LINEAR DYNAMICAL SYSTEMS, APPROXIMATION

Approximation is an important methodology in science and engineering. In this essay we will review a theory of approximation of linear dynamical systems that has a number of desirable features. Main ingredients of this theory are: the 2-norms used to measure the quantities involved, in particular the Hankel norm, the infinity norm, and a set of invariants called the Hankel singular values. The main tool for the construction of approximants is the all-pass dilation (unitary extension) of the original dynamical system.

Two basic and desirable features of an approximation theory are the existence of:

- a constructive methodology for obtaining reduced-complexity models
- an appropriate quantization of the approximation error, in other words, the estimation of some measure of the error between the given high-complexity model and the derived reduced-complexity model(s)

The importance of the second item cannot be overemphasized: In approximating a system, one wishes to have some idea of what has been eliminated.

Often an additional desirable feature consists in

- looking for reduced-complexity models within a specified class, e.g. the class of stable systems

For instance, in case the nominal model is stable, if the intended use of a reduced-complexity model is in open-loop, it is imperative that the latter also be stable.

We will deal with linear, time-invariant, discrete- and continuous-time, finite-dimensional systems described by convolution sums or integrals. In addition, we will only consider the approximation of *stable* systems. Most of the approximation methods available—for example, Padé approximation—fail to satisfy the requirements listed above. We will discuss the theory of *Hankel-norm approximation* and the related *approximation by balanced truncation.* The size of systems—including error systems—is measured in terms of appropriately defined 2-norms, and complexity is measured in terms of the (least) number of state variables (i.e., first-order differential or difference equations) needed to describe the system.

This approach to model reduction has an interesting history. For operators in finite-dimensional spaces (matrices), the problem of optimal approximation by operators of lower rank is solved by the Schmidt–Mirsky theorem (see Ref. 1). Although this is *not* a convex problem and consequently conventional optimization methods do not apply, it can be solved explicitly by using an ad hoc tool: the *singular value decomposition* (SVD) of the operator. The solution involves the *truncation* of *small* singular values.

In the same vein, a linear dynamical system can be represented by means of a structured (Hankel) linear operator in appropriate infinite dimensional spaces. The Adamjan–Arov–Krein (AAK) theory (see Refs. 2 and 3) generalizes the Schmidt–Mirsky result to dynamical systems, that is to structured operators in infinite-dimensional spaces. This is also known as *optimal approximation in the Hankel norm.* The original setting of the AAK theory was functional analytic. Subsequent developments due to Glover (4) resulted in a simplified linear algebraic framework. This setting made the theory quite transparent by providing explicit formulae for the quantities involved.

The Hankel norm approximation problem can be addressed and solved both for discrete- and continuous-time systems. However, the following is a fact: The discrete-time case is closer to that of finite-dimensional operators and to the Schmidt–Mirsky result. Therefore, the intuitive understanding of the results in this case is more straightforward. In the continuous-time case, on the other hand, while the interpretation of the results is less intuitive than for the discrete-time case, the corresponding formulas turn out to be *simpler* (see Remark 3). Because of this dichotomy we will first state the results in the discrete-time case, trying to make connections and draw parallels with the Schmidt–Mirsky theorem (see section entitled "The Schmidt–Mirsky Theorem and the AAK Generalization"). The numerous formulas for constructing approximants, however, will be given for the continuous-time case (see section entitled "Construction of Approximants").

The Hankel-norm approximation, as well as model reduction by balanced truncation, inherits an important property from the Schmidt–Mirsky result: Unitary operators form the building blocks of this method and cannot be approximated; a multiple singular value indicates that the original operator contains a unitary part. This unitary part has to either be included in the approximant as a whole or discarded as a whole; it must not be truncated. The same holds for the extension to Hankel-norm approximation. The fundamental building blocks are all-pass systems, and a multiple Hankel singular value indicates the existence of an all-pass subsystem. In the approximation, this subsystem will be either eliminated or included as a whole; it must not be truncated. Consequently, it turns out that the basic operation involved in constructing approximants is *all-pass dilation* (unitary extension)—that is, the extension of the number of states of the original system so that the aggregate becomes all-pass (unitary) [see Main Theorem (2.3)].

The article contains three main sections. The first, entitled "The Schmidt–Mirsky Theorem and the AAK Generalization," reviews 2-norms and induced 2-norms in finite dimensions and states the Schmidt–Mirsky result. The second half of this section presents the appropriate generalization of these concepts for linear discrete-time systems. An operator which is intrinsically attached to a linear system Σ is the *convolution operator* $\mathcal{S}_\Sigma$. Therefore, the problem of optimal approximation in the 2-induced norm of this operator arises naturally. For reasons explained in the section entitled "Approximation of Σ in the 2-Induced Norm of the Convolution Operator," however, it is currently not possible to solve this problem, except in some special cases [see Remark 4(b)]. A second operator, the *Hankel operator* $\mathcal{H}_\Sigma$, is obtained by restricting the domain and the range of the convolution operator. The 2-induced norm of $\mathcal{H}_\Sigma$ is the *Hankel norm* of Σ. It turns out that the optimal approximation problem is solvable in the Hankel norm of Σ. This is the AAK result, stated in the section entitled "The AAK Theorem." The main result of optimal and suboptimal approximation in the Hankel norm is presented next. The fundamental construction involved is the *all-pass dilation* (unitary extension) $\hat{\Sigma}$ of Σ.

The most natural norm for the approximation problem is the 2-induced norm of the convolution operator, which is equal to the infinity norm of the associated rational transfer function (i.e., the maximum of the amplitude Bode plot). However, only a different problem can be solved, namely the approximation in the 2-induced norm of the Hankel operator (which is different from the convolution operator). Although the problem solved is not the same as the original one, the singular values of the Hankel operator turn out to be important invariants. Among other things, they provide *a priori computable bounds*—both upper and lower—for the infinity norm of the error systems.

The first part of the section entitled "Construction of Approximants" presents the fundamentals of continuous-time linear systems. The convolution and the Hankel operators are introduced together with the grammians and the computation of the singular values of the Hankel operator $\mathcal{H}_\Sigma$. These are followed by Lyapunov equations, inertia results, and state-space characterizations of all-pass systems—all important ingredients of the theory. Subsequently, various formulas for optimal and suboptimal approximants are given. These formulae are presented for continuous-time systems since (as mentioned earlier) they are simpler than their discrete-time counterparts. First comes an input–output approach applicable to single-input single-output systems (see section entitled "Input–Output Construction Method for Scalar Systems"); it involves the solution of a polynomial equation which is straightforward to set up and solve. The remaining formulas are all state-space-based. The following cases are treated (in increasing complexity): square systems, suboptimal case; square systems, optimal case; general systems, suboptimal case. These are followed by the important section entitled "Error Bounds of Optimal and Suboptimal Approximants"; these bounds concern the 2-induced norms of both the Hankel and the convolution operators of error systems. The section entitled "Balanced Realizations and Balanced Model Reduction," presents the closely related approximation method by balanced truncation. Again error bounds are available, although balanced approximants satisfy no optimality properties.

This article concludes with three examples and a brief overview of some selected recent developments, including an overview of the approximation problem for *unstable* systems.

Besides the original sources, namely Refs. 2–4, parts of the material presented below can also be found in books (5,6) and in lecture notes (7). For the mathematical background needed we refer to Ref. 8.

THE SCHMIDT–MIRSKY THEOREM AND THE AAK GENERALIZATION

In this section we will first state the Schmidt–Mirsky theorem concerned with the approximation of finite-dimensional unstructured operators in the 2-norm, which is the operator norm induced by the vector Euclidean norm. Then, discrete-time linear dynamical systems are introduced together with the associated *convolution operator;* this is a structured operator defined on the space of square summable sequences. It is argued in the section entitled "Approximation of Σ in the 2-Induced Norm of the Convolution Operator" that due to the fact that the convolution operator has a continuous spectrum, it is not clear how the Schmidt–Mirsky result might be generalized in this case. However, roughly speaking by restricting the domain and the range of the convolution operator, the *Hankel operator* is obtained. It turns out that the Schmidt–Mirsky result can be generalized in this case; this is the famous AAK theorem followed by the section entitled "The Main Result."

2-Norms and Induced 2-Norms in Finite Dimensions

The Euclidean or 2-norm of $x \in \mathbb{R}^n$ is defined as

$$\|x\|_2 := \sqrt{x_1^2 + \cdots + x_n^2}$$

Given the linear map $A:\mathbb{R}^n \to \mathbb{R}^m$, the norm induced by the Euclidean norm in the domain and range of A is the *2-induced* norm:

$$\|x\|_{2\text{-ind}} := \sup_{x\neq 0} \frac{\|Ax\|_2}{\|x\|_2} \tag{1}$$

It readily follows that

$$\frac{\|Ax\|_2^2}{\|x\|_2^2} = \frac{x^*A^*Ax}{x^*x} \leq \lambda_{\max}(A^*A)$$

where $(\cdot)^*$ denotes complex conjugation and transposition. By choosing x to be the eigenvector corresponding to the largest eigenvalue of A^*A, the above upper bound is attained, and hence the 2-induced norm of A is equal to the square root of the largest eigenvalue of A^*A:

$$\|A\|_{2\text{-ind}} = \sqrt{\lambda_{\max}(A^*A)} \tag{2}$$

The $m \times n$ matrix A can also be considered as an element of the $(m+n)$-dimensional space $\mathbb{R}^{m+n}$. The Euclidean norm of A in this space is called the *Frobenius norm:*

$$\|A\|_F := \sqrt{\operatorname{trace}(A^*A)} = \sqrt{\sum_{i,j} A_{i,j}^2} \tag{3}$$

The Frobenius norm is not an induced norm. It satisfies $\|A\|_{2\text{-ind}} \leq \|A\|_F$.

The SVD and the Schmidt–Mirsky Theorem

Consider a rectangular matrix $A \in \mathbb{R}^{n\times m}$; let the eigenvalue decomposition of the symmetric matrices A^*A and AA^* be

$$A^*A = VS_VV^*, AA^* = US_UU^*$$

where U, V are (square) orthogonal matrices of size n, m, respectively (i.e., $UU^* = I_n$, $VV^* = I_m$). Furthermore S_V, S_U are diagonal, and assuming that $n \leq m$ we have

$$\begin{aligned} S_V &= \operatorname{diag}(\sigma_1^2, \ldots, \sigma_n^2, 0, \ldots, 0), \\ S_U &= \operatorname{diag}(\sigma_1^2, \ldots, \sigma_n^2), \qquad \sigma_i \geq \sigma_{i+1} \geq 0 \end{aligned}$$

Given the orthogonal matrices U, V and the nonnegative real numbers $\sigma_1, \ldots, \sigma_n$, we have

$$A = USV^*$$

where S is a matrix of size $n \times m$ with σ_i on the diagonal and zeros elsewhere. This is the *singular value decomposition* (*SVD*) of A; σ_i, $i = 1, \ldots, n$, are the *singular values* of A, while the columns u_i, v_i of U, V are the *left, right singular vectors* of A, respectively. As a consequence, the *dyadic decomposition* of A follows:

$$A = \sum_{i=1}^{n} \sigma_i u_i v_i^* \tag{4}$$

This is a decomposition in terms of rank one matrices; the rank of the sum of any k terms is k. In terms of the singular values the Frobenius norm of A defined by Eq. (3), is

$$\|A\|_F = \sqrt{\sigma_1^2 + \cdots + \sigma_n^2}$$

The following central problem can now be addressed.

PROBLEM 1. OPTIMAL LOW-RANK APPROXIMATION. Given the finite matrix A, find a matrix X of the same size but lower rank such tht the 2-induced norm of the error $E := A - X$ is minimized.

The solution of this problem is provided by the Schmidt–Mirsky theorem (see, e.g., Ref. 1, page 208):

Theorem 1. Schmidt–Mirsky. Given is the matrix A of rank n. For all matrices X of the same size and rank at most $k < n$, there holds:

$$\|A - X\|_{2\text{-ind}} \geq \sigma_{k+1}(A) \tag{5}$$

The lower bound $\sigma_{k+1}(A)$ of the error is attained by X_*, which is obtained by truncating the dyadic decomposition of A, to the leading k terms:

$$X_* := \sum_{i=1}^{k} \sigma_i u_i v_i^* \tag{6}$$

Remark 1. (a) The importance of this theorem lies in the fact that it establishes a relationship between the rank k of the approximant, and the $(k + 1)$st largest singular value of A.

(b) The minimizer X_* given above is not unique, since each member of the family of approximants

$$X(\eta_1, \ldots, \eta_k) := \sum_{i=1}^{k} (\sigma_i - \eta_i) u_i v_i^*, \qquad 0 \leq \eta_i \leq \sigma_{k+1}, i = 1, \ldots, k \tag{7}$$

attains the lower bound, namely $\sigma_{k+1}(A)$.

(c) The problem of minimizing the 2-induced norm of $A - X$ over all matrices X of rank at most k, is a nonconvex optimization problem, since the rank of the sum (or of a linear combination) of two rank k matrices is, in general, not k. Therefore, there is little hope of solving it using conventional optimization methods.

(d) If the error in the above approximation is measured in terms of the Frobenius norm, the lower bound in Eq. (5) is replaced by $\sqrt{\sigma_k^2+1 + \cdots + \sigma_n^2}$. In this case, X_* defined by Eq. (6) is the unique optimal approximant of A, which has rank k.

2-Norms and Induced 2-Norms in Infinite Dimensions

The 2-norm of the infinite sequence $x:\mathbb{Z} \to \mathbb{R}$ is

$$\|x\|_2 := \sqrt{\cdots + x(-1)^2 + x(0)^2 + x(1)^2 + \cdots}$$

The space of all sequences over $\mathbb{Z}$ which have finite 2-norm is denoted by $\ell_2(\mathbb{Z})$; ℓ_2 is known as the *Lebesgue space* of square-summable sequences. Similarly, for x defined over the negative or positive integers $\mathbb{Z}_-$, $\mathbb{Z}_+$, the corresponding spaces of sequences having finite 2-norm are denoted by $\ell_2(\mathbb{Z}_-)$, $\ell_2(\mathbb{Z}_+)$.

The 2-norm of the matrix sequence $X : \mathbb{Z} \to \mathbb{R}^{p\times m}$ is defined as

$$\|X\|_2 := \sqrt{\cdots + \|X(-1)\|_F^2 + \|X(0)\|_F^2 + \|X(1)\|_F^2 + \cdots}$$

The space of all $p \times m$ sequences having finite 2-norm is denoted by $\ell_2^{p\times m}(\mathbb{Z})$. Given the linear map $A : X \to Y$, where X, Y are subspaces of $\ell_2(\mathbb{Z})$, the norm induced by the 2-norm in the domain and range is the *2-induced* norm, defined by Eq. (1); again Eq. (2) holds.

