Multivariate polynomial interpolation with disturbed data

Claudia Fassino, Università di Genova, Italy,
H. M. Möller, Universität Dortmund, Germany
June 14, 2014

Abstract

Given a finite set of points $X \subset \mathbb{R}^n$, one may ask for polynomials $p$ belonging to a subspace $V$, which attain given values at the points of $X$. We focus on subspaces $V$ of $\mathbb{R}[x_1, \ldots, x_n]$, generated by low order monomials. Such $V$ were computed by the BM-algorithm, which is essentially based on an LU-decomposition. In this paper we present a new algorithm based on the numerical more stable QR-decomposition. If $X$ contains only points disturbed by measurement or rounding errors, the homogeneous interpolation problem is replaced by the problem of finding (normalized) polynomials minimizing $\sum_{u \in X} p(u)^2$. Such polynomials can be found as byproduct in the QR-decomposition of the new algorithm.

1 Introduction

Given a set of points $X := \{u_1, \ldots, u_m\} \subset \mathbb{R}^n$, and a set of data $\{\alpha_1, \ldots, \alpha_m\} \subset \mathbb{R}$, the multivariate polynomial interpolation problem consists in finding a polynomial $p \in \mathbb{R}[x_1, \ldots, x_n]$ satisfying

$$p(u_i) = \alpha_i, \quad i = 1, \ldots, m.$$  

For a survey on multivariate polynomial interpolation see [7]. One major difficulty is to find for a given finite point set $X$ an $m$-dimensional subspace $V \subset \mathbb{R}[x_1, \ldots, x_n]$ which is unisolvent, i.e., in which there is for every data set $\{\alpha_1, \ldots, \alpha_m\}$ exactly one interpolating polynomial $p$. We focus on arbitrarily distributed points $u_i$ and arbitrary $m$. A linear algebra argument gives that $V$ is unisolvent if and only if $V \oplus I(X) = \mathbb{R}[x_1, \ldots, x_n]$, where $I(X)$ is the ideal of all polynomials $p$ solving the homogeneous interpolation problem

$$p(u_i) = 0, \quad i = 1, \ldots, m.$$  

A polynomial set $L := \{\ell_1, \ldots, \ell_m\}$ is a basis of an unisolvent $V$ if and only if the $\ell_i + I(X)$ constitute a basis of the factor ring $\mathbb{R}[x_1, \ldots, x_n]/I(X)$. By introducing a linear order $<_T$ among the power products (terms) $x_1^{i_1} \cdots x_n^{i_n}$, one
can have a partial order among polynomials. The BM-algorithm introduced first in [3] is based on the idea to select in every equivalence class \( \ell + I(\mathbb{X}) \) a least polynomial w.r.t. this partial order. It is used and refined in several articles in Computer Algebra like [6] and [11]. In interpolation theory, this approach has been studied by de Boor and Ron in [2]. They called their concept least interpolation. Later, also Sauer pursued this approach in [14] under the name minimal interpolation with minimal monomials.

The knowledge of \( I(\mathbb{X}) \), the complement of \( V \), is also of its own interest. In the bivariate case \( n = 2 \) one needs sometimes to find a low order curve \( p = 0 \) passing through the \( m \) points \( u_i \). For \( n > 2 \) the problem might be to find one or some algebraic surfaces \( p_j = 0 \) through the given \( m \) points. In both cases one has to determine the polynomials in the ideal \( I(\mathbb{X}) \) up to a certain degree. Unfortunately, in applications the problem is often that the points of \( \mathbb{X} \) are only approximations to certain unknown points \( \hat{u}_i \) with \( \|\hat{u}_i - u_i\|_2 \leq \varepsilon \) for \( i = 1, \ldots, m \). Hence there is possibly no low degree polynomial \( p \) vanishing in the \( m \) points \( u_1, \ldots, u_m \), i.e. \( p(u_i) = 0, i = 1, \ldots, m \), but a low degree \( \hat{p} \) vanishing in the unknown \( \hat{u}_1, \ldots, \hat{u}_m \). In this case methods of Computer Algebra like [3] fail to determine \( \hat{p} \). On the other hand, in the least squares method one expects that \( \hat{p} \) or at least a good approximation to \( \hat{p} \) minimizes \( \sum_{i=1}^m p(u_i)^2 \), assumed that all polynomials \( p \) under consideration are scaled in the same way and assumed that \( \varepsilon \) is small enough. For this reason we are interested in polynomials \( p \) such that \( \sum_{i=1}^m p(u_i)^2 \) is small, called by several authors ([5], [9], [15]) “almost vanishing polynomials”.

In the recent literature, the properties of the almost vanishing polynomials have been analyzed. Among the others, in [15], using the notion of H-bases, an algorithm is presented for computing a set of polynomials which generates “almost” the so called approximate ideal of accuracy \( \varepsilon \), that is the set of all almost vanishing polynomials w.r.t. \( \varepsilon \). In [9] an algorithm, based on the singular value decomposition, is presented for computing the “\( \varepsilon \)-approximate vanishing ideal”, that is the ideal generated by a given set of almost vanishing polynomials. In [1] and [5] algorithms for computing a set of almost vanishing polynomials are introduced, following the scheme of the BM algorithm and explicitly using an estimation of the data error. An actual overview over the existing literature can be found in [12]. This thesis also contains many well described numerical programs and includes methods using singular values, while we focus on the less expensive QR-decomposition.

In the present paper, we construct at first a unisolvent \( \mathbb{R} \)-linear space \( V \) following the ideas of [3] assuming that the interpolation points and the computations are exact. The main tool is that we order linearly the set of terms (power products)

\[ T := \{x_1^{i_1} \cdots x_n^{i_n} \mid i_1, \ldots, i_n \in \mathbb{N}_0\} \]

by a so called admissible term order \( <_T \) and consider successively (for increasing terms \( t \)) the \( \mathbb{R} \)-linear spaces \( \text{span}_\mathbb{R}\{\tau \in T \mid \tau \leq_T t\} \). The computation
ends when the first linear space $V$ is found with $V \oplus I(X) = \mathbb{R}[x_1, \ldots, x_n]$. Then an ideal basis for $I(X)$ can be computed. If the linear order is degree compatible, then we detect early low degree polynomials vanishing in the given points.

Section 3 contains two algorithms for computing such unisolvent $V$ and a (linear) basis for it. The first one is essentially the old BM-algorithm based on successive LU-decompositions of matrices $A_k$, having as rows the first $k$ evaluation vectors of a basis of $V$ at $X$. The second one uses the more stable QR-decomposition of matrices $B_k$, where $B_k$ is the transposed of $A_k$. In both cases, the successive decompositions are realized by a recursive procedure. The QR-decomposition allows a least squares interpretation as shown in Section 4. We denote by $V_t$ the subset of $\text{span}_{\mathbb{R}} \{ \tau \in T \mid \tau \leq_T t \}$ consisting in all $p$ which have in its term expansion $1 \in \mathbb{R}$ as cofactor of $t$. If $M_t$ is the minimum of $\sum_{i=1}^m p(u_i)^2$ over all $p \in V_t$, then $\sqrt{M_t}$ is the diagonal element corresponding to $t$ in the upper triangular matrix $R$ of the QR-decomposition of $B_m$. If there is in $V_t$ a polynomial $p$ which has small values $|p(u_i)|$ then $M_t$ is small. Hence the size distribution on the diagonal of $R$ indicates where to find solutions of the homogeneous interpolation problem and basis elements for the unisolvent $\mathbb{R}$-linear space $V$. Section 4 ends with an upper bound for $\sum_{i=1}^m p(u_i)^2$ if $p$ has zeros in $\hat{u}_i$ with $||u_i - \hat{u}_i||_2 \leq \varepsilon$ for $i = 1, \ldots, m$. We conclude with a series of examples.

2 The basic algorithm

In this section we begin with some elementary definitions and results from Computer Algebra. Readers who are not familiar with such techniques and results can find more details and examples in textbooks like [4].

In the following $P := \mathbb{R}[x_1, \ldots, x_n]$ denotes the polynomial ring under consideration. By $T$ we denote the set of power products (terms),

$$T := \{ x_1^{i_1} \cdots x_n^{i_n} \mid i_1, \ldots, i_n \in \mathbb{N}_0 \} .$$

A linear order in $T$ is called admissible if it satisfies

$$1 = x_1^0 \cdots x_n^0 <_T t \quad \text{for all } t \in T \setminus \{1\} ,
\quad t_1 <_T t_2 \Rightarrow tt_1 <_T tt_2 \quad \text{for all } t, t_1, t_2 \in T .$$

An easy consequence is $t_1 <_T tt_1$ for arbitrary $t, t_1 \in T, t \neq 1$. Hence if $t_1$ divides properly $t_2$, i.e. $t_2 = tt_1$ with $t \neq 1$, then $t_1 <_T t_2$. Every admissible term order is a well-order, such that $\min T'$ exists for every non-empty subset $T'$ of $T$. In the following, we fix an admissible term order $<_T$. 

