MINLIP for the Identification of Monotone Wiener Systems

Kristiaan Pelckmans

Division of Systems and Control, Department of Information Technology
Uppsala University, Box 337, SE-751 05, Uppsala, Sweden

Abstract

This paper studies the MINLIP estimator for the identification of Wiener systems consisting of a sequence of a linear FIR dynamical model, and a monotonically increasing (or decreasing) static function. Given $T$ observations, this algorithm boils down to solving a convex quadratic program with $O(T)$ variables and inequality constraints, implementing an inference technique which is based entirely on model complexity control, a technique which is dubbed as Einstein's razor. This technique is proven to yield here almost consistent estimates in case the given data is suitably rich (local Persistently Exciting of sufficient order), and if the samples obey a 'true' monotone Wiener system satisfying some regularity conditions. It is then indicated how to extend the method in order to cope with noisy data, and empirical evidence is given in order to support the claim of efficiency.

Key words: System Identification, Wiener Systems, Nonlinear Systems, Convex Optimization

1 INTRODUCTION

The identification of Wiener systems has been considered in many papers since the 1970s. Different existing approaches could roughly be divided in methods using (i) invertible nonlinearities (reducing to Hammerstein identification), (ii) correlation based approaches exploiting stochastic properties of the signals [1,2], (iii) approximate (recursive) PEM approaches providing a well-established framework for convergence analysis [4,5] and [6], (iv) subspace based approaches [7]. For a general overview see the survey [8] and the recent edited book [9] for many additional pointers and case-studies.

The MINLIP method takes a different angle towards the identification problem, as compared to the ideas implemented in most of those works: the identification problem is not formulated as an error minimization problem (as classically, see e.g. [10]), but rather intends to find the simplest model which does interpolate the samples (approximately). The approach as such implements the idea captured by the saying 'Make things as simple as possible, but no simpler', sometimes dubbed as Einstein's razor.

This approach is in spirit closely related to the Structural Risk Minimization approach as in [11] and the algorithm of the Support Vector Machine (SVM). Such different perspective to the modeling problem not only appeals to different theoretical results, but results as well in a computational efficient approach. Specifically, this paper shows how to phrase the identification problem as a convex Quadratic Program of size proportional to the number of samples. A similar method was used in [12] in the area of survival analysis.

While in general the literature on the identification of Wiener systems is considerable, often theoretical understanding of proposed techniques is restricted to exposition of an appropriate technical implementation. Notable exceptions are given in [5], [8] and [3]. The first work considers a Recursive Prediction Error Method (RPEM) of general Wiener models, and convergence properties are derived using the ODE framework as in [13]. This approach makes considerable assumptions on the stochastic mechanisms underlying the signals used for (recursive) identification. The smoothing approach described in [8] exploits as well a stringent stochastic assumption of the involved signals, and asserts basically that the Wiener system can be identified by di-

---

* The author acknowledges financial support of European Research Council under Seventh Framework Program and within Advanced Grant no. 247035 "Systems and Signals Tools for Estimation and Analysis of Mathematical Models in Endocrinology and Neurology".

Email address: kp@it.uu.se (Kristiaan Pelckmans).

This is a rewritten, refined and extended version of the accepted CDC 2010 submission 'On the Identification of Monotone Wiener Systems'.

2 The full quote is even more adequate: "It can scarcely be denied that the supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation of a single datum of experience." A. Einstein, 1933.
rectly averaging out the nonlinear effect. Although powerful concentration inequalities lie on the basis of this approach, no argument is given that this method applies for Wiener systems which are more complex (realistic) than the academic examples presented in those papers. Formally, this work studies the identification of nonlinear dynamic models which belong to the following model class.

**Definition 1 (FIR Wiener Model \((f, \alpha)\)** A Wiener model consists of a sequence of (i) a linear dynamical model characterized by an impulse response function \(H(q^{-1})\) (here \(q^{-1}\) is the backshift operator as classically applied on the input signal \(\{u_t\}_t\)), and (ii) a static nonlinear function \(f: \mathbb{R} \to \mathbb{R}\) (see Fig. 1). If the signals \(\{u_t\}_t\) and \(\{y_t\}_t\) follow such a model with 'true' subsystem \(H_0\) and 'true' function \(f_0\), we can write

\[
y_t = f_0(H_0(q^{-1})(u_t)),
\]

and we say that the observations come from the Wiener system \((H_0, f_0)\). For a FIR-Wiener model of order \(d\), one considers a Finite Impulse Response (FIR) parametrization of the linear subsystem, or \(H(q^{-1}) = a_1q^{-1} + \cdots + a_dq^{-d}\). We denote such model (in the context of this paper) shortly as the \((f, \alpha)\)-Wiener model. Now since \(a\) and the domain of \(f\) can be rescaled arbitrarily, it is convenient to impose \(\|a\|_2 = 1\), which does avoid identifiability issues. If the signals \(\{u_t\}_t\) and \(\{y_t\}_t\) obey such a model with 'true' function \(f_0\) and 'true' parameters \(a_0\), or

\[
y_t = f_0\left(\sum_{k=1}^d a_{0,k}u_{t-k}\right) = f_0(a_0^T u_t).
\]

