Memorandum COSOR 94-38

Optimal Hankel Norm Identification
of Dynamical Systems

S. Weiland
A.A. Stoorvogel

Eindhoven, November 1994
The Netherlands
Optimal Hankel Norm Identification of Dynamical Systems

Siep Weiland
Department of Electrical Engineering
Eindhoven University of Technology
P.O. Box 513
5600 MB Eindhoven
The Netherlands
E-mail: s.weiland@ele.tue.nl

Anton A. Stoorvogel*
Department of Mathematics
Eindhoven University of Technology
P.O. Box 513
5600 MB Eindhoven
The Netherlands
E-mail: wscoas@win.tue.nl

November 1994

Abstract

The problem of optimal approximate system identification is addressed with a newly defined measure of misfit between observed time series and linear time-invariant models. The behavioral framework is used as a suitable axiomatic setting for a non-parametric introduction of system complexity and a notion of misfit of dynamical systems which is independent of system representations. The misfit function introduced here is characterized in terms of the induced norm of a Hankel operator associated with the data and a co-inner kernel representation of a model. Two optimal approximate identification problems are considered in this framework. New conceptual algorithms are proposed for optimal approximate identification of time series.

Key Words

System identification, approximate modeling, Hankel operators, behavioral theory, linear systems.

*The research of dr. A.A. Stoorvogel has been made possible by a fellowship of the Royal Netherlands Academy of Sciences and Arts.
1 Introduction

The problem of identifying models from observed time series is of paramount importance in identification theory and is at the basis of many investigations in descriptive sciences and control and estimation theory. A key issue in model identification is to find a compromise between model accuracy and model complexity. Accurate models are characterized by their ability to reproduce observed time series of a data generating system with high precision. The complexity of a dynamical system is an indicator of the size (or dimension) of the number of independent time series which are compatible with the laws defining the system. Following Popper's philosophy on assuming falsifiability as a criterion for assessing model quality, it is logical to formalize a framework in which simple and accurate models are considered most useful because they explain observed data well and are more likely to falsify other observations than complex models.

It is the purpose of this paper to present new criteria for model accuracy and model complexity and to study their implications for the approximate model identification problem. We depart from the standard setting of the identification problem by refraining from a-priori assumptions on the input-output structure of the to be identified system. Roughly speaking, model complexity is defined in terms of the total degree of freedom to uniquely specify a trajectory of the system. Autonomous models are viewed as the least complex systems irrespective of their state space dimension. A natural distance measure between data and model is introduced which is independent of representations of models. It is shown that this non-parametric misfit measure corresponds to the Hankel norm of an operator which is associated with the data and with co-inner kernel representations of the model. With this characterization, the misfit between data and model can actually be computed. Two optimal approximate identification problems are formulated which consider complexity and misfit as optimization criteria. Algorithms are given to compute optimal approximate models by performing optimal Hankel norm reductions.

We believe that the identification methods discussed here have important implications not only for applications in system identification, but also for problems related to data reduction, model approximation, the analysis of model uncertainty and robust control.

The paper is organized as follows. In section 2 we define a model class of linear shift-invariant complete and closed subsets of $l_2$. It is shown that systems in this class can be represented both as the kernel as well as the image of a rational operator. These operators allow for interpretations both in the time and the frequency domain. The equivalence class of all kernel representations defining the same system is characterized. In section 3 the identification problem is addressed. It is shown that for an important class of time series the least complex models which are able to reproduce the data are autonomous. All models which do not falsify the data are derived from such an autonomous model in a straightforward way using a simple parametrization. Definitions for system complexity and misfit between model and data are given in section 4. System complexity is characterized in terms of kernel representations and the misfit function is characterized as the induced norm of a Hankel operator which is defined by the data and whose domain of definition is the orthogonal complement of the set of trajectories compatible with a model. Two types of optimal approximate modeling problems are formalized and partially solved in section 5. Conceptual algorithms are discussed in this section. The general identification problem to minimize misfit under a complexity constraint is further considered in section 6. A suboptimal solution is presented and the optimization problem is reformulated and characterized as an algebraic problem and a functional analytic problem. Conclusions of the paper are deferred to section 7.
The notation used in the paper is standard. For any subset of integers \( T \subseteq \mathbb{Z} \) we denote by \( l_2(T, \mathbb{R}^q) \) the set of functions \( w : T \to \mathbb{R}^q \) for which
\[
\| w \|_2^2 := \sum_{t \in T} \| w(t) \|^2 < \infty.
\]
Here, \( \| \cdot \| \) is the standard Euclidean norm in \( \mathbb{R}^q \). If dimensions of objects are understood from the context and if \( T = \mathbb{Z}_+ := \mathbb{Z} \cap [1, \infty) \) then we write \( l_2^+(\mathbb{Z}_+, \mathbb{R}^q) \) for \( l_2(\mathbb{Z}_+, \mathbb{R}^q) \). Let \( L_2 \) denote the space of all functions defined on \( \mathbb{C} \) which are square summable on the unit circle \( \{ z \in \mathbb{C} | |z| = 1 \} \). Any such function can be expanded as a power series \( \sum_{t \in \mathbb{Z}} w(t) z^{-t}, \; z \in \mathbb{C}, \) with \( w \in l_2(\mathbb{Z}, \mathbb{R}^q) \). Let \( \mathcal{H}^+_2 \) and \( \mathcal{H}^-_2 \) denote the Hardy spaces of square summable functions on the unit circle with analytic continuation outside and inside the unit circle, respectively. Similarly, \( \mathcal{H}^\infty_2 \) and \( \mathcal{H}^\infty_\infty \) denote the Hardy spaces of complex valued functions which are bounded on the unit circle with analytic continuation in \( |z| < 1 \) and \( |z| > 1 \), respectively. We will use the prefix \( R \) to denote rational elements of Hardy spaces, i.e., \( R\mathcal{H}^+_2, \; R\mathcal{H}^\infty_2, \) etc. Clearly, \( \mathcal{H}^+_2 \) and \( \mathcal{H}^-_2 \) are Hilbert spaces with their natural norms and inner products and \( L_2 = \mathcal{H}^-_2 \oplus \mathcal{H}^+_2 \). We will write \( \Pi_+ \) and \( \Pi_- \) for the canonical projections of \( L_2 \) onto \( \mathcal{H}^+_2 \) and \( \mathcal{H}^-_2 \), respectively.

### 2 Rational representations of dynamical systems

Following the framework of Willems [6, 8], a model, or a system, is any subset \( B \) of the power set \( W^T \) of all functions \( w : T \to W \). That is, a system is viewed as a collection of trajectories \( B \subseteq W^T \). A system \( B \) is called an \( l_2 \) system if \( B \subseteq l_2(T, W) \). Throughout this paper we consider discrete time systems defined for positive time, i.e., it is assumed that \( T = \mathbb{Z}_+ = \{ t \in \mathbb{Z} | t > 0 \} \).

A few basic concepts of the behavioral framework need to be reviewed here. A system \( B \) is called linear if \( B \) is a linear subspace of \( W^T \), it is said to be shift invariant if \( \sigma B \subseteq B \). Here, \( \sigma \) denotes the usual shift \( (\sigma w)(t) = w(t + 1) \). A system is said to be complete if \( w_{[1, t]} \in B_{[1, t]} \) for all \( t > 0 \) implies that \( w \in B \). A system \( B \) is said to be autonomous if there exists a \( t > 0 \) such that the mapping \( \pi_T : B \to B_{[1, t]} \), defined by the restriction \( \pi_T(w) := w_{[1, t]} \), is injective. Trajectories of an autonomous system are therefore uniquely determined once their values are known on a sufficiently long time interval. A system is called controllable if for all \( w_1, w_2 \in B \) and all \( t_0 \geq 0 \), there exists \( t_1 \geq t_0 \) and \( w \in B \) such that:
\[
w(t) = \begin{cases} w_1(t), & t \leq t_0 \\ w_2(t - t_1 + t_0), & t \geq t_1 \end{cases}
\]
A controllable system has therefore the property that any trajectory in its behavior can be steered in finite time to any other trajectory in the behavior of the system.

