52nd IEEE Conference on Decision and Control
December 10-13, 2013. Florence, Italy

Hankel Based Maximum Margin Classifiers: A Connection Between Machine Learning and Wiener Systems Identification

F. Xiong Y. Cheng O. Camps M. Sznaier C. Lagoa

Abstract— **This paper considers the problem of non-parametric identification of Wiener systems in cases where there is no a-priori available information on the dimension of the output of the linear dynamics. Thus, it can be considered as a generalization to the case of dynamical systems of non-linear manifold embedding methods recently proposed in the machine learning community. A salient feature of this framework is its ability to exploit both positive and negative examples, as opposed to classical identification techniques where usually only data known to have been produced by the unknown system is used. The main result of the paper shows that while in principle this approach leads to challenging non-convex optimization problems, tractable convex relaxations can be obtained by exploiting a combination of recent developments in polynomial optimization and matrix rank minimization. Further, since the resulting algorithm is based on identifying kernels, it uses only information about the covariance matrix of the observed data (as opposed to the data itself). Thus, it can comfortably handle cases such as those arising in computer vision applications where the dimension of the output space is very large (since each data point is a frame from a video sequence with thousands of pixels).**

I. Introduction and motivation

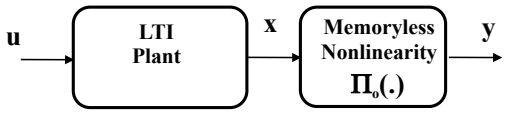


Fig. 1: Wiener System Model

Wiener systems, consisting of the cascade of a Linear Time Invariant (LTI) plant and a memoryless nonlinearity as shown in Figure 1, are ubiquitous in domains ranging from communications to biology and computer vision [7], [4], [16]. Further, these systems provide tractable approximations to more general nonlinear control problems. Thus, during the past few years a substantial research effort has been devoted to the problem of identifying these systems from experimental data, under various assumptions on the available a-priori information. This effort has lead to a large number of approaches, which can be roughly classified into statistical (see for instance [24], [6], [9], [10], [3], [1], [2], [18], [13]) and set membership ([5], [26] and references therein). A salient feature of these approaches is that the dimension n_x of $\mathbf{x}$, the output of the linear portion of the system, is assumed to be known. However, in many cases of practical interest, this information is not a-priori available. Examples of this situation are computer vision applications such as target tracking or activity recognition [14], [25], where the output $\mathbf{y}$ consists of the vectorized frames of a video sequence, and $\mathbf{x}$ is a small set of independent parameters that encapsulate the correlations between the different pixels. In these cases the dimension of $\mathbf{x}$ must also be identified from the experimental data, a situation that cannot be handled by existing Wiener systems identification techniques.

On the other hand, non-linear manifold embedding, that is finding low dimensional parsimonious non-linear projections of high dimensional correlated data, is a classical problem in the machine learning community, where a wealth of methods has been developed during the past few years. These include Locally Linear Embeddings (LLE) [20], semi-definite embedding (FSDE)[23], Global coordination of local linear models (GCM) [19] and Dynamic Global Coordinate Model (DGCM) [15]. While these methods have proved very efficient in handling static data, most do not exploit dynamical information, encapsulated in the temporal ordering of the data, and thus may fail to capture the underlying temporal dynamics. An exception is the paper [25], proposing an embedding algorithm for dynamic data that minimizes a combination of the dimension of the embedding manifold and the McMillan degree of the underlying dynamics, by recasting the problem into a rank minimization form subject to semidefinite constraints. However, no identifiability issues are taken into consideration when selecting the embedding manifold.

The present paper seeks a rapprochement between systems identification and machine learning techniques. Our goal is, starting from experimental input output data, to find an embedding manifold such that the data can indeed be explained as a trajectory of a Wiener system, and to identify its linear and non-linear portions. A salient feature of the proposed approach (common in machine learning, but to the best of our knowledge hitherto not used in the identification community), is the ability to use of both positive and negative samples, that is experimental data generated both by the system to be identified and by other systems. This is a situation commonly encountered in applications such as activity classification, where sample clips of different activities are available, or in tracking, where often a segmentation separating the target of interest from other targets and the background is known. In addition, motivated by the approaches in [21], [25] and kernel-based classification methods in machine learning, rather than working with the

This work was supported in part by NSF grant ECCS–0901433; AFOSR grant FA9559–12–1–0271; and DHS grant 2008-ST-061-ED0001. F. Xiong, Y. Cheng, O. Camps and M. Sznaier are with the Department of Electrical and Computer Engineering, Northeastern University, Boston, MA 02115. C. Lagoa is with the Department of Electrical Engineering, Penn State University, University Park, PA 16802.

978-1-4673-5717-3/13/$31.00 ©2013 IEEE

potentially high dimensional data $\mathbf{y}$, the proposed algorithm uses the inner products $\langle \mathbf{y}_i, \mathbf{y}_j \rangle$, resulting in substantial dimensionality reduction in the matrices involved. The main result of the paper shows that in this context, the problem of jointly finding the embedding manifold and the linear dynamics can be recast into a polynomial optimization over a semi-algebraic set (a set defined by a collection of polynomial inequalities). In turn, the use of recent results from polynomial optimization allows for relaxing this problem to a tractable convex optimization. These results are illustrated with an example showing that the use of both positive and negative samples substantially improves the quality of the models identified from noisy data.

II. Preliminaries

For ease of reference, in this section we summarize the notation used in the paper and recall some results on sparse polynomial optimization that play a key role in establishing the main result of this paper.

