A projection algorithm on the set of polynomials with two bounds

Martin Campos Pinto, Frédérique Charles, Bruno Després, Maxime Herda

A PROJECTION ALGORITHM ON THE SET OF POLYNOMIALS WITH TWO BOUNDS

M. CAMPOS PINTO, F. CHARLES, B. DESPRÉS, AND M. HERDA

Abstract. The motivation of this work stems from the numerical approximation of bounded functions by polynomials satisfying the same bounds. The present contribution makes use of the recent algebraic characterization found in [B. Després, *Numer. Algorithms*, 76(3), (2017)] and [B. Després and M. Herda, *Numer. Algorithms*, 77(1), (2018)] where an interpretation of monovariate polynomials with two bounds is provided in terms of a quaternion algebra and the Euler four-squares formulas. Thanks to this structure, we generate a new nonlinear projection algorithm onto the set of polynomials with two bounds. The numerical analysis of the method provides theoretical error estimates showing stability and continuity of the projection. Some numerical tests illustrate this novel algorithm for constrained polynomial approximation.

Keywords. Positive polynomials, Chebyshev polynomials, Quadratic programming, Quaternions.

MSC2010 subject classification. 65D15, 41A29, 90C20.

1. Introduction

Given $n \in \mathbb{N}$ we let P_n be the set of univariate polynomials of degree less or equal to n, and set by convention $P_{-1} = \{0\}$. A central result is the Lukàcs Theorem [9, Sec. 1.21] which characterizes polynomials with one lower bound. Specifically, let $P_n^+ \subset P_n$ be the subset of positive (or nonnegative) polynomials on the segment $[0,1]$, namely

$$P_n^+ \ := \ \{p \in P_n, \ \text{ such that } \ 0 \leqslant p(x) \ \text{ for all } \ x \in [0,1]\}.$$

In this article we will consider the case of even degrees. The extension to odd degrees is essentially a question of technical matters, with no new ideas with respect to the material presented in this work.

Theorem 1.1 (Even degree [9])**.** *Take $n \in \mathbb{N}$ and $p \in P_{2n}^+$. Then there exists $a \in P_n$ and $b \in P_{n-1}$ such that $p(x) = a^2(x) + b^2(x)w(x)$ with weight $w(x) = x(1-x)$.*

The problem considered in this article is the design and analysis of a nonlinear projection algorithm onto the set of polynomials with one lower bound and one upper bound,

$$U_{2n} \ := \ \{p \in P_n, \ \text{ such that } \ 0 \leqslant p(x) \leqslant 1 \ \text{ for all } \ x \in [0,1]\}.$$

Our approach is based on the observation that we have

$$U_{2n} = \{p \in P_{2n}^+ \mid 1 - p \in P_{2n}^+\}.$$

MH acknowledges support by the Labex CEMPI (ANR-11-LABX-0007-01) and the Labex SMP (ANR-10-LABX-0098).

Given that there already exist projection algorithms on P_{2n}^+ (see [2, 5]), our present objective is to design a nonlinear algorithm that maps a pair $(p_0, 1-p_1) \in P_{2n}^+ \times P_{2n}^+$ into U_{2n}. To do so we will first describe a specific parametrization of the set U_{2n} that heavily relies on the four-squares identity of Euler [6, p. 54]. This theoretical framework will then be used to build a practical algorithm for bounded polynomial approximation. To our knowledge, this work is the first attempt to use the algebraic structure of Euler's identity to build an algorithm with such advanced properties.

The organization is as follows. In the next section we introduce some elementary concepts and notation, and we specify some of the aforementioned algebraic properties: the quaternion structure is recalled, its expression in the Chebychev basis is given and some norms are defined. In Section 3 we then specify our approximation problem with two bounds and define the nonlinear projection algorithm: it is an extension of the theoretical decomposition method from [3, 4] with a new nonlinear correction step. In Section 4 we perform the numerical analysis of the method and state in Theorem 4.2 a continuity or stability result. Finally in the last section we illustrate the method with some simple numerical tests.

2. Notations and basic properties of U_{2n}

2.1. **Representation of polynomials with two bounds.** A polynomial p belongs to U_{2n} if and only if $p \in P_{2n}^+$ and $1-p \in P_{2n}^+$. Define the set of quadruplets

$$\mathcal{Q}_n \ := \ P_n \times P_{n-1} \times P_n \times P_{n-1}.$$

By Theorem 1.1, for any $p \in U_{2n}$, there is a quadruplet $q = (a, b, c, d) \in \mathcal{Q}_n$ such that $a^2(x) + b^2(x)\, w(x) + c^2(x) + d^2(x)\, w(x) = 1$. It is convenient to define the function $M : \mathcal{Q}_n \to P_{2n}$ by

$$M(q)(x) \ := \ a^2(x) + b^2(x)\, w(x) + c^2(x) + d^2(x)\, w(x)\,, \tag{1}$$

and the set

$$\mathcal{U}_n = \{q \in \mathcal{Q}_n, \ \text{ such that } \ M(q) = 1\}\,.$$

The function $(a, b, c, d) \mapsto a^2 + b^2 w$ maps $\mathcal{U}_n$ onto U_{2n}, so it is sufficient to characterize $\mathcal{U}_n$ to get a characterization of the set of polynomials U_{2n}.

A central tool will be a recent factorization result recalled in Theorem 2.1 below, that involves a multiplication law on quadruplets based on Euler's four-square identity [6]. Given two elements $r = (\alpha, \beta, \gamma, \delta)$ and $q = (a, b, c, d)$ in $\mathcal{Q}_\infty = \cup_{n\in\mathbb{N}} \mathcal{Q}_n$, we define $rq := (A, B, C, D) \in \mathcal{Q}_\infty$ with

$$\left\{ \begin{array}{rcccccccc} A &=& \alpha\, a &-& \beta\, b\, w &-& \gamma\, c &-& \delta\, d\, w\,, \\ B &=& \beta\, a &+& \alpha\, b &-& \delta\, c &+& \gamma\, d\,, \\ C &=& \gamma\, a &+& \delta\, b\, w &+& \alpha\, c &-& \beta\, d\, w\,, \\ D &=& \delta\, a &-& \gamma\, b &+& \beta\, c &+& \alpha\, d\,. \end{array} \right. \tag{2}$$

Note that this is actually a modified version of Euler's four-square identity, where the signs are different. The sign convention adopted here will make it simpler to describe $\mathcal{Q}_\infty$ by quaternions. The neutral element for this multiplication law is $(1, 0, 0, 0)$, and every element of $\mathcal{U}_\infty = \cup_{n\in\mathbb{N}} \mathcal{U}_n$ has an inverse. Indeed, define the conjugate of $q = (a, b, c, d)$ in $\mathcal{Q}_\infty$ by

$$\overline{q} \ = \ (a, -b, -c, -d).$$

Then a direct application of formula (2) yields

$$\overline{q}q \;=\; q\overline{q} \;=\; (M(q),0,0,0), \qquad \forall\; q \in \mathcal{Q}_\infty.$$

In particular,

$$\overline{q}q \;=\; q\overline{q} \;=\; (1,0,0,0), \qquad \forall\; q \in \mathcal{U}_\infty$$

so that $\mathcal{U}_\infty$ has indeed a non-commutative group structure. Note that $\overline{\overline{q}} = q$ and that $\overline{r\,q} \;=\; \overline{q}\,\overline{r}$. Moreover M is a morphism, namely $M(qr) \;=\; M(q)M(r)$ for any quadruplets q and r in $\mathcal{Q}_\infty$. With an additional natural addition defined by $(\alpha,\beta,\gamma,\delta) + (a,b,c,d) \;=\; (\alpha+a,\beta+b,\gamma+c,\delta+d)$ and a scalar multiplication $\lambda(a,b,c,d) \;=\; (\lambda a,\lambda b,\lambda c,\lambda d)$, $\mathcal{Q}_\infty$ is a non-commutative $\mathbb{R}$-algebra which inherits all its algebraic properties from the quaternions. Indeed if one represents the quadruplet (a,b,c,d) by the following quaternion-valued formal function $a+ib\sqrt{w}+jc+kd\sqrt{w}$, then the usual quaternions operations based on the relations $i^2 \;=\; j^2 \;=\; k^2 \;=\; ijk \;=\; -1$ coincide with those introduced here on our polynomial quadruplets. In this sense, the equality holds

$$(a,b,c,d) = a + ib\sqrt{w} + jc + kd\sqrt{w} \in \mathcal{Q}_n.$$

The interest of this algebraic formalism lies in the following factorization result.

Theorem 2.1 ([3, 4])**.** *Let $n \in \mathbb{N}$. For any $q \in \mathcal{U}_n$ there is $e \in \mathcal{U}_1$ such that $eq \in \mathcal{U}_{n-1}$. As a consequence, any quadruplet $q \in \mathcal{U}_n$ admits a factorization in at most n elements e_1, e_2, …, e_n of $\mathcal{U}_1$*

$$q \;=\; e_1\, e_2 \,\ldots\, e_n\,. \tag{3}$$

The structure of the proof [3, 4] is as follows. One starts from $q \in \mathcal{U}_n$ and shows that there exists $e_1 \in \mathcal{U}_1$ such that $\overline{e_1}q \in \mathcal{U}_{n-1}$. The construction of e_1 is explicit and based on the examination of the two dominant coefficients of each of the four polynomial components of q. The proof is ended by iteration on n, $n-1$, …

On the basis of this result, one has a constructive characterization of polynomials with bounds. The question addressed in the present work is the evaluation of this structure for algorithmic purposes. Since (3) is a very nonlinear formula, it is not easy to handle. However, in the rest of this article, we will show that it is possible to obtain an efficient nonlinear projection onto $\mathcal{U}_n$ using this structure.

2.2. **Chebychev basis.** It is well-known that Chebychev polynomials enjoy good stability properties which make them suitable for numerical algorithms [8]. Indeed some preliminary tests [4] for the application of Theorem 2.1 have confirmed that the monomial basis may suffer from very poor numerical accuracy for high order polynomials. Our findings are also that Chebychev bases are well adapted to the expression of Euler's four-square formula (2) along their coefficients. These reasons explain why only Chebychev bases are considered in this work for algorithmic purposes.

The shifted Chebychev polynomials of the first kind are the only polynomials such that

$$T_n\left(\frac{\cos(\theta)+1}{2}\right) = \cos(n\theta), \qquad \theta \in \mathbb{R},\; n \in \mathbb{N}.$$

The polynomial T_n is of degree n and the definition actually extends to negative indices, as $T_{-n} = T_n$. The shifted Chebychev polynomials of the second kind are

the only polynomials such that

$$U_n\left(\frac{\cos(\theta)+1}{2}\right) = \frac{2\,\sin(n\theta)}{\sin(\theta)}, \qquad \theta \in \mathbb{R}\backslash\pi\mathbb{Z},\ n \in \mathbb{N}.$$

Now U_n is of degree $n-1$, recalling our convention $P_{-1} = \{0\}$, and again we may extend the definition to negative indices: one has $U_{-n} = -U_n$. Note that the shifted Chebychev polynomials of second kind are usually defined without the factor 2, and with an index that is the degree of the polynomial. Our notation will allow to simplify some of the subsequent computations.

