CORE
Metadata, citation and similar papers at core.ac.uk
Provided by Infoscience - École polytechnique fédérale de Lausanne

Filtering in Legendre Spectral Methods

J.S. Hesthaven* and R. M. Kirby†
*Division of Applied Mathematics, Brown University, Box F, Providence, RI 02912, USA.
†School of Computing, University of Utah, Salt Lake City, UT 84112, USA.
E-mail: Jan.Hesthaven@brown.edu; kirby@cs.utah.edu

We discuss the impact of modal filtering in Legendre spectral methods, both on accuracy and stability. For the former, we derive sufficient conditions on the filter to recover high order accuracy away from points of discontinuity. Computational results confirm that less strict necessary conditions appear to be adequate. We proceed to discuss a instability mechanism in polynomial spectral methods and prove that filtering suffices to ensure stability. The results are illustrated by computational experiments.

Key Words: Spectral Methods; Filtering; Stabilization; Legendre Polynomials

1. INTRODUCTION

While the advantages of the use of spectral and pseudospectral methods for solving partial differential equations with smooth solutions are widely acknowledged [3, 18, 8], the prospects of using such methods for problems with non-smooth solutions remain more controversial. This can be attributed mainly to the appearance of the Gibbs phenomenon, appearing when approximating non-smooth solutions using polynomials, and the often detrimental effect this has on the stability of the computational scheme. Thus, it is often perceived that spectral methods are too sensitive and lack robustness to allow the modeling of problems of realistic complexity which, by their very nature, most often are dominated by under-resolved and unresolved dynamics.

The literature is rich with ideas for overcoming this lack of robustness, all centered around the idea of introducing sufficient dissipation of the high modes without sacrificing the accuracy. The exact process of doing so is less obvious as too much dissipation clearly destroys the accuracy of the solution.

Recently, the use of spectral filters to regain robustness has received considerable attention due to the effectiveness of this approach and the low computational cost of applying such filters compared to alternatives, e.g., limiting. The filters are characterized by modifying the expansion coefficients of the solution, thus leading to a global modification of the solution.

For Fourier spectral and pseudospectral methods, early results showed the promise of this for linear problems [21] and with experimental evidence for nonlinear problems [20]. The first rigorous analysis of the approximation properties of the filtered expansion is offered in [27], showing the potential of recovering spectral convergence anywhere away from the point(s) of discontinuity. A simple proof of the stabilizing effect of this approach for linear problems can be found in [13]. An overview of these results can be found in [16].

The use of filters in spectral and pseudospectral methods based on orthogonal polynomials [5, 2, 9, 24, 17], appears to be equally powerful and is being used increasingly to enable the modeling of complex time-dependent phenomena.

However, apart from Chebyshev-based methods, which one can consider as a special case of Fourier methods, there is no substantial theory to support the use of filters for general polynomial methods. Needless to say, given the extensive results in the literature, there is little reason to doubt that filtering also works in this case.

In this paper we attempt to shed some light on the use of filtering in Legendre spectral and pseudospectral methods. We first discuss some numerical experiments in detail to established some qualitative understanding of the impact of filtering in terms of accuracy and stability. This sets the stage for a second look at the impact of filtering, resulting in the derivation of sufficient, but likely not necessary, conditions for accuracy improvements. The analysis is loosely based on the previous work in [27] although some new developments are needed. This analysis elucidates the nonuniform impact of the filter and shows how properties of the filter function, e.g., smoothness, are key components of the accuracy improvements. This analysis is, to the best of our knowledge, the first to yield a quantitative, although partial, understanding of how the filter works in Legendre spectral methods. We subsequently discuss how the filter can be utilized as a stabilization in time dependent problems, confirming numerical results both offered here and found widely in the literature.

We would like to emphasize that spectral filtering, as considered here, is only one of several different, but related ways of improving accuracy and stability of spectral and pseudospectral methods. In particular, the physical space filters [11, 12, 26] and spectral vanishing viscosity methods [25, 22, 23] are closely related. A discussion of these methods and their relations can be found in [13, 16], and we shall not discuss this further. The Gegenbauer reconstruction method [12] can likewise be considered as a filter, albeit with focus on accuracy rather than stability. Other stabilization techniques sometimes used are dealiasing [3] and over integration [19].

