Padé-Legendre Interpolants for Gibbs Reconstruction

J.S. Hesthaven[†], S.M. Kaber[‡], and L. Lurati[†]

[†] *Division of Applied Mathematics, Brown University, Box F, Providence, RI 02912, USA*
[‡] *Laboratoire Jacques-Louis Lions, Université Pierre & Marie Curie, 75252 Paris cedex 05, France*

E-mail: Jan.Hesthaven@brown.edu; kaber@ann.jussieu.fr; laural@dam.brown.edu

Dedicated to our friend and mentor, Prof David Gottlieb, on the occasion of his 60^{th} birthday

We discuss the use of Padé-Legendre interpolants as an approach for the postprocessing of data contaminated by Gibbs oscillations. A fast interpolation based reconstruction is proposed and its excellent performance illustrated on several problems. Almost non-oscillatory behavior is shown without knowledge of the position of discontinuities. Then we consider the performance for computational data obtained from nontrivial tests, revealing some sensitivity to noisy data. A domain decomposition approach is proposed as a partial resolution to this and illustrated with examples.

1. INTRODUCTION

The nonuniform pointwise convergence, known as the Gibbs phenomenon, of polynomial approximations to a discontinuous function is a well known and much studied phenomenon, see e.g. [13] and references therein. Among the consequences of the Gibbs phenomenon is the lack of convergence at the jump with an overshoot of approximately 9% of the jump size, a global $\mathcal{O}(N^{-1})$ convergence rate in mean, and a steepness of the approximation right at the jump being proportional to the order, N, of the polynomial expansion.

The literature is rich with methods trying to reduce or even eliminate these problems. The perhaps simplest approach is that of modal filtering, essentially modifying the expansion to make it converge more rapidly in regions sufficiently far away from the discontinuities [13, 16, 22]. An alternative approach is physical space filtering using mollifiers [14, 21], yielding similar behavior, the latter requiring approximate information about the location of the discontinuity. Both methods, however, do not overcome the lack of convergence at the point of discontinuity. To achieve this, information about the exact shock location is needed. With this, the Gibbs phenomenon can be completely resolved [13], albeit this approach has considerable practical problems, i.e. the need to know the location of the discontinuity

exactly and convergence and conditioning problems at high orders of approximation [3, 10].

In this work we discuss the use of rational functions, Padé-Legendre interpolants, as a tool for postprocessing of discontinuous functions represented by classic orthogonal polynomials. As rational functions are richer than simple polynomial representations, one can hope that the impact of the discontinuity will be less severe and, thus, that one can use this as a postprocessing tool to reduce the impact of the Gibbs phenomena in polynomial expansions.

Similar efforts has been pursued recently, see e.g. [5, 6, 8, 9, 19]. However, most of these previous efforts have dealt with exact expansions and simple functions. An exception is [5] in which a rational approximation was used to post-process the pseudo-spectral Fourier solution of Burgers' equation and an incompressible Boussinesq convection flow. Furthermore, most previous work is based on Fourier-Padé methods [5, 6, 9]. Although [8, 19] also deal with functions based on orthogonal polynomials as in this paper, they consider only simple functions and denominators of very low order.

In this work we shall discuss a number of different aspects. First, we shall not assume knowledge of expansions but rather of point values, leading to interpolatory rational functions. We discuss the interpolants in detail and present an efficient algorithm for computing the Padé interpolant of a function. Secondly, we shall investigate the behavior of these interpolants for problems with inexact or noisy data, e.g., reconstructions based on computed data only.

In Section 2 we shall recall basic properties of the Legendre polynomials, Gauss quadratures, and associated interpolation polynomials while Section 3 defines the rational Padé-Legendre interpolants. This sets the stage for Section 4 where we discuss the computational construction of the Padé-Legendre interpolants, leading to an algorithm that requires only point values of the function, in contrast with [6, 8, 20] which require knowledge of the expansion coefficients. In Section 5 we illustrate the performance of these interpolants for the postprocessing of polynomial representations of functions with limited regularity. We also discuss in detail the impact of noise in the data, e.g., due to computations, and propose a multi-domain approach to the reconstruction in such cases. We extend the test cases to include the postprocessing of pseudo-spectral solutions of both Burgers' equation and the 1-D shock entropy problem. Section 6 contains a few concluding remarks.

2. LEGENDRE POLYNOMIALS AND EXPANSIONS

Let $L^2(–1,1)$ be the space of measurable functions, u, such that the integral $\int_{-1}^{1} |u(x)|^2 dx$ is finite. Equipped with the scalar product

$$\langle u, v\rangle := \int_{-1}^{1} u(x)v(x)dx, \tag{1}$$

$L^2(–1,1)$ is a Hilbert space. The norm derived from this scalar product (1) is denoted $\|u\| := \sqrt{\langle u, u\rangle}$. A Hilbert basis of $L^2(–1,1)$ is given by the Legendre polynomials $P_n(x)$. These polynomials of order n are defined by

$$\forall (n,m) \in \mathbb{N}\times\mathbb{N}, \quad \langle P_n, P_m\rangle = \frac{1}{n+1/2}\delta_{n,m}, \tag{2}$$

with δ being the Kronecker symbol. Every function in $L^2(-1,1)$ has an L^2-convergent expansion in Legendre polynomials

$$\mathfrak{L}(u) := \sum_{n=0}^{\infty} \hat{u}_n P_n,$$

where $\hat{u}_n$ is the n-th Legendre coefficient of u

$$\hat{u}_n := \frac{1}{\|P_n\|^2}\langle u, P_n\rangle = (n+\frac{1}{2})\int_{-1}^{1} u(x)P_n(x)\,dx. \tag{3}$$

For $N \in \mathbb{N}$, one defines the truncated series

$$\mathfrak{L}_N(u) := \sum_{n=0}^{N} \hat{u}_n P_n.$$

By the orthogonality properties of the Legendre polynomials, $\mathfrak{L}_N(u) - u$ is orthogonal to $\mathbb{P}_N$:

$$\forall p \in \mathbb{P}_N \quad : \quad \langle u - \mathfrak{L}_N(u), p\rangle = 0 \tag{4}$$

where we denote by $\mathbb{P}_N$ the set of algebraic polynomials of degree less than or equal to N.

The rate of convergence of the error $\|\mathfrak{L}_N(u) - u\|$ is related solely to the smoothness of the function u, e.g., if u belongs to the Sobolev space $H^s(-1,1)$, $s \geq 0$, there exists a constant c (depending solely on s) such that (consult [2] for instance)

$$\forall N \in \mathbb{N} \quad : \qquad \|\mathfrak{L}_N(u) - u\| \,\leq\, cN^{-s}\|u\|_{H^s}.$$

Hence for smooth u, the polynomial $\mathfrak{L}_N(u)$ is a very accurate approximation of u as N goes to ∞.

However, for discontinuous functions or functions only belonging to H^s with $s < 1/2$, the expansion converges slowly and lacks pointwise convergence, a manifestation of the Gibbs phenomenon.

2.1. Legendre Quadratures

For a given $N \in \mathbb{N}$, we shall consider the Gauss quadrature

$$\int_{-1}^{1} \varphi(x)dx = \sum_{j=1}^{M} \varphi(\tilde{x}_j)\tilde{\omega}_j + \sum_{j=0}^{N-M} \varphi(x_j)\omega_j. \tag{5}$$

where the M nodes, $\tilde{x}_j$, are predefined, leaving a total of $2N - M + 2$ degrees of freedom to maximize the accuracy of the summation.