Chebychev polynomials enjoy natural orthogonality properties [1, 9]. Define the scalar products

$$\langle f,\, g\rangle_T \;=\; \int_0^1 f(x)\, g(x)\, w(x)^{-1/2}\,\mathrm{d}x\,, \qquad \langle f, g\rangle_U \;=\; \int_0^1 f(x)\, g(x)\, w(x)^{1/2}\,\mathrm{d}x.$$

Then for any $(i,j) \in \mathbb{Z}^2\backslash\{(0,0)\}$, one has $\langle T_i,\, T_j\rangle_T \;=\; \langle U_i,\, U_j\rangle_U \;=\; \frac{\pi}{2}\,\delta_{ij}$ where δ_{ij} is the Kronecker symbol and $\langle T_0,\, T_0\rangle_T \;=\; \pi$. These formulas are established by noticing that the weight is such that $w\big(\frac{\cos(\theta)+1}{2}\big) = \frac{\sin^2(\theta)}{4}$.

Remark 2.2. *One has the identity for all* $n \in \mathbb{N}$

$$1 = T_n(x)^2 + U_n(x)^2 w(x).$$

It underlines that the Lukàcs decomposition of a polynomial is non unique.

2.2.1. *A Chebychev basis for the set* $\mathcal{U}_n$. Any $(a,b,c,d) \in \mathcal{U}_n$ admits a Chebychev representation

$$\begin{aligned} a(x) &= \sum_{i=0}^{n} a_i\, T_i(x)\,, & c(x) &= \sum_{i=0}^{n} c_i\, T_i(x)\,, \\ b(x) &= \sum_{i=1}^{n} b_i\, U_i(x)\,, & d(x) &= \sum_{i=1}^{n} d_i\, U_i(x)\,, \end{aligned} \tag{4}$$

with

$$a_i = \frac{2-\delta_{i0}}{\pi}\,\langle a(x), T_i(x)\rangle_T\,, \quad c_i = \frac{2-\delta_{i0}}{\pi}\,\langle c(x), T_i(x)\rangle_T\,, \quad i \in \{0,1,\dots,n\}, \tag{5}$$

and

$$b_i = \frac{2}{\pi}\,\langle b(x), U_i(x)\rangle_U\,, \quad d_i = \frac{2}{\pi}\,\langle d(x), U_i(x)\rangle_U\,, \quad i \in \{1,\dots,n\}. \tag{6}$$

It will be convenient to extend these coefficients for all $i \in \mathbb{Z}$, setting $a_i = c_i = 0$ or $b_i = d_i = 0$ when i is outside of the above ranges.

The coefficients of the product (2) of two quadruplets r and q can be expressed quite handily in the Chebychev basis. Indeed, as a consequence of the De Moivre formulas, for any $(i,j) \in \mathbb{Z}^2$ one has

$$T_iT_j = \frac{T_{i-j}+T_{i+j}}{2}, \quad U_iU_jw = \frac{T_{i-j}-T_{i+j}}{2}, \quad U_iT_j = \frac{U_{i+j}+U_{i-j}}{2}. \tag{7}$$

It is useful to consider the sign function $\mathrm{sgn}(x) = 1$ for $x > 0$, $\mathrm{sgn}(x) = -1$ for $x < 0$ and $\mathrm{sgn}(0) = 0$.

Lemma 2.3. *The coefficients of the polynomials in* (2) *can be expressed as*

(8)
$$\begin{cases} 2\,A_k &= \displaystyle\sum_{i+j=k} (\alpha_i\,a_j + \beta_i\,b_j - \gamma_i\,c_j + \delta_i\,d_j) \;+ \sum_{|i-j|=k} (\alpha_i\,a_j - \beta_i\,b_j - \gamma_i\,c_j - \delta_i\,d_j), \\ 2\,B_k &= \displaystyle\sum_{i+j=k} (\beta_i\,a_j + \alpha_i\,b_j - \delta_i\,c_j + \gamma_i\,d_j) \;+ \sum_{|i-j|=k} (\beta_i\,a_j + s_{ij}\,\alpha_i\,b_j - \delta_i\,c_j + s_{ij}\,\gamma_i\,d_j), \\ 2\,C_k &= \displaystyle\sum_{i+j=k} (\gamma_i\,a_j - \delta_i\,b_j + \alpha_i\,c_j + \beta_i\,d_j) \;+ \sum_{|i-j|=k} (\gamma_i\,a_j + \delta_i\,b_j + \alpha_i\,c_j - \beta_i\,d_j), \\ 2\,D_k &= \displaystyle\sum_{i+j=k} (\delta_i\,a_j - \gamma_i\,b_j + \beta_i\,c_j + \alpha_i\,d_j) \;+ \sum_{|i-j|=k} (\delta_i\,a_j - s_{ij}\,\gamma_i\,b_j + \beta_i\,c_j + s_{ij}\,\alpha_i\,d_j), \end{cases}$$

where $s_{ij} = \operatorname{sgn}(j-i)$.

Proof. Obtained from (2) and the De Moivre formulas (7). □

Lemma 2.4. *Take* $q \in \mathcal{Q}_n$. *One can write* $M(q) = \sum_{i=0}^{2n} M(q)_i T_i$ *which is expressed with the Chebychev basis of the first kind only. The dominant coefficient is*

$$M(q)_{2n} \;=\; \frac{1}{2}(a_n^2 \,-\, b_n^2 \,+\, c_n^2 \,-\, d_n^2) \tag{9}$$

and the next one is

$$M(q)_{2n-1} \;=\; \begin{cases} a_n\,a_{n-1} \,-\, b_n\,b_{n-1} \,+\, c_n\,c_{n-1} \,-\, d_n\,d_{n-1} & \text{if } \; n \geqslant 2\,, \\ 2\,a_1\,a_0 \,+\, 2\,c_1\,c_0 & \text{if } \; n = 1\,. \end{cases} \tag{10}$$

Proof. The expansion of $M(q)$ along the Chebyshev basis shows products $T_\alpha T_\beta$ and products $U_\alpha U_\beta$. The De Moivre formulas (7) yield that all products can be expanded along the T_γ solely. Direct computations yield the coefficients $M(q)_{2n}$ and $M(q)_{2n-1}$. In formula (10), the case $n = 1$ comes from the term δ_{i0} in (5). □

For later use we define $\mathcal{U}_n^{(i)} \subset \mathcal{Q}_n$ as the subset of quadruplets q such that the $2i$ dominant coefficients of $M(q)$ vanish,

$$\mathcal{U}_n^{(i)} = M^{-1}\big(P_{2n-2i}\big) \cap \mathcal{Q}_n \tag{11}$$

Obviously, $q \in \mathcal{Q}_n$ is in $\mathcal{U}_n^{(i)}$ if and only if $M(q)_{2n-2i+1} = \cdots = M(q)_{2n} = 0$, and in particular $q \in \mathcal{U}_n^{(n)}$ iff $M(q) \in \mathbb{R}$. Thus one has the embeddings

$$\mathcal{U}_n \subset \mathcal{U}_n^{(n)} \subset \cdots \subset \mathcal{U}_n^{(1)} \subset \mathcal{Q}_n.$$

2.3. **Metrics.** The continuity properties of the projection algorithm defined in the next section will be analyzed with convenient norms which are defined below.

For any real polynomial p, we consider its weighted L^1 norm

$$\|p\| := \int_0^1 |p(x)| \frac{\mathrm{d}x}{\sqrt{w(x)}}.$$

For quadruplets $q = (a, b, c, d) \in \mathcal{Q}_n$, we define a specific norm $|||\cdot|||$

$$|||q|||^2 := \|M(q)\| \;=\; \int_0^1 a^2 w^{-\frac{1}{2}} + \int_0^1 b^2 w^{\frac{1}{2}} + \int_0^1 c^2 w^{-\frac{1}{2}} + \int_0^1 d^2 w^{\frac{1}{2}}\,. \tag{12}$$

The orthogonality of Chebychev polynomials yields the Plancherel-like equality

$$|||q|||^2 \;=\; \pi\,(|a_0|^2 + |c_0|^2) + \frac{\pi}{2} \sum_{i=1}^{n} (|a_i|^2 + |b_i|^2 + |c_i|^2 + |d_i|^2). \tag{13}$$

Since M is a morphism and $M(e) = 1$ for $e \in \mathcal{U}_\infty$, one has

$$|||eq||| \;=\; |||q||| \quad \text{for any} \quad e \in \mathcal{U}_\infty,\ q \in \mathcal{Q}_\infty. \tag{14}$$

This last property is very useful when dealing with the decomposition formulas of Theorem 2.1.

3. The projection algorithm

In order to motivate our projection algorithm we consider the problem of computing a polynomial approximation with two bounds to some given function f assuming that, as a preliminary step, we are able to construct two polynomials with one bound, $p_0 \in P_{2n}^+$ and $p_1 \in 1 - P_{2n}^+$, which both approximate f in some sense,

$$p_0 = a^2 + b^2 w \approx f \qquad \text{and} \qquad p_1 = 1 - c^2 - d^2 w \approx f.$$

By construction, the polynomial p_0 is non negative and the polynomial p_1 is less than 1. The point is that this preliminary step is doable: for example we refer to [2, 5] where effective algorithms are proposed to compute polynomial approximations with one bound. The method [2] is restricted to monovariate polynomials, while [5] is more general and adresses multivariate polynomials. In the numerical section we shall use a third different method described in the appendix. In all cases, one ends up with a quadruplet $q = (a, b, c, d) \in \mathcal{Q}_n$ such that

$$M(q) = a^2 + b^2 w + c^2 + d^2 w = p_0 + 1 - p_1 \approx 1.$$

In particular, the quadruplet q may not be in $\mathcal{U}_n$, so that neither p_0 or p_1 are in U_{2n}. The numerical illustrations at the end show it is indeed the case. Our objective is then to construct an algorithm which projects $q = (a, b, c, d)$ into $\tilde{q} = (\widetilde{a}, \widetilde{b}, \widetilde{c}, \widetilde{d}) \in \mathcal{U}_n$ and thus provides a polynomial approximation

$$\tilde{p} := \widetilde{a}^2 + \widetilde{b}^2 w = 1 - \widetilde{c}^2 - \widetilde{d}^2 w \approx f \quad \text{ with two bounds,} \quad \tilde{p} \in U_{2n}.$$

To do so we will use the iterative decomposition technique developped in the theoretical proof of [3, 4] with an additional correction step.