3
Definition 2.1 For arbitrary polynomials \( p \in \mathcal{P} \setminus \{0\} \), \( p = \sum_{i=1}^{m} c_i t_i \) with \( c_i \in \mathbb{R} \setminus \{0\} \) and \( t_1 <_T t_2 <_T \ldots <_T t_m \), we call \( \text{lt}(p) := t_m \) leading term, \( \text{lc}(p) := c_m \) leading coefficient of \( p \), and \( \text{lm}(p) := c_m t_m \) leading monomial.

Definition 2.2 Let \( I \subseteq \mathcal{P} \) be an ideal. By \( \text{lt}(I) \) we denote the set of leading terms
\[
\text{lt}(I) := \{ \text{lt}(p) \mid 0 \neq p \in I \}.
\]
If \( \mathcal{F} \subset \mathcal{P} \) is a finite or infinite set, then \( \langle \mathcal{F} \rangle \) denotes the least ideal containing \( \mathcal{F} \),
\[
\langle \mathcal{F} \rangle = \left\{ \sum_{i=1}^{m} g_i f_i \mid \{f_1, \ldots, f_m\} \subseteq \mathcal{F}, g_1, \ldots, g_m \in \mathcal{P} \right\}.
\]
If \( \mathcal{F} \) is finite, then it is called basis of the ideal \( \langle \mathcal{F} \rangle \). The set \( \{g_1, \ldots, g_r\} \subset I \) is called a Gröbner basis of an ideal \( I \), if \( \langle \text{lt}(I) \rangle = \langle \text{lt}(g_1), \ldots, \text{lt}(g_r) \rangle \).

A Gröbner basis of an ideal \( I \) is also a basis of \( I \), [4].

Definition 2.3 If \( I \subset \mathcal{P} \) is an ideal, then
\[
\mathcal{N} := \{ t \in T \mid t \notin \text{lt}(I) \}
\]
is called normal set or set of standard monomials.

In [4], the standard monomials \( t \in \mathcal{N} \) are also called basis monomials, because the set of cosets \( \{t + I \mid t \in \mathcal{N} \} \) is a basis of the factor ring \( \mathcal{P}/I \).

In multivariate interpolation, the name lower set is used for \( \mathcal{N} \). In connection with border bases, e.g. in [10], the set of terms \( \mathcal{N} \) is also called order ideal.

Definition 2.4 For \( \mathbb{X} := \{u_1, \ldots, u_m\} \subset \mathbb{R}^n \) the vanishing ideal \( I(\mathbb{X}) \) is defined by
\[
I(\mathbb{X}) := \{ p \in \mathcal{P} \mid p(u_i) = 0, \ i = 1, \ldots, m \}.
\]
The following Proposition holds ([4], Th. 2.10).

Proposition 2.5 \( I(\mathbb{X}) \) is an ideal. Its normal set \( \mathcal{N} \) satisfies \( |\mathcal{N}| = |\mathbb{X}| \), where \( |U| \) denotes the number of elements of the set \( U \).

It is easily shown that a set \( \mathcal{N} \subseteq T \) is the normal set of an ideal if and only if
\[
t \in T, t_i \in \mathcal{N}, t|t_i \Rightarrow t \in \mathcal{N}, \tag{1}
\]
where \( t|t_i \) means that \( t \) divides \( t_i \). One says that the normal set is closed under division. An old result of Macaulay gives that for every finite subset \( \mathcal{N} \subset T \), which is closed under division, there is a point set \( \mathbb{X} \) such that \( \mathcal{N} \) is the normal set of \( I(\mathbb{X}) \), see [13].
Example 2.1 Let $X := \{(0,0), (1,1), (1,-1)\}$. If $<_T$ is the Lex-order

$$x^iy^j <_T x^ky^\ell \iff i < k \text{ or } i = k \text{ and } j < \ell ,$$

then $I(X) \subset \mathbb{R}[x,y]$ has as Gröbner basis $G = \{y^3 - y, x - y^2\}$ and hence $(\text{lt}(I(X))) = (y^3, x)$ and $N = \{1, y, y^2\}$. If $<_T$ is the DegLex-order,

$$x^iy^j <_T x^ky^\ell \iff i + j < k + \ell \text{ or } i + j = k + \ell \text{ and } i < k ,$$

then $I(X)$ has as Gröbner basis $\{y^2 - x, xy - y, x^2 - x\}$, hence $N = \{1, y, x\}$.

Proposition 2.6 Let $N$ be the normal set of an ideal $I \subset \mathcal{P}$. Then $t^* \in \text{lt}(I)$ if and only if

$$\exists \ p \in \text{span}_R \{t \in N \mid t <_T t^*\} : t^* - p \in I.$$

Proof. Let $t^* \in \text{lt}(I)$. Then there is a $p = \sum c_it_i$ with $t_i \in N$ and $c_i \in \mathbb{R}\setminus\{0\}$ such that $t^* - p \in I$, since the cosets $t + I$, $t \in N$, constitute a basis of $\mathcal{P}/I$. By construction $t_i \not\in \text{lt}(I)$ but $\text{lt}(t^* - p) \in \text{lt}(I)$. Hence $t^* = \text{lt}(t^* - p)$ which implies $t^* >_T t_i$. Conversely, if $t^* - p \in I$ where $p \in \text{span}_R \{t \in N \mid t <_T t^*\}$, then $\text{lt}(t^* - p) = t^*$ which is in $\text{lt}(I)$ because $t^* - p \in I$. □

The polynomial $p$ in Proposition 2.6 is uniquely determined by $t^* \in \text{lt}(I)$ and $N$. It is called normal form of $t^*$. For vanishing ideals the criterion in Proposition 2.6 can be reformulated using vectors.

Proposition 2.7 Let $X := \{u_1, \ldots, u_m\}$ and let $N$ be the normal set of $I(X)$. Defining $t(X) := (t(u_1), \ldots, t(u_m))$ for $t \in T$, then $t^* \in \text{lt}(I(X))$ holds if and only if

$$t^*(X) \in \text{span}_R \{t(X) \mid t \in N, t <_T t^*\} .$$

(2)

In addition, if $N := \{t_1, \ldots, t_m\}$ with $t_1 <_T t_2 <_T \ldots <_T t_m$, then

$$t_k = \min \{ t \in T \mid \dim(\text{span}_R \{t(X), \ldots, t_{k-1}(X), t(X)\}) = k \} .$$

(3)

Proof. The equivalence of $t^* \in \text{lt}(I(X))$ and (2) holds because of Proposition 2.6 and

$$t^* - \sum_i c_it_i \in I \iff t^*(X) - \sum_i c_it_i(X) = 0 .$$

The vectors $t_1(X), \ldots, t_m(X)$ are linearly independent because otherwise a $t_k \in N$ belongs to $\text{lt}(I(X))$ by (2) contradicting $N \cap \text{lt}(I(X)) = \emptyset$. This together with (2) gives (3). □

If the normal set $N$ of an ideal $I$ is known, the minimal elements of $T \setminus N$ w.r.t. the relation $a|b$ ($a$ divides $b$) are easily determined. Let $M$ be the set of these minimal elements. By Dickson’s Lemma [4], $M$ is finite. By
construction $\langle M \rangle = \langle 1t(I) \rangle$ and $\langle M' \rangle \neq \langle 1t(I) \rangle$ for every proper subset $M'$ of $M$, but $\langle I \rangle = \langle 1t(I) \rangle$ for every $M' \subset T$ containing $M$. For instance

$$\tilde{M} := \{ x, t \mid t \in \mathcal{N}, \ 1 \leq i \leq n \} \setminus \mathcal{N}$$

contains $M$ and is hence also a basis of $\langle 1t(I) \rangle$.

By Proposition 2.6, there is for every $t \in T \setminus \mathcal{N}$ a $p$ in $\text{span} \mathcal{N}$ such that $t - p \in I$. Hence $\{ t - p \mid t \in \mathcal{M}, \ p \in \text{span} \mathcal{N} \}$ and also $\{ t - p \mid t \in \tilde{M}, \ p \in \text{span} \mathcal{N} \}$ are Gröbner bases. The first one is a so called reduced Gröbner basis, the latter a border basis.

Criterion (2) allows to decide by methods of linear algebra that $t^* \in \langle 1t(I(X)) \rangle$ or its negation $t^* \in \mathcal{N}$ holds true, once all elements $t \in \mathcal{N}$ with $t <_T t^*$ are known. Since we consider the terms $t^*$ in increasing order, we do not need to test a term $t \in \langle 1t(I(X)) \rangle$, which is a proper multiple of an element of $M$, because the latter is considered earlier. Therefore we deal in the following only with the finite sets $\mathcal{N}$ and $\mathcal{M}$. Then criterion (2) reduces to the test $t^* \in \mathcal{M}$ or its negation $t^* \in \mathcal{N}$. This is the central idea for the BM-algorithm, which is in the basic version as follows.