Linear, Discrete-Time Systems

A linear, time-invariant, finite-dimensional, discrete-time system Σ is defined as follows:

$$\Sigma : \begin{array}{l} x(t+1) = Ax(t) + Bu(t) \\ y(t) = Cx(t) + Du(t) \end{array}, \qquad t \in \mathbb{Z}$$

$x(t) \in \mathbb{R}^n$ is the value of the *state* at time t, $u(t) \in \mathbb{R}^m$ and $y(t) \in \mathbb{R}^p$ are the values of the input and output at time t, respectively; and A, B, C, and D are constant maps. For simplicity, we will use the notation

$$\Sigma := \left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right) \in \mathbb{R}^{(n+p)\times(n+m)} \tag{8}$$

The *dimension* (or *order*) of the system is n: dim $\Sigma = n$. We will denote the unit step function by $\mathbb{I}$ ($\mathbb{I}(t) = 1$ for $t > 0$, and zero otherwise) and the Kronecker delta by δ ($\delta(0) = 1$, and zero otherwise). The *impulse response* h_Σ of Σ is

$$h_\Sigma(t) = CA^{t-1}B\mathbb{I}(t) + D\delta(t) \tag{9}$$

To Σ we associate the *convolution operator* $\mathcal{S}_\Sigma$ which maps inputs u into outputs y:

$$\mathcal{S}_\Sigma : u \longmapsto y, \qquad \text{where } y(t) = (h_\Sigma * u)(t) := \sum_{\tau=-\infty}^{t} h_\Sigma(t-\tau)u(\tau), t \in \mathbb{Z} \tag{10}$$

This convolution sum can also be written in matrix notation:

$$\begin{pmatrix} \vdots \\ y(-2) \\ y(-1) \\ \hline y(0) \\ y(1) \\ \vdots \end{pmatrix} = \underbrace{\left(\begin{array}{cccc|ccc} \ddots & & & & & & \\ & \cdots & h(0) & & & & \\ & \cdots & h(1) & h(0) & & & \\ \hline & \cdots & h(2) & h(1) & h(0) & & \\ & \cdots & h(3) & h(2) & h(1) & h(0) & \\ & & \vdots & \vdots & \vdots & \vdots & \ddots \end{array}\right)}_{\mathcal{S}_\Sigma} \begin{pmatrix} \vdots \\ u(-2) \\ u(-1) \\ \hline u(0) \\ u(1) \\ \vdots \end{pmatrix} \tag{11}$$

$\mathcal{S}_\Sigma$ has (block) *Toeplitz* and lower-triangular structure. The rational $p \times m$ matrix function

$$H_\Sigma(z) := \sum_{t=0}^{\infty} h_\Sigma(t)z^{-t} = C(zI - A)^{-1}B + D = \left[\frac{p_{ij}(z)}{q_{ij}(z)}\right], \quad 1 \le i \le p,\ 1 \le j \le m$$

is called the *transfer function* of Σ. The system is *stable* if the eigenvalues of A are inside the unit disk in the complex plane, that is, $|\lambda_i(A)| < 1$; this condition is equivalent to the impulse response being an ℓ_2 matrix sequence: $h_\Sigma \in \ell_2^{p\times m}(\mathbb{Z})$. If Σ is not stable, its impulse response does not have a finite 2-norm. However, if A has no eigenvalues on the unit circle, the impulse response can be reinterpreted so that it *does* have a finite 2-norm. This is done next. Let T be a basis change in the state space such that the matrices A, B, C can be partitioned as

$$A = \begin{pmatrix} A_+ & 0 \\ 0 & A_- \end{pmatrix}, \qquad B = \begin{pmatrix} B_+ \\ B_- \end{pmatrix}, \qquad C - (C_+ \quad C_-)$$

where the eigenvalues of A_+ and A_- are inside and outside of the unit disk, respectively. In contrast to Eq. (9), the ℓ_2-impulse response denoted by $h_{\Sigma 2}$ of Σ is defined as follows:

$$h_{\Sigma 2} := h_{2+} + h_{2-} \tag{12}$$

where

$$h_{2+}(t) := C_+A_+^tB_+\mathbb{I}(t) + \delta(t)D$$
$$h_{2-}(t) := C_-A_-^tB_-\mathbb{I}(-t)$$

where as before $\mathbb{I}$ is the unit step function and δ is the Kronecker symbol. Accordingly, we will write $\Sigma_2 = \Sigma_+ + \Sigma_-$. Notice that the algebraic expression for he transfer function remains the same in both cases:

$$H_{\Sigma 2}(z) = H_{2+}(z) + H_{2-}(z) = C_+(zI - A_+)^{-1}B_+ + D + C_-(zI - A_-)^{-1}B_- = H_\Sigma(z)$$

What is modified is the region of convergence of H_Σ from

$$|\lambda_{\max}(A_-)| < |z| \quad \text{to} \quad |\lambda_{\max}(A_+)| < |z| < |\lambda_{\min}(A_-)|$$

This is equivalent to trading the lack of stability (poles outside the unit disk) for the lack of causality ($h_{\Sigma 2}$ is nonzero for negative time). Thus a system with poles both inside and outside of the unit disk (but not *on* the unit circle) will be interpreted as a possibly antistable ℓ_2 system Σ_2, by defining the impulse response to be nonzero for negative time. Consequently $h_{\Sigma 2} \in \ell_2(\mathbb{Z})$. The matrix representation of the corresponding convolution operator is

$$\mathcal{S}_{\Sigma 2} = \left(\begin{array}{ccc|ccc} \ddots & \vdots & \vdots & \vdots & \vdots & \\ \cdots & h_{2+}(0) & h_{2-}(-1) & h_{2-}(-2) & h_{2-}(-3) & \cdots \\ \cdots & h_{2+}(1) & h_{2+}(0) & h_{2-}(-1) & h_{2-}(-2) & \cdots \\ \hline \cdots & h_{2+}(2) & h_{2+}(1) & h_{2+}(0) & h_{2-}(-1) & \cdots \\ \cdots & h_{2+}(3) & h_{2+}(2) & h_{2+}(1) & h_{2+}(0) & \cdots \\ & \vdots & \vdots & \vdots & \vdots & \ddots \end{array}\right) \tag{13}$$

Notice that $\mathcal{S}_{\Sigma 2}$ has block Toeplitz structure, but is no longer lower triangular. All discrete-time systems Σ with no poles on the unit circle will be interpreted as ℓ_2 systems Σ_2. For simplicity of notation, however, they will still be denoted by Σ. They are composed of two subsystems: (a) the stable and causal part Σ_+ and (b) the stable but anti-causal part Σ_-: $\Sigma = \Sigma_+ + \Sigma_-$.

For *square* ℓ_2 systems—that is, ℓ_2 systems having the same number of inputs and outputs $p = m$—the class of *all-pass* ℓ_2 systems is defined as follows: For all pairs u, y satisfying Eq. (10), there holds

$$\|y\|_2 = \alpha\|u\|_2 \tag{14}$$

for some fixed positive constant α. This is equivalent to

$$\mathcal{S}_\Sigma^*\mathcal{S}_\Sigma = \alpha^2 I \Leftrightarrow H_\Sigma^*(z^{-1})H_\Sigma(z) = \alpha^2 I_m, \quad |z| = 1$$

This last condition says that the transfer function (scaled by α) is a unitary matrix on the unit circle.

Approximation of Σ in the 2-Induced Norm of the Convolution Operator

Let Σ be an ℓ_2 system. The convolution operator $\mathcal{S}_\Sigma$ can be considered as a map:

$$\mathcal{S}_\Sigma : \ell_2^m(\mathbb{Z}) \to \ell_2^p(\mathbb{Z})$$

The 2-induced norm of Σ is defined as the 2-induced norm of $\mathcal{S}_\Sigma$:

$$\|\Sigma\|_{2\text{-ind}} := \|\mathcal{S}_\Sigma\|_{2\text{-ind}} = \sup_{u\neq 0}\frac{\|\mathcal{S}_\Sigma u\|_2}{\|u\|_2}$$

Due to the equivalence between the time and frequency domains (the Fourier transform is an isometric isomorphism), this norm can also be defined in the frequency domain. In particular,

$$\|\Sigma\|_{2\text{-ind}} = \sup_{|z|=1}\sigma_{\max}H_\Sigma(z) =: \|H_\Sigma\|_{\ell_\infty}$$

This latter quantity is known as the ℓ_∞ norm of H_Σ. If the system is single-input single-output, it is the supremum of the amplitude Bode plot. In the case where Σ is stable, H_Σ is analytic outside the unit disk, and the following holds:

$$\|\Sigma\|_{2\text{-ind}} = \sup_{|z|\geq 1}\sigma_{\max}H_\Sigma(z) =: \|H_\Sigma\|_{h_\infty}$$

This is known as the h_∞ norm of H_Σ. For simplicity, we will use the notation

$$\|\Sigma\|_{2\text{-ind}} := \|\mathcal{S}_\Sigma\|_{2\text{-ind}} = \|H_\Sigma\|_\infty \tag{15}$$

It will be clear from the context whether the subscript ∞ stands for the ℓ_∞ norm of the h_∞ norm. If Σ is all-pass—that is, Eq. (14) is satisfied—then $\|\Sigma\|_{2\text{-ind}} = \|H_\Sigma\|_\infty = \alpha$.

Our aim is the generalization of the Schmidt–Mirsky result for an appropriately defined operator. It is most natural to explore the possibility of optimal approximation of Σ in the 2-induced norm of the convolution operator $\mathcal{S}_\Sigma$. In this regard the following result holds.

Proposition 1. Let $\sigma_* := \inf_{|z|=1}\ \sigma_{min}(H_\Sigma(z))$, while $\sigma^* := \sup_{|z|=1}\sigma_{max}(H_\Sigma(z))$. Every σ in the interval $[\sigma_*, \sigma^*]$ is a singular value of the operator $\mathcal{S}_\Sigma$.

We conclude that since the singular values of $\mathcal{S}_\Sigma$ form a continuum, it is not a priori clear how the Schmidt–Mirsky result might be generalized to the convolution operator of Σ.

Remark 2. Despite the above conclusion, the approach discussed below yields as byproduct the solution of this problem in the special case of the one-step model reduction; see Remark 4(b).

Approximation of Σ in the 2-Induced Norm of the Hankel Operator

The next attempt to address this approximation problem is by defining a *different* operator attached to the system Σ. Recall that $\ell_2^m(\mathcal{I})$ denotes the space of square-summable sequences of vectors, defined on the interval $\mathcal{I}$, with entries in $\mathbb{R}^m$. Given the *stable* and *causal* system Σ, the following operator is defined by restricting the domain and the range of the convolution operator, Eq. (10):

$$\mathcal{H}_\Sigma : \ell^m(\mathbb{Z}_-) \longrightarrow \ell^p(\mathbb{Z}_+)$$
$$u_- \longmapsto y_+, \quad \text{where } y_+(t) = \sum_{\tau=-\infty}^{-1} h_\Sigma(t-\tau)u_-(\tau), \quad t \geq 0 \tag{16}$$

$\mathcal{H}_\Sigma$ is called the *Hankel operator* of Σ. Its matrix representation in the canonical bases is

$$\begin{pmatrix} y(0)\\ y(1)\\ y(2)\\ \vdots \end{pmatrix} = \underbrace{\begin{pmatrix} h(1) & h(2) & h(3) & \cdots\\ h(2) & h(3) & h(4) & \cdots\\ h(3) & h(4) & h(5) & \cdots\\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}}_{\mathcal{H}_\Sigma} \begin{pmatrix} u(-1)\\ u(-2)\\ u(-3)\\ \vdots \end{pmatrix} \tag{17}$$

Thus the Hankel operator of Σ maps *past* inputs into *future* outputs. It has a number of properties given next. The first one for single-input single-output systems is due to Kronecker.

Proposition 2. Given the system Σ defined by Eq. (8), the rank of H_Σ is at most n. The rank is exactly n if, and only if, the system Σ is reachable and observable. Furthermore, if Σ is stable, $\mathcal{H}_\Sigma$ has a finite set of nonzero singular values.

In order to compute the singular values of the Hankel operator, we define the *reachability* matrix $\mathcal{R}$ and the *observability* matrix $\mathcal{O}$:

$$\mathcal{R}(A,B) = [B \quad AB \quad A^2B \quad \cdots \quad], \qquad \mathcal{O}(C,A) = [\mathcal{R}(A^*,C^*)]^*$$

both of these matrices have rank at most n. The *reachability* and *observability grammians* are

$$\mathcal{P} := \mathcal{R}\mathcal{R}^* = \sum_{t\geq 0} A^t BB^*(A^*)^t, \qquad \mathcal{Q} := \mathcal{O}^*\mathcal{O} = \sum_{t\geq 0}(A^*)^t C^*CA^t \tag{18}$$

The quantities $\mathcal{P}$ and $\mathcal{Q}$ are $n \times n$ symmetric, positive semidefinite matrices. They are positive definite if and only if Σ is reachable and observable. The grammians are (the unique) solutions of the linear matrix equations:

$$A\mathcal{P}A^* + BB^* = \mathcal{P}, \qquad A^*\mathcal{Q}A + C^*C = \mathcal{Q} \tag{19}$$

which are called *discrete-time Lyapunov* or *Stein* equations.

The (nonzero) singular values σ_i of $\mathcal{H}_\Sigma$ are the square roots of the (nonzero) eigenvalues λ_i of $\mathcal{H}_\Sigma^*\mathcal{H}_\Sigma$. The key to this computation is the fact that the Hankel operator can be factored in the product of the observability and the reachability matrices:

$$\mathcal{H}_\Sigma = \mathcal{O}(C,A)\mathcal{R}(A,B)$$

For $u_i \in \ell_2^m(\mathbb{Z}_-)$, $u_i \neq 0$, there holds

$$\begin{aligned}\mathcal{H}_\Sigma^*\mathcal{H}_\Sigma u_i &= \sigma_i^2 u_i \Leftrightarrow \mathcal{R}^*\mathcal{O}^*\mathcal{O}\mathcal{R}u_i \\ &= \sigma_i^2 u_i \Leftrightarrow \underbrace{\mathcal{R}\mathcal{R}^*}_{\mathcal{P}}\underbrace{\mathcal{O}^*\mathcal{O}}_{\mathcal{Q}}\mathcal{R}u_i = \sigma_i^2\mathcal{R}u_i\end{aligned}$$

Thus, u_i is an eigenfunction of $\mathcal{H}_\Sigma^*\mathcal{H}_\Sigma$ corresponding to the nonzero eigenvalue σ_i^2 iff $\mathcal{R}\,u_i \neq 0$ is an eigenfunction of the product of the grammians $\mathcal{P}\mathcal{Q}$:

$$\sigma_i^2(\mathcal{H}_\Sigma) = \lambda_i(\mathcal{H}_\Sigma^*\mathcal{H}_\Sigma) = \lambda_i(\mathcal{P}\mathcal{Q}) \tag{20}$$

Proposition 3. The nonzero singular values of the Hankel operator $\mathcal{H}_\Sigma$ associated with the stable system Σ are the square roots of the eigenvalues of the product of the grammians $\mathcal{P}\mathcal{Q}$.

Definition 1. The *Hankel singular values* of the stable system Σ as in Eq. (10), denoted by

$\sigma_1(\Sigma) > \cdots > \sigma_q(\Sigma)$ with multiplicity

$$r_i, i = 1, \ldots, q, \sum_{i=1}^{q} r_i = n \tag{21}$$

are the singular values of $\mathcal{H}_\Sigma$ defined by Eq. (16). The *Hankel norm* of Σ is the largest Hankel singular value:

$$\|\Sigma\|_H := \sigma_1(\Sigma)$$

The Hankel operator of a not necessarily stable ℓ_2 system Σ is defined as the Hankel operator of its stable and causal part $\Sigma_+ : \mathcal{H}_\Sigma := \mathcal{H}_{\Sigma+}$.

Thus, the Hankel norm of a system having poles both inside and outside the unit circle is defined to be the Hankel norm of its causal and stable part. In general,

$$\|\Sigma\|_{\text{2-ind}} \geq \|\Sigma\|_H \tag{22}$$

An important property of the Hankel singular values of Σ is that twice their sum provides an upper bound for the 2-induced norm of Σ (see also the section entitled "Error Bounds of Optimal and Suboptimal Approximants").

