2011 50th IEEE Conference on Decision and Control and
European Control Conference (CDC-ECC)
Orlando, FL, USA, December 12-15, 2011

A Simple Recursive Algorithm for Learning a Monotone Wiener System

Kristiaan Pelckmans, kp@it.uu.se, Liang Dai liang.dai@it.uu.se
Division of Systems and Control, Department of Information Technology
Uppsala University, Box 337, SE-751 05, Uppsala, Sweden

***Abstract*— This paper studies a recursive identification method (i.e. an adaptive filter, or online learning algorithm) - termed the RANKTRON - for learning a Monotone Wiener model from observed input-output pairs. Such a model consists of a sequence of an unknown Linear Time-Invariant (LTI) dynamic model, followed by an unknown monotone (in- or decreasing) static nonlinear function. The main contribution is the introduction of a technical argument which establish worst-case performance of the proposed algorithm. The same tool is then used to derive properties in case the Monotone Wiener assumption only holds approximatively, and to the case where the output nonlinearity is a quantization function. An application of the RANKTRON is reported for the identification of a 20e order LTI based on quantized observations, using a mere O(1000) samples.**

I. Introduction

Consider a monotone Wiener model as depicted in Fig. (1). Here H_0 denotes an LTI dynamical system, and $f_0 : \mathbb{R} \to \mathbb{R}$ denotes a monotonically in- or decreasing static nonlinear function. The term $e = \{e_t\}_t$ can represent arbitrary model residuals, e.g. due to measurement errors or model misspecification. Such models can represent quantization-, saturation or transformation effects (see e.g. [1]). Recursive identification (i.e. adaptive filtering or online learning) applied to this case tries to recover (or approximate) both H_0 and f_0 from incremental sets of observed input-output pairs. Such problem (where instead f_0 does not need to be monotone) was considered in [1], [2], [3], proposing an approximate Recursive Prediction Error Method (RPEM). The analysis in this series of paper extends the stochastic ODE analysis considered in [4]. It essentially exploits a list of stochastic conditions of the involved signals, and requires the assumption that a *true* model (i.e. the system) belongs to the studied model-class. A nonparametric approach (not assuming such a *true* model) and its corresponding analysis was proposed in [5]. The analysis here is essentially based on probabilistic concentration inequalities, exploiting stochastic assumptions on the input signals, see also [6]. One could find many more works in literature dealing with the problem of recursive Wiener estimation [7], but the gradient-free approach as presented here is not described before.

The algorithm in this paper - referred to as the RANKTRON algorithm - is inspired instead by the well-known PERCEPTRON learning rule, excelling both in simplicity and power, see e.g. [8] for a survey. The analysis adopts a deterministic (non-stochastic, worst-case) framework. The presented study of the RANKTRON owes directly to the mistake-bound of the PERCEPTRON, as given by Block and Novikov (see e.g. [9] for a contemporary formulation and citations). As will be indicated in Subsection II.E the algorithm is related to the PRANK algorithm given in [12].

This manuscript is organized as follows. Section II formalizes the problem setup, proposes the RANKTRON algorithm and states the theoretical guarantees. Section III illustrated the practical working and compares with existing techniques. Section IV concludes this paper.

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".

II. Recursive Identification

A. Monotone Wiener Systems and Models

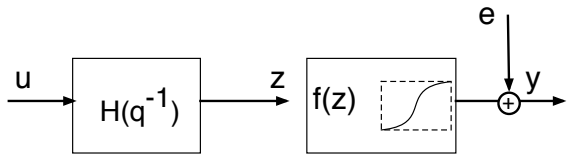

Fig. 1. *Representation of a Monotone Wiener systems with measurement noise. The function $f : \mathbb{R} \to \mathbb{R}$ is assumed to be monotonically in- or decreasing. Neither H nor f is assumed to be invertible.*

The following definitions fix the class of Monotone Wiener Systems under consideration.

Definition 1 (Wiener System) *A Wiener system (H_0, f_0) consists of a sequence of (i) a linear dynamical model $H_0(q^{-1})$ (here q^{-1} is the backshift operator as classically) applied to the* input *signal $\{u_t\}_t$, followed by (ii) a static nonlinear function $f_0 : \mathbb{R} \to \mathbb{R}$ (see Fig. 1). If the signals $\{u_t\}_t$ and $\{y_t\}_t$ follow such a system exactly, we can write*

$$y_t = f_0\Big(H_0(q^{-1})(u_t)\Big),\ \forall t, \tag{1}$$

and we say that the observations come from the Wiener system (H_0, f_0). If we merely approximate this system with (H, f) and some (small) slacks $\{e_t\}_t$ we write

$$y_t = f\Big(H(q^{-1})(u_t)\Big) + e_t,\ \forall t. \tag{2}$$

If H could be represented as a Finite Impulse Response (FIR) with $d > 0$ coefficients - or $H(q^{-1}) = h_1 + h_2 q^{-1} + \cdots + h_d q^{-d+1}$- such model is denoted (in the context of this paper) as the $(\mathbf{h}, f)$-Wiener model, or

$$y_t = f\left(\sum_{k=1}^{d} h_k u_{t-k+1}\right) + e_t = f(\mathbf{h}^T \mathbf{u}_t) + e_t,\ \forall t, \tag{3}$$

where $\mathbf{h} = (h_1, \ldots, h_d)^T \in \mathbb{R}^d$ *and we define* $\mathbf{u}_t = (u_t, u_{t-1}, \ldots, u_{t-d+1})^T \in \mathbb{R}^d$. *We will denote the set of possible observations as* $\mathcal{O} = \{(\mathbf{u}_t, y_t)\}_t \subseteq \mathbb{R}^d \times \mathbb{R}$.