**Basic BM-algorithm**

Input: $X := \{ u_1, \ldots, u_m \} \subset \mathbb{R}^n$.

Output: The normal set $\mathcal{N}$ and the set $\mathcal{M}$ of minimal terms in $\langle 1t(I(X)) \rangle$ w.r.t. the relation $a | b$.

Calculation:

$\mathcal{N} := \{ 1 \}; \ \mathcal{M} := \{ \}; \ L := \{ x_1, \ldots, x_n \}$;

while $L \neq \{ \}$ do

$t^* := \text{min}_{<_T} L$;
$L := L \setminus \{ t^* \}$;

if $t^*(X) \notin \text{span} \mathcal{R} \{ t(X) \mid t \in \mathcal{N} \}$  % The crucial boolean expression

then $\mathcal{N} := \mathcal{N} \cup \{ t^* \}$;

for $i = 1$ to $n$ do if no $t \in L$ divides $x_i t^*$ then $L := L \cup \{ x_i t^* \}$

else $\mathcal{M} := \mathcal{M} \cup \{ t^* \}$

end if
end while;

return $\mathcal{N}$, $\mathcal{M}$.

The basic BM-algorithm gives just the term sets $\mathcal{N}$ and $\mathcal{M}$ and not yet polynomials. However, the knowledge of the normal set alone is sufficient for finding an appropriate interpolation space for interpolation in $X = \{ u_1, \ldots, u_m \} \subset \mathbb{R}^n$. In the univariate case the situation is simple. Here, the set $\mathcal{N}$ is always generated by the terms $1, x, \ldots, x^{m-1}$. They span a linear space in which the interpolation problem in the $m$ points is uniquely soluble. In the multivariate case, we have the following.
Theorem 2.8 Let $\mathcal{N}$ be the normal set for the vanishing ideal $I(\{u_1, \ldots, u_m\})$ with $u_i \neq u_j$ for $i \neq j$. Then $\text{span}_{\mathbb{R}} \mathcal{N}$ is unisolvent, i.e., for arbitrary $\alpha_1, \ldots, \alpha_m \in \mathbb{R}$ there exists exactly one $p \in \text{span}_{\mathbb{R}} \mathcal{N}$ such that $p(u_i) = \alpha_i$, $i = 1, \ldots, m$.

Let $\mathcal{N} = \{t_1, \ldots, t_m\}$ with $t_1 <_T t_2 <_T \cdots <_T t_m$. If $\text{span}_{\mathbb{R}} \{t'_1, \ldots, t'_m\}$ with $t'_1 <_T t'_2 <_T \cdots <_T t'_m$ is also unisolvent, then $t_i \leq_T t'_i$ for $i = 1, \ldots, m$.

Proof. The $m$ vectors $t(X) \in \mathbb{R}^m$, $t \in \mathcal{N}$, are linearly independent by Proposition 2.7. Hence they are a basis of $\mathbb{R}^m$. This is equivalent to the fact that the interpolation problem $p \in \text{span}_{\mathbb{R}} \mathcal{N}$, $p(u_i) = \alpha_i$, $i = 1, \ldots, m$, has always a unique solution. The remaining follows by equation (3). \hfill \Box

Some authors require for an interpolating polynomial $p$ that its degree is not greater than the degree of the polynomial $f$ which is interpolated. If the term order $<_T$ is chosen such that lower degree terms are less w.r.t. $<_T$, then this requirement results from the following.

Corollary 2.9 Let $\mathcal{N}$ be as in Theorem 2.8 and let $f \in \mathcal{P}$. The interpolating polynomial $p \in \text{span}_{\mathbb{R}} \mathcal{N}$ with $p(u_i) = f(u_i)$, $i = 1, \ldots, m$, satisfies $\mathbf{1}t(p) \leq_T \mathbf{1}t(f)$.

Proof. Assume that in the contrary $\mathbf{1}t(f) <_T \mathbf{1}t(p)$. Then $\mathbf{1}t(p-f) = \mathbf{1}t(p)$. But $p-f \in I(\{u_1, \ldots, u_m\}) =: I$ such that $\mathbf{1}t(p-f) \in \mathbf{1}t(I)$ whereas $\mathbf{1}t(p) \in \mathcal{N} = T \setminus \mathbf{1}t(I)$, a contradiction. \hfill \Box

3 The complete BM-algorithm

The complete BM-algorithm extends the basic one by computing a reduced Gröbner basis for $I(X)$, a basis for $\text{span}_{\mathbb{R}} \mathcal{N}$, and modifies the criterion $t^* \notin \mathcal{M}$. We present here two variants of the complete algorithm. One is the original one of [3] and uses implicitly an LU-decomposition. The second one is based on a QR-decomposition. When in the basic BM-algorithm the term $t^*$ is considered, then one has already computed

$$\{t_1, \ldots, t_{s-1}\} := \{t \in \mathcal{N} \mid t <_T t^*\}.$$

The decision $t^* \notin \mathcal{M}$, i.e., $t^* \in \mathcal{N}$, uses equation (2) of Proposition 2.7. If we denote by $A_{s-1}$ the $(s-1) \times m$-matrix with rows $t_1(X), \ldots, t_{s-1}(X)$, then $A_{s-1}$ has full row rank $s-1$. Hence $t^* \in \mathcal{N}$ holds iff

$$\text{rank} \begin{bmatrix} A_{s-1} \\ t^*(X) \end{bmatrix} = s.$$
In the two variants of the BM-algorithm, we use the recursion

\[
A_s = \begin{bmatrix}
A_{s-1} & \cdots & \cdots & \cdots \\
\vdots & \ddots & \ddots & \ddots \\
\vdots & \cdots & \ddots & \cdots \\
t_s(X)
\end{bmatrix}
\]

and transform \(A_s\) or its transpose \(A_s^T\) respectively, to an upper triangular matrix using that the corresponding upper triangular matrix for \(A_{s-1}\), or \(A_{s-1}^T\) respectively, is already computed.

In [3], polynomials \(q_1, \ldots, q_m\) are constructed, which can be considered as the multivariate analogous of the Newton polynomials \(n_1, \ldots, n_m \in \mathbb{R}[x]\) defined by \(n_k \in \text{span}\{1, \ldots, x^{k-1}\}, \text{lcm}(n_k) = 1\), and \(n_k(u_{\nu}) = \ldots = n_k(u_{\nu_{s-1}}) = 0\). These polynomials were computed successively in the algorithm. When one considers \(t^*\) and has already found \(t_1, \ldots, t_{s-1} \in \mathcal{N}\), then \(q_1, \ldots, q_{s-1}\) are already computed, satisfying \(q_i \in \text{span}_\mathbb{R}\{t_1, \ldots, t_i\}, \text{lcm}(q_i) = 1\), and

\[
q_i(u_{\nu}) = \ldots = q_i(u_{\nu_{s-1}}) = 0, \quad q_i(u_{\nu}) \neq 0.
\]

Then one defines \(q\) recursively starting with \(q^{(0)} := t^*\) and then

\[
q^{(i+1)} := q^{(i)} - \frac{q^{(i)}(u_{\nu_i})}{q_i(u_{\nu})} q_i \quad \text{for} \quad i = 0, \ldots, s - 1.
\]

\(q := q^{(s)}\) has as representation \(q = t^* + \sum_{i=1}^{s-1} \ell_i t_i\) with some \(\ell_i \in \mathbb{R}\) and \(q(u_{\nu_i}) = 0\) for \(i = 1, \ldots, s - 1\). If \(q(u_{\nu_i}) = 0\) for all \(i = 1, \ldots, m\), then \(t^* \in \text{lt}(I(\mathbb{K}))\) and \(q\) is an element of the reduced Gröbner basis. Otherwise, there is an \(u_{\nu_s} \in U \setminus \{\nu_1, \ldots, \nu_{s-1}\}\) such that \(q(u_{\nu_s}) \neq 0\). In this case \(t_s := t^* \in \mathcal{N}\) and \(q_s := q\).