3.1. Definition of the projection.

The design principle of the algorithm is to follow the iterative factorization structure developed in the proof of Theorem 2.1. Since this procedure is applied to a quadruplet that is *not* in the set $\mathcal{U}_n$, the key issue is to design a correction step that effectively allows to perform each iterative factorization. Thus our construction involves two functions that will be properly described below, see Definitions 3.4, 3.7 and 3.9.

- *The new correction function* $\chi_n : \mathcal{Q}_n \to \mathcal{U}_n^{(1)}$ is a projection onto $\mathcal{U}_n^{(1)}$, see (11). From $q \in \mathcal{Q}_n$ it creates $\widehat{q} = \chi_n(q)$ by modifying only the two dominant coefficients of the four polynomials constituting q, in order for the two dominant coefficients of $M(\widehat{q})$ to vanish.
- *The factorization function* $\phi_n : \mathcal{U}_n^{(1)} \to \mathcal{U}_1$, which from a corrected quadruplet $\widehat{q}$ explicitly builds an element $e = \phi_n(\widehat{q}) \in \mathcal{U}_1$ such that $e\widehat{q} \in \mathcal{Q}_{n-1}$. It relies on a technical adaptation of the proof of Theorem 2.1.

The structure of the algorithm is then as follows.

Definition 3.1. *The projection onto $\mathcal{U}_n$ is defined by the factorized formula*

$$\Pi_n : \left\{ \begin{array}{ccc} \mathcal{Q}_n & \longrightarrow & \mathcal{U}_n \\ q & \longmapsto & \overline{e_1}\,\overline{e_2}\,\ldots\,\overline{e_n}\,r_0 \end{array} \right. \tag{15}$$

where each factor is computed iteratively, setting $q_n := q$ and for $i = 0, \ldots, n-1$,

$$\left\{ \begin{array}{rcll} \widehat{q}_{n-i} & := & \chi_{n-i}(q_{n-i}) & \in \mathcal{U}_n^{(1)}\,, \\ e_{i+1} & := & \phi_{n-i}(\widehat{q}_{n-i}) & \in \mathcal{U}_1\,, \\ q_{n-(i+1)} & := & e_{i+1}\widehat{q}_{n-i} & \in \mathcal{Q}_{n-(i+1)}. \end{array} \right. \tag{16}$$

Here, χ_{n-i} is the correction function defined in Def. 3.4 and 3.7, ϕ_{n-i} is the factorization function defined in Def. 3.9 and $e_{i+1}\widehat{q}_{n-i}$ is a quaternion product. Finally the last term $r_0 \in \mathcal{U}_0$ in (15) is defined as

$$r_0 \;:=\; \left\{ \begin{array}{ll} q_0/M(q_0)^{1/2} & \text{if } \; q_0 \neq 0\,, \\ (1,0,0,0) & \text{otherwise}. \end{array} \right.$$

3.2. **The correction function $\chi_n : \mathcal{Q}_n \to \mathcal{U}_n^{(1)}$ for $n \geqslant 2$.** Let $q = (a,b,c,d) \in \mathcal{Q}_n$ and let us define $\chi_n(q) := \widehat{q} = (\widehat{a},\widehat{b},\widehat{c},\widehat{d}) \in \mathcal{U}_n^{(1)}$. The polynomials $\widehat{a}$, $\widehat{b}$, $\widehat{c}$ and $\widehat{d}$ are defined by changing only the dominant coefficients of (a,b,c,d) in the Chebychev basis (4). This is performed as follows.

The low order coefficients remain unchanged, namely

$$\widehat{a}_i = a_i\,, \quad \widehat{b}_i = b_i\,, \quad \widehat{c}_i = c_i\,, \quad \widehat{a}_i = a_i \quad \text{for all } i \leqslant n-2$$

In order for $\widehat{q}$ to be an element of $\mathcal{U}_n^{(1)}$, the remaining high order coefficients must satisfy the algebraic relations (9)-(10)

$$\left\{ \begin{array}{l} \widehat{a}_n^2 \,-\, \widehat{b}_n^2 \,+\, \widehat{c}_n^2 \,-\, \widehat{d}_n^2 \;=\; 0\,, \\ \widehat{a}_n\,\widehat{a}_{n-1} \,-\, \widehat{b}_n\,\widehat{b}_{n-1} \,+\, \widehat{c}_n\,\widehat{c}_{n-1} \,-\, \widehat{d}_n\,\widehat{d}_{n-1} \;=\; 0\,. \end{array} \right. \tag{17}$$

Since we desire $\chi_n(q)$ to be as close as possible to q, we decide to project

$$X = (a_n, a_{n-1}, b_n, b_{n-1}, c_n, c_{n-1}, d_n, d_{n-1})^t \tag{18}$$

onto the algebraic manifold $\mathcal{V} \subset \mathbb{R}^8$ defined by (17). The problem is thus reduced to building a projection $\chi : \mathbb{R}^8 \to \mathcal{V}$.

The mathematical issue is that the Euclidean projection on this non-convex set cannot be properly defined. Indeed if one denotes by $\|\cdot\|$ the euclidean norm in $\mathbb{R}^8$ the following quadratically constrained quadratic program

$$\inf_{Y \in \mathcal{V}} \; \frac{1}{2}\|X - Y\|^2\,, \tag{19}$$

may have multiple solutions. Via a dual convex nonlinear program, we are nevertheless able to explicitly compute at least one solution, which reveals sufficient for our algorithmic purposes. Once we are provided with a suitable candidate written as

$$\chi(X) = \left(\widehat{a}_n, \widehat{a}_{n-1}, \widehat{b}_n, \widehat{b}_{n-1}, \widehat{c}_n, \widehat{c}_{n-1}, \widehat{d}_n, \widehat{d}_{n-1}\right)^t\,,$$

we may gather the coefficients to determine $\chi_n(q)$. This will properly stated in Definition 3.4.

3.2.1. *A dual convex program.* The set of constraints of the optimization problem (17) is written as

$$\mathcal{V} \;=\; \{Y \in \mathbb{R}^8 \;\text{ such that }\; Y^t A Y \;=\; Y^t B Y \;=\; 0\}$$

with symmetric block diagonal matrices

$$A = \mathrm{diag}(S, -S, S, -S) \in \mathcal{M}_8(\mathbb{R}), \quad B = (T, -T, T, -T) \in \mathcal{M}_8(\mathbb{R})$$

where

$$S = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad T = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

The set $\mathcal{V}$ is also called the correction manifold in the following. The Lagrangian associated to (19) is

$$L(Y, \lambda, \mu) \;=\; \frac{1}{2}\left(\|X - Y\|^2 + \lambda\, Y^t A Y + \mu\, Y^t B Y\right).$$

The triplets (Y, λ, μ) satisfying the first order optimality condition $\nabla L \;=\; 0$ are those satisfying $Y \in \mathcal{V}$ and

$$M_{\lambda,\mu} Y \;=\; X$$

with

$$M_{\lambda,\mu} \;=\; I + \lambda\, A + \mu\, B\,. \tag{20}$$

The conditions of invertibility of $M_{\lambda,\mu}$ reduce to the invertibility of $I \pm (\lambda S + \mu T)$. The four eigenvalues of the symmetric matrix $M_{\lambda,\mu} = M^t_{\lambda,\mu} \in \mathcal{M}_8(\mathbb{R})$ are $1 \pm (\lambda \pm \sqrt{|\lambda|^2 + 4|\mu|^2})/2$ and $1 \pm (\lambda \mp \sqrt{|\lambda|^2 + 4|\mu|^2})/2$. It is natural to define the open and convex set

$$\mathcal{D} = \{(\lambda, \mu) \in \mathbb{R}^2 \text{ such that } M_{\lambda,\mu} > 0\} = \{|\lambda| + \mu^2 < 1\}.$$

This set is bounded with boundary $\partial\mathcal{D} = \{|\lambda| + \mu^2 = 1\}$. Moreover, on $\mathcal{D}$ it holds

$$I \pm (\lambda S + \mu T) \geqslant 0, \qquad \text{hence} \qquad \|\lambda S + \mu T\| \leqslant 1$$

in the matrix 2-norm over $\mathbb{R}^2$, which results in a uniform bound

$$\|M_{\lambda,\mu}\| \leqslant 2, \qquad (\lambda, \mu) \in \mathcal{D}$$

in the matrix 2-norm over $\mathbb{R}^8$. Let us now consider the dual optimization problem

$$(\lambda^*(X), \mu^*(X)) \in \arg \inf_{(\lambda,\mu)\in\mathcal{D}} G_X(\lambda, \mu) \tag{21}$$

with

$$G_X(\lambda, \mu) = X^t\, M^{-1}_{\lambda,\mu}\, X\,. \tag{22}$$

The function G_X enjoys the following nice property.

Lemma 3.2. *Assume G_X has a critical point $(\lambda^*, \mu^*) \in \mathcal{D}$, in the sense that $\nabla G_X(\lambda^*, \mu^*) = 0$. Then $Y^* = M^{-1}_{\lambda^*,\mu^*} X$ is in the correction manifold $\mathcal{V}$.*

Proof. One has the differential formula $dM^{-1} = -M^{-1} dM M^{-1}$ which holds for matrices $M > 0$. So an explicit calculation shows that

$$\partial_\lambda G_X(\lambda^*, \mu^*) = -X^t M^{-1}_{\lambda^*,\mu^*} A M^{-1}_{\lambda^*,\mu^*} X = -Y^{*t} A Y^* = 0.$$

Similarly $\partial_\mu G_X(\lambda^*, \mu^*) = -Y^{*t} B Y^* = 0$, hence $Y^* \in \mathcal{V}$. In particular, $\mathcal{V} \neq \varnothing$. □

The following result shows that, generically, (λ^*, μ^*) exists and is a global minimum of the functional G_X.