Lemma 1. Given the stable system Σ with Hankel singular values σ_i, $i = 1, \ldots, n$ (multiplicities included), the following holds true: $\|\Sigma\|_{\text{2-ind}} \leq 2(\sigma_1 + \cdots + \sigma_n)$.

The AAK Theorem. Consider the stable systems Σ and Σ' of dimension n and k, respectively. By Proposition 2, $\mathcal{H}_\Sigma$ has rank n and $\mathcal{H}_{\Sigma'}$ has rank k. Therefore, the Schmidt–Mirsky theorem implies that

$$\|\mathcal{H}_\Sigma - \mathcal{H}_{\Sigma'}\|_{\text{2-ind}} \geq \sigma_{k+1}(\mathcal{H}_\Sigma) \tag{23}$$

The question which arises is to find the infimum of the above norm, given the fact that the approximant is structured (block Hankel matrix): $\inf_{\Sigma'}\|\mathcal{H}_\Sigma - \mathcal{H}_{\Sigma'}\|_{\text{2-ind}}$. A remarkable result due to Adamjan, Arov, and Krein, code-named *AAK result,* asserts that this lower bound is indeed attained for some Σ' or dimension k. The original sources for this result are Refs. 2 and 3.

Theorem 2. AAK Theorem. Given the $\ell_2^{p\times m}(\mathbb{Z}_+)$ sequence of matrices $h = (h(t))_{t>0}$, such that the associated Hankel matrix $\mathcal{H}$ has finite rank n, there exists an $\ell_2^{p\times m}(\mathbb{Z}_+)$ sequence of matrices $h_* = (h_*(t))_{t>0}$, such that the associated Hankel matrix $\mathcal{H}_*$ has rank k and in addition

$$\|\mathcal{H} - \mathcal{H}_*\|_{\text{2-ind}} = \sigma_{k+1}(\mathcal{H}) \tag{24}$$

If $p = m = 1$, the optimal approximant is unique.

The result says that every stable and causal system Σ can be optimally approximated by a stable and causal system Σ_* of lower dimensions; the optimality is with respect to the 2-induced norm of the associated Hankel operator (see Fig. 1).

The Main Result. In this section we will present the main result. As it turns out, one can consider both suboptimal and optimal approximants within the same framework. Actually, as shown in the sections entitled "State-Space Construction for Square Systems: Suboptimal Case" and "State-Space Construction for Square Systems: Optimal Case," the formulas for suboptimal approximants are simpler than their optimal counterparts.

PROBLEM 2. Given a stable system Σ, we seek approximants Σ_* satisfying

$$\sigma_{k+1}(\Sigma) \leq \|\Sigma - \Sigma_*\|_H \leq \epsilon < \sigma_k(\Sigma)$$

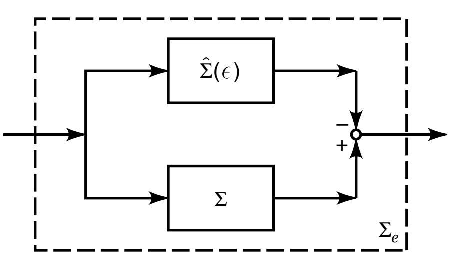


Figure 1. Construction of approximants.

This is a generalization of Problem 1, as well as the problem solved by the AAK theorem. The concept introduced in the next definition is the key to its solution.

Definition 2. Let Σ_e be the parallel connection of Σ and $\hat{\Sigma}$: $\Sigma_e := \Sigma - \hat{\Sigma}$. If Σ_e is an all-pass system with norm ϵ, $\hat{\Sigma}$ is called an ϵ-all-pass dilation of Σ.

As a consequence of the inertia result of the section entitled "The Grammians, Lyapunov Equations, and an Inertia Result," the all-pass dilation system has the following crucial property.

Main Lemma 1. Let $\hat{\Sigma}$ be an ϵ-all-pass dilation of Σ, where ϵ satisfies Eq. (25). It follows that $\hat{\Sigma}$ has exactly k poles inside the unit disk, that is, $\dim \Sigma_+ = k$.

We also restate the analog of the Schmidt–Mirsky result [Eq. (23)], applied to dynamical systems:

Proposition 4. Given the stable system Σ, let Σ' have at most k poles inside the unit disk. Then

$$\|\Sigma - \Sigma'\|_H \geq \sigma_{k+1}(\Sigma)$$

This means that the 2-induced norm of the Hankel operator of the difference between Σ and Σ' is no less than the $(k+1)$st singular value of the Hankel operator of Σ. Finally, recall that if a system has both stable and unstable poles, its Hankel norm is that of its stable part. We are now ready for the main result which is valid for both discrete- and continuous-time systems.

Theorem 3. Let $\hat{\Sigma}$ be an ϵ-all-pass dilation of the linear, stable, discrete- or continuous-time system Σ, where

$$\sigma_{k+1}(\Sigma) \leq \epsilon < \sigma_k(\Sigma) \tag{25}$$

It follows that $\hat{\Sigma}_+$ has exactly k stable poles and consequently

$$\sigma_{k+1}(\Sigma) \leq \|\Sigma - \hat{\Sigma}\|_H < \epsilon \tag{26}$$

In case $\sigma_{k+1}(\Sigma) = \epsilon$,

$$\sigma_{k+1}(\Sigma) = \|\Sigma - \hat{\Sigma}\|_H$$

Proof. The result is a consequence of the following sequence of equalities and inequalities:

$$\sigma_{k+1}(\Sigma) \leq \|\Sigma - \hat{\Sigma}_+\|_H = \|\Sigma - \hat{\Sigma}\|_H \leq \|\Sigma - \hat{\Sigma}\|_\infty = \epsilon$$

The first inequality on the left side is a consequence of Main Lemma 1, the equality follows by definition, the second inequality follows from Eq. (22), and the last equality holds by construction, since $\Sigma - \hat{\Sigma}$ is ϵ-all-pass.

Remark 3. (a) For $\epsilon = \sigma_1(\Sigma)$, the above theorem yields the solution of the *Nehari problem,* namely to find the best anti-stable approximant of Σ in the 2-induced norm of the convolution operator (i.e., the ℓ_∞ norm).

(b) We are given a stable system Σ and seek to compute an approximant in the same class (i.e., stable). In order to achieve this, the construction given above takes us outside this class of systems, since the all-pass dilation system $\hat{\Sigma}$ has poles both inside and outside the unit circle. In terms of matrices, we start with a system whose convolution operator $\mathcal{S}_\Sigma$ is a (block) lower triangular Toeplitz matrix. We then compute a (block) Toeplitz matrix $\mathcal{S}_{\hat{\Sigma}}$, which is no longer lower triangular, such that the difference $\mathcal{S}_\Sigma - \mathcal{S}_{\hat{\Sigma}}$ is unitary. It then follows that the lower left-hand portion of $\mathcal{S}_{\hat{\Sigma}}$, which is the Hankel matrix $\mathcal{H}_{\hat{\Sigma}}$, has rank r and approximates the Hankel matrix $\mathcal{H}_\Sigma$, so that the 2-norm of the error satisfies Eq. (25).

(c) The suboptimal and optimal approximants can be constructed using explicit formulae. For continuous-time systems, see the section entitled "Construction Formulas for Hankel-Norm Approximants."

CONSTRUCTION OF APPROXIMANTS

The purpose of this section is to present, and to a certain extent derive, formulas for suboptimal and optimal approximants in the Hankel norm. Because of Theorem 3, all we need is the ability to construct all-pass dilations of a given system. To this goal, the first subsection is dedicated to the presentation of important aspects of the theory of linear, continuous-time systems; these facts are used in the second subsection. The closely related approach to system approximation by balanced truncation is briefly discussed in the section entitled "Balanced Realizations and Balanced Model Reduction."

Linear, Continuous-Time Systems

$\mathcal{L}_2$ Linear Systems. For continuous-time functions, let

$$\mathcal{L}^n(\mathcal{I}) := \{f : \mathcal{I} \to \mathbb{R}^n, \mathcal{I} \subset \mathbb{R}\}$$

Frequent choices of $\mathcal{I}$: $\mathcal{I} = \mathbb{R}$, $\mathcal{I} = \mathbb{R}_+$ or $\mathcal{I} = \mathbb{R}_-$. The 2-norm of a function f is

$$\|f\|_2 := \left(\int_{t\in\mathcal{I}} \|f(t)\|_2^2\, dt\right)^{1/2}, \qquad f \in \mathcal{L}^n(\mathcal{I})$$

The corresponding $\mathcal{L}_2$ space of square-integrable functions is

$$\mathcal{L}_2^n(\mathcal{I}) := \{f \in \mathcal{L}^n(\mathcal{I}),\ \|f\|_2 < \infty\}$$

The 2-norm of the matrix function $F : \mathcal{I} \to \mathbb{R}^{p\times m}$, is defined as

$$\|F\|_2 := \left(\int_{t\,\in\,\mathcal{I}} \|F(t)\|_F^2\, dt\right)^{1/2}$$

where the subscript "F" denotes the Frobenius norm. The space of all $p \times m$ matrix functions having finite 2-norm is denoted by $\mathcal{L}_2^{p\times m}(\mathcal{I})$. Let $A : X \to Y$, where X and Y are subspaces of $\mathcal{L}_2^q(\mathbb{R})$, for some q. The 2-induced norm of A is defined as in Eq. (1), and Eq. (2) holds true.

We will consider linear, finite-dimensional, time-invariant, continuous-time systems described by the following set of differential and algebraic equations:

$$\Sigma: \quad \begin{aligned} \frac{dx(t)}{dt} &= Ax(t) + Bu(t), \\ y(t) &= Cx(t) + Du(t) \end{aligned} \qquad t \in \mathbb{R}$$

where $u(t) \in \mathbb{R}^m$, $x(t) \in \mathbb{R}^n$, and $y(t) \in \mathbb{R}^p$ are the values of the input, state, and output at time t, respectively. The system will be abbreviated as

$$\Sigma = \left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right) \in \mathbb{R}^{(n+p)\times(n+m)} \tag{27}$$

In analogy to the discrete-time case, we define the continuous-time unit step function $\mathbb{I}(t)$ ($\mathbb{I}(t) = 1$ for $t \geq 0$, and zero otherwise) and the delta distribution δ. The impulse response of this system is

$$h_\Sigma(t) = Ce^{At}B\mathbb{I}(t) + \delta(t)D \tag{28}$$

while the transfer function is

$$H_\Sigma(s) = C(sI - A)^{-1}B + D$$

Unless A has all its eigenvalues in the left half-plane (LHP), the impulse response h_Σ is not square-integrable—that is, does not belong to the space $\mathscr{L}_2^{p\times m}(\mathbb{R})$. In this case we shall say that the system is not an $\mathscr{L}_2$ system. If, however, A has no eigenvalues on the $j\omega$ axis, it can be interpreted as an $\mathscr{L}_2$ system by appropriate redefinition of the impulse response. As in the discrete-time case, let there be a state transformation such that

$$A = \begin{pmatrix} A_+ & 0 \\ 0 & A_- \end{pmatrix}, \qquad B = \begin{pmatrix} B_+ \\ B_- \end{pmatrix}, \qquad C = (C_+ \quad C_-)$$

where the eigenvalues of A_+ and A_- are in the LHP and right half-plane (RHP), respectively. The $\mathscr{L}_2$-impulse response is defined as follows:

$$h_{\Sigma 2} := h_{2+} + h_{2-} \tag{29}$$

where

$$\begin{aligned} h_{2+}(t) &:= C_+e^{A_+t}B_+\mathbb{I}(t) + \delta(t)D \\ h_{2-}(t) &:= C_-e^{A_-t}B_-\mathbb{I}(-t) \end{aligned}$$

As in the discrete-time case,

$$\begin{aligned} H_{\Sigma 2}(s) = H_{2+}(s) + H_{2-}(s) &= C_+(sI - A_+)^{-1}B_+ \\ &+ D + C_-(sI - A_-)^{-1}B_- = H_\Sigma(s) \end{aligned}$$

This is equivalent to trading the lack of stability (poles in the RHP) to the lack of causality ($h_{\Sigma 2}$ is nonzero for negative time). What is modified is the region of convergence from

$$\mathscr{Re}(\lambda_{\max}(A_-)) < \mathscr{Re}(s) \quad \text{to} \quad \mathscr{Re}(\lambda_{\max}(A_+)) < \mathscr{Re}(s) < \mathscr{R}(\lambda_{\min}(A_-))$$

From now on, all continuous-time systems Σ with no poles on the imaginary axis will be interpreted as $\mathscr{L}_2$ systems; to keep the notation simple, they will be denoted by Σ instead of Σ_2. The stable and causal subsystem will be denoted by Σ_+, and the stable but anti-causal one will be denoted by Σ_-: $\Sigma = \Sigma_+ + \Sigma_-$.

The Convolution Operator. The convolution operator associated to Σ is defined as follows:

$$\mathscr{S}_\Sigma : u \longmapsto y,$$
$$\text{where } y(t) = (h_\Sigma * u)(t) := \int_{-\infty}^{\infty} h_\Sigma(t-\tau)u(\tau)\,d\tau, \qquad t \in \mathbb{R} \tag{30}$$

Corresponding to Proposition 1 we have the following:

Proposition 5. Let $\sigma_* := \inf_{\omega\in\mathbb{R}} \sigma_{\min}(H_\Sigma(j\omega))$, while $\sigma^* := \sup_{\omega\in\mathbb{R}} \sigma_{\max}(H_\Sigma(j\omega))$. Every σ in the interval $[\sigma_*, \sigma^*]$ is a singular value of the operator $\mathscr{S}_\Sigma$.

Since the singular values of $\mathscr{S}_\Sigma$ form a continuum, the same conclusion as in the discrete-time case follows (see also Remark 2). Again, the 2-induced norm of Σ turns out to be the infinity norm of the transfer function H_Σ:

$$\|\Sigma\|_{2\text{-ind}} := \|\mathscr{S}_\Sigma\|_{2\text{-ind}} = \|H_\Sigma\|_\infty \tag{31}$$

If Σ is all-pass [Eq. (14)], then $\|\Sigma\|_{2\text{-ind}} = \|H_\Sigma\|_\infty = \alpha$.

The Grammians, Lyapunov Equations, and an Inertia Result. For stable systems (i.e., $\mathscr{R}e(\lambda_i(A)) < 0$), the following quantities are defined:

$$\mathscr{P} := \int_0^\infty e^{At}BB^*e^{A^*t}\,dt, \qquad \mathscr{Q} := \int_0^\infty e^{A^*t}C^*Ce^{At}\,dt \tag{32}$$

They are called the *reachability* and *observability grammians* of Σ, respectively. By definition, $\mathscr{P}$ and $\mathscr{Q}$ are positive semidefinite. It can be shown that reachability and observability of Σ are equivalent to the positive definiteness of each one of these grammians. As in the discrete-time case, the grammians are the (unique) solutions of the linear matrix equations

$$A\mathscr{P} + \mathscr{P}A^* + BB^* = 0, \qquad A^*\mathscr{Q} + \mathscr{Q}A + C^*C = 0 \tag{33}$$

which are known as the continuous-time *Lyapunov equations* [cf. Eq. (19)]. Such equations have a remarkable property known as *inertia* result: There is a relationship between the number of eigenvalues of A and (say) $\mathscr{P}$ in the LHP, RHP, and the imaginary axis. More precisely, the *inertia* of $A \in \mathbb{R}^{n\times n}$ is

$$\text{in}\,(A) := \{\nu(A), \delta(A), \pi(A)\} \tag{34}$$

where $\nu(A)$, $\delta(A)$, and $\pi(A)$ are the number of eigenvalues of A in the LHP, on the imaginary axis and in the RHP, respectively.

