Difference algebra and system identification

Christian Lyzell, Torkel Glad, Martin Enqvist and Lennart Ljung

Linköping University Post Print

N.B.: When citing this work, cite the original article.

Difference Algebra and System Identification

Christian Lyzell, Torkel Glad, Martin Enqvist, Lennart Ljung

Division of Automatic Control, Department of Electrical Engineering, Linköping University, SE-58183, Sweden

Abstract

The framework of differential algebra, especially Ritt's algorithm, has turned out to be a useful tool when analyzing the identifiability of certain nonlinear continuous-time model structures. This framework provides conceptually interesting means to analyze complex nonlinear model structures via the much simpler linear regression models. One difficulty when working with continuous-time signals is dealing with white noise in nonlinear systems. In this paper, *difference* algebraic techniques, which mimic the differential algebraic techniques, are presented. Besides making it possible to analyze discrete-time model structures, this opens up the possibility of dealing with noise. Unfortunately, the corresponding discrete-time identifiability results are not as conclusive as in continuous time. In addition, an alternative elimination scheme to Ritt's algorithm will be formalized and the resulting algorithm is analyzed when applied to a special form of the NFIR model structure.

Key words: System identification, Identifiability, Ritt's algorithm.

1 Introduction

A trend in the optimization community is to try to find convex reformulations of initially non-convex optimization problems, either by relaxations or by algebraic manipulations, for which efficient and robust algorithms already exist.

The *prediction-error* approach (see, for instance, Ljung, 1999) to parameter estimation for time-invariant model structures often involves solving a non-convex optimization problem. Since there are usually many local optima, it may be difficult to guarantee that the global optimum will be found when applying a local search method.

A necessary condition for the uniqueness of the global optimum in a parameter estimation problem is that the chosen model structure is *globally identifiable*. In continuous time, Ljung and Glad (1994) were able to show that a nonlinear model structure, containing polynomial nonlinearities, is globally identifiable if and only if it can be rewritten as a linear regression model. The implications of this result for system identification are twofold. First of all, instead of analyzing the properties of a globally identifiable nonlinear model structure, one can work with an equivalent linear regres-

sion model, which significantly simplifies the analysis. Second, since an algorithm that performs the transformations of the nonlinear model to an equivalent linear regression model is provided, this also implies that the parameter estimation problem becomes trivial once the transformations have been made. The current paper concerns the discrete-time version of the above result.

The proof of the identifiability result in Ljung and Glad (1994) is based on differential algebraic techniques and, in particular, Ritt's algorithm. This theory was founded by Ritt (1950) and then further developed by Seidenberg (1956) and Kolchin (1973). An introduction to differential algebra from a control perspective is given in Fliess and Glad (1994), Fliess (1990) and further control related applications are presented in Fliess et al. (1995). Algorithmic aspects of Ritt's algorithm are discussed in Diop and Fliess (1991a), Diop and Fliess (1991b), Diop (1989), Diop (1991), Glad (1992) and Glad (1997). For nondifferential problems an alternative to Ritt's theory of characteristic sets is given by Gröbner bases, see, for instance, Buchberger (1976) and Becker and Weispfenning (1993). A comparison of the two approaches can be found in Forsman (1991, Chapter 2).

The goal of this paper is to generalize Ritt's algorithm to systems of difference equations. This problem has been addressed earlier by Kotsios (2001) by introducing algebraic operations on the δ-operator first ap-

Email addresses: lyzell@isy.liu.se (Christian Lyzell), torkel@isy.liu.se (Torkel Glad), maren@isy.liu.se (Martin Enqvist), ljung@isy.liu.se (Lennart Ljung).

pearing in Kotsios and Kalouptsidis (1993). However, this makes it difficult to compare the resulting algorithm with the continuous-time version presented in, for instance, Glad (1997). In this presentation, a more direct approach to the generalization of Ritt's algorithm to discrete time with as few alterations from the continuous-time version as possible will be given. Besides making it possible to analyze discrete-time model structures, this opens up the possibility of dealing with noise in a system identification setting.

Example 1. Consider the model structure defined by

$$y(t) = \theta u(t) + \theta^2 u(t-1) + e(t), \tag{1}$$

where $y(t)$ is the output, $u(t)$ is the input, and $e(t)$ is the noise, respectively. Now, consider the problem of estimating the unknown parameter θ in (1) by minimizing the *least-squares prediction-error*

$$V_N(\theta, Z^N) \triangleq \frac{1}{N} \sum_{t=1}^{N} \left(y(t) - \hat{y}(t|\theta) \right)^2, \tag{2}$$

where the *predictor* $\hat{y}(t|\theta)$ is given by

$$\hat{y}(t|\theta) \triangleq \theta u(t) + \theta^2 u(t-1).$$

A dataset $Z^N = (u(t), y(t))_{t=0}^{99}$ has been generated from (1) using the input $u(t) = \sin(t) + \sin(t/3)$, the parameter value $\theta = 0.9$ and the noise $e(t) = 0$. The solid line in Figure 1 illustrates the cost function (2) as a quartic polynomial in the parameter θ, which is clearly non-convex. Thus, to be able to find the global
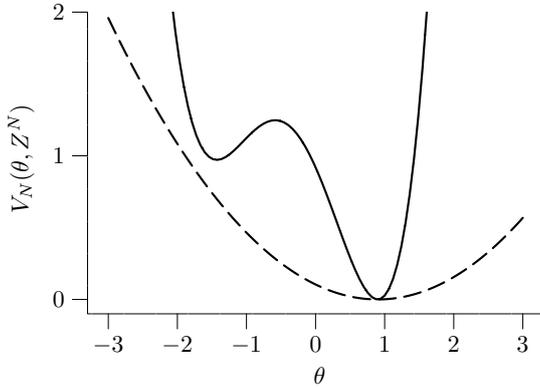

Fig. 1. The cost function (2) for the problem of estimating the parameter θ in (1) via minimizing the cost function (2). The solid line represents using (1) directly and the dashed line using the reformulated description (4).

optimum via a local search, one must have some a priori knowledge regarding the value of the parameter θ.

A common method to find an initial estimate of θ is to use *over-parametrization* (see, for instance, Bai, 1998).

This method is based on introducing $\tilde{\theta}_1 \triangleq \theta$ and $\tilde{\theta}_2 \triangleq \theta^2$, so that the model defined by (1) is linear in the new parameters, which may now be estimated using standard least-squares techniques. Since one is only interested in the estimate of $\tilde{\theta}_1$, one can just ignore the resulting estimate of $\tilde{\theta}_2$ or reconcile it with the estimate of $\tilde{\theta}_1$. Here, we propose a different technique.

Instead of introducing new parameters, let us, inspired by the differential algebra (see, for instance, Ljung and Glad, 1994), time shift (1):

$$y(t-1) = \theta u(t-1) + \theta^2 u(t-2) + e(t-1). \tag{3}$$

By examining the equations, we see that multiplying (1) by $u(t-2)$ and (3) by $u(t-1)$, and then subtracting the result, we obtain

$$
\begin{aligned}
u(t-2)y(t) &- u(t-1)y(t-1) \\
&= (u(t-2)u(t) - u^2(t-1))\theta \\
&\quad + u(t-2)e(t) - u(t-1)e(t-1),
\end{aligned} \tag{4}
$$

which is a linear regression model. Now, the least-squares cost function for this formulation, given the same noise-free dataset as above, is illustrated by the dashed line in Figure 1, which is a quadratic polynomial in the parameter θ with the minimum coinciding with the global minimum of the least-squares prediction-error cost function.