- where \(a_0 \in \mathbb{R}^d\) and \(\|a_0\|_2 = 1\) and we define \(u_t = (u_{t-1}, \ldots, u_{t-d})^T \in \mathbb{R}^d\); we say that the observations come from the Wiener system \((a_0, f_0)\). We will denote the set of possible observations as \(S = \{(u_t, y_t)\}_t \subseteq \mathbb{R}^d \times \mathbb{R}\).

We specialize further to a subset of this class as follows, schematically illustrated in Fig. 1.

**Definition 2 (Monotone FIR Wiener Model \((f, \alpha)\)** A FIR-Wiener model \((f, \alpha)\) is called monotone if \(f : \mathbb{R} \to \mathbb{R}\) is monotonically increasing (but not necessarily invertible) such that \(y_t = f(a^T u_t)\). We define the Monotone Wiener model class formally as

\[
\mathcal{F} = \{(f, \alpha) \mid f: \mathbb{R} \to \mathbb{R} \text{ Monotonically increasing, } a \in \mathbb{R}^d, \|a\|_2 = 1\}.
\]

Note that by similarity one has \((f, \alpha) = (f', -\alpha)\), where \(f(z) = f'(-z)\) for all \(z \in \mathbb{R}\). Now \(f'\) is monotonically decreasing, explaining why we can omit the denominator 'increasing' in the nomenclature. As argued before, this class of monotone FIR-Wiener models can capture such different effects as (1) quantized output measurements, (2) saturation effects of the sensor, and (3) handling of general bijective transformations of the output scaling (cfr. the temperature scale of Celsius versus Fahrenheit), amongst others.

This paper is organized as follows. Section 2 formulates the MINLIP estimator, gives the underlying rationale, derives almost consistency of the estimates, and extends the estimator to the case where noise is present in the data. Section 3 then gives nontrivial numerical examples of the identification technique. Section 4 concludes with some open questions.

### 2 Identification of Monotone Wiener Systems

#### 2.1 The Realizable Case

The rationale behind the technique goes as follows. Identification of a Wiener system from observations is traditionally obtained by solving

\[
\min_{a, f : \|a\|_2 = 1} J(f, \alpha) = \sum_{t=d+1}^T (f(a^T u_t) - y_t)^2.
\]

We refer to this formulation as to a prediction error method for Wiener models - abbreviated here as WPPEM - and it will be mainly this approach that we will contrast the proposed method against. Note that formally this approach is an Output-Error Method (OEM). It can only be called a PEM method when stringent stochastic assumptions can be made on the noise, see e.g. [6] for a discussion of this issue. In general, such a formulation is hard to solve as the unknowns interact directly with the unknown function \(f\). As a result one typically resorts to an iterative scheme or general purpose nonlinear optimization routine. The practical procedures lack generality and robustness for different reasons: (i) depending on the form (or parametrization) of \(f\), ill-conditioning of the problem may arise or even gradient information may not exist; (ii) the problem can often be stuck in local minima, which can be arbitrary bad; (iii) procedures are highly depending on the exact representation of the unknown \(f\). In general, such procedures are therefore not easily scalable to more complex settings.

However, if one can safely assume that an \((a_0, f_0)\) interpolates the given samples exactly as in eq. (2), the following
approach is applicable
\[
\min_{a, f : \|a\|_1=1} J'(f, a) = c(f)
\]
s.t. \( y_t = f(a^T u_t), \forall t = d + 1, \ldots, T, \) (5)

where \( c : \{ f \} \rightarrow \mathbb{R} \) is an appropriate complexity term. This is a formalization of the saying ‘Make things as simple as possible, but no simpler’, sometimes dubbed as Einstein’s razor. The rationale is that \( a \) - which is a misspecification of \( a_0 \) - requires an \( f' \) with model complexity \( c(f') > c(f_0) \) in order to interpolate the data. This is often the case for smoothness complexity terms: if \( a \) were misspecified some values \( \{a^T(u_i - u_j)\} \) would decrease, necessitating a less smooth \( f \) in order to have \( y_i = f(a^T u_i) \) and \( y_j = f(a^T u_j) \). Specifically, MINLIN formalizes such smoothness property in terms of the Lipschitz constant of \( f \). This Lipschitz constant is defined as the minimal \( L \) such that
\[
|y_i - y_j| \leq L|z_i - z_j| \forall i, j,
\]
when a function is only defined as the mapping of samples \( \{z_i\} \), to \( \{f(z_i) = y_i\} \). This is equivalent to the ordinary Lipschitz constant of the piecewise linear interpolation of \( \{(z_i, y_i)\} \). A formal derivation of why this complexity term is appropriate is given in the next subsection. Now, since the form of \( f \) is not important (nor known) except for (i) the order preserving monotonicity property, and (ii) the corresponding Lipschitz constant, the problem (5) where \( c(f) = L \) can be rephrased as
\[
\min_{a, L} L^2 \text{ s.t. } \|a\|_2 = 1,
\]
\[
|y_i - y_j| \leq L|a^T(u_i - u_j)|, \forall i = d + 1, \ldots, T. \tag{7}
\]