Definition 2 (Lipschitz Monotone FIR Wiener System) *A FIR-Wiener system* $(\mathbf{h}_0, f_0)$ *is called* monotone *if* $f_0 : \mathbb{R} \to \mathbb{R}$ *is monotonically in- or decreasing (but not necessarily invertible), or*

$$(z - z')(f_0(z) - f_0(z')) \geq 0, \ \forall z, z' \in \mathbb{R}, \tag{4}$$

and $y_t = f_0(\mathbf{h}_0^T \mathbf{u}_t)$ *for all* t. *Moreover,* $(\mathbf{h}_0, f_0, L_0)$ *for* $L_0 < \infty$ *is a Lipschitz Monotone FIR Wiener System if additionally*

$$|y - y'| \leq L_0 \left|\mathbf{h}_0^T(\mathbf{u} - \mathbf{u}')\right|, \forall (\mathbf{u}, y), (\mathbf{u}', y') \in \mathcal{O}, \tag{5}$$

where $0 < \|\mathbf{h}_0\|_2 < \infty$.

Note that this definition gives properties of a dataset $\mathcal{O} = \{(\mathbf{u}_t, y_t)\}$, not necessarily of the system $(\mathbf{h}_*, f_*)$ underlying the data. Specifically, such system is identifiable from input-output behavior $\mathcal{O}$ only up to the (i) gain of the intermediate signals $\{z_t\}_t$, and (ii) the 'direction' of the nonlinearity (i.e. whether f_0 is monotonically in- or decreasing). Formally

Proposition 1 (Identifiability) *We consider the class of models consisting of Monotone FIR Wiener systems*

$$\begin{aligned}\mathbb{M}_L = \big\{(\mathbf{h}, f) : \ \mathbf{h} \in \mathbb{R}^d, \|\mathbf{h}\|_2 = 1, f : \mathbb{R} \to \mathbb{R}, \\ (y - y') \leq L(\mathbf{u} - \mathbf{u}')^T \mathbf{h}, \ \forall (\mathbf{u}, y), (\mathbf{u}', y') \in \mathcal{O}, y > y'\big\},\end{aligned} \tag{6}$$

where f *is monotonically* **in***creasing. Then* $\mathbb{M}_L$ *describes any Lipschitz monotone FIR Wiener system of order* $< d$.

B. The Basic RANKTRON Algorithm for Learning $\mathbf{h}_0$

In case one is only interested in recovering a parameter $\mathbf{h}$ from samples, the following recursion will work nicely. Let $\mathbf{h}_{(0)} = 0_m$, and

$$\mathbf{h}_{(t)} = \mathbf{h}_{(t-1)} + (y_t - y_{t'})(\mathbf{u}_t - \mathbf{u}_{t'}), \tag{7}$$

where $t' < t$ is defined such that

$$(y_t - y_{t'})(\mathbf{u}_t - \mathbf{u}_{t'})^T \mathbf{h}_{(t-1)} \leq 0. \tag{8}$$

Let $M_t \subset \{1, \ldots, t\}$ denote the indices s where a mistake against $s' < s$, and a corresponding update was made. This recursion leads to the RANKTRON algorithm, to which a more general version is spelled out in the next subsections. The naming 'RANKTRON' comes from the fact that a monotone function f_0 preserved the ordering or the 'ranking' of data, this being the essential reason as to why this simple algorithm comes with useful worst-case guarantees. The resulting algorithm will not make too large a cumulative mistakes as seen next

Lemma 1 (Mistake Bound) *Assume that* $\{(\mathbf{u}_t, y_t)\}_t$ *is such that there exists a* $\mathbf{h}_0 \in \mathbb{R}^d$ *with* $\|\mathbf{h}_0\|_2 = 1$ *and a monotone increasing* $f_0 : \mathbb{R} \to \mathbb{R}$ *which has Lipschitz constant* L_0, *such that* $y_t = f_0(\mathbf{h}_0^T \mathbf{u}_t)$ *for all* $t = 1, 2, \ldots$. *Then*

$$\sum_{s \in M_t} (y_s - y_{s'})^2 \leq 2L_0^2 r_u^2, \tag{9}$$

where $r_u > 0$ *is such that* $r_u > \max_{t=1,2,\ldots} \|\mathbf{u}_t\|_2$.

Proof: Unfolding the recursion gives

$$\mathbf{h}_{(t)} = \sum_{s \in M_t} (y_s - y_{s'})(\mathbf{u}_s - \mathbf{u}_{s'}). \tag{10}$$

The idea is to consider evolution of the quantity $\mathbf{h}_{(t)}^T \mathbf{h}_0$:

$$\mathbf{h}_{(t)}^T \mathbf{h}_0 = \sum_{s \in M_t} (y_s - y_{s'})(\mathbf{u}_s - \mathbf{u}_{s'})^T \mathbf{h}_0 \geq \frac{1}{L_0} \sum_{s \in M_t} (y_s - y_{s'})^2. \tag{11}$$

Conversely, from Cauchy-Schwarz' inequality we have that

$$\mathbf{h}_{(t)}^T \mathbf{h}_0 \leq \|\mathbf{h}_{(t)}\|_2 \|\mathbf{h}_0\|_2 \tag{12}$$

By construction we have that $\|\mathbf{h}_0\|_2 = 1$. Moreover we have in case no mistake was committed at iteration t that $\mathbf{h}_{(t)} = \mathbf{h}_{(t-1)}$. If a mistake was committed we have that

$$\begin{aligned}&\mathbf{h}_{(t)}^T \mathbf{h}_{(t)} = \\ &\left(\mathbf{h}_{(t-1)} + (y_t - y_{t'})(\mathbf{u}_t - \mathbf{u}_{t'})\right)^T \left(\mathbf{h}_{(t-1)} + (y_t - y_{t'})(\mathbf{u}_t - \mathbf{u}_{t'})\right) \\ &\leq \mathbf{h}_{(t-1)}^T \mathbf{h}_{(t-1)} + (y_t - y_{t'})^2 \|\mathbf{u}_t - \mathbf{u}_{t'}\|_2^2,\end{aligned} \tag{13}$$

and since $(y_t - y_{t'})(\mathbf{u}_t - \mathbf{u}_{t'})^T \mathbf{h}_{(t)} \leq 0$, and since $(\mathbf{u}_t, y_t)$ commits a mistake agains $(\mathbf{u}_{t'}, y_{t'})$ by construction. Hence

$$\|\mathbf{h}_{(t)}\|_2^2 \leq 2r_u^2 \sum_{s \in M_t} (y_s - y_{s'})^2. \tag{14}$$

Combining eq. (11), (12) and (14) yields the result. ∎

This basic reasoning is now extended to derive an algorithm which comes up with actual predictions.