The discrete-time model structures that will be studied in this paper are those that can be written as a system of difference equations in the form

$$f_i\big(z(t), \dots, z(t-N), \theta\big) = 0, \quad i = 1, \dots, n, \tag{5}$$

where f_i are polynomials. The variable z contains several components

$$z(t)^{\mathrm{T}} = (y(t)^{\mathrm{T}}, u(t)^{\mathrm{T}}, e(t)^{\mathrm{T}}, x(t)^{\mathrm{T}}), \tag{6}$$

where y is the output, u is the input, e is the noise and x consists of internal variables.

This paper is a development of the material presented in Lyzell et al. (2009) and is outlined as follows: In Section 2, the basic algebraic framework of this paper is presented with the main result being the generalization of Ritt's algorithm to systems of difference equations. The ensuing Section 3 discusses the generalization of the identifiability results from Ljung and Glad (1994) to discrete time. Furthermore, the implications of noise in the parameter estimation problem for certain nonlinear model structures are investigated in Section 4. This section contains also a formulation

of an alternative approach to Ritt's algorithm for a special form of the *nonlinear finite impulse response* (NFIR) model structure and a numerical example is given. Finally, some conclusions are drawn in Section 5.

2 Algebraic concepts

In this section, we are interested in formalizing the algebra concerning solutions to systems of polynomial equations, where the polynomials depend only on a finite number of variables (which are themselves elements of one or several time-dependent signals), that is, the solution to systems of *difference equations*. For example, the polynomial $x(t)^3 + 1$ in the variable $x(t)$ is a function f of the sequence $X_t = (x(t-\tau))_{\tau=0}^\infty$ which maps X_t to the polynomial $X_{t,1}^3 + 1$ (where $X_{t,1}$ is the first element in the sequence X_t). The solution to the difference equation $f = 0$, that is $x(t)^3 + 1 = 0$, is thus a sequence $(x(\tau))_{\tau=-\infty}^\infty$ satisfying $f(X_t) = 0$ for all t. In general, the solution to a system of difference equations will be found by algebraic manipulations involving the backward shift operator q^{-1}, which applied to our example polynomial results in $q^{-1}f(X_t) = x(t-1)^3 + 1$. Thus, the time-shifted polynomial $q^{-1}f$ is a function of the second element $X_{t,2}$ of the sequence X_t in the same way as f is a function of $X_{t,1}$. Thus, the two polynomials $f(X_t)$ and $q^{-1}f(X_t)$ are considered to be different. For the sake of notational convenience, from here on the argument of the polynomials will be left out. Starting with the basics of difference algebra for time-dependent signals, we will move on to polynomials and finally an algorithm is presented for systems of difference polynomials.

2.1 Signal shifts

As discussed above, we are interested in systems described by polynomials in time-dependent variables and their time shifts. From here on, $\mathbb{N}$ is defined as the set of all nonnegative integers and $\mathbb{Z}_+$ as the set of all positive integers, respectively. The shifts will be denoted by

$$u^{(k)} \triangleq q^{-k}u(t) = u(t-k), \quad k \in \mathbb{N}, \tag{7}$$

where q is the forward time shift operator and the order of the displacement is given in the parenthesis [1]. To simplify the notation in the examples to come, we also introduce $\dot{u} \triangleq u^{(1)}$ and $\ddot{u} \triangleq u^{(2)}$. A fundamental concept for the algorithmic aspects of differential algebra is *ranking*. This is a *total ordering* (see, for instance, Lang, 2002) of all variables and their derivatives. In

---

[1] Alternatively, we may use shifts forwards in time $u^{(k)} \triangleq q^k u(t)$ for all $k \in \mathbb{N}$ and the same theory applies.

the discrete-time case it corresponds to a total ordering of all time-shifted variables.

Definition 2. A binary operator $\prec$, which is a total ordering satisfying

(i) $u^{(\mu)} \prec u^{(\mu+\sigma)}$
(ii) $u^{(\mu)} \prec y^{(\nu)} \implies u^{(\mu+\sigma)} \prec y^{(\nu+\sigma)}$,

for all $\mu \in \mathbb{N}$ and $\nu, \sigma \in \mathbb{Z}_+$, is called an *ordering of the signals* u and y and we say that u is *ranked* lower than y if $u \prec y$.

There are many possible choices of orderings of signals. For example, let u and y be two signals. Then two possible orderings are

$$u \prec y \prec \dot{u} \prec \dot{y} \prec \ddot{u} \prec \ddot{y} \prec \cdots \tag{8a}$$
$$u \prec \dot{u} \prec \ddot{u} \prec \cdots \prec y \prec \dot{y} \prec \ddot{y} \prec \cdots \tag{8b}$$

The latter ordering will often be written in short $u^{(\cdot)} \prec y^{(\cdot)}$. Let us turn our attention to polynomials with variables that are time-shifted signals. These polynomials will be used to represent difference equations as discussed in the introduction to this section.

2.2 Polynomials

As with signals, polynomials can also be time-shifted. In fact, the polynomial $f^{(\sigma)}$ is the result when all variables in the polynomial f have been shifted $\sigma \in \mathbb{N}$ time steps. Even though most of the results and definitions in this section apply to polynomials in static variables, the formulations will focus on polynomials in time-dependent variables.

To illustrate the algebraic concepts below in a simple manner, using the notation introduced in (7), let

$$f \triangleq \dot{u}y + \ddot{u}^3 \dot{y}^2, \tag{9a}$$
$$g \triangleq \dot{y}^2 + \ddot{y}, \tag{9b}$$
$$h \triangleq u + \dot{u}^2, \tag{9c}$$

be polynomials in the sequences $U_t = \left(u(t+\tau)\right)_{\tau=0}^\infty$ and $Y_t = \left(y(t+\tau)\right)_{\tau=0}^\infty$.

To be able to order polynomials we need to find the highest ranked variable in the polynomial.

Definition 3. The highest ranked time-shifted variable in a, possibly time-shifted, polynomial f is called the *leader* and is denoted by ℓ_f. The *degree* of a variable x in f is the highest exponent of x that occurs in f and is denoted by $\deg_x f$.

The polynomial (9a), with the ordering (8a), has the leader $\ell_f = \ddot{u}$ with $\deg_{\ddot{u}} f = 3$ and if the ordering is changed to (8b), the leader is given by $\ell_f = \dot{y}$ with $\deg_{\dot{y}} f = 2$. Thus, the leader depends on both the polynomial at hand and the ordering of the time-shifted signals that is used. The ranking of polynomials can now be defined.

Definition 4. Let f and g be two polynomials with leaders ℓ_f and ℓ_g, respectively. We say that f is *ranked lower* than g, denoted $f \prec g$, if either $\ell_f \prec \ell_g$ or if $\ell_f = \ell_g$ and $\deg_{\ell_f} f < \deg_{\ell_f} g$. If $\ell_f = \ell_g$ and $\deg_{\ell_f} f = \deg_{\ell_f} g$, we say that f and g have *equal ranking* and write $f \sim g$.

The polynomials f and g in (9), with the ordering (8a), have the leaders $\ell_f = \ddot{u}$ and $\ell_g = \ddot{y}$, respectively. Since $\ddot{u} \prec \ddot{y}$, according to (8a), it follows that $f \prec g$.

In this section, we will be dealing with the elimination of variables in time-shifted polynomials. In this context the following concept is important.

Definition 5. Let f be a polynomial with leader ℓ_f. A polynomial g is said to be *reduced* with respect to f, if there is no positive time shift of ℓ_f in g and if $\deg_{\ell_f} g < \deg_{\ell_f} f$.

Using the ordering (8b) with the polynomials f and g in (9), the leaders are given by $\ell_f = \dot{y}$ and $\ell_g = \ddot{y}$, respectively. Thus, in this case, f is reduced with respect to g but not vice versa. The above concepts (ranking and reduced) are related as follows.