A. Notation and Definitions

$\mathbb{R}, \mathbb{N}$	set of real number and non negative integers
$\mathbf{x}, \mathbf{M}$	a vector in $\mathbb{R}^n$ (matrix in $\mathbb{R}^{n\times m}$)
$\mathbf{M} \succeq \mathbf{N}$	the matrix $\mathbf{M} - \mathbf{N}$ is positive semidefinite.
$\mathbf{H}_y^{m,n}$	Hankel matrix associated with a vector sequence $\mathbf{y}(.)$: $\mathbf{H}_y^{m,n} \doteq \begin{bmatrix} \mathbf{y}_0 & \mathbf{y}_1 \cdots & \mathbf{y}_m \\ \mathbf{y}_1 & \mathbf{y}_2 \cdots & \mathbf{y}_m \\ \vdots & \vdots \ddots & \vdots \\ \mathbf{y}_n & \mathbf{y}_{n+1} \cdots & \mathbf{y}_{m+n-1} \end{bmatrix}$

In the sequel the indexes m, n may be omitted when clear from the context.

B. The problem of moments

Next, we recall results from polynomial optimization required to recast the Wiener identification problem into a tractable convex optimization form. Let $K \subset \mathbb{R}^{n_x}$ be a compact semi-algebraic set defined by a collection of polynomial inequalities of the form $g_k(\mathbf{x}) \geq 0, k = 1, \cdots, d$, that is,

$$K = \{\mathbf{x} \,|\, \mathbf{x} \in \mathbb{R}^{n_x}\, g_k(\mathbf{x}) \geq 0,\, k = 1, \cdots, d\} \tag{1}$$

and consider the problem of minimizing a multivariate polynomial $p = \sum_\alpha p_{\boldsymbol{\alpha}} x^{\boldsymbol{\alpha}}$ over the set K. As shown in [12], this problem is equivalent to $\min_\mu \mathcal{E}(p)$ where $\mathcal{E}$ denotes expectation and μ denotes the set of all Borel measures supported in K, or equivalently

$$\begin{aligned} p^* \quad &= \min_{\mathbf{m}} \textstyle\sum_\alpha p_{\boldsymbol{\alpha}} m_{\boldsymbol{\alpha}} \\ &\text{subject to } \exists \mu, \text{ supported in } K \text{ such that} \\ &m_{\boldsymbol{\alpha}} = \mathcal{E}_\mu(x^{\boldsymbol{\alpha}}) \end{aligned} \tag{2}$$

In turn, [8] existence of such a representing measure μ is equivalent to positive semidefiniteness of the (infinite) moment $\mathbf{M}(\mathbf{m})$ and localizing $\mathbf{L}(g_k\mathbf{m})$ matrices, from where it follows that an equivalent convex, albeit infinite dimensional reformulation of (2) is given by:

$$\begin{aligned} p^* = \quad \min_{\mathbf{m}} \quad & \textstyle\sum_\alpha p_{\boldsymbol{\alpha}} m_{\boldsymbol{\alpha}} \\ \text{s.t.} \quad & \mathbf{M}(\mathbf{m}) \succeq 0, \\ & \mathbf{L}(g_k\mathbf{m}) \succeq 0, k = 1, \ldots, d, \end{aligned} \tag{3}$$

A truncated version of this problem involving moments of order up to $2N$ is given by:

$$\begin{aligned} p_N^* = \quad \min_{\mathbf{m}} \quad & \textstyle\sum_\alpha p_{\boldsymbol{\alpha}} m_{\boldsymbol{\alpha}} \\ \text{s.t.} \quad & \mathbf{M}_N(\mathbf{m}) \succeq 0, \\ & \mathbf{L}_N(g_k\mathbf{m}) \succeq 0, k = 1, \ldots, d, \end{aligned} \tag{4}$$

$$\begin{aligned} \text{where } \mathbf{M}_N(\mathbf{m})(i,j) =& m_{\alpha^{(i)}+\alpha^{(j)}},\, \forall i,\, j = 1, \cdots, S_N \\ \mathbf{L}_N(g_k\mathbf{m})(i,j) =& \sum_\beta g_{k,\beta^{(l)}} m_{\beta^{(l)}+\alpha^{(i)}+\alpha^{(j)}}, \\ & \forall i,\, j = 1, \cdots, S_{N-\lfloor \frac{\delta_k}{2} \rfloor} \end{aligned} \tag{5}$$

and $S_N = \binom{N+n_x}{n_x}$(e.g. the number of moments in $\mathbb{R}_x^n$ up to order N). The main result of [12] shows that $p_N^* \uparrow p^*$, monotonically, thus providing a hierarchy of convergent relaxations.

C. Sparse polynomial optimization

In many cases of practical interest, both the polynomial objective and the constraints that define the set K exhibit a sparse structure that can be used to reduce the computational complexity entailed in solving the (truncated) problems (4). The following property plays a key role in exploiting this structure:

Definition 1: [11]: Assume that the polynomial p can be partitioned into $p = p_1 + \cdots + p_d$ such that each p_k and that the constraints g_k that define the set K contain only variables indexed by elements of some subset $I_k \subset \{1, \cdots, n\}$. If there exists a reordering $I_{k'}$ of I_k such that for every $k' = 1, \cdots, d-1$:

$$I_{k'+1} \cap \cup_{j=1}^{k'} I_j \subset I_s \quad \text{for some } s \leq k' \tag{6}$$

then the *running intersection property* is satisfied.

