Optimal Hankel norm identification of dynamical systems

Citation for published version (APA):

Document status and date:
Published: 01/01/1996

Document Version:
Publisher’s PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:
• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher’s website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license above, please follow below link for the End User Agreement:
www.tue.nl/taverne

Take down policy
If you believe that this document breaches copyright please contact us at:
openaccess@tue.nl
providing details and we will investigate your claim.

Download date: 04. Sep. 2020
Optimal Hankel Norm Identification
of Dynamical Systems

Siep Weiland and Anton A. Stoovogel

July, 1996

Measurement and Control Systems
Internal Report, 96 I/03

Reprint of paper submitted to Automatica, July 1996

Eindhoven, July 1996
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

Paper submitted to Automatica, July 1996

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 allows for 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. Sets of Pareto optimal models are defined as feasible tradeoffs between complexity and misfit of models and it is shown how classes of Pareto optimal models are obtained as exact models of compressed data sets obtained from Hankel norm approximations of data matrices. This leads to new conceptual algorithms for optimal approximate identification of time series.

Keywords

System identification, Approximate modeling, Hankel operators, Behavioral theory, Linear systems.

*Part of this research has been made possible by a grant from the European Community for the Systems Identification and Modeling Network (SIMONET).
†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, 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 time series of a data generating system with high precision whereas the complexity of a model is an indicator of the size (or dimension) of the family of independent time series which a model is able to explain. Intuitively, it seems logical to formalize a system identification framework in which simple and accurate models are preferred above complex and imprecise models. With well defined notions of complexity and misfit, the central issue in system identification is therefore the question of approximating data with accurate and low complexity 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. Contrary to the prevailing approaches in this area, we will neither assume an input-output structure on the observed data nor on the class of candidate models. The motivation for such an input-output independent approach stems from the practical consideration that in many modeling problems the input-output (or causality) structure of a set of selected variables may be unknown or unclear. In this paper we will adopt the behavioral framework [9, 10, 11] to discuss deterministic identification problems in which input and output variables are not displayed explicitly or, stated otherwise, play an entirely symmetric role. This symmetric treatment of system variables allows a symmetric way of treating noise characteristics. This means that in principle noise can be added to all system variables and not only to outputs of the system. However, we will not make specific assumptions on noise, but rather interpret noise as a specific part of the data which results from errors in the approximation of data by a model.

The contribution of this paper can be summarized as follows. A notion of misfit between an observed time series and a linear time-invariant model is introduced which is independent of representations of models. The misfit resembles a data weighted gap distance between the model and the behavior of a specific linear time-invariant model which is exact for the given data. As such, the minimization of the misfit function can be seen as a model approximation problem. It is shown that this misfit measure can be expressed as 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. Together with a suitable notion of model complexity, the misfit function defines for any observed time series a set of Pareto optimal models which distinguish the optimal achievable levels of complexity and misfit. Roughly speaking, a model is said to be Pareto optimal with respect to a given time series if all models of smaller complexity have a larger misfit (i.e., are less accurate) and all models of smaller misfit have larger complexity. In this paper we will characterize Pareto optimal models for a given data set and show that these models are obtained by performing optimal Hankel norm reductions. It is a key result of this paper that all optimal approximate models of given complexity and tolerated misfit correspond to the class of exact models of data sets obtained from Hankel norm approximations of the original data.

The paper is organized as follows. Some notation is introduced at the end of this section. Notions of complexity and misfit of dynamical systems are introduced in Section 2. In this section we further state the main problem of interest by formalizing Pareto optimal models. Model representations are discussed in Section 3. Characterizations of model complexity and misfit in terms of kernel representations are provided in this section. We proceed with the theoretically important problem of exact identification in Section 4. Main results on approximate system identification are presented in Section 5. Conclusions are formulated in Section 6.

The notation used in the paper is standard. For any integer \( q > 0 \) and subset of integers \( T \subseteq \mathbb{Z} \) we denote by \( \mathcal{L}_2(T, \mathbb{R}^q) \) 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 we write \( \mathcal{L}_2^+ \) for \( \mathcal{L}_2(\mathbb{Z}^+, \mathbb{R}^q) \) and \( \mathcal{L}_2^- \) for \( \mathcal{L}_2(\mathbb{Z}^-, \mathbb{R}^q) \). Here, \( \mathbb{Z}^+ = \{ t \in \mathbb{Z} \mid t \geq 0 \} \) and \( \mathbb{Z}^- \) is the complement of \( \mathbb{Z}^+ \).
of $\mathbb{Z}_+$ in $\mathbb{Z}$. Let $L_2$ denote the space of all complex valued functions which are square integrable on the unit circle $\mathbb{D} := \{z \in \mathbb{C} \mid |z| = 1\}$. Any such function $\hat{w}$ can be expanded as a Laurent series

$$\hat{w}(z) = \sum_{t=-\infty}^{\infty} w(t)z^{-t} \quad (1.1)$$

with $z \in \mathbb{C}$ and $w \in \ell_2$. The Hardy space $H^+_\infty$ consists of all square integrable functions on $\mathbb{D}$ with analytic continuation outside the unit circle (including $\infty$). The orthogonal complement of $H^+_\infty$ in $L_2$ will be denoted $H^-_\infty$. The canonical projections from $L_2$ to $H^+_\infty$ and $H^-_\infty$ will be denoted by $\Pi_+$ and $\Pi_-$, respectively. $H^+_\infty$ and $H^-_\infty$ 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., $RH^+_\infty$, $RH^-_\infty$, etc. It is assumed that $\ell_2$, $\ell^+_\infty$, $\ell^-_\infty$, $L_2$, $H^+_\infty$ and $H^-_\infty$ are equipped with their natural inner products which are all denoted by $\langle \cdot, \cdot \rangle$. The context will indicate which inner product is meant.

## 2 Problem formulation

Consider a finite set of real valued multivariable time-series

$$w_i : T \rightarrow \mathbb{R}^q, \quad i = 1, \ldots, N \quad (2.1)$$

where $N > 0$ is the number of observed time-series, $T \subseteq \mathbb{Z}$ is the time set and $\mathbb{R}^q$ with $q > 0$ is the space in which the observed variables take their values. This paper concerns the fundamental problem to find optimal models which explain the data (2.1). Here, by a model we will mean a linear time-invariant dynamical system; optimal will mean the best possible trade-off between the objectives of low complexity and high accuracy models; and explaining data will be understood in the sense that the observed time-series $w_i$, $i = 1, \ldots, N$, admits a decomposition

$$w_i = w'_i + w''_i \quad (2.2)$$

where $w'_i$ represents the part of $w$ which satisfies the constraints imposed by a linear time-invariant dynamical system and $w''_i$ represents the noise, the unexplained part of the observed time-series. We will make these concepts precise in the next few subsections.

