the destination point. For example, for uncertainty as in Fig. 1.7(b) we have that an upper bound on the frequency response of ΔY_{r} is monotonically decreasing function of ω such that $|\Delta(\mathrm{j}\omega)Y_{\mathrm{r}}(\mathrm{j}\omega)| \leq 0.1$ for all $\omega > 6.44$. In any case, we should aim at increasing the frequency range in which the sensitivity magnitude is small. This is, in turn, equivalent to increasing the target crossover as much as possible while preserving stability and reasonable stability margins under every possible incarnation of P. It is also worth emphasizing that $S(0) = 0$ under any stabilizing R, because of the integral action in the plant. Hence, $\lim_{t\to\infty} e_{y_{\mathrm{r}}}(t) = 0$ for every y_{r} converging to a constant, despite modeling uncertainty.

e_d: The change with respect to the open-loop $e_d = Pd$ is again the addition of the sensitivity function. Hence, reducing $|S(\mathrm{j}\omega)|$ at low frequencies reduces the effect of the load disturbance. But because the plant itself has an unbounded static gain, zero steady-state error requires an integral action in the controller as well. Thus, a policy to reduce e_d would be similar to that for reducing $e_{y_{\mathrm{r}}}$, with the additional requirement to have an integral action in the controller.

e_n: This component of the error signal is an artefact of closing the feedback loop, it does not exist in open-loop control. In many situations, the spectrum of the measurement noise is dominated by high frequencies. This implies upper bounds on the choice of the crossover frequency and requirements on the high-frequency roll-off of the controller. Although the effect of n is not studied below, some specifications are inspired by its presence.

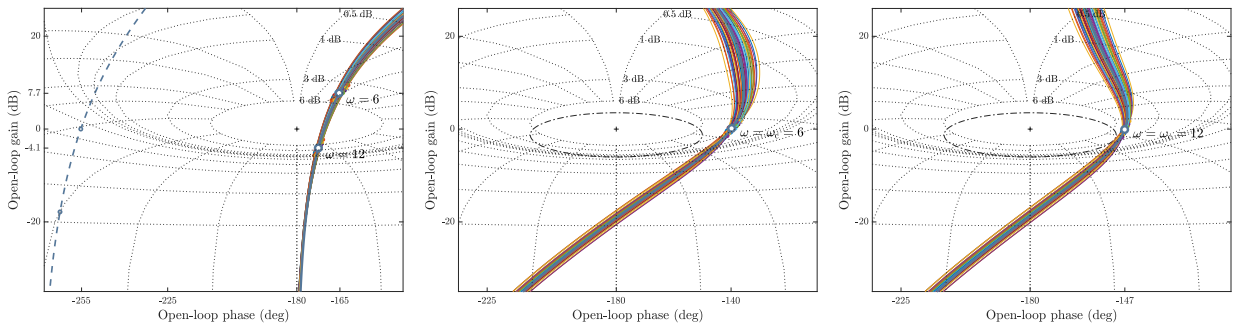
With these considerations in mind, consider the design of the feedback part of the controller, R, for the system in Fig. 1.9 with the following specifications in terms of the nominal plant P_0 from (1.22):

1. an integral action in $R(s)$,
2. a high-frequency roll-off of at least 1 for $R(s)$,
3. a modulus margin[3] of above 0.5, i.e. $|S(\mathrm{j}\omega)| < 2$ for all ω,
4. the control sensitivity magnitude $|T_{\mathrm{c}}(\mathrm{j}\omega)| < 1$ for all ω,
5. as large crossover frequency ω_{c} of the nominal loop P_0R as possible under the requirements above.

Attaining the first requirement is trivial. The second requirement is motivated by the need to render the magnitude of the complementary and control sensitivity frequency responses small at high frequencies to reduce the effect of the measurement noise. Its fulfillment is also technically simple, we just need to end up with a strictly proper $R(s)$. The third one requires the nominal loop to be sufficiently far from the critical point, which will be pursued via adding enough phase margin by the lead component. The specification on $|T_{\mathrm{c}}(\mathrm{j}\omega)|$ aims at avoiding high control effort and is the main source of conflict with the crossover requirement in the last item above. So the design is a sequence of iterations, in which we increase or decrease the target crossover frequency and see whether the resulting control sensitivity is strictly contractive.

Although specifications are formulated for the *design model* P_0, we need to take into account modeling uncertainty. To this end, design steps will be verified for a family of plants, those under the conditions of Fig. 1.7(b), both during checking frequency-domain properties of the designed loops and in time-domain

[3]The modulus margin μ_{m} is the shortest Euclidean distance of the frequency response from the critical point on the Nyquist plane, see §7.3.1 for more details.



(a) $P(\mathrm{j}\omega)$ and $P_0(\mathrm{j}\omega)R_{\mathrm{int}}(\mathrm{j}\omega)$ (dashed) (b) $L(\mathrm{j}\omega)$ for $\omega_{\mathrm{c}} = 6$ (c) $L(\mathrm{j}\omega)$ for $\omega_{\mathrm{c}} = 12$

Fig. 1.10: Nichols charts of the plant and designed loops (dash-dotted line is the 6 dB inverse M-circle)

simulations. The Nichols plots of those plants are depicted by thin solid lines in Fig. 1.10(a). Hollow dots represent two selected frequencies, both of them will be eventually our crossover choices. It can be seen that the frequency responses from the studied class are relatively dense, not far from the nominal one shown by the thick line in Fig. 1.10. As such, nominal margins and crossover are expected to be close to those for all these plants.

Technically, the controller design is done in the following steps. First, the first two requirements above can be guaranteed by augmenting a strictly proper integrator of the form $R_{\mathrm{int}}(s) = k/s$ to the plant and then designing a proper remaining part of the controller. In doing so, we choose the gain k so that the crossover frequency of the resulted augmented plant $P_0 R_{\mathrm{int}}$ is exactly the chosen ω_{c}. This yields

$$R_{\mathrm{int}}(s) = \frac{\omega_{\mathrm{c}}/|P_0(\mathrm{j}\omega_{\mathrm{c}})|}{s}.$$

But the integrator in this controller adds additional 90° of phase lag, leading to loops like that shown by the dashed line in Fig. 1.10(a) for all crossover candidates in $[1, 20]$, which can safely be considered the required range with the uncertain Δy_{r} negligible above $\omega = 6.44$ [rad/sec] and the plant bandwidth below $\omega = 12.54$ [rad/sec]. By the Nyquist arguments, we now need then at least 75° of phase lead at $\omega = \omega_{\mathrm{c}}$ to render the closed-loop system stable. In fact we need more to have a reasonable phase margin. This implies that a single lead controller, whose phase lead is below 90°, would not be enough. A general[4] second-order lead tuned for a given crossover frequency ω_{c} can be written in the form

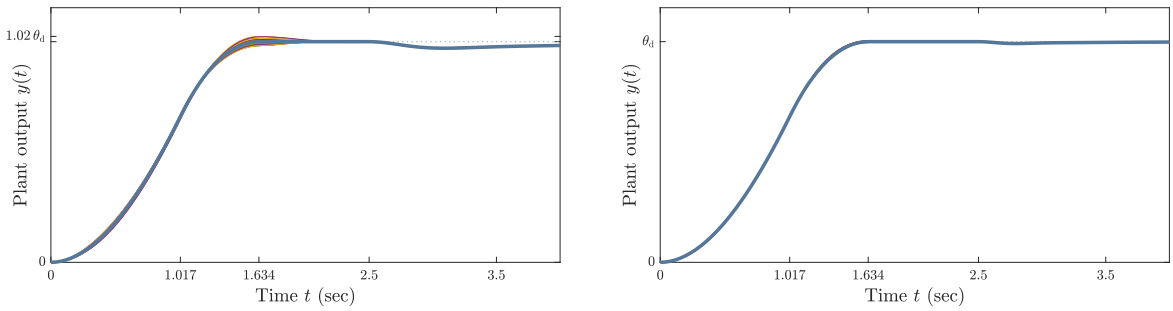
$$R_{\mathrm{lead}}(s) = \frac{\alpha s^2 + 2\zeta\sqrt{\alpha}\omega_{\mathrm{c}} s + \omega_{\mathrm{c}}^2}{s^2 + 2\zeta\sqrt{\alpha}\omega_{\mathrm{c}} s + \alpha\omega_{\mathrm{c}}^2} \quad \text{for some } \alpha > 1 \text{ and } \zeta \in \left[\frac{1}{\sqrt{2}}, \sqrt{2}\right]. \tag{1.32}$$

Its maximal phase lead, always at $\omega = \omega_{\mathrm{c}}$, is $180 - 2\arctan(2\zeta\sqrt{\alpha}/(\alpha - 1))$ [deg]. It grows as ζ decreases and as α increases. An increase of α also implies a decrease of the static gain of R_{lead} and an increase of its high-frequency gain. In other words, α may be thought of as the cost of phase lead in this controller. A small damping factor ζ implies that the phase lead is concentrated in a more narrow frequency band. In our system, because of modeling uncertainty, we would prefer to keep this frequency band wide enough.

The procedure described above is technically simple, although the need to tune three parameters, ω_{c}, α, and ζ, could make it time consuming. The design may be simplified by limiting the consideration to damping factors slightly above one and aiming at a phase margin of about 40°. After some trial and error process, the choice $\omega_{\mathrm{c}} = 6$ [rad/sec] was found suitable, accompanied by $\alpha = 20$ and $\zeta = 3/\sqrt{5} \approx 1.342$. After rounding coefficients, the overall controller

$$R(s) = R_{\mathrm{int}}(s)R_{\mathrm{lead}}(s) = \frac{50(s+3)(s+0.6)}{s(s+60)(s+12)}.$$

[4] A yet more general form would take different damping factors in the numerator and denominator, resulting in *skew notches*.



(a) Design with $\omega_c = 6$

(b) Design with $\omega_c = 12$

Fig. 1.11: Closed-loop controlled command responses y, disturbance τ_e satisfies $\tau_e(t) = 0.1\mathbb{1}(t - 2.5)$

The frequency responses of the resulting loops are shown in Fig. 1.10(b), where the nominal loop is represented by the thick line and loops for other possible plants from the considered set are depicted by thin lines. We can see that the closed-loop system is always stable, with the phase margin about 40° for every plant from the set, which is reasonably large. Checking the closed-loop control sensitivity magnitudes separately, we could see that the bound $|T_c(j\omega)| < 1$ holds for all frequencies, as requited. Likewise, plotting the sensitivity functions for all plants from the studied class would reveal that they are all high-pass filters with the cutoff frequencies[5] above $\omega = 3.2$ [rad/sec] and $|S(j\omega)| < 1.767 < 2$. In fact, the latter could be seen already from the *inverse M-circles* on the open-loop Nichols chart. The dash-dotted line in Fig. 1.10(b) represents one such inverse circle for the 6 dB level, which evidently does not touch loop frequency responses. The closed-loop sensitivity cutoff bound suggests that the effect of modeling uncertainty on the command response is suppressed at dominant frequencies of y_r. This suppression shows up clearly in the closed-loop command responses in Fig. 1.11(a), where the thick line again represents the nominal case and thin lines represent all others, all in the time interval $[0, 2.5]$. These responses are considerably closer to the nominal response than in the open-loop case, which demonstrates clearly advantages of feedback in reducing effects of modeling uncertainty.

Another advantage of feedback can be seen in responses to a step torque disturbance applied at $t = 2.5$. Comparing closed-loop responses in Fig. 1.11(a) with those in Fig. 1.8(b) under open-loop control, we can see that the effect of τ_e on the former is substantially smaller. Moreover, the presence of an integral action in R guarantees that y converges to y_r in steady state, which is a qualitative difference from the diverging open-loop results.

At the end, note that if the requirement on the control sensitivity is relaxed, then even tighter closed-loop results can be obtained. For example, let us double the crossover frequency, to $\omega_c = 12$ [rad/sec]. Keeping the design logic and the parameters α and ζ in (1.32) untouched, we end up with the loop in Fig. 1.10(c) for the controller

$$R(s) = \frac{380(s+6)(s+1.2)}{s(s+120)(s+24)},$$

for which $\max_\omega |T_c(j\omega)| = 3.5$. Although stability margins of this design are smaller than those for $\omega_c = 6$, we still have that $|S(j\omega)| < 2$ for all frequencies, which is seen via the inverse M-circle, depicted by the dash-dotted line in Fig. 1.10(c). As a result of the increase in ω_c we have a wider suppression region of the sensitivity functions, whose cutoff frequencies are now above $\omega = 6.1$ [rad/sec]. And then the time responses of all plants from the family in Fig. 1.7(b) are almost indistinguishable, see Fig. 1.11(b). The sensitivity to torque disturbances is also lower than in the less aggressive design in Fig. 1.11(a).

[5] By the *cutoff frequency* of a high-pass $F(s)$ we understand the largest ω_{coff} such that $|F(j\omega)| \le -3$ dB, $\forall |\omega| \le \omega_{\text{coff}}$.

Part I

Stand-Alone Systems

Chapter 2

Static Systems

Static systems are systems in which the relation between input and output signals (more precisely, between external signals) is *memoryless*. Loosely speaking, this means that the output of the system at any time instance depends on its input at the same time instance only and does not depend on past or future inputs. As such, the time dependence can be dropped in studying static systems and they can be viewed from a pure algebraic (frozen time) perspective. The goal of this chapter is to introduce basic properties of static MIMO systems, those related to processing sizes and spatial directions of involved signals.

2.1 Frozen-time signals and static systems

Throughout this chapter signals are viewed as elements of a finite-dimensional vector space $\mathbb{F}^n$ for $n \in \mathbb{N}$. The notation $\mathbb{F}$ is just a placeholder for either $\mathbb{R}$ or $\mathbb{C}$, the result apply to both unless specifically mentioned. Linear I/O systems are then linear operators from $\mathbb{F}^m$ to $\mathbb{F}^p$, denoted $\mathbb{F}^m \to \mathbb{F}^p$, for suitable dimensions. It is convenient to think of such signals and systems as vectors and matrices, respectively, i.e. as tables, accompanied by corresponding manipulation rules.

Specifically, a signal $x \in \mathbb{F}^n$ can always be presented in the vector form

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \sum_{i=1}^{n} x_i e_i,$$

where e_i stands for the ith standard basis in $\mathbb{F}^n$ and x_i are the corresponding coordinates. Such a representation usually reflects the nature behind components of multi-dimensional signals, with each coordinate representing some specific physical quantity (position, voltage, pressure, etc) of different parts of signals.

Consider now a linear system $G : \mathbb{F}^m \to \mathbb{F}^p$. Denoting by e_i and $\hat{e}_j$ the ith standard basis in $\mathbb{F}^m$ and the jth standard basis in $\mathbb{F}^p$, respectively, and by g_{ij} the ith coordinate of $Ge_j \in \mathbb{F}^p$ in $\{\hat{e}_1, \dots, \hat{e}_p\}$, the relation $y = Gu$ reads

$$\begin{aligned} y &= G(u_1 e_1 + \cdots + u_m e_m) = u_1(Ge_1) + \cdots + u_m(Ge_m) && \text{(by linearity)} \\ &= u_1(g_{11}\hat{e}_1 + \cdots + g_{p1}\hat{e}_p) + \cdots + u_m(g_{1m}\hat{e}_1 + \cdots + g_{pm}\hat{e}_p) \\ &= (g_{11}u_1 + \cdots + g_{1m}u_m)\hat{e}_1 + \cdots + (g_{p1}u_1 + \cdots + g_{pm}u_m)\hat{e}_p = y_1\hat{e}_1 + \cdots + y_p\hat{e}_p. \end{aligned}$$

This implies that the action of G can be presented in the following table (matrix) form:

$$\begin{bmatrix} y_1 \\ \vdots \\ y_p \end{bmatrix} = \begin{bmatrix} g_{11}u_1 + \cdots + g_{1m}u_m \\ \vdots \\ g_{p1}u_1 + \cdots + g_{pm}u_m \end{bmatrix} =: \begin{bmatrix} g_{11} & \cdots & g_{1m} \\ \vdots & & \vdots \\ g_{p1} & \cdots & g_{pm} \end{bmatrix} \begin{bmatrix} u_1 \\ \vdots \\ u_m \end{bmatrix}. \tag{2.1}$$

The matrix in the right-hand side of (2.1) is the matrix representation (or the *system matrix*) of G in the standard bases of $\mathbb{F}^m$ and $\mathbb{F}^p$ (see §A.2.4).

Strictly speaking, vector representations of signals and matrix representations of systems are not signals and systems themselves. For example, these representations change with the change of bases, see §2.1.1 below for details. Nonetheless, we henceforth use the same notation to denote signals / systems and their vector / matrix representations and frequently interchange these notions. In particular, in many cases static systems $\mathbb{F}^m \to \mathbb{F}^p$ will be treated as $p \times m$ matrices and the expression $y = Gu$, which is meant to denote that y is the output of the system G under the input u, will be understood as the matrix multiplication like that in (2.1). This ambiguity normally does not give rise to any problem. At the same time, vectors and matrices may be easier to grasp than abstract elements of vector spaces and mappings.

It follows from (2.1) that the (i, j)th element of a matrix G, g_{ij}, can be interpreted as the transmission between the jth element of the input and the ith element of the output of the corresponding system G. Denote by $g_{\bullet j} \in \mathbb{F}^{p\times 1}$ and $g_{i\bullet} \in \mathbb{F}^{1\times m}$ the jth column and the ith row of G, respectively, so that

$$G = \begin{bmatrix} g_{\bullet 1} & \cdots & g_{\bullet m} \end{bmatrix} = \begin{bmatrix} g_{1\bullet} \\ \vdots \\ g_{p\bullet} \end{bmatrix}.$$

To interpret the columns of G, note that $y = Gu$ can be written as

$$y = g_{\bullet 1}u_1 + \cdots + g_{\bullet m}u_m,$$

so that $g_{\bullet j}$ represents the actuation of the jth element of the input vector by G, hence the *actuator interpretation*. Similarly, the rows of G can be interpreted via the relation

$$y = \begin{bmatrix} g_{1\bullet}u \\ \vdots \\ g_{p\bullet}u \end{bmatrix}.$$

Thus, $g_{i\bullet}$ can be thought of as the *sensor* for the ith measurement channel.

Remark 2.1 (matrix terminology). The following definitions and special matrices are used throughout these notes. A matrix $G \in \mathbb{F}^{p\times m}$ is said to be *square* if $m = p$, *tall* if $p > m$, and *fat* if $p < m$. A matrix G is called *upper (lower) triangular* if its elements $g_{ij} = 0$ whenever $i > j$ $(i < j)$ and *diagonal* if $g_{ij} = 0$ whenever $i \neq j$. The *matrix trace* of a square matrix $G \in \mathbb{F}^{m\times m}$ is defined as $\operatorname{tr}(G) := \sum_{i=1}^{m} g_{ii} \in \mathbb{F}$. ∇

2.1.1 Basis change and similarity transformations

The visualization of signals via their coordinates in the standard basis in $\mathbb{F}^n$ is merely a matter of convention (and convenience). We may consider any other basis in $\mathbb{F}^n$ if it suits us and think of signal components as coordinates in this basis (after all, coordinates in any basis completely determine a signal). This change of viewpoint is done via the *coordinate change* procedure described below.

Let $\{v_1, \ldots, v_n\}$ be a (non-standard) basis on $\mathbb{F}^n$. Any x is uniquely decomposed as

$$x = \tilde{x}_1 v_1 + \cdots + \tilde{x}_n v_n = \begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix} \begin{bmatrix} \tilde{x}_1 \\ \vdots \\ \tilde{x}_n \end{bmatrix} =: T\tilde{x}, \tag{2.2}$$

where $\tilde{x}_i$, $i \in \mathbb{Z}_{1..n}$, are the coordinates of x in this basis. Since all vectors v_i are linearly independent, the matrix $T \in \mathbb{F}^{n\times n}$ is invertible and therefore $\tilde{x} = T^{-1}x$ is unique. We thus ended up with a new vector,

$\tilde{x}$, whose elements are the coordinates of x in $\{v_1, \ldots, v_n\}$. It is important to realize that x and $\tilde{x}$ are different vectors (tables), their elements are in general different. However, they represent the same signal, the difference is in the viewpoint.

Example 2.1. A frequently used basis on $\mathbb{C}^n$ is $\{\phi_1, \phi_2, \ldots, \phi_n\}$, where

$$\phi_i := \frac{1}{n}\begin{bmatrix} 1 \\ (\mathrm{e}^{\mathrm{j}2\pi/n})^{i-1} \\ \vdots \\ (\mathrm{e}^{\mathrm{j}2\pi(n-1)/n})^{i-1} \end{bmatrix}, \quad i \in \mathbb{Z}_{1..n}.$$

If $x \in \mathbb{F}^n$ and x_i are its coordinates in the standard basis, then the coordinates $\tilde{x}_i \in \mathbb{C}$ of x in $\{\phi_1, \ldots, \phi_n\}$ are the discrete Fourier transform (DFT) coefficients of the sequence $\{x_i\}_{i \in \mathbb{Z}_{1..n}}$. In effect, this is a decomposition of a sequence into elementary n-periodic harmonics. Viewing a signal in terms of its DFT coefficients is informative in various applications, for example, in spectral analysis. ◊

To see the effect of a coordinate change on a system matrix $G \in \mathbb{F}^{p \times m}$, let us change the input and output bases as $u \to \tilde{u} := T_u^{-1} u$ and $y \to \tilde{y} := T_y^{-1} y$ for some nonsingular matrices $T_u \in \mathbb{F}^{m \times m}$ and $T_y \in \mathbb{F}^{p \times p}$, whose columns form new bases in the input and output spaces of G. In this case the corresponding matrix relation reads

$$\tilde{y} = T_y^{-1} y = T_y^{-1} G u = T_y^{-1} G T_u \tilde{u}$$

and the matrix $T_y^{-1} G T_u \in \mathbb{F}^{p \times m}$ is the matrix representation of G in these new coordinates. If G is square and both input and output coordinates are transformed to the same basis formed by the columns of a nonsingular matrix T, the resulting matrix in the transformed coordinates, $T^{-1} G T$, is called *similar* to G. This similarity transformation is again just a change of viewpoint and preserves many properties of G.

Example 2.2. Consider a system $G : \mathbb{R}^3 \to \mathbb{R}^3$, whose matrix representation in the standard bases is

$$G = \frac{1}{3}\begin{bmatrix} 5 & -1 & -1 \\ -1 & 5 & -1 \\ -1 & -1 & 5 \end{bmatrix}.$$

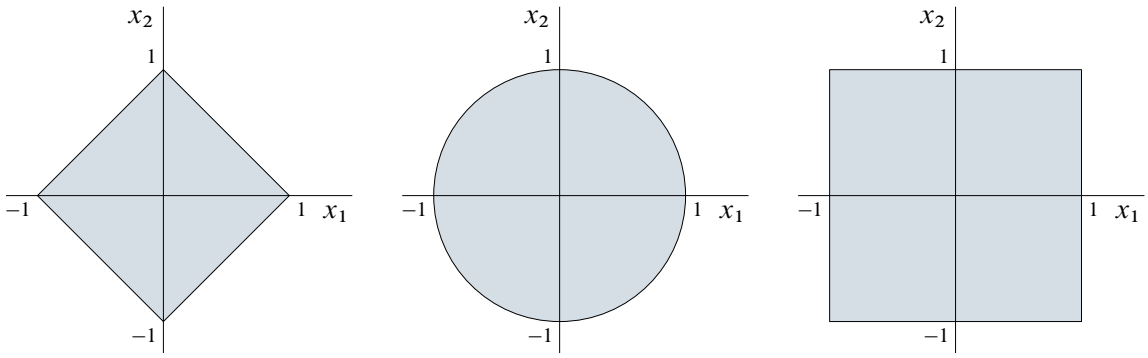
Let us now view both its input and its output signals via their DTF coefficients, i.e. via the coordinates in the basis $\{\phi_1, \ldots, \phi_n\}$ considered in Example 2.1. The corresponding matrices

$$T = \frac{1}{2}\begin{bmatrix} 2 & 2 & 2 \\ 2 & -1+\mathrm{j}\sqrt{3} & -1-\mathrm{j}\sqrt{3} \\ 2 & -1-\mathrm{j}\sqrt{3} & -1+\mathrm{j}\sqrt{3} \end{bmatrix} \quad \text{and} \quad T^{-1} = \frac{1}{6}\begin{bmatrix} 2 & 2 & 2 \\ 2 & -1-\mathrm{j}\sqrt{3} & -1+\mathrm{j}\sqrt{3} \\ 2 & -1+\mathrm{j}\sqrt{3} & -1-\mathrm{j}\sqrt{3} \end{bmatrix},$$

so that

$$T^{-1} G T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}.$$

The diagonal structure of this matrix representation means that the processing of DFT coefficients by G is decoupled. Such a decoupling simplifies the understanding of properties of G and plays an important role in the analysis later on. ◊



(a) $\mathcal{B}_1$ (b) $\mathcal{B}_2$ (c) $\mathcal{B}_\infty$

Fig. 2.1: Unit balls in $\mathbb{R}^2$

2.2 Size matters

As discussed in Section 1.4, control is essentially about making errors small under admissibly-sized control effort. It is obviously important to quantify the meanings of "small" and "admissibly-sized" here. In other words, we have to define metrics by which a small error can be distinguished from a large one and the admissibility of a control effort can be measured. This task is straightforward in the SISO case, where the absolute value of the signal is an ultimate measure of its size and the absolute value of the 1×1 matrix representation of systems of interest is an ultimate measure of their gains. But extending these definitions to MIMO systems is no longer unambiguous and might not even be sufficient.

2.2.1 Signal (vector) norms

By the *size* of a signal $x \in \mathbb{F}^n$ we naturally understand its norm, see §A.1.2. But there are many non-trivially different norms if $n > 1$. The Hölder vector norms, or q-norms, on $\mathbb{F}^n$ are defined as

$$\|x\|_q := \left(\sum_{i=1}^{n} |x_i|^q\right)^{1/q} \quad (1 \le q \le \infty). \tag{2.3}$$

Some frequently used special cases are $\|x\|_1 = \sum_{i=1}^{n} |x_i|$, $\|x\|_\infty = \max_{1 \le i \le n} |x_i|$, and, especially,

$$\|x\|_2 = \left(\sum_{i=1}^{n} |x_i|^2\right)^{1/2}. \tag{2.4}$$

The latter, known as the *Euclidean norm*, is the best known and particularly important because $\mathbb{F}^n$ endowed by it is an inner product space (see below). For this reason, we often omit the subscript "2" from the notation of the Euclidean norm, so that $\|\cdot\|$ stands for $\|\cdot\|_2$ when applied to elements of $\mathbb{F}^n$ hereafter. Note that the quantity $\|x\|_q$ as in (2.3) are sometimes referred to as a norm of x also for $q \in [0, 1)$. But this is not accurate, because such a $\|x\|_q$ does not satisfy the triangle inequality and thus does not qualify as a norm.

It may be useful to define the set of all vectors the norm (size) of which is "small" in some sense. To this end the notion of the unit ball may be used. The *unit ball* $\mathcal{B}_q$ associated with the q-norm on $\mathbb{F}^n$ is defined as

$$\mathcal{B}_q := \{x \mid \|x\|_q \le 1\}.$$

This $\mathcal{B}_q$ is not a subspace of $\mathbb{F}^n$, e.g. it is not closed under the scalar multiplication. The shadowed areas in Fig. 2.1 show unit balls in $\mathbb{R}^2$ determined by some Hölder norms. The boundary of each of these areas

comprises all *unit vectors* in the given metric, i.e. all vectors whose norms equal one. The areas in Fig. 2.1 show clearly that the notion of "smallness" depends on the choice of the metric. For example, the area covered by $\mathcal{B}_\infty$ is twice as large as the area covered by $\mathcal{B}_1$. The choice of a suitable metric depends thus on the concrete application.

Although vector norms are different, they are all comparable. Specifically, all norms on finite-dimensional spaces are *equivalent*, in the sense that given any two vector norms, say $\|\cdot\|_a$ and $\|\cdot\|_b$, there are constants $\gamma_2 \geq \gamma_1 > 0$, which are independent of x, such that

$$\gamma_1 \|x\|_b \leq \|x\|_a \leq \gamma_2 \|x\|_b. \tag{2.5}$$

For example, it is known that $\|x\|_r \leq \|x\|_q \leq n^{1/q-1/r}\|x\|_r$ whenever $q < r$ for all $x \in \mathbb{F}^n$. Among other things, the equivalence of norms means that a bounded vector in one norm remains bounded in any other.

2.2.2 System (matrix) norms

The size of a system can be defined via the maximal amplification that it can provide over all possible input signals. This is the logic of the notion of the *induced norm* of matrices. Specifically, the norm induced by the q vector norm is (see also §A.2)

$$\|G\|_q := \sup_{u \in \mathbb{F}^m, u \neq 0} \frac{\|Gu\|_q}{\|u\|_q} = \sup_{\|u\|_q = 1} \|Gu\|_q = \sup_{\|u\|_q \in \mathcal{B}_q} \|Gu\|_q.$$

In other words, the induced norm is the largest, in terms of its q-norm, vector contained in $G\mathcal{B}_q$ (by linearity, such a vector should lie on the boundary of $G\mathcal{B}_q$). Induced norms for some particular Hölder norms can be calculated explicitly, like

$$\|G\|_1 = \max_{1 \leq j \leq m} \sum_{i=1}^{p} |g_{ij}|, \qquad \text{(column sum)} \tag{2.6a}$$

$$\|G\|_2 = \sqrt{\lambda_{\max}(G'G)}, \qquad \text{(spectral norm)} \tag{2.6b}$$

$$\|G\|_\infty = \max_{1 \leq i \leq p} \sum_{j=1}^{m} |g_{ij}|, \qquad \text{(row sum)} \tag{2.6c}$$

where $\lambda_{\max}$ stands for the maximal eigenvalue (see §2.3.3). It follows from the definition of the induced norm that $\|G_2 G_1 u\|_q \leq \|G_2\|_q \|G_1 u\|_q \leq \|G_2\|_q \|G_1\|_q \|u\|_q$ for all u. Hence,

$$\|G_2 G_1\|_q \leq \|G_2\|_q \|G_1\|_q, \tag{2.7}$$

which is known as the *sub-multiplicative property* of norms. Among q-norms, the spectral norm will be most frequently used throughout these notes. Hence, the subscript "2" will be frequently omitted from its notation and $\|\cdot\|$ should be understood as $\|\cdot\|_2$ when applied to elements from $\mathbb{F}^{p \times m}$.

There may be matrix norms that are *not* induced. For example,

$$\|G\|_{\mathrm{F}} := \sqrt{\mathrm{tr}(G'G)} = \left(\sum_{i=1}^{m} \|Ge_i\|^2\right)^{1/2} = \left(\sum_{i=1}^{p}\sum_{j=1}^{m} |g_{ij}|^2\right)^{1/2} \qquad \text{(Frobenius norm)}, \tag{2.8}$$

which is the matrix version of the Hilbert-Schmidt operator norm, also satisfies all norm conditions on p. 182 and can thus be used as a measure of the size of G. Although this norm is not induced in any signal metric on $\mathbb{F}^m$ and $\mathbb{F}^q$ (the proof of this is not quite trivial), it still can be interpreted in terms of inputs and outputs of the corresponding system. Namely, the Frobenius norm can be thought of as the average, modulo the scaling factor $1/m$, of the Euclidean norms of the responses to standard basis inputs e_i.

Similarly to vector norms, all matrix norms are equivalent in the sense (2.5). For example, it can be shown that

$$\|G\| \le \|G\|_{\mathrm{F}} \le \sqrt{\operatorname{rank}(G)}\,\|G\|,$$

where $\operatorname{rank}(G)$ is its rank, defined at the end of §2.3.1.

It may appear natural to think of a norm of G as its gain. This is indeed the ultimate measure of the smallness of G, for the size of no its output can exceed $\|G\|_q\|u\|_q$. However, norms no longer capture the whole picture in the MIMO case, as illustrated by the example below.

Example 2.3. Let

$$P = \begin{bmatrix} 1+\alpha & 1-\alpha \\ -1+\alpha & -1-\alpha \end{bmatrix}$$

for some $\alpha \in [0, 1]$, which is the plant in (1.18). It can be verified that $\|P\| = 2$, independently of α as long as $|\alpha| \le 1$. Thus, this system can amplify input signals up to a factor of 2, provided the size of signals is measured by their Euclidean norm. This observation can be illustrated by considering inputs of the form $u = \left[\begin{smallmatrix} u_0 \\ u_0 \end{smallmatrix}\right]$, whose $\|u\| = \sqrt{2}|u_0|$. In this case

$$y = \begin{bmatrix} 1+\alpha & 1-\alpha \\ -1+\alpha & -1-\alpha \end{bmatrix}\begin{bmatrix} 1 \\ 1 \end{bmatrix} u_0 = \begin{bmatrix} 2 \\ 2 \end{bmatrix} u_0 \quad \Longrightarrow \quad \|y\| = \sqrt{8}|u_0| = 2\|u\|.$$

In other words, the gain of P is indeed 2 for this class of inputs. But this is not the case for some other inputs. Take for instance $u = \left[\begin{smallmatrix} u_0 \\ -u_0 \end{smallmatrix}\right]$, whose $\|u\| = \sqrt{2}|u_0|$ as well. In this case

$$y = \begin{bmatrix} 1+\alpha & 1-\alpha \\ -1+\alpha & -1-\alpha \end{bmatrix}\begin{bmatrix} 1 \\ -1 \end{bmatrix} u_0 = \begin{bmatrix} 2\alpha \\ 2\alpha \end{bmatrix} u_0 \quad \Longrightarrow \quad \|y\| = \sqrt{8}\alpha|u_0| = 2\alpha\|u\|.$$

If $\alpha \ll 1$, then $\|y\| \ll 2\|u\|$, meaning that the gain of P might be substantially smaller than 2. Moreover, if $\alpha = 0$, then no $u = \left[\begin{smallmatrix} u_0 \\ -u_0 \end{smallmatrix}\right]$ passes the system at all. This implies that the induced norm notion is too limited to characterize amplification / alleviation properties of systems comprehensively. ◊

2.3 Direction matters as well

In light of the discussion in Example 2.3, it is important to identify the property of input signals that determines the amplification of a system. This property, which has no counterpart for scalar-valued signals, is the notion of signal (spatial) direction. By the *direction* of a signal $x \in \mathbb{F}^n$ we understand the 1-dimensional subspace $\operatorname{span}(x) \subset \mathbb{F}^n$. We say that x and y are *co-directed* if their directions coincide, i.e. if $y = \alpha x$ for some nonzero $\alpha \in \mathbb{F}$. All input signals of the form $\left[\begin{smallmatrix} u_0 \\ u_0 \end{smallmatrix}\right]$ in Example 2.3, for which the gain of the system was equivalent to $\|P\|$, are co-directed, with the direction $\operatorname{span}\left(\left[\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right]\right)$. Likewise, all inputs of the form $\left[\begin{smallmatrix} u_0 \\ -u_0 \end{smallmatrix}\right]$ amplified by 2α are also co-directed, now with the direction $\operatorname{span}\left(\left[\begin{smallmatrix} 1 \\ -1 \end{smallmatrix}\right]\right)$. This observation is general. By linearity, the amplification along the same direction scales linearly with the size of inputs.

Directions can be compared using the notion of the *inner product*, see §A.1.2. A possible inner product on $\mathbb{F}^n$ is

$$\langle x, y\rangle_2 := \sum_{i=1}^{n} \overline{y}_i x_i. \tag{2.9}$$

As a matter of fact, the Euclidean vector norm (2.4) is induced by this inner product, viz. $\|x\|^2 = \langle x, x\rangle_2$. To have it consistent with the notation used for the Euclidean norm, the subscript is normally dropped when the inner product (2.9) is used. A mismatch between the directions of signals x and y can be characterized by their normalized inner product $\varkappa_{xy} = \langle x, y\rangle/(\|x\|\|y\|) \in [-1, 1]$, see (A.2) on p. 183 for a precise definition. It is readily seen that for co-directed signals $|\varkappa_{xy}| = 1$. Also, $\varkappa_{xy} = 0$ iff the signals x and y

are *orthogonal*, denoted $x \perp y$, i.e. $\langle x, y\rangle = 0$. This may justify the interpretation of $\varkappa_{xy}$ as the cosine of the *angle* between the directions of x and y, which is the standard cosine for signals on $\mathbb{R}^2$.

Remark 2.2 (adjoint system). With the help of the inner product notion, the *adjoint* of a $p \times m$ matrix G is the $m \times p$ matrix G' such that $\langle Gu, y\rangle = \langle u, G'y\rangle$ for all $u \in \mathbb{F}^m$ and $y \in \mathbb{F}^p$. By (2.9) and (2.1),

$$\begin{aligned}\langle Gu, y\rangle &= \overline{y_1}(g_{11}u_1 + \cdots + g_{1m}u_m) + \cdots + \overline{y_p}(g_{p1}u_1 + \cdots + g_{pm}u_m)\\ &= \overline{(\overline{g_{11}}y_1 + \cdots + \overline{g_{p1}}y_p)}u_1 + \cdots + \overline{(\overline{g_{1m}}y_1 + \cdots + \overline{g_{pm}}y_p)}u_m\\ &= \left\langle u, \begin{bmatrix} \overline{g_{11}} & \cdots & \overline{g_{p1}} \\ \vdots & & \vdots \\ \overline{g_{1m}} & \cdots & \overline{g_{pm}} \end{bmatrix} y \right\rangle,\end{aligned}$$

so that the (i, j)th element of G' is $\overline{g_{ji}}$. If the elements of G are real, G' is merely its *transpose*. The columns (rows) of G are now the rows (columns) of G' modulo the complex conjugation of their elements. Thus, the rows of G can now be thought as actuators of G' and the columns of G—as sensors of G'. ▽

Remark 2.3 (more matrix terminology). A matrix satisfying $G' = G$ is said to be *Hermitian* (or *symmetric*, if it is real). A matrix satisfying $G' = -G$ is called *skew-Hermitian* (or *skew-symmetric*, if it is real). Skew-Hermitian matrices must have pure imaginary diagonal elements (zero, if the matrix is real). A square matrix G is said to be *normal* if $G'G = GG'$. Hermitian matrices are obviously normal, but there are also non-Hermitian normal matrices, like $\left[\begin{smallmatrix} \alpha & \beta \\ -\beta & \alpha \end{smallmatrix}\right]$. ▽

In the remainder of this section direction-related structural properties of MIMO systems are discussed.

2.3.1 Kernel and image spaces

Basic structural notions of linear systems are their kernel and image spaces (see §A.2). Given a matrix (system) $G \in \mathbb{F}^{p\times m}$, its *kernel* (null space) is defined as $\ker G := \{u \in \mathbb{F}^m \mid Gu = 0\}$ and *image* (range) as $\operatorname{Im} G := \{y \in \mathbb{F}^p \mid y = Gu \text{ for some } u \in \mathbb{F}^m\}$. Thus, $\ker G$ can be viewed as the space of all directions of input signals that do not pass G and $\operatorname{Im} G$ is can be viewed as the space of all possible directions of output signals. It is readily seen that the image of a matrix is the span of its columns, $\operatorname{Im} G = \operatorname{span}(g_{\bullet 1}, \ldots, g_{\bullet m})$. We say that the kernel of G is *trivial* if $\ker G = \{0\}$, i.e. if all columns of G are linearly independent.

The kernel of a system has interesting interpretations from the actuation and sensing viewpoints discussed in Section 2.1. Specifically, from the actuation point of view the kernel can be thought of as characterizing the *freedom of choice* for the input u to produce a desired output. Namely, $Gu_1 = Gu_2$ iff $u_1 - u_2 \in \ker G$. Indeed, by linearity we have that $Gu_1 = Gu_2 \iff G(u_1 - u_2) = 0$, whence the claim follows immediately. This property can be reformulated as follows: let u_0 be any vector such that the desired output $y_{\mathrm{d}} = Gu_0$, then the set of all inputs reaching the same y_{d} is given by $u = u_0 + u_{\mathrm{N}}$, where $u_{\mathrm{N}} \in \ker G$ but otherwise *arbitrary*. The term "set of all inputs" is understood here in the "iff" sense, i.e.

($\Longrightarrow$) $y_{\mathrm{d}} = G(u_0 + u_{\mathrm{N}})$ for every $u_{\mathrm{N}} \in \ker G$ and

($\Longleftarrow$) every u for which $y_{\mathrm{d}} = Gu$ can be presented in the form $u = u_0 + u_{\mathrm{N}}$ for some $u_{\mathrm{N}} \in \ker G$.

It is clear now that an input reaching a given output from $\operatorname{Im} G$ is unique iff $\ker G$ is trivial. Taking the sensor viewpoint, $\ker G$ can be viewed as the characterization of the *ambiguity* in u for a given measurement $y_{\mathrm{m}} = Gu$. Indeed, y_{m} can be produced by every input of the form $u_0 + u_{\mathrm{N}}$, where $u_{\mathrm{N}} \in \ker G$. Thus, an input signal can be unambiguously reconstructed from a measurement iff $\ker G = \{0\}$. As a matter of fact, because outputs of Gu necessarily lie in $\operatorname{Im} G$, the latter space characterizes the set of *consistent* measurements, i.e. the measurements that can be "explained" by some input.

Proposition 2.1. $(\ker G)^\perp = \operatorname{Im} G'$ *and* $(\operatorname{Im} G)^\perp = \ker G'$ *for all* $G \in \mathbb{F}^{p\times m}$.

Proof. We know that $u \in \ker G \iff Gu = 0 \iff 0 = \langle Gu, y\rangle = \langle u, G'y\rangle$ for all $y \in \mathbb{F}^p$. Thus, $u \in \ker G$ iff it is orthogonal to all vectors from $\operatorname{Im} G'$, which proves the first statement. The proof of the second statement goes similarly. □

The first statement above actually implies that $\mathbb{F}^m = \ker G \oplus \operatorname{Im} G'$, i.e. any input signal can be decomposed into a part belonging to $\ker G$ (which does not show up in the output) and an orthogonal part belonging to $\operatorname{Im} G'$ (which fully passes to the output). Similarly, the second statement of Proposition 2.1 implies that $\mathbb{F}^p = \operatorname{Im} G \oplus \ker G'$, i.e. that signals, which cannot be reached by via u, are those annihilated by the adjoint system.

Proposition 2.1 can be used to connect dimensions of the kernel and image spaces.

Proposition 2.2. *If* $G \in \mathbb{F}^{p\times m}$, *then* $\dim(\operatorname{Im} G) + \dim(\ker G) = m$.

Because $\dim(\operatorname{Im} G)$ is the dimension of the subspace reachable by inputs and $\dim(\ker G)$ is the dimension of the input subspace "filtered out" by G, Proposition 2.2 can be thought of as the principle of *conservation of dimensions*. Namely, each dimension is either crushed to zero or ends up in the output.

The dimension of the image space is called the *rank* of G. Obviously, $\operatorname{rank}(G)$ is the number of linearly independent columns of G. Less obvious, yet still true, is the fact that $\operatorname{rank}(G)$ also equals the number of linearly independent rows of G. This implies that $\operatorname{rank}(G) = \operatorname{rank}(G')$ and that $\operatorname{rank}(G) \le \min\{p, m\}$. If the equality holds in the latter expression, then G is said to have full rank. A square matrix is nonsingular iff it has full rank.

2.3.2 Diagonal matrices

The kernel and image spaces shed some light on how various input signal directions are processed by systems and what output signal directions are attainable. But they do not explain much about how systems amplify signals. To gain more insight, let us consider a very special class of systems, known as diagonal.

Arguably, diagonal matrices defined in Remark 2.1 constitute the next simplest class of matrices after scaled identity matrices. Considered as a system, a matrix $G \in \mathbb{F}^{m\times m}$ is *diagonal* if

$$Ge_i = g_i e_i, \quad \forall i \in \mathbb{Z}_{1..m}, \tag{2.10}$$

for some scalars $g_i \in \mathbb{F}$. In other words, systems described by diagonal matrices do not change directions for inputs co-directed with the standard basis. It is easy to see that the elements of diagonal matrices indeed satisfy $g_{ij} = 0$ whenever $i \neq j$. We denote diagonal matrices as $G = \operatorname{diag}\{g_1, \dots, g_m\}$ or even as the shorter form $G = \operatorname{diag}\{g_i\}$. It is readily seen that $I = \operatorname{diag}\{1, \dots, 1\}$. Operations on diagonal matrices can be carried out in terms of individual diagonal elements. For example, it is easy to see that $\operatorname{diag}\{g_i\}' = \operatorname{diag}\{\overline{g_i}\}$ and $\operatorname{diag}\{g_i\}$ is invertible iff $g_i \neq 0$ for all $i \in \mathbb{Z}_{1..m}$, with $\operatorname{diag}\{g_i\}^{-1} = \operatorname{diag}\{1/g_i\}$. Also, it can be shown that induced norms of a diagonal matrix are $\|\operatorname{diag}\{g_i\}\|_q = \max_i |g_i|$ for all $q \ge 1$.

A diagonal $m \times m$ system can be seen as a collection of m independent (decoupled) SISO systems. All inputs, whose direction is $\operatorname{span}(e_i)$, are amplified by $|g_i|$ then. Moreover, the response to a general input in this case is $y = G(\upsilon_1 e_1 + \cdots + \upsilon_m e_m) = \upsilon_1 g_1 e_1 + \cdots + \upsilon_m g_m e_m$, from which, by the orthonormal property of the standard basis,

$$\|y\|^2 = \sum_{i=1}^{m} |g_i|^2 |\upsilon_i|^2.$$

Therefore, $|g_i|$ is the gain of G for the ith standard basis coordinate with respect to all inputs. Thus, the knowledge of gains and their indices lets us know in what input directions we can expect larger amplifications and in what directions the amplification is lower. We also know in what output directions shall we expect amplified / attenuated signals. This completes the picture in the diagonal case.

2.3.3 Eigenvalues and eigenvectors

It is worth remembering at this point that the diagonal structure is *basis dependent*. In other words, a similarity transformation might destroy the decoupling structure of diagonal matrices. At the same time, one may expect that the reverse path, i.e. changing bases to convert an arbitrary system to diagonal, can be exploited (see Example 2.2).

Taking into account (2.2), finding a basis in which a system G is diagonal can be carried out via finding a *nonzero* $u \in \mathbb{F}^m$ such that

$$Gu = \lambda u \tag{2.11}$$

for some $\lambda \in \mathbb{F}$. This equality can be rewritten as $(\lambda I - G)u = 0$, whence the existence of the required u is equivalent to the existence of a scalar $\lambda \in \mathbb{F}$ such that the matrix $\lambda I - G$ is singular. The latter question always has an affirmative answer. In fact, there are exactly m such scalars, λ_i, $i \in \mathbb{Z}_{1..m}$, which are called *eigenvalues* of G. Any $u \neq 0$ for which (2.11) holds for a given λ is called the *eigenvector* associated with this eigenvalue.

Eigenvalues are the solutions of the *characteristic equation*

$$\det(\lambda I - G) =: \chi_G(\lambda) = 0.$$

The set of all eigenvalues of G is called its *spectrum* and denoted $\operatorname{spec}(G)$. Matrices whose spectrum is located in the open left half-plane $\mathbb{C} \setminus \bar{\mathbb{C}}_0$ are called *Hurwitz* and those with the spectrum in the open unit disk $\mathbb{D}$ are called *Schur*. It is easy to see that similarity transformations do not affect the spectrum, i.e. $\operatorname{spec}(G) = \operatorname{spec}(T^{-1}GT)$ for all nonsingular T. The maximal modulus of eigenvalues is called the *spectral radius*, denoted $\rho(G) := \max_{1\le i\le m}|\lambda_i| \ge 0$. If all eigenvalues of G are real, we denote the largest and the smallest of them as $\lambda_{\max}(G)$ and $\lambda_{\min}(G)$, respectively. Eigenvalues of real matrices are not necessarily real. Still, it can be shown that the eigenvalues of Hermitian matrices are real, so that the definition of the matrix spectral norm on p. 27 always makes sense. As a matter of fact, eigenvalues of skew-Hermitian matrices are always located on the imaginary axis. If λ_i is a root of multiplicity ν_i of $\chi_G(\lambda)$, we say that λ_i is an eigenvalue of G of *algebraic multiplicity* ν_i. If $\nu_i = 1$, we say that the eigenvalue λ_i is *simple*, otherwise we say that λ_i is a *repeated* eigenvalue of G. The *geometric multiplicity*, μ_i, of λ_i is defined as the dimension of $\ker(\lambda_i I - G)$ or, equivalently (cf. Proposition 2.2), as $m - \operatorname{rank}(\lambda_i I - G)$. Algebraic and geometric multiplicities need not to be the same (in fact, the latter is always smaller than or equal to the former). For example, both I_2 and $\left[\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right]$ have an eigenvalue $\lambda = 1$ of an algebraic multiplicity 2, as the characteristic polynomial is $(\lambda - 1)^2$ in both cases, whereas the geometric multiplicity is 2 for I_2 and 1 for $\left[\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right]$, because $\ker\left[\begin{smallmatrix} 0 & -1 \\ 0 & 0 \end{smallmatrix}\right] = \operatorname{span}(e_1)$. The smallest $\iota_i \in \mathbb{N}$ such that the dimension of $\ker(\lambda_i I - G)^{\iota_i}$ equals its algebraic multiplicity ν_i is called the *index* of the eigenvalue λ_i. This index equals the dimension of the largest Jordan block associated with λ_i and does not exceed the geometric multiplicity of λ_i, i.e. we have that $\iota_i \le \mu_i \le \nu_i$.

Returning to (2.11), eigenvectors show the directions that are not altered by G and the corresponding eigenvalues can be thought of as the (scalar) gains of G along these directions. If G has m linearly independent eigenvectors u_i, they form a basis in $\mathbb{F}^m$ (known as the *eigenbasis*). The matrix representation of G in the eigenbasis, $T^{-1}GT$ for $T = \left[\begin{smallmatrix} u_1 & \cdots & u_m \end{smallmatrix}\right]$, is diagonal. However, the utility of such a diagonalization is quite limited for the following two reasons.

1. Not every matrix can be diagonalized this way. First, the procedure does not apply to non-square matrices. Second, not every square matrix has m linearly independent eigenvectors. For example, $\left[\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right]$ has only one eigenvector corresponding to its double eigenvalue at $\lambda = 1$, $u = e_1$, and is thus not diagonalizable. In general, there are m independent eigenvectors iff the geometric multiplicity of each eigenvalue equals its algebraic multiplicity.

2. Even if a matrix is diagonalizable, the corresponding basis is not necessarily orthogonal. As such, we can know what is the gain for inputs in the direction of a single eigenvector, but not necessarily how a general input is amplified. Indeed, consider a general input signal $u = \upsilon_1 u_1 + \cdots + \upsilon_m u_m$, decomposed with respect to an eigenbasis $\{u_1, \ldots, u_m\}$. The output is $y = \upsilon_1 \lambda_1 u_1 + \cdots + \upsilon_m \lambda_m u_m$, but now $\|y\|^2 \neq \sum_i |g_i|^2 |\lambda_i|^2$. In other words, $|\lambda_i|$ are not necessarily gains of G in whatever sense. This is easy to see from the system $\left[\begin{smallmatrix} 1 & \alpha \\ 0 & 0 \end{smallmatrix}\right]$, whose eigenvalues at $\lambda = 0$ and $\lambda = 1$ may be arbitrarily smaller than its spectral norm (worst-case gain), which equals $\sqrt{1 + |\alpha|^2}$.

Thus, eigenvectors do not offer an informative basis for the analysis of system gains, unless the eigenbasis of the system matrix is orthonormal. The latter happens iff the matrix is normal, see the definition in Remark 2.3.

There is still an unexploited degree of freedom in the quest for having a diagonalization, suitable for the analysis of system gains. Namely, we may have different bases changes for input and output signals. This direction will eventually lead to the required result, see §2.3.5 below. But before some preliminary definitions are needed.

2.3.4 Unitary matrices

Signals do not preserve their size in all directions when pass through system (equivalently, when their coordinates are changed) in general. However, some systems do preserve input size in all directions. A matrix (system) $G \in \mathbb{F}^{m \times m}$ such that $\|Gu\| = \|u\|$ for all $u \in \mathbb{F}^m$ is said to be *unitary* (or norm preserving). Clearly, unitary matrices transform any element from the unit ball $\mathcal{B}_2$ in the Euclidean norm into another element from $\mathcal{B}_2$.

With the help of the polarization identity (see p. 183), it can be seen that a matrix G is unitary iff $\langle u, v \rangle = \langle Gu, Gv \rangle$ for all $u, v \in \mathbb{F}^m$. In other words, G is unitary iff $\langle u, v \rangle = \langle u, G'Gv \rangle$ or, equivalently, iff $\langle u, (I - G'G)v \rangle = 0$ for all $u, v \in \mathbb{F}^m$. This, in turn, implies G is unitary iff $G'G = I$. It follows by similar arguments that a matrix G is unitary iff $GG' = I$. These facts imply that both rows and columns of an $m \times m$ unitary matrix constitute *orthonormal bases* in $\mathbb{F}^m$. Moreover, unitary matrices can be defined via the equality $G^{-1} = G'$.

Examples of unitary matrices are the plain rotation and reflection matrices

$$R_\theta = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \quad \text{and} \quad Q_\theta = \begin{bmatrix} \cos 2\theta & \sin 2\theta \\ \sin 2\theta & -\cos 2\theta \end{bmatrix}, \tag{2.12}$$

respectively. The former rotates every vector from $\mathbb{R}^2$ by the angle θ counterclockwise (cf. the equality $\langle R_\theta u, u \rangle = \cos\theta(u_1^2 + u_2^2) = \|u\|^2 \cos\theta$). The latter reflects every such vector about the line passing through the origin at the angle θ with respect to the abscissa. In general, a unitary matrix G is said to be a rotation matrix if $\det(G) = 1$ and every unitary matrix on $\mathbb{R}^{m \times m}$ can be presented as a combination of rotation and reflection matrices.

2.3.5 Singular value decomposition

The result below, presented without the proof, establishes that all matrices can be diagonalized by orthonormal basis changes of their input and output spaces.

Theorem 2.3 (The Singular Value Decomposition)**.** *Given any $G \in \mathbb{F}^{p \times m}$, there are unitary*

$$U = \begin{bmatrix} u_1 & u_2 & \cdots & u_p \end{bmatrix} \in \mathbb{F}^{p \times p} \quad \textit{and} \quad V = \begin{bmatrix} v_1 & v_2 & \cdots & v_m \end{bmatrix} \in \mathbb{F}^{m \times m}$$

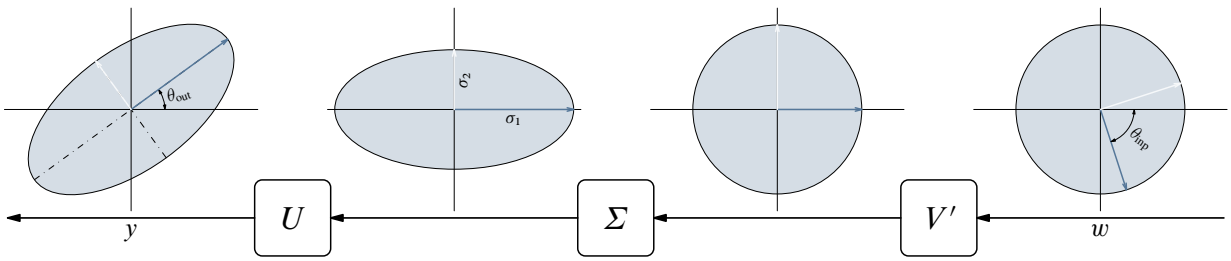


Fig. 2.2: SVD in $\mathbb{R}^2$ for rotation U and V matrices ($\theta_{\text{inp}} = -2\pi/5$, $\sigma_1 = \sqrt{2}$, $\sigma_2 = 1/\sqrt{2}$, and $\theta_{\text{out}} = \pi/5$)

and a diagonal (with either the last row, or the last column, or both empty)

$$\Sigma = \begin{bmatrix} \sigma_1 & & 0 & 0 \\ & \ddots & & \vdots \\ 0 & & \sigma_{\min\{p,m\}} & 0 \\ 0 & \cdots & 0 & 0 \end{bmatrix} \in \mathbb{R}_+^{p\times m}, \quad \textit{with } \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\min\{p,m\}} \geq 0,$$

such that

$$G = U\Sigma V' = \sum_{i=1}^{\min\{p,m\}} \sigma_i\, u_i v_i'. \tag{2.13}$$

The scalars $\sigma_i \geq 0$ defined in Theorem 2.3 are called the *singular values* of G (if the affiliation might be ambiguous, we write $\sigma_i(G)$) and the vectors u_i and v_i are called the left and right *singular vectors* of G, respectively. Singular values and vectors are associated with certain eigensystem problems. To see these relations, observe that (2.13) implies that $Gv_i = \sigma_i u_i$ for all $i = 1, \ldots, \min\{p, m\}$ and, likewise, that $G'u_i = \sigma_i v_i$. Hence,

$$G'Gv_i = \sigma_i^2 v_i \quad \text{and} \quad GG'u_i = \sigma_i^2 u_i,$$

so that σ_i^2 is an eigenvalue of both $G'G$ and GG', u_i is the corresponding eigenvector of GG', and v_i is the corresponding eigenvector of $G'G$. Clearly, left and right singular vectors coincide iff G is normal.

Remark 2.4 (plane geometry of SVD). When considered on $\mathbb{R}^2$, SVD has a neat geometrical interpretation. To see this, consider SVD for

$$G = \begin{bmatrix} \cos\theta_{\text{out}} & -\sin\theta_{\text{out}} \\ \sin\theta_{\text{out}} & \cos\theta_{\text{out}} \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} \begin{bmatrix} \cos\theta_{\text{inp}} & \sin\theta_{\text{inp}} \\ -\sin\theta_{\text{inp}} & \cos\theta_{\text{inp}} \end{bmatrix}.$$

This G maps the unit ball $\mathcal{B}_2$ into the interior of the ellipse with semi-axes of σ_1 and σ_2 rotated by the angle θ_{out}, see Fig. 2.2. On its way of doing this, G effectively performs the following transformations:

1. V' rotates the unit ball by $-\theta_{\text{inp}}$, then
2. Σ scales the rotated unit ball to an ellipse, and finally
3. U rotates the resulting ellipse once again by θ_{out}.

We may thus always think of U and V as rotations (or reflections) and of Σ as scaling. ▽

SVD can help to understand the structure of the response of $G : w \mapsto y$ to a general input w. To see that, let by $w_i = v_i'w$ be the ith coordinate of w in the input basis $\{v_1, \ldots, v_m\}$. By the second equality of (2.13) and the orthonormality of $\{v_1, \ldots, v_p\}$,

$$y = Gw = \sum_{i=1}^{\min\{p,m\}} \sigma_i u_i v_i' w = \sum_{i=1}^{\min\{p,m\}} \sigma_i w_i u_i.$$

Thus, the scalars $\sigma_i w_i$ are the coordinates of y in the output basis $\{u_1, \dots, u_p\}$. In other words, when the input and output bases are changed to those of the right and left singular vectors, respectively, the system becomes a collection of $\min\{p, m\}$ independent (decoupled) SISO systems, whose gains are the singular values. This analogy with diagonal systems discussed in §2.3.2 can be continued via the relation

$$\|y\|^2 = \sum_{i=1}^{\min\{p,m\}} \sigma_i^2 |w_i|^2, \tag{2.14}$$

which uses the fact that the basis $\{u_1, \dots, u_p\}$ is orthonormal. Unlike diagonal matrices, inputs co-directed with the input basis vector v_i change their direction to u_i in the output of the system.

The maximal and minimal singular values, denoted by $\overline{\sigma}(G)$ and $\underline{\sigma}(G)$, respectively, determine then the largest and smallest gains of a matrix G. This is established by the following result.

Proposition 2.4. *Given a matrix $G \in \mathbb{F}^{p\times m}$,*

$$\overline{\sigma}(G) = \max_{\|w\|=1} \|Gw\|$$

(i.e. $\overline{\sigma}(G) = \|G\|$) with a maximizing $w = v_1$. If, in addition, G is tall, i.e. $p \geq m$, then

$$\underline{\sigma}(G) = \min_{\|w\|=1} \|Gw\|$$

with a minimizing $w = v_m$. Moreover, if $p = m$ and $\det(G) \neq 0$, then $\underline{\sigma}(G) = 1/\|G^{-1}\|$.

Proof. The condition $\|w\| = 1$ in terms of the coordinates w_i of the input w in the orthonormal basis $\{v_1, \dots, v_m\}$ reads $|w_1|^2 + \cdots + |w_m|^2 = 1$. It then follows from (2.14) that

$$\|Gw\| \leq \overline{\sigma}(G)\sqrt{|w_1|^2 + \cdots + |w_{\min\{p,m\}}|^2} \leq \overline{\sigma}(G)\sqrt{|w_1|^2 + \cdots + |w_m|^2} = \overline{\sigma}(G)$$

and the equality is obviously achieved if $w_1 = 1$ and $w_2 = \cdots = w_{\min\{p,m\}} = 0$, i.e. if $w = v_1$. This proves the first statement. Note that the maximizing w is not unique. Obviously, the same bound is attained with $w = -v_1$. Moreover, if $\sigma_1 = \cdots = \sigma_k$ for some $k \in \mathbb{Z}_{2..m}$, then $w = w_1 v_1 + \cdots + w_k v_k$ also yields $\|Gw\| = \overline{\sigma}(G)$ for all coefficients w_i such that $|w_1|^2 + \cdots + |w_k|^2 = 1$.

Now, if G is tall, then $p \geq m$ and (2.14) yields

$$\|Gw\| \geq \underline{\sigma}(G)\sqrt{|w_1|^2 + \cdots + |w_m|^2} = \underline{\sigma}(G).$$

The equality is achieved here if $w_1 = \cdots = w_{m-1} = 0$ and $w_m = 1$, i.e. if $w = v_m$. Similarly to the maximizing input case, this choice is not unique.

The last statement follows by the fact that $\overline{\sigma}(G^{-1}) = \overline{\sigma}(V\Sigma^{-1}U') = 1/\underline{\sigma}(G)$. □

With the insight offered by SVD, we may better understand the MIMO examples of Section 1.5.

Example 2.4. Return to Example 1.1 on p. 10. The SVD of plant (1.18) for any $\alpha \in [0, 1]$ is

$$P = \left(\frac{1}{\sqrt{2}}\begin{bmatrix} 1 & -1 \\ -1 & -1 \end{bmatrix}\right)\begin{bmatrix} 2 & 0 \\ 0 & 2\alpha \end{bmatrix}\left(\frac{1}{\sqrt{2}}\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}\right)'. \tag{2.15}$$

The parameter α thus affects one of its gains without altering gain directions. If $\alpha = 0$, one gain vanishes. Hence, the output in the direction $\operatorname{span}(u_2) = \operatorname{span}\left(\left[\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right]\right)$ is always zero, no matter what input signals are applied. An implication of this is that only references $y_r \in \mathbb{R}^2 \ominus \operatorname{span}(u_2) = \operatorname{span}(u_1) = \operatorname{span}\left(\left[\begin{smallmatrix} -1 \\ 1 \end{smallmatrix}\right]\right)$ can be tracked, which agrees with the conclusion of Example 1.1. At the same time, disturbance signals d directed as $\operatorname{span}(v_2) = \operatorname{span}\left(\left[\begin{smallmatrix} -1 \\ 1 \end{smallmatrix}\right]\right)$ cannot affect the controlled output, which is also what we saw.

If $\alpha > 0$, we have that $u = P^{-1}r$. Consider then the gains and their direction in P^{-1}. We have that

$$\begin{aligned} P^{-1} = V\Sigma^{-1}U' &= \left(\frac{1}{\sqrt{2}}\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}\right)\begin{bmatrix} 1/2 & 0 \\ 0 & 1/(2\alpha) \end{bmatrix}\left(\frac{1}{\sqrt{2}}\begin{bmatrix} 1 & -1 \\ -1 & -1 \end{bmatrix}\right)' \\ &= \left(\frac{1}{\sqrt{2}}\begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix}\right)\begin{bmatrix} 1/(2\alpha) & 0 \\ 0 & 1/2 \end{bmatrix}\left(\frac{1}{\sqrt{2}}\begin{bmatrix} -1 & 1 \\ -1 & -1 \end{bmatrix}\right)', \end{aligned}$$

which implies that as α decreases, one gain increases. The output direction of this increase is $\operatorname{span}(v_2) = \operatorname{span}\left(\left[\begin{smallmatrix} -1 \\ 1 \end{smallmatrix}\right]\right)$, which is exactly what we can see in (1.20). From the latter equation we can also see that if y_{r} is in the direction of $\operatorname{span}\left(\left[\begin{smallmatrix} -1 \\ 1 \end{smallmatrix}\right]\right)$, the parameter α does not affect the control signal. This could have been seen without calculating the control input explicitly, just by noticing that the input direction of the gain $1/2\alpha$, which is $\operatorname{span}(u_1) = \operatorname{span}\left(\left[\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right]\right)$, is perpendicular to $\operatorname{span}\left(\left[\begin{smallmatrix} -1 \\ 1 \end{smallmatrix}\right]\right)$. ◊

Example 2.5. Consider now Example 1.2 on p. 12 for $\alpha = 0$. What makes it so special are the facts that *(i)* the input and output directions corresponding to the only nonzero plant gain in (2.15) happen to be perpendicular and *(ii)* the controller, $R = kI$, does not alter directions of its input signals. Thus, the resulting plant input is $k(r - y) + d$ and its feedback component is always in $\operatorname{span}(u_1) = \operatorname{span}\left(\left[\begin{smallmatrix} -1 \\ 1 \end{smallmatrix}\right]\right)$, no matter what are y_{r} and d. But this direction is filtered out by the plant, because $v_1'u_1 = 0$. In other words, the system *effectively acts in open loop*. The control input can then be calculated as

$$k\big(r - P(k(r - y) + d)\big) = k\big(r - P(kr + d)\big) = k(I - kP)r - kPd,$$

agreeing with that in Example 1.2 and explaining it. These explanations suggest that the knowledge of directional properties of the plant can be important in designing feedback controllers. ◊

The Frobenius matrix norm of G can also be expressed in terms of its singular values,

$$\|G\|_{\mathrm{F}}^2 := \operatorname{tr}(G'G) = \operatorname{tr}(V\Sigma'U'U\Sigma V') = \operatorname{tr}(V\Sigma^2V') = \operatorname{tr}(\Sigma^2V'V) = \operatorname{tr}(\Sigma^2),$$

where the facts that $\Sigma' = \Sigma$ and $\operatorname{tr}(G_1G_2) = \operatorname{tr}(G_2G_1)$ are used. Thus, we end up with

$$\|G\|_{\mathrm{F}}^2 = \sum_{i=1}^{\min\{p,m\}} \sigma_i^2. \tag{2.16}$$

The singular value decomposition is a numerically reliable[1] tool of concretizing structural properties of matrices. The following result can be used towards this end.

Proposition 2.5. *If singular values of $G \in \mathbb{F}^{p\times m}$ satisfy $\sigma_1 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_{\min\{p,m\}} = 0$ for some $r \leq \min\{p, m\}$, then*

- $\operatorname{Im} G = \operatorname{span}(u_1, \ldots, u_r)$;
- $\operatorname{rank}(G) = r$;
- $\ker G = \operatorname{span}(v_{r+1}, \ldots, v_m)$.

Proof. Bring in full column rank matrices $U_r := \begin{bmatrix} u_1 & \cdots & u_r \end{bmatrix}$ and $V_r := \begin{bmatrix} v_1 & \cdots & v_r \end{bmatrix}$ and a nonsingular matrix $\Sigma_r := \operatorname{diag}\{\sigma_1, \ldots, \sigma_r\}$. The SVD of G can then be rewritten as

$$G = U_r\Sigma_rV_r'.$$

Since the rows of V' are independent and Σ_r is nonsingular, $\operatorname{Im} V_r' = \operatorname{Im}\Sigma_r = \mathbb{F}^r$. Hence, $\operatorname{Im} G = \operatorname{Im} U_r = \operatorname{span}(u_1, \ldots, u_r)$ and $\dim(\operatorname{Im} G) = r$. This proves the first two statements.

Now, since columns of U_r are independent and Σ_r is nonsingular, $\ker U_r = \ker U_r\Sigma_r = \{0\}$. Hence, $\ker G = \ker V_r' = (\operatorname{Im} V_r)^\perp$ (the latter follows by Proposition 2.1). Yet the orthogonality of V_r implies that $(\operatorname{Im} V_r)^\perp = \operatorname{span}(v_{r+1}, \ldots, v_m)$. This proves the last statement. □

[1] There are numerically efficient and stable algorithms of calculating the SVD.

It follows from the proof of Proposition 2.5 that any $p \times m$ matrix G having rank r can be factorized as

$$G = FH, \qquad \text{where } F \in \mathbb{F}^{p\times r} \text{ and } H \in \mathbb{F}^{r\times m} \text{ are some full rank matrices.} \tag{2.17}$$

We just may pick $F = U_r \Sigma_r$ and $H = V_r'$, although this choice is obviously not unique (it is unique up to multiplications by $r \times r$ nonsingular matrices, $G = FMM^{-1}H$). A factorization of form (2.17) is called the *full rank factorization* (or full rank decomposition) of A.

We conclude this section with another application of the singular value decomposition. Suppose we want to approximate some G it with a "simpler" matrix. One way to quantize the notion "simple" is via the rank of the approximation. Indeed, the rank, which is the dimension of the image, can be thought of as a measure of richness of the corresponding mapping. Moreover, low rank approximations might require less memory to be stored. For example, applying a low rank factorization, like that in (2.17), we end up with $(p + m)r$ elements, which is less that mp elements required to store G if $r < 0.5 \min\{p, m\}$. The following result presents a complete solution to the problems of approximating a matrix by a lower rank matrix for both spectral and Frobenius norms used to measure the the approximation performance.

Theorem 2.6. *Given $G \in \mathbb{F}^{p\times m}$ and its SVD in form* (2.13). *For every $1 \le l < \min\{p, m\}$ we have:*

- $\min_{\operatorname{rank}(H)\le l} \|G - H\| = \sigma_{l+1}$
- $\min_{\operatorname{rank}(H)\le l} \|G - H\|_F = \sqrt{\sum_{i=l+1}^{\min\{p,m\}} \sigma_i^2}$

with the minimizing $H = G_l := \sum_{i=1}^{l} \sigma_i u_i v_i'$ in both cases.

Proof. It follows from Proposition 2.5 that G_l is indeed a rank l matrix. Define

$$\tilde{G} := G - G_l = \sum_{i=l+1}^{\min\{p,m\}} \sigma_i u_i v_i',$$

where the last equality follows from (2.13). It is readily seen that the last expression above is the SVD of $\tilde{G}$ and therefore

$$\|\tilde{G}\| = \sigma_{l+1} \quad \text{and} \quad \|\tilde{G}\|_{\mathrm{F}}^2 = \sum_{i=l+1}^{\min\{p,m\}} \sigma_i^2$$

(the former follows from Proposition 2.4 and the latter follows from (2.16)). Thus, we only need to prove that $\|G - H_l\| \ge \|\tilde{G}\|$ and $\|G - H_l\|_{\mathrm{F}} \ge \|\tilde{G}\|_{\mathrm{F}}$ for all H_l such that $\operatorname{rank}(H_l) \le l$.

Spectral norm: Following the definitions in the proof of Proposition 2.5, introduce partial matrices U_{l+1}, V_{l+1}, and Σ_{l+1} and define

$$M_{l+1} := U_{l+1}'(G - H_l)V_{l+1} = \Sigma_{l+1} - U_{l+1}' H_l V_{l+1} =: \Sigma_{l+1} - \tilde{H}_l \in \mathbb{F}^{(l+1)\times(l+1)}.$$

Clearly, $\operatorname{rank}(\tilde{H}_l) \le l$ too and $\|M_{l+1}\| \le \|U_{l+1}\|\|G - H_l\|\|V_{l+1}\| = \|G - H_l\|$. Because $\tilde{H}_l$ has reduced rank, there always exists $0 \ne \eta \in \ker \tilde{H}_l \subset \mathbb{F}^{l+1}$. In this case we have:

$$\|M_{l+1}\eta\| = \|\Sigma_{l+1}\eta\| \ge \sigma_{l+1}\|\eta\| \qquad \text{(by Proposition 2.4).}$$

Thus, $\|G - H_l\| \ge \|M_{l+1}\| \ge \sigma_{l+1} = \|\tilde{G}\|$ for every H_l such that $\operatorname{rank}(H_l) \le l$.

Frobenius norm: Consider the matrix

$$M := U'(G - H_l)V = \Sigma - U' H_l V =: \Sigma - \tilde{H}_l,$$

where $\operatorname{rank}(\tilde{H}_l) \leq l$. Let $\{\tilde{v}_1, \ldots, \tilde{v}_{m-l}\}$ be a set of orthonormal linearly independent elements of $\ker \tilde{H}_l$ (if $\operatorname{rank}(\tilde{H}_l) = l$, then this is an orthonormal basis of $\ker \tilde{H}_l$). The matrix

$$\tilde{V} := \begin{bmatrix} \tilde{v}_1 & \cdots & \tilde{v}_{m-l} \end{bmatrix} \in \mathbb{F}^{m\times(m-l)}$$

satisfies $\tilde{V}'\tilde{V} = I_{m-l}$ (so that $\|\tilde{V}\|_{\mathrm{F}}^2 = m - l$ and also each its row, $\tilde{v}_{i\bullet}$, satisfies $\tilde{v}_{i\bullet}\tilde{v}_{i\bullet}' \leq 1$) and $\tilde{H}_l\tilde{V} = 0$ and we have:

$$\|G - H_l\|_{\mathrm{F}}^2 = \|M\|_{\mathrm{F}}^2 \geq \|M\tilde{V}\|_{\mathrm{F}}^2 = \|\Sigma\tilde{V}\|_{\mathrm{F}}^2 = \sum_{i=1}^{m} \sigma_i^2\alpha_i$$

(with some abuse of notation we assume that $\sigma_i = 0$ even if $\min\{p, m\} < i \leq m$), where $\alpha_i := \|\tilde{v}_{i\bullet}\|^2 \in [0, 1]$. Because $\|\tilde{V}\|_{\mathrm{F}}^2 = m - l$, we have that

$$\sum_{i=1}^{m} \alpha_i = m - l \quad \text{or, equivalently,} \quad \sum_{i=1}^{l} \alpha_i = \sum_{i=l+1}^{m} (1 - \alpha_i) =: \sum_{i=l+1}^{m} \beta_i$$

with $\beta_i \in [0, 1]$ too. Thus,

$$\begin{aligned}\|\Sigma\tilde{V}\|_{\mathrm{F}}^2 &= \sum_{i=1}^{l} \sigma_i^2\alpha_i + \sum_{i=l+1}^{m} \sigma_i^2\alpha_i \\ &\geq \sigma_l^2 \sum_{i=1}^{l} \alpha_i + \sum_{i=l+1}^{m} \sigma_i^2\alpha_i = \sum_{i=l+1}^{m} (\sigma_l^2 - \sigma_i^2)\beta_i + \sum_{i=l+1}^{m} \sigma_i^2 \geq \sum_{i=l+1}^{m} \sigma_i^2\end{aligned}$$

and we just showed that $\|G - H_l\|_{\mathrm{F}}^2 \geq \|\tilde{G}\|_{\mathrm{F}}^2$.

This completes the proof. □

2.4 Systems as a modeling tool

Hitherto, we saw that some properties of a static system (matrix) G can be characterized by two subspaces associated with it, $\ker G$ and $\operatorname{Im} G$. The purpose of this section is to exploit the reverse direction, namely, the use of systems to *generate* subspaces. We shall also see that static systems can be used to *shape* metrics in $\mathbb{F}^n$. These aspects of linear systems are extensively used in control applications to concretize the otherwise somewhat abstract notion of the "subspace" and to form metrics more suitable for our purposes than the standard (e.g. Hölder) norms.

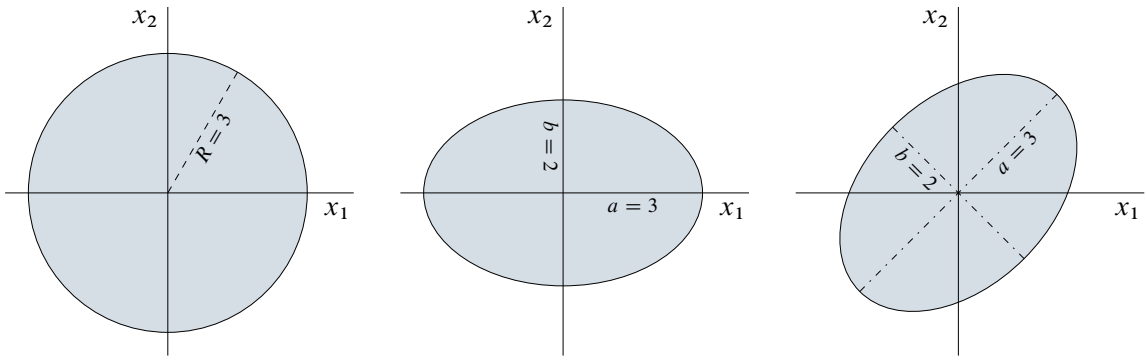
The following Proposition, which is technically quite trivial, is an important conceptual result.

Proposition 2.7. *A set $\mathcal{S} \subset \mathbb{F}^n$ is a subspace iff either of the following conditions holds:*

- *there are an integer $d \leq n$ and a full-rank matrix $S_{\mathrm{I}} \in \mathbb{F}^{n\times d}$ such that $\mathcal{S} = \operatorname{Im} S_{\mathrm{I}}$*
- *there are an integer $d \leq n$ and a full-rank matrix $S_{\mathrm{K}} \in \mathbb{F}^{(n-d)\times n}$ such that $\mathcal{S} = \ker S_{\mathrm{K}}$*

Moreover, this d is the dimension of $\mathcal{S}$.

Proof. As both image and kernel are subspaces, the "if" part in both cases is immediate, as well as the fact that $\dim(\mathcal{S}) = d$. To show the "only if" part, let $\mathcal{S}$ be a d-dimensional subspace and $\{s_1, \ldots, s_d\}$ be its basis. By the very definition of the span, $\mathcal{S} = \operatorname{Im} S_{\mathrm{I}}$ for $S_{\mathrm{I}} = \begin{bmatrix} s_1 & \ldots & s_d \end{bmatrix}$. Likewise, if $\{s_{d+1}, \ldots, s_n\}$ is a basis of the $(n - d)$-dimensional space $\mathcal{S}^\perp$, a required S_{K} is $S_{\mathrm{K}} = \begin{bmatrix} s_{d+1} & \ldots & s_n \end{bmatrix}'$. □



(a) $\|x\|_a = \sqrt{\left(\frac{x_1}{3}\right)^2 + \left(\frac{x_2}{3}\right)^2}$ (b) $\|x\|_b = \sqrt{\left(\frac{x_1}{3}\right)^2 + \left(\frac{x_2}{2}\right)^2}$ (c) $\|x\|_c = \sqrt{\left(\frac{x_1+x_2}{3\sqrt{2}}\right)^2 + \left(\frac{x_1-x_2}{2\sqrt{2}}\right)^2}$

Fig. 2.3: Unit balls for exotic norms in $\mathbb{R}^2$

Proposition 2.7, at least from the viewpoint of an engineer, renders subspaces a concrete notion. Once a matrix S_{I} (or S_{K}) is chosen, we no longer need to keep in mind constraints imposed by $\mathcal{S}$, the *model* of $\mathcal{S}$ (i.e. S_{I} or S_{K}) takes care of it. For example, consider the "freedom of choice" characterization of the system kernel discussed in §2.3.1. Remember, that if we know one particular input, say u_0, producing a desired output $y_{\mathrm{d}} = Gu_0$ for a given system $G : \mathbb{F}^m \to \mathbb{F}^p$, then all inputs producing y_{d} are given by $u_0 + u_{\mathrm{N}}$ for any $u_{\mathrm{N}} \in \ker G$. Let now $G_{\mathrm{N}} \in \mathbb{F}^{m \times \mathrm{rank}(G)}$ be a matrix such that $\ker G = \operatorname{Im} G_{\mathrm{N}}$ (it can be constructed from singular vectors of G using Proposition 2.5). In this case, u satisfies $y_{\mathrm{d}} = Gu$ iff $u = u_0 + G_{\mathrm{N}} v$, where $v \in \mathbb{F}^{\mathrm{rank}(G)}$ is arbitrary. The advantage of this characterization is that the free parameter, which is now v, is unconstrained. All constraints are accounted for by the model of $\ker G$, i.e. the matrix G_{N}.

Another use of matrices as a modeling tool is in shaping metrics. The standard Hölder metrics from §2.2.1 might not be sufficiently rich for all situations. For example, it might be appropriate to define the set of all "small" signals not as $\{x \mid \|x\|_q \leq 1\}$, but rather as $\{x \mid \|x\|_q \leq \gamma\}$ for some $\gamma > 0$. This is just a matter of scaling, so we in principle may redefine norms via an appropriate scaling of x. The norm $\|\cdot\|_a$ for $q = 2$ and $\gamma = 3$ and the corresponding unit ball are shown in Fig. 2.3(a). In other applications it might happen that we need to scale different components of x differently to reflect their relative importance. This requires introducing another norm, like $\|\cdot\|_b$, whose unit ball has an elliptic shape, see Fig. 2.3(b). Such a unit ball effectively declares that the direction of e_2 is more "important" that the direction of e_1. Yet another possibility is different scaling along to some other basis, for example $\|\cdot\|_c$, shown in Fig. 2.3(c), reflects scaling along with the orthonormal basis $\left\{\frac{1}{\sqrt{2}}\left[\begin{smallmatrix}1\\1\end{smallmatrix}\right], \frac{1}{\sqrt{2}}\left[\begin{smallmatrix}1\\-1\end{smallmatrix}\right]\right\}$ and results in a rotated elliptic disc.

The introduction of different metrics for different needs is an unwieldy solution. A more elegant way, which facilitates a unified treatment, is to generate signals with required metric by a signal with a standard and easily treatable metric. To this end, define the *weighted unit ball*

$$G\mathcal{B}_q := \left\{x \mid x = Gu, \|u\|_q \leq 1\right\}$$

for some matrix G. With this notation, it is readily seen that the unit ball in the "a" metric in Fig. 2.3(a) is merely $3\mathcal{B}_2$, the unit ball in the "b" metric in Fig. 2.3(b) is $\left[\begin{smallmatrix}3&0\\0&2\end{smallmatrix}\right]\mathcal{B}_2$, and the unit ball in the "c" metric in Fig. 2.3(c) is $\frac{1}{\sqrt{2}}\left[\begin{smallmatrix}3&-2\\3&2\end{smallmatrix}\right]\mathcal{B}_2$. Thus, in all three cases presented in Fig. 2.3 we may use the same $\mathcal{B}_2$ to generate the regions of "small" signals.

Remark 2.5. It is readily verified that

$$\mathcal{B}_\infty = \sqrt{2} R_{\pi/4} \mathcal{B}_1,$$

where $R_{\pi/4}$ is the plain rotation matrix defined in (2.12) for the angle $\theta = \pi/4$. This suggests that the $\mathbb{F}^n$ norms $\|\cdot\|_1$ and $\|\cdot\|_\infty$ are, in a sense, transposable. ∇

Chapter 3

Dynamic Systems

Dynamic systems are systems that are not static. This means that relations between external signals must involve some memory effects. In other words, outputs of a dynamical system at a time instance t_c may depend not only on system inputs at t_c, but also on the inputs at other time instances, preceding or following t_c. This implies that the time evolution or the frequency dependence play crucial roles in the analysis of dynamical systems, which can no longer be analyzed in frozen time.

In this chapter some basic facts on continuous-time signals and analog dynamic systems are collected. Both time- and transformed-domain perspectives are presented. Although the notes are mainly concerned with finite-dimensional systems, wider perspectives are given in some cases.

3.1 Continuous-time signals

Time-dependent signals were introduced in Section 1.1 on a conceptual level. In this section we present them from a more formal perspective. Mathematically, a signal is a function assigning to each element from its *domain* one element from its *codomain*. Throughout these notes domains are normally assumed to be the whole real axis $\mathbb{R}$, interpreted as continuous time, and codomains are the n-dimensional real space $\mathbb{R}^n$ for some $n \in \mathbb{N}$. As such, a signal x is viewed as a mapping between $\mathbb{R}$ and $\mathbb{R}^n$, denoted $x : \mathbb{R} \to \mathbb{R}^n$. This choice reflects analog signals frequently encountered in control applications. Still, different domains and / or codomains may be considered, often under only minor alterations. The value of x at a given time instance $t \in \mathbb{R}$ is denoted as $x(t)$. This $x(t) \in \mathbb{R}^n$, i.e. is a finite-dimensional vector like those studied in Chapter 2, and thus shall not be confused with the signal x itself. We say that a signal is scalar-valued if $n = 1$ and vector-valued otherwise. Similarly to frozen-time signals, time-dependent signals are convenient to visualize in terms of coordinates in the standard basis of $\mathbb{R}^n$. With this logic, x_i stands for its ith component in the standard basis, which is a scalar-valued signal.

The set of signals $\mathbb{R} \to \mathbb{R}^n$ constitute a vector space, with obvious, frozen-time, definitions of the addition and multiplication by scalar, operations. Namely, $x = y + z$ implies that $x(t) = y(t) + z(t)$ for all t and, given $\alpha \in \mathbb{R}$, $x = \alpha y$ implies that $x(t) = \alpha y(t)$, also for all t. Another operation on continuous-time signals that we need is the *time shift* $\mathcal{S}_\tau$, acting on a signal x as

$$(\mathcal{S}_\tau x)(t) = x(t + \tau), \quad \forall t \in \mathbb{R} \tag{3.1}$$

for a given $\tau \in \mathbb{R}$. This operation is important in the analysis of dynamic systems.

3.1.1 Normed time-domain signal spaces

Admissible signals in each specific situation are convenient to formalize via belonging them to normed spaces, where the notion of the size of a signal is quantified. This is of primary importance in quantifying

specifications to control systems, like what is meant by the requirements that errors are "small," control signals are "affordable," etc. Commonly used normed spaces on signals $\mathbb{R} \to \mathbb{R}^n$ are the Lebesgue spaces

$$L_q^n(\mathbb{R}) := \left\{ x : \mathbb{R} \to \mathbb{R}^n \;\middle|\; \|f\|_q := \left(\int_{\mathbb{R}} \|x(t)\|_q^q \mathrm{d}t \right)^{1/q} < \infty \right\} \tag{3.2}$$

under $q \geq 1$, where $\|x(t)\|_q$ is the Hölder vector norm on $\mathbb{R}^n$ introduced in (2.3). These spaces are Banach, i.e. complete. We often drop the signal domain and use the notation L_q^n, or even L_q when the dimension of the signal domain is irrelevant of clear from the context. The domain of the signal is specified for spaces when it is different from $\mathbb{R}$. The quantity $\|x\|_q$ defined in (3.2) is a norm and referred to as the L_q-norm of a signal x. Frequently used special cases are those for $q = 1, 2, \infty$. The latter corresponds to

$$L_\infty^n(\mathbb{R}) := \left\{ x : \mathbb{R} \to \mathbb{R}^n \mid \|x\|_\infty := \sup_{t \in \mathbb{R}} \|f(t)\|_\infty < \infty \right\} \tag{3.3}$$

and comprises all uniformly bounded signals. We thus say that x is *bounded* if $x \in L_\infty$.

Spaces L_q are infinite dimensional, i.e. there is no finite basis on them. Therefore, L_q-norms are not equivalent, in the sense of definition (2.5). This implies that a signal belonging to one space does not necessarily belong to another. For example, the step function $\mathbb{1}$ is clearly bounded, so $\mathbb{1} \in L_\infty$. Yet it is neither absolutely nor square integrable on $\mathbb{R}$, meaning that $\mathbb{1} \notin L_1$ and $\mathbb{1} \notin L_2$. Another example is the sine cardinal, defined as $\mathrm{sinc}(t) := \sin(t)/t$, with $\mathrm{sinc}(0) = 1$. It is bounded, $\|\mathrm{sinc}\|_\infty = 1$, so belongs to L_∞, and $\|\mathrm{sinc}\|_2 = \sqrt{\pi}$, so belongs to L_2 as well. However,

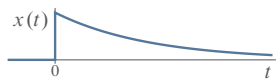
$$\begin{aligned} \int_{\mathbb{R}} |\mathrm{sinc}(t)| \mathrm{d}t &= \sum_{i \in \mathbb{Z}} \int_{(i-1)\pi}^{i\pi} \left| \frac{\sin(t)}{t} \right| \mathrm{d}t = 2 \sum_{i \in \mathbb{N}} \int_{(i-1)\pi}^{i\pi} \frac{|\sin(t)|}{t} \mathrm{d}t \\ &> 2 \sum_{i \in \mathbb{N}} \int_{(i-1)\pi}^{i\pi} \frac{|\sin(t)|}{i\pi} \mathrm{d}t = 2 \sum_{i \in \mathbb{N}} \frac{1}{i\pi} \int_{(i-1)\pi}^{i\pi} |\sin(t)| \mathrm{d}t = \frac{4}{\pi} \sum_{i \in \mathbb{N}} \frac{1}{i} = \infty, \end{aligned} \tag{3.4}$$

so that $\mathrm{sinc} \notin L_1$.

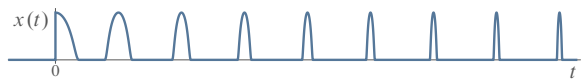
The space

$$L_2^n(\mathbb{R}) := \left\{ x : \mathbb{R} \to \mathbb{R}^n \;\middle|\; \|x\|_2 := \left(\int_{\mathbb{R}} \|x(t)\|^2 \mathrm{d}t \right)^{1/2} < \infty \right\}, \tag{3.5}$$

or simply L_2, is of special importance. It comprises finite-energy signals (understood as $\|x\|_2^2$) and its prevalence is perhaps mainly motivated by favorable mathematical properties. Belonging to L_2 requires either a sufficiently fast (but not as fast as belonging to L_1, we saw that in the sinc example above) decay of the signal at $t \to \pm\infty$ or sufficiently small support of its non-decaying parts, e.g.



or



.

In fact, L_2-signals need not be even bounded. For example, the unbounded x such that $x(t) = \mathrm{e}^{-|t|}/\sqrt[4]{|t|}$ has $\|x\|_2 = \sqrt[4]{2\pi}$ and therefore belongs to L_2. Two more spaces that we shall need later on are

$$L_{2+}^n := \left\{ x \in L_2^n(\mathbb{R}) \mid x(t) = 0 \text{ if } t < 0 \right\} \quad \text{and} \quad L_{2-}^n := \left\{ x \in L_2^n(\mathbb{R}) \mid x(t) = 0 \text{ if } t > 0 \right\},$$

which are subspaces of L_2. The space L_2 is a Hilbert space, i.e. in addition to sizes we can talk about *angles* between signals via the notion of the *inner product*,

$$\langle x, y \rangle_2 := \int_{\mathbb{R}} [y(t)]' x(t) \mathrm{d}t,$$

which can be thought of as the cosine of the angle between x and y, scaled by $\|x\|_2 \|y\|_2$ (cf. (A.2)). The inner product is particularly important in relation to the notion of *orthogonality*, which, in turn, plays a key role in various optimization procedures. We say that two signals x and y are orthogonal, denoted $x \perp y$, if $\langle x, y \rangle_2 = 0$. For example, every $x \in L_{2+}$ and $y \in L_{2-}$ are orthogonal, just because their supports are disjoint. Hence the relation $L_2 = L_{2+} \oplus L_{2-}$, saying that all $x \in L_2$ can be decomposed as $x = x_+ + x_-$ into two orthogonal signals $x_+ \in L_{2+}$ and $x_- \in L_{2-}$.

3.1.2 Laplace and Fourier transforms

Signals are naturally perceived as time-domain phenomena, this is how we are accustomed to sense them in many situations. However, it may be useful to look at signals from other viewpoints as well, in various transformed domains. In the control literature the Laplace- and Fourier-domain analyses are prevalent in the analysis of signals supported on infinite (or semi-infinite) time intervals.

The two-sided (bilateral) Laplace transform $X = \mathfrak{L}\{x\}$ of a continuous-time signal $x : \mathbb{R} \to \mathbb{R}^n$ is the signal $X : \mathrm{RoC} \subset \mathbb{C} \to \mathbb{C}^n$ such that

$$X(s) = \int_{\mathbb{R}} x(t)\mathrm{e}^{-st}\,\mathrm{d}t. \tag{3.6}$$

The *region of convergence* (RoC) of the Laplace transform is a subset of the complex plane at which the integral in (3.6) is convergent. To guarantee the absolute convergence for a given $s \in \mathbb{C}$, the signal $\exp_{-\operatorname{Re} s} x$ must belong to L_1. If $\operatorname{Re} s > 0$ ($\operatorname{Re} s < 0$), this effectively requires that x does not grow "too fast" as $t \to +\infty$ ($t \to -\infty$) and decays "sufficiently fast" as $t \to -\infty$ ($t \to +\infty$). These conditions are rather draconian for signals supported in the whole $\mathbb{R}$. However, signals of interest frequently have their support only on semi-axes, $\mathbb{R}_+$ or $\mathbb{R}_-$. For signals with support in $\mathbb{R}_+$, the RoC is typically the right half-plane $\mathbb{C}_\alpha := \{s \in \mathbb{C} \mid \operatorname{Re} s > \alpha\}$ for some $\alpha \in \mathbb{R} \cup \{\pm\infty\}$. For example, $(\mathfrak{L}\{\mathbb{1}\})(s) = 1/s$ and its RoC is $\mathbb{C}_0$. The case of $\alpha = -\infty$ corresponds to $\mathrm{RoC} = \mathbb{C}$, which happens, for example, when x is a bounded function with a finite support. The case of $\alpha = +\infty$ corresponds to $\mathrm{RoC} = \varnothing$, which happens, for example, for the signal x such that $x(t) = \mathrm{e}^{t^2}\mathbb{1}(t)$. RoCs for signals supported in the negative semi-axis are left half-planes. Distributions (generalized functions) can also be transformed, for instance the Laplace transform of the Dirac delta $(\mathfrak{L}\{\delta\})(s) = 1$, with $\mathrm{RoC} = \mathbb{C}$.