What remains is organized as follows. In Sec. 2 we recall the Legendre polynomials and their properties as well as basic properties of Legendre expansions of general functions. Section 3 offers some experimental evidence of the impact of filtering, both on accuracy and stability of Legendre spectral methods. This sets the stage for Sec. 4 where we revisit the filtered expansions and obtain partial but rigorous results for the filter induced accuracy improvements possible and derive sufficient conditions for arbitrary accuracy away from discontinuities. In this section we also explain the stabilizing effect of the filter when solving time dependent problems using Legendre spectral methods. Section 5 contains a few concluding remarks.

2. LEGENDRE POLYNOMIALS AND EXPANSIONS

We focus the attention on polynomial expansions of the form

$$u(x) = \sum_{n=0}^{\infty} \hat{u}_n P_n(x) \quad , \quad x \in [-1,1] \quad . \tag{1}$$

Here, $P_n(x)$ represents the n'th order Legendre polynomial, defined as the polynomial solution to the Sturm-Liouville problem

$$\mathcal{L}P_n(x) = \frac{d}{dx}(1-x^2)\frac{d}{dx}P_n(x) = -\lambda_n P_n(x) \quad , \tag{2}$$

where

$$\lambda_n = n(n+1) \quad .$$

The Legendre polynomials are normalized such that

$$P_n(\pm 1) = (\pm 1)^n \quad .$$

Furthermore, we have the center value

$$P_n(0) = (-1)^m 2^{-2m} \begin{pmatrix} 2m \\ m \end{pmatrix} \quad , \quad P'_{n+1}(0) = (n+1)P_n(0) \quad , \tag{3}$$

for $n = 2m$, while $P_n(0) = 0$ for n being odd as a consequence of the symmetry

$$P_n(x) = (-1)^n P_n(-x) \quad . \tag{4}$$

We introduce the inner-product and the associated norm

$$(u,v) = \int_{-1}^{1} u(x)v(x)\,dx \quad , \quad \|u\| = \sqrt{(u,u)} \quad ,$$

and use this to define the usual spaces

$$L^2[-1,1] = \{u \,|\, \|u\| < \infty\} \quad ,$$

and the associated higher Sobolev norms

$$H^p[-1,1] = \left\{ u \in L^2 \,|\, \|u\|^2_{H^p[-1,1]} = \sum_{i=0}^{p} \|u^{(i)}\|^2 < \infty \right\} \quad .$$

As the Legendre polynomials satisfy Eq.(2), we have

$$(P_n, P_m) = \delta_{nm}\gamma_n \quad , \quad \gamma_n = \frac{2}{2n+1} \quad . \tag{5}$$

Thus, for all $u(x) \in L^2$ we recover the expansion coefficients, $\hat{u}_n$, in Eq.(1) of the form

$$\hat{u}_n = \frac{1}{\gamma_n}\left(u, P_n\right) \quad . \tag{6}$$

Rather than evaluating the above integral exactly, one can use the Legendre-Gauss-Lobatto quadrature

$$\tilde{u}_n = \frac{1}{\tilde{\gamma}_n}\sum_{i=0}^{N} u(x_i)P_n(x_i)w_i \quad , \tag{7}$$

where (x_i, w_i) represent the Legendre-Gauss-Lobatto quadrature nodes and weights, respectively (see e.g. [3]). The quadrature is exact if $u(x)P_n(x)$ is a polynomial of degree $2N-1$ or less. For general functions, $\hat{u}_n \neq \tilde{u}_n$, recognized as the aliasing error. In the present context, this is not essential and we shall not make an effort to distinguish between these two sets of expansion coefficients (see [16] for a discussion).

For later use, let us define the discrete inner product and associated $L^2[-1,1]$-equivalent discrete norm as

$$\left[u_N, v_N\right]_N = \sum_{i=0}^{N} u_N(x_i)v_N(x_i)w_i \quad , \quad \|u_N\|_N^2 = \left[u_N, u_N\right]_N \quad .$$

In computational methods, e.g., spectral methods, one is concerned with the truncated expansion

$$u_N(x) = \sum_{n=0}^{N} \hat{u}_n P_n(x) \quad , \quad x \in [-1,1] \quad , \tag{8}$$

and how it behaves as N increases. In other words, we wish to understand how $u - u_N$, measured in some appropriate norm, decays when increasing N.