C. The RANKTRON for learning $(\mathbf{h}_0, f_0)$

Let us now consider the slightly more involved case where both $\mathbf{h}_0$ and f_0 has to be estimated in order to make predictions for new samples $\mathbf{u}$. Although no (parametric) assumptions of f_0 will be required, the current approach will use (internally) the following representation of a monotonically increasing nonlinearity.

Definition 3 (Piecewise Linear Reconstruction of f_0**)** *If* $|\mathcal{R}_m| < 2$, *set* $f_{\mathcal{R}_m}(z) = 0$ *for all* $z \in \mathbb{R}$. *Given samples* $\mathcal{R}_m = \{(z_k, y^k)\}_{k=1}^m \subset \mathbb{R} \times \mathbb{R}$ *with* $|\mathcal{R}_m| \geq 2$, *then the piecewise* linear *function* $f_{\mathcal{R}_m}$ *interpolating these samples is defined as*

$$f_{\mathcal{R}_m}(z) = \frac{z - z_{\underline{z}(z)}}{z_{\overline{z}(z)} - z_{\underline{z}(z)}} \left(y_{\overline{z}(z)} - y_{\underline{z}(z)}\right) + y_{\underline{z}(z)}, \tag{15}$$

where we define $\overline{z}(z) = \arg\min_i (z_i \in \mathcal{R}_m : z_i \geq z)$ *and* $\underline{z}(z) = \arg\max_i (z_i \in \mathcal{R}_m : z_i \leq z)$. *In case either* $\overline{z}(z)$ *or* $\underline{z}(z)$ *is empty, define* $\overline{z}(z) = \arg\max_i (z_i \in \mathcal{R}_m)$ *or* $\underline{z}(z) = \arg\min_i (z_i \in \mathcal{R}_m)$ *respectively.*

Direct manipulation shows that this function is monotonically increasing and continuous. Moreover, if the samples

$\mathcal{R}_m = \{(z_k, y^k)\}_k \subset \mathbb{R} \times \mathbb{R}$ satisfy $|y - y'| \leq L_0|z - z'|$ for all $(z, y), (z', y') \in \mathcal{R}_m$, for a constant $L_0 < \infty$, then the function $f_{\mathcal{R}_m}$ is Lipschitz with constant L_0 as well. Let $\mathcal{R}_m = \{(z_k, y^k)\}_{k=1}^m$ be m reference points such that $z_k > z_l$ in case $k, l = 1, \ldots, m$ and $y^k \geq y^l$. This condition implies that the corresponding function $f_{\mathcal{R}_m}$ is monotone increasing. In the algorithm we will fix the values y^k such that they span the range of the function f of interest. To make this formal, assume there is a range $[\underline{f}, \overline{f}]$ such that all possible outcomes y_t fall into this interval. Then we will fix

$$y^k = \underline{f} + \left(\frac{k-1}{m-1}\right)(\underline{f} - \overline{f}), \tag{16}$$

such that $\min_{y^k \neq y^{k'}} |y^k - y^{k'}| = \rho$ where $\rho > 0$ equals $\rho = \frac{1}{m-1}(\underline{f} - \overline{f})$. The algorithm will then figure out the corresponding values of $\{z_k\}_k$ adaptively. We will drop the subscript '$_m$' and index the reference set by $_{(t)}$, with t denoting the iteration when it was computed. The resulting RANKTRON algorithm is spelled out in alg. (1). Here, we define $\underline{y}(y)$ and $\overline{y}(y)$ as

$$\begin{cases} \underline{y}(y) = \arg\max_{k=1,\ldots,m}\{y^k \leq y\} \\ \overline{y}(y) = \arg\min_{k=1,\ldots,m}\{y^k \geq y\}, \end{cases} \tag{17}$$

and $\underline{y}(y) = y^1$ or $\underline{y}(y) = y^m$ if the sets are empty.

Algorithm 1 The RANKTRON

Require: Let $m \geq 1$. Let $\mathcal{R}_{(0)} = \{(z_{(0),k}, y^k)\}_{k=1}^m$ with $z_{(0),k} = \frac{(\underline{f} - \overline{f})}{2}$ for all $k = 1, \ldots, m$. Let $\mathbf{h}_{(0)} = 0_d$.

for t=1,2,... **do**

(1) A prediction of y_t based on a vector $\mathbf{u}_t$ is computed as

$$\hat{y}_t = f_{\mathcal{R}_{(t-1)}}\left(\mathbf{u}_t^T \mathbf{h}_{(t-1)}\right). \tag{18}$$

(2) The corresponding loss can be computed as

$$\ell_t = (y_t - \hat{y}_t)^2. \tag{19}$$

(3) Based on this loss $\mathbf{h}_{(t-1)}$ and $\mathcal{R}_{(t-1)}$ are updated as follows. Let the indices $c_{t,k} \in [0, 1]$ be defined as

$$c_{t,k} = \begin{cases} 0 & k < \underline{y}(\hat{y}_t) \\ 0 & k > \overline{y}(\hat{y}_t) \\ \frac{y^{k+1} - \hat{y}_t}{y^{k+1} - y^k} & k = \underline{y}(\hat{y}_t) \\ \frac{\hat{y}_t - y^{k'-1}}{y^{k'} - y^{k'-1}} & k' = \overline{y}(\hat{y}_t). \end{cases} \tag{20}$$