Let \( \mathbb{B}_2 \) denote the class of all linear, shift invariant, complete and closed subsets of \( l_2(\mathbb{Z}_+, \mathbb{R}^q) \). A model \( B \in \mathbb{B}_2 \) can be represented as the kernel (or null space) of a rational operator. Let \( \Theta \in R\mathcal{H}^\infty_\infty \) have dimension \( g \times q \) and let \( \sigma \) denote the shift \( (\sigma w)(t) = w(t + 1) \). Associate with \( \Theta \) an operator \( \Theta(\sigma) : l_2(\mathbb{Z}_+, \mathbb{R}^q) \to l_2(\mathbb{Z}_+, \mathbb{R}^q) \) as follows. Write
\[
\Theta(\sigma) := \Theta_0 + \Theta_1 \sigma + \cdots + \Theta_k \sigma^k + \cdots
\]
where $\Theta_k \in \mathbb{R}^{k \times q}$ are constant real matrices which uniquely define the Laurent series expansion $\Theta(z) = \sum_{k=0}^{\infty} \Theta_k z^k$ with $z \in \mathbb{C}$. Then $\Theta(\sigma) : w \mapsto \hat{v} := \Theta(\sigma)w$ is well defined by the convolution of the sequences $\{\Theta\}_{k}$ and $w(k)$. Consider the equation

$$\Theta(\sigma)w = 0 \quad (2.2)$$

which defines a system $B$ by putting

$$B = B_{\ker}(\Theta) := \{w \in l_2^+ | \Theta(\sigma)w = 0\}. \quad (2.3)$$

Any such $\Theta$ is said to be a kernel representation of $B$. Note that $B_{\ker}(\Theta)$ is a linear, time invariant and closed subset of $l_2(\mathbb{Z}_+, \mathbb{R}^d)$. We call this a “full rank” kernel representation if $\Theta$ has full rank as a rational matrix.

Let $\hat{B}$ be defined as the image of $B$ under the Laplace transform, i.e., $\hat{w} \in \hat{B}$ if and only if $\hat{w}(z) = \sum_{t \geq 0} w(t)z^{-t}$ for $w \in B$ and $z \in \mathbb{C}$. As stated before, $\Pi_+$ denotes the canonical projection $\Pi_+: l_2 \to \mathcal{H}_2^+$. Then (2.3) is equivalently described in the frequency domain by

$$\hat{B} = \{\hat{w} \in \mathcal{H}_2^+ | \Pi_+\Theta \hat{w} = 0 \} = \ker \Pi_+\Theta =: \hat{B}_{\ker}(\Theta).$$

Remark 2.1 At this stage it is important to note that a kernel representation $\Theta \in \mathcal{RH}_\infty^-$ of a dynamical system $B \in \mathbb{R}_2$ can be interpreted as either a time domain operator acting on $l_2^+$ signals or as the $\mathcal{H}_2^+$ kernel of $\Pi_+\Theta$ viewed as a multiplicative operator. We emphasize that in the present setting initial values of signals $w \in B_{\ker}(\Theta)$ are not constrained to be zero. Elements of $l_2^+$ should therefore not be thought of as signals $w \in l_2(\mathbb{Z}, \mathbb{R}^d)$ which are zero for non-positive $t \in \mathbb{Z}$. Signals on $\mathbb{C}$ with transient phenomena are naturally considered in this framework. Moreover, autonomous systems appear naturally in the above description of dynamical systems. As an example, the rational function $\Theta(z) = (z - 0.5)/(z + 2) \in \mathcal{RH}_\infty^-$ has expansion (2.1) with

$$\Theta_k = \begin{cases} -1/4 & \text{for } k = 0 \\ (-5/4)(-1/2)^k & \text{for } k \geq 1 \end{cases}$$

which yields that the $l_2^+$ kernel of $\Theta(\sigma)$ consists of exponentials of the form $w_1(t) := \lambda(1/2)^t$ with $t > 0$ and $\lambda \in \mathbb{R}$. This clearly defines a non-trivial autonomous system. The $\mathcal{H}_2^+$ kernel of $\Pi_+\Theta$ equals $\hat{B}_{\ker}(\Theta) = \{\lambda(z - 0.5)^{-1} \mid \lambda \in \mathbb{R}, z \in \mathbb{C}\}$. Note that the $\mathcal{H}_2^+$ kernel of $\Theta$ (when viewed as a multiplicative operator) contains only the trivial function $\hat{w} = 0$ which, for this reason, is a less appealing representation to describe dynamical systems. Also, $\Theta(\sigma)$ contains only $w = 0$ in its kernel when $l_2(\mathbb{Z}, \mathbb{R})$ is taken as its domain of definition.

Kernel representations of systems are never unique. A characterization of non-uniqueness of kernel representations will be used in the sequel and is given in the following theorem.

Theorem 2.2 For $i = 1, 2$, let $\Theta_i \in \mathcal{RH}_\infty^-$ be a full rank kernel representation of $B_i = B_{\ker}(\Theta_i)$. Then $B_1 = B_2$ if and only if there exist a unit $^1U \in \mathcal{RH}_\infty^-$ such that $\Theta_2 = U\Theta_1$.

Proof. (if) Suppose that $\Theta_2 = U\Theta_1$ for some unit $U \in \mathcal{RH}_\infty^-$. Let $w \in B_1$ and observe that $\hat{\nu} := \Theta_1\hat{w} \in \mathcal{H}_2^-$. Then also $U\hat{\nu} = U\Theta_1\hat{w} = \Theta_2\hat{w} \in \mathcal{H}_2^-$ which implies that $\Pi_+\Theta_2\hat{w} = 0$. Hence

$^1$A unit $U \in \mathcal{RH}_\infty^-$ is a square matrix with entries in $\mathcal{RH}_\infty^-$ whose inverse $U^{-1}$ exists and also belongs to $\mathcal{RH}_\infty^-$. 
2 RATIONAL REPRESENTATIONS OF DYNAMICAL SYSTEMS

\( w \in B_2 \) from which we conclude that \( B_1 \subseteq B_2 \). The reverse inclusion follows from a similar argument by using that \( \Theta_1 = U^{-1}\Theta_2 \) with \( U^{-1} \in \mathcal{R}_{\infty} \).

(only if) We first show that \( B_1 \subseteq B_2 \Rightarrow \Theta_2 = U_1\Theta_1 \) for some \( U_1 \in \mathcal{R}_{\infty} \). To see this, first observe that

\[
\hat{B}_{\ker}(\Theta_1) = \{ \hat{w} \in \mathcal{H}_2^+ | (\Theta_1\hat{w}, \hat{v}) = 0 \text{ for all } \hat{v} \in \mathcal{H}_2^+ \} = \\
= \{ \hat{w} \in \mathcal{H}_2^+ | (\hat{w}, \Theta_1^{-1}\hat{v}) = 0, \text{ for all } \hat{v} \in \mathcal{H}_2^+ \} = \\
= (\text{im } \Theta_1^{-1})^\perp \quad (2.4)
\]

where \( \Theta_1^{-1} : \mathcal{H}_2^+ \rightarrow \mathcal{H}_2^+ \) is the dual operator \( \Theta_1^{-1}(z) = \Theta_1(z^{-1}) \in \mathcal{R}_{\infty}^+ \). Therefore, \( B_1 \subseteq B_2 \) implies that \( B_2 \subseteq B_1 \) which in turn implies that \( \text{im } \Theta_2^\perp \subseteq \text{im } \Theta_1^\perp \).