Insight into this can be gained by recalling Parsevals's identity,

$$\|u\|^2 = \sum_{n=0}^{\infty} \gamma_n \left(\hat{u}_n\right)^2 \quad ,$$

implying that

$$\|u - u_N\|^2 = \sum_{n=N+1}^{\infty} \gamma_n \left(\hat{u}_n\right)^2 \quad ,$$

i.e., the accuracy depends solely on the decay of the expansion coefficients, $\hat{u}_n$, and the behavior of γ_n, given in Eq.(5). Repeated integration by parts of Eq.(6) yields $(n \neq 0)$

$$\hat{u}_n = \frac{1}{(-\lambda_n)^q}\frac{1}{\gamma_n}\left(\mathcal{L}^q u, P_n\right) \quad .$$

Recall that $\mathcal{L}$ essentially is a 2nd order operator, i.e., if $u(x) \in H^{2q}[-1,1]$, we can combine these results to obtain ($q \geq 0$) [4]

$$\|u - u_N\| \leq CN^{-2q} \|u\|_{H^{2q}[-1,1]} \quad .$$

Likewise, we get the point wise estimate [4] ($q > 1/2$)

$$\|u - u_N\|_{L^\infty[-1,1]} \leq CN^{-2q+1} \|u\|_{H^{2q}[-1,1]} \quad . \tag{9}$$

Clearly, the smoothness of the solution as indicated by the decay of the expansion coefficients is the main source of accuracy.

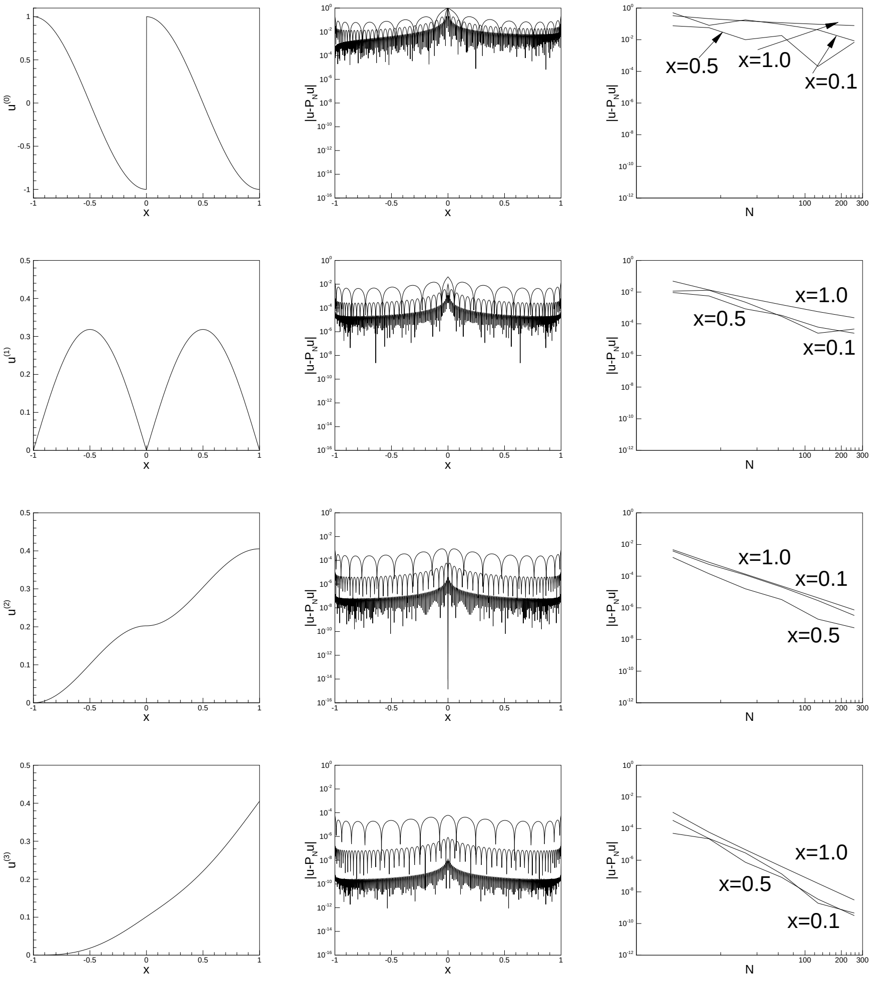

FIG. 1. Left column shows the first four functions in the sequence, $u^{(i)}$, defined in Eq.(10). In the middle column we show the point wise error of the truncated Legendre expansion for each function for increasing values of truncation, exemplified by $N = 16$, $N = 64$, and $N = 256$. In the right column is shown the point wise error at three points, $x = 0.1; 0.5; 1.0$, in the interval.