Note that the exact form of \( f \) does not follow directly, but is implicitly modeled in the problem formulation. After a change of variable and exploiting the ordering relations between the samples, the MINLIN estimator is obtained:

**Definition 3 (MINLIN)** Given an ordered set of samples \( \{(u_i, y_i)\}_{i=1}^T \subset \mathbb{R}^d \times \mathbb{R} \) indexed such that \( y_{(i-1)} \leq y_{(i)} \) for all \( i = d + 1, \ldots, T \), then our (rescaled) estimate \( \hat{a}_T \) follows by solving
\[
\hat{a}_T = \arg\min_a a^T a \text{ s.t. }
\]
\[
(y_{(i)} - y_{(i-1)}) \leq a^T(u_{(i)} - u_{(i-1)}), \forall i = d + 1, \ldots, T, \tag{8}
\]

where the estimated function \( \hat{f} \) is specified only implicitly as in Proposition 1. Moreover, the Lipschitz constant is recovered as \( L = a^T_T \hat{a}_T \).

This problem can be cast as a convex Quadratic Program (QP) with \( T - d \) linear constraints and \( d \) unknowns. This problem can be solved efficiently with contemporarily available solvers \(^3\) in most mathematical packages \(^4\). The \( a_T \) which minimizes this constrained objective is our estimate of a (rescaled) version of the parameters of the FIR system \( H(q) \). The problem is written in matrix notation as
\[
a_T = \min_{a \in \mathbb{R}^d} a^T a \text{ s.t. } (\Delta a) a \geq \Delta y, \tag{9}
\]

where \( u = (u_{(d)}, \ldots, u_{(T)}) \in \mathbb{R}^{T-d+1} \) is a Hankel matrix (up to sorting), \( y = (y_{(d)}, \ldots, y_{(T)}) \in \mathbb{R}^{T-d+1} \) is an ordered vector and
\[
\Delta = \begin{bmatrix}
-1 & 1 & 0 \\
0 & -1 & 1 \\
\end{bmatrix} \in \mathbb{R}^{T-d+1} \times (T-d+1).
\]

2.2 Almost Consistency of MINLIN

This subsection characterizes how well the estimate approaches the true impulse response, in case suitable assumptions on the data and the static nonlinearity hold. Particularly, we assume that the observations arise from a true Monotone Wiener model \((f_0, a_0)\) with FIR system of given order \( d \) for the dynamic part, and when no noise perturbs the observations. This problem is already non-trivial as (i) the proposed model is essentially nonlinear and non-parametric, and (ii) the method is not based directly on minimizing a mismatch between the model and the observations. The main outcome is that almost consistent estimates are given when the data satisfies a condition of local Persistency of Excitation (PE), and the true monotone static nonlinearity is smooth around its steepest part. The latter condition is closely related to the modulus of continuity, playing a crucial role in the analysis of nonlinear estimators in theoretical statistics \([15]\). This result is referred to here as **almost consistency**, as guarantees accuracy of the estimates only up to a small (but often non-zero) approximation term. In a sense, this is the best one could hope for here because of two reasons: (i) The model is non-parametric (or semi-parametric) as the static monotone function of the model cannot be expressed straightforwardly in terms of a (small number of) parameters. In that respect, a finite dataset contains never enough information in order to reconstruct this system exactly. (ii) A finite dataset can never be locally exciting in every (arbitrary small) neighborhood, but can only guarantee this condition for all localities which are sufficiently large. In other words, when a finite dataset is given, the samples cannot be expected to sufficiently rich in all arbitrary small balls around all possible inputs: such

\(^3\) In our experiments we use the solver available at http://www.mosek.com.

\(^4\) In practice, it is possible to run this method in a matter of seconds on a standard laptop for datasets of size up to 20.000 samples.
balls will only scarcely contain a sample tout court! Since the model is nonlinear, such localized richness of the data has to be assumed in some form. The analogue for estimating a Linear Time Invariant (LTI) FIR model goes as follows: assume the system can be described exactly as a FIR model of given order (smaller than) \( d \), then the appropriate condition for exact consistency of the least squares estimate is whether the data is (globally) PE to order at least \( d \) (see e.g. [10]). Since we assume that there is no noise in the data, there is no approximation to be made in that case.