Lemma 3.3. *For any $X \in \mathbb{R}^8$, the function $G_X : \mathcal{D} \to \mathbb{R}_+$ is convex and C^1. Moreover there is a dense open subset $\mathcal{S} \subset \mathbb{R}^8$ such that whenever $X \in \mathcal{S}$, the function G_X tends to $+\infty$ on the boundary of $\mathcal{D}$ (namely, it is coercive).*

Proof. The convexity stems from the non-negativity of $M_{\lambda,\mu}$ since for $\alpha, \beta \in \mathbb{R}^2$

$$\begin{pmatrix}\alpha\\ \beta\end{pmatrix}^t \mathrm{Hess}\, G_X(\lambda,\mu) \begin{pmatrix}\alpha\\ \beta\end{pmatrix} = X^t M_{\lambda,\mu}^{-1}(\alpha A + \beta B) M_{\lambda,\mu}^{-1}(\alpha A + \beta B) M_{\lambda,\mu}^{-1} X \geqslant 0.$$

By explicitly inverting $M_{\lambda,\mu}$ and using the notation (18), one has that

$$\begin{aligned} G_X(\lambda,\mu) \;=\; & \frac{(a_n - \mu a_{n-1})^2}{1+\lambda-\mu^2} + a_{n-1}^2 + \frac{(b_n + \mu b_{n-1})^2}{1-\lambda-\mu^2} + b_{n-1}^2 \\ & + \frac{(c_n - \mu c_{n-1})^2}{1+\lambda-\mu^2} + c_{n-1}^2 + \frac{(d_n + \mu d_{n-1})^2}{1-\lambda-\mu^2} + d_{n-1}^2. \end{aligned}$$

This shows that G_X is C^1 on $\mathcal{D}$ and goes to $+\infty$ on $\partial\mathcal{D} = \{(\lambda,\mu) \text{ s.t. } |\lambda| + \mu^2 = 1\}$ as soon as the terms between parenthesis do not vanish (uniformly in μ). It is the case for

$$X \in \mathcal{S} \;=\; \{a_n c_{n-1} \neq a_{n-1} c_n \;\text{ and }\; b_n d_{n-1} \neq b_{n-1} d_n\} \subset \mathbb{R}^8.$$

The set $\mathcal{S}$ is an open and dense subset of $\mathbb{R}^8$. □

At this point, for any X in the dense set $\mathcal{S}$ of Lemma 3.3, we know that the dual optimization problem admits at least one solution $(\lambda^*, \mu^*) \in \mathcal{D}$ that is a critical point of G_X. We can then define

$$\begin{aligned} \chi(X): \quad & \mathcal{S} \longrightarrow \mathcal{V}, \\ & X \longmapsto \chi(X) = M^{-1}_{\lambda^*(X),\mu^*(X)} X \end{aligned} \tag{23}$$

where $(\lambda^*(X), \mu^*(X))$ is a global minimizer of G_X obtained by a given convex optimization method. Of course, the definition of χ may vary since there are possibly several global minima (the precise implementation is detailed in Section 5). Also by perturbation around $\mathcal{S}$, the function χ can defined

$$\chi(X): \quad \mathbb{R}^8 \longrightarrow \mathcal{V}$$

with the same restrictions concerning the choice of the minimizer which is non unique as well and the choice of the perturbation. Regardless of these choices, we may now state the complete definition of χ_n when $n \geqslant 2$.

Definition 3.4. *The function $\chi_n : \mathcal{Q}_n \to \mathcal{U}_n^{(1)}$ takes $q = (a(x),\, b(x),\, c(x),\, d(x))$ as argument and returns*

$$\chi_n(q) = \widehat{q} = (\widehat{a}(x),\, \widehat{b}(x),\, \widehat{c}(x),\, \widehat{d}(x))$$

with

$$\widehat{a}_i = a_i\,, \quad \widehat{b}_i = b_i\,, \quad \widehat{c}_i = c_i\,, \quad \widehat{d}_i = d_i \quad \textit{for all} \;\; i \leqslant n-2$$

and

$$\left(\widehat{a}_n, \widehat{a}_{n-1}, \widehat{b}_n, \widehat{b}_{n-1}, \widehat{c}_n, \widehat{c}_{n-1}, \widehat{d}_n, \widehat{d}_{n-1}\right) = \chi\left(a_n, a_{n-1}, b_n, b_{n-1}, c_n, c_{n-1}, d_n, d_{n-1}\right)$$

where the projection χ is defined in (21)-(23).

3.2.2. *Properties of the non convex optimization problem.*

Proposition 3.5. *The function χ has values in the correction manifold $\mathcal{V}$, and*

(i) it is nonincreasing in the euclidean norm of $\mathbb{R}^8$, namely

$$\|\chi(X)\| \leqslant \|X\|;$$

(ii) it satisfies the estimate

$$\|X - \chi(X)\| \leqslant 2^{\frac{3}{4}}\|X\|^{1/2}(\left|X^tAX\right| + \left|X^tBX\right|)^{1/4}$$

where the right hand side vanishes for $X \in \mathcal{V}$;

(iii) it is idempotent, i.e. $\chi \circ \chi = \chi$.

These estimates are uniform with respect to the choice of the minimizer in (21).

Proof. Let $X \in \mathcal{S}$ and $Y^* = \chi(X) = M^{-1}_{\lambda^*,\mu^*}X$ as defined in Lemma 3.2. We know that $Y^* \in \mathcal{V}$.

(i) One has

$$X^tY^* = Y^{*t}M_{\lambda^*,\mu^*}Y^* = Y^{*t}(I + \lambda^*A + \mu^*B)Y^* = \|Y^*\|^2$$

which yields the first estimate $\|Y^*\| \leqslant \|X\|$.

(ii) A Taylor formula with integral remainder expansion yields

$$G_X(\lambda^*,\mu^*) = G_X(0,0) - (\lambda^*\, X^tAX \,+\, \mu^*\, X^tBX)$$

$$+2\int_0^1 X^tM^{-1}_{s\lambda^*,s\mu^*}(\lambda^*A + \mu^*B)M^{-1}_{s\lambda^*,s\mu^*}(\lambda^*A + \mu^*B)M^{-1}_{s\lambda^*,s\mu^*}X(1-s)\,\mathrm{d}s.$$

Since $G_X(\lambda^*,\mu^*) \leqslant G_X(0,0)$ and the matrices commute

$$M^{-1}_{s\lambda^*,s\mu^*}(\lambda^*A + \mu^*B) = (\lambda^*A + \mu^*B)M^{-1}_{s\lambda^*,s\mu^*},$$

one has the inequality

$$2\int_0^1 (Z^*)^tM^{-3}_{s\lambda^*,s\mu^*}Z^*(1-s)\mathrm{d}s \;\leqslant\; (\lambda^*\, X^tAX \,+\, \mu^*\, X^tBX)$$

where $Z^* = (\lambda^*A + \mu^*B)X$. It yields

$$\|Z^*\|^2 = 2\int_0^1 (M^{-3/2}_{s\lambda^*,s\mu^*}Z^*)^tM^3_{s\lambda^*,s\mu^*}M^{-3/2}_{s\lambda^*,s\mu^*}Z^*(1-s)\mathrm{d}s$$

$$\leqslant 2\int_0^1 \|M_{s\lambda^*,s\mu^*}\|^3 \left\|M^{-3/2}_{s\lambda^*,s\mu^*}Z^*\right\|^2 (1-s)\mathrm{d}s$$

$$\leqslant 2\int_0^1 2^3 \left\|M^{-3/2}_{s\lambda^*,s\mu^*}Z^*\right\|^2 (1-s)\mathrm{d}s$$

$$\leqslant 2^4\int_0^1 (Z^*)^tM^{-3}_{s\lambda^*,s\mu^*}Z^*(1-s)\mathrm{d}s$$

$$\leqslant 2^3(\lambda^*\, X^tAX + \mu^*\, X^tBX).$$

Using $|\lambda^*| + (\mu^*)^2 < 1$, one gets the technical bound

$$\|Z^*\| \leqslant 2^{\frac{3}{2}}(\left|X^tAX\right| + \left|X^tBX\right|)^{1/2}.$$

By definition of Y^* one has $X - Y^* = (\lambda^* A + \mu^* B)Y^*$. So

$$\begin{aligned} \|X - Y^*\|^2 &= {Y^*}^t(\lambda^* A + \mu^* B)(X - Y^*) \\ &= {Y^*}^t(\lambda^* A + \mu^* B)X \\ &= {Y^*}^t Z^* \leqslant \|Y^*\|\,\|Z^*\|. \end{aligned}$$

So $\|X - Y^*\| \leqslant \|Y^*\|^{\frac{1}{2}}\|Z^*\|^{\frac{1}{2}}$. One concludes with (i) and the previous technical bound.

(iii) The estimate in (ii) yields $\|\chi \circ \chi(X) - \chi(X)\| = 0$ since $\chi(X) \in \mathcal{V}$.

The proof is ended. □

Corollary 3.6. *Let $n \geqslant 2$. The correction function $\chi_n : \mathcal{Q}_n \to \mathcal{U}_n^{(1)}$ satisfies*

(i) $\quad |\!|\!|\chi_n(q)|\!|\!| \leqslant |\!|\!|q|\!|\!|, \qquad q \in \mathcal{Q}_n,$

(ii) $\quad |\!|\!|q - \chi_n(q)|\!|\!| \leqslant\ C\,|\!|\!|q|\!|\!|^{1/2}(|M(q)_{2n}| + |M(q)_{2n-1}|)^{1/4}, \quad q \in \mathcal{Q}_n,$

(iii) $\quad \chi_n \circ \chi_n = \chi_n.$

for some constant $C > 1$.

Proof. These properties follow from Proposition 3.5, observing that the non zero coefficients of $q - \chi_n(q)$ coincide with those of $X - \chi(X)$: using (13) this gives

$$|\!|\!|q|\!|\!|^2 - |\!|\!|\chi_n(q)|\!|\!|^2 = \frac{\pi}{2}\big(\|X\|^2 - \|\chi(X)\|^2\big) \geqslant 0$$

and for estimate (ii) we use $|M(q)_{2n}| + |M(q)_{2n-1}| = \frac{1}{2}\big(\,|X^t A X| + |X^t B X|\,\big)$. □

3.3. **The correction function $\chi_1 : \mathcal{Q}_1 \to \mathcal{U}_1^{(1)}$.** For $n = 1$, the correction function χ_n needs a specific definition. Indeed, in order for $\widehat{q} = \chi_1(q)$ to be in $\mathcal{U}_1^{(1)}$, the following relations must hold

$$\left\{ \begin{array}{l} \widehat{a}_1^2 - \widehat{b}_1^2 + \widehat{c}_1^2 - \widehat{d}_1^2 = 0, \\ \widehat{a}_1\,\widehat{a}_0 + \widehat{c}_1\,\widehat{c}_0 = 0, \end{array} \right. \tag{24}$$

and they slightly differ from the previous ones (17). However the method and results are essentially the same. Specifically, (24) define a slightly different set of constraints

$$\widetilde{\mathcal{V}} = \{\tilde{Y} = (\widehat{a}_1, \widehat{a}_0, \widehat{b}_1, \widehat{c}_1, \widehat{c}_0, \widehat{d}_1) \in \mathbb{R}^6 \ \text{ such that } \ \tilde{Y}^t \tilde{A} \tilde{Y} = \tilde{Y}^t \tilde{B} \tilde{Y} = 0\}$$

with symmetric block diagonal matrices

$$\tilde{A} = \mathrm{diag}(S, -1, S, -1) \in \mathcal{M}_6(\mathbb{R}), \quad \tilde{B} = (T, 0, T, 0) \in \mathcal{M}_6(\mathbb{R}).$$

This leads to the dual optimization problem

$$(\lambda^*(\tilde{X}), \mu^*(\tilde{X})) \in \arg \inf_{(\lambda,\mu)\in\widetilde{\mathcal{D}}} \tilde{G}_{\tilde{X}}(\lambda, \mu) \qquad \text{with} \qquad \tilde{G}_{\tilde{X}}(\lambda, \mu) = \tilde{X}^t\, \tilde{M}_{\lambda,\mu}^{-1}\, \tilde{X} \tag{25}$$

with a matrix $\tilde{M}_{\lambda,\mu} = I + \lambda\tilde{A} + \mu\tilde{B}$ and a bounded convex domain now defined as $\widetilde{\mathcal{D}} = \{(\lambda, \mu) \in \mathbb{R}^2 : \mu^2 - 1 \leqslant \lambda \leqslant 1\}$. Thus we define

$$\begin{array}{rl} \tilde{\chi}: & \mathbb{R}^6 \longrightarrow \widetilde{\mathcal{V}}, \\ & \tilde{X} \longmapsto \tilde{\chi}(\tilde{X}) = \tilde{M}^{-1}_{\lambda^*(\tilde{X}),\mu^*(\tilde{X})}\tilde{X} \end{array} \tag{26}$$

where $(\lambda^*(\tilde{X}), \mu^*(\tilde{X}))$ is the global minima of the convex and coercive nonlinear program (25) obtained by a given optimization method.