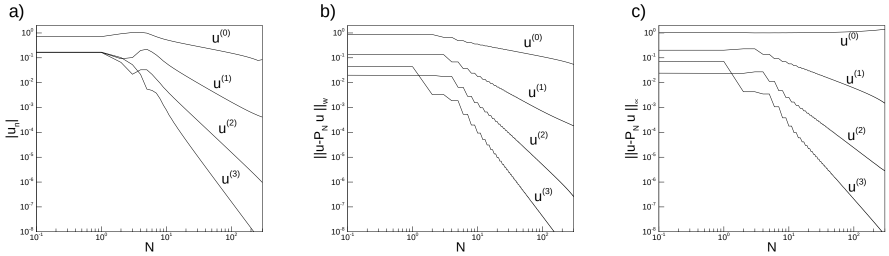

FIG. 2. In a) we show the envelope, i.e., the upper bound, of $|\hat{u}_n|$ for the first four functions in the sequence, $u^{(i)}$, defined in Eq.(10). In b) we show the associated L^2 error while c) shows the corresponding L^∞ error.

To illustrate this, consider an example to which we shall return again later. Define the sequence of functions

$$u^{(0)}(x) = \begin{cases} -\cos(\pi x) & x \in [-1,0] \\ \cos(\pi x) & x \in]0,1] \end{cases} \quad , \quad u^{(i)}(x) = \int_{-1}^{x} u^{(i-1)}(s)\, ds \quad . \qquad (10)$$

Note in particular that this sequence is constructed such that $u^{(q)} \in H^q[-1,1]$, i.e., it serves to understand the behavior of the expansion as a function of the truncation, N, and the regularity of the function being approximated.

To familiarize ourselves with these functions and the properties of the associated expansions, we show in Fig. 1 the first four functions as well as the point wise behavior of the error for the truncated expansions of these functions. As expected, we see a clear relation between the convergence rate and the regularity of the solution. To avoid complications we approximate the expansion coefficients, Eq.(6), by those computed using very high order Gaussian quadratures, Eq.(7).

This behavior is furthermore confirmed in Fig. 2 which illustrates the decay of the expansion coefficients, $\hat{u}_n$, as well as the associated mean and point wise error for increasing values of N. Inspection confirms the convergence estimates outlined above.

3. FILTERING OF POLYNOMIAL EXPANSIONS – A FIRST LOOK

As illustrated above, if the function being approximated possesses significant regularity we can expect the spectral expansion to be highly efficient for the representation of the function and its spatial derivatives. In other words, only relatively few terms are needed to produce a very accurate approximation. On the other hand, for problems with limited regularity the situation is a bit more complex. Unfortunately, this is generally the case for most interesting problems where the limited regularity is caused by the solution itself or by a lack of resolution. This causes a global deterioration in the accuracy and possible a lack of point wise convergence in the case of a true discontinuity.

However, from the previous discussion we also appreciate that the decay of the expansion coefficients, $\hat{u}_n$, is intimately related to the accuracy. Thus, one could ask whether it is possible to modify (e.g. attenuate in some prescribed way) the expansion coefficients in such a way as to improve the accuracy of the truncated expansion. This is the basic idea of modal or spectral filtering.

The question to consider is, of course, exactly how such a modification should be made and attempt to seek an understanding of the consequences of modifying the expansion coefficients. On one hand, one wishes to improve on the accuracy away from the point, $x = c$, where $u(x)$ looses smoothness. On the other hand, one does not wish to make matters worse away from $x = c$. Finally, if $u(x)$ is indeed already smooth, the impact of the filter should be minimal and should not destroy the convergence rate.

We shall consider filtered expansions of the kind

$$\mathcal{F}_N u_N(x) = \sum_{n=0}^{N} \sigma\left(\frac{n}{N}\right) \hat{u}_n P_n(x) \quad , \tag{11}$$

where $\sigma(\eta)$ is a real filter function, the specification of which is a central element. As we shall discuss below, defining the filter function in the following way appears to ensure convergence everywhere away from $x = c$ and guarantees that convergence is not destroyed for smooth functions.