Throughout, it will be assumed that the time set $T = \mathbb{Z}_+$ and that the observed time-series are square summable, i.e., $w_i \in \ell_2(\mathbb{Z}_+, \mathbb{R}^q)$ for $i = 1, \ldots, N$. In that case, the $z$-transform $\hat{w}_i$ of $w_i$ as defined in (1.1) exist for all $i = 1, \ldots, N$ and the data (2.1) can be represented in the frequency domain by defining

$$W(z) := [\hat{w}_1(z) \quad \hat{w}_2(z) \quad \cdots \quad \hat{w}_N(z)] \quad (2.3)$$

which is an element of the Hardy space $H^+_\infty$. It is clear that the $z$-transform defines a bijection between $W$ and the data matrix $[w_1 \quad \cdots \quad w_N]$.

### 2.1 Dynamical systems

Following the framework of Willems [9, 11] a model, or a (discrete time) dynamical system is a subset $B$ of the set of all functions $w : T \rightarrow S$ defined on a time set $T \subseteq \mathbb{Z}$ and taking values in a signal space $S$. That is, a dynamical system is a subset of time series $B \subseteq S^T$. Throughout this paper we consider discrete time systems defined for positive time, i.e., $T = \mathbb{Z}_+$ and with finite dimensional real valued signal spaces $S = \mathbb{R}^q$, with $q > 0$ denoting the number of components in the time series. Of special interest will be the class of $\ell_2$ systems which is defined as follows.
Definition 2.1 A (discrete time) dynamical system \( B \) is called an \( \ell_2 \)-system if \( B \) is a closed\(^1 \) subset of \( \ell_2^+ \).

A system \( B \) is called linear if \( B \) is a real linear subspace of \( W^T \), it is said to be time-invariant if \( \sigma B \subseteq B \). Here, \( \sigma \) denotes the usual shift \( \sigma [w](t) = w(t+1) \). A dynamical system is said to be \( \ell_2 \)-complete if the collection of all finite truncations of system trajectories define the system, that is if

\[
\{ w \in \ell_2^+ \mid w_{[0,t]} \in B_{[0,t]} \text{ for all } t \in \mathbb{Z}_+ \} \implies \{ w \in B \}.
\]

(See also [11]). We will focus on the model class of linear, time-invariant and \( \ell_2 \)-complete dynamical systems and denote this model class by \( \mathcal{B}_2 \), i.e.

\[
\mathcal{B}_2 := \{ B \subseteq \ell_2^+ \mid B \text{ is linear, time-invariant and } \ell_2 \text{-complete} \}.
\]

Since \( \ell_2 \)-completeness of subsets \( B \subseteq \ell_2^+ \) implies closedness of \( B \) in the standard \( \ell_2 \) topology, we have that \( \mathcal{B}_2 \) defines a model class of \( \ell_2 \) systems. The interest in this specific model class stems from the fact that elements in \( \mathcal{B}_2 \) admit simple representations in terms of rational operators (see Theorem 3.2 below) and, more importantly, due to the isomorphism between the Hilbert spaces \( \ell_2^+ \) and \( \mathcal{H}_2^+ \) elements in \( \mathcal{B}_2 \) can equivalently be considered in the frequency domain by defining, for each \( B \in \mathcal{B}_2 \),

\[
\hat{B} := \{ \hat{w} \in \mathcal{H}_2^+ \mid w \in B \}
\]

where \( \hat{w} \) is the z-transform defined in (1.1). Clearly, \( \hat{B} \) is a linear subspace of \( \mathcal{H}_2^+ \) which is shift-invariant in the sense that \( \hat{\sigma} \hat{B} \subseteq \hat{B} \) where \( [\hat{\sigma} \hat{w}](z) = z \hat{w}(z) - z \hat{w}(\infty) \).

Further, a system \( B \in \mathcal{B}_2 \) is said to be autonomous if there exists a \( t > 0 \) such that the mapping \( \pi_t : B \to B_{[0,t]} \), defined by the restriction \( \pi_t(w) := w_{[0,t]} \), is injective.

2.2 Model complexity

In order to formalize a feasible 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. Roughly speaking, the complexity of a model is a measure for the number of degrees of freedom to specify the trajectories which the model conceives. The more degrees of freedom a model has, the more trajectories this model may generate, the higher its complexity will be. We choose a measure of complexity which reflects the number of independent time series which a system can generate in a time window of given length.

Definition 2.2 The complexity \( c(B) \) of a dynamical system \( B \in \mathcal{B}_2 \) is the pair of integers \( c(B) = (m, n) \) which satisfy

\[
\dim(B_{[0,t-1]}) = mt + n \tag{2.4}
\]

for all \( t \geq n \).

The numbers \( m \) and \( n \) are well defined in this way (see e.g. [7] pp. 16 or [10] Theorem 25) and have simple system theoretic interpretations. \( m \) 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 \( B \). One can therefore think of \( m \) as the dimension of the input-space of the model \( B \). The number \( n \) corresponds to the degree of freedom due to initial conditions of \( B \). That is, \( n \) equals the state space dimension in any minimal state space representation of \( B \).

It seems reasonable to declare a model \( B_1 \) less complex than a model \( B_2 \) if it allows less trajectories per time window. That is, if

\[
\dim(B_1_{[0,t]}) < \dim(B_2_{[0,t]})
\]

\(^1\)Closed in the \( \ell_2 \) topology.
for all time $t \geq 0$. Due to the characterization (2.4) this motivates a lexicographic ordering on system complexities, i.e., if $c(B_1) = (m_1, n_1)$ and $c(B_2) = (m_2, n_2)$ then

$$c(B_1) \leq c(B_2) :\Leftrightarrow \begin{cases} m_1 < m_2 \\
1 = m_2 \text{ and } n_1 \leq n_2.
\end{cases}$$

In particular, this ordering leads to the property that

$$B_1 \subseteq B_2 \implies c(B_1) \leq c(B_2)$$

and implies that autonomous systems $(m = 0)$ are always less complex than non-autonomous systems.

2.3 The misfit

Consider the data matrix (2.3) and an element $B \in \mathbb{B}_2$. The orthogonal complement of $B$ in $\ell^+_2$ is denoted $B^\perp$ and is defined as

$$B^\perp := \{ v \in \ell^+_2 \mid \langle w, v \rangle = 0 \text{ for all } w \in B \}.$$ 

The orthogonal complement $B^\perp$ will be referred to as the set of laws of the model $B$. $\bar{B}^\perp$ is the set of $z$-transformed elements of $B^\perp$ or, equivalently, the orthogonal complement of $\bar{B}$ in $\mathcal{H}_2^\perp$. The discrepancy between model and data is defined as follows.