Then

$$\begin{cases} \mathbf{h}_{(t)} = \mathbf{h}_{(t-1)} + (y_t - \hat{y}_t)\mathbf{u}_t \\ z_{(t),k} = z_{(t-1),k} - (y_t - \hat{y}_t)c_{t,k} \quad \forall k = 1, \ldots, m \end{cases} \tag{21}$$

and let $\mathcal{R}_{(t)} = \{(z_{(t),k}, y^k)\}_{k=1}^m$.

end for

Proposition 2 *By construction of the algorithm we have that for any $t = 1, 2, \ldots$ that*

$$\hat{y}_t = \sum_{k=1}^m c_{t,k} y^k \tag{22}$$

and that

$$\hat{y}_t = f_{\mathcal{R}_0}\left(\sum_{k=1}^m c_{t,k} z_k^0\right). \tag{23}$$

The performance of this algorithm is expressed in terms of the cumulative loss

$$L_n = \sum_{t=1}^n \ell_t = \sum_{t=1}^n (y_t - \hat{y}_t)^2. \tag{24}$$

In order to introduce ideas we will assume that there exists a 'true' $\mathbf{h}_0 \in \mathbb{R}^d$ and a set $\mathcal{R}_0 = \{(z_k^0, y^k)\}_k$ such that one has for all $t = 1, 2, \ldots$ that

$$f_{\mathcal{R}_0}(\mathbf{u}_t^T \mathbf{h}_0) = y_t. \tag{25}$$

In other words, here we assume that the data satisfies exactly a noiseless monotone FIR Wiener system with nonlinearity which can be expressed as a piecewise linear function. In the next subsection we will relax this stringent assumption. The corresponding Lipschitz constant is given as

$$\frac{1}{L_0} = \max_{k \neq k'} \frac{z_k^0 - z_{k'}^0}{y^k - y^{k'}} \tag{26}$$

Hence, we have that for all $z > z'$ that

$$f_{\mathcal{R}_0}(z) - f_{\mathcal{R}_0}(z') \leq L_0(z - z'). \tag{27}$$

Theorem 1 (Mistake Bound of the RANKTRON) *Given a dataset $\{(\mathbf{u}_t, y_t)\}_t$ such that there exists a $\mathbf{h}_0 \in \mathbb{R}^d$ and a monotone increasing $f_{\mathcal{R}_0}$ such that $y_t = f_{\mathcal{R}_0}(\mathbf{u}_t^T \mathbf{h}_0)$ for all $t = 1, 2, \ldots$. Assume that $f_{\mathcal{R}_0}$ has $L_0 > 0$ as in (27). For any t we have that*

$$\sum_{s=1}^t (y_s - \hat{y}_s)^2 \leq L_0^2 \left(1 + \sum_{k=1}^m (z_k^0)^2\right)(r_{\mathbf{u}}^2 + m), \tag{28}$$

where $r_u > 0$ is such that $r_u > \max_{t=1,2,\ldots} \|\mathbf{u}_t\|_2$.

This is a surprising result as the bound does not depend on the number of samples t which are examined. It means that if seeing n samples, the worst-case average mistake is $O(1/n)$, implying linear convergence. The bound however is strongly dependent on the number of piecewise linearities (knots) m, and degrades in $O(d^2)$ where d is the FIR model order.

Proof: The key idea to the proof is to encode the samples as follows. Let $\tilde{\mathbf{u}}_t \in \mathbb{R}^{d+m}$ be defined as

$$\tilde{\mathbf{u}}_t = (\mathbf{u}_t^T, 0, \ldots, 0)^T \in \mathbb{R}^{d+m}. \tag{29}$$

Let us define a representation of the actual prediction as

$$\tilde{\mathbf{v}}_t = (0_d^T, c_{t,1}, \ldots, c_{t,m})^T \in \mathbb{R}^{d+m}. \tag{30}$$

Equivalently we extend the vector $\mathbf{h}_{(t)}$ as $\tilde{\mathbf{h}}_{(t)} \in \mathbb{R}^{d+m}$ as

$$\tilde{\mathbf{h}}_{(t)} = \left(\mathbf{h}_{(t)}^T, z_{(t),1}, \ldots, z_{(t),m}\right)^T \in \mathbb{R}^{d+m}. \tag{31}$$

Then we have that

$$\tilde{\mathbf{u}}_t^T \tilde{\mathbf{h}}_{(t-1)} = \mathbf{u}_t^T \mathbf{h}_{(t-1)}, \quad \tilde{\mathbf{v}}_t^T \tilde{\mathbf{h}}_{(t-1)} = \sum_{k=1}^{m} c_{t,k} z_{(t-1),k}, \quad (32)$$

and that

$$\tilde{\mathbf{u}}_t^T \tilde{\mathbf{h}}_0 = \mathbf{u}_t^T \mathbf{h}_0, \quad \tilde{\mathbf{v}}_t^T \tilde{\mathbf{h}}_0 = \sum_{k=1}^{m} c_{t,k} z_k^0. \quad (33)$$

Then we have that $(\tilde{\mathbf{u}}_t - \tilde{\mathbf{v}}_t)^T \tilde{\mathbf{h}}_{(t-1)} = 0$ and that $(\tilde{\mathbf{u}}_t - \tilde{\mathbf{v}}_t)^T \tilde{\mathbf{h}}_0 \propto (y_t - \hat{y}_t)$. Here we have used the fact that $f_{\mathcal{R}_0}$ is monotonically increasing with Lipschitz constant L_0 . The idea of working with extended vectors was successfully employed in [9] to derive similar guarantees for a extension of the PERCEPTRON towards an ordinal regression context. Unfolding the recursion of the algorithm gives that

$$\tilde{\mathbf{h}}_{(t)} = \sum_{s=1}^{t} (y_s - \hat{y}_s)(\tilde{\mathbf{u}}_s - \tilde{\mathbf{v}}_s), \quad (34)$$