This is in matrix notation, using \(q_i = t_i + \sum_{k=1}^{t-1} \ell_{ik} t_k\) with \(\ell_{ik} \in \mathbb{R}\),

\[
\begin{pmatrix}
1 & 0 & \cdots & 0 & 0 \\
\ell_{21} & 1 & \cdots & \cdots & \cdots \\
\vdots & \ddots & \ddots & \ddots & \ddots \\
\ell_{s-1,1} & \cdots & \ell_{s-1,s-2} & 1 & 0 \\
\ell_1 & \cdots & \ell_{s-2} & \ell_{s-1} & 1
\end{pmatrix}
\begin{pmatrix}
t_1(u_{\nu_1}) & t_1(u_{\nu_2}) & \cdots & t_1(u_{\nu_m}) \\
\vdots & \vdots & \ddots & \ddots \\
t_{s-1}(u_{\nu_1}) & t_{s-1}(u_{\nu_2}) & \cdots & t_{s-1}(u_{\nu_m}) \\
t^*(u_{\nu_1}) & t^*(u_{\nu_2}) & \cdots & t^*(u_{\nu_m})
\end{pmatrix}
= \begin{pmatrix}
q_1(u_{\nu_1}) & q_1(u_{\nu_2}) & \cdots & q_1(u_{\nu_{s-1}}) & q_1(u_{\nu_s}) & \cdots & q_1(u_{\nu_m}) \\
0 & q_2(u_{\nu_2}) & \cdots & q_2(u_{\nu_{s-1}}) & q_2(u_{\nu_s}) & \cdots & q_2(u_{\nu_m}) \\
\vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\
0 & \cdots & 0 & q_{s-1}(u_{\nu_{s-1}}) & q_{s-1}(u_{\nu_s}) & \cdots & q_{s-1}(u_{\nu_m}) \\
0 & \cdots & 0 & \cdots & 0 & q(u_{\nu_s}) & q(u_{\nu_m})
\end{pmatrix}
\]

At the end, for \(t^* = t_m\) and \(s = m\), this identity is \(MA_m P = Q_m\), where \(P\) is the permutation matrix which permutes \(i\)-th and \(\nu_i\)-th column, \(Q_m\) an upper triangular \(m \times m\)-matrix, and \(M\) a quadratic lower diagonal matrix with diagonal \((1, 1, \ldots, 1)\), such that \(M^{-1}Q_m\) is the LU-decomposition of \(A_m P\).

The number of floating point operations (flops), i.e. the number of floating point additions and multiplications, for computing the unknowns \(\ell_1, \ldots, \ell_{s-1}\)
is about $2sm - s^2$. If $m$ is much greater than $s$, the linear increase in $m$ indicates a good performance. This explains the success of the BM-algorithm in exact arithmetic for finding a low degree polynomial in $I(X)$. However, when the points $u_1, \ldots, u_m$ are only approximately known or when floating point arithmetic is used, then the decision that all $q(u_i)$, $i = s, \ldots, m$, are zero is no more decidable and a more stable variant of the BM-algorithm should be preferred.

The QR-decomposition applied to matrices $A_{s-1}$ can do the job. We prefer to apply the QR-decomposition to $A_{s-1}^T$, the transpose of $A_{s-1}$, because we have then a better interpretation of intermediate results. Assuming again, that the term $t^*$ is under consideration in the basic BM-algorithm and we already have

$$\{t_1, \ldots, t_{s-1}\} := \{t \in \mathcal{N} \mid t <_T t^*\},$$

we define

$$B_{s-1} := A_{s-1}^T = \begin{pmatrix} t_1(u_1) & t_2(u_1) & \cdots & t_{s-1}(u_1) \\ t_1(u_2) & t_2(u_2) & \cdots & t_{s-1}(u_2) \\ \vdots & \vdots & \ddots & \vdots \\ t_1(u_m) & t_2(u_m) & \cdots & t_{s-1}(u_m) \end{pmatrix}.$$ 

Then $B_{s-1}$ has full column rank. Considering now the previously introduced vectors $t(X)$ as column vectors, $B_{s-1}$ has $t_1(X), \ldots, t_{s-1}(X)$ as columns. And the criterion $t^* \in \mathcal{N}$ needed in the basic BM-algorithm is equivalent to rank$[B_{s-1}:t^*(X)] = s$. If $B_{s-1} = Q_{s-1}R_{s-1}$ with an orthogonal matrix $Q_{s-1} \in \mathbb{R}^{m \times m}$ and an upper triangular matrix $R_{s-1} = (r_{ij}^{(s-1)}) \in \mathbb{R}^{m \times (s-1)}$, then

$$Q_{s-1}^T[B_{s-1}:t^*(X)] = [R_{s-1}:Q_{s-1}^Tt^*(X)] = \begin{pmatrix} r_{11}^{(s-1)} & r_{12}^{(s-1)} & \cdots & r_{1,s-1}^{(s-1)} & b_1 \\ 0 & r_{22}^{(s-1)} & \cdots & r_{2,s-1}^{(s-1)} & b_2 \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & r_{s-1,s-1}^{(s-1)} & b_{s-1} \\ 0 & \cdots & 0 & 0 & b_s \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & b_m \end{pmatrix}$$

with $b = Q_{s-1}^Tt^*(X)$. The extended matrix $[B_{s-1}:t^*(X)]$ has rank $s$ iff the last $m - s + 1$ entries of $b$ are not simultaneously 0. In this case, a column vector $w \in \mathbb{R}^m$ exists, which has the first $s - 1$ entries equal to 0 and its remaining entries are determined such that the vector

$$(I - \frac{2}{w^T w}ww^T)Q_{s-1}^Tt^*(X)$$

has a positive $s$-th entry and the forthcoming entries at position $s + 1, \ldots, m$ are 0. In matrix-vector notation, there is a Householder matrix

$$H_w = I - \frac{2}{w^T w}ww^T \quad (H_w \text{ is symmetric and orthogonal!})$$

9
such that \(H_wQ^T_{s-1}[B_{s-1}:t^*(\mathbb{X})]\) is an upper triangular matrix \(R_s\) with full column rank. Since \(Q_{s-1}H_w\) is orthogonal, \((Q_{s-1}H_w)R_s\) is the QR-decomposition of the extended matrix \(B_s := [B_{s-1}:t^*(\mathbb{X})]\) with \(t_s = t^*\).

If the last \(m - s + 1\) entries of \(Q^T_{s-1}t^*(\mathbb{X})\) are simultaneously zero, then the last column \(Q^T_{s-1}t^*(\mathbb{X})\) depends linearly on the columns of upper triangular matrix \(Q^T_{s-1}B_{s-1} = R_{s-1}\). Backward substitution gives the solution \(c := (c_1, \ldots, c_{s-1})^T\) of \(R_{s-1}c = Q^T_{s-1}t^*(\mathbb{X})\). Then the polynomial

\[
p := t^* - \sum_{i=1}^{s-1} c_i t_i
\]

vanishes in the points \(u_1, \ldots, u_m\), i.e., \(p \in I(\mathbb{X})\) with \(1 t(p) = t^*\).

Following [8], the determination of the vector \(w\) of the Householder matrix \(H_w\) requires about \(3(m - s)\) flops and the multiplication about \(2(m - s)^2\) flops. The complexity here is much higher than for the LU-decomposition. On the other hand, the QR-decomposition is more stable and the central key in the following least squares method.

### 4 The least squares method

In many applications, one is interested in a point set \(\hat{\mathbb{X}} := \{\hat{u}_1, \ldots, \hat{u}_m\} \subset \mathbb{R}^n\), but due to rounding or measurement errors only an approximate point set \(\mathbb{X} := \{u_1, \ldots, u_m\}\) is known. Therefore it is impossible to find the vanishing ideal \(I(\hat{\mathbb{X}})\), but only polynomials \(p\) which are very close to polynomials \(\hat{p} \in I(\hat{\mathbb{X}})\). Intuitively, a polynomial \(p^*\) is close to an element of \(I(\hat{\mathbb{X}})\), if all values \(p^*(u_i)\) are close to 0. We therefore make the following definition.

**Definition 4.1** Let \(V_{t^*} := \{p \in \mathcal{P} \mid 1 t(p) = t^*\}\). We say that a polynomial \(p^* \in V_{t^*}\) is a best approximation in the set \(V_{t^*}\) if

\[
\sum_{j=1}^{m} p^*(u_j)^2 \leq \sum_{j=1}^{m} p(u_j)^2 \quad \text{if} \quad p \in V_{t^*}.
\]

(Note that all polynomials in \(V_{t^*}\) have leading coefficient 1.)

If one adds a polynomial \(q\) with \(q(u_j) = 0\) for \(j = 1, \ldots, m\) and \(1 t(q) < T t^*\) to a \(p \in V_{t^*}\), then the square sum is unchanged,

\[
\sum_{j=1}^{m} p(u_j)^2 = \sum_{j=1}^{m} (p(u_j) + q(u_j))^2.
\]

Also \(p + q \in V_{t^*}\). Therefore if \(\{t_1, \ldots, t_{s-1}\} := \{t \in \mathcal{N} \mid t < T t^*\}\),

we may omit all terms \(t < T t^*\) with \(t \not\in \{t_1, \ldots, t_{s-1}\}\) and obtain a unique \(p^* \in \text{span}\{t_1, \ldots, t_{s-1}, t^*\}\) with

\[
\sum_{j=1}^{m} p^*(u_j)^2 = \sum_{j=1}^{m} p(u_j)^2 \quad \text{if} \quad p \in V_{t^*}.
\]  \(\text{(4)}\)
The finding of the optimal \( p^* \in \text{span}\{t_1, \ldots, t_{s-1}, t^*\} \cap V_t \) is the least squares problem

\[
\min_{c \in \mathbb{R}^{s-1}} \| B_{s-1}c - t^*(X) \|^2_2 = \| B_{s-1}c^* - t^*(X) \|^2_2,
\]

(5)

with column vectors \( c = (c_1, \ldots, c_{s-1})^T \), \( c^* = (c_1^*, \ldots, c_{s-1}^*)^T \in \mathbb{R}^{s-1} \). The vector \( c^* \) is uniquely determined in (5) because \( B_{s-1} \) has full column rank, and so \( p^* = t^* - \sum_{i=1}^{s-1} c_i^* t_i \). In the case \( t^* \in 1t(I(X)) \) there is a \( p^* = \sum c_i^* t_i \) such that \( t^* - p^* \in I(X) \), see Proposition 2.6. The vector \( c^* = (c_1^*, \ldots, c_{s-1}^*)^T \), inserted in (5), shows that the minimum is 0. These results can be summarized in the following theorem.