Assume that the observed system obeys the relation given in (2) with a fixed (but unknown) monotonically increasing function \( f_0 : \mathbb{R} \rightarrow \mathbb{R} \) which is Lipschitz monotone with constant \( L_0 < \infty \), and parameter vector \( a_0 \in \mathbb{R}^d \) with \( |a_0|_2 = 1 \). Those are referred to as the true function \( f_0 \) and the true parameters \( a_0 \) respectively. Then, the question is addressed whether if we see enough data, the MINLIP estimator \( \hat{a}_T \) will equal \( a_0 \) up to a scaling constant. Sometimes it will be convenient to rewrite the MINLIP estimator (8) as the following minimax problem:

\[
\ell = \max_{\|a\|_2 = 1} \inf_{(u,y),(u',y') \in S,y>y'} \frac{a^T(u-u')}{{y-y'}}. \tag{11}
\]

where \( \ell \geq \frac{1}{m} \) by construction of \( L_0 \). In order to characterize the solution, the following two conditions are needed.

**Definition 4 (f is \((L_0,g)\)-Lipschitz on \( S' \subseteq \mathbb{R} \))** The function \( f \) is said to be \((L_0,g)\)-Lipschitz on \( S' \subseteq \mathbb{R} \) for a decreasing, positive function \( g : \mathbb{R}^+ \rightarrow \mathbb{R}^+ \) with \( g(0) = 1 \) if:

(A) one has for all \( z,z' \in S' \) that

\[
(f(z) - f(z')) \leq L_0(z-z'). \tag{12}
\]

(B) there exists a \( z \in S' \) and \( z' \in S' \) such that

\[
(f(z) - f(z')) = L_0(z-z'), \tag{13}
\]

(C) one has for this \( z \), for any \( \epsilon > 0 \) and \( z'' \in S' \) where \( |z-z''| \leq \epsilon \) that

\[
|f(z) - f(z'')| \geq g(|z-z''|)L_0|z-z''|. \tag{14}
\]

Hence \( g \) denotes how `smooth' the constant \( L \) decays in a neighborhood of \( z \) where the actual Lipschitz constraint is met (that is, a slower decaying function \( g \) indicates a higher smoothness). In particular, a value \( h(\epsilon) = 1 \) implies that the function \( f \) is linear with slope \( L_0 \) in this neighborhood. Such characterization is illustrated for \( f(z) = \tanh(z) \) in Fig. (2) for a smoothness function \( g(\epsilon) = 1/(1+ce) \).

**Definition 5 (\( \epsilon \)-Local Persistently Exciting)** We say that a set \( S \subseteq \mathbb{R}^d \) is \( \epsilon \)-local persistent exciting of order \( m \) for \( \epsilon > 0 \) iff for any vector \( u \in S \), there exist \( m \) vectors \( u_1, \ldots, u_m \in S \) with \( \{u - u_k\}_{k=1}^m \) linearly independent vectors and

\[
\|u - u_k\|_2 \leq \epsilon, \forall k = 1, \ldots, m. \tag{15}
\]

This definition can be seen as a local version of Persistence of Excitation (PE), see e.g. [10] for the classical definition of PE. In case both Def. 4 and Def. 5 hold, then almost consistency is implied as proven next.

Fig. 2. Schematic illustration of the \((L_0, g)\)-Lipschitz property of a function \( f \) with \( g(\epsilon) = \frac{1}{1+\epsilon} \). There should be a sample \( z \) where the Lipschitz constant \( L_0 \) is attained, and in the \( \epsilon \)-neighborhood of this sample \( z \) the Lipschitz-property shouldn’t decay too fast, e.g. in the neighborhood of \( z \) the function behave almost linearly.

**Theorem 1 (Almost Consistency)** Fix \( \epsilon > 0 \) and consider the \((f_0, a_0)\)-monotone Wiener system as in Def. 2, with corresponding observations in \( S \). If \( f_0 : \mathbb{R} \rightarrow \mathbb{R} \) is \((L_0, g)\)-Lipschitz and monotone on the set \( \{(z,y) : z = a_0^T u : u \in S\} \) and \( S \) is \( \epsilon \)-local PE, then

\[
\hat{a}_T^T a_0 \geq g(\epsilon), \tag{16}
\]

where \( \hat{a}_T \) is the renormalized estimate of MINLIP as in (8), that is \( a_T \propto \hat{a}_T \) and \( \|\hat{a}_T\|_2 = 1 \).