Lemma 6. *Let f and g be two polynomials. If $f \prec g$ under some ordering, then f is also reduced with respect to g under that ordering.*

Proof. If $f \prec g$, then either $\ell_f \prec \ell_g$ or $\ell_f = \ell_g$ and $\deg_{\ell_f} f < \deg_{\ell_g} g$. In the former case, then it follows from Definition 2 that $\ell_f \prec \ell_g \prec \ell_g^{(\sigma)}$ for all $\sigma \in \mathbb{Z}_+$. Thus, f does not depend on $\ell_g^{(\sigma)}$ for any $\sigma \in \mathbb{N}$ and in particular $0 = \deg_{\ell_g} f < \deg_{\ell_g} g$. Hence, f must be reduced with respect to g. The latter case follows in a similar fashion. $\square$

That the two concepts are not equivalent is easily seen if one chooses the ordering (8a) with the simple polynomials $f = y$ and $g = u$. Since f does not depend on the variable u, it holds that f is reduced with respect to g, but $f \nprec g$ since $\ell_g \prec \ell_f$.

Before we continue providing the tool needed to reduce polynomials with respect to each other, some additional concepts of difference polynomials are needed.

These will not play a major role in what follows but are used to guarantee the existence of solutions to the resulting reduced sets of polynomials.

Definition 7. The *separant* S_f of a polynomial f is the partial derivative of f with respect to the leader, while the *initial* I_f is the coefficient of the highest power of the leader in f.

The polynomial f in (9a) have, under the ordering (8a), the leader $\ell_f = \ddot{u}$. This implies that the separant is given by $S_f = 3\ddot{u}^2 \dot{y}^2$ and the initial $I_f = \dot{y}^2$. The tool needed to reduce polynomials with respect to each other is a variant of the standard polynomial division.

Lemma 8. *Let f and g be polynomials in the variable x of degree m and n, respectively, written in the form*

$$f = a_m x^m + \cdots + a_0 \quad and \quad g = b_n x^n + \cdots + b_0,$$

where $m \geq n$. Then there exist polynomials $Q \neq 0$, $\bar{Q}$ and R such that

$$Qf = \bar{Q}g + R,$$

where $\deg_x R < n$. Furthermore, with Q given by b_n^{m-n+1}, then $\bar{Q}$ and R are unique.

Proof. See, for instance, Mishra (1993, page 168). $\square$

The suggested choice of Q in the lemma above is the initial I_g to the power of $m - n + 1$. It is also worth noting that the polynomials f and g in Lemma 8 may be multivariable, which implies that the coefficients of the variable x are polynomials in the remaining variables. Now, let us illustrate the concept of pseudo-division on the polynomials f and g in (9), where the variable in question is $\ddot{y}$. Since $I_g = 1$, it holds that

$$f = \ddot{u}^3 g + (-\ddot{u}^3 \ddot{y} + \dot{u}y),$$

so that $Q = 1$, $\bar{Q} = \ddot{u}^3$ and $R = -\ddot{u}^3 \ddot{y} + \dot{u}y$, which satisfies Lemma 8 if we notice that $\deg_{\ddot{y}} R = 0$. During the algebraic simplifications, it is important that the solutions to the original system of polynomial equations are preserved. To this end, the following results show how pseudo-division can be used to eliminate variables while preserving the solution.

Lemma 9. *Let g be a polynomial with leader ℓ_g and let f be a polynomial containing $\ell_g^{(\sigma)}$ for some $\sigma \in \mathbb{N}$. Then there exist polynomials R and Q such that*

(i) R does not contain $\ell_g^{(\sigma)}$.
(ii) Every solution of $f = 0$, $g = 0$ is also a solution of $R = 0$, $g = 0$.

4

(iii) Every solution of $R = 0$, $g = 0$ with $Q \neq 0$ is also a solution of $f = 0$, $g = 0$.

Proof. Let m and n be the degrees of f and $g^{(\sigma)}$ as polynomials in $\ell_g^{(\sigma)}$, respectively. On one hand, if $m \geq n$, then pseudo-division according to Lemma 8 yields polynomials $Q_1 \neq 0$, $\bar{Q}_1$ and R_1 such that

$$Q_1 f = \bar{Q}_1 g^{(\sigma)} + R_1, \tag{10}$$

with $\deg_{\ell_g^{(\sigma)}} R_1 < n \leq m$. If R_1 still depends on the variable $\ell_g^{(\sigma)}$, then further pseudo-division yields polynomials $Q_2 \neq 0$, $\bar{Q}_2$ and R_2 such that

$$Q_2 g^{(\sigma)} = \bar{Q}_2 R_1 + \tilde{R}_2, \tag{11}$$

with $\deg_{\ell_g^{(\sigma)}} \tilde{R}_2 < \deg_{\ell_g^{(\sigma)}} R_1$. Combining (11) and (10) yields

$$Q_1 \bar{Q}_2 f = (Q_2 + \bar{Q}_1 \bar{Q}_2) g^{(\sigma)} + R_2,$$

where we have defined $R_2 \triangleq -\tilde{R}_2$. Continuing in this manner, by recursive application of Lemma 8, one may always construct polynomials Q, $\bar{Q}$ and R such that

$$Q f = \bar{Q} g^{(\sigma)} + R, \tag{12}$$

with $\deg_{\ell_g^{(\sigma)}} R = 0$, that is, R does not depend on the variable $\ell_g^{(\sigma)}$. Now, assume that such polynomials have been constructed. Then the first statement in the lemma holds true. For the second statement, rewrite (12) as

$$R = Q f - \bar{Q} g^{(\sigma)}.$$

Since $g^{(\sigma)}$ is just g with all variables time shifted $\sigma \geq 0$ steps, it must hold that $g^{(\sigma)} = 0$ whenever $g = 0$. Thus, $R = 0$ whenever $f = 0$ and $g = 0$. For the final statement, rewrite (12), assuming $Q \neq 0$, as

$$f = \frac{1}{Q}\big(\bar{Q} g^{(\sigma)} + R\big).$$

Since $g = 0$ implies $g^{(\sigma)} = 0$, it follows that $f = 0$ whenever $R = 0$ and $g = 0$, which concludes the proof for the case when $m \geq n$. On the other hand, if $m < n$, then Lemma 8 yields polynomials $Q \neq 0$, $\bar{Q}$ and R such that

$$Q g^{(\sigma)} = \bar{Q} f + R$$

with $\deg_{\ell_g^{(\sigma)}} R < m$ and similar arguments as in the former case can be applied. $\square$

In the proof of the Lemma 9 above, we see that it is possible that several pseudo-divisions are needed

in the case $\sigma > 0$ to eliminate $\ell_g^{(\sigma)}$ from the polynomial f. This is the main difference between the continuous-time case and the discrete-time case. In the continuous-time case, $g^{(\sigma)}$ is affine in $\ell_g^{(\sigma)}$ (due to the product rule of differentiation) and thus only one pseudo-division is needed to eliminate $\ell_g^{(\sigma)}$ from f. In the discrete-time case, time shifting a polynomial does not change its degree with respect to a variable, that is, $\deg_{\ell_g} g = \deg_{\ell_g^{(\sigma)}} g^{(\sigma)}$ and several pseudo-divisions might be needed.

Now, the following extension of Lemma 9 provides (as we will see later) the main step in our algorithm.