It can be shown that for problems that satisfy the *running intersection property*, it is possible to construct a hierarchy of semidefinite programs of smaller size. Specifically, partition the objective function $\{p_j\}_{j=1}^d$ according to the sets $\{I_k\}$ and consider the problem:

$$\begin{aligned} p_N^* = \quad \min_{\mathbf{m}} \quad & \textstyle\sum_{j=1}^l \sum_{\alpha(j)} p_{j,\boldsymbol{\alpha}(j)} m_{\boldsymbol{\alpha}(j)} \\ \text{s.t.} \quad & \mathbf{M}_N(\mathbf{m}_{I_k}) \succeq 0, k = 1, \ldots, d \\ & \mathbf{L}_N(g_k\mathbf{m}_{I_k}) \succeq 0, k = 1, \ldots, d \end{aligned} \tag{7}$$

where $p_{j,\boldsymbol{\alpha}(j)}$ is the coefficient of the $\boldsymbol{\alpha}(j)^{th}$ monomial in the polynomial p_j, $\mathbf{M}_N(\mathbf{m}_{I_k})$ denote the moment matrix and $\mathbf{L}_N(g_k\mathbf{m}_{I_k})$ the localizing matrix for the subset of variables in I_k. Then, as shown in [11] $p_N^* \uparrow p^*$. It is worth emphasizing that for the case of generic polynomials and constraints, an N^{th} order relaxation requires considering moments and localization matrices containing $O(n^{2N})$ variables. On the other hand, if the running intersection property holds, it is possible to define d sets of smaller sized

matrices each containing variables only in I_k (i.e. number of variables is $O(\kappa^{2N})$, where κ is the maximum cardinality of I_k). In many practical applications, including the one considered in this paper, $\kappa \ll n$. Hence, exploiting the sparse structure substantially reduces the number of variables in the optimization (and hence the computational complexity), while still providing convergent relaxations.

D. Problem Statement

Consider the interconnection shown in Figure 1, where $\mathbf{u} \in \mathbb{R}^{n_u}$ and $\mathbf{y} \in \mathbb{R}^{n_y}$ denote a known input/output pair and where the internal signal $\mathbf{x} \in \mathbb{R}^{n_x}$ is unmeasurable. Further, we will assume that the only a priori information available about the non-linearity is a bound on its gain and the fact that it is smooth (but not necessarily globally invertible) everywhere. In this context, the problem of interest in this paper can be precisely stated as follows:

Problem 1: Given:

1.- n_+ output sequences $\{\mathbf{y}_t^{i,+}\}$, $i = 1, \ldots, n_+$ generated by the system under consideration in response to a known, given input $\{\mathbf{u}_t\}$ and unknown initial conditions.

2.- n_- sequences $\{\mathbf{y}_t^{i,-}\}$, $i = 1, \ldots, n_-$ generated by systems known to be different from the one to be identified, in response to the same input $\{\mathbf{u}_t\}$.

3.- Some weak a priori assumptions on the nonlinearity $\Pi(.) \in \mathcal{NL}$.

Find n_x, the internal signal sequence $\{\mathbf{x}_j^{i,+}\}$, and the coefficients of an Auto-Regressive Model (ARX) of the form

$$a_o \mathbf{x}_t^{(i,+)} + \sum_{k=1}^{n_a} a_k \mathbf{x}_{t-k}^{(i,+)} + \sum_{k=1}^{n_b} \mathbf{B}_k \mathbf{u}_{t-k} = 0 \qquad (8)$$

such that $\mathbf{y}_t^{i,+} = \Pi(\mathbf{x}_t^{i,+})$ for some admissible nonlinearity $\Pi(.) : \mathbb{R}^{n_x} \to \mathbb{R}^{n_y}$, $\Pi(.) \in \mathcal{NL}$

In the sequel, we will solve this problem by recasting it first into a polynomial optimization form which in turn will be relaxed to a sequence of convex optimization problems. To this effect, we recall the following result relating the dimension of the ARX model to the rank of a Hankel matrix constructed from the input/output data:

Lemma 1: Consider the Hankel matrices $\mathbf{H_u}, \mathbf{H_x}$ associated with an input/output sequence $(\mathbf{u}, \mathbf{x})$ and let $\mathbf{H_u^\perp}$ denote the right annihilator of $\mathbf{H_u}$ (e.g. a matrix whose columns span the kernel of $\mathbf{H_u}$). Then the order of the minimal model of the form (8) that interpolates the data is $n_a = \mathrm{rank}(\mathbf{H_x H_u^\perp})$.

Proof: Follows from Theorem 2 in [17] ∎

Thus, Problem 1 can be solved by finding sequences $\{\mathbf{x}_t^{i,+}\}$ such that (i) $\mathbf{y}_t^{i,+} = \Pi(\mathbf{x}_t^{i,+})$, (ii) the corresponding Hankel matrices satisfy $\mathrm{rank}(\mathbf{H}_{\mathbf{x}^{i,+}} \mathbf{H_u^\perp}) \leq n_a$, and (iii) all the sequences $\{\mathbf{x}_t^{i,+}\}$ lie in the same n_a dimensional subspace (spanned by the solutions of (8)).

Lemma 2: Given sequences $\{\mathbf{x}_t^{i,+}, \mathbf{u}_t\}$ form the corresponding Hankel matrices $\mathbf{H}_{\mathbf{x}^{i,+}}$ and $\mathbf{H_u}$, with $n_a > \mathrm{rank}(\mathbf{H_u})$ columns. Assume that there exists a vector $\mathbf{w}$ such that

$$\mathbf{H}_{\mathbf{x}^{i,+}} \mathbf{H_u^\perp} \mathbf{w} = 0; \text{for } i = 1, \ldots, n_+ \qquad (9)$$

Finally, let $\mathbf{v} \doteq \mathbf{H_u^\perp} \mathbf{w}$. Then the sequences $\{\mathbf{x}_t^{i,+}\}$ lie in the n_a dimensional space orthogonal to $\mathbf{v}$.