Definition 3.7. *The function $\chi_1 : \mathcal{Q}_1 \to \mathcal{U}_1^{(1)}$ takes $q = (a_1 T_1(x)+a_0,\, b_1 U_1,\, c_1 T_1(x)+ c_0,\, d_1 U_1)$ as argument and returns*

$$\chi_1(q) = \widehat{q} = (\widehat{a}_1\, T_1(x) + \widehat{a}_0, \widehat{b}_1\, U_1, \widehat{c}_1\, T_1(x) + \widehat{c}_0, \widehat{d}_1\, U_1)$$

with $(\widehat{a}_1, \widehat{a}_0, \widehat{b}_1, \widehat{c}_1, \widehat{c}_0, \widehat{d}_1) = \tilde{\chi}(a_1, a_0, b_1, c_1, c_0, d_1)$ and $\tilde{\chi}$ defined by (25-26).

The function χ_1 has the same properties as χ_n for $n \geqslant 2$. In particular the results of Corollary 3.6 can be established also for $n = 1$. We state this as a proposition for later reference.

Proposition 3.8. *Let $n \geqslant 1$. The correction function $\chi_n : \mathcal{Q}_n \to \mathcal{U}_n^{(1)}$ satisfies*

(i) $\quad |\!|\!| \chi_n(q) |\!|\!| \leqslant |\!|\!| q |\!|\!|, \qquad\qquad q \in \mathcal{Q}_n,$

(ii) $\quad |\!|\!| q - \chi_n(q) |\!|\!| \leqslant\ C\, |\!|\!| q |\!|\!|^{1/2} (|M(q)_{2n}| \,+\, |M(q)_{2n-1}|)^{1/4}, \quad q \in \mathcal{Q}_n,$

(iii) $\quad \chi_n \circ \chi_n = \chi_n.$

for some constant $C > 1$.

3.4. **The factorization function $\phi_n : \mathcal{U}_n^{(1)} \to \mathcal{U}_1$ for $n \geqslant 1$.**

Definition 3.9. *The factorization function $\phi_n : \mathcal{U}_n^{(1)} \to \mathcal{U}_1$ takes $\widehat{q} = (\widehat{a}, \widehat{b}, \widehat{c}, \widehat{d})$ as argument. If $\widehat{a}_n^2 + \widehat{c}_n^2 = 0$, it returns $\phi_n(\widehat{q}) = (1, 0, 0, 0)$. Otherwise it is defined as follows.*

Case $n \geqslant 2$: *then $\phi_n(\widehat{q}) = K\, (\alpha_1 T_1 + \alpha_0,\ \beta_1\, U_1,\ \gamma_1\, T_1 + \gamma_0,\ \delta_1\, U_1)$ where*

$$\alpha_1 \,=\, \widehat{a}_n\,, \qquad \beta_1 \,=\, -\,\widehat{b}_n\,, \qquad \gamma_1 \,=\, -\,\widehat{c}_n\,, \qquad \delta_1 \,=\, -\,\widehat{d}_n\,, \tag{27}$$

$$\alpha_0 = \frac{\widehat{a}_{n-1}}{2} \;-\; \frac{\widehat{b}_n\, \widehat{b}_{n-1} + \widehat{d}_n\, \widehat{d}_{n-1}}{2(\widehat{a}_n^2 \,+\, \widehat{c}_n^2)}\, \widehat{a}_n \;+\; \frac{\widehat{b}_n\, \widehat{d}_{n-1} \,-\, \widehat{d}_n\, \widehat{b}_{n-1}}{2(\widehat{a}_n^2 \,+\, \widehat{c}_n^2)}\, \widehat{c}_n\,, \tag{28}$$

$$\gamma_0 = -\,\frac{\widehat{c}_{n-1}}{2} \;+\; \frac{\widehat{b}_n\, \widehat{b}_{n-1} + \widehat{d}_n\, \widehat{d}_{n-1}}{2(\widehat{a}_n^2 \,+\, \widehat{c}_n^2)}\, \widehat{c}_n \;+\; \frac{\widehat{b}_n\, \widehat{d}_{n-1} \,-\, \widehat{d}_n\, \widehat{b}_{n-1}}{2(\widehat{a}_n^2 \,+\, \widehat{c}_n^2)}\, \widehat{a}_n\,, \tag{29}$$

and $K = \big(\alpha_0^2 + \gamma_0^2 + \frac{1}{2}(\alpha_1^2 + \beta_1^2 + \gamma_1^2 + \delta_1^2)\big)^{-1/2}$ which is correctly defined since $\alpha_1^2 + \gamma_1^2 > 0$.

Case $n = 1$: *then $\phi_1(\widehat{q}) = K\widehat{q}$ with*

$$K = M(\widehat{q})^{-1/2} = \left(\widehat{a}_0^2 + \widehat{c}_0^2 + \frac{1}{2}(\widehat{a}_1^2 + \widehat{b}_1^2 + \widehat{c}_1^2 + \widehat{d}_1^2) \right)^{-1/2}.$$

Remark 3.10. *If $\widehat{a}_n^2 + \widehat{c}_n^2 = 0$ then $\widehat{q} \in \mathcal{U}_n^{(1)}$. So by (17) (or (24) if $n = 1$), one has also $\widehat{b}_n^2 + \widehat{d}_n^2 = 0$ and thus $\widehat{q} \in \mathcal{Q}_{n-1}$. This explains why these cases are distinguished in the definition.*

Proposition 3.11. *For all $\widehat{q} \in \mathcal{U}_n^{(1)}$, one has $\phi_n(\widehat{q}) \in \mathcal{U}_1$ and $\phi_n(\widehat{q})\widehat{q} \in \mathcal{Q}_{n-1}$.*

Proof. If $n = 1$, then since $\widehat{q} \in \mathcal{U}_1^{(1)}$, one has $M(\widehat{q})_2 = M(\widehat{q})_1 = 0$, so clearly $\phi_1(\widehat{q}) \in \mathcal{U}_1$ and $\phi_1(\widehat{q})\widehat{q} \in \mathcal{Q}_0$. Consider the product formulas (8). Regardless of the values of α_0 and γ_0, the product $(A, B, C, D) = (\alpha, \beta, \gamma, \delta)\, \widehat{q}$ is such that

$B_{n+1} = C_{n+1} = D_{n+1} = 0$; thanks to (17) one also has $A_{n+1} = 0$. The next coefficients of (A, B, C, D) are

$$\begin{aligned}
2\,A_n &= \Big(\widehat{a}_n\,\widehat{a}_{n-1} - \widehat{b}_n\,\widehat{b}_{n-1} + \widehat{c}_n\,\widehat{c}_{n-1} - \widehat{d}_n\,\widehat{d}_{n-1}\Big) + 2\,(\alpha_0\,\widehat{a}_n - \gamma_0\,\widehat{c}_n)\,,\\
2\,B_n &= \Big(-\,\widehat{b}_n\,\widehat{a}_{n-1} + \widehat{a}_n\,\widehat{b}_{n-1} + \widehat{d}_n\,\widehat{c}_{n-1} - \widehat{c}_n\,\widehat{d}_{n-1}\Big) + 2\,(\alpha_0\,\widehat{b}_n + \gamma_0\,\widehat{d}_n)\,,\\
2\,C_n &= \Big(-\,\widehat{c}_n\,\widehat{a}_{n-1} + \widehat{d}_n\,\widehat{b}_{n-1} + \widehat{a}_n\,\widehat{c}_{n-1} - \widehat{b}_n\,\widehat{d}_{n-1}\Big) + 2\,(\gamma_0\,\widehat{a}_n + \alpha_0\,\widehat{c}_n)\,,\\
2\,D_n &= \Big(-\,\widehat{d}_n\,\widehat{a}_{n-1} + \widehat{c}_n\,\widehat{b}_{n-1} - \widehat{b}_n\,\widehat{c}_{n-1} + \widehat{a}_n\,\widehat{d}_{n-1}\Big) + 2\,(-\gamma_0\,\widehat{b}_n + \alpha_0\,\widehat{d}_n)\,.
\end{aligned}$$

Thanks to the choice of α_0 and γ_0 in (28) and (29), all these coefficients vanish too. To simplify the computation of B_n and D_n, notice that since $\widehat{a}_n^2 + \widehat{c}_n^2 = \widehat{b}_n^2 + \widehat{d}_n^2$, the coefficients α_0 and γ_0 rewrite

$$\alpha_0 = \frac{\widehat{a}_{n-1}}{2} \;+\; \frac{\widehat{c}_n\,\widehat{d}_{n-1} - \widehat{a}_n\,\widehat{b}_{n-1}}{2(\widehat{b}_n^2 \,+\, \widehat{d}_n^2)}\,\widehat{b}_n \;-\; \frac{\widehat{a}_n\,\widehat{d}_{n-1} \,+\, \widehat{c}_n\,\widehat{b}_{n-1}}{2(\widehat{b}_n^2 \,+\, \widehat{d}_n^2)}\,\widehat{d}_n\,,$$

and

$$\gamma_0 = -\frac{\widehat{c}_{n-1}}{2} \;+\; \frac{\widehat{c}_n\,\widehat{d}_{n-1} \,-\, \widehat{a}_n\,\widehat{b}_{n-1}}{2(\widehat{b}_n^2 \,+\, \widehat{d}_n^2)}\,\widehat{d}_n \;+\; \frac{\widehat{a}_n\,\widehat{d}_{n-1} + \widehat{c}_n\,\widehat{b}_{n-1}}{2(\widehat{b}_n^2 \,+\, \widehat{d}_n^2)}\,\widehat{b}_n\,.$$