Definition 2.3 The misfit between a model $B \in \mathbb{B}_2$ and the data $W$ is defined as

$$d(B, W) := \sup \left\{ \frac{|Wx, v|}{\|x\| \|v\|} \mid v \in B^\perp, x \in \mathcal{H}_2^\perp \right\} \tag{2.5}$$

where the inner product is the usual inner product in $\mathcal{L}_2$ and the norms are the standard norms in the Hilbert spaces $\mathcal{H}_2^\perp$ and $\mathcal{H}_2^\perp$.

This misfit criterion has the following interpretation. A model $B \in \mathbb{B}_2$ is assumed to be a closed, linear and time-invariant subset of $\ell^+_2$. This means that linear combinations and left-shifted versions of any (finite) set of trajectories $\{w_i \in B \mid i = 1, \ldots, N\}$, also belong to $B$. If we aim to model the data (2.1) using linear, time-invariant models, it seems therefore reasonable to consider all linear combinations and left-shifts of the time-series (2.1). In particular,

$$B_0 := \text{span}\{\sigma^i w_i \mid i = 1, \ldots, N; t \in \mathbb{Z}_+\} \tag{2.6}$$

defines a linear, left-shift-invariant $\ell_2$ system which is exact for the data (2.1) in the sense that $w_i \in B$ for $i = 1, \ldots, N$. By viewing $W$ as a multiplicative operator mapping elements of $\mathcal{H}_2^\perp$ into elements of $\mathcal{L}_2$, the set (2.6) can be generated as the image of $W$. Indeed, if we write $W = \Pi_- W + \Pi_+ W$ then the image of $W$ is naturally decomposed into the orthogonal sets $\mathcal{W}_- := \Pi_- W \mathcal{H}_2^\perp$ and $\mathcal{W}_+ := \Pi_+ W \mathcal{H}_2^\perp$. The latter, $\mathcal{W}_+$, is clearly a shift-invariant subspace of $\mathcal{H}_2^\perp$ with the property that $w_i \in \mathcal{W}_+$ for $i = 1, \ldots, N$. Its closure $\overline{\mathcal{W}_+} = \mathcal{B}_0$. Since $B^\perp$ consists of all continuous linear functionals on $\ell_2$ which vanish at $B$, the values of $(w, v)$ with normalized elements $w \in B_0$ and $v \in B^\perp$ indicate to what extent the laws implied by the model $B$ fail to explain the exact model $B_0$ of the data. The misfit (2.5) is therefore the maximum discrepancy between laws $v \in B^\perp$ implied by accepting the model $B$ and data points in the exact linear time-invariant model $B_0$ explaining the data.

Remark 2.4 Note that $d(B, W) \geq 0$ and $d(B, W) = 0$ if and only if $w_i \in B$ for $i = 1, \ldots, N$. Further, the implication

$$B_1 \subseteq B_2 \Rightarrow d(B_1, W) \geq d(B_2, W)$$
is immediate from the definition and, using the notation above, \( d(B_0, W) = 0 \). Finally, it is important to observe that the misfit (2.5) is independent of representations of systems \( B \in \mathbb{B}_2 \). A characterization of the misfit in terms of system representations is one of the central results of the next section.

The misfit (2.5) can equivalently be viewed as a distance measure between \( B_0 \) and a model \( B \in \mathbb{B}_2 \). To see this, consider the \( \ell_2 \) system \( B_0 \) defined above and equip \( B_0 \) with the data weighted norm

\[
\| w \|_w := \inf \left\{ \| x \| \mid x \in \mathcal{H}_2^w, \ w = \Pi_w W x \right\}.
\]

The following proposition is then immediate from Definition 2.3.

**Proposition 2.5** If \( B \in \mathbb{B}_2 \) and \( W \in \mathcal{H}_2^+ \) then

\[
d(B, W) = \sup \left\{ \frac{\langle w, v \rangle}{\| w \|_w \| v \|} \mid v \in B^\perp, \ w \in B_0 \right\} \tag{2.7}
\]

**Remark 2.6** It is interesting to point out the relation between the expression (2.7) and the gap between the two closed subspaces \( B \in \mathbb{B}_2 \) and \( B_0 \in \mathcal{H}_2^+ \). The gap [4, 5] is defined as

\[
g(B, B_0) := \sup \left\{ \frac{\langle w, v \rangle}{\| w \|_w \| v \|} \mid v \in B^\perp, \ w \in B_0 \right\} \tag{2.8}
\]

By Proposition 2.5, the misfit \( d \) therefore incorporates a weighting along the principal axes of the image of \( \Pi_w W \) under \( \mathcal{H}_2^+ \) and can be interpreted as a *data-weighted gap*. We would like to emphasize the importance of this data-weighting for applications of the misfit (2.5) in system identification. Indeed, if we consider the example of an observed scalar valued time series \( w(t) = \lambda_i^1 + \epsilon \lambda_i^2 \), with \( t \in \mathbb{Z}_+, |\lambda_i| < 1, i = 1, 2 \) and \( \epsilon > 0 \), then \( w \in \mathcal{H}_2^+ \) and the linear span of the time series \( \lambda_i^1 \) and \( \lambda_i^2 \) defines the exact model \( B_0 \) in (2.6). It has complexity \( c(B_0) = (0, 2) \). A lower order approximate model obtained by minimizing the gap (2.8) over all \( B \in \mathbb{B}_2 \) with complexity \( c(B) = (0, 1) \) would therefore result in a model which does not depend on the value of \( \epsilon \). That is, for small \( \epsilon > 0 \), the relative weight of the component \( \lambda_i^1 \) of the data sequence \( w \) is not discriminated by the gap criterion. However, in the misfit-criterion (2.5) it is.

### 2.4 Optimal models

The mainstream approach to the approximate modeling problem amounts to minimizing a misfit criterion subject to a complexity constraint on the class of admissible models. An equally interesting paradigm amounts to finding models with minimal complexity subject to a maximal tolerated level of misfit. Both approaches involve a trade-off between the desirable objectives of low complexity and small misfit. In the context of the notions of complexity and misfit introduced in this section these objectives can be combined by considering a notion of Pareto optimality.

**Definition 2.7** A model \( B^w \in \mathbb{B}_2 \) is called Pareto optimal for the data \( W \) if for all \( B \in \mathbb{B}_2 \) the implications

\[
c(B) \leq c(B^w) \implies d(B, W) \geq d(B^w, W)
\]

\[
d(B, W) \leq d(B^w, W) \implies c(B) \geq c(B^w)
\]

hold simultaneously.

Hence, a model is Pareto optimal with respect to the data (2.1) if all models of smaller complexity have larger misfit, and all models of smaller misfit have larger complexity. The main result of this paper is a characterization of Pareto optimal models for a data set \( W \). We will show that Pareto optimal models can be
obtained as exact models, i.e. models with zero misfit, of a data set \( W' \) which is derived from \( W \). The main implication of this result is that a family of optimal approximate models can be easily generated as the class of exact models of a related data set.