Proposition 6. Let A and $X = X^*$ satisfy the Lyapunov equation: $AX + XA^* = R$, where $R \geq 0$. If the pair (A, R) is reachable, then the inertia of A is equal to the inertia of X: $\text{in}(A) = \text{in}(X)$.

Main Lemma 1 is based on this inertia result.

Remark 3. One of the reasons why continuous-time formulas are simpler than their discrete-time counterparts is that the Lyapunov equations [Eqs. (34)] are linear in A, while the discrete-time Lyapunov equations [Eqs. (19)] are quadratic in A.

All-Pass $\mathscr{L}_2$ Systems. *All-pass* systems are a subclass of $\mathscr{L}_2$ systems. As indicated in Theorem 3, all-pass systems play an important role in Hankel norm approximation. The following characterization is central for the construction of suboptimal and optimal approximants (see sections entitled “State-Space Construction for Square Systems: Suboptimal Case” and “State-Space Construction for Square Systems: Optimal Case”). The proof is by direct computation; for details see Ref. 7.

Proposition 7. Given is Σ as in Eq. (27) with $p = m$. The following statements are equivalent.

- Σ is α-all-pass.
- For all input–output pairs (u, y) satisfying $y = h_\Sigma * u$, there holds: $\|y\|_2 = \alpha\|u\|_2$.
- $H_\Sigma^*(-j\omega)H_\Sigma(j\omega) = \alpha^2 I_m$
- There exists $\mathscr{Q} = \mathscr{Q}^* \in \mathbb{R}^{n\times n}$, such that the following equations are satisfied:

$$\begin{aligned} A^*\mathscr{Q} + \mathscr{Q}A + C^*C &= 0 \\ \mathscr{Q}B + C^*D &= 0 \\ D^*D &= \alpha^2 I_m \end{aligned} \tag{35}$$

- The solutions $\mathscr{P}$ and $\mathscr{Q}$ of the Lyapunov equations

$$A\mathscr{P} + \mathscr{P}A^* + BB^* = 0, \qquad A^*\mathscr{Q} + \mathscr{Q}A + C^*C = 0$$

  satisfy $\mathscr{P}\mathscr{Q} = \alpha^2 I_n$, and in addition we have $D^*D = \alpha^2 I_m$.

The Continuous-Time Hankel Operator. In analogy to the discrete-time case, we define the operator $\mathscr{H}_\Sigma$ of the stable system Σ which maps *past inputs* into *future outputs:*

$$\begin{aligned} &\mathscr{H}_\Sigma : \mathscr{L}_2^m(\mathbb{R}_-) \longrightarrow \mathscr{L}_2^p(\mathbb{R}_+) \\ &\qquad u_- \longmapsto y_+, \text{ where } y_+(t) = \int_{-\infty}^0 h_\Sigma(t-\tau)u_-(\tau)\,d\tau, \\ &\qquad\qquad t \geq 0 \end{aligned} \tag{36}$$

$\mathscr{H}_\Sigma$ is called the *Hankel operator* of Σ. Unlike in the discrete-time case, however (see Eq. (17)), $\mathscr{H}_\Sigma$ has *no* matrix representation.

It turns out that just as in Eq. (20), it can be shown that the nonzero singular values of the continuous-time Hankel operator are the eigenvalues of the product of the two grammians:

$$\sigma_i^2(\mathscr{H}_\Sigma) = \lambda_i(\mathscr{H}_\Sigma^*\mathscr{H}_\Sigma) = \lambda_i(\mathscr{P}\mathscr{Q}) \tag{37}$$

The Hankel singular values of $\mathscr{H}_\Sigma$ and their multiplicities will be denoted as in Eq. (21). Finally,

$$\|\Sigma\|_{\mathrm{H}} = \|\Sigma_+\|_{\mathrm{H}} \leq \|\Sigma\|_{\text{2-ind}}$$

Lemma 1 can be strengthened in the continuous-time case (see also the section entitled “Error Bounds of Optimal and Suboptimal Approximants”).

Lemma 2. Let the *distinct* Hankel singular values of the stable system Σ be σ_i, $i = 1, \ldots, q$; following holds: $\|\Sigma\|_{\text{2-ind}} \leq 2(\sigma_1 + \cdots + \sigma_q)$.

A Transformation Between Continuous- and Discrete-Time Systems. As mentioned in the introductory paragraphs of this article, the formulas for the construction of approximants will be given for the continuous-time case because they are generally simpler than their discrete-time counterparts. Thus, in order to apply the formulas to discrete-time systems, the system will have to be transformed to continuous-time first, and at the end the continuous-time optimal or suboptimal approximants obtained will have to be transformed back to discrete-time approximants.

One transformation between continuous- and discrete-time systems is given by the bilinear transformation $z = (1 + s)/(1 - s)$ of the complex plane onto itself. The resulting relationship between the transfer function $H_c(s)$ of a continuous-time system and that of the corresponding discrete-time system $H_d(z)$ is

$$H_c(s) = H_d\left(\frac{1+s}{1-s}\right)$$

The state space maps

$$\Sigma_c := \left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right), \qquad \Sigma_d := \left(\begin{array}{c|c} F & G \\ \hline H & J \end{array}\right)$$

are related as given in Table 1. Furthermore, the proposition that follows states that the Hankel and infinity norms remain unchanged by this transformation.

Proposition 8. Given the stable continuous-time system Σ_c with grammians $\mathscr{P}_c$ and $\mathscr{Q}_c$, let Σ_d with grammians $\mathscr{P}_d$ and $\mathscr{Q}_d$, be the discrete-time system obtained by means of the transformation given above. It follows that $\mathscr{P}_c = \mathscr{P}_d$ and $\mathscr{Q}_c = \mathscr{Q}_d$. Furthermore, this bilinear transformation also preserves the infinity norms (i.e., the 2-induced norms of the associated convolution operators): $\|\Sigma_c\|_\infty = \|\Sigma_d\|_\infty$.

Table 1. Transformation Formulas

Continuous-Time		Discrete-Time
A, B, C, D	$z = \dfrac{1+s}{1-s}$	$F = (I+A)(I-A)^{-1}$ $G = \sqrt{2}(I-A)^{-1}B$ $H = \sqrt{2}C(I-A)^{-1}$ $J = D + C(I-A)^{-1}B$
$A = (F+I)^{-1}(F-I)$ $B = \sqrt{2}(F+I)^{-1}G$ $C = \sqrt{2}H(F+I)^{-1}$ $D = J - H(F+I)^{-1}G$	$s = \dfrac{z-1}{z+1}$	F, G, H, J

Construction Formulas for Hankel-Norm Approximants

We are ready to give some of the formulas for the construction of suboptimal and optimal Hankel-norm approximants. As mentioned earlier, all formulas describe the construction of all-pass dilation systems [see Eq. (3)]. We will concentrate on the following cases (in increasing degree of complexity):

- An input–output construction applicable to scalar systems. Both optimal and suboptimal approximants are treated. The advantage of this approach is that the equations can be set up in a straightforward manner using the numerator and denominator polynomials of the transfer function of the given system (see the section entitled "Input–Output Construction Method for Scalar Systems").
- A state-space-based construction method for suboptimal approximants (see the section entitled "State-Space Construction for Square Systems: Suboptimal Case") and for optimal approximants (see the section entitled "State-Space Construction for Square Systems: Optimal Case") of square systems.
- A state-space-based parameterization of *all* suboptimal approximants for general (i.e., not necessarily square) systems (see the section entitled "General Case: Parameterization of All Suboptimal Approximants").
- The optimality of the approximants is with respect to the Hankel norm (2-induced norm of the Hankel operator). The section entitled "Error Bounds of Optimal and Suboptimal Approximants" gives an account of error bounds for the infinity norm of approximants (2-induced norm of the convolution operator).
- The section entitled "Balanced Realizations and Balanced Model Reduction" discusses model reduction by balanced truncation which uses the same ingredients as the Hankel norm model reduction theory. The approximants have a number of interesting properties, including the existence of error bounds for the infinity norm of the error. However, no optimality holds.

Input–Output Construction Method for Scalar Systems. Given the polynomials $a = \sum_{i=0}^{\alpha} a_i s^i$, $b = \sum_{i=0}^{\beta} b_i s^i$, and $c = \sum_{i=0}^{\gamma} c_i s^i$ satisfying $c(s) = a(s)b(s)$, the coefficients of the product c are a linear combination of those of b:

$$\boldsymbol{c} = \mathbb{T}(a)\boldsymbol{b}$$

where $\boldsymbol{c} := (c_\gamma \quad c_{\gamma-1} \ . \ . \ . \ c_1 \quad c_0)^* \in \mathbb{R}^{\gamma+1}$, $\boldsymbol{b} := (b_\beta \ . \ . \ . \ b_0)^* \in \mathbb{R}^{\beta+1}$, and $\mathbb{T}(a)$ is a Toeplitz matrix with first column $(a_\alpha \ . \ . \ . \ a_0 \quad 0 \ . \ . \ . \ 0)^* \in \mathbb{R}^{\gamma+1}$, and first row $(a_\alpha \quad 0 \ . \ . \ . \ 0) \in \mathbb{R}^{1\times(\beta+1)}$. We will also define the sign matrix

$$\mathbb{K} = \mathrm{diag}(\qquad \ldots \qquad , 1, -1, 1)$$

of appropriate size. Given a polynomial a with real coefficients, the polynomial c^* is defined as $c(s)^* := c(-s)$. This means that

$$\boldsymbol{c}^* = \mathbb{K}\boldsymbol{c}$$

The basic construction given in Theorem 3 hinges on the construction of an ϵ-all-pass dilation $\hat{\Sigma}$ of Σ. Let

$$H_\Sigma(s) = \frac{p(s)}{q(s)}, \qquad H_{\hat{\Sigma}}(s) = \frac{\hat{p}(s)}{\hat{q}(s)}$$

We require that the difference $H_\Sigma - H_{\hat{\Sigma}} = H_{\Sigma_e}$ be ϵ-all-pass. Therefore, the problem is as follows: Given ϵ and the polynomials p and q such that $\deg(p) \le \deg(q) := n$, find polynomials $\hat{p}$ and $\hat{q}$ of degree at most n such that

$$\frac{p}{q} - \frac{\hat{p}}{\hat{q}} = \epsilon \frac{q^*\hat{q}^*}{q\hat{q}} \Leftrightarrow p\hat{q} - q\hat{p} = \epsilon q^* \hat{q}^* \tag{38}$$

This polynomial equation can be rewritten as a matrix equation involving the quantities defined above:

$$\mathbb{T}(p)\hat{\boldsymbol{q}} - \mathbb{T}(q)\hat{\boldsymbol{p}} = \epsilon\mathbb{T}(q^*)\hat{\boldsymbol{q}}^* = \epsilon\mathbb{T}(q^*)\mathbb{K}\hat{\boldsymbol{q}}$$

Collecting terms we have

$$(\mathbb{T}(p) - \epsilon\mathbb{T}(q^*)\mathbb{K}, \qquad -\mathbb{T}(q))\begin{pmatrix}\hat{\boldsymbol{q}}\\ \hat{\boldsymbol{p}}\end{pmatrix} = 0 \tag{39}$$

The solution of this set of linear equations provides the coefficients of the ϵ-all pass dilation system $\hat{\Sigma}$. Furthermore, this system can be solved for both the suboptimal $\epsilon \ne \sigma_i$ and the optimal $\epsilon = \sigma_i$ cases. We will illustrate the features of this approach by means of a simple example. For an alternative approach along similar lines, see Ref. 9.

Example. Let Σ be a second-order system, that is, $n = 2$. If we normalize the coefficient of the highest power of $\hat{q}$, that is, $\hat{q}_2 = 1$, we obtain the following system of equations:

$$\underbrace{\begin{pmatrix} 0 & 0 & q_2 & 0 & 0 \\ p_2 - \epsilon q_2 & 0 & q_1 & q_2 & 0 \\ p_1 + \epsilon q_1 & p_2 + \epsilon q_2 & q_0 & q_1 & q_2 \\ p_0 - \epsilon q_0 & p_1 - \epsilon q_1 & 0 & q_0 & q_1 \\ 0 & p_0 + \epsilon q_0 & 0 & 0 & q_0 \end{pmatrix}}_{W(\epsilon)} \begin{pmatrix} \hat{q}_1 \\ \hat{q}_0 \\ \hat{p}_2 \\ \hat{p}_1 \\ \hat{p}_0 \end{pmatrix} = -\begin{pmatrix} p_2 + \epsilon q_2 \\ p_1 - \epsilon q_1 \\ p_0 + \epsilon q_0 \\ 0 \\ 0 \end{pmatrix}$$

This can be solved for all ϵ which are not roots of the equation $\det W(\epsilon) = 0$. The latter is a polynomial equation of second degree; there are thus two values of ϵ, ϵ_1, and ϵ_2, for which the determinant of W is zero. It can be shown that the roots of this determinant are the *eigenvalues* of the Hankel operator $\mathcal{H}_\Sigma$; since in the single-input single-output case $\mathcal{H}_\Sigma$ is self-adjoint (symmetric), the absolute values of ϵ_1 and ϵ_2, are the singular values of $\mathcal{H}_\Sigma$. Thus both suboptimal and optimal approximants can be computed this way. (See also the first example of the section entitled "Examples.")