This means that the smoother the function \( f_0 \) is towards its steepest part, the better estimates we get. In particular, when \( f_0 \) is (almost) linear - we only need global PE to have exact estimates \( a_T \propto a_0 \). This notion is similar to the classical modulus of continuity of a nonlinear estimator, which is often used in order to characterize the properties of statistical estimators, see e.g. [15].

**Proof:** Let the Lipschitz constant be achieved in the sample \( (u,y), (u',y') \in S \), such that

\[
(y - y') = L_0(u - u')^T a_0. \tag{17}
\]

As the set \( S \) is \( \epsilon \)-local persistent exciting of order \( d \), one can find for the sample \( (u,y) \in S \) \( d \) vectors \( u_1, \ldots, u_d \) contained in \( S \) such that the vectors \( (u-u_1), \ldots, (u-u_d) \) are linearly independent and have norm smaller than \( \epsilon \). This implies that one can rewrite \( a_0, a_T \in \mathbb{R}^d \) (i.e. the true parameter vector and the optimal estimate associated to (8)) as

\[
a_0 = \sum_{k=1}^d \alpha_k (u - u_k) \quad \text{and} \quad a_T = \sum_{k=1}^d \alpha_k (u - u_k)
\]

where \( \alpha, \alpha_0 \in \mathbb{R}^d \), and we let \( (u - u_k) = \sigma_k(u - u_k) \).
\[
\|u_k\|_2 \text{ with } \sigma_k \in \{-1, 1\} \text{ such that } \|u - u_k\|_2 = 1 \text{ and } \frac{u - u_k}{a_0} \geq 0 \text{ for all } k = 1, \ldots, d. \text{ Define as previously for each } k = 1, \ldots, d, \text{ the constant } L_k \in \mathbb{R}_+ \text{ such that } (y - y_k) = L_k(u - u_k) - a_0 \text{, where } \mathcal{L}_k \leq L_0 \text{ by construction.} \text{ Now define the matrix } \mathcal{D}_u \in \mathbb{R}^{d \times d} \text{ such that } \\
\mathcal{D}_u \mathcal{D}_u = \begin{pmatrix}
(u - u_1), (u - u_2), \ldots, (u - u_d)
\end{pmatrix}^T \text{ such that } \mathcal{D}_u a_0 \geq 0, \text{ and the matrix } L \in \mathbb{R}^{d \times d} \text{ as } L = \text{diag}(L_1, \ldots, L_d). \text{ Then we have in matrix notation that } a_0 = \mathcal{D}_u^T a_0 \text{ and } a_T = \mathcal{D}_u^T a. \text{ By construction of the MINLIP estimator we have that } \mathcal{L}_u \mathcal{D}_u a_0 = \mathcal{L}_u \mathcal{D}_u a_T \leq \mathcal{L}_u \mathcal{D}_u a_0 \text{, where } \ell \text{ is the minimal value obtained in (11). As such } \frac{\mathcal{L}_u \mathcal{D}_u a_0 \leq \mathcal{D}_u \mathcal{D}_u a}{\mathcal{D}_u a_0.} \text{ Then one has }
\begin{align*}
\mathcal{D}_u^T a_T &= a_0^T (\mathcal{D}_u \mathcal{D}_u a) a_0 \geq \frac{1}{\ell} \mathcal{D}_u^T L (\mathcal{D}_u \mathcal{D}_u a) a_0 \\
&\geq \frac{g(\ell) L_0}{\ell} a_0^T (\mathcal{D}_u \mathcal{D}_u a) a_0 \\
&\geq g(\ell) a_0^T (\mathcal{D}_u \mathcal{D}_u a) a_0 = g(\ell) a_0^T a_0 = g(\ell),
\end{align*}
\]}

since \( \ell \leq L_0. \) This proofs the result.

### 2.3 Identification with Noisy Data

There are a number of different ways one can model noise in the class of monotone Wiener systems. A first one is to consider noise on the measured outputs (or measurement noise). One may argue that this model is not a very realistic assumption in case the observations are quantized versions of the output of the linear system. That is, once the signal is quantized (and transmitted), it can often be measured without error. On the other hand, it is often not clear which noise model of a quantized signal (with a finite number of different levels) fits the application (a Gaussian distribution would not make much sense here). Another assumption one can make is that noise occurs in the signal \( \{u_t\}_t \), but as this results in coloring of the noise by the unknown linear subsystem, this model is not adopted as yet. A third alternative is that the noise comes in between the linear subsystem and the monotone static function (see Fig. 3). As in the following no restrictive assumptions (as whiteness of the noise signal is assumed), this could be seen as uncertainty coming in in the model by under-modeling of the linear system. This is the view which underlies the following definitions.