Definition 3.1. The filter function, $\sigma(\eta) \in \mathsf{C}^p : \mathsf{R}^+ \to [0, 1]$, $p > 1$, has the following properties

$$\sigma(\eta) : \begin{cases} \sigma(0) = 1 & \\ \sigma^{(k)}(0) = 0 & k = 1...p-1 \\ \sigma(\eta) = 0 & \eta \geq 1 \\ \sigma^{(k)}(1) = 0 & k = 1...p-1 \end{cases} .$$

We shall call this a p'th order filter.

In the following we consider two different filter functions. The first one is

$$\sigma_O(\eta) = 1 - \frac{\Gamma(2p)}{\Gamma(p)^2} \int_0^\eta [t(1-t)]^{p-1} \, dt \quad , \tag{12}$$

which is a $2p-1$'th order Birkhoff-Hermite interpolating polynomial, defining $\sigma_O(\eta)$, and $\Gamma(x)$ is the Gamma function. This filter was first proposed in [27] and we shall subsequently refer to it as the optimal filter due to its close connection to Def 3.1.

The filter function, $\sigma_O(\eta)$, is illustrated in Fig. 3a for different values of p. Note that as the order, p, increases, $\sigma_O(\eta)$ approaches an inverted Heaviside function, centered at $\eta = 0.5$. The strong condition on smoothness at $\sigma(1)$ implies that a substantial part of the high modes are getting modified, leaving only a small fraction of the low modes almost unchanged.

This observation is one of the key reasons for the widespread use of the exponential filter, $\sigma_E(\eta)$, defined as

$$\sigma_E(\eta) = \exp\left(-\alpha \eta^p\right) \quad . \tag{13}$$

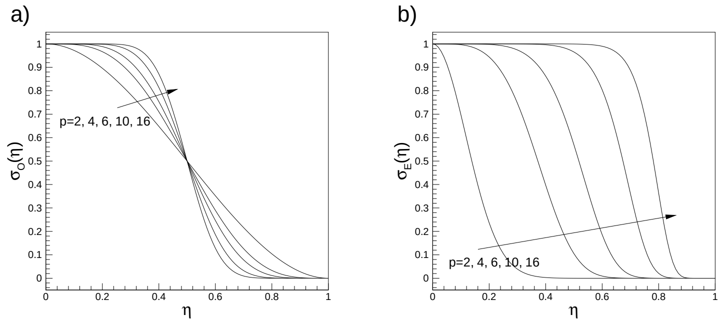

FIG. 3. In a) we show filter function, $\sigma_O(\eta)$ for the optimal filter, Eq.(12) for increasing values of p. b) shows a similar sequence for the exponential filter, $\sigma_E(\eta)$, given in Eq.(13).

The parameter, α, measures the modification of the maximum mode, i.e., $\sigma_E(1) = \exp(-\alpha)$. Typically, $\alpha = -\log(\varepsilon_M)$ where ε_M is the machine accuracy. Other choices can also be used, e.g., decreasing α implies a smaller modification of the high modes.

In Fig. 3b we illustrate this filter for increasing values of p. We note in particular that a significantly larger part of the modes remains essentially unchanged for larger values of p when compared with the optimal filter, Eq.(12).

The exponential filter does not, however, conform to the definition of the filter function given in Definition 3.1. In particular we have that

$$|\sigma_E^{(k)}(1)| \simeq (\alpha p)^k \exp(-\alpha) \ \ ,$$

which can be very far from zero for large values of k and p.

3.1. Improving Accuracy

Let us consider the impact of the filter on the point wise accuracy as a function of the order, p, of the filter, the length, N, of the expansion, and the regularity of the function being approximated.

We first apply the optimal filter, Eq.(12), on the polynomial representations of the four test functions shown in Fig. 1. In Fig. 4 we show the point wise error associated with three different filter orders for the four test functions.

A few observations are worth making. For a function of fixed regularity (rows in Fig. 4), filtering can dramatically improve the accuracy of the expansion away from the point of discontinuity. Also, increasing N, increases the size of the regions away the non-smooth point where the filtering is efficient. On the other hand, the order of the filter may also impact the accuracy unfavorably, as illustrated for the $p = 2$ filter (first column in Fig. 4) which seems to limit the convergence rate regardless of the regularity of the function. However, if p is sufficiently large, this does not seem to adversely affect the point wise error.