State-Space Construction for Square Systems: Suboptimal Case. In this section we will discuss the construction of approximants in the case where $m = p$. Actually this is no loss of generality because if one has to apply the algorithm to a nonsquare system, additional rows or columns of zeros can be added so as to make the system square. For details we refer to Ref. 4. Consider the system Σ as in Eq. (27), with $\mathcal{R}e(\lambda_i(A)) < 0$ and $m = p$. For simplicity it is assumed that the D-matrix of Σ is zero. Compute the grammians $\mathcal{P}$ and $\mathcal{Q}$ by solving the Lyapunov equations [Eqs. (34)]. We are looking for

$$\hat{\Sigma} = \left(\begin{array}{c|c} \hat{A} & \hat{B} \\ \hline \hat{C} & \hat{D} \end{array}\right)$$

such that the parallel connection of Σ and $\hat{\Sigma}$ denoted by Σ_e, is ϵ-all-pass. First, note that Σ_e has the following state-space representation:

$$\Sigma_e = \left(\begin{array}{cc|c} A & & B \\ & \hat{A} & \hat{B} \\ \hline C & -\hat{C} & -\hat{D} \end{array}\right) \tag{40}$$

According to the last characterization of Proposition 7, Σ_e is ϵ-all-pass iff the corresponding grammians and D matrices satisfy

$$\mathcal{P}_e\mathcal{Q}_e = \epsilon^2 I_{2n}, \qquad \hat{D}^*\hat{D} = \epsilon^2 I_m \tag{41}$$

This implies that $\hat{D}/\epsilon$ is a unitary matrix of size m. The key to the construction of $\mathcal{P}_e$ and $\mathcal{Q}_e$, is the *unitary dilation* of $\mathcal{P}$ and $\mathcal{Q}$. Consider the simple case where the grammians are scalar $\mathcal{P} = p$, $\mathcal{Q} = q$, and $\epsilon = 1$. In this case we are looking for filling in the ? so that

$$\begin{pmatrix} p & ? \\ ? & ? \end{pmatrix}\begin{pmatrix} q & ? \\ ? & ? \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

assuming that $pq \neq 1$. It readily follows that one solution is

$$\begin{pmatrix} p & 1-pq \\ 1-pq & -q(1-pq) \end{pmatrix}\begin{pmatrix} q & 1 \\ 1 & -\dfrac{p}{1-pq} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{42}$$

In the general (nonscalar) case, this suggests defining the quantity:

$$\Gamma := \epsilon^2 I_n - \mathcal{P}\mathcal{Q} \tag{43}$$

Assuming that ϵ is not equal to any of the eigenvalues of the product $\mathcal{P}\mathcal{Q}$ (i.e., to any of the singular values of the Hankel operator $\mathcal{H}_\Sigma$), Γ is invertible. Keeping in mind that $\mathcal{P}$ and $\mathcal{Q}$ do not commute, the choice of $\mathcal{P}_e$ and $\mathcal{Q}_e$ corresponding to Eq. (42) is

$$\mathcal{P}_e = \begin{pmatrix} \mathcal{P} & \Gamma \\ \Gamma^* & -Q\Gamma \end{pmatrix}, \qquad \mathcal{Q}_e = \begin{pmatrix} \mathcal{Q} & I_n \\ I_n & -\Gamma^{-1}\mathcal{P} \end{pmatrix} \tag{44}$$

Once the matrices $\mathcal{P}_e$ and $\mathcal{Q}_e$ for the dilated system Σ_e have been constructed, the next step is to construct the matrices $\hat{A}$, $\hat{B}$, and $\hat{C}$ of the dilation system $\hat{\Sigma}$. According to Eq. (35) of Proposition 6, the Lyapunov equation $A_e^*\mathcal{Q}_e + \mathcal{Q}_e A_e + C_e^*C_e = 0$ and the equation $\mathcal{Q}_e B_e + C_e^* D_e = 0$ have to be satisfied. Solving these two equations for the unknown quantities, we obtain

$$\begin{aligned} \hat{A} &= -A^* + C^*(C\mathcal{P} - \hat{D}B^*)\Gamma^{-*} \\ \hat{B} &= -\mathcal{Q}B + C^*\hat{D} \\ \hat{C} &= (C\mathcal{P} - \hat{D}B^*)\Gamma^{-*} \end{aligned} \tag{45}$$

where $\Gamma^{-*} := (\Gamma^*)^{-1}$. There remains to show is that $\hat{A}$ has exactly k stable eigenvalues. From the Lyapunov equation for $\mathcal{Q}_e$ it also follows that $\hat{\mathcal{Q}} = -\Gamma^{-1}\mathcal{P}$. By construction [see Eq. (43)], Γ has k positive and $n - k$ negative eigenvalues; the same holds true for $\hat{\mathcal{Q}}$ (i.e., in $(\hat{\mathcal{Q}}) = \{k, 0, n - k\}$). Furthermore, the pair $\hat{C}$, $\hat{A}$ is observable, because otherwise A has eigenvalues on the $j\omega$ axis, which is a contradiction to the original assumption that A is stable. Then by Proposition 6, the inertia of $-\hat{\mathcal{Q}}$ is equal to that of $\hat{A}$, which completes the proof.

Corollary 1. The system $\hat{\Sigma}$ has dimension n and the dilated system Σ_e has dimension $2n$. The stable subsystem $\hat{\Sigma}_+$ has dimension k.

State-Space Construction for Square Systems: Optimal Case. The construction of the previous section can be extended to include the optimal case. This section presents the formulae which first appeared in Section 6 of Ref. 4.

In this case we need to construct the all-pass dilation Σ_e for ϵ equal to the lower bound in Eq. (25), namely $\sigma_{k+1}(\Sigma)$. Thus, σ_{k+1} is an eigenvalue of the product of the grammians $\mathcal{P}\mathcal{Q}$ of multiplicity, say r. There exists a basis change in the state space such that

$$\mathcal{P} = \begin{pmatrix} I_r\sigma_{k+1} & 0 \\ 0 & \mathcal{P}_2 \end{pmatrix}, \qquad \mathcal{Q} = \begin{pmatrix} I_r\sigma_{k+1} & 0 \\ 0 & \mathcal{Q}_2 \end{pmatrix} \tag{46}$$

The *balancing transformation* Eq. (60) discussed in the section entitled "Balanced Realizations and Balanced Model Reduction" accomplishes this goal.

Clearly, *only* the pair $\mathcal{P}_2$ and $\mathcal{Q}_2$ needs to be dilated. As in the suboptimal case, explicit formulae for $\hat{\Sigma}$ can be obtained by using Eq. (35) of Proposition 7. Partition A, B, and C conformally with $\mathcal{P}$ and $\mathcal{Q}$:

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}, \qquad B = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix}, \qquad C = (C_1 \quad C_2)$$

where $A_{11} \in \mathbb{R}^{r\times r}$, B_1, $C_1^* \in \mathbb{R}^{r\times m}$, and I_r denotes the $r \times r$ identity matrix. The (1, 1) block of the Lyapunov equations yields $B_1B_1^* = C_1^*C_1$; this implies the existence of a unitary matrix U of size m, such that

$$B_1U = C_1^*, \qquad UU^* = I_m$$

Using Eqs. (45) we construct an all-pass dilation of the subsystem (A_{22}, B_{21}, C_{12}); solving the corresponding equations we obtain

$$\begin{aligned} \hat{D} &= \sigma_{k+1}U \\ \hat{B} &= -\mathcal{Q}_2B_2 + C_2^*\hat{D} \\ \hat{C} &= (C_2\mathcal{P}_2 - \hat{D}B_2^*)(\Gamma_2^*)^{-1} \\ \hat{A} &= -A_{22}^* + C_2^*\hat{C} \end{aligned} \tag{47}$$

where $\Gamma_2 := \sigma_{k+1}^2 - \mathscr{P}_2\mathscr{Q}_2 \in \mathbb{R}^{(n-r)\times(n-r)}$. Hence unlike the suboptimal case, $\hat{D}$ is not arbitrary. In the single-input single-output case, $\hat{D}$ is completely determined by the above relationship (is either $+1$ or -1) and hence there is a unique optimal approximant. This is not true for systems with more than one input and/or more than one output. Furthermore, in the *one-step reduction* case (i.e., $\sigma_{k+1} = \sigma_n$) the all-pass dilation system $\hat{\Sigma}$ has no antistable part; that is, it is the *optimal* Hankel-norm approximant.

Corollary 2. (a) In contrast to the suboptimal case, in the optimal case, $\hat{\Sigma}$ has dimension r, and the dilated system Σ_e has dimension $2n - r$. Furthermore, the stable subsystem Σ_+ has dimension $n - r$. (b) In the case where $\sigma_{k+1} = \sigma_n$, since $\hat{\Sigma} = \hat{\Sigma}_+$, $\Sigma - \hat{\Sigma}$ is stable and all-pass with norm σ_n.

General Case: Parameterization of All Suboptimal Approximants. Given the system Σ as in Eq. (27), let ϵ satisfy Eq. (25). The following rational matrix $\Theta(s)$ has size $(p + m) \times (p + m)$:

$$\Theta(s) := I_{p+m} - C_\Theta(sI - A_\Theta)^{-1}\mathscr{Q}_\Theta^{-1}C_\Theta^* J \tag{48}$$

where

$$C_\Theta := \begin{pmatrix} C & 0 \\ 0 & B^* \end{pmatrix} \in \mathbb{R}^{(p+m)\times 2n}$$
$$A_\Theta := \begin{pmatrix} A & 0 \\ 0 & -A^* \end{pmatrix} \in \mathbb{R}^{2n\times 2n}$$
$$\mathscr{Q}_\Theta := \begin{pmatrix} \mathscr{Q} & \epsilon I_n \\ \epsilon I_n & \mathscr{P} \end{pmatrix} \in \mathbb{R}^{2n\times 2n},$$
$$J := \begin{pmatrix} I_p & 0 \\ 0 & -I_m \end{pmatrix} \in \mathbb{R}^{(p+m)\times(p+m)}$$

Putting these expressions together we obtain

$$\Theta(s) = \left(\begin{array}{c|c} \Theta_{11} & \Theta_{12} \\ \hline \Theta_{21} & \Theta_{22} \end{array}\right) := \left(\begin{array}{c|c} I_p + C(sI - A)^{-1}\Gamma^{-1}\mathscr{P}C^* & \epsilon C(sI - A)^{-1}\Gamma^{-1}B \\ \hline -\epsilon B^*(sI + A^*)^{-1}\Gamma^{-*}C^* & I_m - B^*(sI + A^*)^{-1}\mathscr{Q}\Gamma^{-1}B \end{array}\right) \tag{49}$$

By construction, Θ is J-unitary on the $j\omega$ axis; that is, $\Theta^*(-j\omega)J\Theta(j\omega) = J$. Define

$$\begin{pmatrix} \Phi_1(\Delta) \\ \Phi_2(\Delta) \end{pmatrix} := \Theta(s)\begin{pmatrix} \Delta(s) \\ I_m \end{pmatrix} = \begin{pmatrix} \Theta_{11}(s)\Delta(s) + \Theta_{12}(s) \\ \Theta_{21}(s)\Delta(s) + \Theta_{22}(s) \end{pmatrix} \tag{50}$$

The proof of the following result can be found in Chap. 24 of Ref. 10. Recall Theorem 3.

Theorem 4. $\hat{\Sigma}$ is an ϵ-all-pass dilation Σ if and only if

$$H_\Sigma - H_{\hat{\Sigma}} = \Phi_1(\Delta)\Phi_2(\Delta)^{-1} \tag{51}$$

where $\Delta(s)$ is a $p \times m$ anti-stable contraction (all poles in the right half-plane and $\|\Delta(s)\|_\infty < 1$).

Error Bounds of Optimal and Suboptimal Approximants. Given is the stable $m \times m$ system Σ, having Hankel singular values σ_i and multiplicity r_i, $i = 1, \ldots, q$ [see Eq. (21)]. Let $\hat{\Sigma}$ be the ϵ-all-pass dilation of Σ, where $\epsilon = \sigma_q$ (the smallest singular value). Following Corollary 2, the dimension of $\hat{\Sigma}_+$ is $n - q$, which implies that $\hat{\Sigma} = \hat{\Sigma}_+$; that is, it is *stable*. Thus by the same corollary, for a *one-step* reduction, the all-pass dilation system has *no* unstable poles; for simplicity we will use the notation

$$\Sigma_q := \hat{\Sigma}(\sigma_q)$$

Thus $\Sigma - \Sigma_q$ is *all-pass* of magnitude σ_q, and consequently we have

$$\sigma_i(\Sigma_q) = \sigma_i(\Sigma), \qquad i = 1, \ldots, n - r_q \tag{52}$$

We now approximate Σ_q by Σ_{q-1} through a one-step optimal Hankel-norm reduction. By successive application of the same procedure, we thus obtain a sequence of all-pass systems Σ_i, $i = 1, \ldots, q$, and a system Σ_0 consisting of the matrix D_0, such that the transfer function of Σ is decomposed as follows:

$$H_\Sigma(s) = D_0 + H_1(s) + \cdots + H_q(s) \tag{53}$$

where $H_k(s)$ is the transfer function of the stable all-pass system Σ_k having dimension $2n - \sum_{i=k}^{q} r_i$, $k = 1, \ldots, q$; the dimension of the partial sums $\sum_{i=1}^{k} H_i(s)$ is equal to $\sum_{i=1}^{k} r_i$, $k = 1, 2, \ldots, q$. Thus Eq. (53) is the equivalent of the dyadic decomposition Eq. (4), for Σ.

From the above decomposition we can derive the following upper bound for the $\mathscr{H}_\infty$ norm of Σ. Assuming that $H_\Sigma(\infty) = D = 0$, we obtain

$$\|H_\Sigma(s) - D_0\|_\infty \le \underbrace{\|H_1(s)\|_\infty}_{\sigma_1} + \cdots + \underbrace{\|H_q(s)\|_\infty}_{\sigma_q}$$

Evaluating the above expression at infinity yields

$$\|D_0\|_2 \le \sigma_1 + \cdots + \sigma_q$$

Thus, combining this inequality with $\|H_\Sigma(s)\|_\infty \le \|D_0\|_2 + \sigma_1 + \cdots + \sigma_q$, yields the bound given in Lemma 2:

$$\|H_\Sigma(s)\|_\infty \le 2(\sigma_1 + \cdots + \sigma_q) \tag{54}$$

This bound can be sharpened by computing an appropriate $D_0 \in \mathbb{R}^{m\times m}$ as in Eq. (53):

$$\|H_\Sigma(s) - D_0\|_\infty \le \sigma_1 + \cdots + \sigma_q \tag{55}$$

Finally, we state a result of Ref. 4, on the Hankel singular values of the stable part Σ_{e+} of the all-pass dilation Σ_e. It will be assumed for simplicity that each Hankel singular value has multiplicity one: $r_i = 1$. Assume that $\Sigma_{e+} = \Sigma - \hat{\Sigma}_+$, where $\hat{\Sigma}_+$ is an optimal Hankel approximant of Σ of dimension k; the following holds: $\sigma_1(\Sigma_{e+}) = \cdots = \sigma_{2k+1}(\Sigma_{e+}) = \sigma_{k+1}(\Sigma)$, $\sigma_{2k+2}(\Sigma_{e+}) = \sigma_1(\hat{\Sigma}_-) \le \sigma_{k+2}(\Sigma), \ldots, \sigma_{n+k}(\Sigma_{e+}) = \sigma_{n-k-1}(\hat{\Sigma}_-) \le \sigma_n(\Sigma)$. Using these inequalities we obtain an error bound for the $\mathscr{H}_\infty$ norm of the error of Σ with a degree k optimal approximant $\hat{\Sigma}_+$. First, note that

$$\delta := \|\Sigma_- - D_0\|_\infty \le \sigma_1(\hat{\Sigma}_-) + \cdots + \sigma_{n-k-1}(\hat{\Sigma}_-)$$

This implies $\|\Sigma - \hat{\Sigma}_+ - D_0\|_\infty \leq \sigma_{k+1}(\Sigma) + \delta$. Combining all previous inequalities we obtain the bounds:

$$\sigma_{k+1} \leq \|\Sigma - \hat{\Sigma}_+ - D_0\|_\infty \leq \sigma_{k+1}(\Sigma) + \sigma_1(\hat{\Sigma}_-) + \cdots + \sigma_{n-k-1}(\hat{\Sigma}_-) \leq \sigma_{k+1}(\Sigma) + \cdots + \sigma_n(\Sigma) \tag{56}$$

The left-hand-side inequality in Eq. (25), and the above finally yield upper and lower bounds on the $\mathcal{H}_\infty$ norm of the error:

$$\sigma_{k+1} \leq \|\Sigma - \hat{\Sigma}_+\|_\infty \leq 2(\sigma_{k+1} + \cdots + \sigma_n) \tag{57}$$

The first upper bound in Eq. (56) above is the tightest; but both a D term and the singular values of the antistable $\hat{\Sigma}_-$ part of the all-pass dilation are needed. The second upper bound in Eq. (56) is the second-tightest; it only requires knowledge of the optimal D term. Finally the upper bound in Eq. (57) is the least tight among these three. It is, however, the most useful, since it can be determined a priori, that is, before the computation of the approximants, once the singular values of the original system Σ are known.