**Definition 6 (Noisy FIR Wiener Model \((f, a)\))** A FIR Wiener model consists of a sequence of \((i)\) a linear dynamical model characterized by an impulse response function \(H(q^{-1})\) applied on the input signal \(\{u_t\}_t\), \((ii)\) a static nonlinear function \(f : \mathbb{R} \to \mathbb{R}\), and \((iii)\) a sequence of ‘noise’ terms \(\{e_t\}_t\). If the signals \(\{u_t\}_t\), \(\{y_t\}_t\) and \(\{e_t\}_t\) follow such a model with ‘true’ subsystem \(H_0\) and ‘true’ function \(f_0\), we can write
\[
y_t = f_0\left(H_0(q^{-1})(u_t) + e_t\right),
\]

![Fig. 3. Schematic representation of a noisy monotone Wiener system. This paper adopts the setting that noise comes in after the linear dynamic part (capturing model mismatch), and right before application of the static nonlinearity (or quantization).](image)

If the signals \(\{u_t\}_t, \{y_t\}_t, \text{ and } \{e_t\}_t\) obey a Wiener FIR-model with ‘true’ function \(f_0\) and ‘true’ parameters \(a_0\), or \(y_t = f_0\left(\sum_{k=1}^d a_k u_{t-k} + e_t\right)\) where \(a_0 \in \mathbb{R}^d\) and \(\|a_0\|_2 = 1\).

Given time-series \(\{u_t\}_t, \text{ and } \{y_t\}_t\), referred to as ‘input’ and ‘output’. Again, let \(\{(u_t, y_{t,i})\}_j \in \mathbb{R}^d\) be a dataset containing \(T-d+1\) samples. Let this set be reindexed as \(\{(u_{ij}, y_{ij})\}_j \in \mathbb{R}^d\) where \(y_{ij} \leq y_{ij}\) for all \(d < i < j \leq T\). Then adopting the noisy model (2) suggests modification of the standard MINLIP (see eq. (8)) as
\[
\min_{a,c} \frac{1}{2} a^T a + \frac{\gamma}{2} \sum_{t=1}^T c_t \text{ s.t. } \forall i = d + 1, \ldots, T:
\]
\[
y_{ij} - y_{ij-i} \leq (a^T u_{ij} + e_{ij}) - (a^T u_{ij-i} + e_{ij-i}),
\]

where the fixed regularization parameter \(\gamma > 0\) trades the Lipschitz based regularization term and the penalization of the residuals. The choice of penalizing the absolute loss of the residuals is inspired by (i) robustness considerations and (ii) non-stochastic nature of the residuals where a worst-case approach is more suited. The tuning of this constant can be done with an appropriate model selection criterion as cross-validation. As before, this optimization problem can be solved efficiently as a convex quadratic program (QP) using \(O(T)\) unknowns and inequality constraints.

### 3 Empirical Evidence

This section spells out a number of artificial yet challenging case studies. A main argument which is made here is that although the underlying system \(H_0\) may be represented as a fractional polynomial, it is often useful to consider a FIR model consisting of a large number of tapped delays (say \(d = O(100)\)) which can approach (the impulse response function of) \(H_0\) arbitrarily close. In this way, one does not have to specify explicitly model order or delay of the model. We will refer to this approach as an over-parametrization. In order to choose the number \(d\) of tapped delays in the FIR model, one may perform a non-parametric analysis of the impulse response of the data (neglecting nonlinear effects as yet). It is found that MINLIP is especially appropriate for such an over-parametrization approach and doesn’t lose much efficiency as it builds in explicitly a mechanism of
model complexity control and regularization, dealing with ill-posedness problems often present in such context.

The monotone Wiener systems from which the data is generated take the following form. The linear subsystem $H_0(q^{-1})$ are represented as fractional polynomial models as

$$
H_0(q^{-1}) = \frac{B(q^{-1})}{A(q^{-1})} = \frac{b_0 + b_1q^{-1} + \ldots + b_{2m_z}q^{-2m_z}}{1 + a_1q^{-1} + \ldots + a_{2m_p}q^{-2m_p}},
$$

where $2m_z > 0$ and $2m_p > 0$ denote the orders of the polynomials $A(q^{-1})$ and $B(q^{-1})$. Those polynomials are chosen such that they have $m_z$ and $m_p$ conjugate pairs of zeros and poles respectively. In this example, we set $m_z = 2$ and $m_p = 20$. The conjugate poles and conjugate zeros are uniformly at random picked (strictly) inside the unit circle (see Fig. 4 for an example). In general, we see that a FIR representation of $d = 200$ is sufficient to capture the dynamics of such a system. The output nonlinearity is fixed as $f_0 : \mathbb{R} \to \mathbb{R}$ where for $x \in \mathbb{R}$ one has

$$
f_0(x) = 2 + \tanh(5x + 2) + 0.5 \tanh(5x - 3).
$$

This function is somewhat challenging as it cannot be described as a simple saturation function, is not symmetric around any point, and has an almost zero gradient in $x = 0$. Then a monotone Wiener system is constructed as

$$
y_t = f_0(gH_0(q^{-1})u_t), \forall t = 1, \ldots, T,
$$

where the gain $g > 0$ is chosen such that the values $\{gH_0(q^{-1})u_t\}_t$ have a unit standard deviation. The estimates of MINLIP on a time-series of length $T=450$, $500, 550, 600$ - taken from the Wiener System of Fig. (4), Fig. (5) - is displayed in Fig. (5). Here a FIR approximation of $d = 200$ is used, capturing the dynamics of the system $H_0$ reasonably well (see Fig. (4.a)).