We can extend \( \Theta_1 \) from a multiplicative operator from \( \mathcal{H}_2^+ \) to a multiplicative operator from \( \mathbb{R}(s) \) to \( \mathbb{R}(s) \). We claim that also for this extended operator there holds \( \text{im } \Theta_2^\perp \subseteq \text{im } \Theta_1^\perp \). Indeed, suppose contrary to what we want to prove that there exists \( \hat{v} \in \mathbb{R}(s) \) such that \( \Theta_2^\perp \hat{v} \notin \text{im } \Theta_1^\perp \). Let \( \hat{u} \in \mathbb{R}(s) \) be invertible and such that \( \hat{u}\hat{v} \in \mathcal{H}_2^+ \). By assumption, there exists \( h \in \mathcal{H}_2^+ \) such that \( \Theta_2^\perp \hat{u}\hat{v} = \Theta_1^{-1}h \). But then \( \Theta_2^\perp \hat{v} = \Theta_1^{-1}h\hat{u}^{-1} \) with \( \hat{u}^{-1} \in \mathbb{R}(s) \) which yields a contradiction.

\( \Theta_1^\perp \) has full column rank as a rational matrix. Hence, there exists a rational matrix \( \Theta_1^L \) such that \( \Theta_1 = \Theta_1^L\Theta_1^{-1} \). Then it is easy to check that the inclusion of the images over the field \( \mathbb{R}(s) \) implies that \( \Theta_2^\perp = \Theta_2^L U_1^\perp \) where \( U_1^\perp := \Theta_1^L \Theta_2^\perp \). It remains to show that \( U_1 \) is antistable, i.e., \( U_1 \in \mathcal{R}_{\infty}^+ \).

Since \( \text{im } \Theta_2^\perp \subseteq \text{im } \Theta_1^\perp \) with \( \Theta_1^\perp \) viewed as an operator from \( \mathcal{H}_2^+ \) to \( \mathcal{H}_2^+ \), for all \( \hat{v}_1 \in \mathcal{H}_2^+ \) there exists \( \hat{v}_2 \in \mathcal{H}_2^+ \) such that \( \Theta_2^\perp \hat{v}_2 = \Theta_1^{-1}\hat{v}_1 \). Note that \( \Theta_2^\perp(U_1^\perp \hat{v}_2 - \hat{v}_1) = 0 \). Since \( \Theta_2^\perp \) has full column rank as a rational matrix, this implies \( U_1^\perp \hat{v}_2 = \hat{v}_1 \in \mathcal{H}_2^+ \). Since \( \hat{v}_2 \) is an arbitrary element of \( \mathcal{H}_2^+ \) we find that \( U_1^\perp \) maps \( \mathcal{H}_2^+ \) to \( \mathcal{H}_2^+ \) and hence \( U_1 \) is in \( \mathcal{R}_{\infty}^+ \).

Similarly, from the reverse inclusion \( B_2 \subseteq B_1 \) we infer that \( \exists U_2 \in \mathcal{R}_{\infty}^+ \) such that \( \Theta_1 = U_2\Theta_2 \). Therefore, \( \Theta_2 = U_1 U_2 \Theta_2 \) and \( \Theta_1 = U_2 U_1 \Theta_1 \) which implies that \( U_1 U_2 = U_2 U_1 = I \). In particular, it follows that \( U_1 \) and \( U_2 \) are square matrices and units in \( \mathcal{R}_{\infty}^+ \) which yields the result.

Besides kernel representations, systems with behavior \( B \in \mathbb{B}_2 \) also admit image representations. However, for details on image representations in this framework we refer to [11]. If \( B \in \mathbb{B}_2 \) has complexity\(^2\) \( c(B) = (m, n) \), then \( \hat{B} \) admits an image representation in the frequency domain:

\[
\hat{B} = \left\{ \Pi_+\Psi s \mid s = \begin{pmatrix} s_1 \\ s_2 \end{pmatrix} \in \ell_2^m, \quad s_2 \in (\mathcal{H}_2^-)^q - m \right\}
\]

where \( \Psi \in \mathcal{R}_{\infty}^+ \) is a square matrix of full rank which admits a partitioning \( \Psi = [\Psi_c \quad \Psi_a] \) where \( \Psi_c \in \mathcal{R}_{\infty}^+ \) is of dimension \( q \times m \) and \( \Psi_a \in \mathcal{R}_{\infty}^+ \) is of dimension \( q \times (q - m) \). The matrices \( \Psi_c \) and \( \Psi_a \) constitute a decomposition of \( \hat{B} \) in a controllable and an autonomous part.

Precisely, \( \hat{B} = \hat{B}_c + \hat{B}_a \) where \( \hat{B}_c = \Pi_+\Psi_c \ell_2^m \) is the controllable part of \( \hat{B} \) and \( \hat{B}_a = \Pi_+\Psi_a \mathcal{H}_2^- \) is a (non-unique) autonomous part of \( \hat{B} \). In particular, we infer that

- \( B \) is autonomous if and only if

\[
\hat{B} = \left\{ \Pi_+\Psi_a s \mid s \in \mathcal{H}_2^- \right\}
\]

with \( \Psi_a \in \mathcal{R}_{\infty}^+ \) a square matrix of full rank.

\(^2\)A formal definition of system complexity is given in section 4.
• $\mathcal{B}$ is controllable if and only if

$$\mathcal{B} = \{ \Pi_+ \Psi_c s \mid s \in \mathcal{L}_2 \}$$

with $\Psi_c \in \mathcal{RH}_{2\infty}$ a (possibly) non-square matrix.

Using the fact that a kernel representation is unique up to multiplication we can easily see that there exists a kernel representation with $\Theta \in \mathcal{RH}_{2\infty}$ a co-inner matrix\(^3\). We can then construct an image representation from such a co-inner kernel representation.

Suppose that $\mathcal{B} = \mathcal{B}_{\text{ker}}(\Theta)$ with $\Theta$ co-inner. If $\mathcal{B}$ is autonomous (i.e., $\Theta$ is square) there holds

$$\mathcal{B} = \{ \Pi_+ \Theta^- s \mid s \in \mathcal{H}_2^- \}$$

Otherwise first construct $\Theta_c$ such that:

$$\begin{pmatrix} \Theta \\ \Theta_c \end{pmatrix}$$

is a square and inner matrix in $\mathcal{RH}_{2\infty}$. If $\mathcal{B}$ is controllable then

$$\mathcal{B} = \{ \Pi_+ \Theta_c^- s \mid s \in \mathcal{L}_2 \}$$

and, in general, we get:

$$\mathcal{B} = \left\{ \Pi_+ \left( \Theta_c^- \Theta^- \right) \begin{pmatrix} s_1 \\ s_2 \end{pmatrix} \mid s_1, s_2 \in (\mathcal{H}_2^-)^{n-m} \right\}.$$

### 3 Data and unfalsified models

Consider a finite set of $q$ dimensional time-series

$$\tilde{w}_i : T \rightarrow W, \ i = 1, \ldots, n \quad (3.1)$$

where $n > 0$ is the number of observed time-series, $T \subseteq \mathbb{R}$ is the time set and $W := \mathbb{R}^q$ is the signal space in which the observed variables take their values. A fundamental problem is to find an optimal model which explains this data.

A system $\mathcal{B}$ is unfalsified by the data (3.1) if $\tilde{w}_i \in \mathcal{B}$ for $i = 1, \ldots, n$. A system $\mathcal{B}_1$ is said to be more powerful than $\mathcal{B}_2$ if $\mathcal{B}_1 \subseteq \mathcal{B}_2$. Hence, the more powerful a system is, the more trajectories it refutes, the more likely it is able to falsify observations.