Proof: Follows from the fact that, for each i, the corresponding sequence $\{\mathbf{x}_t^{i,+}\}$ satisfies

$$\begin{bmatrix} \mathbf{x}_k^{i,+} & \mathbf{x}_{k+1}^{i,+} & \ldots & \mathbf{x}_{k+n_a-1}^{i,+} \end{bmatrix} \mathbf{v} = 0, k = 1, \ldots$$

∎

From Lemmas 1 and 2 it follows that Problem 1 is equivalent to:

Problem 2: Find scalars n_x and n, vector sequences $\mathbf{x}^{i,+} \in \mathbb{R}^{n_x}$ and a vector $\mathbf{w} \in \mathbb{R}^n$ such that

$$\begin{array}{l} \|\mathbf{H}_{\mathbf{x}^{i,+}} \mathbf{H_u^\perp} \mathbf{w}\|_2^2 = 0 \; i = 1, \ldots, n_+ \\ \|\mathbf{H}_{\mathbf{x}^{i,-}} \mathbf{H_u^\perp} \mathbf{w}\|_2^2 > 0 \; i = 1, \ldots, n_- \\ \mathbf{y} = \Pi(\mathbf{x}) \text{for some } \Pi \in \mathcal{NL} \end{array} \qquad (10)$$

where the inequality above guarantees separation between the system to be identified and those known to be different from it, thus avoiding trivial solutions. Once suitable sequences $\mathbf{x}^{i,+}$ have been found, the parameters of the linear model can be recovered through a subspace identification step, and the nonlinearity can be found for instance by interpolation or using non-linear regression methods.

A potential problem with the formulation above is that it is sensitive to noise. In addition, in many practical applications n_y is very large ($\mathcal{O}(10^3)$ or higher) and the corresponding n_x is not necessarily small, leading to both computational complexity and memory problems. To circumvent these difficulties, in the next section we present a practical algorithm, motivated in part by classification techniques used in the machine learning community.

III. Maximum Margin Hankel Classifiers

In this section we present an identification algorithm, based on the ideas outlined in the section above, but modified to handle both noise and large data sets. To this effect, we start by modifying (10) to:

$$\begin{array}{l} \|\mathbf{H}_{\mathbf{x}^{i,+}} \mathbf{H_u^\perp} \mathbf{w}\|_2^2 \leq \xi_i \; \xi_i > 0, i = 1, ..., n_+ \\ \|\mathbf{H}_{\mathbf{x}^{i,-}} \mathbf{H_u^\perp} \mathbf{w}\|_2^2 + \xi_i \geq 1, \; \xi_i > 0 \; i = 1, ..., n_- \\ \mathbf{y} = \Pi(\mathbf{x}) \text{for some } \Pi \in \mathcal{NL} \end{array} \qquad (11)$$

that is, we seek a vector that is close to the intersection of the null spaces of the Hankel matrices corresponding to the samples known to have been originated from the system to be identified (this is equivalent to adding a fitting error or noise term with norm ξ_i to the left hand side of (8)). At the same time, the right hand side in the inequality above imposes that this vector is far from the null spaces of the Hankel matrices corresponding to the negative examples[1]. Finally, inspired by the maximum margin classification ideas, we will seek to minimize a combination of the ℓ_2 norm of $\mathbf{w}$ and the ℓ_1 norm of the vector of fitting errors[2] leading to the following optimization problem:

[1]The right hand side can be set to one, without loss of generality, by simply scaling the vector $\mathbf{w}$.

[2]$\|\mathbf{w}\|_2$ is related to the margin between positive and negative classes, while minimizing the ℓ_1 norm of $\boldsymbol{\xi}$ allows for the existence of mislabeled samples.

Problem 3:

$$\begin{array}{ll}\min_{\mathbf{x},\mathbf{w},\xi_i\geq 0} & \frac{1}{2}||\mathbf{w}||_2^2 + C\sum \xi_i \\ \text{subject to} & \mathbf{w}^T(\mathbf{H}_\mathbf{u}^\perp)^T\mathbf{G}_i\mathbf{H}_\mathbf{u}^\perp\mathbf{w} \leq \xi_i \quad ,\forall \mathbf{G}_i \in \mathbf{G}_+ \\ & \mathbf{w}^T(\mathbf{H}_\mathbf{u}^\perp)^T\mathbf{G}_i\mathbf{H}_\mathbf{u}^\perp\mathbf{w} + \xi_i \geq 1 \quad ,\forall \mathbf{G}_i \in \mathbf{G}_- \\ & \mathbf{y}_i = \Pi(\mathbf{x}_i) \text{ some } \Pi(.) \in \mathcal{NL}\end{array} \tag{12}$$

where we have defined $\mathbf{G}_i \doteq \mathbf{H}_i^T\mathbf{H}_i$. Note that the problem above is not convex, even in cases where the last constraint is convex, due to the terms involving the products of the elements of $\mathbf{G}_i$ and the vector $\mathbf{w}$. Nevertheless, these constraints are third degree multivariate polynomials in the unknowns $\mathbf{G}, \mathbf{w}, \boldsymbol{\xi}$. Thus, if $\mathcal{NL}$ also admits a semi-algebraic description, then Problem 3 reduces to a polynomial optimization over a semi-algebraic set, which, in principle, can be solved using the techniques outlined in section II-B.