Fig. 4. An example of a system $H_0$ randomly generated, with $m_p = 20$ and $m_z = 2$. Panel (a) shows the resulting impulse response (and hence a FIR approximation) up to lag $d = 200$. The blue dashed line gives the estimate of eq. (4), the red dashed-dotted line gives the estimate given by MINLIP. Panel (b) displays the conjugate poles and conjugate zeros in the complex domain using a pole-zero plot of the system. Observe (i) the presence of a considerable (non-zero) delay, and (ii) the fact that some poles are located close to the unit circle. This makes a inverse modeling approach unfeasible.

The following 6 different identification methods were implemented to benchmark the MINLIP against:

1. (LS $'x-y'$) In order to provide a (naive) lower-bound to the performance, a FIR identification technique based on a Least Squares (LS) argument was implemented on the signals $\{u_t\}_t$ and $\{y_t\}_t$ directly, neglecting the Wiener structure altogether.

2. (LS $'x-z'$) In order to get an upper-bound on the performance of the identification technique, an ARX identification technique was implemented based on the (latent) intermediate signal $\{z_t = H_0(q^{-1})u_t\}_t$.

3. (WPEM FIR) We consider a FIR model structure of sufficiently high order (here $d = 200$) such that (21) can be represented fairly well, and we let the corresponding FIR coefficients act directly as unknowns. The nonlinearity is represented as a piecewise function based on 20 fixed knots which were optimally tuned to the example at hand. Global optimization on both FIR coefficients as well as on the unknowns of the nonlinearity is performed by the Broyden-Fletcher-Goldfarb-Shannon (BFGS) method implemented in MATLAB in the fminunc function.

4. (WPEM ARX) Here we implement the same approach now based on a class of ARX models representing the optimal predictors corresponding to (21). Now, we let the poles and zeros of this model class act as unknowns directly. As before, the nonlinearity is expressed as a piecewise linear function with fixed grid points. Global optimization is performed by the BFGS method.

5. (Greblicki2002) The smoothing approach as described in [8] is implemented as well. This method works directly on a FIR overparameterization of the (linear sub-) system, and gives reasonable estimates when sufficiently many input-samples following a stochastic (approximately white) Gaussian process are provided.

6. (Bai2006) The last approach we benchmark against is the technique described in [14, 3] using prior knowledge of the monotonicity of the output function. This
technique is implemented by solving
\[
    \min_a \sum_{j=d+1}^{T} \left( \text{sign}(y(j) - y(j-1)) - \text{sign}(u(j) - u(j-1))^T a) \right)
\]
(24)
which is in our experiments solved by the BFGS method. We found that in order to make this
approach to work well one needs to resort to a smooth proxy 'sign' of the discrete function
'sign' = \( I(z > 0) - I(z < 0) \), making gradient information available at most unknowns in
the search space. In many cases solving this problem takes substantially more resources (CPU-power, memory)
compared to the other techniques.

Accuracy of an estimate is expressed in terms of the angle
response of the estimate. As such, measuring performance based on prediction
dictions requires additional estimation of the static nonlin-
erly. Thirdly, the current method concentrates at first instance
(hence cannot play a role in the quality measure), and the

If the systems \( H_0 \) and \( H_T \) have impulse response vector
\( h_0 \) and \( h \) respectively, this coefficient can be written as the
Pearson correlation coefficient between those two vectors, or
\[
    \rho_h = \frac{\sum_{i=0}^{\infty} H_0(q^{-1})\delta_i H_T(q^{-1})\delta_i}{\| H_0 \|_2 \| H_T \|_2}.
\]
(25)
If the systems \( H_0 \) and \( H_T \) have impulse response vector
\( h_0 \) and \( h \) respectively, this coefficient can be written as the
Pearson correlation coefficient between those two vectors, or
\[
    \rho_h = \frac{\sum_{i=0}^{\infty} H_0(q^{-1})\delta_i H_T(q^{-1})\delta_i}{\| H_0 \|_2 \| H_T \|_2}.
\]
(25)