The most powerful unfalsified model in $\mathbb{B}_2$, based on the data (3.1), is a model $\mathcal{B}_{\text{MPUM}} \in \mathbb{B}_2$ which is unfalsified by (3.1) and which is more powerful than any other unfalsified model $\mathcal{B} \in \mathbb{B}_2$. It has been shown (see [6, 8]) that for any data set (3.1) with $\tilde{w}_i \in l_2(\mathbb{Z}_+, W)$ the most powerful unfalsified model $\mathcal{B}_{\text{MPUM}} \in \mathbb{B}_2$ for (3.1) exists and is unique.

Let $W$ denote the Laplace transform of the data matrix $[\tilde{w}_1, \ldots, \tilde{w}_n]$, i.e.,

$$W(z) := \tilde{W}(1)z^{-1} + \tilde{W}(2)z^{-2} + \ldots, \quad (3.2)$$

where $\tilde{W}(t) = [\tilde{w}_1(t), \ldots, \tilde{w}_n(t)]$, $t \in \mathbb{Z}_+$. The Laplace transform then defines a bijection between the data (3.1) and $W$. We will make the following assumption on the data

---

\(^3\)A matrix $\Theta \in \mathcal{RH}_{2\infty}$ is called co-inner if it is norm preserving on the orthogonal complement of its kernel or, equivalently, if $\Theta \Theta^- = I$ where $\Theta^-(z) := \Theta^T(z^{-1})$.
4 THE APPROXIMATE MODELING PROBLEM

Assumption 3.1 \( W \in \mathcal{RH}_2^+ \).

Remark 3.2 Typical examples of these data sets include finite length measurements, finite sets of frequency response measurements and polynomial exponential time-series. A study of data sets of this type recently appeared in [2]. See [9] for a methodology to approximate data sets by polynomial-exponential time series which satisfy assumption 3.1 using risk minimization techniques.

The exact modeling problem amounts to characterizing the set of all kernel representations of unfalsified models of the data (3.1). Let \( \mathcal{M} \) denote the family of all kernel representations of (3.1), i.e.,

\[
\mathcal{M} := \{ \Theta \in \mathcal{RH}_\infty^- | \Pi_+ \hat{\Theta} \hat{w}_i = 0, \ i = 1, \ldots, n \}.
\]

The following theorem gives a complete characterization of all unfalsified models of (3.1) [1].

Theorem 3.3 Suppose that assumption 3.1 holds and suppose that \( W = \Theta_{\text{MPUM}}^{-1} \psi_{\text{MPUM}} \) is a left coprime factorization over \( \mathcal{RH}_\infty^- \) of \( W \). Then

1. \( \Theta_{\text{MPUM}} \in \mathcal{M} \)
2. \( B_{\text{MPUM}} = B_{\text{ker}}(\Theta_{\text{MPUM}}) \)
3. \( B_{\text{MPUM}} \) is an autonomous system.

Moreover,

\[
\mathcal{M} = \{ \Theta | \Theta = \Lambda \Theta_{\text{MPUM}} \text{ with } \Lambda \in \mathcal{RH}_\infty^- \}
\]

Theorem 3.3 has the interpretation that all unfalsified models (2.3) for the data (3.1) are generated from the most powerful unfalsified \( B_{\text{MPUM}} \) by left-multiplication of a kernel representation \( \Theta_{\text{MPUM}} \) of \( B_{\text{MPUM}} \) by elements \( \Lambda \in \mathcal{RH}_\infty^- \).

Note that \( \Theta_{\text{MPUM}} \) is a square \( q \times q \) rational matrix which, by theorem 2.2, is uniquely determined up to pre-multiplication by units in \( \mathcal{RH}_\infty^- \). Observe also that the corresponding autonomous behavior \( B_{\text{ker}}(\Theta) \) is a finite dimensional subspace of \( \mathcal{H}_2^+ \).

4 The approximate modeling problem

In order to formalize the approximate modeling problem we will specify the complexity of a system \( B \in \mathcal{B}_2 \) together with a criterion of misfit which expresses to what extent a model fails to explain a given data set (3.1). Approximate modeling then amounts to finding minimal misfit models which satisfy a complexity constraint or, alternatively, finding minimal complexity models subject to a tolerated misfit constraint between model and data.
4 THE APPROXIMATE MODELING PROBLEM

4.1 Model complexity

The more trajectories a system can generate, the higher its complexity. We formalize this by first introducing the relative complexity of a system viewed over time segments of length \( t > 0 \) [7].

**Definition 4.1** The relative complexity \( c_{\text{rel}} \) of a dynamical system \( B \in \mathcal{B}_2 \) is defined as \( c_{\text{rel}} : \mathbb{Z}_+ \times \mathcal{B}_2 \rightarrow \mathbb{R}_+ \) with

\[
c_{\text{rel}}(t, B) := \frac{\dim(\mathcal{B}_{[1,t]})}{t}.
\]

Clearly, for \( B \in \mathcal{B}_2 \) the function \( c_{\text{rel}}(\cdot, B) \) is monotone non-increasing. Therefore, the limit \( \lim_{t \to \infty} c_{\text{rel}}(t, B) \) exists. System complexity is defined in terms of the limit behavior of \( c_{\text{rel}} \) as follows.

**Definition 4.2** The complexity \( c \) of a dynamical system \( B \in \mathcal{B}_2 \) is a mapping \( c : \mathcal{B}_2 \rightarrow [0, q] \times [0, \infty) \) defined as \( c(B) := (m(B), n(B)) \) where

\[
m(B) := \lim_{t \to \infty} c_{\text{rel}}(t, B)
\]

\[
n(B) := \lim_{t \to \infty} \frac{t(c_{\text{rel}}(t, B) - m(B))}{t}
\]

**Remark 4.3** It has been shown in [7] that the complexity is a well defined integer valued function from \( \mathcal{B}_2 \) to \([0, q] \times [0, \infty)\). Intuitively, \( m(B) \) indicates the dimension of the subspace of the external signal space \( \mathbb{R}^q \) which is not restricted in any time interval by the laws of \( B \). Its complement \( p(B) := q - m(B) \) can be interpreted as the minimal number of independent laws which define \( B \). One can therefore think of \( m(B) \) as the dimension of the input-space and \( p(B) \) as the dimension of the output space in any input-output representation of \( B \). The number \( n(B) \) denotes the rate of convergence of the relative complexity \( c_{\text{rel}}(\cdot, B) \) to \( m(B) \) and corresponds to the dimension of the space of initial conditions of \( B \). That is, \( n(B) \) equals the state space dimension in any minimal state space representation of \( B \).

It has been shown in [7] (Theorem 25) that \( m(B) \) and \( n(B) \) are well defined for any \( B \in \mathcal{B}_2 \). There exists an elegant relation between the complexity of a system \( B \in \mathcal{B}_2 \) and the number and degree of independent laws which constitute a kernel representation of \( B \). Consider all \( \Theta \in \mathcal{R}\mathcal{H}_\infty^\cdot \) such that \( B = B_{\text{ker}}(\Theta) \). An element \( \Theta^* \in \mathcal{R}\mathcal{H}_\infty^- \) in this set is said to be a minimal kernel representation of \( B \) if (i) \( \Theta^* \) has full rank and (ii) it has minimal McMillan degree in the set of all kernel representations of \( B \).

**Proposition 4.4** [7] Let \( B \in \mathcal{B}_2 \) and suppose that \( \Theta^* \) is a minimal kernel representation of \( B \). Then its complexity \( c(B) = (m(B), n(B)) \) where \( m(B) = q - \text{rank}(\Theta^*) \) and \( n(B) \) is the McMillan degree of \( \Theta^* \).