The Fourier transform $X = \mathfrak{F}\{x\}$ of a continuous-time signal $x : \mathbb{R} \to \mathbb{R}^n$ is the signal $X : \mathrm{j}\mathbb{R} \to \mathbb{C}^n$ such that

$$X(\mathrm{j}\omega) = \int_{\mathbb{R}} x(t)\mathrm{e}^{-\mathrm{j}\omega t}\,\mathrm{d}t, \tag{3.7}$$

where $\omega \in \mathbb{R}$ is called the (angular) frequency and measured in radians per time unit (e.g. per second). We may be inclined to see this transform as a special case of the Laplace transform, in which s is only allowed to be on the imaginary axis $\mathrm{j}\mathbb{R}$. This is indeed the case if $\mathrm{j}\mathbb{R} \subset \mathrm{RoC}$, but might be more delicate otherwise. The Fourier transform is well defined for signals from L_1 (plus some mild technical assumptions, effectively nonrestrictive "in the wild"). In that case the inverse Fourier transform,

$$x(t) = \frac{1}{2\pi}\int_{\mathbb{R}} X(\mathrm{j}\omega)\mathrm{e}^{\mathrm{j}\omega t}\,\mathrm{d}\omega, \tag{3.8}$$

yields the original x. By the Plancherel theorem, the transform may be extended to functions from L_2, with a weaker convergence, where the right-hand side of (3.8) converges to x only in the L_2-norm. A wider class of functions can be treated by allowing the distribution formalism and taking some liberties with convergence. For example, $(\mathfrak{F}\{\delta\})(\mathrm{j}\omega) = 1$ and $(\mathfrak{F}\{\mathbb{1}\})(\mathrm{j}\omega) = 1/(\mathrm{j}\omega) + \pi\delta(\omega)$. Note that the latter is not $\mathfrak{L}\{\mathbb{1}\}$ at $s = \mathrm{j}\omega$, because the RoC of the Laplace transform does not include the imaginary axis.

The Fourier transform is defined on a substantially narrower class of signals than the Laplace transform. However, unlike the latter, the former is more tangible. Indeed, the inverse Fourier transform (3.8) shows that x is a superposition of harmonic signals $\exp_{\mathrm{j}\omega}$ with frequencies ω and the value of the Fourier transform at a frequency ω, $X(\mathrm{j}\omega)$, is the weight of the harmonic $\exp_{\mathrm{j}\omega}$ in x. This is why $\mathfrak{F}\{x\}$ is called the *frequency-domain* representation of x or its *spectrum*. The spectrum thus offers a valuable insight into properties of signals. If the spectrum is dominated by low frequencies, we may expect the signal to vary slowly in the time domain. Fast signals may be expected to have their spectra concentrated in high frequencies. This insight is vital for classical control methods, as well as in many other fields, to separate the treatment of signals that act simultaneously but have different spectral properties, see §1.4.3 for an example.

A remarkable, and useful, property of the Fourier and Laplace transforms is that they preserve (appropriately defined) angles and sizes. Namely, let $L_2(\mathrm{j}\mathbb{R})$ denote the space of square integrable functions $\mathrm{j}\mathbb{R} \to \mathbb{C}^n$, similarly to (3.5). The Fourier transform is a unitary mapping $L_2(\mathbb{R}) \to L_2(\mathrm{j}\mathbb{R})$ such that

$$\langle x, y\rangle_2 = \langle X, Y\rangle_2 := \frac{1}{2\pi}\int_{\mathbb{R}} [Y(\mathrm{j}\omega)]' X(\mathrm{j}\omega)\,\mathrm{d}\omega \tag{3.9}$$

and, consequently, $\|x\|_2 = \|X\|_2 := \sqrt{\langle X, X\rangle_2}$ (the scaling is introduced to render it unitary). This result is known as the Parseval's theorem in the engineering literature. Likewise, by the Paley–Wiener theorem, the Laplace transform is a unitary mapping $L_{2+} \to H_2$ preserving the inner product, where

$$H_2^n := \left\{ X : \mathbb{C}_0 \to \mathbb{C}^n \;\middle|\; X(s) \text{ is holomorphic in } \mathbb{C}_0 \text{ and } \|X\|_2 := \left(\sup_{\sigma>0} \frac{1}{2\pi}\int_{\mathbb{R}} \|X(\sigma + \mathrm{j}\omega)\|^2 \mathrm{d}\omega\right)^{1/2} < \infty \right\}. \tag{3.10}$$

The Hardy space H_2^n, or simply H_2, defined by (3.10) plays an important role in frequency-domain analyses, so some clarifications are in order. Although the imaginary axis is not in the domain of functions in it, the boundary function $\tilde{X}$ such that $\tilde{X}(\mathrm{j}\omega) := \lim_{\sigma\downarrow 0} X(\sigma + \mathrm{j}\omega)$ exists for almost all $\omega \in \mathbb{R}$ and is such that $\|\tilde{X}\|_2 = \|X\|_2$ (hence, $\tilde{X} \in L_2(\mathrm{j}\mathbb{R})$). It is then customary to identify functions from H_2 with their boundary functions in $L_2(\mathrm{j}\mathbb{R})$ and regard H_2 as a closed subspace of $L_2(\mathrm{j}\mathbb{R})$ that inherits the inner product defined in (3.9) and, consequently, the $L_2(\mathrm{j}\mathbb{R})$ norm. Thus, by the H_2-norm of $X \in H_2^n$ we shall understand

$$\|X\|_2 = \left(\frac{1}{2\pi}\int_{\mathbb{R}} \|X(\mathrm{j}\omega)\|^2 \mathrm{d}\omega\right)^{1/2}.$$

The orthogonal complement of H_2 in $L_2(\mathrm{j}\mathbb{R})$, denoted by $H_2^\perp$, consists then of functions holomorphic in $\mathbb{C} \setminus \overline{\mathbb{C}}_0$ and square integrable over all vertical lines in it. The Laplace transform is then a unitary mapping $L_{2-} \to H_2^\perp$. Moreover, we have that $L_2(\mathrm{j}\mathbb{R}) = H_2 \oplus H_2^\perp$, which is the transformed domain counterpart of the time-domain relation $L_2(\mathbb{R}) = L_{2+} \oplus L_{2-}$.

3.2 Linear systems in time domains

Like signals, the notions of systems and their mathematical models were already introduced in Section 1.1 on a conceptual level. The purpose of this and next sections is to elaborate on more analytical notions related to I/O systems, like the impulse response, transfer functions, and so on. We consider only systems on L_2, which somewhat dominate the literature because of their Hilbert space relations, although systems on general L_q spaces, as well as on an interval $\mathbb{I} \subset \mathbb{R}$, can be considered as well. The exposition is attempted to be not overly technical, concentrating rather on underlying ideas. For that reason, the issues of convergence, continuity, and the like are left beyond the scope.

From the mathematical point of view, linear continuous-time systems with m-dimensional inputs and p-dimensional outputs are understood as linear operators $G : \mathfrak{D}_G \subset L_2^m \to L_2^p$ for some domain $\mathfrak{D}_G$. The linearity means that the superposition property holds, see (A.4) on p. 186. A general class of linear systems $G : u \mapsto y$ may be described by their *kernel representation*

$$y(t) = \int_{\mathbb{R}} g(t, s)u(s)\,\mathrm{d}s, \tag{3.11}$$

where the *impulse response* (or kernel) $g : \mathbb{R}^2 \to \mathbb{R}^{p\times m}$ of G. The jth column of $g(t, s)$ is the response at the time instance t of G to the input signal $e_j \mathcal{S}_{-s}\delta$, i.e. the Dirac delta applied at the time instance s in the direction of $e_j \in \mathbb{R}^m$. To explain relation (3.11), assume for simplicity that $m = 1$ and note that

$$u = \int_{\mathbb{R}} (\mathcal{S}_{-s}\delta)u(s)\,\mathrm{d}s \quad \Longrightarrow \quad u(t) = \int_{\mathbb{R}} \delta(t - s)u(s)\,\mathrm{d}s$$

for all continuous signals u. Hence, the relation $y = Gu$ reads

$$y(t) = \Big(G\Big(\int_{\mathbb{R}} (\mathcal{S}_{-s}\delta)u(s)\,\mathrm{d}s\Big)\Big)(t) = \int_{\mathbb{R}} \Big(G\big((\mathcal{S}_{-s}\delta)u(s)\big)\Big)(t)\,\mathrm{d}s = \int_{\mathbb{R}} \big(G(\mathcal{S}_{-s}\delta)\big)(t)u(s)\,\mathrm{d}s, \tag{3.12}$$

which yields (3.11) because $g := G(\mathcal{S}_{-s}\delta)$. The last two equalities in (3.12) use the linearity property of G, viz. its additivity and homogeneity parts, respectively. If $m > 1$, then the same arguments apply to the decomposition $u = \sum_{i=1}^{m} e_i u_i$. It is worth emphasized that (3.12) is not a formal proof, for it implicitly considers only continuous inputs and assumes that all involved integrals converge. Yet the level of technicalities required to derive (3.11) rigorously goes beyond the scope of these notes. Zealots of mathematical rigor are referred to [34] for details.

The impulse response in (3.11) is not necessarily a function. It may involve Dirac deltas even for simple systems. For example, if G is the unit gain, i.e. such that $Gu = u$, then $g(t,s) = \delta(t-s)$. A sufficiently general class of impulse responses is

$$g(t,s) = \tilde{g}(t,s) + \sum_{i\in\mathbb{Z}} g_i(t)\delta(t-s-\kappa_i(t)) \tag{3.13}$$

for a locally bounded (i.e. bounded on any bounded subset of $\mathbb{R}^2$) function $\tilde{g} : \mathbb{R}^2 \to \mathbb{R}^{p\times m}$, locally bounded $g_i : \mathbb{R} \to \mathbb{R}^{p\times m}$, and the increasing, at each $t \in \mathbb{R}$, sequence $\{\kappa_i(t)\}_{i\in\mathbb{Z}}$, such that $\kappa_0(t) \equiv 0$ and there is a constant $\epsilon > 0$, independent of i and t, such that $\kappa_i(t) - \kappa_{i-1}(t) \geq \epsilon$ for all $i \in \mathbb{Z}$.

Several examples of continuous-time systems $u \mapsto y$ on L_2 are presented below.

- The *integrator* G_{int} acts as $\dot{y}(t) = u(t)$ assuming $\lim_{t\to-\infty} y(t) = 0$. The term integrator stems from the relation $y(t) = \int_{-\infty}^{t} u(s)\,\mathrm{d}s$. Its impulse response g_{int} has $g_{\text{int}}(t,s) = \mathbb{1}(t-s)$.
- Another system accumulating the past without forgetting is $G_{\text{dint},\mu}$, acting as $y(t) = y(t-\mu) + u(t)$ for some $\mu > 0$. Its impulse response $g_{\text{dint},\mu}$ has $g_{\text{dint},\mu}(t,s) = \sum_{i\in\mathbb{N}} \delta(t-s-(i-1)\mu)$.
- The *finite-memory integrator* $G_{\text{fmint},\mu}$ acts according to $y(t) = \int_{t-\mu}^{t} u(s)\,\mathrm{d}s$ for some $\mu > 0$ and has the impulse response $g_{\text{fmint},\mu}$ such that $g_{\text{fmint},\mu}(t,s) = \mathbb{1}(t-s) - \mathbb{1}(t-s-\mu)$.
- The *τ-delay operator* D_τ is defined as $y(t) = u(t-\tau)$ for $\tau > 0$. In other words, this is exactly the reciprocal shift operator, $D_\tau = \mathcal{S}_{-\tau}$. Its impulse response d_τ has $d_\tau(t,s) = \delta(t-s-\tau)$.
- The *ideal low-pass filter* F_{ILP} with bandwidth ω_{b} has $f_{\text{ILP}}(t,s) = (\omega_{\text{b}}/\pi)\,\mathrm{sinc}(\omega_{\text{b}}(t-s))$.

A system G is said to be *stable* if $\mathfrak{D}_G = L_2^m$ and $\|G\| := \sup_{\|u\|_2=1} \|Gu\|_2 < \infty$. Otherwise, G is referred to as *unstable*. The quantity $\|G\|$ defined above is called the *L_2-induced norm* of G. Both G_{int} and $G_{\text{dint},\mu}$ are unstable. To see this, apply $u = \mathbb{1}_{[0,T]} \in L_2$, which results in $(G_{\text{int}}u)(t) = t\,\mathbb{1}_{[0,T]}(t) + T\,\mathbb{1}(t-T)$ and $G_{\text{dint},\mu}u = \mathbb{1}$, neither of which belongs to L_2. Thus, not every L_2 signal is in the domains of G_{int} and $G_{\text{dint},\mu}$ (we shall discuss domains of such systems later on). The finite-memory integrator $G_{\text{fmint},\mu}$ is stable as an operator $L_2 \to L_2$. Indeed, in this case

$$\begin{aligned}\|y\|_2^2 &= \int_{\mathbb{R}} \Big\| \int_{t-\mu}^{t} u(s)\,\mathrm{d}s \Big\|^2 \mathrm{d}t \\ &\leq \mu \int_{\mathbb{R}} \int_{t-\mu}^{t} \|u(s)\|^2 \mathrm{d}s\,\mathrm{d}t = \mu \int_{\mathbb{R}} \int_{0}^{\mu} \|u(t-s)\|^2 \mathrm{d}s\,\mathrm{d}t = \mu \int_{0}^{\mu} \int_{\mathbb{R}} \|u(t)\|^2 \mathrm{d}t\,\mathrm{d}s = \mu^2 \|u\|_2^2,\end{aligned}$$

where the inequality follows by the Cauchy–Schwarz inequality (A.1) on p. 182 (it becomes the equality if u is piecewise-constant and does not change its sign) and the second equality in the bottom line follows by Tonelli's theorem. Therefore, $y \in L_2$ for all $u \in L_2$ and $\|G_{\text{fmint},\mu}\| = \mu < \infty$. The τ-delay operator D_τ is also stable, which follows from the fact that $\|D_\tau u\|_2 = \|u\|_2$ for every $u \in L_2$. It is not obvious, but nevertheless true, that F_{ILP} is stable as an operator on L_2.

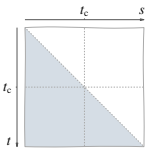
A system $G : \mathfrak{D}_G \subset L_2^m \to L_2^p$ is called *left invertible* if there is $G^+ : \mathfrak{D}_{G^+} \subset L_2^p \to L_2^m$ such that $G^+Gu = u$ for all $u \in \mathfrak{D}_G$ such that $Gu \in \mathfrak{D}_{G^+}$. A system is said to be *right invertible* if there is $G^+ : \mathfrak{D}_{G^+} \subset L_2^p \to L_2^m$ such that $GG^+u = u$ for all $u \in \mathfrak{D}_{G^+}$ such that $G^+u \in \mathfrak{D}_G$. Left / right inverses are not unique in general. If there is a *stable* right (left) inverse G^+, the system is said to be *stably right (left) invertible*. If $p = m$, a system is left invertible iff it is right invertible and left and right inverses are unique and coincide, they are denoted G^{-1} then. A stable square system G such that G^{-1} is also stable is dubbed *bi-stable*.

Another important property of I/O systems is causality. Loosely speaking, a system is called causal if its output at every time instance t_c can only depend on the past and present inputs, up to and including the very same time instance t_c. Formally, we say that G is *causal* if for every $t_c \in \mathbb{R}$ we have that $y(t) = 0$ for all $t \le t_c$ whenever $u(t) = 0$ for all $t \le t_c$. Such a requirement can be expressed as

$$\int_{\mathbb{R}} g(t,s)u(s)\,\mathrm{d}s = \int_{t_c}^{\infty} g(t,s)u(s)\,\mathrm{d}s = 0, \quad \forall t < t_c$$

and all admissible u. This condition, in turn, reads

$$g(t,s) = 0 \text{ whenever } s > t \qquad \text{i.e. } g(t,s) : \tag{3.14}$$



and requires that $g_i = 0$ for all $i < 0$ in (3.13). All examples considered above except F_{ILP} are causal. The delay operator D_τ becomes non-causal if redefined for $\tau < 0$. In fact, D_τ is *anti-causal* in this case, meaning that its output at each time instance can only depend on the future inputs. The ideal low-pass filter F_{ILP} is non-causal.

The *adjoint* $G' : L_2^p \to L_2^m$ of a stable $G : L_2^m \to L_2^p$ is defined the standard way, via the inner product relation $\langle Gu, y\rangle_2 = \langle u, G'y\rangle_2$. We then have:

$$\begin{aligned}\int_{\mathbb{R}} [y(t)]'(Gu)(t)\,\mathrm{d}t &= \int_{\mathbb{R}} [y(t)]' \int_{\mathbb{R}} g(t,s)u(s)\,\mathrm{d}s\,\mathrm{d}t = \int_{\mathbb{R}}\int_{\mathbb{R}} \big[[g(t,s)]'y(t)\big]'u(s)\,\mathrm{d}s\,\mathrm{d}t \\ &= \int_{\mathbb{R}}\int_{\mathbb{R}} \big[[g(s,t)]'y(s)\big]'\,\mathrm{d}s\,u(t)\,\mathrm{d}t.\end{aligned}$$

Thus, the adjoint of G is the system, whose impulse response at (t,s) equals $[g(s,t)]'$, i.e. it takes both transposing g and interchanging its arguments. This result can be extended to unstable systems as well. It follows from (3.14) that the adjoint of a causal system is anti-causal.

A linear system G is called *time invariant* (abbreviated LTI) if

$$G\mathcal{S}_\tau = \mathcal{S}_\tau G, \quad \forall \tau \in \mathbb{R}. \tag{3.15}$$

This definition effectively says that a delayed input produces a delayed, but otherwise unchanged, output. Otherwise, G is said to be *time varying*. If (3.15) holds for $\tau = T > 0$ but not for $\tau \in (0,T)$, then G is dubbed *T-periodic*. If G is LTI, then (3.12) can be continued as

$$y(t) = \int_{\mathbb{R}} \big(G(\mathcal{S}_{-s}\delta)\big)(t)u(s)\,\mathrm{d}s = \int_{\mathbb{R}} \big(\mathcal{S}_{-s}(G\delta)\big)(t)u(s)\,\mathrm{d}s = \int_{\mathbb{R}} (G\delta)(t-s)u(s)\,\mathrm{d}s,$$

meaning that only the response of G to the Dirac delta applied at $t = 0$ matters. The same conclusion can be drawn from (3.11), in terms of which the time invariance property reads

$$\int_{\mathbb{R}} g(t,s)u(s-\tau)\,\mathrm{d}s = (G\mathcal{S}_\tau u)(t) = (\mathcal{S}_\tau Gu)(t) = \int_{\mathbb{R}} g(t-\tau,s)u(s)\,\mathrm{d}s = \int_{\mathbb{R}} g(t-\tau,s-\tau)u(s-\tau)\,\mathrm{d}s$$

and implies that a system is LTI iff $g(t,s) = g(t-\tau, s-\tau)$ for all $t, s, \tau \in \mathbb{R}$, so that $g(t,s) = g(t-s, 0)$. We thus do not need the impulse response of an LTI G to be a function of two independent variables and treat it as $g : \mathbb{R} \to \mathbb{R}^{p\times m}$ with $g(t) = g(t,0)$. In that case (3.11) can be rewritten as

$$y(t) = \int_{\mathbb{R}} g(t-s)u(s)\,\mathrm{d}s, \tag{3.16}$$

which is known as the *convolution integral*, denoted $y = g * u$. An LTI system G is causal iff $g(t) = 0$ whenever $t < 0$, in which case the upper integration limit in (3.16) can be taken as t.

Remark 3.1. The equality $g(t,s) = g(t-\tau, s-\tau)$ is sometimes used as the definition of time invariance. The notion of time invariance is then readily extendible to systems defined on finite time intervals, where the shift operator struggles and LTI systems defined that way do not satisfy the relation $G\mathcal{S}_\tau = \mathcal{S}_\tau G$. ∇

Remark 3.2 (L_∞ stability). It may be hard to evaluate the L_2-stability of an LTI G in terms of properties of its impulse response g. The Laplace-domain representation of G are more useful toward that end, as will be discussed in the next section. Curiously, the L_∞-stability, defined similarly, but for the L_∞ signal norms, has a direct connection with g. Namely, a system is L_∞-stable iff $g \in L_1^{p\times m}$, where the matrix version of L_1 is defined exactly as (3.2) modulo replacing the vector 1-norm with the matrix one from (2.6a). Here we assume, sloppily, that $\delta \in L_1$. The quantity $\|G\|_{\mathcal{A}} := \|g\|_1$, known as the $\mathcal{A}$-norm of G, is then the L_∞-induced norm of G. It is readily seen that G_{int} and $G_{\text{dint},\mu}$ are L_∞-unstable, while $G_{\text{fmint},\mu}$ and D_τ are L_∞-stable, exactly like in the L_2 case. Remarkably, the ideal low-pass filter F_{ILP} is L_∞-unstable, cf. (3.4), even though it is L_2-stable. This is actually a general result that L_∞ stability implies L_2 stability, but not vice versa. Hence, the L_2 *in*stability of LTI systems can be verified via their impulse responses. ∇

3.3 LTI systems in transformed domains

The convolution representation of linear time-invariant systems is particularly appealing from the Laplace and Fourier transforms viewpoints. A key is the property of these transforms to turn convolution integrals into plain products, namely,

$$y = g * u \iff \mathfrak{L}\{y\} = \mathfrak{L}\{g\}\mathfrak{L}\{u\} \iff \mathfrak{F}\{y\} = \mathfrak{F}\{g\}\mathfrak{F}\{u\} \tag{3.17}$$

whenever involved signals have overlapping RoCs of their Laplace transforms or are Fourier transformable. The relations above are advantageous, for multiplication operators are easier to analyze than integral ones. This section introduces the main characteristics of LTI systems in transformed domains.

3.3.1 Frequency response

To start with, consider an LTI system G whose impulse response $g \in L_1^{p\times m}$, i.e. that the matrix 1-norm of $g(t)$ is integrable. This necessarily implies that $g_{\bullet j} \in L_1^p$ for every $j \in \mathbb{Z}_{1..m}$, so that g has a well-defined Fourier transform. In this case every Fourier transformable input u results in a Fourier transformable $y = Gu$ such that

$$Y(\mathrm{j}\omega) = G(\mathrm{j}\omega)U(\mathrm{j}\omega) \tag{3.18}$$

for all ω, which is the last equality of (3.17).

The Fourier transform $G(\mathrm{j}\omega)$ of the impulse response of G is called its *frequency response*. It can be interpreted in terms of the response of the system to harmonic inputs, like $u = u_\omega \exp_{\mathrm{j}\omega}$, in which case

$$y(t) = \int_{\mathbb{R}} g(t-\tau)u_\omega \mathrm{e}^{\mathrm{j}\omega\tau}\mathrm{d}\tau = \int_{\mathbb{R}} g(\tau)u_\omega \mathrm{e}^{\mathrm{j}\omega(t-\tau)}\mathrm{d}\tau = \int_{\mathbb{R}} g(\tau)\mathrm{e}^{-\mathrm{j}\omega\tau}\mathrm{d}\tau\, u_\omega \mathrm{e}^{\mathrm{j}\omega t} = [G(\mathrm{j}\omega)u_\omega]\mathrm{e}^{\mathrm{j}\omega t}.$$

Thus, the response to a harmonic input is also a harmonic signal with the same frequency, whose amplitude, phase, and direction are shaped by $G(\mathrm{j}\omega)$. This result may be used to identify frequency responses of stable systems from the response of stable LTI systems to harmonic excitation. The frequency response is also useful in analyzing performance of LTI systems. It follows from (3.18) that $G(\mathrm{j}\omega)$ determines the way in which a system alters the spectrum of its input. We may then aim at shaping the frequency response of systems of interest to attenuate or amplify required frequencies in their outputs. For example, the ideal low-pass filter introduced in the previous section has the frequency response $F_{\mathrm{ILP}}(\mathrm{j}\omega) = \mathbb{1}(\omega+\omega_{\mathrm{b}}) - \mathbb{1}(\omega-\omega_{\mathrm{b}})$, which is the unit height rectangular pulse. This implies that any harmonic signal $\exp_{\mathrm{j}\omega}$ passes it unaltered if $\omega \in [-\omega_{\mathrm{b}}, \omega_{\mathrm{b}}]$ and does not pass at all otherwise (hence, the term).

Although relation (3.18) is purely algebraic, one should be careful in manipulating systems via their frequency responses. Apparent hazards lie in difficulties to handle exponentially growing signals (common in unstable phenomena) and to trace causality in the Fourier domain. To illustrate these difficulties, consider the unity feedback closed-loop system in Fig. 1.4(c) with $d = n = 0$. Let the plant P be LTI with the impulse response satisfying $p(t) = -2\,\mathrm{e}^{-t}\,\mathbb{1}(t)$ and the controller R be static with $r = \delta$. This plant is causal and stable and acts in the time domain as

$$y(t) = -2\int_{-\infty}^{t} \mathrm{e}^{-(t-s)}u(s)\,\mathrm{d}s \iff \dot{y}(t) = -y(t) - 2u(t).$$

The controller acts in the time domain as $u = y_{\mathrm{r}} - y$, so the closed-loop system $G : y_{\mathrm{r}} \mapsto y$ satisfies

$$\dot{y}(t) = -y(t) - 2\big(y_{\mathrm{r}}(t) - y(t)\big) = y(t) - 2y_{\mathrm{r}}(t) \iff y(t) = -2\int_{-\infty}^{t} \mathrm{e}^{t-s}u(s)\,\mathrm{d}s.$$

This is a *causal* (the closed-loop system should remain causal by the very nature of this feedback interconnection) and *unstable* system. The latter can be seen by the fact that the impulse response of the closed-loop system, which satisfies $g(t) = -2\mathrm{e}^{t}\,\mathbb{1}(t)$, is not in L_1 (cf. Remark 3.2). At the same time, the frequency responses of P and R are $-2/(1+\mathrm{j}\omega)$ and 1, respectively, so the closed-loop system in the frequency domain acts as

$$Y(\mathrm{j}\omega) = P(\mathrm{j}\omega)U(\mathrm{j}\omega) = P(\mathrm{j}\omega)\big(R(\mathrm{j}\omega) - Y(\mathrm{j}\omega)\big) \implies G(\mathrm{j}\omega) = \frac{P(\mathrm{j}\omega)}{1+P(\mathrm{j}\omega)} = \frac{2}{1-\mathrm{j}\omega}.$$

The inverse Fourier transform of $G(\mathrm{j}\omega)$ above satisfies $g(t) = 2\,\mathrm{e}^{t}\,\mathbb{1}(-t)$, which corresponds to a *stable* and *anti-causal* system. But this would be an erroneous conclusion.

3.3.2 Transfer functions

Analyzing LTI systems in the Laplace transform domain is a more reliable means to deal with system interconnections, provided we keep track of regions of convergence. The Laplace transform also turns convolutions into algebraic relations. By the second relation of (3.17),

$$Y(s) = G(s)U(s) \tag{3.19}$$

for any s in the RoC of both $U(s)$ and $G(s)$, where the $p \times m$ function $G(s)$ is the Laplace transform of the impulse response $g(t)$ of G. This $G(s)$ is called the *transfer function* of the LTI system G. As the Laplace transform is applicable to a wider class of signals, relation (3.19) holds for unstable systems as well, with a "dirty" analytic continuation trick to define $G(s)$ beyond its RoC. The transfer functions of the examples considered in Section 3.2 are

$$G_{\mathrm{int}}(s) = \frac{1}{s}, \quad G_{\mathrm{dint},\mu}(s) = \sum_{i\in\mathbb{N}} \mathrm{e}^{-s(i-1)T} = \frac{1}{1-\mathrm{e}^{-sT}}, \quad G_{\mathrm{fmint},\mu}(s) = \frac{1-\mathrm{e}^{-s\mu}}{s}, \quad \text{and} \quad D_{\tau}(s) = \mathrm{e}^{-s\tau},$$

whose RoCs are $\mathbb{C}_0$, $\mathbb{C}_0$ (think of geometric series), $\mathbb{C}$, and $\mathbb{C}$, respectively. The transfer function of the ideal low-pass filter is a shaky business though. Its impulse response can be split as

$$f_{\mathrm{ILP}}(t) = f_{\mathrm{ILP,c}}(t) + f_{\mathrm{ILP},\bar{\mathrm{c}}}(t) := \frac{\omega_{\mathrm{b}}}{\pi}\operatorname{sinc}(\omega_{\mathrm{b}}t)\mathbb{1}(t) + \frac{\omega_{\mathrm{b}}}{\pi}\operatorname{sinc}(\omega_{\mathrm{b}}t)\mathbb{1}(-t).$$

While $\mathfrak{L}\{f_{\mathrm{ILP,c}}\}$ has $\mathbb{C}_0$ as its RoC, $\mathfrak{L}\{f_{\mathrm{ILP},\bar{\mathrm{c}}}\}$ has it in $\mathbb{C}\setminus\bar{\mathbb{C}}_0$ (as neither of them belongs to L_1). These two regions are not intersecting.

Returning to the example considered in §3.3.1, the transfer functions of the plant and the controller there are $P(s) = -2/(s+1)$ and $R(s) = 1$, with RoCs in $\mathbb{C}_{-1}$ and $\mathbb{C}$, respectively. The closed-loop transfer function

$$G(s) = \frac{P(s)}{1+P(s)} = \frac{2}{-s+1}.$$

If considered out of context, this transfer function is ambiguous in defining the impulse response of G. Specifically, it may correspond to g satisfying either $g(t) = -2\,\mathrm{e}^{t}\mathbb{1}(t)$, if the RoC is the right half-plane $\mathbb{C}_1$, or $g(t) = 2\,\mathrm{e}^{-t}\mathbb{1}(-t)$, if the RoC is the left half-plane $\mathbb{C}\setminus\bar{\mathbb{C}}_1$. But in the current context, with the RoC of $P(s)$ in a right half-plane, only the former option is possible. Thus, the closed-loop system is a causal unstable system with the impulse response $g = -2\exp_1\mathbb{1}$, which is the right conclusion.

The line of reasoning above is common in feedback control applications, which consider almost exclusively causal systems. Thus, possible RoCs are some right half-planes (sometimes, the whole $\mathbb{C}$). To keep consistency among regions of convergence, signals are then also assumed to have support in $\mathbb{R}_+$. And once regions of convergence are agreed, LTI systems can be represented by their transfer functions and manipulated algebraically. For example, the parallel and cascade interconnections of two compatibly dimensioned systems G_1 and G_2 have the transfer functions $G_1(s)+G_2(s)$ and $G_2(s)G_1(s)$, respectively, and the inverse of G has $G^{-1}(s)$ as its transfer function. We even use the same notation for systems and their transfer functions, especially if the Laplace variable is dropped, and frequently interchange these notions.

The L_2-stability can also be analyzed in terms of transfer functions. Namely, it is a known, albeit nontrivial, result that an LTI G is causal and L_2-stable iff its transfer function is holomorphic and bounded in $\mathbb{C}_0$, i.e. iff $G \in H_\infty^{p\times m}$, where

$$H_\infty^{p\times m} := \left\{G : \mathbb{C}_0 \to \mathbb{C}^{p\times m} \mid G \text{ is holomorphic in } \mathbb{C}_0 \text{ and } \|G\|_\infty := \sup_{s\in\mathbb{C}_0}\|G(s)\| < \infty\right\}, \tag{3.20}$$

and $\|G(s)\|$ is the matrix spectral norm on $\mathbb{C}^{p\times m}$ defined by (2.6b). In other words, given a transfer function $G(s)$, then $GH_2^m \subset H_2^p$ iff $G \in H_\infty^{p\times m}$. If $p = m$ and not only $G(s)$, but also its inverse $G^{-1} \in H_\infty$, then we have that $GH_2^m = H_2^m$, i.e. any output $y \in L_{2+}$ has an input $u \in L_{2+}$ producing it. Regarding the examples from Section 3.2,

- $G_{\mathrm{int}} \notin H_\infty$: $1/s$ is holomorphic in $\mathbb{C}_0$ (remember, the imaginary axis is not a part of this region), but it is not bounded there, because no upper bound on $1/|s|$ exist as $s \to 0$ along any path in $\mathbb{C}_0$;
- $G_{\mathrm{dint},\mu} \notin H_\infty$ for the same reasons;
- $G_{\mathrm{fmint},\mu} \in H_\infty$: $(1-\mathrm{e}^{-s\mu})/s$ is holomorphic in $\mathbb{C}_0$ and bounded there, because

$$\begin{aligned}
\left|G_{\mathrm{fmint},\mu}\left(\frac{\sigma+\mathrm{j}\omega}{\mu}\right)\right|^2 &= \mu^2\left|\frac{1-\mathrm{e}^{-(\sigma+\mathrm{j}\omega)}}{\sigma+\mathrm{j}\omega}\right|^2 = \mu^2\frac{1-2\mathrm{e}^{-\sigma}\cos\omega+\mathrm{e}^{-2\sigma}}{\sigma^2+\omega^2}\\
&= \mu^2\left(\frac{1-\mathrm{e}^{-\sigma}}{\sigma}\right)^2 - \mu^2\frac{4\omega^2\mathrm{e}^{-\sigma}}{\sigma^2(\sigma^2+\omega^2)}\left(\sinh^2\left(\frac{\sigma}{2}\right) - 2\frac{1-\cos\omega}{\omega^2}\left(\frac{\sigma}{2}\right)^2\right)\\
&\le \mu^2\left(\frac{1-\mathrm{e}^{-\sigma}}{\sigma}\right)^2 - \mu^2\frac{4\omega^2\mathrm{e}^{-\sigma}(\sinh^2(\sigma/2)-(\sigma/2)^2)}{\sigma^2(\sigma^2+\omega^2)} \le \mu^2\left(\frac{1-\mathrm{e}^{-\sigma}}{\sigma}\right)^2\\
&< \mu^2
\end{aligned}$$

for all $\sigma > 0$, where the inequalities $2(1-\cos\omega)/\omega^2 \le 1$ and $\sinh^2 x > x^2$ for all $x \ne 0$ were used;

- $D_\tau \in H_\infty$: $\mathrm{e}^{-s\tau}$ is holomorphic in $\mathbb{C}_0$ (in fact, entire) and bounded there, $|\mathrm{e}^{-(\sigma+\mathrm{j}\omega)\tau}| = |\mathrm{e}^{-\sigma\tau}| < 1$.

Of these transfer functions only $G_{\mathrm{dint},\mu}(s)$ is stably invertible, as $G^{-1}_{\mathrm{dint},\mu}(s) = 1 - \mathrm{e}^{-s\mu}$ does belong to H_∞. The transfer function of the closed-loop system in the example from the beginning of this section is also not in H_∞, because $2/(-s+1)$ is not holomorphic at $s = 1$.

The H_∞ space is a Banach (complete normed) space. Yet its norm in the form defined by (3.20) is not frequently used because it is neither readily calculable (cf. the calculations for $G_{\mathrm{fmint},\mu}$ above) nor intuitively interpretable. There is a more tangible form though. Like in the case of the H_2 signal space defined by (3.10), it can be shown that every $G \in H_\infty$ has a unique *boundary function* $\tilde{G} \in L_\infty(\mathrm{j}\mathbb{R})$ such that $\tilde{G}(\mathrm{j}\omega) = \lim_{\sigma\downarrow 0} G(\sigma + \mathrm{j}\omega)$ for almost all ω, where

$$L_\infty^{p\times m}(\mathrm{j}\mathbb{R}) := \left\{ \tilde{G} : \mathrm{j}\mathbb{R} \to \mathbb{C}^{p\times m} \mid \|\tilde{G}\|_\infty := \operatorname{ess\,sup}_{\omega\in\mathbb{R}} \|\tilde{G}(\mathrm{j}\omega)\| < \infty \right\}, \tag{3.21}$$

and $\|\tilde{G}\|_\infty = \|G\|_\infty$. It is customary to identify G with $\tilde{G}$ and regard H_∞ as a closed subspace of $L_\infty(\mathrm{j}\mathbb{R})$, in which case

$$\|G\|_\infty = \operatorname*{ess\,sup}_{\omega\in\mathbb{R}} \|G(\mathrm{j}\omega)\|. \tag{3.22}$$

It should be emphasized that this equality holds only for functions $G \in H_\infty$ as defined by (3.20). For example, the quantity on the right-hand side of (3.22) for $G(s) = 1/(s-1)$ equals 1. But this $G \notin H_\infty$, so (3.22) makes no sense for it. In the SISO case $\|G\|_\infty$ equals the peak of the magnitude frequency response of $G(\mathrm{j}\omega)$, which is the maximal gain of the stable G for all possible single-harmonic inputs. Moreover, the H_∞-norm of the transfer function $G(s)$ is the L_2-induced norm of the system G, i.e. its square is the energetic gain of G in the sense that it equals the upper limit on the output energy attainable by all possible unit-energy inputs.

Remark 3.3 (Poisson integral). An interesting, and sometimes useful, mathematical fact is that the boundary function $\tilde{G} \in L_\infty(\mathrm{j}\mathbb{R})$ of $G \in H_\infty$ completely determines G. Namely, at each $s \in \mathbb{C}_0$ we have that

$$G(s) = \frac{1}{\pi}\int_{\mathbb{R}} \tilde{G}(\mathrm{j}\omega)\frac{\operatorname{Re} s}{(\operatorname{Re} s)^2 + (\operatorname{Im} s - \omega)^2}\mathrm{d}\omega. \tag{3.23}$$

This relation is known as the *Poisson integral formula*. ▽

Remark 3.4 (stability and system poles). The exhaustive characterization of stable and causal systems as those having transfer functions in H_∞ is a powerful property. Among other things, it can lead to the conclusion that the sheer absence of poles in the closed right half-plane $\overline{\mathbb{C}}_0$ is not always an indicator of stability. To see that, consider an LTI system G with the transfer function

$$G(s) = \frac{1}{s + 1 + s\mathrm{e}^{-s}}. \tag{3.24}$$

The characteristic equation of $G(s)$ is $s + 1 + s\mathrm{e}^{-s} = 0$ or, equivalently, $\mathrm{e}^{-s} = -(s+1)/s$. Assuming that $s = \sigma + \mathrm{j}\omega$ is its root, the magnitude equality yields that

$$\mathrm{e}^{-\sigma} = \left|1 + \frac{1}{\sigma + \mathrm{j}\omega}\right| = \left|\left(1 + \frac{\sigma}{\sigma^2+\omega^2}\right) - \mathrm{j}\frac{\omega}{\sigma^2+\omega^2}\right| \geq \left|1 + \frac{\sigma}{\sigma^2+\omega^2}\right|.$$

First, consider the case $\sigma = 0$. The first equality above reads then $1 = 1 + 1/|\omega|$, which holds for none $\omega \in \mathbb{R}$. Now, let $\sigma > 0$. The inequality above yields then that $\mathrm{e}^{-\sigma} > 1$, which is again impossible. Thus, $G(s)$ has no poles in the closed RHP. Nevertheless, this $G \notin H_\infty$. To see this, note that $s + 1 + s\mathrm{e}^{-s}$ is an entire function with an infinite number of roots. As such [24, Thm. 10.18], these roots do not accumulate and there must exist a sequence $\{s_i\} \in \mathbb{C} \setminus \overline{\mathbb{C}}_0$ such that

$$s_i + 1 + s_i\mathrm{e}^{-s_i} = 0 \quad \text{and} \quad \lim_{i\to\infty} |s_i| = \infty.$$

Define then yet another sequence, $\{\tilde{s}_i\} \in \mathbb{C}_0$ with $\tilde{s}_i = -s_i$. The values of $G(s)$ at each $\tilde{s}_i$ satisfy

$$G(\tilde{s}_i) = \frac{1}{1 - s_i - s_i \mathrm{e}^{s_i}} = \frac{1}{1 - s_i + s_i^2/(1 + s_i)} = 1 + s_i.$$

Thus, we have a sequence in $\mathbb{C}_0$ at which $\lim_{i\to\infty}|G(\tilde{s}_i)| = \infty$. Hence, $G \notin H_\infty$ indeed. This, in turn, implies that the system G is L_2-unstable. ∇

Another important system space that can be expressed in terms of transfer functions is

$$H_2^{p\times m} := \left\{ G : \mathbb{C}_0 \to \mathbb{C}^{p\times m} \,\middle|\, G(s) \text{ is holomorphic in } \mathbb{C}_0 \right.$$
$$\left. \text{and } \|G\|_2 := \left(\sup_{\sigma>0} \frac{1}{2\pi} \int_{\mathbb{R}} \|G(\sigma + \mathrm{j}\omega)\|_{\mathrm{F}}^2 \mathrm{d}\omega \right)^{1/2} < \infty \right\}, \quad (3.25)$$

which is effectively a version of the H_2 signal space defined by (3.10) for matrix-valued signals. Like in the vector-valued case, every $G \in H_2$ has a unique boundary function $\tilde{G} \in L_2(\mathrm{j}\mathbb{R})$, where

$$L_2^{p\times m}(\mathrm{j}\mathbb{R}) := \left\{ \tilde{G} : \mathrm{j}\mathbb{R} \to \mathbb{C}^{p\times m} \,\middle|\, \|\tilde{G}\|_2 := \left(\frac{1}{2\pi} \int_{\mathbb{R}} \|\tilde{G}(\mathrm{j}\omega)\|_{\mathrm{F}}^2 \mathrm{d}\omega \right)^{1/2} < \infty \right\},$$

such that $\|G\|_2 = \|\tilde{G}\|_2$. The space H_2 is regarded then as a closed subspace of $L_2(\mathrm{j}\mathbb{R})$. By Parseval's theorem $G \in L_2(\mathrm{j}\mathbb{R})$ iff the impulse response of the corresponding system $g \in L_2(\mathbb{R})$ and by the Paley–Wiener theorem $G \in H_2$ iff its impulse response $g \in L_{2+}$, i.e. G is also causal. Thus, H_2 is the space of transfer functions of causal systems, whose impulse responses have finite energy, i.e. $\|g\|_2^2 < \infty$. A consequence of this is that a system G has an $L_2(\mathrm{j}\mathbb{R})$ frequency response or an H_2 transfer function only if $g_i = 0$ for all $i \in \mathbb{Z}$ in (3.13). For the examples from Section 3.2,

- $G_{\mathrm{int}} \notin H_2$: $1/s$ is holomorphic in $\mathbb{C}_0$, but

$$\frac{1}{2\pi} \int_{\mathbb{R}} |G_{\mathrm{int}}(\sigma + \mathrm{j}\omega)|^2 \mathrm{d}\omega = \frac{1}{2\pi} \int_{\mathbb{R}} \frac{\mathrm{d}\omega}{\sigma^2 + \omega^2} = \frac{1}{2\sigma}$$

  has no bound over $\sigma > 0$ (alternatively, and simpler, its impulse response $\mathbb{1} \notin L_{2+}$);
- $G_{\mathrm{dint},\mu} \notin H_2$ for similar reasons (the integral actually diverges for every σ then);
- $G_{\mathrm{fmint},\mu} \in H_2$: $(1 - \mathrm{e}^{-s\mu})/s$ is holomorphic in $\mathbb{C}_0$ and

$$\frac{1}{2\pi} \int_{\mathbb{R}} |G_{\mathrm{fmint},\mu}(\sigma + \mathrm{j}\omega)|^2 \mathrm{d}\omega = \frac{1}{2\pi} \int_{\mathbb{R}} \frac{1 - 2\mathrm{e}^{-\sigma\mu}\cos(\omega\mu) + \mathrm{e}^{-2\sigma\mu}}{\sigma^2 + \omega^2} \mathrm{d}\omega = \frac{1 - \mathrm{e}^{-2\sigma\mu}}{2\sigma} < \mu$$

  for all $\sigma > 0$ (alternatively, and simpler, $\|\mathbb{1}_{[0,\mu]}\|_2 = \sqrt{\mu} < \infty$);
- $D_\tau \notin H_2$: $\mathrm{e}^{-s\tau}$ is holomorphic in $\mathbb{C}_0$ (in fact, entire), but

$$\frac{1}{2\pi} \int_{\mathbb{R}} |D_\tau(\sigma + \mathrm{j}\omega)|^2 \mathrm{d}\omega = \frac{\mathrm{e}^{-2\sigma\tau}}{2\pi} \int_{\mathbb{R}} \mathrm{d}\omega = \infty$$

  for all $\sigma > 0$ and all $\tau \geq 0$ (alternatively, it follows from the fact that $D_\tau\delta \notin L_{2+}$ for all $\tau \geq 0$).

Stability is not necessarily related to H_2. We already saw an example, $D_\tau(s)$, of a transfer function, which belongs to H_∞, but not to H_2. An opposite situation is also possible. For instance, consider a causal LTI system G having the impulse response $g = \operatorname{sinc} \mathbb{1}$. This is a truncated and scaled version of the ideal low-pass filter. It can be verified that $\|G\|_2 = \|g\|_2 = \sqrt{\pi/2} < \infty$ in this case, so $G \in H_2$. But its transfer function $G(s) = \arctan(1/s)$ does not belong to H_∞ as it is unbounded on any path in $\mathbb{C}_0$ approaching $s = \pm\mathrm{j}$. Hence, this $G \notin H_\infty$ and thus the system is unstable.

The H_2 space finds its use mainly in measuring performance. The H_2-norm of error systems is the cost function of choice in many optimal control and estimation problems, like LQG and Wiener / Kalman filtering. The reason for this is twofold. First, $L_2(\mathrm{j}\mathbb{R})$ and H_2 are a Hilbert space, with the inner product

$$\langle G_1, G_2\rangle_2 := \frac{1}{2\pi}\int_{\mathbb{R}} \operatorname{tr}\big([G_2(\mathrm{j}\omega)]' G_1(\mathrm{j}\omega)\big)\,\mathrm{d}\omega = \int_{\mathbb{R}} \operatorname{tr}\big([g_2(t)]' g_1(t)\big)\,\mathrm{d}t. \tag{3.26}$$

This renders treatments of corresponding optimization problems simpler, with the neat Projection Theorem as a key tool. Second, the H_2-norm has transparent deterministic and stochastic time-domain interpretations. The former we already saw, it is the energy of the impulse response. From the stochastic point of view, the squared H_2-norm of a system equals the steady-state variance of its response to a unit-intensity white Gaussian input.

A $p \times m$ transfer function $G(s)$ is said to be *proper* if

$$\exists \alpha \geq 0 \text{ such that } \sup_{s\in\mathbb{C}_\alpha} \|G(s)\| < \infty. \tag{3.27}$$

Clearly, any H_∞ transfer function is proper, so causal and stable systems must have proper transfer functions. All examples considered above are proper. Examples of non-proper transfer functions are $s^2/(s+1)$ and $D_\tau(s)$ for $\tau < 0$. The former corresponds to an unstable systems, because its domain includes only differentiable functions from L_2, and the latter corresponds to a non-causal system. Properness is sometimes confused with the boundedness of the frequency responses at high frequencies. Namely, some references call $G(s)$ proper if there is a frequency ω_a such that $\sup_{\omega>\omega_\mathrm{a}}\|G(\mathrm{j}\omega)\| < \infty$. This is not always the same as (3.27). Indeed, $G_{\mathrm{dint},\mu}(s) = 1/(1-\mathrm{e}^{-sT})$ would not be regarded proper by the definition based on the frequency response, although it is proper by (3.27). A transfer function is called *strictly proper* if there is $\alpha \geq 0$ such that

$$\lim_{|s|\to\infty, s\in\mathbb{C}_\alpha} \|G(s)\| = 0. \tag{3.28}$$

It can be shown that any H_2 transfer function is strictly proper. Of the examples above, only $D_\tau(s)$ is not strictly proper. The transfer function $G(s)$ defined by (3.24) is also strictly proper. Strictly proper transfer functions might not correspond to finite-bandwidth frequency responses. The same $G(s)$ from (3.24) is an example of that.

The transfer function of the anti-causal adjoint G' of a causal LTI system G can be derived from the property of the impulse response of G' discussed on p. 44. Namely, let g be the impulse response, having support in $\mathbb{R}_+$, of G. The impulse response of G' satisfies then $[g(-t)]'$. It has support in $\mathbb{R}_-$ and its Laplace transform

$$\mathfrak{L}\{g'\} = \int_{\mathbb{R}} [g(-t)]' \mathrm{e}^{-st}\,\mathrm{d}t = \Big[\int_{\mathbb{R}} g(t)\mathrm{e}^{-(-\overline{s})t}\,\mathrm{d}t\Big]' = [G(-\overline{s})]'$$

with its RoC expected to be in $\mathbb{C}\setminus\overline{\mathbb{C}}_\alpha$ for some $\alpha \in \mathbb{R}$. The transfer function above, known as the *conjugate* transfer function of $G(s)$, is denoted as

$$G^\sim(s) := [G(-\overline{s})]'. \tag{3.29}$$

If coefficients of $G(s)$ are real, with some abuse of notation we may think of the conjugate transfer function as $G^\sim(s) = [G(-s)]'$, effectively treating the $[\cdot]'$ notation as the mere transpose. The conjugate transfer function becomes the standard adjoint when considered on the imaginary axis, i.e. $G^\sim(\mathrm{j}\omega) = [G(\mathrm{j}\omega)]'$.

A transfer functions $G \in H_\infty^{p\times m}$ is called *inner* (*co-inner*) if $G^\sim(s)G(s) = I_m$ ($G(s)G^\sim(s) = I_p$). Clearly, a transfer function can be inner (co-inner) only if $p \geq m$ ($p \leq m$). Of the examples considered above, only the delay system D_τ has an inner, and co-inner, transfer function, $(\mathrm{e}^{-s\tau})^\sim \mathrm{e}^{-s\tau} = \mathrm{e}^{s\tau}\mathrm{e}^{-s\tau} = 1$.

Another example of a co-inner transfer function is $0.5\left[\ \mathrm{e}^{-s\tau}\ \ (-s+a)/(s+a)\ \right]$ for any $a > 0$. Every system with an inner transfer function is an isometry on L_2, in a sense that it is a (causal) operator $L_2 \to L_2$ such that $\|Gu\| = \|u\|$ for all $u \in L_2$. This can be seen via the relation

$$\|Gu\|_2^2 = \|GU\|_2^2 = \langle GU, GU\rangle_2 = \langle G^{\sim}GU, U\rangle_2 = \langle U, U\rangle_2^2 = \|U\|_2^2 = \|u\|_2^2,$$

where the time-domain norm is on $L_2(\mathbb{R})$ and the frequency-domain norm is on $L_2(\mathrm{j}\mathbb{R})$. Likewise, if G has a co-inner transfer function, its (anti-causal) adjoint G' is an isometry on L_2. In the square case ($p = m$) systems with inner / co-inner transfer functions can be thought of as a dynamic counterpart of unitary matrices introduced in §2.3.4. The following result is important in numerous problems.

Proposition 3.1. *Let $W_{\mathrm{i}}(s)$ and $W_{\mathrm{ci}}(s)$ be an inner and a co-inner transfer functions, respectively. The following statements hold true:*

1. $\|G\|_\infty = \|W_{\mathrm{i}} G W_{\mathrm{ci}}\|_\infty$ *for all* $G \in H_\infty$,
2. $\|G\|_2 = \|W_{\mathrm{i}} G W_{\mathrm{ci}}\|_2$ *for all* $G \in H_2$.

Proof. Follows from the definitions of the corresponding norms and the facts that the spectral (used in the H_∞ case) and Frobenius (used in the H_2 case) norms of a matrix M can be expressed in terms of both $M'M$ and MM'. □

3.3.3 Coprime factorization of transfer functions over H_∞

Two integers n and m are said to be *coprime* if their greatest common divisor is 1. Alternatively, it follows from the Euclidean algorithm that n and m are coprime iff there exist integers x and y such that $xm + yn = 1$. This criterion can be generalized to polynomials, polynomial matrices and, further, to stable transfer functions. It is also handy to use the criterion above as the definition of coprimeness.

Functions $M \in H_\infty^{m\times m}$ and $N \in H_\infty^{p\times m}$ of $s \in \mathbb{C}$ having the same number of columns are said to be (strongly) *right coprime over* H_∞ if there are functions $X \in H_\infty^{m\times m}$ and $Y \in H_\infty^{m\times p}$ such that

$$\left[\ X(s)\ \ Y(s)\ \right]\begin{bmatrix} M(s) \\ N(s) \end{bmatrix} = X(s)M(s) + Y(s)N(s) = I_m. \tag{3.30}$$

The equality above is sometimes called the *Bézout equality* and the transfer matrices $X(s)$ and $Y(s)$ are called the *left Bézout coefficients* for $M(s)$ and $N(s)$. Similarly, functions $\tilde{M} \in H_\infty^{p\times p}$ and $\tilde{N} \in H_\infty^{p\times m}$ of $s \in \mathbb{C}$ having the same number of rows are said to be (strongly) *left coprime over* H_∞ if there are functions $\tilde{X} \in H_\infty^{p\times p}$ and $\tilde{Y} \in H_\infty^{m\times p}$, called the *right Bézout coefficients* for $\tilde{M}(s)$ and $\tilde{N}(s)$, such that

$$\left[\ \tilde{M}(s)\ \ \tilde{N}(s)\ \right]\begin{bmatrix} \tilde{X}(s) \\ \tilde{Y}(s) \end{bmatrix} = \tilde{M}(s)\tilde{X}(s) + \tilde{N}(s)\tilde{Y}(s) = I_p. \tag{3.31}$$

It should be clear that if $m = p = 1$, then M and N are right coprime iff they are left coprime, with the same left and right Bézout coefficients. Conditions (3.30) and (3.31) effectively say that

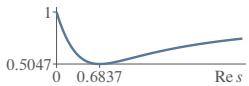
$$\begin{bmatrix} M \\ N \end{bmatrix} \in H_\infty^{(p+m)\times m} \quad \text{and} \quad \left[\ \tilde{M}\ \ \tilde{N}\ \right] \in H_\infty^{p\times(p+m)}$$

are left and right invertible, respectively, in H_∞. We may expect that this requires the non-singularity of the values of those functions over their whole domain $\mathbb{C}_0$. Indeed, an outcome of the celebrated, and highly nontrivial, corona theorem is that M and N are right coprime and $\tilde{M}$ and $\tilde{N}$ are left coprime iff

$$\inf_{s\in\mathbb{C}_0} \underline{\sigma}\left(\begin{bmatrix} M(s) \\ N(s) \end{bmatrix}\right) > 0 \quad \text{and} \quad \inf_{s\in\mathbb{C}_0} \underline{\sigma}\left(\left[\ \tilde{M}(s)\ \ \tilde{N}(s)\ \right]\right) > 0, \tag{3.32}$$

respectively. Thus, we may conclude that $M(s) = 1/(s+1)$ and $N(s) = s\mathrm{e}^{-s}/(s+1)$ are not coprime, because both vanish in $\mathbb{C}_\alpha$ as $\alpha \to \infty$, whereas $M(s) = s/(s+1)$ and $N(s) = \mathrm{e}^{-s}/(s+1)$ are coprime, because

$$\underline{\sigma}\left(\begin{bmatrix} M(s) \\ N(s) \end{bmatrix}\right) = \frac{\sqrt{\mathrm{e}^{-2\operatorname{Re} s} + |s|^2}}{|1+s|} \geq \frac{\sqrt{\mathrm{e}^{-2\operatorname{Re} s} + (\operatorname{Re} s)^2}}{1+\operatorname{Re} s} =$$



$$> \frac{1}{2}, \quad \forall s \in \mathbb{C}_0$$

in this case (the first inequality follows by the fact that $1 + \operatorname{Re} s > \sqrt{\mathrm{e}^{-2\operatorname{Re} s} + (\operatorname{Re} s)^2}$ whenever $\operatorname{Re} s > 0$). As a matter of fact, the Bézout coefficients for the last example are $X(s) = 1 + (1 - \mathrm{e}^{-s})/s = 1 + G_{\mathrm{fmint},1}(s)$ and $Y(s) = 1$.

There is a wide class[1] of systems G, whose transfer functions $G(s)$ can be represented as

$$G(s) = N(s)M^{-1}(s) = \tilde{M}^{-1}(s)\tilde{N}(s) \tag{3.33}$$

for right coprime $M, N \in H_\infty$ and left coprime $\tilde{M}, \tilde{N} \in H_\infty$ such that $M^{-1}(s)$ and $\tilde{M}^{-1}(s)$ are proper. These representations are called a *right coprime factorization* (*rcf*) and a *left coprime factorization* (*lcf*) of $G(s)$ over H_∞, respectively. Hereafter, we refer to the factors M and $\tilde{M}$ as *denominators* of corresponding coprime factorizations and to N and $\tilde{N}$ as their *numerators*. Possible coprime factors of the first four examples introduced in Section 3.2 are as follows (*rcf*'s are assumed below, they are the same as *lcf*'s in the SISO case):

- $G_{\mathrm{int}}(s) = 1/(s+a)\cdot(s/(s+a))^{-1}$ for any $a > 0$, with the Bézout coefficients $X(s) = 1$ and $Y(s) = a$;
- $G_{\mathrm{dint},\mu}(s) = 1 \cdot (1 - \mathrm{e}^{-s\mu})^{-1}$, with the Bézout coefficients $X(s) = 1$ and $Y(s) = \mathrm{e}^{-s\mu}$;
- $G_{\mathrm{fmint},\mu}(s) = (1 - \mathrm{e}^{-s\mu})/s \cdot 1^{-1}$, with the Bézout coefficients $X(s) = 1$ and $Y(s) = 0$;
- $D_\tau(s) = \mathrm{e}^{-s\tau} \cdot 1^{-1}$, with the Bézout coefficients $X(s) = 1$ and $Y(s) = 0$.

The construction of coprime factors and the corresponding Bézout coefficients is particularly simple for stable systems, as could be seen above. Indeed, if $G \in H_\infty$, we can always choose $M(s) = I$ and $N(s) = G(s)$, which are coprime because $X(s) = I$ and $Y(s) = 0$ are left Bézout coefficients for them. The situation in the unstable case is less straightforward though.

The notion of the coprime factorization plays an important role in the stability analysis of feedback systems and designing stabilizing controllers. It is also useful in analyzing properties of MIMO systems. These issues will be studied in Chapters 6 and 7. In the remainder of this section we concentrate on general properties of the coprime factors and their use in characterizing domains of unstable systems, which was promised on p. 43.

Coprime factorizations are not unique, as could be seen in the construction of coprime factors of $G_{\mathrm{int}}(s)$ above. Yet a simple connection between different right (left) coprime factorizations of the same transfer function exists, as shown in the result below.

Proposition 3.2. *If $N_1(s)M_1^{-1}(s) = N_2(s)M_2^{-1}(s)$ and $\tilde{M}_1^{-1}(s)\tilde{N}_1(s) = \tilde{M}_2^{-1}(s)\tilde{N}_2(s)$ are rcf's and lcf's, respectively, then*

$$\begin{bmatrix} M_2(s) \\ N_2(s) \end{bmatrix} = \begin{bmatrix} M_1(s) \\ N_1(s) \end{bmatrix} U(s) \quad \textit{and} \quad \begin{bmatrix} \tilde{M}_2(s) & \tilde{N}_2(s) \end{bmatrix} = \tilde{U}(s) \begin{bmatrix} \tilde{M}_1(s) & \tilde{N}_1(s) \end{bmatrix}$$

for some bi-stable $U(s)$ and $\tilde{U}(s)$, i.e. square and such that that $U, U^{-1}, \tilde{U}, \tilde{U}^{-1} \in H_\infty$.

[1] This class effectively covers all systems of interest in feedback control, because plants not belonging to it are not stabilizable by feedback [26]. An example of unstabilizable plants is $s\mathrm{e}^{-s}$, which indeed does not have strongly coprime factors.

Proof. Only the *rcf* case will be proved. The *lcf* case follows by similar arguments.

Define $U_a := M_1^{-1}M_2$. It can be verified that

$$\begin{bmatrix} M_1 \\ N_1 \end{bmatrix} U_a = \begin{bmatrix} M_2 \\ N_1 M_1^{-1} M_2 \end{bmatrix} = \begin{bmatrix} M_2 \\ N_2 M_2^{-1} M_2 \end{bmatrix} = \begin{bmatrix} M_2 \\ N_2 \end{bmatrix}.$$

Let now the transfer functions $X_1 \in H_\infty$ and $Y_1 \in H_\infty$ be left Bézout coefficients for M_1 and N_1 (they exist by the coprimeness of M_1 and N_1). Pre-multiplying the expression above by $\begin{bmatrix} X_1 & Y_1 \end{bmatrix}$ we obtain that $U_a = X_1 M_2 + Y_1 N_2 \in H_\infty$. Similarly, $(X_2 M_1 + Y_2 N_1) U_a = I$ for any left Bézout coefficients X_2 and Y_2 for M_2 and N_2. Hence, $U_a^{-1} = X_2 M_1 + Y_2 N_1 \in H_\infty$ and $U = U_a$ does satisfy the conditions of the Proposition. □

If U is bi-stable, then $U(s)$ is bounded and nonsingular at every $s \in \mathbb{C}_0$. Hence, if the denominator M of a *rcf* of a system G has a singular $M(s_0)$ at some $s_0 \in \mathbb{C}_0$, then Proposition 3.2 implies that so do denominators of all other *rcf*'s of G. Similar arguments apply to the denominators of *lcf*'s. Yet singular points of denominators may be associated with the instability of G. The arguments above suggest then that the (in)stability of a system may be concluded from the (in)stability of the inverse of denominators of its coprime factorizations. The following result proves that intuition formally and is quite useful.

Proposition 3.3. *If $G(s) = N(s)M^{-1}(s) = \tilde{M}^{-1}(s)\tilde{N}(s)$ are rcf and lcf, respectively, then*

$$G \in H_\infty \iff M^{-1} \in H_\infty \iff \tilde{M}^{-1} \in H_\infty.$$

Proof. It is obvious that the stability of M^{-1} implies that of G. To show the other direction, assume that $G \in H_\infty$. In this case, $XM + YN = I$ implies that $M^{-1} = X + YG \in H_\infty$. This shows that the first equivalence holds true. The equivalence for the denominator in the *lcf* follows by similar arguments. □

The denominator of a *rcf* of G completely determines domains of LTI systems.

Proposition 3.4. *If $G : \mathfrak{D}_G \subset L_2^m \to L_2^p$ is LTI and such that its transfer function admits a rcf over H_∞ of the form $G(s) = N(s)M^{-1}(s)$, then $\mathfrak{D}_G = ML_2^m := \{u \mid u = Mv \text{ for some } v \in L_2^m\} = \operatorname{Im} M$.*

Proof. Denote $\mathcal{V} := ML_2^m$. Because $M \in H_\infty$, the space $\mathcal{V} \subset L_2^m$. Because $G\mathcal{V} = NL_2^m$ and $N \in H_\infty$, we have that $\mathcal{V} \subset \mathfrak{D}_G$. Now, pick an arbitrary $u_0 \in \mathfrak{D}_G$, i.e. consider $u_0 \in L_2^m$ such that $y_0 = Gu_0 \in L_2^p$. To prove that $\mathfrak{D}_G \subset \mathcal{V}$ we need to show that $v_0 := M^{-1}u_0 \in L_2^m$. But we know that

$$\begin{bmatrix} u_0 \\ y_0 \end{bmatrix} = \begin{bmatrix} M \\ N \end{bmatrix} v_0 \in L_2^{m+p} \implies Xu_0 + Yy_0 = v_0 \in L_2^m$$

because $X, Y \in H_\infty$. Thus, $\mathfrak{D}_G = \mathcal{V}$. □

If $G \in H_\infty$, it follows from Proposition 3.2 that M is bi-stable, so $ML_2^m = L_2^m$, as expected. Otherwise, ML_2^m is a proper subspace of L_2^m. For example, $\mathfrak{D}_{G_{\text{int}}} = s/(s+1)\, L_2$ contains only signals, whose Fourier transforms vanish at $\omega = 0$, and $\mathfrak{D}_{G_{\text{dint},\mu}} = (1 - \mathrm{e}^{-s\mu})\, L_2$ contains only signals, whose Fourier transforms vanish at $\omega = 2k\pi/T$ for all $k \in \mathbb{Z}$.

Now, note that

$$\begin{bmatrix} X & Y \\ -\tilde{N} & \tilde{M} \end{bmatrix} \begin{bmatrix} M & -\tilde{Y} \\ N & \tilde{X} \end{bmatrix} = \begin{bmatrix} I_m & Y\tilde{X} - X\tilde{Y} \\ 0 & I_p \end{bmatrix}$$

or, equivalently (remember (B.16a)),

$$I_{m+p} = \begin{bmatrix} X & Y \\ -\tilde{N} & \tilde{M} \end{bmatrix} \begin{bmatrix} M & -\tilde{Y} \\ N & \tilde{X} \end{bmatrix} \begin{bmatrix} I_m & -Y\tilde{X} + X\tilde{Y} \\ 0 & I_p \end{bmatrix} = \begin{bmatrix} X & Y \\ -\tilde{N} & \tilde{M} \end{bmatrix} \begin{bmatrix} M & -\tilde{Y} + M(X\tilde{Y} - Y\tilde{X}) \\ N & \tilde{X} + N(X\tilde{Y} - Y\tilde{X}) \end{bmatrix}.$$

Therefore, if $\tilde{X}$ and $\tilde{Y}$ are right Bézout coefficients for $\tilde{M}$ and $\tilde{N}$, then so are $\tilde{Y} - M(X\tilde{Y} - Y\tilde{X})$ and $\tilde{X} + N(X\tilde{Y} - Y\tilde{X})$. The latter, in turn, means that the Bézout coefficients can always be chosen so that

$$\begin{bmatrix} X(s) & Y(s) \\ -\tilde{N}(s) & \tilde{M}(s) \end{bmatrix} \begin{bmatrix} M(s) & -\tilde{Y}(s) \\ N(s) & \tilde{X}(s) \end{bmatrix} = \begin{bmatrix} I_m & 0 \\ 0 & I_p \end{bmatrix}. \tag{3.34}$$

Such transfer functions are called the *doubly coprime factorization* of G in $H_\infty^{p\times m}$ and they exists whenever so do some *rcf* and *lcf* of G. It should be clear from (3.34) that doubly coprime factors are such that

$$\begin{bmatrix} X(s) & Y(s) \\ -\tilde{N}(s) & \tilde{M}(s) \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} M(s) & -\tilde{Y}(s) \\ N(s) & \tilde{X}(s) \end{bmatrix}$$

are invertible in H_∞, i.e. they are bi-stable.

3.4 Real-rational transfer functions and their properties

Up to this point we did not dig into the MIMO nature of studied systems. Apart from superficial aspects, like marking input and output dimensions explicitly at times, using the identity matrices instead of the scalar identity and norms instead of absolute values, or writing inverses as M^{-1} rather than $1/M$, the arguments were "gender-neutral." It is time to scrutinize properties of MIMO transfer functions.

Throughout this section, and mostly throughout the rest of the notes, we consider the class of *real-rational* transfer functions, which are $p \times m$ transfer functions $G(s)$ each element of which, $G_{ij}(s)$, is the quotient of two finite polynomials of s with real coefficients. Such transfer functions correspond to systems described by *ordinary differential equations*. The impulse responses of systems with real-rational transfer functions are a superposition of a finite number of exponential and polynomial functions of t and at most one weighted Dirac delta, which follows by applying the inverse Laplace transform to a real-rational $G(s)$.

Properties of LTI systems are noticeably simplified when their transfer functions are real rational. A real-rational transfer function $G(s)$ is then proper iff $\|G(\infty)\| < \infty$ or, equivalently, iff the degree of the numerator of each element $G_{ij}(s)$ does not exceed that of its denominator. A transfer function is strictly proper iff $G(\infty) = 0$ or, equivalently, iff the degree of the numerator of each elements $G_{ij}(s)$ is less than that of its denominator. We also say that a square $G(s)$ is *bi-proper* if $\det(G(\infty)) \neq 0$. Subsets of the H_∞ and H_2 spaces introduced in §3.3.2 consisting of real-rational functions are denoted RH_∞ and RH_2, respectively. It can be shown that RH_∞ (RH_2) comprises all proper (strictly proper) transfer functions, whose elements have no poles in the closed right half-plane $\overline{\mathbb{C}}_0$, thus, $RH_2 \subset RH_\infty$. Consequently, an LTI system with a real-rational transfer function is stable iff its transfer function is proper and has no poles in the closed right half-plane, which is a familiar criterion. Also, the L_∞ and L_2 stability notions are equivalent for such systems. Finally, it can be shown (a constructive proof is presented in §4.3.1) that proper real-rational systems always admit *lcf* and *rcf* over H_∞ or, more precisely, over RH_∞ then.

3.4.1 Poles, zeros, and degree: diagonal case

The concepts of poles and zeros of transfer functions play an important role in classical control. The purpose of this and the next subsections is to extend these classical notions to MIMO systems. The extension is not trivial and evinces some qualitative differences between SISO and MIMO systems.

Similarly to the discussion in §2.3.2, we start studying MIMO transfer functions with the diagonal case. This class of transfer functions is relatively simple and intuitive, yet already captures important differences from SISO systems. Like in the static case, a square $m \times m$ transfer function $G(s)$ is called diagonal if

$G(s)e_i = e_i G_i(s)$ for all $i \in \mathbb{Z}_{1..m}$, for some SISO transfer functions $G_i(s)$. In other words, diagonal

$$G(s) = \begin{bmatrix} G_1(s) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & G_m(s) \end{bmatrix}.$$

Diagonal transfer functions may be thought of as merely a collection of m *independent* SISO systems (called subsystems). As such, properties of diagonal systems are readily deducible from those of each of their subsystems. In particular, *poles* (*zeros*) of $G(s)$ can be defined as the union of poles (zeros) of each $G_i(s)$, where $i = 1, \ldots, m$. With these definitions, poles of diagonal MIMO systems are still points in $\mathbb{C}$ where $G(s)$ is not defined and zeros are the points at which $G(s)$ becomes singular.

It is worth emphasizing that each diagonal element of $G(s)$ is an independent system. Hence, even if $G_i(s)$ and $G_j(s)$ have poles (zeros) at the very same point of the complex plane, these are *different* poles (zeros). For example, $G(s) = \operatorname{diag}\{1/s, 1/s\}$ has *two* poles at the origin because these poles belong to independent subsystems. Accordingly, the *degree* of $G(s)$ must be defined as the sum of the degrees of each of its subsystems $G_i(s)$ (provided, of course, that the transfer function of each subsystem is irreducible). An important consequence of this discussion is that diagonal systems may have *uncancellable* poles and zeros at the same point. Indeed, nothing prevents one subsystem to have a pole at a zero of another subsystem. For example, $G(s) = \operatorname{diag}\{(s-1)/s, s/(s-1)\}$ has both poles and zeros at $s = 0$ and $s = 1$. Yet these poles and zeros cannot be canceled because they belong to independent subsystems. The last example also shows clearly that $\det(G(s))$ might not be a good means for determining poles and zeros of $G(s)$ in the MIMO case. Indeed, $\det\big(\operatorname{diag}\{(s-1)/s, s/(s-1)\}\big) = 1$, which does not contain any information about poles and zeros of this $G(s)$.

The discussion in the previous paragraph suggests that the location of system poles and zeros alone does not provide sufficient information about properties of these poles and zeros. In the diagonal case, the location information can be complemented by associating each pole and zero with the corresponding subsystem(s). This should identify poles and zeros exhaustively. However, the mere association with subsystems is not extendible to more general, not necessarily diagonal, systems.

A more suitable alternative in this respect is the introduction of the notion of poles and zeros *directions* (or spatial directions). We may say that a pole p_k of the ith subsystem $G_i(s)$ is a pole of $G(s)$ with the pole direction $\operatorname{span}(e_i)$. If p_k is *also* a pole of the jth subsystem $G_j(s)$, we say that it is a pole of $G(s)$ with the direction $\operatorname{span}(e_i, e_j)$. This definition can be routinely extended to the case where p_k is a pole of μ_k subsystems. We then say that p_k is a pole of $G(s)$ with a *geometric multiplicity* of μ_k. Note that the geometric multiplicity notion does not account for the multiplicity of poles in each subsystem, which might also be important. By analogy with the corresponding definitions for matrix eigenvalues, we may then call the multiplicity of a pole in the ith subsystem its ith *partial multiplicity* and then the sum of partial multiplicities—the *algebraic multiplicity* of this pole. To complete the picture, the notions of directions and multiplicities can be associated with zeros in a similar fashion. If there is a pole and a zero at the same point $s_0 \in \mathbb{C}$, a pole-zero cancellation between them takes place iff the intersection of their directions is non-empty, i.e. iff their directions are not orthogonal. Returning to $G(s) = \operatorname{diag}\{(s-1)/s, s/(s-1)\}$, we can see that it has a single pole at the origin, whose direction is $\operatorname{span}(e_1)$, and a single zero at the origin, whose direction is $\operatorname{span}(e_2)$. Because $\operatorname{span}(e_1) \perp \operatorname{span}(e_2)$, there is no cancellation between them, which we already saw. Similar arguments obviously apply to the pole and zero of that $G(s)$ at $s = 1$.

3.4.2 Poles, zeros, and degree: general case

To extend the definitions of the previous subsection to the general (not necessarily diagonal) case, some preliminaries are required. The notion of the *normal rank* is a generalization of the matrix rank notion to transfer functions. Given a transfer matrix $G(s)$, its normal rank is $\operatorname{nrank}(G(s)) := \max_{s\in\mathbb{C}} \operatorname{rank}(G(s))$. It

can be shown that if $G(s)$ is proper, then $\operatorname{rank}(G(s)) = \operatorname{nrank}(G(s))$ for all but a finitely many s. A square polynomial matrix $U(s)$ is said to be *unimodular* if $\det(U(s)) = \text{const} \neq 0$. In other words, the inverse of a unimodular polynomial matrix exists and is a polynomial matrix too. Unimodular polynomial matrices can be thought of as polynomial matrices having no zeros. Finally, we say that a polynomial $\beta(s)$ *divides* a polynomial $\alpha(s)$ if $\alpha(s)/\beta(s)$ is polynomial as well.

We are now in the position to state the following important result, showing that any transfer function can be diagonalized by unimodular basis changes.

Theorem 3.5 (Smith–McMillan form)**.** *Given a $p \times m$ transfer function $G(s)$ having* $\operatorname{nrank}(G(s)) = r$ *for some $r \leq \min\{p, m\}$, there are unimodular polynomial matrices $U(s)$ and $V(s)$ such that*

$$U(s)G(s)V(s) = \begin{bmatrix} \alpha_1(s)/\beta_1(s) & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & \alpha_r(s)/\beta_r(s) & 0 \\ 0 & \cdots & 0 & 0 \end{bmatrix}, \tag{3.35}$$

where $\alpha_i(s)$ divides $\alpha_{i+1}(s)$, $\beta_{i+1}(s)$ divides $\beta_i(s)$, and $\alpha_i(s)$ and $\beta_i(s)$ are coprime at every $i \in \mathbb{Z}_{1..r}$.

Since polynomial matrices $U(s)$ and $V(s)$ are unimodular, they affect neither the singularities of $G(s)$ nor its rank. Poles and zeros of $G(s)$ can then be defined in terms of the polynomials $\alpha_i(s)$ and $\beta_i(s)$. Namely, the roots of the polynomials

$$\phi_{\mathrm{p}}(s) := \prod_{i=1}^{r} \beta_i(s) \quad \text{and} \quad \phi_{\mathrm{z}}(s) := \prod_{i=1}^{r} \alpha_i(s) \tag{3.36}$$

are called the *poles* and the *transmission zeros* (or simply zeros) of $G(s)$, respectively. It can be seen from (3.35) that $p_i \in \mathbb{C}$ is a pole of $G(s)$ iff it is a singularity of $G(s)$, i.e. a point which is a pole of at least one entry of $G(s)$. If $p_i \in \mathbb{C}$ is a root of $\beta_{\mu_i}(s)$ and not a root of $\beta_{\mu_i+1}(s)$ for some $\mu_i \leq r$, then we say that the pole p_i has a *geometric multiplicity* of μ_i. The multiplicity of p_i in each $\beta_j(s)$ is said to be its jth *partial multiplicity* and the multiplicity of p_i in $\phi_{\mathrm{p}}(s)$ is said to be its *algebraic multiplicity*. Similarly, if $z_i \in \mathbb{C}$ is a root of $\alpha_{r-\mu_i+1}(s)$ and not a root of $\alpha_{r-\mu_i}(s)$ for some $\mu_i \leq r$, then we say that the transmission zero z_i has a *geometric multiplicity* of μ_i. The multiplicity of z_i in each $\alpha_j(s)$ is said to be its jth *partial multiplicity* and the multiplicity of z_i in $\phi_{\mathrm{z}}(s)$ is said to be its *algebraic multiplicity*. Finally,

$$n := \sum_{i=1}^{r} \deg(\beta_i(s)) = \deg(\phi_{\mathrm{p}}(s)) \tag{3.37}$$

is called the *McMillan degree* (or degree) of $G(s)$.

Let $p_i \in \mathbb{C}$ be a pole of $G(s)$, whose geometric multiplicity is μ_i. Motivated by the decomposition in (3.35), by the input direction of p_i we understand the subspace

$$\operatorname{pdir}_{\mathrm{i}}(G, p_i) = \left(\operatorname{Im} V(p_i) \begin{bmatrix} e_{\mu_i+1} & \cdots & e_m \end{bmatrix}\right)^{\perp} = \ker \begin{bmatrix} e'_{\mu_i+1} \\ \vdots \\ e'_m \end{bmatrix} [V(p_i)]' \subset \mathbb{C}^m, \tag{3.38a}$$

where e_i is the ith standard basis in $\mathbb{C}^m$. By the output direction of p_i we then understand the subspace

$$\operatorname{pdir}_{\mathrm{o}}(G, p_i) = \ker \begin{bmatrix} \tilde{e}'_{\mu_i+1} \\ \vdots \\ \tilde{e}'_p \end{bmatrix} U(p_i) = \left(\operatorname{Im}[U(p_i)]' \begin{bmatrix} \tilde{e}_{\mu_i+1} & \cdots & \tilde{e}_p \end{bmatrix}\right)^{\perp} \subset \mathbb{C}^p, \tag{3.38b}$$

where $\tilde{e}_i$ stands for the ith standard basis in $\mathbb{C}^p$. The dimension of both the input and the output pole directions equals the geometric multiplicity of the pole, μ_i.

By analogy with (3.38), define the input and output directions of a transmission zero $z_i \in \mathbb{C}$ of $G(s)$ with the geometric multiplicity μ_i as

$$\mathrm{zdir_i}(G, z_i) := \operatorname{Im} V(z_i) \begin{bmatrix} e_{r-\mu_i+1} & \cdots & e_m \end{bmatrix} \subset \mathbb{C}^m \tag{3.39a}$$

and

$$\mathrm{zdir_o}(G, z_i) := \operatorname{Im}[U(z_i)]' \begin{bmatrix} \tilde{e}_{r-\mu_i+1} & \cdots & \tilde{e}_p \end{bmatrix} \subset \mathbb{C}^p, \tag{3.39b}$$

respectively. The dimensions of the input and output zero directions are $\mu_i+m-r \geq \mu_i$ and $\mu_i+p-r \geq \mu_i$, respectively. Thus, the dimension of the input direction of z_i exceeds its multiplicity whenever $r < m$. The latter might happen if $G(s)$ is "fat," i.e. $p < m$, or has a defective normal rank. If either of these conditions holds, $G(s)$ has a nontrivial null space at every s at which it is defined and $\mathrm{zdir_i}(G, z_i)$ contains "artifacts" not directly related to z_i. This part can, in principle, be separated from the "zero-related" null space, although this direction is not pursued below, see [14, Sec. A.4] for details. The same applies to the output zero direction if $r < p$.

If s_0 is both a pole and a zero of $G(s)$ having geometric multiplicities μ_p and μ_z, respectively, then $\mu_\mathrm{p} + \mu_\mathrm{z} \leq r$ and (3.38) and (3.39) yield that $\mathrm{pdir_i}(G, s_0) \perp \mathrm{zdir_i}(G, s_0)$ and $\mathrm{pdir_o}(G, s_0) \perp \mathrm{zdir_o}(G, s_0)$. This is why these pole and zero do not cancel each other.

Example 3.1. Consider the 2×2 transfer function

$$G(s) = \frac{1}{s} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}.$$

The nonsingular constant matrices

$$U(s) = \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix} \quad \text{and} \quad V(s) = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}$$

are obviously unimodular and

$$U(s)G(s)V(s) = \begin{bmatrix} 1/s & 0 \\ 0 & 0 \end{bmatrix}$$

is in the form of Theorem 3.5. Thus, the McMillan degree of $G(s)$ is 1 and it has a single pole at $s = 0$ and no transmission zeros. The directions of this pole are

$$\mathrm{pdir_i}(G, 0) = \ker \begin{bmatrix} 0 & 1 \end{bmatrix} [V(0)]' = \operatorname{span}\left(\begin{bmatrix} 1 \\ 1 \end{bmatrix} \right) \quad \text{and} \quad \mathrm{pdir_o}(G, 0) = \ker \begin{bmatrix} 0 & 1 \end{bmatrix} U(0) = \operatorname{span}\left(\begin{bmatrix} 1 \\ 1 \end{bmatrix} \right).$$

As a matter of fact, these directions coincide with the directions defined by the first right and left singular vectors of $\operatorname{Res}(G(s), 0)$. ◊

Example 3.2. Consider the 2×2 transfer function

$$G(s) = \begin{bmatrix} 1 & 1/s \\ 0 & 1 \end{bmatrix}.$$

The polynomial matrices

$$U(s) = \begin{bmatrix} 1 & 0 \\ s & -1 \end{bmatrix} \quad \text{and} \quad V(s) = \begin{bmatrix} 0 & 1 \\ 1 & -s \end{bmatrix}$$

are unimodular and

$$U(s)G(s)V(s) = \begin{bmatrix} 1/s & 0 \\ 0 & s \end{bmatrix}$$

is in the form of Theorem 3.5. Hence, the McMillan degree of $G(s)$ is 1 and it has a single pole at $s = 0$ and a single transmission zero at $s = 0$ too. The directions of the pole are

$$\mathrm{pdir_i}(G, 0) = \ker \begin{bmatrix} 0 & 1 \end{bmatrix} [V(0)]' = \mathrm{span}\left(\begin{bmatrix} 0 \\ 1 \end{bmatrix}\right) \quad \text{and} \quad \mathrm{pdir_o}(G, 0) = \ker \begin{bmatrix} 0 & 1 \end{bmatrix} U(0) = \mathrm{span}\left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}\right)$$

and of the zero are

$$\mathrm{zdir_i}(G, 0) = \mathrm{Im}\, V(0) \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \mathrm{span}\left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}\right) \quad \text{and} \quad \mathrm{zdir_o}(G, 0) = \mathrm{Im}[U(0)]' \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \mathrm{span}\left(\begin{bmatrix} 0 \\ 1 \end{bmatrix}\right).$$

As expected, corresponding pole and zero directions are orthogonal. ◊

Example 3.3. Consider the 2×3 transfer function

$$G(s) = \begin{bmatrix} 1/(s+1) & 0 & (s-1)/((s+1)(s+2)) \\ -1/(s-1) & 1/(s+2) & 1/(s+2) \end{bmatrix}.$$

Let

$$U(s) = \frac{1}{6} \begin{bmatrix} 3 & 3 \\ s^3 - s^2 - 4s - 2 & s^3 - s^2 - 4s + 4 \end{bmatrix} \quad \text{and} \quad V(s) = \frac{1}{6} \begin{bmatrix} 2(s-2) & -6(s-1) & -3(s-1) \\ 4 & -24 & -6(s+2) \\ 0 & 6 & 3(s+2) \end{bmatrix},$$

which are unimodular because $\det(U(s)) = 1/2$ and $\det(V(s)) = 1$. It can be verified that

$$U(s)G(s)V(s) = \begin{bmatrix} 1/((s^2-1)(s+2)) & 0 & 0 \\ 0 & (s-1)/(s+2) & 0 \end{bmatrix},$$

which is thus the Smith–McMillan form of $G(s)$. It is thus clear that $G(s)$ has the McMillan degree $n = 4$, poles at $\{-2, -2, -1, 1\}$ and a transmission zero at $\{1\}$. All poles and zeros except the pole at $s = -2$ have multiplicity 1. The pole at $s = -2$ has a geometric multiplicity of 2 and each of its partial multiplicities equals 1. The directions of the poles are

$$\mathrm{pdir_i}(G, 1) = \ker \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} [V(1)]' = \mathrm{span}\left(\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}\right), \quad \mathrm{pdir_o}(G, 1) = \ker \begin{bmatrix} 0 & 1 \end{bmatrix} U(1) = \mathrm{span}\left(\begin{bmatrix} 0 \\ 1 \end{bmatrix}\right),$$

$$\mathrm{pdir_i}(G, -1) = \ker \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} [V(-1)]' = \mathrm{span}\left(\begin{bmatrix} -1 \\ 0 \\ 2 \end{bmatrix}\right),$$

$$\mathrm{pdir_o}(G, -1) = \ker \begin{bmatrix} 0 & 1 \end{bmatrix} U(-1) = \mathrm{span}\left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}\right),$$

$$\mathrm{pdir_i}(G, -2) = \ker \begin{bmatrix} 0 & 0 & 1 \end{bmatrix} [V(-2)]' = \mathrm{span}\left(\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}\right), \quad \text{and} \quad \mathrm{pdir_o}(G, -2) = \mathbb{C}^2.$$

The directions of the zero, skipping details, are

$$\mathrm{zdir_i}(G, 1) = \mathrm{span}\left(\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}\right) \quad \text{and} \quad \mathrm{zdir_o}(G, 1) = \mathrm{span}\left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}\right).$$

Again, corresponding directions of the pole and the zero at $s = 1$ are orthogonal. ◊

Intuitively, we may expect that an input, whose spatial direction is orthogonal to the input direction of a pole of $G(s)$ at $s = s_i$, does not excite that pole. In other words, that if $u \perp \mathrm{pdir_i}(G, p_i)$, then the transfer function $G(s)u$ has no poles at $s = p_i$. Examples above seem to confirm that. For instance, $u = e_1$ is orthogonal to $\mathrm{pdir_i}(G, -2)$ for the transfer function $G(s)$ in Example 3.3. And indeed,

$$G(s)u = \begin{bmatrix} 1/(s+1) \\ -1/(s-1) \end{bmatrix}$$

has no poles at $s = -2$ in this case. However, this conclusion is only true if all partial multiplicities of p_i are 1. Otherwise, the situation is more involved, as can be seen from the example below.

Example 3.4. Consider the 2×2 transfer function

$$G(s) = \begin{bmatrix} 1/s & 1/s^2 \\ 0 & 1/s \end{bmatrix},$$

whose Smith–McMillan form is

$$\begin{bmatrix} 1 & 0 \\ -s & 1 \end{bmatrix} G(s) \begin{bmatrix} 0 & -1 \\ 1 & s \end{bmatrix} = \begin{bmatrix} 1/s^2 & 0 \\ 0 & 1 \end{bmatrix}.$$

Thus, it has a pole at the origin, whose geometric multiplicity is 1 and partial multiplicity is 2, and no transmission zeros. The directions of the pole, according to (3.38), are

$$\mathrm{pdir_i}(G, 0) = \ker \begin{bmatrix} 0 & 1 \end{bmatrix} [V(0)]' = \mathrm{span}\left(\begin{bmatrix} 0 \\ 1 \end{bmatrix}\right) \quad \text{and} \quad \mathrm{pdir_o}(G, 0) = \ker \begin{bmatrix} 0 & 1 \end{bmatrix} U(0) = \mathrm{span}\left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}\right).$$

Now, $e_1 \perp \mathrm{pdir_i}(G, 0)$, but

$$G(s)e_1 = \begin{bmatrix} 1/s \\ 0 \end{bmatrix}$$

still has a pole at the origin. However, this pole is single now, meaning that u does cancel one pole. This never happens if $u \notin \mathrm{span}(e_1)$, i.e. u is not orthogonal to $\mathrm{pdir_i}(G, 0)$. Indeed, all such vectors, modulo the multiplication by a scalar, are of the form $\left[\begin{smallmatrix} \alpha \\ 1 \end{smallmatrix}\right]$ for some $\alpha \neq 0$. In this case,

$$G(s)u = \begin{bmatrix} (\alpha s + 1)/s^2 \\ 1/s \end{bmatrix} = \begin{bmatrix} 1 & -\alpha \\ -s & \alpha s + 1 \end{bmatrix}^{-1} \begin{bmatrix} 1/s^2 \\ 0 \end{bmatrix} = \begin{bmatrix} \alpha s + 1 & \alpha \\ s & 1 \end{bmatrix} \begin{bmatrix} 1/s^2 \\ 0 \end{bmatrix}$$

always has a double pole at the origin. ◊

A problem with definitions expressed in terms of the Smith–McMillan form is that this form might be hard to compute, especially for high-dimensional transfer functions. Alternative approaches to compute the quantities above may thus be preferable. The use of the state-space representation of $G(s)$ studied in the next chapter yields numerically efficient and conceptually clear way to characterize poles, zeros, and degree of MIMO transfer functions, see Section 4.3.2. Still, simpler algorithms are available directly in terms of transfer functions, as well as when mild simplifying assumptions on poles and zeros are imposed. Below some of these results, which might help in hand calculations, are presented.

Proposition 3.6. *Let $G(s)$ be a transfer function with* $\mathrm{nrank}(G(s)) = r$. *The following statements hold true.*

1. *The polynomial $\phi_{\mathrm{p}}(s)$ in* (3.36) *is the least common denominator of all nonzero minors of $G(s)$ of all orders provided that all common poles and zeros in each of these minors were canceled.*

2. *The polynomial* $\phi_z(s)$ *in* (3.36) *is the greatest common divisor of all the numerators of all* r*-order minors of* $G(s)$ *provided that these minors have been adjusted to have* $\phi_p(s)$ *as their denominators.*

Example 3.5. Consider again the transfer function from Example 3.3. Its nonzero minors of order 1 are:

$$\frac{1}{s+1}, \quad \frac{s-1}{(s+1)(s+2)}, \quad -\frac{1}{s-1}, \quad \frac{1}{s+2}, \quad \text{and} \quad \frac{1}{s+2}$$

and the minors of order 2 are:

$$-\frac{s-1}{(s+1)(s+2)^2}, \quad \frac{2}{(s+1)(s+2)}, \quad \text{and} \quad \frac{1}{(s+1)(s+2)}.$$

The least common denominator of all these transfer functions is then

$$\phi_p(s) = (s+2)^2(s+1)(s-1) = (s+2)^2(s^2-1),$$

which corresponds to the McMillan degree 4 with the poles $\{-2, -2, -1, 1\}$, exactly as in Example 3.3. Now, to find the system zeros, rewrite the three minors of order 2 above in the form:

$$-\frac{(s-1)^2}{\phi_p(s)}, \quad \frac{2(s+2)(s-1)}{\phi_p(s)}, \quad \text{and} \quad \frac{(s+2)(s-1)}{\phi_p(s)}.$$

The common factor of their numerators is the zero polynomial

$$\phi_z(s) = s-1$$

corresponding to a single zero at $s = 1$, again in complete agreement with Example 3.3. ◊

The analysis of singular points of MIMO transfer functions is substantially simplified, both conceptually and computationally, if they are either poles or zeros, but not both. The following result can be derived directly from (3.38) and (3.39).

Proposition 3.7. *Let* $G(s)$ *be a* $p \times m$ *real-rational proper transfer function.*

1. *If* $z_i \in \mathbb{C}$ *is not a pole of* $G(s)$*, then it is a transmission zero of* $G(s)$ *iff* $\operatorname{rank}\big(G(z_i)\big) < \operatorname{nrank}\big(G(s)\big)$*. Furthermore,* $\operatorname{nrank}(G(s)) - \operatorname{rank}(G(z_i))$ *equals then the geometric multiplicity of the zero at* z_i *and* $\operatorname{zdir}_i(G, z_i) = \ker G(z_i)$ *and* $\operatorname{zdir}_o(G, z_i) = \ker[G(z_i)]'$*.*
2. *If* $p = m = \operatorname{nrank}(G(s))$ *and* $p_i \in \mathbb{C}$ *is not a transmission zero of* $G(s)$*, it is a pole of* $G(s)$ *iff* $\det\big(G^{-1}(p_i)\big) = 0$*. Furthermore,* $m - \operatorname{rank}(G^{-1}(p_i))$ *equals then the geometric multiplicity of the pole at* p_i *and* $\operatorname{pdir}_i(G, p_i) = \ker[G^{-1}(p_i)]'$ *and* $\operatorname{pdir}_o(G, p_i) = \ker G^{-1}(p_i)$*.*

The results of Proposition 3.7 might no longer be true in the case when $s_0 \in \mathbb{C}$ is both a pole and a zero of $G(s)$. For example, the transfer function in Example 3.2 is square and invertible, with

$$G^{-1}(s) = \begin{bmatrix} 1 & -1/s \\ 0 & 1 \end{bmatrix}.$$

But $\det(G^{-1}(s)) = 1$, i.e. is nonzero at the pole of $G(s)$ at $s = 0$, which also its transmission zero.

Remark 3.5 (poles of non-square systems). The result of the second item of Proposition 3.7 may be relevant in the analysis of general systems, not only square ones. For example, Proposition 3.3 effectively says that all unstable poles of a real-rational system are those of the inverse of a denominator of its coprime factorization. Denominators, in both *lcf*'s and *rcf*'s, are always square, always stable, and always of a full normal rank. Hence, it follows from Proposition 3.7 that all unstable poles of $G(s)$ are zeros of any of its denominators. ▽

3.A Discrete-time signals and systems

Although these notes are mostly devoted to continuous-time control systems, their discrete-time counterparts are simpler mathematically and more tangible in some respects. This appendix aims at shedding light on the underlying ideas behind the kernel and state-space representations of linear systems via systems operating on discrete signals. Because of the above-mentioned purpose, the exposition skips many aspects of discrete systems that do not help to understand corresponding properties of continuous-time systems.

3.A.1 Discrete-time signals in time domain

An n-dimensional discrete-time signal f is understood as

$$f : \mathbb{Z} \to \mathbb{F}^n \qquad \text{or} \qquad f : \mathbb{I} \to \mathbb{F}^n \quad \text{for some } \mathbb{I} \subset \mathbb{Z}.$$

Its value at a time instance t is denoted as $f[t]$, where square brackets aim at distinguishing discrete-time signals from their continuous-time counterparts (where parentheses are used for independent variables). Discrete counterparts of L_q signal spaces are

$$\ell_q^n(\mathbb{I}) := \Big\{ f : \mathbb{I} \to \mathbb{F}^n \,\Big|\, \|f\|_q := \Big(\sum_{t\in\mathbb{I}} \|f[t]\|_q^q\Big)^{1/q} < \infty \Big\} \tag{3.40}$$

(or simply $\ell_q(\mathbb{I})$ / ℓ_q), where $\|f\|_q$ is called the ℓ_q-norm of a signal f. Important particular cases are ℓ_1, ℓ_∞, and

$$\ell_2^n(\mathbb{I}) := \Big\{ f : \mathbb{I} \to \mathbb{F}^n \,\Big|\, \|f\|_2 := \Big(\sum_{t\in\mathbb{I}} \|f[t]\|^2\Big)^{1/2} < \infty \Big\}. \tag{3.41}$$

Also, we may need

$$\ell_{2+}^n(\mathbb{Z}) := \big\{ f \in \ell_2^n(\mathbb{Z}) \mid f[t] = 0 \text{ if } t < 0 \big\} \quad \text{and} \quad \ell_{2-}^n(\mathbb{Z}) := \big\{ f \in \ell_2^n(\mathbb{Z}) \mid f[t] = 0 \text{ if } t \geq 0 \big\}.$$

They are Hilbert spaces with the inner product

$$\langle f_1, f_2 \rangle_2 = \sum_{t\in\mathbb{I}} (f_2[t])' f_1[t].$$

The orthogonality notion and the relation $\ell_2(\mathbb{Z}) = \ell_{2+}(\mathbb{Z}) \oplus \ell_{2-}(\mathbb{Z})$ are the same as in continuous time.

Unlike continuous-time signal spaces, the spaces $\ell_q(\mathbb{I})$ may be finite dimensional. This happens if $|\mathbb{I}| < \infty$, where $|\mathbb{I}|$ stands for the cardinality (the number of elements) of the set $\mathbb{I}$. A possible orthonormal basis on such spaces is the sequence of Kronecker delta functions, $\{\delta[t]\}_{t\in\mathbb{I}}$. If $\mathbb{I}$ is infinite, the space $\ell_q(\mathbb{I})$ is infinite dimensional. Norms are then not equivalent, similarly to the continuous-time case.

3.A.2 Discrete-time systems, their kernel representation and system matrices

We first address discrete systems operating over finite time intervals, because the underlying signal spaces are finite dimensional. So let $\mathbb{I} \subset \mathbb{Z}$ be such that $|\mathbb{I}| < \infty$. Consider a bounded linear system (operator) $G : \ell_2^m(\mathbb{I}) \to \ell_2^p(\mathbb{I})$ and denote its input signal by u. It is readily seen that any $u \in \ell_2^m(\mathbb{I})$ can be presented in the form

$$u[t] = \sum_{s\in\mathbb{I}} \delta[t-s]u[s],$$

where δ is the Kronecker delta. This is effectively the decomposition of u on the basis of $\{\delta[t]\}$, cf. (A.3), albeit coordinates are vectors in $\mathbb{R}^m$ here. By linearity, the action of G on this superposition of pulses can be written as

$$y[t] = (Gu)[t] = \Big(G \sum_{s\in\mathbb{I}} \delta[\cdot - s]u[s]\Big)[t] = \sum_{s\in\mathbb{I}} (G\delta[\cdot - s])[t]u[s] = \sum_{s\in\mathbb{I}} g[t,s]u[s], \tag{3.42}$$

where the function

$$g[t,s] := (G\delta[\cdot - s])[t] : \mathbb{I}^2 \to \mathbb{R}^{p\times m},$$

called the *impulse response* of G, is the response of G, at the time instance t, to the unit pulse[2] applied at the time instance s. Form (3.42) is known as the *kernel representation* of G. For example, let $G_{\text{int}} : u \mapsto y$ be defined via $y[t] = y[k-1] + u[t]$. This system is known as the *integrator*. Its kernel $g_{\text{int}}[t,s] = \mathbb{1}[t-s]$, where $\mathbb{1}[t]$ is the discrete step and the kernel representation is $y[t] = \sum_{s=t_0}^{t} u[s]$, where t_0 denotes the first element of $\mathbb{I}$.

Extending the arguments above to the case when the interval $\mathbb{I}$ is unbounded, like $\mathbb{Z}$ or $\mathbb{Z}_+$ might be delicate. First, the class of operators $\ell_2(\mathbb{I}) \to \ell_2(\mathbb{I})$ is then more restrictive. For example, the integrator G_{int} does not belong to that class, which can be seen by applying the input $u = \delta \in \ell_2(\mathbb{Z})$, resulting in the output $y = \mathbb{1} \notin \ell_2(\mathbb{Z})$. Thus, we have to define linear systems as operators $G : \mathfrak{D}_G \subset \ell_2^m(\mathbb{I}) \to \ell_2^p(\mathbb{I})$ for some *domain* $\mathfrak{D}_G$. Returning to the integrator example, its domain is an (open) subspace of $\ell_2(\mathbb{I})$ comprising of functions u, which can be presented as $u[t] = v[t] - v[t-1]$ for an arbitrary $v \in \ell_2(\mathbb{Z})$. It follows from the triangle inequality that any such $u \in \ell_2(\mathbb{Z})$ and then $y[t] = v[t]$ is obviously an $\ell_2(\mathbb{Z})$ function. Second, even if defined over proper subspaces, the use of superposition arguments for infinite sums in (3.42) requires certain care with the convergence. Still, the class of systems that can be described by the kernel representation (3.42) is sufficient for all practical purposes[3]

A linear system G is said to be *stable* if $\mathfrak{D}_G = \ell_2^m(\mathbb{I})$ and $\|G\| := \sup_{\|u\|=1} \|Gu\| < \infty$. The norm defined by the last expression is called the *ℓ_2-induced norm* of G. The integrator is clearly *unstable* by this definition, which is intuitive. We say that a system $G : u \mapsto y$ is *causal* (*strictly causal*) if $y[t]$ may depend only on $u[s]$ for $s \le t$ ($s < t$). In terms of the kernel representation causality is equivalent to the condition that $g[t,s] = 0$ whenever $s > t$. Strict causality requires in addition that the *feedthrough terms* $g[t,t] = 0$ for all $t \in \mathbb{I}$. The integrator G_{int} is clearly causal, but not strictly causal.

Assuming $\mathbb{I} = \mathbb{Z}_{t_0..t_h}$, equation (3.42) can be equivalently presented in the matrix form

$$\underbrace{\begin{bmatrix} y[t_0] \\ y[t_0+1] \\ \vdots \\ y[t_1] \end{bmatrix}}_{[\![y]\!]} = \underbrace{\begin{bmatrix} g[t_0,t_0] & g[t_0,t_0+1] & \cdots & g[t_0,t_1] \\ g[t_0+1,t_0] & g[t_0+1,t_0+1] & \cdots & g[t_0+1,t_1] \\ \vdots & \vdots & & \vdots \\ g[t_1,t_0] & g[t_1,t_0+1] & \cdots & g[t_1,t_1] \end{bmatrix}}_{[\![G]\!]} \underbrace{\begin{bmatrix} u[t_0] \\ u[t_0+1] \\ \vdots \\ u[t_1] \end{bmatrix}}_{[\![u]\!]}. \tag{3.43}$$

The matrix $[\![G]\!]$ built on the impulse responses as in (3.43) is called the *system matrix* of G and is the matrix representation of G in the orthonormal basis $\{\delta[t]\}$ of $\ell_2(\mathbb{I})$, see §A.2.4. Note that we use the same notation for stacked vectors and system matrices; the meaning, as well as the interval, are normally clear from the context. Causal systems have block lower-triangular system matrices, with zero gray blocks.

System matrices and stacked signals for systems acting on infinite horizons are infinite. For example, if $\mathbb{I} = \mathbb{Z}$, the relation $[\![y]\!] = [\![G]\!][\![u]\!]$ reads

$$\begin{bmatrix} \vdots \\ y[-2] \\ y[-1] \\ \hline y[0] \\ y[1] \\ \vdots \end{bmatrix} = \left[\begin{array}{cccc:ccc} & \vdots & \vdots & & \vdots & \vdots & \\ \cdots & g[-2,-2] & g[-2,-1] & & g[-2,0] & g[-2,1] & \cdots \\ \cdots & g[-1,-2] & g[-1,-1] & & g[-1,0] & g[-1,1] & \cdots \\ \hdashline \cdots & g[0,-2] & g[0,-1] & & g[0,0] & g[0,1] & \cdots \\ \cdots & g[1,-2] & g[1,-1] & & g[1,0] & g[1,1] & \cdots \\ & \vdots & \vdots & & \vdots & \vdots & \end{array}\right] \begin{bmatrix} \vdots \\ u[-2] \\ u[-1] \\ \hline u[0] \\ u[1] \\ \vdots \end{bmatrix}. \tag{3.44}$$

[2]More precisely, the jth column of $g[t,s]$ is the response of G, at the time instance t, to the input $e_j\delta$ applied at the time instance s, where e_j is the jth standard basis on $\mathbb{R}^m$.

[3]See [25] for a more general representation and more detailed explanations.

It is worth emphasizing that doubly-infinite matrices as above are qualitatively different from their finite (and even semi-infinite) counterparts. To demonstrate that, consider the *backward (right) shift* operator $S : \ell_2^m(\mathbb{I}) \to \ell_2^m(\mathbb{I})$, which is defined as $(Su)[t] = u[t-1]$, for all $t \in \mathbb{I}$ (if $\mathbb{I}$ is left-bounded, by say t_0, then it is assumed that $(Su)[t_0] = 0$). Its impulse response is $s[t,r] = \delta[t-r-1]I_m$. If $\mathbb{I} = \mathbb{Z}_{t_0..t_1}$, the system matrix of the backward shift operator,

$$[\![S]\!] = \begin{bmatrix} 0 & \cdots & 0 & 0 \\ I & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & I & 0 \end{bmatrix},$$

is singular. At the same time, if considered as an operator on $\ell_2(\mathbb{Z})$, the system matrix of S is unitary,

$$[\![S]\!]^{-1} = \left[\begin{array}{cc:cccc} & \vdots\ \vdots & \vdots & \vdots & & \\ \cdots & 0\ \ 0 & 0 & 0 & \cdots \\ \cdots & I\ \ 0 & 0 & 0 & \cdots \\ \hdashline \cdots & 0\ \ I & 0 & 0 & \cdots \\ \cdots & 0\ \ 0 & I & 0 & \cdots \\ & \vdots\ \vdots & \vdots & \vdots & \end{array}\right]^{-1} = \left[\begin{array}{cc:cccc} & \vdots\ \vdots & \vdots & \vdots & & \\ \cdots & 0\ \ I & 0 & 0 & \cdots \\ \cdots & 0\ \ 0 & I & 0 & \cdots \\ \hdashline \cdots & 0\ \ 0 & 0 & I & \cdots \\ \cdots & 0\ \ 0 & 0 & 0 & \cdots \\ & \vdots\ \vdots & \vdots & \vdots & \end{array}\right] = [\![S]\!]' = [\![\tilde{S}]\!],$$

where $\tilde{S}$ is the *forward (left) shift* operator acting as $(\tilde{S}u)[t] = u[t+1]$ (if $\mathbb{I}$ is right-bounded, by say t_1, then it is assumed that $(\tilde{S}u)[t_1] = 0$). In fact, $[\![S]\!]$ is singular even if considered on $\mathbb{I} = \mathbb{Z}_+$, in which case it is injective (i.e. $\ker S = \{0\}$), but not surjective (i.e. $\operatorname{Im} S \neq \ell_2(\mathbb{Z}_+)$).

Manipulations with discrete-time systems can be carried out in terms of their system matrices. For example, it follows directly from (3.43) (or from (3.44)) that the cascade parallel interconnections of compatibly dimensioned systems G_1 and G_2 satisfy $[\![G_2G_1]\!] = [\![G_2]\!][\![G_1]\!]$ and $[\![G_1 + G_2]\!] = [\![G_1]\!] + [\![G_2]\!]$. Likewise, $[\![G^{-1}]\!] = [\![G]\!]^{-1}$. It should also be clear from the definition that $\|x\| = \|[\![x]\!]\|$, i.e. the ℓ_2-norm of a signal x equals the Euclidean norm of its stacked vector. As such, $\|G\| = \|[\![G]\!]\|$, i.e. the ℓ_2-induced norm of a system equals the spectral norm of its system matrix.

A linear discrete system G on $\ell_2(\mathbb{Z})$ is called *shift invariant* (abbreviated LSI) if $GS = SG$. Similarly to the continuous-time time invariance, this definition effectively says that a delayed input produces a delayed, but otherwise unchanged, output. In terms of the kernel representation (3.42), this reads as

$$(GS)[t] = \sum_{s\in\mathbb{Z}} g[t,s]u[s-1] = \sum_{s\in\mathbb{Z}} g[t,s+1]u[s] = \sum_{s\in\mathbb{Z}} g[t-1,s]u[s] = (SG)[t].$$

Hence, shift invariance is equivalent to the kernel constraint $g[t,s] = g[t-1,s-1]$, which should hold for all $t,s \in \mathbb{Z}$. This, in turn, implies that the impulse response of an LSI system satisfies

$$g[t,s] = g[t-1,s-1] = g[t-2,s-2] = \cdots = g[t-s,0].$$

Thus, LSI systems are completely characterized by their response to the input impulse applied at the time instance $s = 0$, denoted $g[t] := g[t,0]$. The matrices $g[t] \in \mathbb{R}^{p\times m}$ are also known as the *Markov parameters* of G. With this convention, (3.42) can be presented as the *convolution sum*

$$y[t] = \sum_{s\in\mathbb{Z}} g[t-s]u[s]. \tag{3.45}$$

The backward shift operator S is obviously shift invariant, by the very definition, which can also be seen via the fact that its impulse response has $s[t,r] = \delta[t-r-1]$, i.e. is a function of $t-r$. The integrator G_{int}

defined above, whose impulse response satisfies $g_{\text{int}}[t,s] = \mathbb{1}[t-s]$, is clearly shift invariant too. System matrices of LSI systems have the block Toeplitz structure, like

$$\begin{bmatrix} \vdots \\ y[-2] \\ y[-1] \\ \hdashline y[0] \\ y[1] \\ \vdots \end{bmatrix} = \left[\begin{array}{ccc:ccc} & \vdots & \vdots & \vdots & \vdots & \\ \cdots & g[0] & g[-1] & g[-2] & g[-3] & \cdots \\ \cdots & g[1] & g[0] & g[-1] & g[-2] & \cdots \\ \hdashline \cdots & g[2] & g[1] & g[0] & g[-1] & \cdots \\ \cdots & g[3] & g[2] & g[1] & g[0] & \cdots \\ & \vdots & \vdots & \vdots & \vdots & \end{array}\right] \begin{bmatrix} \vdots \\ u[-2] \\ u[-1] \\ \hdashline u[0] \\ u[1] \\ \vdots \end{bmatrix}, \tag{3.46}$$

where elements along each block diagonal are equivalent. An LSI system G is causal iff $g[t] = 0$ whenever $t < 0$ and strictly causal iff $g[t] = 0$ whenever $t \leq 0$. Causal systems have upper block-triangular system matrices.

3.A.3 State space representation of finite-dimensional LSI systems

The notion of *state* plays an important role in systems theory, facilitating an economical representation of dynamical systems, their efficient implementation, elegant solutions to many control design problems, et cetera. The state variable appears naturally in many applications, like the "position-velocity" pair for the mechanical system in Fig. 1.2(a) or the "charge-current" pair for the RLC circuit in Fig. 1.2(b), and is conventionally introduced as such. Still, in many situations a state might have no tangible physical meaning. The goal of this section is to introduce the notion on an abstract, system-theoretic, level from scratch.

A key, and highly nontrivial, property of many systems of interest is that the relation between the past and the future has relatively low complexity. Specifically, consider a *causal shift-invariant* system[4] G on $\ell_2(\mathbb{Z})$ and fix some time instance t_c. It follows from (3.46) that the outputs $y[t]$ for $t \geq t_c$ satisfy

$$\begin{bmatrix} y[t_c] \\ y[t_c+1] \\ y[t_c+2] \\ \vdots \end{bmatrix} = \underbrace{\begin{bmatrix} g[1] & g[2] & g[3] & \cdots \\ g[2] & g[3] & g[4] & \cdots \\ g[3] & g[4] & g[5] & \cdots \\ \vdots & \vdots & \vdots & \end{bmatrix}}_{[\![\mathfrak{H}_G]\!]} \begin{bmatrix} u[t_c-1] \\ u[t_c-2] \\ u[t_c-3] \\ \vdots \end{bmatrix} + \underbrace{\begin{bmatrix} g[0] & 0 & 0 & \cdots \\ g[1] & g[0] & 0 & \cdots \\ g[2] & g[1] & g[0] & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}}_{[\![\mathfrak{T}_G]\!]} \begin{bmatrix} u[t_c] \\ u[t_c+1] \\ u[t_c+2] \\ \vdots \end{bmatrix}. \tag{3.47}$$

The *Hankel operator* $\mathfrak{H}_G$ associated with G connects the past inputs with the present and future inputs at each time instance. It turns out that there is a large class of systems (perhaps, all systems whose response can be calculated) for which the rank of $\mathfrak{H}_G$, and therefore the rank of the infinite-dimensional matrix $[\![\mathfrak{H}_G]\!]$, are finite. This class is called *finite-dimensional systems* and the rank of the Hankel operator, say n, is called the *state dimension* or the *order* of the system G. In this case the system matrix of the corresponding Hankel operator can be factorized as

$$[\![\mathfrak{H}_G]\!] = \begin{bmatrix} O_1 \\ O_2 \\ O_3 \\ \vdots \end{bmatrix} \begin{bmatrix} R_1 & R_2 & R_3 & \cdots \end{bmatrix} =: O_G R_G \tag{3.48}$$

($O_i \in \mathbb{R}^{p\times n}$ and $R_i \in \mathbb{R}^{n\times m}$ for each $i \in \mathbb{N}$), where O_G and R_G are rank-n matrices having n columns and n rows, respectively, cf. the full rank decomposition in (2.17). Matrices O_i and R_i are unique up to the post-multiplication by a nonsingular matrix $M \in \mathbb{R}^{n\times n}$ and the pre-multiplication by M^{-1}, respectively.

[4]The arguments can be extended to non-causal and time-varying systems, but this would render the notation bulky, see [7].

Example 3.6. Consider an LTI causal system G having the impulse response $g[t] = \alpha^{t-1}\mathbb{1}[t-1]$ for some $\alpha \in \mathbb{R}$. Using the convention that $\alpha^0 = 1$ for all $\alpha \in \mathbb{R}$, the case $\alpha = 0$ corresponds to $G = S$. The case $\alpha = 1$ then yields the delayed integrator, $G = G_{\text{int}}S$. Whatever α is, the corresponding Hankel matrix,

$$[\![\mathfrak{H}_G]\!] = \begin{bmatrix} 1 & \alpha & \alpha^2 & \cdots \\ \alpha & \alpha^2 & \alpha^3 & \cdots \\ \alpha^2 & \alpha^3 & \alpha^4 & \cdots \\ \vdots & \vdots & \vdots & \end{bmatrix} = \begin{bmatrix} 1 \\ \alpha \\ \alpha^2 \\ \vdots \end{bmatrix} \begin{bmatrix} 1 & \alpha & \alpha^2 & \cdots \end{bmatrix},$$

has rank 1 and is indeed in form (3.48). ◊

The n-dimensional vector

$$x[t_c] := R_G \begin{bmatrix} u[t_c-1] \\ u[t_c-2] \\ u[t_c-3] \\ \vdots \end{bmatrix} = \sum_{s=-\infty}^{t_c-1} R_{t_c-s}u[s] \tag{3.49}$$

is then called the *state vector* of G at the time instance t_c and is unique up to the multiplication by a nonsingular matrix $M \in \mathbb{R}^{n\times n}$. The state vector may be thought of as a *history accumulator*. Indeed, by (3.47)–(3.49),

$$\begin{bmatrix} y[t_c] \\ y[t_c+1] \\ y[t_c+2] \\ \vdots \end{bmatrix} = O_G x[t_c] + [\![\mathfrak{T}_G]\!] \begin{bmatrix} u[t_c] \\ u[t_c+1] \\ u[t_c+2] \\ \vdots \end{bmatrix},$$

so the knowledge of $x[t_c]$ is sufficient to account for the effect of the whole input history up to t_c on future outputs. This suggests that causal systems can be treated as operators on $\ell_2(\mathbb{Z}_+)$, rather than on $\ell_2(\mathbb{Z})$, with the *initial condition* $x[0]$ explaining the effect of prehistoric inputs ($t = 0$ is chosen as a starting point merely by convenience). Having a starting point may be intuitive in control applications, this is the time moment where the control law is applied.

It can be shown[5] that (3.48) and the Hankel structure of $[\![\mathfrak{H}_G]\!]$ guarantee the existence of matrices $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, and $C \in \mathbb{R}^{p\times n}$ such that

$$O_G = \begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \end{bmatrix} \quad \text{and} \quad R_G = \begin{bmatrix} B & AB & A^2B & \cdots \end{bmatrix}$$

(infinite observability and reachability matrices). A consequence of this structure is that the state vector can be propagated recursively. To see this, consider

$$\begin{aligned} x[t+1] &= \begin{bmatrix} B & AB & A^2B & \cdots \end{bmatrix} \begin{bmatrix} u[t] \\ u[t-1] \\ u[t-2] \\ \vdots \end{bmatrix} = A \begin{bmatrix} B & AB & A^2B & \cdots \end{bmatrix} \begin{bmatrix} u[t-1] \\ u[t-2] \\ u[t-3] \\ \vdots \end{bmatrix} + Bu[t] \\ &= Ax[t] + Bu[t]. \end{aligned}$$

[5] See [3, Sec.4.4] for details on the realization theory.

Furthermore, the output at every time instance is a linear function of the current state and input. Indeed, because $g[t] = CA^{t-1}B$ for all $t \in \mathbb{N}$, equality (3.45) reads

$$y[t] = C\sum_{s=-\infty}^{t-1} A^{t-s-1}Bu[s] + g[0]u[t] = Cx[t] + g[0]u[t].$$

Thus, if only the response at nonnegative time instances is of interest, any linear causal shift-invariant n-order $\ell_2(\mathbb{Z})$ system G can be viewed as an operator $G : \mathfrak{D}_G \subset \mathbb{R}^n \times \ell_2^m(\mathbb{Z}_+) \to \ell_2^p(\mathbb{Z}_+)$ and described in terms of four matrices A, B, C, and $D := g[0] \in \mathbb{R}^{p\times m}$ and an initial condition vector $x_0 \in \mathbb{R}^n$ by the following recursion:

$$\begin{cases} x[t+1] = Ax[t] + Bu[t], \quad x[0] = x_0 \\ \quad y[t] = Cx[t] + Du[t]. \end{cases} \tag{3.50}$$

Description (3.50) is called the *state representation* of G and the quadruple (A, B, C, D) is called the *state-space realization* of G. We frequently assume that the initial condition is zero, i.e. that $x[0] = 0$, which may take place in many situations when the system starts from an equilibrium and the linear model is written in terms of deviation variables. In this case we may consider causal LSI systems as operators $G : \mathfrak{D}_G \subset \ell_2(\mathbb{Z}_+) \to \ell_2(\mathbb{Z}_+)$.

Chapter 4

State-Space Techniques for LTI Systems

THE STATE-SPACE DESCRIPTION is a conceptually convenient and well-suited for numerical manipulations way to represent finite-dimensional systems. It leads to elegant structural properties (like the separation principle) of controllers and equally suited for description of both SISO and MIMO systems. This chapter studies basic properties of state-space realizations for LTI continuous-time systems.

4.1 Basic definitions and properties

Like in the discrete-time case introduced in §3.A.3, the notion of state for continuous-time systems is associated with the *Hankel operator* $\mathfrak{H}_G : \mathfrak{D}_{\mathfrak{H}_G} \subset L_2(-\infty, t_c) \to L_2(t_c, \infty)$, which connects past (with respect to any given t_c) inputs of a causal G and its future outputs as follows:

$$y(t) = \int_{-\infty}^{t_c} g(t-s)u(s)\,\mathrm{d}s, \quad \forall t > t_c.$$

If the rank of the infinite-dimensional Hankel operator is finite, say n, the system is said to be *finite dimensional* and n is its dimension (order). Note that $\mathfrak{H}_G$ might have a finite rank only if $g_i = 0$ for all $i > 0$ in (3.13). As causality requires that $g_i = 0$ for $i < 0$ as well, we may only have $g_0 \neq 0$ for causal finite-dimensional systems. But the $g_0\delta(t)$ term, which is responsible for the instantaneous connection between $u(t)$ and $y(t)$, is not a part of the Hankel operator. If the Hankel operator has rank n, its kernel can be factorized [8, Ch. IX] as $\tilde{g}(t-s) = \tilde{g}_{\mathrm{O}}(t)\tilde{g}_{\mathrm{R}}(-s)$ for some $\tilde{g}_{\mathrm{O}} : \mathbb{R} \to \mathbb{R}^{p\times n}$ and $\tilde{g}_{\mathrm{R}} : \mathbb{R} \to \mathbb{R}^{n\times m}$. The output of $\mathfrak{H}_G$ can then be rewritten as

$$y(t) = \tilde{g}_{\mathrm{O}}(t-t_c)\int_{-\infty}^{t_c} \tilde{g}_{\mathrm{R}}(t_c-s)u(s)\,\mathrm{d}s + g_0u(t) = \tilde{g}_{\mathrm{O}}(t-t_c)x(t_c) + g_0u(t), \quad \forall t > t_c$$

where the n-dimensional vector

$$x(t) := \int_{-\infty}^{t} \tilde{g}_{\mathrm{R}}(t-s)u(s)\,\mathrm{d}s \tag{4.1}$$

is called the *state vector* of G at time t. The state vector may be interpreted as the history accumulator. The latter can be seen by rewriting the mapping $G : u \mapsto y$ given by (3.16) for $t \geq t_c$ as

$$y(t) = \int_{\mathbb{R}} \tilde{g}(t-s)u(s)\,\mathrm{d}s + g_0u(t) = \tilde{g}_{\mathrm{O}}(t-t_c)x(t_c) + \int_{t_c}^{t} \tilde{g}(t-s)u(s)\,\mathrm{d}s + g_0u(t) \tag{4.2}$$

which shows that the only information about the input history up to the "starting point" $t = t_c$ that we need is the state vector $x(t_c)$. This fact facilitates the treatment of causal LTI systems as operators on the non-negative semi-axis $\mathbb{R}_+$.

The kernel of any rank-n Hankel operator associated with an LTI system at a time instance t is of the form $\tilde{g}(t) = C\mathrm{e}^{At}B$ for some matrices $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, and $C \in \mathbb{R}^{p\times n}$. In this case, we can choose $\tilde{g}_{\mathrm{O}}(t) = C\mathrm{e}^{At}$ and $\tilde{g}_{\mathrm{R}}(s) = \mathrm{e}^{As}B$ and then

$$\dot{x}(t) = \frac{\mathrm{d}}{\mathrm{d}t}\int_{-\infty}^{t} \mathrm{e}^{A(t-s)}Bu(s)\mathrm{d}s = A\int_{-\infty}^{t} \mathrm{e}^{A(t-s)}Bu(s)\mathrm{d}s + \mathrm{e}^{A(t-s)}Bu(s)\big|_{s=t} = Ax(t) + Bu(t).$$

Moreover, it then follows from (4.2) that $y(t) = Cx(t) + Du(t)$, where $D := g_0 \in \mathbb{R}^{p\times m}$. Thus, any finite-dimensional causal LTI G can be described as an operator $\mathfrak{D}_G \subset \mathbb{R}^n \times L_2^m(\mathbb{R}_+) \to L_2^p(\mathbb{R}_+)$ by the following set of equations:

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t), \quad x(0) = x_0 \\ y(t) = Cx(t) + Du(t). \end{cases} \tag{4.3}$$

Description (4.3) is called the *state representation* of G and the quadruple (A, B, C, D) is called its *state-space realization*. We frequently assume that the initial condition is zero, i.e. that $x(0) = 0$, which may take place in many situations when the system starts from an equilibrium and the linear model is written in terms of deviation variables. We may then consider systems as operators $\mathfrak{D}_G \subset L_{2+}^m \to L_{2+}^p$. The variable $\tilde{x} = Tx$ for any nonsingular matrix $T \in \mathbb{R}^{n\times n}$ is also a legitimate state vector, whose realization $(TAT^{-1}, TB, CT^{-1}, D)$ is dubbed *similar* to (A, B, C, D), with $\tilde{x}(0) = Tx_0$.

It follows from the arguments preceding (4.3) that the impulse response of a finite-dimensional LTI G in terms of its state-space realization has

$$g(t) = D\delta(t) + C\mathrm{e}^{At}B\mathbb{1}(t), \tag{4.4}$$

which can also be derived from the solution of (4.3). By the time-differentiation property of the Laplace transform, or transforming $g(t)$ in (4.4) directly, the transfer function of an LTI G in terms of its state space realization is

$$G(s) = D + C(sI - A)^{-1}B =: \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right] \qquad \left(\text{meaning } \begin{array}{c} \\ sx \leftarrow \\ y \leftarrow \end{array} \begin{array}{c} \begin{array}{cc} x & u \\ \downarrow & \downarrow \end{array} \\ \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right] \end{array}\right) \tag{4.5}$$

and it is real-rational and proper. Both the impulse response and the transfer function are invariant under similarity transformation, i.e. for every nonsingular T

$$D\delta(t) + CT^{-1}\mathrm{e}^{TAT^{-1}t}TB\mathbb{1}(t) = D\delta(t) + C\mathrm{e}^{At}B\mathbb{1}(t) \quad \text{and} \quad \left[\begin{array}{c|c} TAT^{-1} & TB \\ \hline CT^{-1} & D \end{array}\right] = \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right].$$

It is readily seen that $G(\infty) = D$, so that $G(s)$ is strictly proper iff $D = 0$ and a square $G(s)$ is bi-proper iff $\det(D) \neq 0$.

4.1.1 Operations on transfer functions in terms of state-space realizations

One of advantages of the analyzing LTI dynamical systems in the Laplace domain is that algebraic manipulations over transfer functions are performed in the same manner as corresponding manipulations over static matrices. Addition, multiplication, inverse, et cetera are carried out s-wise, whereas corresponding manipulations in terms of the impulse responses are substantially less transparent. In this subsection we shall see that manipulations over transfer functions can be efficiently and transparently performed in terms of their state-space realizations. Toward this end, we make use of the time-domain equations in (4.3) and signal flow relations between operands. Below, some expressions are derived for transfer functions

$$G_1(s) = \left[\begin{array}{c|c} A_1 & B_1 \\ \hline C_1 & D_1 \end{array}\right] \quad \text{and} \quad G_2(s) = \left[\begin{array}{c|c} A_2 & B_2 \\ \hline C_2 & D_2 \end{array}\right]$$

of systems $G_1 : u_1 \mapsto y_1$ and $G_2 : u_2 \mapsto y_2$, which can be equivalently be described as

$$G_1 : \begin{cases} \dot{x}_1(t) = A_1 x_1(t) + B_1 u_1(t) \\ y_1(t) = C_1 x_1(t) + D_1 u_1(t) \end{cases} \quad \text{and} \quad G_2 : \begin{cases} \dot{x}_2(t) = A_2 x_2(t) + B_2 u_2(t) \\ y_2(t) = C_2 x_2(t) + D_2 u_2(t) \end{cases}$$

in terms of their input, output, and state vectors (below zero initial conditions are assumed, because this is how transfer functions are defined).

Addition (parallel interconnection) The sum $G(s) = G_1(s) + G_2(s)$ can be interpreted as the transfer function of the parallel interconnection $G : u \mapsto y$ of G_1 and G_2. This reads $u_1 = u_2 = u$ and $y = y_1 + y_2$. The state-space equations then can be united to result in

$$G_1 + G_2 : \begin{cases} \begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} u(t) \\ y(t) = \begin{bmatrix} C_1 & C_2 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + (D_1 + D_2) u(t) \end{cases}$$

or, equivalently, in the formula

$$\left[\begin{array}{c|c} A_1 & B_1 \\ \hline C_1 & D_1 \end{array}\right] + \left[\begin{array}{c|c} A_2 & B_2 \\ \hline C_2 & D_2 \end{array}\right] = \left[\begin{array}{cc|c} A_1 & 0 & B_1 \\ 0 & A_2 & B_2 \\ \hline C_1 & C_2 & D_1 + D_2 \end{array}\right]. \tag{4.6}$$

Multiplication (cascade interconnection) The product $G(s) = G_2(s)G_1(s)$ can be seen as the transfer function of the cascade interconnection of G_1 and G_2. In other words, $u_1 = u$, $u_2 = y_1$, and $y = y_2$. This defines the state-space equation

$$G_2 G_1 : \begin{cases} \begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} A_1 & 0 \\ B_2 C_1 & A_2 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} B_1 \\ B_2 D_1 \end{bmatrix} u(t) \\ y(t) = \begin{bmatrix} D_2 C_1 & C_2 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + D_2 D_1 u(t) \end{cases}$$

or, equivalently, the formula

$$\left[\begin{array}{c|c} A_2 & B_2 \\ \hline C_2 & D_2 \end{array}\right] \left[\begin{array}{c|c} A_1 & B_1 \\ \hline C_1 & D_1 \end{array}\right] = \left[\begin{array}{cc|c} A_1 & 0 & B_1 \\ B_2 C_1 & A_2 & B_2 D_1 \\ \hline D_2 C_1 & C_2 & D_2 D_1 \end{array}\right] = \left[\begin{array}{cc|c} A_2 & B_2 C_1 & B_2 D_1 \\ 0 & A_1 & B_1 \\ \hline C_2 & D_2 C_1 & D_2 D_1 \end{array}\right], \tag{4.7}$$

where the last expression is obtained by merely swapping x_1 and x_2.

Inverse If G defines the relation $y = Gu$, its inverse should define the relation $u = G^{-1} y$. Clearly, $G^{-1}(s)$ is proper iff $G(s)$ is bi-proper, i.e. iff $\det(D) \neq 0$. In this case, the output equation in (4.3) rewrites as $u(t) = D^{-1} y(t) - D^{-1} C x(t)$. Substituting this equality into the state equation yields

$$G^{-1} : \begin{cases} \dot{x}(t) = A x(t) + B D^{-1} y(t) - B D^{-1} C x(t) = (A - B D^{-1} C) x(t) + B D^{-1} y(t) \\ u(t) = -D^{-1} C x(t) + D^{-1} y(t). \end{cases}$$

In other words,

$$\left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right]^{-1} = \left[\begin{array}{c|c} A - B D^{-1} C & B D^{-1} \\ \hline -D^{-1} C & D^{-1} \end{array}\right] \tag{4.8}$$

and it exists iff $\det(D) \neq 0$.

Partial fraction expansion The problem of splitting $G_2(s)G_1(s) = H_1(s) + H_2(s)$ for some $H_1(s)$ and $H_2(s)$ such that the "A" matrices of H_i match those of G_i, $i = 1, 2$, can be thought of as one step in the MIMO version of the partial fraction expansion procedure. Taking into account (4.7) and (4.6), this problem boils down to transforming a block-triangular matrix to a block-diagonal one. To this end the Roth's removal rule (see Proposition B.1) can be used. Namely, there is a similarity transformation between the "A" matrices in (4.7) and (4.6) iff the Sylvester equation

$$XA_1 - A_2X + B_2C_1 = 0$$

is solvable in X. A sufficient condition for the latter (cf. Section B.1, p. 189) is $\operatorname{spec}(A_1) \cap \operatorname{spec}(A_2) = \varnothing$, which could be expected from the theory of partial fraction decomposition. Anyhow, if a required X exists, the similarity transformation applied to the last realization in (4.7) is $T = \left[\begin{smallmatrix} I & X \\ 0 & I \end{smallmatrix}\right]$ and it yields

$$\left[\begin{array}{cc|c} A_2 & B_2C_1 & B_2D_1 \\ 0 & A_1 & B_1 \\ \hline C_2 & D_2C_1 & D_2D_1 \end{array}\right] = \left[\begin{array}{cc|c} A_2 & 0 & B_2D_1 + XB_1 \\ 0 & A_1 & B_1 \\ \hline C_2 & D_2C_1 - C_2X & D_2D_1 \end{array}\right],$$

so

$$\left[\begin{array}{c|c} A_2 & B_2 \\ \hline C_2 & D_2 \end{array}\right]\left[\begin{array}{c|c} A_1 & B_1 \\ \hline C_1 & D_1 \end{array}\right] = \left[\begin{array}{c|c} A_1 & B_1 \\ \hline D_2C_1 - C_2X & 0 \end{array}\right] + \left[\begin{array}{c|c} A_2 & B_2D_1 + XB_1 \\ \hline C_2 & 0 \end{array}\right] + D_2D_1 \tag{4.9}$$

(the feedthrough term D_2D_1 can be a part of either of the dynamic terms on the right-hand side).

4.2 Structural properties

In this section the controllability, observability and some other relevant state-space concepts will be reviewed. These notions play a central role in the state-space control theory. Because the notes are concerned mainly with frequency-domain methods, the state-space formalism is essentially exploited as a tool (analytical and computational) for dealing with transfer functions. For this reason, controllability and observability are studied here as structural properties of state-space realizations of transfer functions rather than the ability to steer the state to an arbitrary position by the input or to reconstruct the whole state vector via measuring the output.

4.2.1 Controllability and stabilizability

We say that the matrix pair $(A, B) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times m}$ is *controllable* if the eigenvalues of $A + BK$ can be freely assigned by a suitable choice of $K \in \mathbb{R}^{m\times n}$, with the obvious restriction that complex eigenvalues of $A + BK$ are in conjugate pairs ('K' is for Kalman here). Otherwise (A, B) is said to be *uncontrollable*. Controllability can be analyzed by numerous methods, some of them are presented in the following result.

Theorem 4.1. *The following statements are equivalent:*

1. *The pair (A, B) is controllable.*
2. *The matrix $\left[\begin{array}{cc} A - sI & B \end{array}\right]$ has full rank $\forall s \in \mathbb{C}$ (the PBH [Popov–Belevich–Hautus] test).*
3. *The matrix*
$$W_c(t) := \int_0^t \mathrm{e}^{As} BB' \mathrm{e}^{A's} \,\mathrm{d}s$$
is positive definite for all $t > 0$ (the Gramian-based test).
4. *The controllability matrix*
$$M_c := \left[\begin{array}{cccc} B & AB & \dots & A^{n-1}B \end{array}\right]$$
has full rank (i.e. $\operatorname{rank}(M_c) = n$).

To prove this theorem, we need a couple of technical results, which are important in their own right.

Lemma 4.2. $\operatorname{Im} W_c(t) = \operatorname{Im} M_c$ *for every* $t > 0$.

Proof. By Proposition 2.1 on p. 29 and the fact that $W_c(t)$ is symmetric, the statement of the lemma is equivalent to the condition that $\ker W_c(t) = \ker M_c'$. It is readily seen that $W_c(t) \geq 0$ for all t. Hence,

$$W_c(t)\eta = 0 \iff \eta' W_c(t)\eta = 0 \iff \int_0^t \|\eta' e^{As} B\|^2 \, ds = 0 \iff \eta' e^{As} B = 0, \quad \forall s \in [0, t].$$

Because e^{As} is analytic, the latter condition is equivalent to $\eta'(e^{As})^{(i)} B|_{s=0} = 0$ for all $i \in \mathbb{Z}_+$. Thus, $\eta \in \ker W_c(t)$ for some $t \in \mathbb{R}$ iff $\eta' \begin{bmatrix} B & AB & A^2B & \cdots \end{bmatrix} = 0$.

Now, it follows by the Cayley–Hamilton theorem that A^{n+j} is a linear combination of A^i, $i \in \mathbb{Z}_{0..n-1}$, for all $j \in \mathbb{Z}_+$. Hence, $\ker \begin{bmatrix} B & AB & A^2B & \cdots \end{bmatrix}' = \ker M_c'$. This completes the proof. □

Lemma 4.3. *If* $\operatorname{rank} W_c(t) = r < n$, *then there is a unitary matrix* U_c *such that*[1]

$$(U_c A U_c', U_c B) = \left(\begin{bmatrix} A_c & \times \\ 0 & A_{\bar{c}} \end{bmatrix}, \begin{bmatrix} B_c \\ 0 \end{bmatrix} \right),$$

where the pair $(A_c, B_c) \in \mathbb{R}^{r \times r} \times \mathbb{R}^{r \times m}$ *is such that* $\int_0^t e^{A_c s} B_c B_c' e^{A_c' s} \, ds > 0$.

Proof. Suppose that A is Hurwitz. It follows from the proof of Lemma 4.2 that $\ker W_c(t)$ is independent of t. Thus, it is sufficient to prove the result for $P := W_c(\infty) \geq 0$, which satisfies the Lyapunov equation

$$AP + PA' + BB' = 0 \tag{4.10}$$

(see Section B.1) and referred to as the *controllability Gramian* of (A, B). If $\operatorname{rank} P = r < m$, there is a unitary U_c such that $U_c P U_c' = \left[\begin{smallmatrix} P_c & 0 \\ 0 & 0 \end{smallmatrix}\right]$ for some $r \times r$ matrix $P_c > 0$. Now,

$$(U_c A U_c', U_c B) = \left(\begin{bmatrix} A_c & A_{12} \\ A_{21} & A_{\bar{c}} \end{bmatrix}, \begin{bmatrix} B_c \\ B_2 \end{bmatrix} \right),$$

where the partition is compatible to that of $U_c P U_c'$. The Lyapunov equation (4.10) reads then

$$\begin{bmatrix} A_c P_c + P_c A_c' + B_c B_c' & P_c A_{21}' + B_c B_2' \\ A_{21} P_c + B_2 B_c' & B_2 B_2' \end{bmatrix} = 0.$$

Its $(2, 2)$ entry yields $B_2 = 0$, the off-diagonal entries yield $A_{21} = 0$ (P_c is nonsingular), which implies that A is block-triangular, so A_c is Hurwitz. Hence, the Lyapunov equation in the $(1, 1)$ entry is solved by $P_c = \lim_{t \to \infty} \int_0^t e^{A_c s} B_c B_c' e^{A_c' s} \, ds > 0$ and then the integral under the limit is nonsingular for all $t > 0$.

If A is not Hurwitz, there is a sufficiently large $\alpha > 0$ such that $A - \alpha I$ is Hurwitz. It follows from the proof of Lemma 4.2 that the kernel of $W_c(t)$ does not change under the replacement $A \to A - \alpha I$. Moreover, such a replacement does not change the triangular structure of the matrix in the statement of the lemma either. Hence, the result remains valid for a general A. □

Proof of Theorem 4.1. We start with showing the equivalence of conditions 2–4.

$2 \iff 3$: First, prove that $2 \Longleftarrow 3$. Assume, on the contrary, that $\operatorname{rank} \begin{bmatrix} A - sI & B \end{bmatrix} = n$, $\forall s \in \mathbb{C}$, but $\operatorname{rank} W_c(t) = r < n$. By Lemma 4.3, there is a unitary U_c such that

$$U_c \begin{bmatrix} A - sI & B \end{bmatrix} \begin{bmatrix} U_c' & 0 \\ 0 & I \end{bmatrix} = \left[\begin{array}{cc:c} A_c - sI_r & \times & B_c \\ 0 & A_{\bar{c}} - sI_{n-r} & 0 \end{array}\right].$$

It is readily seen that the rank of this matrix drops at every eigenvalue of $A_{\bar{c}}$, which is a contradiction.

[1] The notation "×" is used hereafter to denote irrelevant blocks.

Now, prove that 2 $\implies$ 3. Again on the contrary, let rank $W_c(t) = n$, but rank $\begin{bmatrix} A - s_0 I & B \end{bmatrix} < n$ at some $s_0 \in \mathbb{C}$. There is then $\eta_0 \neq 0$ such that $\eta_0' \begin{bmatrix} A - s_0 I & B \end{bmatrix} = 0$. Hence, $\eta_0' A = s_0 \eta_0'$ and $\eta_0' B = 0$. This implies that $\eta_0' e^{At} B = e^{s_0 t} \eta_0' B = 0$ for all t, so $\eta' W_c(t) = 0$, which is a contradiction.

3 $\iff$ 4 : Follows by Lemma 4.2.

Next, show that 1 implies 2 (which, in turn, yields that 1 also implies 3 and 4):

1 $\implies$ 2 : Assume, on the contrary, that (A, B) is controllable, but rank $\begin{bmatrix} A - s_0 I & B \end{bmatrix} < n$ at some $s_0 \in \mathbb{C}$. There is then $\eta_0 \neq 0$ such that $\eta_0' A = s_0 \eta_0'$ and $\eta_0' B = 0$. Therefore, $\eta_0'(A + BK) = s_0 \eta_0'$, implying that $s_0 \in \operatorname{spec}(A + BK)$ regardless of K. Hence, (A, B) is not controllable, which is a contradiction.

The final step is to show that 4 (and, therefore, 2 and 3) implies 1:

4 $\implies$ 1 : The proof is only outlined below, details are somewhat bulky, but otherwise unenlightening. If $m = 1$, then $M_c \in \mathbb{R}^{n \times n}$ is nonsingular and Ackermann's formula,

$$K = -e_n' M_c^{-1} \chi_{A+BK}(A),$$

yields K, for which $\det(sI - A - BK) = \chi_{A+BK}(s)$ for any monic polynomial $\chi_{A+BK}(s)$ of degree n. If $m > 1$, then for any nonzero vector $b_0 \in \operatorname{Im} B$ we can construct $K_0 \in \mathbb{R}^{m \times n}$ such that $(A + BK_0, b_0)$ is controllable, see [10], and we can again apply Ackermann's formula and then shift the result by K_0.

This completes the proof. □

Remark 4.1 (controllability and similarity transformations). The controllability property is invariant under similarity transformations. Indeed, if $(\tilde{A}, \tilde{B}) := (TAT^{-1}, TB)$ for some nonsingular T, then

$$A + BK = T^{-1}\tilde{A}T + T^{-1}\tilde{B}K = T^{-1}(\tilde{A} + \tilde{B}\tilde{K})T,$$

where $\tilde{K} := KT^{-1}$ is bijective to K. Furthermore, $\tilde{W}_c(t)$ and the controllability matrix $\tilde{M}_c$ associate with $(\tilde{A}, \tilde{B})$ are

$$\tilde{W}_c(t) = TW_c(t)T' \quad \text{and} \quad \tilde{M}_c = TM_c, \tag{4.11}$$

where $W_c(t)$ and M_c are their counterparts associated with (A, B). ▽

Several intermediate results used in course of proving Theorem 4.1 are of independent interest. First, we saw that the PBH test can fail only at $s \in \operatorname{spec}(A)$. Eigenvalues of A at which this happens are called *uncontrollable modes* of (A, B). This is because these modes remain the eigenvalues of $A + BK$ for every K, see the "1 $\implies$ 2" part of the proof of Theorem 4.1. In other words, uncontrollable modes of A cannot be affected via B. Unstable uncontrollable modes are particularly important, as their presence implies that the matrix $A + BK$ cannot be made Hurwitz. Such eigenvalues are thus called *unstabilizable modes*. A pair (A, B) is then called *stabilizable* if there is K such that $A + BK$ is Hurwitz. Obviously, any controllable pair (A, B) is stabilizable as well. The inverse is not necessarily true, a stabilizable pair may have uncontrollable modes provided that these modes all lie in $\overline{\mathbb{C}}_0$. This is formalized in the following result, which is an immediate consequence of Theorem 4.1.

Proposition 4.4. *The following statements are equivalent:*

1. *The pair (A, B) is stabilizable.*
2. *The matrix $\begin{bmatrix} A - sI & B \end{bmatrix}$ has full row rank $\forall s \in \overline{\mathbb{C}}_0$.*

Second, the transformation in Lemma 4.3 is convenient to visualize the fact that uncontrollable modes cannot be affected through B. The proposition below slightly reformulates Lemma 4.3.

Proposition 4.5 (controllable decomposition)**.** *There is a nonsingular matrix T_c such that*

$$(T_c A T_c^{-1}, T_c B) = \left(\begin{bmatrix} A_c & \times \\ 0 & A_{\bar{c}} \end{bmatrix}, \begin{bmatrix} B_c \\ 0 \end{bmatrix} \right), \tag{4.12}$$

where the pair (A_c, B_c) is controllable and $\operatorname{spec}(A_{\bar{c}})$ *comprises all uncontrollable modes of (A, B). Moreover, the similarity transformation matrix T_c brings (A, B) to form* (4.12) *iff*

$$T_c W_c(t) T_c' = \begin{bmatrix} \tilde{W}_c(t) & 0 \\ 0 & 0 \end{bmatrix}$$

for some $\tilde{W}_c(t) > 0$ and all $t > 0$.

Proof. The first part was constructively proved in Lemma 4.3. It also proves the "if" part of the second statement. The "only if" part can be shown by explicitly constructing $W_c(t)$ for a realization as that in the right-hand side of (4.12). □

The special form of (A, B) in (4.12) is called the *controllable decomposition*. It is clear that (A, B) is stabilizable iff $A_{\bar{c}}$ is Hurwitz. Proposition 4.5 also suggests an algorithm for constructing a controllable decomposition. Namely, all we need is to find a transformation that block-diagonalizes $W_c(t)$ at any $t > 0$.

4.2.2 Observability and detectability

The notions of observability and detectability are dual to the controllability and stabilizability, respectively. We say that the matrix pair $(C, A) \in \mathbb{R}^{p\times n} \times \mathbb{R}^{n\times n}$ is *observable* if the eigenvalues of $A + LC$ can be freely assigned by a suitable choice of $L \in \mathbb{R}^{n\times p}$ and that the pair (C, A) is *detectable* if there exists an L such that the matrix $A + LC$ is Hurwitz ('L' is for Luenberger here). Some observability tests are presented below.

Proposition 4.6. *The following statements are equivalent:*

1. *The pair (C, A) is observable.*
2. *The matrix* $\begin{bmatrix} A - sI \\ C \end{bmatrix}$ *has full column rank $\forall s \in \mathbb{C}$.*
3. *The matrix*
$$W_o(t) := \int_0^t \mathrm{e}^{A'\tau} C' C \mathrm{e}^{A\tau} \mathrm{d}\tau$$
is positive definite for any $t > 0$.
4. *The observability matrix*
$$M_o := \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix}$$
has full rank (i.e. $\operatorname{rank}(M_o) = n$*).*
5. *The pair (A', C') is controllable.*

Proof. The equivalence 1. $\iff$ 5. follows from the very fact that $(A + LC)' = A' + C'L'$ has the same spectrum as $A + LC$. The rest follows from Theorem 4.1. □

The notion of unobservable modes can be introduced on the basis of the PBH condition of Proposition 4.6, much in parallel to the stabilizability notion in the previous subsection. Namely, an eigenvalue of A whose eigenvector v satisfies $Cv = 0$ is called its *unobservable mode*. The pair (C, A) is detectable iff it has no unobservable modes in $\overline{\mathbb{C}}_0$.

Remark 4.2 (observability and similarity transformations). Like the controllability case discussed in Remark 4.1, observability is invariant under similarity transformations. Moreover, the observability-related matrices for $(\tilde{C}, \tilde{A}) := (CT^{-1}, TAT^{-1})$ are

$$\tilde{W}_{\mathrm{o}}(t) = T^{-\prime} W_{\mathrm{o}}(t) T^{-1} \quad \text{and} \quad \tilde{M}_{\mathrm{o}} = M_{\mathrm{o}} T^{-1}, \tag{4.13}$$

where $W_{\mathrm{o}}(t)$ and M_{o} are their counterparts associated with (C, A). ▽

Similarity transformation can be used to visualize unobservable modes. The following counterpart of Proposition 4.5 introduces the *observable decomposition*.

Proposition 4.7 (observable decomposition)**.** *There is a nonsingular matrix T_o such that*

$$(CT_o^{-1}, T_o A T_o^{-1}) = \left(\begin{bmatrix} C_o & 0 \end{bmatrix}, \begin{bmatrix} A_o & 0 \\ \times & A_{\bar{o}} \end{bmatrix} \right), \tag{4.14}$$

where the pair (C_o, A_o) is observable and $\mathrm{spec}(A_{\bar{o}})$ *comprises all unobservable modes of (C, A). Moreover, the similarity transformation matrix T_o brings (C, A) to form* (4.14) *iff*

$$T_o^{-\prime} W_o(t) T_o^{-1} = \begin{bmatrix} \tilde{W}_o(t) & 0 \\ 0 & 0 \end{bmatrix}$$

for some $\tilde{W}_o(t) > 0$ and all $t > 0$.

Like in the controllability analysis, the form of Proposition 4.7 in the stable case is shown up through the matrix $Q := W_{\mathrm{o}}(\infty) \geq 0$, which is called the *observability Gramian* of (C, A). The observability Gramian is the unique solution of the Lyapunov equation

$$A'Q + QA + C'C = 0. \tag{4.15}$$

4.2.3 Kalman canonical decomposition and minimality

Consider now the transfer function $G(s) = D + C(sI - A)^{-1}B$ and assume that the pair (A, B) is not controllable. By Propositions 4.5, there is a similarity transformation bringing $G(s)$ to the form

$$G(s) = \left[\begin{array}{cc|c} A_{\mathrm{c}} & \times & B_{\mathrm{c}} \\ 0 & A_{\bar{\mathrm{c}}} & 0 \\ \hline C_{\mathrm{c}} & \times & D \end{array} \right]$$

where the pair $(A_{\mathrm{c}}, B_{\mathrm{c}})$ is controllable. Now, making use of (B.16a), we have

$$\begin{aligned} G(s) &= D + \begin{bmatrix} C_{\mathrm{c}} & \times \end{bmatrix} \left(sI - \begin{bmatrix} A_{\mathrm{c}} & \times \\ 0 & A_{\bar{\mathrm{c}}} \end{bmatrix} \right)^{-1} \begin{bmatrix} B_{\mathrm{c}} \\ 0 \end{bmatrix} \\ &= D + \begin{bmatrix} C_{\mathrm{c}} & \times \end{bmatrix} \begin{bmatrix} (sI - A_{\mathrm{c}})^{-1} & \times \\ 0 & (sI - A_{\bar{\mathrm{c}}})^{-1} \end{bmatrix} \begin{bmatrix} B_{\mathrm{c}} \\ 0 \end{bmatrix} \\ &= D + C_{\mathrm{c}}(sI - A_{\mathrm{c}})^{-1} B_{\mathrm{c}}. \end{aligned}$$

In other words, uncontrollable modes of state-space realizations do not affect the transfer function. Similar arguments, based on Propositions 4.7 and equality (B.16b), yield that unobservable modes of state-space realizations do not affect the transfer function either. These two cases can be united to end up with the following fundamental result.

Theorem 4.8 (Kalman canonical decomposition)**.** *There is a nonsingular matrix T such that*

$$G(s) = \left[\begin{array}{c|c} TAT^{-1} & TB \\ \hline CT^{-1} & D \end{array}\right] = \left[\begin{array}{cccc|c} A_{c\bar{o}} & \times & \times & \times & B_{c\bar{o}} \\ 0 & A_{co} & 0 & \times & B_{co} \\ 0 & 0 & A_{\bar{c}\bar{o}} & \times & 0 \\ 0 & 0 & 0 & A_{\bar{c}o} & 0 \\ \hline 0 & C_{co} & 0 & C_{\bar{c}o} & D \end{array}\right] = \left[\begin{array}{c|c} A_{co} & B_{co} \\ \hline C_{co} & D \end{array}\right], \tag{4.16}$$

where the pair (A_{co}, B_{co}) is controllable and the pair (C_{co}, A_{co}) is observable, so that the modes of A_{co}, $A_{c\bar{o}}$, $A_{\bar{c}o}$, and $A_{\bar{c}\bar{o}}$ are controllable-and-observable, controllable-but-unobservable, observable-but-uncontrollable, and uncontrollable-and-unobservable modes of the triple (C, A, B), respectively.

Proof. First, carry out the controllable decomposition with a transformation matrix T_1. Second, apply the observable decomposition to the controllable part and swap the blocks in (4.14) to end up with an upper triangular controllable "A" matrix, which shall give another transformation matrix, say T_2, and

$$G(s) = \left[\begin{array}{ccc|c} A_{\mathrm{c\bar{o}}} & \times & \times & B_{\mathrm{c\bar{o}}} \\ 0 & A_{\mathrm{co}} & \times & B_{\mathrm{co}} \\ 0 & 0 & A_{\bar{\mathrm{c}}} & 0 \\ \hline 0 & C_{\mathrm{co}} & \times & D \end{array}\right]. \tag{4.17}$$

Without loss of generality (see arguments in the proof of Lemma 4.3), assume that A_{co} and $A_{\bar{\mathrm{c}}}$ are Hurwitz and consider the pair

$$\left(\begin{bmatrix} C_{\mathrm{co}} & \times \end{bmatrix}, \begin{bmatrix} A_{\mathrm{co}} & \times \\ 0 & A_{\bar{\mathrm{c}}} \end{bmatrix} \right). \tag{4.18}$$

The Lyapunov equation for the corresponding observability Gramian is

$$\begin{bmatrix} A'_{\mathrm{co}} & 0 \\ \times & A'_{\bar{\mathrm{c}}} \end{bmatrix} \begin{bmatrix} Q_{\mathrm{co}} & Q_{12} \\ Q'_{12} & Q_{22} \end{bmatrix} + \begin{bmatrix} Q_{\mathrm{co}} & Q_{12} \\ Q'_{12} & Q_{22} \end{bmatrix} \begin{bmatrix} A_{\mathrm{co}} & \times \\ 0 & A_{\bar{\mathrm{c}}} \end{bmatrix} + \begin{bmatrix} C'_{\mathrm{co}} \\ \times \end{bmatrix} \begin{bmatrix} C_{\mathrm{co}} & \times \end{bmatrix} = 0.$$

Because $(C_{\mathrm{co}}, A_{\mathrm{co}})$ is observable, by construction, $Q_{\mathrm{co}} > 0$ and the similarity transformation

$$T_3 = \begin{bmatrix} I & 0 & 0 \\ 0 & I & Q_{\mathrm{co}}^{-1} Q_{12} \\ 0 & 0 & I \end{bmatrix}$$

keeps all "named" blocks in (4.17) intact and at the same time renders the observability Gramian of the transformed version of (4.18) block-diagonal (cf. (B.14a)), i.e.

$$\begin{bmatrix} A'_{\mathrm{co}} & 0 \\ A'_{23} & A'_{\bar{\mathrm{c}}} \end{bmatrix} \begin{bmatrix} Q_{\mathrm{co}} & 0 \\ 0 & Q_3 \end{bmatrix} + \begin{bmatrix} Q_{\mathrm{co}} & 0 \\ 0 & Q_3 \end{bmatrix} \begin{bmatrix} A_{\mathrm{co}} & A_{23} \\ 0 & A_{\bar{\mathrm{c}}} \end{bmatrix} + \begin{bmatrix} C'_{\mathrm{co}} \\ C'_{\bar{\mathrm{c}}} \end{bmatrix} \begin{bmatrix} C_{\mathrm{co}} & C_{\bar{\mathrm{c}}} \end{bmatrix} = 0$$

for some A_{23} and $C_{\bar{\mathrm{c}}}$. It is readily seen that Q_3 is the observability Gramian of $(C_{\bar{\mathrm{c}}}, A_{\bar{\mathrm{c}}})$. Carrying out yet another observable decomposition (and the permutation as above) on it, we obtain the fourth transformation matrix, say T_4, and bring $(C_{\bar{\mathrm{c}}}, A_{\bar{\mathrm{c}}})$ to the form of the last two blocks in (4.16). The very same transformation renders the first block of A_{23}, the one having the same number of columns as $A_{\bar{\mathrm{c}}\bar{\mathrm{o}}}$, zero. This follows from the last Lyapunov equation above, with $C_{\bar{\mathrm{c}}} = \begin{bmatrix} 0 & C_{\bar{\mathrm{c}}\mathrm{o}} \end{bmatrix}$. Thus, $T = T_4 T_3 T_2 T_1$ is what we need.

Finally, the controllability and observability properties of the eigenvalues of each diagonal block of TAT^{-1} and the last equality in (4.16) follow by straightforward algebra. □

Motivated by this result, uncontrollable and unobservable modes of state-space realizations are called their *hidden modes*. It follows from Theorem 4.8 that hidden modes of the state-space realization indeed do not affect the transfer function. This means that any realization that contains hidden modes is non-minimal in the sense that there exists another realization of the same transfer function with lower-dimensional "A" matrix. Motivated by this, a state-space realization (A, B, C, D) is said to be *minimal* if the dimension of A is smallest possible among all realizations of a given system. We shall see in §4.3.2 that this notion is a key in expressing "external" properties of transfer functions in terms of "internal" properties of their realizations. Meanwhile, the following result characterizes minimality in terms of verifiable properties.

Theorem 4.9. *A realization (A, B, C, D) is minimal iff (A, B) is controllable and (C, A) is observable.*

Proof. The "only if" part follows directly from Theorem 4.8. So it suffice to prove the "if" part only. Assume, on the contrary, that the dimension of A is n, (A, B) is controllable, (C, A) is observable, but the realization is not minimal. In other words, there exist matrices A_r, B_r, and C_r so that

$$\left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right] = \left[\begin{array}{c|c} A_\mathrm{r} & B_\mathrm{r} \\ \hline C_\mathrm{r} & D \end{array}\right]$$

and the dimension of A_r is $n_\mathrm{r} < n$. It follows from the equality above that $C\mathrm{e}^{At}B = C_\mathrm{r}\mathrm{e}^{A_\mathrm{r}t}B_\mathrm{r}$ for all t. Therefore,

$$C\mathrm{e}^{A\sigma}\mathrm{e}^{A\tau}B = C_\mathrm{r}\mathrm{e}^{A_\mathrm{r}\sigma}\mathrm{e}^{A_\mathrm{r}\tau}B_\mathrm{r}$$

for all σ and τ. Pre-multiplying both sides of this equality by $\mathrm{e}^{A'\sigma}C'$ and post-multiplying by $B'\mathrm{e}^{A'\tau}$ yield:

$$\mathrm{e}^{A'\sigma}C'C\mathrm{e}^{A\sigma}\mathrm{e}^{A\tau}BB'\mathrm{e}^{A'\tau} = \mathrm{e}^{A'\sigma}C'C_\mathrm{r}\mathrm{e}^{A_\mathrm{r}\sigma}\mathrm{e}^{A_\mathrm{r}\tau}B_\mathrm{r}B'\mathrm{e}^{A'\tau}.$$

After integrating both sides from 0 to t over both σ and τ we get:

$$W_\mathrm{o}(t)W_\mathrm{c}(t) = \int_0^t \mathrm{e}^{A'\sigma}C'C_\mathrm{r}\mathrm{e}^{A_\mathrm{r}\sigma}\,\mathrm{d}\sigma \int_0^t \mathrm{e}^{A_\mathrm{r}\tau}B_\mathrm{r}B'\mathrm{e}^{A'\tau}\,\mathrm{d}\tau =: W_{\mathrm{r}1}(t)W_{\mathrm{r}2}(t).$$

Since (A, B) is controllable and (C, A) is observable, $\mathrm{rank}(W_\mathrm{o}W_\mathrm{c}) = n$. On the other hand, $W_{\mathrm{r}1}(t) \in \mathbb{R}^{n\times n_\mathrm{r}}$ and $W_{\mathrm{r}2}(t) \in \mathbb{R}^{n_\mathrm{r}\times n}$. Hence, $\mathrm{rank}(W_{\mathrm{r}1}W_{\mathrm{r}2}) \le n_\mathrm{r} < n$, which is a contradiction. □

Although a minimal realization is not unique, any two minimal realizations are tightly connected, as is shown by the following result.

Theorem 4.10. *Any two minimal realizations of a finite-dimensional LTI system are similar.*

Proof. Consider two minimal realizations (A, B, C, D) and $(\tilde{A}, \tilde{B}, \tilde{C}, D)$ of an LTI system G. Our goal is to shown that there is a nonsingular matrix T such that $(\tilde{A}, \tilde{B}, \tilde{C}, D) = (TAT^{-1}, TB, CT^{-1}, D)$. By minimality, the controllability matrices associated with the realizations above, M_c and $\tilde{M}_\mathrm{c}$, have full row ranks and the observability matrices M_o and $\tilde{M}_\mathrm{o}$ have full column ranks. If both realizations represent the same system, then $C\mathrm{e}^{At}B = \tilde{C}\mathrm{e}^{\tilde{A}t}\tilde{B}$ for all t, which is equivalent to the condition $CA^iB = \tilde{C}\tilde{A}^i\tilde{B}$ for all $i \in \mathbb{Z}_+$ (see the proof of Lemma 4.2). This, in turn, implies that the Hankel matrices

$$M_\mathrm{o}M_\mathrm{c} = \tilde{M}_\mathrm{o}\tilde{M}_\mathrm{c} \quad \text{and} \quad M_\mathrm{o}AM_\mathrm{c} = \tilde{M}_\mathrm{o}\tilde{A}\tilde{M}_\mathrm{c}, \tag{4.19}$$

which are key equalities to proceed. Introduce $T := (\tilde{M}_\mathrm{o}'\tilde{M}_\mathrm{o})^{-1}\tilde{M}_\mathrm{o}'M_\mathrm{o}$ and $S := M_\mathrm{c}\tilde{M}_\mathrm{c}'(\tilde{M}_\mathrm{c}\tilde{M}_\mathrm{c}')^{-1}$, which are well defined by the minimality of $(\tilde{A}, \tilde{B}, \tilde{C}, D)$. The first equality in (4.19) yields that $TS = I$, so $T = S^{-1}$, and also that $TM_\mathrm{c} = \tilde{M}_\mathrm{c}$ and $M_\mathrm{o}S = M_\mathrm{o}T^{-1} = \tilde{M}_\mathrm{o}$. It then follows from structures of the controllability and observability matrices that $TB = \tilde{B}$ and $CT^{-1} = \tilde{C}$. Finally, the second equality of (4.19) yields that $TAS = TAT^{-1} = \tilde{A}$. Hence, T is the required similarity transformation matrix. □

4.2.4 Constructing minimal realizations: Gilbert's realization

One can frequently face with the need to build a minimal state-space realization for a given transfer function. Of course, there always exists a possibility to transform a given non-minimal realization to the Kalman canonical form and then to extract hidden modes by elimination. This option, however, might be computationally consuming and introduce additional numerical errors. Moreover, the resulting system would almost certainly loose any structure of the original transfer function. In the SISO case, there are numerous algorithms for constructing minimal realizations directly from transfer functions. For example, any canonical realization, like companion and observer forms, is minimal provided the transfer function contains no pole / zero cancellations. The situation is more complicated in the MIMO case, where the construction of a minimal state-space realization might not be trivial.

Below, we consider a relatively simple algorithm, which is applicable only to systems having simple poles. Let $G(s)$ be a $p \times m$ proper transfer function. It can always be presented as

$$G(s) = \frac{1}{d(s)} N_G(s),$$

where $N_G(s)$ is a polynomial matrix and $d(s)$ is a monic scalar polynomial, which is the least common denominator of the entries $G_{ij}(s)$ of $G(s)$. The following result can then be formulated.

Theorem 4.11 (Gilbert's realization)**.** *Let $d(s) = (s - a_1) \cdots (s - a_r)$ with $a_j \neq a_i$ whenever $i \neq j$ and $G_i := \mathrm{Res}(G(s), a_i) := \lim_{s \to a_i} (s - a_i) G(s)$ have rank $n_i \leq \min\{p, m\}$. The $(\sum_{i=1}^r n_i)$-order realization*

$$\left[\begin{array}{ccc|c} a_1 I_{n_1} & & 0 & B_1 \\ & \ddots & & \vdots \\ 0 & & a_r I_{n_r} & B_r \\ \hline C_1 & \cdots & C_r & G(\infty) \end{array}\right], \tag{4.20}$$

where $B_i \in \mathbb{R}^{n_i \times m}$ and $C_i \in \mathbb{R}^{p \times n_i}$ are full rank matrices such that $G_i = C_i B_i$, is a minimal realization of $G(s)$.

Proof. Because $G(s)$ is proper, it can be decomposed as

$$G(s) = G(\infty) + \sum_{i=1}^{r} \frac{1}{s - a_i} G_i = G(\infty) + \sum_{i=1}^{r} \frac{1}{s - a_i} C_i B_i .$$

It is then an immediate consequence of (4.5) that (4.20) is a realization of $G(s)$. To prove its minimality, we need to prove that realization (4.20) is both controllable and observable. Let us start with controllability. Suppose, on the contrary, that a_i is an uncontrollable mode of (4.20). Therefore, applying the PBH test, there should exist a nonzero $\eta \in \mathbb{C}^{\sum_{i=1}^r n_i}$ such that

$$\begin{aligned} 0 &= \begin{bmatrix} \eta_1' & \cdots & \eta_i' & \cdots & \eta_r' \end{bmatrix} \begin{bmatrix} (a_1 - a_i) I_{n_1} & & & & & B_1 \\ & \ddots & & & & \vdots \\ & & 0 \cdot I_{n_i} & & & B_i \\ & & & \ddots & & \vdots \\ & & & & (a_r - a_i) I_{n_r} & B_r \end{bmatrix} \\ &= \begin{bmatrix} (a_1 - a_i)\eta_1' & \cdots & 0 & \cdots & (a_r - a_i)\eta_r' & \sum_{j=1}^r \eta_j B_j \end{bmatrix}. \end{aligned}$$

Since $a_j - a_i \neq 0$ for all $j \neq i$, the condition above reads $\eta_j = 0$ for all $j \neq i$ and then $\eta_i' B_i = 0$. Yet the latter condition, together with the fact that B_i is a full rank matrix, implies that $\eta_i = 0$ as well. Therefore, $\eta = 0$, which is a contradiction. Thus (4.20) is controllable. The proof of the observability is similar. □

4.3 Properties of transfer functions via state-space realizations

Various properties of transfer functions can be expressed in terms of their state-space realizations. Several examples of that are provided below. In particular, we study how to construct coprime factorizations (§4.3.1), characterize poles and zeros of transfer functions (§4.3.2), and compute their H_2 and H_∞ norms (§4.3.4). The section also briefly discusses the celebrated KYP lemma in §4.3.5.

4.3.1 Coprime factorizations

The notion of the coprime factorization over H_∞ was introduced in §3.3.3. It was mentioned there that constructing coprime factors might not be trivial if $G(s)$ is unstable. The result below shows that the task is simplified if $G(s)$ is real-rational and given in terms of its state-space realization (A, B, C, D). Namely, the problem is effectively reduced to those of designing a stabilizing state feedback and a stable observer.

Proposition 4.12. *Let (A, B, C, D) be stabilizable and detectable. The transfer functions*

$$\begin{bmatrix} X(s) & Y(s) \\ -\tilde{N}(s) & \tilde{M}(s) \end{bmatrix} = \left[\begin{array}{c|cc} A+LC & B+LD & -L \\ \hline -K & I & 0 \\ -C & -D & I \end{array}\right] \tag{4.21a}$$

and

$$\begin{bmatrix} M(s) & -\tilde{Y}(s) \\ N(s) & \tilde{X}(s) \end{bmatrix} = \left[\begin{array}{c|cc} A+BK & B & -L \\ \hline K & I & 0 \\ C+DK & D & I \end{array}\right], \tag{4.21b}$$

for arbitrary K and L such that $A+BK$ and $A+LC$ are Hurwitz, constitute a double coprime factorization, as in (3.34)*, of $G(s) = D + C(sI - A)^{-1}B$.*

Proof. It follows from (4.5) that poles of a transfer function are contained in the spectrum of the "A" matrix of any of its state-space realizations. Hence, the transfer functions in (4.21) may have poles only in $\mathbb{C} \setminus \bar{\mathbb{C}}_0$. Because these transfer functions are also proper, they all belong to RH_∞. Now, it follows directly from (4.8) that

$$\begin{bmatrix} X(s) & Y(s) \\ -\tilde{N}(s) & \tilde{M}(s) \end{bmatrix}^{-1} = \begin{bmatrix} M(s) & -\tilde{Y}(s) \\ N(s) & \tilde{X}(s) \end{bmatrix},$$

which agrees with (3.34) and *(a)* proves that $N(s)$ and $M(s)$ are right coprime over RH_∞, *(b)* proves that $\tilde{N}(s)$ and $\tilde{M}(s)$ are left coprime over RH_∞, and *(c)* gives an explicit construction of the corresponding Bézout coefficients. To prove that $N(s)$ and $M(s)$ constitute a *rcf* of $G(s)$, use (4.8) and (4.7) to get

$$\begin{aligned} N(s)M^{-1}(s) &= \left[\begin{array}{c|c} A+BK & B \\ \hline C+DK & D \end{array}\right] \left[\begin{array}{c|c} A+BK & B \\ \hline K & I \end{array}\right]^{-1} = \left[\begin{array}{c|c} A+BK & B \\ \hline C+DK & D \end{array}\right] \left[\begin{array}{c|c} A & B \\ \hline -K & I \end{array}\right] \\ &= \left[\begin{array}{cc|c} A & 0 & B \\ -BK & A+BK & B \\ \hline -DK & C+DK & D \end{array}\right] = \left[\begin{array}{cc|c} A & 0 & B \\ 0 & A+BK & 0 \\ \hline C & C+DK & D \end{array}\right] = \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right] = G(s), \end{aligned}$$

where the second equality in the second line follows by applying the similarity transformation $\left[\begin{smallmatrix} I & 0 \\ -I & I \end{smallmatrix}\right]$ and the third equality there follows by the fact that the eigenvalues of $A + BK$ are uncontrollable (cf. the discussion at the beginning of §4.2.3). Finally, $\tilde{N}(s)$ and $\tilde{M}(s)$ constitute a *lcf* of $G(s)$ because $\tilde{N}(s)M(s) = \tilde{M}(s)N(s)$, which, in turn, follows from the "$(2, 1)$" part of (3.34). □

Remark 4.3 (signal-based inversion). The algebraic way to prove that $NM^{-1} = G$ is straightforward, but somewhat boring. A more elegant approach would be to manipulate input and output signals. To see how it works, note that

$$\begin{bmatrix} u \\ y \end{bmatrix} = \begin{bmatrix} M \\ N \end{bmatrix} v \quad \Longrightarrow \quad \begin{bmatrix} v \\ y \end{bmatrix} = \begin{bmatrix} M^{-1} \\ NM^{-1} \end{bmatrix} u$$

whenever M is invertible. In other words, the system NM^{-1} can be obtained by swapping the input v with the first output u (as a byproduct, we also have M^{-1}). This action, performed via state-space equations, reads

$$\begin{bmatrix} M \\ N \end{bmatrix} : \begin{cases} \dot{x}(t) = (A + BK)x(t) + Bv(t) \\ u(t) = Kx(t) + v(t) \\ y(t) = (C + DK)x(t) + Dv(t) \end{cases} \quad \Longrightarrow \quad \begin{bmatrix} M^{-1} \\ NM^{-1} \end{bmatrix} : \begin{cases} \dot{x}(t) = Ax(t) + Bu(t) \\ v(t) = -Kx(t) + u(t) \\ y(t) = Cx(t) + Du(t) \end{cases},$$

which is arguably more intuitive than the algebraic transformations in the proof of Proposition 4.12. ∇

The choice of the parameters K and L is clearly not unique. This fact can be exploited to end up with factorizations having some favorable properties. For example, K can be chosen so that $\left[\begin{smallmatrix} M(s) \\ N(s) \end{smallmatrix}\right]$ is inner (always possible, see Section 9.A), $M(s)$ is inner (possible if $G(s)$ has no pure imaginary poles), or $N(s)$ is inner (possible if $G(s)$ has no pure imaginary zeros and is left invertible). Such choices, as well as their duals in terms of L, are useful in solving various optimization problems.

4.3.2 Poles, zeros, and degree

It follows from Cramer's rule that poles of the transfer function $G(s) = D + C(sI - A)^{-1}B$ belong to $\operatorname{spec}(A)$ (this was already used in the proof of Proposition 4.12). Motivated by this, we say that $p_i \in \mathbb{C}$ is a *pole of the realization* (A, B, C, D) if $p_i \in \operatorname{spec}(A)$. Clearly, these poles are invariant under similarity transformations. The following fundamental result shows that the relation between poles of a transfer function and those of its realization is indeed strong.

Theorem 4.13. *The McMillan degree of $G(s)$ equals the order of its minimal realization (A, B, C, D) and the set of poles of $G(s)$ coincides with* $\operatorname{spec}(A)$.

Proof (outline). It follows from Theorem 4.8 that hidden modes of a realization do not affect the transfer function. We thus may assume that the realization (A, B, C, D) is minimal and only need to prove that every eigenvalue of A is a pole of $G(s)$, multiplicities counted. The proof (omitted because it is bulky) follows from the explicit construction of a minimal realization from the Smith–McMillan form in [13]. □

Pole directions of transfer functions were defined in §3.4.2 via the Smith–McMillan form (3.35). Connections between this form and state-space realizations requires higher-level polynomial matrix techniques, which are not studied in these notes (see [23, Ch. 4] for details). Some insight can be gained by looking into the Smith–McMillan form of the transfer function $(sI - J_{0,n})^{-1}$, where $J_{0,n} \in \mathbb{R}^{n\times n}$ is the Jordan block of size n associated with 0,

$$J_{0,n} := \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & \cdots & 0 & 0 \end{bmatrix}$$

It can be verified by direct substitution that the Smith–McMillan form of $(sI - J_{0,n})^{-1}$ is

$$U_0(s)(sI - J_{0,n})^{-1}V_0(s) = \operatorname{diag}\{1/s^n, I_{n-1}\}, \tag{4.22}$$

where the $n \times n$ unimodular polynomial matrices

$$U_0(s) := \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & -s & 1 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & -s & 1 & \cdots & 0 & 0 \\ -s & 1 & 0 & \cdots & 0 & 0 \end{bmatrix} \quad \text{and} \quad V_0(s) := \begin{bmatrix} 0 & 0 & \cdots & 0 & -1 \\ 0 & 0 & \cdots & -1 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & -1 & \cdots & 0 & 0 \\ 1 & s & \cdots & s^{n-2} & s^{n-1} \end{bmatrix}$$

(cf. Example 3.4 on p. 59). It follows from (4.22) that the geometric multiplicity of the pole of $(sI - J_{0,n})^{-1}$ at $s = 0$ is one regardless of n, which is not quite intuitive. Also, it is readily seen that

$$\ker \begin{bmatrix} e_2' \\ \vdots \\ e_n' \end{bmatrix} [V_0(0)]' = \operatorname{span}(e_n) = \ker J_{0,n}' \quad \text{and} \quad \ker \begin{bmatrix} e_2' \\ \vdots \\ e_n' \end{bmatrix} U_0(0) = \operatorname{span}(e_1) = \ker J_{0,n}.$$

cf. (3.38).

These arguments can be extended to Jordan blocks associated with any p_i by replacing $s \to s - p_i$ and to a general A via transforming it to the Jordan normal form. Hence, the geometric multiplicity of every pole in $(sI - A)^{-1}$, as per the definition on p. 56, equals the geometric multiplicity of the corresponding eigenvalue of A, as per the definition on p. 31. It may then appear natural to define the input and output directions of a pole p_i of the realization (A, B, C, D) as

$$\operatorname{pdir}_{\mathrm{i}}(G, p_i) = B' \ker(p_i I - A)' \subset \mathbb{C}^m \quad \text{and} \quad \operatorname{pdir}_{\mathrm{o}}(G, p_i) = C \ker(p_i I - A) \subset \mathbb{C}^p. \tag{4.23}$$

Remark 4.4 (pole directions in Gilbert's realization). Determining pole directions of systems given in Gilbert's realization from Theorem 4.11 is particularly simple. Indeed, in this case the eigenvectors of A are the standard basis in $\mathbb{C}^n$, so that

$$\operatorname{pdir}_{\mathrm{i}}(G, a_i) = \operatorname{Im} B_i' \quad \text{and} \quad \operatorname{pdir}_{\mathrm{o}}(G, a_i) = \operatorname{Im} C_i \tag{4.23$'$}$$

for realization (4.20), which follows by a direct inspection. ∇

Introduce now the polynomial matrix

$$R_G(s) := \begin{bmatrix} A - sI_n & B \\ C & D \end{bmatrix}, \tag{4.24}$$

which is called the *Rosenbrock system matrix* (RSM) of the realization (A, B, C, D). The following technical result sheds light on the relation between $R_G(s)$ and the corresponding transfer function.

Lemma 4.14. *If (A, B, C, D) is a realization of G, then*

$$\operatorname{rank}(R_G(s_0)) = n + \operatorname{rank}(G(s_0)), \quad \forall s_0 \notin \operatorname{spec}(A).$$

and then $\operatorname{nrank}(R_G(s)) = n + \operatorname{nrank}(G(s))$.

Proof. The result follows from either one of the relations

$$R_G(s) = \begin{bmatrix} A - sI & 0 \\ C & G(s) \end{bmatrix} \begin{bmatrix} I & -(sI - A)^{-1}B \\ 0 & I \end{bmatrix} = \begin{bmatrix} I & 0 \\ -C(sI - A)^{-1} & I \end{bmatrix} \begin{bmatrix} A - sI & B \\ 0 & G(s) \end{bmatrix}, \tag{4.25}$$

which are straightforward to verify, and the fact that the normal rank of a polynomial matrix differs from its rank only at a finite number of points. The result can also be proved via the observation that $G(s)$ is the Schur complement of $A - sI$ in $R_G(s)$. □

Lemma 4.14 suggests that there may be a relation between system zeros and points at which the corresponding RSM loses its normal rank. We then call every $z_i \in \mathbb{C}$ at which $\operatorname{rank}(R_G(z_i)) < \operatorname{nrank}(R_G(s))$ an *invariant zero of the realization* (A, B, C, D). Because

$$\begin{bmatrix} TAT^{-1} - sI & TB \\ CT^{-1} & D \end{bmatrix} = \begin{bmatrix} T & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} A - sI & B \\ C & D \end{bmatrix} \begin{bmatrix} T^{-1} & 0 \\ 0 & I \end{bmatrix},$$

invariant zeros, like poles of the realization, are not affected by similarity transformations. The following result establishes a connection between invariant zeros of a realization and the transmission zeros of the corresponding transfer function, defined in §3.4.2.

Theorem 4.15. *Invariant zeros of a realization (A, B, C, D) comprise all its hidden modes and the transmission zeros of $D + C(sI - A)^{-1}B$.*

Proof. Since invariant zeros are invariant under similarity transformations, we may consider the Kalman canonical realization as in (4.17). It is then readily verified that $R_G(s)$ loses its normal rank at all hidden modes. We thus only need to prove that invariant zeros of a minimal realization coincide with transmission zeros of the corresponding transfer function. This claim is a direct consequence of the first item of Proposition 3.7 and Lemma 4.14 in the case when invariant zeros are not poles of the realization. The proof in the general case is quite technical and thus omitted, see [12, §6.5.3] for details. □

Hence, the problem of calculating transmission zeros of a transfer function $G(s)$ can be reduced to the problem of finding points at which the RSM associated with a minimal realization of $G(s)$ drops rank. This is a so-called *generalized eigenvalue problem* associated with the matrix pencil

$$R_G(s) = \begin{bmatrix} A & B \\ C & D \end{bmatrix} - s \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}$$

and we need only finite generalized eigenvalues here.

Zero directions defined by (3.39) can also be expressed in terms of RSMs. Consider first the case when $z_i \in \mathbb{C}$ is a transmission zero of $G(s)$ given in terms of its minimal realization (A, B, C, D), but not its pole (so that $z_i \notin \operatorname{spec}(A)$). By Proposition 3.7, in this case $\operatorname{zdir}_{\mathrm{i}}(G, z_i) = \ker G(z_i)$. Given any nonzero $u_i \in \ker G(z_i)$, the first equality in (4.25) yields

$$\begin{aligned} 0 &= \begin{bmatrix} A - z_i I & 0 \\ C & G(z_i) \end{bmatrix} \begin{bmatrix} 0 \\ u_i \end{bmatrix} = \begin{bmatrix} A - z_i I & B \\ C & D \end{bmatrix} \begin{bmatrix} I & (z_i I - A)^{-1}B \\ 0 & I \end{bmatrix} \begin{bmatrix} 0 \\ u_i \end{bmatrix} \\ &= \begin{bmatrix} A - z_i I & B \\ C & D \end{bmatrix} \begin{bmatrix} (z_i I - A)^{-1} B u_i \\ u_i \end{bmatrix}. \end{aligned}$$

This shows that any vector from $\operatorname{zdir}_{\mathrm{i}}(G, z_i)$ must belong to $\begin{bmatrix} 0 & I_m \end{bmatrix} \ker R_G(z_i)$. The other direction is also true. To see that, suppose now that $\left[\begin{smallmatrix} x_i \\ u_i \end{smallmatrix}\right] \in \ker R_G(z_i)$. This implies, again by the first equality in (4.25), that

$$0 = \begin{bmatrix} A - z_i I & B \\ C & D \end{bmatrix} \begin{bmatrix} x_i \\ u_i \end{bmatrix} = \begin{bmatrix} A - z_i I & 0 \\ C & G(z_i) \end{bmatrix} \begin{bmatrix} \tilde{x}_i \\ u_i \end{bmatrix}, \tag{4.26}$$

where $\tilde{x}_i := x_i - (z_i I - A)^{-1} B u_i$. Because $z_i \notin \operatorname{spec}(A)$, the first block row above entails $\tilde{x}_i = 0$ and the second one reads $G(z_i)u_i = 0$, which is what we need. Similar arguments can be applied to the output zero direction, which leads to the following relations:

$$\operatorname{zdir}_{\mathrm{i}}(G, z_i) = \begin{bmatrix} 0 & I_m \end{bmatrix} \ker R_G(z_i) \subset \mathbb{C}^m \quad \text{and} \quad \operatorname{zdir}_{\mathrm{o}}(G, z_i) = \begin{bmatrix} 0 & I_p \end{bmatrix} \ker [R_G(z_i)]' \subset \mathbb{C}^p. \tag{4.27}$$

The situation turns more complicated if an invariant zero $z_i \in \operatorname{spec}(A)$. It can still be shown (perhaps) that the directions in (4.27) coincide with those in (3.39) in the general case as well. The argument below

supports this claim circumstantially. Assume that there is $\left[\begin{smallmatrix} x_i \\ u_i \end{smallmatrix}\right] \neq 0$ such that the first equality in (4.26) holds true. By the assumed observability of (C, A) we can conclude that $u_i \neq 0$. But then the equality $(z_i I - A)x_i = Bu_i \iff u_i' B' = x_i'(z_i I - A)'$, yields that $u_i' B' \eta = 0$ for all $\eta \in \ker(z_i I - A)'$. This, in turn, implies that $u_i \perp \mathrm{pdir_i}(G, z_i)$, which agrees with the discussion about the orthogonality of pole and zero directions on p. 57.

Example 4.1. Consider the transfer function from Example 3.1 on p. 57,

$$G(s) = \frac{1}{s}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} = \frac{1}{s}\begin{bmatrix} 1 \\ 1 \end{bmatrix}\begin{bmatrix} 1 & 1 \end{bmatrix}.$$

By Theorem 4.11, its minimal realization is

$$G(s) = \left[\begin{array}{c|cc} 0 & 1 & 1 \\ \hline 1 & 0 & 0 \\ 1 & 0 & 0 \end{array}\right],$$

which has order 1. This realization has one pole at the origin and because $\ker(s-0)|_{s=0} = \mathbb{C}$, we have that

$$\mathrm{pdir_i}(G, 0) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}\mathbb{C} = \mathrm{span}\left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}\right) \quad \text{and} \quad \mathrm{pdir_o}(G, 0) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}\mathbb{C} = \mathrm{span}\left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}\right).$$

The Rosenbrock system matrix for this system,

$$R_G(s) = \begin{bmatrix} -s & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix},$$

is such that $\mathrm{rank}(R_G(s)) = 2$ for all $s \in \mathbb{C}$. Hence, the system has no zeros. All these results agree with those in Example 3.1. ◊

Example 4.2. The transfer function from Example 3.2 on p. 57,

$$G(s) = \begin{bmatrix} 1 & 1/s \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \frac{1}{s}\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \frac{1}{s}\begin{bmatrix} 1 \\ 0 \end{bmatrix}\begin{bmatrix} 0 & 1 \end{bmatrix}.$$

Its minimal Gilbert's realization is

$$G(s) = \left[\begin{array}{c|cc} 0 & 0 & 1 \\ \hline 1 & 1 & 0 \\ 0 & 0 & 1 \end{array}\right]$$

and has order 1. This realization has one pole at the origin and because $\ker(s-0)|_{s=0} = \mathbb{C}$, we have that

$$\mathrm{pdir_i}(G, 0) = \begin{bmatrix} 0 \\ 1 \end{bmatrix}\mathbb{C} = \mathrm{span}\left(\begin{bmatrix} 0 \\ 1 \end{bmatrix}\right) \quad \text{and} \quad \mathrm{pdir_o}(G, 0) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}\mathbb{C} = \mathrm{span}\left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}\right).$$

The Rosenbrock system matrix for this system,

$$R_G(s) = \begin{bmatrix} -s & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

has full normal rank and $\det(R_G(s)) = -s$. Hence, the system has a zero at the origin, whose multiplicity is 1 (because $\mathrm{rank}(R_G(0)) = 2$). It is readily verified that

$$\ker R_G(0) = \mathrm{span}\left(\left[\begin{array}{c} -1 \\ \hdashline 1 \\ 0 \end{array}\right]\right) \quad \text{and} \quad \ker[R_G(0)]' = \mathrm{span}\left(\left[\begin{array}{c} -1 \\ \hdashline 0 \\ 1 \end{array}\right]\right),$$

so that

$$\operatorname{zdir}_{\mathrm{i}}(G,0) = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \ker R_G(0) = \operatorname{span}\left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}\right)$$

and

$$\operatorname{zdir}_{\mathrm{o}}(G,0) = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \ker[R_G(0)]' = \operatorname{span}\left(\begin{bmatrix} 0 \\ 1 \end{bmatrix}\right).$$

All these results again agree with those in Example 3.2. ◊

Example 4.3. Consider now the transfer function from Example 3.3 on p. 58,

$$G(s) = \begin{bmatrix} 1/(s+1) & 0 & (s-1)/((s+1)(s+2)) \\ -1/(s-1) & 1/(s+2) & 1/(s+2) \end{bmatrix}.$$

The least common denominator of its entries is $d(s) = (s-1)(s+1)(s+2)$ and the residues of its roots are

$$\operatorname{Res}(G(s),1) = \begin{bmatrix} 0 \\ -1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}, \quad \operatorname{Res}(G(s),-1) = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & -2 \end{bmatrix}, \quad \operatorname{Res}(G(s),-2) = \begin{bmatrix} 0 & 0 & 3 \\ 0 & 1 & 1 \end{bmatrix}.$$

Hence, its minimal realization by Theorem 4.11 is

$$G(s) = \left[\begin{array}{c:c:cc|ccc} 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ \hdashline 0 & -1 & 0 & 0 & 1 & 0 & -2 \\ \hdashline 0 & 0 & -2 & 0 & 0 & 0 & 3 \\ 0 & 0 & 0 & -2 & 0 & 1 & 1 \\ \hline 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 & 0 & 0 \end{array}\right]$$

which has order 4. Because this is Gilbert's realization, we use formulae (4.23′) on p. 80 to end up with

$$\operatorname{pdir}_{\mathrm{i}}(G,1) = \operatorname{Im}\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \operatorname{span}\left(\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}\right), \quad \operatorname{pdir}_{\mathrm{o}}(G,1) = \operatorname{Im}\begin{bmatrix} 0 \\ -1 \end{bmatrix} = \operatorname{span}\left(\begin{bmatrix} 0 \\ 1 \end{bmatrix}\right),$$

$$\operatorname{pdir}_{\mathrm{i}}(G,-1) = \operatorname{Im}\begin{bmatrix} 1 \\ 0 \\ -2 \end{bmatrix} = \operatorname{span}\left(\begin{bmatrix} -1 \\ 0 \\ 2 \end{bmatrix}\right), \quad \operatorname{pdir}_{\mathrm{o}}(G,-1) = \operatorname{Im}\begin{bmatrix} 1 \\ 0 \end{bmatrix} = \operatorname{span}\left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}\right),$$

$$\operatorname{pdir}_{\mathrm{i}}(G,-2) = \operatorname{Im}\begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 3 & 1 \end{bmatrix} = \operatorname{span}\left(\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}\right), \quad \text{and} \quad \operatorname{pdir}_{\mathrm{o}}(G,-2) = \operatorname{Im} I_2 = \mathbb{C}^2.$$

It can be verified that the Rosenbrock system matrix for this system,

$$R_G(s) = \begin{bmatrix} 1-s & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & -1-s & 0 & 0 & 1 & 0 & -2 \\ 0 & 0 & -2-s & 0 & 0 & 0 & 3 \\ 0 & 0 & 0 & -2-s & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix},$$

has full rank, 6, at $s = 0$. Hence, $\operatorname{nrank}(R_G(s)) = 6$. Solving the corresponding generalized eigenvalue problem, we find only one finite generalized eigenvalue, that at $s = 1$. Now,

$$\ker R_G(1) = \operatorname{span}\left(\begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ \hdashline 0 \\ 3 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ -3 \\ 3 \\ 1 \\ \hdashline 0 \\ 0 \\ 3 \end{bmatrix}\right) \quad \text{and} \quad \ker[R_G(1)]' = \operatorname{span}\left(\begin{bmatrix} -3 \\ 3 \\ 2 \\ 0 \\ \hdashline 6 \\ 0 \end{bmatrix}\right),$$

which yields

$$\operatorname{zdir_i}(G, 1) = \operatorname{span}\left(\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}\right) \quad \text{and} \quad \operatorname{zdir_o}(G, 1) = \operatorname{span}\left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}\right).$$

All these results are in accordance with those computed in Example 3.3, again. This is especially noteworthy with regard to the zero directions, because the zero does coincide with a pole. ◊

SISO zeros can be interpreted as complex points z_i such that the input u satisfying $u(t) = \mathrm{e}^{z_i t}\mathbb{1}(t)$ is filtered out by the system. Such an interpretation can be extended to the MIMO case as well. To this end, let z_i be an invariant zero of the realization (A, B, C, D) and consider an input of the form $u(t) = u_i \mathrm{e}^{z_i t}\mathbb{1}(t)$ for some nonzero $u_i \in \mathbb{C}^m$. If the Sylvester equation $-x_i z_i + A x_i + B u_i = 0$ is solvable in $x_i \in \mathbb{C}^n$, then the output in the Laplace domain can be split as

$$Y(s) = G(s) u_i \frac{1}{s - z_i} = \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right]\left[\begin{array}{c|c} z_i & 1 \\ \hline u_i & 0 \end{array}\right] = -\left[\begin{array}{c|c} A & x_i \\ \hline C & 0 \end{array}\right] + \left[\begin{array}{c|c} z_i & 1 \\ \hline C x_i + D u_i & 0 \end{array}\right],$$

by (4.9). The Sylvester equation above is solvable for all u_i if $z_i \notin \operatorname{spec}(A)$ and may be solvable if $z_i \in \operatorname{spec}(A)$ and $u_i \perp \operatorname{pdir_i}(G, z_i)$, see the discussion on p. 82. Assuming that x_i exists, the expansion of $Y(s)$ above yields

$$y(t) = -C\mathrm{e}^{At} x_i \mathbb{1}(t) + (C x_i + D u_i)\mathrm{e}^{z_i t}\mathbb{1}(t).$$

The first term in this expression equals the response of the system (4.3) to the initial condition $x_0 = -x_i$ and can be thought of as the transient component of the response to the chosen u. The second term is then the "steady-state" effect of the input. It is filtered out by the system iff $C x_i + D u_i = 0$, which, in turn, happens iff $u_i \in \operatorname{zdir_i}(G, z_i)$, cf. the first equality in (4.26). As a matter of fact, if the initial condition of the system is $x(0) = x_i$ and the input satisfies $u(t) = u_i \mathrm{e}^{z_i t}\mathbb{1}(t)$ for $\left[\begin{smallmatrix} x_i \\ u_i \end{smallmatrix}\right] \in \ker R_G(z_i)$, then the output $y = 0$, i.e. the system does not respond to this combination of initial conditions and input at all.

By similar arguments, it can be shown that if the output of G having a zero at z_i is measured by a sensor with the transfer function $1/(s - z_i)\, y_i'$, then the measured signal does not contain a dynamical effect of the sensor iff $y_i \in \operatorname{zdir_o}(G, z_i)$.

4.3.3 Realization poles and invariant zeros in terms of coprime factors

Coprime factors over RH_∞ of finite-dimensional MIMO systems play, in a sense, roles of the numerator and denominator of SISO transfer functions. We may then expect that poles and zeros of a transfer function are related to zeros of its coprime factors. This is indeed the case when poles and invariant zeros of state-space realizations are considered. The result below also connects pole and zero directions of transfer functions with zero directions of coprime factors.

Proposition 4.16. *Let (A, B, C, D) be a stabilizable and detectable realization of a finite-dimensional LTI system G and the coprime factors $\tilde{M}(s)$, $\tilde{N}(s)$, $M(s)$, and $N(s)$ of $G(s)$ be constructed as in* (4.21)*. The following statements hold true regardless the choice of the stabilizing gains K and L in* (4.21)*:*

- $z_i \in \mathbb{C}$ *is an invariant zero of* $\tilde{M}$ *iff it is a realization pole of* G*, with* $\mathrm{zdir_i}(\tilde{M}, z_i) = \mathrm{pdir_o}(G, z_i)$*;*
- $z_i \in \mathbb{C}$ *is an invariant zero of* M *iff it is a realization pole of* G*, with* $\mathrm{zdir_o}(M, z_i) = \mathrm{pdir_i}(G, z_i)$*;*
- $z_i \in \mathbb{C}$ *is an invariant zero of* $\tilde{N}$ *iff it is an invariant zero of* G*, with* $\mathrm{zdir_i}(\tilde{N}, z_i) = \mathrm{zdir_i}(G, z_i)$*;*
- $z_i \in \mathbb{C}$ *is an invariant zero of* N *iff it is an invariant zero of* G*, with* $\mathrm{zdir_o}(N, z_i) = \mathrm{zdir_o}(G, z_i)$.

Proof. Key elements of the proof are the relations

$$R_{\tilde{M}}(s) = \begin{bmatrix} A + LC - sI & L \\ C & I \end{bmatrix} = \begin{bmatrix} A - sI & L \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ C & I \end{bmatrix},$$

$$R_M(s) = \begin{bmatrix} A + BK - sI & B \\ K & I \end{bmatrix} = \begin{bmatrix} I & B \\ 0 & I \end{bmatrix} \begin{bmatrix} A - sI & 0 \\ K & I \end{bmatrix},$$

$$R_{\tilde{N}}(s) = \begin{bmatrix} A + LC - sI & B + LD \\ C & D \end{bmatrix} = \begin{bmatrix} I & L \\ 0 & I \end{bmatrix} \begin{bmatrix} A - sI & B \\ C & D \end{bmatrix} = \begin{bmatrix} I & L \\ 0 & I \end{bmatrix} R_G(s),$$

and

$$R_N(s) = \begin{bmatrix} A + BK - sI & B \\ C + DK & D \end{bmatrix} = \begin{bmatrix} A - sI & B \\ C & D \end{bmatrix} \begin{bmatrix} I & 0 \\ K & I \end{bmatrix} = R_G(s) \begin{bmatrix} I & 0 \\ K & I \end{bmatrix},$$

which are readily verifiable. Because all matrix factors above are square and nonsingular, the equivalence of invariant zeros of the coprime factors with invariant zeros and realization poles of G follow immediately. Consider now input zero directions of $\tilde{M}$. Because

$$\ker \begin{bmatrix} A - z_i I & L \\ 0 & I \end{bmatrix} = \begin{bmatrix} I \\ 0 \end{bmatrix} \ker(A - z_i I) \quad \Longrightarrow \quad \ker R_{\tilde{M}}(z_i) = \begin{bmatrix} -I \\ C \end{bmatrix} \ker(z_i I - A),$$

whence the direction statement of the first item follows by comparing the second equality of (4.23) with the first equality of (4.27). The direction statement in the second item follows by dual arguments. Finally, the direction statements in the last two items follow from the obvious facts that $\ker R_{\tilde{N}}(z_i) = \ker R_G(z_i)$ and $\ker[R_N(z_i)]' = \ker[R_G(z_i)]'$. □

It should be emphasized that the realizations of $\tilde{M}$, $\tilde{N}$, M, N in (4.21) need not be minimal even if the original realization of G is minimal, there might be stable cancellations in them. Hence, invariant zeros of either of the coprime factors might not be related to transmission zeros of the corresponding transfer functions. Rather, they could be hidden modes of $\tilde{M}$, $\tilde{N}$, M, or N. Still, no unstable cancellations can occur, so that all invariant zeros of N and $\tilde{N}$ in $\overline{\mathbb{C}}_0$ coincide and are necessarily transmission zeros of both $N(s)$ and $\tilde{N}(s)$. Likewise, all invariant zeros of M and $\tilde{M}$ in $\overline{\mathbb{C}}_0$ coincide and are necessarily transmission zeros of both $M(s)$ and $\tilde{M}(s)$ and are thus the poles of $M^{-1}(s)$ and $\tilde{M}^{-1}(s)$ (cf. Proposition 3.3).

An advantage of dealing with $\tilde{M}$ or M instead of G itself in the analysis of poles lies in the fact that the denominators are always square and bi-proper, which may simplify their analysis. An advantage of dealing with $\tilde{N}$ or N instead of G itself in the analysis of zeros is in the possibility to avoid subtle situations of poles and zeros at the very same points. Indeed, although invariant zeros of $\tilde{N}$ and N do not depend on L and K, their poles do. If the original realization of G is minimal, we can always assign eigenvalues of $A + LC$ or $A + BK$ to be different from zeros of G. Moreover, in many situations only unstable, i.e. those in $\overline{\mathbb{C}}_0$, zeros are of interest, so any coprime factors over RH_∞ can be used to analyze those zeros via the rank drop of $\tilde{N}(s)$ and $N(s)$ rather than higher-dimensional RSMs.

4.3.4 Computing system norms

In this subsection the computation of the H_2 and H_∞ norms of a stable G in terms of its state-space realization $G(s) = D + C(sI - A)^{-1}B$ having a *Hurwitz* A is studied.

Proposition 4.17. *If $D = 0$, then*

$$\|G\|_2^2 = \operatorname{tr}(B'QB) = \operatorname{tr}(CPC'),$$

where Q and P are the observability and controllability Gramians of (C, A) and (A, B), respectively.

Proof. The impulse response of G satisfies $g(t) = C\mathrm{e}^{At}B\mathbb{1}(t)$, see (4.4). It follows from (3.26) that

$$\|G\|_2^2 = \int_{\mathbb{R}} \operatorname{tr}\big(g(t)'g(t)\big)\mathrm{d}t = \int_{\mathbb{R}_+} \operatorname{tr}\big(B'\mathrm{e}^{A't}C'C\mathrm{e}^{At}B\big)\mathrm{d}t = \operatorname{tr}\bigg(B'\int_{\mathbb{R}_+} \mathrm{e}^{A't}C'C\mathrm{e}^{At}\,\mathrm{d}t\,B\bigg),$$

which yields the first equality of the proposition. Now, by the property $\operatorname{tr}(M_1M_2) = \operatorname{tr}(M_2M_1)$,

$$\|G\|_2^2 = \int_{\mathbb{R}} \operatorname{tr}\big(g(t)g(t)'\big)\mathrm{d}t = \int_{\mathbb{R}_+} \operatorname{tr}\big(C\mathrm{e}^{At}BB'\mathrm{e}^{A't}C'\big)\mathrm{d}t = \operatorname{tr}\bigg(C\int_{\mathbb{R}_+} \mathrm{e}^{At}BB'\mathrm{e}^{A't}\,\mathrm{d}t\,C'\bigg),$$

which yields the second equality of the proposition. □

Proposition 4.17 gives an efficient algorithm to compute the H_2 norm. The Lyapunov equations for computing P and Q, (4.10) and (4.15), are linear and can thus be solved in a finite number of steps.

Unlike the H_2 case, there are no closed-form formulae for the H_∞ norm of a stable LTI system. Rather, an iterative search procedure can be organized on every step of which one has to check whether $\|G\|_\infty < \gamma$ for a given $\gamma > 0$. The basis for this procedure is given by the result below.

Proposition 4.18. *If $G \in RH_\infty$, then $\|G\|_\infty < \gamma$ for a given $\gamma > 0$ iff $\|D\| < \gamma$ and the matrix*

$$H_G := \begin{bmatrix} A & 0 \\ C'C & -A' \end{bmatrix} - \begin{bmatrix} B \\ C'D \end{bmatrix}(\gamma^2 I - D'D)^{-1}\begin{bmatrix} -D'C & B' \end{bmatrix}$$

has no pure imaginary eigenvalues.

Proof. It follows from equality (3.22), Theorem A.2, and the fact that $G \in RH_\infty$ that

$$\|G\|_\infty < \gamma \iff \gamma^2 I - [G(\mathrm{j}\omega)]'G(\mathrm{j}\omega) > 0, \quad \forall \omega \in \mathbb{R} \cup \{\pm\infty\}.$$

Because $G(\infty) = D$, the condition on the feedthrough term in the statement of the proposition is necessary. Therefore, in what follows we assume that $D'D < \gamma^2 I_m$. But in that case we have that $\|G\|_\infty < \gamma$ iff $\Phi(s) := \gamma^2 I - G^\sim(s)G(s)$ has no pure imaginary transmission zeros, by mere continuity of its frequency response. It follows from the definition of the conjugate transfer function in (3.29) that

$$G(s) = \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right] \implies G^\sim(s) = \left[\begin{array}{c|c} -A' & C' \\ \hline -B' & D' \end{array}\right].$$

Hence,

$$\Phi(s) = \gamma^2 I - \left[\begin{array}{c|c} -A' & C' \\ \hline -B' & D' \end{array}\right]\left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right] = \left[\begin{array}{cc|c} A & 0 & B \\ C'C & -A' & C'D \\ \hline -D'C & B' & \gamma^2 I - D'D \end{array}\right],$$

by (4.7). Because A is Hurwitz, the realization above has no pure imaginary poles. This, together with Theorem 4.15, implies that $\Phi(s)$ has no pure imaginary zeros iff its realization above has no pure imaginary invariant zeros, regardless its minimality. Taking into account that $\operatorname{nrank}(\Phi(s)) = \operatorname{rank}(\Phi(\infty)) = m$, we have the following condition:

$$\operatorname{rank}\left(\left[\begin{array}{cc:c} A - \mathrm{j}\omega I_n & 0 & B \\ C'C & -A' - \mathrm{j}\omega I_n & C'D \\ \hdashline -D'C & B' & \gamma^2 I_m - D'D \end{array}\right]\right) = 2n + m, \quad \forall \omega \in \mathbb{R}.$$

It follows from (B.14b) that the latter condition is equivalent to the full rank of the Schur complement of $\gamma^2 I_m - D'D$ in it, which reads $\operatorname{rank}(H_G - \mathrm{j}\omega I_{2n}) = 2n$, $\forall \omega \in \mathbb{R}$, and yields the second statement of the proposition. □

Remark 4.5 (what happens at $\gamma = \|G\|_\infty$). It follows from the proof of Proposition 4.18 that two situations are possible if $\|G\|_\infty = \gamma$. It might happen that $\|D\|$ is still strictly smaller than γ, in which case H_G is well defined and must have at least a pair of jω-axis eigenvalues. In fact, such eigenvalues should be repeated. This can be seen via the property that $H_G = -H_G'$, which, in turn, implies that $\lambda \in \operatorname{spec}(H_G)$ iff so does $-\lambda$ too. Thus, as γ approaches its lower bound, a pair of eigenvalues, one in $\mathbb{C} \setminus \overline{\mathbb{C}}_0$ and another one in $\mathbb{C}_0$, approach the very same point at the imaginary axis. Then the absolute values of those eigenvalues are the very frequencies at which $\|G(\mathrm{j}\omega)\| = \gamma$. It should be emphasized that the mere presence of pure imaginary eigenvalues in H_G, even repeated ones, does not say that $\|G\|_\infty = \gamma$. It just implies that $\|G(\mathrm{j}\omega)\| = \gamma$ at those frequencies. Now, if $\|D\| = \overline{\sigma}(D) = \gamma$, it is still possible that $\|G(\mathrm{j}\omega)\| = \gamma$ at some other frequencies ω (in which case $\operatorname{nrank}(\Phi(s)) = m$) or even at all $\omega \in \mathbb{R}$ (in which case $\operatorname{nrank}(\Phi(s)) < m$, this happens, for example, if $G(s)$ is inner). ∇

Using the result of Proposition 4.18, the H_∞ norm can be found by a bisection algorithm. To this end, we select an upper bound γ_{u} and a lower bound γ_{l} (e.g. $\gamma_{\mathrm{l}} = \overline{\sigma}(D)$) for $\|G\|_\infty$ and then check whether $\|G\|_\infty < (\gamma_{\mathrm{l}} + \gamma_{\mathrm{u}})/2$ (in this case $\gamma_{\mathrm{u}} \to (\gamma_{\mathrm{l}} + \gamma_{\mathrm{u}})/2$) or not (in this case $\gamma_{\mathrm{l}} \to (\gamma_{\mathrm{l}} + \gamma_{\mathrm{u}})/2$). The iterations are repeated until the relative error $1 - \gamma_{\mathrm{l}}/\gamma_{\mathrm{u}} \in (0, 1)$ falls within a required tolerance level.

4.3.5 KYP lemma

The section is concluded with the celebrated KYP (Kalman–Yakubovich–Popov) lemma. It connects a class of frequency-domain inequalities with linear matrix inequalities, aka LMIs.

Theorem 4.19 (KYP). *Consider a $p \times m$ real-rational transfer function $G(s) = D + C(sI - A)^{-1}B$ and a matrix $M_{KYP} = M_{KYP}' \in \mathbb{R}^{(m+p)\times(m+p)}$. If $A \in \mathbb{R}^{n\times n}$ has no pure imaginary eigenvalues, then*

$$\begin{bmatrix} [G(\mathrm{j}\omega)]' & I_m \end{bmatrix} M_{KYP} \begin{bmatrix} G(\mathrm{j}\omega) \\ I_m \end{bmatrix} < 0 \tag{4.28}$$

for all $\omega \in \mathbb{R} \cup \{\pm\infty\}$ iff there is $X = X' \in \mathbb{R}^{n\times n}$ such that

$$\begin{bmatrix} C' & 0 \\ D' & I_m \end{bmatrix} M_{KYP} \begin{bmatrix} C & D \\ 0 & I_m \end{bmatrix} + \begin{bmatrix} I_n & A' \\ 0 & B' \end{bmatrix} \begin{bmatrix} 0 & X \\ X & 0 \end{bmatrix} \begin{bmatrix} I_n & 0 \\ A & B \end{bmatrix} < 0. \tag{4.29}$$

Proof. To simplify the exposition, assume that $C = I$ and $D = 0$. Because

$$\begin{bmatrix} G(\mathrm{j}\omega) \\ I \end{bmatrix} = \begin{bmatrix} D + C(\mathrm{j}\omega I - A)^{-1}B \\ I \end{bmatrix} = \begin{bmatrix} C & D \\ 0 & I \end{bmatrix} \begin{bmatrix} (\mathrm{j}\omega I - A)^{-1}B \\ I \end{bmatrix},$$

this assumption entails no loss of generality, we can always redefine

$$M_{\mathrm{KYP}} \to \begin{bmatrix} C' & 0 \\ D' & I_m \end{bmatrix} M_{\mathrm{KYP}} \begin{bmatrix} C & D \\ 0 & I_m \end{bmatrix} =: \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix}.$$

(4.28) $\Longleftarrow$ (4.29) : It is readily seen that $AG(s) + B = A(sI - A)^{-1}B + B = s(sI - A)^{-1}B$. Hence,

$$\begin{aligned} &\begin{bmatrix} [G(\mathrm{j}\omega)]' & I \end{bmatrix} \begin{bmatrix} I & A' \\ 0 & B' \end{bmatrix} \begin{bmatrix} 0 & X \\ X & 0 \end{bmatrix} \begin{bmatrix} I & 0 \\ A & B \end{bmatrix} \begin{bmatrix} G(\mathrm{j}\omega) \\ I \end{bmatrix} \\ &\qquad = -B'(\mathrm{j}\omega I + A')^{-1} \begin{bmatrix} I & -\mathrm{j}\omega I \end{bmatrix} \begin{bmatrix} 0 & X \\ X & 0 \end{bmatrix} \begin{bmatrix} I \\ \mathrm{j}\omega I \end{bmatrix} (\mathrm{j}\omega I - A)^{-1}B = 0 \end{aligned}$$

for all X. Inequality (4.28) follows then via post- and pre-multiplying (4.29) by $\left[\begin{smallmatrix} G(\mathrm{j}\omega) \\ I \end{smallmatrix}\right]$ and its adjoint, respectively.

(4.28) $\Longrightarrow$ (4.29) : Define

$$\Phi(s) := \begin{bmatrix} G^{\sim}(s) & I \end{bmatrix} M_{\text{KYP}} \begin{bmatrix} G(s) \\ I \end{bmatrix} = \left[\begin{array}{cc|c} A & 0 & B \\ M_{11} & -A' & M_{12} \\ \hline M_{21} & -B' & M_{22} \end{array}\right],$$

where the last equality follows by state-space constructions similar to those used to derive the realization of $\Phi(s)$ in the proof of Proposition 4.18. Inequality (4.28) reads then $\Phi(\mathrm{j}\omega) < 0$, for all ω. This requires $M_{22} < 0$, which is assumed hereafter. Repeating the arguments of the proof of Proposition 4.18, (4.28) implies that $\Phi(s)$ has no pure imaginary zeros, which, together with the assumed property that $\operatorname{spec}(A) \cap \mathrm{j}\mathbb{R} = \varnothing$, is equivalent to the absence of pure imaginary eigenvalues of

$$H_M = \begin{bmatrix} A & 0 \\ M_{11} & -A' \end{bmatrix} - \begin{bmatrix} B \\ M_{12} \end{bmatrix} M_{22}^{-1} \begin{bmatrix} M_{21} & -B' \end{bmatrix} = \begin{bmatrix} A - BM_{22}^{-1}M_{21} & BM_{22}^{-1}B' \\ M_{11} - M_{12}M_{22}^{-1}M_{21} & -A' + M_{12}M_{22}^{-1}B' \end{bmatrix}.$$

This is a Hamiltonian matrix, see (B.11) on 193.

Assume first that (A, B) is stabilizable, so that $(A - BM_{22}^{-1}M_{21}, BM_{22}^{-1}B')$ is stabilizable as well. By Theorem B.6, there is a stabilizing solution $\tilde{X} = \tilde{X}'$ to the Riccati equation

$$(A - BM_{22}^{-1}M_{21})'\tilde{X} + \tilde{X}(A - BM_{22}^{-1}M_{21}) - M_{11} + M_{12}M_{22}^{-1}M_{21} + \tilde{X}BM_{22}^{-1}B'\tilde{X} = 0$$

such that $A_X := A - BM_{22}^{-1}M_{21} + BM_{22}^{-1}B'\tilde{X}$ is Hurwitz. Consider now the Lyapunov equation

$$A_X'Y + YA_X = (A - BM_{22}^{-1}M_{21})'Y + Y(A - BM_{22}^{-1}M_{21}) + YBM_{22}^{-1}B'\tilde{X} + \tilde{X}BM_{22}^{-1}B'Y = -I,$$

which has a unique solution $Y = Y' > 0$. It is readily seen that the matrix $X = Y - \tilde{X}$ satisfies

$$A'X + XA + M_{11} - (XB + M_{12})M_{22}^{-1}(B'X + M_{21}) = -I - YBM_{22}^{-1}B'Y < 0.$$

By Lemma B.8 together with the fact that $M_{22} < 0$, the latter inequality is equivalent to

$$\begin{bmatrix} A'X + XA + M_{11} & XB + M_{12} \\ B'X + M_{21} & M_{22} \end{bmatrix} = M_{\text{KYP}} + \begin{bmatrix} I_n & A' \\ 0 & B' \end{bmatrix} \begin{bmatrix} 0 & X \\ X & 0 \end{bmatrix} \begin{bmatrix} I_n & 0 \\ A & B \end{bmatrix} < 0,$$

i.e. we did find $X = X'$ such that (4.29) holds true.

If (A, B) is not stabilizable, we can construct a similar realization of G of a form similar to that in (4.12), but with only unstabilizable modes isolated. In this case a sought X can be chosen to be block diagonal with the block corresponding to the stabilizable part constructed as above and that corresponding to the unstabilizable part, $A_{\bar{\mathrm{s}}}$, which does not affect (4.28), constructed via the Lyapunov equation $A_{\bar{\mathrm{s}}}'X_2 + X_2A_{\bar{\mathrm{s}}} = -\alpha I$ for a sufficiently large $\alpha > 0$. The details are left as an exercise.

Note that the non-strict counterparts of inequalities (4.28) and (4.29) are also equivalent, provided (A, B) is controllable, see [22] for a proof. □

Theorem 4.19 establishes that an infinite set of inequalities based on the frequency response of a real-rational system can be verified via a finite number of LMIs. This is an important relation, both from the numerical and conceptual viewpoints. Linear-matrix inequalities are convex and can be efficiently solved via interior-point and bundle methods, with plenty of software available. The reduction to matrix inequalities in terms of state-space realizations also gives a valuable insight into properties of frequency responses and their connections with time-domain properties of dynamical systems. LMIs can then be connected with algebraic Riccati equations, facilitates a deeper analysis of their properties.

The KYP lemma is a fairly general result. For instance, it includes the H_∞-norm bound as its special case with the choice

$$M_{\text{KYP}} = \begin{bmatrix} I_p & 0 \\ 0 & -\gamma^2 I_m \end{bmatrix},$$

for which the inequality in (4.28) reads $\gamma^2 I - [G(\mathrm{j}\omega)]' G(\mathrm{j}\omega) > 0$, for all ω. If, in addition, A is Hurwitz, then this inequality is equivalent to $\|G\|_\infty < \gamma$. With this choice of M_{KYP} inequality (4.29) reads

$$\begin{bmatrix} A'X + XA + C'C & XB + C'D \\ B'X + D'C & D'D - \gamma^2 I \end{bmatrix} < 0. \tag{4.29$_{\text{BRL}}$}$$

In fact, it can be shown that $G \in RH_\infty$ and $\|G\|_\infty < \gamma$ iff there is $X = X' > 0$ such that the inequality above holds. In other words, there is no need to assume that A is Hurwitz if the positive definiteness of X is required. This result is known as the *bounded-real lemma*. The equivalence of this condition and the condition of Proposition 4.18 can be seen in the proof of the KYP lemma. One more equivalent form can be stated as the solvability of the ARE

$$A'X + XA + C'C + (XB + C'D)(\gamma^2 I - D'D)^{-1}(B'X + D'C) = 0$$

under $\overline{\sigma}(D) < \gamma^2$, which can be viewed via considering the Schur complement of the $(2,2)$ element of (4.29$_{\text{BRL}}$).

Another important particular case of Theorem 4.19 corresponds to the choice

$$M_{\text{KYP}} = -\begin{bmatrix} 0 & I_m \\ I_m & 0 \end{bmatrix}$$

under $p = m$, which turns (4.28) into the equality $G(\mathrm{j}\omega) + [G(\mathrm{j}\omega)]' > 0$ for all ω. This is a MIMO counterpart of the condition $\operatorname{Re} G(\mathrm{j}\omega) > 0$, i.e. that the frequency response $G(\mathrm{j}\omega)$ is located entirely in the right half-plane of the Nyquist plot. The LMI (4.29) then turns

$$\begin{bmatrix} A'X + XA & XB - C' \\ B'X - C & -D - D' \end{bmatrix} < 0, \tag{4.29$_{\text{PRL}}$}$$

with $X = X' > 0$ only if A is Hurwitz. This is known as the *strict positive-real lemma*. This property is related to the *passivity* property of $L_2^m \to L_2^m$ systems, which reads as the condition that $\langle Gu, u\rangle_2 > 0$ for all nonzero $u \in L_2^m$, and is important in many applications. For instance, passive systems constitute a convenient class for feedback stabilization as a feedback interconnection of two passive systems is stable, see Theorem 6.2 on p. 115, even in the nonlinear case. This property is routinely used in the control of flexible mechanical systems, telerobotics, adaptive control, et cetera.

Remark 4.6 (finite-frequency KYP lemma). If the inequality in (4.28) is only required to hold for $|\omega| \le \omega_0$ with a given frequency $\omega_0 > 0$, then inequality (4.29) should be replaced with

$$\begin{bmatrix} C' & 0 \\ D' & I_m \end{bmatrix} M_{\text{KYP}} \begin{bmatrix} C & D \\ 0 & I_m \end{bmatrix} + \begin{bmatrix} I_n & A' \\ 0 & B' \end{bmatrix} \begin{bmatrix} \omega_0^2 Y & X \\ X & -Y \end{bmatrix} \begin{bmatrix} I_n & 0 \\ A & B \end{bmatrix} < 0 \tag{4.30}$$

for some $X = X'$ and $Y = Y' > 0$. This is still an LMI, in both its parameters, so it can be efficiently solved numerically. This neat result was proved by Iwasaki, Meinsma, and Fu [11], so perhaps it should be referred to as the *IMF lemma*. ▽

4.4 Model order reduction by balanced truncation

A tradeoff between accuracy and complexity is one of fundamental dilemmas in engineering in general and in control in particular. On the one hand, complex models might be required to describe physical phenomena accurately and complex (high-order) controller might be needed to satisfy required performance specifications. On the other hand, complex models are harder to deal with and high-order controllers are harder to implement, because their implementation might demand more expensive hardware and be less reliable comparing with low-order controllers. It is thus important to have methods, enabling us to replace complex models with their simpler approximations without compromising performance too much. Having such methods, one can either reduce the complexity of the plant prior to controller design or reduce that of the resulted controller (or both).

In the realm of LTI systems, complexity is almost exclusively measured by the order of corresponding models[2]. Accordingly, complexity reduction is known as the *model order reduction*. An abstract order reduction problem can be posed as follows:

- given an n-order $p \times m$ LTI system G and $n_{\mathrm{r}} < n$, find an n_{r}-order $p \times m$ LTI system G_{r}, which is "close" to G.

Throughout this section, the "closeness" of G and G_{r} will be measured by the H_∞ norm of the difference between their transfer functions, i.e. $\|G - G_{\mathrm{r}}\|_\infty$. Of course, any other norm and some other measures of the difference between G and G_{r} may make sense depending on the situation.

Performance-wise, a natural approach to derive a reduced-order G_{r} would be to solve $\min_{G_{\mathrm{r}}} \|G - G_{\mathrm{r}}\|_\infty$. However, there might still be no computationally reliable solution to this problem in general. For that reason, the treatment in this section will be based on simpler mode-truncation arguments. These arguments may also we viewed as an example of the use of abstract results discussed in Section 4.2 to a more concrete application.

4.4.1 How minimal is minimal realization

Some poles of transfer functions might affect input/output relations in the system less than others. Classical control employs the notion of *dominance* to explain this situation. Intuitively, poles that are far left in the complex plane or those nearly canceled by zeros should have a limited effect on the system properties. The following simple examples support this intuition quantitatively:

Example 4.4. Consider

$$G(s) = \frac{1}{(s+1)(\tau s+1)} \quad \text{for some } \tau \in (0,1).$$

This transfer function has degree 2. However, if $\tau \ll 1$, the pole of $G(s)$ at $s = -1/\tau$ becomes almost unnoticeable relatively to that at $s = -1$. Hence, the first-order transfer function $G_{\mathrm{r}}(s) = 1/(s+1)$ can be expected to be an accurate representation of the system. This can be seen from

$$G(s) - G_{\mathrm{r}}(s) = -\frac{\tau s}{(s+1)(\tau s+1)} \quad \Longrightarrow \quad \|G - G_{\mathrm{r}}\|_\infty = \frac{\tau}{1+\tau},$$

which vanishes as $\tau \to 0$. ◊

Example 4.5. Consider

$$G(s) = \frac{2s+1}{(s+1)((2-\epsilon)s+1)} \quad \text{for some } \epsilon \in (0,1).$$

[2]In the nonlinear case one may also be interested in limiting the complexity of involved static nonlinear functions.

This transfer function has degree 2 and its pole at $1/(\epsilon-2) \in (-1,-1/2)$ is closer to the origin than that at -1. However, if $\epsilon \ll 1$, then the former pole is almost canceled by the zero at $s=-1/2$. Not surprisingly, the reduced order approximation $G_{\mathrm{r}}(s) = 1/(s+1)$ is accurate for small ϵ, as can be seen from

$$G(s) - G_{\mathrm{r}}(s) = \frac{\epsilon s}{(s+1)((2-\epsilon)s+1)} \qquad \Longrightarrow \qquad \|G - G_{\mathrm{r}}\|_\infty = \frac{\epsilon}{3-\epsilon},$$

which also vanishes as $\epsilon \to 0$. ◊

A naïve approach to the order reduction of LTI systems would be to look for "far left" poles or "closely" located poles and zeros and cancel them. This is straightforward, at least conceptually, in the SISO case. However, in the MIMO case the directions of poles and zeros should also be taken into account, we already know that poles and zeros at the same location do not necessarily cancel each other. The need to account for pole / zero directions would complicate the procedure considerably. In the case of near cancellations, this difficulty can be circumvented by considering alternative indications of the "closeness" between poles and zeros. For example, motivated by the discussed in Section 4.2.3 we may expect that the effect of "almost" hidden (i.e. "almost" uncontrollable and / or unobservable) modes on the properties of G is negligible. This observation can be supported by the examples below:

Example 4.6. Consider the system from Example 4.5. Its possible state-space realization is

$$G(s) = \left[\begin{array}{cc|c} -(3-\epsilon)/(2-\epsilon) & 1 & 2/(2-\epsilon) \\ -1/(2-\epsilon) & 0 & 1/(2-\epsilon) \\ \hline 1 & 0 & 0 \end{array}\right].$$

This realization is minimal and its observability matrix $M_{\mathrm{o}} = \left[\begin{smallmatrix} 1 & 0 \\ -(3-\epsilon)/(2-\epsilon) & 1 \end{smallmatrix}\right]$ is well-conditioned. Yet its controllability matrix,

$$M_{\mathrm{c}} = \frac{1}{(2-\epsilon)^2}\begin{bmatrix} 4-2\epsilon & -4+\epsilon \\ 2-\epsilon & -2 \end{bmatrix},$$

is almost singular for small ϵ, because $\det(M_{\mathrm{c}}) = -\epsilon/(2-\epsilon)^3$. ◊

This example suggests that the presence of "almost" hidden modes may indicate the presence of closely located poles and zeros which are "almost" canceled. The reduction of such modes should not then incur significant errors. Natural questions then are how such "almost" hidden modes can be discovered in state-space realizations and how they can be reduced?

To answer these questions, return to Proposition 4.5, which shows that non-controllable modes and their directions can be seen from the null space of the associated controllability Gramian P. Extending this reasoning to the case of a near-singular P seems to be natural. Indeed, assume that $P > 0$, i.e. that the realization is controllable, but some of its singular values, say $\sigma_{n_{\mathrm{r}}+1}$ to σ_n are "almost" zero. This means, that by partitioning the state-space realization of G accordingly, i.e. as

$$G(s) = \left[\begin{array}{c:c|c} A_{11} & A_{12} & B_1 \\ \hdashline A_{21} & A_{22} & B_2 \\ \hline C_1 & C_2 & D \end{array}\right]$$

with $A_{11} \in \mathbb{R}^{n_{\mathrm{r}} \times n_{\mathrm{r}}}$, we should end up with "almost zero" A_{21} and B_2. One might then be tempted to eliminate these "almost uncontrollable" modes and would expect that the resulting n_{r}-order transfer function

$$G_{\mathrm{r}}(s) = \left[\begin{array}{c|c} A_{11} & B_1 \\ \hline C_1 & D \end{array}\right]$$

is close to $G(s)$. Yet the result might fall short of this expectation as shown below.

Example 4.7. Consider the transfer function

$$G(s) = \frac{18}{5s^2 + 12s + 9} = \left[\begin{array}{cc|c} -2 & -1/\alpha & 1 \\ \alpha & -2/5 & \alpha \\ \hline -1 & 1/\alpha & 0 \end{array}\right],$$

which is true for all $\alpha \neq 0$. The controllability Gramian of this realization is

$$P = \mathrm{diag}\{0.25, 1.25\alpha^2\}.$$

Thus, this realization can be made almost uncontrollable by selecting α small enough. Moreover, the subblocks $A_{21} = B_2 = \alpha$ are then also small. Yet the truncation of the second state yields the reduced-order transfer function

$$G_{\mathrm{r}}(s) = \left[\begin{array}{c|c} -2 & 1 \\ \hline -1 & 0 \end{array}\right] = -\frac{1}{s+2},$$

which is anything but a good approximation of $G(s)$ since

$$\|G - G_{\mathrm{r}}\|_\infty = 2.5 > \|G - 0\|_\infty = \|G\|_\infty = 2.$$

The source of the problem becomes apparent when we check the observability Gramian, which is

$$Q = \mathrm{diag}\{0.25, 1.25/\alpha^2\}.$$

This shows that for small α the second state becomes "over-observable," in a sense. ◊

Example 4.7 shows clearly that the controllability (or observability) Gramian alone cannot serve as an accurate indication of the relative importance of the system modes in the input/output behavior. Yet Example 4.7 also suggests a remedy. Indeed, if the "degrees" of controllability and observability of each mode were *balanced* (equalized), then the situation above would never occur. Such a balancing is indeed possible and is studied in the next subsection.

4.4.2 Balanced realization and Hankel singular values

Consider the state-space realization (4.5) and denote by P and Q its controllability and observability Gramians, respectively. Consider also another realization of G, $(\tilde{A}, \tilde{B}, \tilde{C}, D) := (TAT^{-1}, TB, CT^{-1}, D)$, where T is a nonsingular similarity transformation matrix, which has $\tilde{P}$ and $\tilde{Q}$ as its Gramians. As follows from (4.11) and (4.13) on pp. 72 and 74, respectively, the following relations take place:

$$\tilde{P} = TPT' \quad \text{and} \quad \tilde{Q} = T^{-\prime}QT^{-1}.$$

The eigenvalues (and, therefore, the singular values) of both Gramians are not preserved under similarity transformations. This fact confirms the previous conclusion that the singular values of either P or Q alone cannot be used to decide whether a part of the system dynamics is negligible or not. Yet further inspection shows that

$$\tilde{P}\tilde{Q} = T(PQ)T^{-1},$$

i.e. the spectrum of the product of the controllability and observability Gramians *is* invariant under similarity transformations.

Because $\mathrm{spec}(PQ) = \mathrm{spec}(Q^{1/2}PQ^{1/2})$, we may expect that PQ is also diagonalizable by an appropriate similarity transformation. This is indeed true. To see this, let the realization (4.5) be minimal, so that both P and Q are nonsingular. Because these matrices are also symmetric, there are unitary U_{c} and

U_o such that $P = U_c \Sigma_c U_c'$ and $Q = U_o \Sigma_o U_o'$ for some diagonal $\Sigma_c > 0$ and $\Sigma_o > 0$. Construct the nonsingular $H = \Sigma_o^{1/2} U_o' U_c \Sigma_c^{1/2}$ and bring in its SVD, $H = U_H \Sigma_H V_H'$. Then, defining

$$T_H = \Sigma_H^{-1/2} U_H' \Sigma_o^{1/2} U_o', \tag{4.31}$$

it is a matter of straightforward algebra so see that

$$T_H P Q T_H^{-1} = \Sigma_H^2.$$

Note that similar result can be obtained in the case when the realization (4.5) is not minimal, yet the derivations are more involved then.

Let $\Sigma_H = \mathrm{diag}\{\sigma_1 I_{n_1}, \ldots, \sigma_l I_{n_l}\}$ for some $\sigma_1 > \cdots > \sigma_l \geq 0$ and $\sum_{i=1}^{l} n_i = n$. The numbers σ_i, $i = 1, \ldots, l$, are called the *Hankel singular values* of the system. The maximal of the Hankel singular values, $\sigma_1 = \sqrt{\rho(PQ)}$, is called the *Hankel norm* of G and is denoted as $\|G\|_{\mathrm{H}}$. This is actually the induced norm of the Hankel operator $\mathfrak{H}_G : L_2(\mathbb{R}_-) \to L_2(\mathbb{R}_+)$ associated with the system. This operator was defined at the beginning of Section 4, see also §B.1.2 on p. 190 for more details on the Hankel norm.

In the context of model reduction, the Hankel singular values do reflect the relative importance of system modes. "Small" σ_i implies that n_i modes of G almost do not affect the system and thus can be eliminated. Yet the knowledge of σ_i is not sufficient to propose a constructive model reduction procedure. The modes corresponding to small Hankel singular values should be detected as well. To this end the following result is important.

Theorem 4.20. *Let $G \in RH_\infty$ and* (4.5) *be its minimal realization. There exists a state transformation T such that the Gramians P and Q of the realization $(TAT^{-1}, TB, CT^{-1}, D)$ satisfies*

$$P = Q = \Sigma := \mathrm{diag}\{\sigma_1 I_{n_1}, \ldots, \sigma_l I_{n_l}\},$$

where $\sigma_1 > \cdots > \sigma_l > 0$ are the Hankel singular values of G.

Proof. Take $T = T_H$, where T_H is defined by (4.31). Then the equality above can be verified by the direct substitution. □

The realization of Theorem 4.20 is called the *balanced realization*, reflecting the fact that the controllability and observability Gramians are equally emphasized. Consequently, the modes corresponding to small diagonal elements of P can now be regarded as less important and, therefore, as elimination candidates.

Example 4.8. Consider the transfer function from Example 4.7. It is clear from the Gramians calculated there that its realization proposed in Example 4.7 becomes balanced if $\alpha = 1$ and the components of the state vector are permutated, i.e.

$$G(s) = \left[\begin{array}{cc|c} -2/5 & 1 & 1 \\ -1 & -2 & 1 \\ \hline 1 & -1 & 0 \end{array}\right]$$

The corresponding Gramians $P = Q = \mathrm{diag}\{1.25, 0.25\}$. As a matter of fact, removing the less dominant second state results now in $G_r(s) = 1/(s + 0.4)$, for which $\|G - G_r\|_\infty = 0.5$. ◊

4.4.3 Balanced truncation

Consider now a stable G and assume that its realization of the form (4.5) is balanced as described in Theorem 4.20. Motivated by Example 4.8, the modes corresponding to the smallest Hankel singular values are natural elimination candidates. However, there are a couple of points we have to carry about. First, we

definitely want the truncated system to be stable too. So the question is whether stability of the reduced-order model can be guaranteed and under what conditions. Second, in order to decide how many modes can be truncated without incurring significant errors, we would like to have some a priori bounds on the model reduction error. These issues are addressed in the following result, whose proof is omitted.

Theorem 4.21. *Let G be a stable system, whose balanced realization is partitioned as*

$$G(s) = \left[\begin{array}{cc|c} A_{11} & A_{12} & B_1 \\ A_{21} & A_{22} & B_2 \\ \hline C_1 & C_2 & D \end{array}\right]$$

and the Gramians $P = Q = \Sigma = \operatorname{diag}\{\Sigma_{11}, \Sigma_{22}\}$ so that

$$\Sigma_{11} := \operatorname{diag}\{\sigma_1 I_{n_1}, \ldots, \sigma_r I_{n_r}\} \quad \textit{and} \quad \Sigma_{22} := \operatorname{diag}\{\sigma_{r+1} I_{n_{r+1}}, \ldots, \sigma_l I_{n_l}\}$$

for $\sigma_1 > \cdots > \sigma_r > \sigma_{r+1} > \cdots > \sigma_l$. The truncated system G_r with the transfer function

$$G_r(s) := \left[\begin{array}{c|c} A_{11} & B_1 \\ \hline C_1 & D \end{array}\right]$$

is balanced, with $P_r = Q_r = \Sigma_{11}$, stable, and such that

$$\|G - G_r\|_\infty \le 2(\sigma_{r+1} + \cdots + \sigma_l). \tag{4.32}$$

Moreover, if $r = l - 1$, then the bound is achieved, i.e. $\|G - G_{l-1}\|_\infty = 2\sigma_l$.

Note, that Σ_{11} and Σ_{22} in Theorem 4.21 should not have diagonal elements in common. This is a key limitation to guarantee the stability of the truncated system as can be seen from the following example:

Example 4.9. Consider the (inner) transfer function

$$G(s) = \frac{(s-1)^2}{(s+1)^2} = \left[\begin{array}{cc|c} -1+\cos 2\theta & 1-\sin 2\theta & 2\sin\theta \\ -1-\sin 2\theta & -1-\cos 2\theta & 2\cos\theta \\ \hline -2\sin\theta & -2\cos\theta & 1 \end{array}\right].$$

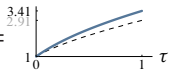
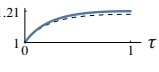
The realization above is balanced, with $\Sigma = I_2$, for all θ. Its "A_{11}" part is unstable whenever $\cos 2\theta = 1$ or, equivalently, $\theta = \pi k$. ◊

It may be of interest to return to the first two examples of §4.4.1 and see how the balanced truncation approach compares with ad hoc cancellations.

Example 4.10. The balanced realization of the system studied in Example 4.4 is

$$G(s) = \frac{1}{(s+1)(\tau s+1)} = \left[\begin{array}{cc|c} -(1-(1+\tau)/\alpha)(\tau+1)/(2\tau) & 2/\alpha & 1/\sqrt{\alpha} \\ -2/\alpha & (1+(1+\tau)/\alpha)(\tau+1)/(2\tau) & 1/\sqrt{\alpha} \\ \hline 1/\sqrt{\alpha} & -1/\sqrt{\alpha} & 0 \end{array}\right],$$

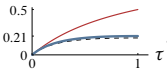
where $\alpha := \sqrt{\tau^2 + 6\tau + 1}$, and the Gramians $P = Q = 0.25\operatorname{diag}\{\alpha/(\tau+1)+1, \alpha/(\tau+1)-1\}$. Thus, the first-order approximation obtained by the balanced truncation is

$$G_r(s) = \frac{k_1}{\tau_1 s + 1}, \quad \text{where } \tau_1 = \frac{\alpha}{2}\left(\frac{\alpha}{\tau+1} + 1\right) =$$  $$\text{and } k_1 = \frac{\tau_1}{\alpha} =$$ 

in this case. Unlike the mechanical elimination of non-dominant modes, the balanced truncation procedure changes both the time constant of the dominant mode and the static gain of the resulted system, especially

when τ is not close to zero. The increase of the time constant is perfectly logical, an extra pole is known to slow down the response. The logic behind the increase of the static gain is not evident. Still, the H_∞ optimal approximation of this G having the same structure as the balanced truncation, which can be found via a brute-force parametric search and whose parameters are shown by the dashed lines above, exhibits the same trend. The advantage of adjusting the time constant and the gain of the approximant can be seen via the resulting approximation error, which is twice the smallest Hankel singular value, i.e.

$$\|G - G_{\mathrm{r}}\|_\infty = \frac{1}{2}\left(\frac{\sqrt{\tau^2 + 6\tau + 1}}{\tau + 1} - 1\right) =$$



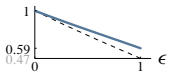
The red line above shows the error for the ad hoc truncation, which was derived in Example 4.4. We can see a substantial improvement of the approximation accuracy of the balanced truncation procedure for every τ in this case. The optimal approximation error (dashed line) is only slightly smaller than that attained via balanced truncation. ◊

Example 4.11. Finding an analytic expression of the balanced realization of the system of Example 4.5, which is

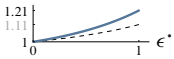
$$G(s) = \frac{2s + 1}{(s + 1)((2 - \epsilon)s + 1)},$$

is more involved. For that reason, only numerical results of the balanced truncation of this system to its first-order approximation are presented:

$$G_{\mathrm{r}}(s) = \frac{k_2}{\tau_2 s + 1}, \quad \text{where } \tau_2 =$$

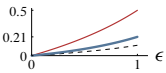


$$\text{and } k_2 =$$



We can again see that the balanced truncation changes both the remaining time constant and the static gain. The time constant decreases now, which is again logical. Indeed, the original system has a second pole, which is slower than the first one at $s = -1$, and a zero at $s = -1/2$, which is even slower. Thus, the zero is more dominant than either of the poles. But the addition of a zero renders the response quicker, which is reflected in $\tau_2 < 1$. The optimal H_∞ approximation (dashed lines) shows similar trends. And, like the previous example, the balanced truncation outperforms the cancellation of close pole and zero:

$$\|G - G_{\mathrm{r}}\|_\infty = \sqrt{\frac{\epsilon^2 - 4\epsilon + 9 - (3 - \epsilon)\sqrt{\epsilon^2 - 2\epsilon + 9}}{2(3 - \epsilon)^2}} =$$



The error derived in Example 4.5 for the cancellation case is shown in the red line above and the optimal error is shown by the dashed line. ◊

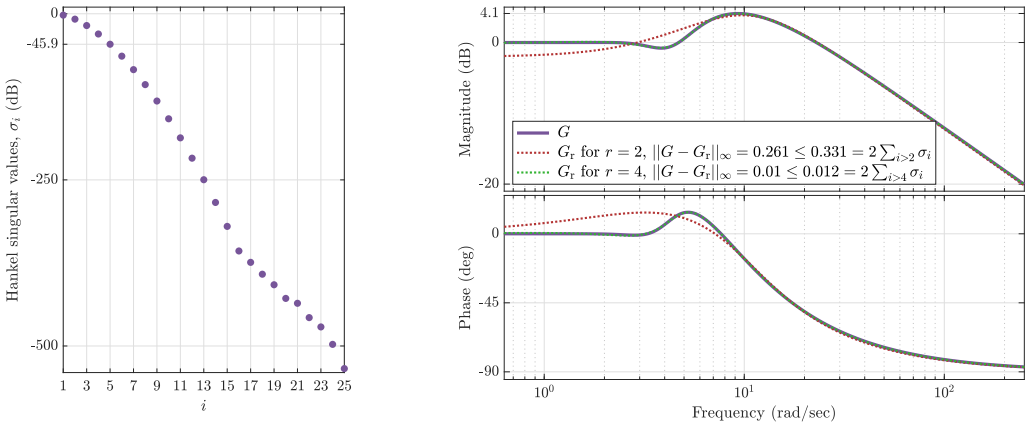
Let us conclude with another example, which illustrate the potential power of the balanced truncation method in simplifying high-order systems.

Example 4.12. Consider a 25-order stable system G with the strictly proper transfer function

$$G(s) = 1 - \left(\frac{s + 1}{s + 2}\right)^{25},$$

having $\|G\|_\infty = 1.6021$. Its Hankel singular values are all different and depicted in Fig. 4.1(a), in dB. It is clearly seen that only a few first Hankel singular values are noticeable, while the other are extremely small. This implies that their truncation shall not visibly affect G. This is indeed the case, as can be seen in Fig. 4.1(b), where the Bode plots of the balanced truncations

$$G_2(s) = \frac{24.513(s + 3.568)}{s^2 + 16.69s + 110.5} \quad \text{and} \quad G_4(s) = \frac{24.986(s + 3.196)(s^2 + 3.165s + 19.48)}{(s^2 + 5.629s + 24.58)(s^2 + 14.69s + 63.86)}$$



(a) Hankel singular values of G

(b) Frequency responses of $G(s)$ and its approximations

Fig. 4.1: Plots for Example 4.12

are depicted by dotted lines. The second-order approximation is not quite accurate, which could be seen from the singular values in Fig. 4.1(a). The fourth-order approximation is very close to the full-order G. The Bode plots of the fifth-order approximation would be indistinguishable from those of G. Note that the actual values of $\|G - G_r\|_\infty$ are smaller than their upper bounds calculated by (4.32). ◊

Remark 4.7 (Hankel norm approximation). It is possible to derive a reduced-order approximation of G by minimizing the Hankel norm of the mismatch $G - G_r$. This problem is solvable and guarantees that the H_∞-norm of the resulted approximation error satisfies

$$\|G - G_r\|_\infty \leq \sigma_{r+1} + \cdots + \sigma_l,$$

i.e. the bound is a half of that of the balanced truncation in (4.32). Obtaining the optimal approximation in this case is more numerically involved than truncating the balanced realization and normally results in a bi-proper G_r. The latter is different from the H_∞ optimal results presented in Examples 4.10 and 4.11, where the strictly proper structure $k/(\tau s + 1)$ was enforced. ▽

Part II

Interconnected Systems

Chapter 5

Interactions Between Systems

To alter the behavior of a system one can connect it with other systems. This approach is in the core of many engineering fields, including, of course, control engineering, where the very goal is to change behaviors. This chapter aims at studying effects of systems interconnections on their properties.

5.1 Basic interconnections and cancellations

Parallel, cascade (series), and feedback interconnections shown in Fig. 5.1 on the next page constitute basic building blocks of systems interactions. The first two were discussed in §4.1.1, although more as a technical tool used later on. In this section we are concerned with properties of systems resulting from these three interconnection of $p_1 \times m_1$ and $p_2 \times m_2$ systems G_1 and G_2 in terms of their *minimal* realizations

$$G_1(s) = \left[\begin{array}{c|c} A_1 & B_1 \\ \hline C_1 & D_1 \end{array}\right] \quad \text{and} \quad G_2(s) = \left[\begin{array}{c|c} A_2 & B_2 \\ \hline C_2 & D_2 \end{array}\right] \tag{5.1}$$

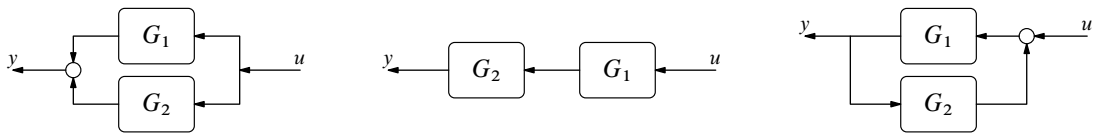
of orders n_1 and n_2, respectively. Intuitively, dynamics of interconnections should be affected by dynamics of each component. A quantitative expression of this would be the preservation of dimensions. In other words, we would expect that the order of interconnections equals the sum of the orders of its components, $n_1 + n_2$. Yet this is not always the case. Situations when the order of an interconnection is strictly less than the sum of the orders of its components are referred to as *cancellations* hereafter. These may be pole-zero cancellations, similar to their SISO counterparts, but may also be cancellations related to a general deficiency of components, like a deficient normal rank. Cancellations are the focus point of this section.

5.1.1 Parallel interconnection

We start with the system $G : u \mapsto y$ in Fig. 5.1(a), for which $m_1 = m_2$ and $p_1 = p_2$ should be assumed. The transfer function of G in terms of its realization is

$$G(s) = \left[\begin{array}{cc|c} A_1 & 0 & B_1 \\ 0 & A_2 & B_2 \\ \hline C_1 & C_2 & D_1 + D_2 \end{array}\right], \tag{5.2}$$

see (4.6). Poles of the realization above are obviously the union of those of the realizations of G_1 and G_2 in (5.1). The question is whether all realization poles are also those of the transfer function $G(s)$. If this is not the case, i.e. if the realization in (5.2) is not minimal, we say that there are cancellations in the interconnection. In the SISO case, cancellations take place whenever $G_1(s)$ and $G_2(s)$ have common poles. In MIMO systems, pole directions play role as well, so the analysis should be more involved. We address cancellations via the controllability and observability properties of (5.2).



(a) parallel, $G_1 + G_2$ (b) cascade, $G_2 G_1$ (c) feedback, $G_1(I - G_2 G_1)^{-1}$

Fig. 5.1: Basic system interconnections

Consider the observability of the realization in (5.2) (controllability is addressed by similar arguments). If it is lost, there should be $\lambda \in \mathbb{C}$ and $\left[\begin{smallmatrix}\eta_1\\ \eta_2\end{smallmatrix}\right] \neq 0$ such that

$$\begin{bmatrix} A_1 - \lambda I & 0 \\ 0 & A_2 - \lambda I \\ C_1 & C_2 \end{bmatrix} \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} (A_1 - \lambda I)\eta_1 \\ (A_2 - \lambda I)\eta_2 \\ C_1\eta_1 + C_2\eta_2 \end{bmatrix} = 0.$$

First, note that $\eta_i \neq 0$, which follows by the assumed observability of (C_i, A_i). Then the first two block rows above imply that λ must be an eigenvalue of both A_1 and A_2 and that η_i must be the corresponding eigenvectors of A_i. As a matter of fact, this means that cancellations in the parallel interconnection can take place only if $G_1(s)$ and $G_2(s)$ have common poles. Now, again by the minimality of the realizations of G_i in (5.1) we know that $C_i\eta_i \neq 0$. Hence, observability is lost iff $C_1\eta_1$ and $C_2\eta_2$ are co-directed for some eigenvectors η_i of A_i. This, in turn, is possible iff

$$C_1 \ker(\lambda I - A_1) \cap C_2 \ker(\lambda I - A_2) \neq \{0\},$$

which is the condition under which (5.2) is not observable. The arguments above, combined with (4.23), can be summarized as follows.

Proposition 5.1. *Suppose that both* (A_1, B_1, C_1, D_1) *and* (A_2, B_2, C_2, D_2) *are minimal. The realization in* (5.2) *is controllable iff*

$$\mathrm{pdir_i}(G_1, \lambda) \cap \mathrm{pdir_i}(G_2, \lambda) = \{0\}$$

and is observable iff

$$\mathrm{pdir_o}(G_1, \lambda) \cap \mathrm{pdir_o}(G_2, \lambda) = \{0\},$$

both for all $\lambda \in \mathrm{spec}(A_1) \cap \mathrm{spec}(A_2)$.

5.1.2 Cascade interconnection

Consider the system $G : u \mapsto y$ in Fig. 5.1(b), for which $p_1 = m_2$ should be assumed. The transfer function of G in terms of its realization is

$$G(s) = \left[\begin{array}{cc|c} A_1 & 0 & B_1 \\ B_2C_1 & A_2 & B_2D_1 \\ \hline D_2C_1 & C_2 & D_2D_1 \end{array}\right], \tag{5.3}$$

see (4.7). Poles of the realization above are again the union of those of the realizations of G_1 and G_2 in (5.1) and we say that there are no cancellations in the cascade interconnection if the poles of the transfer function $G(s)$ are the union of the poles of $G_1(s)$ and $G_2(s)$. Cancellations in the SISO case take place if, and only if, poles of $G_1(s)$ match zeros of $G_2(s)$ or vice versa. Expectably, the MIMO pole-zero cancellations should be affected by the directions of potentially canceled poles and zeros. On top of this, MIMO cancellations might not be related to zeros. For example, let $G_1(s) = \left[\begin{smallmatrix} 1/s & 0 \\ 0 & 1/s \end{smallmatrix}\right]$ and $G_2(s) = \left[\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right]$, which are second- and

zero-order transfer functions, respectively. Being static, $G_2(s)$ has no zeros at all. Nevertheless, $G(s) = G_2(s)G_1(s) = \left[\begin{smallmatrix} 1/s & 1/s \\ 1/s & 1/s \end{smallmatrix}\right]$ is a first-order system, we saw that in Example 3.1 on p. 57. This happened because $G_2(s)$ has a normal rank deficiency, a phenomenon having no SISO counterpart.

So, consider again the observability of the interconnected realization in (5.3). If it is lost, there should be $\lambda \in \mathbb{C}$ and $\left[\begin{smallmatrix} \eta_1 \\ \eta_2 \end{smallmatrix}\right] \neq 0$ such that

$$0 = \begin{bmatrix} A_1 - \lambda I & 0 \\ B_2 C_1 & A_2 - \lambda I \\ D_2 C_1 & C_2 \end{bmatrix} \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} 0 & I & 0 \\ A_2 - \lambda I & 0 & B_2 \\ C_2 & 0 & D_2 \end{bmatrix} \begin{bmatrix} 0 & I \\ A_1 - \lambda I & 0 \\ C_1 & 0 \end{bmatrix} \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix}$$

or, equivalently,

$$(A_1 - \lambda I)\eta_1 = 0 \quad \text{and} \quad \begin{bmatrix} A_2 - \lambda I & B_2 \\ C_2 & D_2 \end{bmatrix} \begin{bmatrix} \eta_2 \\ C_1 \eta_1 \end{bmatrix} = 0.$$

First, $\eta_1 \neq 0$, for otherwise $\eta_2 = 0$ too by the assumed observability of (C_2, A_2). Hence, λ must be an eigenvalue of A_1 and then $C_1\eta_1 \in C_1 \ker(\lambda I - A_1)$ must be a component in the kernel of the Rosenbrock system matrix $R_{G_2}(\lambda)$. Because (C_2, A_2) is observable, we conclude that $0 \neq C_1\eta_1 \in \operatorname{zdir_i}(G_2, \lambda)$.

The arguments above, combined with (4.23) and (4.27), can be summarized in the following result.

Proposition 5.2. *Suppose that both* (A_1, B_1, C_1, D_1) *and* (A_2, B_2, C_2, D_2) *are minimal. The realization in* (5.3) *is controllable iff*

$$\operatorname{pdir_i}(G_2, \lambda) \cap \operatorname{zdir_o}(G_1, \lambda) = \{0\}$$

for all $\lambda \in \operatorname{spec}(A_2)$ *and is observable iff*

$$\operatorname{zdir_i}(G_2, \lambda) \cap \operatorname{pdir_o}(G_1, \lambda) = \{0\},$$

for all $\lambda \in \operatorname{spec}(A_1)$.

It should be emphasized, again, that $\operatorname{zdir_o}(G_1, \lambda)$ and $\operatorname{zdir_i}(G_2, \lambda)$ might be nontrivial even if λ is not a zero of G_1 and G_2, respectively. If $\operatorname{nrank}(R_{G_1}(s)) < n_1 + p_1$, be it because $m_1 < p_1$ or because of its normal rank deficiency, $\operatorname{zdir_o}(G_1, s)$ is nontrivial for all s. Likewise, if $\operatorname{nrank}(R_{G_2}(s)) < n_2 + m_2$ for whatever reason, $\operatorname{zdir_i}(G_2, s)$ is nontrivial for all s too. This is why we refer to the phenomenon of $G(s)$ having its McMillan degree below $n_1 + n_2$ as just "cancellations," rather than "pole-zero cancellations."

5.1.3 Feedback interconnection

Now consider the system $G : u \mapsto y$ in Fig. 5.1(c), for which $p_1 = m_2$ and $p_2 = m_1$ should be assumed. The positive feedback can be considered without loss of generality, the negative feedback corresponds to the replacement $G_2 \to -G_2$. We say that this interconnection is *well posed* if y exists and is unique for all u. The derivation of the realization of this interconnection should start with the realizations of its components,

$$G_1 : \begin{cases} \dot{x}_1(t) = A_1 x_1(t) + B_1 u_1(t) \\ y_1(t) = C_1 x_1(t) + D_1 u_1(t) \end{cases} \qquad \text{and} \qquad G_2 : \begin{cases} \dot{x}_2(t) = A_2 x_2(t) + B_2 u_2(t) \\ y_2(t) = C_2 x_2(t) + D_2 u_2(t) \end{cases}$$

Signal relations resulting in the system in Fig. 5.1(c) are $y = y_1$, $u_2 = y$, and $u_1 = u + y_2$. The output equations above result then in

$$y(t) = C_1 x_1(t) + D_1 u(t) + D_1 y_2(t) \iff (I - D_1 D_2) y(t) = C_1 x_1(t) + D_1 C_2 x_2(t) + D_1 u(t).$$

This equation is obviously solvable for all u if $\det(I - D_1 D_2) \neq 0$. Otherwise, $\operatorname{Im} D_1 \subset \operatorname{Im}(I - D_1 D_2)$ is required. By Proposition 2.1, the latter condition is equivalent to $\ker(I - D_2' D_1') \subset \ker D_1'$, which is wrong

for singular $I - D_1 D_2$ (in fact, $\ker(I - D_2' D_1') \perp \ker D_1'$ then). Hence, the feedback interconnection in Fig. 5.1(c) is well posed iff $I - D_1 D_2$ is invertible, which we assume throughout this section. The output equation of G reads then

$$y(t) = (I - D_1 D_2)^{-1}\big(C_1 x_1(t) + D_1 C_2 x_2(t) + D_1 u(t)\big).$$

The substitution of this expression into the output equation of G_2 with $u_2 = y$ yields

$$y_2(t) = (I - D_2 D_1)^{-1}\big(D_2 C_1 x_1(t) + C_2 x_2(t) + D_2 D_1 u(t)\big).$$

Substituting these expressions to the state equations of G_1 and G_2, we get the realization

$$G : \begin{cases} \begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} A_1 + B_1 \tilde{S} D_2 C_1 & B_1 \tilde{S} C_2 \\ B_2 S C_1 & A_2 + B_2 S D_1 C_2 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} B_1 \tilde{S} \\ B_2 S D_1 \end{bmatrix} u(t) \\ y(t) = S \begin{bmatrix} C_1 & D_1 C_2 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + S D_1 u(t) \end{cases}$$

where

$$S := (I_{m_2} - D_1 D_2)^{-1} \quad \text{and} \quad \tilde{S} := (I_{m_1} - D_2 D_1)^{-1} = I_{m_1} + D_2 S D_1$$

(the last equality follows by the Matrix Inversion Lemma, see Lemma B.7 on p. 195). Hence, we end up with

$$G(s) = \left[\begin{array}{cc|c} A_1 + B_1 D_2 S C_1 & B_1 C_2 + B_1 D_2 S D_1 C_2 & B_1 + B_1 D_2 S D_1 \\ B_2 S C_1 & A_2 + B_2 S D_1 C_2 & B_2 S D_1 \\ \hline S C_1 & S D_1 C_2 & S D_1 \end{array}\right]. \tag{5.4}$$

The realization in (5.4) is substantially simplified if $D_1 D_2 = 0$ and $D_2 D_1 = 0$, for which $S = I$ and $\tilde{S} = I$. In such a case we say that the feedback system in Fig. 5.1(c) *has no algebraic loops*. Clearly, the feedback interconnection is well posed whenever it has no algebraic loops.

Consider now the observability property of the realization in (5.4). To this end, note that the PBH matrix for it is

$$\begin{bmatrix} A_1 + B_1 D_2 S C_1 - \lambda I & B_1 C_2 + B_1 D_2 S D_1 C_2 \\ B_2 S C_1 & A_2 + B_2 S D_1 C_2 - \lambda I \\ S C_1 & S D_1 C_2 \end{bmatrix} = \begin{bmatrix} I & 0 & B_1 D_2 S \\ 0 & I & B_2 S \\ 0 & 0 & S \end{bmatrix} \begin{bmatrix} A_1 - \lambda I & B_1 C_2 \\ 0 & A_2 - \lambda I \\ C_1 & D_1 C_2 \end{bmatrix}.$$

Because the first matrix in the right-hand side above is nonsingular, the realization in (5.4) loses observability iff the "natural" realization of $G_1 G_2$ loses observability. The conditions for the latter are then given by Proposition 5.2 modulo swapping the order of G_1 and G_2.

The controllability analysis follows similar reasoning. Namely,

$$\begin{bmatrix} A_1 + B_1 D_2 S C_1 - \lambda I & B_1 C_2 + B_1 D_2 S D_1 C_2 & B_1 + B_1 D_2 S D_1 \\ B_2 S C_1 & A_2 + B_2 S D_1 C_2 - \lambda I & B_2 S D_1 \end{bmatrix}$$
$$= \begin{bmatrix} A_1 - \lambda I & 0 & B_1 \\ B_2 C_1 & A_2 - \lambda I & B_2 D_1 \end{bmatrix} \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ \tilde{S} D_2 C_1 & \tilde{S} C_2 & \tilde{S} \end{bmatrix},$$

which is derived by $S D_1 = D_1 \tilde{S}$ and $D_2 S = \tilde{S} D_2$. The only point to pay attention is that there is no symmetry between G_1 and G_2 in the feedback interconnection in Fig. 5.1(c) and the PBH controllability matrix of G is related to that of the cascade $G_2 G_1$, rather than $G_1 G_2$ as in the observability case.

The arguments above can be summarized in the following result.

Proposition 5.3. *Suppose that both* (A_1, B_1, C_1, D_1) *and* (A_2, B_2, C_2, D_2) *are minimal and that* $I - D_1 D_2$ *is nonsingular. The realization in* (5.4) *is controllable iff*

$$\mathrm{pdir_i}(G_2, \lambda) \cap \mathrm{zdir_o}(G_1, \lambda) = \{0\}$$

and is observable iff

$$\mathrm{zdir_i}(G_1, \lambda) \cap \mathrm{pdir_o}(G_2, \lambda) = \{0\},$$

both for all $\lambda \in \mathrm{spec}(A_2)$. *In particular, every pole of* $G_2(s)$ *canceled in* $G_2(s)G_1(s)$ *is not controllable in* (5.4) *and every pole of* $G_2(s)$ *canceled in* $G_1(s)G_2(s)$ *is not observable in that realization.*

A qualitative difference between the feedback interconnection and those studied in §5.1.1 and §5.1.2 is that the poles of the realization in (5.4) are typically not those of G_1 and G_2. While the parallel and cascade connections can affect joint dynamics only via cancellations, the feedback connection has more authority over doing that. The ability to alter dynamics is a key property of feedback interconnections, extensively used in control applications. This is especially important in the context of the stability of interconnections, which will be studied in depth in Chapter 6.

However, there might be situations when poles of G_1 are still those of G. This happens if those poles are canceled in the series interconnections of G_1 and G_2, regardless the order. To see this, assume that a pole of G_1 at $s = \lambda$ is canceled in $G_2 G_1$. By Proposition 5.2, $\mathrm{zdir_i}(G_2, \lambda) \cap \mathrm{pdir_o}(G_1, \lambda) \neq \{0\}$ then. In other words, there is $y_\lambda \neq 0$ such that

$$(A_1 - \lambda I)\eta_1 = 0, \quad y_\lambda = C_1 \eta_1 \quad \text{and} \quad \begin{bmatrix} A_2 - \lambda I & B_2 \\ C_2 & D_2 \end{bmatrix} \begin{bmatrix} \eta_2 \\ y_\lambda \end{bmatrix} = 0$$

for some $\eta_1 \neq 0$ and η_2. Consider now the RSM associated with $G_1 G_2 - I$,

$$R_{G_1 G_2 - I}(s) = \begin{bmatrix} A_1 - sI & B_1 C_2 & B_1 D_2 \\ 0 & A_2 - sI & B_2 \\ C_1 & D_1 C_2 & D_1 D_2 - I \end{bmatrix} = \begin{bmatrix} A_1 - sI & B_1 C_2 & B_1 D_2 \\ 0 & A_2 - sI & B_2 \\ C_1 & D_1 C_2 & -S^{-1} \end{bmatrix}.$$

Its normal rank is $n_1 + n_2 + m_2$ because S^{-1} is nonsingular. It is readily verified that

$$R_{G_1 G_2 - I}(\lambda) \begin{bmatrix} \eta_1 \\ \eta_2 \\ y_\lambda \end{bmatrix} = \begin{bmatrix} B_1(C_2 \eta_2 + D_2 y_\lambda) \\ (A_2 - \lambda I)\eta_2 + B_2 y_\lambda \\ y_\lambda + D_1(C_2 \eta_2 + D_2 y_2) - y_\lambda \end{bmatrix} = 0,$$

which means that $R_{G_1 G_2 - I}(s)$ loses its rank at λ (i.e. λ is an invariant zero of $G_1 G_2 - I$). But then the Schur complement of $-S^{-1}$ in $R_{G_1 G_2 - I}(s)$,

$$\begin{bmatrix} A_1 - sI & B_1 C_2 \\ 0 & A_2 - sI \end{bmatrix} + \begin{bmatrix} B_1 D_2 \\ B_2 \end{bmatrix} S \begin{bmatrix} C_1 & D_1 C_2 \end{bmatrix} = \begin{bmatrix} A_1 + B_1 D_2 S C_1 - sI & B_1 C_2 + B_1 D_2 S D_1 C_2 \\ B_2 S C_1 & A_2 + B_2 S D_1 C_2 - sI \end{bmatrix},$$

must be singular at $s = \lambda$. In other words, this λ is a pole of the realization in (5.4). Repeating these arguments for poles of G_1 canceled in $G_1 G_2$ we have the following result.

Proposition 5.4. *Suppose that both* (A_1, B_1, C_1, D_1) *and* (A_2, B_2, C_2, D_2) *are minimal and* $I - D_1 D_2$ *is nonsingular. If* $\lambda \in \mathrm{spec}(A_1)$ *is canceled in* $G_1 G_2$ *or* $G_2 G_1$, *then it is a pole of realization* (5.4). *Moreover, if this canceled* $\lambda \notin \mathrm{spec}(A_2)$, *then it is also a pole of the corresponding transfer function,* $G(s)$.

Proof. It is only left to prove the last statement. If $\lambda \notin \mathrm{spec}(A_2)$, then it is not a pole of $G_2(s)$ and neither its controllability nor its observability in (5.4) can be lost, by Proposition 5.3. Hence, the pole of realization (5.4) is a pole of $G(s)$ as well. □

The condition that λ is not a pole of $G_2(s)$ is sufficient for $G(s)$ to have it as a pole, but not necessary. The examples below demonstrate that it may go both ways in that case.

Example 5.1. Let

$$G_1(s) = \begin{bmatrix} 1 & 1/s \\ 0 & 1 \end{bmatrix} = \left[\begin{array}{c|cc} 0 & 0 & 1 \\ \hline 1 & 1 & 0 \\ 0 & 0 & 1 \end{array}\right] \quad \text{and} \quad G_2(s) = -\begin{bmatrix} 1 & 0 \\ 0 & 1/s \end{bmatrix} = \left[\begin{array}{c|cc} 0 & 0 & 1 \\ \hline 0 & -1 & 0 \\ -1 & 0 & 0 \end{array}\right].$$

We know from Examples 3.2 and 4.2 that $G_1(s)$ has both a pole and a zero at the origin, with

$$\mathrm{pdir_i}(G_1, 0) = \mathrm{zdir_o}(G_1, 0) = \mathrm{span}(e_2) \quad \text{and} \quad \mathrm{pdir_o}(G_1, 0) = \mathrm{zdir_i}(G_1, 0) = \mathrm{span}(e_1).$$

Because $G_2(s)$ is diagonal, it obviously has a pole at the origin and no zeros, with

$$\mathrm{pdir_i}(G_2, 0) = \mathrm{pdir_o}(G_2, 0) = \mathrm{span}(e_2).$$

Hence, by Proposition 5.2 the pole at the origin is canceled in

$$G_2(s)G_1(s) = -\begin{bmatrix} 1 & 1/s \\ 0 & 1/s \end{bmatrix} \quad \left(\text{but not in } G_1(s)G_2(s) = -\begin{bmatrix} 1 & 1/s^2 \\ 0 & 1/s \end{bmatrix}\right).$$

The feedback interconnection of G_1 and G_2 as in Fig. 5.1(c) does not satisfy the condition in the last statement of Proposition 5.4 and, indeed, we have that

$$G(s) = \left[\begin{array}{cc|cc} 0 & -1 & 0 & 1 \\ 0 & -1 & 0 & 1 \\ \hline 1/2 & 0 & 1/2 & 0 \\ 0 & -1 & 0 & 1 \end{array}\right] = \frac{1}{2}\begin{bmatrix} 1 & 1/(s+1) \\ 0 & 2s/(s+1) \end{bmatrix}$$

does not have poles at $s = 0$ (because the realization pole at the origin is not controllable). ◊

Example 5.2. Let now swap G_1 and G_2 from the previous example,

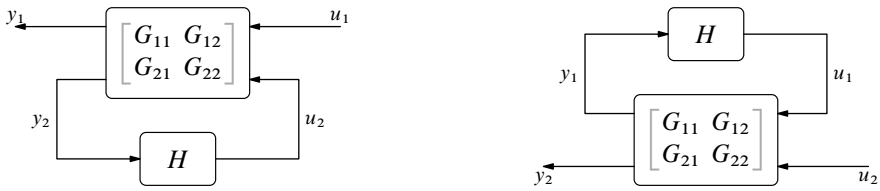
$$G_1(s) = -\begin{bmatrix} 1 & 0 \\ 0 & 1/s \end{bmatrix} = \left[\begin{array}{c|cc} 0 & 0 & 1 \\ \hline 0 & -1 & 0 \\ -1 & 0 & 0 \end{array}\right] \quad \text{and} \quad G_2(s) = \begin{bmatrix} 1 & 1/s \\ 0 & 1 \end{bmatrix} = \left[\begin{array}{c|cc} 0 & 0 & 1 \\ \hline 1 & 1 & 0 \\ 0 & 0 & 1 \end{array}\right],$$

in which case $G_1(s)G_2(s)$ has a cancellation at the origin. And although the condition in the last statement of Proposition 5.4 still does not hold, in this case

$$G(s) = \left[\begin{array}{cc|cc} -1 & 0 & 0 & 1 \\ -1 & 0 & 0 & 0 \\ \hline 0 & -0.5 & -0.5 & 0 \\ -1 & 0 & 0 & 0 \end{array}\right] = \frac{1}{2}\begin{bmatrix} -1 & 1/(s^2+s) \\ 0 & -2/(s+1) \end{bmatrix}$$

does have a pole at the origin. ◊

Propositions 5.3 and 5.4 show that the effects of dynamics of G_1 and G_2 on those of $G : u \mapsto y$ in Fig. 5.1(c) are different. Specifically, assume that there are cancellations in the loop, i.e. in $G_1(s)G_2(s)$ or $G_2(s)G_1(s)$. While canceled modes of the backward path system, G_2, are canceled in G as well, canceled modes of the forward path system, G_1, may remain modes of G, unchanged. This difference is one of the reasons why internal stability, which was mentioned in §1.4.2 and will be studied in §6.1.1, requires the stability of closed-loop systems from all possible inputs to all signals in the loop.



(a) Lower LFT, $\mathcal{F}_{\mathrm{l}}(G, H)$ (b) Upper LFT, $\mathcal{F}_{\mathrm{u}}(G, H)$

Fig. 5.2: Linear fractional transformations (LFTs)

5.2 Linear fractional transformations

A more general interconnection model, which includes the parallel, cascade, and feedback interconnections in Fig. 5.1 as its particular cases, is the *linear fractional transformation* (LFT) presented in Fig. 5.2. Here

$$G = \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix}$$

is a system mapping two (possibly vector) inputs u_1 and u_2 with two (possibly vector) outputs y_1 and y_2. A pair of these inputs and outputs is then connected via yet another system H. The lower LFT, depicted in Fig. 5.2(a) and denoted $\mathcal{F}_{\mathrm{l}}(G, H)$, does it according to the law $u_2 = Hy_2$ and the upper LFT, $\mathcal{F}_{\mathrm{u}}(G, H)$ in Fig. 5.2(b)—according to the law $u_1 = Hy_1$.

The lower LFT is the mapping $\mathcal{F}_{\mathrm{l}}(G, H) : u_1 \mapsto y_1$. To derive it, consider the relation

$$u_2 = Hy_2 = KG_{21}u_1 + HG_{22}u_2 \iff (I - HG_{22})u_2 = HG_{21}u_1.$$

If $I - G_{22}H$ is invertible, then $u_2 = (I - HG_{22})^{-1}HG_{21}u_1$. Thus, $y_1 = \big(G_{11} + G_{12}H(I - G_{22}H)^{-1}G_{21}\big)u_1$ and we end up with

$$\mathcal{F}_{\mathrm{l}}(G, H) = G_{11} + G_{12}(I - HG_{22})^{-1}HG_{21} = G_{11} + G_{12}H(I - G_{22}H)^{-1}G_{21}, \tag{5.5}$$

where the second equality follows from the relation $X(I - YX)^{-1} = (I - XY)^{-1}X$, which is true for all X and Y provided $I - XY$ is invertible. The invertibility of $I - G_{22}H$ (or, equivalently, of $I - G_{22}H$), which guarantees that (5.5) is well defined, is referred to as the *well posedness* condition of $\mathcal{F}_{\mathrm{l}}(G, H)$ and will be discussed later on, in §5.2.1.

The interconnections studied in Section 5.1 can be viewed as special cases of this LFT, e.g.

$$\mathcal{F}_{\mathrm{l}}\left(\begin{bmatrix} G_1 & I \\ I & 0 \end{bmatrix}, G_2\right) = G_1 + G_2, \quad \mathcal{F}_{\mathrm{l}}\left(\begin{bmatrix} 0 & I \\ G_1 & 0 \end{bmatrix}, G_2\right) = G_2G_1,$$

and

$$\mathcal{F}_{\mathrm{l}}\left(\begin{bmatrix} G_1 & G_1 \\ G_1 & G_1 \end{bmatrix}, G_2\right) = G_1 + G_1(I - G_2G_1)^{-1}G_2G_1 = G_1(I - G_2G_1)^{-1}$$

(these choices are not unique). Normally, if $G_{22} = 0$, then the corresponding lower LFT defines an interconnection without feedback. Feedback interconnections correspond to the case of $G_{22} \neq 0$.

The upper LFT is the mapping $\mathcal{F}_{\mathrm{u}}(G, H) : u_2 \mapsto y_2$ for the system in Fig. 5.2(b). Repeating the steps that led to (5.5), we can derive the following expression for it:

$$\mathcal{F}_{\mathrm{u}}(G, H) = G_{22} + G_{21}(I - HG_{11})^{-1}HG_{12} = G_{22} + G_{21}H(I - G_{11}H)^{-1}G_{12}, \tag{5.6}$$

where the invertibility of $I - HG_{11}$ (or, equivalently, of $I - G_{11}H$) is assumed. It is readily seen that the upper LFT turns the lower LFT if the two inputs and the two outputs are swapped. In other words,

$$\mathcal{F}_{\mathrm{u}}(G, H) = \mathcal{F}_{\mathrm{l}}\left(\left[\begin{smallmatrix} 0 & I \\ I & 0 \end{smallmatrix}\right]G\left[\begin{smallmatrix} 0 & I \\ I & 0 \end{smallmatrix}\right], H\right). \tag{5.7}$$

This means that the separation between the lower and upper linear fractional transforms is merely a matter of notational convenience.

Example 5.3. Let $T(s) = D + C(sI - A)^{-1}B$. One can then verify that

$$T(s) = \mathcal{F}_{\mathrm{u}}\left(\begin{bmatrix} A & B \\ C & D \end{bmatrix}, \frac{1}{s}I\right),$$

which is well defined whenever $s \notin \operatorname{spec}(A)$. ◊

Example 5.4. The bilinear (Tustin) transform is given by $s = \gamma(z-1)/(z+1)$. For any $\gamma > 0$ it transforms the unit disk to the left half-plane. Observing that $s = \gamma - 2\gamma/(z+1)$, the Tustin transform can be written as

$$s = \mathcal{F}_{\mathrm{l}}\left(\begin{bmatrix} \gamma & -2 \\ \gamma & -1 \end{bmatrix}, \frac{1}{z}\right),$$

which well defined whenever $z \neq -1$. ◊

The two propositions below present some useful algebraic properties of LFTs.

Proposition 5.5. *Suppose $\mathcal{F}_l(G, H)$ is square and well posed and G_{11} is nonsingular. Then $\mathcal{F}_l(G, H)$ is invertible and*

$$[\mathcal{F}_l(G, H)]^{-1} = \mathcal{F}_l(\hat{G}, H), \quad \textit{where } \hat{G} = \begin{bmatrix} G_{11}^{-1} & -G_{11}^{-1}G_{12} \\ G_{21}G_{11}^{-1} & G_{22} - G_{21}G_{11}^{-1}G_{12} \end{bmatrix}.$$

Similarly, if $\mathcal{F}_u(G, H)$ is square and well posed and G_{22} is nonsingular, then $\mathcal{F}_u(G, H)$ is invertible and

$$[\mathcal{F}_u(G, H)]^{-1} = \mathcal{F}_u(\tilde{G}, H), \quad \textit{where } \tilde{G} = \begin{bmatrix} G_{11} - G_{12}G_{22}^{-1}G_{21} & G_{12}G_{22}^{-1} \\ -G_{22}^{-1}G_{21} & G_{22}^{-1} \end{bmatrix}.$$

Proof. We prove only the lower LFT part, the upper LFT part is proved similarly. The logic is that the sought $[\mathcal{F}_{\mathrm{l}}(G, H)]^{-1}$ is a mapping $y_1 \mapsto u_1$, so what we need is to swap u_1 with y_1. Because H is still supposed to have y_2 as its input and u_2 as its output, these signals remain untouched. The result then follows from the relation

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix}\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \iff \begin{bmatrix} u_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} G_{11}^{-1} & -G_{11}^{-1}G_{12} \\ G_{21}G_{11}^{-1} & G_{22} - G_{21}G_{11}^{-1}G_{12} \end{bmatrix}\begin{bmatrix} y_1 \\ u_2 \end{bmatrix} = \hat{G}\begin{bmatrix} y_1 \\ u_2 \end{bmatrix},$$

which is verified by direct substitution. □

Example 5.5. Consider the LFT in Example 5.3. If D is square and nonsingular, Proposition 5.5 yields

$$T^{-1}(s) = \mathcal{F}_{\mathrm{u}}\left(\begin{bmatrix} A - BD^{-1}C & BD^{-1} \\ -D^{-1}C & D^{-1} \end{bmatrix}, \frac{1}{s}I\right) = D^{-1} - D^{-1}C(sI - A + BD^{-1}C)^{-1}BD^{-1},$$

which agrees with (4.8). ◊

Example 5.6. Consider the LFT from Example 5.4. Using Proposition 5.5, and then (5.7), we have that

$$\frac{1}{s} = \mathcal{F}_{\mathrm{l}}\left(\begin{bmatrix} 1/\gamma & 2/\gamma \\ 1 & 1 \end{bmatrix}, \frac{1}{z}\right) = \mathcal{F}_{\mathrm{u}}\left(\begin{bmatrix} 1 & 1 \\ 2/\gamma & 1/\gamma \end{bmatrix}, \frac{1}{z}\right),$$

which is well defined whenever $z \neq 1$ and indeed yield $1/s = (z+1)/(\gamma(z-1))$. ◊

Proposition 5.6. *If G is invertible and such that G_{12} and G_{21} are square and invertible too, then*

$$T = \mathcal{F}_l(G, H) \iff H = \mathcal{F}_u(G^{-1}, T).$$

Proof. Let $T = \mathcal{F}_\mathrm{l}(G, H)$. Since G is nonsingular, $\left[\begin{smallmatrix} u_1 \\ u_2 \end{smallmatrix}\right] = G^{-1}\left[\begin{smallmatrix} y_1 \\ y_2 \end{smallmatrix}\right]$. Then, taking into account that T is the mapping $u_1 \mapsto y_1$ and H is the mapping $y_2 \mapsto u_2$, we have that H satisfies the following set of equations:

$$\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = G^{-1} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \quad \text{and} \quad y_1 = T\, u_1.$$

This defines exactly $\mathcal{F}_\mathrm{u}(G^{-1}, T)$. Similar arguments yield that $H = \mathcal{F}_\mathrm{u}(G^{-1}, T) \implies T = \mathcal{F}_\mathrm{l}(G, H)$. To complete the proof we only need to show that $\mathcal{F}_\mathrm{l}(G, H)$ is well posed iff $\mathcal{F}_\mathrm{u}(G^{-1}, T)$ is well posed. To show this, assume first that $\mathcal{F}_\mathrm{l}(G, H)$ is well posed, i.e. that $I - G_{22}H$ is nonsingular. The well-posedness of $\mathcal{F}_\mathrm{u}(G^{-1}, T)$ reads then as the non-singularity of $I - \left[\begin{smallmatrix} I & 0 \end{smallmatrix}\right] G^{-1} \left[\begin{smallmatrix} I \\ 0 \end{smallmatrix}\right] T$ or, because $I - M_1 M_2$ is nonsingular iff $I - M_2 M_1$ is nonsingular, of

$$\begin{aligned} I - G^{-1}\begin{bmatrix} I \\ 0 \end{bmatrix} T \begin{bmatrix} I & 0 \end{bmatrix} &= G^{-1}\left(\begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix} - \begin{bmatrix} G_{11} + G_{12}H(I - G_{22}H)^{-1}G_{21} & 0 \\ 0 & 0 \end{bmatrix}\right) \\ &= G^{-1}\begin{bmatrix} -G_{12}H(I - G_{22}H)^{-1}G_{21} & G_{12} \\ G_{21} & G_{22} \end{bmatrix} \\ &= G^{-1}\begin{bmatrix} G_{12} & -G_{12}H(I - G_{22}H)^{-1}G_{21} \\ G_{22} & G_{21} \end{bmatrix}\begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix}. \end{aligned}$$

Now, G_{12} is assumed to be invertible. Its Schur complement in the system in the middle of the right-hand side above,

$$G_{21} + G_{22}H(I - G_{22}H)^{-1}G_{21} = (I - G_{22}H)^{-1}G_{21},$$

is invertible too, because G_{21} is also assumed to be invertible. Hence, $\mathcal{F}_\mathrm{u}(G^{-1}, T)$ is well posed. The other direction follows by similar arguments. □

Example 5.7. Consider the lower LFT in Example 5.6. Because its "G" matrix satisfies the conditions of Proposition 5.6,

$$\frac{1}{z} = \mathcal{F}_\mathrm{u}\left(\begin{bmatrix} -\gamma & 2 \\ \gamma & -1 \end{bmatrix}, \frac{1}{s}\right) = \mathcal{F}_\mathrm{l}\left(\begin{bmatrix} -1 & \gamma \\ 2 & -\gamma \end{bmatrix}, \frac{1}{s}\right).$$

In other words, $1/z = -(s - \gamma)/(s + \gamma)$, as expected. ◊

5.2.1 Well posedness of LFT

The invertibility of $I - HG_{22}$ assumed in the developments above clearly guarantees that $\mathcal{F}_\mathrm{l}(G, H)$ is well defined. Yet this assumption might not be necessary in this context. To see this, consider the static

$$\begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix} = \left[\begin{array}{c:cc} 1 & 0 & 1 \\ \hdashline 1 & \alpha & 0 \\ 0 & 0 & 0 \end{array}\right] \quad \text{and} \quad H = I_2.$$

The matrix $I - HD_{22} = \mathrm{diag}\{1 - \alpha, 1\}$ is singular at $\alpha = 1$. At the same time, $\mathcal{F}_\mathrm{l}(G, H) = 1$ irrespective of α, so it is well defined as a mapping $u_1 \mapsto y_1$ even if $I - HD_{22}$ is singular. Nevertheless, this situation is problematic, because the mappings from u_1 to the "internal" variables u_2 and y_2 might still not be well defined. Indeed, we already saw that u_2 satisfies $(I - HG_{22})u_2 = HG_{21}u_1$, so that for this example

$$\begin{bmatrix} 1 - \alpha & 0 \\ 0 & 1 \end{bmatrix} u_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} u_1.$$

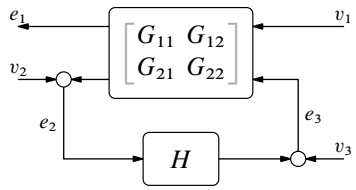


Fig. 5.3: Well-posedness setup for a lower LFT

This relation is not well defined at $\alpha = 1$.

To rule out such situations, the notion of well posedness for the LFT in Fig. 5.2(a) is introduced via adding two auxiliary exogenous inputs v_2 and v_3 as depicted[1] in Fig. 5.3. Supposing that both G and H are well-defined systems, we say that the liner-fractional transformation $\mathcal{F}_l(G, H) : v_1 \mapsto e_1$ is *well posed* if all nine mappings $(e_1, e_2, e_2) \mapsto (v_1, v_2, v_3)$ are well defined. Because these signals are related via

$$\left[\begin{array}{c:cc} I & 0 & -G_{12} \\ \hdashline 0 & I & -G_{22} \\ 0 & -H & I \end{array}\right] \left[\begin{array}{c} e_1 \\ \hdashline e_2 \\ e_3 \end{array}\right] = \left[\begin{array}{c:cc} G_{11} & 0 & 0 \\ \hdashline G_{21} & I & 0 \\ 0 & 0 & I \end{array}\right] \left[\begin{array}{c} v_1 \\ \hdashline v_2 \\ v_3 \end{array}\right],$$

well-posedness is equivalent to an appropriately defined invertibility of the system on the left-hand side of the relation above. This, in turn, boils down to the invertibility of

$$\begin{bmatrix} I & -G_{22} \\ -H & I \end{bmatrix},$$

which is equivalent to the invertibility of either $I - HG_{22}$ or $I - G_{22}H$ (they are the Schur complements of the diagonal blocks). These are exactly the assumptions stated in the beginning of this section.

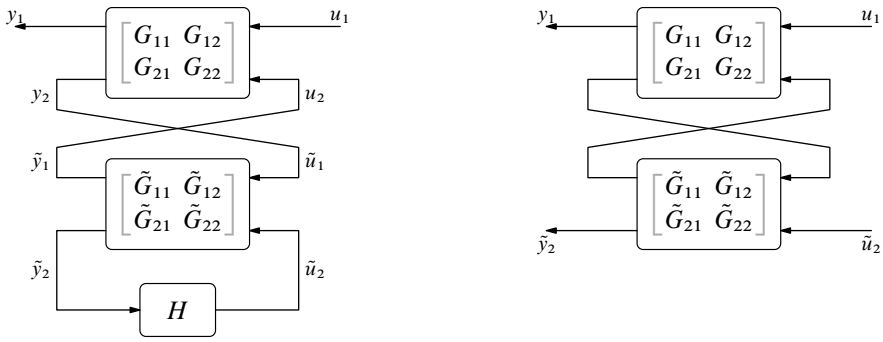
In what follows LFTs are considered over finite-dimensional LTI systems. Such systems are well defined if their (real-rational) transfer functions are proper, so that they admit state-space realizations of form (4.3). A system is then invertible iff its transfer function is bi-proper. Thus, the well-posedness property of LFTs should be understood in this case as the invertibility of the matrix $I - G_{22}(\infty)H(\infty)$ or, equivalently, $I - H(\infty)G_{22}(\infty)$. Similarly to the feedback interconnection studied in §5.1.3, we say that a lower LFT has no algebraic loops if both $G_{22}(\infty)H(\infty) = 0$ and $H(\infty)G_{22}(\infty) = 0$. Linear fractional transformations having no algebraic loops are always well posed.

5.2.2 Redheffer star product

An important property of linear fractional transformations is that they can be nested one into another one. Namely, $\mathcal{F}_l(G, \mathcal{F}_l(\tilde{G}, H))$ is again a lower LFT. This can be seen via the block-diagram in Fig. 5.4. The interconnection of G and $\tilde{G}$, which is obtained by equating $\tilde{u}_1 = y_2$ and $u_2 = \tilde{y}_1$, is known as the *Redheffer star-product*, denoted $G \star \tilde{G}$. The star product is a mapping $(u_1, \tilde{u}_2) \mapsto (y_1, \tilde{y}_2)$. Algebraically, the relations between input and output signals in Fig. 5.4 is

$$\begin{bmatrix} y_1 \\ \tilde{y}_2 \\ y_2 \\ \tilde{y}_1 \end{bmatrix} = \begin{bmatrix} G_{11} & 0 & 0 & G_{12} \\ 0 & \tilde{G}_{22} & \tilde{G}_{21} & 0 \\ G_{21} & 0 & 0 & G_{22} \\ 0 & \tilde{G}_{12} & \tilde{G}_{11} & 0 \end{bmatrix} \begin{bmatrix} u_1 \\ \tilde{u}_2 \\ \tilde{u}_1 \\ u_2 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} y_2 \\ \tilde{y}_1 \end{bmatrix} = \begin{bmatrix} \tilde{u}_1 \\ u_2 \end{bmatrix}.$$

[1] Mind that the other variables are also renamed there, mainly to confuse the adversary.



(a) $\mathcal{F}_{\mathrm{l}}(G, \mathcal{F}_{\mathrm{l}}(\tilde{G}, H))$ (b) $G \star \tilde{G}$

Fig. 5.4: Nested LFTs and the Redheffer star product

Thus, the internal signals $\tilde{u}_1$ and u_2 satisfy

$$\begin{bmatrix} \tilde{u}_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} G_{21} & 0 & 0 & G_{22} \\ 0 & \tilde{G}_{12} & \tilde{G}_{11} & 0 \end{bmatrix} \begin{bmatrix} u_1 \\ \tilde{u}_2 \\ \tilde{u}_1 \\ u_2 \end{bmatrix} \iff \begin{bmatrix} I & -G_{22} \\ -\tilde{G}_{11} & I \end{bmatrix} \begin{bmatrix} \tilde{u}_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} G_{21} & 0 \\ 0 & \tilde{G}_{12} \end{bmatrix} \begin{bmatrix} u_1 \\ \tilde{u}_2 \end{bmatrix},$$

from which, assuming the invertibility of $I - G_{22}\tilde{G}_{11}$ (or, equivalently, of $I - \tilde{G}_{11}G_{22}$),

$$\begin{aligned} G \star \tilde{G} &= \begin{bmatrix} G_{11} & 0 \\ 0 & \tilde{G}_{22} \end{bmatrix} + \begin{bmatrix} 0 & G_{12} \\ \tilde{G}_{21} & 0 \end{bmatrix} \begin{bmatrix} I & -G_{22} \\ -\tilde{G}_{11} & I \end{bmatrix}^{-1} \begin{bmatrix} G_{21} & 0 \\ 0 & \tilde{G}_{12} \end{bmatrix} \\ &= \begin{bmatrix} G_{11} & G_{12}\tilde{G}_{12} \\ 0 & \tilde{G}_{22} \end{bmatrix} + \begin{bmatrix} G_{12}\tilde{G}_{11} \\ \tilde{G}_{21} \end{bmatrix} (I - G_{22}\tilde{G}_{11})^{-1} \begin{bmatrix} G_{21} & G_{22}\tilde{G}_{12} \end{bmatrix}, \end{aligned}$$

where the last equality is obtained by (B.15b). The nested LFT in Fig. 5.4 reads then

$$\mathcal{F}_{\mathrm{l}}(G, \mathcal{F}_{\mathrm{l}}(\tilde{G}, H)) = \mathcal{F}_{\mathrm{l}}(G \star \tilde{G}, H). \tag{5.8a}$$

Likewise,

$$\mathcal{F}_{\mathrm{u}}(G, \mathcal{F}_{\mathrm{u}}(\tilde{G}, H)) = \mathcal{F}_{\mathrm{u}}(\tilde{G} \star G, H), \tag{5.8b}$$

which can be derived by similar arguments.

The transfer function of $G \star \tilde{G}$ can be derived in terms of those of its components,

$$G(s) = \left[\begin{array}{c|cc} A & B_1 & B_2 \\ \hline C_1 & D_{11} & D_{12} \\ C_2 & D_{21} & D_{22} \end{array}\right] \quad \text{and} \quad \tilde{G}(s) = \left[\begin{array}{c|cc} \tilde{A} & \tilde{B}_1 & \tilde{B}_2 \\ \hline \tilde{C}_1 & \tilde{D}_{11} & \tilde{D}_{12} \\ \tilde{C}_2 & \tilde{D}_{21} & \tilde{D}_{22} \end{array}\right],$$

as shown below.

Proposition 5.7. *If* $\det(I - D_{22}\tilde{D}_{11}) \neq 0$*, then*

$$G(s) \star \tilde{G}(s) = \left[\begin{array}{c|c} A_\star & B_\star \\ \hline C_\star & D_\star \end{array}\right],$$

where, denoting $S := (I - D_{22}\tilde{D}_{11})^{-1}$,

$$
\begin{aligned}
A_\star &= \begin{bmatrix} A & B_2 \\ C_2 & D_{22} \end{bmatrix} \star \begin{bmatrix} \tilde{D}_{11} & \tilde{C}_1 \\ \tilde{B}_1 & \tilde{A} \end{bmatrix} = \begin{bmatrix} A & B_2\tilde{C}_1 \\ 0 & \tilde{A} \end{bmatrix} + \begin{bmatrix} B_2\tilde{D}_{11} \\ \tilde{B}_1 \end{bmatrix} S \begin{bmatrix} C_2 & D_{22}\tilde{C}_1 \end{bmatrix}, \\
B_\star &= \begin{bmatrix} B_1 & B_2 \\ D_{21} & D_{22} \end{bmatrix} \star \begin{bmatrix} \tilde{D}_{11} & \tilde{D}_{12} \\ \tilde{B}_1 & \tilde{B}_2 \end{bmatrix} = \begin{bmatrix} B_1 & B_2\tilde{D}_{12} \\ 0 & \tilde{B}_2 \end{bmatrix} + \begin{bmatrix} B_2\tilde{D}_{11} \\ \tilde{B}_1 \end{bmatrix} S \begin{bmatrix} D_{21} & D_{22}\tilde{D}_{12} \end{bmatrix}, \\
C_\star &= \begin{bmatrix} C_1 & D_{12} \\ C_2 & D_{22} \end{bmatrix} \star \begin{bmatrix} \tilde{D}_{11} & \tilde{C}_1 \\ \tilde{D}_{21} & \tilde{C}_2 \end{bmatrix} = \begin{bmatrix} C_1 & D_{12}\tilde{C}_1 \\ 0 & \tilde{C}_2 \end{bmatrix} + \begin{bmatrix} D_{12}\tilde{D}_{11} \\ \tilde{D}_{21} \end{bmatrix} S \begin{bmatrix} C_2 & D_{22}\tilde{C}_1 \end{bmatrix}, \\
D_\star &= \begin{bmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{bmatrix} \star \begin{bmatrix} \tilde{D}_{11} & \tilde{D}_{12} \\ \tilde{D}_{21} & \tilde{D}_{22} \end{bmatrix} = \begin{bmatrix} D_{11} & D_{12}\tilde{D}_{12} \\ 0 & \tilde{D}_{22} \end{bmatrix} + \begin{bmatrix} D_{12}\tilde{D}_{11} \\ \tilde{D}_{21} \end{bmatrix} S \begin{bmatrix} D_{21} & D_{22}\tilde{D}_{12} \end{bmatrix}.
\end{aligned}
$$

Proof. The state equations of G and $\tilde{G}$ are

$$
G : \begin{cases} \dot{x}(t) = Ax(t) + B_1u_1(t) + B_2u_2(t) \\ y_1(t) = C_1x(t) + D_{11}u_1(t) + D_{12}u_2(t) \\ y_2(t) = C_2x(t) + D_{21}u_1(t) + D_{22}u_2(t) \end{cases} \quad \text{and} \quad \tilde{G} : \begin{cases} \dot{\tilde{x}}(t) = \tilde{A}\tilde{x}(t) + \tilde{B}_1\tilde{u}_1(t) + \tilde{B}_2\tilde{u}_2(t) \\ \tilde{y}_1(t) = \tilde{C}_1\tilde{x}(t) + \tilde{D}_{11}\tilde{u}_1(t) + \tilde{D}_{12}\tilde{u}_2(t) \\ \tilde{y}_2(t) = \tilde{C}_2\tilde{x}(t) + \tilde{D}_{21}\tilde{u}_1(t) + \tilde{D}_{22}\tilde{u}_2(t) \end{cases}
$$

The logic of the development below is to equate $y_2 = \tilde{u}_1$ and $u_2 = \tilde{y}_1$ and eliminate these variables. To this end, substitute $\tilde{y}_1$ from the second equation for $\tilde{G}$ to the third equation for G (as u_2) to get

$$
y_2 = C_2x + D_{21}u_1 + D_{22}(\tilde{C}_1\tilde{x} + \tilde{D}_{11}\tilde{u}_1 + \tilde{D}_{12}\tilde{u}_2) = C_2x + D_{21}u_1 + D_{22}(\tilde{C}_1\tilde{x} + \tilde{D}_{11}y_2 + \tilde{D}_{12}\tilde{u}_2),
$$

from which

$$
y_2 = SC_2x + SD_{22}\tilde{C}_1\tilde{x} + SD_{21}u_1 + SD_{22}\tilde{D}_{12}\tilde{u}_2.
$$

Then,

$$
\begin{aligned}
\tilde{y}_1 &= \tilde{C}_1\tilde{x} + \tilde{D}_{11}y_2 + \tilde{D}_{12}\tilde{u}_2 = \tilde{C}_1\tilde{x} + \tilde{D}_{11}(SC_2x + SD_{22}\tilde{C}_1\tilde{x} + SD_{21}u_1 + SD_{22}\tilde{D}_{12}\tilde{u}_2) + \tilde{D}_{12}\tilde{u}_2 \\
&= \tilde{D}_{11}SC_2x + \tilde{S}\tilde{C}_1\tilde{x} + \tilde{D}_{11}SD_{21}u_1 + \tilde{S}\tilde{D}_{12}\tilde{u}_2,
\end{aligned}
$$

where the equality $I + \tilde{D}_{11}SD_{22} = \tilde{S}$ is used (it follows from Lemma B.7 on p. 195). The state-space formula follows then by direct substitution with the use of the fact that $\tilde{D}_{11}S = \tilde{S}\tilde{D}_{11}$. □

The formula of Proposition 5.7 is considerably simplified if both $D_{22}\tilde{D}_{11} = 0$ or $\tilde{D}_{11}D_{22} = 0$. Note also that each matrix of the realization of $G \star \tilde{G}$ is itself a star product. For example,

Example 5.8. Consider the LFTs from Examples 5.3 and 5.6. With the help of (5.8b), we have that

$$
\begin{aligned}
\bar{P}(z) := P\left(\gamma\tfrac{z-1}{z+1}\right) &= \mathcal{F}_{\mathrm{u}}\left(\begin{bmatrix} A & B \\ C & D \end{bmatrix}, \mathcal{F}_{\mathrm{u}}\left(\begin{bmatrix} I & I \\ 2/\gamma\, I & 1/\gamma\, I \end{bmatrix}, \frac{1}{z}\right)\right) \\
&= \mathcal{F}_{\mathrm{u}}\left(\begin{bmatrix} I & I \\ 2/\gamma\, I & 1/\gamma\, I \end{bmatrix} \star \begin{bmatrix} A & B \\ C & D \end{bmatrix}, \frac{1}{z}\right) \\
&= \mathcal{F}_{\mathrm{u}}\left(\begin{bmatrix} (\gamma I + A)(\gamma I - A)^{-1} & \gamma(\gamma I - A)^{-1}B \\ 2C(\gamma I - A)^{-1} & D + C(\gamma I - A)^{-1}B \end{bmatrix}, \frac{1}{z}\right) \\
&= \left[\begin{array}{c|c} (\gamma I + A)(\gamma I - A)^{-1} & \gamma(\gamma I - A)^{-1}B \\ \hline 2C(\gamma I - A)^{-1} & D + C(\gamma I - A)^{-1}B \end{array}\right],
\end{aligned}
$$

which is well defined whenever $\gamma \notin \operatorname{spec}(A)$. The transfer function $\bar{P}(z)$ above is the transfer function of the discrete system $\bar{P}$, obtained from P by Tustin's method. ◊

Chapter 6

Stability of Interconnections

STABILITY is one of fundamental requirements to control systems, which should be met in virtually any application. It is therefore of a great importance to understand, what are stabilizing mechanisms in system interconnections and how they can be exploited to end up with stable controlled behaviors. The goal of this chapter is to shed light on these well-studied questions. We also address the use of stabilization techniques in analyzing steady-state behaviors, which is somewhat less manifest aspects of the stability formalism.

6.1 Closed-loop stability

An important property of feedback is its ability to stabilize controlled dynamics. In this section we define the internal stability notion, which is required to work with feedback interconnections, and study stability criteria that do not hinge on an overly detailed knowledge of involved systems models.

6.1.1 Internal stability

The system in Fig. 6.1(a) on the next page depicts a feedback interconnection of two LTI systems, S_1 and S_2. We say that this interconnection is *internally stable* if all four systems $v_i \mapsto e_j$, $i, j \in \mathbb{Z}_{1..2}$, are stable. The relation between the inputs and outputs in Fig. 6.1(a) can be expressed as

$$\begin{bmatrix} I & -S_2 \\ -S_1 & I \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$$

Thus, assuming the invertibility of the system in the left-hand side above, the system in Fig. 6.1(a) is internally stable iff the system

$$\begin{aligned}
\begin{bmatrix} I & -S_2 \\ -S_1 & I \end{bmatrix}^{-1} &= \left(\begin{bmatrix} I & -S_2 \\ 0 & I \end{bmatrix} - \begin{bmatrix} 0 \\ I \end{bmatrix} S_1 \begin{bmatrix} I & 0 \end{bmatrix} \right)^{-1} \\
&= \begin{bmatrix} I & S_2 \\ 0 & I \end{bmatrix} + \begin{bmatrix} S_2 \\ I \end{bmatrix} S_1 (I - S_2 S_1)^{-1} \begin{bmatrix} I & S_2 \end{bmatrix} \qquad \text{(by Lemma B.7)} \qquad (6.1a) \\
&= \begin{bmatrix} (I - S_2 S_1)^{-1} & S_2 (I - S_1 S_2)^{-1} \\ S_1 (I - S_2 S_1)^{-1} & (I - S_1 S_2)^{-1} \end{bmatrix} \qquad (6.1b)
\end{aligned}$$

is stable. As we study causal L_2 systems, the last requirement is equivalent to the condition that the transfer function of system (6.1) belongs to H_∞.

The reason for considering all possible closed-loop systems, rather than only one of them, lies in the need to rule out stabilization via unstable cancellations in the loop. To understand this better, consider two simple examples.

(a) Symmetric (analysis) form (b) Plant–controller (design) form

Fig. 6.1: Closed-loop internal stability setup

Example 6.1. Let

$$S_1(s) = \frac{1}{s} \quad \text{and} \quad S_2(s) = -\frac{s}{s+1}.$$

It is readily seen that with these choices the $(1,1)$, $(1,2)$, and $(2,2)$ components of (6.1b),

$$\frac{1}{1 - S_2(s)S_1(s)} = \frac{1}{1 - S_1(s)S_2(s)} = \frac{s+1}{s+2} \quad \text{and} \quad \frac{S_2(s)}{1 - S_1(s)S_2(s)} = -\frac{s}{s+2},$$

are H_∞ functions (stable), whereas the $(2,1)$ element,

$$\frac{S_1(s)}{1 - S_1(s)S_2(s)} = \frac{s+1}{s(s+2)},$$

is not. The reason is the cancellation of the unstable pole of $G_1(s)$ at the origin by a zero of $G_2(s)$. The canceled pole still shows up as a pole of the closed-loop transfer function corresponding to the interconnection in Fig. 5.1(c) under $G_1 = S_1$ and $G_2 = S_2$, cf. Proposition 5.4. ◊

The example above is representative for SISO systems. Namely, it is then sufficient to consider only the "off-diagonal" systems $v_1 \mapsto e_2$ and $v_2 \mapsto e_1$. This is also the case if there are unstable cancellations of the same pole in both $S_1(s)S_2(s)$ and $S_2(s)S_1(s)$. But not for general MIMO systems, as shown below.

Example 6.2. Let

$$S_1(s) = \begin{bmatrix} 1 & 1/s \\ 0 & 1 \end{bmatrix} \quad \text{and} \quad S_2(s) = -\begin{bmatrix} s/(s+1) & 0 \\ 0 & 1/s \end{bmatrix}.$$

For these choices

$$\begin{bmatrix} I & -S_2(s) \\ -S_1(s) & I \end{bmatrix}^{-1} = \left[\begin{array}{cc:cc} \frac{s+1}{2s+1} & -\frac{s}{(s+1)(2s+1)} & -\frac{s}{2s+1} & \frac{1}{(s+1)(2s+1)} \\ 0 & \frac{s}{s+1} & 0 & -\frac{1}{s+1} \\ \hdashline \frac{s+1}{2s+1} & \frac{1}{2s+1} & \frac{s+1}{2s+1} & -\frac{1}{s(2s+1)} \\ 0 & \frac{s}{s+1} & 0 & \frac{s}{s+1} \end{array}\right]$$

and all closed-loop transfer functions except that of the $(2,2)$ block are stable. But the system $v_2 \mapsto e_2$ is unstable, hence the interconnection in Fig. 6.1(a) is not internally stable. The reason is again unstable cancellations. Now both $S_1(s)$ and $S_2(s)$ have a pole and a zero at the origin, with

$$\begin{aligned} \mathrm{pdir_i}(S_1, 0) &= \mathrm{zdir_o}(S_1, 0) = \mathrm{span}(e_2), & \mathrm{pdir_o}(S_1, 0) &= \mathrm{zdir_i}(S_1, 0) = \mathrm{span}(e_1), \\ \mathrm{pdir_i}(S_2, 0) &= \mathrm{pdir_o}(S_2, 0) = \mathrm{span}(e_2), & \mathrm{zdir_i}(S_2, 0) &= \mathrm{zdir_o}(S_2, 0) = \mathrm{span}(e_1). \end{aligned}$$

(cf. Example 5.1 on p. 104). Hence, there is an unstable cancellation in $S_2(s)S_1(s)$, but not in $S_1(s)S_2(s)$. The canceled dynamics then still show up in $(I - S_1S_2)^{-1}$. As a matter of fact, this system has also a zero at the origin, with $\mathrm{zdir_i}((I - S_1S_2)^{-1}, 0) = \mathrm{span}(e_1)$ and $\mathrm{zdir_o}((I - S_1S_2)^{-1}, 0) = \mathrm{span}(e_2)$. Because the input zero direction of $(I - S_1(s)S_2(s))^{-1}$ matches the output pole direction of the pole of $S_1(s)$, this pole is canceled in $(I - S_1(s)S_2(s))^{-1}S_1(s)$. Likewise, because the output zero direction of $(I - S_1S_2)^{-1}$ matches the input pole direction of the pole at the origin of S_2, this pole is canceled in $S_2(s)(I - S_1(s)S_2(s))^{-1}$. ◊

In principle, we could consider any one of the four closed-loop systems in Fig. 6.1(a), provided unstable cancellations in both S_1S_2 and S_2S_1 are ruled out. However, keeping track of such cancellations is way more cumbersome than the analysis of the combined system in (6.1).

Remark 6.1. The internal stability of feedback interconnections can be alternatively defined in terms of state-space realizations of involved systems. For example, we may call the system in Fig. 5.1(c) internally stable if all poles of the realization in (5.4) are in $\mathbb{C} \setminus \bar{\mathbb{C}}_0$. Because that realization contains all modes of the system, even the canceled ones, no auxiliary signals are required. Still, the definition of the internal stability notion via the abstract, representation-free, setup in Fig. 6.1(a) appears cleaner. Its potential technical advantage lies in that it is readily applicable to infinite-dimensional systems. ∇

Stability of feedback interconnections is a delicate matter. For example, the feedback interconnection of stable system is not necessarily stable (like with $S_1 = (-s+1)/(s+1)$ and $S_2 = 2$). Stability analyses require then a fairly detailed knowledge of models of S_1 and S_2. Nonetheless, there are situations, in which closed-loop stability can be guaranteed under rather mild and general assumptions on interconnected dynamics. Two of those situations are studied below. The presented results are very general and do not even need to assume that the involved dynamics are finite dimensional.

6.1.2 Small gain theorem

Arguably, the best known, and the most important, result of this kind is the small gain theorem. Roughly, it states that if the loop is stable and contractive, i.e. its gain is smaller than 1, at all frequencies, then the closed loop system is stable as well. This can be seen as the condition that the frequency-response plot of a stable loop is located within the open unit disk $\mathbb{D}$ on the Nyquist plane, in the shaded area in Fig. 6.2(a) on the next page. Obviously, none of such plots can encircle the critical point. The Nyquist stability criterion yields then that the closed-loop system is stable. An algebraic MIMO version of this result is stated below.

Theorem 6.1. *Consider the system in Fig. 6.1(a). If $S_i \in H_\infty$ with $\|S_i\|_\infty = \gamma_i \geq 0$ for $i = 1, 2$, then the closed-loop system is internally stable whenever $\gamma_1\gamma_2 < 1$.*

Proof. Given $M_1 \in \mathbb{C}^{p\times m}$ and $M_2 \in \mathbb{C}^{m\times p}$, by Proposition 2.4 $\|(I - M_2M_1)^{-1}\| = 1/\underline{\sigma}(I - M_2M_1)$ and

$$\underline{\sigma}(I - M_2M_1) = \min_{\|u\|=1} \|(I - M_2M_1)u\| \geq \min_{\|u\|=1} (\|u\| - \|M_2M_1u\|) = 1 - \|M_2M_1\| \geq 1 - \|M_1\|\|M_2\|,$$

whose first inequality follows by the triangle inequality and the second one follows by the sub-multiplicative property of the spectral matrix norm, cf. (2.7). Hence,

$$\sup_{s\in\mathbb{C}_0} \|(I - S_2(s)S_1(s))^{-1}\| = \frac{1}{\inf_{s\in\mathbb{C}_0} \underline{\sigma}(I - S_2(s)S_1(s))} \leq \frac{1}{1-\gamma_1\gamma_2},$$

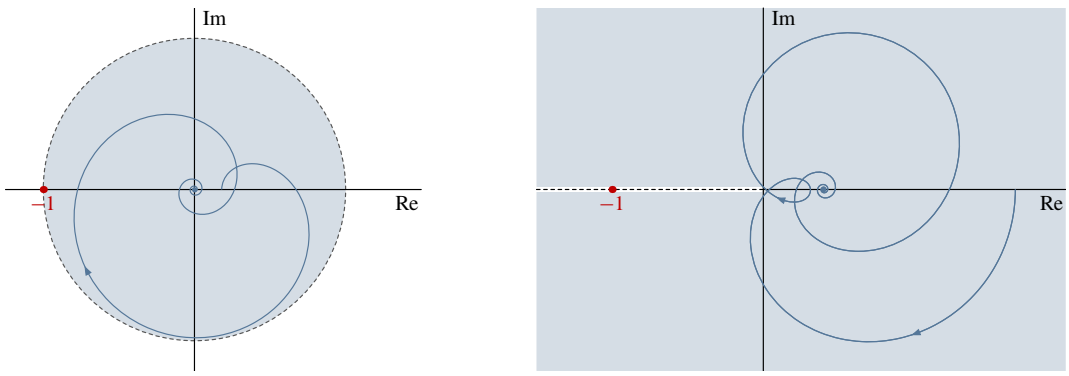
which holds because $\gamma_1\gamma_2 < 1$. Thus, $(I - S_2(s)S_1(s))^{-1}$ is bounded in $\mathbb{C}_0$ and, because both $S_1(s)$ and $S_2(s)$ are holomorphic in that region, is also holomorphic in $\mathbb{C}_0$. Therefore, $(I - S_2S_1)^{-1} \in H_\infty$. This, together with the facts that $S_1 \in H_\infty$ and $S_2 \in H_\infty$, yields the stability of (6.1). □

Remark 6.2 (beyond LTI). The result of Theorem 6.1 remains valid under time-varying and nonlinear S_1 and S_2. All we need is to replace the H_∞ condition with an appropriate definition of stability and the H_∞-norms with whatever induced norms of the involved systems. The proof then uses time-domain arguments. Roughly, it is based on the following two inequalities (true for every $q \geq 1$),

$$\|e_1\|_q = \|v_1 + S_2e_2\|_q \leq \|v_1\|_q + \|S_2e_2\|_q \leq \|v_1\|_q + \gamma_1\|e_2\|_q$$

and

$$\|e_2\|_q = \|v_2 + S_1e_1\|_q \leq \|v_2\|_q + \|S_1e_1\|_q \leq \|v_2\|_q + \gamma_2\|e_1\|_q,$$



(a) Small gain (b) Passivity

Fig. 6.2: Nyquist plane insight into small-gain and passivity analysis philosophies

which, in turn, are direct consequences of the triangle inequality and the definition of the induced norm. The only delicacy is that we do not know a priori that $e_1, e_2 \in L_q$. As such, a rigorous proof starts with considering systems over a finite horizon, say $[0, T]$, showing that there are bounds on norms of e_1 and e_2 independent of T, and then using limit arguments to prove the infinite-horizon case. ∇

6.1.3 Passivity theorem

While the small-gain philosophy is about the loop gain, passivity results are effectively about the loop phase. A phase-centric way to ensure the stability of the closed-loop system in the case when the loop gain has no open right-half place poles is to require that the loop phase is in $(-\pi, \pi)$ [rad]. In other words, the Nyquist plot should never cross the negative real semi-axis. This also ensures that the critical point is not encircled, even if the gain is arbitrarily high. The permitted region on the Nyquist plane (mind that the negative feedback is assumed) is the whole complex plane sans its negative real semi-axis, i.e. $\mathbb{C} \setminus (-\infty, 0)$, see the shaded area in Fig. 6.2(b).

A key concept to formulate such ideas algebraically for the system in Fig. 6.1(a) is the notion of positive realness. This notion has its roots in the circuit theory and is connected with the passivity property of systems, see [2, Sec. 2.7] or [6, Ch. VI]. Given an $m \times m$ system G, we say that its transfer function $G(s)$ is *positive real* (PR) if $G(s) \in \mathbb{R}^{m\times m}$ for all $s \in (0, \infty)$, $G(s)$ is holomorphic in $\mathbb{C}_0$ and $G(s) + [G(s)]' \geq 0$ for all $s \in \mathbb{C}_0$. The latter condition is the MIMO counterpart of $\operatorname{Re} G(s) \geq 0$. PR transfer functions need not be bounded in $\mathbb{C}_0$, so they might not belong to H_∞. For example, $1/s$ is PR because it is holomorphic in $\mathbb{C}_0$ and

$$\frac{1}{s} + \frac{1}{\bar{s}} = 2\frac{\operatorname{Re} s}{|s|^2} > 0, \quad \forall s \in \mathbb{C}_0.$$

A more exotic example is $\tanh(s)$, which can be associated with certain wave equations [5, §3.2]. It has infinitely many pure imaginary poles, at $s = \mathrm{j}(i + 1/2)\pi$ for all $i \in \mathbb{Z}$, but is holomorphic in $\mathbb{C}_0$ and

$$\tanh(s) + \tanh(\bar{s}) = 2\frac{1 - \mathrm{e}^{-4\operatorname{Re} s}}{|1 + \mathrm{e}^{-2s}|^2} > 0, \quad \forall s \in \mathbb{C}_0$$

and is thus PR. Positive-real transfer functions are also not necessarily proper. For example, s is PR. But $1/s^2$ and s^2 are not, because

$$\frac{1}{s^2} + \frac{1}{\bar{s}^2} = 2\frac{(\operatorname{Re} s)^2 - (\operatorname{Im} s)^2}{|s|^4} \quad \text{and} \quad s^2 + \bar{s}^2 = 2\big((\operatorname{Re} s)^2 - (\operatorname{Im} s)^2\big),$$

are both negative for all $s \in \mathbb{C}_0$ such that $|\mathrm{Im}\, s| > \mathrm{Re}\, s$. It can be shown that PR transfer functions may only have simple pure imaginary unstable poles[1] $\mathrm{j}\omega_i$ whose residues $G_i := \lim_{s\to \mathrm{j}\omega_i}(s - \mathrm{j}\omega_i)G(s)$ are finite and satisfy $G_i = G_i' \geq 0$. Also, PR transfer functions cannot have zeros in the open right-half plane. If $\mathrm{rank}(G(s_0) + [G(s_0)]') < \mathrm{normalrank}(G(s) + [G(s)]')$ for some $s_0 \in \mathbb{C}_0$, then $G(s)$ is necessarily not PR. A transfer function $G(s)$ is said to be *strongly positive real* (SPR) if it is PR and there is $\epsilon > 0$ such that $G(s) + [G(s)]' \geq \epsilon I$ for all $s \in \mathbb{C}_0$. None of the transfer functions discussed above is SPR. But $s + 1$ and $(s + 2)/(s + 1)$ are SPR, both with $\epsilon = 2$. SPR transfer functions may not have pure imaginary singularities, but are still not necessarily in H_∞ (like $s + 1$).

Systems with PR and SPR transfer functions have intuitive interpretations in terms of their frequency responses. If $G(s)$ is PR, then

$$G(\mathrm{j}\omega) + [G(\mathrm{j}\omega)]' \geq 0, \quad \forall \omega \in \mathbb{R} \setminus \{\text{ pure imaginary singularities of } G(s)\}$$

and if it is SPR, then $G(\mathrm{j}\omega) + [G(\mathrm{j}\omega)]' \geq \epsilon I$ for some $\epsilon > 0$ and all $\omega \in \mathbb{R}$. In the scalar case, that means that their Nyquist plots are in the closed and open right-half plane, respectively, and the phase of their frequency responses is in the ranges $[-\pi/2, \pi/2]$ and $(-\pi/2, \pi/2)$. This suggests that the negative feedback interconnection of LTI systems with PR and SPR transfer functions has its loop frequency response in the shaded area in Fig. 6.2(b), never crossing the negative real semi-axis, and it thus stable. This is indeed true, under some technical assumptions, and applies to MIMO systems. A key result is presented below. Although its formulation is not symmetric, in the sense that it imposes different conditions on S_1 and S_2, it should be clear that S_1 and S_2 may be interchanged without affecting the result.

Theorem 6.2. *Consider the system in Fig. 6.1(a). If $S_1(s)$ is positive real, $-S_2(s)$ is strongly positive real, and $S_2 \in H_\infty$, then the closed-loop system is internally stable.*

The proof of Theorem 6.2 requires some technical results. The first one of them may be thought of a matrix counterpart of the known fact that the bilinear (Tustin) transform maps a half plane to a unit disk.

Lemma 6.3. *Let $M \in \mathbb{C}^{m\times m}$ and $\delta \in [0, 1]$. The following statements are equivalent:*

1. $M + M' \geq \delta(I + M'M)$,
2. $\det(I + M) \neq 0$ *and* $\|(I - M)(I + M)^{-1}\| \leq \sqrt{(1 - \delta)/(1 + \delta)}$.

Moreover, if either of these conditions holds, then $\|(I + M)^{-1}\| \leq 1/\sqrt{1 + \delta}$.

Proof. We start with showing that the first condition implies that $\det(I + M) \neq 0$ and the last condition. To this end, note that the first condition can be rewritten as $(I + M')(I + M) \geq (1+\delta)(I + M'M) \geq (1+\delta)I$, whence these two conditions follow.

Now, the first statement is equivalent to the condition

$$0 \leq M + M' - \delta(I + M'M) = \frac{1+\delta}{2}\Big(\frac{1-\delta}{1+\delta}(I + M')(I + M) - (I - M')(I - M)\Big).$$

Thus, the first statement holds iff

$$(I + M')^{-1}(I - M')(I - M)(I + M)^{-1} \leq \frac{1-\delta}{1+\delta} I,$$

which is equivalent to the second statement by Theorem A.2. □

The following result is an extension of Lemma 6.3 to transfer functions. It shows that positive-realness is connected with the contraction property via linear fractional transformations and is of independent interest. In fact, the "only if" part of its second item also holds, under mild technical conditions, see [9, Cors. 3.6 and 4.3]. Nevertheless, the formulation below is sufficient for the purposes of this section.

[1] Irrational positive-real transfer functions may also have certain pure imaginary *essential singularities*, see [9, Thm. 3.7].

Proposition 6.4. *If $G(s)$ is PR, then*

- $(I+G)^{-1} \in H_\infty$ *with* $\|(I+G)^{-1}\|_\infty \le 1$
- $(I-G)(I+G)^{-1} \in H_\infty$ *with* $\|(I-G)(I+G)^{-1}\|_\infty \le 1$.

If $G(s)$ is SPR and $G \in H_\infty$, then the non-strict inequalities above can be replaced with the strict ones.

Proof. If $G(s)$ is PR, then $G(s)$ satisfies the first condition of Lemma 6.3 with $\delta = 0$ for all $s \in \mathbb{C}_0$. Together with the assumed holomorphic property of $G(s)$ in $\mathbb{C}_0$ that yields both items of the proposition. If $G(s)$ is SPR, then $G(s) + [G(s)]' \ge \epsilon I$ for some $\epsilon \in (0, 2\|G\|_\infty]$. In this case the first condition of Lemma 6.3 holds for $G(s)$ for all $s \in \mathbb{C}_0$ under $\delta \le \epsilon/(1+\|G\|_\infty^2) \le 1$. If $G \in H_\infty$, then this $\delta > 0$, so that $\sqrt{(1-\delta)/(1+\delta)} < 1$ and $1/\sqrt{1+\delta} < 1$. Hence, we have the strict contractiveness of both $(I+G)^{-1}$ and $(I-G)(I+G)^{-1}$. □

We are now in the position to prove the passivity theorem.

Proof of Theorem 6.2. It is a matter of straightforward algebra to verify that

$$\begin{bmatrix} I & -S_2 \\ -S_1 & I \end{bmatrix}^{-1} = \begin{bmatrix} I & I \\ I & -I \end{bmatrix} \overbrace{\begin{bmatrix} I & (I+S_2)(I-S_2)^{-1} \\ -(I-S_1)(I+S_1)^{-1} & I \end{bmatrix}^{-1}}^{\tilde{T}_{\text{aux}}} \begin{bmatrix} (I-S_2)^{-1} & 0 \\ 0 & -(I+S_1)^{-1} \end{bmatrix}.$$

The first and the last factors in the right-hand side above are stable, the last one by Proposition 6.4. Hence, the stability of $\tilde{T}_{\text{aux}}$ would imply that of the system in Fig. 6.1(a) (but not vice versa, because $I + S_1$ is not necessarily stable). The system $\tilde{T}_{\text{aux}}$ is the closed-loop system associated with the internal stability of Fig. 6.1(a) under the substitutes $S_1 \to \tilde{S}_1 := (I-S_1)(I+S_1)^{-1}$ and $S_2 \to \tilde{S}_2 := -(I+S_2)(I-S_2)^{-1}$. By Proposition 6.4 we have that both these systems are stable, with $\|\tilde{S}_1\|_\infty \le 1$ and $\|\tilde{S}_2\|_\infty < 1$. Hence, the closed-loop system is stable by the small gain theorem (Theorem 6.1). □

The proof of Theorem 6.2 might suggest that the passivity theorem is merely a special case of the small gain theorem. Technically, this may be true. However, having a stability result directly in terms of the positive realness property is important in many applications, because of connections between the positive realness and passivity properties. By Parseval's theorem, a causal system G whose transfer function is PR satisfies $\langle Gu, u\rangle_2 \ge 0$ for all $u \in L_2[0,T]$ and all $T > 0$. This property, known as *passivity*, has energy conservation interpretations in several applications, in particular, in electrical and mechanical networks. As such, the passivity theorem is used to design simple controllers to stabilize systems, which exhibit energy conservation properties.

For example, a PR plant is stabilized by any proportional controller under negative feedback, provided the controller gain is positive definite. This implies that PR plants admit arbitrarily high-gain controllers in theory, cf. the discussion in §1.4.2. This may suggest that having feedback loops with passive elements might not be a realistic situation. Indeed, the passivity property itself is extremely *fragile* in control systems. Infinitesimal loop delays or arbitrarily fast sampling in the loop would destroy it. It thus appears to be taken for granted too frequently, especially at high frequencies. Perhaps some frequency-dependent alternation of passivity and small-gain arguments is a more sensible rationale behind the stability of many practical feedback systems.

Remark 6.3 (beyond LTI). The result of Theorem 6.2 also remains valid under time-varying and nonlinear S_1 and S_2. A fairly general formulation can be found in [6, Sec. VI.5]. To state a version of it, denote by $\|\cdot\|_T$ and $\langle\cdot,\cdot\rangle_T$ the norm and inner product on $L_2[0,T]$. Assume that there are constants ϵ_1, ϵ_2, $\gamma_2 > 0$, and β_i for $i = \{1,2,3\}$ such that

1. $\langle S_1 e_1, e_1\rangle_T \geq \epsilon_1 \|S_1 e_1\|_T^2 + \beta_1$, for all $T \in \mathbb{R}_+$,
2. $\langle e_2, -S_2 e_2\rangle_T \geq \epsilon_2 \|e_2\|_T^2 + \beta_2$ and $\|S_2 e_2\|_T \leq \gamma_2 \|e_2\|_T^2 + \beta_3$, for all $T \in \mathbb{R}_+$.

The closed-loop system in Fig. 6.1(a) is then internally stable if $\epsilon_1 + \epsilon_2 > 0$. The constants β_i, which play no role in the stability condition itself, are required to handle nonlinear systems, in the linear case all β_i may be taken zero. The PR property for S_1 is replaced above with the stronger *strict output passivity*. It becomes the familiar passivity under $\epsilon_1 = 0$, but is more restrictive under $\epsilon_1 > 0$. The first requirement on S_2 is known as the *strict input passivity* and can be viewed as the time-domain counterpart of the SPR property under $\epsilon_2 > 0$. The second requirement on S_2 is just its L_2-stability. With $\epsilon_1 = 0$ the result ($\epsilon_2 > 0$ implies stability) is similar to that in Theorem 6.2. In general, the lack of passivity in one of systems may be compensated by its excess in another one. ▽

6.2 Closed-loop stabilization

Having defined the stability notion, we are now in the position to study the *stabilization* problem associated with the feedback interconnections in Fig. 6.1(a). The basic assumption now is that one of the interconnecting systems is given and another one can be chosen (designed) to render the interconnection internally stable. In this context, it is convenient to switch to the block-diagram in Fig. 6.1(b), in which the given system (the plant) is denoted P and the system to be chosen (the controller / regulator) is denoted R. Henceforth, we mostly assume that the plant is finite dimensional and seek for finite-dimensional controllers having *proper* transfer functions. The internal stability of the system in Fig. 6.1(b) is then equivalent (see §3.3.2) to the condition $T_{\text{aux}} \in RH_\infty$, where

$$T_{\text{aux}} := \begin{bmatrix} I & 0 \\ P & I \end{bmatrix} + \begin{bmatrix} I \\ P \end{bmatrix} R(I - PR)^{-1} \begin{bmatrix} P & I \end{bmatrix} = \begin{bmatrix} (I - RP)^{-1} & R(I - PR)^{-1} \\ P(I - RP)^{-1} & (I - PR)^{-1} \end{bmatrix} \tag{6.2}$$

is the system $\left[\begin{smallmatrix} v_1 \\ v_2 \end{smallmatrix}\right] \mapsto \left[\begin{smallmatrix} e_1 \\ e_2 \end{smallmatrix}\right]$, cf. (6.1). Note that the stability of T_{aux} requires the properness of $T_{\text{aux}}(s)$, see the discussion posterior to Eqn. (3.27) on p. 50. This, in turn, requires the invertibility of $I - R(\infty)P(\infty)$, i.e. the well posedness of the system in Fig. 6.1(b).

6.2.1 All stabilizing controllers: stable plants

The main idea behind characterizing all stabilizing controllers below is to reduce the stability analysis of the four subsystems in T_{aux} to that of only one it subsystem, which is a bijective function of the controller. This reduction is particularly simple and intuitive in the case when the plant itself is stable. We thus start with assuming that $P \in RH_\infty$, postponing the general case to §6.2.2.

If the plant is stable, then

$$\begin{bmatrix} I & 0 \\ -P & I \end{bmatrix} = \begin{bmatrix} I & 0 \\ P & I \end{bmatrix}^{-1}$$

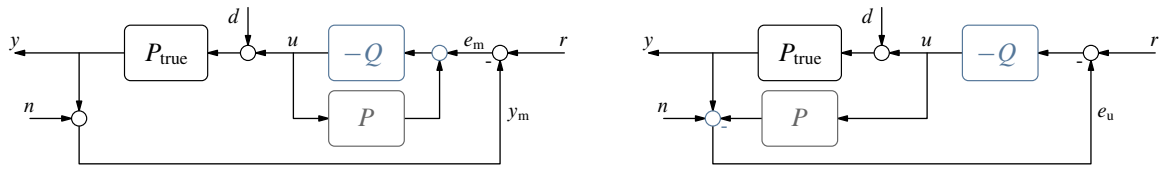
is bi-stable. Hence, the stability of T_{aux} is equivalent to that of

$$\begin{bmatrix} I & 0 \\ -P & I \end{bmatrix} T_{\text{aux}} \begin{bmatrix} I & 0 \\ -P & I \end{bmatrix} = \begin{bmatrix} I & 0 \\ -P & I \end{bmatrix} + \begin{bmatrix} I \\ 0 \end{bmatrix} R(I - PR)^{-1} \begin{bmatrix} 0 & I \end{bmatrix} = \begin{bmatrix} I & R(I - PR)^{-1} \\ -P & I \end{bmatrix}.$$

Three sub-blocks of this system are stable regardless of R and only the $(1, 2)$ sub-block depends on the controller. Thus,

$$T_{\text{aux}} \in RH_\infty \iff T_{\text{c}} := R(I - PR)^{-1} \in RH_\infty,$$

which is intuitive. Indeed, if P is stable, then unstable loop cancellations are only possible between unstable poles of R and the plant. Hence, it is sufficient to consider only the system mapping $v_1 \mapsto e_2$, which can be thought of as the control sensitivity system defined in §1.4.2.



(a) Original form (b) Equivalent IMC form

Fig. 6.3: Q-parametrization in the unity-feedback system (with negative feedback)

An advantage of analyzing T_{c} instead of T_{aux} is that the former is a bijective function of R. In other words, not only R uniquely determines T_{c} (provided the feedback is well posed), but also every T_{c} can be produced by a unique R. The mapping from T_{c} to R can be obtained via various approaches. We use LFT-based arguments below. It is readily seen that

$$T_{\text{c}} = \mathcal{F}_{\text{l}}\left(\begin{bmatrix} 0 & I \\ I & P \end{bmatrix}, R\right).$$

This lower LFT falls into the scope of Proposition 5.6, so we have that

$$R = \mathcal{F}_{\text{u}}\left(\begin{bmatrix} 0 & I \\ I & P \end{bmatrix}^{-1}, T_{\text{c}}\right) = \mathcal{F}_{\text{u}}\left(\begin{bmatrix} -P & I \\ I & 0 \end{bmatrix}, T_{\text{c}}\right) = \mathcal{F}_{\text{l}}\left(\begin{bmatrix} 0 & I \\ I & -P \end{bmatrix}, T_{\text{c}}\right) = T_{\text{c}}(I + PT_{\text{c}})^{-1}.$$

These relations imply that every controller that can be presented in the form $R = Q(I + PQ)^{-1}$ for some stable Q is stabilizing (it results in $T_{\text{c}} = Q$) and every stabilizing R can be presented in that form for a stable Q (actually, $Q = T_{\text{c}}$). These arguments lead to the result below.

Theorem 6.5 (Q-parametrization)**.** *If $P \in RH_\infty$, then R stabilizes the system in Fig. 6.1 iff there exists $Q \in RH_\infty$ such that*

$$R = \mathcal{F}_l\left(\begin{bmatrix} 0 & I \\ I & -P \end{bmatrix}, Q\right) = Q(I + PQ)^{-1}. \tag{6.3}$$

Moreover, R is well posed iff $I + P(\infty)Q(\infty)$ is nonsingular.

A remarkable outcome of the characterization of stabilizing controllers above is that is substantially simplifies resulting closed-loop systems. Indeed, substituting $R(I - PR)^{-1} = Q$ into the first part of (6.2) yields

$$T_{\text{aux}} = \begin{bmatrix} I & 0 \\ P & I \end{bmatrix} + \begin{bmatrix} I \\ P \end{bmatrix} Q \begin{bmatrix} P & I \end{bmatrix} = \begin{bmatrix} I + QP & Q \\ P + PQP & I + PQ \end{bmatrix}.$$

All subsystems above are linear or affine functions of Q, which is an advantage over the fractional dependence on R in (6.2). In a sense, the parametrization of Theorem 6.5 converts a closed-loop problem into an open-loop one, cf. (1.7). This has far-reaching implications on understanding attainable closed-loop maps and controller design methods. Some of these issues will be discussed in Chapter 7.

The structure of the parametrization in (6.3) has an intuitive interpretation in terms of the unity feedback control architecture like that in Fig. 1.4(c) on p. 4. To see that, consider the closed-loop system in Fig. 6.3(a), which is the standard unity-feedback system, whose controller is of the form (6.3). By P_{true} we understand the real plant, which might be different from its model P used by the controller. Mind also that we consider the negative feedback case now. Hence, the sign of the controller in (6.3) should be inverted, which yields the positive feedback in the internal controller loop and the sign inverse of the Q-parameter in the diagrams in Fig. 6.3. The system can then be equivalently presented in the form depicted in Fig. 6.3(b) via elementary block-diagram manipulation rules. If there is no uncertainty, i.e. if $P = P_{\text{true}}$, $d = 0$, and $n = 0$, the signal

fed back vanishes, $e_{\mathrm{u}} = 0$, and we effectively have an open-loop control system, like that in Fig. 1.4(b). The signal e_{u} shows up only if there is uncertainty, when

$$e_{\mathrm{u}} = (P_{\mathrm{true}} - P)u + P_{\mathrm{true}}d + n.$$

This agrees with the understanding that for open-loop stable processes feedback is only needed because of uncertainty. The feedback signal $e_{\mathrm{u}} = y_{\mathrm{m}} - Pu$ can thus be thought of as the uncertainty indicator. As a matter of fact, the architecture in Fig. 6.3(b) is known as the *internal model control* (IMC) setup, see [19] for its comprehensive exposition.

6.2.2 All stabilizing controllers: possibly unstable plants

The transformation used in the previous subsection to decouple three out of four components of the closed-loop system from R cannot be used if $P \notin H_\infty$ because it is unstable then. Yet the idea behind it still applies if properly modified.

To see how, bring in a doubly coprime factorization over RH_∞, like that in (3.34). That is, consider transfer functions $N \in RH_\infty^{p\times m}$, $M \in RH_\infty^{m\times m}$, $\tilde{N} \in RH_\infty^{p\times m}$, and $\tilde{M} \in RH_\infty^{p\times p}$ such that $M(s)$ and $\tilde{M}(s)$ are bi-proper,

$$P = NM^{-1} = \tilde{M}^{-1}\tilde{N},$$

and there are appropriately dimensioned $X, Y, \tilde{X}, \tilde{Y} \in RH_\infty$ such that

$$\begin{bmatrix} X & Y \\ -\tilde{N} & \tilde{M} \end{bmatrix}\begin{bmatrix} M & -\tilde{Y} \\ N & \tilde{X} \end{bmatrix} = \begin{bmatrix} I_m & 0 \\ 0 & I_p \end{bmatrix} \tag{6.4}$$

(this is just a repeated (3.34)). From the construction in §4.3.1 we know that these functions always exist. We can also suppose that $X(s)$ and $\tilde{X}(s)$ are bi-proper as well, cf. (4.21).

Now, the second term on the middle expression of (6.2) can be rewritten as

$$\begin{bmatrix} I \\ P \end{bmatrix} R(I - PR)^{-1}\begin{bmatrix} P & I \end{bmatrix} = \begin{bmatrix} M \\ N \end{bmatrix} M^{-1}R(I - PR)^{-1}\tilde{M}^{-1}\begin{bmatrix} \tilde{N} & \tilde{M} \end{bmatrix}.$$

It follows from (6.4) that

$$\begin{bmatrix} X & Y \\ -\tilde{N} & \tilde{M} \end{bmatrix}\begin{bmatrix} M \\ N \end{bmatrix} = \begin{bmatrix} I \\ 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \tilde{N} & \tilde{M} \end{bmatrix}\begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix}\begin{bmatrix} M & -\tilde{Y} \\ N & \tilde{X} \end{bmatrix}\begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix} = \begin{bmatrix} 0 & I \end{bmatrix}. \tag{6.5}$$

This suggests that the bi-stable factors on the left-hand side of (6.4) could be used instead of $\left[\begin{smallmatrix} I & 0 \\ -P & I \end{smallmatrix}\right]$. This is indeed the case as shown in the proof of the theorem below.

Theorem 6.6 (Q-parametrization)**.** *R stabilizes the system in Fig. 6.1 iff there exists $Q \in RH_\infty$ such that*

$$R = \mathcal{F}_l\left(\begin{bmatrix} -X^{-1}Y & X^{-1} \\ \tilde{M} + \tilde{N}X^{-1}Y & -\tilde{N}X^{-1} \end{bmatrix}, Q\right) = (X + Q\tilde{N})^{-1}(-Y + Q\tilde{M}) \tag{6.6a}$$

$$= \mathcal{F}_l\left(\begin{bmatrix} -\tilde{Y}\tilde{X}^{-1} & M + \tilde{Y}\tilde{X}^{-1}N \\ \tilde{X}^{-1} & -\tilde{X}^{-1}N \end{bmatrix}, Q\right) = (-\tilde{Y} + MQ)(\tilde{X} + NQ)^{-1}. \tag{6.6b}$$

Moreover, R is well posed iff $\tilde{X}(\infty) + N(\infty)Q(\infty)$ (equivalently, $X(\infty) + Q(\infty)\tilde{N}(\infty)$) is nonsingular.

Proof. Because the factors on the left-hand side of (6.4) are bi-stable, by construction,

$$T_{\mathrm{aux}} \in RH_\infty \iff \tilde{T}_{\mathrm{aux}} := \begin{bmatrix} X & Y \\ -\tilde{N} & \tilde{M} \end{bmatrix} T_{\mathrm{aux}} \begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix}\begin{bmatrix} M & -\tilde{Y} \\ N & \tilde{X} \end{bmatrix}\begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix} \in RH_\infty$$

It is readily verified using (6.4) that

$$\begin{bmatrix} X & Y \\ -\tilde{N} & \tilde{M} \end{bmatrix}\begin{bmatrix} I & 0 \\ P & I \end{bmatrix}\begin{bmatrix} M & \tilde{Y} \\ -N & \tilde{X} \end{bmatrix} = \begin{bmatrix} M^{-1} & Y \\ 0 & \tilde{M} \end{bmatrix}\begin{bmatrix} M & \tilde{Y} \\ -N & \tilde{X} \end{bmatrix} = \begin{bmatrix} I - YN & Y\tilde{X} + M^{-1}\tilde{Y} \\ -\tilde{M}N & \tilde{M}\tilde{X} \end{bmatrix},$$

so that, taking into account (6.5),

$$\tilde{T}_{\text{aux}} = \begin{bmatrix} I - YN & Y\tilde{X} \\ -\tilde{M}N & \tilde{M}\tilde{X} \end{bmatrix} + \begin{bmatrix} I \\ 0 \end{bmatrix}\left(M^{-1}\tilde{Y} + M^{-1}R(I - PR)^{-1}\tilde{M}^{-1}\right)\begin{bmatrix} 0 & I \end{bmatrix}. \tag{6.7}$$

The first term on the right-hand side above is stable irrespective of R, which affects only the $(1, 2)$ sub-block of $\tilde{T}_{\text{aux}}$. Therefore, we have that

$$T_{\text{aux}} \in RH_\infty \iff Q := M^{-1}\tilde{Y} + M^{-1}R(I - PR)^{-1}\tilde{M}^{-1} = \mathcal{F}_{\text{u}}\left(\begin{bmatrix} P & \tilde{M}^{-1} \\ M^{-1} & M^{-1}\tilde{Y} \end{bmatrix}, R\right) \in RH_\infty.$$

Now, because

$$\begin{aligned}
\begin{bmatrix} P & \tilde{M}^{-1} \\ M^{-1} & M^{-1}\tilde{Y} \end{bmatrix} &= \begin{bmatrix} 0 & \tilde{M}^{-1} \\ M^{-1} & 0 \end{bmatrix}\begin{bmatrix} I & \tilde{Y} \\ \tilde{N} & I \end{bmatrix} \\
&= \begin{bmatrix} 0 & M \\ \tilde{M} & 0 \end{bmatrix}^{-1}\begin{bmatrix} I + \tilde{Y}(I - \tilde{N}\tilde{Y})^{-1}\tilde{N} & -\tilde{Y}(I - \tilde{N}\tilde{Y})^{-1} \\ -(I - \tilde{N}\tilde{Y})^{-1}\tilde{N} & (I - \tilde{N}\tilde{Y})^{-1} \end{bmatrix}^{-1} && \text{(by (B.15a))} \\
&= \left(\begin{bmatrix} I + \tilde{Y}\tilde{X}^{-1}\tilde{M}^{-1}\tilde{N} & -\tilde{Y}\tilde{X}^{-1}\tilde{M}^{-1} \\ -\tilde{X}^{-1}\tilde{M}^{-1}\tilde{N} & \tilde{X}^{-1}\tilde{M}^{-1} \end{bmatrix}\begin{bmatrix} 0 & M \\ \tilde{M} & 0 \end{bmatrix}\right)^{-1} && \text{(by (6.4))} \\
&= \begin{bmatrix} -\tilde{Y}\tilde{X}^{-1} & M + \tilde{Y}\tilde{X}^{-1}N \\ \tilde{X}^{-1} & -\tilde{X}^{-1}N \end{bmatrix}^{-1}
\end{aligned}$$

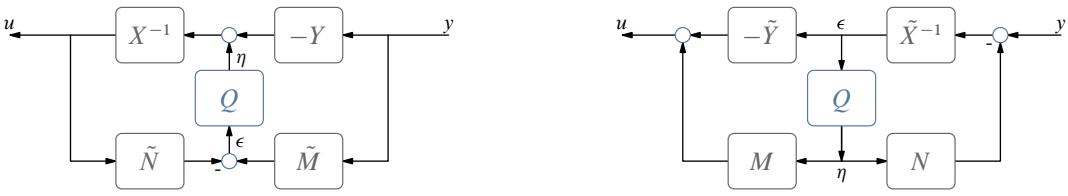
is invertible and so are its $(1, 2)$ and $(2, 1)$ sub-blocks, it follows from Proposition 5.6 that the mapping $R \mapsto Q$ is bijective whenever $I + P(\infty)R(\infty)$ is nonsingular. Repeating then the arguments used to prove Theorem 6.5 we end up with the LFT in (6.6b). Opening up this LFT, we have that

$$\begin{aligned}
R &= -\tilde{Y}\tilde{X}^{-1} + (M + \tilde{Y}\tilde{X}^{-1}N)Q(I + \tilde{X}^{-1}NQ)^{-1}\tilde{X}^{-1} \\
&= \left(-\tilde{Y}\tilde{X}^{-1}(\tilde{X} + NQ) + (M + \tilde{Y}\tilde{X}^{-1}N)Q\right)(\tilde{X} + NQ)^{-1} = (-\tilde{Y} + MQ)(\tilde{X} + NQ)^{-1},
\end{aligned}$$

which is exactly the second equality in (6.6b). The formulae in (6.6a) follow then by straightforward algebra and (6.4). □

The characterization of the set of all stabilizing controllers for the system in Fig. 6.1 presented in Theorem 6.6 is known as the *Youla–Kučera parametrization*. In the particular case of $P \in RH_\infty$ the Youla–Kučera parametrization matches that of Theorem 6.5. This follows from the possibility to choose $N = \tilde{N} = P$, $M = X = I_m$, $\tilde{M} = \tilde{X} = I_p$, and $Y = \tilde{Y} = 0$ in that case. The controller corresponding to the trivial choice $Q = 0$, i.e. $R = -\tilde{Y}\tilde{X}^{-1} = -X^{-1}Y$, is sometimes called the central controller of this parametrization. Because the Bézout coefficients (as well as their respective coprime factors) are not unique, neither is the central controller. Still, it might be convenient to have a controller, around which the whole parametrization is built and with respect to which the free parameter Q is analyzed.

To interpret the effect of the free parameter Q on the central controller, consider the block-diagrams in Fig. 6.4. They present two forms of the Youla–Kučera parametrization, those given by the *lcf* (6.6a) and by the *rcf* (6.6b). The signals y and u stand for the measured plant output and the control input generated by the controller. In the nominal case, when the plant model is accurate and no exogenous inputs, like disturbances and measurement noise, affect the plant, these signals are related as $y = NM^{-1}u = \tilde{M}^{-1}\tilde{N}u$.



(a) In the *lcf* form (6.6a) (b) In the *rcf* form (6.6b)

Fig. 6.4: Youla–Kučera parametrization of stabilizing controllers

Fig. 6.4(a): The input to Q is $\epsilon = \tilde{M}y - \tilde{N}u$. In the nominal case, when $\tilde{M}y = \tilde{N}u$, we have that $\epsilon = 0$. In this situation only the central controller acts. The signal ϵ can then be viewed as an indicator of the mismatch between expected and actual behavior of the control system, be it due to modeling uncertainty or due to the effect of exogenous signals (disturbances). The block Q is activated only if this indicator generates an "uncertainty alert" notification.

Fig. 6.4(b): The signal η generated by the block Q affects both the plant input, via adding $M\eta$ to it, and the measured output, via subtracting $N\eta$ from it. If the plant model is perfectly known, the input injection $M\eta$ propagates to y as the signal $PM\eta = N\eta$, which is canceled out by $-N\eta$ coming directly from Q. Thus, Q affects the controller behavior only if there is a mismatch between the model and the actual plant. However, unlike the setup in Fig. 6.4(a), exogenous signals affecting the plant have no affect on this process.

It should be emphasized that these interpretations are based on idealized scenarios. Modeling uncertainties and disturbances must be present in any realistic setup, so the effect of Q is always perceptible.

Remark 6.4 (non-uniqueness of Bézout coefficients). As already mentioned above, the non-uniqueness of the doubly coprime factorization (6.4) may affect the central controller in (6.6). But it does not offer any additional degree of freedom to the Q-parametrization. To see that, consider another choice of coprime factors of P, say N_1, M_1, $\tilde{N}_1$, and $\tilde{M}_1$. By Proposition 3.2, there are bi-stable U and $\tilde{U}$ such that $N_1 = NU$, $M_1 = MU$, $\tilde{N}_1 = \tilde{U}\tilde{N}$, and $\tilde{M}_1 = \tilde{U}\tilde{M}$. Bézout coefficients of the new factorization resulting in (6.4) satisfy then

$$\begin{bmatrix} I & 0 \\ 0 & \tilde{U} \end{bmatrix} \begin{bmatrix} X_1 & Y_1 \\ -\tilde{N} & \tilde{M} \end{bmatrix} \begin{bmatrix} M & -\tilde{Y}_1 \\ N & \tilde{X}_1 \end{bmatrix} \begin{bmatrix} U & 0 \\ 0 & I \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}$$

or, equivalently,

$$\begin{bmatrix} U & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} X_1 & Y_1 \\ -\tilde{N} & \tilde{M} \end{bmatrix} \begin{bmatrix} M & -\tilde{Y}_1 \\ N & \tilde{X}_1 \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & \tilde{U} \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}$$

Hence, we may have that $Y_1 = U^{-1}Y$, $X_1 = U^{-1}X$, $\tilde{Y}_1 = \tilde{Y}\tilde{U}^{-1}$, and $\tilde{X}_1 = \tilde{X}\tilde{U}^{-1}$. In this case, (6.6b) in terms of the new factorization reads

$$\begin{aligned} R &= (-\tilde{Y}_1 + M_1Q_1)(\tilde{X}_1 + N_1Q_1)^{-1} = (-\tilde{Y}\tilde{U}^{-1} + MUQ_1)(\tilde{X}\tilde{U}^{-1} + NUQ_1)^{-1} \\ &= (-\tilde{Y} + MUQ_1\tilde{U})(\tilde{X} + NUQ_1\tilde{U})^{-1} \end{aligned}$$

and we return to the original parametrization for any given Q with $Q_1 = U^{-1}Q\tilde{U}^{-1}$, which is admissible because $U^{-1}, \tilde{U}^{-1} \in RH_\infty$. Thus, a different choice of coprime factors of P merely results in a different choice of Q to end up with the same stabilizing controller in form (6.6b). Similar arguments apply to form (6.6a) of the parametrization. ▽

Remark 6.5 (coprime factors of stabilizing controllers). It is readily verified that

$$(X + Q\tilde{N})M + (-Y + Q\tilde{M})(-N) = I \quad \text{and} \quad \tilde{M}(\tilde{X} + NQ) + (-\tilde{N})(-\tilde{Y} + MQ) = I$$

for all Q. This implies that the right-hand sides of (6.6a) and (6.6b) constitute *rcf* and *lcf* of stabilizing R over RH_∞, respectively. Another consequence of the Bézout equalities above is that every pair of Bézout coefficient of coprime factorizations of the plant constitute coprime factors of a stabilizing controller for it. Thus, constructing Bézout coefficients is effectively equivalent to constructing a stabilizing controller (this fact is used in the proof of Theorem 6.8 below). This analogy works both ways. Say, if we have a simple stabilizing controller, then its coprime factors can serve as Bézout coefficients for a coprime factorization of the plant. ▽

Like in the stable case, the Youla–Kučera parametrization can be used to simplify the closed-loop system T_{aux}. To this end, rewrite (6.7) as

$$\tilde{T}_{\text{aux}} = \begin{bmatrix} I \\ 0 \end{bmatrix} \begin{bmatrix} I & 0 \end{bmatrix} + \begin{bmatrix} Y \\ \tilde{M} \end{bmatrix} \begin{bmatrix} -N & \tilde{X} \end{bmatrix} + \begin{bmatrix} I \\ 0 \end{bmatrix} Q \begin{bmatrix} 0 & I \end{bmatrix}.$$

It is then a matter of tedious but straightforward algebra to see that

$$T_{\text{aux}} = \begin{bmatrix} MX & -MY \\ NX & I - NY \end{bmatrix} + \begin{bmatrix} M \\ N \end{bmatrix} Q \begin{bmatrix} \tilde{N} & \tilde{M} \end{bmatrix} = \begin{bmatrix} MX + MQ\tilde{N} & -MY + MQ\tilde{M} \\ NX + NQ\tilde{N} & I - NY + NQ\tilde{M} \end{bmatrix}.$$

This is again an affine function of Q, which simplifies performance analyses of the closed-loop system, as well as controller design procedures, see Chapter 7 for further details.

Remark 6.6 (stability and domains of systems in the loop). The first output of T_{aux},

$$e_1 = M\big(-Yv_1 + Xv_2 + Q(\tilde{M}v_1 + \tilde{N}v_2)\big),$$

is the input signal entering P in Fig. 6.1(b). The formula above implies that $e_1 \in \mathfrak{D}_P$ for all $v_1, v_2 \in L_2$, cf. Proposition 3.4. In other words, every stabilizing controller ensures that the plant input is always in its domain, which is expectable. Likewise, the second output of T_{aux} is

$$e_2 = (NX + NQ\tilde{N})v_1 + (I - NY + NQ\tilde{M})v_2 = (\tilde{X} + NQ)(\tilde{N}v_1 + \tilde{M}v_2),$$

where the relations $NX = \tilde{X}\tilde{N}$ and $I - NY = \tilde{X}\tilde{M}$ resulting from (6.4) are used. Taking into account the discussion in Remark 6.5, this relation implies that any stabilizing controller should ensure that its input signal $e_2 \in \mathfrak{D}_R$. ▽

The generator of all stabilizing controllers in the Youla–Kučera parametrization can be constructed via a state-space realization of P. Suppose

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

and that (A, B) is stabilizable and (C, A) is detectable (i.e. no minimality assumption is required). We aim at constructing a state-space realization of the 2×2 generator of the LFT in (6.6b). A key observation toward this end is that

$$\begin{bmatrix} u \\ \epsilon \end{bmatrix} = J \begin{bmatrix} y \\ \eta \end{bmatrix} := \begin{bmatrix} -\tilde{Y}\tilde{X}^{-1} & M + \tilde{Y}\tilde{X}^{-1}N \\ \tilde{X}^{-1} & -\tilde{X}^{-1}N \end{bmatrix} \begin{bmatrix} y \\ \eta \end{bmatrix} \iff \begin{bmatrix} u \\ y \end{bmatrix} = \begin{bmatrix} M & -\tilde{Y} \\ N & \tilde{X} \end{bmatrix} \begin{bmatrix} \eta \\ \epsilon \end{bmatrix},$$

so that the logic of Remark 4.3 on p. 79 can be used. Namely, (4.21b) reads

$$\begin{bmatrix} M & -\tilde{Y} \\ N & \tilde{X} \end{bmatrix} : \begin{cases} \dot{\hat{x}}(t) = (A + BK)\hat{x}(t) + B\eta(t) - L\epsilon(t) \\ u(t) = K\hat{x}(t) + \eta(t) \\ y(t) = (C + DK)\hat{x}(t) + D\eta(t) + \epsilon(t) \end{cases}$$

where K and L are any matrices such that $A + BK$ and $A + LC$ are Hurwitz. It follows from the output equation above that $\epsilon = -(C + DK)\hat{x} + y - D\eta$. Hence, a realization of J is that between the rearranged signals,

$$J : \begin{cases} \dot{\hat{x}}(t) = (A + BK + LC + LDK)\hat{x}(t) - Ly(t) + (B + LD)\eta(t) \\ u(t) = K\hat{x}(t) + \eta(t) \\ \epsilon(t) = -(C + DK)\hat{x}(t) + y(t) - D\eta(t) \end{cases} \tag{6.8}$$

and the first equality in (6.6b) reads

$$R(s) = \mathcal{F}_l\left(\left[\begin{array}{c|cc} A + BK + LC + LDK & -L & B + LD \\ \hline K & 0 & I \\ -C - DK & I & -D \end{array}\right], Q(s)\right) \tag{6.9}$$

with the well-posedness requirement $\det(I - Q(\infty)D) \neq 0$.

Remark 6.7 (Q-parametrization of general controllers). The arguments used to derive the parametrizations in (6.6), as well as that in (6.9), are based on transforming I/O signals in Fig. 6.1 and did not touch the controller R itself. Hence, they apply almost literally to the case when R is not constrained to be LTI or finite dimensional. This is why a sheer replacement of $Q \in RH_\infty$ in Theorem 6.6 with any causal system—possibly nonlinear, time varying, or infinite dimensional—bounded as an operator on L_2 does not change the result. The only subtlety here is that the well-posedness condition should be reformulated if $P(\infty) \neq 0$, perhaps via ideas from [30, Ch. 4]. We then end up with an exhaustive parametrization of all nonlinear time-varying infinite-dimensional controllers for finite-dimensional LTI plants. ▽

6.2.3 All stabilizing controllers based on a given one

Return to the parametrization of all stabilizing controllers in (6.9). To gain an additional insight into it, rearrange the state equation in (6.8) as

$$\begin{aligned} \dot{\hat{x}}(t) &= A\hat{x}(t) + B\big(K\hat{x}(t) + \eta(t)\big) - L\big(y(t) - C\hat{x}(t) - D(K\hat{x}(t) + \eta(t))\big) \\ &= A\hat{x}(t) + Bu(t) - L\big(y(t) - C\hat{x}(t) - Du(t)\big). \end{aligned}$$

This is the conventional observer equation. The signal $\epsilon = y - C\hat{x} - Du = C(x - \hat{x})$ is an indicator of the mismatch between the plant state x and its estimation $\hat{x}$, obtained from the measurement equation $y = Cx + Du$ of the plant P (ϵ is known as the *innovations process* in the Kalman filtering theory). The control signal $u = K\hat{x} + \eta$ is then almost the standard observer-based control law, modulo the addition of the "correction" signal $\eta = Q\epsilon$.

The central controller of (6.9), the one with $Q = 0$ and thus $\eta = 0$, is the classical observer-based controller, which is known to stabilize the plant. It means that all stabilizing controllers in this case can be constructed as an extension of a given stabilizing controller. It may be of interest to extend this characterization to an arbitrary given stabilizing controllers, say C_0. This is indeed possible and to derive such a parametrization a preliminary technical result, which is of independent interest, is required.

Lemma 6.7. *The following conditions are equivalent:*

1. *R internally stabilizes P,*
2. $\begin{bmatrix} M & -N_R \\ -N & M_R \end{bmatrix}^{-1} \in RH_\infty$,
3. $\begin{bmatrix} \tilde{M} & -\tilde{N} \\ -\tilde{N}_R & \tilde{M}_R \end{bmatrix}^{-1} \in RH_\infty$,

4. $(\tilde{M}_R M - \tilde{N}_R N)^{-1} \in RH_\infty$,
5. $(\tilde{M} M_R - \tilde{N} N_R)^{-1} \in RH_\infty$,

where components of the transfer functions above are coprime factors of the plant $P = NM^{-1} = \tilde{M}^{-1}\tilde{N}$ *and the controller* $R = N_R M_R^{-1} = \tilde{M}_R^{-1}\tilde{N}_R$ *over* RH_∞.

Proof. With these factorizations, the system T_{aux} defined by (6.2) can be expressed as follows:

$$T_{\text{aux}} := \begin{bmatrix} I & -R \\ -P & I \end{bmatrix}^{-1} = \begin{bmatrix} M & 0 \\ 0 & M_R \end{bmatrix} \begin{bmatrix} M & -N_R \\ -N & M_R \end{bmatrix}^{-1} \tag{6.10a}$$

$$= \begin{bmatrix} \tilde{M}_R & -\tilde{N}_R \\ -\tilde{N} & \tilde{M} \end{bmatrix}^{-1} \begin{bmatrix} \tilde{M}_R & 0 \\ 0 & \tilde{M} \end{bmatrix} \tag{6.10b}$$

$$= \begin{bmatrix} M & 0 \\ N & I \end{bmatrix} \begin{bmatrix} (\tilde{M}_R M - \tilde{N}_R N)^{-1} & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} \tilde{M}_R & \tilde{N}_R \\ 0 & I \end{bmatrix}, \tag{6.10c}$$

where the last equality follows by (B.14a) on p. 195 and the relation $I - RP = \tilde{M}_R^{-1}(\tilde{M}_R M - \tilde{N}_R N)M^{-1}$.

$1 \iff 2$: Let $\tilde{Y}, \tilde{X}$ and $\tilde{Y}_R, \tilde{X}_R$ be the Bézout coefficients for N, M and N_R, M_R, respectively. It is readily verified that

$$\begin{bmatrix} \tilde{M}_R & -\tilde{Y} \\ -\tilde{Y}_R & \tilde{M} \end{bmatrix} \begin{bmatrix} M & -N_R \\ -N & M_R \end{bmatrix} + \begin{bmatrix} \tilde{X} - \tilde{M}_R & \tilde{Y} + \tilde{N}_R \\ \tilde{N} + \tilde{Y}_R & -\tilde{M} + \tilde{X}_R \end{bmatrix} \begin{bmatrix} M & 0 \\ 0 & M_R \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}.$$

This proves that the factors in (6.10a) constitute a *rcf* of T_{aux} over RH_∞. Hence, the sought equivalence follows by Proposition 3.3.

$1 \iff 3$: Follows by similar arguments from the left coprimeness of the factors in (6.10b).

$1 \implies 4 \wedge 5$: Follows by the already proved fact that $1 \iff 2 \wedge 3$ and the relation

$$\begin{bmatrix} M & -N_R \\ -N & M_R \end{bmatrix}^{-1} \begin{bmatrix} \tilde{M}_R & \tilde{N}_R \\ \tilde{N} & \tilde{M} \end{bmatrix}^{-1} = \begin{bmatrix} (\tilde{M}_R M - \tilde{N}_R N)^{-1} & 0 \\ 0 & (\tilde{M} M_R - \tilde{N} N_R)^{-1} \end{bmatrix},$$

which can be verified by a straightforward algebra.

$4 \implies 1$: Follows by (6.10c).

The last two items also prove that $5 \implies 1$, which completes the proof. □

Now, suppose that a finite-dimensional LTI controller R_0 internally stabilizes the plant. Bring in its coprime factorizations over RH_∞,

$$R_0 = N_{R_0} M_{R_0}^{-1} = \tilde{M}_{R_0}^{-1} \tilde{N}_{R_0}$$

for appropriately dimensional $N_{R_0}, M_{R_0}, \tilde{N}_{R_0}, \tilde{M}_{R_0} \in RH_\infty$. The set of all stabilizing controllers having this R_0 as its central controller is given by the following result.

Theorem 6.8 (Q-parametrization from R_0)**.** *R stabilizes the system in Fig. 6.1 iff there exists $Q \in RH_\infty$ such that*

$$R = \mathcal{F}_l\left(\begin{bmatrix} R_0 & I \\ I & -P(I - R_0 P)^{-1} \end{bmatrix}, \tilde{M}_{R_0}^{-1} Q M_{R_0}^{-1} \right) \tag{6.11a}$$

$$= \big(N_{R_0} + (\tilde{M}_{R_0} - \tilde{N}_{R_0} P)^{-1} Q\big)\big(M_{R_0} + P(\tilde{M}_{R_0} - \tilde{N}_{R_0} P)^{-1} Q\big)^{-1} \tag{6.11b}$$

$$= \big(-\tilde{M}_{R_0} + Q(M_{R_0} - P N_{R_0})^{-1} P\big)^{-1} \big(\tilde{N}_{R_0} + Q(M_{R_0} - P N_{R_0})^{-1}\big) \tag{6.11c}$$

and such that $I + P(\infty)(\tilde{M}_{R_0}^{-1}(\infty) Q(\infty) M_{R_0}^{-1}(\infty) - C_0(\infty))$ is nonsingular.

Proof. The idea behind the proof is to construct right (left) coprime factors of P having $-\tilde{N}_{R_0}$ and $\tilde{M}_{R_0}$ ($-N_{R_0}$ and M_{R_0}) as the Bézout coefficients for them, so that the formulae of Theorem 6.6 can be used. To this end, start with any coprime factorizations $P = \tilde{M}^{-1}\tilde{N} = NM^{-1}$. By Lemma 6.7, the stability of the closed-loop system implies that $U_0 := \tilde{M}_{R_0}M - \tilde{N}_{R_0}N$ and $\tilde{U}_0 := \tilde{M}M_{R_0} - \tilde{N}N_{R_0}$ are bi-stable. By Proposition 3.2, MU_0^{-1} and NU_0^{-1} ($\tilde{U}_0^{-1}\tilde{M}$ and $\tilde{U}_0^{-1}\tilde{N}$) are then also right (left) coprime factors of P. But then

$$\begin{bmatrix} \tilde{U}_0^{-1}\tilde{M} & -\tilde{U}_0^{-1}\tilde{N} \\ -\tilde{N}_{R_0} & \tilde{M}_{R_0} \end{bmatrix} \begin{bmatrix} M_{R_0} & NU_0^{-1} \\ N_{R_0} & MU_0^{-1} \end{bmatrix} = I$$

and we have our sought factorizations, cf. (6.4). By (6.6b), all stabilizing controllers can then be characterized as $(N_{R_0} + MU_0^{-1}Q)(M_{R_0} + NU_0^{-1}Q)^{-1}$ or $(\tilde{M}_{R_0} + Q\tilde{U}_0^{-1}\tilde{N})^{-1}(\tilde{N}_{R_0} + Q\tilde{U}_0^{-1}\tilde{M})$. The formulae in (6.11) follow then by straightforward algebra. □

Remark 6.8 (all stabilized plants). The stability setup in Fig. 6.1 is symmetric with respect to the plant and the controller in it. In other words, P and R can be interchanged. As a result, if we know one plant, say P_0, stabilized by a given controller R, we can exhaustively characterize all plants stabilized by that very controller. This class can be obtained by "mirroring" the formula of Theorem 6.8 as

$$P = \mathcal{F}_{\mathrm{u}}\left(\begin{bmatrix} -R(I - P_0R)^{-1} & I \\ I & P_0 \end{bmatrix}, \tilde{M}_0^{-1}QM_0^{-1}\right),$$

where M_0 and $\tilde{M}_0$ are the denominators of coprime factorizations of P_0 over RH_∞ (the upper LFT is used merely for aesthetic reasons). Characterizations of this kind are used in some closed-loop identification algorithms, where the task is to identify a more accurate plant model from closed-loop experiments via adjusting Q. Perhaps, the parametrization above can also be used in some control design problems. ▽

6.2.4 Extensions

Up to this point only finite-dimensional LTI systems having real-rational transfer functions were considered. Yet apart from somewhat more involved well-posedness conditions, all arguments of §6.2.2 apply to infinite-dimensional systems, whose transfer functions are irrational, literally.

A potential obstacle in exploiting this direction might be the construction of doubly coprime factorizations of irrational transfer functions. There are general-purpose methods, expressing coprime factors, or counterparts to the state-space formula in (6.9), in terms of some operator equations. However, the resulted controllers might be non-transparent and hard to implement. Still, for some relatively simple, yet quite practical, classes of infinite-dimensional problems the arguments of §6.2.2 can be applied in a neat way. One such problem is studied below.

Consider first the following abstract class of infinite-dimensional LTI plants:

$$P = P_{\mathrm{fd}} + \Pi, \tag{6.12}$$

where P_{fd} is a finite-dimensional LTI system having a real-rational transfer function $P_{\mathrm{fd}}(s)$ and $\Pi \in H_\infty$ but otherwise unconstrained. This class of systems effectively contains all systems having a finite number of unstable poles. The following result shows that a doubly coprime factorization of P can be constructed from that of its finite-dimensional part.

Lemma 6.9. *Let $P_{fd} = N_{fd}M_{fd}^{-1} = \tilde{M}_{fd}^{-1}\tilde{N}_{fd}$, where $N_{fd}, M_{fd}, \tilde{N}_{fd}, \tilde{M}_{fd} \in RH_\infty$ satisfy*

$$\begin{bmatrix} X_{fd} & Y_{fd} \\ -\tilde{N}_{fd} & \tilde{M}_{fd} \end{bmatrix} \begin{bmatrix} M_{fd} & -\tilde{Y}_{fd} \\ N_{fd} & \tilde{X}_{fd} \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}$$

for appropriately dimensional RH_∞ Bézout coefficients. Then

$$\begin{bmatrix} X & Y \\ -\tilde{N} & \tilde{M} \end{bmatrix} := \begin{bmatrix} X_{fd} - Y_{fd}\Pi & Y_{fd} \\ -\tilde{N}_{fd} - \tilde{M}_{fd}\Pi & \tilde{M}_{fd} \end{bmatrix} \quad \textit{and} \quad \begin{bmatrix} M & -\tilde{Y} \\ N & \tilde{X} \end{bmatrix} := \begin{bmatrix} M_{fd} & -\tilde{Y}_{fd} \\ N_{fd} + \Pi M_{fd} & \tilde{X}_{fd} - \Pi \tilde{Y}_{fd} \end{bmatrix}$$

constitute a doubly coprime factorization of $P = P_{fd} + \Pi = NM^{-1} = \tilde{M}^{-1}\tilde{N}$ over H_∞.

Proof. Follows from the relation

$$\begin{bmatrix} \tilde{M} & -\tilde{N} \\ Y & X \end{bmatrix} \begin{bmatrix} \tilde{X} & N \\ -\tilde{Y} & M \end{bmatrix} = \begin{bmatrix} X_{\mathrm{fd}} & Y_{\mathrm{fd}} \\ -\tilde{N}_{\mathrm{fd}} & \tilde{M}_{\mathrm{fd}} \end{bmatrix} \begin{bmatrix} I & 0 \\ -\Pi & I \end{bmatrix} \begin{bmatrix} I & 0 \\ \Pi & I \end{bmatrix} \begin{bmatrix} M_{\mathrm{fd}} & -\tilde{Y}_{\mathrm{fd}} \\ N_{\mathrm{fd}} & \tilde{X}_{\mathrm{fd}} \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}$$

and the facts that $M = M_{\mathrm{fd}}$ and $\tilde{M} = \tilde{M}_{\mathrm{fd}}$ are bi-proper. □

The transparency of the coprime factors formulae pays off in the controller architecture. Applying the parametrization of Theorem 6.6, all controllers stabilizing this P can be presented as

$$\begin{aligned} R &= (-\tilde{Y} + MQ)(\tilde{X} + NQ)^{-1} = (-\tilde{Y}_{\mathrm{fd}} + M_{\mathrm{fd}}Q)(\tilde{X}_{\mathrm{fd}} + N_{\mathrm{fd}}Q + \Pi(-\tilde{Y}_{\mathrm{fd}} + M_{\mathrm{fd}}))^{-1} \\ &= (-\tilde{Y}_{\mathrm{fd}} + M_{\mathrm{fd}}Q)(\tilde{X}_{\mathrm{fd}} + N_{\mathrm{fd}}Q)^{-1}\big(I + \Pi(-\tilde{Y}_{\mathrm{fd}} + M_{\mathrm{fd}})(\tilde{X}_{\mathrm{fd}} + N_{\mathrm{fd}}Q)^{-1}\big)^{-1} \\ &= R_{\mathrm{fd}}(I + \Pi R_{\mathrm{fd}})^{-1}, \end{aligned}$$

where $R_{\mathrm{fd}} = (-\tilde{Y}_{\mathrm{fd}} + M_{\mathrm{fd}}Q)(\tilde{X}_{\mathrm{fd}} + N_{\mathrm{fd}}Q)^{-1}$. This is the feedback interconnection of $G_1 = R_{\mathrm{fd}}$ and $G_2 = -\Pi$ in the setup presented in Fig. 5.1(c). Here R_{fd} is the parametrization of all stabilizing controllers for the finite-dimensional part of the plant, P_{fd}. Thus, altering a plant by a stable system connected in parallel can always be counteracted by adding the same stable system as an internal feedback to the controller. This is a frequently used trick, known as *loop shifting*.

This result has several useful applications. The best known of them is perhaps that to the stabilization of dead-time systems. Let

$$P(s) = P_0(s)\mathrm{e}^{-\tau s} = \left[\begin{array}{c|c} A & B \\ \hline R & 0 \end{array}\right] \mathrm{e}^{-\tau s}$$

for some delay $\tau > 0$. What we need is to transform this P to form (6.12). This is a straightforward task in the stable case. Indeed,

$$P(s) = P_0(s) - P_0(s)(1 - \mathrm{e}^{-\tau s})$$

is what we need because $\Pi = -P_0(1 - \mathrm{e}^{-\tau s}) \in H_\infty$ then. Adding this $-P_0(1 - \mathrm{e}^{-\tau s})$ in feedback with a finite-dimensional controller designed for the delay-free system P_0 results in the celebrated *Smith controller* proposed in [27], which is a well-understood controller architecture for dead-time systems.

If P_0 is unstable, the split of P as in (6.12) is still possible. Arguably, the easiest way to see that is via the impulse response of P (derived by delaying (4.4)),

$$\begin{aligned} p(t) &= C\mathrm{e}^{A(t-\tau)}B\mathbb{1}(t-\tau) = C\mathrm{e}^{A(t-\tau)}B\mathbb{1}(t) - C\mathrm{e}^{A(t-\tau)}B\mathbb{1}_{[0,\tau]}(t) \\ &= C\mathrm{e}^{-A\tau}\,\mathrm{e}^{At}B\mathbb{1}(t) - C\mathrm{e}^{A(t-\tau)}B\mathbb{1}_{[0,\tau]}(t) =: p_{\mathrm{fd}}(t) + \pi(t). \end{aligned}$$

The first term above is the impulse response of a finite-dimensional system P_{fd}, whose transfer function

$$P_{\mathrm{fd}}(s) = \left[\begin{array}{c|c} A & B \\ \hline C\mathrm{e}^{-A\tau} & 0 \end{array}\right] = \left[\begin{array}{c|c} A & \mathrm{e}^{-A\tau}B \\ \hline C & 0 \end{array}\right]. \tag{6.13}$$

The second term corresponds to an FIR (finite impulse response) system Π whose transfer function

$$\Pi(s) = \mathfrak{L}\{\pi\} = -C\int_0^\tau \mathrm{e}^{A(t-\tau)}\mathrm{e}^{-st}\,\mathrm{d}t\,B$$

does belong to H_∞. To see that, we need to show that $\Pi(s)$ is holomorphic and bounded in $\mathbb{C}_0$. As $\Pi(s)$ is an entire function of s (integral of an exponential function), the first part is obvious. Now, for all $s \in \mathbb{C}_0$

$$\|\Pi(s)\| \le \int_0^\tau |\mathrm{e}^{-st}| \|C\mathrm{e}^{A(t-\tau)}B\| \mathrm{d}t < \int_0^\tau \|C\mathrm{e}^{-At}B\| \mathrm{d}t < \infty,$$

where the second inequality follows by the fact that $|\mathrm{e}^{-st}| = \mathrm{e}^{-t\,\mathrm{Re}\,s} < 1$ for all $\mathrm{Re}\,s > 0$ and $t \ge 0$. Hence, $\Pi(s)$ is bounded in $\mathbb{C}_0$ and thus belongs to H_∞. The FIR Π above, known as the *modified Smith predictor*, was proposed in [29]. As a matter of fact, $\Pi(s)$ can be alternatively presented as

$$\Pi(s) = C(\mathrm{e}^{-A\tau} - \mathrm{e}^{-\tau s}I)(sI - A)^{-1}B,$$

whose all singularities at the eigenvalues of A are removable. The case of $A = 0$ and $B = C = 1$, when $\Pi(s) = (1 - \mathrm{e}^{-\tau s})/s$, gives the finite-memory integrator $G_{\mathrm{fmint},\tau}$ studied in Chapter 3.

6.3 Open-loop stabilization

Unlike the feedback interconnection, (open-loop) parallel and series interconnections can only alter joint dynamics via canceling some of their parts. These cancellations are not limited to cancellations between poles and zeros, they might reflect directional properties or even more complicated phenomena. Unstable dynamics can be stabilized via cancellations. For example, if $G_1(s) = -1/(s(s+1))$, placing in parallel with it $G_2(s) = 1/s$ yields a stable system with the transfer function $G(s) = G_1(s) + G_2(s) = 1/(s+1)$. A less trivial example is the (unstable) system with the transfer function $G_1(s) = 1/(s + 1 + s\mathrm{e}^{-s})$ studied in Remark 3.4 on p. 48. It can be stabilized by connecting it in series with any finite-dimensional low-pass filter, e.g. $G_2(s) = 1/(s+1)$, as in that case one can show that $G_2G_1 \in H_\infty$. Although no poles of $G_1(s)$ are canceled here, poles are not the cause of its instability. Of course, stabilization via open-loop cancellations would be quite fragile (arbitrarily small mismatches ruin it) and would not guarantee internal stability. Therefore, this is by no means a practical stabilization method.

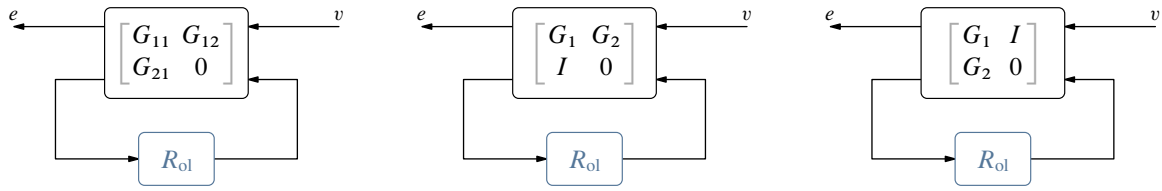
Nonetheless, the ideas behind it are useful in shaping steady-state behavior of controlled systems. To clarify this statement, consider the open-loop tracking problem in the configuration of Fig. 1.4(b) for a stable plant with the transfer function $P(s) = 1/(s+1)$ and a stable controller R to be designed to reduce the tracking error $e = y_\mathrm{r} - y$. The transfer function of the system $y_\mathrm{r} \mapsto e$ there is $T_\mathrm{e}(s) = 1 - R(s)/(s+1)$ and it is stable whenever so is R, i.e. there are no constraints on $R(s)$ apart from the obvious requirement to be stable. But if we also need to have a zero steady-state error for the step $y_\mathrm{r}(t) = \mathbb{1}(t)$, which reads

$$\lim_{t\to\infty} e(t) = \lim_{s\to 0} sT_\mathrm{e}(s)\frac{1}{s} = T_\mathrm{e}(0) = 1 - R(0)$$

by the final value theorem, then the condition $R(0) = 1$ must hold. This is an additional (interpolation) constraint to be imposed on R. At the same time, consider the problem of stabilizing $T_\mathrm{e}W = W - PRW$, where the (unstable) weight $W(s) = 1/s$. It is readily seen that

$$T_\mathrm{e}(s)W(s) = \frac{1}{s} - \frac{R(s)}{s(s+1)} = \frac{1 - R(0)}{s} - \frac{R(s) - R(0)(s+1)}{s(s+1)}.$$

The second term on the right-hand side above has a removable singularity at the origin and can be shown to be an RH_∞ function whenever $R \in RH_\infty$. This implies that all instabilities of $T_\mathrm{e}W$ are in the first term above. Hence, $T_\mathrm{e}W$ is stable iff $R(0) = 1$, which is the same interpolation constraint as that guaranteeing the zero steady-state error to a step reference, discussed above. In other words, the steady-state requirement in this case can be cast as a stability requirement for an appropriately chosen (unstable) weighting function



(a) General two-sided setup (b) One-sided "tracking" setup (c) One-sided "estimation" setup

Fig. 6.5: Open-loop stabilization configurations

W. This W is not a system existent "in the flesh," whose unstable dynamics should be stabilized. Rather, it is a fictitious system, whose sole purpose is to shape the error transfer function at certain points in the complex plane.

Thus, if considered in an appropriate context, stabilization by cancellations has no hidden hazards in it. A general open-loop stabilization setup is presented in Fig. 6.5(a), where systems G_{ij} are given and and a *stable* $R_{\rm ol}$ is to be selected to render the "error" system $G_{\rm e} : v \mapsto e$ stable as well. The zero G_{22} block in Fig. 6.5(a) is what makes it open loop, with the error system

$$G_{\rm e} = G_{11} + G_{12} R_{\rm ol} G_{21}$$

being an affine function of $R_{\rm ol}$. We are not concerned with internal signals here, so they are not named. The problem represented by Fig. 6.5(a) is dubbed the *two-sided* setup and might be rather complicated, see [14, Ch. 3]. For that reason, it shall not be studied here in its full generality. Rather, its special, *one-sided*, versions presented in Figs. 6.5(b) and 6.5(c) are considered. Their error systems, $G_1 + G_2 R_{\rm ol}$ and $G_1 + R_{\rm ol} G_2$, can be thought of as representing feedforward tracking and estimation problems, respectively (more details will be discussed in Chapter 7). These setups are, in a sense, "transpose" to each other and a solution to one of them can easily be derived from the other by algebraic duality. So only the "estimation" version is addressed below in details.

6.3.1 Stabilization in one-sided setting

Consider the interconnection in Fig. 6.5(c). Our task is to select $R_{\rm ol} \in RH_\infty$ such that

$$G_{\rm e} = G_1 + R_{\rm ol} G_2 \tag{6.14}$$

belongs to RH_∞ too. The problem is trivial if G_1 and G_2 are themselves stable, in which case any stable $R_{\rm ol}$ does the trick. If only G_2 is unstable, its instabilities can be canceled by $R_{\rm ol}$ via matching the directions of every unstable pole p_i of $G_2(s)$ by zero directions of $R_{\rm ol}(p_i)$, see §5.1.2. If G_1 is unstable, its unstable dynamics should be canceled by those of $R_{\rm ol} G_2$. Because $R_{\rm ol}$ is required to be stable, we must have every instability of G_1 present in G_2. The task of $R_{\rm ol}$ is then reshape the output directions of unstable poles of G_2 to match those of G_1, see §5.1.1. Thus, the stabilization problem is effectively to shape $R_{\rm ol}(p_i)$ at every unstable pole p_i of both $G_1(s)$ and $G_2(s)$. These can be viewed as problems of characterizing stable systems from given interpolations constraints on their transfer functions.

It should be emphasized that an appropriate shaping of $R_{\rm ol}$ is not always possible, even if all unstable poles of $G_1(s)$ are also those of $G_2(s)$. For example, let $G_1(s) = 1/s I_2$ and $G_2(s) = \operatorname{diag}\{1/s, 1/(s^2+1)\}$. The error system

$$G_{\rm e}(s) = \begin{bmatrix} (1 + R_{{\rm ol},11}(s))/s & R_{{\rm ol},12}(s)/(s^2+1) \\ R_{{\rm ol},21}(s)/s & 1/s + R_{{\rm ol},22}(s)/(s^2+1) \end{bmatrix}.$$

To stabilize its $(1,1)$ and $(2,1)$ elements, we need $R_{{\rm ol},11}(0) = -1$ and $R_{{\rm ol},21}(0) = 0$, which cancel their only instabilities at the origin. The $(1,2)$ element is stable iff $R_{{\rm ol},12}(\pm {\rm j}) = 0$, which cancels its unstable

poles at $s = \pm \mathrm{j}$. But there is no way to cancel the pole at the origin in the $(2, 2)$ element by a stable $R_{\mathrm{ol},22}$ (the poles at $s = \pm \mathrm{j}$ are canceled if $R_{\mathrm{ol},22}(\pm \mathrm{j}) = 0$). In general, we may expect that a stabilizing R_{ol} exists if input directions of every unstable mode of G_1 is contained in those of G_2. Otherwise, some instabilities of the former might be excited even if their counterparts of the latter are not, in which case R_{ol} cannot "see" them and cannot counteract.

Thus, the stabilization of the system in Fig. 6.5(c) boils down to two main technical issues. First, we should know how to characterize the "containment" of the directions of unstable poles of $G_1(s)$ in those of $G_2(s)$. Second, we should know how to generate (preferably, all) stable systems interpolating given points in the complex plane, directions counting. Addressing both these issues is substantially simplified by the use of coprime factorization machinery.

To this end, let $G_2 = N_2 M_2^{-1} = \tilde{M}_2^{-1} \tilde{N}_2$ constitute a doubly coprime factorization of G_2 with appropriate Bézout coefficients $X_2, Y_2, \tilde{X}_2, \tilde{Y}_2 \in RH_\infty$. The lemma below gives necessary and sufficient stabilizability conditions for the problem in Fig. 6.5(c).

Lemma 6.10. *The following conditions are equivalent:*

1. *there is $R_{ol} \in RH_\infty$ stabilizing G_e in* (6.14),
2. *$G_1 M_2 \in RH_\infty$ (equivalently, $\mathfrak{D}_{G_1} \subset \mathfrak{D}_{G_2}$, see Proposition 3.4),*
3. *the combined system admits a lcf of the form*

$$\begin{bmatrix} G_1 \\ G_2 \end{bmatrix} = \begin{bmatrix} I & \tilde{M}_1 \\ 0 & \tilde{M}_2 \end{bmatrix}^{-1} \begin{bmatrix} \tilde{N}_1 \\ \tilde{N}_2 \end{bmatrix} \tag{6.15}$$

for some $\tilde{N}_1, \tilde{M}_1 \in RH_\infty$.

Proof.

$1 \iff 2$: To prove that $1 \implies 2$, assume that $R_{\mathrm{ol}} \in RH_\infty$ stabilizes $G_{\mathrm{e}} = G_1 + R_{\mathrm{ol}} N_2 M_2^{-1}$. This implies that $G_{\mathrm{e}} M_2 = G_1 M_2 + R_{\mathrm{ol}} N_2$ or, equivalently, $G_1 M_2 = G_{\mathrm{e}} M_2 - R_{\mathrm{ol}} N_2 \in RH_\infty$. To prove that $1 \impliedby 2$, assume that $G_1 M_2 \in RH_\infty$ and pick $R_{\mathrm{ol}} = -G_1 M_2 Y_2$. Then

$$G_{\mathrm{e}} = G_1 - G_1 M_2 Y_2 N_2 M_2^{-1} = G_1 - G_1 M_2 (I - X_2 M_2) M_2^{-1} = G_1 M_2 X_2 \in RH_\infty.$$

$1 \iff 3$: If the factorization in (6.15) exist, then the choice $R_{\mathrm{ol}} = \tilde{M}_1$ is stabilizing because $G_{\mathrm{e}} = \tilde{N}_1$ then. If there is $R_{\mathrm{ol}} \in RH_\infty$ for which $G_{\mathrm{e}} \in RH_\infty$, then (6.15) holds with $\tilde{N}_1 = G_{\mathrm{e}}$ and $\tilde{M}_1 = R_{\mathrm{ol}}$. It is only left to show that the factorization in (6.15) is coprime. This follows by the relation

$$\begin{bmatrix} I & \tilde{M}_1 \\ 0 & \tilde{M}_2 \end{bmatrix} \begin{bmatrix} I & -\tilde{M}_1 \tilde{X}_2 - \tilde{N}_1 \tilde{Y}_2 \\ 0 & \tilde{X}_2 \end{bmatrix} + \begin{bmatrix} \tilde{N}_1 \\ \tilde{N}_2 \end{bmatrix} \begin{bmatrix} 0 & \tilde{Y}_2 \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}, \tag{6.16}$$

which is the related Bézout equality constructed from the Bézout coefficients of the left coprime $\tilde{M}_2$ and $\tilde{N}_2$.

The relation $2 \iff 3$ is then obviously true. □

The second condition of Lemma 6.10 can be viewed as a way to formalize the "containment" mentioned above. It is intuitive and easy to verify. For instance, in the example considered above we can always select $M_2(s) = \mathrm{diag}\{s/(s+1), (s^2+1)/(s+1)^2\}$, so that $G_1(s) M_2(s) = \mathrm{diag}\{1/(s+1), (s^2+1)/(s(s+1)^2)\}$ is unstable. However, this condition is less self-contained in characterizing all stabilizing R_{ol}'s (a *lcf* of G_2 would be required to that end). The last condition of Lemma 6.10 may appear less intuitive, but it leads to the following useful result.

Theorem 6.11. *If a lcf of G_1 and G_2 of form* (6.15) *exists, then $R_{ol} \in RH_\infty$ stabilizes the system in Fig. 6.5(c) iff there is $Q \in RH_\infty$ such that*

$$R_{ol} = \tilde{M}_1 + Q\tilde{M}_2 \tag{6.17}$$

and then $G_e = \tilde{N}_1 + Q\tilde{N}_2$ is the set of all attainable stable error systems $v \mapsto e$.

Proof. Obviously, this $R_{\rm ol} \in RH_\infty$ if so does Q. The first row of (6.15) reads $G_2 = \tilde{N}_1 - \tilde{M}_1\tilde{M}_2^{-1}\tilde{N}_2$. Hence, if $R_{\rm ol}$ is of form (6.17), then $G_{\rm e} = \tilde{N} + Q\tilde{N}_2 \in RH_\infty$, whence the sufficiency of (6.17) follows. To show necessity, let $G_{\rm e} = G_1 - R_{\rm ol,0}G_2 \in RH_\infty$ for some $R_{\rm ol,0} \in RH_\infty$. Define $Q_0 := (R_{\rm ol,0} - \tilde{M}_1)\tilde{M}_2^{-1}$. It is readily verified that it satisfies the equality

$$\begin{bmatrix} I & Q_0 \end{bmatrix}\begin{bmatrix} I & \tilde{M}_1 & \tilde{N}_1 \\ 0 & \tilde{M}_2 & \tilde{N}_2 \end{bmatrix} = \begin{bmatrix} I & R_{\rm ol,0} & G_{\rm e} \end{bmatrix}.$$

Post-multiplying this equality by the Bézout coefficients from (6.16) yields

$$\begin{bmatrix} I & Q_0 \end{bmatrix} = \begin{bmatrix} I & -\tilde{M}_1\tilde{X}_2 - \tilde{N}_1\tilde{Y}_2 + R_{\rm ol,0}\tilde{X}_2 + G_{\rm e}\tilde{Y}_2 \end{bmatrix} \in RH_\infty.$$

Hence, this $R_{\rm ol,0}$ is in form (6.17) for $Q = Q_0$. □

Example 6.3. To illustrate the results above, consider the problem of reconstructing a signal v from its version passed via a communication channel, whose transfer function $H(s) = (-s+1)/(s+1)$. The measurement equation can then be defined as $y = Hv$ and our goal is to design a stable reconstructor $R_{\rm ol} : y \mapsto \hat{v}$ rendering the reconstruction error $e = v - \hat{v}$ small. This requirement can be expressed in terms of the error system connecting v with e, i.e. $1 - R_{\rm ol}H$. Because H is not stably invertible, we cannot expect to have the zero error. But we may require it to be zero for the DC component of v. The latter reads as the condition $1 - R_{\rm ol}(0)H(0) = 0$ or, equivalently, as the requirement $R_{\rm ol}(0) = 1/H(0) = 1$. To reformulate this requirement as a stabilization problem, introduce the weight $W_v(s) = 1/s$ and assume that $v = W_v\tilde{v}$ for some fictitious $\tilde{v} \in L_2$. The error system $G_{\rm e} : \tilde{v} \mapsto e := v - \hat{v}$ becomes then

$$G_{\rm e} = (1 - R_{\rm ol}H)W_v$$

and the condition $G_{\rm e} \in RH_\infty$ is equivalent to $1 - R_{\rm ol}(0)H(0) = 0$. This error system is in form (6.14) for $G_1 = W_v$ and $G_2 = -HW_v$, for which

$$G(s) = \begin{bmatrix} G_1(s) \\ G_2(s) \end{bmatrix} = \begin{bmatrix} 1/s \\ (s-1)/(s(s+1)) \end{bmatrix}.$$

It is perhaps not hard to guess that a possible denominator of a *rcf* of this G_2 is $M(s) = s/(s+1)$ and then $G_1(s)M(s) = 1/(s+1)$. Hence, the second condition of Lemma 6.10 holds true and we may expect to be able to construct a *lcf* of G in form (6.15). With some educated guesses, its possible choice is

$$G(s) = \begin{bmatrix} 1 & 1 \\ 0 & s/(s+1) \end{bmatrix}^{-1}\begin{bmatrix} 2/(s+1) \\ (s-1)/(s+1)^2 \end{bmatrix},$$

which are indeed left coprime, as can be seen from the Bézout equality

$$\begin{bmatrix} 1 & 1 \\ 0 & s/(s+1) \end{bmatrix}\begin{bmatrix} 1 & -1 \\ 0 & (s+3)/(s+1) \end{bmatrix} + \begin{bmatrix} 2/(s+1) \\ (s-1)/(s+1)^2 \end{bmatrix}\begin{bmatrix} 0 & -1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Thus, the sets of all "stabilizing" reconstructors $R_{\rm ol}$ and corresponding error systems from Theorem 6.11 have

$$R_{\rm ol}(s) = 1 + Q(s)\frac{s}{s+1} \quad \text{and} \quad G_{\rm e}(s) = \frac{2}{s+1} + Q(s)\frac{s-1}{(s+1)^2}$$

for an arbitrary $Q \in RH_\infty$. All these $R_{\rm ol}$'s indeed satisfy the required condition $R_{\rm ol}(0) = 1$ and can thus be viewed as a parametrization of all stable transfer functions with the unit static gain. ◊

The parametrization of all stabilizing $R_{\rm ol}$'s in (6.17) relies on a *lcf* of form (6.15). We show below that this factorization can be readily constructed from a joint state-space realization of G_1 and G_2. To this end, bring in a realization

$$G(s) = \begin{bmatrix} G_1(s) \\ G_2(s) \end{bmatrix} = \left[\begin{array}{c|c} A & B \\ \hline C_1 & D_1 \\ C_2 & D_2 \end{array}\right] \tag{6.18}$$

The following result yields an algorithm for constructing the parametrization of Theorem 6.11.

Lemma 6.12. *If the realization of G in* (6.18) *is stabilizable, then G admits a left coprime factorization of form* (6.16) *iff* (C_2, A) *is detectable, in which case*

$$\begin{bmatrix} \tilde{M}_1(s) & \tilde{N}_1(s) \\ \tilde{M}_2(s) & \tilde{N}_2(s) \end{bmatrix} = \left[\begin{array}{c|cc} A + L_2C_2 & L_2 & B + L_2D_2 \\ \hline C_1 & 0 & D_1 \\ C_2 & I & D_2 \end{array}\right] \tag{6.19}$$

for any L_2 such that $A + L_2C_2$ is Hurwitz.

Proof. The sufficiency of the detectability of (C_2, A) follows by choosing $L = \begin{bmatrix} 0 & L_2 \end{bmatrix}$ in realization (4.21a) of $\tilde{M}$ in a general *lcf* $G = \tilde{M}^{-1}\tilde{N}$ (the stabilizability of (A, B) is required to construct corresponding Bézout coefficients).

To prove necessity, assume that G can be factorized as in (6.15). Bring in arbitrary realizations

$$\begin{bmatrix} \tilde{M}_1(s) \\ \tilde{M}_2(s) \end{bmatrix} = \left[\begin{array}{c|c} A_M & B_M \\ \hline C_{M1} & D_{M1} \\ C_{M2} & D_{M2} \end{array}\right] \quad \text{and} \quad \begin{bmatrix} \tilde{N}_1(s) \\ \tilde{N}_2(s) \end{bmatrix} = \left[\begin{array}{c|c} A_N & B_N \\ \hline C_{N1} & D_{N1} \\ C_{N2} & D_{N2} \end{array}\right],$$

such that A_M and A_N are Hurwitz. Because $\tilde{M}_2(s)$ is the denominator of a coprime factorization of $G_2(s)$, D_{M2} must be square and nonsingular and we can assume that $D_{M2} = I$ without loss of generality (otherwise, the scaling $\tilde{M}_2 \to D_{M2}^{-1}\tilde{M}_2$ and $\tilde{N}_2 \to D_{M2}^{-1}\tilde{N}_2$ does the trick). In this case equality (6.15) reads

$$\begin{aligned} G(s) &= \left[\begin{array}{c|cc} A_M & 0 & B_M \\ \hline C_{M1} & I & D_{M1} \\ C_{M2} & 0 & I \end{array}\right]^{-1} \left[\begin{array}{c|c} A_N & B_N \\ \hline C_{N1} & D_{N1} \\ C_{N2} & D_{N2} \end{array}\right] = \left[\begin{array}{c|cc} A_M - B_MC_{M2} & 0 & -B_M \\ \hline C_{M1} - D_{M1}C_{M2} & I & -D_{M1} \\ C_{M2} & 0 & I \end{array}\right] \left[\begin{array}{c|c} A_N & B_N \\ \hline C_{N1} & D_{N1} \\ C_{N2} & D_{N2} \end{array}\right] \\ &= \left[\begin{array}{cc|c} A_N & 0 & B_N \\ -B_MC_{N2} & A_M - B_MC_{M2} & -B_MD_{N2} \\ \hline C_{N1} - D_{M1}C_{N2} & C_{M1} - D_{M1}C_{M2} & D_{N1} + D_{M1}D_{N2} \\ C_{N2} & C_{M2} & D_{N2} \end{array}\right]. \end{aligned}$$

To prove the statement of the Lemma, it is now sufficient to show that the realization above is detectable from its second block output. To this end, bring in the corresponding PBH observability equality:

$$\begin{bmatrix} A_N - \lambda I & 0 \\ -B_MC_{N2} & A_M - B_MC_{M2} - \lambda I \\ C_{N2} & C_{M2} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} I & 0 & 0 \\ 0 & I & -B_M \\ 0 & 0 & I \end{bmatrix} \begin{bmatrix} A_N - \lambda I & 0 \\ 0 & A_M - \lambda I \\ C_{N2} & C_{M2} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0.$$

Because A_M and A_N are Hurwitz, this equality can hold for no $\lambda \in \overline{\mathbb{C}}_0$. □

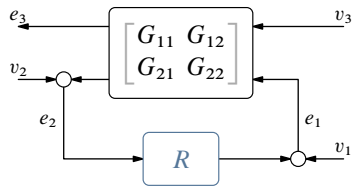


Fig. 6.6: General LFT internal stability setup

6.4 Internal stability in the LFT setting

This chapter is wound up with a study of the internal stability for the LFT setup depicted in Fig. 6.6. This is a generalization of the internal stability notion for the system in Fig. 6.1 and it shall play an important role in the performance analyses in the next chapter.

Following the logic of §6.1.1, we say that the system in in Fig. 6.6 is *internally stable* if all nine systems $(v_1, v_2, v_3) \mapsto (e_1, e_2, e_3)$ are stable. The relation between its inputs and outputs can be described by the following equation:

$$\begin{bmatrix} I & -R & 0 \\ -G_{22} & I & 0 \\ -G_{12} & 0 & I \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ e_3 \end{bmatrix} = \begin{bmatrix} I & 0 & 0 \\ 0 & I & G_{21} \\ 0 & 0 & G_{11} \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}.$$

Hence, its internal stability is equivalent to the stability of

$$\begin{aligned} T_{\text{aux}} &:= \begin{bmatrix} I & -R & 0 \\ -G_{22} & I & 0 \\ -G_{12} & 0 & I \end{bmatrix}^{-1} \begin{bmatrix} I & 0 & 0 \\ 0 & I & G_{21} \\ 0 & 0 & G_{11} \end{bmatrix} = \left(\begin{bmatrix} I & 0 & 0 \\ -G_{22} & I & 0 \\ -G_{12} & 0 & I \end{bmatrix} - \begin{bmatrix} 0 & R & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \right)^{-1} \begin{bmatrix} I & 0 & 0 \\ 0 & I & G_{21} \\ 0 & 0 & G_{11} \end{bmatrix} \\ &= \begin{bmatrix} I & 0 & 0 \\ G_{22} & I & G_{21} \\ G_{12} & 0 & G_{11} \end{bmatrix} + \begin{bmatrix} I \\ G_{22} \\ G_{12} \end{bmatrix} R(I - G_{22}R)^{-1} \begin{bmatrix} G_{22} & I & G_{21} \end{bmatrix}, \end{aligned} \tag{6.20}$$

where the Matrix Inversion Lemma (Lemma B.7) is used to derive the last equality.

The analysis of (6.20) also follows the logic of §6.2.2. We are seeking for bi-stable transformations decoupling the R-independent terms in the second term of (6.20). However, there is a qualitative difference between (6.2) and (6.20). The R-independent terms of the former, see (6.7), are always stable. This is not necessarily true for (6.20). Hence, the stabilization result below includes restrictive conditions on coprime factorizations of the composed system G.

Theorem 6.13. *There is an internally stabilizing R iff there are coprime factorizations of the form*

$$\begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix} = \begin{bmatrix} I & \tilde{M}_{12} \\ 0 & \tilde{M}_{22} \end{bmatrix}^{-1} \begin{bmatrix} \tilde{N}_{11} & \tilde{N}_{12} \\ \tilde{N}_{21} & \tilde{N}_{22} \end{bmatrix} = \begin{bmatrix} N_{11} & N_{12} \\ N_{21} & N_{22} \end{bmatrix} \begin{bmatrix} I & 0 \\ M_{21} & M_{22} \end{bmatrix}^{-1} \tag{6.21}$$

with right coprime $\tilde{N}_{22}$ and $\tilde{M}_{22}$ and left coprime N_{22} and M_{22}. If factorizations as above exist, then R internally stabilizes the system in Fig. 6.6 iff it internally stabilizes the system in Fig. 6.1 under $P = G_{22}$.

Proof. The first step is to prove the necessity of (6.21). To this end, bring in a doubly coprime factorization of G_{22}, i.e. transfer functions $N_{22}, M_{22}, \tilde{N}_{22}, \tilde{M}_{22} \in RH_\infty$ such that $M_{22}(s)$ and $\tilde{M}_{22}(s)$ are bi-proper,

$$G_{22} = N_{22}M_{22}^{-1} = \tilde{M}_{22}^{-1}\tilde{N}_{22},$$

and there are appropriately dimensioned $X_{22}, Y_{22}, \tilde{X}_{22}, \tilde{Y}_{22} \in RH_\infty$ such that

$$\begin{bmatrix} X_{22} & Y_{22} \\ -\tilde{N}_{22} & \tilde{M}_{22} \end{bmatrix} \begin{bmatrix} M_{22} & -\tilde{Y}_{22} \\ N_{22} & \tilde{X}_{22} \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}.$$

From the construction in §4.3.1 we know that these function always exist. Now,

$$T_{\text{aux}}\begin{bmatrix} M_{22} & \tilde{Y}_{22} & 0 \\ -N_{22} & \tilde{X}_{22} & 0 \\ 0 & 0 & I \end{bmatrix} = \begin{bmatrix} M_{22} & \tilde{Y}_{22} & 0 \\ 0 & \tilde{M}_{22}^{-1} & G_{21} \\ G_{12}M_{22} & G_{12}\tilde{Y}_{22} & G_{11} \end{bmatrix} + \begin{bmatrix} I \\ G_{22} \\ G_{12} \end{bmatrix} R(I - G_{22}R)^{-1}\begin{bmatrix} 0 & \tilde{M}_{22}^{-1} & G_{21} \end{bmatrix}$$

is stable iff T_{aux} is stable. Because the first column of the system above does not depend on R, there is a stabilizing R *only if* $G_{12}M_{22} \in RH_\infty$. By Lemma 6.10, there are then $\tilde{N}_{12}, \tilde{M}_{12} \in RH_\infty$ such that

$$\begin{bmatrix} G_{12} \\ G_{22} \end{bmatrix} = \begin{bmatrix} G_{12}M_{22} \\ N_{22} \end{bmatrix} M_{22}^{-1} = \begin{bmatrix} I & \tilde{M}_{12} \\ 0 & \tilde{M}_{22} \end{bmatrix}^{-1} \begin{bmatrix} \tilde{N}_{12} \\ \tilde{N}_{22} \end{bmatrix},$$

and the later is a *lcf*. Moreover, taking into account (6.16) we can construct Bézout coefficients for the corresponding doubly coprime factorization as follows:

$$\left[\begin{array}{c:cc} X_{22} & 0 & Y_{22} \\ \hdashline -\tilde{N}_{12} & I & \tilde{M}_{12} \\ -\tilde{N}_{22} & 0 & \tilde{M}_{22} \end{array}\right] \left[\begin{array}{c:cc} M_{22} & 0 & -\tilde{Y}_{22} \\ \hdashline G_{12}M_{22} & I & -\tilde{M}_{12}\tilde{X}_{22} - \tilde{N}_{12}\tilde{Y}_{22} \\ N_{22} & 0 & \tilde{X}_{22} \end{array}\right] = \left[\begin{array}{c:cc} I & 0 & 0 \\ \hdashline 0 & I & 0 \\ 0 & 0 & I \end{array}\right].$$

Next, consider

$$\begin{bmatrix} X_{22} & Y_{22} & 0 \\ -\tilde{N}_{12} & \tilde{M}_{12} & I \\ -\tilde{N}_{22} & \tilde{M}_{22} & 0 \end{bmatrix} T_{\text{aux}} = \begin{bmatrix} M_{22}^{-1} & Y_{22} & Y_{22}G_{21} \\ 0 & \tilde{M}_{12} & G_{11} + \tilde{M}_{12}G_{21} \\ 0 & \tilde{M}_{22} & \tilde{M}_{22}G_{21} \end{bmatrix} + \begin{bmatrix} M_{22}^{-1} \\ 0 \\ 0 \end{bmatrix} R(I - G_{22}R)^{-1}\begin{bmatrix} G_{22} & I & G_{21} \end{bmatrix},$$

which is again stable iff T_{aux} is stable. Because the last two rows of this system are independent of R,

$$\begin{bmatrix} G_{11} + \tilde{M}_{12}G_{21} \\ \tilde{M}_{22}G_{21} \end{bmatrix} = \begin{bmatrix} I & \tilde{M}_{12} \\ 0 & \tilde{M}_{22} \end{bmatrix} \begin{bmatrix} G_{11} \\ G_{21} \end{bmatrix}$$

must be stable for a stabilizing R to exist. Combining two necessary conditions above, we have the necessity of the first (left) factorization in (6.21), where $\tilde{N}_{11} = G_{11} + \tilde{M}_{12}G_{21}$ and $\tilde{N}_{21} = \tilde{M}_{22}G_{21}$. The necessity of the right factorization in (6.21) follows by dual arguments. To complete the proof of necessity, we only need now to show that the factorizations in (6.21) are coprime. This follows by the explicit construction of the Bézout equalities,

$$\left[\begin{array}{cc:cc} I & 0 & 0 & 0 \\ X_{21} & X_{22} & 0 & Y_{22} \\ \hdashline -\tilde{N}_{11} & -\tilde{N}_{12} & I & \tilde{M}_{12} \\ -\tilde{N}_{21} & -\tilde{N}_{22} & 0 & \tilde{M}_{22} \end{array}\right] \left[\begin{array}{cc:cc} I & 0 & 0 & 0 \\ M_{21} & M_{22} & 0 & -\tilde{Y}_{22} \\ \hdashline N_{11} & N_{12} & I & \tilde{X}_{12} \\ N_{21} & N_{22} & 0 & \tilde{X}_{22} \end{array}\right] = \left[\begin{array}{cc:cc} I & 0 & 0 & 0 \\ 0 & I & 0 & 0 \\ \hdashline 0 & 0 & I & 0 \\ 0 & 0 & 0 & I \end{array}\right], \tag{6.22}$$

where $\tilde{N}_{11} = G_{11} + \tilde{M}_{12}G_{21}$, $\tilde{N}_{12} = G_{12}M_{22}$, $\tilde{N}_{21} = \tilde{M}_{22}G_{21}$, and $\tilde{X}_{12} = -\tilde{M}_{12}\tilde{X}_{22} - G_{12}M_{22}\tilde{Y}_{22}$ are the factors associated with the *lcf* in (6.21) and $N_{11} = G_{11} + G_{12}M_{21}$, $N_{21} = \tilde{M}_{22}G_{21}$, $N_{12} = G_{12}M_{22}$, and $X_{21} = -X_{22}M_{21} - Y_{22}\tilde{M}_{22}G_{12}$ are those associated with the *rcf* there (the construction of $\tilde{M}_{12}$ and M_{21} is less explicit, see the proof of Lemma 6.10).

To prove sufficiency of (6.21) and properties of R, define

$$\tilde{T}_{\text{aux}} := \begin{bmatrix} X_{22} & 0 & Y_{22} \\ -\tilde{N}_{12} & I & \tilde{M}_{12} \\ -\tilde{N}_{22} & 0 & \tilde{M}_{22} \end{bmatrix} \begin{bmatrix} I & 0 & 0 \\ 0 & 0 & I \\ 0 & I & 0 \end{bmatrix} T_{\text{aux}} \begin{bmatrix} 0 & I & 0 \\ 0 & 0 & -I \\ I & 0 & 0 \end{bmatrix} \begin{bmatrix} I & 0 & 0 \\ M_{21} & M_{22} & -\tilde{Y}_{22} \\ N_{21} & N_{22} & \tilde{X}_{22} \end{bmatrix},$$

which is again stable iff T_{aux} is stable, because both external factors in it are bi-stable (follows from (6.22)). By tedious but otherwise straightforward algebra it can be shown that

$$\tilde{T}_{\text{aux}} = \begin{bmatrix} -X_{21} & I & 0 \\ \tilde{N}_{11} & 0 & 0 \\ \tilde{N}_{21} & 0 & 0 \end{bmatrix} - \begin{bmatrix} Y_{22} \\ \tilde{M}_{12} \\ \tilde{M}_{22} \end{bmatrix} \begin{bmatrix} N_{21} & N_{22} & \tilde{X}_{22} \end{bmatrix} - \begin{bmatrix} I \\ 0 \\ 0 \end{bmatrix} Q \begin{bmatrix} 0 & 0 & I \end{bmatrix},$$

where $Q := M_{22}^{-1}\tilde{Y}_{22} + M_{22}^{-1}R(I - G_{22}R)^{-1}\tilde{M}_{22}^{-1}$. Hence, the stability of $\tilde{T}_{\text{aux}}$ is equivalent to that of Q, which is exactly what we had in the proof of Theorem 6.6, cf. (6.7). But the stabilization of Q depends only on G_{22} and is always possible. This proves the sufficiency of (6.21) and the last statement. □

Despite looking rather technical, the stabilizability condition of Theorem 6.13 are intuitive. Indeed, the equalities in (6.21) read

$$\begin{aligned} \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix} &= \begin{bmatrix} \tilde{N}_{11} - \tilde{M}_{12}\tilde{M}_{22}^{-1}\tilde{N}_{21} & \tilde{N}_{12} - \tilde{M}_{12}\tilde{M}_{22}^{-1}\tilde{N}_{22} \\ \tilde{M}_{22}^{-1}\tilde{N}_{21} & \tilde{M}_{22}^{-1}\tilde{N}_{22} \end{bmatrix} = \begin{bmatrix} \tilde{N}_{11} & \tilde{N}_{12} \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} \tilde{M}_{12} \\ -I \end{bmatrix} \tilde{M}_{22}^{-1} \begin{bmatrix} \tilde{N}_{21} & \tilde{N}_{22} \end{bmatrix} \\ &= \begin{bmatrix} N_{11} - N_{12}M_{22}^{-1}M_{21} & N_{12}M_{22}^{-1} \\ N_{21} - N_{22}M_{22}^{-1}M_{21} & N_{22}M_{22}^{-1} \end{bmatrix} = \begin{bmatrix} N_{11} & 0 \\ N_{21} & 0 \end{bmatrix} - \begin{bmatrix} N_{12} \\ N_{22} \end{bmatrix} M_{22}^{-1} \begin{bmatrix} M_{21} & -I \end{bmatrix}, \end{aligned}$$

i.e. all unstable poles of the composed system are those of either $\tilde{M}_{22}(s)$ and $M_{22}(s)$, which are the denominators of left and right coprime factors of G_{22}, respectively. But we know, see Proposition 3.3, that the inverses of the denominators of coprime factorizations contain all unstable poles of the system itself. Hence, conditions of Theorem 6.13 effectively say that *all unstable modes of G must be present in G_{22}*, around which the feedback loop is closed. Consequently, in the open-loop case studied in Section 6.3, in which $G_{22} = 0$, internal stabilization is possible only if all remaining G_{ij} are stable themselves.

Substituting $R(I - G_{22}R)^{-1} = -\tilde{Y}_{22}\tilde{M}_{22} + M_{22}Q\tilde{M}_{22}$ into (6.20), all stable systems $T_{33} : v_3 \mapsto e_3$ can be characterized as

$$\begin{aligned} T_{33} &= G_{11} - G_{12}\tilde{Y}_{22}\tilde{M}_{22}G_{21} + G_{12}M_{22}Q\tilde{M}_{22}G_{21} = \tilde{N}_{11} - (\tilde{M}_{12} + N_{12}Y_{22})\tilde{M}_{22}^{-1}\tilde{N}_{21} + N_{12}Q\tilde{N}_{21} \\ &= \tilde{N}_{11} + \tilde{X}_{12}\tilde{N}_{21} + N_{12}Q\tilde{N}_{21}, \end{aligned}$$

where the equalities $M_{22}^{-1}\tilde{Y}_{22} = Y_{22}\tilde{M}_{22}^{-1}$ and $\tilde{X}_{12}\tilde{M}_{22} = -(\tilde{M}_{12} + N_{12}Y_{22})$, which follow from (6.22), are used. This is again an affine function of Q.

Arguably, the simplest way to verify the condition of Theorem 6.13 for a general G is via its state-space realization. Specifically, bring in a *minimal* realization of the composite system,

$$\begin{bmatrix} G_{11}(s) & G_{12}(s) \\ G_{21}(s) & G_{22}(s) \end{bmatrix} = \left[\begin{array}{c|cc} A & B_1 & B_2 \\ \hline C_1 & D_{11} & D_{12} \\ C_2 & D_{21} & D_{22} \end{array}\right].$$

The following result offers simple ways to verify the conditions of Theorem 6.13 and construct all stabilizing controllers for the system in Fig. 6.6.

Proposition 6.14. *The conditions of Theorem 6.13 hold iff (A, B_2) is stabilizable and (C_2, A) is detectable. If these conditions hold, then all proper stabilizing controllers for the system in Fig. 6.6 are given by*

$$R(s) = \mathcal{F}_l\left(\left[\begin{array}{c|cc} A + B_2K + LC_2 + LD_{22}K & -L & B_2 + LD_{22} \\ \hline K & 0 & I \\ -C_2 - D_{22}K & I & -D_{22} \end{array}\right], Q(s)\right), \tag{6.23}$$

where K and L are any matrices such that $A + B_2K$ and $A + LC_2$ are Hurwitz and $Q \in RH_\infty$ and is such that $\det(I - Q(\infty)D_{22}) \neq 0$*, but otherwise arbitrary.*

Proof. The stabilizability conditions can be proved in line with the proof of Lemma 6.12. Formula (6.23) is merely a notational adjustment to (6.9), because R only needs to stabilize G_{22}. □

Chapter 7

Performance and the Standard Problem

Optimization-based approaches to control system design comprise essentially two basic stages. In the first stage, a control problem is formulated as a problem of minimizing a norm ("size") of some specially built closed-loop system. Then, in the second stage, the controller is produced by solving this optimization problem subject to a suitably defined constraint (like stability). It is usually convenient to formulate the optimization problem in the first stage in a unified fashion, so that it can be solved using a general purpose machinery. Such a unified optimization setup, known as *the standard problem*, is the focus point of this chapter. Specifically, we shall see how to cast some (more or less) standard approaches to the optimization-based controller design as standard problems and discuss some aspects of the related *generalized plant* paradigm.

7.1 The setup, main definitions, and stability

The standard problem corresponds to the lower LFT configuration depicted in Fig. 7.1 on the next page. The system G there is known as the generalized plant and R is referred to as the controller. Our goal is then to design R to reduce the effect of w on z in the closed-loop system $\mathcal{F}_{\rfloor}(G, R) : w \mapsto z$. The generalized plant has two inputs, w and u, and two outputs, z and y, whose meaning is spelt out below.

- $\boldsymbol{w}$ is dubbed the *exogenous input* and contains all exogenous signals, whose effect is important for the problem at hand. These may be a reference signal, a load disturbance, measurement noise, et cetera. These signals might also be fictitious (normalized) signals forming the actual signals of interest.
- $\boldsymbol{z}$ is dubbed the *regulated output* and contains signals that are required to be kept "small" (in whatever sense). These may be deviations from a required behavior, like tracking or estimation errors, actuator signals and suchlike, weighted to focus their relative importance and important aspects of them.
- $\boldsymbol{y}$ is the *measured output*, through which the controller acquires the information about the effect of w on the system behavior.
- $\boldsymbol{u}$ is the *control input*, which is the signal generated by the controller and through which the effect of w on z can be affected.

Control and estimation problems considered throughout this chapter shall clarify these definitions.

The generalized plant G contains dynamics of a controlled plant itself, sensors, actuators, weighing functions (see below), and even some fixed parts of the controller (e.g. the integral action). It is conventionally presented by the joint dynamics of its components, like

$$G(s) = \begin{bmatrix} G_{zw}(s) & G_{zu}(s) \\ G_{yw}(s) & G_{yu}(s) \end{bmatrix} = \left[\begin{array}{c|cc} A & B_w & B_u \\ \hline C_z & D_{zw} & D_{zu} \\ C_y & D_{yw} & D_{yu} \end{array}\right] \tag{7.1}$$

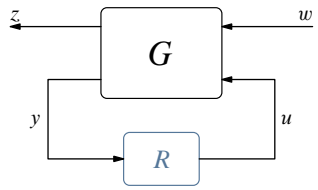


Fig. 7.1: The "standard problem"

for input and output partitions compatible with those in Fig. 7.1.

Because stability is normally a vital property of control systems, in most cases we require the controller to internally stabilize the system in Fig. 7.1 in the sense discussed in Section 6.4. Consequently, we assume that the realization (A, B_u, C_y, D_{yu}) is stabilizable and detectable or, equivalently, that G admits a doubly coprime factorization over RH_∞ of the form

$$G = \begin{bmatrix} N_{zw} & N_{zu} \\ N_{yw} & N_{yu} \end{bmatrix} \begin{bmatrix} I & 0 \\ M_{yw} & M_{yu} \end{bmatrix}^{-1} = \begin{bmatrix} I & \tilde{M}_{zu} \\ 0 & \tilde{M}_{yu} \end{bmatrix}^{-1} \begin{bmatrix} \tilde{N}_{zw} & \tilde{N}_{zu} \\ \tilde{N}_{yw} & \tilde{N}_{yu} \end{bmatrix}, \tag{7.2}$$

with the corresponding Bézout coefficients

$$\left[\begin{array}{cc:cc} I & 0 & 0 & 0 \\ X_{yw} & X_{yu} & 0 & Y_{yu} \\ \hdashline -\tilde{N}_{zw} & -\tilde{N}_{zu} & I & \tilde{M}_{zu} \\ -\tilde{N}_{yw} & -\tilde{N}_{yu} & 0 & \tilde{M}_{yu} \end{array}\right] \left[\begin{array}{cc:cc} I & 0 & 0 & 0 \\ M_{yw} & M_{yu} & 0 & -\tilde{Y}_{yu} \\ \hdashline N_{zw} & N_{zu} & I & \tilde{X}_{zu} \\ N_{yw} & N_{yu} & 0 & \tilde{X}_{yu} \end{array}\right] = \left[\begin{array}{cc:cc} I & 0 & 0 & 0 \\ 0 & I & 0 & 0 \\ \hdashline 0 & 0 & I & 0 \\ 0 & 0 & 0 & I \end{array}\right], \tag{7.3}$$

cf. (6.21) and (6.22). The class of all stabilizing controllers is

$$R = (-\tilde{Y}_{yu} + M_{yu}Q)(\tilde{X}_{yu} + N_{yu}Q)^{-1} = (X_{yu} + Q\tilde{N}_{yu})^{-1}(-Y_{yu} + Q\tilde{M}_{yu}) = \mathcal{F}_l(J, Q), \tag{7.4}$$

where

$$J(s) = \left[\begin{array}{c|cc} A + B_uK_u + L_yC_y + L_yD_{yu}K_u & -L_y & B_u + L_yD_{yu} \\ \hline K_u & 0 & I \\ -C_y - D_{yu}K_u & I & -D_{yu} \end{array}\right] \tag{7.5}$$

and Q is any stable system such that $\det(I - Q(\infty)D_{yu}) \neq 0$. With this class of admissible controllers, the set of all closed-loop maps $T_{zw} : w \mapsto z$ can be characterized as

$$T_{zw} = \tilde{N}_{zw} + \tilde{X}_{zu}\tilde{N}_{yw} + N_{zu}Q\tilde{N}_{yw}. \tag{7.6}$$

An informative state-space realization of the systems on the right-hand side of (7.6) is

$$\begin{bmatrix} \tilde{N}_{zw}(s) + \tilde{X}_{zu}(s)\tilde{N}_{yw}(s) & N_{zu}(s) \\ \tilde{N}_{yw}(s) & 0 \end{bmatrix} = \left[\begin{array}{cc|cc} A + B_uK_u & -L_yC_y & -L_yD_{yw} & B_u \\ 0 & A + L_yC_y & B_w + L_yD_{yw} & 0 \\ \hline C_z + D_{zu}K_u & C_z & D_{zw} & D_{zu} \\ 0 & C_y & D_{yw} & 0 \end{array}\right], \tag{7.7}$$

which actually equals $G(s) \star J(s)$ and as such can be derived by the formulae of Proposition 5.7 (and applying the similarity transformation $\left[\begin{smallmatrix} 0 & I \\ I & -I \end{smallmatrix}\right]$). The $(2, 2)$ element of the realization above contains only uncontrollable (those of $A + L_yC_y$) and unobservable (those of $A + B_uK_u$) modes and is thus indeed zero.

7.2 Some (familiar) H_2 problems

The H_2 version of the standard problem in Fig. 7.1 consists in designing a stabilizing R that minimizes the H_2-norm of the system $T_{zw} : w \mapsto z$. This section aims at showing how a couple of classical control and filtering problems (see [15, 1] for their explication) can be cast as standard H_2 problems.

7.2.1 LQR

The linear-quadratic regulator problem, aka LQR, studies systems of the form

$$\dot{x}(t) = Ax(t) + Bu(t), \quad x(0) = x_0$$

and aims at minimizing the cost function

$$\mathcal{J}_{\text{LQR}} := \int_{\mathbb{R}_+} \big(x'(t)Qx(t) + u'(t)Ru(t)\big)\mathrm{d}t \tag{7.8}$$

for some $Q = Q' \geq 0$ and $R = R' > 0$. Although the problem itself does not assume any particular measurement equation, it is well known that the optimal control law can be written in the form of a static state feedback. Thus, we shall assume that the whole state vector $x(t)$ is perfectly measurable.

To cast this problem in the form presented in Fig. 7.1, we shall define its inputs and outputs first. To this end, note that $\mathcal{J}_{\text{LQR}}$ is the L_2-norm of the signal

$$z = \begin{bmatrix} Q^{1/2}x \\ R^{1/2}u \end{bmatrix}.$$

This is a natural candidate for the regulated output. To determine w, note that the only influence on the system not related to our actions is the initial condition. But the effect of the initial condition x_0 is equivalent to the effect of the Dirac delta impulse applied via the "B" matrix x_0. In other words, the state dynamics can be equivalently described as

$$\dot{x}(t) = Ax(t) + x_0\delta(t) + Bu(t), \quad x(0) = 0.$$

This implies that the LQR problem boils down to minimizing the L_2-norm of z under an impulse input. But that is exactly the deterministic interpretation of the H_2-norm of the controlled system discussed on p. 50. Taking into account that the measured output is x and the control input is u, we end up with the generalized plant

$$G(s) = G_{\text{LQR}}(s) := \left[\begin{array}{c|c:c} A & x_0 & B \\ \hline Q^{1/2} & 0 & 0 \\ 0 & 0 & R^{1/2} \\ \hdashline I & 0 & 0 \end{array}\right]. \tag{7.9}$$

The solution to the LQR problem is currently well understood, see [15, 1]. It is based on one algebraic Riccati equation, has a static state-feedback structure, and possesses several attractive properties, like the infinite gain margin and a phase margin of at least 60°. Another known property is that the solution of the corresponding H_2 optimization problem does not depend on x_0. Hence, it can be replaced with any B_w, including the case of $B_w = I$.

Yet another generalization may be introduced by considering the generalized plant

$$G(s) = G_{\text{LQR}'}(s) := \left[\begin{array}{c|c:c} A & B_w & B_u \\ \hline C_z & 0 & D_{zu} \\ \hdashline I & 0 & 0 \end{array}\right], \tag{7.9$'$}$$

which corresponds to the cost function

$$\mathcal{J}_{\text{LQR}'} = \int_{\mathbb{R}_+} \begin{bmatrix} x'(t) & u'(t) \end{bmatrix} \begin{bmatrix} C_z'C_z & C_z'D_{zu} \\ D_{zu}'C_z & D_{zu}'D_{zu} \end{bmatrix} \begin{bmatrix} x(t) \\ u(t) \end{bmatrix} \mathrm{d}t. \tag{7.8$'$}$$

The cost in (7.8) corresponds to the particular case of $C_z'D_{zu} = 0$. A nonzero $C_z'D_{zu}$, which penalizes a cross-coupling between x and u, might be handy in handling cost functions including a physically meaningful variable $y = C_yx$ and its derivatives. Indeed, $\dot{y}(t) = C_yAx(t) + C_yBu(t)$ and, unless $B'C_y'C_yA = 0$, there is a cross-coupling in $\dot{y}'\dot{y}$.

7.2.2 Steady-state Kalman–Bucy filtering

Let

$$\begin{cases} \dot{x}(t) = Ax(t) + n_x(t) \\ y(t) = Cx(t) + n_y(t) \end{cases} \tag{7.10}$$

where n_x and n_y are independent zero-mean white Gaussian processes having covariances

$$\mathrm{E}[n_x(t)n_x'(s)] = Q_x\delta(t-s) \quad \text{and} \quad \mathrm{E}[n_y(t)n_y'(s)] = Q_y\delta(t-s)$$

for some $Q_x = Q_x' \geq 0$ and $Q_y = Q_y' > 0$, where $\mathrm{E}[v]$ denotes the expected value of a random variable v. A stationary version of the celebrated Kalman–Bucy filtering problem can be posed as the task of generating a reconstruction (estimate) $\hat{x}(t)$ of the state vector $x(t)$ using measurements $y(t)$ that minimize

$$\mathcal{J}_{\mathrm{KBF}} := \lim_{t\to\infty} \mathrm{E}\big[\|x(t) - \hat{x}(t)\|^2\big],$$

which is the steady-state variance of the reconstruction error $x - \hat{x}$. It is assumed that the reconstructor $y \mapsto \hat{x}$, known as *filter*, is a stable and causal LTI system. Stability is required because filtering is an open-loop operation. The time invariance of the filter, which result in the time invariance of the whole system, renders the problem meaningful. Otherwise nothing would prevent a filter to start acting at an arbitrarily late time instance, which would not affect $\mathcal{J}_{\mathrm{KBF}}$. A more conventional Kalman–Bucy filtering formulation aims at minimizing the error variance at every time instance and in general results in a time-varying filter. However, if the process in (7.10) is time invariant, parameters of such time-varying filters tend to converge to their steady-state values rapidly. Hence, the steady-state version may not be unduly limiting in such situations performance-wise, while is clearly advantageous from the implementation viewpoint.

To formulate the standard problem, note that the system of interest from the performance viewpoint is that connecting the exogenous signals n_x and n_y with the reconstruction error $z = x - \hat{x}$. It is known [15, §1.11.3] that the asymptotic variance of the response of an LTI system $G : u \mapsto y$ to a zero-mean white input with $\mathrm{E}[u(t)u'(s)] = Q\delta(t-s)$ equals $\lim_{t\to\infty} \mathrm{E}[\|y(t)\|^2] = \|GQ^{1/2}\|_2^2$. This implies that the H_2-norm is the right performance measure here and that the input channels of the system of interest should be scaled by $\mathrm{diag}\{Q_x^{1/2}, Q_y^{1/2}\}$. This yields the generalized plant

$$G(s) = G_{\mathrm{KBF}}(s) := \left[\begin{array}{c|cc:c} A & Q_x^{1/2} & 0 & 0 \\ \hline I & 0 & 0 & -I \\ \hdashline C & 0 & Q_y^{1/2} & 0 \end{array}\right], \tag{7.11}$$

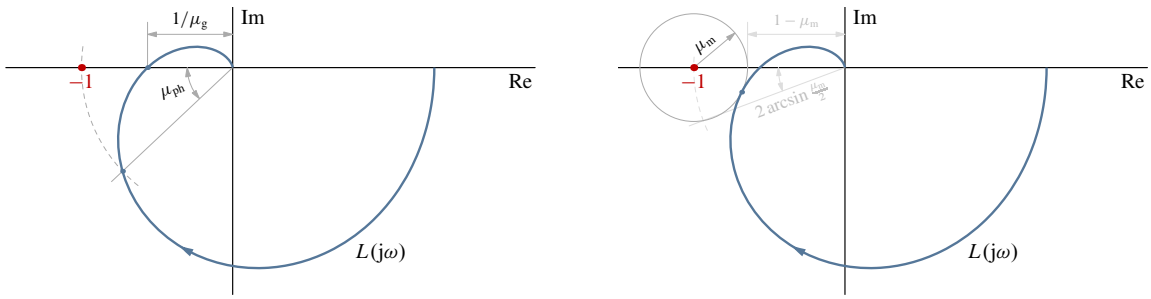
whose measured output is y and whose control input is the filter output $\hat{x}$. The exogenous input w in this case is any zero-mean unit-intensity, i.e. normalized, white process such that

$$\begin{bmatrix} n_x(t) \\ n_y(t) \end{bmatrix} = \begin{bmatrix} Q_x^{1/2} & 0 \\ 0 & Q_y^{1/2} \end{bmatrix} w(t).$$

The Kalman–Bucy filtering problem is then to design a stabilizing R minimizing the H_2-norm of the system $w \mapsto z$, which is a special case of the standard H_2 problem. Because the problem is open loop (as $G_{yu} = 0$), the internal stability requires A to be Hurwitz and constrains the filter R to be stable itself. We can also accommodate unstable processes by dropping the internal stability requirement and requiring R to be stable explicitly. This situation can be handles with the help the machinery in Section 6.3.

The problem can be faintly generalized by considering the generalized plant

$$G(s) = G_{\mathrm{KBF}'}(s) := \left[\begin{array}{c|cc} A & B_w & 0 \\ \hline C_z & 0 & -I \\ \hdashline C_y & D_{yw} & 0 \end{array}\right]. \tag{7.11$'$}$$



(a) Gain (μ_g) and phase (μ_{ph}) margins

(b) Modulus margin (μ_m)

Fig. 7.2: Stability margins in the Nyquist plane

It corresponds to the assumption that the exogenous signals in (7.10) have covariances

$$\mathrm{E}\left[\begin{bmatrix} n_x(t) \\ n_y(t) \end{bmatrix}\begin{bmatrix} n_x'(s) & n_y'(s) \end{bmatrix}\right] = \begin{bmatrix} B_w B_w' & B_w D_{yw}' \\ D_{yw} B_w' & D_{yw} D_{yw}' \end{bmatrix}\delta(t-s),$$

i.e. it allows mutual dependencies between n_x and n_y if $D_{yw}B_w' \neq 0$. Also, the regulated signal in (7.11′) aims at estimating a subset of the state vector, $v = C_z x$. Yet the latter generalization actually changes nothing, as the optimal solution is known to be of the form $\hat{v} = C_v \hat{x}$, where $\hat{x}$ is the optimal estimate of the whole stave vector x. Further generalizations include relaxations of the causality requirement. If R is assumed to have access to a finite preview of y, the problem is dubbed the *fixed-lag smoothing*. If the whole future behavior of y is available, we have a *fixed-interval smoothing* problem.

7.3 Some (less familiar) H_∞ problems

The H_∞ version of the standard problem is more recent[1] and less conspicuous in introductory control texts. Nevertheless, there is a number of H_∞ problems that are directly connected to classical frequency-domain control ideas and thus should be easier to grasp based on loop-shaping insight. Thus, this section not only introduces some of these problems and shows how to cast them in the form of Fig. 7.1, but also discusses their potential in revealing intrinsic limitations of feedback control. Another emphasis in this section is placed on the need to be-careful-what-you-wish-for in posing control design problems. Optimization, and the H_∞ approach is not an exception, might find unexpected ways to exploit loopholes even in seemingly well-chosen cost functions to produce poor to meaningless optimal controllers. This should be always taken this into consideration in formulating optimization-based problems.

7.3.1 Maximum attainable modulus margin

Stability margins play a prominent role in classical frequency-domain design methods. They quantify the proximity of the loop frequency-response plot to the critical point $(-1, 0)$ on the Nyquist plane. The "far from the critical point" requirement is essential both to avoid resonances in closed-loop frequency responses (and thus have smoother transients) and to reduce the sensitivity of the controlled system to modeling mismatches.

Historically, the best known margins are the gain and phase margins, μ_g and μ_{ph}, shown in Fig. 7.2(a), which are relatively simple to calculate also from the Bode plot. Another margin, which is less common but not less informative, is the *modulus margin*, μ_m. It is defined as the shortest (Euclidean) distance from

[1] Developments of the H_∞ control theory started in the early '80s, with the appeal of the seminal work of Zames [33].

the critical point to the Nyquist plot, see Fig. 7.2(b). In other words, $\mu_{\rm m} = \inf_{\omega\in\mathbb{R}}|1 + L(\mathrm{j}\omega)|$, so that $1/\mu_{\rm m} = \sup_{\omega\in\mathbb{R}}|S(\mathrm{j}\omega)|$, where $S := 1/(1 + L)$ is the sensitivity function, which is obviously assumed to be stable. Hence, we have that

$$\mu_{\rm m} = 1/\|S\|_\infty$$

and minimizing the H_∞-norm of the sensitivity function is effectively the problem of maximizing the modulus margin.

Remark 7.1 ($\mu_{\rm g}$ and $\mu_{\rm ph}$ warranties via $\mu_{\rm m}$). It is a matter of simple plane geometry to show that if $\mu_{\rm m} \le 1$, then

$$\mu_{\rm g} \ge \frac{1}{1-\mu_{\rm m}} \quad \text{and} \quad \mu_{\rm ph} \ge 2\arcsin\frac{\mu_{\rm m}}{2}.$$

We shall see below that the condition $\mu_{\rm m} \le 1$ holds in any meaningful problem. In the particular case of $\mu_{\rm m} = 1$, we have the familiar [1, Sec. 5.4] stability margins of the LQR controller. ∇

To form the standard problem for minimizing $\|S\|_\infty$, consider the unity-feedback system in Fig. 1.4(c). The sensitivity function corresponds to the system from $y_{\rm r} \mapsto e$ there. Setting the exogenous signal as $y_{\rm r}$, the regulated signal—as e, the measured output—also as e and u in its standard role, we end up with

$$G(s) = G_{\rm MM}(s) = \begin{bmatrix} I & -P(s) \\ I & -P(s) \end{bmatrix}. \tag{7.12}$$

Bring in a doubly coprime factorization of $P = NM^{-1} = \tilde{M}^{-1}\tilde{N}$. Because $G_{yu} = -P$ here, we have that $N_{yu} = -N$, $\tilde{N}_{yu} = -\tilde{N}$, and the corresponding Bézout coefficients $Y_{yu} = -Y$ and $\tilde{Y}_{yu} = -\tilde{Y}$, which should be taken into account in all related equations. The stabilizability condition (7.2) always holds for this G:

$$G = \begin{bmatrix} I & -N \\ I & -N \end{bmatrix}\begin{bmatrix} I & 0 \\ 0 & M \end{bmatrix}^{-1} = \begin{bmatrix} I & -I \\ 0 & \tilde{M} \end{bmatrix}^{-1}\begin{bmatrix} 0 & 0 \\ \tilde{M} & -\tilde{N} \end{bmatrix}$$

with

$$\left[\begin{array}{cc:cc} I & 0 & 0 & 0 \\ Y & X & 0 & -Y \\ \hdashline 0 & 0 & I & -I \\ -\tilde{M} & \tilde{N} & 0 & \tilde{M} \end{array}\right]\left[\begin{array}{cc:cc} I & 0 & 0 & 0 \\ 0 & M & 0 & \tilde{Y} \\ \hdashline I & -N & I & \tilde{X} \\ I & -N & 0 & \tilde{X} \end{array}\right] = \left[\begin{array}{cc:cc} I & 0 & 0 & 0 \\ 0 & I & 0 & 0 \\ \hdashline 0 & 0 & I & 0 \\ 0 & 0 & 0 & I \end{array}\right].$$

It then follows from (7.4) and the first equality in (7.6) that all stabilizing controllers are parametrized as

$$R = (\tilde{Y} + MQ)(\tilde{X} - NQ)^{-1} = (X - Q\tilde{N})^{-1}(Y + Q\tilde{M}) \tag{7.13}$$

and all attainable stable sensitivity functions are parametrized as

$$T_{zw} = S = (\tilde{X} - NQ)\tilde{M} \tag{7.14}$$

for an arbitrary $Q \in RH_\infty$ such that $\tilde{X}(\infty) - N(\infty)Q(\infty)$ is nonsingular. This form is quite convenient for the analysis of properties of S, as will be shown in the examples below.

Example 7.1. Let

$$P(s) = \frac{s - z_1}{s + 1}$$

for some $z_1 \in \mathbb{R}$. In this case we can always choose $M = \tilde{M} = X = \tilde{X} = 1$ and $N = \tilde{N} = P$. Hence, all attainable stable sensitivity transfer functions are

$$S(s) = 1 - \frac{s - z_1}{s + 1}\,Q(s)$$

for $Q \in RH_\infty$ such that $Q(\infty) \ne 1$. This is essentially an *open-loop plant inversion* problem like that considered in §1.4.1. There are two situations, depending on the sign of the plant zero:

1. If $z_1 < 0$, i.e. the plant is minimum phase, then the obvious choice $Q(s) = (s+1)/(s-z_1)$ is optimal, as it renders $S = 0$. But this choice is not admissible, because it violates the well-posedness condition. Still, we can approximate this ideal yet infeasible Q by altering its high-frequency gain slightly almost without sacrificing performance. For example, $Q(s) = ((1-\epsilon)s+1)/(s-z_1)$ is feasible for all $\epsilon \neq 0$ and yields $S(s) = \epsilon s/(s+1)$, whose H_∞ norm approaches zero as $\epsilon \to 0$.
2. If the plant is nonminimum phase ($z_1 \geq 0$), then we can no longer stably invert $N(s)$. In fact, the condition $S(z_1) = 1$ must hold regardless Q, because $Q(z_1)$ is bounded whenever $Q \in RH_\infty$. Taking into account the original definition of the H_∞ norm in (3.20) on p. 47, this implies that $\|S\|_\infty \geq 1$. In fact, this bound is attained by the trivial choice $Q = 0$, which happens to be unique.

As a matter of fact, all stable plants with bi-proper transfer functions can be treated similarly. ◊

Arguments of Example 7.1 can be extended to general stable plants with strictly proper transfer functions as well. Specifically, if P is stable, then all stable sensitivity functions are given by

$$S(s) = 1 - P(s)Q(s)$$

for an arbitrary $Q \in RH_\infty$ (because $P(\infty) = 0$, the controller is always well posed). Because $P(s)Q(s)$ is always strictly proper then, $S(\infty) = 1$ and the optimal $\|S\|_\infty = 1$, again attained by opening the loop. In other words, the minimum modulus margin for stable strictly proper plants is $\mu_\mathrm{m} = 1$ and it is attained by $Q = 0$. Of course, this is a senseless controller. This suggests that the problem of sheer minimizing $\|S\|_\infty$ is not the problem to rely on in controller design. Still, having a calculable lower bound on the modulus margin has its value, it helps to tell hard from easy plants from the stability margins point of view.

The situation becomes less trivial for unstable plants, as opening the loop is not an option here.

Example 7.2. Let now

$$P(s) = \frac{s - z_1}{s^2 - 1}$$

for some $z_1 \neq 1$. A possible choice of coprime factors in this case is

$$M(s) = \tilde{M}(s) = \frac{s-1}{s+a} \quad \text{and} \quad N(s) = \tilde{N}(s) = \frac{s - z_1}{(s+a)(s+1)} = P(s)M(s)$$

for any $a > 0$. This $M(s)$ is the lowest-order proper transfer function containing the unstable pole of $P(s)$ as its zero. The corresponding Bézout coefficients are

$$X(s) = \tilde{X}(s) = \frac{s + (az_1 + 2z_1 + a)/(z_1 - 1)}{s+1} \quad \text{and} \quad Y(s) = \tilde{Y}(s) = -\frac{2(a+1)}{z_1 - 1}.$$

Hence, all attainable stable sensitivity transfer functions are

$$S(s) = \left(\frac{s + (az_1 + 2z_1 + a)/(z_1 - 1)}{s+1} - \frac{s - z_1}{(s+1)(s+a)} Q(s) \right) \frac{s-1}{s+a}$$

for any $Q \in RH_\infty$. An important observation in analyzing this system is that the factor $(s-1)/(s+a)$ becomes co-inner (defined on p. 50) if $a = 1$. This choice is convenient because, by Proposition 3.1, the multiplication of a transfer function by a co-inner function from the right does not change its H_∞ norm. Hence, if $a = 1$, then $\|S\|_\infty = \|S_\mathrm{eq}\|_\infty$, where

$$S_\mathrm{eq}(s) := \frac{s + (3z_1 + 1)/(z_1 - 1)}{s+1} - \frac{s - z_1}{(s+1)^2} Q(s).$$

This is again an open-loop plant inversion problem, again with two qualitatively different situations, depending on the sign of the plant zero:

1. If $z_1 < 0$, then the only constraint imposed on S_{eq} is that $S_{\text{eq}}(\infty) = 1$ and it yields the lower bound $\|S_{\text{eq}}\|_\infty \geq 1$. This norm can always be reached, although the solution might not be unique. Indeed, the (nontrivial) choice

$$Q(s) = \frac{\tilde{X}(s) - \tilde{X}(\infty)}{N(s)} = \left(\frac{s + (3z_1 + 1)/(z_1 - 1)}{s + 1} - 1\right)\frac{(s+1)^2}{s - z_1} = \frac{2(z_1 + 1)}{z_1 - 1}\,\frac{s+1}{s - z_1} \tag{7.15}$$

yields $S_{\text{eq}}(s) = 1$, which obviously attains the lower bound. But $Q = 0$ is also an admissible solution if $-1 \leq z_1 < 0$, in which case $|(3z_1 + 1)/(z_1 - 1)| \leq 1$ and $|\tilde{X}(\mathrm{j}\omega)|$ is a monotonically increasing function of ω, approaching 1 as $\omega \to \infty$.

2. If $z_1 \geq 0$, then there are two constraints that $S_{\text{eq}}(s)$ must satisfy,

$$S_{\text{eq}}(\infty) = 1 \quad \text{and} \quad S_{\text{eq}}(z_1) = \tilde{X}(z_1) = \frac{1}{\tilde{M}(z_1)} = \frac{z_1 + 1}{z_1 - 1}$$

(if $z_1 = 0$, the fact that all components are real-rational is used to justify the second constraint). The first of these constraints is associated with the strict properness of $N(s)$ and the second one—with its RHP zero. Hence,

$$\|S_{\text{eq}}\|_\infty \geq \max\left\{1, \left|\frac{z_1 + 1}{z_1 - 1}\right|\right\} = \frac{z_1 + 1}{|z_1 - 1|}$$

(because $z_1 + 1 \geq |z_1 - 1|$ for all $z_1 \geq 0$). Thus, the constraint imposed by the nonminimum-phase zero of $N(s)$ is more restrictive than that imposed by its "zero at infinity." Forget for a moment about the latter and apply the same logic of the choice of Q as that used in the previous item to choose

$$Q(s) = \frac{\tilde{X}(s) - \tilde{X}(z_1)}{N(s)} = \left(\frac{s + (3z_1 + 1)/(z_1 - 1)}{s + 1} - \frac{z_1 + 1}{z_1 - 1}\right)\frac{(s+1)^2}{s - z_1} = -\frac{2}{z_1 - 1}(s + 1).$$

This Q results in the static $S_{\text{eq}}(s) = (z_1 + 1)/(z_1 - 1)$, which obviously attains the corresponding lower bound. Yet this $Q(s)$ is non-proper, so it should be modified. A simple choice is

$$Q(s) = -\frac{2}{z_1 - 1}\,\frac{s + 1}{\epsilon s + 1} \tag{7.16}$$

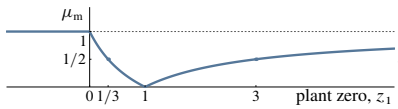
for some $\epsilon > 0$, which guarantees

$$\|S_{\text{eq}}\|_\infty < \frac{3\epsilon + 1}{\epsilon + 1}\,\frac{z_1 + 1}{|z_1 - 1|}$$

(the derivations are tedious). Hence, the performance attained by the non-proper $Q(s)$ above can be recovered by a proper $Q(s)$ arbitrarily close if ϵ is sufficiently small.

Summarizing, the supremal modulus margin attainable in this case is

$$\sup_{\text{stabilizing } R} \mu_{\text{m}} = \begin{cases} 1 & \text{if } z_1 < 0 \\ |z_1 - 1|/(z_1 + 1) & \text{if } z_1 \geq 0 \end{cases} =$$



.

This technical result is actually quite intuitive. It attests to the well-known thesis that nonminimum-phase zeros render feedback stabilization harder and especially so if such zeros are close to unstable plant poles.

As a matter of fact, the controller in (7.13) corresponding to Q from (7.16) has the transfer function

$$R(s) = -\frac{2(s+1)((1 + 2\epsilon)s + 1)}{\epsilon(z_1 - 1)s^2 + (z_1 + 1 + (3z_1 + 1)\epsilon)s + z_1 + 1} \xrightarrow{\epsilon \to 0} -\frac{2}{z_1 + 1}(s + 1).$$

This controller cancels the stable pole of the plant at $s = -1$, becomes unstable if $z_1 \in [0, 1)$, its static gain $R(0) = -2/(z_1 + 1)$, and its high-frequency gain $R(\infty) = -2(\epsilon^{-1} + 2)/(z_1 - 1)$ grows as $\epsilon \to 0$. This is perhaps not a controller one would choose in any meaningful formulation. Effectively, it succeeds only in one thing, the increase of the modulus margin. The controller for the minimum-phase case corresponding to Q in (7.15), $R(s) = 2(s + 1)/(s - z_1)$, is not quite impressive either. An important point to take note in this respect is that an optimal controller is not necessarily a "good" one. $\diamondsuit$

Arguments of Example 7.2 can be applied to general unstable plants with strictly proper nonminimum-phase transfer functions as well. To see this, let $P(s)$ have $n_{\text{rhpp}} \in \mathbb{N}$ poles in $\mathbb{C}_0$, say at $s = p_i$, and no poles on the imaginary axis. In this case we can always construct a coprime factorization of the plant with the co-inner $\tilde{M}(s) = \prod_i^{n_{\text{rhpp}}}(s - p_i)/(s + p_i)$. This again renders $\|S\|_\infty = \|\tilde{X} - NQ\|_\infty$. If $P(s)$ has a zero $z_j \in \mathbb{C}_0$, then so does $N(s)$ and $\|S\|_\infty \geq |\tilde{X}(z_j)|$. Now, at every zero of $N(s)$ we have $\tilde{X}(z_j) = 1/\tilde{M}(z_j)$, which follows from the Bézout equality $\tilde{M}\tilde{X} + \tilde{N}\tilde{Y} = 1$. Thus, $\|S\|_\infty \geq \max_j 1/|\tilde{M}(z_j)|$. In other words,

$$\mu_{\text{m}} \leq \min_{z_j \in \mathbb{C}_0} \prod_{i=1}^{n_{\text{rhpp}}} \left| \frac{z_j - p_i}{z_j + p_i} \right| < 1. \tag{7.17}$$

If $P(s)$ has only one nonminimum-phase zero, then this bound is tight. Otherwise, it might be conservative. But even then it supports the conventional wisdom that a plant is hard to control if its transfer function has poles and zeros in the open right half-plane $\mathbb{C}_0$ in a close proximity to each other.

The solution logic in the MIMO case is similar, although the results are richer.

Example 7.3. Consider

$$P(s) = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \frac{1}{s-1} + \begin{bmatrix} \alpha & -2z_1/(z_1 - 1) \\ -\alpha\beta & \beta \end{bmatrix} \frac{1}{s+1}$$

for some $\alpha \in \mathbb{R}$, $\beta \in \mathbb{R} \setminus \{0\}$, and $z_1 \in \mathbb{R}_+ \setminus \{1\}$. This system has an unstable pole at $s = 1$ with

$$\text{pdir}_{\text{i}}(P, 1) = \text{span}\left(\begin{bmatrix} 0 \\ 1 \end{bmatrix}\right) \quad \text{and} \quad \text{pdir}_{\text{o}}(P, 1) = \text{span}\left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}\right)$$

and a (nonminimum-phase) zero at $s = z_1$ with

$$\text{zdir}_{\text{i}}(P, z_1) = \text{span}\left(\begin{bmatrix} 1 \\ \alpha \end{bmatrix}\right) \quad \text{and} \quad \text{zdir}_{\text{o}}(P, z_1) = \text{span}\left(\begin{bmatrix} \beta \\ 1 \end{bmatrix}\right).$$

Possible doubly coprime factors of this $P(s)$ with inner $M(s)$ and $\tilde{M}(s)$ are

$$\begin{bmatrix} X & Y \\ -\tilde{N} & \tilde{M} \end{bmatrix} = \left[\begin{array}{cc:cc} 1 & 0 & 0 & 0 \\ 0 & 1 + \frac{z_1+1}{z_1-1}\frac{2}{s+1} & 2 & 2/\beta \\ \hdashline -\frac{\alpha(s-1)}{(s+1)^2} & \frac{(z_1+1)s-3z_1+1}{(z_1-1)(s+1)^2} & \frac{s-1}{s+1} & 0 \\ \frac{\alpha\beta}{s+1} & -\frac{\beta}{s+1} & 0 & 1 \end{array}\right]$$

and

$$\begin{bmatrix} M & -\tilde{Y} \\ N & \tilde{X} \end{bmatrix} = \left[\begin{array}{cc:cc} 1 & 0 & 0 & 0 \\ 0 & \frac{s-1}{s+1} & -2 & -\frac{2}{\beta}\frac{s-1}{s+1} \\ \hdashline \frac{\alpha}{s+1} & -\frac{(z_1+1)s-3z_1+1}{(z_1-1)(s+1)^2} & 1 + \frac{z_1}{z_1-1}\frac{4}{s+1} & 2\frac{(z_1+1)s-3z_1+1}{\beta(z_1-1)(s+1)^2} \\ -\frac{\alpha\beta}{s+1} & -\frac{\beta(s-1)}{(s+1)^2} & -\frac{2\beta}{s+1} & \frac{s^2+3}{(s+1)^2} \end{array}\right]$$

(the Bézout coefficients above are not required for the discussion to follow). Because $\tilde{M}$ is co-inner, we again have that $\|S\|_\infty = \|S_{\text{eq}}\|_\infty$, where $S_{\text{eq}} := \tilde{X} - NQ$. The transfer function $N(s)$ is not invertible

only at $s = \infty$, where it vanishes, and $s = z_1$, where its rank drops to 1. Hence, $S_{\text{eq}}(s)$ is constrained at those points. The constraint at infinity is not different from what we had in the SISO case of Example 7.2, namely, $S_{\text{eq}}(\infty) = \tilde{X}(\infty) = I$, so that $\|S_{\text{eq}}\|_\infty \geq 1$. However, the constraint at the nonminimum-phase zero is qualitatively different. We know from §4.3.3 that the unstable zeros of $P(s)$ and $N(s)$ and their output directions coincide. Hence, $S_{\text{eq}}(z_1)$ is only constrained along $\text{zdir}_\text{o}(P, z_1)$, whereas no constraints are imposed on it along $\mathbb{C}^2 \ominus \text{zdir}_\text{o}(P, z_1)$. In other words, we only have to account for

$$\eta' S_{\text{eq}}(z_1) = \eta' \tilde{X}(z_1) = \eta' \tilde{M}^{-1}(z_1), \quad \text{for any } \eta \in \text{zdir}_\text{o}(P, z_1),$$

where the second equality follows by $\tilde{X} = \tilde{M}^{-1} - P\tilde{Y}$. Thus, we have that

$$\|S\|_\infty \geq \|S_{\text{eq}}(z_1)\| \geq \max_{\eta \in \text{zdir}_\text{o}(P, z_1), \|\eta\|=1} \|\eta' \tilde{M}^{-1}(z_1)\|.$$

As $\text{zdir}_\text{o}(P, z_1))$ is a one-dimensional space, the only unity vector in it, modulo the sign, is $\eta = \left[\begin{smallmatrix} \cos\theta_\text{o} \\ \sin\theta_\text{o} \end{smallmatrix}\right]$, where $\theta_\text{o} = \arccos \beta/\sqrt{1+\beta^2} \in (0, \pi)$ is the angle between $\text{pdir}_\text{o}(P, 1)$ and $\text{zdir}_\text{o}(P, z_1)$, cf. (A.2) on p. 183. It is then readily seen that

$$\|S\|_\infty \geq \gamma_{\text{opt}} := \sqrt{\left(\frac{z_1+1}{z_1-1}\right)^2 \cos^2\theta_\text{o} + \sin^2\theta_\text{o}} \in \left(1, \frac{z_1+1}{|z_1-1|}\right).$$

Because $\gamma_{\text{opt}} > 1$, it can be shown that this lower bound is tight, meaning there are $Q \in H_\infty$ approaching it arbitrarily close.

It is informative to compare this γ_{opt} with the optimal performance attained in the SISO case studied in Example 7.2, which is $(z_1 + 1)/|z_1 - 1|$. The MIMO performance level, γ_{opt}, is always smaller and approaches the SISO performance as $\theta_\text{o} \to \pi/2$ (for $\beta \to \infty$). At the same time, as $\theta_o \to 0$ (for $\beta \to 0$), γ_{opt} approaches the performance level of the SISO control in the stable case studied in Example 7.1. This says that system is easier to control if output directions of unstable pole and zero are further apart. In other words, in MIMO systems not only relative positions of poles and zeros in $\mathbb{C}$, but also their spatial alignments are important.

As a matter of fact, the sensitivity function considered above is the output sensitivity $S_\text{o} = (I + PR)^{-1}$, cf. the discussion in the beginning of Section 1.5. If the input sensitivity $S_\text{i} = (I + RP)^{-1}$ was considered, the angle between *input* directions of the unstable pole and zero of $P(s)$ would be relevant. ◊

7.3.2 Weighted sensitivity

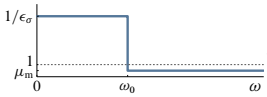
The problem of minimizing the sensitivity function uniformly over all frequencies offers an insight into intrinsic limitations of feedback systems, but is not a way to produce meaningful controllers. This is because stability margins is not the only requirement to control systems. A better justified approach to handle $S(\text{j}\omega)$ would be to impose different requirements on it in different frequency ranges. For example, it is normally required to reduce $|S(\text{j}\omega)|$ at low frequencies within a required closed-loop bandwidth, see the discussion in the beginning of §1.4.3 and the design in §1.A.4. Combining this requirement with that on the modulus margin, which should hold over all frequencies, we may consider the following requirements:

$$|S(\text{j}\omega)| \leq \begin{cases} \epsilon_\sigma & \text{if } \omega \leq \omega_0 \\ 1/\mu_\text{m} & \text{otherwise} \end{cases} \tag{7.18}$$

for some $\epsilon_\sigma < 1$, $\mu_\text{m} < 1$, and $\omega_0 > 0$, which may be viewed as tuning parameters. These are simplified requirements, one may add more requirements in more frequency ranges, like rendering $S(\text{j}\omega) = 0$ at a number of selected frequencies, like $\omega = 0$, et cetera. Nonetheless, (7.18) does capture the essence of the idea and is thus sufficiently general for our purposes.

Requirements (7.18) can be cast as a standard H_∞ problem. To this end, introduce a system W_σ having the frequency response

$$|W_\sigma(\mathrm{j}\omega)| = \begin{cases} 1/\epsilon_\sigma & \text{if } \omega \le \omega_0 \\ \mu_\mathrm{m} & \text{otherwise} \end{cases} =$$



. (7.19)

Bound (7.18) is then equivalent to the condition $|W_\sigma(\mathrm{j}\omega)S(\mathrm{j}\omega)| \le 1$ for all ω. Equivalently, if the natural condition $W_\sigma S \in H_\infty$ holds, this can be expressed as the H_∞-norm bound

$$\|W_\sigma S\|_\infty \le 1. \tag{7.20}$$

The question of the existence of a stabilizing controller guaranteeing (7.20) can be addressed via minimizing $\|W_\sigma S\|_\infty$. Clearly, a required controller exists iff the minimum attainable norm is smaller than or equal to 1. The existence of such a controller can be resolved conclusively in the framework of the standard H_∞ problem.

The H_∞ problem with the cost $\|W_\sigma S\|_\infty$ for a given weighting function W_σ is known as the *weighted sensitivity* problem. It can be formulated for a wide class of weighting functions, not necessarily those having the frequency response as in (7.19). The corresponding generalized plant is

$$G(s) = G_\mathrm{ws}(s) = \begin{bmatrix} W_\sigma(s) & -W_\sigma(s)P(s) \\ I & -P(s) \end{bmatrix}, \tag{7.21}$$

which can be constructed by similar arguments to those leading to (7.12). Important is that the minimization here is a tool to attain a feasible solution for (7.20) rather than a goal per se, so the central question is whether the system $T_{zw} = W_\sigma S$ corresponding to (7.21) is contractive.

It follows from Theorem 6.13, cf. the discussion on p. 134, that this system is internally stabilizable only if all its instabilities are contained in $G_{yw} = -P$. This entails that the condition $W_\sigma \in H_\infty$ should be imposed on the weighting function. But the system is always internally stabilizable as

$$G = \begin{bmatrix} W_\sigma & -W_\sigma N \\ I & -N \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & M \end{bmatrix}^{-1} = \begin{bmatrix} I & -W_\sigma \\ 0 & \tilde{M} \end{bmatrix}^{-1} \begin{bmatrix} 0 & 0 \\ \tilde{M} & -\tilde{N} \end{bmatrix}$$

is the coprime factorization required in Theorem 6.13 (the corresponding Bézout coefficients can be constructed similarly to their counterparts for G in (7.12)). The set of all stabilizing controllers in this case is still that in (7.13), yielding the affine family of attainable stable

$$T_{zw} = W_\sigma S = W_\sigma(\tilde{X} - NQ)\tilde{M} \tag{7.22}$$

in terms of an arbitrary $Q \in RH_\infty$ such that $\tilde{X}(\infty) - N(\infty)Q(\infty)$ is invertible.

Before we discuss the solution to the weighted sensitivity problem, some additional clarifications about properties of the weighting function are in order.

- There might be situations when weighting functions with pure imaginary poles are required, like if there is a need to impose an integral action in the controller via requiring $S(0) = 0$. In such situations, we have to relax the internal stability requirement and use ideas from Section 6.3. Addressing such situations goes beyond the scope of the notes though.
- Without loss of generality we may limit our attention to weighting functions having no zeros in $\mathbb{C}_0$. This is because $|W_\sigma(\mathrm{j}\omega)| = |W_\sigma(\mathrm{j}\omega)\,(\mathrm{j}\omega - z_1)/(\mathrm{j}\omega + z_1)|$, so we can replace any RHP zero with its mirror in the LHP without affecting the left-hand side of (7.20).

- We also assume that $W_\sigma(s)$ has no pure imaginary zeros, i.e. that $|W_\sigma(\mathrm{j}\omega)| \neq 0$ for all ω. This should not appreciably limit the class of considered problems. Indeed, $|W_\sigma(\mathrm{j}\omega_1)| = 0$ effectively implies that $|S(\mathrm{j}\omega_1)|$ is not a part of the optimization problem. Because $|S(\mathrm{j}\omega_1)|$ must be bounded anyway (as $S \in H_\infty$), adding a very small penalty on it does not change the problem.

Thus, in what follow we assume that $W_\sigma(s)$ has neither poles no zeros in $\bar{\mathbb{C}}_0$ and is bi-proper.

The problem of minimizing the H_∞ norm of (7.22) is then not quite different from that of (7.14). Considering for the sake of simplicity the SISO case, the first step is again to select a co-inner $\tilde{M}$, which is possible whenever $P(s)$ has no pure imaginary poles, and then to look for constraints imposed upon $T_{\mathrm{eq}} := W_\sigma \tilde{X} - W_\sigma N Q$ by non-invertible components of $W_\sigma N$. The assumptions on W_σ above imply that these are exactly non-invertible components of N. Two cases should again be considered separately.

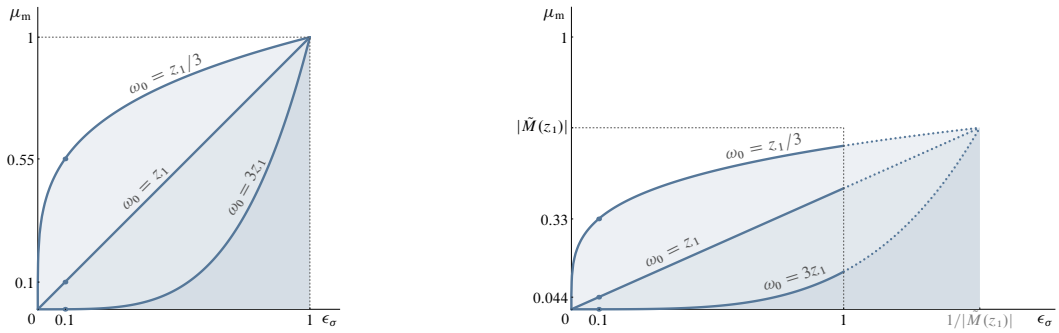
1. If $P(s)$ is minimum phase, then so is $W_\sigma(s)N(s)$ and, loosely speaking, the only constraint on the closed-loop transfer function is $T_{\mathrm{eq}}(\infty) = W_\sigma(\infty)\tilde{X}(\infty) = W_\sigma(\infty) = \mu_{\mathrm{m}}$. This constraint expectably agrees with (7.20) if $\mu_{\mathrm{m}} \leq 1$. But it depends neither on ϵ_σ nor on ω_0, meaning that any low-frequency performance level over any frequency band can be attained.

2. If $P(s)$ has unstable zeros, say at $s = z_i$, those are the only unstable zeros of $W_\sigma(s)N(s)$ and we have that
$$T_{\mathrm{eq}}(z_i) = W_\sigma(z_i)\tilde{X}(z_i) = W_\sigma(z_i)/\tilde{M}(z_i).$$
Thus, there is a stabilizing controller rendering $\|T_{\mathrm{eq}}\|_\infty = \|W_\sigma S\|_\infty \leq 1$ only if $|W_\sigma(z_i)/\tilde{M}(z_i)| \leq 1$, which, in turn, is equivalent to the condition, cf. (7.17),
$$|W_\sigma(z_i)| \leq |\tilde{M}(z_i)| = \prod_{j=1}^{n_{\mathrm{rhpp}}} \left| \frac{z_i - p_j}{z_i + p_j} \right| \leq 1 \tag{7.23}$$
at every nonminimum-phase zero z_i of the plant. This condition becomes also sufficient if there is only one nonminimum-phase zero of $P(s)$, say at $s = z_1$ (we again ignore a potential non-properness of $1/N(s)$ which is almost always safe to do).

Performance limitations

Let us first focus on implications of condition (7.23), which depends on all parameters of the original constraint (7.18)—the low-frequency performance level $\epsilon_\sigma < 1$, the modulus margin $\mu_{\mathrm{m}} < 1$, and the bandwidth ω_0. The analysis of (7.23) is hindered by the fact that the weighting function in (7.19) is defined in terms of its magnitude frequency response, whereas we mostly need to analyze its magnitude at points in $\mathbb{C}_0$. In mathematical terms, we need to extrapolate the absolute value of an H_∞ function $W_\sigma(s)$ from the absolute values of its boundary function. It is known, see Remark 3.3, that an H_∞ function can be recovered from its boundary function via the Poisson integral formula (3.23). The magnitude of the frequency response does not define an H_∞ function unambiguously. For example, 1 and $(s-1)/(s+1)$ are both in H_∞ and both have the unity magnitude of their frequency responses. Still, a class of H_∞ functions, known as *outer* (roughly, minimum-phase) functions, can be recovered from the magnitude of their frequency responses. This fact is established in the following technical lemma, see [21, Sec. 1.3].

Lemma 7.1. *If $\phi(\omega) : \mathbb{R} \to \mathbb{R}$ satisfies $\int_{\mathbb{R}} |\phi(\omega)|/(1+\omega^2)\,\mathrm{d}\omega < \infty$, then the (outer) function*
$$f(s) = \exp\left(\frac{1}{\pi} \int_{\mathbb{R}} \phi(\omega) \left(\frac{\operatorname{Re} s}{(\operatorname{Re} s)^2 + (\operatorname{Im} s - \omega)^2} - \mathrm{j}\left(\frac{\operatorname{Im} s - \omega}{(\operatorname{Re} s)^2 + (\operatorname{Im} s - \omega)^2} + \frac{\omega}{1+\omega^2} \right) \right) \mathrm{d}\omega \right)$$
belongs to H_∞ and is such that $\lim_{\sigma \downarrow 0} \ln|f(\sigma + \mathrm{j}\omega)| = \phi(\omega)$ for almost every ω.



(a) Stable plant

(b) Unstable plant, $|\tilde{M}(z_1)| = 2/3$

Fig. 7.3: ϵ_σ–μ_m tradeoff (waterbed effect) for different ω_0/z_1; admissible regions are below lines

Having this result, we are in a position to analyze (7.23). An immediate consequence of Lemma 7.1 is that any outer (minimum phase, in engineering terms) transfer function $W_\sigma(s)$ satisfies

$$|W_\sigma(s)| = \exp\left(\frac{1}{\pi}\int_{\mathbb{R}} \ln|W_\sigma(\mathrm{j}\omega)| \frac{\operatorname{Re} s}{(\operatorname{Re} s)^2 + (\operatorname{Im} s - \omega)^2}\mathrm{d}\omega\right) = \exp\left(\frac{1}{\pi}\int_{\mathbb{R}} \ln|W_\sigma(\mathrm{j}\omega)|\, \mathrm{d}\arctan\frac{\omega + \operatorname{Im} s}{\operatorname{Re} s}\right)$$

at every $s \in \mathbb{C}_0$. For the sake of simplicity, assume that there is only one unstable zero of $P(s)$, say at $s = z_1 > 0$. Thus, taking into account (7.19), we have that

$$|W_\sigma(z_1)| = \exp\left(-\frac{2\ln\epsilon_\sigma}{\pi}\int_0^{\omega_0} \mathrm{d}\arctan\frac{\omega}{z_1} + \frac{2\ln\mu_m}{\pi}\int_{\omega_0}^{\infty} \mathrm{d}\arctan\frac{\omega}{z_1}\right)$$

from which

$$|W_\sigma(z_1)| = \frac{(\mu_m)^{1-\beta_z}}{(\epsilon_\sigma)^{\beta_z}}, \quad \text{where } \beta_z := \frac{2}{\pi}\arctan\frac{\omega_0}{z_1} \in (0, 1) \tag{7.24}$$

(this β_z is a strictly increasing function of the normalized bandwidth ω_0/z_1). $|W_\sigma(z_1)|$ in (7.24) is a decreasing function of ϵ_σ, an increasing function of μ_m, and, because $\epsilon_\sigma \in (0, 1)$ and $\mu_m \in (0, 1)$, an increasing function of ω_0. These properties support the intuition that the problem becomes harder as ϵ_σ decreases and μ_m and ω_0 increase. The fact that the required bandwidth is normalized by the position of the nonminimum-phase zero of the plant also agrees with the conventional wisdom that limitations on the achievable closed-loop bandwidth become more severe as the unstable zero approaches the real axis.

Reachable ϵ_σ and μ_m as functions of the intended bandwidth are visualized in Fig. 7.3. Shaded areas there show regions in the ϵ_σ–μ_m plane in which inequality (7.23) can hold for various ω_0/z_1 and under various $|\tilde{M}(z_1)|$ on the right-hand side of (7.23). First, consider the case when the plant is stable, for which $\tilde{M} = 1$, see Fig. 7.3(a). We know from the analysis in §7.3.1 that any modulus margin $\mu_m \le 1$ can be attained here if this is the only design goal. Fig. 7.3(a) shows that this is no longer the case if, in addition, we need to reduce the sensitivity magnitude below some level in a finite low-frequency band. For example, if $\epsilon_\sigma = 0.1$, then the attainable modulus margin drops from about 0.55 when $\omega_0 = z_1/3$ to virtually zero when $\omega_0 = 3z_1$. This implies that as we push $|S(\mathrm{j}\omega)|$ down in $\omega \in [0, \omega_0]$, it necessarily grows outside this frequency band, up to $1/\mu_m$. That phenomenon is known as the *waterbed effect*, which is fairly self-explanatory. The situation is similar in the unstable plant case shown in Fig. 7.3(b), the only difference is that we could not attain the upper bound on μ_m, which equals now $|\tilde{M}(z_1)| < 1$, if $\epsilon_\sigma < 1$.

Note that the level curves in the plots in Fig. 7.3 are straight lines if $\omega_0 = z_1$, convex curves with a vertical slope at $\epsilon_\sigma \downarrow 0$ if $\omega_0 < z_1$, and concave curves with a horizontal slope at $\epsilon_\sigma \downarrow 0$ if $\omega_0 > z_1$. Hence, if the bandwidth requirement does not exceed the zero, the system is easier to control in a sense that a slight relaxation of the low-frequency magnitude requirement yields a significant modulus margin improvement.

If the bandwidth requirement goes beyond the size of the RHP zero, then similar improvements of μ_{m} require a significantly larger concession in terms of the performance. This agrees with a classical rule of thumb, limiting the closed-loop bandwidth by the location of the nonminimum-phase zero of the system.

Optimal controller: design and properties

The performance bounds discussed above were obtained assuming that the frequency response of the weighting function is (7.19). This frequency response corresponds to an infinite-dimensional system. As a result, an infinite-dimensional Q, and then R, are required to approach these bounds, which is not practical. Normally, finite-dimensional weights are used to design finite-dimensional controllers (designing finite-dimensional controllers for infinite-dimensional problem data is a substantially more complicated, and less transparent, task). It is customary to capture the spirit of requirement (7.18) by simple real-rational weights, whose frequency responses are close to (7.19).

A possible, albeit not unique, direction here is to use normalized Butterworth polynomials. These are Hurwitz polynomials $B_n(s)$ of order $n \in \mathbb{N}$, whose frequency response magnitude

$$|B_n(\mathrm{j}\omega)| = \sqrt{1+\omega^{2n}} \tag{7.25}$$

is a monotonically increasing function of ω, approximating $\max\{1, \omega^n\}$ well. For example, $B_1(s) = s+1$ and $B_2(s) = s^2 + \sqrt{2}s + 1$. We may consider the bi-proper and minimum-phase transfer function of the form

$$W_{\sigma,n}(s) = k\frac{B_n(s/\omega_2)}{B_n(s/\omega_1)}, \tag{7.26}$$

whose three parameters are tuned to approximate (7.19), as a candidate weighting function. The choice of its parameters might be based on the conditions

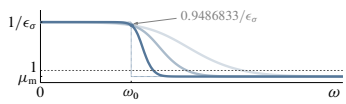
$$W_{\sigma,n}(0) = k = \frac{1}{\epsilon_\sigma}, \quad W_{\sigma,n}(\infty) = k\frac{\omega_1^n}{\omega_2^n} = \mu_{\mathrm{m}}, \quad \text{and} \quad |W_{\sigma,n}(\omega_0)| = k\frac{\omega_1^n}{\omega_2^n}\sqrt{\frac{\omega_2^{2n}+\omega_0^{2n}}{\omega_1^{2n}+\omega_0^{2n}}} = \frac{\alpha}{\epsilon_\sigma}$$

for some $\alpha \in (0,1)$ determining what we compromise in terms of performance at frequencies close to ω_0. These equalities yield

$$k = \frac{1}{\epsilon_\sigma}, \quad \omega_1 = \left(\frac{\alpha^2 - (\mu_{\mathrm{m}}\epsilon_\sigma)^2}{1-\alpha^2}\right)^{1/2n}\omega_0, \quad \text{and} \quad \omega_2 = \left(\frac{\alpha^2/(\mu_{\mathrm{m}}\epsilon_\sigma)^2 - 1}{1-\alpha^2}\right)^{1/2n}\omega_0 > \omega_1, \tag{7.27}$$

which corresponds to the following frequency responses:

$$|W_{\sigma,n}(\mathrm{j}\omega)| =$$



if $\alpha = \sqrt{0.9} \approx 0.95$ and $n \in \{1, 2, 5\}$ (with the logarithmic abscissa scale).

With this introduction, we are in a position to consider simple design examples.

Example 7.4. Return to Example 7.1 with the plant

$$P(s) = \frac{s - z_1}{s+1}$$

and assume that it is nonminimum phase, i.e. that $z_1 > 0$. Let $\epsilon_\sigma = 0.1$ and $\mu_{\mathrm{m}} = 0.5$, with the bandwidth being our tuning parameter. Select a second-order Butterworth weight with $\alpha = \sqrt{0.9}$, which is

$$W_{\sigma,2}(s) = \frac{0.5s^2 + 5.47\omega_0 s + 30\omega_0^2}{s^2 + 2.45\omega_0 s + 3\omega_0^2} \tag{7.28}$$

in this case. Condition (7.23) with this weighting function yields $\omega_0 \leq 0.0912 z_1$. As a matter of fact, this is more conservative, by more than a factor of four, than the bound $\omega_0 \leq 0.38 z_1$ resulting from (7.23) for the original weighting function with the magnitude frequency response (7.19). Choosing then the maximal $\omega_0 = 0.0912 z_1$ and following the arguments of Example 7.1, all stable closed-loop transfer functions are

$$T_{zw}(s) = \frac{0.5s^2 + 0.499 z_1 s + 0.249 z_1^2}{s^2 + 0.223 z_1 s + 0.0249 z_1^2}\left(1 - \frac{s - z_1}{s + 1}\, Q(s)\right)$$

for $Q \in RH_\infty$ such that $Q(\infty) \neq 1$. The H_∞-optimal Q for this problem is

$$Q(s) = \frac{W_{\sigma,2}(s) - W_{\sigma,2}(z_1)}{W_{\sigma,2}(s)P(s)} = -\frac{(s+1)(s + 0.448 z_1)}{s^2 + 0.998 z_1 s + 0.498 z_1^2}$$

and then the optimal controller

$$R(s) = \frac{0.5(s+1)(s + 0.448 z_1)}{s^2 + 0.223 z_1 s + 0.0249 z_1^2}.$$

The order of this controller is two, which is actually one short of the order of the generalized plant (7.21). This property is common for H_∞ optimal solutions. Another common property of this controller, this time regarding weighted sensitivity H_∞ problems, is that it cancels the stable pole of the plant. Canceling stable plant poles might not be desirable if $P(s)$ has lightly damped poles, but the weighted sensitivity formulation does not have an intuitive control over that.

Consider now the control effort required to attain (7.20). To this end, we need to analyze the the control sensitivity function $T_c = SR = (\tilde{Y} + MQ)\tilde{M}$. With the chosen controller,

$$T_c(s) = \frac{(s+1)(s + 0.448 z_1)}{s^2 + 0.998 z_1 s + 0.498 z_1^2}.$$

Both its static gain, $|T_c(0)| = 0.9/z_1$, and the peak of its magnitude frequency response,

$$\|T_c\|_\infty = \sqrt{0.5 + 0.405/z_1^2 + \sqrt{0.2907 + 1.1717/z_1^4}} > 1.0194,$$

grow without bound as the zero z_1 approaches the origin. This implies that not only the problem becomes harder for smaller z_1, but also that attaining the required performance level requires higher control effort in that case. The magnitude of T_c can be reduced by relaxing the requirements, but there is no direct parameter that can affect it in the weighted sensitivity problem either. ◊

7.3.3 Mixed sensitivity

Control effort can be addressed by imposing explicit constraints on the gain of the control sensitivity frequency response, similarly to what was done with the sensitivity frequency response in (7.18). A possible choice, which is sufficiently simple while still reflects basic requirements, is

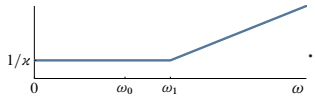
$$|T_c(\mathrm{j}\omega)| \leq \varkappa \min\{1, (\omega_1/\omega)^\nu\} \tag{7.29}$$

for some $\varkappa > 0$ reflecting constraints on the gain of the closed-loop mapping $y_r \mapsto u$ and $\nu \in \mathbb{Z}_+$ reflecting requirements on the controller roll-off at high frequencies (understood as frequencies $\omega > \omega_1$). This is again a constraint on the contractiveness of the H_∞-norm, this time of the weighted control sensitivity in

$$\|W_\varkappa T_c\|_\infty \leq 1$$

for

$$|W_\varkappa(\mathrm{j}\omega)| = \frac{\max\{1, (\omega/\omega_1)^\nu\}}{\varkappa} =$$



. (7.30)

When this constraint complements (7.20), we have the problem of designing a stabilizing controller, which guarantees

$$\|W_\sigma S\|_\infty \le 1 \quad \text{and} \quad \|W_\varkappa T_\mathrm{c}\|_\infty \le 1. \tag{7.31}$$

This is the so-called *multidisk problem* and it is different from the standard H_∞ problem. This is because (7.31) requires that

$$\begin{bmatrix} W_\sigma(\mathrm{j}\omega)S(\mathrm{j}\omega) \\ W_\varkappa(\mathrm{j}\omega)T_\mathrm{c}(\mathrm{j}\omega) \end{bmatrix} \in \mathcal{B}_\infty, \quad \forall\omega$$

i.e. belongs to the unit ball in the ∞-Hölder vector norm on $\mathbb{C}^2$, see Fig. 2.1(c) on p. 26, whereas the underlying vector norm for the H_∞ space is the Euclidean (2-Hölder) norm. The solution to the multidisk problem is substantially more complicated than that of a standard H_∞ problem.

A workaround is to note that $\mathcal{B}_2 \subset \mathcal{B}_\infty$, cf. the areas in Figs. 2.1(b) and 2.1(c), so that (7.31) holds whenever

$$\left\| \begin{bmatrix} W_\sigma S \\ W_\varkappa T_\mathrm{c} \end{bmatrix} \right\|_\infty \le 1. \tag{7.32}$$

The problem of designing a stabilizing controller guaranteeing this condition for some stable weighting functions W_σ and $W_\varkappa$ (their frequency responses might be different from those in (7.19) and (7.30)) is known as the *mixed sensitivity problem*. The mixed sensitivity formulation introduces some conservatism, because the failure of attaining (7.32) does not necessarily implies that of (7.31), unless the minimal attainable norm of the left hand side of (7.32) is above $\sqrt{2}$.

At the same time, the mixed sensitivity problem is a special case of the standard H_∞ problem and as such can be solved by standard tools, like those presented in §7.A.2. To see this, return to the unity-feedback system in Fig. 1.4(c) and note that T_c corresponds to the system $y_\mathrm{r} \mapsto u$ there, like S—to the system $y_\mathrm{r} \mapsto e$. Thus suggests that we may still select the exogenous signal as y_r and the regulated signal—as $\left[\begin{smallmatrix} W_\sigma e \\ W_\varkappa u \end{smallmatrix}\right]$, the measured output—as e, and u in its standard role. This yields the generalized plant

$$G(s) = G_\mathrm{MS}(s) = \left[\begin{array}{c:c} W_\sigma(s) & -W_\sigma(s)P(s) \\ 0 & W_\varkappa(s) \\ \hdashline I & -P(s) \end{array}\right], \tag{7.33}$$

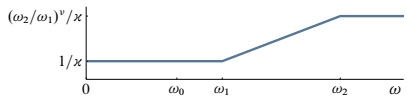
whose closed-loop system $w \mapsto z$ is exactly the system on the left-hand side of (7.32). Because $W_\varkappa$ is assumed to be stable, all unstable poles of this $G(s)$ are indeed those of $P(s)$, so that the system is stabilizable. Because G_{yu} in (7.33) is the same as that in (7.12), all stabilizing controllers for the mixed sensitivity problem are again given by (7.13) and all stable closed-loop systems are parametrized as

$$T_{zw} = \begin{bmatrix} W_\sigma & 0 \\ 0 & -W_\varkappa \end{bmatrix} \left(\begin{bmatrix} \tilde{X} \\ -\tilde{Y} \end{bmatrix} - \begin{bmatrix} N \\ M \end{bmatrix} Q \right) \tilde{M},$$

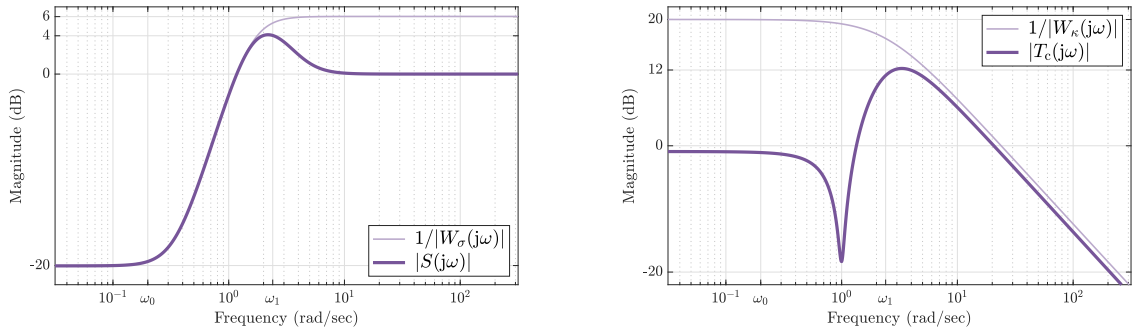
which is again an affine function of Q.

Remark 7.2 (proper $W_\varkappa(s)$). The requirement on $W_\varkappa$ to be stable actually conflicts with (7.30), which grows unbounded as $\omega \to \infty$. A simple workaround is to replace it with something like

$$|W_\varkappa(\mathrm{j}\omega)| = \frac{\min\{(\omega_2/\omega_1)^\nu, \max\{1, (\omega/\omega_1)^\nu\}\}}{\varkappa} =$$



(7.30′)



(a) Sensitivity S and its weight W_σ

(b) Control sensitivity T_c and its weight $W_\varkappa$

Fig. 7.4: Plots for Example 7.5

for a sufficiently large $\omega_2 > \omega_1$. This function can be approximated by the bi-stable

$$W_{\varkappa,\nu}(s) = \frac{B_\nu(s/\omega_1)}{\varkappa B_\nu(s/\omega_2)}, \tag{7.34}$$

where $B_\nu(s)$ is the normalized Butterworth polynomial of order ν, whose frequency response gain is as in (7.25). A more elegant approach, which does not involve approximations, is described in [18]. ∇

Solving the standard problem corresponding to the generalized plant (7.33) in an analytic form is no longer informative and reasonably compact. Yet the problem can be solved via the state-space procedure presented in §7.A.2. Technical assumptions there hold whenever W_σ and $W_\varkappa$ are bi-stable, so the problem is normally well posed. The following example .

Example 7.5. Consider a stable lightly damped plant with the transfer function

$$P(s) = \frac{1}{s^2 + 0.1s + 1}$$

and the following requirements to its closed-loop frequency responses:

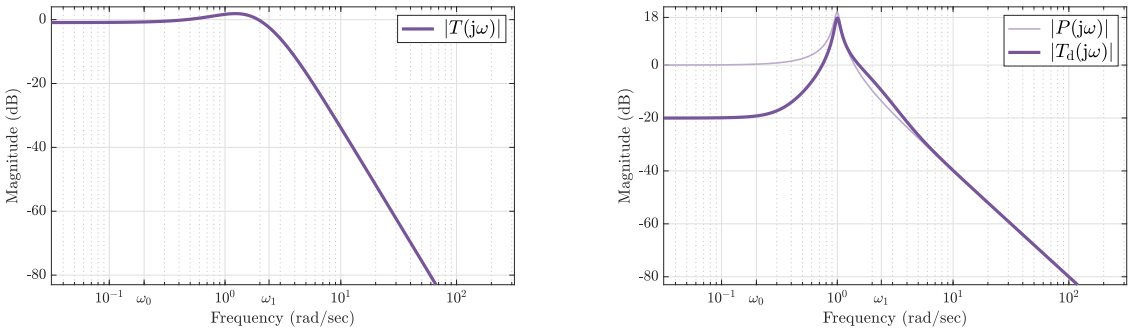
1. the sensitivity magnitude does not exceed $\epsilon_\sigma = 0.1$ over the frequency band $[0, \omega_0]$ for some ω_0,
2. the modulus margin $\mu_\mathrm{m} \geq 0.5$,
3. the control sensitivity is bounded by $\varkappa = 10$ and decays with the roll-off 1 after $\omega_1 = 2.4$ [rad/sec].

Our tuning parameter is ω_0, which we endeavor to increase as long as condition (7.32) holds. The sensitivity weighting function W_σ is chosen as the second-order weight of form (7.26) with $\alpha = \sqrt{0.9}$. This is exactly the choice of Example 7.4 presented by (7.28). The control sensitivity weight is a first-order transfer function of form (7.34) with $\omega_2 = 1000\omega_1$, which is sufficiently high. Thus, our choice is

$$W_{\varkappa,1}(s) = \frac{100(s + 2.4)}{s + 2400}.$$

A simple trial and error procedure yields the maximal attainable $\omega_0 = 0.2$ [rad/sec]. The resulted sensitivity functions are presented in Fig. 7.4 and are located below the plots of the reciprocal to the corresponding weighting function (thin light lines). In fact, the attained modulus margin, $\mu_\mathrm{m} \approx 0.623$, is even better than that required. The resulting controller has the transfer function

$$R(s) = \frac{15367(s + 2400)(s + 0.9364)(s^2 + 0.1s + 1)}{(s + 1.774 \cdot 10^6)(s^2 + 0.4896s + 0.1198)(s^2 + 6.372s + 17.97)}.$$



(a) Complementary sensitivity T

(b) Disturbance sensitivity T_{d}

Fig. 7.5: More plots for Example 7.5

One of its poles is located very far left, which is a technical property caused by numerical inaccuracies. This pole can be safely removed. There is also a zero located at the very pole of the weight $W_{\varkappa,1}(s)$ above at $s = 2400$. This is a general property, so if this pole is chosen to be sufficiently far left, the corresponding controller zero can also be safely removed. As a matter of fact, the Hankel singular values of this controller are $\{5.3784, 1.9576, 1.5738, 0.4788, 0.0043\}$, also suggesting that a fourth-order approximation would be almost the same. We can then perform the balanced truncation procedure studied in Section 4.4. But in this case a naïve cancellation of the far-left pole and zero,

$$R(s) = \frac{20.787(s + 0.9364)(s^2 + 0.1s + 1)}{(s^2 + 0.4896s + 0.1198)(s^2 + 6.372s + 17.97)},$$

yields practically the same result and also does not alter the static gain of the controller, $R(0) = 9.04$.

Another general property of the optimal mixed sensitivity controller is that it cancels all stable poles of the plant, similarly to what we saw in the weighted sensitivity optimization. This might be problematic in the case when the plant has lightly-damped poles, like what we have in the present example. The problem is already hinted at in the control sensitivity plot in Fig. 7.4(b), where a visible notch at the plant resonance around 1 [rad/sec] is present. This implies that the controller effectively "shuts its eyes" at that frequency and is thus inactive in dampening the resonance. A more perceptible picture is obtained via inspecting two remaining closed-loop frequency responses, which are the complementary sensitivity function T and the disturbance sensitivity T_{d} presented in Fig. 7.5. The cancellation of lightly-damped plant poles results in that the resonance peak of the plant remains present in T_{d}, see Fig. 7.5(b). This phenomenon is a result of the exclusion of T_{d} from the mixed sensitivity cost. Consequently, it happens that it pays off for the optimization procedure to cancel them. To avoid this trait, the disturbance sensitivity should also be made a part of the cost function, perhaps with its own weight. However, this would further increase the number of design parameters, rendering the weights harder to tune. ◊

7.3.4 Concluding remarks

In this section we saw how several simple, perhaps simplistic, problems motivated by classical frequency-domain design ideas can be cast as optimal control problems with the H_∞ cost. Some traits of the described techniques are worth emphasizing, as they are representative of the whole family of methods.

- The H_∞ system norm, which is uniform over all frequencies, might appear "dumb" from the classical control perspective, where different frequency ranges play different roles. Nevertheless, it becomes a powerful tool for shaping system gains over different frequencies, i.e. "cleverly" selective, by the use of frequency-dependent weighting functions. Adding such weights is similar to the use of matrix

weights to shape spatial properties in Section 2.4. This idea, which is one of the cornerstones of modern optimization-based approaches, rendering them a powerful analysis and design tool.

- A significant advantage of the optimization-based approaches discussed in this section is that they succeed in *separating* hardly formalizable yet technically simple stage of the specification (weighting) selection from the technically more difficult stage of the controller design for given specifications. The latter stage, which is frequently most difficult in the classical design, is now well formalized, with the internal stability granted "for free." Consequently, a negative answer ($\gamma_{\text{opt}} \geq 1$) means that there exist no stabilizing controllers achieving required performance level. This is a clear advantage over the classical loop-shaping methods, where the failure to find an admissible controller does not imply that such a controller does not exist for given specifications.
- The optimal attainable performance level is treated in many situations just as the *indicator* of whether required specifications can be met, rather than as the "best achievable performance." In other words, the optimization here serves as a technical tool and by no means as a design goal.
- The use of optimization-based methods, both H_2 and H_∞, in MIMO systems is not substantially different from that in SISO systems. Of course, the choice of weighting functions becomes more complicated, with more design parameters and more properties to take into consideration (like spatial directions). But the technical side is the same, except perhaps for a handful of problems where analytic solutions are available. In fact, optimization-based methods offer natural generalizations of some SISO notions, like stability margins, which would be highly nontrivial otherwise.
- With all these advantages, "you get what you pay for," they say, and they are right. Optimization is not a panacea and there is no way to squeeze all requirements to any real-life control problem into one cost function. Optimization-based approaches should thus be used consciously. There is no universal control design method and no universal recipe for selecting weighting functions. We saw that optimization procedures might be efficient in finding weaknesses in the cost function to produce optimal yet poor controllers. This should be always remembered and any solution must be tested carefully, especially from viewpoints that are not explicitly included in the cost.

As a matter of disclosure, my design method of choice is H_∞ loop shaping, which is sufficiently simple yet powerful and efficient in many situations. This method is discussed in Chapter 9.

7.A State-space solutions to standard problems

This appendix collects solutions to the H_2 and H_∞ problems for the setup in Fig. 7.1 and the generalized plant given in terms of its state-space realization as in (7.1). The goal here is to provide a handy reference, rather than ideas and techniques for solving these problems. For this reason, no comprehensive proofs are provided. But then the results are presented in almost the most general forms, without imposing simplifying assumptions and without the need to carry out intermediate transformations. This deviates from what is conventionally done in the literature.

Both the H_2 and the H_∞ versions of the standard problem require the following assumptions on the parameters of the realization in (7.1):

$\mathcal{A}_1$: the pair (A, B_u) is stabilizable,

$\mathcal{A}_2$: the pair (C_y, A) is detectable,

$\mathcal{A}_3$: the realization (A, B_u, C_z, D_{zu}) has no invariant zeros in $\mathrm{j}\mathbb{R}$ and $D'_{zu}D_{zu} > 0$,

$\mathcal{A}_4$: the realization (A, B_w, C_y, D_{yw}) has no invariant zeros in $\mathrm{j}\mathbb{R}$ and $D_{yw}D'_{yw} > 0$.

Assumptions $\mathcal{A}_{1,2}$ are obviously necessary, otherwise no stabilizing controller exists, by Proposition 6.14. Assumptions $\mathcal{A}_{3,4}$ are technical and are required to guarantee the solvability of two involved algebraic Riccati equations. Often, although not always, they are necessary for the corresponding optimization problems to be well defined.

7.A.1 The H_2 standard problem

The standard H_2 problem for the system in Fig. 7.1 can be posed as the design on an internally stabilizing R, which minimizes $\|\mathcal{F}_l(G, R)\|_2$. Its solution is based on the following two algebraic Riccati equations:

$$A'X + XA + C_z'C_z - (XB_u + C_z'D_{zu})(D_{zu}'D_{zu})^{-1}(B_u'X + D_{zu}'C_z) = 0 \tag{7.35a}$$

and

$$AY + YA' + B_wB_w' - (YC_y' + B_wD_{yw}')(D_{yw}D_{yw}')^{-1}(C_yY + D_{yw}B_w') = 0, \tag{7.35b}$$

whose solutions X and Y are said to be stabilizing if $A + B_uK_u$ and $A + L_yC_y$ are Hurwitz, respectively, where

$$K_u := -(D_{zu}'D_{zu})^{-1}(B_u'X + D_{zu}'C_z) \quad \text{and} \quad L_y := -(YC_y' + B_wD_{yw}')(D_{yw}D_{yw}')^{-1}. \tag{7.36}$$

If $\mathcal{A}_{1\text{–}4}$ hold true, then stabilizing solutions always exist, are unique, and such that $X = X' \geq 0$ and $Y = Y' \geq 0$. Furthermore, $X > 0$ iff (A, B_u, C_z, D_{zu}) has no invariant zeros in $\mathbb{C} \setminus \bar{\mathbb{C}}_0$ and $Y > 0$ iff (A, B_w, C_y, D_{yw}) has no invariant zeros in $\mathbb{C} \setminus \bar{\mathbb{C}}_0$.

Theorem 7.2. *Let $\mathcal{A}_{1\text{–}4}$ hold true and D_{zw} be such that* $\operatorname{Im} D_{zw} \subset \operatorname{Im} D_{zu}$ *and* $\ker D_{yw} \subset \ker D_{zw}$. *If all stabilizing controllers are presented in the form*

$$R(s) = \mathcal{F}_l\left(\left[\begin{array}{c|cc} A + B_uK_u + L_yC_y + L_yD_{yu}K_u & -L_y & B_u + L_yD_{yu} \\ \hline K_u & 0 & I \\ -C_y - D_{yu}K_u & I & -D_{yu} \end{array}\right], Q(s)\right)$$

with K_u and L_y as in (7.36), *then* $\|\mathcal{F}_l(G, C)\|_2^2 = \gamma_{\text{opt}} + \|D_{zw} + D_{zu}QD_{yw}\|_2^2$, *where*

$$\gamma_{\text{opt}} := \operatorname{tr}(B_w'XB_w) + \operatorname{tr}(C_zYC_z') + \operatorname{tr}(XAY + YA'X).$$

The unique controller attaining γ_{opt} is produced by $Q = -(D_{zu}'D_{zu})^{-1}D_{zu}'D_{yw}D_{yw}'(D_{yw}D_{yw}')^{-1}$.

Some remarks are in order:

Remark 7.3 (solution properties). If $D_{zw} = 0$, then the optimal R, which corresponds to $Q = 0$, is an observer-based controller (cf. the discussion at the beginning of §6.2.3), comprised of the LQR state feedback with the gain K_u and the Kalman–Bucy filter with the gain L_y. This separation is remarkable and not quite obvious. At the same time, the optimal cost is not just a sum of the LQR (the first term in the expression for γ_{opt}) and the Kalman–Bucy (the second term) costs. It also contains the coupling term $\operatorname{tr}(XAY + YA'X)$, which might be both positive and negative, depending of properties of A. This can be explained via rewriting the optimal cost as

$$\gamma_{\text{opt}} = \operatorname{tr}(B_w'XB_w) + \operatorname{tr}(D_{zu}K_uYK_u'D_{zu}').$$

The first term above is still the LQR cost. The second term is the cost of estimating $v = D_{zu}K_ux$ from the measured y. The signal $D_{zu}K_ux$ is the contribution of the LQR control law $u = K_ux$ to the regulated signal $z = C_zx + D_{zw}w + D_{zu}u$. ▽

Remark 7.4 (AREs). The Hamiltonian matrices, see (B.9), associated with the AREs (7.35) are

$$H_X = \begin{bmatrix} A & 0 \\ -C_z'C_z & -A' \end{bmatrix} - \begin{bmatrix} B_u \\ -C_z'D_{zu} \end{bmatrix} (D_{zu}'D_{zu})^{-1} \begin{bmatrix} D_{zu}'C_z & B_u' \end{bmatrix}$$

and

$$H_Y = \begin{bmatrix} A' & 0 \\ -B_wB_w' & -A \end{bmatrix} - \begin{bmatrix} C_y' \\ -B_wD_{yw}' \end{bmatrix} (D_{yw}D_{yw}')^{-1} \begin{bmatrix} D_{yw}B_w' & C_y \end{bmatrix}.$$

Their $(1,2)$ parts (the "R" part of (B.7)) are negative semi-definite, so that Theorem B.6 applies and the connections of their solvability with $\mathcal{A}_{1\text{–}4}$ can be derived. ▽

7.A.2 The H_∞ standard problem

The standard H_∞ problem for the system in Fig. 7.1 can be posed as the design on an internally stabilizing R, which renders $\|\mathcal{F}_l(G,R)\|_\infty < \gamma$ for a given $\gamma > 0$. Its solution is based on two algebraic Riccati equations too, now of the form

$$\begin{aligned} A'X + XA + C_z'C_z - \big(X \begin{bmatrix} B_w & B_u \end{bmatrix} + C_z' \begin{bmatrix} D_{zw} & D_{zu} \end{bmatrix}\big) \\ \times \begin{bmatrix} D_{zw}'D_{zw} - \gamma^2 I & D_{zw}'D_{zu} \\ D_{zu}'D_{zw} & D_{zu}'D_{zu} \end{bmatrix}^{-1} \left(\begin{bmatrix} B_w' \\ B_u' \end{bmatrix} X + \begin{bmatrix} D_{zw}' \\ D_{zu}' \end{bmatrix} C_z \right) = 0 \end{aligned} \tag{7.37a}$$

and

$$\begin{aligned} AY + YA' + B_wB_w' - \big(Y \begin{bmatrix} C_z' & C_y' \end{bmatrix} + B_w \begin{bmatrix} D_{zw}' & D_{yw}' \end{bmatrix}\big) \\ \times \begin{bmatrix} D_{zw}D_{zw}' - \gamma^2 I & D_{zw}D_{yw}' \\ D_{yw}D_{zw}' & D_{yw}D_{yw}' \end{bmatrix}^{-1} \left(\begin{bmatrix} C_z \\ C_y \end{bmatrix} Y + \begin{bmatrix} D_{zw} \\ D_{yw} \end{bmatrix} B_w' \right) = 0, \end{aligned} \tag{7.37b}$$

whose solutions are said to be stabilizing if $A + B_wK_w + B_uK_u$ and $A + L_zC_z + L_yC_y$ are Hurwitz, where

$$\begin{bmatrix} K_w \\ K_u \end{bmatrix} := -\begin{bmatrix} D_{zw}'D_{zw} - \gamma^2 I & D_{zw}'D_{zu} \\ D_{zu}'D_{zw} & D_{zu}'D_{zu} \end{bmatrix}^{-1} \left(\begin{bmatrix} B_w' \\ B_u' \end{bmatrix} X + \begin{bmatrix} D_{zw}' \\ D_{zu}' \end{bmatrix} C_z \right) \tag{7.38a}$$

and

$$\begin{bmatrix} L_z & L_y \end{bmatrix} := -\big(Y \begin{bmatrix} C_z' & C_y' \end{bmatrix} + B_w \begin{bmatrix} D_{zw}' & D_{yw}' \end{bmatrix}\big) \begin{bmatrix} D_{zw}D_{zw}' - \gamma^2 I & D_{zw}D_{yw}' \\ D_{yw}D_{zw}' & D_{yw}D_{yw}' \end{bmatrix}^{-1}. \tag{7.38b}$$

Assumptions $\mathcal{A}_{1\text{–}4}$ are necessary for the solvability of these equations, but might not be sufficient if γ is not large enough. Moreover, even if AREs (7.37) do admit stabilizing solutions, these solutions might not be positive semi-definite, also depending on γ. At the same time, null spaces of X and Y do not depend on γ. We still have that $\det(X) \neq 0$ iff (A, B_u, C_z, D_{zu}) has no invariant zeros in $\mathbb{C} \setminus \bar{\mathbb{C}}_0$ and $\det(Y) \neq 0$ iff (A, B_w, C_y, D_{yw}) has no invariant zeros in $\mathbb{C} \setminus \bar{\mathbb{C}}_0$. Also, the AREs in (7.37) reduce to the corresponding H_2 AREs in (7.35) as $\gamma \to \infty$.

Theorem 7.3. *If $\mathcal{A}_{1\text{–}4}$ hold true, then the standard H_∞ problem is solvable iff*

(a) $\max\{\|(I - D_{zu}(D_{zu}'D_{zu})^{-1}D_{zu}')D_{zw}\|, \|D_{zw}(I - D_{yw}'(D_{yw}D_{yw}')^{-1}D_{yw})\|\} < \gamma$,

(b) there is a stabilizing solution X to ARE (7.37a) *such that $X = X' \geq 0$,*

(c) there is a stabilizing solution Y to ARE (7.37b) *such that $Y = Y' \geq 0$,*

(d) $\rho(XY) < \gamma^2$.

In this case $Z_\gamma := (I - \gamma^{-2}YX)^{-1}$ *is well defined and, denoting* $\tilde{B}_u := B_u + L_z D_{zu} + L_y D_{yu}$ *and* $\tilde{C}_y := C_y + D_{yw}K_w + D_{yu}K_u$*, all* γ*-suboptimal controllers are given by*

$$R(s) = \mathcal{F}_l\left(\left[\begin{array}{c|cc} A + B_w K_w + B_u K_u + Z_\gamma L_y \tilde{C}_y & -Z_\gamma L_y & Z_\gamma \tilde{B}_u \\ \hline K_u & 0 & I \\ -\tilde{C}_y & I & -D_{yu} \end{array}\right], Q(s)\right)$$
$$= \mathcal{F}_l\left(\left[\begin{array}{c|cc} A + L_z C_z + L_y C_y + \tilde{B}_u K_u Z_\gamma & -L_y & \tilde{B}_u \\ \hline K_u Z_\gamma & 0 & I \\ -\tilde{C}_y Z_\gamma & I & -D_{yu} \end{array}\right], Q(s)\right)$$

for any $Q \in H_\infty$ *such that* $\|D_{zw} + D_{zu}QD_{yw}\|_\infty < \gamma$.

Some remarks are in order:

Remark 7.5 (solution properties). Although this fact is less evident than in the H_2 case, the central suboptimal controller, that corresponding to $Q = 0$, is also observer based. The resulting control signal is also the H_∞ suboptimal estimate of the signal $v = D_{zu}u_{\text{fi}}$, where

$$u_{\text{fi}} := -(D_{zu}'D_{zu})^{-1}D_{zu}'D_{zw}(w - K_w x) + K_u x = -(D_{zu}'D_{zu})^{-1}(D_{zu}'D_{zw}w + (B_u'X + D_{zu}'C_z)x)$$

and $u = u_{\text{fi}}$ would attain the performance level γ if both the exogenous input w and the plant state x were measurable (known as the *full-information* problem, requires that $\|(I - D_{zu}(D_{zu}'D_{zu})^{-1}D_{zu}')D_{zw}\| < \gamma$ and condition (b) of Theorem 7.3 holds). As a matter of fact, $w_{\text{worst}} = K_w x$ is the worst-case exogenous input, which is the most problematic from the H_∞ performance viewpoint in full-information control. In contrast to the H_2 (Kalman–Bucy) case, parameters of the H_∞ estimator do depend on the signal it estimates, so the formulae are more involved and nontrivial transformations are required to decouple the estimator ARE, which depends on the state-feedback gain, from K_u. The coupling condition (d) of Theorem 7.3 is actually a remnant of this procedure. ▽

Remark 7.6 (what if $\|D_{zw}\| < \gamma$). The existence of Q such that $\|D_{zw} + D_{zu}QD_{yw}\|_\infty < \gamma$ is guaranteed by condition (a) in Theorem 7.3 (follows by Parrott's theorem [20]). If $\|D_{zw}\| < \gamma$, then condition (a) holds regardless D_{zu} and D_{yw} and the constraint on Q is simplified, rendering $Q = 0$ feasible. Moreover, if we replace $-D_{yu}$ with $-D_{yu} - D_{yw}D_{zw}'(\gamma^2 I - D_{zw}D_{zw}')^{-1}D_{zu}$ in the $(2, 2)$ feedthrough term of the generator of all γ-suboptimal controllers, then it can be shown that Q must satisfy $\|S_u Q S_y\|_\infty < \gamma$, where S_u and S_y are any matrices satisfying

$$S_u'S_u = D_{zu}'(I - \gamma^{-2}D_{zw}D_{zw}')^{-1}D_{zu} \quad \text{and} \quad S_y S_y' = D_{yw}(I - \gamma^{-2}D_{zw}'D_{zw})^{-1}D_{yw}'$$

(they are both nonsingular). ▽

Remark 7.7 (AREs). The Hamiltonian matrices, see (B.9), associated with the AREs (7.37) are

$$H_X = \begin{bmatrix} A & 0 \\ -C_z'C_z & -A' \end{bmatrix} - \begin{bmatrix} B_w & B_u \\ -C_z'D_{zw} & -C_z'D_{zu} \end{bmatrix} \begin{bmatrix} D_{zw}'D_{zw} - \gamma^2 I & D_{zw}'D_{zu} \\ D_{zu}'D_{zw} & D_{zu}'D_{zu} \end{bmatrix}^{-1} \begin{bmatrix} D_{zw}'C_z & B_w' \\ D_{zu}'C_z & B_u' \end{bmatrix}$$

and

$$H_Y = \begin{bmatrix} A' & 0 \\ -B_w B_w' & -A \end{bmatrix} - \begin{bmatrix} C_z' & C_y' \\ -B_w D_{zw}' & -B_w D_{yw}' \end{bmatrix} \begin{bmatrix} D_{zw}D_{zw}' - \gamma^2 I & D_{zw}D_{yw}' \\ D_{yw}D_{zw}' & D_{yw}D_{yw}' \end{bmatrix}^{-1} \begin{bmatrix} D_{zw}B_w' & C_z \\ D_{yw}B_w' & C_y \end{bmatrix}$$

and their $(1, 2)$ parts are sign definite only for $\gamma \to \infty$, in general. Hence, the results of Theorem B.6 can no longer be used. As a matter of fact, as γ decreases, the stabilizing solutions of H_∞ AREs normally increase, to the point where one or several eigenvalues change their sign via infinity. Thus, the coupling condition typically violates first. Still, it might be advantageous to work with the pseudo-inverses of X and Y. The null spaces of X and Y do not depend on γ and the eigenvalues of their pseudo-inverses change sign via zero crossings as γ decreases. ▽

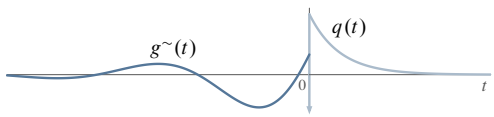


Fig. 7.6: Nehari extension of $G^{\sim}(s) = [(s/3+1)/(s^2+2s/3+1)]^{\sim}$, the arrow represents the Dirac δ

7.A.3 The Nehari extension problem

The last problem considered in this Appendix is actually not a standard problem. Nevertheless, it plays an important role in solving the H_∞ standard problem and is of interest by itself.

Let $G \in RH_\infty$ be strictly proper and given by its stable realization $G(s) = C(sI - A)^{-1}B$ (i.e. A is Hurwitz). The *Nehari extension problem* (or just the Nehari problem) is the problem of finding the closest extension of the stable and anti-causal adjoint G' of this G to the class of stable and causal systems, see Fig. 7.6. The metric used is the $L_\infty(\mathrm{j}\mathbb{R})$ system norm defined by (3.21) on p. 48, which makes the problem nontrivial (the solution in the $L_2(\mathrm{j}\mathbb{R})$ metric would be zero). In formal terms, it is the problem of finding

$$\gamma_{\mathrm{N}} := \inf_{Q \in RH_\infty} \|G^{\sim} + Q\|_\infty,$$

because $G^{\sim}(s)$ is the transfer function of the adjoint G'.

To formulate the solution to the Nehari problem, introduce the controllability and observability Gramians of G, W_{c} and W_{o}, respectively, satisfying the Lyapunov equations

$$AW_{\mathrm{c}} + W_{\mathrm{c}}A' + BB' = 0 \quad \text{and} \quad W_{\mathrm{o}}A + A'W_{\mathrm{o}} + C'C = 0$$

(cf. (4.10) and (4.15)). The following result gives both the distance and all "sufficiently close" RH_∞ functions to $G^{\sim}$.

Theorem 7.4. *$\gamma_N = \|G\|_H = \sqrt{\rho(W_c W_o)}$. Given $\gamma > \gamma_N$, all $Q \in RH_\infty$ such that $\|G^{\sim} + Q\|_\infty \le \gamma$ are given by*

$$Q(s) = \mathcal{F}_l\left(\left[\begin{array}{c|cc} A - V_\gamma W_c C'C & V_\gamma W_c C' & V_\gamma B \\ \hline -B'W_o & 0 & I \\ -C & I & 0 \end{array}\right], \tilde{Q}(s)\right) = \mathcal{F}_l\left(\left[\begin{array}{c|cc} A - BB'W_o V_\gamma & W_c C' & B \\ \hline -B'W_o V_\gamma & 0 & I \\ -CV_\gamma & I & 0 \end{array}\right], \tilde{Q}(s)\right)$$

for an arbitrary $\tilde{Q} \in RH_\infty$ such that $\|\tilde{Q}\|_\infty \le \gamma$, where $V_\gamma := (\gamma^2 I - W_c W_o)^{-1}$ is well defined.

Proof that $\gamma_N \ge \|G\|_H$ (for it's fun). Because $\|G^{\sim} + Q\|_\infty = \|G + Q^{\sim}\|_\infty$, we may look for the distance from $L_\infty(\mathrm{j}\mathbb{R}) \setminus H_\infty$ to a given $G \in RH_\infty$, which is more convenient for the notational conventions adopted in the notes. For any signal $v \in L_{2-}$ we have that $Q'v \in L_{2-}$ too. Hence,

$$((G + Q')v)_+ = (Gv)_+$$

for all such Q's, where x_+ denotes the orthogonal projection of $x \in L_2$ onto L_{2+} (i.e. $x_+(t) = x(t)$ for $t \ge 0$ and $x_+(t) = 0$ for $t < 0$). Taking into account that the $L_\infty(\mathrm{j}\mathbb{R})$ norm of the frequency response of a system is the induced norm of the corresponding operator $L_2 \to L_2$, we have that

$$\begin{aligned} \|G + Q^{\sim}\|_\infty &= \sup_{v \in L_2, \|v\|_2 = 1} \|(G + Q')v\|_2 \ge \sup_{v \in L_{2-}, \|v\|_2 = 1} \|(G + Q')v\|_2 \\ &\ge \sup_{v \in L_{2-}, \|v\|_2 = 1} \|((G + Q')v)_+\|_2 = \sup_{v \in L_{2-}, \|v\|_2 = 1} \|(Gv)_+\|_2 = \|G\|_{\mathrm{H}}, \end{aligned}$$

see (B.5) on p. 190. The Gramian expression follows then by Proposition B.4. Hence, $\gamma_{\mathrm{N}} \ge \|G\|_{\mathrm{H}}$.

The proof that this bound is tight and the derivation of the parametrization are way more technical and thus omitted. □

Note that as $\gamma \downarrow \gamma_{\mathrm{N}}$, the matrix V_γ becomes singular. Still, the generator of all solutions in Theorem 7.4 is well defined and only loses at least one of its poles. For example, consider the optimal extension problem for the second-order $G(s) = (s/3 + 1)/(s^2 + 2s/3 + 1)$. The impulse response of its adjoint is shown in Fig. 7.6 by the dark line. In this case $\gamma_{\mathrm{N}} = 1$ and the optimal extension is unique and has the first-order transfer function

$$Q(s) = -\frac{s}{s+1}.$$

Its impulse response, $q(t) = -\delta(t) + \mathrm{e}^{-t}\mathbb{1}(t)$, is shown in Fig. 7.6 by the pale line.

Chapter 8

Model Uncertainty and Robustness

NOBODY'S PERFECT and models of controlled processes are not exceptions. As was already discussed in Section 1.2, mathematical models are merely approximations, more or less accurate, of described phenomena. Any meaningful control analysis should take this fundamental fact into account. This is particularly important in feedback control, because feedback can both alleviate effects of modeling inaccuracies and aggravate them, to the level of losing stability. One day, this chapter will present basic ideas on describing model uncertainty and coping with their effect on system stability and performance.

Chapter 9

H_∞ Loop Shaping Design Method

Loop shaping is a conceptually simple yet powerful design method in the frequency domain. The idea (see §1.4.3) is to express closed-loop design objectives in terms of requirements on the open-loop transfer function (loop gain), which is a linear function of the controller and independent of involved signals and loop components. The controller is then designed to shape the loop gain differently in different frequency ranges. However, the constraints of the loop phase near crossover frequency (stability and stability margins requirements) complicate the loop shaping procedure considerably, especially for systems with right half-plane poles and zeros and in the MIMO case. This chapter studies an optimization-based twist on this theme, dubbed H_∞ loop shaping. The approach, put forward by McFarlane and Glover [17], follows classical loop shaping guidelines in the choice of the control objectives and casts the stability and "far from the critical point" requirements as a special H_∞ optimization problem. This yields a relatively simple design method, capable of guaranteeing certain important characteristics (stability, robustness, et cetera) and intuitive to adjust, with not too many tuning parameters.

9.1 The setup and loop-shaping guidelines

Consider the block-diagram[1] in Fig. 9.1 on the next page. It represents the so-called 2-degrees-of-freedom (2DOF) control architecture originated in [16], which combines open- and closed-loop control configurations presented in Fig. 1.4(b) and 1.4(c), respectively. The $p \times m$ plant model $P = NM^{-1}$, where N and M are its right coprime factors, is supposed to be given and $m \times p$ systems R and F, with $F \in H_\infty$, are design parameters (controllers). We also assume hereafter that $\operatorname{nrank}(P(s)) = p$, i.e. that the plant is *not underactuated*. It is readily seen that the controlled signals y and u in this setup equal

$$\begin{bmatrix} y \\ u \end{bmatrix} = \begin{bmatrix} N \\ M \end{bmatrix} F y_\mathrm{r} + \begin{bmatrix} S_\mathrm{o} & T_\mathrm{d} \\ T_\mathrm{c} & T_\mathrm{i} \end{bmatrix} \begin{bmatrix} d_\mathrm{o} \\ d_\mathrm{i} \end{bmatrix} + \begin{bmatrix} T_\mathrm{o} \\ T_\mathrm{c} \end{bmatrix} n \tag{9.1}$$

for

$$\begin{bmatrix} S_\mathrm{o} & T_\mathrm{d} \\ T_\mathrm{c} & T_\mathrm{i} \end{bmatrix} := \begin{bmatrix} I \\ R \end{bmatrix} (I - PR)^{-1} \begin{bmatrix} I & P \end{bmatrix} \quad \text{and} \quad T_\mathrm{o} := (I - PR)^{-1} PR = S_\mathrm{o} - I$$

(cf. (1.19) on p. 10). A remarkable property of this relation that the effect of the (measured) reference signal y_r on the system behavior depends only on the feedforward component of the controller, F, whereas effects of all other (unmeasured) signals—only on the feedback component of the controller, R. This explains the term and, more importantly, makes it possible to address tracking requirements separately from those of disturbance attenuation (and robustness). The former is solved in open loop and is thus less sensitive to the

[1] The positive feedback form is chosen solely for aesthetic reasons, negative feedback corresponds to changing the sign of R.

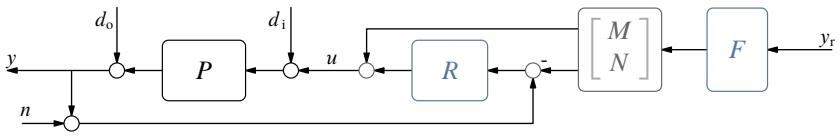


Fig. 9.1: 2-degrees-of-freedom control setup

approach taken to choose a stable F. Selecting R is a closed-loop design problem, which renders it more complicated and potentially hazardous, as an inapt design might give rise to a substantial performance deterioration or even to instability. For this reason, in this chapter we concentrate on the design of R and disregard reference tracking properties of the system.

Remark 9.1 (tracking and modeling uncertainty). The complete separation between the tracking and feedback properties holds only under the perfect match between the plant and the systems M and N used to process the reference signal in Fig. 9.1. If this is not the case, i.e. if $P \neq NM^{-1}$, then instead of (9.1) we have

$$\begin{bmatrix} y \\ u \end{bmatrix} = \begin{bmatrix} N \\ M \end{bmatrix} F y_r + \begin{bmatrix} S_o & T_d \\ T_c & T_i \end{bmatrix} \left(\begin{bmatrix} -N \\ M \end{bmatrix} F y_r + \begin{bmatrix} d_o \\ d_i \end{bmatrix} \right) + \begin{bmatrix} T_o \\ T_c \end{bmatrix} n, \tag{9.1$'$}$$

where the system

$$\begin{bmatrix} S_o & T_d \\ T_c & T_i \end{bmatrix} \begin{bmatrix} -N \\ M \end{bmatrix} F = \begin{bmatrix} I \\ R \end{bmatrix} (I - PR)^{-1}(PM - N)F$$

reflects the modeling mismatch. If a "size" of $(I - PR)^{-1}(PM - N)F$ is small, which depends on the robustness properties of the closed-loop system, we may expect that tracking performance is still largely separated from other properties of the system. ▽

Disturbance attenuation and low noise sensitivity performance can be expressed, somewhat simplistically, as the following frequency-dependent requirements:

- $\|S_o(j\omega)\| \ll 1$ at frequencies where the spectrum of d_o is concentrated (typically, low frequencies),
- $\|S_d(j\omega)\| \ll 1$ at frequencies where the spectrum of d_i is concentrated (typically, all frequencies),
- $\|T_o(j\omega)\| \ll 1$ at frequencies where the spectrum of n is concentrated (typically, high frequencies).

It is implicitly assumed hereafter that various components of the disturbance signals are normalized, so we shall expect them to have roughly equal intensities. In addition, it is normally required to

- keep $\|T_c(j\omega)\|$ and $\|T_i(j\omega)\|$ not too large at all $\omega \in \mathbb{R}$,

where the proportions are determined by the scaling of the control signal, and

- avoid sharp resonance peaks in closed-loop frequency responses

to prevent oscillating transients in the closed-loop system. In the classical SISO loop shaping, the first three of these requirements are translated to those on the gain of the open-loop frequency response, the fourth requirement is affected via adjusting the crossover frequency, and the fifth requirement is expressed as the "far from the critical point" endeavor and quantified by stability margins. In the remainder of this section conceptually straightforward MIMO generalizations of the gain requirements are discussed.

Define the $p \times p$ *output loop transfer function*

$$L_o(s) := P(s)R(s),$$

which is obtained by breaking the feedback loop in Fig. 9.1 at the plant output. Clearly, $S_o = (I - L_o)^{-1}$ and $T_o = (I - L_o)^{-1}L_o$ then. Because $\|M\| = \overline{\sigma}(M) = 1/\underline{\sigma}(M^{-1})$ for any nonsingular M, we have that

$\|S_o(j\omega)\| = 1/\underline{\sigma}(I - L_o(j\omega))$. Also, given a nonsingular M such that $I - M$ is nonsingular too, it follows by Proposition 2.4 on p. 34 and the reverse triangle inequality that

$$\underline{\sigma}(I - M) = \min_{\|u\|=1} \|(I - M)u\| \geq \min_{\|u\|=1} |\|Mu\| - \|u\|| = |\underline{\sigma}(M) - 1|.$$

Hence,

$$\|S_o(j\omega)\| \leq \frac{1}{\underline{\sigma}(L_o(j\omega)) - 1}$$

provided $\underline{\sigma}(L_o(j\omega)) > 1$. This, in turn, implies that $\|S_o(j\omega)\| \ll 1$ whenever $\underline{\sigma}(L_o(j\omega)) \gg 1$, i.e. the lowest loop gain is sufficiently high.

Next,

$$\|T_d(j\omega)\| \leq \|S_o(j\omega)\|\|P(j\omega)\| \leq \frac{\overline{\sigma}(P(j\omega))}{\underline{\sigma}(L_o(j\omega)) - 1}.$$

Hence, a high loop gain also helps to attenuate input (load) disturbances. Certain care should be taken at frequencies, where the plant itself has a high gain, which is the case for lightly-damped systems, for instance. In such situations, extra requirements on the loop gain should normally be imposed. At the same time, load disturbances may be attenuated even under a low loop gain at frequencies where the plant gain is low itself. Thus, the requirement $\underline{\sigma}(L_o(j\omega)) \gg 1$ in the context of load disturbance attenuation should only be considered in the intersection of the spectrum of d_i and the bandwidth of the plant, which are typically low frequencies.

Increasing the loop gain is not helpful for reducing $\|T_o(j\omega)\|$ though. Indeed, if $\underline{\sigma}(L_o(j\omega)) > 0$, then

$$\|T_o(j\omega)\| = \|(I - L_o^{-1}(j\omega))^{-1}\| = \frac{1}{\underline{\sigma}(I - L_o^{-1}(j\omega))} \geq \frac{1}{\underline{\sigma}(L_o^{-1}(j\omega)) + 1} = \frac{1}{1 + 1/\overline{\sigma}(L_o(j\omega))}$$
$$\geq \frac{1}{1 + 1/\underline{\sigma}(L_o(j\omega))}$$

and an increase of $\underline{\sigma}(L_o(j\omega))$ raises this lower bound. What can help is a decrease of the loop gain. To see that, note that

$$\|T_o(j\omega)\| \leq \|(I - L_o(j\omega))^{-1}\|\|L_o(j\omega)\| = \frac{\overline{\sigma}(L_o(j\omega))}{\underline{\sigma}(I - L_o(j\omega))} \leq \frac{\overline{\sigma}(L_o(j\omega))}{1 - \underline{\sigma}(L_o(j\omega))} \leq \frac{\overline{\sigma}(L_o(j\omega))}{1 - \overline{\sigma}(L_o(j\omega))},$$

provided $\overline{\sigma}(L_o(j\omega)) < 1$. Thus, $\|T_o(j\omega)\| \ll 1$ whenever $\overline{\sigma}(L_o(j\omega)) \ll 1$, i.e. the highest loop gain is sufficiently low.

Summarizing the arguments above, the first three requirements to the closed-loop frequency responses can be translated to the following requirements to the loop gains:

- $\underline{\sigma}(L_o(j\omega)) \gg 1$ at frequencies where the spectrum of d_o is concentrated (typically, low frequencies) and where the spectrum of d_i is concentrated within the bandwidth of P (also low frequencies),
- $\overline{\sigma}(L_o(j\omega)) \ll 1$ at frequencies where the spectrum of n is concentrated (typically, high frequencies).

These are MIMO counterparts of the standard SISO loop gain shaping guidelines briefly outlined in §1.4.3. They are relatively straightforward to grasp and not overly difficult to attain, much like what we know in the SISO loop-shaping design.

Connections between the control effort and the crossover frequency, and even the definition of the latter, are less transparent in the MIMO case. Still, these connections are not overly quantitative in the SISO case either and the underlying idea remains roughly the same. Namely, any increase of the frequency band in which the loop gain is high beyond the plant bandwidth requires higher control efforts. The extension of the "far from the critical point" requirement to MIMO loops, where the very notion of phase is not well defined, is less obvious and lies at the core of the H_∞ loop shaping approach.

9.2 Principles of H_∞ loop shaping

There are many really good reasons to keep the Nyquist plot of the loop far from the critical point $(-1, 0)$ (equivalently, the Nichols plot far from the critical points $(180 \pmod{360}, 0)$). One of them is particularly relevant in the context of extending the idea to MIMO loops. It is the need to render the feedback system less sensitive to model inaccuracies. Indeed, if the Nyquist plot is too close to the critical point, small alterations to the loop frequency response might lead to a change in its winding number around the critical point, which, in turn, implies that the closed-loop system becomes unstable. Thus, the distance of the Nyquist plot, in whatever meaningful sense, from the critical point is essentially a robustness indicator. But robustness, as should be discussed in Chapter 8, is a property extendible to MIMO systems, in fact in an analytic manner.

This observation is in the core of the H_∞ loop shifting approach. Its essence is to split the problem into two phases. In the first, the magnitude of the loop is shaped using weighting functions placed in series with the plant. In the second phase, a controller is designed to maximize an appropriately chosen robustness measure, which is used as a success indicator. If the second phase is successful, the weights are added to the controller, also in series. If not, weights are reselected and the procedure repeats. A more detailed exposition of these phases is described below.

1. Let W_o and W_i be weighting functions such that the magnitude frequency response of $P_\mathrm{msh} := W_\mathrm{o} P W_\mathrm{i}$ reflects our requirements to the magnitude frequency response of the loop and its crossover frequency. In the SISO case, we can keep either one of these weights identity without loss of generality. In the MIMO case, there might be situations when different input or output channels should be scaled differently, in which situations we need both W_o and W_i. Also, it might be conceptually convenient to use W_i to shape low-frequency properties and W_o to shape high-frequency properties, just because the former will then process the control signals, while the latter—measured signals. Technically, the only limitation on the weights is that the cascade $P_\mathrm{msh}(s)$ has no unstable cancellation and is proper. Thus, we may consider the unstable PID weight $W_\mathrm{i}(s) = k_\mathrm{p}(1 + k_\mathrm{i}/s + \tau_\mathrm{d} s)$ or suchlike and a low-pass filter as W_o. The choice of the weights is thus technically simple.

2. This phase consists in formulating and solving an H_∞ robust stabilization problem to ensure the closed-loop stability and robustness of the resulting system. Success is "measured" by the reciprocal of the attainable H_∞-norm of the corresponding standard problem (although should be always validated by analyzing the resulted closed-loop system). If this phase is deemed successful, i.e. if the success indicator is sufficiently large, the (sub)optimal controller, say R_rob, forms the basis of the resulting controller for our problem. Specifically, the final controller is $R = W_\mathrm{i} R_\mathrm{rob} W_\mathrm{o}$. This choice renders $L_\mathrm{o} = P W_\mathrm{i} R_\mathrm{rob} W_\mathrm{o}$, which is, of course, different from the intended $W_\mathrm{o} P W_\mathrm{i}$ (and even from the designed $W_\mathrm{o} P W_\mathrm{i} R_\mathrm{rob}$, unless W_o commutes with the rest). Still, the resulting closed-loop system is always stable and we may expect that if the robust stabilization is successful, then R_rob alters the loop gain mainly in the crossover region, which is its role in the whole process.

It should be emphasized that the whole procedure depends heavily on the robust stabilization problem of choice in the second phase of the design. Being based on the H_∞ optimization techniques, the outcome of this problem possesses all traits of optimization-based solutions discussed in §7.3.4. On the credit side, we can be sure that there is no other controller rendering the system "more robust" with respect to the chosen measure. On the debit side, the optimal solution might exploit loopholes in the problem formulation to produce poor controllers. This means that the choice of the robust stability problem is of a major importance in the success of the whole procedure.

To understand the choice of the optimization problem in the H_∞ loop-shaping procedure suggested by McFarlane and Glover, note that the H_∞ norm of essentially every closed-loop system can be interpreted in terms of an appropriately defined robustness measure. We already saw in §7.3.1 that the sensitivity

function is connected with the modulus margin $\mu_{\rm m}$. In more abstract terms, it is known that

- $\|S_{\rm i}\|_\infty$ quantifies robustness in the inverse input multiplicative uncertainty model $P(I+\Delta)^{-1}$,
- $\|S_{\rm o}\|_\infty$ quantifies robustness in the inverse output multiplicative uncertainty model $(I+\Delta)^{-1}P$,
- $\|T_{\rm i}\|_\infty$ quantifies robustness in the input multiplicative uncertainty model $P(I+\Delta)$,
- $\|T_{\rm o}\|_\infty$ quantifies robustness in the output multiplicative uncertainty model $(I+\Delta)P$,
- $\|T_{\rm c}\|_\infty$ quantifies robustness in the additive uncertainty model $P+\Delta$,
- $\|T_{\rm d}\|_\infty$ quantifies robustness in the inverse additive uncertainty model $(I+P\Delta)^{-1}P$

(the meaning of this quantification is that the closed-loop system is stable for all $\Delta \in H_\infty$ and such that $\|\Delta\|_\infty \le \alpha$ iff the H_∞-norm of the corresponding transfer function is smaller than $1/\alpha$, this can be proved via the small gain theorem, see Theorem 6.1). The uncertainty model is normally recommended to be chosen to reflect dominant uncertainty sources in the real plant. But this is not a right criterion for loop shaping. Rather, the choice should result in a balanced optimization problem, where "dirty tricks" of canceling dominant lightly damped or slow plant poles and zeros do not pay off.

The sought balancing can be achieved if all four closed-loop transfer functions shaping the response to $d_{\rm i}$ and $d_{\rm o}$ in (9.1), aka the Gang of Four, are included in the cost with equal weights. Specifically, the problem solved in the second phase is

$$\underset{R_{\rm rob}}{\text{minimize}} \left\| \begin{bmatrix} R_{\rm rob} \\ I \end{bmatrix} (I - P_{\rm msh}R_{\rm rob})^{-1} \begin{bmatrix} I & P_{\rm msh} \end{bmatrix} \right\|_\infty, \tag{9.2}$$

which can be named the *balanced sensitivity* problem. This optimization is not expected to encourage stable yet bad pole-zero cancellations between $P_{\rm msh}(s)$ and $R_{\rm rob}(s)$. This is because whatever stable poles or zeros of the plant are canceled by the controller, they still show up as poles in at least one of the block entries in the cost. This problem has a concrete robustness measure associated with it too. This is the robustness with respect to uncertainty in the so-called normalized coprime factorization of the plant, in the form $(\tilde{M}_{\rm msh}+\tilde{\Delta}_M)^{-1}(\tilde{N}_{\rm msh}+\tilde{\Delta}_N)$ for $\|\begin{bmatrix} \tilde{\Delta}_M & \tilde{\Delta}_N \end{bmatrix}\|_\infty \le \alpha$ and with an appropriate scaling on the left coprime factors $\tilde{M}_{\rm msh}$ and $\tilde{N}_{\rm msh}$ of $P_{\rm msh}$, see Lemma 9.1 on p. 173. This robust stability setup is not likely to reflect any practical situation, but it has good system-theoretical justifications being associated with the gap metric [28]. Incidentally, and atypically for H_∞ problems, the solution to (9.2) is quite simple as well. In particular, the optimal attainable performance level, say $\gamma_{\rm min}$, can be obtained analytically, rather than derived via iterations. Technical details are discussed in Section 9.A.

With the attainable performance level $\gamma_{\rm min}$ of (9.2) in hand, the success indicator in the second phase of the H_∞ loop shaping procedure is chosen as the smallest destabilizing uncertainty size, which is

$$\epsilon_{\rm max} = \frac{1}{\gamma_{\rm min}} \in (0,1) \tag{9.3}$$

(as a matter of fact, $\epsilon_{\rm max}^2 = 1 - \|\begin{bmatrix} \tilde{N}_{\rm msh} & \tilde{M}_{\rm msh} \end{bmatrix}\|_{\rm H}$ for a special, normalized, *lcf* of $P_{\rm msh}$). A "sufficiently large" $\epsilon_{\rm max}$ should indicate that the magnitude shape introduced in the first phase does not conflict with the crossover region requirements too much. If $\epsilon_{\rm max} \ll 1$, we should be alarmed and reconsider the weights selected in the first phase; there is a good chance that the associated loop gain requirements are not realistic and some relaxations, be it with respect to gain requirements or the crossover, are needed. The question of what indicator to regard as "sufficiently large" is open to interpretation. The level $\epsilon_{\rm max} \ge 0.25$ may perhaps be regarded as adequate in many situations. In some cases a higher robustness level can be achieved. For example, if $P_{\rm msh}(s)$ is positive real, see §6.1.3, then $\epsilon_{\rm max} \ge \sqrt{0.5} \approx 0.707$, which confirms yet again that systems with positive-real transfer functions are easy to control. It is not against reason to regard $\epsilon_{\rm max} < 0.1$ as inadequate, although even then the resulting loop should be inspected.

9.3 Design case studies

Below we consider several academic SISO examples to illustrate traits of the H_∞ loop shaping methodology. All results are derived with the formulae of Section 9.A at the end of this chapter.

9.3.1 Double integrator

We start with a couple of examples where an analytic solution can be derived. Consider first the H_∞ loop shaping procedure for the double integrator $P(s) = 1/s^2$ and the static weight $W(s) = \tilde{\omega}_c^2$, where $\tilde{\omega}_c > 0$ is an intended crossover frequency. With this choice

$$P_{\text{msh}}(s) = \frac{\tilde{\omega}_c^2}{s^2} = \left[\begin{array}{cc|c} 0 & 1 & 0 \\ 0 & 0 & \tilde{\omega}_c \\ \hline \tilde{\omega}_c & 0 & 0 \end{array}\right]$$

indeed has its crossover frequency $\omega_c = \tilde{\omega}_c$.

To solve now the balanced sensitivity problem (9.2), we need AREs (9.12). The control Riccati (9.12a) in this case reads

$$\begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} X + X \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} - X \begin{bmatrix} 0 & 0 \\ 0 & \tilde{\omega}_c^2 \end{bmatrix} X + \begin{bmatrix} \tilde{\omega}_c^2 & 0 \\ 0 & 0 \end{bmatrix} = 0$$

with the stabilizing solution

$$X = \begin{bmatrix} \sqrt{2}\,\tilde{\omega}_c & 1 \\ 1 & \sqrt{2}/\tilde{\omega}_c \end{bmatrix} > 0.$$

Likewise, the filtering ARE (9.12b) reads

$$\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} Y + Y \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} - Y \begin{bmatrix} \tilde{\omega}_c^2 & 0 \\ 0 & 0 \end{bmatrix} Y + \begin{bmatrix} 0 & 0 \\ 0 & \tilde{\omega}_c^2 \end{bmatrix} = 0$$

and its stabilizing solution

$$Y = \begin{bmatrix} \sqrt{2}/\tilde{\omega}_c & 1 \\ 1 & \sqrt{2}\,\tilde{\omega}_c \end{bmatrix} > 0.$$

Thus,

$$\gamma_{\min}^2 = 1 + \rho\left(\begin{bmatrix} \sqrt{2}/\tilde{\omega}_c & 1 \\ 1 & \sqrt{2}\,\tilde{\omega}_c \end{bmatrix}\begin{bmatrix} \sqrt{2}\,\tilde{\omega}_c & 1 \\ 1 & \sqrt{2}/\tilde{\omega}_c \end{bmatrix}\right) = 1 + \rho\left(\begin{bmatrix} 3 & 2\sqrt{2}/\tilde{\omega}_c \\ 2\sqrt{2}\,\tilde{\omega}_c & 3 \end{bmatrix}\right) = 4 + 2\sqrt{2}$$

and, by (9.3)

$$\epsilon_{\max} = \sqrt{\frac{1}{2} - \frac{\sqrt{2}}{4}} \approx 0.3827, \tag{9.4}$$

which is reasonably high. Remarkably, it does not depend on $\tilde{\omega}_c$, which can be explained by the fact that the phase of $P_{\text{msh}}(\mathrm{j}\omega)$ does not depend on ω either.

We then calculate the controller corresponding to $\gamma = \epsilon_{\max}$. To this end, we need to calculate the matrix Z_γ defined in Theorem 9.2 at the optimal performance

$$Z_{\gamma_{\min}} = \frac{2+\sqrt{2}}{4}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \frac{2-\sqrt{2}}{4}\begin{bmatrix} 3 & 2\sqrt{2}/\tilde{\omega}_c \\ 2\sqrt{2}\,\tilde{\omega}_c & 3 \end{bmatrix} = (\sqrt{2}-1)\begin{bmatrix} 1 & -1/\tilde{\omega}_c \\ -\tilde{\omega}_c & 1 \end{bmatrix}$$

Substituting this matrix into (9.14′), we obtain the first-order transfer function

$$R_{\text{rob}}(s) = -\frac{(1+\sqrt{2})s + \tilde{\omega}_c}{s + \tilde{\omega}_c(1+\sqrt{2})} \tag{9.5}$$

(mind that feedback is positive). This is the classical lead controller tuned for the crossover frequency $\tilde{\omega}_c$, which provides a phase lead of 45° at this frequency and has $|R_{rob}(\infty)/R_{rob}(0)| = 3 + 2\sqrt{2} \approx 15.3\,\text{dB}$.

After moving the weight to the controller, we end up with

$$R(s) = -\tilde{\omega}_c^2 \frac{(1+\sqrt{2})s + \tilde{\omega}_c}{s + \tilde{\omega}_c(1+\sqrt{2})},$$

which is the controller to implement. Because $|R(j\tilde{\omega}_c)| = \tilde{\omega}_c^2$, the crossover frequency of the resulting loop $L(s) = P(s)R(s)$ is exactly as intended, i.e. $\omega_c = \tilde{\omega}_c$, which is not typical. The classical stability margins of the resulted loop are $\mu_g = \infty$, $\mu_{ph} = 45°$, and $\mu_m \approx 0.692$.

9.3.2 Triple integrator

Let us now repeat the steps considered in the previous case study with the triple integrator $P(s) = 1/s^3$, which is a more challenging system to control, as more than a 90° phase lead is required. Select again a static weight, now of the form $W(s) = \tilde{\omega}_c^3$, where $\tilde{\omega}_c > 0$ is an intended crossover frequency. With this choice

$$P_{msh}(s) = \frac{\tilde{\omega}_c^3}{s^3} = \left[\begin{array}{ccc|c} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & \sqrt{\tilde{\omega}_c^3} \\ \hline \sqrt{\tilde{\omega}_c^3} & 0 & 0 & 0 \end{array}\right]$$

indeed has its crossover frequency at $\omega = \tilde{\omega}_c$.

AREs (9.12) still can be solved analytically, with the stabilizing solutions

$$X = \begin{bmatrix} 2\tilde{\omega}_c^2 & 2\tilde{\omega}_c & 1 \\ 2\tilde{\omega}_c & 3 & 2/\tilde{\omega}_c \\ 1 & 2/\tilde{\omega}_c & 2/\tilde{\omega}_c^2 \end{bmatrix} > 0 \quad \text{and} \quad Y = \begin{bmatrix} 2/\tilde{\omega}_c^2 & 2/\tilde{\omega}_c & 1 \\ 2/\tilde{\omega}_c & 3 & 2\tilde{\omega}_c \\ 1 & 2\tilde{\omega}_c & 2\tilde{\omega}_c^2 \end{bmatrix} > 0$$

and

$$\epsilon_{max} = \sqrt{\frac{1}{2} - \frac{\sqrt{2}}{3}} \approx 0.1691, \tag{9.6}$$

which is about 44% of what we had in the double integrator case. It does not depend on $\tilde{\omega}_c$, again, which also follows by the fact that the phase of $P_{msh}(j\omega)$ does not depend on ω.

The controller corresponding to $\gamma = \epsilon_{max}$ calculated by (9.14′) is then the second-order transfer function

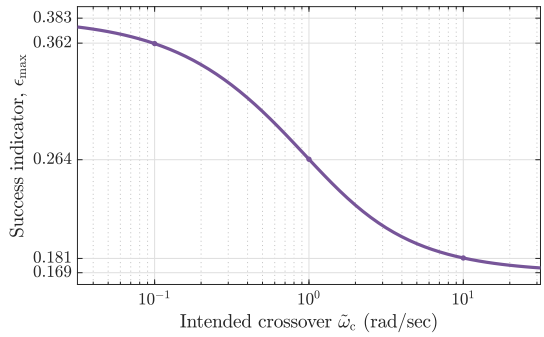
$$R_{rob}(s) = -\frac{(1+\sqrt{2})^2 s^2 + (2+\sqrt{2})\tilde{\omega}_c s + \tilde{\omega}_c^2}{s^2 + (2+\sqrt{2})\tilde{\omega}_c s + (1+\sqrt{2})^2\tilde{\omega}_c^2}. \tag{9.7}$$

This is a complex second-order lead controller tuned for the crossover frequency $\tilde{\omega}_c$, actually, of the form (1.32) for the choices $\alpha = (1+\sqrt{2})^2 \approx 5.828$ and $\zeta = 1/\sqrt{2}$. It provides a phase lead of $\approx 109°$ at $\tilde{\omega}_c$ and has $|R_{rob}(\infty)/R_{rob}(0)| = (1+\sqrt{2})^4 \approx 30.6\,\text{dB}$. As a matter of fact, this $R_{rob}(s) = \beta B_2(s/\omega_1)/B_2(s/\omega_2)$, where $B_2(s)$ is the normalized Butterworth polynomial of order 2, see the discussion on p. 148, $\beta = 3 - 2\sqrt{2} \approx 0.1716$, $\omega_1 = \tilde{\omega}_c\beta$ and $\omega_2 = \tilde{\omega}_c/\beta$.

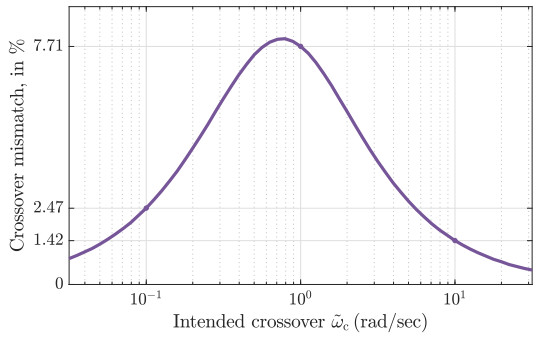
After moving the weight to the controller, we end up with

$$R(s) = -\tilde{\omega}_c^2 \frac{(1+\sqrt{2})^2 s^2 + (2+\sqrt{2})\tilde{\omega}_c s + \tilde{\omega}_c^2}{s^2 + (2+\sqrt{2})\tilde{\omega}_c s + (1+\sqrt{2})^2\tilde{\omega}_c^2},$$

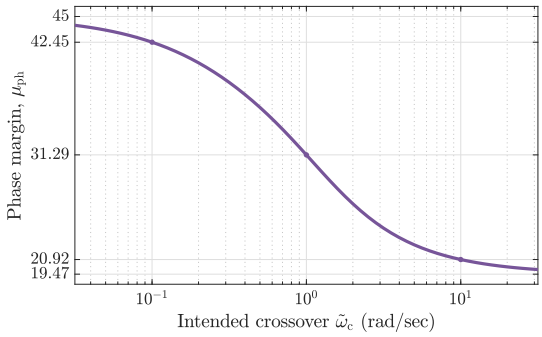
which is the controller to implement. Because $|R(j\tilde{\omega}_c)| = \tilde{\omega}_c^2$, the crossover frequency of the resulting loop $L(s) = P(s)R(s)$ is again exactly as intended. The classical stability margins of the resulted loop, $\mu_g \approx 2.29$, $\mu_{ph} \approx 19°$, and $\mu_m \approx 0.33$, are significantly lower than those in the double integrator case.

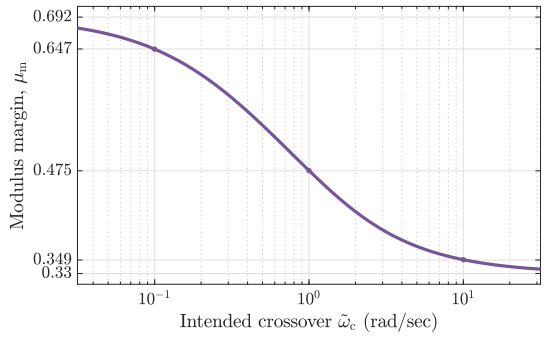


(a) Second-phase success indicator, $\epsilon_{\max}$



(b) Resulting crossover mismatch, $1 - \omega_c/\tilde{\omega}_c$





(c) Phase and modulus margins

Fig. 9.2: Performance measures for the example in §9.3.3

9.3.3 Servo control of a DC motor

Let now

$$P(s) = \frac{1}{s(s+1)},$$

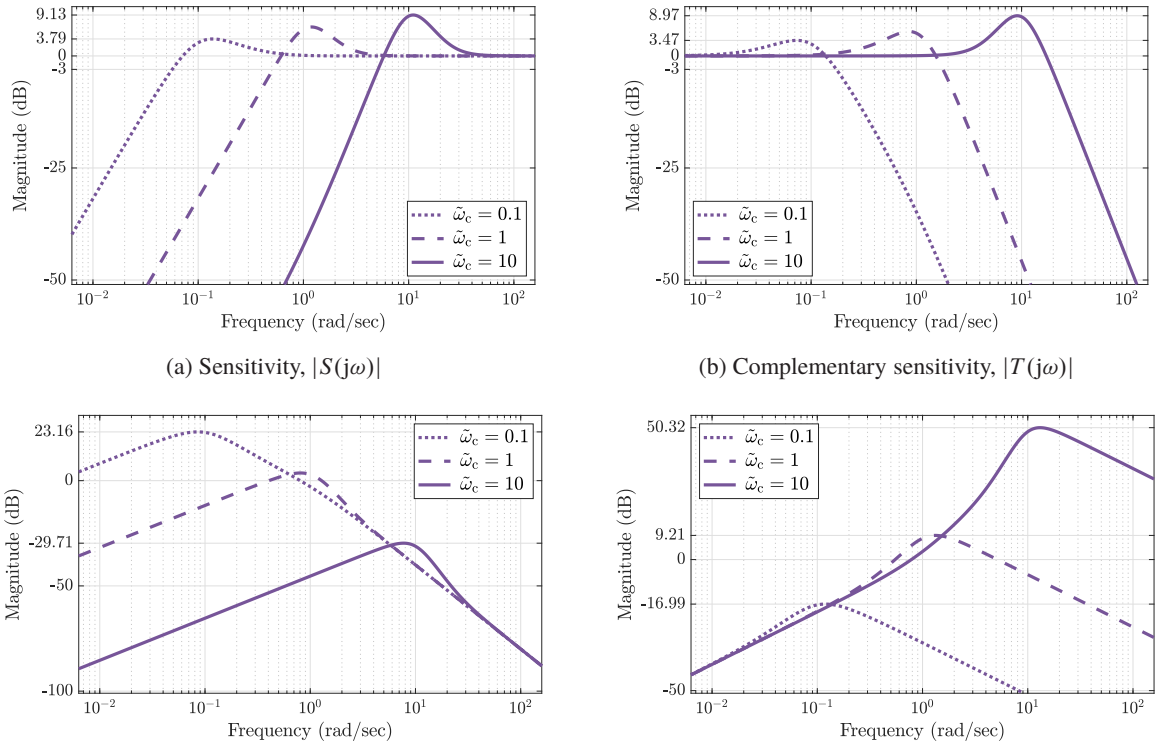
which can be thought of as a model of a DC motor connected with a rigid mechanical load, see §1.A.1, in particular, Eqn. (1.22) on p. 13. The goal here is to design a controller with an integral action (to reduce slow load disturbances d_i), to have a high-frequency roll-off, and to have a desired crossover. The bandwidth of the plant itself is $\omega_b = 1$ [rad/sec] and it has a unit gain at the frequency $\omega = 0.786$ [rad/sec].

A choice of the weighting function for the requirements above is $W(s) = k/s$, where $k = \tilde{\omega}_c^2\sqrt{1+\tilde{\omega}_c^2}$ for an intended crossover $\tilde{\omega}_c > 0$. With this choice,

$$P_{\text{msh}}(s) = \frac{\tilde{\omega}_c^2\sqrt{1+\tilde{\omega}_c^2}}{s^2(s+1)}.$$

If $\tilde{\omega}_c$ is placed well below the plant bandwidth, the phase lag of the pole at $s = -1$ around $\omega = \tilde{\omega}_c$ is negligible and we effectively have the same double integrator from §9.3.1. As $\tilde{\omega}_c$ increases, the problem is expected to become harder, because the phase lag due to the plant pole will become noticeable. If $\tilde{\omega}_c$ grows well beyond the plant bandwidth, the problem should approach that for the triple integrator considered in §9.3.2. These arguments are supported by the success indicator $\epsilon_{\max}$. It is expectably a decreasing function of $\tilde{\omega}_c$ then, starting from 0.3827 as in (9.4) and approaching 0.1691 as in (9.6) as $\tilde{\omega}_c$ grows, see Fig. 9.2(a). Other stability margins similarly deteriorate as $\tilde{\omega}_c$ increases, which is seen in the plots in Fig. 9.2(c).

The controllers produced in the second phase of the H_∞ loop-shaping procedure also migrate, in a sense, from the first-order lead as in (9.5) for relatively small $\tilde{\omega}_c$ to the complex second-order lead as in



(a) Sensitivity, $|S(\mathrm{j}\omega)|$

(b) Complementary sensitivity, $|T(\mathrm{j}\omega)|$

(c) Disturbance sensitivity, $|T_{\mathrm{d}}(\mathrm{j}\omega)|$

(d) Control sensitivity, $|T_{\mathrm{c}}(\mathrm{j}\omega)|$

Fig. 9.3: Closed-loop frequency-responses for the example in §9.3.3

(9.7) as $\tilde{\omega}_{\mathrm{c}}$ grows. Controllers $R = WR_{\mathrm{rob}}$ for three choices of $\tilde{\omega}_{\mathrm{c}}$ are presented below:

$$R(s) = -\frac{0.025912(s+0.04072)(s+1)}{s(s+0.2819)(s+0.9603)}, \qquad \text{if } \tilde{\omega}_{\mathrm{c}} = 0.1$$

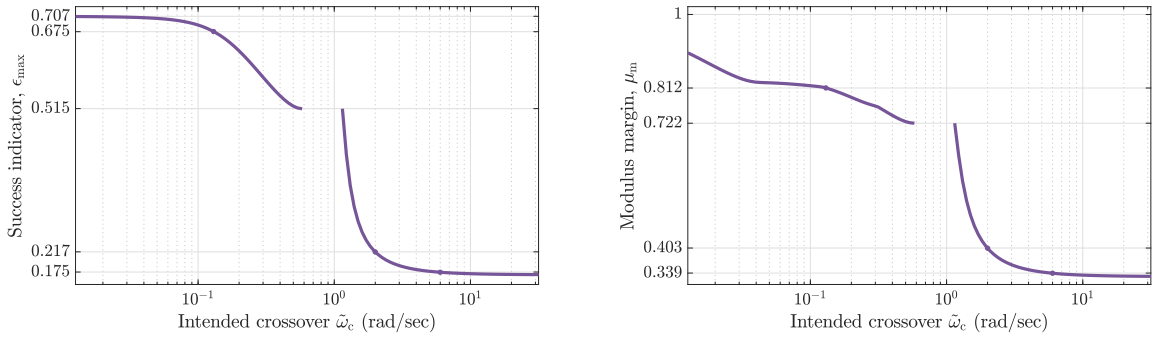
$$R(s) = -\frac{5.1584(s+0.4733)(s+0.9477)}{s(s^2+3.614s+5.968)}, \qquad \text{if } \tilde{\omega}_{\mathrm{c}} = 1$$

$$R(s) = -\frac{5448.9(s^2+6.512s+19.16)}{s(s^2+33.62s+563.1)}, \qquad \text{if } \tilde{\omega}_{\mathrm{c}} = 10$$

They all are strictly proper and contain integral action, as required. The controller for $\tilde{\omega}_{\mathrm{c}} = 0.1$ can afford canceling the stable pole of the plant as this pole is fast comparing to the required crossover. Moreover, it has a close pole at $s = -0.96$, so the whole controller is effectively an integrator with a first-order lead. The controller for $\tilde{\omega}_{\mathrm{c}} = 1$ does not cancel the stable pole of the plant. Its zeros at $s = 0.47$ and $s = 0.95$ attract two root loci, so that the stable pole of the plant is shifted, rather than canceled. The controller for $\tilde{\omega}_{\mathrm{c}} = 10$ is not even close to canceling the plant pole at $s = -1$, apparently because this pole is rather slow relatively to the intended bandwidth.

Note that the actual crossover frequency ω_{c}, that of the resulted loop PWR_{rob}, is now affected by the control design in the second phase of the H_∞ loop-shaping procedure. Its relative deviation from the intended $\tilde{\omega}_{\mathrm{c}}$ is presented in Fig. 9.2(b). Still, this deviation is always below 8%, meaning that in this case robust stability does not conflict with the crossover requirement too much.

The closed-loop frequency responses are shown in Fig. 9.3. Their trends are quite expectable. As $\tilde{\omega}_{\mathrm{c}}$ grows, the closed-loop bandwidth increases (both if defined in terms of the sensitivity function, see Fig. 9.3(a), and if defined in terms of the complementary sensitivity, see Fig. 9.3(b)), disturbance attenuation improves (see Fig. 9.3(c)), but all this comes at the expense of higher control effort (see Fig. 9.3(d)).



(a) Second-phase success indicator, $\epsilon_{\max}$

(b) Modulus margin, μ_m

Fig. 9.4: Performance measures for the example in §9.3.4

The design can be refined by adding a lead element to the weighting function. This, in a sense, helps the loop-shaping procedure to end up with a more robust design in the second phase. For example, if the lead element $(3s + \tilde{\omega}_c)/(s + 3\tilde{\omega}_c)$ was added to the I weight $W(s)$, the success indicator would lie in $(0.268, 0.542)$, rather than in $(0.109, 0.383)$ as in Fig. 9.2(a). The resulting stability margins and closed-loop frequency responses would improve as well, at the expense of a slightly slower high-frequency decay of the controller gain. Further refinements can be brought via adding a low-pass filter…

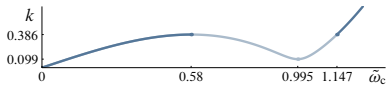
9.3.4 Lightly-damped system from Example 7.5

Consider now a plant with the transfer function

$$P(s) = \frac{1}{s^2 + 0.1s + 1},$$

which was studied in Example 7.5 on p. 151. The goal here remains the same as in §9.3.3: design a controller with an integral action, a high-frequency roll-off, and a desired crossover. The bandwidth of the plant itself is $\omega_b = 1.551$ [rad/sec]. We again select $W(s) = k/s$, now with

$$k = \tilde{\omega}_c\sqrt{1 - 1.99\tilde{\omega}_c^2 + \tilde{\omega}_c^4} =$$

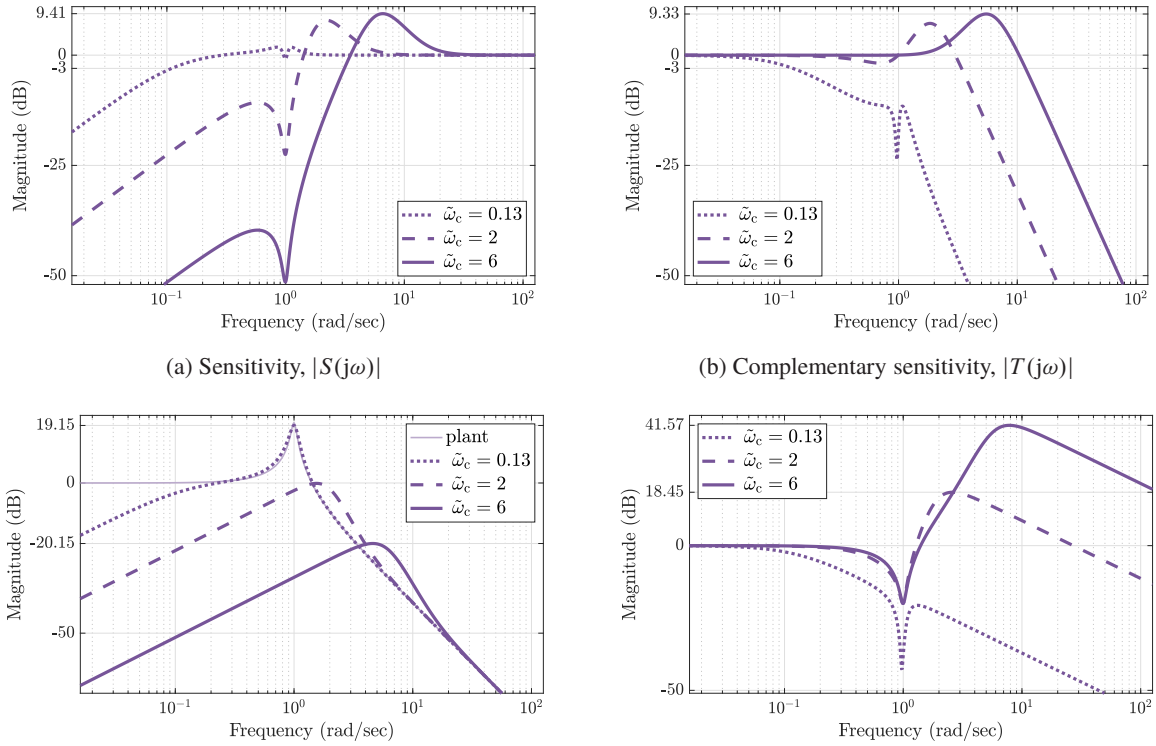


for an intended crossover $\tilde{\omega}_c > 0$. With this choice,

$$P_{\text{msh}}(s) = \frac{\tilde{\omega}_c\sqrt{1 - 1.99\tilde{\omega}_c^2 + \tilde{\omega}_c^4}}{s(s^2 + 0.1s + 1)}$$

satisfies $|P_{\text{msh}}(\mathrm{j}\tilde{\omega}_c)| = 1$. If $\tilde{\omega}_c \in (0, 0.58) \cup (1.147, \infty)$, this is the only crossover frequency. Otherwise, there are three and the smallest of them is below 0.58 anyway. Thus, the choice of $\tilde{\omega}_c \in (0.58, 1.147)$, around the resonance peak, would produce the same gain k of the weighting function $W(s)$ as some smaller $\tilde{\omega}_c$ and there is no loss of generality in considering only $\tilde{\omega}_c \in (0, 0.58) \cup (1.147, \infty)$.

The success indicator $\epsilon_{\max}$ and the modulus margin of the resulting design are presented in Fig. 9.4 (the actual crossover is complicated around the plant resonance, so its values are not presented). It is seen that the success indicator is very high if $\tilde{\omega}_c$ is below the plant resonance at $\omega = 0.9975$, but deteriorates rapidly afterwards. This can be explained by observing that at low $\tilde{\omega}_c$'s the resonance of P is outside the frequencies of interest and P_{msh} is essentially a single integrator. The latter is an easy system to control, with the huge $\epsilon_{\max} = \sqrt{0.5} \approx 0.7071$. After the resonant frequency, the phase of $P_{\text{msh}}(\mathrm{j}\omega)$ quickly decays to $-270°$, rendering P_{msh} close to the triple integrator from §9.3.2, whose $\epsilon_{\max} \approx 0.1691$.



(a) Sensitivity, $|S(\mathrm{j}\omega)|$

(b) Complementary sensitivity, $|T(\mathrm{j}\omega)|$

(c) Disturbance sensitivity, $|T_\mathrm{d}(\mathrm{j}\omega)|$

(d) Control sensitivity, $|T_\mathrm{c}(\mathrm{j}\omega)|$

Fig. 9.5: Closed-loop frequency-responses for the example in §9.3.3

Consider now design outcomes for three choices of $\tilde{\omega}_\mathrm{c} \in \{0.13, 2, 6\}$. The resulting controllers are

$$R(s) = -\frac{0.1396(s^2 - 0.03291s + 0.9464)}{s(s^2 + 0.6051s + 1.129)}, \qquad \text{if } \tilde{\omega}_\mathrm{c} = 0.13$$

$$R(s) = -\frac{27.083(s^2 + 0.8226s + 0.8323)}{s(s^2 + 5.725s + 16.88)}, \qquad \text{if } \tilde{\omega}_\mathrm{c} = 2$$

$$R(s) = -\frac{1184.8(s^2 + 3.453s + 6.361)}{s(s^2 + 20.1s + 202.4)}, \qquad \text{if } \tilde{\omega}_\mathrm{c} = 6$$

The controller for $\tilde{\omega}_\mathrm{c} = 0.13$ is actually nonminimum phase[2] and its zeros at $s = 0.0165 \pm \mathrm{j}0.9727$ nearly cancel the lightly-damped plant poles at $s = -0.05 \pm \mathrm{j}0.9987$. This near-cancellation can be explained by the irrelevance of the resonant frequency, which lies almost a decade above the intended crossover, for the design in this case. As high loop gain requirements overtake the resonance, canceling the lightly-damped poles of $P(s)$ does not pay off any longer and the controllers designed with $\tilde{\omega}_\mathrm{c} = 2$ and $\tilde{\omega}_\mathrm{c} = 6$ have their zeros, $s = -0.4113 \pm \mathrm{j}0.8143$, $s = -1.7264 \pm \mathrm{j}1.8386$, further away from the plant poles.

The effect of canceling (and not canceling) lightly-damped poles can be clearly seen in the disturbance sensitivity plots in Fig. 9.5(c), by comparing closed-loop frequency response with that of the plant itself (shown by the thin light line there). The design with $\tilde{\omega}_\mathrm{c} = 0.13$ practically does not alter the plant response to load disturbances above that frequency, including the resonance. As $\tilde{\omega}_\mathrm{c}$ passes the resonant frequency, the closed-loop disturbance sensitivity at the latter is substantially lower than that of the plant.

In other respects, the trends of the plots in Fig. 9.5 are similar to those in the previous example. It is perhaps worth emphasizing that the H_∞ loop-shaping design with $\tilde{\omega}_\mathrm{c} = 2$ is roughly compatible with the

[2]For no apparent reason, mirroring its zeros over the imaginary axis would result in virtually the same design.

mixed sensitivity design of Example 7.5 in properties of the sensitivity and control sensitivity functions, see Figs. 7.4 and 7.5. At the same time, it reduces the disturbance sensitivity peak by almost 20 dB, i.e. by an order of magnitude.

9.A Balanced sensitivity problem in H_∞

This section presents technical steps required to solve the balanced sensitivity problem of form (9.2) and its state-space solution. Throughout this part we consider the problem of minimizing the H_∞-norm of the system

$$T_{zw} = \begin{bmatrix} T_{\mathrm{c}} & T_{\mathrm{i}} \\ S_{\mathrm{o}} & T_{\mathrm{d}} \end{bmatrix} = \begin{bmatrix} R_{\mathrm{rob}} \\ I \end{bmatrix} (I - P_{\mathrm{msh}} R_{\mathrm{rob}})^{-1} \begin{bmatrix} I & P_{\mathrm{msh}} \end{bmatrix}, \tag{9.8}$$

i.e. we drop the subscripts associated with the phases of H_∞ loop shaping.

Remark 9.2 (symmetry). Consider the system

$$\bar{T}_{zw} := \begin{bmatrix} I \\ P_{\mathrm{msh}} \end{bmatrix} (I - R_{\mathrm{rob}} P_{\mathrm{msh}})^{-1} \begin{bmatrix} R_{\mathrm{rob}} & I \end{bmatrix} = \begin{bmatrix} T_{\mathrm{c}} & S_{\mathrm{i}} \\ T_{\mathrm{o}} & T_{\mathrm{d}} \end{bmatrix} = \begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix} + T_{zw}.$$

It can be shown that $\|\bar{T}_{zw}(\mathrm{j}\omega)\| = \|T_{zw}(\mathrm{j}\omega)\|$ for all $\omega \in \mathbb{R}$ at which $\det(I + P_{\mathrm{msh}}(\mathrm{j}\omega) R_{\mathrm{rob}}(\mathrm{j}\omega)) \neq 0$. Thus, the use of S_{o} and T_{i} in (9.8) does not entail any asymmetry between input and output sensitivity and complementary sensitivity functions. ▽

The closed-loop system in (9.8) corresponds to the standard problem in Fig. 7.1 with the generalized plant

$$G(s) = G_{\mathrm{BS}}(s) = \left[\begin{array}{cc:c} 0 & 0 & I \\ I & P_{\mathrm{msh}}(s) & P_{\mathrm{msh}}(s) \\ \hdashline I & P_{\mathrm{msh}}(s) & P_{\mathrm{msh}}(s) \end{array}\right], \tag{9.9}$$

which can be obtained as the system $\left[\begin{smallmatrix} d_{\mathrm{o}} \\ d_{\mathrm{i}} \end{smallmatrix}\right] \mapsto \left[\begin{smallmatrix} u \\ y \end{smallmatrix}\right]$ in the setup of Fig. 9.1. It is always stabilizable, as follows from Theorem 6.13 via the following construction of coprime factorizations of G:

$$G = \left[\begin{array}{cc:c} 0 & -I & M \\ I & 0 & N \\ \hdashline I & 0 & N \end{array}\right] \left[\begin{array}{cc:c} I & 0 & 0 \\ 0 & I & 0 \\ \hdashline 0 & -I & M \end{array}\right]^{-1} = \left[\begin{array}{cc:c} I & 0 & 0 \\ 0 & I & -I \\ \hdashline 0 & 0 & \tilde{M} \end{array}\right]^{-1} \left[\begin{array}{cc:c} 0 & 0 & I \\ 0 & 0 & 0 \\ \hdashline \tilde{M} & \tilde{N} & \tilde{N} \end{array}\right]$$

(the corresponding Bézout coefficients are straightforward to generate from those of P_{msh}). Using the Youla–Kučera parametrization from (7.4), we then end up with

$$T_{zw} = \left(\begin{bmatrix} -\tilde{Y} \\ \tilde{X} \end{bmatrix} + \begin{bmatrix} M \\ N \end{bmatrix} Q \right) \begin{bmatrix} \tilde{M} & \tilde{N} \end{bmatrix}, \tag{9.10}$$

which is again an affine function of the Q-parameter.

The minimization of $\|T_{zw}\|_\infty$ is a particular case of the standard H_∞ problem, whose solution is presented in §7.A.2. As such, the solution can be in principle derived via Theorem 7.3. However, the balanced sensitivity problem possesses some special properties, worth examining separately. This is the subject of the remainder of this section.

The solution discussed below exploits the freedom in the choice of the coprime factors of P_{msh}. To motivate a special doubly coprime factorization of choice below, recall the arguments used in Example 7.2 on p. 141 to choose a co-inner denominator of the plant. In the same vein, it may make sense to look for a *lcf* $P_{\mathrm{msh}} = \tilde{M}^{-1}\tilde{N}$ with a co-inner $\begin{bmatrix} \tilde{M}(s) & \tilde{N}(s) \end{bmatrix}$, because this also eliminates this factor from the analysis of $\|T_{zw}\|_\infty$ by virtue of Proposition 3.1 on p. 51. The result below generalizes this choice to other parts of a doubly coprime factorization of P_{msh} (the notation $G^\sim$ stands for the conjugate of G defined in (3.29)).