Remark 4. (a) In a similar way a bound for the infinity norm of suboptimal approximants can be obtained. Given Σ, let $\hat{\Sigma}$ be an all-pass dilation satisfying Eq. (25). Then, similarly to Eq. (56), the following holds:

$$\sigma_{r+1} \leq \|\Sigma - \hat{\Sigma}_+\|_\infty \leq 2(\epsilon + \sigma_{r+1} + \cdots + \sigma_n)$$

(b) For a one-step Hankel-norm reduction, $\Sigma - \hat{\Sigma}$ is all-pass; that is,

$$\|\Sigma - \hat{\Sigma}\|_H = \|\Sigma - \hat{\Sigma}\|_\infty = \sigma_q$$

which shows that in this case $\hat{\Sigma}$ is an optimal approximant not only in the Hankel norm but in the infinity norm as well. Thus for this special case, and despite Propositions 1 and 5, the Hankel-norm approximation theory yields a solution for the optimal approximation problem in the 2-norm of the convolution operator $\mathcal{S}_\Sigma$; recall that this is the problem we initially wished to solve.

Balanced Realizations and Balanced Model Reduction

A model reduction method which is closely related to the Hankel-norm approximation method is approximation by *balanced truncation.* This involves a particular realization of a linear system Σ given by Eq. (27), called *balanced realization* (the D matrix is irrelevant in this case). We start by explaining the rational behind this particular state-space realization.

The Concept of Balancing. Consider a stable system Σ with positive definite reachability and observability grammians $\mathcal{P}$ and $\mathcal{Q}$. It can be shown that the *minimal energy* required for the transfer of the state of the system from 0 to some final state x_r is

$$x_r^* \mathcal{P}^{-1} x_r \tag{58}$$

Similarly, the largest *observation energy* produced by observing the output, when the initial state of the system is x_o and the input is zero ($u = 0$), is equal to

$$x_o^* \mathcal{Q} x_o \tag{59}$$

We conclude that the states which are difficult to reach—that is, those which require a large amount of energy to reach—are in the span of the eigenvectors of the reachability grammian $\mathcal{P}$ corresponding to small eigenvalues. Moreover, the states which are difficult to observe—that is, those which yield small amounts of observation energy—are those which lie in the span of the eigenvectors of the observability grammian $\mathcal{Q}$ corresponding to small eigenvalues as well.

This observation suggests that one way to obtain reduced order models is by eliminating those states which are difficult to reach and/or difficult to observe. However, states which are difficult to reach may not be difficult to observe, and vice versa; as simple examples show, these properties are basis-dependent. This suggests the need for a basis in which states which are difficult to reach are simultaneously difficult to observe, and vice versa. From these considerations, the question arises: Given a continuous- or discrete-time stable system Σ, does there exist a basis of the state space in which states which are *difficult to reach* are also *difficult to observe?*

The answer to this question is affirmative. The transformation which achieves this goal is called a *balancing transformation.* Under an equivalence transformation T (basis change in the state space: $\tilde{x} := Tx$, $\det T \neq 0$) we have $\tilde{A} = TAT^{-1}$, $\tilde{B} = TB$, and $\tilde{C} = CT^{-1}$, while the grammians are transformed as follows:

$$\tilde{\mathcal{P}} = T\mathcal{P}T^*, \qquad \tilde{\mathcal{Q}} = T^{-*}\mathcal{Q}T^{-1} \Rightarrow \tilde{\mathcal{P}}\tilde{\mathcal{Q}} = T(\mathcal{P}\mathcal{Q})T^{-1}$$

The problem is to find T, $\det T \neq 0$, such that the transformed grammians $\tilde{\mathcal{P}}$ and $\tilde{\mathcal{Q}}$ are equal. This will ensure that the states which are difficult to reach are precisely those which are difficult to observe.

Definition 3. The stable system Σ is balanced iff $\mathcal{P} = \mathcal{Q}$. Σ is principal-axis balanced iff

$$\mathcal{P} = \mathcal{Q} = S := \mathrm{diag}(\sigma_1, \ldots, \sigma_n)$$

The existence of a balancing transformation is guaranteed.

Lemma 3. Balancing Transformation. Given the stable system Σ and the corresponding grammians $\mathcal{P}$ and $\mathcal{Q}$, let the matrices R, U, and S be defined by $\mathcal{P} =: R^*R$ and $R\mathcal{Q}R^* =: US^2U^*$. A (principal axis) balancing transformation is given by

$$T := S^{1/2}U^*R^{-*} \tag{60}$$

where (as before) $R^{-*} =: (R^*)^{-1}$

To verify that T is a balancing transformation, it follows by direct calculation that $T\mathcal{P}T^* = S$ and $T^{-*}\mathcal{Q}T^{-1} = S$. We also note that if the Hankel singular values are distinct (i.e., have multiplicity one), balancing transformations $\hat{T}$ are determined from T given above, up to multiplication by a *sign* matrix L—that is, a diagonal matrix with ± 1 on the diagonal: $\hat{T} = LT$.

Model Reduction. Let Σ be balanced with grammians equal to S, and partition:

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}, \qquad S = \begin{pmatrix} S_1 & 0 \\ 0 & S_2 \end{pmatrix}, \qquad B = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix}, \qquad C = (C_1 \quad C_2) \tag{61}$$

The systems

$$\Sigma_i := \left(\begin{array}{c|c} A_{ii} & B_i \\ \hline C_i & \end{array}\right), \qquad i = 1, 2$$

are called reduced-order systems obtained form Σ by balanced truncation. These have certain guaranteed properties. However, these properties are different from discrete- and continuous-time systems. Hence we state two theorems. For a proof see, for example, Ref. 6.

Theorem 5. Balanced Truncation: Continuous-Time Systems. Given the stable (no poles in the closed right half-plane) continuous-time system Σ, the reduced-order systems Σ_i, $i = 1, 2$, obtained by balanced truncation have the following properties:

1. Σ_i, $i = 1. 2$, satisfy the Lyapunov equations $A_{ii}S_i + S_iA_{ii}^* + B_iB_i^* = 0$, and $A_{ii}^*S_i + S_iA_{ii} + C_i^*C_i = 0$. Furthermore, A_{ii}, $i = 1, 2$, have no eigenvalues in the open right half-plane.
2. If S_1 and S_2 have no eigenvalues in common, both Σ_1 and Σ_2 have no poles on the imaginary axis and are, in addition, reachable, observable, and balanced.
3. Let the distinct singular values of Σ be σ_i, with multiplicities m_i, $i = 1, \ldots, q$. Let Σ_1 have singular values σ_i, $i = 1, \ldots, k$, with the corresponding multiplicities m_i, $i = 1, \ldots, k$, $k < q$. The $\mathcal{H}_\infty$ norm of the difference between the full-order system Σ and the reduced-order system Σ_1 is upper-bounded by twice the sum of the neglected Hankel singular values:

$$\|\Sigma - \Sigma_1\|_\infty \le 2(\sigma_{k+1} + \cdots + \sigma_q) \tag{62}$$

If the smallest singular value is truncated—that is, $S_2 = \sigma_q I_{r_q}$—equality holds.

Theorem 6. Balanced Truncation: Discrete-Time Systems. Given the stable, discrete-time system (no poles in the complement of the open unit disk) Σ, the reduced-order systems Σ_i obtained by balanced truncation have the following properties:

1. Σ_i, $i = 1, 2$, have no poles in the closed unit disk; these systems are, in general, not balanced.
2. If $\lambda_{\min}(S_1) > \lambda_{\max}(S_2)$, Σ_1 is, in addition, reachable and observable.
3. The h_∞ norm of the difference between full- and reduced-order models is upper-bounded by twice the sum of the neglected Hankel singular values, multiplicities included:

$$\|\Sigma - \Sigma_1\|_\infty \le 2\,\text{trace}(S_2)$$

Remark 5. (a) The last part of the above theorems says that if the neglected singular values are small, then the Bode plots of Σ and Σ_1 are guaranteed to be close in the $\mathcal{H}_\infty$ norm. The difference between part 3 for continuous- and discrete-time systems above is that the multiplicities of the neglected singular values do not enter in the upper bound for continuous-time systems.

(b) Proposition 8 implies that the bilinear transformation between discrete- and continuous-time systems preserves balancing (see also "A Simple Discrete-Time Example" below).

(c) Let Σ_{hank} and Σ_{bal} be the reduced-order systems obtained by one step Hankel-norm approximation, and one step balanced truncation, respectively. It can be shown that $\Sigma_{\text{hank}} - \Sigma_{\text{bal}}$ is all-pass with norm σ_q; it readily follows that

$$\|\Sigma - \Sigma_{\text{bal}}\|_\infty \le 2\sigma_q \quad \text{and} \quad \|\Sigma - \Sigma_{\text{bal}}\|_H \le 2\sigma_q$$

A consequence of the above inequalities is that the error for reduction by balanced truncation can be upper-bounded by means of the singular values of Σ as given in Theorem 5. Furthermore, this bound is valid both for the $\mathcal{H}_\infty$ norm and for the Hankel norm.

(d) Every linear, time-invariant, continuous-time and stable system Σ, can be expressed in balanced canonical form. In the generic case (distinct singular values) this canonical form is given in terms of $2n$ positive numbers, namely the singular values $\sigma_i > 0$, and some $b_i > 0$, as well as n signs $s_i = \pm 1$, $i = 1, \ldots, n$. The quantities $\lambda_i := s_i\sigma_i$ are called signed singular values of Σ; they satisfy $H_\Sigma(0) = 2(\lambda_1 + \cdots + \lambda_n)$. For details on balanced canonical forms we refer to the work of Ober—for example, Ref. 11.

EXAMPLES

In this section we will illustrate the results presented above by means of three examples. The first deals with a simple second-order continuous-time system. The purpose is to compute the limit of suboptimal Hankel-norm approximants as ϵ tends to one of the singular values. The second example discusses the approximation of a discrete-time system [third-order finite impulse response (FIR) system] by balanced truncation and Hankel-norm approximation. The section concludes with the approximation of the four classic analog filters (Butterworth, Chebyshev 1, Chebyshev 2, and Elliptic) by balanced truncation and Hankel-norm approximation.

A Simple Continuous-Time Example

Consider the system Σ given by Eq. (27), where $n = 2$, $m = p = 1$, and

$$A = -\begin{pmatrix} \dfrac{1}{2\sigma_1} & \dfrac{1}{\sigma_1 + \sigma_2} \\ \dfrac{1}{\sigma_1 + \sigma_2} & \dfrac{1}{2\sigma_2} \end{pmatrix}, \qquad B = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad C = (1 \quad 1), \qquad D = 0$$

where $\sigma_1 > \sigma_2$. This system is in *balanced canonical form;* this means that the grammians are $\mathcal{P} = \mathcal{Q} = \text{diag}(\sigma_1, \sigma_2) := S$; this canonical form is a special case of the forms discussed in Ref. 11.

We wish to compute the suboptimal Hankel-norm approximants for $\sigma_1 > \epsilon > \sigma_2$. Then we will compute the limit of this family for $\epsilon \to \sigma_2$ and $\epsilon \to \sigma_1$, and we will show that the system obtained is indeed the optimal approximant. From Eq. (43), $\Gamma = \epsilon^2 I_2 - S^2 = \mathrm{diag}(\epsilon^2 - \sigma_1^2, \epsilon^2 - \sigma_2^2)$; the inertia of Γ is $\{1, 0, 1\}$; furthermore, from Eq. (45) we obtain

$$\hat{A} = \begin{pmatrix} \dfrac{\epsilon - \sigma_1}{2\sigma_1(\epsilon + \sigma_1)} & \dfrac{\epsilon - \sigma_1}{(\sigma_1 + \sigma_2)(\epsilon + \sigma_2)} \\ \dfrac{\epsilon - \sigma_2}{(\sigma_1 + \sigma_2)(\epsilon + \sigma_1)} & \dfrac{\epsilon - \sigma_2}{2\sigma_2(\epsilon + \sigma_2)} \end{pmatrix},$$

$$\hat{B} = \begin{pmatrix} \epsilon - \sigma_1 \\ \epsilon - \sigma_2 \end{pmatrix},$$

$$\hat{C} = \begin{pmatrix} \dfrac{-1}{\epsilon + \sigma_1} & \dfrac{-1}{\epsilon + \sigma_2} \end{pmatrix},$$

$$\hat{D} = \epsilon$$

Since the inertia of $\hat{A}$ is equal to the inertia of $-\Gamma$, $\hat{A}$ has one stable and one unstable poles (this can be checked directly by noticing that the determinant of $\hat{A}$ is negative). As $\epsilon \to \sigma_2$ we obtain

$$\hat{A} = \begin{pmatrix} \dfrac{\sigma_2 - \sigma_1}{2\sigma_1(\sigma_1 + \sigma_2)} & \dfrac{\sigma_2 - \sigma_1}{2\sigma_2(\sigma_1 + \sigma_2)} \\ 0 & 0 \end{pmatrix}, \qquad \hat{B} = \begin{pmatrix} \sigma_2 - \sigma_1 \\ 0 \end{pmatrix},$$

$$\hat{C} = \begin{pmatrix} \dfrac{-1}{\sigma_1 + \sigma_2} & \dfrac{-1}{2\sigma_2} \end{pmatrix}, \qquad \hat{D} = \sigma_2$$

This system is not reachable but observable (i.e., there is a pole-zero cancellation in the transfer function). A state-space representation of the reachable and observable subsystem is

$$\bar{A} = \frac{\sigma_2 - \sigma_1}{2\sigma_1(\sigma_1 + \sigma_2)}, \quad \bar{B} = \sigma_2 - \sigma_1, \quad \bar{C} = \frac{-1}{\sigma_1 + \sigma_2}, \quad \bar{D} = \sigma_2$$

Equations (45) depend on the choice of $\hat{D}$. If we choose it to be $-\epsilon$, the limit still exists and gives a realization of the optimal system which is equivalent to $\bar{A}$, $\bar{B}$, $\bar{C}$, $\bar{D}$ given above.

Finally, if $\epsilon \to \sigma_1$, after a pole-zero cancellation, we obtain the following reachable and observable approximant:

$$\bar{A} = \frac{\sigma_1 - \sigma_2}{2\sigma_1(\sigma_1 + \sigma_2)}, \quad \bar{B} = \sigma_1 - \sigma_2, \quad \bar{C} = \frac{-1}{\sigma_1 + \sigma_2}, \quad \bar{D} = \sigma_1$$

This is the best antistable approximant of Σ—that is, the Nehari solution [see Remark 3(a)].

A Simple Discrete-Time Example

In this section we will consider a third-order discrete-time FIR system described by the transfer function:

$$H(z) = \frac{z^2 + 1}{z^3} \tag{63}$$

We will first consider approximation by balanced truncation. The issue here is to examine balanced truncation first in discrete-time and then in continuous-time, using the bilinear transformation of the section entitled "A Transformation Between Continuous- and Discrete-Time Systems," and compare the results. Subsequently, Hankel-norm approximation will be investigated.