We remark that any Pareto optimal model \( B^{\text{opt}} \) with complexity \( c(B^{\text{opt}}) = (m, n) \) and misfit \( d(B^{\text{opt}}, W) = \sigma \) provides a solution to two related identification problems:

IP-1: Given data represented by \( W \in \mathcal{H}_2^+ \) together with a pair of integers \( (m, n) \), find models \( B \in \mathcal{B}_2 \) of complexity \( c(B) \leq (m, n) \) such that the misfit \( d(B, W) \) is minimized.

IP-2: Given data represented by \( W \in \mathcal{H}_2^+ \) together with a misfit level \( \sigma \geq 0 \), find models \( B \in \mathcal{B}_2 \) of misfit \( d(B, W) \leq \sigma \) such that the complexity \( c(B) \) is minimized.

3 Rational representations of \( \ell_2 \) systems

The above identification problems were formalized independent of system representations. This has an obvious advantage, as optimal models will not depend on specific parameterizations of model sets. However, for computational purposes system representations will play a crucial role. In this section we first discuss representations of systems \( B \in \mathcal{B}_2 \). Specifically, the model set \( \mathcal{B}_2 \) is characterized as the set of systems which can be represented as the kernel of a rational operator. We then show that complexity and misfit admit simple characterizations in terms of kernel representations.

3.1 Kernel representations

Let \( \Theta \in \mathcal{R}\mathcal{H}_\infty \). Corresponding to \( \Theta \) we define a continuous linear operator \( \Theta : \mathcal{H}_2^+ \to \ell_2 \) defined by the multiplication

\[
(\Theta w)(z) = \Theta(z)w(z)
\]

where \( z \in \mathbb{C} \). We associate with each such \( \Theta \) a model \( \hat{B} \) which consists of those \( w \in \mathcal{H}_2^+ \) for which \( \Theta w \) belongs to \( \mathcal{H}_2^+ \). Formally,

\[
\hat{B}_\omega(\Theta) := \{ w \in \mathcal{H}_2^+ \mid \Theta w \in \mathcal{H}_2^+ \} = \{ w \in \mathcal{H}_2^+ \mid \Pi_+ \Theta w = 0 \} = \ker \Pi_+ \Theta
\]

(3.1)

Obviously, \( \hat{B}_\omega(\Theta) \) is a linear shift-invariant and closed subset of \( \mathcal{H}_2^+ \). Its inverse \( z \)-transform \( B_\omega(\Theta) \) belongs to \( \mathcal{B}_2 \) and coincides with the kernel of a convolution operator. Specifically, if

\[
\Theta(z) = \sum_{k=-\infty}^{\infty} \Theta_k z^{-k}
\]

is the Laurent series expansion of \( \Theta \) with \( z \in \mathbb{C} \) and \( \Theta_k \in \mathbb{R}^{x \times y} \) then, as \( \Theta \in \mathcal{R}\mathcal{H}_\infty \), \( \Theta_k = 0 \) for \( k > 0 \) and the map \( \Theta(\sigma) : \ell_2^+ \to \ell_2^+ \) defined by the convolution

\[
[\Theta(\sigma)w](t) := \sum_{k=0}^{\infty} \Theta_{t-k} w(k).
\]

is well defined and

\[
B_\omega(\Theta) := \ker \Theta(\sigma)
\]

(3.2)

is a time domain representation of (3.1). A subset \( B \subseteq \ell_2^+ \) [or \( \hat{B} \subseteq \mathcal{H}_2^+ \)] is said to have a kernel representation if there exists \( \Theta \in \mathcal{R}\mathcal{H}_\infty \) such that \( B = B_\omega(\Theta) \) [or \( \hat{B} = \hat{B}_\omega(\Theta) \)].
Remark 3.1 We emphasize that in the present setting the values of signals \( w \in \mathbb{B}_w(\Theta) \) are not constrained to be zero at time instances \( t \in \mathbb{Z}_- \). Elements of \( \ell_2^+ \) should therefore not be thought of as signals \( w \in \ell_2 \) which vanish for \( t \in \mathbb{Z}_- \). It is a feature of the present setting that autonomous systems, systems with transient phenomena, non-controllable systems and systems with non-zero initial conditions are naturally considered.

The following result is of crucial importance as it precisely characterizes the model class \( \mathbb{B}_2 \) as those systems which admit rational kernel representations. For a proof we refer to [14] or [15].

**Theorem 3.2** The following statements are equivalent

1. \( B \in \mathbb{B}_2 \)
2. \( B \) admits a rational kernel representation, i.e. there exists \( \Theta \in \mathcal{RH}_\infty^- \) such that \( B = B_w(\Theta) \).

A characterization of subsystems and system equivalence is given in the following result.

**Theorem 3.3** For \( i = 1, 2 \), let \( \Theta_i \in \mathcal{RH}_\infty^- \) be a full rank kernel representation of \( B_i = B_w(\Theta_i) \). Then

1. \( B_1 \subseteq B_2 \) if and only if there exists \( U \in \mathcal{RH}_\infty^- \) such that \( \Theta_2 = U \Theta_1 \).
2. \( B_1 = B_2 \) if and only if there exists a unit \( U \in \mathcal{RH}_\infty^- \) such that \( \Theta_2 = U \Theta_1 \).

**Proof.** See [14] or [15]. \( \square \)

An operator \( \Theta \in \mathcal{RH}_\infty^- \) is said to be co-inner if it is norm preserving on the orthogonal complement of its kernel or, equivalently, if \( \Theta \Theta^* = I \) where \( \Theta^*(z) := \Theta^T(z^{-1}) \).

**Corollary 3.4** Every \( B \in \mathbb{B}_2 \) admits a kernel representation \( B_w(\Theta) \) where \( \Theta \in \mathcal{RH}_\infty^- \) is co-inner.

**Proof.** By theorem 3.2 every \( B \in \mathbb{B}_2 \) can be represented as \( B = B_w(\Theta) \) where \( \Theta \in \mathcal{RH}_\infty^- \) has full rank. It is well known [8] that any such \( \Theta \) can be factorized as \( \Theta = \Theta_s \Theta_i \) where \( \Theta_s \in \mathcal{RH}_\infty^- \) is a square matrix with \( \text{rank}(\Theta_s(\lambda)) \) constant for all \( |\lambda| \geq 1 \) and \( \Theta_i \) is co-inner. The result then follows from Theorem 3.3 by observing that \( U := \Theta_s^{-1} \) is a unit in \( \mathcal{RH}_\infty^- \). \( \square \)

### 3.2 Characterization of system complexity

There exists a simple relation between the complexity of a system and the complexity of its representation. Let the complexity of a kernel representation \( \Theta \in \mathcal{RH}_\infty^- \) of \( B \in \mathbb{B}_2 \) be its McMillan degree \( \text{deg}(\Theta) \).