Thus $\phi_n(\widehat{q})\widehat{q} \in \mathcal{Q}_{n-1}$. Finally $\phi_n(\widehat{q}) \in \mathcal{U}_1$ since

$$\begin{aligned}
M\left(K(\alpha,\beta,\gamma,\delta)\right)(x) \;=\; & \frac{K^2}{2}\,(\widehat{a}_n^2 \,-\, \widehat{b}_n^2 \,+\, \widehat{c}_n^2 \,-\, \widehat{d}_n^2)\,T_2(x)\\
& + \frac{K^2}{2}\,(\widehat{a}_n\widehat{a}_{n-1} \,-\, \widehat{b}_n\widehat{b}_{n-1} \,+\, \widehat{c}_n\,\widehat{c}_{n-1} \,-\, \widehat{d}_n\,\widehat{d}_{n-1})\,T_1(x)\\
& + K^2\left(\alpha_0^2 + \gamma_0^2 + \frac{1}{2}(\widehat{a}_n^2 \,+\, \widehat{b}_n^2 \,+\, \widehat{c}_n^2 \,+\, \widehat{d}_n^2)\right)T_0 \;=\; 1\,.
\end{aligned}$$

This ends the proof. □

Remark 3.12. *The factorization built in the previous proof provides a constructive proof of Theorem 2.1. Indeed if $M(q) = 1$, one can check that the correction step is not active in the projection algorithm (16), i.e., $\widehat{q}_{n-i} = q_{n-i}$ for all i. One recovers the decomposition formulas of Theorem 2.1.*

4. Error estimates

With the material developed above, one can now use the projection and consider $\Pi_n(q) = (\tilde{a}, \tilde{b}, \tilde{c}, \tilde{d}) \in \mathcal{U}_n$ for $q = (a, b, c, d) \in \mathcal{Q}_n$. But for practical purposes, which ultimately is our concern, such a procedure would have little interest if the difference $q - \Pi_n(q)$ was large. It is precisely the purpose of this section to analyze this difference.

Since the projection algorithm is very nonlinear, one can expect technical difficulties in proving sharp error estimates. In what follows, we explain how the various estimates and properties already obtained combine to show some continuity properties of the projection Π_n.

In order to quantify the distance to $\mathcal{U}_n$, we define the difference

$$\varepsilon(q) = M(q) - 1. \tag{30}$$

The main theoretical result of this work is as follows.

Theorem 4.1. *Let $n \in \mathbb{N}$ and $H > 0$. For any quadruplet $q \in \mathcal{Q}_n$ satisfying $|||q||| \leqslant H$, one has*

$$|||q - \Pi_n(q)||| \;\leqslant\; (n+1)\, C(H)\, \max\left\{ \|\varepsilon(q)\|, \|\varepsilon(q)\|^{2^{-(2n+1)}} \right\}.$$

for some constant $C(H) > 0$ depending only on H.

It is instructive to reformulate Theorem 4.1 in terms of polynomials rather than in terms of quaternions.

Corollary 4.2 (of Theorem 4.1)**.** *Let $n \in \mathbb{N}$, $H > 0$ and $q = (a, b, c, d) \in \mathcal{Q}_n$ an arbitrary quadruplet satisfying $|||q||| \leqslant H$. Note $p_0 = a^2 + b^2 w$ and $p_1 = 1 - c^2 - d^2 w$ and consider $(\tilde{a}, \tilde{b}, \tilde{c}, \tilde{d}) = \Pi_n(q)$. There exists a constant $C(H) > 0$ such that the polynomial with two bounds*

$$\tilde{p} := \tilde{a}^2 + \tilde{b}^2 w = 1 - \tilde{c}^2 - \tilde{d}^2 w \;\in U_{2n}$$

satisfies

$$\|p_0 - \tilde{p}\| \;\leqslant\; (n+1)\, C(H)\, \max\left\{ \|p_0 - p_1\|, \|p_0 - p_1\|^{2^{-(2n+1)}} \right\}.$$

Proof. Using the definition of the norms $\|\cdot\|$, $|||\cdot|||$ and two Cauchy-Schwarz inequalities, we write

$$\begin{aligned}
\|p_0 - \tilde{p}\| &= \|(a+\tilde{a})(a-\tilde{a}) + (b+\tilde{b})(b-\tilde{b})w\| \\
&= \textstyle\int_0^1 |(a+\tilde{a})(a-\tilde{a})w^{-\frac{1}{2}} + (b+\tilde{b})(b-\tilde{b})w^{\frac{1}{2}}| \\
&\leqslant \|(a+\tilde{a})^2 + (b+\tilde{b})^2 w\|^{\frac{1}{2}} \|(a-\tilde{a})^2 + (b-\tilde{b})^2 w\|^{\frac{1}{2}} \\
&\leqslant |||q + \Pi_n(q)|||\; |||q - \Pi_n(q)||| \\
&\leqslant (H + |||\Pi_n(q)|||)\; |||q - \Pi_n(q)|||.
\end{aligned}$$

The result follows by combining the estimate of Theorem 4.1 with the equality $\varepsilon(q) = M(q) - 1 = p_0 - p_1$ and the observation that $|||\Pi_n(q)||| = \|M(\Pi_n(q))\|^{1/2} = \|1\|^{1/2} = \sqrt{\pi}$.

□

To prove Theorem 4.1 we begin by establishing a couple of elementary estimates.

Proposition 4.3. *There is a constant $C > 1$ such that for any integer $n \geqslant 1$ and any $q \in \mathcal{Q}_n$, the nonlinear correction operator χ_n satisfies*

$$|||q - \chi_n(q)||| \;\leqslant\; C\, |||q|||^{1/2}\, \|\varepsilon(q)\|^{1/4} \tag{31}$$

as well as

$$\|\varepsilon(\chi_n(q))\| \;\leqslant\; C\,(1 + |||q|||^{3/2})\, \|\varepsilon(q)\|^{1/4}. \tag{32}$$

Proof. Let $q = (a, b, c, d)$ and $\widehat{q} = (\widehat{a}, \widehat{b}, \widehat{c}, \widehat{d}) = \chi_n(q)$. The first estimate follows from Proposition 3.8, and the observation that the i-th coefficients of $M(q)$ and $\varepsilon(q) = M(q) - 1$ in the (T_n) Chebyshev basis coincide for $i \geqslant 1$, thus

$$|M(q)_i| = |\varepsilon(q)_i| = \tfrac{2}{\pi} |\langle \varepsilon(q), T_i \rangle_T| \leqslant \tfrac{2}{\pi} \|\varepsilon(q)\| \|T_i\|_{L^\infty(0,1)} = \tfrac{2}{\pi} \|\varepsilon(q)\|.$$

For the second estimate we compute

$$\begin{aligned}\|\varepsilon(\widehat{q})\| &= \|\widehat{a}^2 + w\widehat{b}^2 + \widehat{c}^2 + w\widehat{d}^2 - 1\| \\ &= \|\varepsilon(q) + (\widehat{a}+a)(\widehat{a}-a) + w(\widehat{b}+b)(\widehat{b}-b) + (\widehat{c}+c)(\widehat{c}-c) + w(\widehat{d}+d)(\widehat{d}-d)\| \\ &\leqslant \|\varepsilon(q)\| + |||q+\widehat{q}||| \, |||q-\widehat{q}||| \\ &\leqslant C\big(\|\varepsilon(q)\|^{3/4} + |||q+\widehat{q}||| \, |||q|||^{1/2}\big)\|\varepsilon(q)\|^{1/4}\end{aligned}$$

where the first inequality is obtained like in the proof of Corollary 4.2, and the second one is (31). Finally estimate (32) is obtained by using $|||\widehat{q}||| \leqslant |||q|||$ from Proposition 3.8, and the bound $\|\varepsilon(q)\| \leqslant \|1\| + \|M(q)\| = \sqrt{\pi} + |||q|||^2$. □

With the estimates of Proposition 4.3 in hand, we can now prove Theorem 4.1.

Proof of Theorem 4.1. For $q \in \mathcal{Q}_n$ we write $\Pi_n(q) = \overline{e_1}\,\overline{e_2} \ldots \overline{e_n}\, r_0$, according to Definition 3.1. Using that $|||eq||| = |||q|||$ for $e \in \mathcal{U}_1$, see (14), one notes that

$$|||q - \Pi_n(q)||| \;=\; |||e_n\, e_{n-1} \ldots e_1\, q_n - r_0||| \,. \tag{33}$$

Denoting next $q_n = q$ and $q_{n-(i+1)} = e_{i+1}\,\chi_{n-i}(q_{n-i})$ as in Definition 3.1, we write a telescopic decomposition

$$\begin{aligned}q_0 = e_n\chi_1(q_1) &= e_n\, q_1 + e_n\,(\chi_1(q_1) - q_1) = \cdots \\ &= e_n\, e_{n-1} \ldots e_1\, q_n + \sum_{i=0}^{n-1} e_n\, e_{n-1} \ldots e_{i+1}\,(\chi_{n-i}(q_{n-i}) - q_{n-i})\end{aligned}$$

rewritten as

$$e_n\, e_{n-1} \ldots e_1\, q_n - r_0 = -\left(\sum_{i=0}^{n-1} e_n\, e_{n-1} \ldots e_{i+1}\,(\chi_{n-i}(q_{n-i}) - q_{n-i})\right) + (q_0 - r_0)\,.$$

The identity (33) and the triangular inequality yield

$$|||q - \Pi_n(q)||| \leqslant \sum_{i=0}^{n-1} |||e_n\, e_{n-1} \ldots e_{i+1}\,(q_{n-i} - \chi_{n-i}(q_{n-i}))||| \;+\; |||q_0 - r_0|||\,.$$