Hence we have by definition of the minimal Lipschitz property of $f_{\mathcal{R}_0}$ that

$$\tilde{\mathbf{h}}_0^T \tilde{\mathbf{h}}_{(t)} \geq \frac{1}{L_0} \sum_{s=1}^{t} (y_s - \hat{y}_s)^2. \quad (35)$$

Conversely, we have the following inequality

$$\tilde{\mathbf{h}}_0^T \tilde{\mathbf{h}}_{(t)} \leq \|\tilde{\mathbf{h}}_0\|_2 \|\tilde{\mathbf{h}}_{(t)}\|_2, \quad (36)$$

by application of Cauchy-Schwarz' inequality. Furthermore it is not too hard to bound the contribution of the two terms on the righthand side of the inequality. At first we have that

$$\tilde{\mathbf{h}}_0^T \tilde{\mathbf{h}}_0 = 1 + \sum_{k=1}^{m} (z_k^0)^2. \quad (37)$$

Secondly, the definition of the recursion in eq. (34) gives

$$\begin{aligned}
\tilde{\mathbf{h}}_{(t)}^T \tilde{\mathbf{h}}_{(t)} &= \Big(\tilde{\mathbf{h}}_{(t-1)} + (y_t - \hat{y}_t)(\tilde{\mathbf{u}}_t - \tilde{\mathbf{v}}_t)\Big)^T \\
&\quad \Big(\tilde{\mathbf{h}}_{(t-1)} + (y_t - \hat{y}_t)(\tilde{\mathbf{u}}_t - \tilde{\mathbf{v}}_t)\Big) \\
&= \tilde{\mathbf{h}}_{(t-1)}^T \tilde{\mathbf{h}}_{(t-1)} + (y_t - \hat{y}_t)^2 \left(\mathbf{u}_t^T \mathbf{u}_t + \sum_{k=1}^{m} c_{t,k}^2\right) \\
&\quad + 2(y_t - \hat{y}_t)(\tilde{\mathbf{u}}_t - \tilde{\mathbf{v}}_t)^T \tilde{\mathbf{h}}_{(t-1)} \\
&= \tilde{\mathbf{h}}_{(t-1)}^T \tilde{\mathbf{h}}_{(t-1)} + (y_t - \hat{y}_t)^2 \left(\mathbf{u}_t^T \mathbf{u}_t + \sum_{k=1}^{m} c_{t,k}^2\right), \quad (38)
\end{aligned}$$

where the last equality follows as by construction the term $(y_t - \hat{y}_t)(\tilde{\mathbf{u}}_t - \tilde{\mathbf{v}}_t)^T \tilde{\mathbf{h}}_{(t-1)} = 0$ as indicated in eq. (32). Unfolding the recursion gives that

$$\tilde{\mathbf{h}}_{(t)}^T \tilde{\mathbf{h}}_{(t)} \leq \big(r_{\mathbf{u}}^2 + m\big) \sum_{s=1}^{t} (y_s - \hat{y}_s)^2. \quad (39)$$

Combining the inequalities eq. (35), (36), (37) and eq. (39) gives the result. ■

D. Regret Bound for Individual Sequences

Let us now see what happens if the Monotone FIR Wiener model only holds approximatively. That is, consider the data $\mathcal{O} = \{(\mathbf{u}_t, y_t)\}_t$ where $y_t = f_{\mathcal{R}_0}(\mathbf{u}_t^T \mathbf{h}_0) + e_t$ with some small terms e_t. We will see that the performance guarantee only degrades gracefully in terms of the size of the errors. That is, the better the data follows a monotone Wiener model, the better the performance of the RANKTRON applied to this data. The mistake bound derived previously in the no-noise case is given now as

Corollary 1 (Regret Bound) *Given $\delta \geq 0, \mathbf{h}_* \in \mathbb{R}^d$ and a monotonic increasing function $f_{\mathcal{R}_*} : \mathbb{R} \to \mathbb{R}$ such that*

$$\big|y_t - f_{\mathcal{R}_*}(\mathbf{u}_t^T \mathbf{h}_*)\big| \leq \delta, \forall t = 1, 2, \ldots, \quad (40)$$

and $f_{\mathcal{R}_*}$ is Lipschitz with constant $L_* > 0$. Then

$$\sum_{s=1}^{t} (y_s - \hat{y}_s)^2 \leq \left(\delta\sqrt{t} + \sqrt{\left(1 + \sum_{k=1}^{m} (z_k^0)^2\right)\sqrt{(r_{\mathbf{u}}^2 + m)}}\right)^2. \quad (41)$$

This result is still reasonably tight in case $\delta \geq 0$ is small. *Proof:* The main idea is to introduce auxiliary variables $\bar{y}_t = f_{\mathcal{R}_*}(\mathbf{u}_t^T \mathbf{h}_*)$ for all $t = 1, 2, \ldots$, and to work with those as follows. First, we have using the definitions $\tilde{\mathbf{h}}_{(t)}, \tilde{\mathbf{h}}_*, \tilde{\mathbf{u}}_t, \tilde{\mathbf{v}}_t$ as before that

$$\tilde{\mathbf{h}}_{(t)} = \sum_{s=1}^{t} (y_s - \hat{y}_s)(\tilde{\mathbf{u}}_s - \tilde{\mathbf{v}}_s), \quad (42)$$