For $M = 0$, one recovers the maximally accurate classic Gauss quadrature in which case the $N+1$ nodes are given as the roots of $P_{N+1}(x)$. Another important case is that of $M = 2$ and $\tilde{x}_j = \pm 1$ in which case the remaining $N - 1$ nodes are found as the roots of $P'_N(x)$. All computed points are, in both cases, entirely inside the computational domain, i.e.,

$$-1 < x_0 < x_1 < \cdots < x_{N-M-1} < x_{N-M} < 1 \ .$$

and satisfy a symmetry relation

$$\forall j = 0, \ldots, N-M: \qquad x_j = -x_{N-M-j}.$$

Once the integration nodes are known, the weights can be found by requiring that the quadrature be exact for all polynomials up to order N. However, the true power of the Gaussian quadrature emerges when recalling that (5) is in fact exact for all polynomials $\varphi \in \mathbb{P}_{2N+1-M}$.

Associated with the quadrature formula (5) is the discrete scalar product defined for all continuous functions φ and ψ by

$$\langle \varphi, \psi \rangle_N = \sum_{j=1}^{M} \varphi(\tilde{x}_j)\psi(\tilde{x}_j)\tilde{\omega}_j + \sum_{j=0}^{N-M} \varphi(x_j)\psi(x_j)\omega_j, \tag{6}$$

and the associated discrete norm, $\|\varphi\|_N^2 = \langle \varphi, \varphi \rangle_N$. Due to the accuracy of the Gauss quadrature, if the function $\varphi\psi$ belongs to $\mathbb{P}_{2N+1-M}$ then the discrete and continuous inner product coincide. However, for $M = 2$, this implies that for $\varphi\psi \in \mathbb{P}_{2N}$, the Gauss Lobatto quadrature is no longer exact, i.e., the discrete and continuous norms are not identical.

2.2. Interpolation

Let us in the following refer to x_j as a generic node, be it specified or computed as part of the quadrature construction. Associated with the nodes, $(x_j)_{j=0}^N$, are the Lagrange interpolants $(\ell_j)_{j=0}^N$ defined for $j = 0, \cdots, N$ by

$$\ell_j \in \mathbb{P}_N \quad \text{and} \quad \ell_j(x_k) = \delta_{j,k}, \quad \forall k = 0, \cdots, N. \tag{7}$$

The interpolation of a function u at the quadrature points takes the simple form

$$\mathcal{I}_N(u) := \sum_{j=0}^{N} u(x_j)\ell_j(x).$$

For a smooth u, the polynomial $\mathcal{I}_N(u)$ is a very accurate approximation of u, e.g., if u belongs to the Sobolev space $H^s(–1, 1)$ with $s > \frac{1}{2}$ (see [2]) there exists a constant c such that for all $N \in \mathbb{N}$

$$\|\mathcal{I}_N(u) - u\| \ \leq \ cN^{-s+1/2}\|u\|_{H^s}. \tag{8}$$

An alternative, or dual, form of $\mathcal{I}_N(u)$ is given as

$$\mathcal{I}_N(u) = \sum_{j=0}^{N} \tilde{u}_j P_j,$$

with the discrete Legendre coefficients $\tilde{u}_j$ defined by

$$\tilde{u}_j := \widehat{\mathcal{I}_N(u)}_j = \frac{\langle \mathcal{I}_N(u), P_j \rangle_N}{\|P_j\|_N^2}. \tag{9}$$

For the Gauss quadrature, the accuracy of the quadrature formula implies

$$\tilde{u}_j = (j+1/2)\langle \mathcal{I}_N(u), P_j\rangle_N = (j+1/2)\langle u, P_j\rangle_N. \tag{10}$$

This is, however, not the case for the Gauss-Lobatto quadrature, although the difference is minimal for most practical purposes [15].

3. THE PADÉ-LEGENDRE INTERPOLATION

Given integers N, M, and L, we seek to define a rational function $\mathcal{P}/\mathcal{Q}$ which interpolates a given function u at $N+1$ collocation points. We shall assume that $\mathcal{P} \in \mathbb{P}_M$ and $\mathcal{Q} \in \mathbb{P}_L$.

The main reason for seeking the interpolation of a function, u, by a rational function $\mathcal{R} = \mathcal{P}/\mathcal{Q}$ is based on the hope that the poles of $\mathcal{R}$ are close (in the complex plane) to the singularities of u. In that case the rational approximation can capture the "structure" of u and, thus, lead to an improved representation of u. Furthermore, as the polynomial representation is a special case of the rational form, no deterioration is expected for smooth problems.

Assume that $N \geq 1$. Then

Definition 3.1. Given integers M and L, the pair of polynomials $(\mathcal{P}, \mathcal{Q}) \in \mathbb{P}_M \times \mathbb{P}_L$ is said to be a solution of the (N, M, L) Padé-Legendre interpolation problem of a given function u if $\mathcal{Q}$ has a constant sign on $[-1, 1]$, i.e.,

$$\forall x \in [-1, 1] : \qquad \mathcal{Q}(x) > 0 \tag{11}$$

and

$$\forall \varphi \in \mathbb{P}_N : \qquad \langle \mathcal{P} - \mathcal{Q}u, \varphi\rangle_N = 0. \tag{12}$$

Equation (12) forms a linear system of $N+1$ equations and $M+L+2$ unknowns (the coefficients of $\mathcal{P}$ and $\mathcal{Q}$ in some polynomial basis). This system always has a non trivial solution if $M + L \geq N$.

Equation (12) is motivated by previous work in which $\mathcal{P} - \mathcal{Q}u$ is required to be orthogonal to $\mathbb{P}_N$ using a continuous inner product. We refer to [8] and [20] for Legendre expansions, to [19] for Chebyshev expansions, to [5] for Fourier expansions, and to [17] for the general Jacobi case, although previous work utilizing orthogonal polynomials have only considered the case with L being very small.

If (11)-(12) has a solution, we define the rational function

$$\mathcal{R}(u) := \frac{\mathcal{P}}{\mathcal{Q}}. \tag{13}$$

Let us be a bit more precise about the interpolation properties of $\mathcal{R}(u)$.

Remark. [Interpolation] Suppose there exists a solution $(\mathcal{P}, \mathcal{Q})$ of the (N, M, L) Padé-Legendre interpolation problem (11)-(12). Taking in (12) $\varphi \in \mathbb{P}_N$ to be a Lagrange polynomial ℓ_j based on the quadrature points x_j we get the relations

$$\forall j = 0, \cdots, N \quad : \qquad (\mathcal{P} - \mathcal{Q}u)(x_j) = 0.$$

Since $\mathcal{Q}(x_j) \neq 0$, the rational function $\mathcal{R}(u)$ interpolates u at x_j i.e.,

$$\forall j = 0, \cdots, N \quad : \qquad \mathcal{R}(u)(x_j) = u(x_j). \tag{14}$$

The converse is of course likewise true: if $\mathcal{R}(u) = \dfrac{\mathcal{P}}{\mathcal{Q}}$ with positive denominator, satisfies the interpolation properties (14), then the pair $(\mathcal{P}, \mathcal{Q})$ is a solution of the $(N, \text{degree}(\mathcal{P}), \text{degree}(\mathcal{Q}))$ Padé-Legendre interpolation problem.

The main question is the existence of a pair $(\mathcal{P}, \mathcal{Q})$. Let us first consider the simple case $L = 0$, i.e., the denominator is a (non zero) constant. If $N = M$, the numerator equals this constant times $\mathcal{I}_N(u)$, i.e., the interpolation of u at the grid points. This is known to be a bad approximation if the function u is not regular enough as shown in Eq.(8). If $M < N$ there could be no solution and if $M > N$, there are an infinite number of solutions, $\mathcal{P}$. In this work, we shall only concern ourselves with the cases $M \geq 1$, $M \leq N$, and $L \geq 0$.