Remark 1: An important feature of the formulation above is the fact that the entries of the matrix $\mathbf{G}_i$ depend only on the inner products $\mathbf{x}_i^T\mathbf{x}_j$ rather than on the actual data $\mathbf{x}$, leading to substantial complexity reduction when n_x is not small, since the corresponding polynomials (and hence the associated moment matrices) involve far less variables.

A. Incorporating a-priori information on the nonlinearity

So far, we have not made any assumptions regarding the nonlinearity $\Pi(.)$. In the absence of information about its structure, it is reasonable to impose constraints on the (local) Lipschitz constant of $\Pi^{-1}(.)$, that is, impose that

$$\|\mathbf{x}_r - \mathbf{x}_s\|_2^2 \leq L_i\|\mathbf{y}_r - \mathbf{y}_s\|_2^2, \ \forall (r,s) \in \mathcal{N}_i \tag{13}$$

for some suitably defined neighborhoods $\mathcal{N}_i$. Defining the Kernel (or Gram) matrix by its submatrices

$$\mathbf{K}_{i,j} = \begin{bmatrix} \mathbf{x}_j^{(i)T}\mathbf{x}_j^{(i)} & \mathbf{x}_j^{(i)T}\mathbf{x}_{j+1}^{(i)} & \cdots & \mathbf{x}_j^{(i)T}\mathbf{x}_{j+c}^{(i)} \\ \mathbf{x}_{j+1}^{(i)T}\mathbf{x}_j^{(i)} & \mathbf{x}_{j+1}^{(i)T}\mathbf{x}_{j+1}^{(i)} & \cdots & \mathbf{x}_{j+1}^{(i)T}\mathbf{x}_{j+c}^{(i)} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_{j+c}^{(i)T}\mathbf{x}_j^{(i)} & \mathbf{x}_{j+c}^{(i)T}\mathbf{x}_{j+2}^{(i)} & \cdots & \mathbf{x}_{j+c}^{(i)T}\mathbf{x}_{j+c}^{(i)} \end{bmatrix}$$

where c is a design parameter, and noting that $\mathbf{G}_i = \mathbf{H}_i^T\mathbf{H}_i = \sum_{j=1}^{T-c+1}\mathbf{K}_{i,j}$, it follows that (12), with the additional constraint (13) can be rewritten as:

$$\begin{array}{l}\min_{\mathbf{K},\mathbf{w},\xi} \frac{1}{2}||\mathbf{w}||_2^2 + C\sum\xi_i - \lambda\,\mathrm{Tr}(\mathbf{K}) \\ \text{subject to: } \mathbf{w}^T(\mathbf{H}_\mathbf{u}^\perp)^T\mathbf{G}_i\mathbf{H}_\mathbf{u}^\perp\mathbf{w} \leq \xi_i \quad ,\forall \mathbf{G}_i \in \mathbf{G}_+ \\ \mathbf{w}^T(\mathbf{H}_\mathbf{u}^\perp)^T\mathbf{G}_i\mathbf{H}_\mathbf{u}^\perp\mathbf{w} + \xi_i \geq 1 \quad ,\forall \mathbf{G}_i \in \mathbf{G}_- \\ \mathbf{G}_i = \sum_{j=1}^{T-c+1}\mathbf{K}_{i,j} \qquad \sum k_{ij} = 0 \\ k_{ii} + k_{jj} - 2k_{ij} \leq L_i||\mathbf{y}_i - \mathbf{y}_j||^2, \forall (i,j) \in \mathcal{N}_i \\ \mathbf{K} \succeq 0, \ \xi_i \geq 0\end{array} \tag{14}$$

where the constraint $\sum k_{ij} = 0$ enforces translationally invariant embeddings and where the additional term $-\lambda\,\mathrm{Tr}(\mathbf{K})$ in the objective seeks to favor lower dimensional embeddings [23], [25][3]. As before, the problem above can be solved by using moments-based polynomial optimization techniques.

[3]The relative value of the penalty weights C and λ determines the trade-off between fidelity of the identified system to the experimental data and the dimension of the embedding manifold.

B. Reducing the computational complexity

While the approach outlined above works well for small size problems, its computational complexity grows quickly with the number of data points, since the number of free variables in $\mathbf{K}$ is (number of data points)^2. To circumvent this difficulty, motivated by the approach proposed in Weinberger et al. [22] we will parameterize $\mathbf{K}$ in terms of a much smaller matrix $\tilde{\mathbf{K}}$ as follows

$$\mathbf{K} = \mathbf{Q}\tilde{\mathbf{K}}\mathbf{Q}^T \tag{15}$$

where the fixed matrix $\mathbf{Q}$ captures the local geometry of the nonlinearity. Intuitively, the idea is to approximate the products $\mathbf{y}_i^T\mathbf{y}_j$ as combinations of the products of a relatively few number of "landmarks", encapsulated in $\mathbf{Q}$ and locally linearize $\Pi(.)$ so that the same description applies to the corresponding products $\mathbf{x}_i^T\mathbf{x}_j$. It can be shown (see [22] for details) that in this context an optimal choice of $\mathbf{Q}$, in the sense that it minimizes the ℓ_2 norm of the reconstruction error, can be obtained by selecting, for each measurement $\mathbf{y}_i$ up to k nearest neighbors and solving the following optimization problem:

$$\zeta(\mathbf{P}) = \sum_i ||\mathbf{y}_i - \sum_j P_{ij}\mathbf{y}_j|| \tag{16}$$

subject to $\sum_j P_{ij} = \mathbf{1}$ and $P_{ij} = 0$ if $\mathbf{y}_j$ is not a k nearest neighbor of $\mathbf{y}_i$. Assuming, by reordering points if necessary, that the landmarks correspond to the first n_ℓ points of the set $\{\mathbf{y}_i\}$, it can be shown that the optimal $\mathbf{Q}$ is given by