We impose a lexicographic ordering on system complexities. That is, \( c(B_1) \leq c(B_2) \) if either \( m(B_1) < m(B_2) \) or \( m(B_1) = m(B_2) \) and \( n(B_1) \leq n(B_2) \). Since for autonomous systems \( m(B) = 0 \), this ordering implies that autonomous systems are always regarded as less complex than non-autonomous systems. Further, a system \( B_1 \) with state space dimension \( n(B_1) \) is considered more complex than \( B_2 \) with \( n(B_2) < n(B_1) \) if the input dimensions are equal, i.e., if \( m(B_1) = m(B_2) \).
4.2 Misfit

The discrepancy between model and data is formalized by a misfit function which is independent of specific parametrizations of the model.

Consider an element \( v \in \mathcal{H}_2^+ \) and associate with it the continuous linear functional \( \langle \cdot, v \rangle : \mathcal{H}_2^+ \rightarrow \mathbb{R} \) defined by the natural inner product \( \langle w, v \rangle \) in \( \mathcal{H}_2^+ \). Then \( v \) naturally induces a law \( \langle w, v \rangle = 0 \) on \( \mathcal{H}_2^+ \) and if for all \( w \in \mathcal{B} \) we have \( \langle w, v \rangle = 0 \), then \( v \) is referred to as a law compatible with the system \( \mathcal{B} \). For a given system \( \mathcal{B} \in \mathbb{B}_2 \), the error index of a law \( v \in \mathcal{H}_2^+ \) is the functional

\[
e_B(v) := \sup_{w \in \mathcal{B}} \frac{\langle w, v \rangle}{\|w\|_2 \|v\|_2}.
\]

Note that \( e_B(v) = 0 \) if and only if \( v \) is a law compatible with \( \mathcal{B} \). Further, \( \ker(e_B) = \mathcal{B}^\perp \) is the set of all laws compatible with \( \mathcal{B} \). Here,

\[
\mathcal{B}^\perp := \{ v \in \mathcal{H}_2^+ \mid \langle w, v \rangle = 0 \text{ for all } w \in \mathcal{B} \}
\]

is the orthogonal complement in \( \mathcal{H}_2^+ \) of \( \mathcal{B} \).

**Definition 4.5** The misfit between a model \( \mathcal{B} \) and the data \( W \) is defined as

\[
d(\mathcal{B}, W) := \sup \left\{ \frac{\langle Wx, v \rangle}{\|x\| \|v\|_2} \mid v \in \mathcal{B}^\perp, \ x \in \mathcal{H}_2^- \right\} \quad (4.1)
\]

where the inner product is the usual inner product in \( L_2 \) and the norms are the standard norms in the Hilbert spaces \( \mathcal{H}_2^+ \) and \( \mathcal{H}_2^- \).

In (4.1), the data matrix \( W \) is viewed as an operator mapping elements of \( \mathcal{H}_2^- \) into elements of \( L_2 \). Writing \( W = \Pi_- W + \Pi_+ W \), then the image of \( W \) is naturally decomposed in the orthogonal sets \( \mathcal{W}_- := \Pi_- W \mathcal{H}_2^- \) and \( \mathcal{W}_+ := \Pi_+ W \mathcal{H}_2^- \). Since \( W \in \mathcal{H}_2^+ \), it follows that \( \mathcal{W}_+ \) is finite dimensional and, in fact, \( \mathcal{W}_+ = \mathcal{B}_{\text{MPUM}} \). The misfit criterion can therefore equivalently be viewed as a distance measure between the most powerful unfalsified model implied by the data \( W \) and a model \( \mathcal{B} \in \mathbb{B}_2 \).

This is established in the following proposition which is immediate from definition 4.5.

**Proposition 4.6** Suppose that \( \mathcal{B} \in \mathbb{B}_2 \) and \( W \in \mathcal{H}_2^+ \). Then

\[
d(\mathcal{B}, W) = \sup \left\{ \frac{\langle w, v \rangle}{\|v\| \|w\|_W} \mid v \in \mathcal{B}^\perp, \ w \in \mathcal{B}_{\text{MPUM}} \right\} \quad (4.2)
\]

where

\[
\|w\|_W := \inf \left\{ \|x\| \mid x \in \mathcal{H}_2^-, \ w = \Pi_+ W x \right\}
\]

The norm \( \| \cdot \|_W \) has therefore the interpretation of a 'data weighted' norm on the elements of the most powerful unfalsified model implied by \( W \).

**Remark 4.7** We can interpret the misfit (4.1) as a measure of how far the principal axes of the data matrix \( W \) violate laws of the form \( \langle w, v \rangle = 0 \) where \( v \in \ker(e_B) = \mathcal{B}^\perp \) specify the laws which are implied by the model \( \mathcal{B} \).
Remark 4.8 Note that $d(B, W) \geq 0$ and $d(B, W) = 0$ if and only if $B$ is unfalsified by the data. Moreover, it is important to observe that the misfit 4.1 is independent of representations of systems $B \in \mathbb{B}_2$.

Remark 4.9 It is interesting to point out the relation between the expression (4.2) and the gap between the linear closed subspaces $B \subseteq l^+_2$ and the optimal model $B_{MPUM} \subseteq l^+_2$ of the data. The gap $[3, 4]$ is defined as

$$g(B, B_{MPUM}) := \sup \left\{ \frac{\langle w, v \rangle}{\| v \| \| w \|} \mid v \in \hat{B}^+, w \in \hat{B}_{MPUM} \right\}$$

where the inner product and norms are taken in $\mathcal{H}_2^+$. Note that the misfit measure $d$ incorporates a weighting along the principal axes of the data matrix $W$, whereas the gap depends only on the image of $\Pi + W$ under $\mathcal{H}_2^-$.

Example. As an example, let $q = 1$ and let the data be given by $\tilde{w}_1(t) = \lambda_1^t$, $\tilde{w}_2(t) = \varepsilon \lambda_2^t$, where $|\lambda_1| < 1$, $|\lambda_2| < 1$, $t \in \mathbb{Z}_+$ and $\varepsilon > 0$. Then $W(z) = [(z - \lambda_1)^{-1} \varepsilon (z - \lambda_2)^{-1}]$ and the corresponding most powerful unfalsified model is obviously given by

$$B_{MPUM} = \left\{ c_1 \lambda_1^t + c_2 \lambda_2^t \mid c_1, c_2 \in \mathbb{R}, t \in \mathbb{Z}_+ \right\} = \text{span} \{ \lambda_1^t, \lambda_2^t \}.$$ 

Hence, for any $\varepsilon \neq 0$, $B_{MPUM}$ does not discriminate between its basis components $\tilde{w}_1 = \lambda_1^t$ and $\tilde{w}_2 = \varepsilon \lambda_2^t$ which determine the model. Consequently, in any model approximation procedure a first order approximate model $B = \text{span} \lambda_1^t$ is independent of the value of $\varepsilon$, whereas from an identification point of view it seems logical to explicitly take this value into account as part of an approximate identification routine. 

The following result is crucial in the sequel and relates the misfit (4.1) to the Hankel norm of a specific operator. Let $G$ be a matrix with entries which are analytic on the unit circle $\{ z \in \mathbb{C} \mid |z| = 1 \}$ and associate with $G$ the following norms

$$\| G \|_{H^+} := \sup_{x \in \mathcal{H}_2^+} \frac{\| \Pi_+ G x \|}{\| x \|}$$

$$\| G \|_{H^-} := \sup_{x \in \mathcal{H}_2^-} \frac{\| \Pi_- G x \|}{\| x \|}.$$ 

Then $\| G \|_{H_+}$ is the induced norm of the Hankel operator $\Gamma^+_G := \Pi_+ G$ viewed as a multiplicative operator from $\mathcal{H}_2^+$ to $\mathcal{H}_2^+$. Similarly, $\| G \|_{H_-}$ is the induced norm of the Hankel operator $\Gamma^-_G := \Pi_- G$ mapping $\mathcal{H}_2^+$ to its range in $\mathcal{H}_2^-$. The norms are, respectively, referred to as the positive and negative Hankel norms associated with $G$. The following theorem is the main result of this section and shows that the misfit function (4.1) can in fact be computed as the Hankel norm of an operator.