Thus uniqueness can easily be controlled with some restrictions on the parameters M and L.

PROPOSITION 3.1 (Uniqueness). *Assume $M + L \leq N$. Then a solution of* (11)-(12) *is unique in the sense that it defines a unique rational approximation (13).*

Proof. If $(\mathcal{P}_1, \mathcal{Q}_1)$ and $(\mathcal{P}_2, \mathcal{Q}_2)$ are two solutions, then by the Remark above

$$\forall j = 0 \cdots, N \quad : \qquad \frac{\mathcal{P}_1}{\mathcal{Q}_1}(x_j) = \frac{\mathcal{P}_2}{\mathcal{Q}_2}(x_j) = u(x_j).$$

The polynomial $\mathcal{P}_1\mathcal{Q}_2 - \mathcal{P}_2\mathcal{Q}_1 \in \mathbb{P}_{M+L}$ vanishes at the $N + 1$ different points x_j. The assumption $M + L \leq N$ implies that $\mathcal{P}_1\mathcal{Q}_2 - \mathcal{P}_2\mathcal{Q}_1 \equiv 0$. □

4. COMPUTATION OF THE RATIONAL APPROXIMATION

In the following we shall discuss in some detail the actual computation of the rational interpolant. From now on we shall assume that

$$N = M + L. \tag{15}$$

Remark. Assuming (15), we know that there exists a non trivial solution $(\mathcal{P}, \mathcal{Q})$ of the Padé-Legendre interpolation problem (12). If the denominator $\mathcal{Q}$ never vanishes then we deduce from Proposition 3.1, that the rational approximant $\mathcal{R}(u)$ is unique.

The main difficulty lies in the computation of the denominator. Once this is done, the computation of the numerator is straightforward.

Let us assume for a moment that the denominator $\mathcal{Q}$ has been computed. Taking $\varphi = P_n$ in (12) with $n = 0, \cdots, N$, we get

$$\langle \mathcal{Q}u, P_n \rangle_N = \langle \mathcal{P}, P_n \rangle_N = \|P_n\|_N^2 \tilde{p}_n,$$

from which the Legendre coefficients of $\mathcal{P}$ follows directly. Thus, once $\mathcal{Q}$ is known, $\mathcal{P}$ follows by a quadrature.

For the computation of the denominator, observe first that assumption (15) implies

$$2M + L = M + N \leq 2N \tag{16}$$

Taking $\varphi = P_n$ in (12) with $n = M+1, \cdots, L+M$, we get

$$\langle \mathcal{Q}u, P_n \rangle_N = \langle \mathcal{P}, P_n \rangle_N \ .$$

If, for simplicity, we restrict the attention to the use of Gauss quadratures, we have

$$\langle \mathcal{Q}u, P_n \rangle_N = \int_{-1}^{1} \mathcal{P}(x) P_n(x)\, dx \ .$$

By the orthogonality properties of the Legendre polynomials we recover

$$\langle \mathcal{Q}u, P_n \rangle_N = 0, \qquad \forall n = M+1, \cdots, M+L. \tag{17}$$

If the approach is based on the Gauss-Lobatto quadrature, a slight modification is required for $n = M + L$ but the outcome is essentially the same.

Hence the problem of finding the denominator can be stated as follows: find $\mathcal{Q} \in \mathbb{P}_L$ such that the L discrete Legendre coefficients of the function $\mathcal{Q}u$

$$\left(\widetilde{(\mathcal{Q}u)}_n \right)_{n=M+1}^{M+L}$$

vanish.

Let us write the denominator

$$\mathcal{Q} = \sum_{m=0}^{L} \tilde{q}_m \varphi_m \in \mathbb{P}_L, \tag{18}$$

the polynomials φ_m being either P_m or x^m. To compute the $L+1$ coefficients of the denominator $\mathcal{Q}$ from relations (17), we have at our disposal L equations, i.e., a non trivial solution of this linear system always exists. Inserting the expansion (18) into (17), we end up with a linear system to solve to determine the denominator

$$n = M+1, \cdots, M+L \Longrightarrow \widetilde{(\mathcal{Q}u)}_n = 0.$$

Let us define the vector $q^{(L)} = (\tilde{q}_0, \cdots, \tilde{q}_L)^T$ and the $L \times (L+1)$ matrix with entries depending on the function u, the basis $(\varphi_m)_{m=0}^L$ and the parameters N, M, and L

$$A = \begin{bmatrix} \langle u\varphi_0, P_{M+1} \rangle_N & \cdots & \langle u\varphi_L, P_{M+1} \rangle_N \\ \vdots & & \vdots \\ \langle u\varphi_0, P_{M+L} \rangle_N & \cdots & \langle u\varphi_L, P_{M+L} \rangle_N \end{bmatrix}. \tag{19}$$

The linear system to be solved is $Aq^{(L)} = 0$.

Remark. In the following two cases, A is the null matrix.

1. If u is a polynomial of $\mathbb{P}_{M-L}$, then A is the null matrix. Thus, if a general function u is replaced by $u+v$ with $v \in \mathbb{P}_{M-L}$, the matrix A is unchanged.

We only consider in this work functions u with "high modes" which is typically the case for discontinuous functions.

2. If u vanishes at all the grid points, the matrix of the system is also the null matrix. As discussed below, shifting u as $u+\lambda$ will not change this.

Remark. If $u(x) = \ell_i(x)$, i.e., the Lagrange polynomial associated with x_i, then A is a rank-1 matrix. For $\varphi_m = x^m$, this is immediate and for $\varphi_m = P_m$ it is a consequence of the three-term recurrence relation associated with the Legendre polynomials. This rules out the use of a Padé interpolant of the usual Lagrange polynomial as the basic building block for a spectral method [15].

The matrix A is the product of three matrices $A = BCD$ with $B \in \mathbb{R}^{L,N+1}$, $C \in \mathbb{R}^{N+1,N+1}$, and $D \in \mathbb{R}^{N+1,L+1}$ defined by

$$\begin{cases} B_{i,k} = P_{M+i}(x_k)\omega_k & 1 \le i \le L, \quad 0 \le k \le N. \\ C_{k,j} = u(x_k)\delta_{k,j} & 0 \le k \le N, \ 0 \le j \le N. \\ D_{k,j} = \varphi_j(x_k) & 0 \le k \le N, \ 0 \le j \le L. \end{cases}$$

The matrices B and D depend only on the parameters N, M and L :

$$B = \begin{bmatrix} P_{M+1}(x_0)\omega_0 & \cdots & P_{M+1}(x_N)\omega_N \\ \vdots & & \vdots \\ P_{M+L}(x_0)\omega_0 & \cdots & P_{M+L}(x_N)\omega_N \end{bmatrix},$$

D is a Vandermonde matrix

$$D = \begin{bmatrix} \varphi_0(x_0) & \cdots & \varphi_L(x_0) \\ \vdots & & \vdots \\ \varphi_0(x_N) & \cdots & \varphi_L(x_N) \end{bmatrix}$$

and $C = diag(u(x_0), \cdots, u(x_N))$ is a function of u and N. We list below some basic properties of the three matrices.

If $u > 0$, the matrix C is clearly regular. If u changes sign, shifting it by a factor $\lambda \in \mathbb{R}$ such that $u + \lambda > 0$ does not help. This is a consequence of the next Proposition from which we deduce that $A(u+\lambda) = A(u) + \lambda BD = A(u)$.