$$\mathbf{Q} = \begin{bmatrix} \mathbf{I} \\ -(\boldsymbol{\Phi}_{22})^{-1}\boldsymbol{\Phi}_{21} \end{bmatrix} \tag{17}$$

where $\boldsymbol{\Phi}_{ij}$ denotes the blocks of $\boldsymbol{\Phi} \doteq (\mathbf{I}-\mathbf{P})^T(\mathbf{I}-\mathbf{P})$, partitioned so that its top left $n_\ell \times n_\ell$ block corresponds to the landmarks. Replacing $\mathbf{K}$ in terms of $\tilde{\mathbf{K}}$ in (14) leads to a polynomial optimization problem with substantially fewer variables. Further computational complexity reduction can be achieved by rewriting the objective function and constraints in (14) in terms of $\mathbf{W} \doteq \mathbf{v}\mathbf{v}^T$, where $\mathbf{v} = \mathbf{H}_\mathbf{u}^\perp\mathbf{w}$. Using the fact that $\|\mathbf{v}\|_2 = \|\mathbf{w}\|_2$ and dropping the constraint $\mathrm{rank}(\mathbf{W}) = 1$ leads to:

$$\begin{array}{l}\min_{\tilde{\mathbf{K}},\mathbf{W},\xi} \frac{1}{2}\,\mathrm{Tr}(\mathbf{W}) + C\sum\xi_i - \lambda\,\mathrm{Tr}(\mathbf{Q}\tilde{\mathbf{K}}\mathbf{Q}^T) \\ \text{subject to } \mathrm{Tr}(\mathbf{W}\mathbf{G}_i) \leq \xi_i \quad ,\forall \mathbf{G}_i \in \mathbf{G}_+ \\ \mathrm{Tr}(\mathbf{W}\mathbf{G}_i) + \xi_i \geq 1 \quad ,\forall \mathbf{G}_i \in \mathbf{G}_- \\ \mathbf{W} \succeq 0 \qquad \sum k_{ij} = 0 \\ k_{ii} + k_{jj} - 2k_{ij} \leq (1+\epsilon)||\mathbf{y}_i - \mathbf{y}_j||^2, \forall \eta_{ij} = 1 \\ \tilde{\mathbf{K}} \succeq 0, \ \xi_i \geq 0\end{array} \tag{18}$$

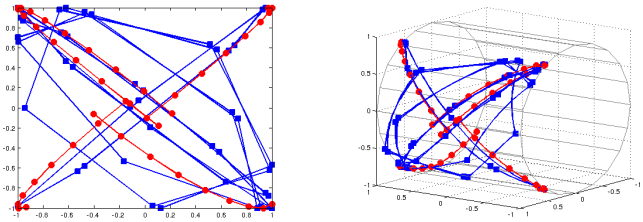
When compared against (14), (18) involves only second order polynomials, rather than third, (except for the constraint $\mathbf{W} \succeq 0$). In principle, this last constraint requires enforcing non-negativity of polynomials of degree up to n_w^2. However, in the case of an atomic measure μ with a single point mass (corresponding to rank 1 moment matrices), then $\mathbf{W} \succeq 0 \iff \mathcal{E}_\mu(\mathbf{W}) \succeq 0$. It follows that in this case a moments based relaxation of (18) is given by the algorithm outlined below, using a reweighed heuristics to seek for

solutions corresponding to rank 1 moment matrices.

Algorithm III-B: Wiener Id via Hankel Classifiers

1) **Input:** C, λ, ϵ, $n_{neighbor}$ number of spatial neighbors and n_ℓ number of landmarks.
2) **Initialization:** Compute $\mathbf{Q}$, set $\mathbf{M}_N^0 = \mathbf{I}$, $\delta^{(0)} = 0$, $k = 0$.
3) **Solve:**
   $\min_{\mathbf{M}_N^{(k+1)}} \ell(\mathbf{m}) + \mathrm{Tr}((\mathbf{M}_N^{(k)} + \delta^{(k)}\mathbf{I})^{-1}\mathbf{M}_N^{(k+1)})$
   subject to
   $\mathbf{M}_N(\mathbf{m}) \succeq 0$, $\mathbf{L}_N(\mathbf{m}) \succeq 0$, $\mathcal{E}_\mu(\mathbf{W}) \succeq 0$
   where $\ell(\mathbf{m})$ is the linear functional of the moments corresponding to the objective in (18), $\mathbf{M}_N(\mathbf{m})$ is the truncated moment matrix containing moments of order up to $2N$, and $\mathbf{L}_N(\mathbf{m})$ is the moments localizing matrix corresponding to the constraints.
4) Terminate if $\mathrm{rank}(\mathbf{M_N}) = 1$ or reach the maximum iteration. Otherwise, $\delta^{(k+1)} = \sigma_2(\mathbf{M_N})$ and repeat step 3).

Remark 2: Note that the polynomial cost function in (18) can be partitioned as $\sum_{l=1}^{n_+ + n_-} p_l$, where $p_l = \frac{1}{n_+ + n_-}\mathrm{Tr}(\mathbf{W}) + \xi_i - \lambda\,\mathrm{Tr}(\mathbf{K}_l)$ and $\mathbf{K}_l$ is the kernel matrix corresponding to the sequence $\mathbf{x}^{(l)}$. Since the corresponding constraints g_l depend only on the variables $\{\mathbf{W}, \tilde{\mathbf{K}}, \xi_l\}$, it follows that the running intersection property is satisfied. Thus we can replace the overall moment matrix with c smaller ones that depend only on $\{\mathbf{W}, \tilde{\mathbf{K}}, \xi_l\}$, significantly reducing the computational complexity in cases when n_y is large (i.e. video sequences).