Using again (14) and the fact that $\bar{r}_0 q_0 = M(q_0)^{1/2}$ (still from Definition 3.1), one gets

$$\begin{aligned}|||q - \Pi_n(q)||| &\leqslant \sum_{i=0}^{n-1} |||q_{n-i} - \chi_{n-i}(q_{n-i})||| \;+\; |||\overline{r_0}q_0 - (1,0,0,0)||| \\ &\leqslant C\sum_{i=0}^{n-1} |||q_{n-i}|||^{3/2}\,\|\varepsilon(q_{n-i})\|^{1/4} \;+\; |M(q_0)^{1/2} - 1|\end{aligned}$$

where the last inequality uses (31). Since the correction functions χ_{n-i} are non-increasing in the $|||\cdot|||$ norm (see again Proposition 3.8), one has $|||q_{n-i}||| \leqslant H$ and thus

$$|||q - \Pi_n(q)||| \;\leqslant\; C(H)\sum_{i=0}^{n-1} \|\varepsilon(q_{n-i})\|^{1/4} \;+\; C\,\|\varepsilon(q_0)\|^{1/2}$$

where we have also used that $|\sqrt{M} - 1| \leqslant \sqrt{|M-1|}$ for all $M \geqslant 0$. Using the morphism property $M(e_i\widehat{q}) = M(\widehat{q})$ and Estimate (32), we finally write

$$\|\varepsilon(q_{n-i})\| = \|\varepsilon(\chi_{n-i+1}(q_{n-i+1}))\| \leqslant C(H)\|\varepsilon(q_{n-i+1})\|^{1/4} \leqslant \cdots \leqslant C(H)\|\varepsilon(q)\|^{1/4^i}$$

which, combined with the previous estimate, yields

$$\|\!|q - \Pi_n(q)|\!\| \;\leqslant\; C(H) \sum_{i=0}^{n-1} \|\varepsilon(q)\|^{1/4^i} \;+\; C(H)\|\varepsilon(q)\|^{1/4^{n+1/2}}.$$

This is enough to conclude. □

5. Numerical illustration

To illustrate the properties of our projection algorithm we have implemented a global polynomial approximation method. Given some data $(x_r, y_r)_{r=1,\dots,2n+1}$, our method builds a polynomial with two bounds, $\tilde{p} \in U_{2n}$, such that the values $(\tilde{p}(x_r))_r$ are a good approximation to $(y_r)_r$. For this purpose we begin by interpolating the data $(x_r, y_r)_r$ by their Lagrange polynomial $p \in P_{2n}$, and use p as an effective target function. Ne note that in general p may be outside of the desired bounds.

The method is divided in three stages.

- In the first stage, one computes a polynomial approximation with one lower bound, $p_0 = a^2 + b^2 w \in P_{2n}^+$. The goal is to compute explicitly a, b and not just p_0. Several methods related to this problem have been proposed by the authors in previous contributions [5, 2]. Here, we use another technique described in Appendix A.
- In the second stage we apply the same method as in the first stage to the data $(x_r, 1 - y_r)_{r=1,\dots,2n+1}$. This yields another polynomial $1 - p_1 = c^2 + d^2 w \in P_{2n}^+$ and hence a second approximation p_1 to the data, now with one upper bound.
- The third stage consists of applying the projection algorithm defined in Section 3. From the polynomials (a, b, c, d) this computes $(\tilde{a}, \tilde{b}, \tilde{c}, \tilde{d}) = \Pi_n(a, b, c, d)$ and provides a polynomial approximation with two bounds $\tilde{p} = \tilde{a}^2 + \tilde{b}^2 w$, as described in Corollary 4.2. The minimization of the dual convex problem which is necessary to compute χ_n is performed with a Newton conjugate gradient trust-region algorithm [7]. In our tests, the minimum is reached between 2 and 5 iterations. This operation is repeated n times (see Definition 3.1). The cost of one iteration does not depend on n.

In the following, we take $n = 5$ so that we are looking for approximations of degree 10. On the horizontal axis the values correspond to Chebyshev nodes, $x_r = 0.0051$, 0.0452, 0.1221, 0.2297, 0.3591, 0.5000, 0.6409, 0.7703, 0.8779, 0.9548, 0.9949.

In the first three test cases we choose different values of $(y_r)_r$, so that the corresponding Lagrange polynomials p have larger amplitudes and exceed the desired bounds. The goal here is to compare qualitatively p with the projected polynomial $\tilde{p}$, in order to witness the quality of the projection. The last test case is an experimental error analysis. We project a series of polynomials at given distances from the set U_{2n} and compare the numerical convergence rate with the theoretical result of Theorem 4.1.

5.1. **First test case.** In this first test case we choose $y_r = 0.1500$, 0.2402, 0.1101, 0.0997, 0.9062, 0.5877, 0.5548, 0.1095, 0.8883, 0.6343 and 0.3360. Althougt $y_r \in (0, 1)$, the Lagrange polynomial p may not be within the bounds. Indeed it exceeds the bounds for $x \approx 0.2$ and $x \approx 0.9$. The results are displayed in Figure 1. One observes that as expected $p_0 \geqslant 0$, $p_1 \leqslant 1$. Finally the projected polynomial is truly between 0 and 1, and seems to be a satisfactory approximation of p.

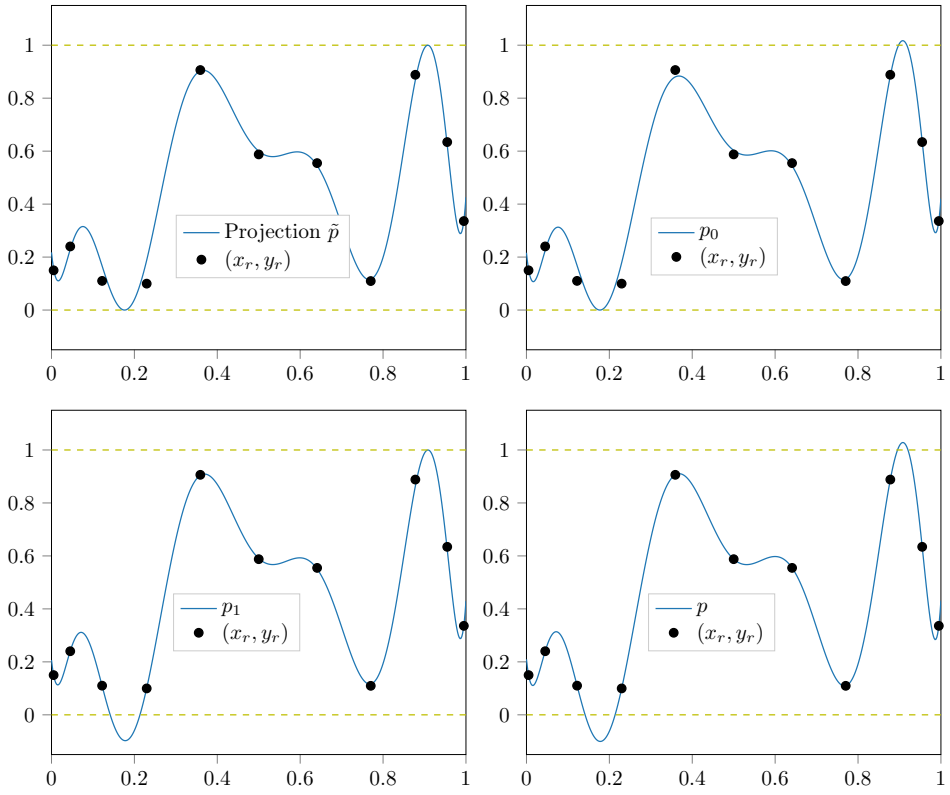

FIGURE 1. **First test case:** Bottom right: Lagrange polynomial p; Bottom left: Upper bound Lukacs approximation; Top right: Lower bound Lukacs approximation; Top left: Projection $\tilde{p}$; Even if the polynomial p_0 and p_1 are marginally out of bounds, a perfect satisfaction of the bounds is observed for $\tilde{p}$.

5.2. **Second test case.** Now $y_r = 0.3326,\ 0.5950,\ -0.0938,\ -0.1245,\ 0.5431,\ 0.8908,\ 1.1076,\ -0.0181,\ 0.5964,\ 0.4571$ and -0.1833. The results are displayed on Figure 2. Despite the large overshoot and undershoot of p_0 and p_1 respectively, one sees that the projected polynomial yields a satisfactory approximation of p.

5.3. **Third test case.** In this third test case $y_r = 0.0114,\ -0.5135,\ 1.3829,\ -0.0664,\ 0.5856,\ -0.5031,\ 0.8059,\ -0.2111,\ 0.9622,\ 1.0676$ and 1.2445. This is a much more severe test in terms of accuracy since the violation of the upper and lower bounds are extreme, and indeed of similar amplitude than the bounds themselves. However we observe in Figure 3 a perfect behavior in terms of satisfaction of the bounds for the projected polynomial, moreover the qualitative profile of the curve seems to be preserved.

5.4. **Fourth test case: error analysis.** In this last numerical test, we want to discuss the estimate of Theorem 4.1 numerically. To proceed we start by defining

$$y_r(t) = t(y_r - \bar{y}) + \bar{y}\,,$$

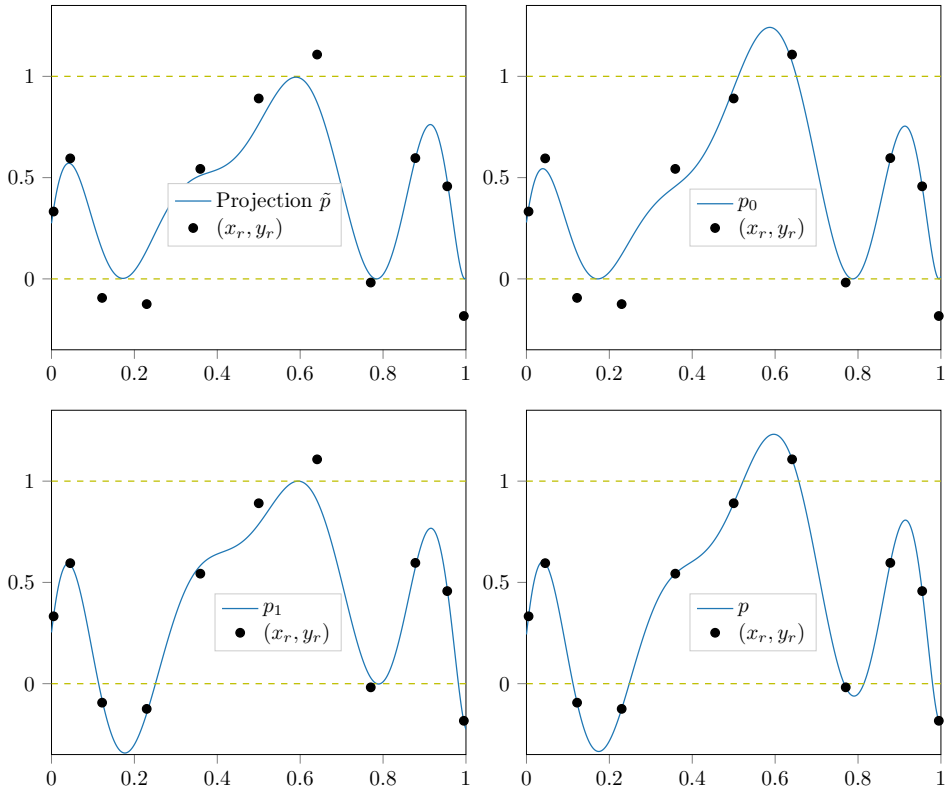

FIGURE 2. **Second test case:** Bottom right: Lagrange polynomial p; Bottom left: Upper bound Lukacs approximation; Top right: Lower bound Lukacs approximation; Top left: Projection $\tilde{p}$; Even if the polynomial p_0 and p_1 are slightly out of bounds, a perfect satisfaction of the bounds is observed for $\tilde{p}$.