**Theorem 3.5** Let \( \Theta \in \mathcal{RH}_\infty^- \) be a co-inner kernel representation of \( B \in \mathbb{B}_2 \). Then the following statements are equivalent

1. \( c(B) = (m, n) \).
2. \( m = q - \text{rank}(\Theta) \) and \( n = \text{deg}(\Theta) \).

The proof of Theorem 3.5 is rather involved and beyond the scope of this paper. We refer to [14] for details or [10] (Section 22) for similar results in terms of polynomial representations.

---

2 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 \( \mathcal{RH}_\infty^- \).
Remark 3.6 Recall that the number $q$ in Theorem 3.5 is the dimension of the signal space of interest (as defined in (2.1)). In view of the interpretation of $m$, the number $p = \text{rank}(\Theta)$ corresponds to the number of outputs of the system $\mathcal{B}_m(\Theta)$. Theorem 3.5 shows that complex systems are represented by (co-inner) kernel representations $\Theta$ of low rank and large McMillan degree. Complex systems therefore satisfy a small number of independent equations, while each equation has a relatively large McMillan degree.

Remark 3.7 It is an interesting consequence of Theorem 3.5 that all co-inner kernel representations of systems $\mathcal{B} \in \mathcal{B}_2$ have equal McMillan degree which is moreover a minimal number in the class of all kernel representations of $\mathcal{B}$.

3.3 Characterization of the misfit

The following result is crucial in the sequel and relates the misfit (2.5) to the Hankel norm of a specific operator. Let $G$ be a matrix whose entries are analytic on the unit circle $\mathbb{D}$, i.e., $G \in \mathcal{L}_\infty$. Corresponding to $G$ one can define the operator norms

$$
\begin{align*}
\| G \|_{H_+} &:= \sup_{x \in \mathcal{H}_2^+} \frac{\| \Pi_+ Gx \|}{\| x \|}, \\
\| G \|_{H_-} &:= \sup_{x \in \mathcal{H}_2^-} \frac{\| \Pi_- Gx \|}{\| x \|}.
\end{align*}
$$

Then $\| G \|_{H_+}$ is the induced norm of the Hankel operator $\Gamma_G^+ := \Pi_+ G \Pi_-$ viewed as a multiplicative operator from $\mathcal{L}_2$ to $\mathcal{H}_2^+$. Similarly, $\| G \|_{H_-}$ is the induced norm of the Hankel operator $\Gamma_G^- := \Pi_- G \Pi_+$ mapping $\mathcal{L}_2$ to its range in $\mathcal{H}_2^\bot$. 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 (2.5) can be computed as the Hankel norm of an operator.

**Theorem 3.8** Let $\Theta \in \mathcal{RH}_\infty$ be a co-inner kernel representation of $\mathcal{B} \in \mathcal{B}_2$. Then

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

where $\Theta^-(z) := \Theta^\top(z^{-1})$.

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

$$
\| \Theta W \|_{H_+} = \| W^- \Theta^- \|_{H_-} = \| W^- \|_{\text{im } \Theta^-} \|_{H_-} = \| W^- \|_{\tilde{\mathcal{B}}},
$$

where we used that $\Theta^-$ is norm preserving and that

$$
\tilde{\mathcal{B}}(\Theta) = \{ \hat{w} \in \mathcal{H}_2^+ \mid \langle \Theta \hat{w}, \hat{v} \rangle = 0 \text{ for all } \hat{v} \in \mathcal{H}_2^+ \}
$$

$$
= \{ \hat{w} \in \mathcal{H}_2^+ \mid \langle \hat{w}, \Theta^\bot \hat{v} \rangle = 0, \text{ for all } \hat{v} \in \mathcal{H}_2^\bot \}
$$

$$
= (\text{im } \Theta^-)^\bot,
$$

together with rationality of $\Theta$ (this yields that $\text{im } \Theta^-$ is closed) implies that $\text{im } \Theta^- = \tilde{\mathcal{B}}$. Deleting the hats, we can write

$$
\| W^- \|_{\tilde{\mathcal{B}}} \|_{H_-} \leq \sup_{v \in \mathcal{B}^+} \frac{\| \Pi_- W^- v \|}{\| v \|} = \sup_{v \in \mathcal{B}^+} \frac{|\langle \Pi_- W^- v, W^- v \rangle|}{\| v \| \| \Pi_- W^- v \|}
$$

$$
\leq \sup_{v \in \mathcal{B}^+, \ x \in \mathcal{H}_2^-} \frac{|\langle x, W^- v \rangle|}{\| v \| \| x \|} = \sup_{v \in \mathcal{B}^+, \ x \in \mathcal{H}_2^-} \frac{|\langle W x, v \rangle|}{\| v \| \| x \|}
$$

$$
= d(\mathcal{B}, W).
$$
The reversed inequality follows by the Schwartz inequality

\[ \|W^\sim\|_{\mathcal{B}^+} = \sup_{v \in \mathcal{B}^+} \frac{\|\Pi_+ W^\sim v\|}{\|v\|} = \sup_{v \in \mathcal{B}^+, x \in \mathcal{N}_2^+} \frac{\|\Pi_+ W^\sim v\|}{\|v\|} \geq \sup_{v \in \mathcal{B}^+, x \in \mathcal{N}_2^+} \frac{|\langle x, W^\sim v \rangle|}{\|v\| \|x\|} = \sup_{v \in \mathcal{B}^+, x \in \mathcal{N}_2^+} \frac{|\langle Wx, v \rangle|}{\|v\| \|x\|} =: d(\mathcal{B}, W). \]

Thus,

\[ \|\Pi_+ \Theta W\|_{\mathcal{H}_2} = \|\Pi_+ W^\sim\|_{\mathcal{B}^+} = \|\Pi_+ W^\sim \Theta^\sim\|_{\mathcal{H}_2} = d(\mathcal{B}, W) \]

as desired. \hfill \Box

Remark 3.9 In common system theoretic language, the Hankel norm is interpreted as the induced norm from \( \ell_2 \)-past inputs to \( \ell_2 \)-future outputs of a stable transfer function. The characterization of the misfit criterion in Theorem 3.8 and more particularly the expression \( d(\mathcal{B}, W) = \|W^\sim\|_{\mathcal{B}^+} \) in the proof shows that the misfit can be interpreted as the induced norm from the \( \ell_2 \) laws of the model \( B \) to the \( \ell_2 \) modeling errors \( e := n^\sim W^\sim v \) with \( v \in \mathcal{B}^+ \).

Remark 3.10 The rationality of \( \Theta \) in Theorem 3.8 is not essential. Since for all \( \Theta \in \mathcal{H}_{\infty} \)

\[ \|W^\sim|_{\text{im } \Theta^\sim}\|_{\mathcal{H}_2} = \|W^\sim|_{\text{im } \Theta^\sim}\|_{\mathcal{H}_2} \]

the argument in the proof of Theorem 3.8 remains valid for the non-rational case.