Approximation by Balanced Truncation. A balanced realization of this system is given by

$$\Sigma_d := \left(\begin{array}{c|c} F & G \\ \hline H & J \end{array}\right) = \left(\begin{array}{ccc|c} 0 & \alpha & 0 & \beta \\ \alpha & 0 & -\alpha & 0 \\ 0 & \alpha & 0 & -\gamma \\ \hline -\beta & 0 & -\gamma & 0 \end{array}\right),$$

$$\alpha = 5^{-1/4}, \qquad \beta = \frac{\sqrt{3\sqrt{5} + 5}}{\sqrt{10}}, \qquad \gamma = \frac{\sqrt{3\sqrt{5} - 5}}{\sqrt{10}}$$

where the reachability and observability grammians are equal and diagonal:

$$\mathcal{P} = \mathcal{Q} = S = \mathrm{diag}(\sigma_1, \sigma_2, \sigma_3), \qquad \sigma_1 = \frac{\sqrt{5} + 1}{2}$$

$$\sigma_2 = 1, \qquad \sigma_3 = \frac{\sqrt{5} - 1}{2}$$

The second- and first-order balanced truncated systems are

$$\Sigma_{d,2} = \left(\begin{array}{c|c} F_2 & G_2 \\ \hline H_2 & J_2 \end{array}\right) = \left(\begin{array}{cc|c} 0 & \alpha & \beta \\ \alpha & 0 & 0 \\ \hline -\beta & 0 & 0 \end{array}\right),$$

$$\Sigma_{d,1} = \left(\begin{array}{c|c} F_1 & G_1 \\ \hline H_1 & J_1 \end{array}\right) = \left(\begin{array}{c|c} 0 & \beta \\ \hline -\beta & 0 \end{array}\right)$$

Notice that $\Sigma_{d,2}$ is *balanced,* but has singular values which are *different* from σ_1 and σ_2. $\Sigma_{d,1}$ is also balanced since $G_1 = -H_2$, but its grammians are not equal to σ_1.

Let Σ_c denote the continuous-time system obtained from Σ_d by means of the bilinear transformation described in the section entitled "A Transformation Between Continuous- and Discrete-Time Systems"

$$\Sigma_c = \left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right) = \left(\begin{array}{ccc|c} -1 - 2\alpha^2 & 2\alpha & 2\alpha^2 & \delta_+ \\ 2\alpha & -1 & -2\alpha & -\sqrt{2} \\ -2\alpha^2 & -2\alpha & -1 + 2\alpha^2 & \delta_- \\ \hline -\delta_+ & \sqrt{2} & \delta_- & 2 \end{array}\right)$$

where $\delta_\pm = \sqrt{2}(\alpha \pm \beta)$. Notice that Σ_c is balanced. We now compute first and second reduced-order systems $\Sigma_{c,1}$ and $\Sigma_{c,2}$ by truncating Σ_c:

$$\Sigma_{c,2} = \left(\begin{array}{c|c} A_2 & B_2 \\ \hline C_2 & D_2 \end{array}\right) = \left(\begin{array}{cc|c} -1 - 2\alpha^2 & 2\alpha & \delta_+ \\ 2\alpha & -1 & -\sqrt{2} \\ \hline -\delta_+ & \sqrt{2} & 2 \end{array}\right)$$

$$\Sigma_{c,1} = \left(\begin{array}{c|c} A_1 & B_1 \\ \hline C_1 & D_1 \end{array}\right) = \left(\begin{array}{c|c} -1 - 2\alpha^2 & \delta_+ \\ \hline -\delta_+ & 2 \end{array}\right)$$

Let $\bar{\Sigma}_{d,2}$ and $\bar{\Sigma}_{d,1}$ be the discrete-time systems obtaining by transforming $\Sigma_{c,2}$ and $\Sigma_{c,1}$ back to discrete-time:

$$\bar{\Sigma}_{d,2} = \left(\begin{array}{c|c} \bar{F}_2 & \bar{G}_2 \\ \hline \bar{H}_2 & \bar{J}_2 \end{array}\right) = \left(\begin{array}{cc|c} 0 & \alpha & \beta \\ \alpha & \alpha^2 & -\alpha\gamma \\ \hline -\beta & \alpha\gamma & -\gamma^2 \end{array}\right)$$

$$\bar{\Sigma}_{d,1} = \left(\begin{array}{c|c} \bar{F}_1 & \bar{G}_1 \\ \hline \bar{H}_1 & \bar{J}_1 \end{array}\right) = \left(\begin{array}{c|c} -\sigma_3/2 & (\alpha + \beta)\sigma_3/2\alpha^2 \\ \hline -(\alpha + \beta)\sigma_3/2\alpha^2 & 2 - (\alpha^2 + \beta)^2/(1 + \alpha^2) \end{array}\right)$$

The conclusion is that $\overline{\Sigma}_{d,2}$ and $\overline{\Sigma}_{d,1}$ are balanced and different from $\Sigma_{d,2}$ and $\Sigma_{d,1}$. It is interesting to notice that the singular value of $\Sigma_{d,1}$ is $\beta^2 = (1 + \sigma_1)/\sqrt{5}$, while that of $\overline{\Sigma}_{d,1}$ is σ_1; β^2 satisfies $\sigma_2 < \beta^2 < \sigma_1$. Furthermore, the singular values of $\Sigma_{d,2}$ are $5\beta^2/4$, $\sqrt{5}\beta^2/4$ which satisfy the following interlacing inequalities:

$$\sigma_3 < \sqrt{5}\beta^2/4 < \sigma_2 < 5\beta^2/4 < \sigma_1$$

The numerical values of the quantities above are $\sigma_1 = 1.618034$, $\sigma_3 = 0.618034$, $\alpha = 0.66874$, $\beta = 1.08204$, $\gamma = 0.413304$, $\delta_+ = 2.47598$, and $\delta_- = .361241$.

Hankel-Norm Approximation. The Hankel operator of the system described by Eq. (63) is

$$\mathcal{H} = \begin{pmatrix} 1 & 0 & 1 & 0 & \\ 0 & 1 & 0 & 0 & \cdots \\ 1 & 0 & 0 & 0 & \\ 0 & 0 & 0 & 0 & \\ & \vdots & & & \ddots \end{pmatrix}$$

The SVD of the 3×3 principal submatrix of $\mathcal{H}$ is

$$\begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix} = \begin{pmatrix} \sqrt{\frac{\sigma_1}{\sqrt{5}}} & 0 & \sqrt{\frac{\sigma_3}{\sqrt{5}}} \\ 0 & 1 & 0 \\ \sqrt{\frac{\sigma_3}{\sqrt{5}}} & 0 & -\sqrt{\frac{\sigma_1}{\sqrt{5}}} \end{pmatrix} \begin{pmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \\ 0 & 0 & \sigma_3 \end{pmatrix} \begin{pmatrix} \sqrt{\frac{\sigma_1}{\sqrt{5}}} & 0 & \sqrt{\frac{\sigma_3}{\sqrt{5}}} \\ 0 & 1 & 0 \\ -\sqrt{\frac{\sigma_3}{\sqrt{5}}} & 0 & \sqrt{\frac{\sigma_1}{\sqrt{5}}} \end{pmatrix} \quad (64)$$

where σ_i, $i = 1, 2, 3$, are as given earlier. It is tempting to conjecture that the optimal second-order approximant is obtained by setting $\sigma_3 = 0$ in Eq. (64). The problem with this procedure is that the resulting approximant does *not* have Hankel structure.

To compute the optimal approximant, the system is first transformed to a continuous-time system using the transformation of the section entitled "A Transformation Between Continuous- and Discrete-Time Systems"; we obtain the transfer function

$$H_c(s) = H_d\left(\frac{1+s}{1-s}\right) = \frac{2(s^3 - s^2 + s - 1)}{(s+1)^3}$$

where H_d is the transfer function defined in Eq. (63). Applying the theory discussed in the section entitled "State-Space Construction for Square Systems: Optimal Case," we obtain the following second-order continuous-time optimal approximant:

$$H_{c,2}(s) = \frac{-(s^2 - 1)}{(1-\sigma_3)s^2 + 2\sigma_1 s + (1 - \sigma_3)}$$

Again using the transformation of the section entitled "A Transformation Between Continuous- and Discrete-Time Systems" we obtain the following discrete-time optimal approximant:

$$H_{d,2}(z) = H_{c,2}\left(\frac{z-1}{z+1}\right) = \frac{z}{z^2 - \sigma_3}$$

Notice that the optimal approximant is *not* an FIR system. It has poles at $\pm\sqrt{\sigma_3}$. Furthermore, the error

$$H_d(z) - H_{d,2}(z) = \sigma_3 \left[\frac{1 - \sigma_3 z^2}{z^3(z^2 - \sigma_3)}\right]$$

is all-pass with magnitude equal to σ_3 on the unit circle [as predicted by Corollary 2(b)]. The corresponding optimal Hankel matrix of rank 2 is

$$\hat{\mathcal{H}} = \begin{pmatrix} 1 & 0 & \sigma_3 & 0 & \sigma_3^2 & \\ 0 & \sigma_3 & 0 & \sigma_3^2 & 0 & \cdots \\ \sigma_3 & 0 & \sigma_3^2 & 0 & \sigma_3^3 & \\ 0 & \sigma_3^2 & 0 & \sigma_3^3 & 0 & \\ \sigma_3^2 & 0 & \sigma_3^3 & 0 & \sigma_3^4 & \\ & \vdots & & & & \ddots \end{pmatrix}$$

In this particular case the 3×3 principal submatrix of $\hat{\mathcal{H}}$ is also an optimal approximant of the corresponding submatrix of $\mathcal{H}$.

$$\begin{pmatrix} 1 & 0 & \sigma_3 \\ 0 & \sigma_3 & 0 \\ \sigma_3 & 0 & \sigma_3^2 \end{pmatrix} = \begin{pmatrix} \sqrt{\frac{\sigma_1}{\sqrt{5}}} & 0 & \sqrt{\frac{\sigma_3}{\sqrt{5}}} \\ 0 & 1 & 0 \\ \sqrt{\frac{\sigma_3}{\sqrt{5}}} & 0 & -\sqrt{\frac{\sigma_1}{\sqrt{5}}} \end{pmatrix} \begin{pmatrix} 1+\sigma_3^2 & 0 & 0 \\ 0 & \sigma_3 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \sqrt{\frac{\sigma_1}{\sqrt{5}}} & 0 & \sqrt{\frac{\sigma_3}{\sqrt{5}}} \\ 0 & 1 & 0 \\ -\sqrt{\frac{\sigma_3}{\sqrt{5}}} & 0 & \sqrt{\frac{\sigma_1}{\sqrt{5}}} \end{pmatrix} \quad (65)$$

Notice that the above decomposition can be obtained from Eq. (64) by making use of the freedom mentioned in Eq. (7). Finally it is readily checked that the Hankel matrix consisting of 1 as (1, 1) entry and 0 everywhere else is the optimal approximant of $\mathcal{H}$ of rank one. The dyadic decomposition [Eq. (53)] of H is

$$H(z) = H_1(z) + H_2(z) + H_3(z) = \sigma_1\left[\frac{1}{z}\right] + \sigma_2\left[\frac{1 - \sigma_3 z^2}{z(z^2 - \sigma_3)}\right] + \sigma_3\left[\frac{-1 + \sigma_3 z^2}{z^3(z^2 - \sigma_3)}\right]$$

Notice that each H_i is σ_i-all-pass, and the degree of H_1 is one, that of $H_1 + H_2$ is two, and finally that of all three summands is three.

A Higher-Order Example

In our last example, we will approximate four well-known types of analog filters by means of balanced truncation and Hankel norm approximation. These are:

1. Σ_B—Butterworth
2. Σ_{C1}—Chebyshev 1: 1% ripple in the pass band (PB)
3. Σ_{C2}—Chebyshev 2: 1% ripple in the stop band (SB)
4. Σ_E—Elliptic: 0.1% ripple both in the PB and in the SB

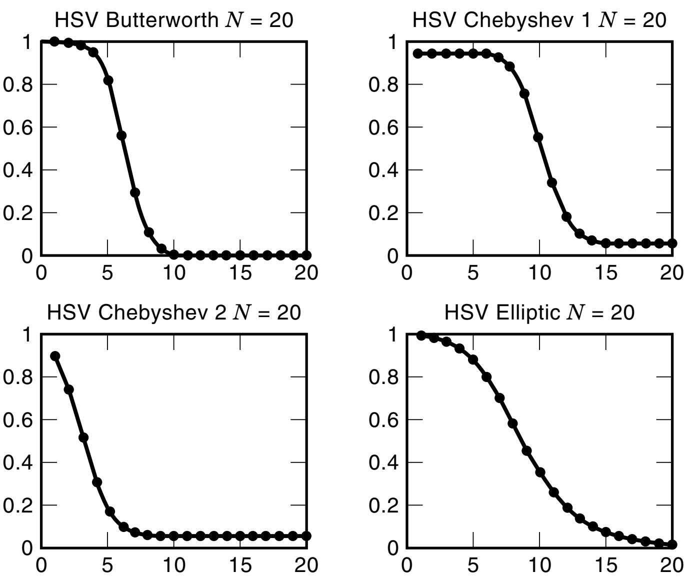
In each case we will consider 20th-order low-pass filters, with pass-band gain equal to 1, and cut-off frequency normalized to 1. Figure 2 shows the Hankel singular values of the full-order models. It follows from these plots that in order to obtain roughly comparable approximation errors, Σ_B and Σ_{C2} will have to be approximated by systems of lower order than Σ_{C1} and Σ_E; we thus choose to approximate Σ_B and Σ_{C2} by *8th-order* models, and we choose to approximate Σ_{C1} and Σ_E by *10th-order* models. It is interesting to observe that for the Chebyshev-2 filter the difference of the singular values, $\sigma_{13} - \sigma_{20}$, is of the order 10^{-7}. Thus Σ_{C2} has an (approximate) 8th-order all-pass subsystem of magnitude 0.05. Consequently, Σ_{C2} cannot be approximated with systems of order 13 through 19. Similarly, the Chebyshev-1 filter has an all-pass subsystem of order 3; consequently approximations of order 1, 2, and 3 are not possible.

The subscript "bal" stands for approximation by balanced truncation, the subscript "hank" stands for optimal Hankel-norm approximation, "FOM" stands for "full-order model," and "ROM" stands for "reduced-order model."

Figure 3 gives the amplitude Bode plots of the error systems and tabulates their $\mathcal{H}_\infty$ norms and upper bounds. We observe that the 10th-order Hankel-norm approximants of Σ_{C1} and Σ_E are not very good in the SB; one way for improving them is to increase the approximation order; another is to compute weighted approximants (see, e.g., Ref. 12). In Table 2, more detailed bounds for the approximants will be given in the case where the approximant contains an optimal D-term "Bound 1" is the first expression on the right-hand side of Eq. (56), and "Bound 2" is the second expression on the right-hand side of the same expression:

$$\begin{aligned}
\text{Bound 11 } &:= \sigma_9(\Sigma) + \sigma_1(\hat{\Sigma}_-) \\
&\quad + \cdots + \sigma_{11}(\hat{\Sigma}_-) \le \sigma_9(\Sigma) \\
&\quad + \cdots + \sigma_{20}(\Sigma) =: \text{ Bound 12} \\
\text{Bound 21 } &:= \sigma_{11}(\Sigma) + \sigma_1(\hat{\Sigma}_-) \\
&\quad + \cdots + \sigma_9(\hat{\Sigma}_-) \le \sigma_{11}(\Sigma) \\
&\quad + \cdots + \sigma_{20}(\Sigma) =: \text{ Bound 22}
\end{aligned}$$

Finally, Figure 4 shows the amplitude Bode plots of the ROMs obtained by Hankel and balanced reductions.