where the values y_r are those of the previous test case (Section 5.3), $\bar{y}$ is their average and $t \in [0,1]$. From x_r and $y_r(t)$ we define the associated Lagrange polynomial p_t. Clearly $p_t = tp + (1-t)\bar{y}$ with p the Lagrange polynomial associated with $(x_r, y_r)_r$. Thus, since $\bar{y} \in [0,1]$ and $p \notin [0,1]$ (see Figure 3), there is some $t_* \in (0,1)$ such that $p^{(t)} \in [0,1]$ if $t \leqslant t_*$. Above the critical value of t the polynomial p_t violates the bounds. We denote by q_t the quaternion corresponding to the Lukacs approximations of p_t and we compare $\|\varepsilon(q_t)\|$ and $\|\!|q_t - \Pi_n(q_t)|\!\|$ for various values of t. In our numerical test we chose $t = 1.$, 0.8, 0.6, 0.5, 0.43, 0.38, 0.35, 0.33, 0.32, 0.31, 0.305, 0.301, 0.298, 0.296, 0.294, 0.293, 0.292, 0.291, 0.2908, 0.2906, 0.2904 and 0.29. The results are showed on Figure 4. The slope is approximately equal to 1 in logarithmic scale which suggests that $\|\!|q_t - \Pi_n(q_t)|\!\| = O(\|\varepsilon(q_t)\|)$. This emphasizes that the error estimate of Theorem 4.1 is probably far from being sharp. Moreover, we see on this test case the convergence and stability of the method.

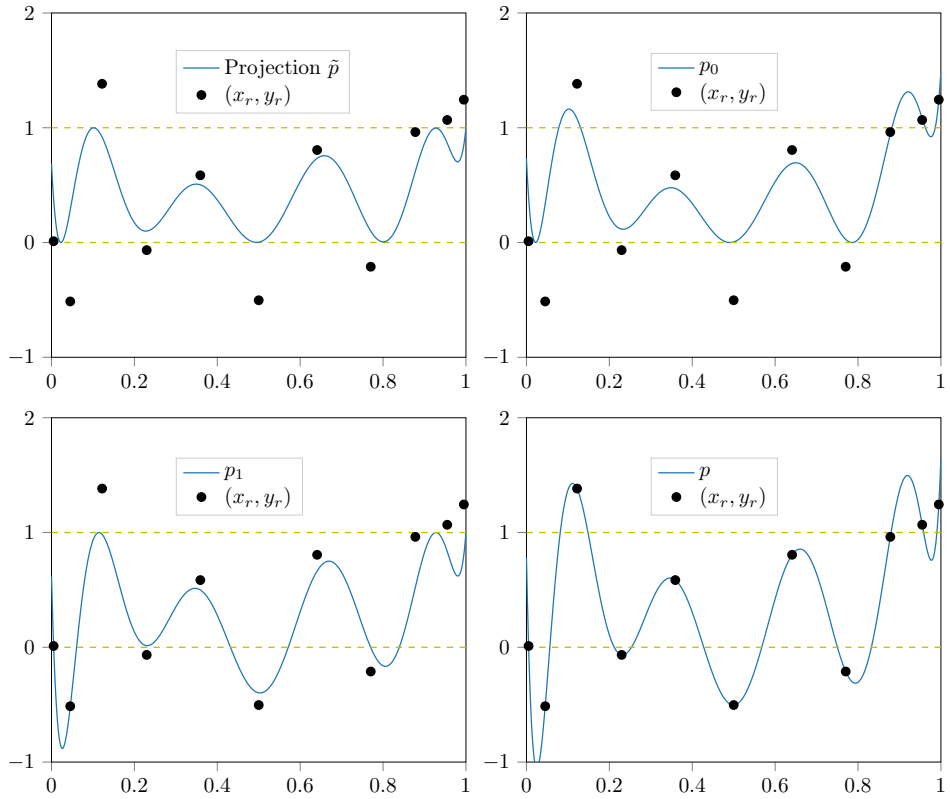

FIGURE 3. **Third test case:** Bottom right: Lagrange polynomial p; Bottom left: Upper bound Lukacs approximation; Top right: Lower bound Lukacs approximation; Top left: Projection $\tilde{p}$; Even if the polynomials p_0 and p_1 are largely out of bounds, a perfect satisfaction of the bounds is observed for $\tilde{p}$.

REFERENCES

[1] Milton Abramowitz and Irene A. Stegun. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55 of *National Bureau of Standards Applied Mathematics Series*. For sale by the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C., 1964.

[2] Martin Campos-Pinto, Frédérique Charles, and Bruno Després. Algorithms For Positive Polynomial Approximation. *SIAM J. Numer. Anal.*, 57(1):148–172, 2019.

[3] B. Despres and M. Herda. Correction to: Polynomials with bounds and numerical approximation [MR3715896]. *Numer. Algorithms*, 77(1):309–311, 2018.

[4] Bruno Després. Polynomials with bounds and numerical approximation. *Numer. Algorithms*, 76(3):829–859, 2017.

[5] Bruno Després and Maxime Herda. Iterative calculation of sum of squares. *arXiv preprint arXiv:1812.02444*, 2018.

[6] Leonhard Euler. Demonstratio theorematis fermatiani omnem numerum sive integrum sive fractum esse summam quatuor pauciorumve quadratorum. *Novi commentarii academiae scientiarum Petropolitanae*, pages 13–58, 1760.

[7] Jorge Nocedal and Stephen J. Wright. *Numerical optimization*. Springer Series in Operations Research and Financial Engineering. Springer, New York, second edition, 2006.

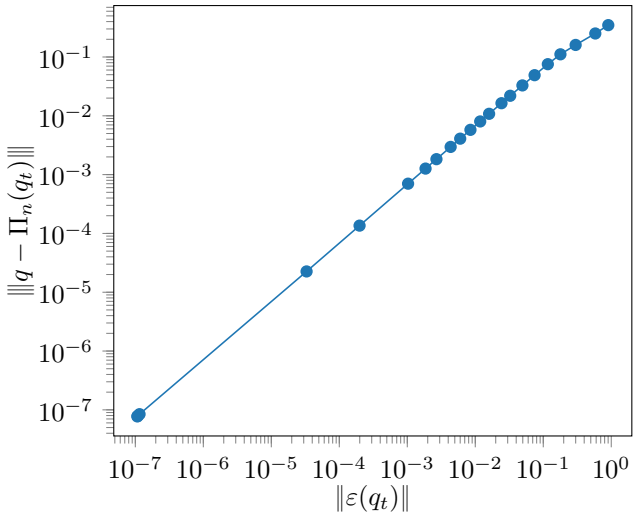

FIGURE 4. **Fourth test case:** Projection error as a function of the error polynomial amplitude $\|\varepsilon(q_t)\|$, which measures the distance of the Lukàcs quaternion q_t to the desired lower and upper bounds. Here the experimental convergence rate ≈ 1 (that is the slope of the line) is much better than what is predicted by Theorem 4.1 which is a slope $\approx \frac{1}{2^{11}}$ for $n = 5$.

[8] Rodrigo B Platte and Lloyd N Trefethen. Chebfun: a new kind of numerical computing. In *Progress in industrial mathematics at ECMI 2008*, pages 69–87. Springer, 2010.

[9] Gábor Szegő. *Orthogonal polynomials*. American Mathematical Society, Providence, R.I., fourth edition, 1975. American Mathematical Society, Colloquium Publications, Vol. XXIII.

APPENDIX A. AN ALGORITHM FOR POSITIVE POLYNOMIAL APPROXIMATION

Here we briefly describe the method used in the numerical tests to compute the positive (or nonnegative) polynomial approximations in Lukacs form. The problem is to find two polynomials $a \in P_n$ and $b \in P_{n-1}$ defining a positive polynomial $p_0(x) = a(x)^2 + b(x)^2 w(x)$ such that given some data $(x_r, y_r)_{r=1,\dots,R}$ (in general with $R = \dim(P_{2n}) = 2n + 1$) the images $(p_0(x_r))_r$ are a good approximation of $(y_r)_r$.

Our algorithm consists in a least-square minimization where a and b are "oscillating polynomials" parametrized by their roots. This parametrization is motivated by the method of [2] where a similar technique has been developped and analysed for positive interpolation.

Mathematically the method relies on the following optimization problem. Find

$$(\alpha^*, \beta^*) \ \in \ \text{argmin}_{\alpha \in \mathbb{R}^{n+1}, \beta \in \mathbb{R}^n} J_t(\alpha, \beta)$$

where the objective function is

$$J(\alpha, \beta) \ = \ \sum_{r=1}^{R} |a[\alpha](x_r)^2 + b[\beta](x_r)^2 w(x_r) - y_r|^2 \, ,$$

with a and b parametrized as follows,

$$a[\alpha](x) \;=\; 2^{n-1}\,\alpha_0 \prod_{i=1}^{n}(x-\alpha_i)\,, \qquad b[\beta](x) \;=\; 2^{n-1}\,\beta_0 \prod_{i=1}^{n-1}(x-\beta_i)\,.$$

The factor 2^{n-1} is taken so that α_0 and β_0 are of the same order as the other components of α and β. Then the approximation polynomial p is defined by

$$p_0(x) = a[\alpha^*](x)^2 + b[\beta^*](x)^2 w(x)$$

The optimization problem is nonlinear and non-convex. However, it can be solved efficiently in practice. Indeed, one can compute explicitely both the gradient and hessian of the functional J. In the numerical tests of Section 5, we used a Newton conjugate gradient trust-region algorithm. The initial couple (α, β) is taken to be appropriate roots of Chebychev polynomials. In this way, the initial polynomials $a[\alpha]$ and $b[\beta]$ are proportional to T_n and U_n, yielding $a[\alpha]^2(x)+b[\beta]^2(x)w(x)$ being some constant polynomial. In all the cases of Section 5, $n = 5$ and the algorithm converges after around 30 iterations.

(M. Campos Pinto, F. Charles, and B. Després) LJLL, Sorbonne Université, CNRS UMR 7598, F-75005, Paris, France

Email address, M. Campos Pinto: campos@ljll.math.upmc.fr

Email address, F. Charles: frederique.charles@ljll.math.upmc.fr

Email address, B. Després: bruno.despres@sorbonne-universite.fr

(M. Herda) Inria, Univ. Lille, CNRS, UMR 8524 – Laboratoire Paul Painlevé, F-59000 Lille, France.

Email address, M. Herda: maxime.herda@inria.fr