The relationship between Hankel norms and Hankel singular values of rational operators is well known [6] and yields the following corollary as an immediate consequence of Theorem 3.8.

Corollary 3.11 If \( B \in \mathbb{B}_2 \) and \( W \in \mathcal{RH}_2^+ \) then

\[ 0 \leq d(\mathcal{B}, W) \leq \sigma_{\text{max}}(W) \]

where \( \sigma_{\text{max}} \) is the maximal Hankel singular value of \( W \).

Proof. Observe that \( 0 \leq d(\mathcal{B}, W) \leq \|W^\sim|_{\text{im } \Theta^\sim}\|_{\mathcal{H}_2} \leq \|W^\sim\|_{\mathcal{H}_2} = \|W\|_{\mathcal{H}_2} = \sigma_{\text{max}}(W). \) \hfill \Box

4 Exact models

In this section we consider the problem of identifying all zero-misfit models for the data (2.1). This exact modeling problem amounts to parameterizing the set of all models \( B \in \mathbb{B}_2 \) which explain the data (2.1) in the sense that \( w_i \in B \) for \( i = 1, \ldots, N \). The importance of this problem will be apparent in the next section where we reduce the approximate modeling problem to a special exact modeling problem.

Let

\[ \mathcal{M} := \{ \Theta \in \mathcal{RH}_\infty^- \mid w_i \in B_{\text{in}}(\Theta) \text{ for } i = 1, \ldots, N \}. \]

denote the family of all kernel representations of models which are compatible with (2.1). Assume that \( W \) is rational, i.e.,
Assumption 4.1 \( W \in \mathcal{RH}_2^+ \).

Typical examples of data sets satisfying assumption 4.1 include finite length time series \( w(0), \ldots, w(T) \) which are either extended with zeros or identified with an exponentially weighted periodic signal of period \( T \). Also, finite sets of frequency response measurements and polynomial exponential time-series satisfy Assumption 4.1. A study of data sets of this type appeared in [2]. We refer to [12] for a methodology to approximate data sets by polynomial-exponential time series which satisfy this assumption using risk minimization techniques.

Recall that an ordered pair \((\Theta, \Psi)\) where \( \Theta, \Psi \in \mathcal{RH}_\infty^- \) is called a left-coprime factorization of \( W \in \mathcal{H}_\infty^+ \) if \( W = \Theta^{-1} \Psi \), \( \det \Theta \neq 0 \) and every common left divisor of \( \Theta \) and \( \Psi \) is a unit in \( \mathcal{RH}_\infty^- \). The set \( M \) is characterized in terms of left coprime factorizations of \( W \) as follows. (See, e.g., [1]).

Theorem 4.2 Suppose that Assumption 4.1 holds. Then there exists a left-coprime factorization \((\Theta_0, \Psi_0)\) of \( W \). Moreover, for all left-coprime factorizations \((\Theta_0, \Psi_0)\) of \( W \) there holds

1. \( \Theta_0 \in M \).
2. \( B_m(\Theta_0) \) is an autonomous system.
3. \( B_0 = B_m(\Theta_0) \) where \( B_0 \) is defined in (2.6).
4. \( M = \{ \Theta \mid \Theta = \Lambda \Theta_0 \text{ with } \Lambda \in \mathcal{RH}_\infty^- \} \)

In particular, the sets \( B_m(\Theta_0) \) and \( \{ \Lambda \Theta_0 \mid \Lambda \in \mathcal{RH}_\infty^- \} \) are independent of the left-coprime factorization \((\Theta_0, \Psi_0)\) of \( W \).

Proof. Existence of left-coprime factorizations of \( W \) has been shown in [8]. Let \((\Theta_0, \Psi_0)\) be any left coprime factorization of \( W \). Since \( \Psi_0 \in \mathcal{RH}_2^+ \) it follows that \( \Pi_+ \Theta_0 W = \Pi_+ \Psi_0 = 0 \) which yields that \( \dot{w}_i = W e_i \in \dot{B}_m(\Theta_0) \) for all \( i = 1, \ldots, N \). Hence \( \Theta_0 \in M \) which proves part 1. If \( B = B_m(\Theta_0) \) and \( \Theta_0 \) is co-inner then Theorem 3.5 and the fact that \( \Theta_0 \) is square and full rank yields that \( B \) has complexity \((m, n) \) with \( m = q - \text{rank}(\Theta_0) = 0 \). Using Theorem 3.3 it is straightforward to see that \( m = 0 \) also if \( \Theta_0 \) is not co-inner. Conclude from this that \( B_m(\Theta_0) \) is autonomous which proves part 2. Using part 1, the inclusion \( B_0 \supseteq B_m(\Theta_0) \) is trivial in the proof of part 3. To prove the converse, recall that

\[ \dot{B}_0 = \Pi_+ W \mathcal{H}_2^+ \]

Since \( W \) is rational the set \( \Pi_+ W \mathcal{H}_2^+ \) is finite dimensional and therefore closed. It therefore suffices to show that \( \Pi_+ W \mathcal{H}_2^+ \subseteq \ker(\Pi_+ \Theta_0) \). To see this, let \( x \in \mathcal{H}_2^- \), and observe that

\[ \Pi_+ \Theta_0 \Pi_+ W x = \Pi_+ \Theta_0[\Theta_0^{-1} \Psi_0 x - \Pi_- \Theta_0^{-1} \Psi_0 x] = \Pi_+ \Psi_0 x - \Pi_+ \Theta_0 y = 0 \]

where we used that both \( \Psi_0 x \) and \( y := \Pi_- W x \) belong to \( \mathcal{H}_2^- \). Since \( x \) is arbitrary, this completes the proof of part 3. Part 4 is an immediate consequence of the first statement of Theorem 3.3, part 3 and the fact that \( \Theta \in M \) if and only if \( B_0 \subseteq B_m(\Theta) \).

Theorem 4.2 has the implication that all exact models for the data (2.1) can be generated from an arbitrary left-coprime factorization of \( W \). It follows from Theorem 3.3 that the set \( B_0 \) is included in any exact model of the data (2.1) satisfying the Assumption 4.1. Like in [1, 10], \( B_0 \) will therefore be referred to as the most powerful unfalsified model associated with the data. Since \( B_0 \) is uniquely defined by the data we write \( B_0 = \text{mpum}(W) \).
### 5 Optimal approximate models

The first main result of this section characterizes the set of all autonomous systems of given complexity and prescribed misfit level.

**Theorem 5.1** Let $B \in \mathbb{B}_2$ and $\sigma > 0$ be given and assume $W \in \mathcal{RH}^+_2$. Let $\sigma_1 \geq \ldots \geq \sigma_n > 0$ denote the ordered Hankel singular values of $W$. Then the following statements are equivalent:

1. $B$ has complexity $c(B) \leq (0, k)$ and misfit $d(B, W) \leq \sigma$.