In Fig. 5 a similar set of examples is presented, although based on the use of the exponential filter, Eq.(13). Although quantitatively different from the results in Fig. 4, the qualitative characteristic are the same in spite of the filter not conforming to Definition 3.1. Indeed, for increasing values of p (columns in Figs. 4-5), the

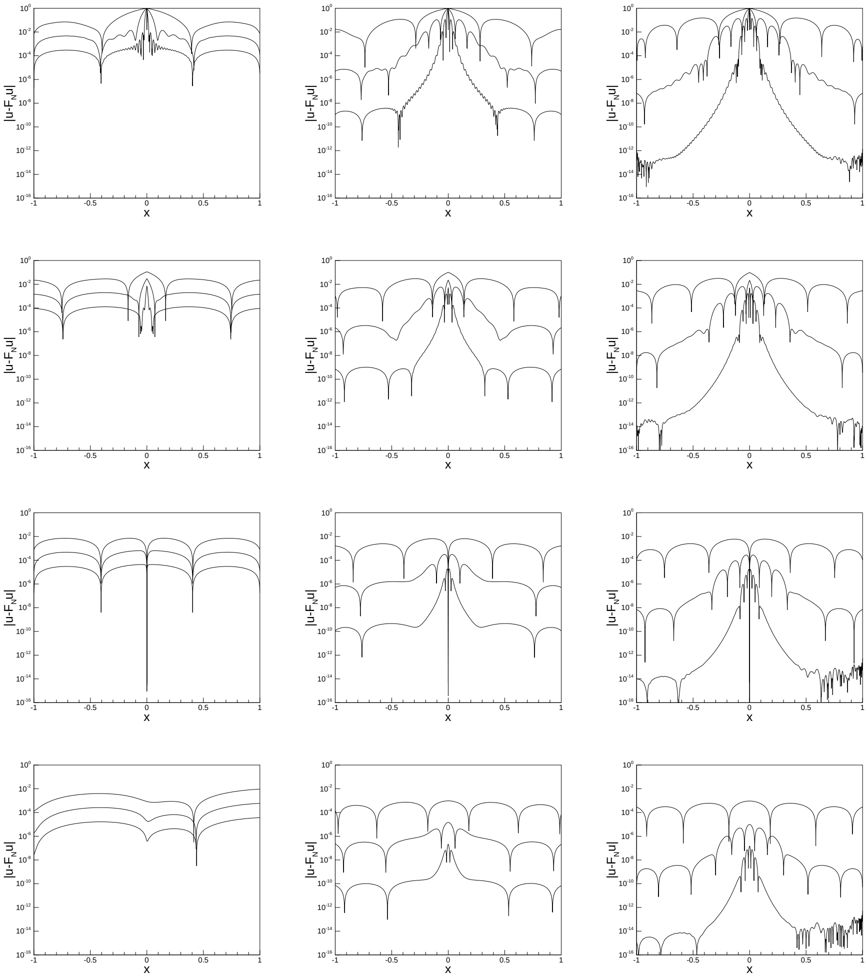

FIG. 4. In the left column is shown the point wise error after the optimal filter, Eq.(12), with $p = 2$ has been applied to the Legendre expansions of the four test functions in Fig. 1, with $N = 16$, $N = 64$ and $N = 256$, for each function. The middle column shows similar results for $p = 6$ while the right column displays the results for $p = 10$.