Theorem 4.10 Let $\Theta \in \mathcal{R} \mathcal{H}_\infty^\infty$ be a co-inner kernel representation of a system $B \in \mathbb{B}_2$. Then

$$d(B, W) = \| \Theta W \|_{H^+} = \| W^\sim \Theta^\sim \|_{H^-}$$ 

4 THE APPROXIMATE MODELING PROBLEM

Proof. First note that with $\Theta$ co-inner,

$$\|\Theta W\|_{H_+} = \|W^\sim \Theta^\sim\|_{H_-} = \|W^\sim\|_{\text{im }\Theta^\sim} = \|W^\sim\|_{B^\perp}^2$$

where we used (2.4) and the fact $\Theta^\sim$ is norm preserving (or inner). We can write

$$\|W^\sim\|_{B^\perp} = \sup_{v \in B^\perp} \frac{\|\Pi_- W^\sim v\|}{\|v\|} = \sup_{v \in B^\perp} \frac{|\langle x, W^\sim v \rangle|}{\|v\| \|x\|} \leq \sup_{v \in B^\perp, x \in \mathcal{H}_2^-} \|v\| \|x\| = d(B, W).$$

The reversed inequality follows by the Schwartz inequality

$$\|W^\sim\|_{B^\perp} = \sup_{v \in B^\perp} \frac{\|\Pi_- W^\sim v\|}{\|v\|} = \sup_{v \in B^\perp, x \in \mathcal{H}_2^-} \|v\| \|x\| \geq \sup_{v \in B^\perp, x \in \mathcal{H}_2^-} \|v\| \|x\| = d(B, W).$$

Thus, $\|\Pi_- \Theta W\|_{H_+} = \|\Pi_- W\|_{B^\perp} \|H_+ = \|\Pi_- W^\sim \Theta^\sim\|_{H_-} = d(B, W).$ \qed

Using theorem 2.2 it is easily seen that every $B \in B_2$ admits a co-inner kernel representation $B = B_{\ker}(\Theta)$. The relationship between (positive) Hankel norms and Hankel singular values is well known [5] and yields the following corollary as an immediate consequence of theorem 4.10.

Corollary 4.11 Let $B$ be a closed subset of $l_2^+$ and let $W \in \mathcal{H}_2^+$ be rational. Then $0 \leq d(B, W) \leq \sigma_{\max}(W)$ where $\sigma_{\max}$ denotes the maximal Hankel singular value of $W$.

4.3 Formal problem statement

The question of approximate model identification usually amounts to minimizing the misfit between data and model subject to a complexity constraint on the class of admissible models. In the context of the above definitions of system complexity and misfit this leads to the following mathematical formulation of the identification problem.

Identification problem P1: Given data represented by $W \in \mathcal{H}_2^+$, together with a tolerated complexity $c_{\text{tol}} = (m_{\text{tol}}, n_{\text{tol}})$ with $m_{\text{tol}} \geq 0$ and $n_{\text{tol}} \geq 0$. Find models $B \in B_2$ of complexity $c(B) \leq c_{\text{tol}}$ such that the misfit $d(B, W)$ is minimized.

We will therefore be interested in all models $B \in B_2$ that minimize the misfit under a complexity constraint. An equally interesting methodology amounts to specifying a maximal tolerated misfit between data and model and searching for a minimal complexity model which meets a misfit constraint.

Identification problem P2: Given data represented by $W \in \mathcal{H}_2^+$ and a tolerated misfit level $d_{\text{tol}} \geq 0$. Find models $B \in B_2$ of misfit $d(B, W) \leq d_{\text{tol}}$ such that the complexity $c(B)$ is minimized.

We will address both problems in the remainder of the paper.
5 Optimal approximate models

We first consider identification problem PI where the tolerated complexity \( c_{\text{tol}} = (0, k) \) for some \( k \geq 0 \). Problem PI then amounts to finding optimal autonomous systems of prescribed dimension which minimize the misfit (4.1). The following theorem provides a complete solution to this problem.

**Theorem 5.1** Let \( W \in \mathcal{H}_2^+ \) be given by (3.2) and suppose that Assumption 3.1 holds. Let \( \sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_k \geq \ldots \geq \sigma_n > 0 \) be the ordered Hankel singular values of \( W \). Let \( W_k \in \mathcal{H}_2^+ \) be an optimal Hankel norm approximation of \( W \) with McMillan degree \( k \) and let \( B_k \) define the most powerful unfalsified model associated with the data \( W_k \).

1. \( B_k \in \mathbb{B}_2 \) is autonomous with complexity \( c(B_k) \leq (0, k) \)
2. \( d(B_k, W) = \sigma_{k+1} \)
3. \( \{c(B'_k) \leq (0, k)\} \Rightarrow \{d(B'_k, W) \geq d(B_k, W)\} \).

Conclude from Theorem 5.1 that

\[ B_k \in \arg \min \{ d(B, W) \mid B \in \mathbb{B}_2, c(B) \leq (0, k) \} \]

In other words, \( B_k \) is a solution of the approximate identification problem PI with tolerated complexity \( c_{\text{tol}} = (0, k) \). The following conceptual and constructive procedure for the computation of a kernel representation of \( B_k \) can be derived from Theorem 5.1.

**Algorithm 1:** Given the data (3.1) and a number \( k \geq 0 \).

1. Define the rational function \( W \in \mathcal{RH}_2^+ \) according to (3.2).
2. Compute an optimal Hankel norm approximant \( W_k \) of \( W \) with McMillan degree \( k \). See [5].
3. Let \( W_k = \Theta_{\text{MPUM}, k}^{-1} \Psi_k \) be a left coprime factorization over \( \mathcal{RH}_\infty^- \) of \( W_k \).
4. Put \( B_k = B_{\text{ker}}(\Theta_{\text{MPUM}, k}) \).

We emphasize the conceptual nature of Algorithm 1. It requires the explicit construction of the high order rational matrix \( W \) as input for an optimal Hankel norm approximation routine. A more direct construction of \( W_k \) on the basis of the original data is a topic of further investigation.