2. $\sigma \geq \sigma_{k+1}$ and there exists $W_k \in \mathcal{RH}^+_2$ such that $\| W - W_k \|_{H_\sigma} \leq \sigma$, $\deg(W_k) \leq k$ and $B = \mpum(W_k)$.

**Proof.** (2 $\Rightarrow$ 1). If $B = \mpum(W_k)$ then by Theorem 4.2, $B$ is autonomous, i.e. $m(B) = 0$. Further

$$m(B) = \dim(B) = \dim \Pi_+ W_k \mathcal{H}_2^+ = \deg(W_k) \leq k$$

which yields that $c(B) \leq (0, k)$. Next, let $\Theta$ be a co-inner kernel representation of $B$. Using Theorem 3.8, we obtain

$$d(B, W) = \| W^\sim \Theta^\sim \|_{H_\sigma} = \| (W^\sim - W_k^\sim) \Theta^\sim \|_{H_\sigma} = \| (W^\sim - W_k^\sim) \|_{\lim \Theta^-} \|_{H_\sigma}$$

$$\leq \| W^\sim - W_k^\sim \|_{H_\sigma} = \| W - W_k \|_{H_\sigma} \leq \sigma$$

which proves the implication.

(1 $\Rightarrow$ 2). Suppose that $c(B) \leq (0, k)$ and $d(B, W) \leq \sigma$. By Corollary 3.4 there exists a co-inner kernel representation $\Theta$ of $B$. Define $\Pi_B$ as the orthogonal projection onto $B$. Clearly, $\Pi_B$ has rank $k$ so that, using Theorem 3.8, we obtain

$$\sigma \geq d(B, W) = \| W^\sim \Theta^\sim \|_{H_\sigma} = \| W^\sim \|_{\lim \Theta^-} \|_{H_\sigma} = \| W^\sim (I - \Pi_B) \|_{H_\sigma} = \| W^\sim - W^\sim \Pi_B \|_{H_\sigma}$$

$$\geq \sigma_{k+1}$$

since $I - \Pi_B$ is an orthogonal projection onto $\im \Theta^-$. The last inequality follows from the fact that, since $\Pi_B$ has rank $k$, $\Pi^\sim - \Pi^\sim \Pi_B$ is, as a rank $k$ approximant of $\Pi^\sim W^\sim$, never closer (in Hankel norm) than the optimal Hankel norm approximant. Hence, necessarily $\sigma \geq \sigma_{k+1}$. Next, from Theorem 6.1 in [6] (or Theorem 20.1.1 in [3]) we infer the existence of $S \in \mathcal{RH}^+_\infty$ such that

$$\sigma \geq d(B, W) = \| W^\sim \Theta^\sim \|_{H_\sigma} = \| W^\sim \Theta^\sim - S \|_{\infty} = \| W^\sim \Theta^\sim - S \Theta \Theta^- \|_{\infty}$$

$$= \| W^\sim \Theta^- - S_+ \Theta^- - S_- \Theta^- \|_{\infty} = \| W^\sim - S_+ - S_- \|_{\infty}$$

$$\geq \| (W^\sim - S_-) \|_{H_\sigma} = \| W - S_- \|_{H_\sigma}$$

(5.1)

Here, $\| \cdot \|_{\infty}$ denotes the $L_\infty$ norm, and $S \Theta$ is decomposed as $S \Theta = S_+ + S_-$ with $S_+ \in \mathcal{RH}^+_\infty$ and $S_- \in \mathcal{RH}^-_\infty$. In (5.1) we used that $\Theta$ is inner, and that the negative Hankel norm of an operator is not larger than the infinity norm of the difference between that operator and any other operator in $\mathcal{H}^+_\infty$ (13), pp. 435). We claim that $W_k := S_-^\sim$ satisfies the conditions mentioned in statement 2. Indeed, $W_k \in \mathcal{RH}^+_2$ and by (5.1)

$$\| W - W_k \|_{H_\sigma} \leq \sigma.\text{ Moreover, because } S^- \in \mathcal{H}_\infty \text{ and } \Theta S^- \in \mathcal{H}_\infty, \text{ we find for all } v \in \mathcal{H}_2^+:$$

$$\Pi_+ \Theta W_k v = \Pi_+ \Theta W_k v = \Pi_+ \Theta (\Theta^- S^- - S_-^\sim) v = \Pi_+ S^- v - \Pi_+ \Theta S_- v = 0.$$
Theorem 5.1 provides necessary and sufficient conditions for the existence of bounded complexity and bounded misfit models for the data $W$. In words, Theorem 5.1 claims that the set of all autonomous systems of order less than or equal to $k$ with misfit less than or equal to $\sigma$ is parameterized by the most powerful unfalsified models of the Hankel norm approximants $W_k$ of $W$ of McMillan degree at most $k$ which satisfy the upperbound $\|W - W_k\|_{H_s} \leq \sigma$. Hence, the class of approximate autonomous models of the data set $W$ is characterized as the class of exact models of reduced data sets $W_k$ that satisfy the conditions mentioned in 2.

It is well-known [3, 6] that whenever $\sigma$ is not equal to one of the Hankel singular values of $W$, the set of all $W_k \in \mathcal{R}H^+_2$ of McMillan degree less than or equal to $k$ which satisfy $\|W - W_k\|_{H_s} \leq \sigma$ can in fact be parameterized by means of a linear fractional transformation of an operator depending on $W$ and a free parameter, $Q \in \mathcal{R}H^+_\infty$ with $\|Q\|_\infty \leq 1$. This parameterization can be used to actually generate the approximate models with complexity less than or equal to $(0, k)$ and misfit less than or equal to $\sigma$.

Theorem 5.2 Let $W \in \mathcal{R}H^+_2$ be given and let $B \in \mathbb{B}_2$ have complexity $c(B) = (0, k)$ and misfit $d(B, W) = \sigma$. Assume moreover that the Hankel singular value $\sigma_k(W) > \sigma_{k+1}(W)$. Then the following statements are equivalent:

1. $B$ is Pareto optimal for the data $W$,
2. $\sigma = \sigma_{k+1}$,
3. there exists an optimal Hankel norm approximant $W_k$ of $W$ of McMillan degree $k$ such that $B = \text{mpum}(W_k)$.

**Proof.** (1 $\Rightarrow$ 2). Suppose that $B$ satisfies the hypothesis. By Theorem 5.1, $\sigma \geq \sigma_{k+1}$. It therefore suffices to show that $B$ is not Pareto optimal if $\sigma > \sigma_{k+1}$. This follows easily by letting $W_k \in \mathcal{R}H^+_2$ be an optimal Hankel norm approximant of $W$ of McMillan degree $k$ and setting $B' := \text{mpum}(W_k)$. Then

$$c(B') = (0, k) = c(B)$$

and

$$d(B', W) = \|W^* - B'^*\|_{H_s} \leq \|W^* - W_k^*\|_{H_s} = \|W - W_k\|_{H_s} = \sigma_{k+1} < \sigma = d(B, W).$$

Hence, $B$ is not Pareto optimal.