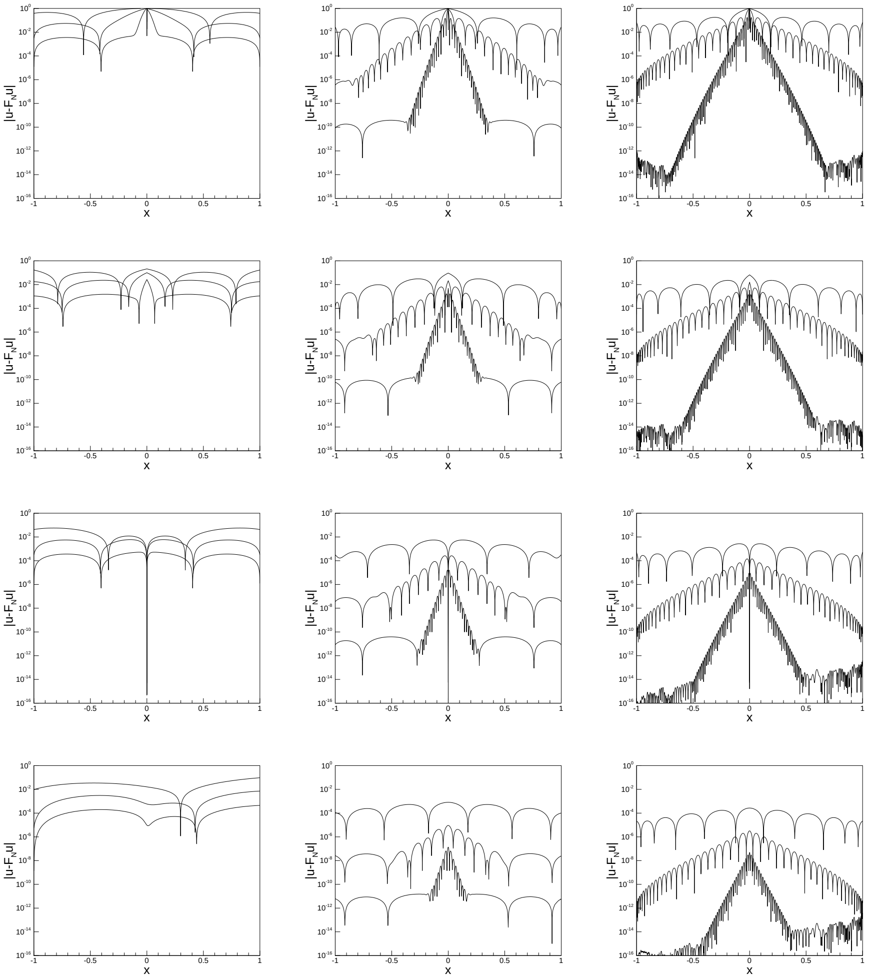

FIG. 5. In the left column is shown the point wise error after the exponential filter, Eq.(13), with $p = 2$ has been applied to the Legendre expansions of the four test functions in Fig. 1, with $N = 16$, $N = 64$ and $N = 256$, for each function. The middle column shows similar results for $p = 6$ while the right column displays the results for $p = 10$.

exponential filter appears superior in terms of the point wise error. Recalling the discussion related to Fig. 3 this is expected and confirms the intuitive understanding of the filter.

To understand in more detail the point wise impact of the filter, we show in Fig. 6 the point wise error at three different points in the domain as a function of the expansion order, N, and the order, p, of the optimal filter, Eq.(12). From this figure it is clear that there is a close relation between the order of the filter and the point wise convergence rate away from the point where the function looses regularity. This seems independent of the regularity of the original function. In contrast to this, very close to the point of discontinuity, the improvements are less dramatic although certainly noticeable.