**Theorem 4.2** The vector \( c^* \) is the solution of the least squares problem (5) if and only if the polynomial \( t^* - \sum_{i=1}^{s-1} c_i^* t_i \) is a best approximation in the set \( V_t \). The minimum is 0 if and only if \( t^* - \sum_{i=1}^{s-1} c_i^* t_i \in I(X) \).

The polynomial of best approximation in the set \( V_t \) is strongly connected with the residual of the least squares problem and the QR-decomposition.

**Theorem 4.3** Let \( \rho_1 := t_1(X) \) and, for \( k = 2, \ldots, s \), let \( t_k \notin 1t(I(X)) \), \( \rho_k \) be the residual of the least squares problem \( \min_c \| B_{k-1}c - t_k(X) \|^2_2 \) and \( p_k \) denote a polynomial of best approximation in the set \( V_{t_k} \). Let \( Q_sR_s \) be a QR-decomposition of \( B_s \), i.e., \( Q_s \) an orthogonal \( m \times m \)-matrix and \( R_s \) an upper triangular matrix with positive diagonal entries. Then, for \( k = 1, \ldots, s \), we have \( \rho_k = (p_k^*(u_1), \ldots, p_k^*(u_m))^T \) and

i) the \( k \)-th column of \( Q_s \) is \( \frac{1}{\| \rho_k \|_2} \rho_k \),

ii) the \( k \)-th diagonal element of \( R_s \) equals \( \| \rho_k \|_2 \), and

iii) \( \| \rho_k \|_2^2 = \sum_{j=1}^m p_k^*(u_j)^2 = \min \left\{ \sum_{j=1}^m p(u_j)^2 \mid p \in V_{t_k} \right\} \).

**Proof.** Applying Theorem 4.2 to \( t^* = t_k \) gives \( p_k^* = t_k - \sum_{i=1}^{k-1} c_i^*(k) t_i \) where the vector \( c^{(k)} := (c_1^{(k)}, \ldots, c_k^{(k)})^T \) satisfies \( \min_c \| B_{k-1}c - t_k(X) \|^2_2 = \| B_{k-1}c^{(k)} - t_k(X) \|^2_2 \). Obviously, the \( j \)-th entry of the residual \( \rho_k = t_k(X) - B_{k-1}c^{(k)} \) is \( t_k(u_j) - \sum_{i=1}^{k-1} c_i^{(k)} t_i(u_j) = p_k^*(u_j) \), \( k = 2, \ldots, s \). For \( k = 1 \) we have \( t_1 = 1 \), \( p_1^* = 1 \), and \( \rho_1 = (1, \ldots, 1)^T \).

The residual \( \rho_k \) is perpendicular to \( W_{k-1} := \text{span}_\mathbb{R}\{t_1(X), \ldots, t_{k-1}(X)\} \) and \( W_k - \rho_k \in W_{k-1} \). By an inductive argument, \( \rho_i^\top \rho_k = 0 \) for \( 1 \leq i < k \leq s \) and \( W_k - \text{span}_\mathbb{R}\{p_1, \ldots, p_k\} \) for \( k = 1, \ldots, s \). In addition, \( \rho_k \neq 0 \) since \( t_k \notin 1t(I(X)) \) for \( k = 1, \ldots, s \). Therefore the matrix \( Q_s \) with columns \( \frac{1}{\| \rho_k \|_2} \rho_k \), \( k = 1, \ldots, s \), is well defined and orthogonal. Since \( t_k(X) - \rho_k \in W_{k-1} = \text{span}_\mathbb{R}\{p_1, \ldots, p_{k-1}\} \), there is an upper triangular matrix \( R_s \) with diagonal entries \( \| \rho_1 \|_2, \ldots, \| \rho_s \|_2 \) such that \( Q_sR_s \) is the QR-decomposition of \( B_s \). Statement iii) follows by \( \rho_k = (p_k^*(u_1), \ldots, p_k^*(u_m))^T \) and the definition of \( p_k^* \).

The BM-algorithm in exact arithmetic can be thus considered as a succession of least squares problems or as an iterative computation of QR-decompositions.
of matrices $B_s$, where the choice of the next $t^*$ is as in the basic BM-algorithm. In the following we also assume that all points of $\mathbb{X}$ are scaled such that they are located in the unit hypercube $C := \{(x_1, \ldots, x_n) \in \mathbb{R}^n \mid -1 \leq x_i \leq 1\}$. This enables us to get informations on the size of the residuals $\rho_k$.

**Theorem 4.4** Let $t_j \in T$ divide $t_k \in T$. If all points $\{u_1, \ldots, u_m\}$ are inside $C$, then $\|\rho_k\|_2 \leq \|\rho_j\|_2$.

**Proof** Let $p_j^* = t_j - \sum_{i=1}^{j-1} c_i^{(j)} t_i$ denote a polynomial of best approximation in $V_{t_j}$ and let $t_k = \tau t_j$ for a $\tau \in T$. Then

$$\|\rho_k\|_2^2 = \min \left\{ \sum_{i=1}^{m} p(u_i)^2 \mid p_k \in V_{t_k} \right\} \leq \sum_{i=1}^{m} (\tau(u_i) t_j(u_i) - \sum_{\ell=1}^{j-1} c^{(j)}_{\ell} \tau(u_i) t_{\ell}(u_i))^2$$

since $\sum_{\ell=1}^{j-1} c^{(j)}_{\ell} \tau t_{\ell} \in \text{span}\{t_1, \ldots, t_{k-1}\}$ and hence $\tau t_j - \sum_{\ell=1}^{j-1} c^{(j)}_{\ell} \tau t_{\ell} \in V_{t_k}$. The condition $u_i \in C$ implies $|\tau(u_i)| \leq 1$ for $i = 1, \ldots, m$. This gives

$$\|\rho_k\|_2^2 \leq \sum_{i=1}^{m} \left( \tau(u_i) t_j(u_i) - \sum_{\ell=1}^{j-1} c^{(j)}_{\ell} \tau(u_i) t_{\ell}(u_i) \right)^2$$

$$= \sum_{i=1}^{m} \left( \tau(u_i)^2 t_j(u_i) - \sum_{\ell=1}^{j-1} c^{(j)}_{\ell} t_{\ell}(u_i) \right)^2$$

$$\leq \sum_{i=1}^{m} \left( t_j(u_i) - \sum_{\ell=1}^{j-1} c^{(j)}_{\ell} t_{\ell}(u_i) \right)^2 = \|\rho_j\|_2^2.$$  

This was to be proved. \(\square\)

If floating point arithmetic is used, then it is unrealistic to assume that a residual $\rho_k$ is 0. But even if this happens and if $\mathbb{X}$ is only an approximation to $\hat{\mathbb{X}}$, the polynomial of best approximation in $V_{t_k}$ being then 0 in all points of $\mathbb{X}$ is possibly not the best approximating polynomial to $I(\hat{\mathbb{X}})$ among all polynomials with leading monomial $t_k$. However, it is to be expected, that the greater

$$\|\rho_k\|_2^2 = \min \left\{ \sum_{j=1}^{m} p(u_j)^2 \mid p \in V_{t_k} \right\},$$

the more unlikely is $t_k \in 1t(I(\hat{\mathbb{X}}))$. Residuals are small, if in case $\|u_i - \hat{u}_i\|_2 \leq \epsilon$ there is a polynomial $p$ zero in $\hat{\mathbb{X}}$ as the following result shows.

**Proposition 4.5** Let $p \in \mathcal{P}$ vanish in the points of $\hat{\mathbb{X}} = \{\hat{u}_1, \ldots, \hat{u}_m\} \subset \mathbb{R}^n$, and let $\mathbb{X} := \{u_1, \ldots, u_m\} \subset \mathbb{R}^n$ be close to $\hat{\mathbb{X}}$, $\|\hat{u}_i - u_i\|_2 \leq \epsilon$, $i = 1, \ldots, m$. Then

$$\sum_{i=1}^{m} p(u_i)^2 \leq \epsilon^2 (M_1^2 + \ldots + M_m^2),$$

where

$$M_i := \max\{\|\nabla p(u)\|_2 \mid \|u - u_i\|_2 \leq \epsilon\}.$$  

Furthermore, $M_i^2 = \|\nabla p(u_i)\|_2^2 + O(\epsilon)$.
Proof. The mean value theorem shows the existence of $\tau_i$ on the line between $u_i$ and $\hat{u}_i$ such that

$$p(u_i) = p(u_i) - p(\hat{u}_i) = \nabla p(\tau_i)(u_i - \hat{u}_i), \quad i = 1, \ldots, m.$$  

Cauchy-Schwarz gives then

$$p(u_i)^2 = (\nabla p(\tau_i)(u_i - \hat{u}_i))^2 \leq \|\nabla p(\tau_i)\|^2_2 \|u_i - \hat{u}_i\|_2^2 \leq M_i^2 \varepsilon^2.$$  

The first assertion follows by summation over all $i$.

Furthermore, let $\eta_i$ be a point such that $M_i = \|\nabla p(\eta_i)\|_2$ and $\|\eta_i - u_i\|_2 \leq \varepsilon$. Since there is a constant $C_{ik}$ such that

$$\left| \left( \frac{\partial p}{\partial x_k}(\eta_i) \right)^2 - \left( \frac{\partial p}{\partial x_k}(u_i) \right)^2 \right| \leq C_{ik} \varepsilon$$

we have

$$M_i^2 = \|\nabla p(\eta_i)\|^2_2 = \sum_{k=1}^n \left( \frac{\partial p}{\partial x_k}(\eta_i) \right)^2 \leq \sum_{k=1}^n \left( \frac{\partial p}{\partial x_k}(u_i) \right)^2 + \varepsilon \sum_{k=1}^n C_{ik}$$

Hence $M_i^2$ equals $\|\nabla p(u_i)\|^2_2$ up to a summand of size $O(\varepsilon)$. \hfill \square

Remark. Denoting by $p(\mathcal{X})$ the vector with entry $p(u_i)$ at position $i$, Proposition 4.5 gives that $p$ has no zero set $\hat{\mathcal{X}}$ close to $\mathcal{X}$ if $\|p(\mathcal{X})\|_2^2 > \varepsilon^2 \sum_{i=1}^m M_i^2$. If $D$ denotes the $m \times n$ matrix with entry $\frac{\partial p}{\partial x_k}(u_i)$ at position $(i,k)$, then $\sum_{i=1}^m M_i^2$ can be estimated, up to an $O(\varepsilon)$ summand, by the square of the Schur (Frobenius) norm of $D$. In fact $\|D\|_F^2 = \sum_{i=1}^m \|\nabla p(u_i)\|^2_2$ and so

$$\sum_{i=1}^m M_i^2 \approx \|D\|_F^2.$$  

In the examples of Section 5, we will use $\|D\|_F^2$ instead of $\sum_{i=1}^m M_i^2$. We expect that $\|p(\mathcal{X})\|_2^2 > \varepsilon^2 \|D\|_F^2$ allows as well to predict that $p$ has no zero set $\hat{\mathcal{X}}$ close to $\mathcal{X}$.

In [5] [Th. 3.5] a result analogous to Prop. 4.5 is shown, using a component-wise analysis: if there exists a set $\hat{\mathcal{X}} = \{\hat{u}_1, \ldots, \hat{u}_m\}$ such that $p(\hat{\mathcal{X}}) = 0$ and $\|\hat{u}_i - u_i\|_\infty < \varepsilon_M$ for each $i = 1, \ldots, m$ then $|p(\mathcal{X})| < \varepsilon$, where the upper bounds hold component-wise. The vector $v_\varepsilon$ can be estimated apart of an $O(\varepsilon_M^2)$ summand as follows

$$|v_\varepsilon| \approx \varepsilon_M \left| I - B(B^TB)^{-1}B^T \right| \sum_{k=1}^n \left| \frac{\partial p}{\partial x_k}(\mathcal{X}) \right| = \varepsilon_M \left| I - B(B^TB)^{-1}B^T \right| |D|_e$$

with $B$ the evaluation matrix of the terms $\{t \in \mathcal{N} \mid t < T \operatorname{LT}(p)\}$ at $\mathcal{X}$, $I$ the identity matrix, and $e = [1, \ldots, 1]^T$. Computing the 2-norm of the estimation...
of $|v_e|$ we obtain an upper bound similar to formula (6) up to a factor $\sqrt{n}$ due to the transformation from two different norms. We have

$$\varepsilon_M \|I - B(B^T B)^{-1}B^T\|_2 \leq \sqrt{n}\varepsilon_M \|I - B(B^T B)^{-1}B^T\|_2 \|D\|_2 \\
\leq \sqrt{n}\varepsilon_M \|I - B(B^T B)^{-1}B^T\|_F \|D\|_F = n\varepsilon_M \|I - B(B^T B)^{-1}B^T\|_2 \|D\|_F$$

Note that, using the singular value decomposition of $B$, it is possible to show that $\|I - B(B^T B)^{-1}B^T\|_2 = 1$. Furthermore, if $\varepsilon$ is the data error estimation in formula (6), we have that $\varepsilon$ is almost equal to $\sqrt{n}\varepsilon_M$ and so

$$\varepsilon_M \|I - B(B^T B)^{-1}B^T\|_2 \|D\|_2 \leq n\varepsilon_M \|D\|_F \approx \sqrt{n}\varepsilon \|D\|_F.$$

We can conclude that upper bound (6) is more precise than the upper bound obtained computing the norm of relation (7) and we suggest to use upper bound (6) when an estimation of the norm of the data error is known while upper bound (7) when an estimation component-wise of the error is known.

This estimation of Proposition 4.5 and Theorem 4.4 suggest to compute the residuals $\rho_k$ for increasing $t_k$ (w.r.t. $<_T$) in the following way. If the first residual $\rho_k$ occurs with a very small euclidean norm compared to the norm of the residuals computed before, then this $t_k$ is the first candidate for $1t(I(X))$. Since $\|\rho_k\|_2 = \|p_k^*(X)\|_2$ by Theorem 4.3 iii), Proposition 4.5 shows when the euclidean norm of $\rho_k$ is not small enough. If there are several residuals with small norm, but $m$ residuals with a significantly greater norm, then the corresponding $m$ terms, say $\{t_{r_1},...,t_{r_m}\}$, are candidates for the normal set. If all points are inside the unit hypercube $C$, then Theorem 4.4 gives that the set $\tilde{N} := \{t_{r_1},...,t_{r_m}\}$ is closed under division. If $X$ is close to $\hat{X}$ and if only small rounding errors occur, then it is to be expected, that $\tilde{N}$ is the normal set of the ideals $I(X)$ and $I(\hat{X})$ and the polynomials of best approximation in $V_t$ for every $t \in T \setminus \tilde{N}$ can be accepted as good approximations to the polynomials of $I(\hat{X})$.

5 Examples

In this section we present two examples in which, starting from a set $X$ of perturbed points, the polynomials of best approximation are computed. In order to detect if a computed polynomial $p^*$ admits a zero set close to $X$, we apply Proposition 4.5, using the “simplified” upper bound, where $\sum_{j=1}^m M_j^2$ is replaced by the (squared) Frobenius norm of $D = \left(\frac{\partial p}{\partial x_i}(u_i)\right)$. All the computations are performed using MatLab and all the coefficients are rounded.

Example 5.1 Let $\hat{X} = \{\hat{u}_1,\ldots,\hat{u}_m\} \subset \mathbb{R}^2$ be a set of exact but “unknown” points on the unit circle $x^2 + y^2 - 1 = 0$. We consider two different cases, varying the component-wise perturbation of the points and varying $m$. 

\begin{center}
14
\end{center}
**First case.** Let \( \hat{\mathcal{X}} \) be the “exact” set of \( m = 8 \) points, where \( \hat{\alpha} = \sqrt{2}/2 \),
\[
\hat{\mathcal{X}} = \{ (1,0), (0,1), (-1,0), (0,-1), (\hat{\alpha},\hat{\alpha}), (-\hat{\alpha},\hat{\alpha}), (-\hat{\alpha},-\hat{\alpha}), (\hat{\alpha},-\hat{\alpha}) \}.
\]
The vanishing ideal \( I(\mathcal{X}) \) is generated by the DegLex-Gröbner basis
\[
g_1 = x^2 + y^2 - 1 \quad g_2 = xy^3 - 0.5xy \quad g_3 = y^5 - 1.5y^3 + 0.5y.
\]
Let \( \mathcal{X} = \{ u_j \mid j = 1, \ldots, 8 \} \) be the set of points obtained perturbing all the coordinates of the points in \( \hat{\mathcal{X}} \) by component-wise errors less than \( 10^{-2} \):
\[
\mathcal{X} = \{ (0.9986, 0.0032), (0.7146, 0.7117), (0.0053, 1.0044), (-0.698, 0.7052), (-0.9972, 0.0028), (-0.7155, -0.7130), (0.0063, -0.9963), (0.7149, -0.7155) \}.
\]
In this case \( \| u_j - \hat{u}_j \|_2 < 1.5 \cdot 10^{-2} =: \varepsilon \). In the DegLex-order, the first eight terms are \( 1 < y < x < y^2 < xy < x^2 < y^3 < xy^2 < y^4 \). Let the \( k \)-th term be denoted by \( R(k,k) \) and let \( R(k,k) \) denote the \( k \)-th diagonal element in the QR-decomposition of \( B_8 = (t_1(\mathcal{X}), \ldots, t_8(\mathcal{X})) \). Then \( \min_c \{ \| B_{k-1}c - t_k(\mathcal{X}) \|_2 \} = R(k,k) \) where \( B_{k-1} = (t_1(\mathcal{X}), \ldots, t_{k-1}(\mathcal{X})) \) by Theorem 4.3. The QR-decomposition gives the following (rounded to four decimal places).

<table>
<thead>
<tr>
<th>( t_k )</th>
<th>( R(k,k) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y^3 )</td>
<td>0.3438</td>
</tr>
<tr>
<td>( y^2 )</td>
<td>1.0008</td>
</tr>
<tr>
<td>( xy^2 )</td>
<td>0.4336</td>
</tr>
<tr>
<td>( y )</td>
<td>2.0064</td>
</tr>
<tr>
<td>( xy )</td>
<td>1.0112</td>
</tr>
<tr>
<td>( 1 )</td>
<td>2.8284</td>
</tr>
<tr>
<td>( x )</td>
<td>2.0031</td>
</tr>
<tr>
<td>( x^2 )</td>
<td>0.0312</td>
</tr>
</tbody>
</table>

Table 1: Terms and diagonal entries in a two-dimensional scheme each

The gap between the element \( R(6,6) = 0.0312 \), associated to \( t_6 = x^2 \), and the other seven diagonal elements suggests that \( p_6^* \), the polynomial of best approximation in \( V_{x^2} \), is close to a polynomial of \( I(\mathcal{X}) \cap V_{x^2} \). The coefficients of \( p_6^* \) are the solution of the least squares problem \( B_{5:0}c = t_6(\mathcal{X}) \),
\[
p_6^* = x^2 - 0.0148xy + 0.9946y^2 - 0.0070x + 0.0035y - 1.0020.
\]

\( p_6^* \) is similar to \( g_1 \) and satisfies the “simplified” upper bound of Proposition 4.5:
\[
\sum_{u \in \mathcal{X}} p_6^*(u)^2 = R(6,6)^2 \approx 9.7495 \cdot 10^{-4} \\
< \varepsilon^2 \sum_{u \in \mathcal{X}} \| \nabla p_6^*(u) \|_2^2 \approx 31.98 \cdot \varepsilon^2 = 7.196 \cdot 10^{-3}.
\]

Now we consider the first 8 terms in the DegLex-order, omitting multiples of \( x^2 \), that is \( 1 < y < x < y^2 < xy < x^3 < xy^2 < y^4 \) (and denote them again by \( t_1 < \ldots < t_8 \)). The diagonal elements \( R(k,k) \) of \( R \) in the QR-decomposition of \( B_8 = (t_1(\mathcal{X}), \ldots, t_8(\mathcal{X})) \) are now
Table 2: Terms and diagonal entries in a two-dimensional scheme each

<table>
<thead>
<tr>
<th>$t_k$</th>
<th>$R(k,k)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y^4$</td>
<td>0.3538</td>
</tr>
<tr>
<td>$y^3$</td>
<td>0.4961</td>
</tr>
<tr>
<td>$y^2$</td>
<td>1.0008</td>
</tr>
<tr>
<td>$y$</td>
<td>2.0064</td>
</tr>
<tr>
<td>1</td>
<td>2.8284</td>
</tr>
</tbody>
</table>

Since there are no large gaps between these values, we expect that \{1, y, x, y^2, xy, y^3, xy^2, y^4\} is the normal set $N$ for $I(\hat{\mathbb{X}})$. Best approximating polynomials in $V_{x,y^3}$ and $V_{y^3}$ respectively are

$p_9^* = xy^3 + 0.0034y^4 + 0.0028xy^2 - 0.0013y^3 - 0.5061xy - 0.0033y^2 + 0.0015x - 0.0016y + 0.0001$, 

$p_{10}^* = y^5 - 0.0053y^4 + 0.0017xy^2 - 1.5067y^3 + 0.0023xy + 0.0028y^2 - 0.0001x + 0.5063y - 0.0015$.

The similarity to $g_2$ and $g_3$ is apparent.

Finally, the following table shows that the “simplified” upper bound of Proposition 4.5 is not satisfied by any polynomial of best approximation $p_k^*$ corresponding to $t_k \in N$, since \|$p_k^*(\mathbb{X})\|_2^2 > \varepsilon^2 \|\nabla p_k^*(\mathbb{X})\|_2^2$, $k = 1, \ldots, 8$.

<table>
<thead>
<tr>
<th>$k$</th>
<th>$|p_k^*(\mathbb{X})|_2^2$</th>
<th>$&gt; \varepsilon^2 |\nabla p_k^*(\mathbb{X})|_2^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
<td>&gt; 0</td>
</tr>
<tr>
<td>2</td>
<td>4.0256</td>
<td>&gt; 1.8000 · 10^{-3}</td>
</tr>
<tr>
<td>3</td>
<td>4.0124</td>
<td>&gt; 1.8000 · 10^{-3}</td>
</tr>
<tr>
<td>4</td>
<td>1.0017</td>
<td>&gt; 3.6234 · 10^{-3}</td>
</tr>
<tr>
<td>5</td>
<td>1.0226</td>
<td>&gt; 1.8091 · 10^{-3}</td>
</tr>
<tr>
<td>6</td>
<td>0.2461</td>
<td>&gt; 3.0624 · 10^{-3}</td>
</tr>
<tr>
<td>7</td>
<td>0.2568</td>
<td>&gt; 1.2571 · 10^{-3}</td>
</tr>
<tr>
<td>8</td>
<td>0.1252</td>
<td>&gt; 1.8165 · 10^{-3}</td>
</tr>
</tbody>
</table>

We conclude that there is no set $\hat{\mathbb{X}}$, close to $\mathbb{X}$ by less than $\varepsilon$, such that $p_k^*(\hat{\mathbb{X}}) = 0$. Hence the terms \{1, y, x, y^2, xy, y^3, xy^2, y^4\} form a normal set.

**Second case.** Let $\hat{\mathbb{X}}$ be a set of $m = 40$ points on the circle $x^2 + y^2 - 1 = 0$ and let $\mathbb{X}$ be a set of points obtained by perturbing the coordinates of each point $u_i \in \hat{\mathbb{X}}$ by less than $10^{-2}$. Then $\|u_i - u_i\| < \sqrt{2} \|u_i - u_i\|_\infty < \sqrt{2} \cdot 10^{-2} =: \varepsilon$ for $i = 1, \ldots, 40$. Following the same strategy and using the same notation as in the first case, we consider the first six terms $1 < y < x < y^2 < xy < x^2$ in the DegLex-order. Let the $k$-th term be denoted by $t_k$ and let $R(k,k)$ denote the $k$-th diagonal element in the QR-decomposition of $B_6 = (t_1(\mathbb{X}), \ldots, t_6(\mathbb{X}))$. Then $\min_c \{\|B_{k-1}c - t_k(\mathbb{X})\|_2\} = R(k,k)$ where $B_{k-1} = (t_1(\mathbb{X}), \ldots, t_{k-1}(\mathbb{X}))$.
by Theorem 4.3. The QR-decomposition gives the following (rounded to four decimal places).

<table>
<thead>
<tr>
<th>$t_k$</th>
<th>$R(k, k)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y^2$</td>
<td>1.8669</td>
</tr>
<tr>
<td>$y$</td>
<td>4.8953</td>
</tr>
<tr>
<td>$x$</td>
<td>6.3241</td>
</tr>
<tr>
<td>$x^2$</td>
<td>3.7284</td>
</tr>
<tr>
<td></td>
<td>0.0620</td>
</tr>
</tbody>
</table>

Table 3: Terms and diagonal entries in a two-dimensional scheme each

Since the element $R(6, 6)$ is considerably smaller than the other ones, we expect $\{1, y, x, y^2, xy\} \subset N$. The term $x^2$ is the leading term of the best approximating polynomial $p_6^* \in V_{x^2}$:

$$p_6^* = x^2 - 0.0029 xy + 1.0032 y^2 - 0.0005 x + 0.0007 y - 1.0037$$

Polynomial $p_6^*$ satisfies the “simplified” upper bound of Proposition 4.5, since

$$\sum_{u \in X} p_6^*(u)^2 \approx 3.844 \cdot 10^{-3}$$

and

$$\varepsilon^2 \sum_{u \in X} \|\nabla p_6^*(u)\|_2^2 = \varepsilon^2 \sum_{(x,y) \in X} (2x - 0.0029 y - 0.0005)^2$$

$$+ (-0.0029 x + 2.0064 y + 0.0007)^2 = 160.96 \cdot \varepsilon^2 \approx 3.22 \cdot 10^{-2}$$

and so it is possible that a zero set of $p_6^*$ close to $X$ by less than $\varepsilon$ exists.

Finally we consider the first 40 terms w.r.t. the DegLex term ordering omitting the multiples of $x^2$.

$$\{ t_k \mid k = 1, \ldots, 40 \} = \{1, y\} \cup \{xy^j, y^{j+2} \mid j = 0, \ldots, 18\} .$$

Computation gives that there is no a large gap between the diagonal elements $R(k, k)$ of $R$ in the QR-decomposition of $B_{40} = (t_1(X), \ldots, t_{40}(X))$. For detecting the elements of the normal set we apply Proposition 4.5 to each best approximating polynomial $p_k^* \in V_{t_k}$, $k = 1, \ldots, 40$. The generic term $t_k$ can be added to the normal set if $p_k^*$ does not satisfy the “simplified” upper bound of Proposition 4.5, that is if

$$\sum_{u \in X} p_k^*(u)^2 > \varepsilon^2 \sum_{u \in X} \|\nabla p_k^*(u)\|_2^2$$

or equivalently if

$$\mu_k^2 := \frac{\sum_{u \in X} p_k^*(u)^2}{\varepsilon^2 \sum_{u \in X} \|\nabla p_k^*(u)\|_2^2} > 1 .$$

In the following figure we display the values $\mu_k$, $k = 1, \ldots, 40$. For $k \leq 16$ the $\mu_k$ are greater than 1 (represented by the horizontal line) and so the terms

$$\{ t_1, \ldots, t_{16} \} = \{1, y\} \cup \{xy^j, y^{j+2} \mid j = 0, \ldots, 6\}$$

can be added to the normal set. For the terms $t_k, k > 16$, the error $\varepsilon$ is not small enough to predict $t_k \in N$. 

17
Example 5.2 The twisted cubic is the curve $\hat{\mathbb{X}} := \{(t^3, t^2, t) \mid t \in \mathbb{R}\}$. Its vanishing ideal $I(\hat{\mathbb{X}})$ is generated by the \textit{DegLex}-Gröbner basis

$$z^2 - y, \ yz - x, \ y^2 - xz.$$ 

The problem is here that the twisted cubic is no zero dimensional variety. The BM-algorithm only applies to zero-dimensional varieties. But if there are enough good approximations to points of the twisted cubic, then one can get good approximations for the low degree polynomials in $I(\hat{\mathbb{X}})$. Let

$$\mathbb{X} := \{u_1, \ldots, u_5\}$$ 

be the following set

$$\mathbb{X} = \{(-0.999912, 1.000029, -1.000039), (-0.124925, 0.249976, -0.499998), (0.000001, 0.000062, 0.000002), (0.125025, 0.250007, 0.500063), (1.000017, 0.99997, 1.000059)\}$$ 

whose points are close to the set $\{\hat{u}_1, \ldots, \hat{u}_5\} \subset \hat{\mathbb{X}}$, consisting of

$$\{-1, 1, 1\}, \{-0.125, 0.25, -0.5\}, \{0, 0, 0\}, \{0.125, 0.25, 0.5\}, \{1, 1, 1\}$$

In this case $\|u_j - \hat{u}_j\|_2 \leq 1.5 \cdot 10^{-4} =: \varepsilon$, $j = 1, \ldots, 5$. In the \textit{DegLex}-order, the first five terms are $1 < z < y < x < z^2$. Let the $k$-th term be denoted by $t_k$ and let $R(k,k)$ denote the $k$-th diagonal element in the QR-decomposition of $B_5 = (t_1(\mathbb{X}), \ldots, t_5(\mathbb{X}))$. Then $\min_c \{\|B_{k-1}c - t_k(\mathbb{X})\|_2\} = R(k,k)$ where $B_{k-1} = (t_1(\mathbb{X}), \ldots, t_{k-1}(\mathbb{X}))$ by Theorem 4.3. The QR-decomposition gives the following (rounded to four decimal places).

<table>
<thead>
<tr>
<th>$t_k$</th>
<th>1</th>
<th>$z$</th>
<th>$y$</th>
<th>$x$</th>
<th>$z^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R(k,k)$</td>
<td>2.2361</td>
<td>1.5812</td>
<td>0.9354</td>
<td>0.4744</td>
<td>0.0001</td>
</tr>
</tbody>
</table>

The terms $\{1, z, y, x\}$ are inserted into $\mathcal{N}$ and the best approximating polynomial $p_5^* \in V_{z^2}$, with leading term $t_5 = z^2$, is computed solving the least
squares problem $B_4 c = t_5(\mathbb{X})$,

$$p_5^* = z^2 - 0.00002x - 1.000126y - 0.00003z + 0.00002 .$$

Analogously, considering the next term $t_6 = yz$ in the $\text{DegLex}$-order, the QR-decomposition of the matrix $(t_1(\mathbb{X}), \ldots, t_4(\mathbb{X}), t_6(\mathbb{X}))$ gives the following (rounded to four decimal places)

$$\begin{array}{|c|c|c|c|c|}
\hline
k & 1 & z & y & x \\
\hline
R(k, k) & 2.2361 & 1.5812 & 0.9354 & 0.4744 & 8.2 \cdot 10^{-6} \\
\hline
\end{array}$$

and so the best approximating polynomial $p_6^* \in V_{yz}$ is computed,

$$p_6^* = yz - 1.0000363x + 0.000048z + 0.000016 .$$

The normal set can be completed adding to $N$ the term $t_7 = xz$ chosen following the $\text{DegLex}$-order, since the diagonal elements of $R$ of the QR-decomposition of the matrix $(t_1(\mathbb{X}), \ldots, t_4(\mathbb{X}), t_7(\mathbb{X}))$ are

$$\begin{array}{|c|c|c|c|c|}
\hline
k & 1 & z & y & x \\
\hline
R(k, k) & 2.2361 & 1.5812 & 0.9354 & 0.4744 & 0.1792 \\
\hline
\end{array}$$

The “simplified” upper bound of Proposition 4.5 shows that there is no set of points close to $\mathbb{X}$ at which the best approximating polynomial $p_k^* \in V_k$, $t_k \in N$, vanish. In fact we have that

$$\sum_{u \in \mathbb{X}} p_k^*(u)^2 \geq \varepsilon^2 \sum_{u \in \mathbb{X}} \|\nabla p_k^*(u)\|_2^2 \quad \forall t_k \in N$$

as reported in the following table

$$\begin{array}{|c|c|}
\hline
k & \|p_k^*(\mathbb{X})\|_2^2 > \varepsilon^2\|\nabla p_k^*(\mathbb{X})\|_2^2 \\
\hline
1 & 5 > 0 \\
2 & 2.5002 > 1.1250 \cdot 10^{-7} \\
3 & 0.8750 > 1.1250 \cdot 10^{-7} \\
4 & 0.2251 > 1.9377 \cdot 10^{-7} \\
7 & 0.0321 > 2.3987 \cdot 10^{-7} \\
\hline
\end{array}$$

Let $t_8 = y^2$ be the next term in the $\text{DegLex}$-order; solving the linear system $(t_1(\mathbb{X}), \ldots, t_4(\mathbb{X}), t_7(\mathbb{X}))c = t_8(\mathbb{X})$ the coefficient vector of polynomial $p_8$, with leading term $t_8$, is computed:

$$p_8 = y^2 - 0.99996xz + 1.05 \cdot 10^{-4}x - 2.85 \cdot 10^{-5}y + 1.63 \cdot 10^{-5}z - 3.14 \cdot 10^{-9}$$

Note that $p_8$ is vanishing at $\mathbb{X}$ (apart of floating point errors). Furthermore, the necessary condition given by Proposition 4.5 for the existence of a set of
points close to $X$ at which $p_5^*$ and $p_6^*$ vanish, is satisfied. In fact, if each $M_j^2$ is approximated by $\|\nabla p_j^*(u_j)\|^2_2$ for $p_5^*$ or by $\|\nabla p_6^*(u_j)\|^2_2$ for $p_6^*$, we have
\[
\sum_{u \in X} p_5^*(u)^2 \approx 10^{-8} < \varepsilon^2 \sum_{u \in X} \|\nabla p_5^*(u)\|^2_2 \approx 3.34 \cdot 10^{-7} ,
\]
\[
\sum_{u \in X} p_6^*(u)^2 \approx 6.8 \cdot 10^{-11} < \varepsilon^2 \sum_{u \in X} \|\nabla p_6^*(u)\|^2_2 \approx 2.17 \cdot 10^{-7} .
\]

References