Hence we have by definition of the minimal Lipschitz property of $f_{\mathcal{R}_*}$ that

$$\begin{aligned}
\tilde{\mathbf{h}}_0^T \tilde{\mathbf{h}}_{(t)} &\geq \frac{1}{L_*} \sum_{s=1}^{t} (\bar{y}_s - \hat{y}_s)^2 - \frac{\delta}{L_*} \sum_{s=1}^{t} |y_s - \hat{y}_s| \\
&\geq \frac{1}{L_*} \sum_{s=1}^{t} (\bar{y}_s - \hat{y}_s)^2 - \frac{\delta\sqrt{t}}{L_*} \sqrt{\sum_{s=1}^{t} (y_s - \hat{y}_s)^2}, \quad (43)
\end{aligned}$$

since $(y_s - \hat{y}_s)(\bar{y}_s - \hat{y}_s) = (y_s - \hat{y}_s)(y_s - y_s + \bar{y}_s - \hat{y}_s) \geq (y_s - \hat{y}_s)^2 - \delta|y_s - \hat{y}_s|$, and by using the inequality relating the 1-norm and the 2-norm. Conversely, we have the following inequality

$$\tilde{\mathbf{h}}_0^T \tilde{\mathbf{h}}_{(t)} \leq \|\tilde{\mathbf{h}}_0\|_2 \|\tilde{\mathbf{h}}_{(t)}\|_2, \quad (44)$$

and by construction we have that $\tilde{\mathbf{h}}_0^T \tilde{\mathbf{h}}_0 = 1 + \sum_{k=1}^{m} (z_k^0)^2$. As before, we also have that

$$\tilde{\mathbf{h}}_{(t)}^T \tilde{\mathbf{h}}_{(t)} \leq \big(r_{\mathbf{u}}^2 + m\big) \sum_{s=1}^{t} (y_s - \hat{y}_s)^2. \quad (45)$$

Then combining eq. (43) and (45) we obtain

$$\begin{aligned}
\sum_{s=1}^{t} (\bar{y}_s - \hat{y}_s)^2 &\leq \delta\sqrt{t} \sqrt{\sum_{s=1}^{t} (y_s - \hat{y}_s)^2} \\
&+ \sqrt{\left(1 + \sum_{k=1}^{m} (z_k^0)^2\right)\sqrt{(r_{\mathbf{u}}^2 + m)}} \sqrt{\sum_{s=1}^{t} (y_s - \hat{y}_s)^2}. \quad (46)
\end{aligned}$$

Rearranging the different terms gives the result. ■

E. The PRANK algorithm for Learning from Quantized Outputs

Let us now consider the special case where the function f_0 is a quantization function, taking a finite (typically a small) $m \geq 2$ number of different values. Let $\{y^k\}_{k=1}^m$ be the set of those different values, such that $y^k < y^l$ if $k < l = 1, \ldots, m$. Then we use the piecewise constant function $q_{\mathcal{R}_m}$ defined as follows

Definition 4 (Piecewise Constant Function $q_{\mathcal{R}_m}$) *Given samples $\mathcal{R}_m = \{(z_k, y^k)\}_{k=1}^m \subset \mathbb{R} \times \mathbb{R}$, define the piecewise constant function $q_{\mathcal{R}_m}$ as*

$$q_{\mathcal{R}_m}(z) = y^{\underline{z}(z)}, \tag{47}$$

where we define as before $\underline{z}(z) = \arg\max_k (z_k \in \mathcal{R}_m : z_k \leq z)$. In case $\underline{z}(z)$ is empty, define $\underline{z}(z) = \arg\min_k (z_k \in \mathcal{R}_m)$.

Then The PRANK algorithm algorithm can be spelled out as in alg. (2).

Algorithm 2 The PRANK Algorithm

Require: Let $\mathcal{R}_{(0)} = \{(0, y^k)\}_{k=1}^m$ and let $\mathbf{h}_{(0)} = 0_d$.

for t=1,2,... **do**

(1) A prediction of y_t based on a vector $\mathbf{u}_t$ is computed as

$$\hat{y}_t = q_{\mathcal{R}_{(t-1)}}\left(\mathbf{u}_t^T \mathbf{h}_{(t-1)}\right). \tag{48}$$

(2) The corresponding loss can be computed as

$$\ell_t = I(y_t \neq \hat{y}_t), \tag{49}$$

with $I(z) = 1$ in case z holds true, and 0 otherwise.

(3) Based on this loss $\mathbf{h}_{(t-1)}$ and $q_{\mathcal{R}'_{(t-1)}}$ are updated as follows. Let the index $k(z)$ be defined such that $z_{(t-1),k(z)} < z_t < z_{(t-1),k(z)+1}$ where $z_t = \mathbf{u}_t^T \mathbf{h}_{(t-1)}$. Then

$$c_t = \begin{cases} \mathbf{e}_{k(z)+1} & \text{if } \; y_t > \hat{y}_t \\ \mathbf{e}_{k(z)} & \text{if } \; y_t < \hat{y}_t \\ 0_d & \text{else,} \end{cases} \tag{50}$$

where $\mathbf{e}_k = (0, \ldots, 1, \ldots, 0)^T \in \{0,1\}^m$ is the kth unit vector. Then if $y_t \neq \hat{y}_t$, let

$$\begin{cases} \mathbf{h}_{(t)} = \mathbf{h}_{(t-1)} + \operatorname{sign}(y_t - \hat{y}_t)\mathbf{u}_t \\ z_{(t),k} = z_{(t-1),k} - \operatorname{sign}(y_t - \hat{y}_t)c_{t,k}, \quad \forall k = 1, \ldots, m, \end{cases} \tag{51}$$

and let $\mathcal{R}_{(t)} = \{(z_{(t),k}, y^k)\}_{k=1}^m$. Else let $\mathbf{h}_{(t)} = \mathbf{h}_{(t-1)}$ and $\mathcal{R}_{(t)} = \mathcal{R}_{(t-1)}$.

end for

Then, the above result can be refined as follows

Corollary 2 (Mistake Bound of the PRANK algorithm) *Assume that there exists a $\rho > 0$, a quantization function $q_{\mathcal{R}_0} : \mathbb{R} \to \{y_1, \ldots y_m\}$ with m levels, as well as a vector $\mathbf{h}_0 \in \mathbb{R}^d$ with $\|\mathbf{h}_0\|_2 = 1$ such that*

$$\left|\mathbf{u}_t^T \mathbf{h}_0 - z_k^0\right| \geq \rho, \forall k = 1, \ldots, m, t = 1, 2, \ldots. \tag{52}$$

Then

$$|M_t| \leq \frac{\left(1 + \sum_{k=1}^m (z_k^0)^2\right)\left(1 + r_{\mathbf{u}}^2\right)}{4\rho^2}, \tag{53}$$

where $|M_t|$ denotes the number of mistakes $|M_t| = \sum_{s=1}^t I(y_s \neq \hat{y}_s)$ the algorithm has committed before or at iteration t.