**Proof.** (of Theorem 5.1.) Part 1 is one of the basic properties of the most powerful unfalsified model. We focus on part 2 and 3. Let \( \Theta_k \) be a co-inner kernel representation of \( B_k \). We obtain:

\[
\| \Theta_k W \|_{H_\infty} = \| \Theta_k (W - W_k) \|_{H_\infty} = \|(W^\sim - W_k^\sim) \Theta_k^\sim \|_{H_\infty} = \|(W^\sim - W_k^\sim)_{|\Theta_k^\sim} \|_{H_\infty} \leq \|(W^\sim - W_k^\sim)\|_{H_\infty} = \sigma_{k+1}
\]

Thus, by theorem 4.10 we have \( d(B_k, W) \leq \sigma_{k+1} \). Let \( \hat{\Theta} \in \mathcal{RH}_\infty^- \) be a co-inner kernel representation of some arbitrary autonomous behavior \( \hat{B} \) of complexity \( c(\hat{B}) \leq k \). Define \( \Pi_{\hat{B}} \) as the orthogonal projection onto \( \hat{B} \). Clearly, \( \Pi_{\hat{B}} \) has rank less than or equal to \( k \). Furthermore,

\[
\| W^\sim_{|\hat{\Theta}} \|_{H_\infty} = \| W^\sim (I - \Pi_{\hat{B}}) \|_{H_\infty}
\]
since \( I - \Pi_{\hat{B}} \) is an orthogonal projection onto \( \text{im} \Theta^\sim \). We get
\[
\|\hat{\Theta}W\|_{H_+} = \|W^\sim \hat{\Theta}^\sim\|_{H_-} = \|W^\sim (I - \Pi_{\hat{B}})\|_{H_-} = \|W^\sim - W^\sim \Pi_{\hat{B}}\|_{H_-} \geq \sigma_{k+1}.
\]
(5.2)
The last inequality follows from the fact that, since \( \Pi_{\hat{B}} \) has rank \( k \), \( \Pi_{\hat{B}}W^\sim \) is, as a rank \( k \) approximant of \( \Pi_{\hat{B}}W^\sim \), never closer (in Hankel norm) than the optimal Hankel norm approximant. It is well known that the optimal Hankel norm approximant results in an error equal to \( \sigma_{k+1} \) and the result follows. (5.2) proves part 3 while (5.1) together with (5.2) for \( \hat{\Theta} = \Theta_k \) proves part 2. \( \Box \)

Next, consider problem P2. A complete solution of this optimal identification problem is given by the following theorem.

**Theorem 5.2** Let \( W \in \mathcal{H}_2^+ \) be given by (3.2) and suppose that Assumption 3.1 holds. Let \( d_{\text{tol}} \geq 0 \) be the maximal tolerated misfit and let \( \sigma_1 \geq \sigma_2 \ldots \geq \sigma_n > 0 \) be the ordered Hankel singular values of \( W \). Define
\[
B_{d_{\text{tol}}} := \begin{cases} 
B_{\text{MPUM}} & \text{if } d_{\text{tol}} < \sigma_n \\
B_k & \text{if } \sigma_k > d_{\text{tol}} \geq \sigma_{k+1} \\
\{0\} & \text{if } d_{\text{tol}} > \sigma_1
\end{cases}
\]
where \( B_{\text{MPUM}} \) is the most powerful unfalsified model of \( W \) and \( B_k \) is the model defined in the last step of Algorithm 1. Then
1. \( d(B_{d_{\text{tol}}}, W) \leq d_{\text{tol}} \).
2. \( B_{d_{\text{tol}}} \in \mathcal{B}_2 \) is autonomous
3. \( B_{d_{\text{tol}}} = \text{arg min} \{ c(B) \mid B \in \mathcal{B}_2, \ d(B, W) \leq d_{\text{tol}} \} \)

**Proof.** The statements are obvious if \( d_{\text{tol}} < \sigma_n \) or \( d_{\text{tol}} > \sigma_1 \). If \( d_{\text{tol}} > 0 \) is such that \( \sigma_k > d_{\text{tol}} \geq \sigma_{k+1} \), then theorem 5.1 states that \( B_{d_{\text{tol}}} \) is autonomous and \( d(B_{d_{\text{tol}}}, W) = \sigma_{k+1} \leq d_{\text{tol}} \) and \( c(B_{d_{\text{tol}}}) \leq (0, k) \). This yields the first two claims. Next, suppose there exists \( B \in \mathcal{B}_2 \) with \( d(B, W) \leq d_{\text{tol}} \) and \( c(B) = (m', k') < c(B_{d_{\text{tol}}}) \). Then, necessarily, \( m' = 0 \) and \( k' < k \) so that, by theorem 5.1, \( d(B, W) \geq \sigma_k > d_{\text{tol}} \). Hence, \( B \) violates the misfit constraint. Conclude that \( B_{d_{\text{tol}}} \) satisfies part 3 of the statement. \( \Box \)

Conclude from theorem 5.2 that \( B_{d_{\text{tol}}} \) is a solution of identification problem P2.

6 Minimizing the misfit – the general case

In this section we reconsider problem P1 for the case where \( 0 < m_{\text{tol}} \leq q \), i.e., we wish to minimize the misfit in the class of non-autonomous systems of given complexity.
6.1 A suboptimal solution

The following result yields a partial solution to problem P1 in the sense that it characterizes a set of models with prescribed complexity and bounded misfit.

**Theorem 6.1** Let \( W \) satisfy assumption 3.1 and let \( B_k \) be the autonomous model defined in theorem 5.1. Then all models \( B \in \mathbb{B}_2 \) with \( B_k \subseteq B \) have misfit \( d(B, W) \leq \sigma_{k+1} \) where \( \sigma_{k+1} \) is the \((k + 1)\)st largest Hankel singular value of \( W \). In particular, if \( B_k = B_{\ker}(\Theta_k) \), then for all \( \Lambda \in \mathcal{RH}_\infty^\text{c} \) the system \( B := B_{\ker}(\Lambda \Theta_k) \) has misfit \( d(B, W) \leq \sigma_{k+1} \).

**Proof.** If \( B_k \subseteq B \) then obviously \( \hat{B} \perp \subseteq \hat{B}_k \). Suppose that \( B_k \) satisfies the hypothesis of the theorem. Then, by definition of the misfit (4.1) there holds

\[
d(B, W) := \sup \left\{ \frac{\| Wx \|}{\| x \|} ; \; v \in \hat{B} \perp , \; x \in \mathcal{H}_2 \right\}
\]

\[
\leq \sup \left\{ \frac{\| Wx \|}{\| x \|} ; \; v \in \hat{B}_k \perp , \; x \in \mathcal{H}_2 \right\}
\]

\[
d(B_k, W) = \sigma_{k+1}
\]

which yields the first claim. The remaining statements are an immediate consequence of the characterization of kernel representations in Theorem 3.3. \( \square \)

**Remark 6.2** If the input dimension \( m > 0 \) is fixed then Theorem 6.1 has the interpretation that the optimal autonomous model \( B_k \) in Theorem 5.1 generates a class of models \( B = B_{\ker}(\Lambda \Theta_{\text{SPUM}, k}) \) parametrized by \( \Lambda \in \mathcal{RH}_\infty^\text{c} \) of rank \( q - m \) which have guaranteed misfit less than or equal to the \((k + 1)\)st singular value of \( W \).

6.2 An algebraic characterization

In the following we give an algebraic characterization of problem P1 using state space representations of models \( B \in \mathbb{B}_2 \). By Proposition 4.4, any \( \Theta \in \mathcal{RH}_\infty^\text{c} \) for which the complexity \( c(B_{\ker}(\Theta)) = (m, k) \) can be written as \( \Theta(z) = C(Iz - A)^{-1}B + D \), with \( p = q - m \),

\[
\begin{pmatrix} A & B \\ C & D \end{pmatrix} \in \mathbb{R}^{(k+p) \times (k+q)}. \tag{6.1}
\]

Associate with the quadruple \((A, B, C, D)\) the equations

\[
\begin{align*}
x(t + 1) &= Ax(t) + Bw(t) \\
v(t) &= Cx(t) + Dw(t)
\end{align*}
\]

which determine the output nulling state space system

\[
B_{\text{on}}(A, B, C, D) := \{ (w, x) \in l_2(\mathbb{Z}_+, \mathbb{R}^{q+k}) \mid (6.2) \text{ is satisfied with } v(t) = 0 \}.
\]