PROPOSITION 4.1. *For $L \le \min(N-1, M+1)$, the matrix BD is the null matrix of $\mathbb{R}^{L,L+1}$.*

Proof. For $n = 1, \cdots, L$ and $m = 0, \cdots, L$, the polynomial $\sum_{j=0}^{N} P_{M+n}\varphi_m$ has degree $M + 2L = L + N \le 2N$. Hence the exactness of the Gauss quadrature formula implies

$$(BD)_{n,m} = \sum_{j=0}^{N} B_{n,j}D_{j,m} = \sum_{j=0}^{N} P_{M+n}(x_j)\varphi_m(x_j)\omega_j = \int_{-1}^{1} P_{M+n}(x)\varphi_m(x)dx \ ,$$

and the result follows from the orthogonality of P_{M+n} and $\mathbb{P}_m$. □

PROPOSITION 4.2. *The matrix B has maximal rank: $rank(B) = L$ and $dim(Null(B)) = M + 1$.*

Proof. The rank of B is the dimension of the linear space spanned by the rows of u. Let us denote by θ_l the l'th row of the matrix B and let $\sum_{l=1}^{L} x_l\theta_l = 0$ be a linear combination of the θ's. Then the polynomial $\sum_{l=1}^{L} x_l P_{M+l} \in \mathbb{P}_{M+L}$ vanishes at the $N+1(> M+L)$ collocation points. Hence it is the null polynomial. It follows from the rank Theorem that $\dim Null(B) = N + 1 - rank(B) = N + 1 - L = M + 1$. □

PROPOSITION 4.3. *Assume $L \leq N$, then*

$$Null(D) = \{0\}, \quad rank(D) = L + 1$$

Proof. For $x \in \mathbb{R}^{L+1}$, $Dx = 0$ implies that

$$\forall j = 0, \cdots, N \quad : \qquad \sum_{l=0}^{L} \varphi_l(x_j)x_l = 0.$$

Hence

$$0 = \sum_{j=0}^{N} |\sum_{l=0}^{L} x_l\varphi_l(x_j)|^2\omega_j = \|\sum_{l=0}^{L} x_l\varphi_l\|_N^2,$$

which implies $x = 0$ and the matrix D is injective. □

5. NUMERICAL TESTS

In this section we apply the Padé-Legendre interpolant to several functions with different degrees of smoothness. We first consider the interpolation of functions where exact values of the function at quadrature points are known. Subsequently, we move on to the application of Padé-Legendre interpolation and reconstruction of data obtained by spectral methods and conclude by providing some very general guidelines for finding a Padé interpolant; i.e., the order of the numerator and denominator.

5.1. Padé-Legendre reconstruction of functions

We first consider the ability of the Padé-Legendre interpolation to reconstruct functions given on exact form. This is clearly the simplest test and shall be used to illustrate basic properties of the scheme.

The general setting for all tests are

- The parameters N, M, and L satisfy the condition (15).
- All the computations have been done using the canonical basis $\varphi_m(x) = x^m$ in the computation of the matrix A in (19). The system $Aq^{(L)} = 0$ is solved by finding the null space of the matrix A, which can be computed by any classical linear algebra toolbox.

– If dim null $A = 1$, the non zero vector $q \in \text{null}\, A$ defines the denominator $\mathcal{Q}$. If the first component of q is non zero, one can ensure $\mathcal{Q}(0) > 0$, which is a minimal requirement.

– If dim null $A \geq 2$, we take any two non zero linearly independent vectors q_1 and q_2 of null A and define $q = \alpha_1 q_1 + \alpha_2 q_2$ such that $\mathcal{Q}(0) > 0$ and $\mathcal{Q}(1) > 0$. This is always possible if the first component of q_1 or that of q_2 is non zero and if the sum of the components of q_1 or that of q_2 is non zero.

- All the graphics are plotted on a uniform grid of 200 points.

As a first simple test, we validate the ability of the scheme to reproduce polynomials. For $u_1(x) = 8\,x^7 - 5.33\,x^6 + 14\,x^4 + x^3 - 9$ with $L = 0$ and $N = M = 7$ we reproduce u_1 perfectly as shown in Table 1.

As a likewise simple second example, we confirm the ability to reproduce rational functions with the test function $u_2(x) = \dfrac{4x^5 - x^4 + x - 1}{x^2 - x + 3}$. Here also, the function is perfectly reproduced once L equals the degree of the denominator as illustrated in Table 1.

For the smooth nonpolynomial function

$$u_3(x) = e^x \sin(2\pi x),$$

Table 1 shows the output of several results. The maximum error is of the same order as $\mathcal{I}_N(u_3) - u_3$ which is very small since u_3 is a smooth function. Note that for $L = 1$ and $L = 2$, the denominator does not degenerate to a constant function.

Consider now a continuous function with a discontinuous derivative,

$$u_4(x) = |x|.$$

We compute the polynomial interpolation $\mathcal{I}_N(u_4)$ and a rational approximation $\mathcal{R}(u_4)$. The three curves u_4, $\mathcal{R}(u_4)$ and $\mathcal{I}(u_4)$ seem to collapse but a zoom around the discontinuity in Fig. 1 shows the precision of the rational approximation. In Fig. 2, we display the pointwise error in log scale: the error decreases very fast at points far from the discontinuous derivative at $x = 0$. Furthermore, the decay rate clearly improves with the order, L, of the denominator. This was found also in [7, 8] where similar results were found for low order of the denominator ($L \leq 2$). The results presented here confirm that this trend continues also for higher degrees of the denominator.

Consider now the discontinuous function

$$u_5(x) = sign(x).$$

The polynomial interpolation illustrates the Gibbs phenomenon, see Fig. 3. The Padé-Legendre reconstruction essentially removes the oscillations except near the discontinuity where the over/under shoots are still present, but severely reduced. In Fig. 4, we display the pointwise error in log scale: the polynomial error is quite uniformly distributed on the whole interval, while the error of the rational interpolant decreases rapidly at points further away from the discontinuity $x = 0$. Furthermore, the decay rate clearly depends on the order, L, of the denominator.

TABLE 1
The error $\max_{-1\le x\le 1} |\left(\mathcal{R}(u_i) - u_i\right)(x)|$ is computed on a uniform grid of 200.

function	N	M	L	$\|\text{Error}\|_\infty$
$u_1(x) = 8\,x^7 - 5.33\,x^6 + 14\,x^4 + x^3 - 9$	7	7	0	1.9380(−12)
$u_2(x) = \dfrac{4x^5 - x^4 + x - 1}{x^2 - x + 3}$	7	6	1	4.7787(−02)
	7	5	2	2.4758(−13)
$u_3(x) = e^x \sin(2\pi x)$	15	15	0	2.1239(−05)
	15	14	1	3.3930(−05)
	15	13	2	8.1673(−06)
	31	31	0	2.0828(−13)
	31	30	1	5.4599(−13)
	31	29	2	2.0473(−13)
$u_4(x) = \|x\|$	31	31	0	2.5993(−02)
	31	29	2	2.0027(−02)
	31	27	4	1.7713(−02)
	63	63	0	1.0879(−02)
	63	61	2	7.9755(−03)
	63	59	4	6.8363(−03)
$u_5(x) = sign(x)$	31	31	0	8.7800(−01)
	31	29	2	8.1496(−01)
	63	63	0	7.6095(−01)
	63	61	2	6.4178(−01)
	127	127	0	5.3290(−01)
	127	125	2	3.5419(−01)
$u_6(x) = \begin{cases} 1 & \text{for } x \in [-1, -0.7[\\ 1 + x + \sin(2\pi e^x) & \text{for } x \in]-0.7, -0.2[\\ x + \sin(2\pi e^x) & \text{for } x \in]-0.2, 0.7[\\ 0 & \text{for } x \in]0.7, 1] \end{cases}$	127	127	0	5.7962(−01)
	127	123	4	2.1880(+00)
	127	121	6	4.1894(−01)

This is fully consistent with the analysis in [17] where it is shown that the maximal overshoot is 0.8% of the jumpsize, i.e., an order of magnitude less than for the pure polynomial approximation. Furthermore, it is conjectured in [8] for the continuous Padé approximation that the order of approximation away from the shock is M^{-L}, which is very similar to what is observed here. Finally, the analysis in [17] shows that if $L \propto M$ then the maximal gradient at $x = 0$ grows like $M^{3/2}$ which is a significant improvement over the classic polynomial result in which the gradient grows only linearly in M. The analysis leading to these results is given in detail in [17].