Lemma 10. *Let f be a polynomial which is not reduced with respect to the polynomial g. Then there exist polynomials R and Q such that*

(i) R is reduced with respect to g.
(ii) Every solution of $f = 0$, $g = 0$ is also a solution of $R = 0$, $g = 0$.
(iii) Every solution of $R = 0$, $g = 0$ with $Q \neq 0$ is also a solution of $f = 0$, $g = 0$.

Proof. If f is not reduced with respect to g, then either f contains some positive time-shift of the leader ℓ_g of g or else f contains ℓ_g to a higher power than g.

In the latter case, Lemma 9 yields polynomials Q, $\bar{Q}$ and R such that
$$Q f = \bar{Q} g + R, \tag{13}$$
where R is reduced with respect to g. Thus, the proof of the statements follows in the same way as in proof of Lemma 9.

In the former case, the polynomial f contains $\ell_g^{(\sigma_1)}$ for some $\sigma_1 \in \mathbb{Z}_+$. Lemma 9 then yields polynomials Q_1, $\bar{Q}_1$ and R_1 satisfying

$$Q_1 f = \bar{Q}_1 g^{(\sigma_1)} + R_1, \tag{14}$$

where R_1 does not contain $\ell_g^{(\sigma_1)}$. If R_1 still contains $\ell_g^{(\sigma_2)}$ for some $\sigma_2 \in \mathbb{N}$ with $\sigma_2 < \sigma_1$, further use of Lemma 9 yields polynomials Q_2, $\bar{Q}_2$ and R_2 satisfying

$$Q_2 R_1 = \bar{Q}_2 g^{(\sigma_2)} + R_2, \tag{15}$$

where R_2 does not contain $\ell_g^{(\sigma_2)}$. Combining (14) and (15) yields

$$Q_2 Q_1 f = Q_2 \bar{Q}_1 g^{(\sigma_1)} + \bar{Q}_2 g^{(\sigma_2)} + R_2.$$

Continuing in this manner, by recursive application of Lemma 9, one can construct polynomials Q and R such

that

$$Qf = \sum_{i=1}^{n} \tilde{Q}_i g^{(\sigma_i)} + R, \tag{16}$$

for some sequence of polynomials $\tilde{Q}_i$, $i = 1, \ldots, n$, where R does not contain $\ell_g^{(\sigma)}$ for any $\sigma \in \mathbb{N}$. Thus, we have constructed a polynomial R that is reduced with respect to g. Since $g = 0$ implies that $g^{(\sigma)} = 0$ for all $\sigma \in \mathbb{N}$, the remaining statements follow in the same way as in the proof of Lemma 9. □

Lemma 10 shows how to reduce one difference equation with respect to another while preserving the solution of both and that the main tool for achieving this is the pseudo-division concept presented in Lemma 8.

Now, let us turn our attention to systems of difference equations.

2.3 Sets of polynomials

Now, we are ready to define the necessary concepts for systems of polynomial difference equations, which will be represented by sets of difference polynomials. The following definition generalizes the concept of reduced polynomials.

Definition 11. A set $\mathscr{A} = \{A_1, \ldots, A_p\}$ of polynomials is called *auto-reduced* if all elements A_i are pairwise reduced with respect to each other. If, in addition, the polynomials $A_1, \ldots, A_p$ in the auto-reduced set $\mathscr{A}$ are in increasing rank, then $\mathscr{A}$ is said to be *ordered*.

Let us once again consider the polynomials f and g in (9), but now with the ordering (8a). Then it is easy to see that f and g are reduced with respect to each other and the set $\{f, g\}$ is auto-reduced. On the other hand, using the ordering (8b), we have already seen that g is not reduced with respect to f and the set $\{f, g\}$ is not auto-reduced. The following definition generalizes the concept of ranking to auto-reduced sets.

Definition 12. Let $\mathscr{A} = \{A_1, \ldots, A_m\}$ and $\mathscr{B} = \{B_1, \ldots, B_n\}$ be two ordered auto-reduced sets. Then $\mathscr{A}$ is said to have a lower *ranking* than $\mathscr{B}$ if either there exists an integer k such that $1 \le k \le \min(m, n)$ satisfying

$$A_j \sim B_j, \quad j = 1, \ldots, k-1$$
$$A_k \prec B_k,$$

or else if $m > n$ and $A_j \sim B_j$ for all $j = 1, \ldots, n$.

The ordering (8a) applied to the polynomials f, g and h in (9) yields two ordered auto-reduced sets of polynomials with two elements, namely $\mathscr{A} \triangleq \{f, g\}$ and

$\mathscr{B} \triangleq \{h, g\}$. Since $h \prec f$, but not vice versa, it holds that $\mathscr{B}$ is ranked lower than $\mathscr{A}$. The concept of reduced polynomials can be generalized as follows.

Definition 13. A *characteristic set* for a given set of time-shifted polynomials is an auto-reduced subset such that no other auto-reduced subset is ranked lower.

Using the same example of ordered auto-reduced sets $\mathscr{A} \triangleq \{f, g\}$ and $\mathscr{B} \triangleq \{h, g\}$ chosen from the set $\{f, g, h\}$ of polynomials (9), it is clear that $\mathscr{B}$ is the characteristic set under the ordering (8a). The basic idea of our algorithm is to reduce a set of polynomials by the use of characteristic sets. In each iteration, the characteristic set is used to reduce the highest ranked polynomial not in the characteristic set to a lower ranked one. Thus, it is important to guarantee that a sequence with decreasing rank is always finite for the algorithm to terminate in a finite number of steps.

Lemma 14. *A sequence of auto-reduced sets, each one ranked lower than the preceding one, can only have finite length.*

In particular, the above statement includes sequences of characteristic sets and is a direct consequence of the following result.

Lemma 15. *A sequence of time-shifted variables from a finite number of signals, each one ranked lower than the preceding one, can only have finite length.*

Proof. Let $y_1, \ldots, y_p$ denote all the variables whose time shifts appear anywhere in the sequence. For each y_j let σ_j denote the order of the first appearing time shift. Then there can be only σ_j lower time-shifted y_j in the sequence and the total number of elements is thus bounded by $\sigma_1 + \cdots + \sigma_p + p$. □

Finally, we are ready to state the elimination algorithm for systems of difference equations in the form (5).

Algorithm 1. Reduces a set of polynomials to a characteristic set.

Input: Two sets of polynomials $\mathscr{F} = \{f_1, \ldots, f_n\}$ and $\mathscr{G} = \emptyset$ with an ordering.

Output: The updated set $\mathscr{F}$ as a characteristic set containing the reduced polynomials and the set $\mathscr{G}$ containing information about the separants, initials and the quotients resulting from the performed pseudo-divisions in the different steps.

1) Compute a characteristic set $\mathscr{A} = \{A_1, \ldots, A_p\}$ of $\mathscr{F}$: Order the polynomials in $\mathscr{F}$ according to their leaders, so that the first polynomial has the lowest ordered leader and initialize $\mathscr{A} \leftarrow \{f_1\}$. For all $f_i \in \mathscr{F}$,

$i = 2, \ldots, n$, test if $\mathscr{A} \cup \{f_i\}$ is auto-reduced and in that case update $\mathscr{A} \leftarrow \mathscr{A} \cup \{f_i\}$.

2) If $\mathscr{F} \setminus \mathscr{A} \neq \emptyset$, where $\setminus$ denotes the set difference, then go to Step 4.

3) Add S_A, I_A for all $A \in \mathscr{A}$ to $\mathscr{G}$ and **stop**.

4) Let f be the highest ranked unreduced polynomial in $\mathscr{F}$ with respect to $\mathscr{A}$ and apply Lemma 10 to get polynomials Q and R such that

$$Q f = \sum_{i=1}^{n} \tilde{Q}_i A^{(\sigma_i)} + R,$$

where A is the highest ordered polynomial in $\mathscr{A}$ such that f is not reduced with respect to A. Update $\mathscr{G} \leftarrow \mathscr{G} \cup \{Q\}$.