The first experiment is based on noiseless data. Figure (6.a)
shows the results of the experiment, where in each iteration
\( T \) samples are generated from a random system \( (f_0, H_0) \),
the different identification algorithms are carried out, and
their respective accuracy is computed. This iteration is per-
formed 100 times for any \( T = 300, 400, 500, \ldots, 1000 \). We see from the results that the MINLIP estimator converges
fast to the best achievable performance (indicated by the
LS 'x' - 'x' approach). The WPEM algorithms in many
cases unreliable results, performing much worse on the aver-
age. Specifically, we find that it is bad practice to combine
the WPEM on (overparametrized) FIR model, perhaps be-
cause global optimization often presents (numerical) prob-
lems when optimizing over so large a set of parameters.
MINLIP however can handle such overparametrization quite

\[
    f_0'(z) = I(z > -0.5) + I(z > 2),
\]
(26)
with the output taking values in the set \( \{0, 1, 2\} \), and the
jumps of the function occurring when \( z = -0.5 \) and \( z = 2 \).
Again, the linear systems \( H_0 \) used to generate the signals
are as in eq. (21), and we benchmark the MINLIP against
the approaches described in the previous subsection (LS, WPEM and BAI). The results are displayed in Fig. (6.b).

Fig. 6. Panel (a): results of the first experiment using noiseless data generated from a monotone Wiener nonlinearity as in eq. (22) and systems \( H_0 \) as in (21). Performance expressed as correlations of \( H_0 \) and \( H_T \) as in (25) are displayed for sample sizes ranging from
\( T=210 \) to \( T=1000 \), using a FIR overparametrization of \( d = 200 \).
The vertical line denotes the place where a least squares technique
would exactly reconstruct \( H_0 \) if \( z_t = H_0(q^{-1})u_t \) were given.
Panel (b): results of the quantization experiment using noiseless
data generated from a quantized nonlinearity \( f_0 \) as in eq. (26)
and systems \( H_0 \) as in (21). Performance expressed as correlations of
\( H_0 \) and \( H_T \) as in (25) are displayed for sample sizes ranging
from \( T=210 \) to \( T=1000 \), using a FIR overparametrization of
\( d = 200 \).
From these results we may suggest a few guidelines. The first is that a naive LS ’\(x - y'\) regression works surprisingly good, and it is not at all trivial to beat this one. The reason the approaches based on global optimization (i.e. WPEM, WPEM.FIR and Bai2006) do not work as good might be that the discrete nature of the identification task translates in a highly non-smooth cost surface given to the optimizer. This experiment however suggests that MINLIP achieves a solution which is often close to the best one could hope for (indicated by the LS ’\(x - z'\) method). The averaging approach proposed in [8] appears fairly robust to the quantization effects as well.

This paragraph reports results achieved with MINLIP in case the intermediate signal \(\{z_t\}\) was perturbed by noise. The experiment is set up as before in Subsection 2.3, but the system becomes now \(y_t = f_0(H_0(g^{-1})a_t + e_t)\) for all \(t = 1, \ldots , T\) where \(f_0\) is as in eq. (22) and \(H_0\) is randomly generated as in (21). The terms \(\{e_t\}\) are zero mean white Gaussian noise with standard deviation \(\sigma_e > 0\). This experiment is conceived in a slightly different manner than before. We consider a fixed number of samples \(T = 500\), and let the Signal-to-Noise Ratio (SNR) vary from 0.1 to 10 (or \(\sigma_e = 0.1, \ldots , 1\)). As indicated in Section III, MINLIP is dependent on the choice of a suitable \(\gamma > 0\), which is in turn depending on the noise variance \(\sigma_e\). As this characteristic is unknown in general in practical applications, the choice of \(\gamma\) is to be made based on a suitable model selection technique. Actually, the problem of model selection in the context of a Wiener model is not covered as such, and prompts new questions related to information criteria, stability and consistency. For now, we use a fixed value of \(\gamma = 10\) which works well in many cases. The results are displayed in graph (7.a). Those results indicate that the MINLIP outperforms the other techniques especially when noise is small compared to the ‘informative’ signal, while all techniques become arbitrarily bad when this ratio grows.

The last experiment fixes an SNR of 3 and studies the performance of the different estimators for signals of increasing length. The evolution of the average accuracy of MINLIP and the competing estimators is given in Fig. (7.b). From this result we see that the behavior is not too different from the noiseless case, except the fact that the WPEM and WPEM.FIR approaches are not very robust to noise, and the approach in (24) is clearly a bad choice in this case and needs additional care.

4 DISCUSSION

This paper discusses the MINLIP estimator applied for identification of monotone Wiener systems. Theoretical analysis implies that this technique gives almost consistent estimates under appropriate conditions. Extensive empirical experiments are given supporting the claim of efficiency of this approach. Many new questions open up in view of this work, perhaps the most interesting is how to extend ideas towards the identification of other block-structured models.

References