	Butterworth	Chebyshev 1	Chebyshev 2	Elliptic
σ_1	9.9998e-01	9.5000e-01	8.9759e-01	9.8803e-01
σ_2	9.9952e-01	9.5000e-01	7.3975e-01	9.8105e-01
σ_3	9.9376e-01	9.5000e-01	5.1523e-01	9.6566e-01
σ_4	9.5558e-01	9.4996e-01	3.0731e-01	9.3602e-01
σ_5	8.2114e-01	9.4962e-01	1.6837e-01	8.8469e-01
σ_6	5.6772e-01	9.4710e-01	9.5205e-02	8.0582e-01
σ_7	2.9670e-01	9.3362e-01	6.3885e-02	6.9997e-01
σ_8	1.1874e-01	8.8268e-01	5.3342e-02	5.7677e-01
σ_9	3.8393e-02	7.5538e-01	5.0643e-02	4.5166e-01
σ_{10}	1.0402e-02	5.5133e-01	5.0103e-02	3.3873e-01
σ_{11}	2.3907e-03	3.3740e-01	5.0014e-02	2.4576e-01
σ_{12}	4.6611e-04	1.8209e-01	5.0002e-02	1.7413e-01
σ_{13}	7.6629e-05	9.8132e-02	5.0000e-02	1.2135e-01
σ_{14}	1.0507e-05	6.3256e-02	5.0000e-02	8.3613e-02
σ_{15}	1.1813e-06	5.2605e-02	5.0000e-02	5.7195e-02
σ_{16}	1.0622e-07	5.0362e-02	5.0000e-02	3.9014e-02
σ_{17}	7.3520e-09	5.0036e-02	5.0000e-02	2.6729e-02
σ_{18}	3.6804e-10	5.0003e-02	5.0000e-02	1.8650e-02
σ_{19}	1.1868e-11	5.0000e-02	5.0000e-02	1.3613e-02
σ_{20}	1.8521e-13	5.0000e-02	5.0000e-02	1.0869e-02

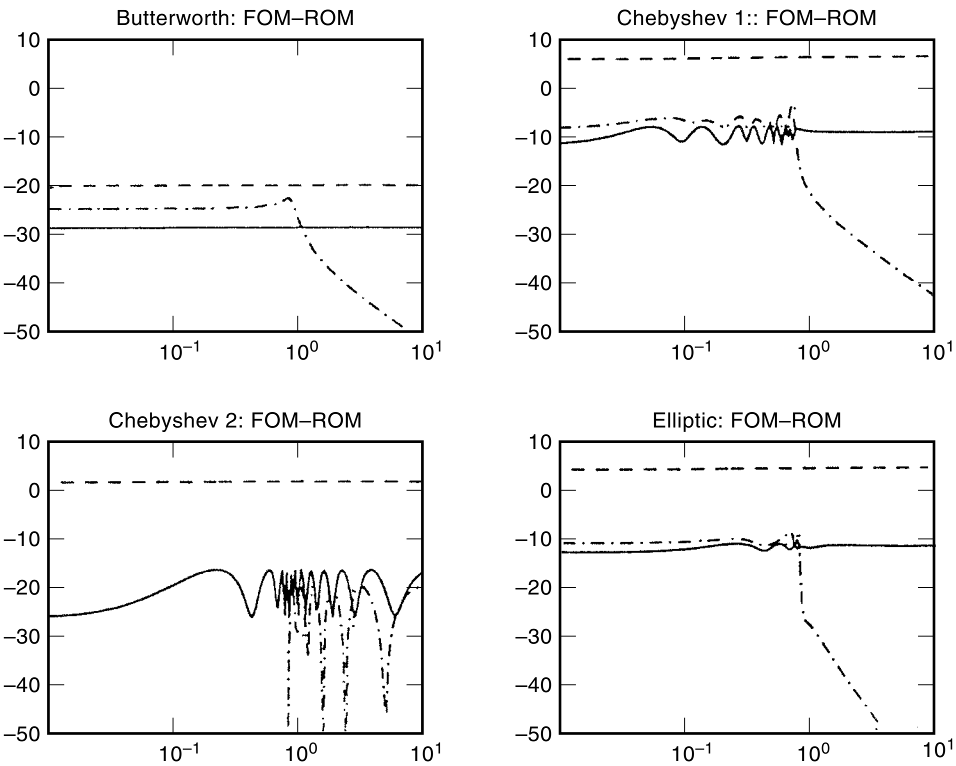
Figure 2. Analog filter approximation: Hankel singular values (curves and numerical values).

CONCLUSION

The computational algorithms that emerged from balancing and Hankel-norm model reduction have found their way to software packages. The Matlab toolboxes, robust control toolbox (13) and μ-toolbox (14), contain m-files which address the approximation problems discussed above. Two such m-files are `sysbal` and `hankmr`; the first is used for balancing and balanced truncation, while the second is used for optimal Hankel-norm approximation (including the computation of the anti-stable part of the all-pass dilation).

We will conclude with a brief discussion of some articles which were written since Glover's seminal paper (4), namely, Refs. 9, 12, 15–20.

Hankel-norm approximation or balanced truncation can be applied to stable systems. The approximants are guaranteed to be close to the original system, within the bounds given in the section entitled "Error Bounds of Optimal and Suboptimal Approximants"; the important aspect of this theory is that these bounds can be computed *a priori*—that is, before computing the approximant. The bounds are in tems of the $\mathcal{H}_\infty$ norm which is a well-defined measure of distance between



	$\mathcal{H}$ •Norm of the Error and Bounds			
	σ_9	$\|\|\Sigma - \Sigma_{\text{hank}}\|\| \bullet$	$\|\|\Sigma - \Sigma_{\text{bal}}\|\| \bullet$	$2(\sigma_9 + \ldots + \sigma_{20})$
Butterworth	0.0383	0.0388	0.0779	0.1035
Chebyshev 2	0.0506	0.1008	0.0999	1.2015
	σ_{11}	$\|\|\Sigma - \Sigma_{\text{hank}}\|\| \bullet$	$\|\|\Sigma - \Sigma_{\text{bal}}\|\| \bullet$	$2(\sigma_{11} + \ldots + \sigma_{20})$
Chebyshev 1	0.3374	0.4113	0.6508	1.9768
Elliptic	0.2457	0.2700	0.3595	1.5818

Figure 3. Analog filter approximation: Bode plots of the error systems for model reduction by optimal Hankel-norm approximation (continuous curves), balanced truncation (dash–dot curves), and the upper bound [Eq. (57)] (dash–dash curves). The table compares the peak values of these Bode plots with the lower and upper bounds predicted by the theory.

stable systems and has a direct relevance in feedback control theory.

When the system is unstable, however, Hankel-norm approximation may not be the right thing to do, since it would not predict closeness in any meaningful way. To bypass this difficulty Ref. 15 proposes approximating unstable systems (in the balanced and Hankel norm sense) using the (normalized) *coprime factors,* derived from the transfer function. Another way to go about reducing unstable systems is by adopting the *gap metric* for measuring the distance between two systems; the gap is a natural measure of distance in the context of feedback control. For details on this approach we refer to the work by Georgiou and Smith (16), and references therein.

The second paper (9) authored by Fuhrmann uses the author's *polynomial models* for a detailed analysis of the Hankel operator with given (scalar) rational symbol (i.e., single-input single-output systems). Furhmann's computations follow a different approach than the one used in Ref. 4. The Schmidt pairs (singular vectors) are explicitly computed and hence the optimal Hankel approximants are obtained. This analysis yields new insights into the problem; we refer to Theorem 8.1 on page 184 of Ref. 9, which shows that the decomposition of the transfer function in the basis provided by the Schmidt pairs of the Hankel operator provides the balanced realization of the associated state space representation. Another advantage is that it suggests the correct treatment of the polynomial equation [Eq. (38)], which yields a set of n linear equa-

Table 2. Various Upper Bounds for the Norm of the Error

Analog Filter	Optimal D	σ_9	$\|\Sigma - \Sigma_{\text{hank}}\|_\infty$	Bound 11	Bound 12
Butterworth	0.0384	0.0384	0.0389	0.0390	0.0517
Chebyshev 2	0.1010	0.0506	0.1008	0.5987	0.6008
	Optimal D	σ_{11}	$\|\Sigma - \Sigma_{\text{hank}}\|_\infty$	Bound 21	Bound 22
Chebyshev 1	0.2988	0.3374	0.4113	0.5030	0.9839
Elliptic	0.2441	0.2458	0.2700	0.3492	0.7909

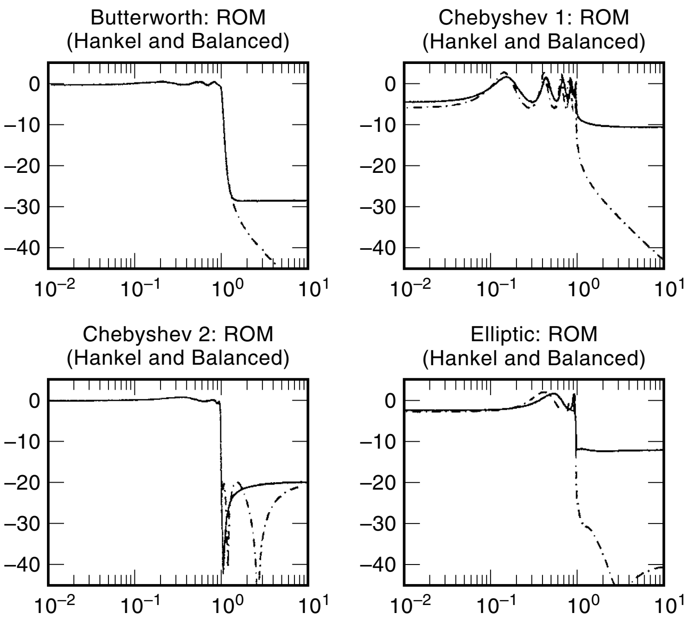


Figure 4. Analog filter approximation: Bode plots of the reduced-order models obtained by balanced truncation (dash–dot curves) and Hankel-norm approximation (continuous curves).

tions, instead of $2n - 1$ as in Eq. (39). Further developments along the same lines are given in Ref. 21.

Weighted Hankel-norm approximation was introduced in Ref. 17. Reference 12 presents a new frequency-weighted balanced truncation method. It comes with explicit $\mathscr{L}_\infty$ norm error bounds. Furthermore, the weighted Hankel-norm problem with anti-stable weighting is solved. These results are applied to $\mathscr{L}_\infty$-norm model reduction by means of Hankel-norm approximation.

Reference 18 discusses a rational interpolation approach to Hankel-norm approximation.

The next article (19) addresses the issue of balanced truncation for second-order form linear systems. These are systems which arise in applications and which are modeled by position and velocity vectors. Thus in simplifying, if one decides to eliminate the variable q_i, the derivative $\dot{q}_i$ should be eliminated as well, and vice versa. Toward this goal, new sets of invariants are introduced and new grammians as well. However, no bounds on the $\mathscr{H}_\infty$ norm of the error are provided.

The question whether in bounds in Eqs. (57) and (62) are tight arises. It can be shown that this property holds for systems having zeros which interlace the poles; an example of such systems are *RC* circuits [this result follows from Eq. (57) and Remark 5(d), after noticing that the Hankel operator in this case is positive definite]. For an elementary exposition of this result we refer to Ref. 22; for a more abstract account, see Ref. 23.

The problems of approximating linear operators in the 2-induced norm, which are (a) finite-dimensional, unstructured and (b) infinite-dimensional structured (Hankel), have been solved. The solutions of these two problems exhibit striking similarities. These similarities suggest the search of a *unifying framework* for the approximation of linear operators in the 2-induced norm. For details see Ref. 24. In particular the problem of approximating a finite-dimensional Hankel matrix remains largely unknown. Some partial results are provided in Ref. 20 which relate to results in Ref. 9.

BIBLIOGRAPHY

1. G. W. Stewart and J. Sun, *Matrix Perturbation Theory,* New York: Academic Press, 1990.
2. V. M. Adamjan, D. Z. Arov, and M. G. Krein, Analytic properties of Schmidt pairs for a Hankel operator and the generalized Schur–Takagi problem, *Math. USSR Sbornik,* **15**: 31–73, 1971.
3. V. M. Adamjan, D. Z. Arov, and M. G. Krein, Infinite block Hankel matrices and related extension problems, *Amer. Math. Soc. Trans.,* **111**: 133–156, 1978.
4. K. Glover, All optimal Hankel-norm approximations of linear multivariable systems and their $\mathscr{L}^\infty$-error bounds, *Int. J. Control,* **39**: 1115–1193, 1984.
5. M. Green and D. J. N. Limebeer, *Linear Robust Control,* Upper Saddle River, NJ: Prentice-Hall, 1995.
6. K. Zhou, J. C. Doyle, and K. Glover, *Robust and Optimal Control,* Upper Saddle River, NJ: Prentice-Hall, 1996.
7. A. C. Antoulas, Lectures on optimal approximation of linear systems, Draft, Dept. Elec. Comp. Eng., Rice Univ., Houston, TX, 1998.
8. P. A. Fuhrmann, *Linear Systems and Operators in Hilbert Space,* New York: McGraw-Hill, 1981.
9. P. A. Fuhrmann, A polynomial approach to Hankel norm and balanced approximations, *Linear Alg. Appl.,* **146**: 133–220, 1991.
10. J. A. Ball, I. Gohberg, and L. Rodman, *Interpolation of Rational Matrix Functions,* Basel: Birkhäuser Verlag, 1990.
11. R. Ober, Balanced parameterization of classes of linear systems, *SIAM J. Control Optim.,* **29**: 1251–1287, 1991.
12. K. Zhou, Frequency-weighted $\mathscr{L}_\infty$ norm and optimal Hankel norm model reduction, *IEEE Trans. Autom. Control,* **40**: 1687–1699, 1995.
13. R. Y. Chiang and M. G. Safonov, *Robust Control Toolbox,* Natick, MA: The MathWorks, 1992.
14. G. J. Balas et al., *μ-analysis and Synthesis Toolbox,* Natick, MA: The MathWorks, 1993.
15. D. G. Meyer, A fractional approach to model reduction, *Proc. Amer. Control Conf.,* Atlanta, 1988, pp. 1041–1047.
16. T. T. Georgiou and M. C. Smith, Approximation in the gap metric: Upper and lower bounds, *IEEE Trans. Autom. Control,* **38**: 946–951, 1993.
17. G. A. Latham and B. D. O. Anderson, Frequency-weighted optimal Hankel normal approximation of stable transfer functions, *Syst. Control Lett.,* **5**: 229–236, 1985.
18. T. Auba and Y. Funabashi, Interpolation approach to Hankel norm model reduction for rational multi-input multi-output systems, *IEEE Trans. Circuits Syst. I; Fundam. Theory Appl.,* **43**: 987–995, 1996.
19. D. G. Meyer and S. Srinivasan, Balancing and model reduction for second-order form linear systems, *IEEE Trans. Autom. Control,* **41**: 1632–1644, 1996.
20. A. C. Antoulas, On the approximation of Hankel Matrices, in U. Helmke, D. Prätzel-Wolters, and E. Zerz, (eds.), *Operators, Systems and Linear Algebra,* Stuttgart: Teubner Verlag, 1997, pp. 17–23.
21. P. A. Fuhrmann and R. Ober, A functional approcah to LQG balancing, *Int. J. Control,* **57**: 627–741, 1993.
22. B. Srinivasan and P. Myszkorowski, Model reduction of systems zeros interlacing the poles, *Syst. Control Lett.,* **30**: 19–24, 1997.

23. R. Ober, On Stieltjes functions and Hankel operators, *Syst. Control Lett.,* **27**: 275–278, 1996.

24. A. C. Antoulas, Approximation of linear operators in the 2-norm, *Linear Algebra Appl.,* special issue on Challenges in Matrix Theory, 1998.

A. C. ANTOULAS
Rice University

LINEAR ELECTRIC COMPONENTS. See LINEAR NETWORK ELEMENTS.

LINEAR FREQUENCY MODULATION. See CHIRP MODULATION.

LINEARLY SEPARABLE LOGIC. See THRESHOLD LOGIC.