5) If $R = 0$ update $\mathscr{F} \leftarrow \mathscr{F} \setminus \{f_k\}$, else $\mathscr{F} \leftarrow (\mathscr{F} \setminus \{f_k\}) \cup \{R\}$ and continue from Step 1.

Algorithm 1 is quite similar to the Ritt-Seidenberg algorithm, see (Ritt, 1950) and (Seidenberg, 1956), and will be referred to as Ritt's algorithm from here on.

Remark 16. The elements added to the set $\mathscr{G}$ in Algorithm 1 are important to ensure the solvability of the resulting system of difference equations and should all be nonzero. The nonvanishing of the separant is used later to ensure that the requirements for the implicit function theorem (see, for instance, Rudin, 1976) are fulfilled.

Remark 17. For those familiar to the continuous-time version of Ritt's algorithm, it is worth noting that Algorithm 1 can be readily generalized to only consider *partially reducedness* in each iteration as described in Fliess and Glad (1994), which may simplify the result when applying Ritt's algorithm.

Some important properties of the proposed algorithm are given below. The proofs of the results are similar to corresponding proofs for the continuous-time case, but with the small distinction remarked upon in connection with Lemma 9.

Theorem 18. *The Algorithm 1 will reach the **stop** after a finite number of steps.*

Proof. The only possible loop is via Step 5 to Step 1. This involves either the removal of a polynomial or its replacement with one that is reduced with respect to $\mathscr{A}$ or has its highest unreduced time-shift removed. If R is reduced, then it is possible to construct a lower auto-reduced set. An infinite loop would thus contradict either Lemma 14 or Lemma 15. □

Theorem 19. *Every solution to the initial set $\mathscr{F}$ in Algorithm 1 is also a solution to the final set. Every solution*

to the final set for which the polynomials of $\mathscr{G}$ are nonzero is also a solution of the initial set.

Proof. The result follows from repeated application of Lemma 10. □

We conclude this section by repeating Example 1, but here the parameter nonlinearity is eliminated using Ritt's algorithm.

Example 20. Consider, once again, the model structure defined by

$$y(t) = \theta u(t) + \theta^2 u(t-1) + e(t), \tag{17}$$

where $y(t)$ is the output, $u(t)$ is the input, and $e(t)$ is the noise, respectively. Let us try to apply Algorithm 1 to reduce (17) to a linear regression model. To this end, it is necessary to express that the parameter θ in (17) is constant, since the difference algebraic framework developed above assumes, by construction, that all signals are time dependent. Therefore, define

$$f_1 \triangleq y - u\theta - \dot{u}\theta^2 - e, \tag{18}$$
$$f_2 \triangleq \theta - \dot{\theta}, \tag{19}$$

where the dot operator denotes the backward shift. Let $\mathscr{F} \triangleq \{f_1, f_2\}$ and $\mathscr{G} = \emptyset$, respectively. The ordering is chosen as

$$u^{(\cdot)} \prec y^{(\cdot)} \prec e^{(\cdot)} \prec \theta^{(\cdot)}.$$

Now, Ritt's algorithm will need two iterations to terminate, but here we will only illustrate the first iteration:

Step 1: The leaders ℓ_{f_1} and ℓ_{f_2} in $\mathscr{F}$ are θ and $\dot{\theta}$, respectively. Since $\ell_{f_1} \prec \ell_{f_2}$, f_1 is reduced with respect to f_2. The other way around, since the polynomial f_2 contains a positive time shift of the leader ℓ_{f_1}, namely $\dot{\ell}_{f_1} = \dot{\theta}$, it is not reduced with respect to f_1, that is, the characteristic set $\mathscr{A}$ is given by $\{f_1\}$.

Step 2: Since $\mathscr{F} \setminus \mathscr{A} = \{f_2\} \neq \emptyset$, we move to step 4.

Step 4: Now, we will make use of the scheme depicted in the proof of Lemma 10 to reduce f_2 with respect to f_1. The division algorithm yields $Q_1 f_2 = \bar{Q}_1 \dot{f}_1 + R_1$, with

$$R_1 = -\dot{y} + \dot{e} + \dot{u}\theta + \ddot{u}\dot{\theta}\theta,$$

$Q_1 = \dot{u} + \ddot{u}\dot{\theta}$, and $\bar{Q}_1 = 1$. Since R_1 still depends on $\dot{\theta}$, yet another division is needed, that is, $Q_2 f_2 = \bar{Q}_2 R_1 + R_2$, with

$$R_2 = \dot{y} - \dot{e} - \dot{u}\theta - \ddot{u}\theta^2,$$

$Q_2 = -\ddot{u}\theta$, and $\bar{Q}_2 = 1$. The polynomial R_2 is not reduced with respect to f_1 and further polynomial division yields $Q_3 R_2 = \bar{Q}_3 f_1 + R_3$, with

$$R_3 = -\ddot{u}y + \ddot{u}e + \dot{u}\dot{y} - \dot{u}\dot{e} + (\ddot{u}u - \dot{u}^2)\theta,$$

$Q_3 = \ddot{u}$, and $\bar{Q}_3 = -\dot{u}$. The polynomial R_3 is reduced with respect to f_1. Putting it all together yields

$$Q f_2 = \tilde{Q}_1 f + \tilde{Q}_2 \dot{f}_1 + R,$$

where $Q = Q_3 Q_2 - Q_1 \bar{Q}_2$, $\tilde{Q}_1 = \bar{Q}_3$, $\tilde{Q}_2 = -Q_3 \bar{Q}_1 \bar{Q}_2$, and $R = R_3$. Finally, update $\mathscr{G} \leftarrow \mathscr{G} \cup \{Q\}$.

Step 5: Update $\mathscr{F} \leftarrow \{f_1, R\}$.

Thus, after the first iteration of Algorithm 1 we have found a new first order polynomial R such that the solution to the system of difference equations corresponding to the polynomial set $\{f_1, R\}$ is the same as the one for the original system given by $\{f_1, f_2\}$. Furthermore, the polynomial R coincides with the polynomial defined by (4) in Example 1.

It is worth noting that in Step 4, two divisions had to be made to eliminate $\dot{\theta}$ from the polynomial f_2. This is one of the differences between the discrete-time and the continuous-time Ritt's algorithm. In continuous time, the derivative always leaves a polynomial which is affine in the leader and only one division is needed for each elimination.

In Ljung and Glad (1994), the differential algebraic tools were used to derive some interesting results on the identifiability of polynomial model structures. Since the discrete-time algorithm is so similar, one could expect to get the same results.

3 Identifiability

In this section, the focus will lie entirely on system identification aspects of the framework presented above. Therefore, in the following, it is assumed that all parameters are constant and that the initial polynomial set $\mathscr{F}$ contains, in addition to the polynomials describing the model structure, the polynomials describing this fact, see Example 20.

In Ljung and Glad (1994), necessary and sufficient conditions for global identifiability of continuous-time model structures were given. Here, we will discuss the generalization of these results to discrete-time model structures. First, let us recall the definition of identifiability.

Definition 21. Let (5) have the solution θ, $z^{\mathrm{T}} = (y^{\mathrm{T}}, u^{\mathrm{T}}, 0, x^{\mathrm{T}})$ and also the solution $\bar{\theta}$, $\bar{z}^{\mathrm{T}} = (\bar{y}^{\mathrm{T}}, \bar{u}^{\mathrm{T}}, 0, \bar{x}^{\mathrm{T}})$.