(a) (b)

Fig. 2: (a) 2D sequences (Red: positive examples; Blue: negative examples) (b) Sequences lifted to 3D.

IV. Illustrative Example

In this section we illustrate the advantages of the proposed method using a simple example with synthetic data. The experimental data consists of twenty time series in $\mathbb{R}^2$ generated using two second order linear models with random initial conditions, and lifted to $\mathbb{R}^3$ by wrapping them on a cylinder surface. The length of all the sequences is 9. Ten positive samples were generated using AR coefficients $\mathbf{r}_+ = [1\ -0.6418\ 1]$, corresponding to a marginally stable system with poles at $p_{1,2}^+ = 0.321 \pm j0.947$ and ten negative samples were generated with $\mathbf{r}_- = [1\ \ -1.95\ 1]$(corresponding to a marginally stable system with poles at $p_{1,2}^- = 0.975 \pm j0.222$). Sample sequences in $\mathbb{R}^2$ are shown in Figure 2(a), where the red and blue trajectories belong to the positive and negative class, respectively. The corresponding lifted sequences in $\mathbb{R}^3$ are shown in Figure 2(b). All the lifted sequences were corrupted with noise uniformly distributed in $[-\sigma\nu, \sigma\nu]$, where ν is the range of the signal and $\sigma = 0.1$. Using this data we ran two sets of experiments. The first set evaluates the performance of both the identification algorithm and of the matrix $\mathbf{W} = \mathbf{v}\mathbf{v}^T$ as a classifier while the second one explores the benefit of including negative examples in the identification process.

Experiment 1: Identification using positive and negative samples. The generated data was divided into a training set with five positive and five negative examples, and a testing set with the remaining sequences. The algorithm III-B was run using the labeled training sequences to estimate the Grammian matrix $\mathbf{K}$ and a rank 1 classifier matrix $\mathbf{W}$, using parameter values $\lambda = 0.1$ and $C = 10^2$ and Hankel matrices with five columns. This led to a 4^{th} order AR model with coefficient $\mathbf{r}_{id} = [0.4490\ \ 0.4448\ \ 0.4464\ \ 0.4461\ \ 0.4497]$, corresponding to a marginally stable system with two poles at $p_{id} = 0.312 \pm j0.951$ and the remaining two at p_{id}^2. Note that two of these poles are very close to the ground truth, while the other two corresponding to the system that explains the ground truth data subsampled by a factor of 2.

Experiment 1 continued: Classification of new samples. The advantages of using the identified AR model as a classifier are illustrated in Table I. In order to classify new, previously unseen samples, the kernel matrix $\mathbf{K}$ learned from the training data was used to embed the new sequences as outlined in section III-B and predict their labels. The results of this process are shown in the columns in Table I. In all cases, the samples are assigned the correct label by using $\mathrm{sgn}(0.5 - \mathrm{Tr}(\mathbf{W}\mathbf{G}_i))$, even in the presence of noise.

TABLE I: Results of running Algorithm III-B with training data consisting of 5 sequences with positive labels and 5 sequences with negative labels.

	$\lambda = 0.1, C = 1 \times 10^2$			
σ	Train $\mathrm{Tr}(\mathbf{W}\mathbf{G}_i)$		Test $\mathrm{Tr}(\mathbf{W}\mathbf{G}_i)$	
	$+$	$-$	$+$	$-$
0.1	0.0031	1.1093	0.0860	1.1671
	0.0005	1.0067	0.0459	2.9623
	0.0005	1.0653	0.0574	0.9321
	0.0009	1.8465	0.1227	1.1849
	0.0017	1.3177	0.0373	1.1137
Mean	0.0013	1.2691	0.0699	1.4720
STD	0.0010	0.3072	0.0311	0.7505

Experiment 2: Benefits of Using Negative Sequences during Training. In this set of experiments, only one labeled example from the negative class was included during training[4]. In

[4]The other four negative training examples are only considered for the isometric constraints, i.e. they do not contribute with misclassification constraints.

TABLE II: Results of running Algorithm III-B with training data consisting of 5 sequences with positive labels, 1 sequence with negative label and 4 sequences with unknown labels (marked with $\diamondsuit$). Sequences that are incorrectly classified are marked with *.

	$\lambda = 0.1, C = 1 \times 10^2$					
σ	Train $+$	Tr($\mathbf{WG}_i$) $-$		Test $+$	Tr($\mathbf{WG}_i$) $-$	
0.1	0.0064	0.2038	$\diamondsuit*$	0.0327	0.2377	$*$
	0.0057	0.1393	$\diamondsuit*$	0.0253	1.3471	
	0.0038	0.1888	$\diamondsuit*$	0.0559	0.1379	$*$
	0.0020	1.1394	$\diamondsuit$	0.0617	0.1383	$*$
	0.0023	1.0000		0.0520	0.1393	$*$
$Mean$	0.0040	0.5343		0.0455	0.4000	
STD	0.0018	0.4399		0.0141	0.4751	

this case, the algorithm III-B led to an AR model with coefficients $\mathbf{r}_{id} = [0.3901\ 0.4558\ 0.4817\ 0.4607\ 0.4424]$ corresponding to an unstable system with poles at $p_{1,2} = 0.264 \pm j0.9833$ and $p_{3,4} = -0.848 \pm j0.612$, which, does not match well the ground truth. Finally, Table II shows that in this case the associated classifier incorrectly labels seven of the negative examples as belonging to the positive class.