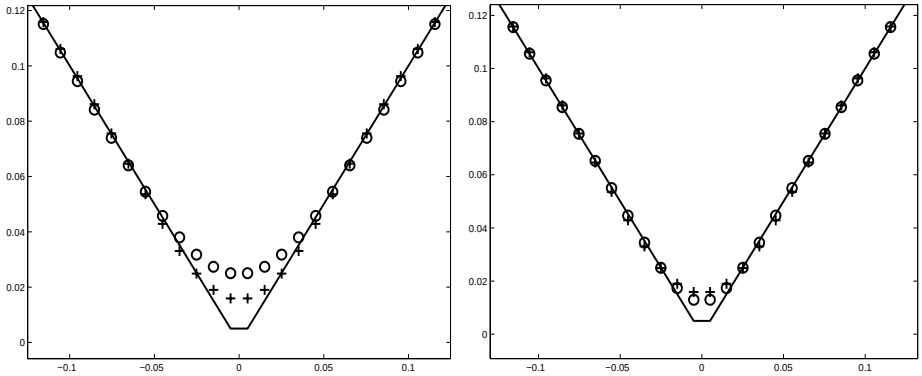

FIG. 1. Polynomial interpolation (+) and rational interpolation (o) of u_4. Left we use $(N, M, L) = (31, 29, 2)$. On the right, $(N, M, L) = (63, 61, 2)$.

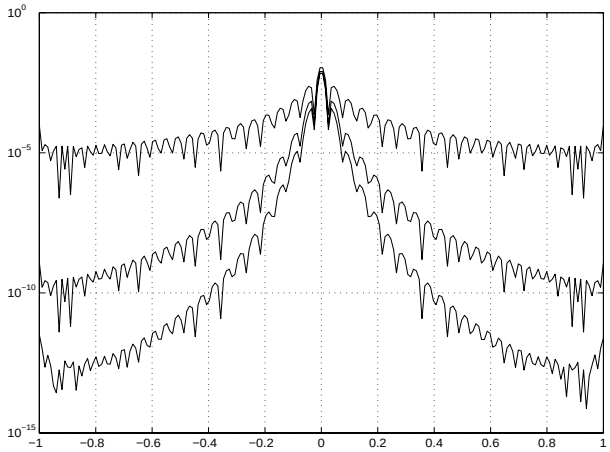

FIG. 2. Logarithm of the pointwise error: $|u_4 - \mathfrak{I}_N(u_4)|$ (top) and $|u - \mathcal{R}_{M,L}(u_4)|$ $N = 63$, $L = 2$ (center) and $L = 4$ (bottom).

In the last example, we consider a discontinuous function with a more complicated shape

$$u_6(x) = \begin{cases} 1 & \text{for } x \in [-1, -0.7[\\ 1 + x + \sin(2\pi e^x) & \text{for } x \in]-0.7, -0.2[\\ x + \sin(2\pi e^x) & \text{for } x \in]-0.2, 0.7[\\ 0 & \text{for } x \in]0.7, 1]. \end{cases}$$

The polynomial interpolation of u_6, based on $N + 1 = 128$ Gauss-Legendre points, is displayed in Fig. 5 which also shows the rational approximation with $L = 4$. The improvement is clear and smoothness is regained except in a narrow region close to the discontinuities. This is expected as we cannot expect to improve the accuracy of the expansion at points of discontinuity. Increasing L to 6, however, produces a very accurate solution even in the neighborhood of the discontinuities as displayed in Fig. 5. The pointwise error in log scale is also shown in Fig. 5,

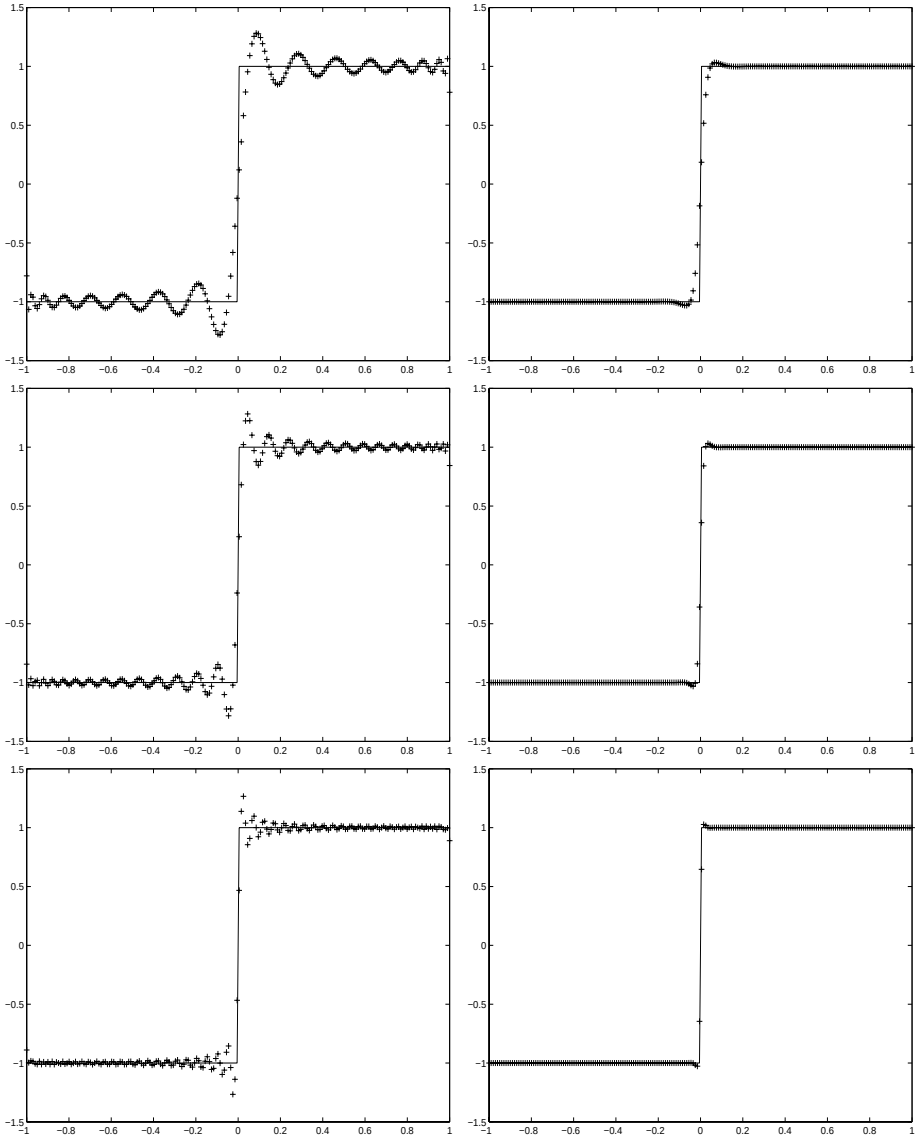

FIG. 3. Interpolation of u_5. In the left column is pure polynomial interpolation with $N = 31$, $N = 63$, and $N = 127$, respectively. The right column represents the Padé-Legendre interpolation with $(N, M, L) = (31, 29, 2)$, $(N, M, L) = (63, 61, 2)$, and $(N, M, L) = (127, 125, 2)$, respectively.

illustrating the clear advantage of even a low order denominator as compared to the pure polynomial interpolation.