If $u = \bar{u}$, $y = \bar{y}$ implies that $\theta = \bar{\theta}$ for all solutions in a given solution set, then the model structure is said to be *identifiable* for that solution set.

Now, since the discrete-time version of Ritt's algorithm is so similar to the continuous-time case, should we not be able to derive equivalent results concerning the identifiability of discrete-time model structures? That is, is a discrete-time model structure, only containing polynomial nonlinearities, globally identifiable if and only if it can be written as a linear regression model?

The first part of this statement is, at this point, quite easy to prove. Namely, if Ritt's algorithm results in a linear regression model, then the corresponding model structure is globally identifiable.

Theorem 22. *If the output of Algorithm 1 contains an expression of the form $Q\theta - P$, where the diagonal matrix Q and the vector P does not depend on θ and x, then the model structure is globally identifiable, provided that $\det Q \neq 0$ for the measured data* [2].

Proof. Since, according to Theorem 19, every solution of the original equations is also a solution of the output equations, it follows that there can only be one value of θ that is consistent with the measured values, provided that $\det Q \neq 0$. $\qquad\square$

Now, let us consider the converse of the above statement. In the continuous-time case, the proof of this fact is based on the following property: If $f(t)$ and $g(t)$ are analytical functions satisfying $f(t)g(t) = 0$ for all $t \in \mathbb{R}$, it holds that either $f(t) = 0$ or $g(t) = 0$ for all $t \in \mathbb{R}$. Unfortunately, this property does not remain valid when the domain changes from $\mathbb{R}$ to $\mathbb{Z}$. A simple example of a discrete-time signal that does not satisfy the above property is the Kronecker delta function

$$\delta(t) \triangleq \begin{cases} 1, & t = 0 \\ 0, & t \in \mathbb{Z} \setminus \{0\} \end{cases}$$

For instance, it holds that $\delta(t)\delta(t-1) = 0$ for all $t \in \mathbb{Z}$ even though $\delta(t)$ and $\delta(t-1)$ are nonzero for $t = 0$ and $t = 1$, respectively. The absence of the above property for discrete-time signals hinders a straightforward generalization of the desired identifiability result. Reasonable assumptions to be able to prove the theorem are yet to be found.

Even though we were not able to provide a general identifiability result, equivalent to that achieved in the

---

[2] The requirement $\det Q \neq 0$ can be interpreted as a condition of *persistence of excitation* of the input signal (see, for instance, Ljung, 1999).

continuous-time case, we are still able to draw some conclusions when the use of Ritt's algorithm does not result in a linear regression model.

Theorem 23. *Let the ranking be given by*

$$u^{(\cdot)} \prec y^{(\cdot)} \prec \theta_1^{(\cdot)} \prec \cdots \prec \theta_m^{(\cdot)} \prec x^{(\cdot)}$$

where x contains any unmeasured variable. Assume that the output of Algorithm 1 is an auto-reduced set with the following form

$$p_0(u, \dot{u}, \ldots, y, \dot{y}, \ldots), \dot{\theta}_1 - \theta_1,$$
$$p_2(u, \dot{u}, \ldots, y, \ddot{y}, \ldots, \theta_1, \theta_2), \ldots$$
$$p_m(u, \dot{u}, \ldots, y, \ddot{y}, \ldots, \theta_1, \ldots, \theta_m), \ldots$$

Furthermore, assume that there exists a solution to this set such that all polynomials in $\mathscr{G}$ from Algorithm 1 are nonzero. Then there are infinitely many values of θ compatible with the u and y of this solution, that is, the system is unidentifiable.

Proof. Fixing the values of u and y in the first polynomial p_0 to those of the given solution means that the corresponding equation is always satisfied. The parameter θ_1 can now be changed to a new arbitrary constant in the second polynomial p_2. If this change is small enough the remaining equations can now be solved successively for the leader due to the nonvanishing of the separants (the *implicit function theorem*, see, for instance, Rudin, 1976). Now, Theorem 19 gives that the solutions corresponding to the polynomial set $\{p_0, \ldots, p_m\}$ are also solutions to the initial set $\mathscr{F}$. Since this initial set contains polynomials $\theta_i - \dot{\theta}_i$ for all $i = 1, \ldots, m$, this implies that the θ_i solutions will be constant. Hence, the result follows. □

The above theorem implies that, in this particular case, if Ritt's algorithm is not able to eliminate the time-shift for at least one of the parameters, not necessarily the first parameter, then the model structure is unidentifiable. Thus, even though we were not able to prove that every globally identifiable model structure may be rewritten as a linear regression model, this indicates that Ritt's algorithm still can be useful when analyzing the identifiability of discrete-time model structures. Ritt's algorithm has some interesting implications for the parameter estimation problem in system identification, since it, in some cases, results in a linear regression.

4 A pragmatic approach

When working with Ritt's algorithm and the framework developed in the sections above, one notices al-
most immediately that the complexity of the resulting linear regression model grows quite rapidly with the number of parameters involved in the original nonlinear model structure. This is especially true in discrete time.

Another drawback using Ritt's algorithm is illustrated in the example below.

Example 24. Consider the model structure defined by

$$y = \theta_1 u + \theta_1 \theta_2 \dot{u} + e, \tag{20}$$

where the dot, as usual, denotes the backward shift. This is a special case of a nonlinearly parametrized *finite impulse response* (FIR) model structure. Now, assume that one is interested in finding estimates of the parameters $\theta = \begin{pmatrix} \theta_1 & \theta_2 \end{pmatrix}^{\mathrm{T}}$ given some input and output data. One possibility would be to use Ritt's algorithm to find a linear regression model from which an estimate can be found via the least-squares method. For the resulting estimate to be unbiased, it is necessary that the regressor and the noise in the linear regression model are uncorrelated. Unfortunately, this is not generally the case for FIR models, like the one defined by (20). Running Algorithm 1 with the ordering

$$e^{(\cdot)} \prec u^{(\cdot)} \prec y^{(\cdot)} \prec \theta_1 \prec \theta_2 \prec \dot{\theta}_1 \prec \dot{\theta}_2,$$

yields a linear regression model which is too lengthy to present in its entirety and only the linear regression for the first parameter $z_1 = \varphi_1 \theta_1 + v_1$ is given:

$$\begin{aligned}
z_1 =\ & u^{(3)} \ddot{u} u \ddot{y} - \ddot{u}^2 \dot{u} \ddot{y} + 2u^{(3)} \ddot{u} \dot{u} \dot{y} - u^{(4)} \dot{u}^2 \dot{y} \\
& - (u^{(3)})^2 u \dot{y} - u^{(3)} \ddot{u}^2 y + u^{(4)} \ddot{u} \dot{u} y, \\
\varphi_1 =\ & 2u^{(3)} \ddot{u} \dot{u}^2 - \ddot{u}^3 \dot{u} - u^{(4)} \dot{u}^3 - (u^{(3)})^2 \dot{u} u + u^{(4)} \ddot{u} \dot{u} u \\
& + u^{(3)} \ddot{u} \dot{u} \ddot{y} - (u^{(3)})^2 u \dot{y} - u^{(3)} \ddot{u}^2 \dot{y} + u^{(4)} u^{(3)} u \dot{y} \\
& + (u^{(3)})^2 \ddot{u} y - u^{(4)} u^{(3)} \ddot{u} y, \\
v_1 =\ & eu^{(4)} \ddot{u} \dot{u} - eu^{(3)} \ddot{u}^2 + 2\dot{e} u^{(3)} \ddot{u} \dot{u} - \ddot{e} \ddot{u}^2 \dot{u} - \dot{e} u^{(4)} \dot{u}^2 \\
& - \dot{e} (u^{(3)})^2 u + \ddot{e} u^{(3)} \ddot{u} u.
\end{aligned}$$

Here we note that the regressors φ_1 are correlated with the noise signal v_1, since $\dot{y}$ has a direct dependence on $\dot{e}$. Thus, the least-squares estimate of the parameters θ may be biased. To find an unbiased estimate, one needs to make use of techniques like *instrumental variables* (see, for instance, Söderström and Stoica, 1983). It should be mentioned that the linear regression model given by Ritt's algorithm for the second parameter is even lengthier than the one for the first parameter presented above.