(2 $\Rightarrow$ 3). Let $\sigma = \sigma_{k+1}$. Using Theorem 5.1 we find that there exists $W_k \in \mathcal{R}H^+_2$ with $\text{deg}(W_k) \leq k$ and $\|W - W_k\|_{H_s} \leq \sigma_{k+1}$ such that $B = \text{mpum}(W_k)$. From the general theory of Hankel norm approximants [6] it is well known that any rank less than or equal to $k$ approximant $W_k$ of $W$ satisfies $\|W - W_k\|_{H_s} \geq \sigma_{k+1}$. Hence $\|W - W_k\|_{H_s} = \sigma_{k+1}$ so that $W_k$ is an optimal Hankel norm approximant of $W$.

(3 $\Rightarrow$ 1). Let $W_k$ be an optimal Hankel norm approximant of $W$ of degree less than or equal to $k$, and let $B = \text{mpum}(W_k)$. Firstly, this implies that if $B' \in \mathbb{B}_2$, $c(B') \leq c(B)$ then $m(B') = 0$, $n(B') \leq k$ and

$$d(B', W) = \|W^* - (I - \Pi_{B'})\|_{H_s} = \|W^* - W^*\Pi_{B'}\|_{H_s} \geq \sigma_{k+1} = d(B, W).$$

Secondly, if $B' \in \mathbb{B}_2$, $c(B') < c(B)$ then $c(B') = (0, k')$ with $k' < k$. By the same argument it follows that

$$d(B', W) \geq \sigma_{k+1} \geq \sigma_{k+1} = d(B, W).$$
which proves the implication
\[ d(B', W) \leq d(B, W) \Rightarrow c(B') \geq c(B). \]
We conclude that \( B \) is Pareto optimal by definition 2.7.

The following conceptual and constructive procedure can be used to obtain models of bounded complexity and bounded misfit. This algorithm is suggested in [6].

**Algorithm 1**: Given the data (2.1) and the numbers \( k \geq 0 \) and \( \sigma > 0 \).

1. Define the rational function \( W \in \mathcal{RH}_2^+ \) according to (2.3).
2. Compute the ordered Hankel singular values \( \sigma_1 \geq \ldots \geq \sigma_n > 0 \) of \( W \) and verify whether \( \sigma > \sigma_{k+1} \).
3. If \( \sigma_k > \sigma > \sigma_{k+1} \) define the augmented operator
   \[ W(z) := \begin{pmatrix} W(z) & 0 \\ 0 & w(z) \end{pmatrix} \]
   where \( w(z) = \sigma_{k+1} \), i.e., \( \| w(z) \|_{H_\infty} = \sigma \). Otherwise take \( w(z) = 0 \).
4. Let \( W_k^{(\sigma)} \) be an optimal Hankel norm approximant of McMillan degree \( k \) to \( W_k \) (see [6]) and let \( W_k \) be the truncation of \( W_k^{(\sigma)} \) to the augmented operator. \( W_k \) will then satisfy \( \| W - W_k \|_{H_\infty} \leq \sigma \).
5. Let \( W_k = \Theta^{-1}_k W_k \) be a left coprime factorization over \( \mathcal{RH}_\infty^+ \) of \( W_k \).
6. Put \( B = B_{\sigma_k}(\Theta_k) \).

This procedure results in models \( B \in \mathbb{B}_2 \) of complexity \( c(B) = (0, k) \) and misfit \( d(B, W) \leq \sigma \). As shown in Theorem 5.2 such a model is Pareto optimal if and only if \( \sigma = \sigma_{k+1} \).

**Remark 5.3** We emphasize the conceptual nature of Algorithm 1. It requires the explicit construction of the high order rational matrix \( W \) as input for a Hankel norm approximation routine. A more direct construction of \( W_k \) on the basis of the original data is a topic of further investigation.

Theorem 5.1 characterizes a class of autonomous systems. This may seem restrictive at first sight but Theorem 3.3 together with the properties of the misfit function mentioned in Remark 2.4 provide a simple parameterization of a class of models with prescribed complexity and bounded misfit.

**Theorem 5.4** Let \( B \) have complexity \( c(B) = (0, k) \) and misfit \( d(B, W) \leq \sigma \). Then all models \( B' \in \mathbb{B}_2 \) with \( B \subseteq B' \) have misfit \( d(B, W) \leq \sigma \). In particular, if \( B = B_{\sigma_k}(\Theta) \), then for all \( \Lambda \in \mathcal{RH}_\infty^+ \) the system
\[ B':=B_{\sigma_k}(\Lambda \Theta) \]
has misfit \( d(B, W) \leq \sigma \).

**Proof.** Immediate from the Definition 2.3, Remark 2.4 and the characterization of subspace inclusions given in Theorem 3.3. 

**Remark 5.5** Theorem 5.4 therefore provides a class of models
\[ \mathbb{B}^\sigma = \{ B' = B_{\sigma_k}(\Lambda \Theta) \mid \Lambda \in \mathcal{RH}_\infty^+ \} \]
of guaranteed misfit level \( \sigma \) which is generated by the autonomous system \( B_{\sigma_k}(\Theta) \). By Theorem 3.5, individual elements \( B = B_{\sigma_k}(\Lambda \Theta) \in \mathbb{B}^\sigma \) have complexity \((m, n)\) with
\[ m = q - \text{rank}(\Lambda \Theta) \]
\[ n = \text{deg}(\Lambda \Theta) \]
Since \( m \) equals the dimension of the input space of \( B \) this class obviously includes non-autonomous systems. We finally remark that these models need not be Pareto optimal if \( m > 0 \).
6 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. The misfit resembles a data weighted gap distance between the model and the most powerful unfalsified model of the given data. It has been shown that the complexity can be expressed in terms of rank properties of normalized kernel representations. The misfit function has been characterized as the Hankel norm of an operator which is associated with the data and a co-inner kernel representation of the model.

These characterizations have been used to parameterize a class of optimal approximate models as exact models of related data sets. It has been shown that these models satisfy a Pareto optimality condition and that these models can be generated by performing optimal Hankel norm reductions on the data matrix.

The approximate modeling procedure discussed here is 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. We believe that the assumption on rationality of $W$ can be replaced by the weaker condition that $W$ is a Hilbert-Schmidt operator without affecting the main results of the paper.

The characterization of Pareto optimal approximate models as exact models of Hankel reduced data sets constitutes the main theoretical result of this paper. The presented algorithm for the computation of these models 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.

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. Generalizations to frequency weighted misfit functions can be easily incorporated in the presented framework but have not been discussed in the present paper.

References