Proof: The proof goes along the same lines as the one set out in the previous Theorem. After definition of $\tilde{\mathbf{u}}_t$, $\tilde{\mathbf{h}}_0$ and $\tilde{\mathbf{h}}_{(t)}$ we note that now we have by assumption that

$$\tilde{\mathbf{h}}_{(t)} = \sum_{s \in M_t} \operatorname{sign}(y_t - \hat{y}_t)(\tilde{\mathbf{u}}_t - \tilde{\mathbf{v}}_t) \tag{54}$$

Such that

$$\tilde{\mathbf{h}}_{(t)}^T \tilde{\mathbf{h}}_0 = \sum_{s \in M_t} \operatorname{sign}(y_t - \hat{y}_t)(\tilde{\mathbf{u}}_t - \tilde{\mathbf{v}}_t)^T \mathbf{h}_0 \geq 2\rho |M_t|. \tag{55}$$

And conversely $\tilde{\mathbf{h}}_0^T \tilde{\mathbf{h}}_{(t)} \leq \|\tilde{\mathbf{h}}_0\|_2 \|\tilde{\mathbf{h}}_{(t)}\|_2$,

$$\tilde{\mathbf{h}}_0^T \tilde{\mathbf{h}}_0 = 1 + \sum_{k=1}^m (z_k^0)^2, \tag{56}$$

and if a mistake was committed at iteration t one has

$$\begin{aligned} \mathbf{h}_{(t)}^T \mathbf{h}_{(t)} &= \mathbf{h}_{(t-1)}^T \mathbf{h}_{(t-1)} \\ &+ 2\operatorname{sign}(y_t - \hat{y}_t)(\tilde{\mathbf{u}}_t - \tilde{\mathbf{v}}_t)^T \tilde{\mathbf{h}}_{(t-1)} + r_{\mathbf{u}}^2 + \sum_{k=1}^m c_{t,k}^2 \\ &\leq \mathbf{h}_{(t-1)}^T \mathbf{h}_{(t-1)} + (r_{\mathbf{u}}^2 + 1), \end{aligned} \tag{57}$$

since one has $\operatorname{sign}(y_t - \hat{y}_t)(\tilde{\mathbf{u}}_t - \tilde{\mathbf{v}}_t)^T \tilde{\mathbf{h}}_{(t-1)} \leq 0$ by definition of $c_{t,k}$ in eq. (50). If no mistake was made $\mathbf{h}_{(t)} = \mathbf{h}_{(t-1)}$ and hence

$$\mathbf{h}_{(t)}^T \mathbf{h}_{(t)} \leq |M_t|(r_{\mathbf{u}}^2 + 1). \tag{58}$$

Combining the inequalities gives the result. ∎

Note that the PRANK algorithm as described here differs slightly from the one given in [12], e.g. in the definition of the loss and the precise update rule.

III. Applications

The working of the algorithm is illustrated in figures (2.a-f), where $y_t = \tanh(u_t + u_{t-1} + u_{t-2})$ for $t = 1, 2, \ldots$ (i.e. $d = 3$). From this simple example, we see that the RANKTRON can already give accurate estimates based on merely a few observations (compare with the examples reported in [3], [5]). A more challenging case is based on the following setup (as described in full detail in [11] where this setup was used for comparison of batch identification techniques of monotone Wiener systems). Consider a randomly generated LTI H_0 consisting of 20 conjugate poles, and 2 conjugate zeros. The nonlinear function is given as

$$f_0(z) = \operatorname{sign}(z + 0.5) + \operatorname{sign}(z - 2). \tag{59}$$

We found that a FIR approximation of $d = 200$ filter-coefficients captures the dynamics of H_0 sufficiently. Signals $\{u_s\}_{s=1}^t$ and $\{y_s\}_{s=1}^t$ are generated satisfying (H_0, f_0) for increasing lengths $t = 210, 230, 250, \ldots, 5000$. The performance of the RANKTRON is contrasted to the following

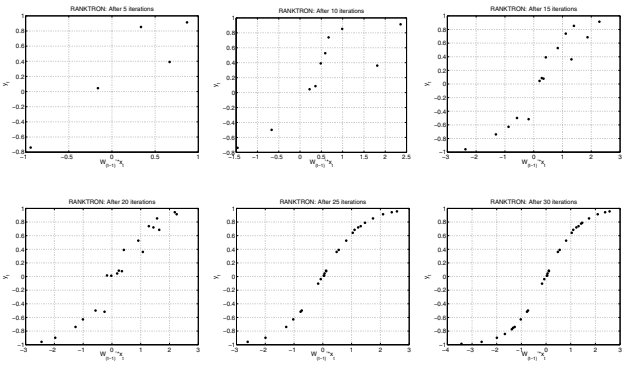

Fig. 2. *Example run of the RANKTRON as in eq. (7) for a dataset based on* $y_t = \tanh(u_t + u_{t-1} + u_{t-2})$. *The panels show respectively the predictions with the current hypothesis* $\mathbf{h}_t$ *after* $t = 7, 12, 17, 22, 27, 32$ *iterations. We see that after processing only 32 samples the RANKTRON gives already a good estimate:* $\mathbf{h}_{32} = (0.9975, 1.0304, 1.0106)^T$, *and the figures show the implicit approximation of* tanh.