In the example above, we noticed that, even for the FIR model structure, one can get correlation between the

regressors and the noise in the resulting linear regression model when using Ritt's algorithm. With this in mind and the complexity of Ritt's algorithm, we will here formalize a more pragmatic approach to eliminate the parameter nonlinearities that might appear. The main idea is to use successive time shifts and pseudo-divisions as in Example 1.

Theorem 25. *Consider the model structure defined by*

$$y = \eta^{\mathrm{T}} f(u,q) + \sum_{i=1}^{m} h_i(\eta,\xi) g_i(u,q) + e, \qquad (21)$$

where y is the output, u is the input, and e is the noise, respectively. The functions f, $(h_i)_{i=1}^{m}$, and $(g_i)_{i=1}^{m}$ may have a nonlinear dependence on the indicated variables, where q denotes the forward shift operator. The unknown parameters $\theta^{\mathrm{T}} = \left(\eta^{\mathrm{T}} \ \xi^{\mathrm{T}} \right)$ are constant, where all parameters that do not appear in any linear term in (21) have been collected in ξ. Under the assumptions above, the model structure (21) can be written as a linear regression model

$$\mathcal{H}_m(u,q) y = \eta^{\mathrm{T}} \varphi_m(u,q) + \mathcal{H}_m(u,q) e, \qquad (22)$$

in the parameter vector η. Here, φ_m is a nonlinear vector-valued function in a finite number of components in u and $\mathcal{H}_m$ is a linear filter whose coefficients have a nonlinear dependence of the components in u, respectively. The components of (22) are defined by

$$\mathcal{H}_{j+1}(u,q) = \tilde{g}_{m-j,j}(\dot{u},q)\mathcal{H}_j(u,q) - \tilde{g}_{m-j,j}(u,q)\mathcal{H}_j(\dot{u},q)q^{-1},$$
$$\varphi_{j+1}(u,q) = \tilde{g}_{m-j,j}(\dot{u},q)\varphi_j(u,q) - \tilde{g}_{m-j,j}(u,q)\varphi_j(\dot{u},q),$$
$$\tilde{g}_{i,j+1}(u,q) = \tilde{g}_{m-j,j}(\dot{u},q)\tilde{g}_{i,j}(u,q) - \tilde{g}_{m-j,j}(u,q)\tilde{g}_{i,j}(\dot{u},q),$$

for $j = 0,\ldots,m-1$, with $\mathcal{H}_0 \triangleq 1$, $\varphi_0 \triangleq f$, and $\tilde{g}_{i,0} \triangleq g_i$, respectively.

Proof. The proof will proceed by eliminating the nonlinear functions $(h_k)_{k=1}^{m}$ from (21) one at a time using time shifts and pseudo-divisions. Starting off by time shifting (21), bearing in mind that the parameters in η and ξ are constant, yields

$$\dot{y} = \eta^{\mathrm{T}} f(\dot{u},q) + \sum_{i=1}^{m} h_i(\eta,\xi) g_i(\dot{u},q) + \dot{e}. \qquad (23)$$

Multiplying (21) by $g_m(\dot{u},q)$ and (23) by $g_m(u,q)$, and then subtracting the results, yields a difference equation in the form

$$\mathcal{H}_1(u,q) y = \eta^{\mathrm{T}} \varphi_1(u,q) + \sum_{i=1}^{m-1} h_i(\eta,\xi) \tilde{g}_{i,1}(u,q) + \mathcal{H}_1(u,q) e,$$

where the filter $\mathcal{H}_1$ and the nonlinear functions φ_1 and $\tilde{g}_{i,1}$ depend on the indicated signal, for instance, $\mathcal{H}_1(u,q) \triangleq g_m(\dot{u},q) - g_m(u,q)q^{-1}$. Now, time shifting the above expression yields

$$\mathcal{H}_1(\dot{u},q)\dot{y} = \eta^{\mathrm{T}} \varphi_1(\dot{u},q) + \sum_{i=1}^{m-1} h_i(\eta,\xi) \tilde{g}_{i,1}(\dot{u},q) + \mathcal{H}_1(\dot{u},q) \dot{e},$$

Thus, cross-multiplying the expressions above with $\tilde{g}_{m-1,1}(\dot{u},q)$ and $\tilde{g}_{m-1,1}(u,q)$, respectively, and subtracting the results, yields a difference equation in the form

$$\mathcal{H}_2(u,q) y = \eta^{\mathrm{T}} \varphi_2(u,q) + \sum_{i=1}^{m-2} h_i(\eta,\xi) \tilde{g}_{i,2}(u,q) + \mathcal{H}_2(u,q) e,$$

where the filter $\mathcal{H}_2$ and the nonlinear functions φ_2 and $\tilde{g}_{i,j}$ depend on the indicated signals, for instance,

$$\mathcal{H}_2(u,q) \triangleq \tilde{g}_{m-1,1}(\dot{u},q)\mathcal{H}_1(u,q) - \tilde{g}_{m-1,1}(u,q)\mathcal{H}_1(\dot{u},q)q^{-1}.$$

Note that the resulting difference equation after the second step is essentially the same as that after the first step. Hence, continuing in this manner, by successive shifts and pseudo-divisions, all the nonlinearities $(h_i)_{i=1}^{m}$ will be eliminated in a total of m steps and the resulting difference equation will be a linear regression model of the form (22). □

There are some conclusions that can be drawn from the result given in Theorem 25. Firstly, the pragmatic elimination scheme applied to an NFIR model structure results in a linear regression model in the parameters appearing linearly in the original NFIR model structure. In particular, if $e(t)$ is white noise with zero mean and independent of the input $u(t)$, then the least-squares estimate of η will be consistent. Secondly, since the resulting noise filter $\mathcal{H}_m$ is known, one may compensate for it using the *weighted least-squares method* (see, for example, Ljung, 1999, page 547). Finally, it is worth noting that neither the regressors φ_m nor the filter $\mathcal{H}_m$ in the resulting linear regression model (22) depend on the nonlinear functions h_i. Thus, these functions do not need to be known when performing the elimination scheme.

Now, let us try out using the elimination scheme constructed in the proof of Theorem 25 on a simple example.

Example 26. Consider the Hammerstein system (see, for instance, Ljung, 1999, page 143) generated by

$$y = (1 + \xi q^{-1})(\eta_1 u + \eta_2 u^3) + e$$
$$= \eta_1 u + \eta_2 u^3 + \eta_1 \xi \dot{u} + \eta_2 \xi \dot{u}^3 + e, \qquad (24)$$

10

where y denotes the output, u denotes the input and e is assume zero mean white noise, respectively. The parameters appearing linearly in (24) are $\eta = \begin{pmatrix} \eta_1 & \eta_2 \end{pmatrix}^{\mathrm{T}}$.

Even for this rather simple example, the application Ritt's algorithm is not practically tractable and the expressions involved become quite complicated. Furthermore, the transformed noise and the regressors will not be independent and a least-squares estimate will therefore become biased.