It is not entirely trivial to see that this set is non-empty. However, for \( w \in l_2^+ \) the state trajectory \( x(t) := -\sum_{j=0}^{\infty} A^{-j-1}Bw(t + j) \) is well defined for \( t > 0 \), belongs to \( l_2^+ \) and satisfies the state equation (6.2). In fact, \( B_{\text{on}}(A, B, C, D) \) is a state space representation of \( B_{\ker}(\Theta) \) in the sense that

\[
B_{\ker}(\Theta) = \{ w \in l_2^+ \mid \exists x \in l_2^+ \text{ such that } (w, x) \in B_{\text{on}}(A, B, C, D) \}.
\]
Co-innerness of $\Theta$ is easily characterized in terms of the state space parameters $(A, B, C, D)$. In fact, $\Theta$ is co-inner whenever

$$
\left( \begin{array}{cc}
A & B \\
C & D
\end{array} \right) \left( \begin{array}{cc}
A & B \\
C & D
\end{array} \right)^{T} = I.
$$

(6.3)

The misfit $d(B_{on}(A, B, C, D), W)$ can be characterized in state space parameters as follows. Suppose that $W(z) = H(zI - F)^{-1}G$ has McMillan degree $n$. Then the rational transfer function $\Theta W$ admits a state space representation

$$
\begin{align*}
\begin{bmatrix}
x_1(t+1) \\
x_2(t+1)
\end{bmatrix} &= \begin{bmatrix}
F & 0 \\
BH & A
\end{bmatrix} \begin{bmatrix}
x_1(t) \\
x_2(t)
\end{bmatrix} + \begin{bmatrix}
G \\
0
\end{bmatrix} u(t) \\
v(t) &= (DH) \begin{bmatrix}
x_1(t) \\
x_2(t)
\end{bmatrix}
\end{align*}
$$

(6.4)

Let $X \in \mathbb{R}^{k \times n}$ be a solution of the Sylvester equation

$$AX - XF + BH = 0.$$

(6.5)

Since the sets of eigenvalues of $A$ and $F$ are disjoint, such an $X$ exists. The positive Hankel norm of $\Theta W$ is equal to the largest Hankel singular value of the transfer function associated with the causal subsystem of (6.4). By performing the state space transformation

$$
\begin{bmatrix}
x_1' \\
x_2'
\end{bmatrix} = \begin{bmatrix}
I & 0 \\
-X & I
\end{bmatrix} \begin{bmatrix}
x_1 \\
x_2
\end{bmatrix}
$$

the causal subsystem of (6.4) is defined by the state space parameters $(F, G, (DH + CX), 0)$. The misfit is then equal to

$$d(B, W) = \|\Theta W\|_{H_k} = \sigma_{\text{max}} \left[ (DH + CX)(Iz - F)^{-1}G \right].$$

where $\sigma_{\text{max}}$ denotes the maximal Hankel singular value. It is well known [5] that the Hankel singular values are obtained as the square roots of the eigenvalues of $L_{c}L_{o}$ where $L_{c}, L_{o} \in \mathbb{R}^{n \times n}$ are the controllability and observability gramians defined as the unique solutions of the following two linear equations:

$$
\begin{align*}
L_{c} &= FL_{c}F^{T} + GG^{T} \\
L_{o} &= F^{T}L_{o}F + (DH + CX)^{T}(DH + CX).
\end{align*}
$$

(6.6)

Choosing a suitable basis of the state space triple $(F, G, H)$ it is possible to a priori scale the controllability gramian $L_{c} = I$. Problem P1 is then equivalent to finding a quadruple $(A, B, C, D)$ of dimensions (6.1) satisfying (6.3) such that the largest eigenvalue of the solution $L_{o}$ of (6.6) is minimal where $X$ is given by the Sylvester equation (6.5).

For such a quadruple $(A, B, C, D)$ the misfit between $B = B_{on}(A, B, C, D)$ and the data $W$ is then equal to

$$d(B, W) = \lambda^{1/2}(L_{o}).$$
6.3 A functional analytic approach

We can also approach the problem studied in the previous section using image representations of models in $\mathbb{B}_2$. We have the following expression for the misfit:

$$d(\mathcal{B}, W) = \sup_{x \in \mathcal{H}_2^-} \sup_{v \in \mathcal{B}^L} \langle Wx, v \rangle$$

For each $x \in \mathcal{H}_2^-$, there exist $w_1 \in \hat{\mathcal{B}}^L \subset \mathcal{H}_2^+$ and $w_2 \in \hat{\mathcal{B}} \subset \mathcal{H}_2^+$ such that $\Pi_+ Wx = w_1 + w_2$. Hence:

$$\sup_{v \in \hat{\mathcal{B}}^L} \langle Wx, v \rangle = \sup_{v \in \hat{\mathcal{B}}^L} \langle \Pi_+ Wx, v \rangle = \sup_{v \in \hat{\mathcal{B}}^L} \langle w_1, v \rangle = \|w_1\|_1$$

We also have:

$$\|w_1\| = \inf_{w_2 \in \mathcal{B}} \|\Pi_+ Wx - w_2\|_1$$

It is natural to restrict attention to controllable systems. A controllable system $\mathcal{B}$ with complexity $c(\mathcal{B}) = (m, k)$ is then characterized by a matrix $\Psi \in \mathcal{RH}_\infty^+$ of dimension $q \times m$ with McMillan degree $k$ in the sense that:

$$\hat{\mathcal{B}} = \{ \Pi_+ \Psi s \mid s \in \mathcal{L}_2 \}$$

The complexity is measured by the number of columns and the McMillan degree of the rational matrix $\Psi$.

The misfit is then given by the following expression:

$$d(\mathcal{B}, W) = \sup_{x \in \mathcal{H}_2^-} \inf_{s \in \mathcal{L}_2} \|\Pi_+ Wx - \Pi_+ \Psi s\|_2$$

Optimizing the above expression over $s$ for fixed $x$ (which is equivalent to solving a linear quadratic control problem) we find that there exists a matrix $S \in \mathcal{RL}_\infty$ such that for fixed $x \in \mathcal{H}_2^-$ the infimum is attained for an optimal $s_x$ given by $s_x := Sx$. Minimizing the misfit over systems $\mathcal{B} \in \mathbb{B}_2$ with complexity $c(\mathcal{B}) = (m, k)$ is then equivalent to:

$$\inf_{\Psi} \inf_{S \in \mathcal{RL}_\infty} \|\Pi_+ W - \Pi_+ \Psi S\|_{\mathcal{H}_-}$$

where $\Psi$ varies over all rational matrices in $\mathcal{H}_2^+$ with $m$ rows and McMillan degree $k$. Although this general problem is again a Hankel norm approximation problem, due to the difficult structure in the approximating system, we were not able to find an explicit solution to this problem.

7 Conclusions

In this paper we developed a method for optimal approximate modeling of time series. Both the complexity and the misfit between model and data have been defined in a representation independent way and we showed that the misfit criterion can be expressed in terms of the Hankel
norm of an operator which is associated with the data and a co-inner kernel representation of the model.

We emphasize that the approximate modeling procedures discussed here are based on the assumption that the Laplace transform of the data is rational and belongs to $\mathcal{H}_2^+$. In the time domain, this includes polynomial-exponential data, impulse responses of linear time-invariant lumped systems and many other signals of practical interest. In particular, finite support time-series or any finite set of frequency responses can be treated in the presented framework. The assumption on rationality of $W$ can be replaced by the condition that $W$ is a Hilbert-Schmidt operator without affecting the main results of the paper.

The presented algorithms are conceptual of nature and are not intended as well suited numerical procedures for computing optimal approximate models. However, it should be emphasized that state space techniques can be used to completely implement the proposed method with purely algebraic manipulations on state space matrices.

We finally remark that the matrix $W_k$ defined in Theorem 5.1 has the interpretation of an approximate data set. The method proposed here has obvious implications for problems related to data reduction. More specifically, data sets consisting of finite samples of signal spectra can be condensed using the described methods.

References


REFERENCES