V. Conclusion

In this paper we introduced an algorithm for non-parametric identification of Wiener systems for cases when there is no a-priori available information on the dimension of the output of the linear dynamics. The proposed framework combines ideas from system identification and maximum margin classifiers using kernel methods to exploit both positive and negative examples. While the proposed framework is based on challenging non-convex optimization problems, tractable convex relaxations (and in some cases exact solutions) can be found by exploiting recent developments in polynomial optimization and matrix rank minimization. Furthermore, since the resulting algorithm is based on identifying kernels, it can easily handle high dimensional data such as video. The benefits of the proposed algorithm were illustrated with a simple example using synthetic data. In particular, the experiments showed that including negative examples in the training data significantly improves the quality of the identified model and its accuracy, when used to classify time series.

References

[1] E. W. Bai. A blind approach to the hammerstein-wiener model identification. *Automatica*, 38:967–979, 2002.

[2] E. W. Bai. Frequency domain identification of wiener models. *Automatica*, 39:1521–1530, 2003.

[3] E.W. Bai. An optimal two-stage identification algorithm for hammerstein-weiner nonlinear systems. *Automatica*, 34:333–338, 1998.

[4] P. Celka and P. Colditz. Nonlinear nonstationary wiener model of infant eeg seizures. *IEEE Transactions on Biomedical Engineering*, 49:556–564, 2002.

[5] V. Cerone and D. Regruto. Parameter bounds evaluation of wiener models with noninvertible polynomial nonlinearities. *Automatica*, 42:1775–1781, 2006.

[6] C.T. Chou, B.R.J. Haverkamp, and M. Verhaegen. Linear and nonlinear system identification using separable least squares. *Eur. J. Control*, pages 116–128, 1999.

[7] S. C. Cripps. *RF Power Amplifiers for Wireless Communications*. Artech House, Norwood, MA, 1999.

[8] Raul E. Curto and Lawrence A. Fialkow. Truncated k-moment problems in several variables. *Journal of Operator Theory*, 54(1):189–226, 2005.

[9] W. Greblicki. Nonparametric identification of wiener systems. *IEEE Transactions on Information Theory*, 38:1487–1493, 1992.

[10] W. Greblicki. Nonparametric approach to wiener system identification. *IEEE Transactions on Circuits and Systems*, 44:538–545, 1997.

[11] J.-B. Lasserre. Convergent SDP-relaxations in polynomial optimization with sparsity. *SIAM J. on Optimization*, 17(3):822–843, October 2006.

[12] J.B. Lasserre. Global optimization with polynomials and the problem of moments. *SIAM J. Optimization*, 11:796–817, 2001.

[13] K. Lia, J.X. Pengan, and E. W. Bai. A two-stage algorithm for identification of nonlinear dynamic systems. *Automatica*, 42(7):1189–1197, 2006.

[14] H. Lim, O. I. Camps, M. Sznaier, and V. Morariu. Dynamic appearance modelling for human tracking. In *IEEE Computer Vision and Pattern Recognition*, pages 751–757, 2006.

[15] R.S. Lin, C.B. Liu, M.H. Yang, N. Ahuja, and S. Levinson. Learning nonlinear manifolds from time series. In *ECCV*, volume LNCS 3952, pages 245–256. Springer-Verlag, 2006.

[16] A. Matveev, X. Hu, R. Frezza, and H. Rehbinder. Observers for systems with implicit output. *Automatic Control, IEEE Transactions on*, 45(1):168 –173, jan 2000.

[17] M. Moonen, B. De Moor, L. Vandenberghe, and J. Vandewalle. On- and off-line identification of linear state-space models. *International Journal of Control*, 49(1):219–232, 1989.

[18] R. Raich, T. Zhou, and M. Viberg. Subspace based approaches for wiener system identification. *IEEE Trans. Aut. Contr.*, 50:1629–1634, 2005.

[19] S. Roweis, L. Saul, and G.E. Hinton. Global coordination of local linear models. *Advances in Neural Information Processing Systems*, 14:889–896, 2001.

[20] Lawrence K. Saul and Sam T. Roweis. Think globally, fit locally: unsupervised learning of low dimensional manifolds. *Journal on Machine Learning Research*, 4:119–155, 2003.

[21] R. Toth, V. Laurain, W. X. Zheng, and K. Poolla. Model structure learning: A support vector machine approach for lpv linear-regression models. In *Proc. 2011 IEEE Conf. Dec. Control*, pages 3192–3197, 2011.

[22] Kilian Q Weinberger, Benjamin D Packer, and Lawrence K Saul. Nonlinear dimensionality reduction by semidefinite programming and kernel matrix factorization. In *Proceedings of the tenth international workshop on artificial intelligence and statistics*, pages 381–388, 2005.

[23] K.Q. Weinberger and L.K. Saul. Unsupervised learning of image manifolds by semidefinite programming. In *CVPR*, volume 2, pages 988 – 995, June 2004.

[24] D. Westwick and M. Verhaegen. Identifying mimo wiener systems using subspace model identification methods. *Signal Processing*, 52:235–258, 1996.

[25] F. Xiong, O. Camps, and M. Sznaier. Low order dynamics embedding for high dimensional time series,. In *2012 IEEE ICCV*, pages 2368–2374, 2012.

[26] B. Yilmaz, M. Ayazoglu, M. Sznaier, and C. Lagoa. Convex relaxations for robust identification of wiener systems and applications. In *Proc. 2011 IEEE Conf. Dec. Control*, pages 2812–2818, 2011.