5.2. Postprocessing of Computational data

In this section results of using Padé-Legendre reconstruction as a postprocessor on computational data are shown. All data sets were computed using spectral methods and since both examples involve the time evolution of a shock, some filtering is used

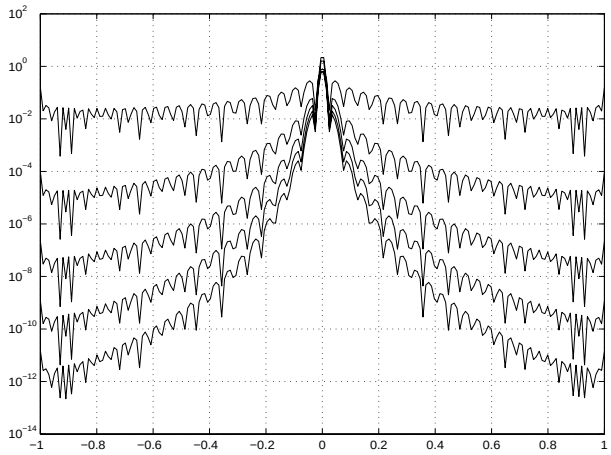

FIG. 4. Function u_5: logarithm of the pointwise error. $N = 63$, $L = 0$ (top), 1, 2, 3 and $L = 4$ (bottom).

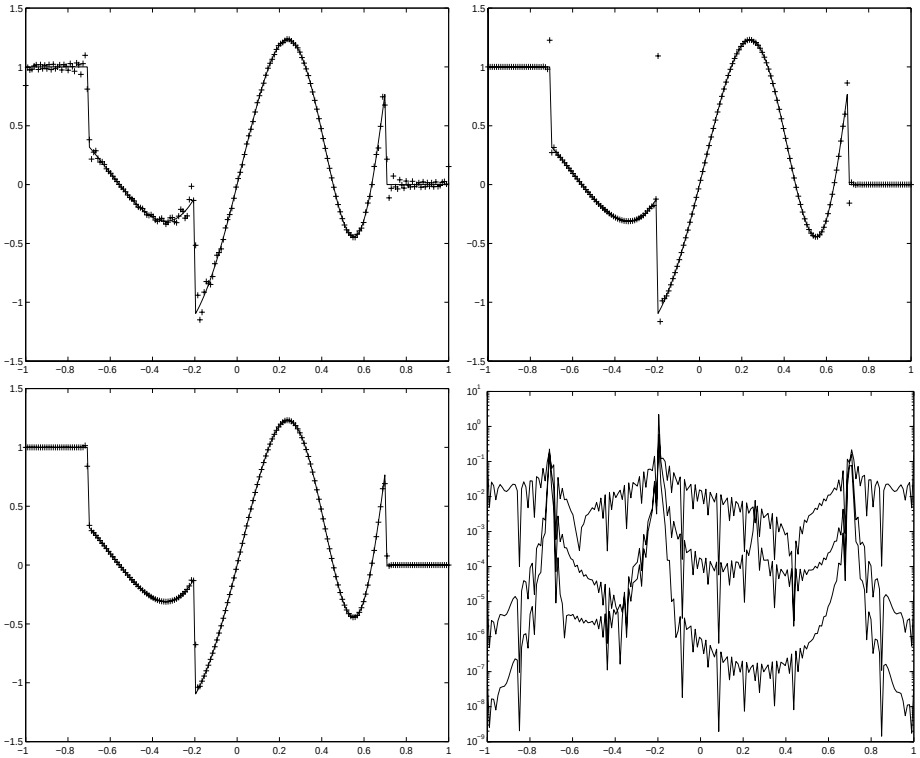

FIG. 5. Interpolation of u_6. Top left is the polynomial interpolation with $N = 127$. Top right is Padé-Legendre interpolation with $(N, M, L) = (127, 123, 4)$ and bottom left with $(N, M, L) = (127, 121, 6)$. On the bottom right is shown the logarithm of the pointwise error for $N = 127$, $L = 0$ (top), $L = 4$ (middle) and $L = 6$ (bottom).

as a method of stabilization [15]. The focus of these examples, however, is not on how the data was obtained, but is to show that Padé-Legendre interpolants can

be used as postprocessors of any data given at the desired quadrature points. The general setting of the simulations are

• The parameters N, M, and L no longer satisfy condition (15). Choices of parameters that do satisfy condition (15) no longer remove Gibb's oscillations. Instead parameters satisfy the looser condition

$$M + L < N. \tag{20}$$

This is done since for computational data, a significant fraction of the high modes are potentially severely polluted (50% or more) and emphasizing these in the reconstruction leads to poor reconstructions as also observed in [5].

• All the computations have been done using the Legendre basis $\varphi_m(x) = P_m$ in the computation of the matrix A in (19). Interpolation of computational data often requires a higher order denominator, in which case the canonical basis is no longer a suitable choice as D becomes severely illconditioned.

• The system $Aq^{(L)} = 0$ is solved using the method detailed below.

– If A_1 denotes the first column of A, $A = [A_1|B_1]$,

$$Aq^{(L)} = 0 \Longleftrightarrow B_1 \begin{pmatrix} q_1 \\ \vdots \\ q_L \end{pmatrix} = -q_0 A_1$$

with the normalization $q_0 = 1$, we have to solve a square system. If B_1 is regular, then A has full rank. Conversely, if A has a full rank, one can eliminate a column of A (say column j) to get an invertible matrix B_j. In this case, we have to fix q_j.

This method may be preferable to the nullspace method from the previous section as it is quicker by a factor of at least ten. This is due to the fact that fixing one coefficient q_0 (or q_j) results in a square linear system which is computationally more efficient to solve than finding the null space of the matrix A which requires a singular value decomposition and this is quite expensive. Since our parameters are now defined by the less constraining condition (20), there are many more combinations of the parameters to explore and speed becomes important.

We first consider the use of the Padé-Legendre interpolant as a post-processor for data from the solution to Burgers' equation

$$u_t + (u^2)_x = 0, \quad x \in [-1, 1] \tag{21}$$

$$u(x, 0) = 0.5 + \sin(\pi x) \tag{22}$$

solved using a stabilized Legendre spectral method [4, 15]. The interpolant is computed using data given at 256 Legendre-Gauss-Lobatto points at time $T = \pi/2$. All graphics are plotted on a uniform grid of 400 points. The raw data provided is shown in Fig. 6 where we also show a good reconstruction. In Fig. 7 we show the pointwise error for $N = 256$, $M = 20$, and increasing values of L, confirming the observations of enhanced convergence rate for the simpler test in the previous section. Increasing L further does not appear to improve the results.