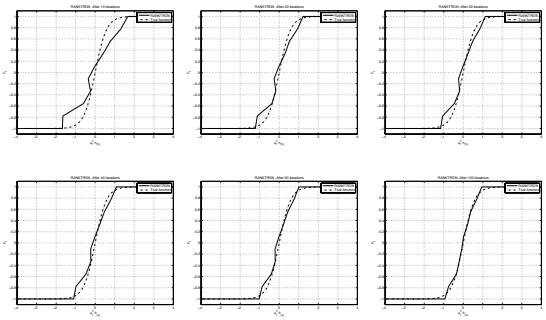

Fig. 3. *Example run of the RANKTRON as in Alg. 1 for the same dataset as before, but now using the piecewise linear reconstruction of the output nonlinearity. Here we parametrize* $f_{\mathcal{R}_m}$ *with* $m = 10$ *knots. Note that the price for such explicit reconstruction is that the algorithm needs twice as many iteration before the same accuracy of* $\mathbf{h}_{(n)}$ *compared to the experiment reported in Fig. 2 is obtained.*

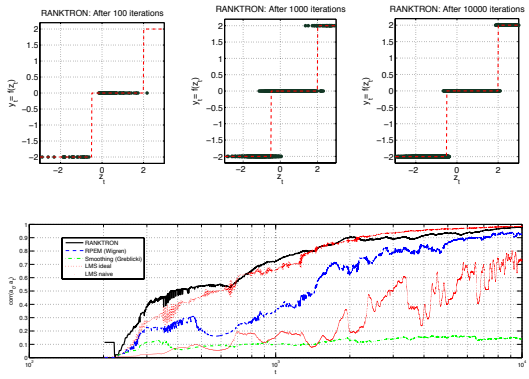

Fig. 4. *Example run on the monotone Wiener system* (H_0, f_0) *as in eq. (59), with a FIR approximation of* $d = 200$. *The first three panels show the estimate of the PRANK algorithm after* $t = 300, 1100, 11000$ *iterations. Panel (d) gives the performances obtained on this task using the different algorithms. Note that the RANKTRON outperforms an RPEM approach based on the given output nonlinearity. A value of* $corr(\mathbf{h}_0, \mathbf{h}_t) = 1$ *indicates a perfect fit. The 'naive' and 'ideal' LMS algorithms give a lower- and upper-bound of the performance of what is achievable with a gradient-based scheme.*

techniques: (1) A 'naive' LMS algorithm applied directly to the signals $\{u_s\}_s$ and $\{y_s\}_s$; (2) An 'ideal' LMS algorithm applied to the signals $\{u_s\}_s$ and $\{H_0(q^{-1})u_s\}_s$; (3) The RPEM method proposed in [1]. Here, a gradient descent algorithm was implemented based on a given smooth approximation of the output nonlinearity f_0. (4) the smoothing approach given in [5]. Those 4 algorithms were carefully tuned with respect to any design parameters (i.e. step parameter, f_0, smoothing factor) to give the best result on this dataset. The RANKTRON algorithm on the other hand was applied without any tuning. The performance is expressed in terms of the correlation between the estimated FIR coefficients and the 'true' impulse response used to generate the (noiseless) data. Note that all mentioned techniques are based on gradient information only.

IV. Conclusions

This paper studied the RANKTRON algorithm for recursive identification of monotone Wiener systems. Advantages are its simplicity and theoretical properties. Experiments indicate the usefulness of the algorithm compared to other (gradient-based) algorithms. Simulations of the different algorithms learns that when data is not so abundant, the simple RANKTRON algorithm performs much better than the version where an explicit function $f_{\mathcal{R}_m}$ approximating f_0 is learned as well. A connection to the PRANK algorithm - issued in the context of online machine learning - is made explicit. Comparison with batch identification algorithms for the same problem (see e.g. [11]) learns that there is much improvement that can be expected by incorporating second order information. Another open question is how one can modify the algorithm and the theoretical argument in case of adopting a parameterized model (IIR, ARMAX, ...).

References

[1] T. Wigren, "Recursive prediction error identification using the nonlinear Wiener model," *Automatica*, vol. 29, no. 4, pp. 1011–1025, 1993.

[2] ——, "Approximate gradients, convergence and positive realness in recursive identification of a class of non-linear systems," *International Journal of Adaptive Control and Signal Processing*, vol. 9, no. 4, 1995.

[3] ——, "Adaptive filtering using quantized output measurements," *IEEE transactions on signal processing*, vol. 46, no. 12, pp. 3423–3426, 1998.

[4] L. Ljung, "Analysis of recursive stochastic algorithms," *IEEE transactions on automatic control*, vol. 22, no. 4, pp. 551–575, 1977.

[5] W. Greblicki, "Recursive identification of Wiener systems," *Int. J. Appl. Math. Comp. Sci*, vol. 11, no. 4, pp. 977–991, 2001.

[6] W. Greblicki and M. Pawlak, *Non-Parametric System Identification*. Cambridge University Press, 2008.

[7] F. Giri, and E.W. Bai, *Block-oriented nonlinear system identification*, Springer, 2010.

[8] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[9] Y. Freund and R. Schapire, "Large margin classification using the perceptron algorithm," *Machine Learning*, vol. 37, no. 3, pp. 277–296, 1999.

[10] J. Kivinen, A. Smola, and R. Williamson, "Online learning with kernels," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2165–2176, 2004.

[11] K. Pelckmans, T. van Waterschoot, and J. Suykens, "Efficient adaptive filtering for smooth linear fir models," *EUSIPCO 2010, Aalborg, Danmark*, 2010.

[12] K. Crammer, Y. Singer, "Pranking with Ranking", *In Advances in Neural Information Processing Systems*, 14, 2001, pp. 64-647.