Now, the pragmatic elimination scheme outlined in the proof of Theorem 25 applied to (24) is quite simple, see also Example 1. First the nonlinearity containing $\eta_2 \xi$ is eliminated by a time-shift of (24) followed by a pseudo-division. Second, the remaining nonlinearity containing $\eta_1 \xi$ is eliminated in a similar manner. The resulting components of the linear regression model (22) are given by

$$\mathcal{H}(u,q) \triangleq \left((u^{(3)})^3 \ddot{u}^3 - u^{(3)} \ddot{u}^5 \right)$$
$$+ \left(u^{(3)} \dot{u}^3 - (u^{(3)})^3 \dot{u} \right) \ddot{u}^2 q^{-1} + \left(\ddot{u}^5 \dot{u} - \ddot{u}^3 \dot{u}^3 \right) q^{-2},$$

and

$$\varphi_{\eta_1}(u,q) \triangleq (u^{(3)})^3 \ddot{u}^2 \dot{u}^2 - \ddot{u}^6 \dot{u} + \ddot{u}^4 \dot{u}^3 - (u^{(3)}) \ddot{u}^2 \dot{u}^4$$
$$- (u^{(3)})^3 \ddot{u}^3 u + u^{(3)} \ddot{u}^5 u,$$
$$\varphi_{\eta_2}(u,q) \triangleq \ddot{u}^6 \dot{u}^3 - \ddot{u}^8 \dot{u} + (u^{(3)})^3 \ddot{u}^2 \dot{u}^4 - u^{(3)} \ddot{u}^2 \dot{u}^6$$
$$- (u^{(3)})^3 \ddot{u}^3 u^3 + u^{(3)} \ddot{u}^5 u^3,$$

where the subscript denotes the corresponding part of the vector-valued function φ.

To evaluate the weighted least-squares parameter estimate performance when using the above linear regression model, a dataset $Z^N = \left(y(t), u(t) \right)_{t=1}^{2,000}$ was generated by simulating (24) with the input signal $u(t) = (1 + 0.5q^{-1})v(t)$, where $v(t)$ was chosen uniformly random in the interval $[-1, 1]$, and the noise signal e was chosen as Gaussian white noise with zero mean and variance 1. The parameter values where chosen as

$$\xi = -0.75, \quad \eta_1 = 0.75, \quad \eta_2 = 0.5,$$

and the *signal to noise ratio* was determined to be approximately $35.6\,\mathrm{dB}$. The estimation experiment was repeated 500 times with different input and noise realizations and resulted in the weighted least-squares mean and standard deviation estimates

$$\hat{\eta}_1 = 0.7513 \pm 0.0347, \quad \hat{\eta}_2 = 0.4995 \pm 0.0245.$$

This indicates that the pragmatic approach may in certain cases be useful when searching for an initial estimate to the prediction-error method.

As pointed out in the example above, the pragmatic elimination scheme may work well enough to be used to find an initial parameter estimate. The scheme is easily implemented using some symbolic software.

It is worth noting that the pragmatic elimination scheme constructed in the proof of Theorem 25 also works for NARX models, that is, when the functions f or $(g_k)_{k=1}^m$ in (21) also depends on the output y. The difference is that one cannot guarantee that the regressor and the noise in the resulting linear regression will be uncorrelated. Thus, the least-squares parameter estimate given a dataset may not be asymptotically unbiased and one might need to make use of some instrumental variables approach (see, for example, Söderström and Stoica, 1983).

5 Conclusions

In this paper a generalization of Ritt's algorithm to systems of difference equations has been presented. Furthermore, the generalization of the identifiability results given in Ljung and Glad (1994) has been discussed. It turned out that it was difficult to do a straightforward generalization and the result could only be partially generalized. As a final result, a pragmatic alternative to Ritt's algorithm was formalized and the effects of the nonlinear transformations of the noise in a parameter estimation perspective was analyzed for a particular class of NFIR models and illustrated via a numerical example. It turned out that the transformed noise signal was the original noise filtered through a known linear time-dependent filter. Thus, if the original noise signal is independent of the input, then the least-squares estimate, using the transformed model structure, will be consistent.

6 Acknowledgment

References

E. Bai. An optimal two-stage identification algorithm for Hammerstein-Wiener nonlinear systems. *Automatica*, 34(3):333–338, 1998.

T. Becker and V. Weispfenning. *Gröbner Bases: A Computational Approach to Commutative Algebra*. Springer-Verlag, New York, 1993.

B. Buchberger. Theoretical basis for the reduction of polynomials to canonical forms. *ACM SIGSAM Bulletin*, 39:19–24, 1976.

S. Diop. A state elimination procedure for nonlinear systems. In *New Trends in Nonlinear Control Theory*, volume 122 of *Lecture Notes in Control and Information Sciences*, pages 190–198. Springer, 1989.

S. Diop. Elimination in control theory. *Mathematics of Control, Signals and Systems*, 4(1):17–32, 1991.

S. Diop and M. Fliess. On nonlinear observability. In *Proceedings of the 1st European Control Conference*, pages 152–157, Grenoble, France, 1991a.

S. Diop and M. Fliess. Nonlinear observability, identifiability and persistent trajectories. In *Proceedings of the 36th IEEE Conference on Decision and Control*, pages 714–719, Brighton, England, 1991b.

M. Fliess. Automatique en temps discret et algèbre aux différences. *Forum Mathematicum*, 2:213–232, 1990.

M. Fliess and T. Glad. An algebraic approach to linear and nonlinear control. In *Essays on Control: Perspectives in the Theory and Its Applications*, volume 14 of *Progress in System and Control Theory*, pages 190–198. Birkhäuser, 1994.

M. Fliess, J. Lévine, P. Martin, and P. Rouchon. Flatness and defect of nonlinear systems: Introductory theory and examples. *International Journal of Control*, 61(6): 1327–1361, 1995.

K. Forsman. *Constructive Commutative Algebra in Nonlinear Control Theory*. PhD thesis, Dissertation No. 261, Linköping University, 1991.

T. Glad. Implementing Ritt's algorithm of differential algebra. In *Proceedings of the 2nd IFAC Symposium on Nonlinear Control Systems Design*, pages 610–614, Bordeux, France, 1992.

T. Glad. Solvability of differential algebraic equations and inequalities: An algorithm. In *European Control Conference*, Brussels, Belgium, 1997.

E. Kolchin. *Differential Algebra and Algebraic Groups*. Academic Press, 1973.

S. Kotsios. An application of Ritt's remainder algorithm to discrete polynomial control systems. *IMA Journal of Mathematical Control and Information*, 18 (18):19–29, 2001.

S. Kotsios and N. Kalouptsidis. The model matching problem for a certain class of nonlinear systems. *International Journal of Control*, 57(3):707–730, 1993.

S. Lang. *Algebra*. Springer-Verlag, third revised edition, 2002.

L. Ljung. *System Identification: Theory for the User*. Prentice Hall, second edition, 1999.

L. Ljung and T. Glad. On global identifiability for arbitrary model parametrizations. *Automatica*, 30 (2):265–276, 1994.

C. Lyzell, T. Glad, M. Enqvist, and L. Ljung. Identification aspects of Ritt's algorithm for discrete-time systems. In *Proceedings of the 15th IFAC Symposium on System Identification*, Saint-Malo, France, July 2009.

B. Mishra. *Algorithmic Algebra*. Texts and Monographs in Computer Science. Springer-Verlag, 1993.

J. F. Ritt. *Differential Algebra*. Dover Publications, 1950.

W. Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, third edition, 1976.

A. Seidenberg. An elimination theory for differential algebra. In *University of California Publications in Mathematics: New Series*, pages 31–66. University of California Press, 1956.

T. Söderström and P. Stoica. *Instrumental Variable Methods for System Identification*. Springer, 1983.