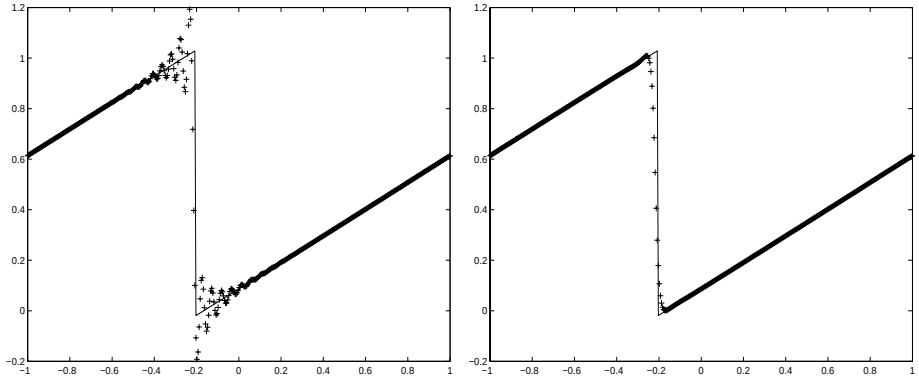

FIG. 6. On the left is shown the purely polynomial solution of Burgers equation with $N = 256$ while the right shows the Padé-Legendre reconstructed solution with $M = 20$ and $L = 8$.

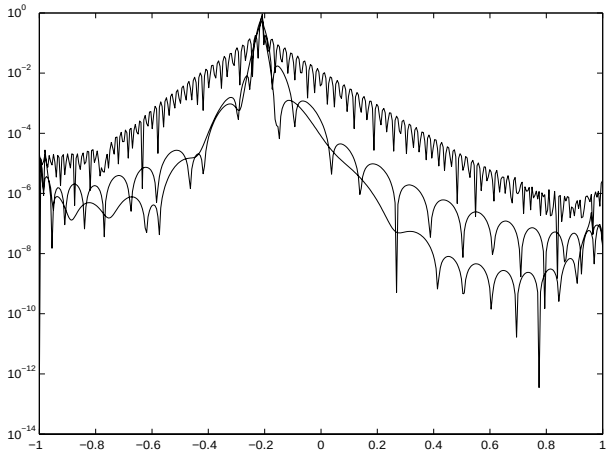

FIG. 7. Pointwise error for the reconstructed solution to Burgers equation. We use $N = 256$ and $M = 20$ for the numerator while the three curves represent from top $L = 0$, $L = 4$, and $L = 8$, respectively.

As a final example, we explore the use of the Padé-Legendre interpolant as a post-processor for the 1-D Shock entropy equations described in [4]. The polynomial interpolant is computed using data given at 256 Legendre-Gauss-Lobatto points. As a comparison for the postprocessed solution, we use a reference solution that was obtained using a ENO scheme on 1200 points [4]. All reconstructions are plotted on a uniform grid of 500 points.

In Fig. 8 we show both the reference solution and the computed polynomial solution, which exhibits Gibb's oscillations to the right of the shock, as well as small inaccuracies in the peaks to the left of the shock. We also show the visually best global reconstruction ($(N, M, L) = (256, 20, 98)$) which effectively removes many of the oscillations to the right of the shock. However, it also introduces some oscillations to the left of the shock that were not present in the original data. This is due entirely to the inexactness of the computed data which makes it difficult to find a global reconstruction which is effective on both sides of the shock. The reconstruction shown in Fig. 8 is the best found among a large ensemble of cases, yet this reconstruction remains unsatisfactory.

However, in attempting to find a global reconstruction one realizes that it is very easy to find reconstructions which work very well on either side of the shock, including the shock itself. This leads to a simple improved algorithm.

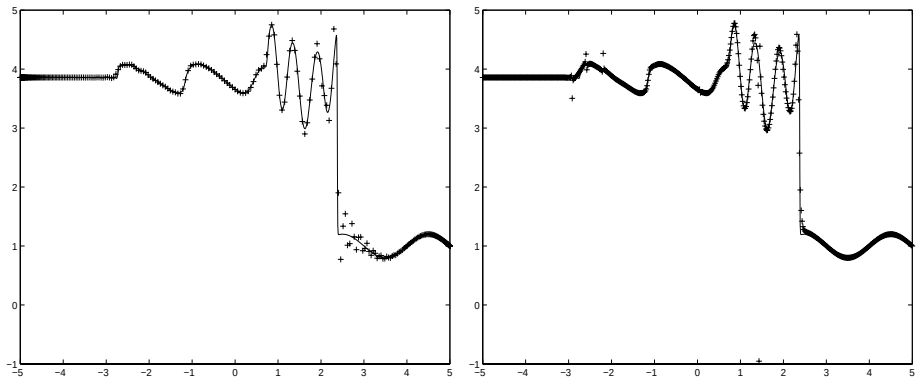

FIG. 8. On the left is shown the purely polynomial solution of the Euler equations for the shock entropy problem with $N = 256$. On the right we show the best obtainable global Padé-Legendre reconstructed solution with $M = 24$ and $L = 98$.

• Seek Padé-Legendre interpolants that resolve either side of the shock well. Examples of these reconstructions are shown in Fig 9.

• Identify the approximate shock location. This does not need to be done accurately, i.e., a 1st order cell location suffices. Any existing method for locating shocks can be used such as [11, 12, 18] as it does not affect the Padé-Legendre reconstruction.

• Patch the two (– or several) reconstructed solutions together across the cell with the shock.

In Fig. 9 we show the pure polynomial polynomial solution with the Gibbs oscillations as well as a patched reconstructed solution, and the two reconstructions used to construct the patched solution. As expected, the patched solution is an excellent approximation to the reference solutions obtained at high computational

cost. In this example, the left reconstruction uses $M = 120$, $L = 16$ and the right side $M = 82$, $L = 20$ and these were "glued" around the computed shock location. The location of the shock was reconstructed using the method detailed in [18].

5.3. Observations on choosing parameters M and L

While computing a Padé interpolant is straightforward and fast, choosing the degree of the numerator M and the denominator L is far from simple. In fact, the relationship amongst the parameters M, L, and N is nonlinear which explains the difficulty in predicting good choices for the parameters (M, L). This is true even for simple exact functions. However, a few conclusions can be drawn from the examples presented here that may serve as guidelines for finding a good reconstruction. Without an exact solution it is hard to define what a good reconstruction is. However, one advantage of the Padé method is that when it does fail to produce a good reconstruction, it is obvious. The reconstruction either does not remove any oscillations present in the original data, or may smooth out some of the oscillations but add others not present in the data itself- a phenomenon which can be seen in the bottom right figure in Fig. 9.

- For exact point values, a simple strategy is to increase the order of the denominator L while satisfying Eq.(15). In general, the rule of thumb is that a more complicated function requires a higher-degree denominator.
- While reconstructing computational data also requires a higher-degree denominator when the data has greater structure, the less restrictive condition, Eq. (20) allows for many more possible choices of (M, L). We choose this condition since computational data is often polluted by numerical damping and insufficient resolution and will require using less modes than the function expansion.
- For all reconstructions, especially those involving computational results, one should exploit the fact that the Padé reconstruction as formulated in sections 4 and 5 is very quick. A good strategy is to compute many reconstructions with various combinations of M and L and then disregard the bad reconstructions, i.e. the reconstructions that have no effect or add extraneous oscillations to the solution. These bad reconstructions can be identified by examining the zeros or minimum values of the denominator, which will lie within or very close to the computational domain.

Further work in understanding the choice of parameters M and L as well as identifying successful approximations is still needed and will increase the ease with which the Padé reconstruction can be used.