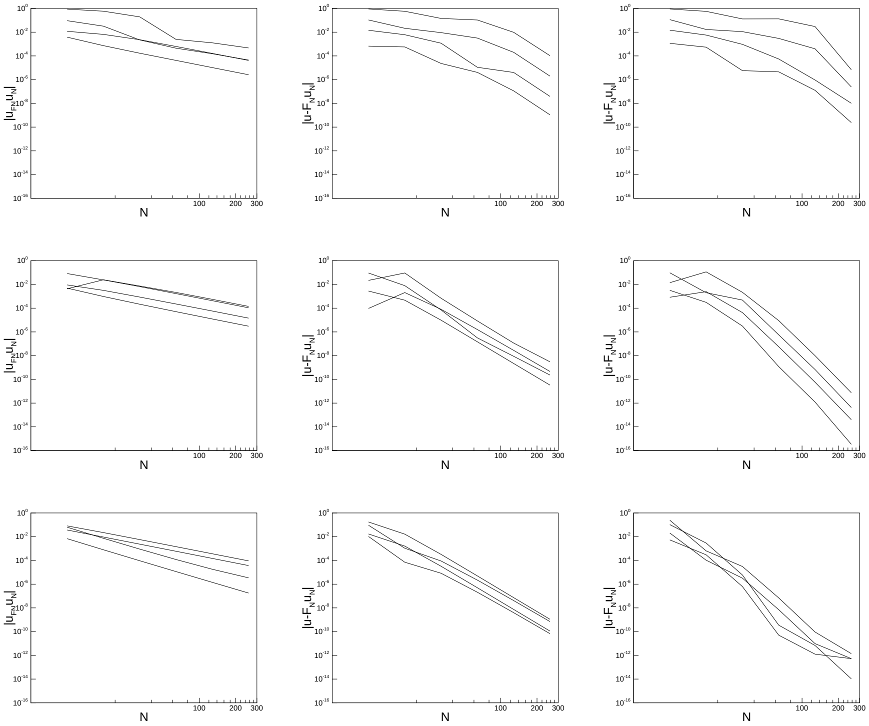

FIG. 6. Point wise errors at specific points in the domain as a function of the length of the expansion, N, the order of the filter, p, used in the optimal filter, Eq.(12), and the regularity of the function being approximated. In each figure are four graphs, corresponding to $u^{(0)}$ to $u^{(3)}$, usually with the former being the top and the latter the bottom graph. The rows corresponds to point wise errors at $x = 0.1$, $x = 0.5$, and $x = 1.0$, respectively, while the columns reflects $p = 2$, $p = 6$, and $p = 10$.

The same set of tests have been repeated with the exponential filter, yielding similar results although the local behavior appears less systematic. For all practical purposes, the results of the two filters are identical.

3.2. Improving Stability

To illustrate the impact of filtering on stability, let us consider the simple problem

$$\frac{\partial u}{\partial t} + a(x)\frac{\partial u}{\partial x} = 0 \ \ , x \in [-1, 1] \ \ . \tag{14}$$

For simplicity we assume that $a(\pm 1)$ is positive but that $a(x)$ in general can change sign inside the domain. To complete the specification we have

$$u(-1, t) = g(t) \ \ , \ \ u(x, 0) = f(x) \ \ .$$

To solve this problem we use a standard Legendre collocation method, although we shall impose the boundary conditions weakly rather than strongly. In other words, we seek a polynomial solution, $u_N(x, t)$, that satisfies

$$\left.\frac{du_N}{dt}\right|_{x_i} + a(x_i)\sum_{j=0}^{N} \mathrm{D}_{ij}u_N(x_j) = -l_0(x_i)\frac{N(N+1)}{4}a(-1)\left[u_N(-1,t)-g(t)\right] \quad (15)$$

where x_i represents the Legendre-Gauss-Lobatto points and D is the differentiation matrix originating from the Lagrange interpolation polynomial, $l_i(x)$, based on x_i, i.e.,

$$l_i(x) = -\frac{(1-x^2)P_N'(x)}{N(N+1)(x-x_i)P_N'(x_i)} \quad , \quad l_i(x_j) = \delta_{ij} \quad , \quad \mathrm{D}_{ij} = \left.\frac{dl_j}{dx}\right|_{x_i} \quad .$$

The exact entries of D can be found in [3, 16]. Also further discussions and a derivation of Eq.(15) can be found in [14, 15, 16], including a proof of stability for $a(x)$ being constant.

Let us consider a specific example for which

$$a(x) = \frac{1}{\pi}\sin(\pi x - 1) \quad .$$

The exact solution to Eq.(14) is in this case [10]

$$u(x,t) = f\left(2\tan^{-1}\left[e^{-t}\tan\left(\frac{\pi x - 1}{2}\right)\right] + 1\right) \quad ,$$

where $u(x,0) = f(x)$ represents the initial condition, e.g., $f(x) = \sin(x)$.

One observes that $\|u\|$ is bounded at all times but that the solution develops a steep gradient around $x = (1-\pi)/\pi \simeq -0.68$ before it finally decays and takes a constant value of $u(x,\infty) = \sin(1)$.

In Fig. 7 we compare the exact and the computed solution, obtained using Legendre collocation method with $N = 256$ and a 4th order Runge-Kutta scheme in time. We clearly see the development of an instability in the unfiltered case (a), resulting in a poor solution everywhere. This instability continues to grow with continued integration.

Applying a weak exponential filter ($p = 16$) removes the stability problems, leaving only traces of Gibbs oscillations close to the steep gradient. Continuing the computation to $t = 10$, corresponding to 35000 time steps, confirms that we have not just postponed the instability but eliminated it. Visually equivalent results can be obtained by using the optimal filter, