Control of Systems with Repeated Scalar Nonlinearities

Yun-Chung Chu

Wolfson College

University of Cambridge

Cambridge, England

A dissertation submitted for
the degree of Doctor of Philosophy

August 1996

last revised
March 1998

To my parents and sister

Preface

I would like to thank my supervisor Professor Keith Glover who has been taking care of me since I came here, both academically and personally, and shown respect for my way and understanding to my stupidity. It has been a pleasure to work with him, and his perseverance with difficult problems and his sense of significant and connected ideas have had a great influence on my work.

I would also like to thank everyone who has made my life here more enjoyable than it might have been, especially to my friend Albert Wong who has constantly helped me deal with various living problems. I have received much help and support from the past and present members of the control group in Cambridge and to all of them I am grateful.

My Ph.D. studies have been financially supported by many organizations which are acknowledged below: the Committee of Vice-Chancellors and Principals "Overseas Research Students Awards Scheme" (92–95), the Friends of Cambridge University in Hong Kong and the Cambridge Commonwealth Trust (92–93), the Croucher Foundation (93–95), the Cambridge Philosophical Society (96), and the Department of Engineering at University of Cambridge (96).

As required by University Statute, I hereby declare that this dissertation is not substantially the same as any that I have submitted for a degree or diploma or other qualification at any other university. I further state that no part of this dissertation has already been or is being concurrently submitted for any such degree, diploma or other qualification. This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration.

Yun-Chung Chu
August 1996, Cambridge

This revised edition corrects some typographical errors that exist in the official copies kept by University of Cambridge and the Croucher Foundation. There is no basic change in its contents.
October 1996, Hong Kong

An insignificant technical error in Chapter 4 was corrected. May 1997, Hong Kong

Typos in (3.31) and Bibliography were corrected. March 1998, Cambridge

Abstract

A class of discrete-time nonlinear systems is studied which is described by a standard linear state-equation except that each state-component is subject to an identical odd 1-Lipschitz nonlinearity. This may represent a class of recurrent neural networks. The emphasis is on the development of an approach that makes the best use of the fact that the nonlinearity on each state-component is the same.

It is shown that the origin of such a system is globally stable, and an upper bound on the ℓ_2-ℓ_2 induced gain can be deduced, if a quadratic Lyapunov/storage function exists whose Hessian matrix is positive definite diagonally dominant. This reduces to finding such a matrix to satisfy appropriate Linear Matrix Inequalities, which is shown computationally tractable.

In synthesis, it is assumed that the controller is also a nonlinear system of this form and has the same nonlinearity as the plant does. With appropriate assumptions on the stabilizability and detectability of the plant, all such controllers that would stabilize the closed-loop are parametrized. A result on performance synthesis is also given.

For model reduction, the approach adopted is close to the conventional balanced truncation for stable linear systems, where the global ℓ_2-ℓ_2 error is the main concern. A model reduction algorithm is proposed based on the theoretical result obtained.

The results are shown valid even when the nonlinearity is dynamic rather than static. Furthermore, using block-diagonal positive definite diagonally dominant matrices they can be extended to the more general case that the nonlinearities on different state-components may be partly but not all identical. It is easy to see that such a system is in fact a constant Linear Fractional Transformation of a diagonal odd 1-Lipschitz nonlinear operator, and the treatment given is an attempt to parallel the conventional treatment of an uncertain linear system described by a Linear Fractional Transformation. This is not trivial, however, due to the nonlinear nature. Still, the Linear Fractional Transformation framework of uncertain linear systems expands into the nonlinear systems to a large extent.

Keywords: nonlinear systems, Linear Fractional Transformations, Linear Matrix Inequalities, state-dependent perturbations, recurrent neural networks, diagonally dominant matrices, diagonal stability, ℓ_2-ℓ_2 induced gain, controller parametrization, gain-scheduling, model reduction, balanced truncation.

Contents

Notation

Special symbols

$:=$	defined as.
$\bullet, \heartsuit, \clubsuit, \spadesuit, \diamondsuit$	dummy variables.
$\triangleleft$	end of definition, lemma, proposition, corollary, etc.
$\blacksquare$	end of proof.
$\otimes$	end of remark or example.

Sets and Functions

Let U, V, W, X, Y be sets and $f : V \to W$, $g : X \to Y$ be functions. (In this dissertation function, map, operator all mean the same thing.)

$V \times X$	the Cartesian product of V and X.
V^n	$\underbrace{V \times V \times \cdots \times V}_{n}$.
$U \cup V$	the union of U and V.
$U \cap V$	the intersection of U and V.
$V \setminus U$	the complement of U in V: $\left\{ v \in V \;\middle|\; v \notin U \right\}$.
$U \subset V$	U is a subset of V, not necessarily proper.
$f(v)$	the value of f at v if $v \in V$, sometimes abbreviated as fv especially in vector, matrix and sequence multiplications.
$f(V)$	the range of f $(\subset W)$.
$f\mid_U$	the restriction of f to $U \subset V$.
$g \circ f$	the composition: $V \to Y$ defined by $g \circ f(v) := g(f(v))$ for $v \in V$ if $f(V) \subset X$, sometimes abbreviated as gf especially in vector, matrix and sequence multiplications.

Sets of numbers

$\mathbb{Z}$	the set of integers.
$\mathbb{Z}_+$	the set of non-negative integers.
$\mathbb{R}^{m \times n}$	the set of $m \times n$ real matrices.
$\mathbb{R}^n$	$\mathbb{R}^{n \times 1}$, i.e. the set of n-dimensional real column vectors.
$\mathbb{R}$	$\mathbb{R}^1$, i.e. the set of real numbers.

Vectors and Matrices

In this dissertation vectors/matrices always refer to vectors/matrices of real numbers unless specified otherwise, but the vector/matrix *notation* may apply to vectors/matrices of other objects such as sequences and operators. So the multiplication is well defined whenever the multiplication of corresponding entries is defined (e.g.

$$\begin{bmatrix} f_1 & f_2 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} := f_1 v_1 + f_2 v_2 := f_1(v_1) + f_2(v_2)$$

$$\begin{bmatrix} g_1 & g_2 \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \end{bmatrix} := g_1 f_1 + g_2 f_2 := g_1 \circ f_1 + g_2 \circ f_2$$

).

Let ζ be a n-dimensional column vector, P be a $m \times n$ matrix, and $P_1, P_2, \ldots, P_l$ be matrices of any dimensions. Their entries belong to certain sets which are not necessarily $\mathbb{R}$.

$\zeta_{(i)}$	the i-th entry/coordinate of ζ.
$p_{(ij)}$	the (i, j)-th entry of P.
$p_{(i)}$	$p_{(i,i)}$ if P is diagonal.
ζ^*, P^*	the transposes of ζ and P.
$\\|\zeta\\|$	the 2-norm of ζ, i.e. $\\|\zeta\\| := \|\zeta\|$ if $\zeta \in \mathbb{R}$, otherwise $\\|\zeta\\| := \sqrt{\sum_{i=0}^{n} \\|\zeta_{(i)}\\|^2}$. If $\zeta \in \mathbb{R}^n$, then $\\|\zeta\\| = \sqrt{\zeta^* \zeta}$.
$\\|P\\|$	the 2-2 induced norm of P, i.e. $\sup_{\zeta \neq 0} \frac{\\|P\zeta\\|}{\\|\zeta\\|}$. If $\zeta \in \mathbb{R}^n$ and $P \in \mathbb{R}^{m \times n}$, then $\\|P\\|$ is the largest singular value of P.
$\mathrm{co}(\{P_1, P_2, \ldots, P_l\})$	the convex hull of $\{P_1, P_2, \ldots, P_l\}$: $$\left\{ \sum_{i=1}^{l} \alpha_i P_i \;\middle|\; \alpha_i \in \mathbb{R}, \alpha_i \geq 0, \sum_{i=1}^{l} \alpha_i = 1 \right\}$$ if $P_1, P_2, \ldots, P_l$ have the same dimension.
$\mathrm{Diag}(P_1, P_2, \ldots, P_l)$	the diagonal augmentation of $P_1, P_2, \ldots, P_l$: $$\begin{bmatrix} P_1 & & & \\ & P_2 & & \\ & & \ddots & \\ & & & P_l \end{bmatrix}$$
$\mathrm{diag}(P)$	$\mathrm{Diag}(p_{(11)}, p_{(22)}, \ldots, p_{(nn)})$ if P is square.
$P \geq 0$ (> 0)	P is positive semidefinite (positive definite), i.e. $P \in \mathbb{R}^{n \times n}$ such that $P = P^*$ and $\zeta^* P \zeta \geq 0$ (> 0) for all $\zeta \in \mathbb{R}^n$, $\zeta \neq 0$.
I_n	the $n \times n$ identity matrix. The n is omitted if the dimension of I is clear or irrelevant. I_0 denotes the empty matrix.

Sequences

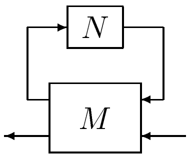
A sequence is a map with domain $\mathbb{Z}_+$. The range of a sequence may be a set of any objects such as real numbers, operators and vectors/matrices of any objects. All sequences in this dissertation are interpreted as time sequences.

Let x, y be sequences.

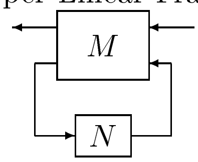
$x_{[k]}$	the value of x at time k.
$\|x\|_K$	$\sqrt{\sum_{k=0}^{K} \|x_{[k]}\|^2}$ $\left(= \sqrt{\sum_{k=0}^{K} x_{[k]}^* x_{[k]}}\right.$ if $x_{[k]} \in \mathbb{R}^n$).
$x \cdot y$	the sequence defined by $(x \cdot y)_{[k]} := x_{[k]} y_{[k]}$ if the multiplication of $x_{[k]}$, $y_{[k]}$ is defined.
$(x \cdot)$	the operator which maps y to $x \cdot y$.
ℓ_{2e}^n	the set of $\mathbb{R}^n$ sequences.
ℓ_{2e}	ℓ_{2e}^1, i.e. the set of real sequences.
ℓ_2^n	$\left\{ x \in \ell_{2e}^n \;\middle|\; \lim_{K\to\infty} \|x\|_K \text{ exists; } \|x\| := \lim_{K\to\infty} \|x\|_K \right\}$.
ℓ_2	ℓ_2^1.

Systems

Let M be a 2-input-channel 2-output-channel system and N be a 1-input-channel 1-output-channel system. (A system may be dynamic or simply a constant matrix.)

$\mathcal{F}_u(M, N)$ — the system 

. If M is linear, $\mathcal{F}_u(M, N)$ is called the upper Linear Fractional Transformation.

$\mathcal{F}_l(M, N)$ — the system 

. If M is linear, $\mathcal{F}_l(M, N)$ is called the lower Linear Fractional Transformation.

Abbreviation

1HLN	single hidden layer feedforward neural network.
GAS	globally asymptotically stable or global asymptotic stability.
IP	Independence Property.
LDI	Linear Differential/Difference Inclusion.
LFT	Linear Fractional Transformation.
LMI	Linear Matrix Inequality.
OF,FI,FC,DF,OE	Output Feedback, Full Information, Full Control, Disturbance Feedforward and Output Estimation problems.
WLOG	without loss of generality.
w.r.t.	with respect to.

Chapter 1

Introduction

1.1 Object of study

The primary object of study in this dissertation is the discrete-time nonlinear system described by

$$x^+ = A\phi(x) + Bu; \quad y = C\phi(x) + Du \tag{1.1}$$

or in matrix form

$$\begin{bmatrix} x^+ \\ y \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} \phi(x) \\ u \end{bmatrix} \tag{1.2}$$

where A, B, C, D are constant (real) matrices, $u \in \ell_{2e}^m$, $y \in \ell_{2e}^p$, $x, x^+ \in \ell_{2e}^n$ such that $x_{[k]}^+ = x_{[k+1]}$, ϕ is a nonlinear[1] function on $x_{(i)}$ and $\phi(x) \in \ell_{2e}^n$ actually means

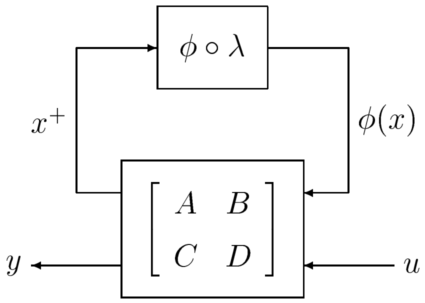
$$\begin{bmatrix} \phi(x_{(1)}) \\ \phi(x_{(2)}) \\ \vdots \\ \phi(x_{(n)}) \end{bmatrix}$$

Note that each coordinate of the state has the same nonlinearity. We shall call a system of this kind a ϕ-system and denote it by $\left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right)_\phi$. (The nomenclature was motivated by Sontag *et al.* See Section 1.2.3 below.) In the most general setting, $\phi : \ell_{2e} \to \ell_{2e}$ is a time-varying dynamic operator. However, we shall present our results in the more familiar setting $\phi : \mathbb{R} \to \mathbb{R}$, and $\phi(x_{(i)}) \in \ell_{2e}$ should be understood as $\phi(x_{(i)})_{[k]} := \phi(x_{(i)[k]})$, i.e. we adopt the convention that if a $\mathbb{R}$ operator operates on a ℓ_{2e} element, it operates at each time instant. It should be emphasized that *all* important results in this dissertation can be generalized to the $\phi : \ell_{2e} \to \ell_{2e}$ setting, and this will be demonstrated in Section 5.2. We also adopt the convention that whenever a $\mathbb{R}$ (or

[1] "Nonlinear" means "not necessarily linear" in this dissertation, and "time-varying" means "not necessarily time-invariant" etc.

ℓ_{2e}) operator is applied to a $\mathbb{R}^n$ (or ℓ_{2e}^n) element it operates coordinatewise, just as $\phi(x) \in \ell_{2e}^n$ should be understood as $\phi(x)_{(i)} := \phi(x_{(i)})$, and we shall not, in notation, distinguish between ϕ as an operator on $\mathbb{R}$ (ℓ_{2e}) and ϕ as an operator on $\mathbb{R}^n$ (ℓ_{2e}^n). If it is necessary to make the dimension of the domain clear then we simply write ϕI_n which may be interpreted as $\phi \circ I_n$ where I_n has no real action but is only to signify the dimension of the domain n.

If $x_{[0]} = 0$, a ϕ-system may be represented by



(1.3)

where $\lambda : \ell_{2e} \to \ell_{2e}$ is the time-delay operator:

$$\lambda(x^+_{(i)})_{[k]} := \begin{cases} x^+_{(i)[k-1]} & \text{if} \quad k > 0 \\ 0 & \text{if} \quad k = 0 \end{cases} \tag{1.4}$$

In view of (1.3), the (zero-initial-state) input-output map: $u \mapsto y$ will be denoted by $\mathcal{F}_u(\begin{bmatrix} A & B \\ C & D \end{bmatrix}, \phi \circ \lambda)$.

For a reason which will become clear in Chapter 2, we shall restrict the class of ϕ to

$$\mathcal{OL}(\mathbb{R}) := \left\{ \phi : \mathbb{R} \to \mathbb{R} \;\middle|\; \forall s, t \in \mathbb{R},\ \ |\phi(s) + \phi(t)| \le |s + t| \right\} \tag{1.5}$$

Note that $\mathcal{OL}(\mathbb{R})$ is actually the set of all nonlinear operators ϕ which are odd (by putting $t = -s$):

$$\forall s \in \mathbb{R},\ \ \phi(-s) = -\phi(s)$$

and 1-Lipschitz (by putting $t = -t$):

$$\forall s, t \in \mathbb{R},\ \ |\phi(s) - \phi(t)| \le |s - t|$$

(In the $\phi : \ell_{2e} \to \ell_{2e}$ setting, the class of ϕ is denoted by $\mathcal{OL}(\ell_{2e})$ which simply speaking is the set of causal ϕ whose restrictions to ℓ_2 are odd and 1-Lipschitz. See Section 5.2 for the precise definition.)

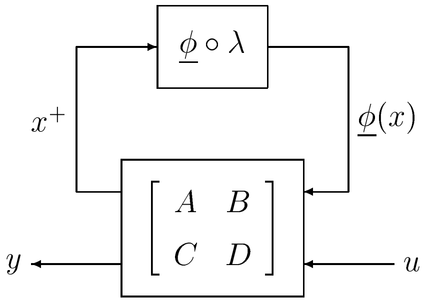
The ϕ-system is a special case of the following nonlinear system, which will also be studied in this dissertation:

$$\begin{bmatrix} x^+ \\ y \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} \underline{\phi}(x) \\ u \end{bmatrix} \tag{1.6}$$

where $\underline{\phi}$ is a *diagonal* nonlinear operator:

$$\underline{\phi}(x) = \begin{bmatrix} \underline{\phi}_{(1)}(x_{(1)}) \\ \underline{\phi}_{(2)}(x_{(2)}) \\ \vdots \\ \underline{\phi}_{(n)}(x_{(n)}) \end{bmatrix} \in \ell_{2e}^n$$

i.e. the nonlinearities on each coordinate of the state are not necessarily identical. We shall call a system of this kind a $\underline{\phi}$-system and denote it by $\left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right)_{\underline{\phi}}$. (1.3) will become



(1.7)

and the input-output map will be denoted by $\mathcal{F}_u(\begin{bmatrix} A & B \\ C & D \end{bmatrix}, \underline{\phi} \circ \lambda)$. We write $\underline{\phi} \in \mathcal{OL}^n(\mathrm{IR})$ if all $\underline{\phi}_{(i)} \in \mathcal{OL}(\mathrm{IR})$.

If we ignore that some $\underline{\phi}_{(i)}$ in the $\underline{\phi}$-system may be identical, studying the $\underline{\phi}$-system is easier than studying the ϕ-system since in the latter we should try to exploit the fact that all $\underline{\phi}_{(i)} = \phi$ are identical. On the other hand, if we would like to acknowledge that some $\underline{\phi}_{(i)}$ in the $\underline{\phi}$-system may be identical, which is essential in obtaining tight results in analysis and doing synthesis for the $\underline{\phi}$-system, it is important to understand the ϕ-system, i.e. the all-$\underline{\phi}_{(i)}$-identical case first. In fact, we shall improve Kaszkurewicz and Bhaya's treatment of the $\underline{\phi}$-system (summarized in Section 1.2.2 below) based on our results on the ϕ-system. See Section 5.1.1 for details.

Whenever the $\bullet$ in $\left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right)_{\bullet}$ is clear or irrelevant, we may simply write the system as $\left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right)$, sometimes even for the linear system $\left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right)_I$. For two systems $\left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right)$, $\left(\begin{array}{c|c} \tilde{A} & \tilde{B} \\ \hline \tilde{C} & \tilde{D} \end{array}\right)$, we write

$$\left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right) \simeq \left(\begin{array}{c|c} \tilde{A} & \tilde{B} \\ \hline \tilde{C} & \tilde{D} \end{array}\right)$$

to mean

$$\mathcal{F}_u(\begin{bmatrix} A & B \\ C & D \end{bmatrix}, \bullet) = \mathcal{F}_u(\begin{bmatrix} \tilde{A} & \tilde{B} \\ \tilde{C} & \tilde{D} \end{bmatrix}, \bullet)$$

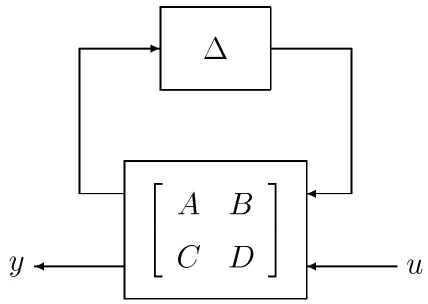
i.e. they have the same (zero-initial-state) input-output map, and in writing this we do *not* imply that their stabilities are the same, i.e. one system may be stable while the other may not be stable, in a sense we shall define in Section 2.2.

It is not necessary that the operator $\underline{\phi}_{(i)}$ (identical or not) and the operator λ must appear as $\underline{\phi}_{(i)} \circ \lambda$. Our treatment presented in this dissertation is readily generalizable to the case where $\underline{\phi}_{(i)}$ and λ are separate. The details are given in Section 5.1.2. This will correspond to the case where $\underline{\phi}_{(i)}$ appears at other possible locations in the state-space description of the system.

1.2 Background

1.2.1 LFT uncertain systems and LMI

Nowadays, a popular approach to represent an uncertain linear system is to use the Linear Fractional Transformation (LFT):



(1.8)

where $\begin{bmatrix} A & B \\ C & D \end{bmatrix}$ is the certain part of the system while Δ collects all the uncertainties and usually takes the form

$$\Delta \in \triangle_{N,M} := \left\{ \mathrm{Diag}(\delta_1 I_{n_{(1)}}, \delta_2 I_{n_{(2)}}, \ldots, \delta_s I_{n_{(s)}}, \Delta_1, \Delta_2, \ldots, \Delta_f) \right\} \tag{1.9}$$

where N, M are $s \times s$, $f \times f$ diagonal matrices of non-negative integers such that

$$\sum_{i=1}^{s} n_{(i)} + \sum_{i=1}^{f} m_{(i)} = n$$

and $\delta_i : \ell_2 \to \ell_2$, $\Delta_i : \ell_2^{m_{(i)}} \to \ell_2^{m_{(i)}}$ are time-varying *linear* dynamic operators. Typically $\Delta \in \mathcal{B}(\triangle_{N,M}) := \left\{ \Delta \in \triangle_{N,M} \;\middle|\; \|\Delta\| \le 1 \right\}$ and $\delta_1 = \lambda$. (The latter means that only $x_{(1)}, x_{(2)}, \ldots, x_{(n_{(1)})}$ are the states of the system in the traditional sense.)

The LFT uncertain system is a rather natural generalization of the linear system, and therefore many results in the linear system can be generalized to the LFT uncertain system. (See [14] for a tutorial.) For example, Beck *et al.*[12, 13, 15] have been working on the realization theory, minimality, and model reduction (via balanced realization) of the LFT uncertain system, and have improved the model reduction result obtained by Wang *et al.*[74]; Lu *et al.*[51] have been working on the stabilizability and detectability, and have parametrized all output-feedback stabilizing controllers for stabilizable and detectable LFT uncertain systems; Packard[60], Apkarian and Gahinet[9, 10], Scorletti and El Ghaoui[71] and Helmersson[36] have been working on the gain-scheduling problem, i.e. to design a controller to achieve some closed-loop performance with the assumption that the uncertainty Δ of the plant is also available to the controller in real time. D'Andrea and Paganini *et al.*[25, 63, 64, 65, 62, 24] have also considered a further abstraction of the LFT uncertain system: the LFT uncertain system in an implicit form, and its connection to the Integral Quadratic Constraint (IQC).

The computational success of the LFT approach relies heavily on the fact that most of the above-mentioned analysis and synthesis problems of the LFT uncertain system can be formulated as optimizations of a linear objective subject to a Linear Matrix Inequality (LMI) constraint. (For a general reference on the LMI, see [22].) The general form of a LMI is

$$F(z) \geq 0 \text{ (or } > 0) \tag{1.10}$$

where F takes symmetric matrix values (so ≥ 0 and > 0 mean positive semidefinite and positive definite respectively) and is affine on the variable z, i.e. if $z \in \mathrm{I\!R}^l$,

$$F(z) = \boldsymbol{F}_0 + \boldsymbol{F}_1 z_{(1)} + \boldsymbol{F}_2 z_{(2)} + \cdots + \boldsymbol{F}_l z_{(l)} \tag{1.11}$$

where $\boldsymbol{F}_i$ are constant symmetric matrices, and we can write (1.10) as

$$F(z) = \boldsymbol{F}_0 + \boldsymbol{F}_1 z_{(1)} + \boldsymbol{F}_2 z_{(2)} + \cdots + \boldsymbol{F}_l z_{(l)} \geq 0 \text{ (or } > 0) \tag{1.12}$$

If we have two LMIs $F_\alpha(z) \geq 0$ and $F_\beta(z) \geq 0$, we can simply combine them into one:

$$\begin{bmatrix} F_\alpha(z) & 0 \\ 0 & F_\beta(z) \end{bmatrix} \geq 0$$

Note that a LMI is a convex constraint, i.e. the feasible set of z is convex. Hence minimizing a linear objective subject to a LMI is a convex optimization problem. In most analysis and synthesis problems of the LFT uncertain system, the variable of a LMI will appear in matrix form. For example, it was shown in [65] that the LFT uncertain system is stable (in some sense) for all $\Delta \in \mathcal{B}(\triangle_{N,M})$ if and only if there exists $P \in \mathcal{C}(\triangle_{N,M})$ such that

$$P > 0; \quad A^*PA - P < 0 \tag{1.13}$$

where $\mathcal{C}(\triangle_{N,M})$ is the set of all matrices commuting with all $\Delta \in \triangle_{N,M}$:

$$\begin{aligned} \mathcal{C}(\triangle_{N,M}) \;&:=\; \left\{ T \in \mathrm{I\!R}^{n\times n} \;\middle|\; T\Delta = \Delta T \text{ for all } \Delta \in \triangle_{N,M} \right\} \\ &=\; \left\{ \mathrm{Diag}(T_1, T_2, \ldots, T_s, \tau_1 I_{m_{(1)}}, \tau_2 I_{m_{(2)}}, \ldots, \tau_f I_{m_{(f)}}) \right\} \end{aligned} \tag{1.14}$$

Of course (1.13) can be written into the general LMI form (1.12) in terms of the entries $p_{(ij)}$ of the matrix P. Hence this is a LMI feasibility problem. Note that such convex optimization problems have unique global minima, and can be solved efficiently from the computational point of view. Furthermore, the recently developed interior-point method[58](and [75] for a survey) to solve these problems has been observed to be very powerful, thus providing a significant improvement over the old ellipsoid algorithm. (For general references on the ellipsoid algorithm, see [20],[1, Ch. 3] and [21, Ch. 14].) Therefore, it has attracted a lot of research efforts from the control community, to formulate the analysis and synthesis problems in terms of LMIs, and led to the development of software packages such as LMI Control Toolbox for MATLAB[31, 32]. With these computational tools, the LFT description of an uncertain system is useful both theoretically and practically.

Of course, this combined LFT/LMI approach to uncertain systems still has its limitations, which mainly come from the class of Δ (1.9). On one hand, δ_i and Δ_i have to be arbitrary time-varying linear dynamic operators on ℓ_2 and $\ell_2^{m_{(i)}}$ so that the results obtained using the LMI approach should be tight. If, for example, δ_i and Δ_i are only (time-invariant) linear algebraic operators on $\mathbb{R}$ and $\mathbb{R}^{m_{(i)}}$, the results given by the LMI approach may be somewhat conservative, since it still uses the same set of matrices (1.14) as $\mathcal{C}(\triangle_{N,M})$. In particular the stability condition (1.13) will become only sufficient but not necessary and sufficient. On the other hand, Δ has to be linear, since the analysis and synthesis of the LFT uncertain system rely heavily on the elements in $\mathcal{C}(\triangle_{N,M})$, which commute with all $\Delta \in \triangle_{N,M}$ and in particular provide sufficient realizations of the system for all analysis and synthesis problems. If Δ has components which are not linear, the set of matrices commuting with it will shrink dramatically. For example, the set of matrices commuting with a general $\phi \in \mathcal{OL}(\mathbb{R})$ is only the set of permutation matrices and diagonal matrices with ± 1 as diagonal entries (and their products, see Section 2.4.1 for details). This set is too small for us to do useful analysis and synthesis. Still, the aim of this dissertation is to extend the class of Δ in the LFT description of uncertain systems to include the nonlinear elements in $\mathcal{OL}(\mathbb{R})$ (in fact $\mathcal{OL}(\ell_{2e})$). The basic tool is still a set of matrices $\{P\}$, which although do not commute with ϕ (i.e. $\exists \zeta,\ P\phi(\zeta) \neq \phi(P\zeta)$) for most $\phi \in \mathcal{OL}(\mathbb{R})$, satisfy

$$\forall \phi \in \mathcal{OL}(\mathbb{R}),\ \ \forall \zeta \in \mathbb{R}^n,\ \ \phi(\zeta)^* P \phi(\zeta) \leq \zeta^* P \zeta \tag{1.15}$$

This set of matrices do *not* provide realizations, but will allow us to generalize certain (but not all!) analysis and synthesis results for the LFT uncertain system to the nonlinear ϕ-system, and in this dissertation we shall see which of those results are generalizable and which are not, and why.

1.2.2 State-dependent perturbations

The $\underline{\phi}$-system where $\underline{\phi}_{(i)}$ belong to a larger class[2] $\mathcal{S} \supset \mathcal{OL}(\mathbb{R})$,

$$\mathcal{S} := \left\{ \phi : \mathbb{R} \to \mathbb{R} \;\middle|\; \forall s \in \mathbb{R},\ \ |\phi(s)| \le |s| \right\} \tag{1.16}$$

was once studied (as a special case of a more general class of uncertain systems) by Kaszkurewicz and Bhaya in [43], where they regarded $\underline{\phi}_{(i)}$ as state-dependent perturbations. Such state-dependent perturbations, for example, can model the signal quantization in digital filters[53, 46, 48]. Kaszkurewicz and Bhaya showed that if there exists a diagonal matrix P such that

$$P > 0; \quad A^* P A - P < 0 \tag{1.17}$$

then the $\underline{\phi}$-system is globally asymptotically stable (GAS). Indeed

$$V(\zeta) := \zeta^* P \zeta \tag{1.18}$$

is a Lyapunov function.

If we compare it to the LFT uncertain system, we see that the set of diagonal matrices is exactly $\mathcal{C}(\triangle_{I_n,I_0})$ where the class of Δ:

$$\triangle_{I_n,I_0} := \{\mathrm{Diag}(\delta_1, \delta_2, \ldots, \delta_n)\} \tag{1.19}$$

takes a special form of (1.9). So to some extent Kaszkurewicz and Bhaya's treatment of the $\underline{\phi}$-system is analogous to the LFT treatment of an uncertain linear system with uncertainty set $\triangle_{I_n,I_0}$. Indeed, for each fixed state-trajectory x, we may imagine $\underline{\phi}_{(i)}(x_{(i)[k]})$, $\underline{\phi}_{(i)} \in \mathcal{S}$ as $\boldsymbol{\delta}_{i[k]} \times x_{(i)[k]}$ where $\boldsymbol{\delta}_{i[k]}$ is a real number with $|\boldsymbol{\delta}_{i[k]}| \le 1$. Then $\underline{\phi}_{(i)} \circ \lambda \rightsquigarrow \delta_i := (\boldsymbol{\delta}_i \,\cdot\,) \circ \lambda$ and $\underline{\phi} \circ \lambda \rightsquigarrow \delta := \mathrm{Diag}(\delta_1, \delta_2, \ldots, \delta_n) \in \mathcal{B}(\triangle_{I_n,I_0})$. Since the stability result holds for the LFT system for all such δ, it also holds for the $\underline{\phi}$-system for all (admissible) state-trajectories x. Note that $\{\delta_i := (\boldsymbol{\delta}_i \,\cdot\,) \circ \lambda\}$ is only a special subclass of the general $\ell_2 \to \ell_2$ operators with norm ≤ 1. In particular $|\boldsymbol{\delta}_i| \le 1$ k-wise. Therefore one may conjecture that the stability condition (1.13) is only sufficient but not necessary. This is true but it turns out that typically the condition is not too conservative. (See Section 2.3.)[3]

Imagining $\underline{\phi}_{(i)}(x_{(i)[k]})$ as $\boldsymbol{\delta}_{i[k]} \times x_{(i)[k]}$ works well only when $\underline{\phi}_{(i)}$ are all independent, or in other words no two $\underline{\phi}_{(i)}$ are known to be identical. It is because $\boldsymbol{\delta}_{i[k]}$ is actually $x_{(i)[k]}$-dependent, and therefore $\underline{\phi}_{(i)} = \underline{\phi}_{(j)}$ does not necessarily imply $\boldsymbol{\delta}_{i[k]} = \boldsymbol{\delta}_{j[k]}$, so δ_i and δ_j will be treated as independent anyway. However, even when we are sure

[2] We use the symbol $\mathcal{S}$ since the condition $|\phi(s)| \le |s|$ is sometimes called the $[-1, 1]$ sector condition.

[3] We are a little bit sloppy about the kinds of stability. Actually (1.13) is about (some sort of) ℓ_2-stability (See [64] for details) which is not exactly GAS. So "sufficient but not necessary" in fact refers to the ℓ_2-stability and "not too conservative" is observed from the ℓ_2-ℓ_2 induced gain (more precisely, from the comparison between γ_d and γ_c in Section 2.3).

that $\underline{\phi}_{(i)}$ are all independent at the beginning, difficulties may arise in further analysis problems or synthesis problems. In analysis, for example, consider two $\underline{\phi}$-systems $G := \left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right)_{\underline{\phi}}$ and $\tilde{G} := \left(\begin{array}{c|c} \tilde{A} & \tilde{B} \\ \hline \tilde{C} & \tilde{D} \end{array}\right)_{\underline{\phi}}$. Sometimes we would like to know whether they are close, by studying the ℓ_2-ℓ_2 induced gain of the error system

$$E := G - \tilde{G} = \left(\begin{array}{cc|c} \tilde{A} & 0 & \tilde{B} \\ 0 & A & B \\ \hline -\tilde{C} & C & -\tilde{D}+D \end{array}\right)_{\left[\begin{smallmatrix} \underline{\phi} & \\ & \underline{\phi} \end{smallmatrix}\right]} \tag{1.20}$$

Note that every $\underline{\phi}_{(i)}$ must appear twice in the nonlinearity of E; and as mentioned above, although studying the ℓ_2-ℓ_2 induced gain of G and $\tilde{G}$ may be converted into studying the ℓ_2-ℓ_2 induced gain of $\mathcal{F}_u(\begin{bmatrix} A & B \\ C & D \end{bmatrix}, \delta)$ and $\mathcal{F}_u(\begin{bmatrix} \tilde{A} & \tilde{B} \\ \tilde{C} & \tilde{D} \end{bmatrix}, \delta)$ respectively, studying the ℓ_2-ℓ_2 induced gain of E cannot be converted into studying the ℓ_2-ℓ_2 induced gain of $\mathcal{F}_u(\left[\begin{array}{cc:c} \tilde{A} & 0 & \tilde{B} \\ 0 & A & B \\ \hdashline -\tilde{C} & C & -\tilde{D}+D \end{array}\right], \begin{bmatrix} \delta & \\ & \delta \end{bmatrix})$. Of course we may treat E as $\mathcal{F}_u(\left[\begin{array}{cc:c} \tilde{A} & 0 & \tilde{B} \\ 0 & A & B \\ \hdashline -\tilde{C} & C & -\tilde{D}+D \end{array}\right], \begin{bmatrix} \delta_1 & & & \\ & \delta_2 & & \\ & & \ddots & \\ & & & \delta_{2n} \end{bmatrix})$, but the ℓ_2-ℓ_2 induced gain determined in this way is typically too conservative to be useful: A conclusion of G, $\tilde{G}$ not close may be drawn only because their having the same nonlinearity $\underline{\phi}$ has not been taken into account. Problems may also arise in synthesis such as gain-scheduling, where certain $\underline{\phi}_{(i)}$ of the plant may appear in the controller and hence more than once in the overall closed-loop system. So, we must conclude that Kaszkurewicz and Bhaya's diagonal approach to the analysis of the $\underline{\phi}$-system has great limitations due to its inability to incorporate the knowledge that certain $\underline{\phi}_{(i)}$ are identical. Their approach is powerful when but only when the $\underline{\phi}_{(i)}$ are state-dependent perturbations which cannot be determined or measured, or used by the controller. The aim of this dissertation is to provide mathematical tools for handling general $\underline{\phi}$-systems in which certain $\underline{\phi}_{(i)}$ known to be identical can be taken account of. The ϕ-system will be studied first and then used as the basic building block for the more general $\underline{\phi}$-system (Section 5.1.1).

Remark 1.1 From (1.20) we see that the error between two $\underline{\phi}$-systems is still a $\underline{\phi}$-system. Similarly the error between two ϕ-systems (with the same ϕ) is still a ϕ-system. In fact for other manipulations such as cascading and feedback connection the resulting system is still of the same form. This nice property, of course, is not common to all classes of nonlinear systems. One way to realize this is that a $\underline{\phi}$-system or ϕ-system can be represented by a LFT as in (1.7) or (1.3), i.e. we can "pull out" all the nonlinearities just as the people "pull out" all the uncertainties of an uncertain linear system so that only a constant matrix remains. So indeed the A, B, C, D-matrices resulting from the parallel or series or feedback connection of two $\underline{\phi}$-systems or ϕ-systems are the same as those resulting from the same manipulation of the linear systems[4] $\otimes$

1.2.3 Recurrent neural networks

A special case of the ϕ-system is a class of recurrent artificial neural networks, which in turns can approximate (in some sense) a more general class of nonlinear systems. Since studying the *global* properties of general nonlinear systems is widely regarded as difficult (at least in the computational sense), this can be a motive of studying the ϕ-system, for which we shall be able to give global results on the stability and the ℓ_2-ℓ_2 induced gain. Of course, even without the approximation property, understanding the *dynamic* behaviour of the recurrent neural network still has its own importance.

The discussion in this section only applies to static $\phi : \mathbb{R} \to \mathbb{R}$ but not to dynamic $\phi : \ell_{2e} \to \ell_{2e}$. When $\phi : \mathbb{R} \to \mathbb{R}$, a ϕ-system may be labelled as a recurrent neural network, where A is the interconnection matrix, B the input weights, C the output weights, D the direct feedforward term which is usually 0, and ϕ the activation function of the neurons. Typical activation functions used in the neural network community are the sign function sgn, the semilinear function (i.e. the standard saturation $\mathrm{sat}(s) := s$ if $|s| \le 1$ and $\mathrm{sat}(s) := \mathrm{sgn}(s)$ if $|s| > 1$), and the hyperbolic tangent tanh. Note that sat and tanh $\in \mathcal{OL}(\mathbb{R})$ but sgn $\notin \mathcal{OL}(\mathbb{R})$. Moreover sgn does not have the desired approximation property, detailed below.

In fact Kaszkurewicz and Bhaya have noticed the connection between the $\underline{\phi}$-system and the recurrent neural network[44, 45]. However, all neurons in an artificial neural network usually have a unified activation function, and therefore identifying the neural network as a ϕ-system is much more appropriate. Indeed, if ϕ is one-one (and hence invertible on its range), the ϕ-system is equivalent to the class of recurrent neural networks studied by Sontag *et al.*[5] (See for example [72]):

$$x^+ = \phi(Ax + Bu); \quad y = Cx \tag{1.21}$$

They called it a σ-system, since they used the symbol σ instead of ϕ to denote the nonlinearity. However we shall use σ to denote the Hankel singular values etc. and therefore ϕ is used to denote the nonlinearity as in Kaszkurewicz and Bhaya's paper[43].

Albertini and Sontag *et al.* have obtained some results on the identifiability[5, 8, 6], observability[7] and forward accessibility[3] of the σ-system, based on different assumptions on the nonlinearity ϕ. Some are applicable to our case where $\phi \in \mathcal{OL}(\mathbb{R})$

[4] just be careful that no state-transformation should be involved.

[5] with some reasonable restriction on the initial state.

and some are not. Basically they have focused on "very nonlinear" ϕ but we shall not exclude the case where $\phi =$ the identity (which corresponds to a linear system). So the treatments are very different. The aim of this dissertation is to incorporate the class of recurrent neural networks in concern into the (extended) LFT framework so that certain well-known analysis and synthesis results can be generalized to it, and our focus will be on the (dynamic) input to output behaviour rather than the zero-input initial state to steady-state output behaviour considered by some researchers in the neural network community.

In [72] Sontag demonstrated how to use a σ-system to approximate the general nonlinear system

$$x^+ = f(x,u); \quad y = h(x) \tag{1.22}$$

where $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$, $h : \mathbb{R}^n \to \mathbb{R}^p$ are continuous. However, we shall *not* quote this result directly. The detailed reason is given in Appendix A. In brief, note that the ϕ-system is a special kind of recurrent neural network in the sense that it does not have any bias terms. A more general recurrent neural network may be

$$\begin{bmatrix} x^+ \\ y \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} \phi(x) \\ u \end{bmatrix} + \begin{bmatrix} \mu \\ \nu \end{bmatrix} \tag{1.23}$$

(where $\mu \in \mathbb{R}^n$, $\nu \in \mathbb{R}^p$) or

$$\begin{bmatrix} x^+ \\ y \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} \phi(x+\mu) - \nu \\ u \end{bmatrix} \tag{1.24}$$

(where $\mu, \nu \in \mathbb{R}^n$). In order to approximate the general nonlinear system (1.22), bias terms are actually needed. The origin being an equilibrium of the general nonlinear system does *not* remove the bias terms automatically. For example, we may choose $\nu = \phi(\mu)$ in (1.24) to keep the origin an equilibrium. Sontag removed the bias terms by augmenting the state, which *deliberately* de-stabilizes the origin of the augmented system and introduces another equilibrium somewhere else in order to simulate the effect of the bias terms. This approach, however, is not compatible with our treatment in later chapters in which the origin is always stable or to be stabilized. So here we give our own approximation result on a slightly restricted class of general nonlinear systems. Still, the treatment is modified from Sontag's[72].

Definition 1.1 A function $\phi : \mathbb{R} \to \mathbb{R}$ is called odd universal[6] if for any closed interval $K \subset \mathbb{R}$ such that $\chi \in K \Leftrightarrow -\chi \in K$, the space of functions

$$\left\{ F_\beta \circ \phi \circ F_\alpha \mid_K \;\middle|\; F_\alpha : \mathbb{R} \to \mathbb{R}^l, F_\beta : \mathbb{R}^l \to \mathbb{R} \text{ are linear}, l \in \mathbb{Z}_+ \right\} \tag{1.25}$$

(i.e. $F_\alpha(\bullet) := \zeta_\alpha \bullet$, $\zeta_\alpha \in \mathbb{R}^l$; $F_\beta(\bullet) := \zeta_\beta^* \bullet$, $\zeta_\beta \in \mathbb{R}^l$) is dense in the space of odd continuous functions: $K \to \mathbb{R}$ in sup-norm. $\triangleleft$

[6] Comparing with the definition of being universal in [72], here F_α, F_β are restricted to *linear* rather than *affine* and therefore $\{F_\beta \circ \phi \circ F_\alpha \mid_K\}$ is only required to be dense in the space of *odd* continuous functions.

For example, the sine function $\sin \in \mathcal{OL}(\mathrm{I\!R})$ is odd universal. In particular we may choose $\zeta_\alpha = \begin{bmatrix} 1 \\ 2 \\ \vdots \\ l \end{bmatrix}$. Then every odd continuous function f on K has a Fourier series which contains only sine terms, and the arithmetic mean of the partial sum of the Fourier series[7] converges uniformly to f. This is Fejér's theorem. (See for example [76, Sec. 5.2].) Another example is the semilinear function $\mathrm{sat} \in \mathcal{OL}(\mathrm{I\!R})$. Note that any odd continuous function f on K may be uniformly approximated by an odd piecewise-affine continuous function, with only a *finite* number of affine pieces[8]. Let $1/\zeta_{\alpha(1)}, 1/\zeta_{\alpha(2)}, \ldots, 1/\zeta_{\alpha(l)}$ be positive real numbers in the ascending order, corresponding to the end-points of the affine pieces with slopes $\eta_{(1)}, \eta_{(2)}, \ldots, \eta_{(l)}$ respectively. Then ζ_β is given by

$$\begin{bmatrix} 1 & 1 & \cdots & 1 \\ 0 & 1 & \cdots & 1 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 \end{bmatrix} \begin{bmatrix} \zeta_{\beta(1)}\zeta_{\alpha(1)} \\ \zeta_{\beta(2)}\zeta_{\alpha(2)} \\ \vdots \\ \zeta_{\beta(l)}\zeta_{\alpha(l)} \end{bmatrix} = \begin{bmatrix} \eta_{(1)} \\ \eta_{(2)} \\ \vdots \\ \eta_{(l)} \end{bmatrix} \tag{1.26}$$

On the other hand, neither $(\bullet)^r$ ($\in \mathcal{OL}(\mathrm{I\!R})$ if and only if $r = 1$) nor $\mathrm{sgn}(\bullet)$ ($\notin \mathcal{OL}(\mathrm{I\!R})$) is odd universal, since $F_\beta \circ \phi \circ F_\alpha(\bullet)$ is only a constant multiple of $(\bullet)^r$ and $\mathrm{sgn}(\bullet)$ respectively. It is interesting to note that, if bias terms are allowed, F_α, F_β may be relaxed from linear to affine, and then sgn will be universal. See Appendix A for details.

Now, if ϕ is odd universal, then for any compact set $K \subset \mathrm{I\!R}^m$ such that $\chi \in K \Leftrightarrow -\chi \in K$, the set of functions

$$\left\{ F_\beta \circ \phi \circ F_\alpha|_K \;\middle|\; F_\alpha : \mathrm{I\!R}^m \to \mathrm{I\!R}^l, F_\beta : \mathrm{I\!R}^l \to \mathrm{I\!R}^p \text{ are linear}, l \in \mathbb{Z}_+ \right\} \tag{1.27}$$

is dense in the set of odd continuous functions: $K \to \mathrm{I\!R}^p$ in sup-norm. To prove this, let $f : K \to \mathrm{I\!R}^p$ be odd continuous. We may work on $f_{(i)}$, $i = 1, 2, \ldots, p$ one by one, and hence may assume without loss of generality (WLOG) $p = 1$. Every continuous function: $K \to \mathrm{I\!R}$ can be approximated uniformly by a m-dimensional trigonometric series[9], and each product term can be expanded as a sum using the standard sum-and-product rule. Hence we obtain

$$f \approx \hat{F}_\beta \circ \cos \circ \hat{F}_\alpha + \tilde{F}_\beta \circ \sin \circ \tilde{F}_\alpha = \tilde{f}_{\text{even}} + \tilde{f}_{\text{odd}} = \tilde{f} \tag{1.28}$$

[7] The partial sum of the Fourier series itself may fail to converge, however. See for example [50, Sec. 4.7-5].

[8] The number of affine pieces depends on the function f and the prescribed error ϵ but can be easily noted to be finite: Since f is continuous, every point in K is inside an open interval within which f changes $< \epsilon$. These open intervals form an open cover of K, which has a finite subcover since K is compact.

[9] Alternatively, one may use a m-dimensional power series as in [72] instead of a trigonometric series to approximate f to prove the result.

on K. Since f is odd, $\tilde{f}_{\text{even}}$ may be ignored: $|f - \tilde{f}| < \epsilon$ on K implies $|f - \tilde{f}_{\text{odd}}| < \epsilon$ on K. At last sin is an odd continuous function, so it can be approximated by ϕ in the sense of Definition 1.1.

$F_\beta \circ \phi \circ F_\alpha$ is called a single hidden layer feedforward neural network (1HLN) with m inputs, l hidden neurons, p outputs and no bias terms. Hence a 1HLN with no bias terms can approximate uniformly any given odd continuous function on any given compact set provided that the number of hidden neurons l is sufficiently large. Now consider the general nonlinear system

$$x^+ = f(x, u); \quad y = h(x) \tag{1.29}$$

where $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$, $h : \mathbb{R}^n \to \mathbb{R}^p$ are odd continuous, i.e. $f(-\zeta, -\eta) = -f(\zeta, \eta)$, $h(-\zeta) = -h(\zeta)$ for all $\zeta \in \mathbb{R}^n$, $\eta \in \mathbb{R}^m$. Note that (1.29) is slightly restricted comparing with (1.22) considered by Sontag, in the sense that f and h should be odd here. WLOG we may assume $y = Cx$, since we can always introduce an additional state $\hat{x} := h(x)$ such that $\hat{x}^+ = h(x^+) = h(f(x, u))$ and $y = \hat{x}$. Note that $h \circ f$ is still odd. Now we use a 1HLN with no bias terms to approximate $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ to obtain

$$x^+ = T\phi(Ax + Bu); \quad y = Cx \tag{1.30}$$

We may assume T has full row rank, since otherwise a state-transformation will change T to $\begin{bmatrix} T_1 \\ 0 \end{bmatrix}$ where T_1 has full row rank, and then we only keep the states corresponding to T_1. Now consider the system

$$\xi^+ = \phi(AT\xi + Bu); \quad y = CT\xi \tag{1.31}$$

Since T has full row rank (i.e. onto), for every $x_{[0]}$ there exists $\xi_{[0]}$ such that $x_{[0]} = T\xi_{[0]}$. Then if (ξ, y) solves (1.31), $(x, y) = (T\xi, y)$ will solve (1.30). Note that (1.31) is a σ-system. Now consider the ϕ-system

$$\varsigma^+ = AT\phi(\varsigma) + Bu; \quad y = CT\phi(\varsigma) \tag{1.32}$$

If we assume $\xi_{[0]}$ is in the range of ϕ (which is reasonable since $\xi_{[k]}$, $k > 0$, will be in the range of ϕ anyway), there exists $\varsigma_{[0]}$ such that $\xi_{[0]} = \phi(\varsigma_{[0]})$. Then similarly if (ς, y) solves (1.32), $(\xi, y) = (\phi(\varsigma), y)$ will solve (1.31). Hence (1.32) is the desired approximant. Note that $T \circ \phi : \varsigma_{[k]} \mapsto x_{[k]}$ is onto and origin-preserving, i.e. each trajectory of (1.30) can be simulated by a trajectory of (1.32), and the origin of (1.30) is GAS if the origin of (1.32) is. Note also that ϕ need not be one-one for (1.32) to approximate (1.31): We were *not* showing that they were *equivalent*.

Such approximation was found to be useful in signal processing[52], for example. However, one should not be too enthusiastic about the approximation property: The approximation is valid on compact subsets of $\mathbb{R}^n \times \mathbb{R}^m$ and for finite time only. Also the approximation results should be interpreted carefully. The state-trajectories being close may mean that all trajectories still tend to somewhere very near the origin but the origin has become locally unstable, or the ℓ_2-ℓ_2 induced gain is still almost the same for large inputs but has increased dramatically for extremely small ones. As we shall be studying the global (in $x_{[0]}$) stability of the origin and the global (in u) ℓ_2-ℓ_2 induced gain in later chapters, perhaps we should consider ourselves studying the dynamic behaviour of a recurrent neural network more than the dynamic behaviour of a general nonlinear system by means of a recurrent neural network. Still, the recurrent neural network we are studying is somewhat special: it has no bias terms and the origin is globally stable or to be globally stabilized (i.e. there will be no other equilibria). On the other hand we shall allow ϕ to be dynamic and uncertain as long as $\phi \in \mathcal{OL}(\ell_{2e})$.

1.3 Outline of the dissertation

The dissertation is organized as follows.

Chapter 2: Analysis

In Chapter 2 we first introduce the positive definite diagonally dominant (pdd) matrices and their connection to the ϕ-system, $\phi \in \mathcal{OL}(\mathrm{I\!R})$. Indeed, most of the results in this dissertation are based on finding pdd matrices subject to appropriate LMIs, which will be shown computationally tractable. The global asymptotic stability and ℓ_2-ℓ_2 induced gain of the ϕ-system are then concluded. The chapter is closed with several remarks on the limitations of the pdd approach which will affect the synthesis and model reduction results in Chapter 3 and 4.

Chapter 3: Synthesis

In Chapter 3 the design of a ϕ-controller for a given ϕ-system is studied. Conditions on parametrization of all stabilizing output-feedback ϕ-controllers are given. The treatment is based on a standard approach in the linear system theory where the Full Information, Full Control, Disturbance Feedforward and Output Estimation problems are solved first. Conditions on closed-loop performance synthesis are also given.

Chapter 4: Model reduction

Chapter 4 is devoted to the model reduction of a stable ϕ-system. The approach adopted is close to the balanced truncation of a stable linear system. The ideas of the observability function and the controllability function are first revised. Then a result on the balanced truncation of the $\underline{\phi}$-system is given, which is however rather weak since $\underline{\phi}_{(i)}$ are assumed different. A stronger model reduction result for the ϕ-system is derived next, and based on it a model reduction algorithm is proposed which also requires a method to minimize the rank of a matrix subject to a LMI or alternatively a pair of good estimates to the observability function and the controllability function.

Chapter 5: Extended results

Chapter 5 extends the basic results obtained in Chapter 2–4 to a more general form of nonlinear systems in three directions. The first one is the general $\underline{\phi}$-system where some but not all $\underline{\phi}_{(i)}$ are identical, the second one is the system with ϕ and λ separate, and the third one is the ϕ-system with a dynamic $\phi \in \mathcal{OL}(\ell_{2e})$ rather than a static $\phi \in \mathcal{OL}(\mathbb{R})$. An interesting example is also given to illustrate certain considerations in applying those basic results.

Chapter 6: Conclusions

We conclude this dissertation with a summary of the results, an outline of its contributions, and several suggestions for future work which are mostly unsolved problems that remain in this dissertation.

Several supplementary results and proofs can be found in the appendices.

Chapter 2

Analysis

2.1 Positive diagonally dominant matrices

In this section we introduce positive diagonally dominant matrices and their properties, which will play an important role in all analysis and synthesis of the ϕ-system. In particular, Lemma 2.2 and Lemma 2.5 are the corner stones of this dissertation in the theoretical and the computational aspects respectively.

Definition 2.1 A matrix P is said to be positive diagonally dominant (pdd for short) if $P > 0$ (positive definite, i.e. P is symmetric such that $\zeta^* P \zeta > 0$ for all $\zeta \neq 0$) and (row) diagonally dominant:

$$\forall i, \quad |p_{(ii)}| \geq \sum_{j \neq i} |p_{(ij)}| \tag{2.1}$$

$\triangleleft$

The following basic tool is needed to prove Lemma 2.2 and 2.3.

Lemma 2.1 If P is pdd and $\forall i, j$,

$$|\alpha_{(ij)} + \alpha_{(ji)}| \leq -(\alpha_{(ii)} + \alpha_{(jj)}) \tag{2.2}$$

then

$$\sum_{i,j} p_{(ij)} \alpha_{(ij)} \leq 0 \tag{2.3}$$

$\triangleleft$

Proof. Note that $p_{(ij)} = p_{(ji)}$. Then

$$\begin{aligned}
2 \sum_{i,j \neq i} p_{(ij)} \alpha_{(ij)} &= \sum_{i,j \neq i} p_{(ij)} (\alpha_{(ij)} + \alpha_{(ji)}) \\
&\leq \sum_{i,j \neq i} |p_{(ij)}| |\alpha_{(ij)} + \alpha_{(ji)}| \\
&\leq -\sum_{i,j \neq i} |p_{(ij)}| (\alpha_{(ii)} + \alpha_{(jj)})
\end{aligned}$$

$$\begin{aligned} &= -2\sum_{i,j\neq i}|p_{(ij)}|\alpha_{(ii)} \\ &\leq -2\sum_i p_{(ii)}\alpha_{(ii)} \end{aligned}$$

where the last inequality follows from $-\alpha_{(ii)} \geq 0$ (from (2.2)), (2.1), and $p_{(ii)} \geq 0$ (since $P > 0$). ∎

The following lemma gives a characterization of pdd matrices.

Lemma 2.2 *P is pdd if and only if $P > 0$ and*

$$\forall\phi \in \mathcal{OL}(\mathbb{R}),\ \ \forall\zeta \in \mathbb{R}^n,\ \ \phi(\zeta)^* P\phi(\zeta) \leq \zeta^* P\zeta \tag{2.4}$$

◁

Proof. [ONLY IF] $\phi \in \mathcal{OL}(\mathbb{R})$ means that $\forall s, t \in \mathbb{R}$,

$$|\phi(s) + \phi(t)| \leq |s + t| \tag{2.5}$$
$$|\phi(s) - \phi(t)| \leq |s - t| \tag{2.6}$$

or equivalently

$$(\phi(s) \pm \phi(t))^2 \leq (s \pm t)^2 \tag{2.7}$$

which implies

$$2|\phi(s)\phi(t) - st| \leq -\left(\left(\phi(s)^2 - s^2\right) + \left(\phi(t)^2 - t^2\right)\right) \tag{2.8}$$

Note that

$$\begin{aligned} \phi(\zeta)^* P\phi(\zeta) - \zeta^* P\zeta &= \sum_{i,j}\phi(\zeta_{(i)})p_{(ij)}\phi(\zeta_{(j)}) - \sum_{i,j}\zeta_{(i)}p_{(ij)}\zeta_{(j)} \\ &= \sum_{i,j}p_{(ij)}(\phi(\zeta_{(i)})\phi(\zeta_{(j)}) - \zeta_{(i)}\zeta_{(j)}) \end{aligned} \tag{2.9}$$

Let $\alpha_{(ij)} = \phi(\zeta_{(i)})\phi(\zeta_{(j)}) - \zeta_{(i)}\zeta_{(j)}$. Then (2.8) implies (2.2) and hence Lemma 2.1 applies.

[IF] We want to show that if $P > 0$ but not diagonally dominant, then $\exists\phi \in \mathcal{OL}(\mathbb{R})$, $\exists\zeta \in \mathbb{R}^n$ such that

$$\zeta^* P\zeta - \phi(\zeta)^* P\phi(\zeta) < 0$$

or equivalently

$$(\zeta - \phi(\zeta))^* P(\zeta + \phi(\zeta)) < 0$$

In fact we may always choose $\phi = \text{sat}$. Since P is not diagonally dominant, $\exists i$, $\exists\epsilon > 0$ such that

$$(2 + \epsilon)p_{(ii)} < \sum_{j\neq i} 2|p_{(ij)}|$$

So we simply let

$$\zeta_{(j)} + \phi(\zeta_{(j)}) = \begin{cases} -2\mathrm{sgn}(p_{(ij)}) & \text{if } \; j \neq i \\ 2+\epsilon & \text{if } \; j = i \end{cases} \tag{2.10}$$

$$\zeta_{(j)} - \phi(\zeta_{(j)}) = \begin{cases} 0 & \text{if } \; j \neq i \\ \epsilon & \text{if } \; j = i \end{cases} \tag{2.11}$$

or in other words

$$\zeta_{(j)} = \begin{cases} -\mathrm{sgn}(p_{(ij)}) & \text{if } \; j \neq i \\ 1+\epsilon & \text{if } \; j = i \end{cases} \tag{2.12}$$

■

Remark 2.1 In general,

$$\forall \phi \in \mathcal{OL}(\mathbb{R}), \;\; \forall \zeta \in \mathbb{R}^n, \;\; \phi(\zeta)^* H_1 \phi(\zeta) \leq \zeta^* H_0 \zeta \tag{2.13}$$

if the separation of H_1 and H_0 is large enough to insert a pdd matrix in between, i.e. there exists a pdd Q such that

$$H_1 \leq Q \leq H_0 \tag{2.14}$$

⊗

An extended version of Lemma 2.2 is the following, which may be of interest but will not be needed in this dissertation.

Lemma 2.3 The followings are equivalent:

[P1] P is pdd.

[P2] $P > 0$, and $\forall \phi \in \mathcal{OL}(\mathbb{R})$, $\forall \zeta \in \mathbb{R}^n$, $\phi(\zeta)^* P \phi(\zeta) \leq \zeta^* P \zeta$.

[P3] $P > 0$, and $\forall \phi \in \mathcal{OL}(\mathbb{R})$, $\forall \zeta \in \mathbb{R}^n$, $\zeta^* P \phi(\zeta) \leq \zeta^* P \zeta$.

[P4] $P > 0$, and $\forall \phi \in \mathcal{OL}(\mathbb{R})$ such that $s \leq t \Rightarrow \phi(s) \leq \phi(t)$ (i.e. monotonically increasing) and $\forall \zeta \in \mathbb{R}^n$, $\phi(\zeta)^* P \phi(\zeta) \leq \zeta^* P \phi(\zeta) \leq \zeta^* P \zeta$.

◁

Proof. In Appendix B. ■

Obviously all diagonal blocks of a pdd matrix are pdd, but less obvious is that so are the Schur complements.

Lemma 2.4 Let $P = \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix}$ be pdd ($P_{21} = P_{12}^*$). Then $P_{11}^{\mho} := P_{22} - P_{21}P_{11}^{-1}P_{12}$ (the Schur complement of P_{11}) and $P_{22}^{\mho} := P_{11} - P_{12}P_{22}^{-1}P_{21}$ (the Schur complement of P_{22}) are both pdd. ◁

Proof. Lemma 2.2 says that $\forall \phi \in \mathcal{OL}(\mathbb{R})$, $\forall \zeta_1, \zeta_2$,

$$\begin{bmatrix} \phi(\zeta_1) \\ \phi(\zeta_2) \end{bmatrix}^* \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} \begin{bmatrix} \phi(\zeta_1) \\ \phi(\zeta_2) \end{bmatrix} \leq \begin{bmatrix} \zeta_1 \\ \zeta_2 \end{bmatrix}^* \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} \begin{bmatrix} \zeta_1 \\ \zeta_2 \end{bmatrix}$$

Now write

$$\begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} = \begin{bmatrix} I & 0 \\ P_{21}P_{11}^{-1} & I \end{bmatrix} \begin{bmatrix} P_{11} & 0 \\ 0 & P_{11}^{\mho} \end{bmatrix} \begin{bmatrix} I & P_{11}^{-1}P_{12} \\ 0 & I \end{bmatrix} \tag{2.15}$$

and let $\zeta_1 = -P_{11}^{-1}P_{12}\zeta_2$ to obtain

$$\heartsuit^* P_{11} \heartsuit + \phi(\zeta_2)^* P_{11}^{\mho} \phi(\zeta_2) \leq \zeta_2^* P_{11}^{\mho} \zeta_2$$

for some $\heartsuit$. Hence

$$\phi(\zeta_2)^* P_{11}^{\mho} \phi(\zeta_2) \leq \zeta_2^* P_{11}^{\mho} \zeta_2$$

for all $\phi \in \mathcal{OL}(\mathbb{R})$ and all ζ_2, and by Lemma 2.2 $P_{11}^{\mho}$ is pdd. The proof for $P_{22}^{\mho}$ is similar. ∎

Remark 2.2 The converse is certainly not true, i.e. P_{11} and its Schur complement both pdd does not imply P pdd. Any positive definite but not pdd 2×2 matrix P is a counterexample. ⊗

The connection to the LMI is given by the following lemma. Obviously specifying a matrix to be pdd is a LMI, so the non-trivial part of the following lemma is actually in the computational complexity.

Lemma 2.5 *P is pdd if and only if $P > 0$ and there exists a symmetric R such that*

$$\begin{aligned} \forall i \neq j, \quad & r_{(ij)} \geq 0, \;\; p_{(ij)} + r_{(ij)} \geq 0 \\ \forall i, \quad & p_{(ii)} \geq \sum_{j \neq i} (p_{(ij)} + 2r_{(ij)}) \end{aligned}$$

which involves only $n(n-1)/2$ variables $r_{(ij)}$ in addition to $p_{(ij)}$ and n^2 inequalities in addition to $P > 0$. ◁

Proof.[1] Sufficiency is obvious since

$$\begin{aligned} p_{(ii)} & \geq \sum_{j \neq i} (p_{(ij)} + 2r_{(ij)}) \\ & = \sum_{j \neq i} (|p_{(ij)} + r_{(ij)}| + |-r_{(ij)}|) \\ & \geq \sum_{j \neq i} |p_{(ij)}| \end{aligned}$$

Necessity follows from choosing

$$r_{(ij)} = \begin{cases} -p_{(ij)} & \text{if } \;\; p_{(ij)} < 0 \\ 0 & \text{if } \;\; p_{(ij)} \geq 0 \end{cases} \tag{2.16}$$

∎

[1]In fact we only decompose $p_{(ij)}$ into its positive part and negative part, i.e. $p_{(ij)} = p_{(ij)}^{+} - p_{(ij)}^{-}$ which is indeed rather standard in solving this kind of problem.

Hence, the existence of a pdd matrix satisfying those LMIs arising from standard analysis and synthesis problems is still a LMI overall, with the additional computational complexity grows reasonably with the dimension n. This guarantees that most theoretical results presented in this dissertation which are based on the pdd matrices will remain computationally tractable.

2.2 Stability

The main objective of this section is to conclude the stability for the ϕ-system (Proposition 2.6). We also present a few tools to be used in the synthesis problems in Chapter 3.

Definition 2.2 The matrix A or the system $\left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right)$ is said to be linearly stable / dd-stable / diagonally stable if there exists a positive definite / pdd / positive diagonal P such that

$$A^*PA - P < 0 \tag{2.17}$$

$\triangleleft$

Certainly diagonal stability implies dd-stability implies linear stability, but in general they are not the same. Numerical examples can be easily found.

Example 2.1 Let

$$A = \begin{bmatrix} 1.93 & -2.11 \\ 0.61 & -0.16 \end{bmatrix}$$

Then A is linearly stable but not dd-stable. If we let

$$A = \begin{bmatrix} 0.25 & -1.39 \\ 0.84 & -0.68 \end{bmatrix}$$

then A is dd-stable but not diagonally stable. A pdd P is given by

$$P = \begin{bmatrix} 1.1626 & -0.6435 \\ -0.6435 & 1.9235 \end{bmatrix}$$

$\otimes$

Each of the three kinds of stability may correspond to the global asymptotic stability of a kind of system. For example, A is linearly stable (or more commonly simply called stable) if and only if (the origin of) the *linear* system $\left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right)_I$ is globally asymptotically stable. For the dd-stability, we have the following

Proposition 2.6 If A is dd-stable, then $\forall \phi \in \mathcal{OL}(\mathbb{R})$, (the origin of) the ϕ-system $\left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right)_{\phi}$ is globally asymptotically stable (i.e. $x^+ = A\phi(x)$ is Lyapunov stable and $\forall x_{[0]}$, $x_{[k]} \to 0$ as $k \to \infty$). $\triangleleft$

Proof. Note that WLOG (2.17) may be strengthened to

$$A^*PA - P \leq -I \tag{2.18}$$

by scaling P. Let $V(\zeta) := \zeta^* P \zeta$. Then (2.18) implies that

$$\begin{aligned} V(x_{[k+1]}) - V(x_{[k]}) &= x^*_{[k+1]} P x_{[k+1]} - x^*_{[k]} P x_{[k]} \\ &= \phi(x_{[k]})^* A^* P A \phi(x_{[k]}) - x^*_{[k]} P x_{[k]} \\ &= \Big(\phi(x_{[k]})^* A^* P A \phi(x_{[k]}) - \phi(x_{[k]})^* P \phi(x_{[k]})\Big) - \\ &\quad\ \Big(x^*_{[k]} P x_{[k]} - \phi(x_{[k]})^* P \phi(x_{[k]})\Big) \\ &\leq -\phi(x_{[k]})^* \phi(x_{[k]}) - \Big(x^*_{[k]} P x_{[k]} - \phi(x_{[k]})^* P \phi(x_{[k]})\Big) \end{aligned}$$

which together with Lemma 2.2 imply that V is a Lyapunov function whose decrement is bounded away from 0 unless $\phi(x_{[k]})^* \phi(x_{[k]}) \to 0$ and $x^*_{[k]} P x_{[k]} - \phi(x_{[k]})^* P \phi(x_{[k]}) \to 0$ (since both terms are non-negative). Hence $x_{[k]} \to 0$ as $k \to \infty$. ■

For the diagonal stability, we have

Proposition 2.7 If A is diagonally stable, then (the origin of) the system

$$\begin{bmatrix} x^+ \\ y \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} \xi \\ u \end{bmatrix}; \quad \forall k, i, \quad |\xi_{[k](i)}| \leq |x_{[k](i)}| \tag{2.19}$$

is globally asymptotically stable. $\triangleleft$

Proof. Note that if P is diagonal, then

$$\xi^*_{[k]} P \xi_{[k]} = \sum_i p_{(i)} |\xi_{[k](i)}|^2 \leq \sum_i p_{(i)} |x_{[k](i)}|^2 = x^*_{[k]} P x_{[k]}$$

Then the rest of the proof is similar to that of Proposition 2.6. ■

Proposition 2.7 is basically the same as the diagonal stability result obtained by Kaszkurewicz and Bhaya in [43]; we only write the system in the more general looking form (2.19) in which we do not care what ξ is as long as it satisfies the constraint $|\xi_{[k](i)}| \leq |x_{[k](i)}|$. This form will be more useful later.

Remark 2.3 It may be of interest to know that the diagonal stability result has a continuous-time counterpart[66, 43]. More precisely, the origin of

$$\dot{x} = A\underline{\phi}(x) \tag{2.20}$$

where $\underline{\phi}_{(i)}$ are smooth enough to guarantee the existence and uniqueness of a solution, $\underline{\phi}_{(i)}(0) = 0$, $\underline{\phi}_{(i)}(s) \times s > 0$ for all nonzero $s \in \mathbb{R}$, and

$$\int_0^s \underline{\phi}_{(i)}(\alpha)d\alpha \to \infty \quad \text{as} \quad |s| \to \infty$$

is GAS if there exists a positive diagonal P such that

$$PA + A^*P < 0 \tag{2.21}$$

A Lyapunov function is given by

$$V(\zeta) := 2\sum_i p_{(i)} \int_0^{\zeta_{(i)}} \underline{\phi}_{(i)}(\alpha)d\alpha \tag{2.22}$$

Several generalizations can be found in [43]. The study of the existence of a positive diagonal solution P to the continuous-time Lyapunov inequality (2.21) and the characterizations of such A has already a long history (see for example [11, 23, 16, 17, 47, 37]), although its discrete-time counterpart has received less attention. A large class of matrices closely related to the diagonal stability is the class of M-matrices[18, Ch. 6] which has been extensively used in the large-scale systems literature[73]. $\otimes$

We close this section with two lemmas about block-triangular A or cascaded systems, which will be useful in the synthesis problems in Chapter 3.

Lemma 2.8 Let $A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$ where A_{12} or $A_{21} = 0$. Then A is linearly stable / dd-stable / diagonally stable if and only if both A_{11} and A_{22} are linearly stable / dd-stable / diagonally stable. $\triangleleft$

Proof. We prove the case $A_{21} = 0$ and the other one is similar. Note that

$$\begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}^* \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix} - \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} < 0$$

if and only if

$$\begin{bmatrix} A_{11} & \heartsuit \\ 0 & A_{22} \end{bmatrix}^* \begin{bmatrix} P_{11} & 0 \\ 0 & P_{11}^{\mho} \end{bmatrix} \begin{bmatrix} A_{11} & \heartsuit \\ 0 & A_{22} \end{bmatrix} - \begin{bmatrix} P_{11} & 0 \\ 0 & P_{11}^{\mho} \end{bmatrix} < 0 \tag{2.23}$$

by (2.15) (where $\heartsuit$ is something independent of P_{22}). Expanding (2.23) we get

$$\begin{bmatrix} A_{11}^*P_{11}A_{11} - P_{11} & \clubsuit \\ \clubsuit^* & \heartsuit^*P_{11}\heartsuit + A_{22}^*P_{11}^{\mho}A_{22} - P_{11}^{\mho} \end{bmatrix} < 0 \tag{2.24}$$

for some $\clubsuit$ independent of P_{22}. Now sufficiency is clear since if P_{11} is such that

$$A_{11}^*P_{11}A_{11} - P_{11} < 0$$

and P_{22} is such that
$$A_{22}^* P_{22} A_{22} - P_{22} < 0$$
then $P = \begin{bmatrix} P_{11} & 0 \\ 0 & P_{22} \end{bmatrix}$ satisfies (2.24) provided that P_{22} $(= P_{11}^{\mho}$ in this case) is chosen to be sufficiently large (by scaling) since $\heartsuit$ and $\clubsuit$ are independent of P_{22}. Necessity also follows easily since (2.24) implies
$$A_{11}^* P_{11} A_{11} - P_{11} < 0$$
and
$$\heartsuit^* P_{11} \heartsuit + A_{22}^* P_{11}^{\mho} A_{22} - P_{11}^{\mho} < 0$$
and hence
$$A_{22}^* P_{11}^{\mho} A_{22} - P_{11}^{\mho} < 0$$
and $P_{11}^{\mho}$ is positive definite / pdd / positive diagonal if P is positive definite / pdd / positive diagonal. (For pdd we have used Lemma 2.4.) ■

Hence, for example, a cascaded system is dd-stable if and only if its individual components are all dd-stable.

Now, let us consider
$$y \longleftarrow \boxed{G} \longleftarrow \boxed{G_1} \longleftarrow u$$
where G is a standard ϕ-system
$$x^+ = A\phi(x) + By_1; \quad y = C\phi(x) + Dy_1$$
(y_1 is the output of G_1), but G_1 is "influenced" by partial states of G as follows: Partition x into $\begin{bmatrix} x_2 \\ x_3 \end{bmatrix}$ then G_1 is described by (either $+$ or $-$)
$$(x_1 \pm x_2)^+ = A_{11}(\phi(x_1) \pm \phi(x_2)) + B_1 u; \quad y_1 = C_1(\phi(x_1) \pm \phi(x_2)) + D_1 u$$
If $\phi \in \mathcal{OL}(\mathbb{R})$, $|\phi(x_{1(\bullet)}) \pm \phi(x_{2(\bullet)})| \le |x_{1(\bullet)} \pm x_{2(\bullet)}|$ and hence G_1 will have the form (2.19). So it would be nice to have A_{11} diagonally stable.

Lemma 2.9 Let $\phi \in \mathcal{OL}(\mathbb{R})$. The system described by
$$\begin{bmatrix} (x_1 \pm x_2)^+ \\ x_2^+ \\ x_3^+ \end{bmatrix} = \begin{bmatrix} A_{11} & 0 & 0 \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix} \begin{bmatrix} \phi(x_1) \pm \phi(x_2) \\ \phi(x_2) \\ \phi(x_3) \end{bmatrix} \tag{2.25}$$
or equivalently
$$\begin{bmatrix} x_1^+ \\ x_2^+ \\ x_3^+ \end{bmatrix} = \begin{bmatrix} A_{11} \mp A_{21} & \pm A_{11} - A_{21} \mp A_{22} & \mp A_{23} \\ A_{21} & \pm A_{21} + A_{22} & A_{23} \\ A_{31} & \pm A_{31} + A_{32} & A_{33} \end{bmatrix} \begin{bmatrix} \phi(x_1) \\ \phi(x_2) \\ \phi(x_3) \end{bmatrix} \tag{2.26}$$

is dd-stable if A_{11} is diagonally stable and $\begin{bmatrix} A_{22} & A_{23} \\ A_{32} & A_{33} \end{bmatrix}$ is dd-stable. ◁

Proof. From the proof of Lemma 2.8 we see that if A_{11} is diagonally stable and $\begin{bmatrix} A_{22} & A_{23} \\ A_{32} & A_{33} \end{bmatrix}$ is dd-stable, then there exists $P = \begin{bmatrix} P_{11} & 0 & 0 \\ 0 & P_{22} & P_{23} \\ 0 & P_{23}^* & P_{33} \end{bmatrix}$ where P_{11} is positive diagonal, $\begin{bmatrix} P_{22} & P_{23} \\ P_{23}^* & P_{33} \end{bmatrix}$ is pdd, such that

$$\begin{bmatrix} A_{11} & 0 & 0 \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix}^* P \begin{bmatrix} A_{11} & 0 & 0 \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix} - P < 0$$

Now it suffices to note that

$$\begin{bmatrix} I & \pm I & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix}^* P \begin{bmatrix} I & \pm I & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix} = \begin{bmatrix} P_{11} & \pm P_{11} & 0 \\ \pm P_{11} & P_{11} + P_{22} & P_{23} \\ 0 & P_{23}^* & P_{33} \end{bmatrix}$$

is pdd. ■

Remark 2.4 Since G_1 precedes G, the influence on G_1 by G is kind of feedback and therefore the stability of the "cascaded" system is not as easy as when G_1 is placed *after* G. In the latter case it should be rather obvious that A_{11} only needs to be dd-stable instead of diagonally stable, i.e. it suffices to have both A_{11} and $\begin{bmatrix} A_{22} & A_{23} \\ A_{32} & A_{33} \end{bmatrix}$ dd-stable, since the A-matrix in (2.25) will then be upper-triangular and so will the A-matrix in (2.26), and both have the same diagonal blocks. ⊗

2.3 ℓ_2-ℓ_2 induced gain

The objective of this section is to give a satisfactory upper bound for the ℓ_2-ℓ_2 induced gain of the ϕ-system.

Definition 2.3 If there exist a positive definite / pdd / positive diagonal P and a non-negative real number γ such that

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^* \begin{bmatrix} P & 0 \\ 0 & I_p \end{bmatrix} \begin{bmatrix} A & B \\ C & D \end{bmatrix} - \begin{bmatrix} P & 0 \\ 0 & \gamma^2 I_m \end{bmatrix} \leq 0 \tag{2.27}$$

then the smallest such γ is called the linear gain / dd-gain / diagonal gain of the matrix $\begin{bmatrix} A & B \\ C & D \end{bmatrix}$ or the system $\left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right)$. ◁

The following result should be obvious.

Proposition 2.10 If $\left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right)$ is linearly stable / dd-stable / diagonally stable, then it has a (finite) linear gain / dd-gain / diagonal gain. $\triangleleft$

As for the stability, each of three kinds of gain may correspond to an upper bound of the ℓ_2-ℓ_2 induced gain of a kind of system. Indeed the linear gain of $\begin{bmatrix} A & B \\ C & D \end{bmatrix}$ is exactly the ℓ_2-ℓ_2 induced gain of the linear system $\left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right)_I$. For the dd-gain, we have

Proposition 2.11 $\forall \phi \in \mathcal{OL}(\mathbb{R})$, the ℓ_2-ℓ_2 induced gain of the ϕ-system $\left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right)_\phi$ is upper bounded by its dd-gain. More precisely, if γ is the dd-gain of $\begin{bmatrix} A & B \\ C & D \end{bmatrix}$, then $x_{[0]} = 0$ implies that

$$\forall K \in \mathbb{Z}_+, \quad \|y\|_K^2 := \sum_{k=0}^{K} y_{[k]}^* y_{[k]} \leq \gamma^2 \sum_{k=0}^{K} u_{[k]}^* u_{[k]} = \gamma^2 \|u\|_K^2$$

and hence

$$\|y\|^2 := \lim_{K\to\infty} \|y\|_K^2 \leq \gamma^2 \|u\|^2$$

if $u \in \ell_2^m$. $\triangleleft$

Proof. Let $V(\zeta) := \zeta^* P \zeta$. Pre- and post-multiplying (2.27) by $\begin{bmatrix} \phi(x_{[k]}) \\ u_{[k]} \end{bmatrix}^*$ and respectively $\begin{bmatrix} \phi(x_{[k]}) \\ u_{[k]} \end{bmatrix}$, we have

$$V(x_{[k+1]}) + y_{[k]}^* y_{[k]} - \phi(x_{[k]})^* P \phi(x_{[k]}) - \gamma^2 u_{[k]}^* u_{[k]} \leq 0$$

or equivalently

$$V(x_{[k+1]}) - V(x_{[k]}) + x_{[k]}^* P x_{[k]} - \phi(x_{[k]})^* P \phi(x_{[k]}) + y_{[k]}^* y_{[k]} - \gamma^2 u_{[k]}^* u_{[k]} \leq 0$$

Then we may apply Lemma 2.2 and take summation from $k = 0$ to any $K \in \mathbb{Z}_+$ to get

$$\begin{aligned} \|y\|_K^2 &\leq \gamma^2 \|u\|_K^2 + V(x_{[0]}) - V(x_{[K+1]}) \\ &\leq \gamma^2 \|u\|_K^2 + V(x_{[0]}) \\ &= \gamma^2 \|u\|_K^2 \end{aligned}$$

since $V(x_{[K+1]}) \geq 0$ and $x_{[0]} = 0$. At last if $u \in \ell_2^m$,

$$\|y\|_K^2 \leq \gamma^2 \|u\|_K^2 \leq \gamma^2 \|u\|^2$$

and hence we simply let $K \to \infty$ to obtain the result. ■

Remark 2.5 Let G be a ϕ-system, $\phi \in \mathcal{OL}(\mathbb{R})$ and $\alpha \in \mathbb{R}$, $\alpha \neq \pm 1$. Since G is nonlinear, αG is not necessarily $= G\alpha$, but they have the same dd-gain. ⊗

For the diagonal gain, the next proposition can be proved similarly.

Proposition 2.12 The ℓ_2-ℓ_2 induced gain of the system (2.19) is upper bounded by the diagonal gain of $\begin{bmatrix} A & B \\ C & D \end{bmatrix}$. ◁

In fact, the system (2.19) belongs to a well-known class of systems called Polytopic Linear Differential/Difference Inclusions (LDIs) (cf. [22, Ch. 4]). Let

$$\mathcal{M} := \left\{ \begin{bmatrix} A\Lambda & B \\ C\Lambda & D \end{bmatrix} \;\middle|\; \Lambda \text{ is a diagonal matrix with } \lambda_{(i)} = \pm 1 \right\} \tag{2.28}$$

and $\mathrm{co}(\mathcal{M})$ denote the convex hull of $\mathcal{M}$. Then (2.19) is equivalent to

$$\begin{bmatrix} x_{[k+1]} \\ y_{[k]} \end{bmatrix} = M_{[k]} \begin{bmatrix} x_{[k]} \\ u_{[k]} \end{bmatrix}; \quad M_{[k]} \in \mathrm{co}(\mathcal{M}) \tag{2.29}$$

Note that if $\mathcal{M}$ is a finite set as in (2.28) then $\mathrm{co}(\mathcal{M})$ is a polytope with its vertices $\in \mathcal{M}$, and it is a standard fact that to prove the stability or ℓ_2-ℓ_2 induced gain of (2.29) it suffices to have a simultaneous Lyapunov or storage function for all the vertices:

Proposition 2.13 If there exists a positive definite P such that

$$\forall M \in \mathcal{M}, \quad M^* \begin{bmatrix} P & 0 \\ 0 & I_p \end{bmatrix} M - \begin{bmatrix} P & 0 \\ 0 & \gamma^2 I_m \end{bmatrix} \leq 0 \tag{2.30}$$

then the ℓ_2-ℓ_2 induced gain of the system (2.29) $\leq \gamma$ ◁

Proof. It suffices to show that (2.30) is equivalent to

$$\forall M \in \mathrm{co}(\mathcal{M}), \quad M^* \begin{bmatrix} P & 0 \\ 0 & I_p \end{bmatrix} M - \begin{bmatrix} P & 0 \\ 0 & \gamma^2 I_m \end{bmatrix} \leq 0 \tag{2.31}$$

then the rest of the proof is obvious. It is easy since using Schur complement we may write the LMI in (2.30) as

$$\begin{bmatrix} \begin{bmatrix} P & 0 \\ 0 & \gamma^2 I_m \end{bmatrix} & M^* \\ M & \begin{bmatrix} P & 0 \\ 0 & I_p \end{bmatrix}^{-1} \end{bmatrix} \geq 0$$

which is linear in M. ■

Proposition 2.13 belongs to the so-called quadratic approach to the LDIs (cf. [22, Ch. 4–6] and the references therein) which provides only sufficient conditions. For necessary and sufficient conditions (see for example [54, 55, 56]) a general convex Lyapunov/storage function instead of a quadratic one may be needed.

Let γ_c denote the smallest γ that can be proved by Proposition 2.13. Certainly $\gamma_c \leq$ the diagonal gain of $\begin{bmatrix} A & B \\ C & D \end{bmatrix}$ since a positive diagonal P satisfying (2.27) also satisfies (2.30) (with $\mathcal{M}$ as defined in (2.28)). So in general as an upper bound of the ℓ_2-ℓ_2 induced gain of the system (2.19), the diagonal gain is not as tight as γ_c. However, computing the diagonal gain is simple, while the number of LMIs in (2.30), i.e. the number of elements in $\mathcal{M}$, will grow as 2^n and therefore computing γ_c for large n is impractical. (Discussion of this issue can be found in [42, 68].) Besides, from the numerical examples presented below the diagonal gain is only slightly looser than γ_c.

A dd-stable ϕ-system may not be diagonally stable. Therefore, it may have a (finite) dd-gain but not a diagonal gain to serve as an upper bound of its ℓ_2-ℓ_2 induced gain. Even when the ϕ-system is diagonally stable, the improvement of the pdd approach over the diagonal approach is still significant. In order to appreciate it, we are now going to give some numerical examples. The `drmodel` function in the Control toolbox of MATLAB is used to generate 22 discrete-time diagonally stable random systems $\left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right)$ with $m = 1$, $n = 4$, $p = 2$. (The random systems generated by `drmodel` have $\|A\| \leq 1$ and therefore are almost always diagonally stable.) The same set of random systems will be used repeatedly to demonstrate other results such as model reduction in Chapter 4, so it is more convenient to number them as Ex. 1–22 just to let one know which particular system we shall be referring to. Let γ_l, γ_{dd}, γ_d denote the linear gain, the dd-gain and the diagonal gain of $\begin{bmatrix} A & B \\ C & D \end{bmatrix}$ respectively, and γ_c be as above. They are computed using the LMI Control toolbox[31, 32] of MATLAB. The results are shown in Table 2.1 on page 34. Theoretically $\gamma_d \geq \gamma_{dd} \geq \gamma_l$ and $\gamma_d \geq \gamma_c \geq \gamma_l$ which are now verified by our numerical results. Table 2.1 also shows that in all 22 cases $\gamma_c \geq \gamma_{dd}$, which is intuitively reasonable since $\left\{ \phi\text{-systems} \;\middle|\; \phi \in \mathcal{OL}(\mathbb{R}) \right\} \subset \left\{ \underline{\phi}\text{-systems} \;\middle|\; \underline{\phi}_{(i)} \in \mathcal{OL}(\mathbb{R}) \right\} \subset \left\{ \text{ systems (2.19) } \right\}$ and the dd-gain, in comparison with γ_c, further takes into account that all $\underline{\phi}_{(i)}$ are identical.

In summary, if all 22 examples are regarded as ϕ-systems, $\phi \in \mathcal{OL}(\mathbb{R})$, then using the diagonal gain to upper-bound the ℓ_2-ℓ_2 induced gain, which has not taken into account that all $\underline{\phi}_{(i)}$ are identical, is typically looser than using the dd-gain. Only in 4 cases out of 22 they may be regarded as identical[2], and in 6 cases the diagonal gain is more then 40 percent looser than the dd-gain. Moreover, if ϕ is locally the identity at the origin ($\phi'(0) = 1$), then the linear gain is a lower bound of the actual ℓ_2-ℓ_2

[2] Two numbers α, β are regarded as numerically the same if $\frac{|\alpha-\beta|}{|\alpha+\beta|/2} \leq 10^{-5}$.

induced gain of the ϕ-system, and we see that in 16 cases out of 22 the upper bound (the dd-gain) and the lower bound (the linear gain) may be regarded as identical, and hence the exact ℓ_2-ℓ_2 induced gain of the ϕ-system is determined (which is furthermore shown to be independent of ϕ provided that $\phi'(0) = 1$ and $\phi \in \mathcal{OL}(\mathbb{R})$). In fact in 21 cases out of 22 the ratio $\gamma_l/\gamma_{dd} > 0.99$. So the dd-gain is typically a tight upper bound of the ℓ_2-ℓ_2 induced gain for these 22 ϕ-systems.[3]

Comparing the diagonal gain with γ_c, we see that there is a rather clear indication that they are typically not the same but very close: only in 6 cases out of 22 we may regard them as identical but the maximum ratio γ_d/γ_c over 22 examples is only 1.0167. So we may simply use the diagonal gain to upper-bound the ℓ_2-ℓ_2 induced gain of the $\underline{\phi}$-system. (In fact, in the $\phi : \ell_{2e} \to \ell_{2e}$ setting $\left\{ \underline{\phi}\text{-systems} \;\middle|\; \underline{\phi}_{(i)} \in \mathcal{OL}(\ell_{2e}) \right\} \not\subset \left\{ \text{systems (2.19)} \right\}$ and therefore γ_c is *not* necessarily an upper bound of the ℓ_2-ℓ_2 induced gain of the $\underline{\phi}$-system, and in that case the diagonal gain must be used.)

2.4 Limitations of the pdd approach

Despite its success in the analysis of the ϕ-system, the pdd approach shows limitations in the controller synthesis and model reduction as will be seen in the coming chapters. It is because the pdd matrices play a more restricted role, namely (2.4), in the ϕ-system $\left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right)_\phi$ than the elements of $\mathcal{C}(\triangle_{N,M})$ in the LFT uncertain system $\mathcal{F}_u(\begin{bmatrix} A & B \\ C & D \end{bmatrix}, \Delta)$. In this section we point out the two most important differences between them.

2.4.1 pdd matrices do not provide system realizations

The first one, roughly speaking, is the lack of system realizations. Let P be pdd and T be invertible such that $T^*T = P$. Then if P is not the identity matrix, T does not commute with ϕ for most $\phi \in \mathcal{OL}(\mathbb{R})$. This is in contrast to the fact that $T \in \mathcal{C}(\triangle_{N,M})$ commutes with all $\Delta \in \triangle_{N,M}$. It is known that, if T is invertible and commutes with Δ, then the LFT uncertain system $\mathcal{F}_u(\begin{bmatrix} A & B \\ C & D \end{bmatrix}, \Delta)$ is equivalent to $\mathcal{F}_u(\begin{bmatrix} TAT^{-1} & TB \\ CT^{-1} & D \end{bmatrix}, \Delta)$ which arises from the linear state-space transformation: $x \mapsto$

[3] In theory, however, the difference between γ_{dd} and γ_l for a general ϕ-system can be arbitrarily large since a linearly stable system need not be dd-stable. A dd-stable ϕ-system that has unity linear gain but arbitrarily large dd-gain is given explicitly in Section 5.1.3, but for that ϕ-system the dd-gain can still be shown to be tight w.r.t. the whole class of ϕ although not to the particular ϕ = the identity. See Section 5.1.3 for details.

Tx. An analog holds for the ϕ-system, but the set of invertible matrices T commuting with the nonlinearity ϕ can be shown to be very small for a general $\phi \in \mathcal{OL}(\mathbb{R})$. As a result we do not have sufficient realizations for the synthesis problem.

Consider the $\underline{\phi}$-system $\left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right)_{\underline{\phi}}$ where $\underline{\phi} \in \mathcal{S}^n$. Then the state-space transformation: $x \mapsto Tx$ leads to $\left(\begin{array}{c|c} TAT^{-1} & TB \\ \hline CT^{-1} & D \end{array}\right)_{T \circ \underline{\phi} \circ T^{-1}}$, and $T \circ \underline{\phi} \circ T^{-1} \in \mathcal{S}^n$ for all $\underline{\phi} \in \mathcal{S}^n$ if and only if T is the product of a permutation matrix and a diagonal matrix. Sufficiency is obvious and necessity can be easily proved as follows: Arbitrarily take two entries $t_{(ij)}$, $t_{(ik)}$ of T such that $|t_{(ij)}| \geq |t_{(ik)}|$. Choose $\underline{\phi}_{(l)} \in \mathcal{S}$ such that $\forall s \in \mathbb{R}$,

$$\underline{\phi}_{(l)}(s) := \begin{cases} s & \text{if} \quad l = j \\ 0 & \text{if} \quad l \neq j \end{cases} \tag{2.32}$$

Since $T \circ \underline{\phi} \circ T^{-1} \in \mathcal{S}^n$, $|(T\underline{\phi}(\zeta))_{(i)}| = |(T\underline{\phi}(T^{-1}T\zeta))_{(i)}| \leq |(T\zeta)_{(i)}|$ for all $\zeta \in \mathbb{R}^n$. Let

$$\zeta_{(l)} = \begin{cases} t_{(ij)} & \text{if} \quad l = j \\ -t_{(ik)} & \text{if} \quad l = k \\ 0 & \text{if} \quad l \neq j, k \end{cases} \tag{2.33}$$

Then

$$t_{(ij)}^2 = |(T\phi(\zeta))_{(i)}| \leq |(T\zeta)_{(i)}| = |t_{(ij)}^2 - t_{(ik)}^2| = t_{(ij)}^2 - t_{(ik)}^2$$

Hence $t_{(ik)} = 0$. So every row of T has at most 1 nonzero entry, but the invertibility of T implies there must be 1 nonzero entry in each row and each column. Therefore T is a combination of permutation and diagonal transformation.

Of course the same argument applies to $\underline{\phi} \in \mathcal{OL}^n(\mathbb{R})$, but not to ϕI_n, $\phi \in \mathcal{OL}(\mathbb{R})$ since the $\underline{\phi}_{(l)}$ chosen in (2.32) are not all identical. Indeed, in the proof of necessity we only used the following property of $\mathcal{S}^n$:

$$\forall \zeta \in \mathbb{R}^n, \quad \forall i, \quad |(T \circ \underline{\phi} \circ T^{-1}(\zeta))_{(i)}| \leq |\zeta_{(i)}| \tag{2.34}$$

and there *does* exist T which is *not* a product of a permutation matrix and a diagonal matrix such that (2.34) holds for all $\underline{\phi} = \phi I_n \in \mathcal{OL}^n(\mathbb{R})$, and this will become an important tool for the synthesis problem. So to prove a similar statement for ϕI_n, $\phi \in \mathcal{OL}(\mathbb{R})$ we have to prove it in a different way. In fact we shall prove a stronger result, in the sense that the phrase "for all ϕ" in the statement may be dropped.

Proposition 2.14 Let $\phi \in \mathcal{OL}(\mathbb{R})$ be such that ϕ is not linear and $\phi'(0)$ exists. Then $T \circ \phi \circ T^{-1} \in \mathcal{OL}^n(\mathbb{R})$ if and only if T is the product of a permutation matrix and a diagonal matrix. ◁

To prove this, we first revise the following basic property of $\phi \in \mathcal{OL}(\mathbb{R})$:

$$\forall s, t \in \mathbb{R}, \quad |\phi(s) + \phi(t)| \leq |s + t|$$

Lemma 2.15 If $\phi'(0)$ exists and

$$\forall s, t \in \mathbb{R}, \quad |\alpha\phi(s) + \beta\phi(t)| \leq |\alpha s + \beta t| \tag{2.35}$$

for some $\alpha \neq 0$, $\beta \neq 0$, $|\alpha| \neq |\beta|$, then ϕ is linear, i.e. $\phi(s) = \phi'(0) \times s$. $\triangleleft$

Proof. WLOG we may assume $|\alpha| < 1$ and $\beta = 1$. Putting $t = -\alpha s$ we get

$$\forall s \in \mathbb{R}, \quad \phi(-\alpha s) = -\alpha\phi(s) \tag{2.36}$$

Now, suppose $\exists s_0 \in \mathbb{R}$ such that $\phi(s_0) \neq \phi'(0) \times s_0$. Let

$$\epsilon = \frac{1}{2}\left|\frac{\phi(s_0)}{s_0} - \phi'(0)\right| > 0 \tag{2.37}$$

Since $\phi'(0)$ exists, $\exists \delta > 0$ such that

$$\left|\frac{\phi(s)}{s} - \phi'(0)\right| < \epsilon$$

for all $s \in (-\delta, \delta)$, $s \neq 0$; but on the other hand (2.36) and (2.37) imply that in any open interval containing $\{0\}$, there exists $\check{s}_0 = (-\alpha)^m s_0 \neq 0$ for some sufficiently large integer m, such that

$$\left|\frac{\phi(\check{s}_0)}{\check{s}_0} - \phi'(0)\right| = 2\epsilon$$

Hence contradiction results. ■

Lemma 2.16 If for some $l \geq 3$,

$$|\phi(s_1) + \phi(s_2) + \cdots + \phi(s_l)| \leq |s_1 + s_2 + \cdots + s_l| \tag{2.38}$$

for all $s_1, s_2, \ldots, s_l \in \mathbb{R}$, then ϕ is linear. $\triangleleft$

Proof. Putting $s_i = 0$ for all i shows that $\phi(0) = 0$. Hence $\phi \in \mathcal{OL}(\mathbb{R})$ by putting $s_i = 0$ for $i \geq 3$. So ϕ is continuous and odd. Putting $s_i = 0$ for $i > 3$ and $s_3 = -s_1 - s_2$ we get

$$\forall s_1, s_2 \in \mathbb{R}, \quad \phi(s_1) + \phi(s_2) = -\phi(-s_1 - s_2) = \phi(s_1 + s_2) \tag{2.39}$$

which together with ϕ continuous imply ϕ linear. (See for example [33, §2.26].) ■

Now we are ready to prove Proposition 2.14.

Proof of Proposition 2.14. Sufficiency is obvious; necessity is proved as follows: Suppose $t_{(ij)} \neq 0$ and $t_{(ik)} \neq 0$. Note that $|(T\phi(\zeta))_{(i)}| = |(T\phi(T^{-1}T\zeta))_{(i)}|$. So $T \circ \phi \circ T^{-1} \in \mathcal{OL}^n(\mathrm{I\!R}) \subset \mathcal{S}^n$ implies that

$$\begin{aligned}
&\forall \zeta_{(j)}, \zeta_{(k)} \in \mathrm{I\!R}, \\
&\quad |t_{(ij)}\phi(\zeta_{(j)}) + t_{(ik)}\phi(\zeta_{(k)})| \\
&= |(T\phi(\zeta))_{(i)}| \qquad \text{by choosing } \zeta_{(l)} = 0 \text{ for } l \neq j, k \\
&\leq |(T\zeta)_{(i)}| \\
&= |t_{(ij)}\zeta_{(j)} + t_{(ik)}\zeta_{(k)}|
\end{aligned} \tag{2.40}$$

Since ϕ is not linear, $|t_{(ij)}| = |t_{(ik)}|$ by Lemma 2.15. Now $T \circ \phi \circ T^{-1} \in \mathcal{OL}^n(\mathrm{I\!R})$ implies that

$$\begin{aligned}
&\forall \zeta_{(j)}, \zeta_{(k)}, \eta_{(j)}, \eta_{(k)} \in \mathrm{I\!R}, \\
&\quad |t_{(ij)}\phi(\zeta_{(j)}) + t_{(ik)}\phi(\zeta_{(k)}) + t_{(ij)}\phi(\eta_{(j)}) + t_{(ik)}\phi(\eta_{(k)})| \\
&= |(T\phi(\zeta))_{(i)} + (T\phi(\eta))_{(i)}| \qquad \text{by choosing } \zeta_{(l)}, \eta_{(l)} = 0 \text{ for } l \neq j, k \\
&\leq |(T\zeta)_{(i)} + (T\eta)_{(i)}| \\
&= |t_{(ij)}\zeta_{(j)} + t_{(ik)}\zeta_{(k)} + t_{(ij)}\eta_{(j)} + t_{(ik)}\eta_{(k)}|
\end{aligned} \tag{2.41}$$

As $|t_{(ij)}| = |t_{(ik)}| \neq 0$ and ϕ is odd, (2.41) is equivalent to

$$\begin{aligned}
&\forall \zeta_{(j)}, \zeta_{(k)}, \eta_{(j)}, \eta_{(k)} \in \mathrm{I\!R}, \\
&\quad |\phi(\zeta_{(j)}) + \phi(\zeta_{(k)}) + \phi(\eta_{(j)}) + \phi(\eta_{(k)})| \\
&\leq |\zeta_{(j)} + \zeta_{(k)} + \eta_{(j)} + \eta_{(k)}|
\end{aligned} \tag{2.42}$$

which contradicts Lemma 2.16. Note that in the proof of necessity we do not really need $\phi \in \mathcal{OL}(\mathrm{I\!R})$; we have only used that ϕ is continuous and odd and $\phi'(0)$ exists. ∎

With Proposition 2.14 the proof of the following corollary is easy.

Corollary 2.17 Let $\phi \in \mathcal{OL}(\mathrm{I\!R})$ be such that ϕ is not linear and $\phi'(0)$ exists. Then $T \circ \phi \circ T^{-1} = \tilde{\phi} I_n$ for some $\tilde{\phi} \in \mathcal{OL}(\mathrm{I\!R})$ (where $\tilde{\phi}$ is not linear and $\tilde{\phi}'(0)$ exists) if and only if T is the product of a permutation matrix and a diagonal matrix whose diagonal entries have the same magnitude, and $\tilde{\phi} = \phi$ if and only if T is the product of a permutation matrix and a diagonal matrix whose diagonal entries are ± 1. ◁

Proof. Sufficiency is obvious. For necessity, it remains to show that

$$\forall s \in \mathrm{I\!R}, \;\; \alpha\phi(\alpha^{-1}s) = \beta\phi(\beta^{-1}s) \tag{2.43}$$

implies $|\alpha| = |\beta|$ for $\alpha, \beta \neq 0$, which follows easily using the same argument as in the proof of Lemma 2.15. ∎

Hence the claim at the beginning of this section that T (invertible) does not commute with ϕ for most $\phi \in \mathcal{OL}(\mathrm{I\!R})$ if T^*T is not the identity matrix is proven. Albertini and

Sontag proved a related result[4, Lemma 4.2] that if ϕ is differentiable and not linear (but not necessarily $\in \mathcal{OL}(\mathbb{R})$) then T (invertible) commutes with ϕ if and only if T is the product of a permutation matrix and a diagonal matrix with diagonal entries ± 1. (Clearly we have tried to relax the condition that ϕ is differentiable to ϕ is only differentiable at 0.) With further conditions imposed on ϕ they proved the stronger result[5, 8] that two σ-systems $(A, B, C)_\phi$ and $(\tilde{A}, \tilde{B}, \tilde{C})_\phi$ have the same zero-initial-state input-output map if and *only if*

$$\tilde{A} = TAT^{-1}; \quad \tilde{B} = TB; \quad \tilde{C} = CT^{-1}$$

for some T which is the product of a permutation matrix and a diagonal matrix with diagonal entries ± 1 (provided that $(A, B, C)_\phi$ and $(\tilde{A}, \tilde{B}, \tilde{C})_\phi$ fall in some generic subclass of σ-systems). Those further conditions imposed on ϕ are 1) ϕ is infinitely differentiable; 2) $\phi'(0) \neq 0$; 3) the i-th derivative of ϕ at 0 is nonzero for infinitely many i. We think 3) is pretty strong but $\sin, \tanh \in \mathcal{OL}(\mathbb{R})$ still satisfy all these conditions[4]. Albertini and Sontag also gave an analogous result on the stronger case about *nonzero*-initial-state input-output map in [6] where the σ-systems are required to be observable[7] and ϕ be odd and satisfy an Independence Property (IP) which says that $\forall l$,

$$1, \quad \phi(\alpha_1 s + \beta_1), \quad \phi(\alpha_2 s + \beta_2), \quad \ldots, \quad \phi(\alpha_l s + \beta_l)$$

are linearly independent as functions of $s \in \mathbb{R}$, where $\alpha_i, \beta_i \in \mathbb{R}$, $\alpha_i \neq 0$ and $(\alpha_i, \beta_i) \neq (\alpha_j, \beta_j)$ if $i \neq j$. Although here no differentiability condition is imposed on ϕ, the condition IP will exclude sat since

$$2\mathrm{sat}(s/2) = \mathrm{sat}(s+1) + \mathrm{sat}(s-1)$$

and furthermore sin since sin is periodic (so $\sin(s)$ and $\sin(s + 2\pi)$ are not linearly independent). tanh will satisfy IP, however.

Remark 2.6 Following the results presented in this section, it should be clear that the idea of realizations is somewhat generalized: In most of the time we do not fix $\underline{\phi}$ (or ϕ) and look for (A, B, C, D), $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$ to realize the same $\underline{\phi}$-system (or ϕ-system). Instead we look for $(A, B, C, D, \underline{\phi})$, $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}, \tilde{\underline{\phi}})$ to realize the same system where $\underline{\phi}$, $\tilde{\underline{\phi}}$ come from the same class. So it is not the realization of a particular $\underline{\phi}$-system, but the realization of the whole class of $\underline{\phi}$-systems we are talking about. In that case T is not exactly required to commute with $\underline{\phi}$. For example, if $\underline{\phi}_{(i)}$ are all different, a permutation matrix does not commute with $\underline{\phi}$, but it is natural to think that renumbering the states is always another possible way to realize the system (although it will also be followed by some natural consequences which may be inconvenient). $\otimes$

[4] It may be of interest to note that the conditions imposed on ϕ for the continuous-time σ-systems are significantly weaker[4].

2.4.2 pdd matrices may not have pdd inverses

The second important difference between the pdd matrices in the ϕ-systems and $\mathcal{C}(\triangle_{N,M})$ in the LFT uncertain systems is that $P \in \mathcal{C}(\triangle_{N,M})$ if and only if $P^{-1} \in \mathcal{C}(\triangle_{N,M})$ for all $P > 0$ while the inverse of a pdd matrix is not necessarily pdd. The implication is severe: a matrix A dd-stable does not imply A^* dd-stable.

Example 2.2 A 2×2 counterexample is *not* possible since the inverse of a 2×2 pdd matrix is always pdd. Let

$$A = \begin{bmatrix} 0.90 & -0.66 & 0.76 \\ 0.66 & -0.14 & 0.26 \\ -0.28 & -0.32 & 1.00 \end{bmatrix}$$

Then A is dd-stable. A pdd P is given by

$$P = \begin{bmatrix} 3.6503 & -2.2605 & -1.0982 \\ -2.2605 & 2.9092 & -0.2975 \\ -1.0982 & -0.2975 & 2.8601 \end{bmatrix}$$

However, A^* is not dd-stable. $\otimes$

As a result, the duality between the systems $\left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right)$ and $\left(\begin{array}{c|c} A^* & C^* \\ \hline B^* & D^* \end{array}\right)$ breaks down in the sense that results on one no longer imply results on the other. For example, finding a L such that $A + LC$ is dd-stable is easy but finding a F such that $A + BF$ is dd-stable is much more difficult. More precisely, we have the following two lemmas.

Lemma 2.18 Let C have full row rank and $C_\perp$ be such that $\begin{bmatrix} C \\ C_\perp \end{bmatrix}$ is invertible and $CC_\perp^* = 0$. Then there exists L such that $A + LC$ is dd-stable if and only if

$$\exists P \ \text{ pdd}, \ \ C_\perp(A^*PA - P)C_\perp^* < 0 \tag{2.44}$$

or equivalently[5] (using Finsler's lemma[30][6])

$$\exists P \ \text{ pdd}, \ \ A^*PA - P - C^*C < 0 \tag{2.45}$$

and a L is given by

$$L = -AP^{-1}C^*(CP^{-1}C^*)^{-1} \tag{2.46}$$

$\triangleleft$

[5]Although mathematically equivalent, numerically (2.44) and (2.47) seem to be more reliable than (2.45) and (2.48). At least we may strengthen (2.44) and (2.47) to

$$\begin{aligned} \exists P \ \text{ pdd}, & \quad C_\perp(A^*PA - P)C_\perp^* & < -I \\ \exists P \ \text{ pdd}, & \quad B_\perp^*(AP^{-1}A^* - P^{-1})B_\perp & < -I \end{aligned}$$

[6]For other applications of Finsler's lemma in control, see for example [67, 49, 19].

Lemma 2.19 Let B have full column rank and $B_\perp$ be such that $\begin{bmatrix} B & B_\perp \end{bmatrix}$ is invertible and $B^* B_\perp = 0$. Then there exists F such that $A + BF$ is dd-stable if and only if

$$\exists P \ \text{pdd}, \quad B_\perp^*(AP^{-1}A^* - P^{-1})B_\perp < 0 \tag{2.47}$$

or equivalently

$$\exists P \ \text{pdd}, \quad AP^{-1}A^* - P^{-1} - BB^* < 0 \tag{2.48}$$

and a F is given by

$$F = -(B^*PB)^{-1}B^*PA \tag{2.49}$$

$\triangleleft$

So clearly (2.44) and (2.45) are convex in P, but in (2.47) and (2.48) the pdd condition is not convex in P^{-1} while the LMI condition is not convex in P, therefore (2.47) and (2.48) are not convex in either P^{-1} or P. Note that the set of matrices having a pdd inverse is not convex, as seen from the following counterexample.

Example 2.3 We have not managed to find a 3×3 counterexample, so a 4×4 counterexample is given. Let

$$P_1 = \begin{bmatrix} 0.50 & -0.14 & -0.30 & -0.24 \\ -0.14 & 0.46 & 0.14 & -0.02 \\ -0.30 & 0.14 & 0.66 & 0.24 \\ -0.24 & -0.02 & 0.24 & 0.46 \end{bmatrix}$$

$$P_2 = \begin{bmatrix} 0.90 & -0.30 & 0.14 & 0.66 \\ -0.30 & 1.00 & -0.68 & -0.46 \\ 0.14 & -0.68 & 0.90 & 0.32 \\ 0.66 & -0.46 & 0.32 & 0.94 \end{bmatrix}$$

Then P_1^{-1}, P_2^{-1} are both pdd. However, $(\frac{P_1+P_2}{2})^{-1}$ is not pdd. $\otimes$

Although many sufficient conditions for a matrix to have a pdd inverse may be derived, none has been found to be particularly useful. The simplest among them, which in addition has a desirable physical meaning, is to force the P^{-1} in (2.47) and (2.48) to be diagonal, and then $A + BF$ is in fact diagonally stable rather than dd-stable. So computationally it is easy to find a L such that $A + LC$ is dd-stable and a F such that $A + BF$ is diagonally stable, but unfortunately we shall want the exact opposite in the synthesis problem: to find a L such that $A + LC$ is diagonally stable and a F such that $A + BF$ is dd-stable. See Chapter 3.

Table 2.1: Comparison of γ_d, γ_c, γ_{dd} and γ_l

Ex.	γ_d	γ_c	γ_{dd}	γ_l
1	4.2612	4.2170	2.9819	2.9819
2	2.6979	2.6690	2.2819	2.2819
3	1.2141	1.2141	1.2141	1.2141
4	2.4801	2.4801	2.4801	2.4801
5	1.7715	1.7698	1.7516	1.7516
6	1.7257	1.7231	1.7211	1.7211
7	4.7369	4.7367	4.7361	4.7361
8	1.3429	1.3429	0.8037	0.7127
9	2.8785	2.8749	2.8209	2.8201
10	2.0328	2.0315	1.9478	1.9478
11	1.4438	1.4436	1.3225	1.3202
12	1.8311	1.8300	1.8290	1.8290
13	1.0145	1.0145	1.0145	1.0145
14	1.8812	1.8738	0.8222	0.8161
15	2.6264	2.6203	2.4879	2.4735
16	4.7092	4.6905	3.3271	3.3271
17	0.7911	0.7911	0.7888	0.7888
18	1.7156	1.7061	1.1278	1.1278
19	4.6011	4.5967	4.5416	4.5408
20	1.8608	1.8608	1.8608	1.8608
21	2.6772	2.6332	2.3915	2.3915
22	9.1276	9.1144	6.2394	6.2393

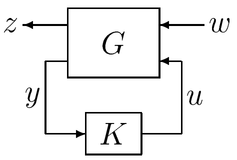
Chapter 3

Synthesis

In this chapter we consider the synthesis problem, i.e. given a ϕ-plant G

$$G \quad : \quad \begin{bmatrix} x^+ \\ z \\ y \end{bmatrix} = \begin{bmatrix} A & B_1 & B_2 \\ C_1 & D_{11} & D_{12} \\ C_2 & D_{21} & D_{22} \end{bmatrix} \begin{bmatrix} \phi(x) \\ w \\ u \end{bmatrix}$$

we design a ϕ-controller K such that the closed-loop



(3.1)

is dd-stable or has a small dd-gain. w, u, z, y are the external input, control input, regulated output and measured output of the plant respectively.

One may wonder why the controller should also take the form of a ϕ-system. On one extreme a simple linear controller may be desirable, while on the other extreme a general nonlinear ϕ-dependent controller may be needed to achieve the best closed-loop performance, and restricting the controller to a ϕ-system may result in some loss of generality. There are two reasons of using a ϕ-controller. The first one is that using a ϕ-controller guarantees that the closed-loop (3.1) is still a ϕ-system (cf. Remark 1.1), and hence the synthesis results will be better integrated with the analysis results presented in Chapter 2. The second one is that from the results presented below we shall see that although a ϕ-controller is ϕ-dependent, the synthesis of a ϕ-controller depends only on the state-space data (A, B, C, D) but not on the nonlinearity ϕ. That means the results presented in this chapter is actually for the situation that ϕ is not available for the synthesis in advance but available to the controller in real time. It is less easy to picture the situation if $\phi \in \mathcal{OL}(\mathbb{R})$, but if we consider $\phi \in \mathcal{OL}(\ell_{2e})$ then the picture is clearer: the nonlinearity ϕ may vary with time so it cannot be known in advance but can be measured in real time. For this reason this kind of synthesis result is sometimes called gain-scheduling[60, 9, 10, 71, 36].

The closed-loop configuration (3.1), which we shall denote by $\mathcal{F}_l(G, K)$, is obviously closely related to the well-known (lower) Linear Fractional Transformation (LFT), but a remark should be given. As a constant matrix,

$$\begin{bmatrix} A & B_1 & B_2 \\ C_1 & D_{11} & D_{12} \\ C_2 & D_{21} & D_{22} \end{bmatrix}$$

is certainly linear in its three inputs $\phi(x)$, w, u. However, as a ϕ-system

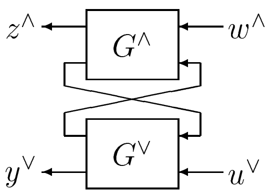
$$G := \left(\begin{array}{c|cc} A & B_1 & B_2 \\ \hline C_1 & D_{11} & D_{12} \\ C_2 & D_{21} & D_{22} \end{array}\right)$$

is *nonlinear* in its inputs w and u. Therefore, for example, $\begin{bmatrix} z \\ y \end{bmatrix} = G \begin{bmatrix} w \\ u \end{bmatrix}$ *cannot* be written as $\begin{bmatrix} z \\ y \end{bmatrix} = \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix} \begin{bmatrix} w \\ u \end{bmatrix}$ (although it can still be written as $\begin{bmatrix} z \\ y \end{bmatrix} = \begin{bmatrix} G_1 \\ G_2 \end{bmatrix} \begin{bmatrix} w \\ u \end{bmatrix}$).

The (Redheffer) star product of two systems

$$G^\wedge := \left(\begin{array}{c|cc} A^\wedge & B_1^\wedge & B_2^\wedge \\ \hline C_1^\wedge & D_{11}^\wedge & D_{12}^\wedge \\ C_2^\wedge & D_{21}^\wedge & D_{22}^\wedge \end{array}\right), \quad G^\vee := \left(\begin{array}{c|cc} A^\vee & B_1^\vee & B_2^\vee \\ \hline C_1^\vee & D_{11}^\vee & D_{12}^\vee \\ C_2^\vee & D_{21}^\vee & D_{22}^\vee \end{array}\right)$$

is defined as



$$= \left(\begin{array}{c|cc} A & B_1 & B_2 \\ \hline C_1 & D_{11} & D_{12} \\ C_2 & D_{21} & D_{22} \end{array}\right) \tag{3.2}$$

where

$$\begin{aligned}
A &:= \begin{bmatrix} A^\wedge + B_2^\wedge (R^\sharp)^{-1} D_{11}^\vee C_2^\wedge & B_2^\wedge (R^\sharp)^{-1} C_1^\vee \\ B_1^\vee (R^\flat)^{-1} C_2^\wedge & A^\vee + B^\vee (R^\flat)^{-1} D_{22}^\wedge C_1^\vee \end{bmatrix} \\
\begin{bmatrix} B_1 & B_2 \end{bmatrix} &:= \begin{bmatrix} B_1^\wedge + B_2^\wedge (R^\sharp)^{-1} D_{11}^\vee D_{21}^\wedge & B_2^\wedge (R^\sharp)^{-1} D_{12}^\vee \\ B_1^\vee (R^\flat)^{-1} D_{21}^\wedge & B_2^\vee + B_1^\vee (R^\flat)^{-1} D_{22}^\wedge D_{12}^\vee \end{bmatrix} \\
\begin{bmatrix} C_1 \\ C_2 \end{bmatrix} &:= \begin{bmatrix} C_1^\wedge + D_{12}^\wedge D_{11}^\vee (R^\flat)^{-1} C_2^\wedge & D_{12}^\wedge (R^\sharp)^{-1} C_1^\vee \\ D_{21}^\vee (R^\flat)^{-1} C_2^\wedge & C_2^\vee + D_{21}^\vee (R^\flat)^{-1} D_{22}^\wedge C_1^\vee \end{bmatrix}
\end{aligned}$$

$$
\begin{aligned}
\begin{bmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{bmatrix} &:= \begin{bmatrix} D_{11}^{\wedge} + D_{12}^{\wedge} D_{11}^{\vee} (R^{\flat})^{-1} D_{21}^{\wedge} & D_{12}^{\wedge} (R^{\sharp})^{-1} D_{12}^{\vee} \\ D_{21}^{\vee} (R^{\flat})^{-1} D_{21}^{\wedge} & D_{22}^{\vee} + D_{21}^{\vee} (R^{\flat})^{-1} D_{22}^{\wedge} D_{12}^{\vee} \end{bmatrix} \\
R^{\flat} &:= I - D_{22}^{\wedge} D_{11}^{\vee} \\
R^{\sharp} &:= I - D_{11}^{\vee} D_{22}^{\wedge}
\end{aligned}
$$

and will be denoted by $G^{\wedge} \star G^{\vee}$.

3.1 Parametrization of all dd-stabilizing output-feedback ϕ-controllers

In this section we parametrize *all* ϕ-controllers K that dd-stabilizes G. This problem is simply called the Output Feedback problem in this section and abbreviated to OF. We follow a standard approach in the linear system theory (See for example [77, Ch. 12]) and solve the OF problem via four special problems: Full Information (FI), Full Control (FC), Disturbance Feedforward (DF) and Output Estimation (OE) in which the ϕ-plant takes respectively the following special structures

$$
\begin{aligned}
G_{\mathrm{FI}} &:= \left(\begin{array}{c|cc} A & B_1 & B_2 \\ \hline C_1 & D_{11} & D_{12} \\ \begin{bmatrix} I \\ 0 \end{bmatrix} & \begin{bmatrix} 0 \\ I \end{bmatrix} & \begin{bmatrix} 0 \\ 0 \end{bmatrix} \end{array} \right) \\
G_{\mathrm{FC}} &:= \left(\begin{array}{c|cc} A & B_1 & \begin{bmatrix} I & 0 \end{bmatrix} \\ \hline C_1 & D_{11} & \begin{bmatrix} 0 & I \end{bmatrix} \\ C_2 & D_{21} & \begin{bmatrix} 0 & 0 \end{bmatrix} \end{array} \right) \\
G_{\mathrm{DF}} &:= \left(\begin{array}{c|cc} A & B_1 & B_2 \\ \hline C_1 & D_{11} & D_{12} \\ C_2 & I & 0 \end{array} \right) \\
G_{\mathrm{OE}} &:= \left(\begin{array}{c|cc} A & B_1 & B_2 \\ \hline C_1 & D_{11} & I \\ C_2 & D_{21} & 0 \end{array} \right)
\end{aligned}
$$

These four special problems in the linear system $\mathcal{H}_\infty$-control theory were first introduced by Doyle *et al.*[28]. Lu *et al.*[51] used the same approach to parametrize all stabilizing output-feedback controllers for the LFT uncertain system. We are now going to use the same approach to parametrize all dd-stabilizing output-feedback ϕ-controllers for the ϕ-system, $\phi \in \mathcal{OL}(\mathbb{R})$. This is not obvious, however. In the linear case FI and FC are respectively equivalent to DF and OE, and FI and DF are respectively dual to FC and OE[1], but these are not totally true in the ϕ-system case due to its nonlinear nature and the breakdown of the system duality (cf. Section 2.4.2). Our parametrization will require that (A, B_2) be dd-stabilizable and (C_2, A) be *diagonally* detectable, i.e. there exists F, L such that $A + B_2F$ is dd-stable and $A + LC_2$ is diagonally stable. Before considering the four special problems, we shall first revise the idea of the observer-based controllers, and immediately why the stability requirement of $A + LC_2$ is strengthened will be hinted.

> **Remark 3.1** It should be pointed out (although obvious) that the stabilization problem in fact only depends on (A, B_2, C_2, D_{22}). Therefore, for example, K dd-stabilizes G_{DF} if and only if it dd-stabilizes G_{OE}. However, they are used to be treated independently since this sheds some insight on the closed-loop structure and the performance synthesis. So we shall follow this traditional way, although we shall not do performance synthesis using the same approach in this dissertation. $\otimes$

3.1.1 Observer-based controllers

In this section we revise the most elementary type of output-feedback controllers, the observer-based controllers. Consider the simplified ϕ-plant

$$G = \left(\begin{array}{c|cc} A & B & B \\ \hline C & D & D \\ C & D & D \end{array}\right) \tag{3.3}$$

or equivalently

$$G \ : \ \begin{bmatrix} x^+ \\ y \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} \phi(x) \\ w + u \end{bmatrix} \tag{3.4}$$

State-feedback dd-stabilization means that

$$u = F\phi(x) \tag{3.5}$$

such that $A + BF$ is dd-stable. When the state of the ϕ-plant is not available, an observer-based controller uses an observer to estimate it, and then state-feedback is based on the estimation. More precisely, an observer is a ϕ-system

$$\tilde{x}^+ = A\phi(\tilde{x}) + B(w + u) + L(C\phi(\tilde{x}) + D(w + u) - y) \tag{3.6}$$

where $\tilde{x}$ tries to simulate x using the same state-space equation with an additional forcing term which is proportional to the difference between the expected plant output based on $\tilde{x}$ and the real output y of the plant. Then the control input to the ϕ-plant is

$$u = F\phi(\tilde{x}) \tag{3.7}$$

[1]See for example [77, Ch. 12] for the exact meaning of these two statements.

The overall system equation will become

$$\begin{bmatrix} x^+ \\ \tilde{x}^+ \\ y \end{bmatrix} = \begin{bmatrix} A & BF & B \\ -LC & A+BF+LC & B \\ C & DF & D \end{bmatrix} \begin{bmatrix} \phi(x) \\ \phi(\tilde{x}) \\ w \end{bmatrix} \tag{3.8}$$

So to dd-stabilize the plant (3.4) using an observer-based controller means to find F, L such that

$$\begin{bmatrix} A & BF \\ -LC & A+BF+LC \end{bmatrix} \tag{3.9}$$

is dd-stable. By Lemma 2.9 it suffices to have $A+LC$ diagonally stable and $A+BF$ dd-stable since (3.8) may be written as

$$\begin{bmatrix} (x-\tilde{x})^+ \\ \tilde{x}^+ \\ y \end{bmatrix} = \begin{bmatrix} A+LC & 0 & 0 \\ -LC & A+BF & B \\ C & C+DF & D \end{bmatrix} \begin{bmatrix} \phi(x)-\phi(\tilde{x}) \\ \phi(\tilde{x}) \\ w \end{bmatrix} \tag{3.10}$$

in which the first "subsystem"

$$(x-\tilde{x})^+ = (A+LC)(\phi(x)-\phi(\tilde{x})) \tag{3.11}$$

represents an error injected into $\tilde{x}$ and y and takes the form (2.19) since $\phi \in \mathcal{OL}(\mathbb{R})$ and hence $|\phi(x_{(i)})-\phi(\tilde{x}_{(i)})| \leq |x_{(i)}-\tilde{x}_{(i)}|$. Therefore it is diagonally stabilized according to Proposition 2.7.

Finding a L such that $A+LC$ is diagonally stable is easy but finding a F such that $A+BF$ is dd-stable is computationally difficult (See the discussion in Section 2.4.2). So computationally we may need to solve instead the sufficient condition that $A+BF$ is diagonally stable. This is rather restrictive, but once this can be done, the parametrization that is to be given below is in terms of a *dd-stable* parameter and *all dd-stabilizing* ϕ-controllers *must* be in that form.

3.1.2 FI and DF problems

Among the four special problems we first consider the FI. The plant is

$$G_{\mathrm{FI}} := \left(\begin{array}{c|cc} A & B_1 & B_2 \\ \hline C_1 & D_{11} & D_{12} \\ \begin{bmatrix} I \\ 0 \end{bmatrix} & \begin{bmatrix} 0 \\ I \end{bmatrix} & \begin{bmatrix} 0 \\ 0 \end{bmatrix} \end{array}\right) \tag{3.12}$$

The following proposition is easy.

Proposition 3.1 If there exists a matrix F such that $A+B_2F$ is dd-stable, then $K_{\mathrm{FI}} := \begin{bmatrix} F & Q \end{bmatrix}$ dd-stabilizes G_{FI} for any dd-stable ϕ-system Q. ◁

Next we consider the DF problem where

$$G_{\mathrm{DF}} := \left(\begin{array}{c|cc} A & B_1 & B_2 \\ \hline C_1 & D_{11} & D_{12} \\ C_2 & I & 0 \end{array}\right) \tag{3.13}$$

Let

$$P_{\mathrm{DF}} := \left(\begin{array}{c|cc} A - B_1C_2 & B_1 & B_2 \\ \hline 0 & 0 & I \\ \begin{bmatrix} I \\ -C_2 \end{bmatrix} & \begin{bmatrix} 0 \\ I \end{bmatrix} & \begin{bmatrix} 0 \\ 0 \end{bmatrix} \end{array}\right) \tag{3.14}$$

Then according to the star product formula (3.2),

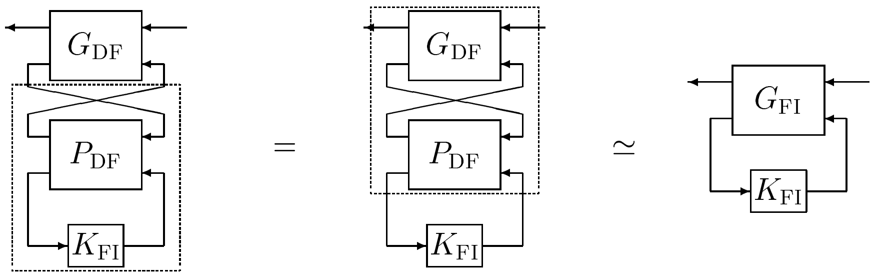
$$G_{\mathrm{DF}} \star P_{\mathrm{DF}} = \left(\begin{array}{cc|cc} A & 0 & B_1 & B_2 \\ B_1C_2 & A - B_1C_2 & B_1 & B_2 \\ \hline C_1 & 0 & D_{11} & D_{12} \\ \begin{bmatrix} 0 \\ C_2 \end{bmatrix} & \begin{bmatrix} I \\ -C_2 \end{bmatrix} & \begin{bmatrix} 0 \\ I \end{bmatrix} & \begin{bmatrix} 0 \\ 0 \end{bmatrix} \end{array}\right) \tag{3.15}$$

which may be written as

$$\begin{bmatrix} (x-\xi)^+ \\ \xi^+ \\ z \\ y \end{bmatrix} = \begin{bmatrix} A - B_1C_2 & 0 & 0 & 0 \\ B_1C_2 & A & B_1 & B_2 \\ C_1 & C_1 & D_{11} & D_{12} \\ \begin{bmatrix} 0 \\ C_2 \end{bmatrix} & \begin{bmatrix} I \\ 0 \end{bmatrix} & \begin{bmatrix} 0 \\ I \end{bmatrix} & \begin{bmatrix} 0 \\ 0 \end{bmatrix} \end{bmatrix} \begin{bmatrix} \phi(x) - \phi(\xi) \\ \phi(\xi) \\ w \\ u \end{bmatrix} \tag{3.16}$$

where x, ξ are the states of G_{DF} and P_{DF} respectively. Hence we get

Proposition 3.2 Suppose that $A - B_1C_2$ is *diagonally* stable. If K_{FI} dd-stabilizes G_{FI}, then $K_{\mathrm{DF}} := \mathcal{F}_l(P_{\mathrm{DF}}, K_{\mathrm{FI}})$ dd-stabilizes G_{DF}. Furthermore,



(3.17)

$$\mathcal{F}_l(G_{\mathrm{DF}}, \mathcal{F}_l(P_{\mathrm{DF}}, K_{\mathrm{FI}})) \quad = \quad \mathcal{F}_l(G_{\mathrm{DF}} \star P_{\mathrm{DF}}, K_{\mathrm{FI}}) \quad \simeq \quad \mathcal{F}_l(G_{\mathrm{FI}}, K_{\mathrm{FI}})$$

$\triangleleft$

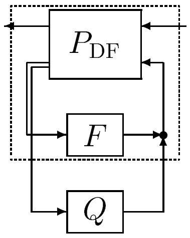
Proof. Clearly the closed-loop A-matrix still has the form

$$\begin{bmatrix} A - B_1C_2 & 0 \\ \heartsuit & \hat{A} \end{bmatrix}$$

where $\hat{A}$ is dd-stable and $A - B_1C_2$ corresponds to $\phi(x) - \phi(\xi)$. Then Lemma 2.9 applies. ■

Notice how diagonal stability is involved instead of merely dd-stability. We have already shown that $K_{\mathrm{FI}} := \begin{bmatrix} F & Q \end{bmatrix}$ is a class of dd-stabilizing ϕ-controllers for G_{FI}, with a dd-stable parameter Q. Hence $\mathcal{F}_l(P_{\mathrm{DF}}, \begin{bmatrix} F & Q \end{bmatrix}) = \mathcal{F}_l(J_{\mathrm{DF}}, Q)$ is a class of dd-stabilizing ϕ-controllers for G_{DF}, where

$$J_{\mathrm{DF}} := \left(\begin{array}{c|cc} A - B_1C_2 + B_2F & B_1 & B_2 \\ \hline F & 0 & I \\ -C_2 & I & 0 \end{array}\right) = \text{ dashed box in} \tag{3.18}$$



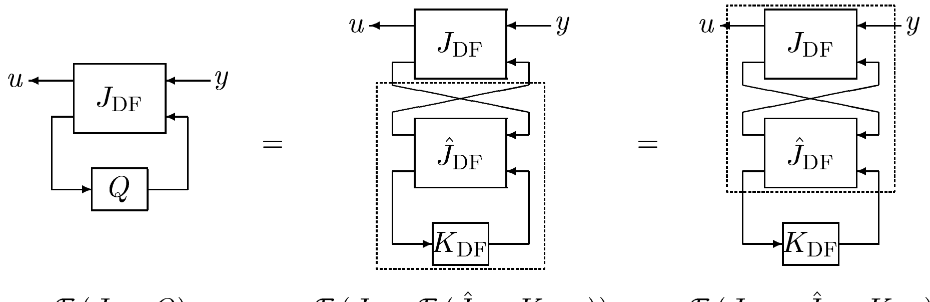
In fact, it parametrizes *all* dd-stabilizing ϕ-controllers for G_{DF}.

Proposition 3.3 If $A - B_1C_2$ is diagonally stable and there exists F such that $A + B_2F$ is dd-stable, then $\mathcal{F}_l(J_{\mathrm{DF}}, Q)$ dd-stabilizes G_{DF} where J_{DF} is as defined in (3.18) and Q is any dd-stable ϕ-system. Also if K_{DF} is a dd-stabilizing ϕ-controller for G_{DF}, then $K_{\mathrm{DF}} \simeq \mathcal{F}_l(J_{\mathrm{DF}}, Q)$ for some dd-stable Q. ◁

Proof. It remains to show the second part. Since K_{DF} dd-stabilizes G_{DF}, it also dd-stabilizes (cf. Remark 3.1)

$$\hat{J}_{\mathrm{DF}} := \left(\begin{array}{c|cc} A & B_1 & B_2 \\ \hline -F & 0 & I \\ C_2 & I & 0 \end{array}\right) \tag{3.19}$$

i.e. $Q := \mathcal{F}_l(\hat{J}_{\mathrm{DF}}, K_{\mathrm{DF}})$ is dd-stable. Then



$$\mathcal{F}_l(J_{\mathrm{DF}}, Q) = \mathcal{F}_l(J_{\mathrm{DF}}, \mathcal{F}_l(\hat{J}_{\mathrm{DF}}, K_{\mathrm{DF}},)) = \mathcal{F}_l(J_{\mathrm{DF}} \star \hat{J}_{\mathrm{DF}}, K_{\mathrm{DF}})$$

where

$$J_{\mathrm{DF}} \star \hat{J}_{\mathrm{DF}} = \left(\begin{array}{cc|cc} A - B_1C_2 + B_2F & -B_2F & B_1 & B_2 \\ -B_1C_2 & A & B_1 & B_2 \\ \hline F & -F & 0 & I \\ -C_2 & C_2 & I & 0 \end{array}\right) \tag{3.20}$$

according to the star product formula (3.2). Now it is easy to check that (input, state, output) $= (\begin{bmatrix} y \\ u \end{bmatrix}, \begin{bmatrix} \xi \\ \xi \end{bmatrix}, \begin{bmatrix} u \\ y \end{bmatrix})$ solves the state-space equation for $J_{\mathrm{DF}} \star \hat{J}_{\mathrm{DF}}$ with $\begin{bmatrix} \xi \\ \xi \end{bmatrix}_{[0]} = 0$. Hence $J_{\mathrm{DF}} \star \hat{J}_{\mathrm{DF}} \simeq \begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix}$ and

$$\mathcal{F}_l(J_{\mathrm{DF}}, Q) \simeq \mathcal{F}_l(\begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix}, K_{\mathrm{DF}}) = K_{\mathrm{DF}}$$

■

Remark 3.2 Although $K_{\mathrm{DF}} \simeq \mathcal{F}_l(J_{\mathrm{DF}} \star \hat{J}_{\mathrm{DF}}, K_{\mathrm{DF}})$, $J_{\mathrm{DF}} \star \hat{J}_{\mathrm{DF}}$ may not be dd-stable even if $A - B_1C_2$ is diagonally stable and $A + B_2F$ is dd-stable. As a result, the dd-stability of K_{DF} and $\mathcal{F}_l(J_{\mathrm{DF}} \star \hat{J}_{\mathrm{DF}}, K_{\mathrm{DF}})$ may not be the same. This does not matter since both will dd-stabilize G_{DF} anyway, but is in contrast to the linear case that $J_{\mathrm{DF}} \star \hat{J}_{\mathrm{DF}}$ is stable if both $A - B_1C_2$ and $A + B_2F$ are. According to Lemma 2.9, $J_{\mathrm{DF}} \star \hat{J}_{\mathrm{DF}}$ is dd-stable if $A - B_1C_2$ is dd-stable and $A + B_2F$ is diagonally stable. So if we want $J_{\mathrm{DF}} \star \hat{J}_{\mathrm{DF}}$ to be dd-stable, the additional condition that $A + B_2F$ is diagonally stable is required. On the other hand, $Q \simeq \mathcal{F}_l(\hat{J}_{\mathrm{DF}} \star J_{\mathrm{DF}}, Q)$, and both are dd-stable if $A - B_1C_2$ is diagonally stable and $A + B_2F$ is dd-stable. (Obviously $K_{\mathrm{DF}} := \mathcal{F}_l(J_{\mathrm{DF}}, Q)$ is a dd-stabilizing ϕ-controller for G_{DF} and hence it will dd-stabilize $\hat{J}_{\mathrm{DF}}$. Alternatively one may work out $\hat{J}_{\mathrm{DF}} \star J_{\mathrm{DF}}$ using the star product formula (3.2) and see by Lemma 2.9 that it is dd-stable.) Hence if the conditions of Proposition 3.3 are satisfied, $\mathcal{F}_l(\hat{J}_{\mathrm{DF}} \star J_{\mathrm{DF}}, Q)$ and Q can be used interchangeably as the dd-stable parameter.

⊗

3.1.3 FC and OE problems

In this section we consider the FC and OE problems. There are noticeable differences from the FI and DF problems. In particular the OE problem is neither exactly equivalent to the FC problem nor exactly dual to the DF problem as simple as in the linear case.

Consider the FC problem where

$$G_{\mathrm{FC}} := \left(\begin{array}{c|cc} A & B_1 & \begin{bmatrix} I & 0 \end{bmatrix} \\ \hline C_1 & D_{11} & \begin{bmatrix} 0 & I \end{bmatrix} \\ C_2 & D_{21} & \begin{bmatrix} 0 & 0 \end{bmatrix} \end{array}\right) \tag{3.21}$$

Again the following proposition is easy.

Proposition 3.4 If there exists a matrix L such that $A + LC_2$ is dd-stable, then $K_{\mathrm{FC}} := \begin{bmatrix} L \\ Q \end{bmatrix}$ dd-stabilizes G_{FC} for any dd-stable ϕ-system Q. $\triangleleft$

However we shall require $A + LC_2$ to be diagonally stable in order to solve the OE problem. Let

$$G_{\mathrm{OE}} := \left(\begin{array}{c|cc} A & B_1 & B_2 \\ \hline C_1 & D_{11} & I \\ C_2 & D_{21} & 0 \end{array}\right) \tag{3.22}$$

and

$$P_{\mathrm{OE}} := \left(\begin{array}{c|cc} A - B_2C_1 & 0 & \begin{bmatrix} I & -B_2 \end{bmatrix} \\ \hline C_1 & 0 & \begin{bmatrix} 0 & I \end{bmatrix} \\ C_2 & I & \begin{bmatrix} 0 & 0 \end{bmatrix} \end{array}\right) \tag{3.23}$$

Then

$$G_{\mathrm{OE}} \star P_{\mathrm{OE}} = \left(\begin{array}{cc|cc} A & B_2C_1 & B_1 & \begin{bmatrix} 0 & B_2 \end{bmatrix} \\ 0 & A - B_2C_1 & 0 & \begin{bmatrix} I & -B_2 \end{bmatrix} \\ \hline C_1 & C_1 & D_{11} & \begin{bmatrix} 0 & I \end{bmatrix} \\ C_2 & C_2 & D_{21} & \begin{bmatrix} 0 & 0 \end{bmatrix} \end{array}\right) \tag{3.24}$$

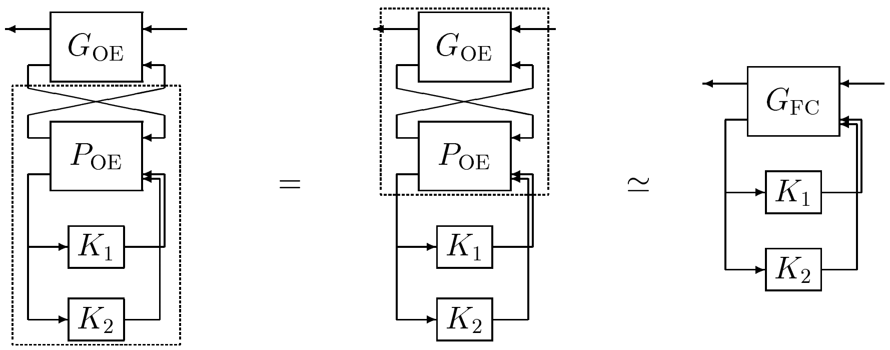
according to the star product formula (3.2), which may also be written as

$$\begin{bmatrix} x^+ \\ (x+\xi)^+ \\ z \\ y \end{bmatrix} = \begin{bmatrix} A - B_2C_1 & B_2C_1 & B_1 & \begin{bmatrix} 0 & B_2 \end{bmatrix} \\ 0 & A & B_1 & \begin{bmatrix} I & 0 \end{bmatrix} \\ 0 & C_1 & D_{11} & \begin{bmatrix} 0 & I \end{bmatrix} \\ 0 & C_2 & D_{21} & \begin{bmatrix} 0 & 0 \end{bmatrix} \end{bmatrix} \begin{bmatrix} \phi(x) \\ \phi(x) + \phi(\xi) \\ w \\ u \end{bmatrix} \tag{3.25}$$

where x, ξ are the states of G_{OE} and P_{OE} respectively.

Proposition 3.5 Suppose that $A - B_2C_1$ is dd-stable and K_{FC} is partitioned as $\begin{bmatrix} K_1 \\ K_2 \end{bmatrix}$. If $\begin{bmatrix} K_1 \\ 0 \end{bmatrix}$ diagonally stabilizes G_{FC} and K_2 is dd-stable, then $K_{\mathrm{OE}} :=$

$\mathcal{F}_l(P_{\mathrm{OE}}, K_{\mathrm{FC}})$ dd-stabilizes G_{OE}. Furthermore,



(3.26)

$$\mathcal{F}_l(G_{\mathrm{OE}}, \mathcal{F}_l(P_{\mathrm{OE}}, K_{\mathrm{FC}})) \;=\; \mathcal{F}_l(G_{\mathrm{OE}} \star P_{\mathrm{OE}}, K_{\mathrm{FC}}) \;\simeq\; \mathcal{F}_l(G_{\mathrm{FC}}, K_{\mathrm{FC}})$$

$\triangleleft$

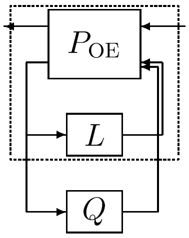
Proof. It suffices to check that the closed-loop A-matrix will take the form

$$\begin{bmatrix} A - B_2C_1 & \heartsuit & \clubsuit \\ 0 & \hat{A} & 0 \\ 0 & \spadesuit & A_{K_2} \end{bmatrix}$$

where A_{K_2} is the A-matrix of K_2 and $\hat{A}$ covering $\phi(x)+\phi(\xi)$ is diagonally stable. Then Lemma 2.9 applies. ∎

Hence if L is such that $A+LC_2$ is diagonally stable, then $\mathcal{F}_l(P_{\mathrm{OE}}, \begin{bmatrix} L \\ Q \end{bmatrix}) = \mathcal{F}_l(J_{\mathrm{OE}}, Q)$ is a class of dd-stabilizing ϕ-controllers for G_{OE}, with a dd-stable parameter Q, where

$$J_{\mathrm{OE}} := \left(\begin{array}{c|cc} A - B_2C_1 + LC_2 & L & -B_2 \\ \hline C_1 & 0 & I \\ C_2 & I & 0 \end{array}\right) = \text{dashed box in} \tag{3.27}$$



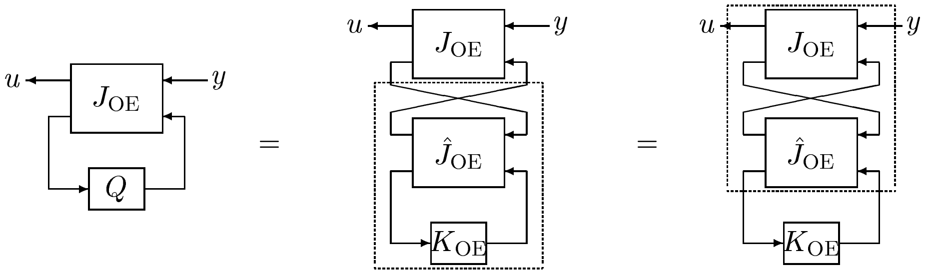
Again, it in fact parametrizes all dd-stabilizing ϕ-controllers for G_{OE}.

Proposition 3.6 If $A - B_2C_1$ is dd-stable and there exists L such that $A + LC_2$ is diagonally stable, then $\mathcal{F}_l(J_{\mathrm{OE}}, Q)$ dd-stabilizes G_{OE} where J_{OE} is as defined in (3.27) and Q is any dd-stable ϕ-system. Also if K_{OE} is a dd-stabilizing ϕ-controller for G_{OE}, then $K_{\mathrm{OE}} \simeq \mathcal{F}_l(J_{\mathrm{OE}}, Q)$ for some dd-stable Q. $\triangleleft$

Proof. The proof is similar to that of Proposition 3.3. Again we only show the second part. Since K_{OE} dd-stabilizes G_{OE}, it also dd-stabilizes

$$\hat{J}_{\mathrm{OE}} := \left(\begin{array}{c|cc} A & -L & B_2 \\ \hline C_1 & 0 & I \\ C_2 & I & 0 \end{array}\right) \tag{3.28}$$

i.e. $Q := \mathcal{F}_l(\hat{J}_{\mathrm{OE}}, K_{\mathrm{OE}})$ is dd-stable. Then



$$\mathcal{F}_l(J_{\mathrm{OE}}, Q) \quad = \quad \mathcal{F}_l(J_{\mathrm{OE}}, \mathcal{F}_l(\hat{J}_{\mathrm{OE}}, K_{\mathrm{OE}},)) \quad = \quad \mathcal{F}_l(J_{\mathrm{OE}} \star \hat{J}_{\mathrm{OE}}, K_{\mathrm{OE}})$$

where

$$J_{\mathrm{OE}} \star \hat{J}_{\mathrm{OE}} = \left(\begin{array}{cc|cc} A - B_2C_1 + LC_2 & -B_2C_1 & L & -B_2 \\ -LC_2 & A & -L & B_2 \\ \hline C_1 & C_1 & 0 & I \\ C_2 & C_2 & I & 0 \end{array} \right) \tag{3.29}$$

according to the star product formula (3.2). Now it is easy to check that $(\text{input}, \text{state}, \text{output}) = (\begin{bmatrix} y \\ u \end{bmatrix}, \begin{bmatrix} \xi \\ -\xi \end{bmatrix}, \begin{bmatrix} u \\ y \end{bmatrix})$ solves the state-space equation for $J_{\mathrm{OE}} \star \hat{J}_{\mathrm{OE}}$ with $\begin{bmatrix} \xi \\ -\xi \end{bmatrix}_{[0]} = 0$. Hence $J_{\mathrm{OE}} \star \hat{J}_{\mathrm{OE}} \simeq \begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix}$ and

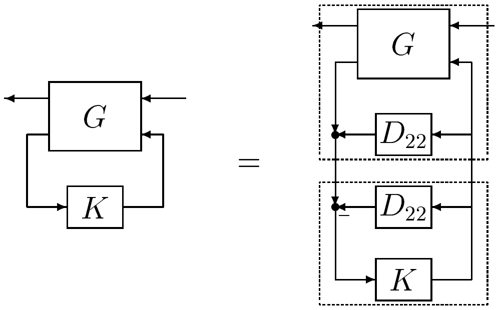
$$\mathcal{F}_l(J_{\mathrm{OE}}, Q) \simeq \mathcal{F}_l(\begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix}, K_{\mathrm{OE}}) = K_{\mathrm{OE}}$$

■

Remark 3.3 Similar to what was stated in Remark 3.2, the conditions of Proposition 3.6 only automatically guarantee the dd-stability of $\hat{J}_{\mathrm{OE}} \star J_{\mathrm{OE}}$ but not of $J_{\mathrm{OE}} \star \hat{J}_{\mathrm{OE}}$. As a result $\mathcal{F}_l(\hat{J}_{\mathrm{OE}} \star J_{\mathrm{OE}}, Q)$, which is $\simeq Q$, is still a dd-stable parameter, but the dd-stability of $\mathcal{F}_l(J_{\mathrm{OE}} \star \hat{J}_{\mathrm{OE}}, K_{\mathrm{OE}})$ and K_{OE} may not be the same: An additional condition that $A - B_2C_1$ is diagonally stable is required. However, even though $\mathcal{F}_l(J_{\mathrm{OE}} \star \hat{J}_{\mathrm{OE}}, K_{\mathrm{OE}})$ and K_{OE} may not have the same dd-stability, they will both dd-stabilize G_{OE} anyway. $\otimes$

3.1.4 OF problem and Closed-loop structure

Now we are ready to solve the OF problem. First of all, note that we may assume WLOG $D_{22} = 0$ since[2]



(3.30)

and hence if K dd-stabilizes G where $D_{22} = 0$, the lower dashed box will dd-stabilize the upper dashed box, i.e. the general plant with D_{22} not necessarily $= 0$. Then the solution for the OF problem is immediate, since the stabilization in fact only depends on (A, B_2, C_2, D_{22}), so the solution for the OE problem is also the solution for the OF problem, and the C_1 in the condition of Proposition 3.6 that $A - B_2C_1$ should be dd-stable is in fact dummy. Alternatively, the solution for the DF problem is also the solution for the OF problem, and the B_1 in the condition of Proposition 3.3 that $A - B_1C_2$ should be diagonally stable is dummy. Hence we obtain the following solution for the OF problem with $D_{22} = 0$.

Proposition 3.7 Suppose that $D_{22} = 0$. If there exist F such that $A + B_2F$ is dd-stable and L such that $A + LC_2$ is diagonally stable, then $\mathcal{F}_l(J, Q)$ dd-stabilizes G where

$$J := \left(\begin{array}{c|cc} A + LC_2 + B_2F & -L & B_2 \\ \hline F & 0 & I \\ -C_2 & I & 0 \end{array}\right) \simeq \left(\begin{array}{c|cc} A + B_2F + LC_2 & L & -B_2 \\ \hline -F & 0 & I \\ C_2 & I & 0 \end{array}\right) \tag{3.31}$$

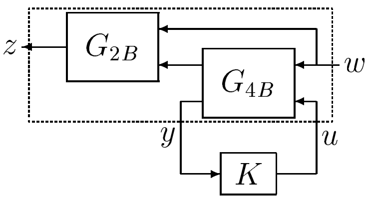
and Q is any dd-stable ϕ-system. Also if K is a dd-stabilizing ϕ-controller for G, then $K \simeq \mathcal{F}_l(J, Q)$ for some dd-stable Q. In particular, Q may be chosen as $\mathcal{F}_l(\hat{J}, K)$ where

$$\hat{J} := \left(\begin{array}{c|cc} A & -L & B_2 \\ \hline -F & 0 & I \\ C_2 & I & 0 \end{array}\right) \tag{3.32}$$

$\triangleleft$

The following interpretation may be given. Suppose that G is decomposed as

$G \;=\;$ dashed box in



(3.33)

[2]with adequate assumption on well-posedness.

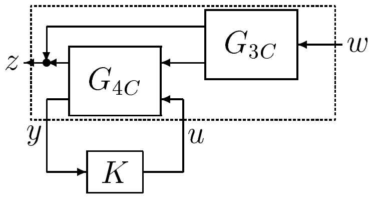
where

$$G_{4B} \;:\; \begin{bmatrix} x_{4B}^+ \\ z_{4B} \\ y \end{bmatrix} = \begin{bmatrix} A & B_1 & B_2 \\ -F & 0 & I \\ C_2 & D_{21} & 0 \end{bmatrix} \begin{bmatrix} \phi(x_{4B}) \\ w \\ u \end{bmatrix}$$

$$G_{2B} \;:\; \begin{bmatrix} x_{4B}^+ \\ z \end{bmatrix} = \begin{bmatrix} A + B_2F & B_1 & B_2 \\ C_1 + D_{12}F & D_{11} & D_{12} \end{bmatrix} \begin{bmatrix} \phi(x_{4B}) \\ w \\ z_{4B} \end{bmatrix}$$

Then G_{4B} is in the OE form and Proposition 3.6 applies. Note that G_{4B} and G_{2B} share the same state. If they are treated as independent, the $=$ in (3.33) should be replaced by $\simeq$ and G_{2B} alone will be dd-stable if $A + B_2F$ is dd-stable. Alternatively, G may be decomposed as

$G \;=\;$ dashed box in



(3.34)

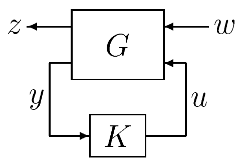
where

$$G_{4C} \;:\; \begin{bmatrix} x_{4C}^+ \\ z_{4C} \\ y \end{bmatrix} = \begin{bmatrix} A & -L & B_2 \\ C_1 & 0 & D_{12} \\ C_2 & I & 0 \end{bmatrix} \begin{bmatrix} \phi(x_{4C}) \\ y_{3C} \\ u \end{bmatrix}$$

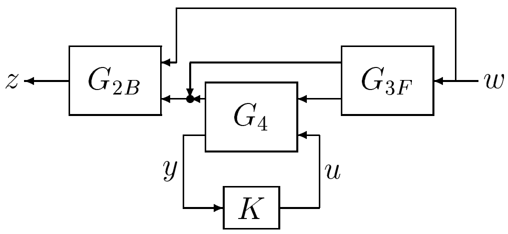
$$G_{3C} \;:\; \begin{bmatrix} (x_{3C} - x_{4C})^+ \\ z_{3C} \\ y_{3C} \end{bmatrix} = \begin{bmatrix} A + LC_2 & B_1 + LD_{21} \\ C_1 & D_{11} \\ C_2 & D_{21} \end{bmatrix} \begin{bmatrix} \phi(x_{3C}) - \phi(x_{4C}) \\ w \end{bmatrix}$$

and $z = z_{3C} + z_{4C}$. Then G_{4C} is in the DF form and Proposition 3.3 applies. Again if G_{4C} and G_{3C} are treated as independent, G_{3C} alone will be diagonally stable if $A + LC_2$ is diagonally stable.

We may further decompose G_{4B} in (3.33) using (3.34) to obtain

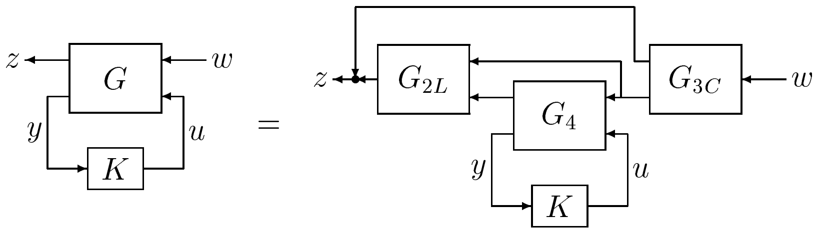


$=$



(3.35)

or decompose G_{4C} in (3.34) using (3.33) to obtain



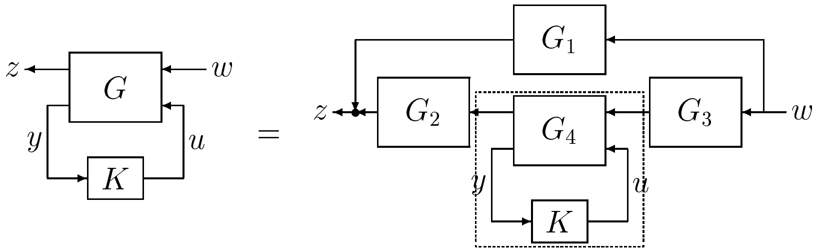
(3.36)

where

$$G_{3F} \quad : \quad \begin{bmatrix} (x_{3F}-x_4)^+ \\ z_{3F} \\ y_{3F} \end{bmatrix} = \begin{bmatrix} A+LC_2 & B_1+LD_{21} \\ -F & 0 \\ C_2 & D_{21} \end{bmatrix} \begin{bmatrix} \phi(x_{3F})-\phi(x_4) \\ w \end{bmatrix}$$

$$G_{2L} \quad : \quad \begin{bmatrix} x_4^+ \\ z \end{bmatrix} = \begin{bmatrix} A+B_2F & -L & B_2 \\ C_1+D_{12}F & 0 & D_{12} \end{bmatrix} \begin{bmatrix} \phi(x_4) \\ w \\ z_4 \end{bmatrix}$$

(x_4 and z_4 are the state and output of G_4 respectively) and G_4 is exactly $\hat{J}$ as defined in (3.32). Hence $\mathcal{F}_l(G_4, K) \simeq Q$ and is dd-stable (cf. Remark 3.2 and 3.3). Further manipulation of (3.35) or (3.36) will result in



(3.37)

where

$$G_1 \quad : \quad \begin{bmatrix} (x_3-x_2)^+ \\ (x_4-x_2)^+ \\ z_1 \end{bmatrix} = \begin{bmatrix} A & B_2F & B_1 \\ -LC_2 & A+B_2F+LC_2 & -LD_{21} \\ C_1 & D_{12}F & D_{11} \end{bmatrix} \begin{bmatrix} \phi(x_3)-\phi(x_2) \\ \phi(x_4)-\phi(x_2) \\ w \end{bmatrix}$$

$$G_2 \quad : \quad \begin{bmatrix} x_2^+ \\ z_2 \end{bmatrix} = \begin{bmatrix} A+B_2F & B_2 \\ C_1+D_{12}F & D_{12} \end{bmatrix} \begin{bmatrix} \phi(x_2) \\ z_4 \end{bmatrix}$$

$$G_3 \quad : \quad \begin{bmatrix} (x_3-x_4)^+ \\ y_3 \end{bmatrix} = \begin{bmatrix} A+LC_2 & B_1+LD_{21} \\ C_2 & D_{21} \end{bmatrix} \begin{bmatrix} \phi(x_3)-\phi(x_4) \\ w \end{bmatrix}$$

($z = z_1 + z_2$) and the dashed box $\simeq Q$. This parallels the well-known fact in the linear case that the closed-loop transfer function can be written as affine in Q: $G_1 + G_2QG_3$ where G_1, G_2, G_3, Q are all stable linear systems. However in our case it should be noted that 1) the closed-loop is not really affine in Q due to the inter-dependence of the states of G_1, G_2, G_3 and G_4; 2) G_1, G_2, G_3 and $\mathcal{F}_l(G_4, K)$, if treated as independent, are not all dd-stable ϕ-systems. Only G_2 and $\mathcal{F}_l(G_4, K)$ are dd-stable ϕ-systems. G_3 is a diagonally stable system of the form (2.19) (since $\phi \in \mathcal{OL}(\mathbb{R})$), and above all G_1 may not be dd-stable. However we may write G_1 as

$$G_1 \quad : \quad \begin{bmatrix} (x_3 - x_4)^+ \\ (x_4 - x_2)^+ \\ z_1 \end{bmatrix} = \begin{bmatrix} A + LC_2 & 0 & B_1 + LD_{21} \\ -LC_2 & A + B_2F & -LD_{21} \\ C_1 & C_1 + D_{12}F & D_{11} \end{bmatrix} \begin{bmatrix} \phi(x_3) - \phi(x_4) \\ \phi(x_4) - \phi(x_2) \\ w \end{bmatrix}$$

Hence if in addition $A + B_2F$ is diagonally stable, G_1 is also a diagonally stable system of the form (2.19). In summary, if both $A + B_2F$ and $A + LC_2$ are diagonally stable, the closed-loop can be written as affine in Q: $G_1 + G_2QG_3$ where G_1, G_3 are diagonally stable system of the form (2.19), G_2 is a diagonally stable ϕ-system, and Q is a dd-stable ϕ-system. The inter-dependence of the states is presented as uncertainties in G_1 and G_3. This is, however, of more interest for the purpose of comparison (to the linear case) rather than application. For practical computation we would prefer to "pull out" all the nonlinearities $\phi \circ \lambda$, a standard LFT technique, so that only a constant LFT remains. A demonstration of this technique is given in Section 3.2.

The closed-loop system is in fact $\mathcal{F}_l(G, K) = \mathcal{F}_l(G, \mathcal{F}_l(J, Q)) = \mathcal{F}_l(G \star J, Q)$ where

$$G \star J = \left(\begin{array}{cc|cc} A & B_2F & B_1 & B_2 \\ -LC_2 & A + B_2F + LC_2 & -LD_{21} & B_2 \\ \hline C_1 & D_{12}F & D_{11} & D_{12} \\ C_2 & -C_2 & D_{21} & 0 \end{array} \right) \tag{3.38}$$

(cf. (3.8)). Clearly

$$\left(\begin{array}{cc|c} A & B_2F & B_2 \\ -LC_2 & A + B_2F + LC_2 & B_2 \\ \hline C_2 & -C_2 & 0 \end{array} \right) \simeq 0$$

Hence the (zero-initial-state) map from the second input of $G \star J$ (i.e. the output of Q) to the second output of $G \star J$ (i.e. the input of Q) is zero provided that $w = 0$. In the linear case this is sufficient for concluding that the closed-loop transfer function is affine in Q. However, since the ϕ-system is nonlinear, the dependence of the second output of $G \star J$ on the second input of $G \star J$ is no longer zero if $w \neq 0$.

For ease of reference, we close this section with the general form of Proposition 3.7 in which D_{22} is not necessarily 0.

Proposition 3.8 If there exist F such that $A + B_2F$ is dd-stable and L such that

$A + LC_2$ is diagonally stable, then $\mathcal{F}_l(J, Q)$ dd-stabilizes G where

$$J := \left(\begin{array}{c|cc} A + B_2F + LC_2 + LD_{22}F & -L & B_2 + LD_{22} \\ \hline F & 0 & I \\ -(C_2 + D_{22}F) & I & -D_{22} \end{array}\right) \tag{3.39}$$

and Q is any dd-stable ϕ-system. Also if K is a dd-stabilizing ϕ-controller for G, then $K \simeq \mathcal{F}_l(J, Q)$ for some dd-stable Q. In particular, Q may be chosen as $\mathcal{F}_l(\hat{J}, K)$ where

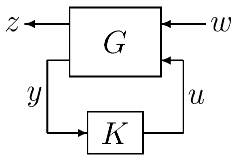
$$\hat{J} := \left(\begin{array}{c|cc} A & -L & B_2 \\ \hline -F & 0 & I \\ C_2 & I & D_{22} \end{array}\right) \tag{3.40}$$

$\triangleleft$

We may now of course use the solution for the OF problem to obtain the parametrizations of *all* dd-stabilizing ϕ-controllers for the special problems FI, FC, DF, OE which will replace the weaker Propositions 3.1, 3.4 in which only a class of ϕ-controllers are parametrized and the less general Propositions 3.3, 3.6 where B_1, C_1 are unnecessarily involved.

3.2 Closed-loop performance synthesis

Having considered the stabilization problem, in this section we consider the performance synthesis: Given a plant G, design a controller K such that the dd-gain of the closed-loop system



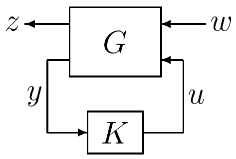
is less than γ. The approach adopted is rather different from the previous section. Instead we shall follow the line developed by Packard[60]. Although our result presented below is partially applicable to the ϕ-system, due to its computational difficulties it is basically useful to the $\underline{\phi}$-system only (i.e. if $\underline{\phi}_{(i)}$ in G are all identical, this cannot been taken into account), and may be considered as the extension of Packard's result on the LFT uncertain system with $\Delta = \delta := \mathrm{Diag}(\delta_1, \delta_2, \ldots, \delta_n) \in \mathcal{B}(\triangle_{I_n, I_0})$ to the nonlinear $\underline{\phi}$-system with $\underline{\phi} := \mathrm{Diag}(\underline{\phi}_{(1)}, \underline{\phi}_{(2)}, \ldots, \underline{\phi}_{(n)}) \in \mathcal{OL}^n(\mathbb{R})$. The extension is not too trivial, however. It is not as obvious as the extension of the result on ℓ_2-ℓ_2 induced gain, where $\underline{\phi}_{(i)} \in \mathcal{S}$ will be sufficient. The complication is due to the fact that the controller K will also have a copy of the nonlinearity $\underline{\phi}$. Packard pointed out that the set of matrices commuting with all $\begin{bmatrix} \delta & \\ & \delta \end{bmatrix}$ is $\begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix}$ where T_{11}, T_{12}, T_{21}, T_{22}

are all diagonal matrices. However, $\begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix}$ (in addition positive definite) does not satisfy

$$\begin{gathered}\forall \underline{\phi}_{(i)} \in \mathcal{OL}(\mathbb{R}), \;\; \forall \zeta_1, \zeta_2 \in \mathbb{R}^n, \\ \begin{bmatrix} \underline{\phi}(\zeta_1) \\ \underline{\phi}(\zeta_2) \end{bmatrix}^* \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} \begin{bmatrix} \underline{\phi}(\zeta_1) \\ \underline{\phi}(\zeta_2) \end{bmatrix} \leq \begin{bmatrix} \zeta_1 \\ \zeta_2 \end{bmatrix}^* \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} \begin{bmatrix} \zeta_1 \\ \zeta_2 \end{bmatrix}\end{gathered} \tag{3.41}$$

The general method to handle arbitrarily repeated nonlinearities will be explained in Section 5.1.1. At this stage it suffices to say that for the $\begin{bmatrix} \underline{\phi} & \\ & \underline{\phi} \end{bmatrix}$ case, $\begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix}$ must be in addition pdd for (3.41) to hold. This results in a slight modification in the matrix dilation part of the proof and the requirement that $\underline{\phi}_{(i)} \in \mathcal{OL}(\mathbb{R})$ rather than merely $\in \mathcal{S}$.

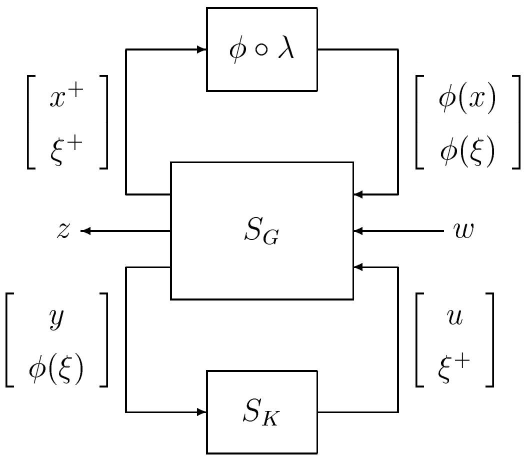
To start with, we consider the ϕ-system



where

$$\begin{aligned} G \;\; &: \;\; \begin{bmatrix} x^+ \\ z \\ y \end{bmatrix} = \begin{bmatrix} A & B_1 & B_2 \\ C_1 & D_{11} & D_{12} \\ C_2 & D_{21} & D_{22} \end{bmatrix} \begin{bmatrix} \phi(x) \\ w \\ u \end{bmatrix} \\ K \;\; &: \;\; \begin{bmatrix} \xi^+ \\ u \end{bmatrix} = \begin{bmatrix} A_K & B_K \\ C_K & D_K \end{bmatrix} \begin{bmatrix} \phi(\xi) \\ y \end{bmatrix} \end{aligned}$$

i.e. we first assume that all $\underline{\phi}_{(i)}$ are identical. Also γ is set to 1 WLOG. Following [60], we "pull out" all the nonlinearities $\phi \circ \lambda$ and write $\mathcal{F}_l(G, K)$ as (assuming that $x_{[0]} = 0$ and $\xi_{[0]} = 0$)



where S_G, S_K are *constant* matrices:

$$\begin{bmatrix} x^+ \\ \xi^+ \\ z \\ y \\ \phi(\xi) \end{bmatrix} = \begin{bmatrix} A & 0 & B_1 & B_2 & 0 \\ 0 & 0 & 0 & 0 & I \\ C_1 & 0 & D_{11} & D_{12} & 0 \\ C_2 & 0 & D_{21} & D_{22} & 0 \\ 0 & I & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \phi(x) \\ \phi(\xi) \\ w \\ u \\ \xi^+ \end{bmatrix} = S_G \begin{bmatrix} \phi(x) \\ \phi(\xi) \\ w \\ u \\ \xi^+ \end{bmatrix}$$

$$\begin{bmatrix} u \\ \xi^+ \end{bmatrix} = \begin{bmatrix} D_K & C_K \\ B_K & A_K \end{bmatrix} \begin{bmatrix} y \\ \phi(\xi) \end{bmatrix} = S_K \begin{bmatrix} y \\ \phi(\xi) \end{bmatrix}$$

So the problem becomes: Given a constant matrix S_G, find a constant matrix S_K such that the dd-gain of $\mathcal{F}_l(S_G, S_K) < 1$, or equivalently

$$\left\| \begin{bmatrix} P^{\frac{1}{2}} & 0 \\ 0 & I_p \end{bmatrix} \mathcal{F}_l(S_G, S_K) \begin{bmatrix} P^{-\frac{1}{2}} & 0 \\ 0 & I_m \end{bmatrix} \right\| < 1$$

for some pdd P. (This is essentially the dd-gain LMI with $\gamma = 1$.) Using (3.30), we may assume that $D_{22} = 0$, and hence

$$\mathcal{F}_l(S_G, S_K) = \begin{bmatrix} A & 0 & B_1 \\ 0 & 0 & 0 \\ C_1 & 0 & D_{11} \end{bmatrix} + \begin{bmatrix} B_2 & 0 \\ 0 & I \\ D_{12} & 0 \end{bmatrix} S_K \begin{bmatrix} C_2 & 0 & D_{21} \\ 0 & I & 0 \end{bmatrix}$$

The following lemma is standard. (See for example [61, Lemma 3.3] or [60, Lemma 5.2].)

Lemma 3.9 Let R, U, V be given matrices such that U has full column rank and V has full row rank. Let $U_\perp$, $V_\perp$ be such that $\begin{bmatrix} U & U_\perp \end{bmatrix}$, $\begin{bmatrix} V \\ V_\perp \end{bmatrix}$ are invertible and $U^* U_\perp = 0$, $V V_\perp^* = 0$. Let $\mathcal{Z}$ be a given set of positive definite matrices. Then

$$\left\| Z^{\frac{1}{2}} (R + UQV) Z^{-\frac{1}{2}} \right\| < 1 \tag{3.42}$$

for some $Z \in \mathcal{Z}$ and some matrix Q if and only if

$$\begin{aligned} V_\perp (R^* Z R - Z) V_\perp^* &< 0 & (3.43) \\ U_\perp^* (R Z^{-1} R^* - Z^{-1}) U_\perp &< 0 & (3.44) \end{aligned}$$

for some $Z \in \mathcal{Z}$, and a Q is given by

$$\begin{aligned} \tilde{R} &:= Z^{\frac{1}{2}} R Z^{-\frac{1}{2}} \\ \tilde{U} &:= Z^{\frac{1}{2}} U (U^* Z U)^{-\frac{1}{2}} \\ \tilde{V} &:= (V Z^{-1} V^*)^{-\frac{1}{2}} V Z^{-\frac{1}{2}} \\ T_1 &:= (U_\perp^* Z^{-1} U_\perp)^{-\frac{1}{2}} U_\perp^* R V_\perp^* (V_\perp Z V_\perp^*)^{-\frac{1}{2}} \end{aligned}$$

$$
\begin{aligned}
T_2 &:= (U^*ZU)^{-\frac{1}{2}}U^*ZRV_\perp^*(V_\perp ZV_\perp^*)^{-\frac{1}{2}} \\
T_3 &:= (U_\perp^*Z^{-1}U_\perp)^{-\frac{1}{2}}U_\perp^*RZ^{-1}V^*(VZ^{-1}V^*)^{-\frac{1}{2}} \\
Q' &:= -T_2(I-T_1^*T_1)^{-\frac{1}{2}}T_1^*(I-T_1T_1^*)^{-\frac{1}{2}}T_3 \\
Q &:= (U^*ZU)^{-\frac{1}{2}}(Q'-\tilde{U}^*\tilde{R}\tilde{V}^*)(VZ^{-1}V^*)^{-\frac{1}{2}}
\end{aligned}
$$

$\triangleleft$

In our case,

$$
R = \begin{bmatrix} A & 0 & B_1 \\ 0 & 0 & 0 \\ C_1 & 0 & D_{11} \end{bmatrix}, \quad U = \begin{bmatrix} B_2 & 0 \\ 0 & I \\ D_{12} & 0 \end{bmatrix}, \quad V = \begin{bmatrix} C_2 & 0 & D_{21} \\ 0 & I & 0 \end{bmatrix}
$$

and $\mathcal{Z}$ is the set of matrices of the form $\begin{bmatrix} P & 0 \\ 0 & I \end{bmatrix}$ where P is pdd. (A little care is needed about the dimension of I as R is not necessarily square.) Since R, U, V have special structures, $U_\perp$ will take the form $\begin{bmatrix} U_{\perp 1} \\ 0 \\ U_{\perp 2} \end{bmatrix}$ and $V_\perp$ will take the form $\begin{bmatrix} V_{\perp 1} & 0 & V_{\perp 2} \end{bmatrix}$, and (3.43), (3.44) will become (where $M = \begin{bmatrix} A & B_1 \\ C_1 & D_{11} \end{bmatrix}$)

$$
\begin{bmatrix} V_{\perp 1} & V_{\perp 2} \end{bmatrix} \left(M^* \begin{bmatrix} X & 0 \\ 0 & I_p \end{bmatrix} M - \begin{bmatrix} X & 0 \\ 0 & I_m \end{bmatrix} \right) \begin{bmatrix} V_{\perp 1}^* \\ V_{\perp 2}^* \end{bmatrix} < 0 \tag{3.45}
$$

$$
\begin{bmatrix} U_{\perp 1}^* & U_{\perp 2}^* \end{bmatrix} \left(M \begin{bmatrix} Y & 0 \\ 0 & I_m \end{bmatrix} M^* - \begin{bmatrix} Y & 0 \\ 0 & I_p \end{bmatrix} \right) \begin{bmatrix} U_{\perp 1} \\ U_{\perp 2} \end{bmatrix} < 0 \tag{3.46}
$$

such that for some matrices X_{12}, X_{22}, Y_{12}, Y_{22},

$$
\begin{bmatrix} X & X_{12} \\ X_{12}^* & X_{22} \end{bmatrix}^{-1} = \begin{bmatrix} Y & Y_{12} \\ Y_{12}^* & Y_{22} \end{bmatrix}
$$

and $\begin{bmatrix} X & X_{12} \\ X_{12}^* & X_{22} \end{bmatrix}$ is pdd. This is basically a matrix dilation problem. If $\begin{bmatrix} X & X_{12} \\ X_{12}^* & X_{22} \end{bmatrix}$ is only required to be positive definite but not necessarily pdd, Packard gave the necessary and sufficient condition that $X - Y^{-1} \geq 0$ (and trivially $X > 0$, $Y > 0$ too). However, how to dilate a matrix to remain pdd is unknown. We can only provide the following

Lemma 3.10 Given $X > 0$, $Y > 0$. If Y^{-1} is pdd and $X - Y^{-1}$ is non-negative diagonal, then there exist X_{12}, X_{22}, Y_{12}, Y_{22} such that

$$\begin{bmatrix} X & X_{12} \\ X_{12}^* & X_{22} \end{bmatrix}^{-1} = \begin{bmatrix} Y & Y_{12} \\ Y_{12}^* & Y_{22} \end{bmatrix}$$

and $\begin{bmatrix} X & X_{12} \\ X_{12}^* & X_{22} \end{bmatrix}$ is pdd. ◁

Proof. Let $X_{12} = X_{22} = X - Y^{-1}$. Since X_{12} is non-negative diagonal, clearly $\begin{bmatrix} X & X_{12} \\ X_{12}^* & X_{22} \end{bmatrix}$ is positive semidefinite diagonally dominant[3] if $X - X_{12}$ and $X_{22} - X_{12}$ are positive semidefinite diagonally dominant. To make $\begin{bmatrix} X & X_{12} \\ X_{12}^* & X_{22} \end{bmatrix}$ positive, we simply delete the columns of X_{12} and rows and columns of X_{22} corresponding to the zero diagonal entries of $X - Y^{-1}$. Then it is easy to check that $X - X_{12}X_{22}^{-1}X_{12}^* = Y^{-1}$, and using the standard matrix inversion formulae (See for example [77, §2.3]) to obtain

$$\begin{bmatrix} X & X_{12} \\ X_{12}^* & X_{22} \end{bmatrix}^{-1} = \begin{bmatrix} Y & -YX_{12}X_{22}^{-1} \\ -X_{22}^{-1}X_{12}^*Y & X_{22}^{-1} + X_{22}^{-1}X_{12}^*YX_{12}X_{22}^{-1} \end{bmatrix} \tag{3.48}$$

■

However, (3.46) is convex in Y but not in Y^{-1}. So the conditions in terms of Y^{-1}, namely Y^{-1} is pdd and $X - Y^{-1}$ is non-negative diagonal, are difficult to use. (Certainly if we could state the condition in terms of Y^{-1}, we would have simply required that $\begin{bmatrix} X & X - Y^{-1} \\ X - Y^{-1} & X - Y^{-1} \end{bmatrix}$ be pdd so that $X - Y^{-1}$ need not be diagonal.) So eventually from the computational point of view, we would further restrict both X and Y to positive diagonal matrices. This strict requirement essentially means that we do not take into account that all $\underline{\phi}_{(i)}$ in G are identical, but only that K and G are using the same copy of $\underline{\phi}$. (See the discussion at the beginning of the section.) Indeed the rank of $X - Y^{-1}$ determines the order of the controller K, and if a diagonal entry of $X - Y^{-1}$ is nonzero, the corresponding $\underline{\phi}_{(i)}$ needs to appear in the controller[4]. We summarize the result in the following proposition, with a general γ.

[3] A matrix P is said to be positive semidefinite diagonally dominant if P is symmetric and

$$\forall i, \ p_{(ii)} \geq \sum_{j \neq i} |p_{(ij)}| \tag{3.47}$$

It is so called since P is indeed positive semidefinite by Geršgorin Theorem (See for example [39, §6.1]).

[4] Conversely, if we *could* choose the rank of $X - Y^{-1}$, then we could choose the order of the controller K, or if we could choose the nonzero diagonal entries of $X - Y^{-1}$, then we could choose which $\underline{\phi}_{(i)}$ of G should appear in the controller. The latter would be useful when some $\underline{\phi}_{(i)}$ of G are known but the others are uncertain.

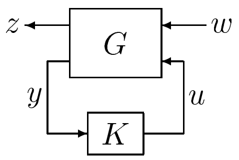
Proposition 3.11 Given

$$G \; : \; \begin{bmatrix} x^+ \\ z \\ y \end{bmatrix} = \begin{bmatrix} A & B_1 & B_2 \\ C_1 & D_{11} & D_{12} \\ C_2 & D_{21} & D_{22} \end{bmatrix} \begin{bmatrix} \underline{\phi}(x) \\ w \\ u \end{bmatrix}$$

where $\check{U} := \begin{bmatrix} B_2 \\ D_{12} \end{bmatrix}$ has full column rank and $\check{V} := \begin{bmatrix} C_2 & D_{21} \end{bmatrix}$ has full row rank. Then there exists

$$K \; : \; \begin{bmatrix} \xi^+ \\ u \end{bmatrix} = \begin{bmatrix} A_K & B_K \\ C_K & D_K \end{bmatrix} \begin{bmatrix} \tilde{\underline{\phi}}(\xi) \\ y \end{bmatrix}$$

such that the ℓ_2-ℓ_2 induced gain of



is less than γ if there exist positive diagonal X, Y such that

$$\check{V}_\perp \left(M^* \begin{bmatrix} X & 0 \\ 0 & I_p \end{bmatrix} M - \begin{bmatrix} X & 0 \\ 0 & \gamma^2 I_m \end{bmatrix} \right) \check{V}_\perp^* \; < \; 0 \tag{3.49}$$

$$\check{U}_\perp^* \left(M \begin{bmatrix} Y & 0 \\ 0 & \gamma^{-2} I_m \end{bmatrix} M^* - \begin{bmatrix} Y & 0 \\ 0 & I_p \end{bmatrix} \right) \check{U}_\perp \; < \; 0 \tag{3.50}$$

$$\begin{bmatrix} X & I \\ I & Y \end{bmatrix} \geq 0 \tag{3.51}$$

where $M = \begin{bmatrix} A & B_1 \\ C_1 & D_{11} \end{bmatrix}$, $\check{U}_\perp$ and $\check{V}_\perp$ are such that $\begin{bmatrix} \check{U} & \check{U}_\perp \end{bmatrix}$, $\begin{bmatrix} \check{V} \\ \check{V}_\perp \end{bmatrix}$ are invertible and $\check{U}^* \check{U}_\perp = 0$, $\check{V} \check{V}_\perp^* = 0$, and $\tilde{\underline{\phi}}$ consists of the $\underline{\phi}_{(i)}$ corresponding to the nonzero diagonal entries of $X - Y^{-1}$. $\triangleleft$

If all $\underline{\phi}_{(i)}$ in G are identical, then performance synthesis via Proposition 3.11 is still possible if (A, B_2) is diagonally stabilizable and (C_2, A) is diagonally detectable, i.e. there exist F, L such that $A + B_2 F$ and $A + LC_2$ are diagonally stable. Of course, in that case the actual achievable closed-loop dd-gain is typically lower than that predicted by Proposition 3.11, or in other words Proposition 3.11 will be somewhat conservative, in two ways: 1) A better controller K is possible but cannot be obtained via diagonal X and Y; 2) Even for the same controller K the closed loop dd-gain is actually lower since the bound given by Proposition 3.11 has not taken into account that all $\underline{\phi}_{(i)}$ in G are in fact identical.

So in general if the $\underline{\phi}$-system is strengthened to the ϕ-system we are not able to give a better synthesis procedure due to our inability to dilate a matrix to remain pdd. But in the special case that the ϕ-plant G has full control, a much better solution may be obtained since matrix dilation will then be unnecessary. Clearly the ultimate difficulty in the synthesis procedure comes from the requirement that Z should satisfy two constraints (3.43) and (3.44) but (3.44) is not convex in Z. So everything is easy if (3.44) can be eliminated, and it suffices to have U invertible, or equivalently $\check{U} := \begin{bmatrix} B_2 \\ D_{12} \end{bmatrix}$ invertible, which is essentially the FC case.

Proposition 3.12 Let G be a ϕ-plant with $\check{U} := \begin{bmatrix} B_2 \\ D_{12} \end{bmatrix}$ invertible. Then there exists a ϕ-controller K such that the dd-gain of $\mathcal{F}_l(G,K) < \gamma$ if there exists a pdd X satisfying (3.49). $\triangleleft$

However, our opinion is that requiring a FC plant is usually rather impractical.

Remark 3.4 Requiring a FI plant is more reasonable, but owing to the breakdown of the system duality (cf. Section 2.4.2), Proposition 3.12 does *not* have a FI counterpart in the sense that Y^{-1} instead of Y in (3.50) will then be required to be pdd, which is computationally difficult. $\otimes$

Chapter 4

Model reduction

In this chapter we present a model reduction algorithm for dd-stable ϕ-systems which borrows and extends certain ideas from the balanced truncation and Hankel-norm approximation of stable linear systems. (For general references see [57, 29, 34, 35] for the continuous-time case and [2, 38] for the discrete-time case.) As will be illustrated by numerical examples, the model reduction algorithm is *not* going to give a satisfactory reduced-order model for *all* ϕ-systems. It rather depends on the data (A, B, C, D). As ϕ-systems are nonlinear, we believe that model reduction (with a small global error) is not always possible and the objective of designing the model reduction algorithm is to identify those crucial cases that the order of the model can be reduced.

4.1 Observability and controllability function

The ideas of the observability function and the controllability function for a general nonlinear system can be found in [70]. Simply speaking, the value of the observability function at $\zeta \in \mathbb{R}^n$ is the amount of output energy generated by the state ζ (when the input is zero) while the value of the controllability function at ζ is the minimum amount of input energy needed to shift the state from 0 to ζ. It is well-known (cf. [57]) that for a continuous-time linear system they are simply $\zeta^* Q \zeta$ and $\zeta^* P^{-1} \zeta$ respectively, where Q is the observability gramian and P is the controllability gramian. For a ϕ-system, we shall only ask for an upper bound of the observability function and a lower bound of the controllability function.

Proposition 4.1 Given a ϕ-system $\left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right)_\phi$, $\phi \in \mathcal{OL}(\mathbb{R})$. Suppose that there exists a pdd Q such that

$$A^* Q A - Q + C^* C \leq 0 \tag{4.1}$$

Then $u = 0$ implies that

$$\|y\|^2 := \sum_{k=0}^{\infty} y_{[k]}^* y_{[k]} \leq x_{[0]}^* Q x_{[0]}$$

$\lhd$

Proof. Let $V(\zeta) := \zeta^* Q \zeta$. Pre- and post-multiplying (4.1) by $\phi(x_{[k]})^*$ and respectively $\phi(x_{[k]})$, we have

$$V(x_{[k+1]}) - \phi(x_{[k]})^* Q \phi(x_{[k]}) + y_{[k]}^* y_{[k]} \leq 0$$

or equivalently

$$V(x_{[k+1]}) - V(x_{[k]}) + x_{[k]}^* Q x_{[k]} - \phi(x_{[k]})^* Q \phi(x_{[k]}) + y_{[k]}^* y_{[k]} \leq 0$$

Then we may apply Lemma 2.2 and take summation from $k = 0$ to any $K \in \mathbb{Z}_+$ to get

$$\|y\|_K^2 \leq V(x_{[0]}) - V(x_{[K+1]}) \leq V(x_{[0]})$$

since $V(x_{[K+1]}) \geq 0$. At last let $K \to \infty$ to obtain the result. $\blacksquare$

Proposition 4.2 Given a ϕ-system $\left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right)_\phi$, $\phi \in \mathcal{OL}(\mathbb{R})$. Suppose that there exists P such that P^{-1} is pdd and

$$APA^* - P + BB^* \leq 0 \tag{4.2}$$

Then $x_{[0]} = 0$ implies that

$$\forall K - 1 \in \mathbb{Z}_+, \quad \|u\|_{K-1}^2 := \sum_{k=0}^{K-1} u_{[k]}^* u_{[k]} \geq x_{[K]}^* P^{-1} x_{[K]}$$

$\lhd$

Proof. Pre- and post-multiplying (4.2) by P^{-1} and then using Schur complements we see that (4.2) is equivalent to

$$\begin{bmatrix} A & B \end{bmatrix}^* P^{-1} \begin{bmatrix} A & B \end{bmatrix} - \begin{bmatrix} P^{-1} & 0 \\ 0 & I \end{bmatrix} \leq 0 \tag{4.3}$$

Let $V(\zeta) := \zeta^* P^{-1} \zeta$. Pre- and post-multiplying (4.3) by $\begin{bmatrix} \phi(x_{[k]}) \\ u_{[k]} \end{bmatrix}^*$ and respectively $\begin{bmatrix} \phi(x_{[k]}) \\ u_{[k]} \end{bmatrix}$, we have

$$V(x_{[k+1]}) - \phi(x_{[k]})^* P^{-1} \phi(x_{[k]}) - u_{[k]}^* u_{[k]} \leq 0$$

or equivalently

$$V(x_{[k+1]}) - V(x_{[k]}) + x_{[k]}^* P^{-1} x_{[k]} - \phi(x_{[k]})^* P^{-1} \phi(x_{[k]}) - u_{[k]}^* u_{[k]} \leq 0$$

Then we may apply Lemma 2.2 and take summation from $k = 0$ to any $K - 1 \in \mathbb{Z}_+$ to get

$$\|u\|_{K-1}^2 \geq V(x_{[K]}) - V(x_{[0]}) = V(x_{[K]})$$

since $x_{[0]} = 0$. ■

Remark 4.1 Note that finding a P with P^{-1} pdd to satisfy

$$APA^* - P + BB^* \leq 0$$

is convex since (4.2) can be converted into (4.3) using Schur complements, which is in contrast to the one we have encountered in the synthesis problem that finding a P with P^{-1} pdd to satisfy

$$APA^* - P - BB^* \leq 0$$

is nonconvex. ⊗

Remark 4.2 Conventionally, the controllability function is defined w.r.t. negative time, and instead of $x_{[0]} = 0$ implies that

$$\forall K - 1 \in \mathbb{Z}_+, \quad \sum_{k=0}^{K-1} u_{[k]}^* u_{[k]} \geq x_{[K]}^* P^{-1} x_{[K]}$$

we should write $x_{[-\infty]} = 0$ implies that

$$\sum_{k=-\infty}^{-1} u_{[k]}^* u_{[k]} \geq x_{[0]}^* P^{-1} x_{[0]}$$

In this way the relation between the Hankel-norm (i.e. the gain from the past input to the future output) and the observability function and the controllability function of the system is clearer. However, we are trying to avoid the use of negative time in this dissertation. Note also that since the ϕ-system is nonlinear, the future output can no longer be written as the sum of the future output caused by the past input and the future output caused by the future input. ⊗

4.2 Balanced truncation

The standard approach in the balanced truncation of a (discrete-time) linear system is to obtain a balanced realization of the linear system first and then truncate the states corresponding to the small Hankel singular values. The balanced realization is obtained by solving the Lyapunov *equations*

$$A^*QA - Q + C^*C = 0 \tag{4.4}$$

$$APA^* - P + BB^* = 0 \tag{4.5}$$

Q and P are called the observability gramian and the controllability gramian of the linear system respectively. If $Q > 0$, $P > 0$, then there exists T such that

$$\begin{aligned} Q &= T^*\Sigma T &(4.6)\\ P^{-1} &= T^*\Sigma^{-1}T &(4.7)\end{aligned}$$

where $\Sigma = \mathrm{Diag}(\sigma_{(1)}, \sigma_{(2)}, \dots, \sigma_{(n)})$ is a positive diagonal matrix, $\sigma_{(1)} \geq \sigma_{(2)} \geq \cdots \geq \sigma_{(n)} > 0$, and $\sigma_{(i)}$, $i = 1, 2, \dots, n$ are called the Hankel singular values of the linear system. Therefore if we make the state-transformation

$$\left(\begin{array}{c|c} A_b & B_b \\ \hline C_b & D_b \end{array}\right)_I := \left(\begin{array}{c|c} TAT^{-1} & TB \\ \hline CT^{-1} & D \end{array}\right)_I = \left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right)_I \tag{4.8}$$

we shall have

$$\begin{aligned} A_b^*\Sigma A_b - \Sigma + C_b^* C_b &= 0\\ A_b\Sigma A_b^* - \Sigma + B_b B_b^* &= 0\end{aligned}$$

Let

$$\begin{bmatrix} A_b & B_b \\ C_b & D_b \end{bmatrix} = \left[\begin{array}{cc:c} A_{b11} & A_{b12} & B_{b1} \\ A_{b21} & A_{b22} & B_{b2} \\ \hdashline C_{b1} & C_{b2} & D_b \end{array}\right] \tag{4.9}$$

where A_{b11} is $r \times r$, B_{b1} is $r \times m$ and C_{b1} is $p \times r$. Then it is known (cf. [2, 38] and [77, §21.8]) that the ℓ_2-ℓ_2 induced gain (i.e. the $\mathcal{H}_\infty$-norm) of the error system $\left(\begin{array}{c|c} A_b & B_b \\ \hline C_b & D_b \end{array}\right)_I - \left(\begin{array}{c|c} A_{b11} & B_{b1} \\ \hline C_{b1} & D_b \end{array}\right)_I$ is bounded by 2 times the sum of *distinct* $\sigma_{(i)}$, $i = r+1, r+2, \dots, n$. In fact this error bound holds even when Q, P only satisfy the Lyapunov *inequalities* (i.e. (4.1) and (4.2)) but not necessarily the equations. Different Q, P (there are a huge number of them) satisfying the Lyapunov inequalities give rise to different T and Σ and hence different reduced-order models and error bounds. Among all those Q, P the solutions to the *equations*, i.e. the observability gramian and the controllability gramian, are the natural choice since they are minimal under the partial order $\leq$ (i.e. any Q, P satisfying the Lyapunov inequalities will have $Q \geq$ the observability gramian and $P \geq$ the controllability gramian). In the case of the LFT uncertain system[13] or the ϕ-system, Q, P are further restricted to some subclass of positive definite matrices and therefore the Lyapunov *equations* may not be feasible, i.e. the inequalities instead of the equations *must* be solved, and there does *not* exist a natural choice of minimal solutions to those inequalities. This greatly complicates the task: For example, for the upper bound of the observability function and the lower bound of the controllability function presented in the previous section, there is *no* natural least upper bound or greatest lower bound. A similar problem exists for the model reduction.

Let us start with a simple case. In [13], Beck relaxed the definition of a balanced system by saying that any system $\left(\begin{array}{c|c} A_b & B_b \\ \hline C_b & D_b \end{array}\right)$ admitting a single diagonal solution to both Lyapunov inequalities (instead of the equations) is balanced. In that case any diagonally stable $\underline{\phi}$-system, $\underline{\phi} \in \mathcal{OL}^n(\mathbb{R})$, has a balanced realization in the following sense: it always admits a diagonal solution to *each* Lyapunov inequality, and therefore T is the product of a permutation matrix and a diagonal matrix and $\underline{\phi} \in \mathcal{OL}^n(\mathbb{R})$ if and only if $T \circ \underline{\phi} \circ T^{-1} \in \mathcal{OL}^n(\mathbb{R})$ (cf. Section 2.4.1). Hence we can get a balanced realization of the system by reselecting $\underline{\phi}$ and reordering the states. Now recall the following lemma from the theory of linear system balanced truncation. It originates in [74].

Lemma 4.3 Given $\begin{bmatrix} A_b & B_b \\ C_b & D_b \end{bmatrix}$. Suppose that $Q = P = \Sigma = \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix}$ solves the Lyapunov inequalities, where Σ_1 is $r \times r$ positive diagonal and $\Sigma_2 = I_{n-r}$. Partition $\begin{bmatrix} A_b & B_b \\ C_b & D_b \end{bmatrix}$ as in (4.9) and let

$$\begin{bmatrix} \hat{A}_{br} & \hat{B}_{br} \\ \hat{C}_{br} & \hat{D}_{br} \end{bmatrix} := \left[\begin{array}{cc:c} A_{b11} & 0 & B_{b1} \\ 0 & 0 & 0 \\ \hdashline C_{b1} & 0 & D_b \end{array}\right] \tag{4.10}$$

Then

$$\left\| \begin{bmatrix} -\Sigma^{\frac{1}{2}} & \Sigma^{\frac{1}{2}} & 0 \\ \Sigma^{-\frac{1}{2}} & \Sigma^{-\frac{1}{2}} & 0 \\ 0 & 0 & I \end{bmatrix} \begin{bmatrix} \hat{A}_{br} & 0 & \hat{B}_{br}/2 \\ 0 & A_b & B_b/2 \\ -\hat{C}_{br} & C_b & -\hat{D}_{br}/2 + D_b/2 \end{bmatrix} \begin{bmatrix} -\Sigma^{\frac{1}{2}} & \Sigma^{\frac{1}{2}} & 0 \\ \Sigma^{-\frac{1}{2}} & \Sigma^{-\frac{1}{2}} & 0 \\ 0 & 0 & I \end{bmatrix}^{-1} \right\| \le 1 \tag{4.11}$$

$\triangleleft$

Proof. See Appendix C. ■

Note that

$$\left(\begin{array}{c|c} A_{be} & B_{be} \\ \hline C_{be} & D_{be} \end{array}\right) := \left(\begin{array}{cc|c} \hat{A}_{br} & 0 & \hat{B}_{br}/2 \\ 0 & A_b & B_b/2 \\ \hline -\hat{C}_{br} & C_b & -\hat{D}_{br}/2 + D_b/2 \end{array}\right) \tag{4.12}$$

is in fact half the error system[1], and therefore (4.11) essentially means that the linear gain of half the error system ≤ 1. If $\Sigma_2 = \sigma_{(n)} I \neq I$, we can still apply Lemma 4.3 to

[1] strictly speaking, the error system/2, which is not necessarily $= \frac{1}{2} \times$ the error system when it is nonlinear, e.g. a $\underline{\phi}$-system or a ϕ-system, but the diagonal gain or the dd-gain is the same anyway. Cf. Remark 2.5.

$\begin{bmatrix} A_b & B_b/\sqrt{\sigma_{(n)}} \\ C_b/\sqrt{\sigma_{(n)}} & D_b/\sigma_{(n)} \end{bmatrix}$ to show that the linear gain of half the error system $\leq \sigma_{(n)}$.

At last we observe that Σ_1 will solve the Lyapunov inequalities for $\begin{bmatrix} A_{b11} & B_{b1} \\ C_{b1} & D_b \end{bmatrix}$, and therefore Lemma 4.3 can be applied again to the reduced-order model. This is how "2 times the sum of distinct $\sigma_{(i)}$" is obtained (and actually we need not arrange $\sigma_{(i)}$ in descending order since Lemma 4.3 does not require $\Sigma_1 \geq I$).

Clearly (4.11) does not only give a bound to the linear gain, but also a bound to the dd-gain since

$$\begin{bmatrix} -\Sigma^{\frac{1}{2}} & \Sigma^{\frac{1}{2}} \\ \Sigma^{-\frac{1}{2}} & \Sigma^{-\frac{1}{2}} \end{bmatrix}^* \begin{bmatrix} -\Sigma^{\frac{1}{2}} & \Sigma^{\frac{1}{2}} \\ \Sigma^{-\frac{1}{2}} & \Sigma^{-\frac{1}{2}} \end{bmatrix} = \begin{bmatrix} \Sigma^{-1}+\Sigma & \Sigma^{-1}-\Sigma \\ \Sigma^{-1}-\Sigma & \Sigma^{-1}+\Sigma \end{bmatrix}$$

is pdd. Indeed, since Σ is diagonal, according to the discussion at the beginning of Section 3.2 we see that the error bound applies even when $\left(\begin{array}{c|c} A_b & B_b \\ \hline C_b & D_b \end{array}\right)$ is only a $\underline{\phi}$-system, $\underline{\phi} \in \mathcal{OL}^n(\mathbb{R})$, but not necessarily a ϕ-system or a linear system. In other words, (4.11) is strong enough to give an error bound not only to the state-truncation of a balanced linear system, but also to the state-truncation of a balanced $\underline{\phi}$-system, and hence to the state-truncation of *any* diagonally stable $\underline{\phi}$-system since a diagonally stable $\underline{\phi}$-system always has a balanced realization. (It is indeed only a state-truncation since T is only the product of a permutation matrix and a diagonal matrix.) Let us summarize the result in the following

Proposition 4.4 Given a $\underline{\phi}$-system $G := \left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right)_{\underline{\phi}}$, $\underline{\phi} \in \mathcal{OL}^n(\mathbb{R})$. Suppose that there exist diagonal Q, P satisfying (4.1), (4.2). Let

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \left[\begin{array}{cc:c} A_{11} & A_{12} & B_1 \\ A_{21} & A_{22} & B_2 \\ \hdashline C_1 & C_2 & D \end{array}\right] \tag{4.13}$$

where A_{11} is $r \times r$, B_1 is $r \times m$ and C_1 is $p \times r$. Then the ℓ_2-ℓ_2 induced gain of

$$\left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right)_{\underline{\phi}} - \left(\begin{array}{c|c} A_{11} & B_1 \\ \hline C_1 & D \end{array}\right)_{\check{\underline{\phi}}}$$

where $\check{\underline{\phi}} := \mathrm{Diag}(\underline{\phi}_{(1)}, \underline{\phi}_{(2)}, \ldots, \underline{\phi}_{(r)})$, is bounded by 2 times the sum of *distinct* $\sigma_{(i)}$, $\sigma_{(i)} := \sqrt{q_{(i)}p_{(i)}}$, $i = r+1, r+2, \ldots, n$. $\triangleleft$

Remark 4.3 Here we see that for a balanced linear system $\left(\begin{array}{c|c} A_b & B_b \\ \hline C_b & D_b \end{array}\right)_I$ the truncation error bound holds even when the states are subject to different nonlinear uncertainties in $\mathcal{OL}(\mathrm{I\!R})$. So although it does not necessarily mean that the error bound is loose, the gap between the bound and the ℓ_2-ℓ_2 induced gain of $\left(\begin{array}{c|c} A_b & B_b \\ \hline C_b & D_b \end{array}\right)_I - \left(\begin{array}{c|c} A_{b11} & B_{b1} \\ \hline C_{b1} & D_b \end{array}\right)_I$ is always large enough to insert the ℓ_2-ℓ_2 induced gain of $\left(\begin{array}{c|c} A_b & B_b \\ \hline C_b & D_b \end{array}\right)_{\underline{\phi}} - \left(\begin{array}{c|c} A_{b11} & B_{b1} \\ \hline C_{b1} & D_b \end{array}\right)_{\check{\underline{\phi}}}$ in between. In particular, 2 times the sum of *all* distinct Hankel singular values is not only a bound on the linear gain but also the diagonal gain of $\left(\begin{array}{c|c} A_b & B_b \\ \hline C_b & 0 \end{array}\right)$. $\otimes$

It is a simple but not a very interesting case since only state-truncation is involved (T is essentially diagonal), unlike the linear case in which there is also "change of coordinates". The model reduction of the ϕ-system will be more interesting, but we first stay in this simple case for a little bit longer in order to illuminate some considerations in choosing the (diagonal) solutions to the Lyapunov inequalities.

Certainly in order to give a tighter error bound, Q, P should be chosen to be small, i.e. no diagonal solutions to the Lyapunov inequalities should be $\leq Q$, P. However, since $\leq$ is only a partial order, minimal elements are not unique unless the Lyapunov *equations* is feasible. A *convenient* choice is the Q, P with minimum trace. Beck chose this in [13]. With more computational efforts it is possible to choose Q, P such that Σ has the minimum trace, but this will not help a lot. The point is that the error bound given by Q, P is not tight anyway in comparison with the balanced truncation of *linear* systems. Suppose that we are going to truncate the i-th state, then we may minimize $q_{(i)}$, $p_{(i)}$ in order to have the *tightest* error bound. This can be done to each state. What cannot be done is to obtain a tight error bound for *all* states from a *single* pair of Q, P. On the other hand, solving a pair of Q, P for each state highly increases the computational efforts. The situation is even more complicated when the number of states to be truncated, $n - r$, is larger than 1. Suppose that we are going to truncate two states: state i and state j. It is not necessarily the best to minimize $2(\sigma_{(i)} + \sigma_{(j)})$ since repeated $\sigma_{(i)}$ are only counted once: If we force $\sigma_{(i)} = \sigma_{(j)}$, $\sigma_{(i)}$ and $\sigma_{(j)}$ may both increase but the error bound is then only $2\sigma_{(i)}$ instead of $2(\sigma_{(i)} + \sigma_{(j)})$. Indeed it is usually the case that if we know which states are to be truncated, forcing those $\sigma_{(i)}$ to be identical will yield a better error bound. Here reveals an interesting conflict between the two roles of Σ: On one hand Σ is used to *differentiate* the importance of the states, to see which states should be kept and which states should be discarded, and therefore $\sigma_{(i)}$ should be as different as possible in order to make the cut clear, while on the other hand Σ is also used to provide the error bound, which is usually tighter if the $\sigma_{(i)}$ of all the discarded states are combined into a single scalar block. One way to resolve this conflict is to use two different Σ for these two roles: We compute a pair of Q, P with minimum trace, for example, to see which states are to be truncated, and then another pair $\tilde{Q}$, $\tilde{P}$ to obtain a tighter error bound. However, the usefulness of the balanced truncation approach will then be much shaded: Unlike for the linear system we can no longer obtain a tight error bound *before* we decide how many and which states are to be truncated. Only a very rough idea of how large the relative error would be can be obtained (from the ratio defined in (4.14) below).

The point is now illustrated with numerical examples. We shall use those 22 examples we have studied in Section 2.3 and treat them as data of $\underline{\phi}$-systems. Suppose that $n-r$ states are to be truncated. The procedure is as follows:

[bt1] Obtain a pair of Q, P with minimum trace to determine which $n-r$ states are to be truncated. 2 times the sum of the corresponding $\sqrt{q_{(i)}p_{(i)}}$ forms an error bound. Reorder the states such that they become the last $n-r$ states, i.e. states $r+1, r+2, \ldots, n$.

[bt2] Next obtain another pair of solutions $\tilde{Q}$, $\tilde{P}$ to the Lyapunov inequalities which minimize $2\sqrt{\tilde{q}_{(r)}\tilde{p}_{(r)}} = 2\sqrt{\tilde{q}_{(r+1)}\tilde{p}_{(r+1)}} = \cdots = 2\sqrt{\tilde{q}_{(n)}\tilde{p}_{(n)}}$. Hence the error bound $2\sqrt{\tilde{q}_{(n)}\tilde{p}_{(n)}}$ is deduced.

[bt3] Obtain the truncated model.

[bt4] Obtain the error system. An upper bound of the ℓ_2-ℓ_2 induced gain of the error system is then computed based on the idea of the dd-gain. (cf. the discussion above or the one at the beginning of Section 3.2. The full details are in Section 5.1.1.)

[bt5] At last the linear gain of the error system is computed and serves as a lower bound of the ℓ_2-ℓ_2 induced gain of the error system.

Practically **[bt2]** is not necessary since the computational complexity of **[bt4]** is not much higher than **[bt2]**, but the objective now is to see that the bound given by **[bt1]** is typically looser than the one given by **[bt2]**, which is in sharp contrast to the state-truncation of a balanced linear system. (When $n-r=1$, **[bt2]** simply minimizes $q_{(n)}$, $p_{(n)}$ independently, but the situation is less obvious when $n-r>1$. However this is a special case of Proposition 4.5 and Corollary 4.6 in the next section and therefore the details are not described here.)

The results are expressed in ratios and their averages over all 22 examples are summarized below:

$n-r$	1	2	3
bound from Q, P/bound from $\tilde{Q}$, $\tilde{P}$	1.5481	2.2496	2.8710
error from **[bt4]**/bound from $\tilde{Q}$, $\tilde{P}$	0.6451	0.6458	0.6148
lower bound from **[bt5]**/error from **[bt4]**	0.6132	0.6899	0.7541

From the table we see that the error bound given by Q, P is looser than the error bound given by $\tilde{Q}$, $\tilde{P}$, and as expected the gap becomes larger as $n-r$ increases (since more $\sigma_{(i)}$ can be combined into a single scalar block). The second row shows that the actual error is lower than the *tighter* error bound by 35–40 percent in average. Note that this is the error of two $\underline{\phi}$-systems but not ϕ-systems: Of course a diagonally stable ϕ-system may be treated as a $\underline{\phi}$-system and truncated in the same way, so if the 22 examples are treated as data of diagonally stable ϕ-systems instead of $\underline{\phi}$-systems, the actual dd-error will be even lower and consequently the upper bound given by $\tilde{Q}$, $\tilde{P}$ will look looser and the lower bound given by the linear gain tighter, and the second and the third row of the above table will become

$n-r$	1	2	3
dd-error from [**bt4**]/bound from $\tilde{Q}$, $\tilde{P}$	0.4696	0.4981	0.4942
lower bound from [**bt5**]/dd-error from [**bt4**]	0.8409	0.8896	0.9348

We have also mentioned that in [**bt1**] it is possible to minimize the trace of Σ rather than the traces of Q, P independently but this will not make much difference. The first row[2] of the table will read

$n-r$	1	2	3
bound from Q, P/bound from $\tilde{Q}$, $\tilde{P}$	1.6469	2.2839	2.8496

The ratios are similar and our conclusion remains unchanged.

If we recall the typical experience in the balanced truncation of linear systems, it is curious that the *same* Σ can be used to differentiate the importance of the states and provide the error bound, i.e. the two conflicting roles can be achieved by a single Σ. We shall see that when the Lyapunov equations are feasible, the bound given by Q, P will typically be much closer to the bound given by $\tilde{Q}$, $\tilde{P}$. Indeed when $n-r=1$, Q (the observability gramian) and P (the controllability gramian) are minimal under the partial order $\leq$, and therefore $Q \leq \tilde{Q}$ and $P \leq \tilde{P}$ which imply $2\sqrt{q_{(n)}p_{(n)}} = 2\sqrt{\tilde{q}_{(n)}\tilde{p}_{(n)}}$, i.e. the two bounds are the same. When $n-r>1$, it is not so sure since repeated $\sqrt{\tilde{q}_{(i)}\tilde{p}_{(i)}}$ are only counted once. So we linearly balance our 22 examples, i.e. treat them as data of linear systems and balance them, so that the Lyapunov equations are made feasible, and then treat them back as $\underline{\phi}$-systems and repeat the state-truncation procedure [**bt1**]–[**bt5**]. Note that they are not the same 22 examples, certainly, but we are going to observe only the average tightness of the bound. The results are

$n-r$	2	3
bound from Q, P/bound from $\tilde{Q}$, $\tilde{P}$	1.0312	1.1457
error from [**bt4**]/bound from $\tilde{Q}$, $\tilde{P}$	0.7363	0.7397
lower bound from [**bt5**]/error from [**bt4**]	0.9185	0.9117

[2] The second and third rows of the table also change slightly because if $\sigma_{(i)}$ are very close, minimizing the trace of Σ instead of the traces of Q, P may possibly cause different states to be truncated but not a much different error. See the later discussion for more details.

The bound given by the solutions to the Lyapunov equations, i.e. the observability gramian Q and the controllability gramian P, is not optimal when $n-r>1$ since it is in average larger than and in no example smaller than the bound given by $\tilde{Q}$, $\tilde{P}$, but it is not a much looser bound as in the previous case, which verifies our experience in the balanced truncation of linear systems that the bound given by the gramians is already sufficient, and therefore based on it we can decide how many and which states are to be truncated. The fact is that when the Lyapunov equations are feasible, the Hankel singular values are typically more widely separate, $\sigma_{(1)} \gg \sigma_{(2)} \gg \cdots \gg \sigma_{(n)}$; otherwise they are much closer to each other, and then forcing $\sigma_{(i)}$ to be identical will yield a much tighter bound. This explains why in the balanced truncation of a linear system where the Lyapunov equations are feasible, a tight error bound can usually be obtained *beforehand*, but in the balanced truncation of a $\underline{\phi}$-system where the Lyapunov equations are in general not feasible, only a loose bound can be obtained beforehand and a tighter bound has to be found out *afterwards*.

Indeed Q, P can tell that a state is more important than another only when the corresponding $\sigma_{(i)}$ are widely separate, especially if we recall that there are no so-called natural solutions to the Lyapunov inequalities if the equations are not feasible. If all $\sigma_{(i)}$ are very close together, state-truncation may not make sense at all. In one of our original 22 examples, Ex. 14, all the $\sigma_{(i)}$ are very close to each other: $\sigma_{\max} = 3.5420$ and $\sigma_{\min} = 3.3213$, and it turns out that truncating *any one* state will result in a dd-error larger than the dd-gain of $\left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right)_{\underline{\phi}} - D$ (i.e. truncating *all* the states)![3] If it is true that $\sigma_{(1)}, \sigma_{(2)}, \ldots, \sigma_{(r)} \gg \sigma_{(r+1)}, \sigma_{(r+2)}, \ldots, \sigma_{(n)}$, then one may check the ratio

$$\rho_{\{r+1,r+2,\ldots,n\}} := \frac{\max_{i\in\{r+1,r+2,\ldots,n\}} \sqrt{q_{(i)}p_{(i)}}}{\max_{i\in\{1,2,\ldots,n\}} \sqrt{q_{(i)}p_{(i)}}} \tag{4.14}$$

to see whether it is acceptably small before the state-truncation is actually performed. From the examples we observe that this ratio is usually a reasonable estimate of the relative error := the dd-gain of the error system/the dd-gain of $\left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right)_{\underline{\phi}} - D$. So if this ratio is already large we should give up the idea of state-truncation. In fact even when $n-r=1$, among our original 22 examples only in two cases this ratio is ≤ 0.1 (0.0511 in Ex. 1 and 0.0521 in Ex. 17) and those two cases are really the only two cases that would have a relative error actually ≤ 0.1 (0.0255 and 0.0336 respectively). These high relative errors are in contrast to the balanced truncation of linear systems, or when the Lyapunov equations are feasible. Among our 22 linearly balanced examples the highest $\rho_{\{r+1,r+2,\ldots,n\}}$ is only 0.0410 which corresponds to the highest relative error

[3]This may also happen to a balanced *linear* system when the Hankel singular values are close, although this is not very typical. For example, in the linearly balanced Ex. 21, $\sigma_{(2)} = 1.4643 \approx \sigma_{(1)} = 1.5841$. Then truncating state 2 gives a (linear) error 2.8362 and truncating state 1 gives an error 2.7735, but truncating both only gives an error 2.3573. In fact the error of truncating only one of them is even larger than the error of truncating all the four states, which is only 2.3915.

of 0.0251. So we see that doing state-truncation for a general $\underline{\phi}$-system is not always possible, unlike for a balanced linear system most of the time it is, and any stable linear system can be balanced. This is not very surprising, however, since a $\underline{\phi}$-system is a nonlinear system after all, and moreover each $\underline{\phi}_{(i)}$ is supposed to be different. Consequently we cannot always find a realization to make the Lyapunov equations feasible as for a linear system, and simply truncating any one $\underline{\phi}_{(i)}$ may probably result in a large error. It is not hard to see that the *form* of a linear system does not specify a special set of directions in the state-space, only the *data* (A, B, C, D) do: for example the set of directions which balance the system; while the *form* of a $\underline{\phi}$-system or a ϕ-system already nominates a special set of state-space directions: the coordinate directions. (For example, Albertini and Sontag have found that the observability of a σ-system is in terms of its so-called coordinate subspace[7].) If these two sets of directions, the coordinate directions and the directions that balance the system are the same, the state-truncation error is just as small as in the linear case; but if these two sets of directions are very different, then state-truncation may not be possible, and there is not much that we can do about it.

4.3 Model reduction via a balanced system

A ϕ-system has very limited realizations (cf. Section 2.4.1) and usually does not have a balanced realization. Of course, Proposition 4.4 may be applied to a diagonally stable ϕ-system — just treat it as a $\underline{\phi}$-system. It is not interesting since only state-truncation is involved. If we examine the whole idea of the balanced truncation with a little bit care, we see that it is in fact not necessary for the original system $\left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right)$ to be equivalent to the balanced system $\left(\begin{array}{c|c} A_b & B_b \\ \hline C_b & D_b \end{array}\right)$, or in other words what we need is only the balanced system $\left(\begin{array}{c|c} A_b & B_b \\ \hline C_b & D_b \end{array}\right)$ but not really a balanced realization of the system $\left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right)$. If we substitute the state-transformation (4.8) back into (4.11), we get

$$\left\| \begin{bmatrix} -\Sigma^{\frac{1}{2}} & \Sigma^{\frac{1}{2}} & 0 \\ \Sigma^{-\frac{1}{2}} & \Sigma^{-\frac{1}{2}} & 0 \\ 0 & 0 & I \end{bmatrix} \begin{bmatrix} \hat{A}_{br} & 0 & \hat{B}_{br}/2 \\ 0 & TAT^{-1} & TB/2 \\ -\hat{C}_{br} & CT^{-1} & -\hat{D}_{br}/2 + D/2 \end{bmatrix} \begin{bmatrix} -\Sigma^{\frac{1}{2}} & \Sigma^{\frac{1}{2}} & 0 \\ \Sigma^{-\frac{1}{2}} & \Sigma^{-\frac{1}{2}} & 0 \\ 0 & 0 & I \end{bmatrix}^{-1} \right\| \leq 1$$

where $\hat{A}_{br}, \hat{B}_{br}, \hat{C}_{br}, \hat{D}_{br}$ are as given in (4.10), or equivalently

$$\left\| \begin{bmatrix} -\Sigma^{\frac{1}{2}}T & \Sigma^{\frac{1}{2}}T & 0 \\ \Sigma^{-\frac{1}{2}}T & \Sigma^{-\frac{1}{2}}T & 0 \\ 0 & 0 & I \end{bmatrix} \begin{bmatrix} \hat{A}_r & 0 & \hat{B}_r/2 \\ 0 & A & B/2 \\ -\hat{C}_r & C & -\hat{D}_r/2 + D/2 \end{bmatrix} \begin{bmatrix} -\Sigma^{\frac{1}{2}}T & \Sigma^{\frac{1}{2}}T & 0 \\ \Sigma^{-\frac{1}{2}}T & \Sigma^{-\frac{1}{2}}T & 0 \\ 0 & 0 & I \end{bmatrix}^{-1} \right\| \leq 1 \tag{4.15}$$

where

$$\left(\begin{array}{c|c} \hat{A}_r & \hat{B}_r \\ \hline \hat{C}_r & \hat{D}_r \end{array} \right) := \left(\begin{array}{c|c} T^{-1}\hat{A}_{br}T & T^{-1}\hat{B}_{br} \\ \hline \hat{C}_{br}T & \hat{D}_{br} \end{array} \right) \tag{4.16}$$

So what we need to do is to make sure that

$$\begin{bmatrix} -\Sigma^{\frac{1}{2}}T & \Sigma^{\frac{1}{2}}T \\ \Sigma^{-\frac{1}{2}}T & \Sigma^{-\frac{1}{2}}T \end{bmatrix}^* \begin{bmatrix} -\Sigma^{\frac{1}{2}}T & \Sigma^{\frac{1}{2}}T \\ \Sigma^{-\frac{1}{2}}T & \Sigma^{-\frac{1}{2}}T \end{bmatrix} = \begin{bmatrix} P^{-1}+Q & P^{-1}-Q \\ P^{-1}-Q & P^{-1}+Q \end{bmatrix}$$

is pdd so that the dd-gain of half the error system ≤ 1, and $\left(\begin{array}{c|c} \hat{A}_r & \hat{B}_r \\ \hline \hat{C}_r & \hat{D}_r \end{array} \right)$ is equivalent to a r-state system (it now looks a n-state system). If both can be done, a reduced-order model can be obtained just as before, although $\left(\begin{array}{c|c} A_b & B_b \\ \hline C_b & D_b \end{array} \right)_\phi$ is by no means equivalent to $\left(\begin{array}{c|c} A & B \\ \hline C & D \end{array} \right)_\phi$.

Proposition 4.5 Given a n-state ϕ-system $G := \left(\begin{array}{c|c} A & B \\ \hline C & D \end{array} \right)_\phi$, $\phi \in \mathcal{OL}(\mathbb{R})$. Suppose that there exist Q, P satisfying

$$A^*QA - Q + C^*C/\sigma \;\leq\; 0 \tag{4.17}$$
$$APA^* - P + BB^*/\sigma \;\leq\; 0 \tag{4.18}$$

such that

$$\begin{bmatrix} P^{-1}+Q & P^{-1}-Q \\ P^{-1}-Q & P^{-1}+Q \end{bmatrix} \tag{4.19}$$

is pdd, and the solutions T, Σ to (4.6), (4.7) can be partitioned as

$$\begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix}, \quad \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix}$$

where T_{11}, Σ_1 are $r \times r$, T_{12} or $T_{21} = 0$ (i.e. T is block-triangular), Σ_1 is positive diagonal and $\Sigma_2 = I_{n-r}$. Let $\begin{bmatrix} A_b & B_b \\ C_b & D_b \end{bmatrix}$ be defined as in (4.8) and partitioned as in

(4.9). Define a r-state ϕ-model

$$G_r := \left(\begin{array}{c|c} A_r & B_r \\ \hline C_r & D_r \end{array}\right)_\phi := \left(\begin{array}{c|c} T_{11}^{-1}A_{b11}T_{11} & T_{11}^{-1}B_{b1} \\ \hline C_{b1}T_{11} & D_b \end{array}\right)_\phi \tag{4.20}$$

Then the dd-gain of the error system $G - G_r \leq 2\sigma$. ◁

Proof. It remains to note that since T is block-triangular, the r-state system $\left(\begin{array}{c|c} A_r & B_r \\ \hline C_r & D_r \end{array}\right) \simeq$ the n-state system $\left(\begin{array}{c|c} \hat{A}_r & \hat{B}_r \\ \hline \hat{C}_r & \hat{D}_r \end{array}\right)$ in (4.16). ■

Remark 4.4 The condition that (4.19) is pdd is stronger than the condition that both Q and P^{-1} are pdd as required in Proposition 4.1 and 4.2. In fact, it is equivalent to the condition that the matrix being composed of the smaller diagonal entries and the larger off-diagonal entries of Q and P^{-1} is still pdd.

Recall that for the balanced truncation of a linear system or a $\underline{\phi}$-system, although $\Sigma_2 = I$ in Lemma 4.3, it does not really matter if Q, P give a $\Sigma_2 = \sigma I \neq I$ since we can scale Q, P by scaling C, B and then apply Lemma 4.3 and scale back the error bound. However, the condition here that (4.19) is pdd is in general not scalable, which must be fulfilled *when* $\Sigma_2 = I$ but *not* σI. This causes σ to appear explicitly in the Lyapunov inequalities (4.17), (4.18). ⊗

Corollary 4.6 Given a n-state ϕ-system $G := \left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right)_\phi$, $\phi \in \mathcal{OL}(\mathbb{R})$. Suppose that there exist Q, P satisfying (4.17), (4.18) such that (4.19) is pdd and

$$Q - P^{-1} = \begin{bmatrix} Q_{11} - (P^{-1})_{11} & 0 \\ 0 & 0 \end{bmatrix} \tag{4.21}$$

where Q_{11}, $(P^{-1})_{11}$ are the top-left $r \times r$ submatrices of Q and P^{-1} respectively. Then there exists a r-state ϕ-model $G_r := \left(\begin{array}{c|c} A_r & B_r \\ \hline C_r & D_r \end{array}\right)_\phi$ such that the dd-gain of the error system $G - G_r \leq 2\sigma$. ◁

Proof. It suffices to note that if (4.21) holds we may choose $T_{12} = 0$ and $\Sigma_2 = I$. ■

Hence the conditions on T and Σ are converted into a condition on $Q - P^{-1}$ which becomes more tractable, and therefore we shall set T_{12} instead of $T_{21} = 0$.

Constraint (4.21) can be easily implemented for numerical calculation, but is not invariant under the permutation of the states. A more general way to state the constraint is that $Q - P^{-1}$ has rank $\leq r$ and its null space contains a $(n - r)$-dimensional coordinate subspace, which can be written as in Corollary 4.7 below.

Corollary 4.7 Given a n-state ϕ-system $G := \left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right)_\phi$, $\phi \in \mathcal{OL}(\mathrm{I\!R})$. Suppose that there exist Q, P satisfying (4.17), (4.18) such that (4.19) is pdd and there exists a diagonal Λ with $\mathrm{rank}(\Lambda) \leq r$ such that

$$\Lambda \geq Q - P^{-1} \geq -\Lambda \tag{4.22}$$

Then there exists a r-state ϕ-model $G_r := \left(\begin{array}{c|c} A_r & B_r \\ \hline C_r & D_r \end{array}\right)_\phi$ such that the dd-gain of the error system $G - G_r \leq 2\sigma$. $\triangleleft$

Remark 4.5 If $\Sigma_1 \geq I$, (4.22) may be replaced by

$$\Lambda \geq Q - P^{-1} \geq 0 \tag{4.23}$$

$\otimes$

Although (4.21) is a convex constraint, the set of matrices having a deficient rank is not convex and therefore finding a deficient-rank Λ to satisfy (4.22) or (4.23) is not a convex problem, which makes the whole model reduction problem computationally difficult. However, solving a deficient-rank solution to a LMI is a well-known problem since it has arisen in many similar contexts such as minimality, model reduction and low-order controller design of LFT uncertain systems. A totally satisfactory solution to this whole class of problems has not been found out yet. David and De Moor have proposed a numerical algorithm[26, 27] to minimize the rank of a matrix subject to a LMI, which is roughly based on the interior-point method. With this algorithm the model reduction problem of a ϕ-system may be posed as follows: Given the error bound $2\sigma_0$, find a minimum-order model G_r of G such that the error is guaranteed to be $\leq 2\sigma_0$. A model reduction algorithm to solve this problem can be:

[mr1] Given the candidate error bound $2\sigma_0$, solve (4.17), (4.18) (with $\sigma = \sigma_0$) and (4.22) for Q, P^{-1}, and a diagonal Λ such that (4.19) is pdd and $\mathrm{rank}(\Lambda)$ is minimized. Note that (4.18) can be converted into a LMI in P^{-1} using Schur complement. This rank minimization problem is not convex and David and De Moor's algorithm would be used. Rearrange the states such that $\Lambda = \begin{bmatrix} \Lambda_1 & 0 \\ 0 & 0 \end{bmatrix}$ and let $r \times r$ be the dimension of Λ_1. In fact, the objective of this step is only to order the states and determine r, but Q, P^{-1} will no longer be used. Another pair of Q, P^{-1} that decide the reduced-order model will be computed in **[mr2]**.

[mr2] Solve (4.17), (4.18), (4.21) for Q, P^{-1}, σ such that (4.19) is pdd and 2σ is minimized. This is a standard LMI optimization problem, i.e. convex. Certainly $2\sigma \leq 2\sigma_0$ and this 2σ is an upper bound of the error.

[mr3] With the Q, P^{-1} obtained in **[mr2]**, solve (4.6), (4.7) for T, Σ such that $T_{12} = 0$ and hence the reduced-order model G_r according to Proposition 4.5.

[mr4] Compute the dd-gain of the error system $G - G_r$, but treat C_r as a variable too when minimizing γ^2. Note that the LMI that computes the dd-gain, (2.27), is in fact convex in P, C, D when A, B are fixed. So when we compute the dd-gain of the error system after we have obtained the reduced-order model, we may reselect C_r (or even D_r if a strictly proper error system is not insisted) to further minimize the dd-error. It turns out that the new C_r will usually be different but not very much different from the original C_r. The effect is somewhat like fine-tuning.

[mr5] At last the linear gain of the error system is computed and serves as a lower bound of the ℓ_2-ℓ_2 induced gain of the error system.

We have said that **[mr2]**, which is derived from Corollary 4.6, is not invariant under state-permutations. More precisely, it depends on which states are the first r states and which states are the last $n - r$ states of G. So **[mr2]** may result in $\mathrm{C}_r^n := \frac{n!}{r!(n-r)!}$ different reduced-order models. Finding out all of them to see which one is the best is not practical and therefore **[mr1]** is used to select one. This is similar to the relation between **[bt2]** and **[bt1]** in the state-truncation procedure, but **[bt2]** was only used to compute the bound but not the reduced-order model, and therefore is not needed in practice, while here **[mr2]** is essential.

Although in practice we would not work out all C_r^n combinations, here in order to understand the performance and limitations of the model reduction algorithm we have done it for all our 22 examples which are now treated as data of ϕ-systems. We ignore **[mr1]** first. Now suppose that $n - r$ states are to be truncated, then we reorder the states in all C_r^n ways and apply **[mr2]**. The best σ is selected[4] and the corresponding reduced-order model obtained (**[mr3]**). The C_r-optimal reduced-order model and its performance are then computed (**[mr4]**) and so is the linear gain of the error system to serve as a lower bound (**[mr5]**) as before. Note that we have chosen the reduced-order model corresponding to the smallest error bound 2σ but not the smallest dd-error, which are not necessarily the same. Since all C_r^n combinations have already been worked out, it is not difficult to find out the latter but we think that the former is more natural because in the whole tradition of balanced truncation or Hankel-norm approximation it is σ but not the actual error that plays the central role of determining the reduced-order model, and we still hope that one day we shall know how to order the states to achieve the best σ directly without working out all C_r^n combinations.

All the results presented below are the average values of all 22 examples. We first observe the tightness of the upper bound 2σ and the lower bound, i.e. the linear gain of the error system:

[4] Certainly Σ_1 is not always $\geq I$ for all C_r^n combinations, and we observe that the best σ does not necessarily correspond to a $\Sigma_1 \geq I$.

$n-r$	1	2	3
dd-error from [**mr4**]/2σ from [**mr2**]	0.5077	0.4850	0.5074
lower bound from [**mr5**]/dd-error from [**mr4**]	0.6955	0.8237	0.8147

Next we observe how much improvement has been made by this model reduction algorithm over the previous simple state-truncation. As we have said a diagonally stable ϕ-system may be treated as a $\underline{\phi}$-system and have the state-truncation procedure [**bt1**]–[**bt5**] applied. In principle the new model reduction method always gives a better error bound, and its average improvement over the 22 examples is summarized below:

$n-r$	1	2	3
2σ from [**mr2**]/bound from [**bt2**]	0.5447	0.5861	0.6221

The actual dd-error need not be smaller, at least theoretically speaking, but the average results of our 22 examples turn out to be

$n-r$	1	2	3
dd-error from [**mr4**]/dd-error from [**bt4**]	0.6083	0.5962	0.6745

(Recall that the dd-error from [**bt4**] should now be the error of the two ϕ-systems but not the $\underline{\phi}$-systems.) The comparison is not totally fair, however, since in the new model reduction algorithm C_r is further optimized. Its effect is shown in the following table:

$n-r$	1	2	3
dd-error when C_r is optimized/ dd-error when C_r is not optimized	0.8691	0.8034	0.8510
dd-error when both C_r and D_r are optimized/ dd-error when only C_r is optimized	0.9729	0.9674	0.9304

So we see that the effect of optimizing C_r is significant: the dd-error is reduced by about 15–20 percent, while the effect of further optimizing D_r is smaller: the dd-error is only further reduced by about 5 percent.

The significance of all the above statistics should not be overestimated. Model reduction of a ϕ-system is not always possible, so in assessing the improvement of the new model reduction algorithm over the simple state-truncation, it is not exactly the right question to ask by how much the error is reduced *in average*, but in *how many cases* model reduction has become from impossible to possible. Cases with the relative error reduced from 100 percent to 50 percent are not significant, while cases with the relative error reduced from 10 percent to 5 percent are. So the above average values were only meant to give a rough idea of the performance of the new model reduction algorithm, and now we are going to have a closer look into those particular cases where the relative error can be small.

Suppose that our aim is no more than 10 percent relative error, as before. (Recall that the relative error is the dd-error/the dd-gain of $\left(\begin{array}{c|c} A & B \\ \hline C & 0 \end{array}\right)$.) Again we suppress **[mr1]** first and use **[mr2]**–**[mr4]** to compute all the cases $r = 0, 1, 2, 3$ for each Ex. (i.e. r is not fixed) and Ex. $= 1, 2, \ldots, 22$. The number of cases that can have a relative error no larger than 10 percent turns out to be 13, and 7 examples are involved. (That means, for certain examples there exist two or more reduced-order models whose relative errors are no larger than 10 percent, and sometimes those reduced-order models may even have different orders r.) So the improvement is 7 out of 22 comparing with 4 out of 22 in the state-truncation of ϕ-systems and 2 out of 22 in the balanced truncation of $\underline{\phi}$-systems.

Now **[mr1]** is employed to see whether it can identify those 7 examples. Note that **[mr1]** can only work on $2\sigma_0$ but not directly on the relative error. Since the error bound 2σ is not very tight, we set $2\sigma_0 = 0.2 \times$ the dd-gain of $\left(\begin{array}{c|c} A & B \\ \hline C & 0 \end{array}\right)$ although our aim is 0.1. 8 out of 22 examples can have $2\sigma \leq 0.2 \times$ the dd-gain of $\left(\begin{array}{c|c} A & B \\ \hline C & 0 \end{array}\right)$ and David and De Moor's algorithm[5] succeeds in 5 examples. In one example, Ex. 6, three reduced-order models of order 3 may have $2\sigma \leq 0.2 \times$ the dd-gain of $\left(\begin{array}{c|c} A & B \\ \hline C & 0 \end{array}\right)$ and David and De Moor's algorithm succeeds in finding one of them, although it is not the one corresponding to the smallest 2σ. This example we still consider as a success. In another example, Ex. 11, a reduced-order model of order 2 may have $2\sigma = 0.1988 \times$ the dd-gain of $\left(\begin{array}{c|c} A & B \\ \hline C & 0 \end{array}\right)$ but David and De Moor's algorithm can only find a reduced-order model of order 3 which has $2\sigma = 0.1793 \times$ the dd-gain of $\left(\begin{array}{c|c} A & B \\ \hline C & 0 \end{array}\right)$, and this example we consider as a failure. In the other two unsuccessful example there exists a reduced-order model of order 3 which may have $2\sigma \leq 0.2 \times$ the dd-gain of $\left(\begin{array}{c|c} A & B \\ \hline C & 0 \end{array}\right)$ (0.1955, 0.1886) but David and De Moor's algorithm fails to identify them. For Ex. 17, a reduced-order model of order 2 is successfully pointed out by David and De Moor's algorithm.

If we go back to the actual dd-error, all 6 reduced-order models give a relative error ≤ 0.1, our original aim, although for Ex. 6 the reduced-order model obtained is not the best one (the relative error is 0.0819 while the best one can be 0.0776). So among all 7 examples which may have a relative error ≤ 0.1, only one (Ex. 20) is missed (whose

[5] See Appendix D for the actual implementation.

$2\sigma = 0.1886$ and relative error 0.0918).

The overall impression of the performance of David and De Moor's algorithm is that when 2σ is close (from below) to $2\sigma_0$, it will probably fail. In other words, some "margin" is needed. This also implies that if in two cases 2σ are close we may not be able to use David and De Moor's algorithm to find out the better case by choosing $2\sigma_0$ in between. It also seems that the "margin" needed is larger when the reduced-order model is to have a smaller order. In conclusion, David and De Moor's algorithm does not always succeed, but as the problem itself is nonconvex, the performance of the algorithm may still be regarded as satisfactory. To use it in our model reduction algorithm, the bound $2\sigma_0$ should not be set too tight, but some "margin" should be allowed for David and De Moor's algorithm to work.

4.4 Using Hankel-norms

In this section an alternative to [**mr1**] is provided for the model reduction algorithm. The main motivation is that although [**mr1**] works all right, its convenience is not comparable to the balanced truncation of linear systems where a list of Hankel singular values can be obtained in advance so that one knows at a glance what will happen in all situations, which greatly facilitates the determination of the order of the reduced-order model. In this section we present a method to obtain a similar list of numbers which will make the model reduction of a ϕ-system much more convenient. Although practically attractive, theoretically it is less convincing and therefore it is only meant to be an alternative to but not a replacement of [**mr1**].

Clearly [**mr1**] is needed mainly because [**mr2**] is not invariant under permutation of the states. If we know how the states should be ordered before [**mr2**] is applied, [**mr1**] is practically not preferred since it is not a convex problem. In the balanced truncation of a linear system or a $\underline{\phi}$-system, we order the states according to $\sigma_{(i)}$ and then the last $n-r$ states are truncated. Here we want something similar: to order the states and then apply [**mr2**]. If the ϕ-system is diagonally stable, one may wonder whether we can combine [**bt1**] with [**mr2**]–[**mr5**], i.e. we order the states according to a pair of diagonal solutions to the Lyapunov inequalities. It is a natural thought: After all this pair of diagonal solutions would provide an error bound. However, besides the facts that not all dd-stable ϕ-systems are diagonally stable and that the error bound will be even looser in this case, there are some other considerations which urge us to find a better method.

The point is that using a pair of diagonal solutions to the Lyapunov inequalities we take it for granted that a natural order of the states of G exists *independent* of r, which is not obvious especially when G is nonlinear. To see this we had better look into some numerical examples. For convenience the following notation is introduced: Let $S_r \subset \{1,2,\ldots,n\}$ have cardinality r and $\bar{S}_{n-r} := \{1,2,\ldots,n\}\backslash S_r$. Define a r-state reduced-order model G_{S_r} of G by reordering the states of G such that state i becomes one of the first r states if $i \in S_r$ or one of the last $n-r$ states if $i \in \bar{S}_{n-r}$ and then applying **[mr2]**–**[mr5]**. Roughly speaking, to say that an order independent of r exists we mean that if $G_{S_0}, G_{S_1}, \ldots, G_{S_n}$ are the best $0, 1, \ldots, n$-state reduced-order models of G then $S_0 \subset S_1 \subset \cdots \subset S_n$.

If our 22 examples are examined it will be found that such an order does not always exist. We take Ex. 6 as an example. The $1, 2, 3$-state reduced-order models corresponding to the smallest σ and dd-errors (which are the same in this case) are $G_{\{2\}}, G_{\{2,3\}}, G_{\{1,3,4\}}$ respectively. Note that $2 \in S_1$ and S_2 but not S_3. The 2σ and the dd-errors for all 3-state reduced-order models are summarized in the following table:

model	$G_{\{2,3,4\}}$	$G_{\{1,3,4\}}$	$G_{\{1,2,4\}}$	$G_{\{1,2,3\}}$
2σ	0.1785	0.1745	0.2907	0.2109
dd-error	0.0886	0.0840	0.1515	0.0992

As an order independent of r does not exist in this example, it is natural to ask what happens if the minimum-trace diagonal solutions to the Lyapunov inequalities are used. The answer is

state i	1	2	3	4
$2\sigma_{(i)}$	1.9232	2.5690	1.3132	2.2885
dd-error if state i is truncated	0.3030	0.4524	0.1521	0.4995

(The error bounds $2\sigma_{(i)}$ are roughly the same if the trace of Σ instead of Q, P is minimized.) Observe that the error bounds are very loose and even much looser when applied to the model reduction algorithm **[mr2]**–**[mr5]**. However, as far as state-truncation is concerned, $\sigma_{(i)}$ give a correct indication of which state should be truncated (i.e. state 3). On the other hand, if it is applied to **[mr2]**–**[mr5]**, $G_{\{1,2,4\}}$ will be the *worst* 3-state reduced-order model! So this example is very illustrative: if T can only be diagonal, or in other words if only state-truncation is allowed, model reduction is in fact impossible in this case since the dd-error of truncating any state is large (at least 0.1521), but if T is relaxed to block-triangular, the dd-error can be greatly reduced and model reduction will become possible. Of course this case is not typical, but in the model reduction of the ϕ-system it is the non-typical case that counts.

One may not be totally convinced by this example since $G_{\{2,3,4\}}$ is almost as good as $G_{\{1,3,4\}}$ (2σ of 0.1785 and 0.1745 and dd-errors of 0.0886 and 0.0840 respectively) and therefore one may argue that we may roughly order the states of G in the sequence $(2, 3, 4, 1)$. So let us take one more example, Ex. 18:

model $\bar{S}_{n-r}$	$G_{\{2,3,4\}}$ $\{1\}$	$G_{\{1,3,4\}}$ $\{2\}$	$G_{\{1,2,3\}}$ $\{4\}$	$G_{\{3,4\}}$ $\{1,2\}$	$G_{\{2,3\}}$ $\{1,4\}$	$G_{\{1,3\}}$ $\{2,4\}$
2σ	1.2087	0.9973	0.3277	1.2129	1.4317	1.3168
dd-error	0.4704	0.4177	0.1353	0.5575	0.6814	0.7420

The best 3-state reduced-order model is $G_{\{1,2,3\}}$ but the best 2-state reduced-order model is $G_{\{3,4\}}$.

In conclusion, we would not like to replace [**mr1**] by [**bt1**], or in other words, use the diagonal solutions to the Lyapunov inequalities to order the states before [**mr2**] is applied, because 1) a dd-stable ϕ-system may not be diagonally stable, 2) the bounds given by the diagonal solutions may be much too loose to be advantageous, and 3) the best order is not independent of r in general. So we are going to propose another method, which is based on the observability function and the controllability function in Section 4.1.

Let Q, P^{-1} pdd satisfy the Lyapunov inequalities (4.1) and (4.3), i.e. $\zeta^* Q \zeta$ is an upper bound of the observability function and $\zeta^* P^{-1} \zeta$ is a lower bound of the controllability function. Define

$$h_{\bar{S}_{n-r}} := \sup_{\substack{\zeta \in \mathbb{R}^n \\ \zeta_{(i)}=0 \text{ if } i \notin \bar{S}_{n-r}}} \sqrt{\frac{\zeta^* Q \zeta}{\zeta^* P^{-1} \zeta}} \tag{4.24}$$

which may be interpreted as an upper bound of the Hankel-norm with ζ restricted to the coordinate subspace associated with $\bar{S}_{n-r}$. Computationally we simply pick the corresponding submatrices of Q, P^{-1} and simultaneously diagonalize them then the largest Hankel singular value is $h_{\bar{S}_{n-r}}$. Note that [**mr2**] implies that the Hankel-norm restricted to $\bar{S}_{n-r}$ is ≤ 1 after the system is scaled by $1/\sigma$, where σ is liked to be small. So the method we propose is: *not* to choose any reduced-order model G_{S_r} to which the corresponding $h_{\bar{S}_{n-r}}$ is already large, provided that $h_{\bar{S}_{n-r}}$ is a reasonable upper bound of the Hankel-norm. More precisely, instead of computing all $\sum_{r=1}^{n} \mathrm{C}_r^n = 2^n - 1$ pairs of Q, P^{-1} satisfying (4.17), (4.18), (4.21) and (4.19) pdd, we only compute *one* pair of pdd Q, P^{-1} satisfying (4.1) and (4.3) but all $2^n - 1$ values of $h_{\bar{S}_{n-r}}$ from them. Although the number of $h_{\bar{S}_{n-r}}$ also grows as $2^n - 1$, computing a $h_{\bar{S}_{n-r}}$ is much much faster than computing a pair of Q, P^{-1} and therefore working out all $2^n - 1$ values is practical for all reasonable values of n. The problem remains is how to choose a suitable pair of pdd solutions Q, P^{-1} to (4.1), (4.3). As before there is no natural choice and now minimizing the trace of P is not a standard LMI problem since P^{-1} is required to be pdd. So instead we minimize the trace of Q and *maximize* the trace of P^{-1}. Again this is only a convenient choice. Having obtained all $h_{\bar{S}_{n-r}}$ we compute the ratio

$$\rho_{\bar{S}_{n-r}} := \frac{h_{\bar{S}_{n-r}}}{h_{\bar{S}_n}} \tag{4.25}$$

Note that this definition is consistent with (4.14). From our 22 examples it is observed that this ratio is usually a reasonable estimate of the relative error := the dd-error/the dd-gain of $\left(\begin{array}{c|c} A & B \\ \hline C & 0 \end{array}\right)$. So for example if we aim at a relative error of 0.1, we may only work out those reduced-order models G_{S_r} that have $\rho_{\bar{S}_{n-r}} \leq 0.1$ instead of all of them using [**mr2**]–[**mr5**]. Take Ex. 6 as an example. All $\rho_{\bar{S}_{n-r}}$ are shown in the following tables (the actual relative errors are only shown for reference; they need not be computed in the current method):

$\bar{S}_{n-r}$	$\{1\}$	$\{2\}$	$\{3\}$	$\{4\}$
$\rho_{\bar{S}_{n-r}}$	0.0919	0.0902	0.1605	0.1135
relative error of G_{S_r}	0.0819	0.0776	0.1400	0.0917

$\bar{S}_{n-r}$	$\{1,2\}$	$\{1,3\}$	$\{1,4\}$	$\{2,3\}$	$\{2,4\}$	$\{3,4\}$
$\rho_{\bar{S}_{n-r}}$	0.1551	0.1609	0.1318	0.1802	0.2084	0.1690
relative error of G_{S_r}	0.1558	0.1558	0.1075	0.1675	0.2190	0.1645

$\bar{S}_{n-r}$	$\{1,2,3\}$	$\{1,2,4\}$	$\{1,3,4\}$	$\{2,3,4\}$
$\rho_{\bar{S}_{n-r}}$	0.2076	0.9308	0.1711	0.2537
relative error of G_{S_r}	0.2204	0.8382	0.1715	0.2346

If we aim at a relative error ≤ 0.1, we may simply reject all $\rho_{\bar{S}_{n-r}} > 0.1$ and then only $G_{\{2,3,4\}}$ and $G_{\{1,3,4\}}$ will be worked out. Of course one may set the acceptance level of $\rho_{\bar{S}_{n-r}}$ a little bit higher if more computations can be afforded. In that case probably $G_{\{1,2,3\}}$ and $G_{\{2,3\}}$ will be worked out as well. At last the most desirable reduced-order model is selected among them.

The result of applying this method to all our 22 examples is as follows: If we only accept those $\rho_{\bar{S}_{n-r}} \leq 0.1$, among all 13 cases (7 examples) that have a relative error ≤ 0.1, 10 cases (6 examples) can be identified. For those examples corresponding to two or more reduced-order models, the best ones are always selected. The only failure is (still) Ex. 20 where $\rho_{\{2\}} = 0.1290$ but the relative error of $G_{\{1,3,4\}}$ is only 0.0918. The method also works out 3 extra cases (2 examples) that actually have a relative error > 0.1.

In summary, a list of numbers $h_{\bar{S}_{n-r}}$ are computed based on an adequate pair of pdd solutions Q, P^{-1} to the Lyapunov inequalities (4.1), (4.3). The objective of this list of numbers is to facilitate various decisions needed for the model reduction. In particular it helps to reject *in advance* a large portion of all C_r^n reduced-order models obtainable from [**mr2**], and allows us to work out those few plausible ones only. The rejection may be based on an acceptance level of $\rho_{\bar{S}_{n-r}}$ and/or the amount of computations we can afford.

4.5 Afterword

There is no obvious way to do model reduction for the ϕ-system. The approach adopted in this chapter follows the line of conventional balanced truncation of linear systems, and recovers those results when the ϕ-system is linearly balanced. The approach has been chosen mainly because it gives us a better intuition of the global ℓ_2-ℓ_2 behaviour of the error system, which is our favourite measure among the others, although strictly speaking, its direct connection is to the observability and controllability function rather than the ℓ_2-ℓ_2 induced gain.

Our model reduction results for the ϕ-system may possibly be applied to the complexity reduction of a static 1HLN, simply by putting $A = 0$. However, it should be realized that the focus of the neural network community on the complexity reduction of a 1HLN is sometimes different, which is basically an interpolation problem: Given a set of input-output patterns, they want to find a static map which fits them. A more complex 1HLN allows a larger degree of freedom, but on the other hand its complexity is liked to be as low as possible since an over-complex 1HLN has some undesirable effect (called low generalizability). Therefore they usually start with a complex 1HLN and then want to do complexity reduction later, during which the given set of input-output patterns is the ultimate measure. This is in contrast to our situation that it is the whole function rather than a finite set of points we would like to approximate and therefore the chance of getting a satisfactory reduced-order model is much lower.

Chapter 5

Extended results

Having presented various results that can be derived from the pdd approach, we are going to show that the applicability of those results can actually be extended to a more general class of nonlinear systems. In Section 5.1 $\underline{\phi}$-systems with partial but not all $\underline{\phi}_{(i)}$ identical are considered, and in Section 5.2 the nonlinearity ϕ is extended from static to dynamic. The understanding of the latter is also essential in constructing cases where the pdd approach can be shown to be tight, as demonstrated in the example in Section 5.1.3 below.

5.1 Systems: General case

5.1.1 General $\underline{\phi}$-systems

Recall that if P is diagonal, then

$$\forall \underline{\phi} \in \mathcal{S}^n, \;\; \forall \zeta \in \mathbb{R}^n, \;\; \underline{\phi}(\zeta)^* P \underline{\phi}(\zeta) \leq \zeta^* P \zeta \tag{5.1}$$

This is basically Kaszkurewicz and Bhaya's diagonal approach to the $\underline{\phi}$-system (Section 1.2.2). Since (5.1) holds no matter $\underline{\phi}_{(i)}$ are identical or not, it is too conservative to do certain kinds of analysis and synthesis where some $\underline{\phi}_{(i)}$ are forced to be identical by nature. On the other hand, we have shown that if P is a pdd matrix,

$$\forall \phi \in \mathcal{OL}(\mathbb{R}), \;\; \forall \zeta \in \mathbb{R}^n, \;\; \phi(\zeta)^* P \phi(\zeta) \leq \zeta^* P \zeta \tag{5.2}$$

Clearly we may combine the two results to obtain a block-diagonal version for the general $\underline{\phi}$-system, where a set of identical $\underline{\phi}_{(i)}$ will correspond to a single block on the diagonal of P.

Definition 5.1 A matrix P and a diagonal nonlinear operator $\underline{\phi}$ are said to have compatible structures if they have the same dimension and $\underline{\phi}_{(i)} \neq \underline{\phi}_{(j)}$ implies $p_{(ij)} = p_{(ji)} = 0$. ◁

Obviously if we rearrange the states of the system to make

$$\underline{\phi} = \mathrm{Diag}(\phi_1 I_{n_{(1)}}, \phi_2 I_{n_{(2)}}, \ldots, \phi_s I_{n_{(s)}}) \tag{5.3}$$

where $\phi_1, \phi_2, \ldots, \phi_s$ are all different, then P will become a block-diagonal matrix

$$P = \mathrm{Diag}(P_1, P_2, \ldots, P_s) \tag{5.4}$$

where P_i are $n_{(i)} \times n_{(i)}$. Hence the following lemma is immediate, which generalizes Lemma 2.2.

Lemma 5.1 P is pdd if and only if $P > 0$ and

$$\forall \zeta \in \mathrm{I\!R}^n, \;\; \underline{\phi}(\zeta)^* P \underline{\phi}(\zeta) \leq \zeta^* P \zeta \tag{5.5}$$

for all $\underline{\phi} \in \mathcal{OL}^n(\mathrm{I\!R})$ having a compatible structure with P. $\triangleleft$

Proof. [IF] is obvious since all $\phi I_n \in \mathcal{OL}^n(\mathrm{I\!R})$ must have compatible structures with P so Lemma 2.2 applies. [ONLY IF] follows easily if we rearrange $\underline{\phi}$ and P to become (5.3) and (5.4) and partition ζ accordingly, then (5.5) becomes

$$\forall \zeta \in \mathrm{I\!R}^n, \;\; \sum_{i=1}^{s} (\phi_i(\zeta_i)^* P_i \phi_i(\zeta_i) - \zeta_i^* P_i \zeta_i) \leq 0$$

$\blacksquare$

So given $\underline{\phi} \in \mathcal{OL}^n(\mathrm{I\!R})$, the basic tool of analysis and synthesis is the set of pdd matrices $\{P\}$ that have compatible structures with $\underline{\phi}$, which have already taken into account that some $\underline{\phi}_{(i)}$ may be identical. Certainly if ϕ_i in (5.3) is the identity for some i, the corresponding P_i only needs to be positive definite but not necessarily pdd. But to avoid further complicating the nomenclature we shall still simply call P a pdd matrix. Also we shall still use the terms "dd-stable", "dd-gain" etc. although P is now restricted to having a compatible structure with $\underline{\phi}$. So for example the general $\underline{\phi}$-system $\left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right)_{\underline{\phi}}$ is said to be dd-stable if there exists a pdd P having a compatible structure with $\underline{\phi}$ such that

$$A^* P A - P < 0$$

and if such a P exists, then all $\underline{\phi}$-systems $\left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right)_{\underline{\phi}}$ with $\underline{\phi}$ having a compatible structure with P are dd-stable.

Note that given the structure of $\underline{\phi}$, the additional requirement that P have a compatible structure with $\underline{\phi}$ does not change the computational tractability of the LMIs. With Lemma 5.1 in hands, results for the ϕ-system extend to the $\underline{\phi}$-system naturally. To be a little bit more careful, the key steps are outlined below. Obviously, results on the stability and the ℓ_2-ℓ_2 induced gain can be easily extended to the $\underline{\phi}$-system. To conclude the stability of the cascaded system, note that if $P = \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix}$ (P pdd) and $\underline{\phi} = \begin{bmatrix} \underline{\phi}_1 & \\ & \underline{\phi}_2 \end{bmatrix}$ have compatible structures, then both P_{11} and $P_{22}^{\mho}$ have a compatible structure with $\underline{\phi}_1$ ($P_{22}^{\mho} = P_{11} - P_{12}P_{22}^{-1}P_{21}$ and $P_{12(i\alpha)}(P_{22}^{-1})_{(\alpha\beta)}P_{21(\beta j)} \neq 0$ implies $\underline{\phi}_{1(i)} = \underline{\phi}_{2(\alpha)} = \underline{\phi}_{2(\beta)} = \underline{\phi}_{1(j)}$), hence Lemma 2.8 will continue to hold.

For the parametrization of all dd-stabilizing output-feedback controllers, the key step is to obtain an extension of Lemma 2.9:

Lemma 5.2 Let $\underline{\phi}_{(i)}, \underline{\phi}_{3(i)} \in \mathcal{OL}(\mathbb{R})$. The system described by

$$\begin{bmatrix} (x_1 \pm x_2)^+ \\ x_2^+ \\ x_3^+ \end{bmatrix} = \begin{bmatrix} A_{11} & 0 & 0 \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix} \begin{bmatrix} \underline{\phi}(x_1) \pm \underline{\phi}(x_2) \\ \underline{\phi}(x_2) \\ \underline{\phi}_3(x_3) \end{bmatrix} \tag{5.6}$$

or equivalently

$$\begin{bmatrix} x_1^+ \\ x_2^+ \\ x_3^+ \end{bmatrix} = \begin{bmatrix} A_{11} \mp A_{21} & \pm A_{11} - A_{21} \mp A_{22} & \mp A_{23} \\ A_{21} & \pm A_{21} + A_{22} & A_{23} \\ A_{31} & \pm A_{31} + A_{32} & A_{33} \end{bmatrix} \begin{bmatrix} \underline{\phi}(x_1) \\ \underline{\phi}(x_2) \\ \underline{\phi}_3(x_3) \end{bmatrix} \tag{5.7}$$

is dd-stable if A_{11} is diagonally stable and $\begin{bmatrix} A_{22} & A_{23} \\ A_{32} & A_{33} \end{bmatrix}$ is dd-stable. $\triangleleft$

Proof. The key point is to note that if $\begin{bmatrix} P_{22} & P_{23} \\ P_{23}^* & P_{33} \end{bmatrix}$ and $\begin{bmatrix} \underline{\phi} & \\ & \underline{\phi}_3 \end{bmatrix}$ have compatible structures, then $\begin{bmatrix} P_{11} & \pm P_{11} & 0 \\ \pm P_{11} & P_{11} + P_{22} & P_{23} \\ 0 & P_{23}^* & P_{33} \end{bmatrix}$ and $\begin{bmatrix} \underline{\phi} & & \\ & \underline{\phi} & \\ & & \underline{\phi}_3 \end{bmatrix}$ have compatible structures too. ■

Lemma 5.2 also tells us that G_{DF} and P_{DF} (and hence J_{DF}) must have the same nonlinearity $\underline{\phi}$, but the nonlinearity of the dd-stable parameter Q ($\underline{\phi}_3$) is "free" as long as only the stability is concerned. It will affect the structure of the closed-loop nonlinearity, however. $\hat{J}_{\mathrm{DF}}$ should also have the same nonlinearity $\underline{\phi}$ as J_{DF} so that $J_{\mathrm{DF}} \star \hat{J}_{\mathrm{DF}} \simeq \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}$. The same applies to OE and OF as well. The LMIs to find F and L are basically intact, with only the additional restriction on the structure of their solutions.

For the model reduction part, results on the observability function, controllability function and the main Proposition 4.5 all extend in a trivial way. Note also that $\begin{bmatrix} P^{-1}+Q & P^{-1}-Q \\ P^{-1}-Q & P^{-1}+Q \end{bmatrix}$ has a compatible structure with $\begin{bmatrix} \underline{\phi} & \\ & \underline{\phi} \end{bmatrix}$ if and only if Q, P^{-1} have a compatible structure with $\underline{\phi}$ if and only if among all T satisfying (4.6), (4.7) there exists one that has a compatible structure with $\underline{\phi}$.

In summary, all the main results presented in this dissertation for the ϕ-system can extend to the $\underline{\phi}$-system, and hence Kaszkurewicz and Bhaya's diagonal approach is greatly enhanced. In fact, it is not hard to see that the theory in the all-$\underline{\phi}_{(i)}$-independent case is rather complete: the difficulties of the ϕ-system in finding a F such that $A + B_2F$ is dd-stable and dilating a matrix to remain pdd are removed automatically. So, when all $\underline{\phi}_{(i)}$ are independent, we know how to estimate the stability and the ℓ_2-ℓ_2 induced gain of the $\underline{\phi}$-system, how to compare whether two $\underline{\phi}$-systems are close, how to parametrize all dd-stabilizing[1] output-feedback controllers, how to do closed-loop performance synthesis, and how to do model reduction by discarding partial less-important nonlinearities. In other words, the whole theory for the LFT uncertain system with the *linear* uncertainty set $\mathcal{B}(\triangle_{I_n,I_0})$ can be adapted to the *nonlinear* $\underline{\phi}$-system provided that $\underline{\phi} \in \mathcal{OL}^n(\mathrm{I\!R})$ ($\mathcal{OL}^n(\ell_{2e})$ as a matter of fact) and all $\underline{\phi}_{(i)}$ are (treated as, at least) independent.

However, the theory for the all-$\underline{\phi}_{(i)}$-independent $\underline{\phi}$-system being rather complete does not imply that it is necessarily effective. For example, if all $\underline{\phi}_{(i)}$ are independent, doing model reduction by discarding any one of them usually results in a significant error, as has been seen in Section 4.2. When all $\underline{\phi}_{(i)}$ are identical, intuitively we would expect that the role of the $n-r$ eliminated nonlinearities can possibly be shared by the remaining r ones since they are the same nonlinear function after all. That is why we have tried to study the model reduction of the ϕ-system although it is much more difficult. In fact, the limitation of the pdd approach to the general $\underline{\phi}$-system (rather than the ϕ-system) should not be overlooked. Saying that two nonlinearities $\underline{\phi}_{(i)}$, $\underline{\phi}_{(j)}$ are independent is in fact much stronger than saying that they are simply different. To realize this, we begin with recalling the following example in Section 1.2.2, where we would like to study the error between two $\underline{\phi}$-systems $G := \left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right)_{\underline{\phi}}$ and

[1]not diagonally stabilizing! however. This is obvious since diagonal stability allows the controller to vanish and therefore a diagonally stabilizing controller simply means that the plant is originally diagonally stable.

$$\tilde{G} := \left(\begin{array}{c|c} \tilde{A} & \tilde{B} \\ \hline \tilde{C} & \tilde{D} \end{array}\right)_{\underline{\phi}} :$$

$$E := G - \tilde{G} = \left(\begin{array}{cc|c} \tilde{A} & 0 & \tilde{B} \\ 0 & A & B \\ \hline -\tilde{C} & C & -\tilde{D} + D \end{array}\right)_{\begin{bmatrix} \underline{\phi} & \\ & \underline{\phi} \end{bmatrix}}$$

The diagonal gain of E does not reflect the error since the two $\underline{\phi}$ are treated as independent. In fact, the diagonal gain of E is at least the diagonal gains of $G - \tilde{D}$ and $D - \tilde{G}$ since one of the $\underline{\phi}$ vanishes does not change the diagonal gain. The dd-gain, on the other hand, takes into account that the two $\underline{\phi}$ are the same. But if $\underline{\phi}_{(i)}$ are all independent, it is still at least the dd-gains of

$$\left(\begin{array}{cc|c} \tilde{a}_{(ii)} & 0 & \tilde{B}_i \\ 0 & a_{(ii)} & B_i \\ \hline -\tilde{C}_i & C_i & -\tilde{D} + D \end{array}\right)_{\begin{bmatrix} \underline{\phi}_{(i)} & \\ & \underline{\phi}_{(i)} \end{bmatrix}}, \quad i = 1, 2, \ldots, n$$

where B_i, $\tilde{B}_i$ are respectively the i-th row of B, $\tilde{B}$ and C_i, $\tilde{C}_i$ are respectively the i-th column of C, $\tilde{C}$. Similarly for a $\underline{\phi}$-system with $\underline{\phi} = \begin{bmatrix} \phi_1 I_{n_{(1)}} & \\ & \phi_2 I_{n_{(2)}} \end{bmatrix}$, the dd-gain, for example, is not only a bound of the ℓ_2-ℓ_2 induced gain when ϕ_1 and ϕ_2 are different, but actually a bound of the ℓ_2-ℓ_2 induced gain when we fix ϕ_1 and let ϕ_2 run through the whole set $\mathcal{OL}$ (in particular the 0 and the $-\phi_2$ in it). So the independency of ϕ_1, ϕ_2 should be interpreted in this way. More generally, to guarantee that the pdd approach will work well, the general $\underline{\phi}$-system must be an *uncertain* system in which at most one nonlinearity is precisely known, and the others must be regarded as arbitrary $\mathcal{OL}$ uncertainties. If two or more of the nonlinearities in the $\underline{\phi}$-system are fixed by design, even when they are totally different, they must cooperate or compete in some way by nature which is out of the scope of the pdd approach and we can hardly imagine that removing or negating any of them will not change the behaviour of the nonlinear system significantly. The diagonal approach cannot recognize that two nonlinearities are identical, so all the nonlinearities must be regarded as independent uncertainties, while the pdd approach can *only* tell whether two nonlinearities are the same or different, so it can account for the interaction between *identical* nonlinearities but not between *different* nonlinearities, and it is effective only when the interaction between different nonlinearities are completely uncertain, or in other words there is at most one precisely known nonlinearity. If the general $\underline{\phi}$-system is in fact not an uncertain system but a nonlinear system with multiple known nonlinearities, the pdd approach is expected to be conservative: Sometimes it can give satisfactory results, but not all the time. An interesting example is given in Section 5.1.3.

5.1.2 Systems with ϕ and λ separate

Recall that in the LFT uncertain system, it is typical to let $\delta_1 = \lambda$ and the other δ_i, Δ_i may not relate to λ directly. For the ϕ-system, it is also possible to separate $\phi \circ \lambda$ into $\begin{bmatrix} \lambda & \\ & \phi \end{bmatrix}$ and write

$$\begin{bmatrix} x^+ \\ y \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} \phi(x) \\ u \end{bmatrix}$$

as

$$\begin{bmatrix} x^+ \\ \xi \\ y \end{bmatrix} = \begin{bmatrix} 0 & A & B \\ I & 0 & 0 \\ 0 & C & D \end{bmatrix} \begin{bmatrix} x \\ \phi(\xi) \\ u \end{bmatrix}$$

Lemma 5.3 A is dd-stable (w.r.t. the structure of ϕ) if and only if $\begin{bmatrix} 0 & A \\ I & 0 \end{bmatrix}$ is dd-stable (w.r.t. the structure of $\begin{bmatrix} \lambda & \\ & \phi \end{bmatrix}$). $\triangleleft$

Proof. Note that

$$\begin{bmatrix} 0 & A \\ I & 0 \end{bmatrix}^* \begin{bmatrix} P_1 & 0 \\ 0 & P_2 \end{bmatrix} \begin{bmatrix} 0 & A \\ I & 0 \end{bmatrix} - \begin{bmatrix} P_1 & 0 \\ 0 & P_2 \end{bmatrix} = \begin{bmatrix} P_2 - P_1 & 0 \\ 0 & A^* P_1 A - P_2 \end{bmatrix}$$

Hence $\begin{bmatrix} 0 & A \\ I & 0 \end{bmatrix}$ is dd-stable if and only if

$$\exists P_1 \text{ positive definite}, \ \ \exists P_2 \text{ pdd}, \ \ A^* P_1 A < P_2 < P_1$$

which is simply equivalent to

$$\exists P_2 \text{ pdd}, \ \ A^* P_2 A < P_2$$

i.e. A is dd-stable. (Cf. Remark 2.1.) ■

Proposition 5.4 If A is dd-stable (w.r.t. the structure of $\begin{bmatrix} \lambda I_{n_x} & \\ & \phi I_{n_\xi} \end{bmatrix}$), then $\forall \phi \in \mathcal{OL}(\mathbb{R})$, (the origin of) the system

$$\begin{bmatrix} x^+ \\ \xi \\ \hdashline y \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} x \\ \phi(\xi) \\ \hdashline u \end{bmatrix} \tag{5.8}$$

is globally asymptotically stable. $\triangleleft$

Proof. The non-trivial part of the proof is to show that if A is dd-stable then the solution (x, ξ, y) to the system (5.8) exists for all $(x_{[0]}, u)$. We defer this part until Section 5.2 in which the proof is given in the general $\phi : \ell_{2e} \to \ell_{2e}$ setting. Now if (5.8) is solvable then A dd-stable implies that there exist P_1 positive definite, P_2 pdd such that

$$A^* \begin{bmatrix} P_1 & 0 \\ 0 & P_2 \end{bmatrix} A - \begin{bmatrix} P_1 & 0 \\ 0 & P_2 \end{bmatrix} \leq \begin{bmatrix} -I & 0 \\ 0 & 0 \end{bmatrix}$$

Pre- and post-multiplying both sides by $\begin{bmatrix} x_{[k]} \\ \phi(\xi_{[k]}) \end{bmatrix}^*$ and respectively $\begin{bmatrix} x_{[k]} \\ \phi(\xi_{[k]}) \end{bmatrix}$ we get

$$x^*_{[k+1]} P_1 x_{[k+1]} - x^*_{[k]} P_1 x_{[k]} + \xi^*_{[k]} P_2 \xi_{[k]} - \phi(\xi_{[k]})^* P_2 \phi(\xi_{[k]}) \leq -x^*_{[k]} x_{[k]}$$

At last Lemma 2.2 is applied to conclude that $V(\zeta) := \zeta^* P_1 \zeta$ is a Lyapunov function. ∎

Clearly the extension is similar for dd-gain and other results. The advantage of separating λ and ϕ is that this more general format allows ϕ to appear at other locations in the state-space description of the system. For example, a linear system with input nonlinearity

$$\begin{bmatrix} x^+ \\ y \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} x \\ \phi(u) \end{bmatrix}$$

may be written as

$$\begin{bmatrix} x^+ \\ \xi \\ y \end{bmatrix} = \begin{bmatrix} A & B & 0 \\ 0 & 0 & I \\ C & D & 0 \end{bmatrix} \begin{bmatrix} x \\ \phi(\xi) \\ u \end{bmatrix}$$

and a linear system with output nonlinearity

$$\begin{bmatrix} x^+ \\ y \end{bmatrix} = \begin{bmatrix} I & \\ & \phi \end{bmatrix} \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} x \\ u \end{bmatrix}$$

may be written as

$$\begin{bmatrix} x^+ \\ \xi \\ y \end{bmatrix} = \begin{bmatrix} A & 0 & B \\ C & 0 & D \\ 0 & I & 0 \end{bmatrix} \begin{bmatrix} x \\ \phi(\xi) \\ u \end{bmatrix}$$

Although we may systematically write any state-space description with nonlinearity ϕ into the form (5.8), its practicality is in fact not very satisfactory since Proposition 5.4 is conservative in the sense that A dd-stable is not only sufficient for the stability of

(5.8) but also sufficient for the stability of

$$\begin{bmatrix} x^+ \\ \xi^+ \\ \hdashline y \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} x \\ \phi(\xi) \\ \hdashline u \end{bmatrix} \tag{5.9}$$

according to Proposition 2.6, which means an one-step delay in ξ does not matter. It is reasonable to suspect that such an implication is too strong, and it may (quite probably) happen that our original system (5.8) is actually stable, but A is far from dd-stable. In fact, it may even happen that A is *not linearly stable.*

Example 5.1 Consider the *linear* system

$$\begin{bmatrix} x^+ \\ \xi_1 \\ \xi_2 \end{bmatrix} = \begin{bmatrix} \alpha & 0 & -\alpha \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x \\ \xi_1 \\ \xi_2 \end{bmatrix} \tag{5.10}$$

where $x_{[k]}$, $\xi_{1[k]}$, $\xi_{2[k]}$, α are all scalars. Then $\begin{bmatrix} \alpha & 0 & -\alpha \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$ is linearly stable[2] if and only if $-\frac{1}{\sqrt{2}} < \alpha < \frac{1}{\sqrt{2}}$. However, (5.10) is in fact equivalent to

$$x^+ = 0 \tag{5.11}$$

which is certainly linearly stable. So for example, the LFT approach will be too conservative to conclude the stability of the uncertain *linear* system

$$\begin{bmatrix} x^+ \\ \xi_1 \\ \xi_2 \end{bmatrix} = \begin{bmatrix} \alpha & 0 & -\alpha \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x \\ (\cos\theta)\cdot\xi_1 \\ (\cos\theta)\cdot\xi_2 \end{bmatrix} \tag{5.12}$$

where $\theta \in \ell_{2e}$ is an uncertainty and $\frac{1}{\sqrt{2}} < |\alpha| < 1$, while on the other hand (5.12) is equivalent to

$$x^+ = \alpha(1 - \cos^2\theta)\cdot x \tag{5.13}$$

which is stable. Clearly the idea is that while the system without delay is stable for all α, introducing delay in $\begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix}$ will de-stabilize the system if $|\alpha|$ is sufficiently large, even when $\theta_{[k]}$ are their nominal values 0 for all k. $\otimes$

The observation in Example 5.1 is not surprising at all if we remember that each δ_i in the LFT uncertain system is in fact a $\ell_2 \to \ell_2$ operator with norm ≤ 1, and as $\|\delta_i \circ \lambda\| \leq \|\delta_i\|\|\lambda\| = \|\delta_i\|$, the stability concluded by the LFT approach will remain unchanged if each δ_i is subject to an arbitrary and independent amount of delay. So

[2] using Jury stability test, for example. (Cf. [59, §3.9].)

writing (5.13) as (5.12), which has effectively decomposed $((1-\cos^2\theta)\cdot)\circ\lambda$ into λ and $(\cos\theta\,\cdot)$, implies that the tolerance must be increased from a single arbitrary amount of delay to *two independent* arbitrary amounts of delays, which is severe. The ϕ-system suffers from the same problem, since as we shall see in Section 5.2 the pdd approach in fact allows ϕ to be dynamic: $\ell_{2e}\to\ell_{2e}$, and $\phi\in\mathcal{OL}(\ell_{2e})$ implies $\phi\circ\lambda\in\mathcal{OL}(\ell_{2e})$. Lemma 5.3 is one of those rare cases that such time delay does not matter since

$$\begin{bmatrix} x^+ \\ \xi \end{bmatrix} = \begin{bmatrix} 0 & A \\ I & 0 \end{bmatrix} \begin{bmatrix} x \\ \phi(\xi) \end{bmatrix}$$

is essentially

$$x_{[k+1]} = A\phi(x_{[k]})$$

and

$$\begin{bmatrix} x^+ \\ \xi^+ \end{bmatrix} = \begin{bmatrix} 0 & A \\ I & 0 \end{bmatrix} \begin{bmatrix} x \\ \phi(\xi) \end{bmatrix}$$

is essentially

$$x_{[k+2]} = A\phi(x_{[k]})$$

so one system is stable if and only if the other, but for a general A-matrix in (5.8) such an “if and only if” result should not be expected.

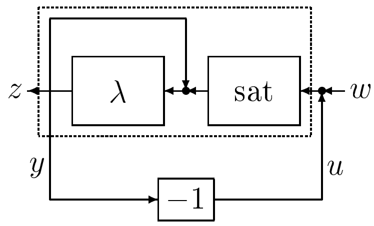
In summary, although separating $\phi\circ\lambda$ in the ϕ-system into $\begin{bmatrix} \lambda & \\ & \phi \end{bmatrix}$ is surely harmless in view of Lemma 5.3, and this more general format allows ϕ to appear at other locations in the state-space description of the system, the gain of doing so is in fact not as large as it first looks. However, when no other effective methods are available to handle such nonlinear systems, the pdd approach is still worth considering as an alternative.

5.1.3 An example

In this section we borrow an example from [69, §8.5] to demonstrate some interesting aspects in applying the results in previous sections. The system is

$$G(\mathrm{sat}) \;\;:\;\; x^+ = x + \mathrm{sat}(w+u); \quad z = y = x \tag{5.14}$$

with unity negative feedback $u=-y$, i.e.



where the dashed box, $G(\text{sat})$, is basically a "summer" with input saturation. $x_{[k]}$, $w_{[k]}$, $u_{[k]}$, $z_{[k]}$, $y_{[k]}$ are all scalars. Naturally the system will be denoted by $\mathcal{F}_l(G(\text{sat}), -1)$. The ℓ_2-ℓ_2 induced gain (from w to z) is known to be $\sqrt{2}$[69, §8.5]. Here we first examine what ℓ_2-ℓ_2 induced gain can be concluded from the pdd approach. (The power of the pdd approach, like the LFT approach to the uncertain linear system, lies in its ability to handle complex systems where analytical solution to the stability or ℓ_2-ℓ_2 induced gain problem is usually difficult. So the purpose of presenting this example is *not* to demonstrate the power of the pdd approach. Indeed this particular system $\mathcal{F}_l(G(\text{sat}), -1)$ can be shown to admit no quadratic Lyapunov/storage function on which the pdd approach is based. As we have said, this example is for the demonstration of some interesting natures in the analysis and synthesis issues only.)

Since sat $\in \mathcal{OL}(\mathbb{R})$, $\mathcal{F}_l(G(\text{sat}), -1)$ may be written as a system with sat and λ separate:

$$\begin{bmatrix} x^+ \\ \xi \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} x \\ \phi(\xi) \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} w; \quad z = x$$

where ϕ is intended to be sat. This will not work, however, since $\begin{bmatrix} 1 & 1 \\ -1 & 0 \end{bmatrix}$ is not linearly stable (both eigenvalues are on the unit circle). Moreover, note that

$$G(\phi) \quad : \quad x^+ = x + \phi(w+u); \quad z = y = x \tag{5.15}$$

in fact has two precisely known "nonlinearities": one identity on x and one ϕ on $(w+u)$, so we know from our discussion at the end of Section 5.1.1 that the pdd approach should not be expected to work well. In particular, it does not distinguish between $\phi = \text{sat}$ and $\phi =$ minus identity, and clearly

$$\mathcal{F}_l(G(-1), -1) \quad : \quad x^+ = x - (w - x) = 2x - w; \quad z = x$$

is unstable. So instead of $G(\phi)$ we study

$$G_a(\phi) \quad : \quad x^+ = a\phi(x/a) + \phi(w+u); \quad z = y = x \tag{5.16}$$

One interpretation is that the state is assumed to be subject to the same kind of saturation too, but with a linear range a times that of the input, and we shall find out how $\mathcal{F}_l(G_a(\phi), -1)$ behaves when $a \gg 1$. In this way "both" nonlinearities (on x and on $(w+u)$) change signs simultaneously, and we observe that ϕ changing from sat to $-$sat does not change the ℓ_2-ℓ_2 induced gain.

We first summarize the results as follows: Suppose $a \geq 1$. Then

- **[a1]** $\mathcal{F}_l(G_a(\phi), -1)$ is stable for all $\phi \in \mathcal{OL}(\mathbb{R})$ if and only if $a < 2$;
- **[a2]** $\mathcal{F}_l(G_a(\phi), -1)$ is stable for all a and indeed has a ℓ_2-ℓ_2 induced gain $\leq a$ for all *monotonically increasing* $\phi \in \mathcal{OL}(\mathbb{R})$.

Both mean that we can hardly use $\mathcal{F}_l(G_a(\phi), -1)$ to approximate $\mathcal{F}_l(G(\mathrm{sat}), -1)$ by letting $a \to \infty$. One simple reason is that the class of ϕ covered by the pdd approach is large although ϕ is only intended to be sat. Indeed both results are tight w.r.t. the class of ϕ, as illustrated below, and we can hardly do better unless we know how to further shrink the class.

To obtain **[a1]**, we write $\mathcal{F}_l(G_a(\phi), -1)$ as

$$\begin{bmatrix} x^+/a \\ \xi^+ \\ z^+ \end{bmatrix} = \begin{bmatrix} 1 & 1/a & 0 \\ -a & -1 & 1 \\ a & 1 & 0 \end{bmatrix} \begin{bmatrix} \phi(x/a) \\ \phi(\xi) \\ w^+ \end{bmatrix} \tag{5.17}$$

where $\xi := w + u = w - x$ and $z^+ = x^+$, so that all the nonlinearities are unified to $\phi \circ \lambda$ in order to minimize the conservativeness of the pdd approach. (5.17) may then be regarded as a ϕ-system

$$\mathcal{F}_l(G_{a+}(\phi), -1) = \left(\begin{array}{cc|c} 1 & 1/a & 0 \\ -a & -1 & 1 \\ \hline a & 1 & 0 \end{array} \right)_\phi \tag{5.18}$$

with input w^+ and output z^+. It is easy to check that $G_a(\phi) \simeq G_{a+}(\phi)$ and hence $\mathcal{F}_l(G_a(\phi), -1)$ and $\mathcal{F}_l(G_{a+}(\phi), -1)$ have the same ℓ_2-ℓ_2 induced gain[3]. The result is that $\mathcal{F}_l(G_{a+}(\phi), -1)$ is dd-stable if and only if $\frac{1}{2} < |a| < 2$. When $a \geq 2$, consider $\phi \in \mathcal{OL}(\mathrm{I\!R})$ such that $\phi(1) = 1$ and $\phi(a) = 0$, then a is an equilibrium of both $\mathcal{F}_l(G(\phi), -1)$ and $\mathcal{F}_l(G_a(\phi), -1)$ (i.e. $x_{[0]} = a$ and $w_{[k]} = 0$ for all k imply $x_{[k]} = a$ for all k) and therefore 0 is not GAS. Note that this ϕ is not monotonically increasing.

To obtain **[a2]**, the key idea is to observe that an odd, 1-Lipschitz and monotonically increasing $\mathrm{I\!R} \to \mathrm{I\!R}$ function ϕ can be written as $\frac{\bullet + \hat{\phi}(\bullet)}{2}$, $\hat{\phi} \in \mathcal{OL}(\mathrm{I\!R})$. So for example

$$\begin{aligned} a\phi(x/a) + \phi(w+u) &= a\left(\frac{x/a + \hat{\phi}(x/a)}{2}\right) + \frac{w + u + \hat{\phi}(w+u)}{2} \\ &= \frac{1}{2}x + \frac{a}{2}\hat{\phi}(x/a) + \frac{1}{2}(w+u) + \frac{1}{2}\hat{\phi}(w+u) \end{aligned}$$

[3] provided that $\phi \in \mathcal{OL}(\mathrm{I\!R})$. In fact this is the only place that we ever need $\phi \in \mathcal{OL}(\mathrm{I\!R})$. For a general $\phi \in \mathcal{OL}(\ell_{2e})$, let $\tilde{\phi}(\bullet) := \lambda\phi((\bullet)^+)$, then $\tilde{\phi} \in \mathcal{OL}(\ell_{2e})$ and $G_a(\phi) \simeq G_{a+}(\tilde{\phi})$. If $\phi \in \mathcal{OL}(\mathrm{I\!R})$, $\phi = \tilde{\phi}$ and hence $\mathcal{F}_l(G_a(\phi), -1)$ and $\mathcal{F}_l(G_{a+}(\phi), -1)$ have the same ℓ_2-ℓ_2 induced gain, otherwise the former $\geq$ the latter in general. Nevertheless $\mathcal{F}_l(G_{a+}(\phi), -1)$ and $\mathcal{F}_l(G_{a+}(\tilde{\phi}), -1)$ have the same dd-gain, so for a general $\phi \in \mathcal{OL}(\ell_{2e})$ the ℓ_2-ℓ_2 induced gain of $\mathcal{F}_l(G_a(\phi), -1)$ is bounded between the ℓ_2-ℓ_2 induced gain and the dd-gain of $\mathcal{F}_l(G_{a+}(\phi), -1)$, and therefore the requirement $\phi \in \mathcal{OL}(\mathrm{I\!R})$ is not very crucial.

and hence $G_a(\phi)$ may be written as

$$\hat{G}_a(\hat{\phi}) \quad : \quad \begin{bmatrix} x^+ \\ x/a \\ \xi \\ z = y \end{bmatrix} = \begin{bmatrix} 1/2 & a/2 & 1/2 & 1/2 \\ 1/a & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ \hat{\phi}(x/a) \\ \hat{\phi}(\xi) \\ w + u \end{bmatrix} \tag{5.19}$$

which is a system with $\hat{\phi}$ and λ separate. This time it can be checked that $\mathcal{F}_l(\hat{G}_a(\hat{\phi}), -1)$ is dd-stable for all nonzero values of a and indeed has a dd-gain $= a$ when $a \geq 1$. To show that the dd-gain is tight, consider

$$\frac{s + \hat{\phi}(s)}{2} = \phi(s) := \begin{cases} 0 & \text{if} \quad |s| < 1 \\ s - \operatorname{sgn}(s) & \text{if} \quad |s| \geq 1 \end{cases} \tag{5.20}$$

Then the input $w = (a + 1, -1, +1, -1, +1, \ldots)$ will produce an output $z = x = (0, a, -a, a, -a, \ldots)$. Hence the ℓ_2-ℓ_2 induced gain of $\mathcal{F}_l(\hat{G}_a(\hat{\phi}), -1)$ for this $\hat{\phi}$ is at least a.

At this stage we may conclude that the pdd approach is conservative w.r.t. the particular choice $\phi =$ sat but is tight w.r.t. the whole class of ϕ. Certainly it may be argued that the analysis results provided by the pdd approach is robust to unforeseen variations in the nonlinearity ϕ. For example, we shall see that if ϕ is a saturation but the width of the linear region is varying with time, then the ℓ_2-ℓ_2 induced gain may be substantially larger than $\sqrt{2}$, but can never exceed the dd-gain since the pdd approach actually copes with $\phi \in \mathcal{OL}(\ell_{2e})$ and therefore such time-variations will be taken care of, and so are those delays that may occur with the saturation nonlinearity. The pdd approach can also handle ϕ whose slopes increase rather than decrease outside the (positive) linear range, e.g. the one in (5.20). However, if unforeseen variations do exist but can be measured in real time, then we can in fact use the materials in Section 3.2 to do gain-scheduling, i.e. instead of simply use the unity negative feedback we design a ϕ-controller[4] $K : y \mapsto u$. Note that when there is no input nonlinearity, i.e. when $\phi =$ the identity, $K = -1$ is indeed the optimal controller, but otherwise the performance of $K = -1$ may deteriorate to **[a1]** and **[a2]**. For example, **[a1]** says that when $K = -1$, the linear range of the state must be less than 2 times the linear range of the input in order to guarantee a stable closed-loop for all $\phi \in \mathcal{OL}(\mathbb{R})$. Naturally one may suspect that it is not only because ϕ but also because a does not appear in the controller $K = -1$. So we are going to design an a-dependent controller K (possibly dynamic) to see whether the closed-loop $\mathcal{F}_l(G_a(\phi), K)$ will have a better performance.

The synthesis results corresponding to **[a1]** and **[a2]** are now summarized as follows: Suppose $a \geq 1$. Then we obtain

[s1] an a-dependent constant-feedback controller K such that $\mathcal{F}_l(G_a(\phi), K)$ is stable for all a and indeed has a ℓ_2-ℓ_2 induced gain $\leq a$ for all $\phi \in \mathcal{OL}(\mathbb{R})$;

[4]Since $G_a(\phi)$ is not exactly a ϕ-system, the controller is not exactly a ϕ-controller but a controller with ϕ and λ separate (with n_λ turns out to be 0).

[s2] an a-dependent gain-scheduling controller K such that $\mathcal{F}_l(G_a(\phi), K)$ has a ℓ_2-ℓ_2 induced gain $\leq \frac{a}{\sqrt{2a-1}} + \epsilon$ for all monotonically increasing $\phi \in \mathcal{OL}(\mathrm{I\!R})$, where ϵ is a small non-negative number according to the numerical results.

As illustrated below, the gain upper bound a in **[s1]** is tight, and if $\mathcal{OL}(\mathrm{I\!R})$ is replaced by $\mathcal{OL}(\ell_{2e})$ in **[s2]**, $\frac{a}{\sqrt{2a-1}}$ is also tight, and furthermore no control u can further reduce this bound and hence the gain-scheduling controller K in **[s2]** is in fact almost optimal.

To obtain **[s1]**, consider

$$G_{a+}(\phi) := \left(\begin{array}{cc|cc} 1 & 1/a & 0 & 0 \\ 0 & 0 & 1 & 1 \\ \hline a & 1 & 0 & 0 \\ a & 1 & 0 & 0 \end{array}\right)_{\phi} \tag{5.21}$$

Since $G_a(\phi) \simeq G_{a+}(\phi)$, designing a K for $G_a(\phi)$ is equivalent to designing a K for $G_{a+}(\phi)$. The synthesis procedure is based on the materials in Section 3.2: We are minimizing γ subject to

$$\begin{bmatrix} C & D \end{bmatrix}_{\perp} \left(M^* \begin{bmatrix} X & 0 \\ 0 & I_p \end{bmatrix} M - \begin{bmatrix} X & 0 \\ 0 & \gamma^2 I_m \end{bmatrix} \right) \begin{bmatrix} C & D \end{bmatrix}_{\perp}^* < 0 \tag{5.22}$$

$$\begin{bmatrix} B \\ D \end{bmatrix}_{\perp}^* \left(M \begin{bmatrix} Y & 0 \\ 0 & \gamma^{-2} I_m \end{bmatrix} M^* - \begin{bmatrix} Y & 0 \\ 0 & I_p \end{bmatrix} \right) \begin{bmatrix} B \\ D \end{bmatrix}_{\perp} < 0 \tag{5.23}$$

$$\begin{bmatrix} X & X_{12} \\ X_{12}^* & X_{22} \end{bmatrix} = \begin{bmatrix} Y & Y_{12} \\ Y_{12}^* & Y_{22} \end{bmatrix}^{-1} \text{ is pdd} \tag{5.24}$$

where $M = \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \left[\begin{array}{cc|c} 1 & 1/a & 0 \\ 0 & 0 & 1 \\ \hline a & 1 & 0 \end{array}\right]$. Recall that we have not yet solved (5.24) in general. Two necessary conditions of (5.24) are X pdd and $\begin{bmatrix} X & I \\ I & Y \end{bmatrix} \geq 0$, while a sufficient condition is Y^{-1} pdd and $X - Y^{-1}$ non-negative diagonal (Lemma 3.10). Since Y is only 2×2, Y^{-1} is pdd if and only if Y is. So we tackle the problem as follows: We first minimize γ subject to (5.22), (5.23) together with X pdd, $\begin{bmatrix} X & I \\ I & Y \end{bmatrix} \geq 0$, and Y pdd; i.e. we only lay $X - Y^{-1}$ non-negative diagonal aside. We keep the Y and then minimize γ again subject to (5.22) together with $X - Y^{-1}$ non-negative diagonal. Of course this time the γ would be larger but if it is not substantially larger we will accept the solution. In fact, in this particular example it turns out that Y^{-1} pdd rather than $X - Y^{-1}$ non-negative diagonal is the more crucial part of the sufficient condition of

(5.24). $X - Y^{-1}$ may be chosen as 0 in the second step without substantially increasing γ, i.e. K may be chosen as static.

Having obtained X and Y, and then $\begin{bmatrix} X & X_{12} \\ X_{12}^* & X_{22} \end{bmatrix}$ using Lemma 3.10, we deduce $S_K := \begin{bmatrix} D_K & C_K \\ B_K & A_K \end{bmatrix}$ using Lemma 3.9 and hence the desired controller K. Finally we get

$$K = -1/a \tag{5.25}$$

Both the dd-gain and the linear gain of

$$\mathcal{F}_l(G_{a+}(\phi), K) = \left(\begin{array}{cc|c} 1 & 1/a & 0 \\ -1 & -1/a & 1 \\ \hline a & 1 & 0 \end{array}\right)_{\phi} \tag{5.26}$$

$= a$ when $a \geq 1$. Hence the dd-gain is tight. So using $K : u = (-1/a)y$ instead of $K : u = -y$ the closed-loop system $\mathcal{F}_l(G_a(\phi), K)$ will become stable for $a \geq 2$ even when ϕ is not necessarily monotonically increasing, and indeed have a ℓ_2-ℓ_2 induced gain $\leq a$.

[**s2**] can be obtained by applying a similar synthesis procedure to $\hat{G}_a(\hat{\phi})$, where

$$M = \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \left[\begin{array}{ccc|c} 1/2 & a/2 & 1/2 & 1/2 \\ 1/a & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ \hline 1 & 0 & 0 & 0 \end{array}\right]$$

and X, Y are restricted to block-diagonal matrices: one 1×1 block corresponding to λ and one 2×2 block corresponding to $\hat{\phi}$. Since the second diagonal block is only 2×2, once again Y^{-1} is pdd if and only if Y is. So in the first step of the synthesis procedure the second diagonal blocks of both X and Y are required to be pdd. In the second step it turns out that $X - Y^{-1}$ may be chosen to be diagonal without substantially increasing γ, and furthermore the first diagonal block of $X - Y^{-1}$ may be chosen as 0. That means again the controller K may be chosen as static since the second diagonal block corresponds to no λ.

The numerical result computed using Lemma 3.9 is not very accurate, usually due to the computation of inverses and square roots of ill-conditioned matrices, and therefore not reported in details. However, having applied the synthesis procedure to $\hat{G}_a(\hat{\phi})$ for different values of a, we are able to deduce that the controller should be given by

$$S_K := \begin{bmatrix} D_K & C_K \\ B_K & A_K \end{bmatrix} = \left[\begin{array}{c|cc} -1 & c_1 & c_2 \\ \hline -1/a & 0 & 0 \\ 1 & -c_1 & -c_2 \end{array}\right]$$

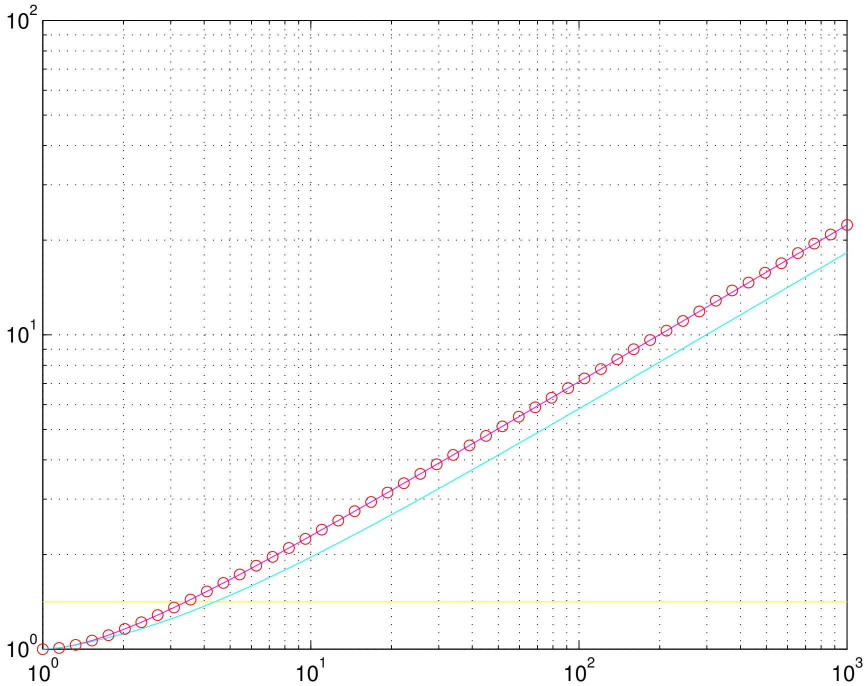


Figure 5.1: The dd-gains (the small circles), $\beta_2(a)$ (the darkest line) and $\beta_1(a)$ (the medium-dark line) for different values of $a \in [1, 1000]$. The lightest line is $\sqrt{2}$.

or equivalently

$$u + c_2\hat{\phi}(u) = -(y + c_1\hat{\phi}(y/a)) \tag{5.27}$$

where $c_2 < 1$ but is very close to 1 and $c_1 \approx c_2 a$. Here we can see an interesting nature of applying the dd-synthesis procedure to a system with ϕ and λ separate: the control law is given implicitly by an equation rather than an input-output function. This flexibility may admit a more complex but powerful control law while on the other hand solving the (nonlinear) equation for the control input u given an admissible y may be computationally expensive when ϕ is complicated. In general the solution u may not be unique but in this particular example the map $K : y \mapsto u$ is well defined since $|c_2| < 1$ implies that $\bullet + c_2\hat{\phi}(\bullet)$ is strictly monotonically increasing and hence is one-one onto $\mathbb{R}$, i.e. invertible on $\mathbb{R}$.

We now choose $c_2 = 0.99$ and $c_1 = c_2 a$. The dd-gain of $\mathcal{F}_l(\hat{G}_a(\hat{\phi}), K)$ as a varies is shown in Figure 5.1 (the small circles). The dd-gain is tight, and indeed K is nearly optimal in the sense that the dd-gain of $\mathcal{F}_l(\hat{G}_a(\hat{\phi}), K)$ is already very close to $\frac{a}{\sqrt{2a-1}}$ (the darkest line) but in fact no control input u satisfying $y_{[0]} = 0 \Rightarrow u_{[0]} = 0$ can ever make the dd-gain of $\hat{G}_a(\hat{\phi})$ from w to z less than $\frac{a}{\sqrt{2a-1}}$. To show this, a $\hat{\phi} \in \mathcal{OL}(\ell_{2e})$ rather than a $\hat{\phi} \in \mathcal{OL}(\mathbb{R})$ is needed. The basic idea of its construction is as follows. Let $\frac{\bullet + \hat{\phi}(\bullet)}{2} = \phi(\bullet) = \boldsymbol{\phi} \cdot \bullet$. If $\boldsymbol{\phi}_{[0]} =$ the identity, then $w_{[0]}$ can bring $x_{[1]}$ to any value (in particular any value $\in [-a, a]$). So effectively we can choose any initial state we like, and the ℓ_2-ℓ_2 induced gain is lower bounded by the (zero-input) initial-state-to-output gain (i.e. $\frac{\sqrt{Q(x_{[0]})}}{\|x_{[0]}\|}$ where Q is the observability function), which can be rather large if $|\boldsymbol{\phi}_{[k]}(\bullet_{[k]})|$, $k > 0$ is bounded and so is the speed of the state's dropping back to 0.

More precisely, let $a \geq 1$ and

$$\boldsymbol{\phi}_{[k]}(s) := \begin{cases} s & \text{if} \quad k = 0 \\ \mathrm{sat}(s) & \text{if} \quad k > 0 \end{cases} \tag{5.28}$$

As said before,

$$w_{[k]} = \begin{cases} a & \text{if} \quad k = 0 \\ 0 & \text{if} \quad k > 0 \end{cases} \tag{5.29}$$

Hence $x_{[1]} = a > 0$, and from $k = 1$ onwards[5]

$$\begin{cases} x_{[k+1]} \geq \min(x_{[k]}, a) - 1 \geq a - k & \text{if} \quad a - k > 0 \\ |x_{[k+1]}| \geq 0 & \text{if} \quad a - k \leq 0 \end{cases}$$

since $|\boldsymbol{\phi}_{[k]}(w_{[k]} + u_{[k]})| = |\boldsymbol{\phi}_{[k]}(u_{[k]})| \leq 1$ for *any* u. So the ℓ_2-ℓ_2 induced gain from w to z is lower bounded by

$$\beta_1(a) := \frac{\sqrt{\sum_{k=0}^{\alpha}(a-k)^2}}{a} = \sqrt{\frac{(\alpha+1)(\alpha(2\alpha+1) + 6a(a-\alpha))}{6a^2}} \tag{5.30}$$

where α is the largest integer $\leq a$. This lower bound $\beta_1(a)$, which is plotted on the same Figure 5.1 (the medium-dark line), is already close to the dd-gain, and in particular $\beta_1(a) \to \infty$ as $a \to \infty$. Since the above argument may also apply to $G(\mathrm{sat})$ with $u = -y$, we see from the choice of ϕ in (5.28) that although the ℓ_2-ℓ_2 induced gain of $\mathcal{F}_l(G(\mathrm{sat}), -1)$ is only $\sqrt{2}$, the initial-state-to-output gain is actually unbounded, and the small ℓ_2-ℓ_2 induced gain in fact relies heavily on the difficulty of bringing the state to a large value by the input due to the input saturation. In contrast, the observability function of a dd-stable system is uniformly bounded by a constant times $\|x_{[0]}\|^2$. The unbounded initial-state-to-output gain implies that $\mathcal{F}_l(G(\mathrm{sat}), -1)$ actually admits *no* quadratic Lyapunov or storage function to prove its stability or ℓ_2-ℓ_2 induced gain, and any approach that restricts the Lyapunov/storage function to the *quadratic* type must fail. The pdd approach is such an approach, and therefore its limited results derivable for $\mathcal{F}_l(G(\mathrm{sat}), -1)$, strictly speaking, should not be explained as simply as the consequence of its conservativeness w.r.t. the particular choice $\phi = \mathrm{sat}$. (A further discussion on this issue can be found in Appendix E.)

A more sophisticated choice of ϕ can tighten the lower bound from $\beta_1(a)$ to $\frac{a}{\sqrt{2a-1}}$. The idea is that the speed of the state's dropping back to 0 is "optimally" bounded rather than only bounded by 1. Let a and w be as before and

$$\boldsymbol{\phi}_{[k]}(s) := \begin{cases} s & \text{if} \quad x_{[k]} = 0 \\ \frac{x_{[k]}}{a}\mathrm{sat}(s/\frac{x_{[k]}}{a}) & \text{if} \quad x_{[k]} \neq 0 \end{cases} \tag{5.31}$$

Hence $x_{[1]} = a > 0$ again, and from $k = 1$ onwards

$$x_{[k+1]} \geq x_{[k]} - \left|\frac{x_{[k]}}{a}\right| \geq a(1 - 1/a)^k$$

[5] can be easily proved by mathematical induction on k.

since $|\boldsymbol{\phi}_{[k]}(w_{[k]}+u_{[k]})| = |\boldsymbol{\phi}_{[k]}(u_{[k]})| \le \left|\frac{x_{[k]}}{a}\right|$ for *any* u. So the ℓ_2-ℓ_2 induced gain from w to z is lower bounded by

$$\beta_2(a) := \frac{\sqrt{\sum_{k=0}^{\infty}(a(1-1/a)^k)^2}}{a} = \frac{a}{\sqrt{2a-1}} \tag{5.32}$$

The dd-gain of $\mathcal{F}_l(\hat{G}_a(\hat{\phi}), K)$ with our choice of K in (5.27) is very close to this lower bound $\beta_2(a)$. The relative error is typically 10^{-5} when $a \le 10$. When $a > 10$ the computation of the dd-gain becomes less accurate, and the relative error fluctuates a lot but is scarcely larger than 10^{-2}. Since the performance bound $\beta_2(a)$ is independent of the control $u_{[k]}$, $k > 0$, we can conclude that K is nearly optimal w.r.t. the whole class of $\hat{\phi}$, i.e. $\mathcal{OL}(\ell_{2e})$. At last, we point out that the choice of ϕ in (5.31) may be replaced by

$$\boldsymbol{\phi}_{[k]}(s) := \begin{cases} s & \text{if} \quad k = 0 \\ (1-1/a)^{k-1}\mathrm{sat}(s/(1-1/a)^{k-1}) & \text{if} \quad k > 0 \end{cases} \tag{5.33}$$

without changing the conclusion.

The interpretation of the nearly optimal controller K is as follows. Writing $\frac{\bullet+\hat{\phi}(\bullet)}{2} = \phi(\bullet)$, we see that the nearly optimal controller law (5.27) is in fact a perturbed version of

$$\phi(u) = -a\phi(y/a) \tag{5.34}$$

and the perturbation is made to invert the ϕ on u. When ϕ is a saturation function, the interpretation of K is rather obvious: $u_{[k]} = -y_{[k]}$ if $|y_{[k]}| \le 1$, otherwise $u_{[k]}$ is as negative as possible so that $w_{[k]}$ needs to be very large in order to keep or further increase the state/output. When ϕ is as in (5.20), K would look very different: $u_{[k]} = -(y_{[k]} - (a-1)\mathrm{sgn}(y_{[k]}))$ if $|y_{[k]}| \ge a$, otherwise $|u_{[k]}|$ is simply kept small (≤ 1) which means that control is essentially unnecessary if $x_{[k]}$ is in the dead-zone. In either case the control strategy is very natural, and it is interesting to note that such different looking strategies can arise from the same control equation (5.27), which is indeed applicable to all $\hat{\phi} \in \mathcal{OL}(\ell_{2e})$.

In this example we have demonstrated the technique that a monotonically increasing $\phi \in \mathcal{OL}(\mathbb{R})$ can be represented by $\frac{\bullet+\hat{\phi}(\bullet)}{2}$, $\hat{\phi} \in \mathcal{OL}(\mathbb{R})$. Indeed it is easy to see that any $\phi \in \mathcal{OL}(\mathbb{R})$ with a slope bounded in some sector can be represented in a similar way, i.e. $\mu \bullet + \nu\hat{\phi}(\bullet)$, $\hat{\phi} \in \mathcal{OL}(\mathbb{R})$ is a function in $\mathcal{OL}(\mathbb{R})$ with a slope bounded in $[\mu-\nu, \mu+\nu]$. So for example, one may ask what upper bound one can get for the ℓ_2-ℓ_2 induced gain of $\mathcal{F}_l(G_a(\phi), -1)$ if the slope of ϕ is not in $[0, 1]$ but in $[\delta, 1]$ for some $\delta > 0$. The answer for $a = 100$ is given in Figure 5.2 in which we see that the dd-gain drops from a rapidly as δ increases: it drops to about one-third when δ is merely 0.02, one-fifth when $\delta = 0.04$ and less than one-tenth when $\delta = 0.1$. Hence we observe that much difference will be made if ϕ is not really flat somewhere somehow.

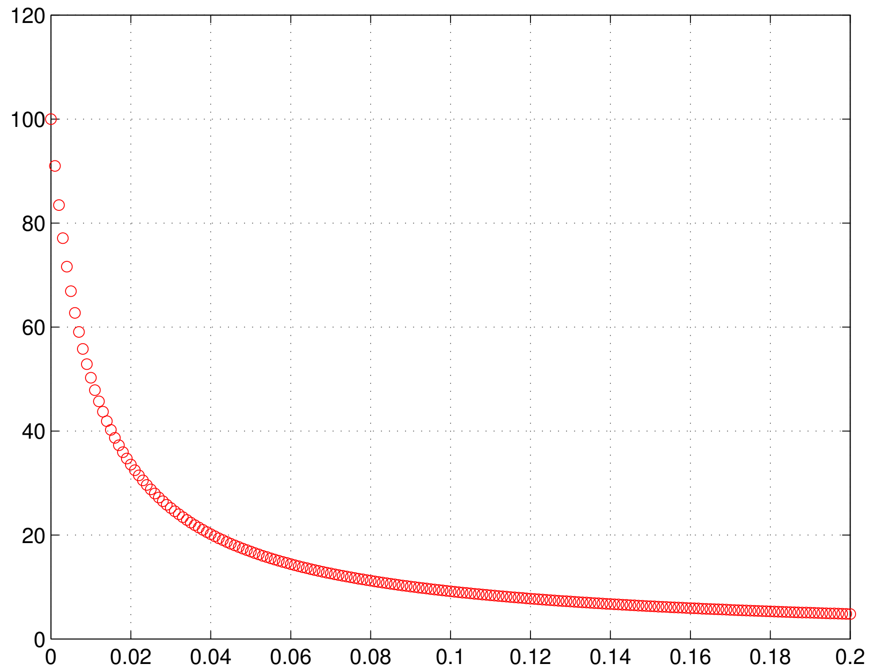


Figure 5.2: The dd-gains (the small circles) for different values of $\delta \in [0, 0.2]$ when $a = 100$.

5.2 ϕ: Dynamic case

The objective of this section is to outline how to generalize all the results from the previous chapters to the $\phi : \ell_{2e} \to \ell_{2e}$ setting. Let

$$\mathcal{OL}(\ell_{2e}) := \left\{ \phi : \ell_{2e} \to \ell_{2e} \text{ causal } \;\middle|\; \forall s, t \in \ell_2,\ \ \|\phi(s) + \phi(t)\| \leq \|s + t\| \right\} \tag{5.35}$$

i.e. $\phi \in \mathcal{OL}(\ell_{2e})$ if and only if ϕ is causal and $\phi\ |_{\ell_2}$ is an odd, 1-Lipschitz, $\ell_2 \to \ell_2$ operator. (ϕ is said to be causal if $\forall K \in \mathbb{Z}_+$, $\mathcal{T}_K \circ \phi = \mathcal{T}_K \circ \phi \circ \mathcal{T}_K$ where $\mathcal{T}_K$ is the truncation operator: $(x_{[0]}, x_{[1]}, \ldots) \mapsto (x_{[0]}, x_{[1]}, \ldots, x_{[K-1]}, x_{[K]}, 0, 0, \ldots)$. The causality is assumed to guarantee the existence and uniqueness of solution $x \in \ell_{2e}^n$, $y \in \ell_{2e}^p$ to the difference equation (1.2) for all $x_{[0]} \in \mathbb{R}^n$, $u \in \ell_{2e}^m$.) It is easy to see that if $\phi \in \mathcal{OL}(\ell_{2e})$, then

$$\forall s, t \in \ell_{2e},\ \ \forall K \in \mathbb{Z}_+,\ \ \|\phi(s) + \phi(t)\|_K \leq \|s + t\|_K \tag{5.36}$$

Hence ϕ is indeed odd and

$$\forall s, t \in \ell_{2e},\ \ \forall K \in \mathbb{Z}_+,\ \ \|\phi(s) - \phi(t)\|_K \leq \|s - t\|_K \tag{5.37}$$

The first and the most important step of the generalization is to obtain the corresponding characterization of pdd matrices.

Lemma 5.5 *P is pdd if and only if $P > 0$ and*

$$\forall \phi \in \mathcal{OL}(\ell_{2e}),\ \ \forall x \in \ell_{2e}^n,\ \ \forall K \in \mathbb{Z}_+,\ \ \sum_{k=0}^{K} \phi(x)_{[k]}^* P \phi(x)_{[k]} \leq \sum_{k=0}^{K} x_{[k]}^* P x_{[k]} \tag{5.38}$$

◁

Proof. The [IF] part is rather obvious since $\forall \boldsymbol{\phi}_{[0]} \in \mathcal{OL}(\mathbb{R})$, $\forall x_{[0]} \in \mathbb{R}^n$, $\phi = ((\boldsymbol{\phi}_{[0]}, 0, 0, 0, 0, \ldots) \cdot) \in \mathcal{OL}(\ell_{2e})$ and $x = (x_{[0]}, 0, 0, 0, 0, \ldots) \in \ell_{2e}$, then (5.38) implies (2.4) and hence Lemma 2.2 applies. For the [ONLY IF] part, $\phi \in \mathcal{OL}(\ell_{2e})$ means that $\forall s, t \in \ell_{2e}$, $\forall K \in \mathbb{Z}_+$,

$$\sum_{k=0}^{K} (\phi(s)_{[k]} \pm \phi(t)_{[k]})^2 \leq \sum_{k=0}^{K} (s_{[k]} \pm t_{[k]})^2 \tag{5.39}$$

which implies

$$2\left|\sum_{k=0}^{K} \left(\phi(s)_{[k]}\phi(t)_{[k]} - s_{[k]}t_{[k]}\right)\right| \leq \\ -\left(\sum_{k=0}^{K} \left(\phi(s)_{[k]}^2 - s_{[k]}^2\right) + \sum_{k=0}^{K} \left(\phi(t)_{[k]}^2 - t_{[k]}^2\right)\right) \tag{5.40}$$

Now $\forall x \in \ell_{2e}^n$, let

$$\alpha_{(ij)} = \sum_{k=0}^{K} \left(\phi(x)_{[k](i)}\phi(x)_{[k](j)} - x_{[k](i)}x_{[k](j)}\right)$$

Then (5.40) implies that $\alpha_{(ij)}$ satisfies (2.2) of Lemma 2.1 and therefore

$$\begin{aligned} & \sum_{k=0}^{K} \left(\phi(x)_{[k]}^* P \phi(x)_{[k]} - x_{[k]}^* P x_{[k]}\right) \\ = & \sum_{k=0}^{K} \left(\sum_{i,j} \phi(x)_{[k](i)} p_{(ij)} \phi(x)_{[k](j)} - \sum_{i,j} x_{[k](i)} p_{(ij)} x_{[k](j)}\right) \\ = & \sum_{i,j} p_{(ij)} \sum_{k=0}^{K} \left(\phi(x)_{[k](i)}\phi(x)_{[k](j)} - x_{[k](i)}x_{[k](j)}\right) \\ = & \sum_{i,j} p_{(ij)}\alpha_{(ij)} \\ \leq & \ 0 \end{aligned}$$

■

With this characterization of pdd matrices, the ℓ_2-stability instead of the GAS of the system would be considered.

Definition 5.2 The system $\left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right)$ is said to be ℓ_2-stable if $x \in \ell_2^n$ for all $x_{[0]} \in \mathbb{R}^n$ and zero input. ◁

An alternative definition analogous to the one which is more commonly used in the LFT framework of uncertain linear systems is given in the proposition below.

Proposition 5.6 The ϕ-system $\left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right)_{\phi}$, $\phi \in \mathcal{OL}(\ell_{2e})$ is ℓ_2-stable if and only if $I - \lambda A\phi\ |_{\ell_2^n} : \ell_2^n \to \ell_2^n$ is invertible. ◁

We are now in a position to present the stability result for the general $\phi : \ell_{2e} \to \ell_{2e}$ system:

Proposition 5.7 If A is dd-stable, then $\forall \phi \in \mathcal{OL}(\ell_{2e})$, the ϕ-system $\left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right)_\phi$ is ℓ_2-stable. $\triangleleft$

Proof. In the proof of Proposition 2.6 we have already shown that

$$V(x_{[k+1]}) - V(x_{[k]}) \leq -\phi(x)^*_{[k]}\phi(x)_{[k]} - \left(x^*_{[k]}Px_{[k]} - \phi(x)^*_{[k]}P\phi(x)_{[k]}\right)$$

Taking summation from $k = 0$ to any $K \in \mathbb{Z}_+$, we get

$$V(x_{[K+1]}) - V(x_{[0]}) \leq -\|\phi(x)\|^2_K - \sum_{k=0}^{K}\left(x^*_{[k]}Px_{[k]} - \phi(x)^*_{[k]}P\phi(x)_{[k]}\right)$$

which implies

$$\|\phi(x)\|^2_K \leq V(x_{[0]}) + \sum_{k=0}^{K}\left(\phi(x)^*_{[k]}P\phi(x)_{[k]} - x^*_{[k]}Px_{[k]}\right) \leq V(x_{[0]})$$

since $V(x_{[K+1]}) \geq 0$ and using Lemma 5.5. Hence $\phi(x) \in \ell_2^n$ by letting $K \to \infty$ and $x \in \ell_2^n$ since $x^+ = A\phi(x)$. ■

The generalizations of the results for the ℓ_2-ℓ_2 induced gain, the observability function and the controllability function are next stated. All are straight-forward.

Proposition 5.8 $\forall \phi \in \mathcal{OL}(\ell_{2e})$, the ℓ_2-ℓ_2 induced gain of the ϕ-system $\left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right)_\phi$ is upper bounded by its dd-gain. More precisely, if γ is the dd-gain of $\begin{bmatrix} A & B \\ C & D \end{bmatrix}$, then $x_{[0]} = 0$ implies that

$$\forall K \in \mathbb{Z}_+, \quad \|y\|^2_K := \sum_{k=0}^{K} y^*_{[k]}y_{[k]} \leq \gamma^2 \sum_{k=0}^{K} u^*_{[k]}u_{[k]} = \gamma^2\|u\|^2_K$$

and hence

$$\|y\|^2 := \lim_{K\to\infty} \|y\|^2_K \leq \gamma^2\|u\|^2$$

if $u \in \ell_2^m$. $\triangleleft$

Proof. The proof of Proposition 2.11 can be easily modified to get

$$\forall K \in \mathbb{Z}_+, \quad \|y\|^2_K \leq \gamma^2\|u\|^2_K + \sum_{k=0}^{K}\left(\phi(x)^*_{[k]}P\phi(x)_{[k]} - x^*_{[k]}Px_{[k]}\right)$$

without applying Lemma 2.2. Now we may apply Lemma 5.5 to obtain the result.

■

Proposition 5.9 Given a ϕ-system $\left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right)_\phi$, $\phi \in \mathcal{OL}(\ell_{2e})$. Suppose that there exists a pdd Q such that

$$A^*QA - Q + C^*C \leq 0 \tag{5.41}$$

Then $u = 0$ implies that

$$\|y\|^2 := \sum_{k=0}^{\infty} y^*_{[k]} y_{[k]} \leq x^*_{[0]} Q x_{[0]}$$

$\triangleleft$

Proof. Similarly the proof of Proposition 4.1 can be easily modified to get

$$\forall K \in \mathbb{Z}_+, \;\; \|y\|_K^2 \leq V(x_{[0]}) + \sum_{k=0}^{K} \left(\phi(x)^*_{[k]} Q \phi(x)_{[k]} - x^*_{[k]} Q x_{[k]}\right)$$

without applying Lemma 2.2. Now we may apply Lemma 5.5 to get

$$\forall K \in \mathbb{Z}_+, \;\; \|y\|_K^2 \leq V(x_{[0]})$$

and let $K \to \infty$ to obtain the result. ∎

Proposition 5.10 Given a ϕ-system $\left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right)_\phi$, $\phi \in \mathcal{OL}(\ell_{2e})$. Suppose that there exists P such that P^{-1} is pdd and

$$APA^* - P + BB^* \leq 0 \tag{5.42}$$

Then $x_{[0]} = 0$ implies that

$$\forall K-1 \in \mathbb{Z}_+, \;\; \|u\|_{K-1}^2 := \sum_{k=0}^{K-1} u^*_{[k]} u_{[k]} \geq x^*_{[K]} P^{-1} x_{[K]}$$

$\triangleleft$

Proof. The proof of Proposition 4.2 can be easily modified to get

$$\forall K-1 \in \mathbb{Z}_+, \;\; \|u\|_{K-1}^2 + \sum_{k=0}^{K-1} \left(\phi(x)^*_{[k]} P^{-1} \phi(x)_{[k]} - x^*_{[k]} P^{-1} x_{[k]}\right) \geq V(x_{[K]})$$

without applying Lemma 2.2. Now we may apply Lemma 5.5 to obtain the result.

∎

In this way the connections between the stability, the ℓ_2-ℓ_2 induced gain etc. of a ϕ-system, $\phi \in \mathcal{OL}(\ell_{2e})$ and its dd-stability, dd-gain etc. are established, and all other results are based on the dd-stability, dd-gain etc. directly. In other words, the extension of the results from the $\phi : \mathbb{R} \to \mathbb{R}$ to the $\phi : \ell_{2e} \to \ell_{2e}$ setting is complete.

Remark 5.1 It should be all right if ϕ is memoryless, i.e. $\phi = (\boldsymbol{\phi} \cdot)$ with $\boldsymbol{\phi}_{[k]} \in \mathcal{OL}(\mathrm{I\!R})$; but if ϕ has memory, then strictly speaking $x_{[K]}$, $K > 0$ is *not* the state of the ϕ-system: The concept of the state is that it sums up what has happened in the past, and therefore the future behaviour of the system depends only on the current state but not directly on the past. Clearly if ϕ has memory then the behaviour of the ϕ-system for $k \geq K$ depends not only on $x_{[K]}$ but also on $x_{[0]}, x_{[1]}, \ldots, x_{[K-1]}$. Since $x_{[K]}$ is not the state of the ϕ-system, the interpretation of the observability function is somewhat difficult: The output energy $\sum_{k=K}^{\infty} y_{[k]}^* y_{[k]}$ is not necessarily bounded by $x_{[K]}^* Q x_{[K]}$ if the input for $k \geq K$ is zero. Instead

$$\sum_{k=K}^{\infty} y_{[k]}^* y_{[k]} \leq x_{[K]}^* Q x_{[K]} - \sum_{k=0}^{K-1} \left(\phi(x)_{[k]}^* Q \phi(x)_{[k]} - x_{[k]}^* Q x_{[k]} \right) \tag{5.43}$$

On the other hand, we see from the proof of Proposition 5.10 that

$$\sum_{k=0}^{K-1} u_{[k]}^* u_{[k]} \geq x_{[K]}^* P^{-1} x_{[K]} - \sum_{k=0}^{K-1} \left(\phi(x)_{[k]}^* P^{-1} \phi(x)_{[k]} - x_{[k]}^* P^{-1} x_{[k]} \right) \tag{5.44}$$

Hence

$$\begin{aligned} \sum_{k=K}^{\infty} y_{[k]}^* y_{[k]} - \sum_{k=0}^{K-1} u_{[k]}^* u_{[k]} \ &\leq \ x_{[K]}^* (Q - P^{-1}) x_{[K]} - \\ &\sum_{k=0}^{K-1} \left(\phi(x)_{[k]}^* (Q - P^{-1}) \phi(x)_{[k]} - x_{[k]}^* (Q - P^{-1}) x_{[k]} \right) \end{aligned} \tag{5.45}$$

So if $\zeta^*(Q - P^{-1})\zeta = 0$ for all $\zeta \in$ the coordinate subspace associated with $\bar{S}_{n-r}$ as required in the model reduction results of Chapter 4, then $\forall K \in \mathbb{Z}_+$, $x_{[0]}, x_{[1]}, \ldots, x_{[K]} \in$ the coordinate subspace associated with $\bar{S}_{n-r}$ and $u_{[k]} = 0$ for $k \geq K$ imply that

$$\sum_{k=K}^{\infty} y_{[k]}^* y_{[k]} \leq \sum_{k=0}^{K-1} u_{[k]}^* u_{[k]}$$

i.e. the Hankel-norm restricted to $\bar{S}_{n-r}$ will be ≤ 1. $\otimes$

At last we show that it is also possible to consider systems with ϕ and λ separate when $\phi \in \mathcal{OL}(\ell_{2e})$.

Proposition 5.11 If A is dd-stable (w.r.t. the structure of $\begin{bmatrix} \lambda I_{n_x} & \\ & \phi I_{n_\xi} \end{bmatrix}$), then $\forall \phi \in \mathcal{OL}(\ell_{2e})$, the system

$$\begin{bmatrix} x^+ \\ \xi \\ y \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} x \\ \phi(\xi) \\ u \end{bmatrix} \tag{5.46}$$

is ℓ_2-stable. $\triangleleft$

To prove Proposition 5.11, we first establish the existence of solution to (5.46):

Lemma 5.12 Suppose that A is dd-stable (w.r.t. the structure of $\begin{bmatrix} \lambda I_{n_x} & \\ & \phi I_{n_\xi} \end{bmatrix}$) and $\phi \in \mathcal{OL}(\ell_{2e})$. Then the solution (x, ξ, y) to the system (5.46) exists for all $(x_{[0]}, u)$. $\triangleleft$

Proof. The lemma is obvious if $\phi = \lambda$, so we may assume $\phi \neq \lambda$ and hence the pdd solution P to the (strict) Lyapunov inequality is block-diagonal $\begin{bmatrix} P_1 & 0 \\ 0 & P_2 \end{bmatrix}$.

Let

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \left[\begin{array}{cc:c} A_{11} & A_{12} & B_1 \\ A_{21} & A_{22} & B_2 \\ \hdashline C_1 & C_2 & D \end{array}\right] \tag{5.47}$$

Then A dd-stable (w.r.t. the structure of $\begin{bmatrix} \lambda I_{n_x} & \\ & \phi I_{n_\xi} \end{bmatrix}$) implies that A_{22} is dd-stable (w.r.t. the structure of ϕI_{n_ξ}). In particular

$$\exists \alpha < 1 \quad \text{s.t.} \quad A_{22}^* P_2 A_{22} - \alpha P_2 \leq 0 \tag{5.48}$$

We want to show that $\forall K \in \mathbb{Z}_+$, given $\xi_{[0]}, \xi_{[1]}, \ldots, \xi_{[K-1]} \in \mathbb{R}^n$, $\forall \eta \in \mathbb{R}^n$, $\exists \xi_{[K]} \in \mathbb{R}^n$ that solves

$$\xi_{[K]} = A_{22} \phi(\xi)_{[K]} + \eta \tag{5.49}$$

(Note that we have used that ϕ is causal.) Clearly this is sufficient for the existence of solution to (5.46).

For convenience let $\boldsymbol{\phi}$ denote the associated $\mathbb{R}^n \to \mathbb{R}^n$ operator: $\xi_{[K]} \mapsto \phi(\xi)_{[K]}$ when $\xi_{[0]}, \xi_{[1]}, \ldots, \xi_{[K-1]}$ are fixed. Now for a $\zeta \in \mathbb{R}^n$ let $\|\zeta\|_{P_2} := \sqrt{\zeta^* P_2 \zeta}$. (5.48) and Lemma 5.5 (since $\phi \in \mathcal{OL}(\ell_{2e})$) imply that

$$\exists \alpha < 1 \quad \text{s.t.} \quad \forall \xi \in \ell_{2e}^n, \ \forall K \in \mathbb{Z}_+, \ \sum_{k=0}^{K} \|A_{22}\phi(\xi)_{[k]}\|_{P_2} \leq \alpha \sum_{k=0}^{K} \|\xi_{[k]}\|_{P_2}$$

Then $\forall \xi_{[K]} \in \mathbb{R}^n$,

$$\frac{\|A_{22}\boldsymbol{\phi}(\xi_{[K]})\|_{P_2}}{\|\xi_{[K]}\|_{P_2}} = \frac{\|A_{22}\phi(\xi)_{[K]}\|_{P_2}}{\|\xi_{[K]}\|_{P_2}} \leq \frac{\sum_{k=0}^{K} \|A_{22}\phi(\xi)_{[k]}\|_{P_2}}{\|\xi_{[K]}\|_{P_2}} \leq \frac{\alpha \sum_{k=0}^{K} \|\xi_{[k]}\|_{P_2}}{\|\xi_{[K]}\|_{P_2}}$$

Hence

$$\limsup_{\|\xi_{[K]}\|_{P_2} \to \infty} \frac{\|A_{22}\boldsymbol{\phi}(\xi_{[K]})\|_{P_2}}{\|\xi_{[K]}\|_{P_2}} \leq \alpha < 1$$

and it follows from [41, Thm. 4.3.15] (which in turn is a consequence of Schauder's fixed point theorem) that $I - A_{22}\boldsymbol{\phi}$ maps $\mathbb{R}^n$ *onto* $\mathbb{R}^n$. This completes the proof.

It may be of interest to know that if $\left\{ \boldsymbol{\phi}(\xi_{[K]}) \;\middle|\; \xi_{[K]} \in \mathbb{R}^n \right\}$ is bounded then the solution to (5.49) and hence (5.46) exists even when A is not dd-stable. This is Schauder's fixed point theorem. On the other hand if A_{22} is diagonally stable and ϕ is memoryless, then Banach's contraction principle tells us that the solution to (5.49) and hence (5.46) exists and is furthermore unique. Note that in general the solution to (5.49) is not unique if A_{22} is only dd-stable even when $\phi \in \mathcal{OL}(\mathbb{R})$. An example is

$$A_{22} = \begin{bmatrix} 0.25 & -1.39 \\ 0.84 & -0.68 \end{bmatrix}$$

i.e. the one given in Example 2.1 that is dd-stable but not diagonally stable. Let $\nu = A_{22}\mu$ such that $\|\mu\| = 1$ and $\|\nu\| = \|A_{22}\| > 1$. It turns out that for this A_{22}, $|\nu_{(i)}| > |\mu_{(i)}|$ and hence it is easy to construct $\phi \in \mathcal{OL}(\mathbb{R})$, $\zeta_1, \zeta_2, \eta \in \mathbb{R}^2$ such that $\zeta_1 - \zeta_2 = A_{22}(\phi(\zeta_1) - \phi(\zeta_2))$ and $\eta = \zeta_i - A_{22}\phi(\zeta_i)$. ■

Now we are ready to prove Proposition 5.11.

Proof of Proposition 5.11. The proposition is obvious if $\phi = \lambda$, so we may assume $\phi \neq \lambda$ and hence the pdd solution P to the (strict) Lyapunov inequality is block-diagonal $\begin{bmatrix} P_1 & 0 \\ 0 & P_2 \end{bmatrix}$. In the proof of Proposition 5.4 we have already shown that

$$V(x_{[k+1]}) - V(x_{[k]}) + \xi_{[k]}^* P_2 \xi_{[k]} - \phi(\xi)_{[k]}^* P_2 \phi(\xi)_{[k]} \leq -x_{[k]}^* x_{[k]}$$

where $V(\zeta) := \zeta^* P_1 \zeta$. Taking summation from $k = 0$ to any $K \in \mathbb{Z}_+$, we get

$$V(x_{[K+1]}) - V(x_{[0]}) + \sum_{k=0}^{K} \left(\xi_{[k]}^* P_2 \xi_{[k]} - \phi(\xi)_{[k]}^* P_2 \phi(\xi)_{[k]} \right) \leq -\|x\|_K^2$$

which implies

$$\|x\|_K^2 \leq V(x_{[0]}) + \sum_{k=0}^{K} \left(\phi(\xi)_{[k]}^* P_2 \phi(\xi)_{[k]} - \xi_{[k]}^* P_2 \xi_{[k]} \right) \leq V(x_{[0]})$$

since $V(x_{[K+1]}) \geq 0$ and using Lemma 5.5. Hence $x \in \ell_2^n$ by letting $K \to \infty$. ■

Chapter 6

Conclusions

A summary of the results obtained in this dissertation is given below, and by this means its contributions are highlighted. Suggestions for future research are listed at the end.

In this dissertation we have studied a class of nonlinear systems described by upper linear fractional transformations with respect to a diagonal odd 1-Lipschitz nonlinear operator $\underline{\phi} \circ \lambda$. Indeed, from the approach we have adopted it is easy to see that the objective of this dissertation is nothing but to extend the LFT framework for the uncertain linear systems to the nonlinear case. For this purpose a method must be (and has been) developed to recognize the significance of repeated entries on the diagonal of $\underline{\phi}$, and then those repeated entries may be grouped into blocks

$$\underline{\phi} = \mathrm{Diag}(\phi_1 I_{n_{(1)}}, \phi_2 I_{n_{(2)}}, \ldots, \phi_s I_{n_{(s)}})$$

analogous to the repeated-scalar blocks in the uncertainty Δ of the linear systems. Therefore, emphasis of this dissertation has been put on the single-block case $\underline{\phi} = \phi I_n$ and the generalization to the multiple-block case is obvious and outlined in Chapter 5, which is in fact no different from the way people did in the linear case for the uncertainty Δ.

This class of nonlinear systems, called the ϕ-systems by us, also has a connection to a class of recurrent neural networks. This class of recurrent neural networks, according to our approach, will have a globally asymptotically stable origin and therefore is not a very general class of recurrent neural networks. So we expect the impact of the work in this dissertation on the area of neural networks will not be as large as on the LFT framework in control. However, our work opens the chance of bringing certain recurrent neural networks under the light of LFT, which should undoubtedly facilitate the solving of some analysis, synthesis and model reduction problems for those recurrent neural networks. In this aspect the contributions of our pdd approach which allows $\underline{\phi}_{(i)}$ to be identical should be significant.

In the LFT framework, the results for an uncertain linear system are obtained via a quadratic Lyapunov/storage function $V(\zeta) := \zeta^* P \zeta$ where P belongs to some class of positive definite matrices and satisfies certain Linear Matrix Inequalities. To be more specific, the class of positive definite matrices is the set of matrices that commute with Δ:

$$P\Delta = \Delta P \tag{6.1}$$

For a ϕ-system, we cannot ask for a P that commutes with ϕ since ϕ is nonlinear. The set of P that commute with ϕ is too small to do decent analysis for the ϕ-system, not to mention synthesis and model reduction. The central idea of this dissertation is that, instead of (6.1), we start with

$$\forall \zeta \in \mathbb{R}^n, \ \ \phi(\zeta)^* P \phi(\zeta) \leq \zeta^* P \zeta \tag{6.2}$$

or

$$\forall x \in \ell_{2e}^n, \ \ \forall K \in \mathbb{Z}_+, \ \ \sum_{k=0}^{K} \phi(x)_{[k]}^* P \phi(x)_{[k]} \leq \sum_{k=0}^{K} x_{[k]}^* P x_{[k]} \tag{6.3}$$

This weaker requirement will admit a much larger set of P than those that commute with ϕ and hence provide sufficient tools for some analysis, synthesis and model reduction problems. On the other hand, the results derivable from this weaker starting point may also be weaker than their linear counterparts.

All results in this dissertation are based on two important facts established in Chapter 2. The first one is that the class of positive definite matrices that satisfy (6.2) or (6.3) for all $\phi \in \mathcal{OL}(\mathbb{R})$ or $\phi \in \mathcal{OL}(\ell_{2e})$ is exactly the class of positive definite diagonally dominant (pdd) matrices. The second one is that the class of pdd matrices can be specified in LMI whose size only grows reasonably with n. These two facts are respectively the theoretical and computational foundations of the work in this dissertation. With them we have successfully concluded the stability and ℓ_2-ℓ_2 induced gain for the ϕ-system in Chapter 2. It has been shown that a dd-stable ϕ-system need not be diagonally stable, and even if it is, its diagonal gain may be substantially larger than its dd-gain. So the improvement of the pdd approach over the diagonal approach is significant. Besides, comparison between two ϕ-systems or analysis of interconnected ϕ-systems would be impossible without the pdd approach.

The close relation of the pdd approach to the LFT approach to the uncertain linear system, and its easy success in analysis of the ϕ-system, may sometimes shadow the fact that a ϕ-system is a *nonlinear* system after all. It is true that a ϕ-system is less general than a general nonlinear system, and in return it allows us to adopt an algebraic LMI formulation to derive global and easily computable rather than local results which should undoubtedly be counted as a strong point, but if one is then cheated by the outlook of the analysis results and thinks that the ϕ-system is always as benign as a linear system, one must greatly err. For example, the LMI for the stability is

$$A^* P A - P < 0$$

which means that if we define a norm in $\mathbb{R}^n$ as $\| \bullet \|_P := \sqrt{\bullet^* P \bullet}$, then $A\Delta$ (for the uncertain linear system) and $A\phi$ (for the ϕ-system) will have induced norm < 1. In the linear case, this is sufficient to conclude that $A\Delta$ is strictly contractive (i.e. Lipschitz with constant < 1) but in the nonlinear case it is *not* sufficient to conclude that $A\phi$ is strictly contractive: ϕ is contractive in $\| \bullet \|$ but not necessarily in $\| \bullet \|_P$ unless P is diagonal! A consequence of this is that although the state-trajectories of a dd-stable ϕ-system will all tend to 0 when there is no input, they may tend to different places depending on the initial states when there is a nonzero constant input if the system is not diagonally stable! Another (related) typical nonlinear behaviour is that the ℓ_2-ℓ_2 induced gain of a ϕ-system is not necessarily equal to the incremental gain, a topic we have not explicitly touched in this dissertation. The reason is that although the ℓ_2-ℓ_2 induced gain of a ϕ-system is bounded by the dd-gain, the incremental gain can only be shown to be bounded by the diagonal gain, and the diagonal approach is not the main focus of this dissertation.

The objective of raising this point is to give an alternative insight into the synthesis or gain-scheduling results in Chapter 3. There are two main results in Chapter 3: One is the parametrization of all dd-stabilizing output feedback ϕ-controllers and the other is on the synthesis of a ϕ-controller to achieve some closed-loop performance. In the parametrization of all dd-stabilizing ϕ-controllers, both pdd and diagonal solutions to Lyapunov inequalities are involved. One interpretation is that the observer part of the controller is in incremental form, and therefore the diagonal approach rather than the pdd approach has to be used to handle that part of the controller. Computationally the results in Chapter 3 are not totally satisfactory. The parametrization of all dd-stabilizing ϕ-controllers involves a solution to the state-feedback Lyapunov inequality whose inverse is required to be pdd, and the closed-loop performance synthesis involves the dilation of a pdd matrix to remain pdd; both are computationally difficult. However, the worst case is no worse than further restricting the matrices in concern to diagonal for computation.

Chapter 4 is on model reduction, a topic that allows more flexibility than any other topics in this dissertation. The algorithm proposed in Chapter 4 follows the line of conventional balanced truncation of linear systems, and to some extent the procedure is similar, although there is a major conceptual difference: For the linear system we find a balanced realization first and then truncate, but for the ϕ-system there is no balanced realization, and the associated balanced system is only an intermediate product that bridges the original ϕ-system and the reduced-order ϕ-model. It is why we called it model reduction via a balanced system rather than balanced truncation in Chapter 4. It should be realized that our model reduction algorithm is only *a* method to do model reduction for the ϕ-system, which has by no means been proven optimal in any rigorous sense. The general lack of methods in obtaining a reduced-order approximant for the ϕ-system should be understood, and we are further convinced that the difficulty in doing model reduction for the ϕ-system is not only due to the lack of methods, but also because a nonlinear system is rarely reducible in essence, and our model reduction algorithm *is* a method that can identify some of those rare cases.

After the basic ideas of the pdd approach have been presented, several extensions of the results are included in Chapter 5. The first and the most obvious one is extending the results from the single-block case ϕI_n to the multiple-block case $\mathrm{Diag}(\phi_1 I_{n_{(1)}}, \phi_2 I_{n_{(2)}}, \ldots, \phi_s I_{n_{(s)}})$ as said at the beginning. The way we did it was exactly the same as the way the linear uncertainty Δ was treated in the LFT approach, therefore it should be noted that it is a way to handle *uncertainties* rather than *known nonlinearities*. As for the LFT uncertain system where only $\delta_1 = \lambda$ is supposed to be certain, for the $\underline{\phi}$-system there should also only be one certain ϕ_i, otherwise the block-diagonal-P approach may be conservative. The second extension included in Chapter 5 is separating ϕ and λ in the ϕ-system to produce a more general form of state-space description, and the last extension is generalizing the nonlinearity ϕ from static to dynamic which leads us nearer to the answer of when the pdd approach will be necessary as well as sufficient.

Overall speaking, the pdd approach provides rich results for the ϕ-system which are closely integrated together. The analysis part is particularly successful while the synthesis and model reduction part still leave room for further improvement. Since ϕ is only required to belong to $\mathcal{OL}(\mathbb{R})$ or $\mathcal{OL}(\ell_{2e})$, but not any other particular information of ϕ is used, the pdd approach may be conservative when ϕ is precisely known (although by no means more conservative than the diagonal approach), while on the other hand the results given by it are robust under admissible variations in ϕ. This kind of tradeoff is typical.

The contributions of this dissertation may be summarized as follows. For the class of discrete-time nonlinear systems in concern, which may be described by linear fractional transformations with respect to a diagonal odd 1-Lipschitz nonlinear operator and represent a class of recurrent neural networks, we have given a rather comprehensive treatment on stability, ℓ_2-ℓ_2 induced gain, synthesis and model reduction based on finding pdd solutions to certain Linear Matrix Inequalities. In particular, we have

- established a sufficient condition for the global asymptotic stability of the origin;
- given a satisfactory upper bound on the ℓ_2-ℓ_2 induced gain for dd-stable ϕ-systems;
- parametrized all dd-stabilizing ϕ-controllers if the system is dd-stabilizable and diagonally detectable;
- presented a way to do performance synthesis for the $\underline{\phi}$-system if the system is diagonally stabilizable and diagonally detectable, and for the ϕ-system if, occasionally, the matrix dilation problem can be solved;
- proposed a model reduction algorithm (two variants: one involves minimizing the rank of a matrix subject to a Linear Matrix Inequality and the other involves choosing pdd solutions to two Lyapunov inequalities) which can identify some dd-stable ϕ-systems where model reduction with small dd-error is possible and provide such a reduced-order model for them.

It is believed that the contributions made by this dissertation in extending the LFT framework from the uncertain linear systems to the above class of nonlinear systems are significant, and it also makes certain contributions to the study of recurrent neural networks.

At last, we point out several related directions which deserve further investigation.

- The first one, naturally, is the necessity of the pdd approach. A typical thought is that if the set of ϕ being considered is large enough then the pdd approach will become necessary, so the question is how large. From the numerical examples presented in this dissertation it seems that the pdd approach is already rather tight once ϕ is allowed to vary with time, i.e. dynamic but memoryless. So we conjecture that $\mathcal{OL}(\ell_{2e})$ is already close to the set we want, but perhaps the assumption on causality has to be dropped somehow. The answer for the necessity of the diagonal stability to the $\underline{\phi}$-system, on the other hand, can be easily deduced from the LFT approach to the uncertain linear system.

- How to find a matrix whose inverse is pdd to solve a LMI is still unknown for $n \geq 3$, and worth trying although it has been shown that the set of matrices having a pdd inverse is not convex for $n \geq 4$. This problem is connected to the state-feedback problem of the ϕ-system and therefore its solution is important. The non-convexity for $n \geq 4$ denies a LMI formulation[1], but if a sufficiently large subset (larger than the set of diagonal matrices, at least) can be formulated in LMI, it is still very useful.

- The problem of dilating a pdd matrix to remain pdd is another unsolved problem that is directly related to the synthesis of the ϕ-system. Its solution is still under investigation.

- Model reduction is surely a topic which deserves more research efforts, but it is difficult to tell in what direction one should try. Model reduction of a ϕ-system that is not dd-stable is a topic that has not been touched by this dissertation, but is certainly worth looking at.

- The behaviours of the system with ϕ and λ separate have not been totally understood. It seems to be the first step towards an implicit description of the ϕ-system. The stability LMI for the uncertain *linear* system is sufficient for the existence and uniqueness of the solution to the state-space description and therefore they are solved simultaneously with the stability problem, but for the *nonlinear* ϕ-system this may not be the case. For the proof of the existence we have, for convenience, assumed causality which can perhaps be removed, but for the uniqueness we have already shown (by counterexample) that the stability LMI is not sufficient. Further investigation to clarify all these is beneficial.

[1] The non-convexity of the set of matrices having a pdd inverse does *not* imply that the set of matrices having a pdd inverse *and* satisfying some additional LMI is not convex, however. So the overall problem may still admit a LMI formulation.

- If the set $\mathcal{OL}(\ell_{2e})$ is regarded as too large to give tight results on a particular ϕ, an approach totally different from the pdd approach may be needed. The ϕ-dependent approach in Appendix E is such an attempt, but it is much less practical than the pdd approach. At this moment, it is difficult to see that any other approach can provide a similarly comprehensive treatment to the ϕ-system as the pdd approach does, but research in this direction may still be helpful. For example, the ϕ-dependent approach in Appendix E has given an improved (over the linear gain) quadratically provable lower estimate of the ℓ_2-ℓ_2 induced gain of the ϕ-system which may help to check the conservativeness of the pdd approach.

It is clear that the success or failure of the pdd approach depends strongly on our ability to manipulate pdd matrices. Most of the above are directly or indirectly related to this.

Appendix A

Approximation properties of recurrent neural networks

In [72] Sontag demonstrated how to use a σ-system to approximate the general nonlinear system

$$x^+ = f(x,u); \quad y = h(x) \tag{A.1}$$

where $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$, $h : \mathbb{R}^n \to \mathbb{R}^p$ are continuous. The result is reproduced below and we shall explain why it is not directly applicable to our case which led to the development of our own approximation result in Section 1.2.3.

Definition A.1 A function $\phi : \mathbb{R} \to \mathbb{R}$ is called universal if for any closed interval $K \subset \mathbb{R}$, the space of functions

$$\left\{ F_\beta \circ \phi \circ F_\alpha \mid_K \quad \middle| \quad F_\alpha : \mathbb{R} \to \mathbb{R}^l, F_\beta : \mathbb{R}^l \to \mathbb{R} \text{ are affine}, l \in \mathbb{Z}_+ \right\} \tag{A.2}$$

(i.e. $F_\alpha(\bullet) := \zeta_\alpha \bullet + \eta_\alpha$, $\zeta_\alpha, \eta_\alpha \in \mathbb{R}^l$; $F_\beta(\bullet) := \zeta_\beta^* \bullet + \eta_\beta$, $\zeta_\beta \in \mathbb{R}^l$, $\eta_\beta \in \mathbb{R}$) is dense in the space of continuous functions: $K \to \mathbb{R}$ in sup-norm. $\lhd$

As mentioned in [72], any function ϕ that is continuous, nonconstant and bounded is universal[40] and any monotonically increasing[1] function ϕ such that both $\lim_{s\to-\infty} \phi(s)$ and $\lim_{s\to+\infty} \phi(s)$ exist is universal. Hence sgn, sat and sin are all universal. On the other hand, $(\bullet)^r$ is not universal.

If ϕ is universal, then for any compact set $K \subset \mathbb{R}^m$, the set of functions

$$\left\{ F_\beta \circ \phi \circ F_\alpha \mid_K \quad \middle| \quad F_\alpha : \mathbb{R}^m \to \mathbb{R}^l, F_\beta : \mathbb{R}^l \to \mathbb{R}^p \text{ are affine}, l \in \mathbb{Z}_+ \right\} \tag{A.3}$$

is dense in the set of continuous functions: $K \to \mathbb{R}^p$ in sup-norm. The proof is similar to the one given in Section 1.2.3 or see [72] for details.

$F_\beta \circ \phi \circ F_\alpha$ is called a single hidden layer feedforward neural network (1HLN) with m inputs, l hidden neurons, p outputs. Hence a 1HLN can approximate uniformly

[1] nonstrict. Strictly monotonically increasing would be said otherwise.

any given continuous function on any given compact set provided that the number of hidden neurons l is sufficiently large. (This is actually a well-known fact in the neural network community.) Note that F_α, F_β are affine instead of linear as in Section 1.2.3 and therefore bias terms may exist. Now consider the general nonlinear system (A.1). WLOG we may assume $y = Cx$, since we can always introduce an additional state $\hat{x} := h(x)$ such that $\hat{x}^+ = h(x^+) = h(f(x,u))$ and $y = \hat{x}$. A 1HLN is used to approximate $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ to obtain

$$x^+ = T\phi(Ax + Bu + \mu) + \nu; \quad y = Cx \tag{A.4}$$

We may assume $T = \begin{bmatrix} T_1 \\ 0 \end{bmatrix}$ with a suitable change of coordinates, where T_1 has full row rank. Partition x, A, ν, C accordingly into $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$, $\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$, $\begin{bmatrix} \nu_1 \\ \nu_2 \end{bmatrix}$ and $\begin{bmatrix} C_1 & C_2 \end{bmatrix}$ to get

$$\begin{aligned} \begin{bmatrix} x_1^+ \\ x_2^+ \end{bmatrix} &= \begin{bmatrix} T_1\phi(\begin{bmatrix} A_{11} & A_{12} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + Bu + \mu) \\ 0 \end{bmatrix} + \begin{bmatrix} \nu_1 \\ \nu_2 \end{bmatrix} \\ y &= \begin{bmatrix} C_1 & C_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \end{aligned} \tag{A.5}$$

Since T has full row rank (i.e. onto), there exists $\hat{\nu}_1$ that solves $T_1\hat{\nu}_1 = \nu_1$, and if we consider the system

$$\begin{aligned} \begin{bmatrix} \xi_1^+ \\ x_2^+ \end{bmatrix} &= \begin{bmatrix} \phi(\begin{bmatrix} A_{11}T_1 & A_{12} \end{bmatrix} \begin{bmatrix} \xi_1 \\ x_2 \end{bmatrix} + Bu + \mu) \\ 0 \end{bmatrix} + \begin{bmatrix} \hat{\nu}_1 \\ \nu_2 \end{bmatrix} \\ y &= \begin{bmatrix} C_1T_1 & C_2 \end{bmatrix} \begin{bmatrix} \xi_1 \\ x_2 \end{bmatrix} \end{aligned} \tag{A.6}$$

we see that for every $x_{1[0]}$ there exists $\xi_{1[0]}$ such that $x_{1[0]} = T_1\xi_{1[0]}$. Note that $(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, y) = (\begin{bmatrix} T_1\xi_1 \\ x_2 \end{bmatrix}, y)$ solves (A.5) if $(\begin{bmatrix} \xi_1 \\ x_2 \end{bmatrix}, y)$ solves (A.6), and (A.6) is of the form

$$\xi^+ = \phi(\hat{A}\xi + \hat{B}u + \hat{\mu}) + \hat{\nu}; \quad y = \hat{C}\xi \tag{A.7}$$

where $\xi = \begin{bmatrix} \xi_1 \\ x_2 \end{bmatrix}$, $\hat{A} = \begin{bmatrix} A_{11}T_1 & A_{12} \\ 0 & 0 \end{bmatrix}$, $\hat{B} = \begin{bmatrix} B \\ 0 \end{bmatrix}$, $\hat{\mu} = \begin{bmatrix} \mu \\ 0 \end{bmatrix}$, $\hat{\nu} = \begin{bmatrix} \hat{\nu}_1 \\ \nu_2 - \phi(0) \end{bmatrix}$ and $\hat{C} = \begin{bmatrix} C_1T_1 & C_2 \end{bmatrix}$. So it remains to remove the bias terms $\hat{\mu}, \hat{\nu}$ if (A.7) is to become a σ-system.

In [72] Sontag removed the bias terms by picking any $s, t \in \mathbb{R}$ such that $s\phi(t) = 1$ and considering the following σ-system

$$\begin{aligned}
\begin{bmatrix} (\xi - \hat{\nu})^+ \\ \xi_3^+ \end{bmatrix} &= \phi(\begin{bmatrix} \hat{A} & (\hat{A}\hat{\nu} + \hat{\mu})s \\ 0 & ts \end{bmatrix} \begin{bmatrix} \xi - \hat{\nu} \\ \xi_3 \end{bmatrix} + \begin{bmatrix} \hat{B} \\ 0 \end{bmatrix} u) \\
y &= \begin{bmatrix} \hat{C} & 0 \end{bmatrix} \begin{bmatrix} \xi \\ \xi_3 \end{bmatrix}
\end{aligned} \tag{A.8}$$

Then (ξ, y) solves (A.7) with initial condition $\xi_{[0]}$ if and only if $(\begin{bmatrix} \xi \\ 1/s \end{bmatrix}, y)$ solves (A.8) with initial condition $\begin{bmatrix} \xi_{[0]} \\ 1/s \end{bmatrix}$. Hence (A.8) is the desired approximant: the map: $\begin{bmatrix} \xi \\ \xi_3 \end{bmatrix}_{[k]} \mapsto x_{[k]}$ is onto and therefore each trajectory of (A.5) can be simulated by a trajectory of (A.8).

Why Sontag's result is not suitable for us is now explained. His way to remove the bias terms deliberately introduces a new equilibrium $\begin{bmatrix} 0 \\ 1/s \end{bmatrix}$ other than the origin in order to simulate the effect of the bias terms. In other words the origin is *meant to be* not GAS. This contradicts our pdd approach which always tries to globally stabilize the origin of the system. Indeed $s\phi(t) = 1$ implies $|st| \geq 1$ since $|t| \geq |\phi(t)|$ for $\phi \in \mathcal{OL}(\mathbb{R})$, which means that the bottom-right block of the A-matrix in (A.8) is made linearly unstable and hence the origin is in fact locally unstable. Therefore we had to develop our own more-restrictive approximation result in Section 1.2.3.

Remark A.1 The approximation of the dynamic system (A.1) via the approximation of the static map f is based on the argument that fixing a finite period of time, the state-trajectories of the dynamic approximant can be arbitrarily close to the original ones provided that their static maps can be arbitrarily close uniformly on a sufficiently large compact set. Practically, however, the number of hidden neurons l is bounded, so in general the static approximant cannot be arbitrarily close to f, and the best static approximant to f need not result in the best (in the ℓ_2-ℓ_2 sense) dynamic approximant to the system (A.1). In fact, it does happen that a good approximant to the static map may result in a poor approximant to the dynamic system. $\otimes$

Appendix B

Proof of Lemma 2.3

[P1] $\Leftrightarrow$ [P2]

This is Lemma 2.2.

[P1] $\Leftrightarrow$ [P3]

[$\Rightarrow$]

$\phi \in \mathcal{OL}(\mathrm{I\!R})$ means that $\forall s, t \in \mathrm{I\!R}$,

$$|\phi(s) \pm \phi(t)| \leq |s \pm t| \tag{B.1}$$

which implies

$$(s \pm t)(\phi(s) \pm \phi(t)) \leq |s \pm t||\phi(s) \pm \phi(t)| \leq (s \pm t)^2 \tag{B.2}$$

and hence

$$|s(\phi(t) - t) + t(\phi(s) - s)| \leq -(s(\phi(s) - s) + t(\phi(t) - t)) \tag{B.3}$$

Note that

$$\begin{aligned} \zeta^* P\phi(\zeta) - \zeta^* P\zeta &= \sum_{i,j} \zeta_{(i)} p_{(ij)} \phi(\zeta_{(j)}) - \sum_{i,j} \zeta_{(i)} p_{(ij)} \zeta_{(j)} \\ &= \sum_{i,j} p_{(ij)} \zeta_{(i)} (\phi(\zeta_{(j)}) - \zeta_{(j)}) \end{aligned} \tag{B.4}$$

Let $\alpha_{(ij)} = \zeta_{(i)}(\phi(\zeta_{(j)}) - \zeta_{(j)})$. Then (B.3) implies (2.2) and hence Lemma 2.1 applies.

[$\Leftarrow$]

We want to show that if $P > 0$ but not diagonally dominant, then $\exists \phi \in \mathcal{OL}(\mathrm{I\!R})$, $\exists \zeta \in \mathrm{I\!R}^n$ such that

$$\zeta^* P\zeta - \zeta^* P\phi(\zeta) < 0$$

or equivalently

$$(\zeta - \phi(\zeta))^* P\zeta < 0$$

As in the [IF] part of the proof of Lemma 2.2, we may always choose $\phi = \mathrm{sat}$. Since P is not diagonally dominant, $\exists i$, $\exists \epsilon > 0$ such that

$$(1+\epsilon)p_{(ii)} < \sum_{j \neq i} |p_{(ij)}|$$

So we simply let

$$\zeta_{(j)} = \begin{cases} -\mathrm{sgn}(p_{(ij)}) & \text{if} \quad j \neq i \\ 1+\epsilon & \text{if} \quad j = i \end{cases} \tag{B.5}$$

$$\zeta_{(j)} - \phi(\zeta_{(j)}) = \begin{cases} 0 & \text{if} \quad j \neq i \\ \epsilon & \text{if} \quad j = i \end{cases} \tag{B.6}$$

to complete the proof.

[P1] $\Leftrightarrow$ [P4]

[$\Rightarrow$]

In view of **[P3]** it remains to show that $\phi(\zeta)^* P\phi(\zeta) \leq \zeta^* P\phi(\zeta)$. Again $\phi \in \mathcal{OL}(\mathbb{R})$ means that $\forall s, t \in \mathbb{R}$,

$$|\phi(s) \pm \phi(t)| \leq |s \pm t| \tag{B.7}$$

which together with $s \leq t \Rightarrow \phi(s) \leq \phi(t)$ implies that

$$(\phi(s) \pm \phi(t))^2 \leq |s \pm t||\phi(s) \pm \phi(t)| = (s \pm t)(\phi(s) \pm \phi(t)) \tag{B.8}$$

Hence

$$|(\phi(s) - s)\phi(t) + (\phi(t) - t)\phi(s)| \leq -((\phi(s) - s)\phi(s) + (\phi(t) - t)\phi(t)) \tag{B.9}$$

Note that

$$\begin{aligned} \phi(\zeta)^* P\phi(\zeta) - \zeta^* P\phi(\zeta) &= \sum_{i,j} \phi(\zeta_{(i)}) p_{(ij)} \phi(\zeta_{(j)}) - \sum_{i,j} \zeta_{(i)} p_{(ij)} \phi(\zeta_{(j)}) \\ &= \sum_{i,j} p_{(ij)} (\phi(\zeta_{(i)}) - \zeta_{(i)}) \phi(\zeta_{(j)}) \end{aligned} \tag{B.10}$$

Let $\alpha_{(ij)} = (\phi(\zeta_{(i)}) - \zeta_{(i)})\phi(\zeta_{(j)})$. Then (B.9) implies (2.2) and hence Lemma 2.1 applies.

[$\Leftarrow$]

This is easy since the ϕ chosen in the [IF] part of the proof of Lemma 2.2 (or in the [$\Leftarrow$] part of the proof of **[P1]** $\Leftrightarrow$ **[P3]**) is sat which is monotonically increasing, thus applicable to this case.

Appendix C

Proof of Lemma 4.3

Lemma 4.3 is rather standard[74]. The proof below is based on [77, §21.8].

Let

$$\hat{M} := \begin{bmatrix} -\Sigma^{\frac{1}{2}} & \Sigma^{\frac{1}{2}} & 0 \\ \Sigma^{-\frac{1}{2}} & \Sigma^{-\frac{1}{2}} & 0 \\ 0 & 0 & I \end{bmatrix} \begin{bmatrix} \hat{A}_{br} & 0 & \hat{B}_{br}/2 \\ 0 & A_b & B_b/2 \\ -\hat{C}_{br} & C_b & -\hat{D}_{br} + D_b \end{bmatrix} \begin{bmatrix} -\Sigma^{\frac{1}{2}} & \Sigma^{\frac{1}{2}} & 0 \\ \Sigma^{-\frac{1}{2}} & \Sigma^{-\frac{1}{2}} & 0 \\ 0 & 0 & I \end{bmatrix}^{-1} \tag{C.1}$$

Putting (4.9), (4.10) and $\Sigma = \begin{bmatrix} \Sigma_1 & 0 \\ 0 & I \end{bmatrix}$ into (C.1) and multiplying everything out we get

$$\hat{M} = \begin{bmatrix} \Sigma_1^{\frac{1}{2}} A_{b11} \Sigma_1^{-\frac{1}{2}} & \frac{1}{2}\Sigma_1^{\frac{1}{2}} A_{b12} & 0 & \frac{1}{2}\Sigma_1^{\frac{1}{2}} A_{b12} & 0 \\ \frac{1}{2} A_{b21} \Sigma_1^{-\frac{1}{2}} & \frac{1}{2} A_{b22} & \frac{1}{2} A_{b21} \Sigma_1^{\frac{1}{2}} & \frac{1}{2} A_{b22} & \frac{1}{2} B_{b2} \\ 0 & \frac{1}{2}\Sigma_1^{-\frac{1}{2}} A_{b12} & \Sigma_1^{-\frac{1}{2}} A_{b11} \Sigma_1^{\frac{1}{2}} & \frac{1}{2}\Sigma_1^{-\frac{1}{2}} A_{b12} & \Sigma_1^{-\frac{1}{2}} B_{b1} \\ \frac{1}{2} A_{b21} \Sigma_1^{-\frac{1}{2}} & \frac{1}{2} A_{b22} & \frac{1}{2} A_{b21} \Sigma_1^{\frac{1}{2}} & \frac{1}{2} A_{b22} & \frac{1}{2} B_{b2} \\ C_{b1} \Sigma_1^{-\frac{1}{2}} & \frac{1}{2} C_{b2} & 0 & \frac{1}{2} C_{b2} & 0 \end{bmatrix} \tag{C.2}$$

which may be written as

$$\begin{bmatrix} 0 & 0 & I & 0 \\ 0 & \frac{1}{\sqrt{2}} I & 0 & 0 \\ I & 0 & 0 & 0 \\ 0 & \frac{1}{\sqrt{2}} I & 0 & 0 \\ 0 & 0 & 0 & I \end{bmatrix} M \begin{bmatrix} I & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{\sqrt{2}} I & 0 & \frac{1}{\sqrt{2}} I & 0 \\ 0 & 0 & I & 0 & 0 \\ 0 & 0 & 0 & 0 & I \end{bmatrix}$$

where

$$M = \begin{bmatrix} 0 & \frac{1}{\sqrt{2}}\Sigma_1^{-\frac{1}{2}} A_{b12} & \Sigma_1^{-\frac{1}{2}} A_{b11} \Sigma_1^{\frac{1}{2}} & \Sigma_1^{-\frac{1}{2}} B_{b1} \\ \frac{1}{\sqrt{2}} A_{b21} \Sigma_1^{-\frac{1}{2}} & A_{b22} & \frac{1}{\sqrt{2}} A_{b21} \Sigma_1^{\frac{1}{2}} & \frac{1}{\sqrt{2}} B_{b2} \\ \Sigma_1^{\frac{1}{2}} A_{b11} \Sigma_1^{-\frac{1}{2}} & \frac{1}{\sqrt{2}} \Sigma_1^{\frac{1}{2}} A_{b12} & 0 & 0 \\ C_{b1} \Sigma_1^{-\frac{1}{2}} & \frac{1}{\sqrt{2}} C_{b2} & 0 & 0 \end{bmatrix} \tag{C.3}$$

Hence $\|\hat{M}\| \leq 1$ if $\|M\| \leq 1$ and we are going to show the latter.

Lemma C.1 If

$$A_b^*\Sigma A_b - \Sigma + C_b^*C_b \leq 0 \tag{C.4}$$
$$A_b\Sigma A_b^* - \Sigma + B_bB_b^* \leq 0 \tag{C.5}$$

then

$$\left\| \begin{bmatrix} A_{b21}\Sigma_1^{-\frac{1}{2}} & A_{b22} \\ \Sigma_1^{\frac{1}{2}}A_{b11}\Sigma_1^{-\frac{1}{2}} & \Sigma_1^{\frac{1}{2}}A_{b12} \\ C_{b1}\Sigma_1^{-\frac{1}{2}} & C_{b2} \end{bmatrix} \right\| \leq 1; \quad \left\| \begin{bmatrix} \Sigma_1^{-\frac{1}{2}}A_{b12} & \Sigma_1^{-\frac{1}{2}}A_{b11}\Sigma_1^{\frac{1}{2}} & \Sigma_1^{-\frac{1}{2}}B_{b1} \\ A_{b22} & A_{b21}\Sigma_1^{\frac{1}{2}} & B_{b2} \end{bmatrix} \right\| \leq 1$$

◁

Proof. (C.4) and (C.5) may be written as

$$\begin{bmatrix} A_b \\ C_b \end{bmatrix}^* \begin{bmatrix} \Sigma & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} A_b \\ C_b \end{bmatrix} - \Sigma \leq 0 \tag{C.6}$$

$$\begin{bmatrix} A_b & B_b \end{bmatrix}^* \Sigma^{-1} \begin{bmatrix} A_b & B_b \end{bmatrix} - \begin{bmatrix} \Sigma^{-1} & 0 \\ 0 & I \end{bmatrix} \leq 0 \tag{C.7}$$

or equivalently

$$\left\| \begin{bmatrix} \Sigma^{\frac{1}{2}}A_b\Sigma^{-\frac{1}{2}} \\ C_b\Sigma^{-\frac{1}{2}} \end{bmatrix} \right\| \leq 1; \quad \left\| \begin{bmatrix} \Sigma^{-\frac{1}{2}}A_b\Sigma^{\frac{1}{2}} & \Sigma^{-\frac{1}{2}}B_b \end{bmatrix} \right\| \leq 1 \tag{C.8}$$

Then the lemma follows by noting that

$$\begin{bmatrix} A_{b21}\Sigma_1^{-\frac{1}{2}} & A_{b22} \\ \Sigma_1^{\frac{1}{2}}A_{b11}\Sigma_1^{-\frac{1}{2}} & \Sigma_1^{\frac{1}{2}}A_{b12} \\ C_{b1}\Sigma_1^{-\frac{1}{2}} & C_{b2} \end{bmatrix} = \begin{bmatrix} 0 & I & 0 \\ I & 0 & 0 \\ 0 & 0 & I \end{bmatrix} \begin{bmatrix} \Sigma^{\frac{1}{2}}A_b\Sigma^{-\frac{1}{2}} \\ C_b\Sigma^{-\frac{1}{2}} \end{bmatrix}$$

$$\begin{bmatrix} \Sigma_1^{-\frac{1}{2}}A_{b12} & \Sigma_1^{-\frac{1}{2}}A_{b11}\Sigma_1^{\frac{1}{2}} & \Sigma_1^{-\frac{1}{2}}B_{b1} \\ A_{b22} & A_{b21}\Sigma_1^{\frac{1}{2}} & B_{b2} \end{bmatrix} = \begin{bmatrix} \Sigma^{-\frac{1}{2}}A_b\Sigma^{\frac{1}{2}} & \Sigma^{-\frac{1}{2}}B_b \end{bmatrix} \begin{bmatrix} 0 & I & 0 \\ I & 0 & 0 \\ 0 & 0 & I \end{bmatrix}$$

■

Lemma C.2 If $X = \begin{bmatrix} X_{11} & X_{12} \\ Z & X_{22} \end{bmatrix}$ has $\|X\| \leq 1$ and $Y = \begin{bmatrix} Y_{11} & Z \\ Y_{21} & Y_{22} \end{bmatrix}$ has $\|Y\| \leq 1$, then

$$N := \begin{bmatrix} 0 & \frac{1}{\sqrt{2}}X_{11} & X_{12} \\ \frac{1}{\sqrt{2}}Y_{11} & Z & \frac{1}{\sqrt{2}}X_{22} \\ Y_{21} & \frac{1}{\sqrt{2}}Y_{22} & 0 \end{bmatrix} \tag{C.9}$$

has $\|N\| \leq 1$. ◁

Proof. Dilate N to

$$\hat{N} := \left[\begin{array}{ccc|c} 0 & \frac{1}{\sqrt{2}}X_{11} & X_{12} & \frac{1}{\sqrt{2}}X_{11} \\ \frac{1}{\sqrt{2}}Y_{11} & Z & \frac{1}{\sqrt{2}}X_{22} & 0 \\ Y_{21} & \frac{1}{\sqrt{2}}Y_{22} & 0 & -\frac{1}{\sqrt{2}}Y_{22} \\ \hline \frac{1}{\sqrt{2}}Y_{11} & 0 & -\frac{1}{\sqrt{2}}X_{22} & -Z \end{array}\right]$$

Then it can be verified that $\|\hat{N}\| \leq 1$ if $\|X\| \leq 1$ and $\|Y\| \leq 1$. ■

Now we can complete the proof by putting

$$\begin{bmatrix} Y_{11} & Z \\ Y_{21} & Y_{22} \end{bmatrix} = \left[\begin{array}{c|c} A_{b21}\Sigma_1^{-\frac{1}{2}} & A_{b22} \\ \hline \Sigma_1^{\frac{1}{2}}A_{b11}\Sigma_1^{-\frac{1}{2}} & \Sigma_1^{\frac{1}{2}}A_{b12} \\ C_{b1}\Sigma_1^{-\frac{1}{2}} & C_{b2} \end{array}\right]$$

$$\begin{bmatrix} X_{11} & X_{12} \\ Z & X_{22} \end{bmatrix} = \left[\begin{array}{c|cc} \Sigma_1^{-\frac{1}{2}}A_{b12} & \Sigma_1^{-\frac{1}{2}}A_{b11}\Sigma_1^{\frac{1}{2}} & \Sigma_1^{-\frac{1}{2}}B_{b1} \\ \hline A_{b22} & A_{b21}\Sigma_1^{\frac{1}{2}} & B_{b2} \end{array}\right]$$

into Lemma C.2 and applying Lemma C.1 to show that M in (C.3) has $\|M\| \leq 1$.

Appendix D

David and De Moor's algorithm

The following is a MATLAB implementation of David and De Moor's algorithm, which was used by [**mr1**] in Chapter 4 for model reduction. (Particular attention is drawn to the usage of MATLAB commands (:) and \ that are employed below.) For theoretical issues about the algorithm please refer to the original paper [27].

The LMI constraint is described by

$$F(z) = \begin{bmatrix} F_1(z) & & & \\ & F_2(z) & & \\ & & \ddots & \\ & & & F_m(z) \end{bmatrix} \geq 0$$

where $z \in \mathbb{R}^l$ is the variable and F_i are affine $(n_i \times n_i)$-matrix valued functions:

$$F_i(z) = \boldsymbol{F}_{i0} + \boldsymbol{F}_{i1} z_{(1)} + \boldsymbol{F}_{i2} z_{(2)} + \cdots + \boldsymbol{F}_{il} z_{(l)}$$

The rank of $F_m(z)$ is to be minimized.

The parameter q in [27] is set to $0.5 + (\sum_{i=1}^{m} n_i)/n_m$. Define

$$\hat{I} := \begin{bmatrix} I_{n_1}(:) \\ I_{n_2}(:) \\ \vdots \\ I_{n_{m-2}}(:) \\ I_{n_{m-1}}(:) \\ (1-q) I_{n_m}(:) \end{bmatrix}$$

The initial guess $z_{[0]}$ is such that $F(z_{[0]}) \geq 0$ which is found by the `feasp` function in the LMI Control toolbox of MATLAB[31].

For $k = 0, 1, \ldots,$ the followings are done: First $F_i(z_{[k]})$ are decomposed into Cholesky factors: $F_i(z_{[k]}) = Z_i(z_{[k]}) Z_i(z_{[k]})^*$ where $Z_i(z_{[k]})$ are lower triangular. Let

$\tilde{F}_{ij}(z_{[k]}) := Z_i(z_{[k]})\backslash \boldsymbol{F}_{ij}/Z_i(z_{[k]})^*$ and

$$\hat{F}(z_{[k]}) := \begin{bmatrix} \tilde{F}_{11}(z_{[k]})(:) & \tilde{F}_{12}(z_{[k]})(:) & \cdots & \tilde{F}_{1l}(z_{[k]})(:) \\ \tilde{F}_{21}(z_{[k]})(:) & \tilde{F}_{22}(z_{[k]})(:) & \cdots & \tilde{F}_{2l}(z_{[k]})(:) \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{F}_{m1}(z_{[k]})(:) & \tilde{F}_{m2}(z_{[k]})(:) & \cdots & \tilde{F}_{ml}(z_{[k]})(:) \end{bmatrix}$$

Then the Newton direction $v(z_{[k]})$ in [27] is given by $v(z_{[k]}) = \hat{F}(z_{[k]})\backslash\hat{I}$; the Newton decrement $\delta(z_{[k]})$ by $\delta(z_{[k]}) = \|\hat{F}(z_{[k]})v(z_{[k]})\|$ (note that $\hat{F}(z_{[k]})v(z_{[k]})$ is a vector); and the Nesterov-Nemirovsky damping factor $\alpha(\delta(z_{[k]}))$ by

$$\alpha(\delta) = \begin{cases} 1/(1+\delta) & \text{if} \quad \delta > 0.25 \\ 1 & \text{if} \quad \delta \leq 0.25 \end{cases}$$

At last z is updated as $z_{[k+1]} = z_{[k]} + \alpha(\delta(z_{[k]}))v(z_{[k]})$.

The terminating criterion given by [27] was tried but found unreliable. So instead the algorithm is terminated manually, usually after it is observed that $\text{rank}(F_m(z))$ has had no change since a large number of iterations ago.

Appendix E

ϕ-dependent approach

Given a particular $\phi \in \mathcal{OL}(\mathbb{R})$, P pdd is only a sufficient condition for

$$\forall \zeta \in \mathbb{R}^n, \quad \phi(\zeta)^* P \phi(\zeta) \leq \zeta^* P \zeta \tag{E.1}$$

which indeed is what we need in proving the results in this dissertation. Although the pdd approach has been observed to work quite satisfactorily, it may happen that one still thinks it conservative and would like to deal with the constraint (E.1) directly. Note that (E.1) is still a convex constraint in P, but it is ϕ-dependent and therefore cannot be converted into a LMI, and numerical algorithms like the older ellipsoid algorithm or the newer interior-point method are not applicable. Clearly the idea of this dissertation is that if a whole set of ϕ, $\mathcal{OL}(\mathbb{R})$ is considered instead of a particular ϕ, then the constraint will be equivalent to P pdd which does not involve ϕ and hence the difficulties arise from its nonlinear nature are removed.

The objective of this appendix is to point out that constraint (E.1) is practically solvable *if* given any P we can detect whether it satisfies (E.1) and find a $\zeta \in \mathbb{R}^n$ such that $\phi(\zeta)^* P \phi(\zeta) > \zeta^* P \zeta$ whenever P does not. Indeed, instead of (E.1) we may further consider

$$\forall \zeta, \quad \underline{\phi}(\zeta)^* H_1 \underline{\phi}(\zeta) \leq \zeta^* H_0 \zeta \tag{E.2}$$

The inequalities associated with the stability, ℓ_2-ℓ_2 induced gain, observability function and controllability function can all be written in this form, which is slightly tighter than separating it into a LMI and (E.1). For example, if the ℓ_2-ℓ_2 induced gain is considered, the actual inequality we need is (E.2) with

$$H_1 = \begin{bmatrix} A & B \\ C & D \end{bmatrix}^* \begin{bmatrix} P & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} A & B \\ C & D \end{bmatrix}, \quad H_0 = \begin{bmatrix} P & 0 \\ 0 & \gamma^2 I \end{bmatrix}, \quad \underline{\phi} = \begin{bmatrix} \phi & \\ & I \end{bmatrix}$$

where γ^2 is to be minimized. In the pdd approach we separate it into $H_1 - H_0 \leq 0$ and (E.1) and then use P pdd to replace the latter, but for the current approach we may handle (E.2) directly. Two methods are described in this appendix.

The first one is a (generalized) ellipsoid algorithm which, for example, can be used to determine an upper bound of the ℓ_2-ℓ_2 induced gain. An outline is given below; the treatment is based on [22, §2.3]. Given $c \in \mathbb{R}^l$, suppose that we want to minimize $c^* z$ subject to

$$\forall \zeta \in \mathbb{R}^n, \quad F(z,\zeta) = \boldsymbol{F}_0(\zeta) + \boldsymbol{F}_1(\zeta) z_{(1)} + \boldsymbol{F}_2(\zeta) z_{(2)} + \cdots + \boldsymbol{F}_l(\zeta) z_{(l)} \geq 0 \tag{E.3}$$

where $\boldsymbol{F}_0 : \mathbb{R}^n \to \mathbb{R}$. Let

$$\mathcal{E}(z_{[0]}, E_{[0]}) := \left\{ \eta \in \mathbb{R}^l \;\middle|\; (\eta - z_{[0]})^* E_{[0]}^{-1} (\eta - z_{[0]}) \leq 1 \right\}$$

be an ellipsoid centred at $z_{[0]}$ which is known to have the optimal z in it[1]. We want to find a plane (the cutting plane) passing through the center $z_{[0]}$ of the ellipsoid: The plane is characterized by its normal vector $g_{[0]} \in \mathbb{R}^l$ which is to be chosen such that the optimal z is in the half-space

$$\mathcal{HS}(z_{[0]}, g_{[0]}) := \left\{ \eta \in \mathbb{R}^l \;\middle|\; g_{[0]}^* (\eta - z_{[0]}) \leq 0 \right\}$$

If $z_{[0]}$ does not satisfy (E.3), $\exists \zeta \in \mathbb{R}^n$ such that $F(z_{[0]}, \zeta) < 0$, then $g_{[0]}$ may be given by $g_{[0](i)} = -\boldsymbol{F}_i(\zeta)$ since $\forall \eta \in \mathbb{R}^l \setminus \mathcal{HS}(z_{[0]}, g_{[0]})$,

$$F(\eta, \zeta) = F(z_{[0]}, \zeta) + \sum_{i=1}^{l} \boldsymbol{F}_i(\zeta)(\eta - z_{[0]}) = F(z_{[0]}, \zeta) - g_{[0]}^* (\eta - z_{[0]}) < 0$$

and hence η is not a feasible point of (E.3) at all, thus surely not the optimal z. If $z_{[0]}$ satisfies (E.3), $g_{[0]}$ may be given by $g_{[0]} = c$ since $\forall \eta \in \mathbb{R}^l \setminus \mathcal{HS}(z_{[0]}, g_{[0]})$,

$$c^* \eta = c^* z_{[0]} + c^* (\eta - z_{[0]}) = c^* z_{[0]} + g_{[0]}^* (\eta - z_{[0]}) > c^* z_{[0]}$$

and hence η is not optimal. Now we know that the optimal z is in the half-ellipsoid $\mathcal{E}(z_{[0]}, E_{[0]}) \cap \mathcal{HS}(z_{[0]}, g_{[0]})$ and we find the minimum-volume ellipsoid that contains this half-ellipsoid:

$$\mathcal{E}(z_{[1]}, E_{[1]}) \supset \mathcal{E}(z_{[0]}, E_{[0]}) \cap \mathcal{HS}(z_{[0]}, g_{[0]})$$

where $z_{[1]}$, $E_{[1]}$ are given by

$$\begin{aligned}
\bar{g}_{[0]} &:= g_{[0]} \frac{1}{\|g_{[0]}\|} \\
E_{[0]} \tilde{g}_{[0]} &:= E_{[0]} \bar{g}_{[0]} \frac{1}{\sqrt{\bar{g}_{[0]}^* E_{[0]} \bar{g}_{[0]}}} \\
z_{[1]} &= z_{[0]} - (E_{[0]} \tilde{g}_{[0]}) \frac{1}{l+1} \\
E_{[1]} &= \left(E_{[0]} - (E_{[0]} \tilde{g}_{[0]})(E_{[0]} \tilde{g}_{[0]})^* \frac{2}{l+1} \right) \frac{l^2}{l^2 - 1}
\end{aligned}$$

[1] Practically it may be difficult to ensure that the optimal $z \in \mathcal{E}(z_{[0]}, E_{[0]})$, but at least a reasonably good suboptimal z should be included. For example, if the ℓ_2-ℓ_2 induced gains of our 22 (diagonally stable) ϕ-systems in Section 2.3 are to be estimated, the diagonal gains and the linear gains and their corresponding P may help to set up the initial ellipsoids $\mathcal{E}(z_{[0]}, E_{[0]})$. Of course one may use the dd-gains instead of the diagonal gains but this will need more computational efforts.

This new ellipsoid will continue to have the optimal z in it so we may apply the same procedure to $\mathcal{E}(z_{[1]}, E_{[1]})$ to obtain $\mathcal{E}(z_{[2]}, E_{[2]})$. In this way a sequence of ellipsoids are obtained whose volumes can be shown to be at most $e^{-\frac{1}{2l}}$ times their predecessors. At last the optimal z will be found.

Clearly the only non-trivial part of the algorithm is to detect whether $z_{[k]}$ satisfies (E.3) and if not, how to find a $\zeta \in \mathbb{R}^n$ such that $F(z_{[k]}, \zeta) < 0$. In the case of estimating the ℓ_2-ℓ_2 induced gain, if H_1 or H_0 happens to be pdd then they satisfy (E.2), but in general the detection is rather difficult. One way to do it is to compute

$$\alpha := \min_{\zeta} \left(\zeta^* H_0 \zeta - \underline{\phi}(\zeta)^* H_1 \underline{\phi}(\zeta) \right) \tag{E.4}$$

If $\alpha < 0$, then H_1, H_0 do not satisfy (E.2) and we find the ζ simultaneously. However, $\zeta^* H_0 \zeta - \underline{\phi}(\zeta)^* H_1 \underline{\phi}(\zeta)$ has in fact a lot of local minima so computationally it is very difficult to guarantee the global minimum α. If the minimization algorithm only arrives at a non-negative local minimum while the actual global minimum $\alpha < 0$, then a false conclusion will be drawn. The problem of the ellipsoid algorithm is that the computation of such a minimizer ζ is needed so many times in the algorithm that no matter how good the minimization algorithm will work, if only it is not perfect then there will still be a significant amount of false iterations which suffice to make the ellipsoid algorithm totally useless. So the ellipsoid algorithm above is only to give a connection between the solvability and detectability of (E.2) but not practically working so far.

The second method, although theoretically sounds stupid, is practically rather effective in estimating the ℓ_2-ℓ_2 induced gain "from below"[2]. Note that for any particular $\zeta \in \mathbb{R}^n$, $F(z, \zeta) = 0$ is a plane in the space of $\{z \in \mathbb{R}^l\}$ and $F(z, \zeta) \geq 0$ is a half-space. So for any finite set $\{\zeta_i\}$, the set

$$\left\{ z \in \mathbb{R}^l \;\middle|\; \forall \zeta_i, \;\; F(z, \zeta_i) \geq 0 \right\}$$

is an *outer* polytopic approximation of the feasible set of (E.3) (in contrast to the *inner* approximation of (E.1) given by the set of pdd matrices). So, for example, the ℓ_2-ℓ_2 induced gain may be estimated from below as follows. We start with simply solving $H_1 - H_0 \leq 0$ (i.e. the linear gain). If H_1, H_0 satisfy (E.2), it will be all right; otherwise $\exists \zeta_1 \in \mathbb{R}^n$ such that $\underline{\phi}(\zeta_1)^* H_1 \underline{\phi}(\zeta_1) > \zeta_1^* H_0 \zeta_1$. Then we solve $H_1 - H_0 \leq 0$ again together with the constraint $\underline{\phi}(\zeta_1)^* H_1 \underline{\phi}(\zeta_1) \leq \zeta_1^* H_0 \zeta_1$. Similarly if H_1, H_0 now satisfy (E.2), stop; otherwise $\exists \zeta_2 \in \mathbb{R}^n$ such that $\underline{\phi}(\zeta_2)^* H_1 \underline{\phi}(\zeta_2) > \zeta_2^* H_0 \zeta_2$ and a further constraint is added to the original set. In this way we hope to improve the outer polytopic approximation of the feasible set successively and hence the estimate of the ℓ_2-ℓ_2 induced gain from below, but the progress has no theoretical guarantee and therefore one may choose to stop at any time.

[2]Strictly speaking it is not a lower bound of the ℓ_2-ℓ_2 induced gain but only a lower bound of its upper bound. This upper bound, however, is believed to be the tightest possible in the quadratic approach.

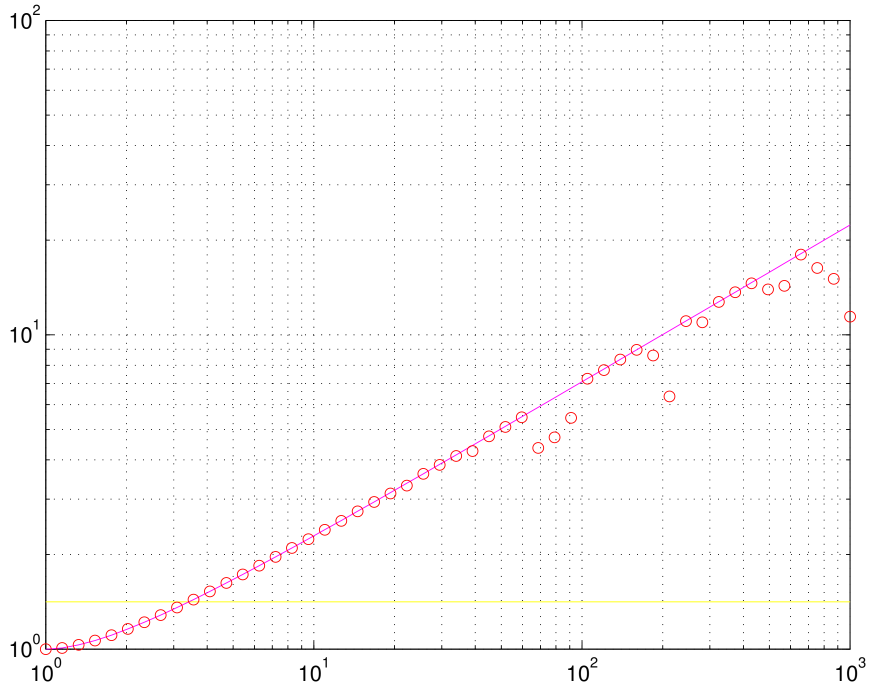


Figure E.1: The lower estimates of the ℓ_2-ℓ_2 induced gains (the small circles) and $\beta_2(a)$ of Section 5.1.3 (the darker line) for different values of $a \in [1, 1000]$. The lighter line is $\sqrt{2}$.

There are two reasons why this method is practical. The first one is rather obvious: the problem is still a standard LMI problem and therefore the efficient interior-point method is applicable. Repeatedly solving the LMI is time consuming but one may choose to stop at any time. Note that the LMI involves no pdd matrices and therefore the computational effort needed for each iteration is much lower than the dd-gain, but the total effort may be higher. The second reason is that it does not "accumulate" the errors caused by the imperfection of the minimization algorithm for (E.4) as the ellipsoid algorithm does. More precisely, if the minimization algorithm fails, the ellipsoid algorithm will *proceed with the error* while the current method will just *stop*, and if we do not trust it but want to continue we simply retry the last minimization, say, with another initial guess to see whether a different conclusion will be drawn.

As a demonstration, the second method is now used to lower estimate the ℓ_2-ℓ_2 induced gain of $\mathcal{F}_l(G_a(\text{sat}), -1)$ which has been studied in Section 5.1.3. We know that $\mathcal{F}_l(G(\text{sat}), -1)$ has a ℓ_2-ℓ_2 induced gain $\sqrt{2}$, but it is also shown in Section 5.1.3 that $\mathcal{F}_l(G(\text{sat}), -1)$ admits no quadratic Lyapunov/storage function and therefore any estimate of the ℓ_2-ℓ_2 induced gain of $\mathcal{F}_l(G_a(\text{sat}), -1)$ using a quadratic Lyapunov/storage function must be unbounded as $a \to \infty$. The minimization algorithm used is the `fmins` function in MATLAB with a random initial guess. If no ζ such that $\underline{\phi}(\zeta)^* H_1 \underline{\phi}(\zeta) > \zeta^* H_0 \zeta$ can be found after `fmins` has been called 32 times with a different random initial guess each time, the whole algorithm will be terminated and the lower estimate of the ℓ_2-ℓ_2 induced gain reported. The results for 50 different values of a from 1 to 1000 are shown in Figure E.1 (the small circles). Curiously it seems that the estimates of the ℓ_2-ℓ_2 induced gains are approaching $\beta_2(a)$ of Section 5.1.3 from below, which is plotted on the same figure for comparison (the darker line). From this it is reasonable to conjecture that no quadratic Lyapunov/storage function can prove a ℓ_2-ℓ_2 induced gain $< \beta_2(a)$ for $\mathcal{F}_l(G_a(\mathrm{sat}), -1)$.

Although in theory the feasible set of (E.2) may require an infinite number of planes to give an exact outer approximation, the average and the maximum cardinality of the sets $\{\zeta_i\}$ that have been used to obtain the lower estimates of the ℓ_2-ℓ_2 induced gains for the 50 different values of a are only 3.62 and 9 respectively. In other words, the numbers of LMIs that have been involved are in fact not that large. However, perhaps this is true only when the dimension of the problem l is small (4 in this case).

Bibliography

[1] M. Akgül. *Topics in Relaxation and Ellipsoidal Methods*, volume 97 of *Research Notes in Mathematics*. Pitman, 1984.

[2] U. M. Al-Saggaf and G. F. Franklin. An error bound for a discrete reduced order model of a linear multivariable system. *IEEE Transactions on Automatic Control*, 32:815–819, 1987.

[3] F. Albertini and P. Dai Pra. Forward accessibility for recurrent neural networks. *IEEE Transactions on Automatic Control*, 40(11):1962–1968, November 1995.

[4] F. Albertini and E. D. Sontag. For neural networks, function determines form. *Neural Networks*, 6:975–990, 1993.

[5] F. Albertini and E. D. Sontag. Identifiability of discrete-time neural networks. In *Proceedings of the European Control Conference*, pages 460–465, Groningen, July 1993.

[6] F. Albertini and E. D. Sontag. Uniqueness of weights for recurrent nets. In *Proceedings of the International Symposium Math. Theory of Networks Syst.*, volume II, pages 599–602, Regensburg, Germany, August 1993.

[7] F. Albertini and E. D. Sontag. State observability in recurrent neural networks. *Systems & Control Letters*, 22(4):235–244, 1994.

[8] F. Albertini, E. D. Sontag, and V. Maillot. Uniqueness of weights for neural networks. In R. J. Mammone, editor, *Artificial Neural Networks for Speech and Vision*, pages 113–125. Chapman and Hall, London, 1993.

[9] P. Apkarian and P. Gahinet. A convex characterization of gain-scheduled $\mathcal{H}_\infty$ controllers. *IEEE Transactions on Automatic Control*, 40(5):853–864, May 1995.

[10] P. Apkarian and P. Gahinet. Erratum to “A convex characterization of gain-scheduled $\mathcal{H}_\infty$ controllers”. *IEEE Transactions on Automatic Control*, 40(9):1681, September 1995.

[11] G. P. Barker, A. Berman, and R. J. Plemmons. Positive diagonal solutions to the Lyapunov equations. *Linear and Multilinear Algebra*, 5:249–256, 1978.

[12] C. Beck. Minimality for uncertain systems and IQCs. In *Proceedings of the 33rd IEEE Conference on Decision and Control*, pages 3068–3073, Lake Buena Vista, Florida, December 1994.

[13] C. Beck. *Model Reduction for Multi-Dimensional and Uncertain Systems.* PhD thesis, California Institute of Technology, 1995.

[14] C. Beck, R. D'Andrea, F. Paganini, W.-M. Lu, and J. Doyle. A state-space theory of uncertain systems. In *Proceedings of the 13th Triennial IFAC World Congress*, pages 291–296, San Francisco, California, July 1996.

[15] C. Beck, J. Doyle, and K. Glover. Model reduction of multi-dimensional and uncertain systems. Submitted to *IEEE Transactions on Automatic Control*, 1995.

[16] A. Berman and D. Hershkowitz. Matrix diagonal stability and its implications. *SIAM J. Algebraic Discrete Methods*, 4(3):377–382, September 1983.

[17] A. Berman and D. Hershkowitz. Characterization of acyclic D-stable matrices. *Linear algebra and its applications*, 58:17–31, April 1984.

[18] A. Berman and R. J. Plemmons. *Nonnegative Matrices in the Mathematical Sciences.* Academic Press, New York, 1979.

[19] J. Bernussou, P. L. D. Peres, and J. C. Geromel. Robust decentralized regulation: A linear programming approach. In *IFAC/IFORS/IMACS Symposium Large Scale Systems: Theory and Applications*, volume 1, pages 135–138, Berlin-GDR, 1989.

[20] R. G. Bland, D. Goldfarb, and M. J. Todd. The ellipsoid method: A survey. *Operations Research*, 29(6):1039–1091, 1981.

[21] S. Boyd and C. Barratt. *Linear Controller Design: Limits of Performance.* Prentice-Hall, 1991.

[22] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan. *Linear Matrix Inequalities in System and Control Theory*, volume 15 of *Studies in Applied Mathematics.* SIAM, Philadelphia, 1994.

[23] G. W. Cross. Three types of matrix stability. *Linear algebra and its applications*, 20:253–263, 1978.

[24] R. D'Andrea. *Implicit Systems and Constrained Optimization.* PhD thesis, California Institute of Technology, 1996.

[25] R. D'Andrea and F. Paganini. Interconnection of uncertain behavioral systems for robust control. In *Proceedings of the 32nd IEEE Conference on Decision and Control*, pages 3642–3647, San Antonio, Texas, December 1993.

[26] J. David and B. De Moor. The opposite of analytic centering for solving minimum rank problems in control and identification. In *Proceedings of the 32nd IEEE Conference on Decision and Control*, pages 2901–2902, San Antonio, Texas, December 1993.

[27] J. David and B. De Moor. Designing reduced order output feedback controllers using a potential reduction method. In *Proceedings of the American Control Conference*, pages 845–849, Baltimore, Maryland, June 1994.

[28] J. C. Doyle, K. Glover, P. P. Khargonekar, and B. A. Francis. State-space solutions to standard $\mathcal{H}_2$ and $\mathcal{H}_\infty$ control problems. *IEEE Transactions on Automatic Control*, 34(8):831–847, 1989.

[29] D. Enns. *Model Reduction for Control System Design.* PhD thesis, Department of Aeronautics and Astronautics, Stanford University, Stanford, California, 1984.

[30] P. Finsler. Über das Vorkommen definiter und semi-definiter Formen in Scharen quadratischer Formen. *Comentarii Mathematici Helvetici*, 9, 1937.

[31] P. Gahinet, A. Nemirovskii, A. J. Laub, and M. Chilali. *The LMI Control Toolbox.* The MathWorks Inc., November 1994. Beta-Release.

[32] P. Gahinet, A. Nemirovskii, A. J. Laub, and M. Chilali. The LMI control toolbox. In *Proceedings of the 33rd IEEE Conference on Decision and Control*, pages 2038–2041, Lake Buena Vista, Florida, December 1994.

[33] B. R. Gelbaum and J. M. H. Olmsted. *Counterexamples in Analysis.* Holden-Day Inc., San Francisco, California, 1964.

[34] K. Glover. All optimal Hankel-norm approximations of linear multivariable systems and their $\mathcal{L}_\infty$-error bounds. *International Journal of Control*, 39:1115–1193, 1984.

[35] K. Glover. A tutorial on Hankel-norm approximation. In J. C. Willems, editor, *Data to Model.* Springer-Verlag, New York, 1989.

[36] A. Helmersson. *Methods for robust gain scheduling.* PhD thesis, Linkoping University, 1995.

[37] D. Hershkowitz and H. Schneider. Scalings of vector spaces and the uniqueness of Lyapunov scaling factors. *Linear and Multilinear Algebra*, 17:203–226, 1985.

[38] D. Hinrichsen and A. J. Pritchard. An improved error estimate for reduced-order models of discrete time system. *IEEE Transactions on Automatic Control*, 35:317–320, 1990.

[39] R. A. Horn and C. R. Johnson. *Matrix analysis.* Cambridge University Press, 1985.

[40] K. Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4:251–257, 1991.

[41] M. C. Joshi and R. K. Bose. *Some Topics in Nonlinear Functional Analysis*. Wiley, 1985.

[42] V. A. Kamenetskii. Absolute stability and absolute instability of control systems with several nonlinear nonstationary elements. *Automation and Remote Control*, 44(12):1543–1552, 1983.

[43] E. Kaszkurewicz and A. Bhaya. Robust stability and diagonal Liapunov functions. *SIAM Journal Matrix Anal. Appl.*, 14(2):508–520, April 1993.

[44] E. Kaszkurewicz and A. Bhaya. On a class of globally stable neural circuits. *IEEE Transactions on Circuits and Systems—I: Fundamental Theory and Applications*, 41(2):171–174, February 1994.

[45] E. Kaszkurewicz and A. Bhaya. Comments on "Necessary and sufficient condition for absolute stability of neural networks". *IEEE Transactions on Circuits and Systems—I: Fundamental Theory and Applications*, 42(8):497–499, August 1995.

[46] E. Kaszkurewicz, A. Hermeto, and L. Hsu. A result on stability of nonlinear discrete time systems and its application to recursive digital filters. In *Proceedings of the 23rd IEEE Conference on Decision and Control*, pages 100–102, Las Vegas, Nevada, December 1984.

[47] E. Kaszkurewicz and L. Hsu. On two classes of matrices with positive diagonal solutions to the Lyapunov equation. *Linear algebra and its applications*, 59:19–27, June 1984.

[48] E. Kaszkurewicz and L. Hsu. Stability conditions for recursive digital filters with signal quantization. Tech. Rep. EE-3, Department of Engineering, COPPE/UFRJ, Rio de Janeiro, Brazil, 1988.

[49] P. P. Khargonekar and M. A. Rotea. Stabilization of uncertain systems with norm bounded uncertainty using control Lyapunov functions. In *Proceedings of the 27th IEEE Conference on Decision and Control*, volume 1, pages 503–507, December 1988.

[50] E. Kreyszig. *Introductory Functional Analysis with Applications*. John Wiley & Sons, 1978.

[51] W.-M. Lu, K. Zhou, and J. C. Doyle. Stabilization of uncertain linear systems: An LFT approach. *IEEE Transactions on Automatic Control*, 41(1):50–65, January 1996.

[52] M. Matthews. *On the uniform approximation of nonlinear discrete-time fading-memory systems using neural network models.* Diss. ETH no. 9635, E.T.H. Zurich, 1992.

[53] W. L. Mills, C. T. Mullis, and R. Roberts. Digital filter realizations without overflow oscillations. *IEEE Transactions on Acoust. Speech Signal Process.*, 26:334–338, 1978.

[54] A. P. Molchanov and E. S. Pyatnitskii. Lyapunov functions that specify necessary and sufficient conditions of absolute stability of nonlinear nonstationary control systems I. *Automation and Remote Control*, 46(3):344–354, 1986.

[55] A. P. Molchanov and E. S. Pyatnitskii. Lyapunov functions that specify necessary and sufficient conditions of absolute stability of nonlinear nonstationary control systems II. *Automation and Remote Control*, 46(4):443–451, 1986.

[56] A. P. Molchanov and E. S. Pyatnitskii. Lyapunov functions that specify necessary and sufficient conditions of absolute stability of nonlinear nonstationary control systems III. *Automation and Remote Control*, 46(5):620–630, 1986.

[57] B. C. Moore. Principal component analysis in linear systems: controllability, observability and model reduction. *IEEE Transactions on Automatic Control*, 26:17–32, 1981.

[58] Yu. Nesterov and A. Nemirovsky. *Interior-point polynomial methods in convex programming*, volume 13 of *Studies in Applied Mathematics*. SIAM, Philadelphia, 1994.

[59] K. Ogata. *Discrete-Time Control Systems*. Prentice-Hall, 1987.

[60] A. Packard. Gain scheduling via linear fractional transformations. *Systems & Control Letters*, 22:79–92, 1994.

[61] A. Packard, K. Zhou, P. Pandey, J. Leonhardson, and G. Balas. Optimal, constant I/O similarity scaling for full-information and state-feedback control problems. *Systems & Control Letters*, 19:271–280, 1992.

[62] F. Paganini. *Sets and Constraints in the Analysis of Uncertain Systems.* PhD thesis, California Institute of Technology, 1995.

[63] F. Paganini, R. D'Andrea, and J. C. Doyle. Behavioral approach to robustness analysis. In *Proceedings of the American Control Conference*, pages 2782–2786, Baltimore, Maryland, June 1994.

[64] F. Paganini and J. Doyle. Analysis of implicit uncertain systems, parts i and ii. CDS Technical Report CIT/CDS 94-018, California Institute of Technology, 1994.

[65] F. Paganini and J. Doyle. Analysis of implicitly defined systems. In *Proceedings of the 33rd IEEE Conference on Decision and Control*, pages 3673–3678, Lake Buena Vista, Florida, December 1994.

[66] S. K. Persidskii. Problem of absolute stability. *Automation and Remote Control*, 12:1889–1895, 1969.

[67] I. R. Petersen and C. V. Hollot. A Ricaati equation approach to the stabilization of uncertain linear systems. *Automatica*, 22(4):397–411, 1986.

[68] E. S. Pyatnitskii and V. I. Skorodinskii. Numerical methods of Lyapunov function construction and their application to the absolute stability problem. *Systems & Control Letters*, 2(2):130–135, August 1982.

[69] B. G. Romanchuk. *Input-Output Analysis of Feedback Loops with Saturation Nonlinearities.* PhD thesis, University of Cambridge, 1995.

[70] J. M. A. Scherpen. Balancing for nonlinear systems. *Systems & Control Letters*, 21(2):143–153, August 1993.

[71] G. Scorletti and L. El Ghaoui. Improved linear matrix inequality conditions for gain scheduling. In *Proceedings of the 34th IEEE Conference on Decision and Control*, New Orleans, Louisiana, December 1995.

[72] E. D. Sontag. Systems combining linearity and saturations, and relations to "neural nets". Report SYCON-92-01, Rutgers Center for Systems and Control, March 1992.

[73] D. D. Šiljak. *Large Scale Dynamic Systems: Stability and Structure.* Elsevier North-Holland, 1978.

[74] W. Wang, J. Doyle, C. Beck, and K. Glover. Model reduction of LFT systems. In *Proceedings of the 30th IEEE Conference on Decision and Control*, Brighton, England, December 1991.

[75] M. Wright. Interior methods for constrained optimization. *Acta Numerica*, 1, 1992.

[76] N. Young. *An introduction to Hilbert space.* Cambridge University Press, 1988.

[77] K. Zhou, J. C. Doyle, and K. Glover. *Robust and Optimal Control.* Prentice-Hall, Englewood Cliffs, New Jersey, 1995.