6. CONCLUSION

In this work we have defined a rational interpolant method based on the knowledge of a function at the Legendre quadrature points and shown a few basic properties of this interpolant. The computation of the interpolant is done very efficiently using the Legendre quadratures and several numerical tests have been done to show the ability of the method to significantly reduce the Gibbs phenomenon.

For data obtained from computations, i.e., data with noise, we found that a direct extension of these techniques is less successful. However, a multi-domain approach, with domains broken by the *approximate* location of the shocks, is shown to work

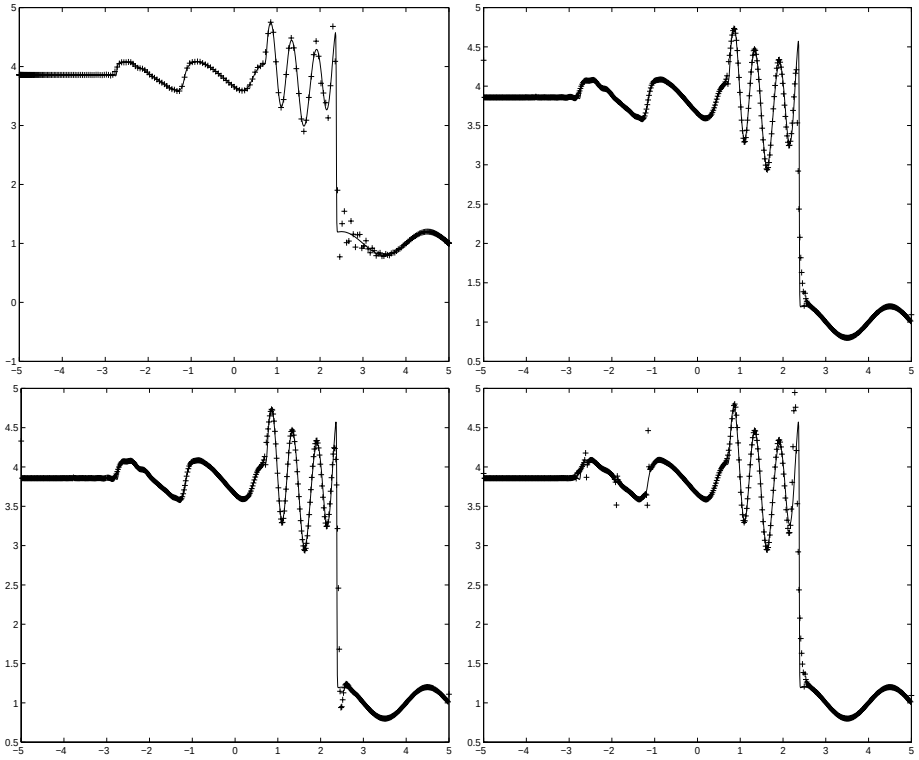

FIG. 9. On the top left is shown the purely polynomial solution of the Euler equations for the shock entropy problem with $N = 256$. On the top right we show the reconstructed solution obtained by two reconstructions patched across the shock. The bottom row of pictures show the reconstructions used in the patching. On the bottom left is the reconstruction with $M = 120, L = 16$ and on the bottom right is the reconstruction corresponding to $M = 82, L = 20$.

very well. The need to only know the shock position approximately is a major advantage over some other reconstruction techniques where the exact location is needed [13].

Another potential advantage of the approach discussed here is its potential generalization to genuine multi-dimensional problems, e.g., on simplices with the grid points being cubature points. We shall explore this in the near future as well as the formulation of a more mathematical description of the encouraging observations made in this work, and further work on the choice of parameters M, L.

ACKNOWLEDGMENT

The work of JSH was partly supported by NSF Career Award DMS0132967, by an NSF International Award NSF-INT 0307475, and by the Alfred P. Sloan Foundation through a Sloan Research Fellowship.

REFERENCES

1. R. Bauer, *Band Filters for determining shock locations*, Ph.D. thesis, Applied Mathematics, Brown University, 1995.

2. C. Bernardi and Y. Maday, *Spectral methods* In Handbook of Numerical Analysis V, North-Holland, 1997.
3. J.P. Boyd, *Trouble with Gegenbauer reconstruction for defeating Gibbs' phenomenon: Runge phenomenon in the diagonal limit of Gegenbauer polynomial approximations*, J. Comput. Physics, in Press (2005).
4. W. S. Don, *Numerical Study of Pseudospectral Methods in Shock Wave Applications*, J. Comput. Physics, **110**(1994), pp. 103-111.
5. W.S. Don, S. M. Kaber, and M. S. Min, *Fourier-Padé Approximations and Filtering for the Spectral Simulations of Incompressible Boussinesq Convection Problem*, Accepted for publications in Mathematics of Computation, 2004.
6. T.A. Driscoll, B. Fornberg, *A Padé-based algorithm for overcoming the Gibbs phenomenon*, Numerical Algorithms, **26**(2001), pp. 77-92.
7. L. Emmel, *Méthode spectrale multidomaine de viscosité évanescente pour des problèmes hyperboliques non linéaires*, Ph.D dissertation, University of Paris 6, 1998.
8. L. Emmel, S.M. Kaber and Y. Maday, *Padé-Jacobi filtering for spectral approximations of discontinuous solutions*, Numerical Algorithms, **33**(2003), pp. 251-264.
9. J.F. Geer, *Rational trigonometric approximations using Fourier series partial sums*, J. Sci. Comput., **10**(1995), pp. 325-356.
10. A. Gelb, *Parameter Optimization and Reduction of Round Off Error for the Gegenbauer Reconstruction Method*, J. Sci. Comput., **20**(2004), pp. 433-459.
11. A. Gelb and E. Tadmor, *Edge Detection from Spectral Data*, Applied Harmonic Analysis, **7**(1999), pp. 101-135.
12. A. Gelb and E. Tadmor, *Detection of Edges in Spectral Data II. Nonlinear Enhancement*, SINUM, **38** (2000), pp. 1389-1408.
13. D. Gottlieb and C.W. Shu, *On the Gibbs Phenomenon and its Resolution*, SIAM Review, **39**(1997), pp. 644-668.
14. D. Gottlieb and E. Tadmor, *Recovering Pointwise Values of Discontinuous Data with Spectral Accuracy*, In Progress and Supercomputing in Computational Fluid Dynamics. Birkhäuser, Boston, 1984. pp. 357-375.
15. D. Gottlieb and J.S. Hesthaven, *Spectral methods for hyperbolic equations*, J. Comput. Applied Math., **128**(2001), pp. 83-131.
16. J.S. Hesthaven and M. Kirby, *Filtering in Legendre Spectral Methods*, 2005 – submitted.
17. J.S. Hesthaven and S.M. Kaber, *Padé-Jacobi Approximants*, 2005 – submitted.
18. S.M. Kaber and H. Vandeven, *Reconstruction d'une fonction discontinue á partir de ses coefficients de Legendre*, C.R.A.S. 317, série I (1993).
19. S.M. Kaber, Y. Maday, *Analysis of some Padé-Chebyshev approximants*, SIAM J. Numerical Analysis, **43**(2004), pp. 437-454.
20. A.C. Matos, *Recursive computation of Padé-Legendre approximants and some acceleration properties*, Numerical Math, **89**(2001), pp. 535-560.
21. J. Tanner and E. Tadmor, *Adaptive Mollifiers - High Resolution Recover of Piecewise Smooth Data from its Spectral Information*, Found. Comput. Math., **2**(2002), pp. 155-189.
22. H. Vandeven, *Family of Spectral Filters for Discontinuous Problems*, J. Sci. Comput., **8**(1991), pp. 159-192.