Lemma 9.1. *There always exists a doubly coprime factorization of P_{msh}, known as* normalized*, such that*

$$\begin{bmatrix} M^\sim & N^\sim \end{bmatrix}\begin{bmatrix} M & -\tilde{Y} \\ N & \tilde{X} \end{bmatrix} = \begin{bmatrix} I & V^\sim \end{bmatrix} \quad \textit{and} \quad \begin{bmatrix} X & Y \\ -\tilde{N} & \tilde{M} \end{bmatrix}\begin{bmatrix} -\tilde{N}^\sim \\ \tilde{M}^\sim \end{bmatrix} = \begin{bmatrix} -V^\sim \\ I \end{bmatrix} \tag{9.11}$$

for some $V \in RH_\infty$ such that $V(s)$ is strictly proper.

Proof. Let $P_{\text{msh}} = N_0 M_0^{-1} = \tilde{M}_0^{-1}\tilde{N}_0$ be some doubly coprime factorization of P_{msh} with corresponding Bézout coefficients X_0, Y_0, $\tilde{X}_0$, and $\tilde{Y}_0$. It is readily seen that

$$\begin{bmatrix} X & Y \\ -\tilde{N} & \tilde{M} \end{bmatrix} := \begin{bmatrix} U_0 & -U_0 W_0 \\ 0 & \tilde{U}_0^{-1} \end{bmatrix}\begin{bmatrix} X_0 & Y_0 \\ -\tilde{N}_0 & \tilde{M}_0 \end{bmatrix}$$

and

$$\begin{bmatrix} M & -\tilde{Y} \\ N & \tilde{X} \end{bmatrix} := \begin{bmatrix} M_0 & -\tilde{Y}_0 \\ N_0 & \tilde{X}_0 \end{bmatrix}\begin{bmatrix} U_0^{-1} & W_0\tilde{U}_0 \\ 0 & \tilde{U}_0 \end{bmatrix}$$

are also doubly coprime factors and Bézout coefficients of P_{msh} for every $U_0, \tilde{U}_0, U_0^{-1}, \tilde{U}_0^{-1}, W_0 \in H_\infty$, cf. Proposition 3.2. Because of coprimeness, we have that

$$[M_0(\mathrm{j}\omega)]' M_0(\mathrm{j}\omega) + [N_0(\mathrm{j}\omega)]' N_0(\mathrm{j}\omega) > 0 \quad \text{and} \quad \tilde{M}_0(\mathrm{j}\omega)[\tilde{M}_0(\mathrm{j}\omega)]' + \tilde{N}_0(\mathrm{j}\omega)[\tilde{N}_0(\mathrm{j}\omega)]' > 0$$

for all $\omega \in \mathbb{R} \cup \{\pm\infty\}$. Hence, there are bistable U_0 and $\tilde{U}_0$ (the spectral factors) such that

$$M_0^\sim M_0 + N_0^\sim N_0 = U_0^\sim U_0 \quad \text{and} \quad \tilde{M}_0\tilde{M}_0^\sim + \tilde{N}_0\tilde{N}_0^\sim = \tilde{U}_0\tilde{U}_0^\sim.$$

These spectral factors are actually unique up to right and left multiplications by unitary matrices. With these choices, $M^\sim M + N^\sim N = I$ and $\tilde{M}\tilde{M}^\sim + \tilde{N}\tilde{N}^\sim = I$, as required by (9.11). Next,

$$\begin{bmatrix} M^\sim & N^\sim \end{bmatrix}\begin{bmatrix} -\tilde{Y} \\ \tilde{X} \end{bmatrix} = \begin{bmatrix} M^\sim & N^\sim \end{bmatrix}\begin{bmatrix} M & -\tilde{Y}_0 \\ N & \tilde{X}_0 \end{bmatrix}\begin{bmatrix} U_0 W_0 \tilde{U}_0 \\ \tilde{U}_0 \end{bmatrix} = U_0 W_0 \tilde{U}_0 - T,$$

where $T := (M^\sim\tilde{Y}_0 - N^\sim\tilde{X}_0)\tilde{U}_0$. We can always decompose $T = T_{\mathrm{s}} + T_{\bar{\mathrm{s}}}$ where $T_{\mathrm{s}} \in RH_\infty$ and $T_{\bar{\mathrm{s}}}(s)$ is strictly proper and has all its poles in the closed right half-plane $\overline{\mathbb{C}}_0$. The choice $W_0 = U_0^{-1} T_{\mathrm{s}} \tilde{U}_0^{-1} \in RH_\infty$ yields then the required left Bézout coefficients under $V = -T_{\bar{\mathrm{s}}}^\sim$.

The final step is to show that the right Bézout coefficients satisfy then the second equality in (9.11). To this end, rewrite its first equality as

$$\begin{bmatrix} M^\sim & N^\sim \end{bmatrix} = \begin{bmatrix} I & V^\sim \end{bmatrix}\begin{bmatrix} M & -\tilde{Y} \\ N & \tilde{X} \end{bmatrix}^{-1} = \begin{bmatrix} I & V^\sim \end{bmatrix}\begin{bmatrix} X & Y \\ -\tilde{N} & \tilde{M} \end{bmatrix}.$$

Post-multiplying both sides by $\begin{bmatrix} -\tilde{N} & \tilde{M} \end{bmatrix}^\sim$ we get

$$0 = \begin{bmatrix} I & V^\sim \end{bmatrix}\begin{bmatrix} X & Y \\ -\tilde{N} & \tilde{M} \end{bmatrix}\begin{bmatrix} -\tilde{N}^\sim \\ \tilde{M}^\sim \end{bmatrix} = \begin{bmatrix} I & V^\sim \end{bmatrix}\begin{bmatrix} Y\tilde{M}^\sim - X\tilde{N}^\sim \\ I \end{bmatrix}$$

and then $V^\sim = X\tilde{N}^\sim - Y\tilde{M}^\sim$, which yields the last piece to (9.11). □

Thus, assume that the factors in (9.10) are normalized and that the Bézout coefficients also satisfy (9.11). Because $\begin{bmatrix} \tilde{M} & \tilde{N} \end{bmatrix}$ is co-inner, it follows from Proposition 3.1 that $\|T_{zw}\|_\infty = \|T_1\|_\infty$, where

$$T_1 := \begin{bmatrix} -\tilde{Y} \\ \tilde{X} \end{bmatrix} + \begin{bmatrix} M \\ N \end{bmatrix} Q.$$

The H_∞-norm of T_1 can then be replaced by its L_∞-norm defined in (3.21). This switch is safe as long as we keep $Q \in H_\infty$. Define

$$U := \begin{bmatrix} M^\sim & N^\sim \\ -\tilde{N} & \tilde{M} \end{bmatrix}.$$

It follows from (9.11) and (3.34) that $U^\sim(s)U(s) = I$, implying that $U(\mathrm{j}\omega)$ is a unitary matrix for every ω. This, in turn, implies that $\|T_1\|_\infty = \|UT_1\|_\infty$. Using the relation

$$UT_1 = \begin{bmatrix} M^\sim & N^\sim \\ -\tilde{N} & \tilde{M} \end{bmatrix} \begin{bmatrix} -\tilde{Y} \\ \tilde{X} \end{bmatrix} + \begin{bmatrix} I \\ 0 \end{bmatrix} Q = \begin{bmatrix} V^\sim + Q \\ I \end{bmatrix}$$

and the definition of the matrix spectral norm in (2.6b), we have that

$$\|U(\mathrm{j}\omega)T_1(\mathrm{j}\omega)\|^2 = 1 + \|[V(\mathrm{j}\omega)]' + Q(\mathrm{j}\omega)\|^2$$

at every ω. Hence, we may conclude that

$$\|T_{zw}\|_\infty = \|UT_1\|_\infty = \sqrt{1 + \|V^\sim + Q\|_\infty^2}$$

for every $Q \in RH_\infty$. But this implies that the problem of selecting $Q \in RH_\infty$ that minimizes $\|T_{zw}\|_\infty$ reduces to the Nehari problem of minimizing $\|V^\sim + Q\|_\infty$. We know from Theorem 7.4 that the latter problem has its minimal performance level at $\|V\|_{\mathrm{H}}$, which is the Hankel norm of V and can be calculated by solving two Lyapunov equations. Hence, we end up with a non-iterative attainable performance formula for the balanced sensitivity problem. Namely,

$$\inf_{\text{stabilizing } R_{\text{rob}}} \left\| \begin{bmatrix} R_{\text{rob}} \\ I \end{bmatrix} (I - P_{\text{msh}} R_{\text{rob}})^{-1} \begin{bmatrix} I & P_{\text{msh}} \end{bmatrix} \right\|_\infty = \sqrt{1 + \|V\|_{\mathrm{H}}^2},$$

where the doubly coprime factors of of P_{msh} are those satisfying (9.11). All admissible controllers can then be constructed in state space using Theorem 7.4.

Remark 9.3 (alternative cost expression). It can be shown, see [17], that the optimal cost for the balanced sensitivity problem can be equivalently expressed as

$$\inf_{\text{stabilizing } R_{\text{rob}}} \left\| \begin{bmatrix} R_{\text{rob}} \\ I \end{bmatrix} (I - P_{\text{msh}} R_{\text{rob}})^{-1} \begin{bmatrix} I & P_{\text{msh}} \end{bmatrix} \right\|_\infty = \frac{1}{\sqrt{1 - \left\| \begin{bmatrix} \tilde{N} & \tilde{M} \end{bmatrix} \right\|_{\mathrm{H}}^2}}.$$

Because $\begin{bmatrix} \tilde{N} & \tilde{M} \end{bmatrix}$ is co-inner, its Hankel norm is always strictly contractive. ▽

To present the resulting state-space formulae, bring in a stabilizable and detectable state-space realization (A, B, C, D) of the plant P_{msh}. We need two algebraic Riccati equations,

$$A'X + XA - (C'D + XB)(I + D'D)^{-1}(D'C + B'X) + C'C = 0, \tag{9.12a}$$

with the corresponding gain matrix $K := -(I + D'D)^{-1}(D'C + B'X)$, and

$$AY + YA' - (BD' + YC')(I + DD')^{-1}(DB' + CY) + BB' = 0, \tag{9.12b}$$

with the corresponding gain matrix $L := -(BD' + YC')(I + DD')^{-1}$. Their solutions are said to be stabilizing if $A + BK$ and $A + LC$ are Hurwitz matrices. Note that these AREs do not depend on γ and are of the H_2 form, with negative semi-definite quadratic terms. These are not typical properties of AREs arising in H_∞ optimization, cf. (7.37). The main result of this section, whose proof is postponed to §9.A.1, is then formulated as follows.

Theorem 9.2. *The minimal attainable performance for the balanced sensitivity H_∞ problem is*

$$\gamma_{min} = \sqrt{1 + \rho(YX)} > 1.$$

Given then any $\gamma > \gamma_{min}$, all γ-suboptimal controllers are given by

$$R_{rob}(s) = \mathcal{F}_l\left(\left[\begin{array}{c|cc} A + BK - Z_\gamma^{-1}YC'(C + DK) & Z_\gamma^{-1}YC' & Z_\gamma^{-1}(B - YCD') \\ \hline -B'X & -D' & I + D'D \\ -C - DK & I & -D \end{array}\right], Q(s)\right), \tag{9.13}$$

for any $Q \in RH_\infty$ such that $\det(I + DQ(\infty)) \neq 0$ *and*

$$\|(I + D'D)^{1/2}Q(I + DD')^{1/2}\|_\infty < \sqrt{\gamma^2 - 1},$$

where $Z_\gamma := (1 - \gamma^{-2})I - \gamma^{-2}YX$.

The *central controller*, the one corresponding to $Q = 0$, is then obviously

$$R_{\text{rob}}(s) = -\left[\begin{array}{c|c} A + BK - Z_\gamma^{-1}YC'(C + DK) & Z_\gamma^{-1}YC' \\ \hline B'X & D' \end{array}\right] \tag{9.14}$$

Note that as $\gamma \downarrow \gamma_{\min}$, the matrix $\lim_{\gamma\downarrow\gamma_{\min}} Z_\gamma = \gamma_{\min}^{-2}(\rho(YX)I - YX)$ becomes singular and the controller above is not well defined. However, this can be resolved via rewriting the central controller as

$$R_{\text{rob}}(s) = -D' - B'X(sZ_\gamma - Z_\gamma A - Z_\gamma BK + YC'(C + DK))^{-1}YC'. \tag{9.14$'$}$$

This is the so-called *descriptor form* of the transfer matrix and it is well defined provided the polynomial matrix $sZ_\gamma - Z_\gamma A - Z_\gamma BK + YC'(C + DK)$ has full normal rank. It can be shown that this is always the case for the central controller above. As a result, the optimal controller is also well defined and, actually, it is a reduced-order controller. The order of the optimal controller equals $\text{rank}(\rho(YX)I - YX)$, i.e. the order of the optimal controller is reduced by the geometric multiplicity of the largest eigenvalue of YX.

9.A.1 Proof of Theorem 9.2

First, some technical facts about the algebraic Riccati equations (9.12) will be established. To this end, define the Hamiltonian matrix

$$H_{(9.12)} = \begin{bmatrix} A & 0 \\ -C'C & -A' \end{bmatrix} - \begin{bmatrix} B \\ -C'D \end{bmatrix}(I + D'D)^{-1}\begin{bmatrix} D'C & B' \end{bmatrix}.$$

It happens that both AREs in (9.12) are associated with it. Indeed, it can be verified that

$$\begin{bmatrix} I & 0 \\ -X & I \end{bmatrix} H_{(9.12)} \begin{bmatrix} I & 0 \\ X & I \end{bmatrix} = \begin{bmatrix} A + BK & -B(I + D'D)^{-1}B' \\ 0 & -(A + BK)' \end{bmatrix} \tag{9.15a}$$

and

$$\begin{bmatrix} I & Y \\ 0 & I \end{bmatrix} H_{(9.12)} \begin{bmatrix} I & -Y \\ 0 & I \end{bmatrix} = \begin{bmatrix} A + LC & 0 \\ -C'(I + DD')^{-1}C & -(A + LC)' \end{bmatrix} \tag{9.15b}$$

for any X and Y satisfying (9.12). The solvability of AREs requires that $H_{(9.12)}$ has no $j\omega$-axis eigenvalues. Straightforward arguments based on the Schur complement notion yields that $\lambda \in \text{spec}(H_{(9.12)})$ iff the polynomial matrix

$$\begin{bmatrix} A - sI & 0 & B \\ -C'C & -A' - sI & -C'D \\ D'C & B' & I + D'D \end{bmatrix} = R_{I + P'_{\text{msh}}P_{\text{msh}}}(s)$$

(R_G stands for the Rosenbrock system matrix of G, defined by (4.24)) looses its rank at $s = \lambda$. Because $I+[P_{\rm msh}(\mathrm{j}\omega)]'P_{\rm msh}(\mathrm{j}\omega) > 0$ for all ω and the realization (A, B, C, D) of $P_{\rm msh}$ is stabilizable and detectable, $I + P_{\rm msh}^{\sim}(s)P_{\rm msh}(s)$ cannot have pure imaginary invariant zeros, so that $R_{I+P'_{\rm msh}P_{\rm msh}}(\mathrm{j}\omega)$ has full rank at all ω. Hence, $H_{(9.12)}$ has no pure imaginary eigenvalues and, by Theorem B.6, both AREs in (9.12) have stabilizing solutions.

The pre-multiplication of (9.15a) and the post-multiplication of (9.15b) by $\left[\begin{smallmatrix} I & Y \\ 0 & I \end{smallmatrix}\right]\left[\begin{smallmatrix} I & 0 \\ X & I \end{smallmatrix}\right]$ equates their left-hand sides and yields the following relation, playing an important role in the analysis below:

$$\begin{bmatrix} I + YX & Y \\ X & I \end{bmatrix}\begin{bmatrix} A + BK & -B(I + D'D)^{-1}B' \\ 0 & -(A + BK)' \end{bmatrix} = \begin{bmatrix} A + LC & 0 \\ -C'(I + DD')^{-1}C & -(A + LC)' \end{bmatrix}\begin{bmatrix} I + YX & Y \\ X & I \end{bmatrix}. \tag{9.16}$$

We also need the following technical result.

Lemma 9.3. *The RH_∞ transfer functions*

$$\begin{bmatrix} X(s) & Y(s) \\ -\tilde{N}(s) & \tilde{M}(s) \end{bmatrix} = \begin{bmatrix} (I + D'D)^{1/2} & (I + D'D)^{-1/2}D' \\ 0 & (I + DD')^{-1/2} \end{bmatrix}\left[\begin{array}{c|cc} A + LC & B + LD & -L \\ \hline -K & I & 0 \\ -C & -D & I \end{array}\right] \tag{9.17a}$$

and

$$\begin{bmatrix} M(s) & -\tilde{Y}(s) \\ N(s) & \tilde{X}(s) \end{bmatrix} = \left[\begin{array}{c|cc} A + BK & B & -L \\ \hline K & I & 0 \\ C + DK & D & I \end{array}\right]\begin{bmatrix} (I + D'D)^{-1/2} & -D'(I + DD')^{-1/2} \\ 0 & (I + DD')^{1/2} \end{bmatrix} \tag{9.17b}$$

constitute a normalized double coprime factorization of P_{msh}, as in Lemma 9.1, with

$$V(s) = -(I + DD')^{-1/2}\left[\begin{array}{c|c} A + LC & (I + YX)B \\ \hline C & 0 \end{array}\right](I + D'D)^{-1/2}. \tag{9.18}$$

Proof. The fact that the transfer functions above constitute a doubly coprime factorization of $P_{\rm msh}$ follows by Propositions 4.12 and 3.2. Denote now $S := (I + D'D)^{-1/2}$ and $\tilde{S} := (I + DD')^{-1/2}$ and note that (9.12b) reads $(A + LC)Y + Y(A + LC)' + (B + LD)(B + LD)' + LL' = 0$. Thus,

$$\begin{bmatrix} X(s) & Y(s) \\ -\tilde{N}(s) & \tilde{M}(s) \end{bmatrix}\begin{bmatrix} -\tilde{N}^{\sim}(s) \\ \tilde{M}^{\sim}(s) \end{bmatrix} = \left[\begin{array}{c|cc} A + LC & B + LD & -L \\ \hline SB'X & S & SD' \\ -\tilde{S}C & -\tilde{S}D & \tilde{S} \end{array}\right]\left[\begin{array}{c|c} -(A + LC)' & C' \\ \hline (B + LD)' & -D' \\ -L' & I \end{array}\right]\tilde{S}$$

$$= \left[\begin{array}{cc|c} A + LC & -(A + LC)Y - Y(A + LC)' & YC' \\ 0 & -(A + LC)' & C' \\ \hline SB'X & SB' & 0 \\ -\tilde{S}C & \tilde{S}CY & \tilde{S}^{-1} \end{array}\right]\tilde{S}$$

and the application of a similarity transformation with the matrix $\left[\begin{smallmatrix} I & -Y \\ 0 & I \end{smallmatrix}\right]$ yields

$$= \left[\begin{array}{cc|c} A + LC & 0 & 0 \\ 0 & -(A + LC)' & C'\tilde{S} \\ \hline SB'X & SB'(I + XY) & 0 \\ -\tilde{S}C & 0 & I \end{array}\right] = \left[\begin{array}{c|c} -(A + LC)' & C'\tilde{S} \\ \hline SB'(I + XY) & 0 \\ 0 & I \end{array}\right],$$

which proves the second equality in (9.11) with V as in (9.18). The first equality in (9.11) can be obtained by similar arguments. □

Now, to derive the formula for the optimal performance of the balanced sensitivity problem we need the controllability and observability Gramians of the realization of V as in (9.18). This is done by the following result.

Lemma 9.4. *The controllability and observability Gramians of V in* (9.18) *are*

$$W_c = Y(I + XY) \quad \textit{and} \quad W_o = (I + XY)^{-1}X,$$

respectively.

Proof. The $(1, 2)$ block of (9.16) reads

$$-(I + YX)B(I + D'D)^{-1}B' - Y(A + BK)' = (A + LC)Y$$

Taking into account that $(A + BK)' = (I + XY)(A + LC)'(I + XY)^{-1}$, which follows from the $(1, 1)$ block of (9.16), we end up with

$$(A + LC)Y(I + XY) + Y(I + XY)(A + LC)' + (I + YX)B(I + D'D)^{-1}B'(I + XY) = 0,$$

meaning that $Y(I + XY)$ is indeed the controllability Gramian of (9.18) (because $A + LC$ is Hurwitz, the Lyapunov equation above has a unique solution).

Now, the $(2, 1)$ block of (9.16) reads

$$X(A + BK) = -(A + LC)'X - C'(I + DD')^{-1}C(I + YX).$$

Taking into account that $A + BK = (I + YX)^{-1}(A + LC)(I + YX)$, we now end up with

$$X(I + YX)^{-1}(A + LC) + (A + LC)'X(I + YX)^{-1} + C'(I + DD')^{-1}C = 0,$$

meaning that $(I + XY)^{-1}X = X(I + YX)^{-1}$ is indeed the observability Gramian of (9.18). □

With this result, the Hankel norm of V is obtained by Proposition B.4 as

$$\|V\|_{\mathrm{H}} = \sqrt{\rho(W_{\mathrm{c}}W_{\mathrm{o}})} = \sqrt{\rho(YX)}.$$

This yields the $\gamma_{\min}$ of Theorem 9.2.

The next step is to characterize all $Q \in RH_\infty$ such that $\|V^\sim + Q\|_{\mathrm{H}}$ for V given by (9.18). This can be done with the help of Theorem 7.4 for

$$A \to A + LC, \quad B \to (I + YX)B(I + D'D)^{-1/2}, \quad C \to -(I + DD')^{-1/2}C,$$

Gramians from Lemma 9.4, and $V_\gamma = \gamma^{-2}Z_\gamma^{-1}$ (mind that γ in Theorem 7.4 equals $\gamma^2 - 1$ in Theorem 9.2). The final controller in (9.13) is finally derived by plugging that LFT to the Youla–Kučera parametrization, based on the coprime factors and their Bézout coefficients in (9.17), and applying the Redheffer start product formula in Proposition 5.7. The details are left as a (tedious) exercise. □

Appendices

Appendix A

Spaces and Operators

THE PURPOSE of this appendix is to collect some basic mathematical definitions required throughout the notes. I'll try to avoid mathematical complications if possible, so the exposition might be somewhat informal at times.

A.1 Vector spaces

A.1.1 Basic definitions

A *field* is a set $\mathbb{F}$ together with two operations, addition $+ : \mathbb{F} \times \mathbb{F} \mapsto \mathbb{F}$ and multiplication $\cdot : \mathbb{F} \times \mathbb{F} \mapsto \mathbb{F}$, which satisfies the conditions presented in Table A.1 for both addition and multiplication for all its elements

	Addition	Multiplication
Commutativity	$\alpha + \beta = \beta + \alpha$	$\alpha \cdot \beta = \beta \cdot \alpha$
Associativity	$(\alpha + \beta) + \gamma = \alpha + (\beta + \gamma)$	$(\alpha \cdot \beta) \cdot \gamma = \alpha \cdot (\beta \cdot \gamma)$
Distributivity	$\alpha \cdot (\beta + \gamma) = \alpha \cdot \beta + \alpha \cdot \gamma$ or	$(\alpha + \beta) \cdot \gamma = \alpha \cdot \gamma + \beta \cdot \gamma$
Identity	$\exists 0 \in \mathbb{F}$ such that $\alpha + 0 = \alpha$	$\exists 1 \in \mathbb{F}$ such that $\alpha \cdot 1 = \alpha$
Inverses	$\exists -\alpha \in \mathbb{F}$ such that $\alpha + (-\alpha) = 0$	$\exists \alpha^{-1} \in \mathbb{F}$ such that $\alpha \cdot \alpha^{-1} = 1$ if $\alpha \neq 0$

Table A.1: Field conditions

α, β, and γ. Because the identity elements for addition and multiplication must be different, every field must have at least two elements. Examples include the complex numbers $\mathbb{C}$, rational numbers $\mathbb{Q}$, and real numbers $\mathbb{R}$, but not the integers $\mathbb{Z}$ (not closed under the inversion).

A *vector space* over a field $\mathbb{F}$, denoted $(\mathcal{V}, \mathbb{F})$ (or simply $\mathcal{V}$, when the field is clear from the context), is a set $\mathcal{V}$ together with two operations, addition $+ : \mathcal{V} \times \mathcal{V} \mapsto \mathcal{V}$ and multiplication $\cdot : \mathbb{F} \times \mathcal{V} \mapsto \mathcal{V}$, which satisfies the conditions presented in Table A.2 for both addition and scalar multiplication for all $u, v, w \in \mathcal{V}$ and $\alpha, \beta \in \mathbb{F}$. It is worth emphasizing that this definition says that any vector space is closed

Addition	Scalar multiplication
$u + v = v + u$	
$(u + v) + w = u + (v + w)$	$(\alpha \cdot \beta) \cdot u = \alpha \cdot (\beta \cdot u)$
$\alpha \cdot (u + v) = \alpha \cdot u + \alpha \cdot v$ and $(\alpha + \beta) \cdot u = \alpha \cdot u + \beta \cdot u$	
$\exists \mathbf{0} \in \mathcal{V}$ such that $u + \mathbf{0} = u$	$0 \cdot u = \mathbf{0}$ and $1 \cdot u = u$

Table A.2: Vector space conditions

under the addition and the multiplication by scalars operations. The boldface notation **0**, used in Table A.2 to distinguish the zero vector from $\mathcal{V}$ from the zero scalar $0 \in \mathbb{F}$, is dropped hereafter. The meaning of "0" is typically clear from the context. Also, the notation "·" is dropped from the scalar multiplication, so that we write $\alpha\beta$ or αu to mean $\alpha \cdot \beta$ or $\alpha \cdot u$, respectively. Examples of vector spaces include the real vector space $(\mathbb{R}^m, \mathbb{R})$ (or simply $\mathbb{R}^m$), the complex vector space $(\mathbb{C}^m, \mathbb{C})$ (or simply $\mathbb{C}^m$), etc.

A.1.2 Measuring sizes and angles

The introduction of the notion "norm" is driven by the need to measure sizes and distances. A function $\|\cdot\| : \mathcal{V} \mapsto \mathbb{R}$ is called a *norm* if it satisfies the following three conditions:

1. $\|v\| \geq 0$ and $\|v\| = 0 \iff v = 0$ (positive definiteness)
2. $\|\alpha v\| = |\alpha|\|v\|, \forall \alpha \in \mathbb{F}$ (homogeneity)
3. $\|u + v\| \leq \|u\| + \|v\|$ (triangle inequality)

for all elements $u, v \in \mathcal{V}$. The conditions above can be interpreted as follows. The first condition says that size cannot be negative and only zero element might have zero size; the second condition states that scaling an element should result in the same scaling of its size; and the third condition is actually a generalization of the Euclidean axiom that the shortest distance between two points is the straight line. Note that there may be many norms for a given vector space $\mathcal{V}$. To distinguish between different norms, they may be indexed, like $\|\cdot\|_q$ norms on $\mathbb{C}^m$ defined in Section 2.2.1. When a vector space is endowed with a norm it becomes a *normed vector space*. A *complete* normed space, i.e. a normed space in which each Cauchy sequence converges to an element of this space, is called a *Banach space*. Examples of normed vector spaces are $\mathbb{C}^m$ with the norm $\|v\| = \sqrt{|v_1|^2 + \cdots + |v_m|^2}$ (called the Euclidean norm), the spaces $C[0, 1]$ of continuous $\mathbb{C}$-valued functions in $[0, 1]$ with the norm $\|v\| = \left(\int_0^1 |v(t)|^2 \mathrm{d}t\right)^{1/2}$, and the space $L_2[0, 1]$ of measurable square-integrable $\mathbb{C}$-valued functions in $[0, 1]$ with the same norm. The first and the third of these normed spaces are Banach, whereas the second one is not, as we can construct a sequence of continuous functions converging to a discontinuous function.

The introduction of the notion of "inner product" may be thought of as driven by the need to measure angles between two vectors. A function $\langle\cdot,\cdot\rangle : \mathcal{V} \times \mathcal{V} \mapsto \mathbb{F}$ is called an *inner product* if it satisfies the following conditions:

1. $\langle v, v\rangle \geq 0$ and $\langle v, v\rangle = 0 \iff v = 0$ (positive definiteness)
2. $\langle \alpha u + \beta v, w\rangle = \alpha\langle u, w\rangle + \beta\langle v, w\rangle, \forall \alpha, \beta \in \mathbb{F}$ (semi-linearity)
3. $\langle u, v\rangle = \overline{\langle v, u\rangle}$ (symmetry)

for all elements $u, v, w \in \mathcal{V}$. It also follows from the first condition that if $\langle u, v\rangle = 0$, $\forall v \in \mathcal{V}$, then $u = 0$. The last two conditions imply that $\langle u, \alpha v + \beta w\rangle = \bar{\alpha}\langle u, v\rangle + \bar{\beta}\langle u, w\rangle$. The third condition (implicitly) requires that $\langle v, v\rangle \in \mathbb{R}$. An example of the inner product is $\langle x, y\rangle := x_1 y_1 + x_2 y_2$ defined on $\mathbb{R}^2$ (this quantity is also called the dot or scalar product). It can be shown that in this case $\langle x, y\rangle = \sqrt{\langle x, x\rangle}\sqrt{\langle y, y\rangle}\cos\theta$, where θ is the angle between the vectors x and y. In other words, the scalar product is the scaled cosine of the angle between two vectors in this case. This interpretation can be continued to the general case. To this end, the following result is required.

Theorem A.1 (Cauchy–Schwarz Inequality)**.** *If $\mathcal{V}$ be a vector space with an inner product $\langle\cdot,\cdot\rangle$, then*

$$|\langle u, v\rangle|^2 \leq \langle u, u\rangle\langle v, v\rangle \tag{A.1}$$

for all $u, v \in \mathcal{V}$.

Proof. First, inequality (A.1) is obviously true for $v = 0$ (in which case it is actually an equality). Assume now that $v \neq 0$, i.e. $\langle v, v\rangle > 0$. Clearly, for any $\alpha \in \mathbb{F}$,

$$0 \leq \langle u - \alpha v, u - \alpha v\rangle = \langle u, u\rangle - \alpha\langle v, u\rangle - \overline{\alpha}\langle u, v\rangle + |\alpha|^2\langle v, v\rangle.$$

Choose $\alpha = \langle u, v\rangle/\langle v, v\rangle$, in which case $\overline{\alpha} = \langle v, u\rangle/\langle v, v\rangle$. The inequality above writes then as

$$0 \leq \langle u, u\rangle - |\langle v, u\rangle|^2/\langle v, v\rangle,$$

whence (A.1) follows immediately. □

Now, define

$$\varkappa_{uv} := \begin{cases} \frac{\langle u,v\rangle}{\sqrt{\langle u,u\rangle}\sqrt{\langle v,v\rangle}} & \text{if } v \neq 0 \text{ and } u \neq 0 \\ 0 & \text{otherwise} \end{cases} \tag{A.2}$$

It follows from the Cauchy–Schwarz inequality that $|\varkappa_{uv}| \leq 1$. Also, for any given $\alpha \in \mathbb{R} \setminus \{0\}$ we have that $\varkappa_{(\alpha u)u} = \operatorname{sign}(\alpha)$. This prompts regarding $\varkappa_{uv}$ as the cosine of the "angle" between u and v. We then may extend the notion of the orthogonality to abstract spaces. Namely, $u, v \in \mathcal{V}$ are said to be *orthogonal*, denoted by $u \perp v$, if $\langle u, v\rangle = 0$, i.e. if the "angle" between them is $\pm\pi/2$.

When a vector space is endowed with an inner product, it becomes an *inner product space*. A complete inner product space is called a *Hilbert space*. All three examples of normed spaces above are also inner product spaces with $\langle u, v\rangle = \overline{v_1}u_1 + \cdots + \overline{v_m}u_m$ for $\mathbb{C}^m$ and $\langle u, v\rangle = \int_0^1 \overline{v(t)}u(t)\,\mathrm{d}t$ for $C[0, 1]$ and $L_2[0, 1]$. A remarkable fact is that the inner product *generates* a norm (i.e. every inner product space is a normed space), namely $\|v\| := \sqrt{\langle v, v\rangle}$. To show that this is indeed a norm, it suffices to show the triangle inequality (the positive definiteness and the homogeneity follow directly from the properties of the inner product). To this end, note that

$$\begin{aligned} \|u + v\|^2 &= \langle u + v, u + v\rangle = \langle u, u\rangle + \langle u, v\rangle + \langle v, u\rangle + \langle v, v\rangle \\ &= \|u\|^2 + \langle u, v\rangle + \overline{\langle u, v\rangle} + \|v\|^2 \\ &\leq \|u\|^2 + 2|\langle u, v\rangle| + \|v\|^2 \leq \|u\|^2 + 2\|u\|\|v\| + \|v\|^2 = (\|u\| + \|v\|)^2, \end{aligned}$$

where the Cauchy–Schwarz inequality was used in the second inequality. Taking the square roots of both sides gives the triangle inequality. The equality here would require that both $\langle u, v\rangle + \overline{\langle u, v\rangle} = 2|\langle u, v\rangle|$ (i.e. $\operatorname{Re}\langle u, v\rangle \geq 0$ and $\operatorname{Im}\langle u, v\rangle = 0$) and $|\langle u, v\rangle| = \|u\|\|v\|$ (i.e. that the "angle" between u and v is zero).

We thus saw that the inner product defines the norm. It turns out that the inner product can be recovered from the corresponding norm:

- $4\langle u, v\rangle = \|u + v\|^2 - \|u - v\|^2 + \mathrm{j}\|u + \mathrm{j}v\|^2 - \mathrm{j}\|u - \mathrm{j}v\|^2$ (polarization identity)

This equality can be proved by substitution.

The following two facts, which hold for any inner product space with the norm generated by the inner product and can be easily checked by direct substitution, are abstract generalizations of the classical geometrical notions:

- if $u \perp v$, then $\|u + v\|^2 = \|u\|^2 + \|v\|^2$ (Pythagoras' theorem)
- $\|u + v\|^2 + \|u - v\|^2 = 2\|u\|^2 + 2\|v\|^2$ (parallelogram law)

Curiously, it can be shown that for any normed space, the norm on which satisfies the parallelogram law, an inner product generating this norm can be defined.

A.1.3 Subspaces and linear combination

A subset $\mathcal{S}$ of a vector space $(\mathcal{V}, \mathbb{F})$ is a *subspace* if $\alpha x + \beta y \in \mathcal{S}$ for all $x, y \in \mathcal{S}$ and all $\alpha, \beta \in \mathbb{F}$. In other words, $\mathcal{S}$ is closed under addition and scalar multiplication and hence it itself is a vector space over $\mathbb{F}$. Clearly, both $\{0\}$ (aka the *zero space*) and the whole $\mathcal{V}$ are subspaces. A subspace $\mathcal{S}$ of $\mathcal{V}$ is said to be *proper* if $\mathcal{S} \neq \mathcal{V}$. Given any two subspaces $\mathcal{S}_1$ and $\mathcal{S}_2$, their *intersection*,

$$\mathcal{S}_1 \cap \mathcal{S}_2 := \big\{x \mid x \in \mathcal{S}_1 \text{ and } x \in \mathcal{S}_2\big\},$$

and *sum*,

$$\mathcal{S}_1 + \mathcal{S}_2 := \big\{x \mid x = x_1 + x_2, x_1 \in \mathcal{S}_1, x_2 \in \mathcal{S}_2\big\},$$

are both subspaces as well (if $\mathcal{S}_1 \cap \mathcal{S}_2 = \{0\}$, then their sum is said to be the *internal direct sum* and is written $\mathcal{S}_1 \oplus \mathcal{S}_2$). A geometric intuition of subspaces is that they are hyperplanes passing through the origin.

Let $\mathcal{S}$ be a subspace of an inner product space $\mathcal{V}$. The *orthogonal complement* of $\mathcal{S}$ in $\mathcal{V}$ is the set

$$\mathcal{S}^\perp := \{x \in \mathcal{V} \mid x \perp y, \forall y \in \mathcal{S}\}.$$

The orthogonal complement is an inner product space itself. Moreover, if $\mathcal{S}$ is closed, then $(\mathcal{S}^\perp)^\perp = \mathcal{S}$ and $\mathcal{V} = \mathcal{S} \oplus \mathcal{S}^\perp$. The latter equality means that any $u \in \mathcal{V}$ can be uniquely decomposed as $u = u_1 + u_2$, where $u_1 \in \mathcal{S}$ and $u_2 \in \mathcal{S}^\perp$ and $\|u\|^2 = \|u_1\|^2 + \|u_2\|^2$.

Given vectors $v_1, \ldots, v_m$, any vector of the form $\alpha_1 v_1 + \cdots + \alpha_m v_m$ for some scalars $\alpha_1, \ldots, \alpha_m$ is a *linear combination* of $v_1, \ldots, v_m$. Any linear combination of elements of a vector space $\mathcal{V}$ belongs to this space. The set of *all* linear combinations of $v_1, \ldots, v_m$ is called *span* of $v_1, \ldots, v_m$ and denoted as

$$\operatorname{span}(v_1, \ldots, v_k) := \big\{v \mid v = \alpha_1 v_1 + \cdots + \alpha_m v_m \text{ for some } \alpha_1, \ldots, \alpha_m \in \mathbb{F}\big\}.$$

Clearly, $\operatorname{span}(v_1, \ldots, v_k)$ is a subspace.

A set of vectors $\{v_1, \ldots, v_m\}$ is said to be *linearly independent* if

$$\alpha_1 v_1 + \cdots + \alpha_m v_m = 0 \quad \Longrightarrow \quad \alpha_1 = \cdots = \alpha_m = 0;$$

otherwise this set is said to be *linearly dependent* (clearly, any set containing 0 is linearly dependent). It follows from the definition that the following three conditions are equivalent:

1. $\{v_1, \ldots, v_m\}$ is linearly independent;
2. coefficients of any linear combination of $v_1, \ldots, v_m$ are uniquely determined, i.e.

$$\alpha_1 v_1 + \cdots + \alpha_m v_m = \beta_1 v_1 + \cdots + \beta_m v_m \quad \Longrightarrow \quad \alpha_i = \beta_i, \ \forall i = 1, \ldots, m;$$

3. no v_i can be expressed as a linear combination of the other vectors from $\{v_1, \ldots, v_m\}$.

A.1.4 Basis and dimension

Let $\mathcal{V}$ be a vector space. A (finite) set of vectors $\{v_1, \ldots, v_m\} \in \mathcal{V}$ is a *basis* for $\mathcal{V}$ if $v_1, \ldots, v_m$ are linearly independent and $\mathcal{V} = \operatorname{span}(v_1, \ldots, v_m)$. In particular, if $\{v_1, \ldots, v_m\}$ is a basis for $\mathcal{V}$, then *every* vector $v \in \mathcal{V}$ can be expressed as a linear combination of $v_1, \ldots, v_m$.

A basis for any nonzero space is not unique. This can be seen by replacing any two vectors v_i and v_j in a basis $\{v_1, \ldots, v_m\}$ with $v_j + v_j$ and $v_j - v_j$. It turns out, however, that *the number of elements* in any basis of $\mathcal{V}$ is the same. This number (i.e. the number of vectors in any basis) is called the *dimension*

of $\mathcal{V}$ and denoted as $\dim \mathcal{V}$. A vector space $\mathcal{V}$ having no basis with finitely many vectors is called *infinite dimensional*.

Since elements of any basis for (an m-dimensional) $\mathcal{V}$ are linearly independent, any vector $v \in \mathcal{V}$ can be uniquely decomposed as a linear combination of a basis for $\mathcal{V}$, i.e.

$$\exists \alpha_i, i = 1, \dots, m, \text{ such that } v = \alpha_1 v_1 + \cdots + \alpha_m v_m.$$

These scalars, i.e. $\alpha_1, \dots, \alpha_m$, are said to be the *coordinates* of v in the given basis.

A basis $\{v_1, \dots, v_m\}$ for an inner product space is said to be *orthogonal* if all its elements are mutually orthogonal, i.e. if $v_i \perp v_j$ whenever $i \neq j$. If, in addition, $\|v_i\| = 1$ for all i, the orthogonal basis is said to be *orthonormal*. An example of the orthonormal basis is the *standard basis* for $\mathbb{C}^m$, $\{e_1, \dots, e_m\}$, where each e_i is defined as the vector whose ith element is 1 and the other elements are 0.

Finding coordinates in an orthonormal basis is a particularly simple task. Indeed, let $\{v_1, \dots, v_m\}$ be an orthonormal basis of an m-dimensional inner product space $\mathcal{V}$. We already know that any vector $v \in \mathcal{V}$ can be uniquely expressed as $v = \alpha_1 v_1 + \cdots + \alpha_m v_m$. Thus,

$$\langle v, v_i \rangle = \sum_{j=1}^{m} \alpha_j \langle v_j, v_i \rangle = \alpha_i \langle v_i, v_i \rangle = \alpha_i$$

or, in other words,

$$v = \sum_{i=1}^{m} \langle v, v_i \rangle v_i. \tag{A.3}$$

This expression immediately leads to the celebrated *Parseval's identity*, saying that any two vectors u and v in $\mathcal{V}$ satisfy

$$\langle u, v \rangle = \sum_{i=1}^{m} \langle u, v_i \rangle \overline{\langle v, v_i \rangle}.$$

In particular, for $u = v$ we have that

$$\|v\|^2 = \sum_{i=1}^{m} |\langle v, v_i \rangle|^2,$$

i.e. the square of the norm on an inner product space equals the sum of squares of the coordinates in an orthonormal basis.

Remark A.1. The analysis above can, in principle, be applied to infinite-dimensional spaces as well, although the math in that case is more delicate. For example, the span of the orthonormal basis might no longer result in $\mathcal{V}$. Rather, only the closure of the span should be equal $\mathcal{V}$. Yet we still may present any element of $\mathcal{V}$ as the infinite-dimensional counterpart of (A.3). Consider, for instance, $\mathcal{V} = L_2[0, 1]$. A possible choice of the orthonormal basis for this space is $\left\{\mathrm{e}^{\mathrm{j}(\theta+2\pi i)t}\right\}_{i \in \mathbb{Z}}$, for any $\theta \in [-\pi, \pi]$. The *Fourier expansion* of an element v of $L_2[0, 1]$ is then

$$v(t) = \sum_{i \in \mathbb{Z}} \alpha_i \mathrm{e}^{\mathrm{j}(\theta+2\pi i)t}, \quad \text{where } \alpha_i = \langle v, v_i \rangle = \int_0^1 \mathrm{e}^{-\mathrm{j}(\theta+2\pi i)t} v(t)\,\mathrm{d}t$$

(best known under $\theta = 0$). One should, however, be careful with the convergence of this Fourier expansion. It might not converge to $v(t)$ pointwise (the *Gibbs phenomenon*), but rather only in the sense of the $L_2[0, 1]$ norm. The reason is that the span of this basis is not $L_2[0, 1]$, as any linear combination of the functions $\mathrm{e}^{\mathrm{j}(\theta+2\pi i)t}$ is continuous, whereas $L_2[0, 1]$ may contain discontinuous functions. ∇

A.2 Linear operators and their properties

Given vector spaces $\mathcal{U}$ and $\mathcal{Y}$, both over a field $\mathbb{F}$, an operator T is a mapping of a vector from $\mathcal{U}$ to a unique vector in $\mathcal{Y}$. An operator is not necessarily defined for each element of $\mathcal{U}$, but rather on its subspace $\mathfrak{D}_T \subset \mathcal{U}$, called the *domain* of T. A compact form of writing the said mapping is $T : \mathfrak{D}_T \subset \mathcal{U} \mapsto \mathcal{Y}$ or in the simplified form $T : \mathcal{U} \to \mathcal{Y}$ if the domain is not important in a given context or is the whole $\mathcal{U}$. An operator T is said to be *linear* if the superposition property holds, i.e. if

$$T(\alpha_1 u_1 + \alpha_2 u_2) = \alpha_1 T u_1 + \alpha_2 T u_2 \tag{A.4}$$

for all $\alpha_1, \alpha_2 \in \mathbb{F}$ and all $u_1, u_2 \in \mathfrak{D}_T$. The superposition is sometimes presented as the combination of the additivity, $T(u_1 + u_2) = Tu_1 + Tu_2$, and homogeneity, $T(\alpha u) = \alpha T u$, properties.

The sum of operators and the multiplication of an operator by a scalar can be easily deduced from the definition. Another operation that can be naturally defined for linear operators is their product (the cascade). Let $T : \mathfrak{D}_T \subset \mathcal{U} \mapsto \mathcal{V}$ and $S : \mathfrak{D}_S \subset \mathcal{V} \mapsto \mathcal{Y}$ be such that the set of all possible outputs of T belongs to $\mathfrak{D}_S$. In this case, $ST : \mathfrak{D}_T \subset \mathcal{U} \mapsto \mathcal{Y}$ is the operator such that $(ST)u = S(Tu)$. The identity operator $I : \mathcal{U} \mapsto \mathcal{U}$ is defined as the operator satisfying $Iu = u$ for all $u \in \mathcal{U}$.

A.2.1 Structural properties

The set of all possible "outputs" of an operator $T : \mathcal{U} \mapsto \mathcal{Y}$ is a subspace of $\mathcal{Y}$, as follows from the very definition of the linear operator. This subspace is called the *image* (or range) of T and denoted as

$$\operatorname{Im} T := \{y \in \mathcal{Y} \mid \exists u \in \mathfrak{D}_T \text{ such that } y = Tu\}.$$

It is also possible to write $T\mathcal{U}$ to describe the image of T. Accordingly, sometimes we may write $T\mathcal{S}$ to denote the set of all possible images of a set $\mathcal{S} \subset \mathcal{U}$ under the linear transformation T, even when $\mathcal{S}$ is not a subspace. A mapping $T : \mathcal{U} \mapsto \mathcal{Y}$ is *surjective* (onto) if $\operatorname{Im} T = \mathcal{Y}$. The dimension of $\operatorname{Im} T$ is called the *rank* of the operator T and is denoted as $\operatorname{rank}(T) := \dim(\operatorname{Im} T)$.

The subset of $\mathcal{U}$, elements of which are nullified by T is also a subspace $\mathcal{U}$. This subspace is called the *kernel* (or the null space) of T. More precisely, it is defined as

$$\ker T := \{u \in \mathfrak{D}_T \mid Tu = 0\}.$$

If the kernel of T is trivial, i.e. $\ker T = \{0\}$, then T is called *injective*. If T is both injective and surjective, it is said to be *bijective*.

An operator $T : \mathcal{U} \mapsto \mathcal{U}$ is *invertible* if there is a mapping $S : \mathcal{U} \mapsto \mathcal{U}$ such that $STu = u = TSu$ for all u from $\mathcal{U}$. This S, which is also a linear operator, is called the *inverse* of T and is denoted as T^{-1}. It is a known result that T is invertible iff it is bijective, i.e. iff $\ker T = \{0\}$ and $\operatorname{Im} T = \mathcal{U}$.

A.2.2 Operators on normed spaces

Consider now linear operators over normed vector spaces. Let $\mathcal{U}$ and $\mathcal{Y}$ be normed spaces with corresponding norms (to simplify the notation, we use the symbol $\|\cdot\|$ for both norms, even though norms on $\mathcal{U}$ and $\mathcal{Y}$ may be different). An operator $T : \mathcal{U} \mapsto \mathcal{Y}$ is said to be *bounded* if $\mathfrak{D}_T = \mathcal{U}$ and there is $\gamma > 0$ such that

$$\|Tu\| \le \gamma \|u\|, \quad \forall u \in \mathcal{U}.$$

If T is bounded, the quantity

$$\|T\| := \sup_{u \in \mathcal{U}, u \neq 0} \frac{\|Tu\|}{\|u\|}$$

is well-defined and can be regarded as its norm (*induced norm*). Alternatively, exploiting the linearity of T, the induced norm can be expressed as

$$\|T\| = \sup_{u \in \mathcal{U}, \|u\|=1} \|Tu\|,$$

Induced norms always satisfy the three conditions for a norm on p. 182. In some situations, the induced norm of operators from $\mathcal{U}$ to $\mathcal{Y}$ are denoted as $\|\cdot\|_{\mathcal{U} \mapsto \mathcal{Y}}$.

A.2.3 Operators on inner product spaces

Additional concepts can be introduced for operators over inner product spaces. One of them is the fundamental notion of the *adjoint* operator. Let $\mathcal{U}$ and $\mathcal{Y}$ be inner product spaces with the inner products $\langle \cdot, \cdot \rangle_{\mathcal{U}}$ and $\langle \cdot, \cdot \rangle_{\mathcal{Y}}$, respectively, and their generated norms and let $T : \mathcal{U} \mapsto \mathcal{Y}$ be a bounded linear operator. There exists a unique operator $T' : \mathcal{Y} \mapsto \mathcal{U}$ satisfying

$$\langle Tu, y \rangle_{\mathcal{Y}} = \langle u, T'y \rangle_{\mathcal{U}}, \quad \forall u \in \mathcal{U}, y \in \mathcal{Y}.$$

It is readily seen that $(T')' = T$, $(\alpha S + \beta T)' = \overline{\alpha} S' + \overline{\beta} T'$, and $(ST)' = T'S'$. It can also be proved that $\|T'\| = \|T\|$ and $\|T'T\| = \|T\|^2$.

An operator $T : \mathcal{U} \mapsto \mathcal{U}$ is called *self-adjoint* if $T = T'$. It follows from the symmetry property of the inner product that if T is self-adjoint, $\langle Tu, u \rangle \in \mathbb{R}$ for all $u \in \mathcal{U}$ even if $\mathcal{U}$ is a space over the complex field $\mathbb{C}$. The converse is also true: if $\mathcal{U}$ is a space over $\mathbb{C}$ and $\langle Tu, u \rangle \in \mathbb{R}$ for all $u \in \mathcal{U}$, then $T = T'$. The induced norm of self-adjoint operators can be calculated as $\|T\| = \sup_{\|u\|=1} \langle Tu, u \rangle$. The last two properties imply that $\langle Tu, u \rangle = 0$ for all $u \in \mathcal{U}$ iff $T = 0$. Motivated by this interpretation of the zero operator, we may define sign definite operators via the sign of the corresponding inner products. A self-adjoint operator T is then said to be positive definite (denoted $T > 0$) if $\langle Tu, u \rangle > 0$ and positive semi-definite (denoted $T \geq 0$) if $\langle Tu, u \rangle \geq 0$ for all $u \neq 0$. With these notions, the comparison of self-adjoint operators $T < S$ and $T \leq S$ should be understood as $S - T > 0$ and $S - T \geq 0$, respectively.

Clearly, $T'T$ is self-adjoint for every $T : \mathcal{U} \mapsto \mathcal{Y}$. Because

$$\langle T'Tu, u \rangle = \langle Tu, Tu \rangle = \|Tu\|^2 \geq 0,$$

the operator $T'T \geq 0$. It turns out that the converse is true as well: if $T : \mathcal{U} \mapsto \mathcal{U}$ is a self-adjoint operator such that $T \geq 0$, it can be factorized as $T = S'S$ for some $S : \mathcal{U} \mapsto \mathcal{Y}$ ($\mathcal{Y}$ may be different from $\mathcal{U}$). Such an S is not unique. There is, however, a unique factor, satisfying $S = S' \geq 0$. This factor is called the *square root* of T and denoted as $T^{1/2}$. Note that $T^{1/2} > 0$ iff $T > 0$.

The notion of sign-definite operators can be used in the calculation of the operator norm, induced by the inner product vector norms. The following result is important.

Theorem A.2. *If $T : \mathcal{U} \mapsto \mathcal{Y}$ is a linear operator over inner product spaces, then*

$$\|T\| < \gamma \iff T'T < \gamma^2 I \iff TT' < \gamma^2 I.$$

Proof. Clearly, $\|T\| < \gamma \iff \|Tu\|^2 < \|\gamma u\|^2 \iff \langle Tu, Tu \rangle < \langle \gamma u, \gamma u \rangle$ for all $u \in \mathcal{U}$. The latter inequality is equivalent to

$$0 < \langle \gamma^2 u, u \rangle - \langle T'Tu, u \rangle = \langle (\gamma^2 I - T'T)u, u \rangle, \quad \forall u \in \mathcal{U},$$

whence the first part follows. The second part follows by the fact that $\|T\| = \|T'\|$. □

A.2.4 Matrix form of linear operators

In some situations, it is convenient to represent linear operators as rectangular tables, known as *matrices*, of elements from $\mathbb{F}$. To this end, assume that $\dim \mathcal{U} = m$ and $\dim \mathcal{Y} = p$ and bring in their bases $\{u_1, \ldots, u_m\}$ and $\{y_1, \ldots, y_p\}$, respectively. Consider now a linear operator $T : \mathcal{U} \mapsto \mathcal{Y}$ and denote by t_{ij} the ith coordinate of Tu_j in the chosen basis of $\mathcal{Y}$ (note that it is uniquely determined). By linearity,

$$\begin{aligned} T(\alpha_1 u_1 + \cdots + \alpha_m u_m) &= \alpha_1 T u_1 + \cdots + \alpha_m T u_m = \alpha_1 \sum_{i=1}^{p} t_{i1} y_i + \cdots + \alpha_m \sum_{i=1}^{p} t_{im} y_i \\ &= \left(\sum_{j=1}^{m} \alpha_j t_{1j} \right) y_1 + \cdots + \left(\sum_{j=1}^{m} \alpha_j t_{pj} \right) y_p = \beta_1 y_1 + \cdots + \beta_p y_p. \end{aligned}$$

This means that the relation between the coordinates α_i and β_i of u and Tu, respectively, can always be written in the form

$$\begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} = \begin{bmatrix} t_{11} & \cdots & t_{1m} \\ \vdots & \ddots & \vdots \\ t_{p1} & \cdots & t_{pm} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_m \end{bmatrix}.$$

The $p \times m$ matrix on the right-hand side above is called the *matrix representation* of T in the given basis and denoted as $[\![T]\!]_{\{u_i\},\{y_i\}}$, or simply $[\![T]\!]$ when the bases are irrelevant or clear from the context. It is worth emphasizing that although the term "operator" is frequently interchanged with the term "operator matrix," strictly speaking $[\![T]\!] \in \mathbb{F}^{p \times m}$ is not $T : \mathcal{U} \mapsto \mathcal{Y}$ itself. While the matrix representation $[\![T]\!]_{\{u_i\},\{y_i\}}$ does depend on the chosen basis of $\mathcal{U}$ and $\mathcal{Y}$, the operator T itself does not.

Still, many properties of matrix representations are coordinate-independent. For example, it is readily seen that T is invertible iff its system matrix $[\![T]\!]$ is invertible. It can also be shown that

$$\operatorname{Im} T = \operatorname{span}(t_{1\bullet}, \ldots, t_{p\bullet}),$$

where $t_{i\bullet}$ stands for the ith row of $[\![T]\!]$ in any bases.

This idea can be extended to operators over infinite-dimensional spaces as well, in which case matrix representations are matrices with infinitely many rows and / or columns. However, a special care should be taken in such extensions for the convergence of infinite sums.

Appendix B

Matrix Equations and Manipulations

MATRIX EQUATIONS play an important role in the state-space analysis and design of finite-dimensional linear systems. Linear matrix equations (Sylvester and Lyapunov equations) are an important analysis tool, whereas quadratic equations (Riccati) are crucial in numerous design problems, like H_2 and H_∞ optimization and robust control. This appendix discusses some basic properties of these equations. It also contains the definition of the Schur complement of block matrices and some related formulae.

B.1 Linear matrix equations (Sylvester & Lyapunov)

Let $A_1 \in \mathbb{R}^{n_1 \times n_1}$, $A_2 \in \mathbb{R}^{n_2 \times n_2}$, and $Q \in \mathbb{R}^{n_1 \times n_2}$. The following linear matrix equation:

$$A_1 X - X A_2 + Q = 0 \tag{B.1}$$

is called the *Sylvester equation*. It has a unique solution $X \in \mathbb{R}^{n_1 \times n_2}$ iff

$$\operatorname{spec}(A_1) \cap \operatorname{spec}(A_2) = \varnothing,$$

i.e. iff none of the eigenvalues of A_1 is also an eigenvalue of A_2. If this condition fails to hold, then the Sylvester equation might have either no solutions or an infinite number of solutions (depending on Q). The following result establishes the connection between the solvability of (B.1) with the block-diagonalizability of certain block-triangular matrices.

Proposition B.1 (Roth's removal rule)**.** *Equation* (B.1) *is solvable iff the matrices*

$$\begin{bmatrix} A_1 & Q \\ 0 & A_2 \end{bmatrix} \quad \textit{and} \quad \begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix}$$

are similar.

An important particular case of the Sylvester equation is the so-called (continuous-time) *Lyapunov equation*, which is defined as

$$AX + XA' + Q = 0 \tag{B.2}$$

for given $A \in \mathbb{R}^{n \times n}$ and $Q \in \mathbb{R}^{n \times n}$. If Q is symmetric (i.e. $Q = Q'$), then the solution X is symmetric too. Furthermore, if A is Hurwitz, i.e. if A has all its eigenvalues in the open left half-plane $\mathbb{C} \setminus \overline{\mathbb{C}}_0$, then

$$X = \int_{\mathbb{R}_+} \mathrm{e}^{At} Q \mathrm{e}^{A't} \, \mathrm{d}t \tag{B.3}$$

exists and is the solution of (B.2). Indeed, it is readily seen that

$$A\,\mathrm{e}^{At}Q\mathrm{e}^{A't} + \mathrm{e}^{At}Q\mathrm{e}^{A't}A' = \frac{\mathrm{d}}{\mathrm{d}t}\big(\mathrm{e}^{At}Q\mathrm{e}^{A't}\big).$$

Thus, if the integral in (B.3) exists,

$$A\int_{\mathbb{R}_+}\mathrm{e}^{At}Q\mathrm{e}^{A't}\,\mathrm{d}t + \int_{\mathbb{R}_+}\mathrm{e}^{At}Q\mathrm{e}^{A't}\,\mathrm{d}t\,A' + Q = \int_{\mathbb{R}_+}\mathrm{d}\big(\mathrm{e}^{At}Q\mathrm{e}^{A't}\big) + Q = 0,$$

where the fact that $\lim_{t\to\infty}\mathrm{e}^{At}Q\mathrm{e}^{A't} = 0$ was used.

B.1.1 Lyapunov equations and stability

The Lyapunov equation plays an important role in control and systems theory. Some examples can be found in §4.2.1, §4.3.4, and §4.4.2. Another example is the connection between the existence of positive definite solution of a Lyapunov equation and the stability of matrices.

Proposition B.2. *A matrix $A \in \mathbb{R}^{n\times n}$ is Hurwitz iff the solution $X \in \mathbb{R}^{n\times n}$ of the Lyapunov equation* (B.2) *satisfies $X = X' > 0$ whenever $Q = Q' > 0$.*

Proof. First, assume that A is Hurwitz. Clearly, (A, Q) is controllable (follows from the non-singularity of Q), so that $X > 0$ by (B.3). Now, let $X > 0$ for some $Q > 0$. Assume that A is not Hurwitz, i.e. that it has an eigenvalue λ such that $\operatorname{Re}\lambda \geq 0$. Denote the corresponding eigenvector by $\eta \neq 0$ and pre- and post-multiply (B.2) by η' and η, respectively. We have:

$$-\eta' Q\eta = \eta'(AX + XA')\eta = \lambda\eta' X\eta + \overline{\lambda}\eta' X\eta = 2\operatorname{Re}\lambda\eta' X\eta.$$

This, in turn, implies that $\operatorname{Re}\lambda\eta' X\eta < 0$, which is a contradiction. □

Proposition B.2 actually says that the stability of A is equivalent to the existence of a matrix $X = X' > 0$ such that

$$AX + XA' < 0. \tag{B.4}$$

Inequality (B.4) belongs to the so-called class of *Linear Matrix Inequalities* (LMI), for the verification of which efficient numerical methods are available. For this reason (B.4) can be considered an alternative to the conventional verification of eigenvalues of A. More important is that the LMI (B.4) can be incorporated into many other analysis and design problems that also reduce to LMIs.

B.1.2 Lyapunov equations and Hankel norm

Another use of the Lyapunov equation in systems analysis is in computing the Hankel norm of causal finite-dimensional LTI systems. Given a stable linear system G, its *Hankel norm* $\|G\|_{\mathrm{H}}$ is defined as

$$\|G\|_{\mathrm{H}} := \sup_{u\in L_{2-},\|u\|_2=1}\|(Gu)_+\|_2, \tag{B.5}$$

where v_+ stands for the orthogonal projection of $v \in L_2$ onto L_{2+}, i.e. it is the induced norm of the Hankel operator $\mathfrak{H}_G : L_2(\mathbb{R}_-) \to L_2(\mathbb{R}_+)$ associated with G.

To compute the Hankel norm of a finite-dimensional LTI system, we need to remember the fundamental (in fact, defining) property of the state vector at any time instance t to accumulate the input history up to t. This implies that the response of G in the time interval $\mathbb{R}_+$ to any input signal having support in $\mathbb{R}_-$ is

completely determined by the state vector of G at $t = 0$, say $x(0) = x_0$. If G has the (minimal) state space realization

$$G(s) = \left[\begin{array}{c|c} A & B \\ \hline C & 0 \end{array}\right],$$

then its output response is $y(t) = C\mathrm{e}^{At}x_0$ and its energy in $\mathbb{R}_+$ is

$$\|Gu\|_2^2 = \int_{\mathbb{R}_+} x_0'\mathrm{e}^{A't}C'C\mathrm{e}^{At}x_0\,\mathrm{d}t = x_0'Qx_0,$$

where $Q = Q' > 0$ is the observability Gramian of the pair (C, A), which satisfies the Lyapunov equation $A'Q + QA + C'C = 0$ (see p. 74).

Now, $x(0)$ depends on u according to the following law:

$$x(0) = \int_{\mathbb{R}_-} \mathrm{e}^{-At}Bu(t)\,\mathrm{d}t. \tag{B.6}$$

It is known that there is an infinite number of inputs satisfying this equation for every $x(0) = x_0$ (mind the controllability of (A, B)). Among all these inputs we are actually interested in that having the minimal energy. Indeed, because for a given x_0 the numerator in (B.5) remains the same, the minimal energy u maximizes the ratio in (B.5). The following lemma yields this u.

Lemma B.3. *Let $P = P' > 0$ be the controllability Gramian of the pair (A, B) satisfying the Lyapunov equation $PA' + AP + BB' = 0$ (see p. 71). The signal*

$$u(t) = u_{min}(t) := B'\mathrm{e}^{-A't}P^{-1}x_0,$$

is the unique minimum-energy input rendering $x(0) = x_0$.

Proof. First, prove that $u_{\min}$ does render $x(0) = x_0$. To this end, substitute it to the right-hand side of (B.6). We have:

$$x(0) = \int_{\mathbb{R}_-} \mathrm{e}^{-At}Bu_{\min}(t)\,\mathrm{d}t = \int_{\mathbb{R}_-} \mathrm{e}^{-At}BB'\mathrm{e}^{-At}\,\mathrm{d}tP^{-1}x_0 = \int_{\mathbb{R}_+} \mathrm{e}^{At}BB'\mathrm{e}^{At}\,\mathrm{d}tP^{-1}x_0 = x_0$$

indeed. Now, define $u_\delta := u - u_{\min}$. It is readily seen that u satisfies (B.6) for $x(0) = x_0$ iff u_δ satisfies

$$\int_{\mathbb{R}_-} \mathrm{e}^{-At}Bu_\delta(t)\,\mathrm{d}t = 0.$$

Hence, if $u \in L_{2-}$ renders $x(0) = x_0$, then its energy is

$$\begin{aligned}\|u\|_2^2 &= \langle u_{\min} + u_\delta, u_{\min} + u_\delta\rangle_2 \\ &= \|u_{\min}\|_2^2 + \|u_\delta\|_2^2 + 2\langle u_\delta, u_{\min}\rangle_2 = \|u_{\min}\|_2^2 + \|u_\delta\|_2^2 + 2x_0'P^{-1}\int_{\mathbb{R}_-} \mathrm{e}^{-At}Bu_\delta(t)\,\mathrm{d}t \\ &= \|u_{\min}\|_2^2 + \|u_\delta\|_2^2.\end{aligned}$$

It is then clear that among all u rendering $x(0) = x_0$, the one with the minimal energy corresponds to $u_\delta = 0$. □

The energy of this input is then

$$\|u_{\min}\|_2^2 = \int_{\mathbb{R}_-} u'(t)u(t)\,\mathrm{d}t = \int_{\mathbb{R}_-} x_0'P^{-1}\mathrm{e}^{-At}BB'\mathrm{e}^{-A't}P^{-1}x_0\,\mathrm{d}t = x_0'P^{-1}x_0.$$

Thus, the Hankel norm can be equivalently calculated via

$$\|G\|_{\mathrm{H}}^2 = \max_{x_0\neq 0} \frac{x_0'Qx_0}{x_0'P^{-1}x_0} = \max_{x_1\neq 0} \frac{x_1'P^{1/2}QP^{1/2}x_1}{x_1'x_1} = \|Q^{1/2}P^{1/2}\|^2 = \rho(P^{1/2}QP^{1/2}) = \rho(PQ),$$

where $\|\cdot\|$ stands for the spectral matrix norm. We thus just proved the following result.

Proposition B.4. *Given a stable and causal finite-dimensional LTI system G with a strictly proper transfer function $G(s)$,*

$$\|G\|_H = \sqrt{\rho(PQ)},$$

where P and Q are the controllability and observability Gramians of G, respectively.

B.2 Quadratic matrix equations (Riccati)

A matrix equation of the form

$$A'X + XA + Q + XRX = 0 \tag{B.7}$$

for some matrices $A \in \mathbb{R}^{n\times n}$, $Q = Q' \in \mathbb{R}^{n\times n}$, and $R = R' \in \mathbb{R}^{n\times n}$ is called the (continuous-time) *algebraic Riccati equation* (ARE). Such equations can be thought of as a matrix extension of the standard quadratic equation $rx^2 + 2ax + q = 0$. Like in the quadratic equation case, a solution X of the ARE of the form (B.7) is not unique. In control applications we are mostly concerned with the so-called *stabilizing* solution, which is defined as the solution for which the matrix $A + RX$ is Hurwitz (stable). The stabilizing solution *is* unique (if exists) as proved below.

Proposition B.5. *If there is a stabilizing solution to* (B.7)*, then it is Hermitian and unique.*

Proof. Let X be a stabilizing solution to (B.7). Denoting $\Delta := X - X'$, we have that

$$\begin{aligned}
0 &= A'X + XA + Q + XRX - (A'X + XA + Q + XRX)' = A'\Delta + \Delta A + XRX - X'RX' \\
&= (A + RX)'\Delta + \Delta(A + RX) + XRX - X'RX' - X'R(X - X') - (X - X')RX \\
&= (A + RX)'\Delta + \Delta(A + RX).
\end{aligned}$$

This is a Lyapunov equation, whose solution by (B.3) is $\Delta = 0$ ($A + RX$ is Hurwitz). Hence, $X = X'$.

Now, assume that there are two stabilizing solutions, $X_1 = X_1'$ and $X_2 = X_2'$, i.e. that

$$A'X_1 + X_1A + Q + X_1RX_1 = 0 \quad \text{and} \quad A'X_2 + X_2A + Q + X_2RX_2 = 0$$

with Hurwitz $A_1 := A + RX_1$ and $A_2 := A + RX_2$. Subtracting the second equation from the first one and denoting $X_\delta := X_1 - X_2$ we have that

$$\begin{aligned}
0 &= A'X_1 + X_1A_1 + Q - (A'X_2 + X_2A_2 + Q) = A'X_\delta + X_1A_1 - X_2A_2 \pm X_2A_1 \\
&= A'X_\delta + X_2(A_1 - A_2) + X_\delta A_1 = A'X_\delta + X_2RX_\delta + X_\delta A_1 \\
&= A_2'X_\delta + X_\delta A_1,
\end{aligned}$$

where the last equality uses the already proved fact that $X_2 = X_2'$. This is a Sylvester equation, whose solution is unique because A_1 and A_2 are Hurwitz. This solution is obviously $X_\delta = 0$, so that $X_1 = X_2$. □

Important for studying AREs, as well as for their numerical solutions, is the fact that solutions of (B.7) can be expressed in terms of the eigenstructure of certain matrices. To see this, note that (B.7) can be equivalently written as

$$\begin{bmatrix} A & R \\ -Q & -A' \end{bmatrix}\begin{bmatrix} I \\ X \end{bmatrix} = \begin{bmatrix} I \\ X \end{bmatrix}(A + RX) \tag{B.8}$$

(here the first row is an obvious equality, whereas the second row is just a rewritten (B.7)). This equation is actually reminiscent of the eigenvalue equation. Connections indeed exist. To see them, denote

$$H_{\text{RIC}} := \begin{bmatrix} A & R \\ -Q & -A' \end{bmatrix} \in \mathbb{R}^{2n\times 2n}. \tag{B.9}$$

Let η be an eigenvector of $A + RX$ corresponding to an eigenvalue λ. Post-multiplying (B.8) by η and denoting $\tilde{\eta} := \left[\begin{smallmatrix} I \\ X \end{smallmatrix}\right]\eta$, we obtain that $H_{\text{RIC}}\tilde{\eta} = \lambda\tilde{\eta}$. This implies that any eigenvalue of the $n \times n$ matrix $A + RX$ is an eigenvalue of the $2n \times 2n$ matrix H_{RIC} as well.

To understand properties of H_{RIC}, introduce the $2n \times 2n$ skew-symmetric matrix

$$J := \begin{bmatrix} 0 & -I \\ I & 0 \end{bmatrix}. \tag{B.10}$$

It is readily verified that $J^{-1} = -J = J'$, so J is also unitary. The direct substitution then shows that

$$J^{-1}H_{\text{RIC}}J = -H'_{\text{RIC}} \tag{B.11}$$

(such matrices are called *Hamiltonian*). This implies that H_{RIC} and $-H'_{\text{RIC}}$ are similar. Hence, λ is an eigenvalue of H_{RIC} iff its imaginary axis mirror, $-\bar{\lambda}$, is an eigenvalue of H_{RIC}.

Now, if $A+RX$ is Hurwitz, H_{RIC} should have n eigenvalues in $\mathbb{C}\setminus\bar{\mathbb{C}}_0$. This, in turn, means that the other n eigenvalues of H_{RIC} must be in $\mathbb{C}_0$ and this partition of the spectrum of H_{RIC} (n eigenvalues in $\mathbb{C}\setminus\bar{\mathbb{C}}_0$ and the other n are in the $\mathbb{C}_0$) is unique. Of course, this partition is possible iff H_{RIC} has no $\mathrm{j}\omega$-axis eigenvalues, which is thus necessary for the ARE (B.7) to have a stabilizing solution. This is not sufficient though. Some insight why is this the case can be gained through the reexamination of the eigenvector structure of H_{RIC}. Indeed, it follows from the formula $\tilde{\eta} = \left[\begin{smallmatrix} \eta \\ X\eta \end{smallmatrix}\right]$ and the fact that η is a (non-zero) eigenvector of $A + RX$ that at least one of the first n components of $\tilde{\eta}$ must be non-zero. This is not necessarily true for all stable (i.e. those corresponding to stable eigenvalues) eigenvectors of Hamiltonian matrices as can be seen in the following simple example:

$$H_{\text{RIC}} = \left[\begin{array}{cc:cc} -1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ \hdashline -1 & -1 & 1 & 0 \\ -1 & -1 & 0 & -1 \end{array}\right], \text{ for which } \operatorname{spec}(H_{\text{RIC}}) = \{\pm 1, \pm\sqrt{2}\} \text{ and } \lambda = -1 \text{ has } \tilde{\eta} = \left[\begin{array}{c} 0 \\ 0 \\ \hdashline 0 \\ 1 \end{array}\right]. \tag{B.12}$$

The corresponding ARE has no stabilizing solutions indeed, which can be easily seen from the fact that the pair $(A, R) = \left(\left[\begin{smallmatrix} -1 & 0 \\ 0 & 1 \end{smallmatrix}\right], \left[\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right]\right)$ is not stabilizable, so that there is no X for which $A + RX$ is stable.

For general R, there might not be an easy criterion, in terms of A, Q, and R, for the existence of a stabilizing solution X to (B.7). The situation is more transparent for sign semi-definite R's as the following result shows.

Theorem B.6. *If either $R \geq 0$ or $R \leq 0$, then there is a stabilizing solution of the ARE* (B.7) *iff H_{RIC} has no $\mathrm{j}\omega$-axis eigenvalues and the pair (A, R) is stabilizable.*

Proof. The necessity of $\operatorname{spec}(H_{\text{RIC}}) \cap \mathrm{j}\mathbb{R} = \varnothing$ was shown before. Hence, we may assume that H_{RIC} has no $\mathrm{j}\omega$-axis eigenvalues and there is a nonsingular $T \in \mathbb{R}^{2n\times 2n}$ such that

$$T^{-1}H_{\text{RIC}}T = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix}^{-1} \begin{bmatrix} A & R \\ -Q & -A' \end{bmatrix} \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} = \begin{bmatrix} H_{\text{s}} & H_{12} \\ 0 & H_{\bar{\text{s}}} \end{bmatrix},$$

where H_{s} is Hurwitz and $H_{\bar{\text{s}}}$ is anti-Hurwitz (one such T, which is orthogonal, can be obtained by the Schur decomposition of H_{RIC}). Equivalently,

$$\begin{bmatrix} A & R \\ -Q & -A' \end{bmatrix} \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} \begin{bmatrix} H_{\text{s}} & H_{12} \\ 0 & H_{\bar{\text{s}}} \end{bmatrix},$$

from which

$$\begin{bmatrix} A & R \\ -Q & -A' \end{bmatrix} \begin{bmatrix} T_{11} \\ T_{21} \end{bmatrix} = \begin{bmatrix} T_{11} \\ T_{21} \end{bmatrix} H_{\text{s}}. \tag{B.13}$$

Now, show that there is a stabilizing solution iff T_{11} is nonsingular. The "if" part follows by construction, as in this case $X = T_{21}T_{11}^{-1}$ is the required solution with $A + RX = T_{11}H_{\mathrm{s}}T_{11}^{-1}$ (can be verified by a direct substitution). To prove the "only if" part, assume that there is a stabilizing solution X. Eqn. (B.8) can then be complemented to

$$\begin{bmatrix} A & R \\ -Q & -A' \end{bmatrix}\begin{bmatrix} I & 0 \\ X & I \end{bmatrix} = \begin{bmatrix} I & 0 \\ X & I \end{bmatrix}\begin{bmatrix} A+RX & R \\ 0 & -(A'+XR) \end{bmatrix}$$

(as a matter of fact, this shows that $A' + XR$ is Hurwitz because the $(2,2)$ block in the last matrix above should contain all n eigenvalues of H_{RIC} in $\mathbb{C}_0$). This implies that

$$\begin{bmatrix} A+RX & R \\ 0 & -(A'+XR) \end{bmatrix} = \begin{bmatrix} I & 0 \\ -X & I \end{bmatrix}\begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix}\begin{bmatrix} H_{\mathrm{s}} & H_{12} \\ 0 & H_{\bar{\mathrm{s}}} \end{bmatrix}\left(\begin{bmatrix} I & 0 \\ -X & I \end{bmatrix}\begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix}\right)^{-1},$$

which, in turn, leads to

$$\begin{bmatrix} A+RX & R \\ 0 & -(A'+XR) \end{bmatrix}\begin{bmatrix} T_{11} \\ T_{21}-XT_{11} \end{bmatrix} = \begin{bmatrix} T_{11} \\ T_{21}-XT_{11} \end{bmatrix} H_{\mathrm{s}}.$$

The second row above reads $(A'+XR)(T_{21}-XT_{11}) + (T_{21}-XT_{11})H_{\mathrm{s}} = 0$, which is a Sylvester equation having the unique solution (as the matrices $A' + XR$ and H_{s} are Hurwitz) $T_{21} - XT_{11} = 0$. This equality implies that if there is a vector $\eta \neq 0$ such that $T_{11}\eta = 0$, then $T\left[\begin{smallmatrix} \eta \\ 0 \end{smallmatrix}\right] = 0$, which is a contradiction because T is nonsingular.

Next, show that T_{11} is nonsingular iff the pair (A, R) is stabilizable. The "only if" part here is straightforward, otherwise $A + RT_{21}T_{11}^{-1} = T_{11}H_{\mathrm{s}}T_{11}^{-1}$ cannot be stable. The "if" part, i.e. the proof that the stabilizability of (A, R) implies $\det(T_{11}) \neq 0$, is more involved. To show it, assume the opposite, i.e. that $\det(T_{11}) = 0$ despite the stabilizability of (A, R). Pick any $0 \neq \eta \in \ker T_{11}$ and pre- and post-multiply the first row of (B.13) by $\eta' T_{21}'$ and η, respectively. This yields (mind the symmetry of $T_{21}'T_{11}$)

$$\eta' T_{21}'(AT_{11} + RT_{21})\eta = \eta' T_{21}' T_{11} H_{\mathrm{s}}\eta = \eta' T_{11}' T_{21} H_{\mathrm{s}}\eta$$

and then $\eta' T_{21}' R T_{21}\eta = 0$. This is the part where the sign definiteness of R is used as then the last equality implies that $RT_{21}\eta = 0$. This, in turn, leads to $T_{11}H_{\mathrm{s}}\eta = 0$ from the first row of (B.13). Since $\eta \in \ker T_{11}$ is arbitrary, we actually just proved that $H_{\mathrm{s}} \ker T_{11} \subset \ker T_{11}$, i.e. that $\ker T_{11}$ is H_{s}-invariant. It is a known fact that any (nonzero) invariant subspace of a matrix M contains at least one eigenvector of M. So there is a vector $0 \neq \zeta \in \ker T_{11}$ such that $H_{\mathrm{s}}\zeta = \lambda\zeta$ for some $\operatorname{Re}\lambda < 0$. Post-multiplying (B.13) by this ζ we have:

$$\begin{bmatrix} R \\ -A' \end{bmatrix} T_{21}\zeta = \begin{bmatrix} 0 \\ \lambda I \end{bmatrix} T_{21}\zeta \iff \begin{bmatrix} R \\ A'+\lambda I \end{bmatrix} T_{21}\zeta = 0 \iff \zeta' T_{21}' \begin{bmatrix} A-(-\lambda)I & R \end{bmatrix} = 0.$$

The detectability of (A, R) then yields (via the PBH test) that the latter necessarily yields $T_{21}\zeta = 0$. As $T_{11}\zeta = 0$ too and T is nonsingular, we have a contradiction. Hence, $\det(T_{11}) \neq 0$.

Thus, we just proved that there is a stabilizing solution X iff (A, R) is stabilizable and this stabilizing $X = T_{21}T_{11}^{-1}$ for any T_{11} and T_{21} satisfying (B.13). This completes the proof. □

B.3 Schur complement and matrix inversion formulae

Let A be a square matrix partitioned as follows:

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},$$

with square A_{11} and A_{22}. If A_{11} is nonsingular, then it can be verified by a direct substitution that

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} I & 0 \\ A_{21}A_{11}^{-1} & I \end{bmatrix} \begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} - A_{21}A_{11}^{-1}A_{12} \end{bmatrix} \begin{bmatrix} I & A_{11}^{-1}A_{12} \\ 0 & I \end{bmatrix}. \tag{B.14a}$$

It follows from this equality that A is nonsingular iff so is the matrix

$$\Delta_{11} := A_{22} - A_{21}A_{11}^{-1}A_{12},$$

which is called the *Schur complement* of A_{11} in A. Similarly, if A_{22} is nonsingular, then

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} I & A_{12}A_{22}^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} A_{11} - A_{12}A_{22}^{-1}A_{21} & 0 \\ 0 & A_{22} \end{bmatrix} \begin{bmatrix} I & 0 \\ A_{22}^{-1}A_{21} & I \end{bmatrix} \tag{B.14b}$$

and the Schur complement of A_{22} is defined as follows:

$$\Delta_{22} := A_{11} - A_{12}A_{22}^{-1}A_{21}.$$

There are many applications of decompositions (B.14). Below two of them are briefly discussed. The first is related to the inversion of block 2×2 matrices. Namely, using the facts that

$$\begin{bmatrix} I & A \\ 0 & I \end{bmatrix}^{-1} = \begin{bmatrix} I & -A \\ 0 & I \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} I & 0 \\ A & I \end{bmatrix}^{-1} = \begin{bmatrix} I & 0 \\ -A & I \end{bmatrix}$$

(can be verified by direct substitution), we have that

$$\begin{aligned} \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}^{-1} &= \begin{bmatrix} I & -A_{11}^{-1}A_{12} \\ 0 & I \end{bmatrix} \begin{bmatrix} A_{11}^{-1} & 0 \\ 0 & \Delta_{11}^{-1} \end{bmatrix} \begin{bmatrix} I & 0 \\ -A_{21}A_{11}^{-1} & I \end{bmatrix} \quad \text{(if } A_{11} \text{ is nonsingular)} \\ &= \left[\begin{array}{c:c} A_{11}^{-1} + A_{11}^{-1}A_{12}\Delta_{11}^{-1}A_{21}A_{11}^{-1} & -A_{11}^{-1}A_{12}\Delta_{11}^{-1} \\ \hdashline -\Delta_{11}^{-1}A_{21}A_{11}^{-1} & \Delta_{11}^{-1} \end{array}\right] \qquad \text{(B.15a)} \\ &= \begin{bmatrix} I & 0 \\ -A_{22}^{-1}A_{21} & I \end{bmatrix} \begin{bmatrix} \Delta_{22}^{-1} & 0 \\ 0 & A_{22}^{-1} \end{bmatrix} \begin{bmatrix} I & -A_{12}A_{22}^{-1} \\ 0 & I \end{bmatrix} \quad \text{(if } A_{22} \text{ is nonsingular)} \\ &= \left[\begin{array}{c:c} \Delta_{22}^{-1} & -\Delta_{22}^{-1}A_{12}A_{22}^{-1} \\ \hdashline -A_{22}^{-1}A_{21}\Delta_{22}^{-1} & A_{22}^{-1} + A_{22}^{-1}A_{21}\Delta_{22}^{-1}A_{12}A_{22}^{-1} \end{array}\right]. \qquad \text{(B.15b)} \end{aligned}$$

The formulae above are particularly simple in the case of general block-triangular matrices:

$$\begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}^{-1} = \begin{bmatrix} A_{11}^{-1} & -A_{11}^{-1}A_{12}A_{22}^{-1} \\ 0 & A_{22}^{-1} \end{bmatrix} \tag{B.16a}$$

and

$$\begin{bmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{bmatrix}^{-1} = \begin{bmatrix} A_{11}^{-1} & 0 \\ -A_{22}^{-1}A_{21}A_{11}^{-1} & A_{22}^{-1} \end{bmatrix}. \tag{B.16b}$$

Relations (B.15a) and (B.15b) also lead to the following useful result.

Lemma B.7 (Matrix Inversion Lemma)**.** *Let $A_{ij} \in \mathbb{F}^{m_i \times m_j}$ for $i, j \in \{1, 2\}$. If A_{11} and A_{22} are nonsingular, then*

$$(A_{11} - A_{12}A_{22}^{-1}A_{21})^{-1} = A_{11}^{-1} + A_{11}^{-1}A_{12}\big(A_{22} - A_{21}A_{11}^{-1}A_{12}\big)^{-1}A_{21}A_{11}^{-1}$$

whenever the inverse in the left-hand side exists.

Proof. The result follows by comparing the $(1,1)$ sub-blocks in (B.15a) and (B.15b). □

Another use of (B.14) is in verifying sign definiteness of symmetric block matrices. Remember, a matrix M is said to be positive definite if $x'Mx > 0$ for all $x \neq 0$, cf. the related discussion on sign definite operators in §A.2.3. So assume that our $A = A'$ and consider the quadratic form

$$q(x_1, x_2) := \begin{bmatrix} x_1' & x_2' \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

Because $q(x_1, 0) = x_1' A_{11} x_1$ and $q(0, x_2) = x_2' A_2 x_2$, this quadratic form is positive for all $x \neq 0$ only if $A_{11} > 0$ and $A_{22} > 0$. But then both Schur complements are well defined and, by (B.14),

$$\begin{aligned} q(x_1, x_2) &= \begin{bmatrix} x_1' + x_2' A_{21} A_{11}^{-1} & x_2' \end{bmatrix} \begin{bmatrix} A_{11} & 0 \\ 0 & \Delta_{11} \end{bmatrix} \begin{bmatrix} x_1 + A_{11}^{-1} A_{12} x_2 \\ x_2 \end{bmatrix} \\ &= (x_1 + A_{11}^{-1} A_{12} x_2)' A_{11} (x_1 + A_{11}^{-1} A_{12} x_2) + x_2' \Delta_{11} x_2 \qquad \text{(B.17a)} \\ &= \begin{bmatrix} x_1' & x_2' + x_1' A_{12} A_{22}^{-1} \end{bmatrix} \begin{bmatrix} \Delta_{22} & 0 \\ 0 & A_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 + A_{22}^{-1} A_{21} x_1 \end{bmatrix} \\ &= x_1' \Delta_{22} x_1 + (x_2 + A_{22}^{-1} A_{21} x_1)' A_{22} (x_2 + A_{22}^{-1} A_{21} x_1) \qquad \text{(B.17b)} \end{aligned}$$

Inspecting (B.17a) and taking into account that $A_{22} > 0$ is necessary, we have that $q(x_1, x_2) > 0$ whenever $\Delta_{11} > 0$. If the latter condition is not true, then there is $x_2 \neq 0$ such that $x_2' \Delta_{11} x_2 \leq 0$. But then the choice $x_1 = -A_{11}^{-1} A_{12} x_2$ renders $q(x_1, x_2) = x_2 \Delta_{11} x_2 \leq 0$ under $x \neq 0$. Hence, $\Delta_{11} > 0$ is not only sufficient, but also necessary for $A > 0$. Similar arguments apply to (B.17b) and we have the following result.

Lemma B.8. *Let $A = A'$. The following conditions are equivalent:*

1. $A > 0$,
2. $A_{11} > 0$ *and* $\Delta_{11} > 0$,
3. $A_{22} > 0$ *and* $\Delta_{22} > 0$.

Lemma B.8 can obviously be used to verify the negative definiteness. One just needs to invert all signs in its conditions.

B.4 Useful matrix relations

Lemma B.9. *Let $A \in \mathbb{F}^{n \times m}$ and $B \in \mathbb{F}^{m \times n}$. Every nonzero eigenvalue of $AB \in \mathbb{F}^{n \times n}$ is also an eigenvalue of $BA \in \mathbb{F}^{m \times m}$.*

Proof. If $\lambda \neq 0$ is an eigenvalue of AB and η its eigenvector, then, $AB\eta = \eta\lambda$ and $\zeta := B\eta \neq 0$. Yet then $BAB\eta = B\eta\lambda$ or, equivalently, $BA\zeta = \zeta\lambda$. □

Lemma B.10. *If $A \in \mathbb{F}^{n \times m}$ and $B \in \mathbb{F}^{m \times n}$ are such that $I_n - AB$ is invertible, then*

$$(I_n - AB)^{-1} A = A (I_m - BA)^{-1} \qquad \text{(B.18)}$$

and $I_m - BA$ is invertible as well.

Proof. The invertibility of $I_n - AB$ is equivalent to the condition that AB does not have eigenvalues at 1. Hence, neither does BA (by Lemma B.9) and the invertibility of $I_m - BA$ follows. Assuming invertibility,

$$(I_n - AB)^{-1} A - A (I_m - BA)^{-1} = (I_n - AB)^{-1} (A(I_m - BA) - (I_n - AB)A)(I_m - BA)^{-1} = 0,$$

which proves (B.18). □

Bibliography

[1] B. D. O. Anderson and J. B. Moore, *Optimal Control: Linear Quadratic Methods*. Englewood Cliffs, NJ: Prentice-Hall, 1989.

[2] B. D. O. Anderson and S. Vongpanitlerd, *Network Analysis and Synthesis: A Modern Systems Theory Approach*. Englewood Cliffs, NJ: Prentice-Hall, 1973.

[3] A. C. Antoulas, *Approximation of Large-Scale Dynamical Systems*. Philadelphia, PA: SIAM, 2005.

[4] M. Athans and P. L. Falb, *Optimal Control: An Introduction to the Theory and Its Applications*. New York, NY: McGraw-Hill, 1966.

[5] R. F. Curtain and K. Morris, "Transfer functions of distributed parameter systems: A tutorial," *Automatica*, vol. 45, no. 5, pp. 1101–1116, 2009.

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

[7] P. Dewilde and A.-J. van der Veen, *Time-Varying Systems and Computations*. Boston, MA: Kluwer Academic Publishers, 1998.

[8] I. Gohberg, S. Goldberg, and M. A. Kaashoek, *Classes of Linear Operators*. Basel, CH: Birkhäuser, 1990, vol. I.

[9] C. Guiver, H. Logemann, and M. R. Opmeer, "Transfer functions of infinite-dimensional systems: positive realness and stabilization," *Math. Control, Signals, Syst.*, vol. 29, pp. 20:1–20:61, 2017.

[10] M. Heymann, "Comments "On pole assignment in multi-input controllable linear systems"," *IEEE Trans. Automat. Control*, vol. 13, no. 6, pp. 748–749, 1968.

[11] T. Iwasaki, G. Meinsma, and M. Fu, "Generalized S-procedure and finite frequency KYP lemma," *Math. Probl. Eng.*, vol. 6, pp. 305–320, 2000.

[12] T. Kailath, *Linear Systems*. Englewood Cliffs, NJ: Prentice-Hall, 1980.

[13] R. E. Kalman, "Irreducible realizations and the degree of a rational matrix," *J. Soc. Indust. Appl. Math.*, vol. 13, no. 2, pp. 520–544, 1965.

[14] M. Kristalny, "Exploiting previewed information in estimation and control," Ph.D. dissertation, Faculty of Mechanical Eng., Technion—IIT, Aug 2010. [Online]. Available: http://leo.technion.ac.il/theses/KristalnyPhD.pdf

[15] H. Kwakernaak and R. Sivan, *Linear Optimal Control Systems*. New York, NY: John Wiley & Sons, 1972.

[16] G. Lang and J. M. Ham, "Conditional feedback systems—a new approach to feedback control," *Trans. American Inst. Electrical Engin., Part II: Applications and Industry*, vol. 74, no. 3, pp. 152–161, 1955.

[17] D. C. McFarlane and K. Glover, *Robust Controller Design Using Normalized Coprime Factor Plant Descriptions*, ser. Lecture Notes in Control and Inform. Sci. Berlin, DE: Springer-Verlag, 1990, vol. 138.

[18] G. Meinsma, "Unstable and nonproper weights in $\mathcal{H}_\infty$ control," *Automatica*, vol. 31, no. 11, pp. 1655–1658, 1995.

[19] M. Morari and E. Zafiriou, *Robust Process Control*. Englewood Cliffs, NJ: Prentice-Hall, 1989.

[20] S. Parrott, "On a quotient norm and the Sz-Nagy-Foias lifting theorem," *J. Funct. Anal.*, vol. 30, pp. 311–328, 1978.

[21] J. R. Partington, *Linear Operators and Linear Systems: An Analytical Approach to Control Theory*. Cambridge, UK: Cambridge University Press, 2004.

[22] A. Rantzer, "On the Kalman–Yakubovich–Popov lemma," *Syst. Control Lett.*, vol. 28, no. 1, pp. 7–10, 1996.

[23] H. H. Rosenbrock, *State-Space and Multivariable Theory*. London, UK: Nelson, 1970.

[24] W. Rudin, *Real and Complex Analysis*, 3rd ed. New York, NY: McGraw-Hill, 1987.

[25] I. W. Sandberg, "Recent representation results for linear system maps: A short survey," *Int. J. Circ. Theor. Appl.*, vol. 35, pp. 497–514, 2007.

[26] M. C. Smith, "On stabilization and the existence of coprime factorizations," *IEEE Trans. Automat. Control*, vol. 34, no. 9, pp. 1005–1007, 1989.

[27] O. J. M. Smith, "Closer control of loops with dead time," *Chem. Eng. Progress*, vol. 53, no. 5, pp. 217–219, 1957.

[28] G. Vinnicombe, *Uncertainty and Feedback: $\mathcal{H}_\infty$ Loop-Shaping and the ν-Gap Metric*. London, UK: Imperial College Press, 2001.

[29] K. Watanabe and M. Ito, "A process-model control for linear systems with delay," *IEEE Trans. Automat. Control*, vol. 26, no. 6, pp. 1261–1269, 1981.

[30] J. C. Willems, *The Analysis of Feedback Systems*. Cambridge, MA: The MIT Press, 1971.

[31] ——, "The behavioral approach to open and interconnected systems," *IEEE Control Syst. Mag.*, vol. 27, no. 6, pp. 46–99, 2007.

[32] B.-F. Wu and E. A. Jonckheere, "A simplified approach to Bode's theorem for continuous-time and discrete-time systems," *IEEE Trans. Automat. Control*, vol. 37, no. 11, pp. 1797–1802, 1992.

[33] G. Zames, "Feedback and optimal sensitivity: Model reference transformations, multiplicative seminorms, and approximate inverses," *IEEE Trans. Automat. Control*, vol. 26, no. 2, pp. 301–320, 1981.

[34] A. H. Zemanian, *Realizability Theory for Continuous Linear Systems*. New York, NY: Academic Press, 1972.

Index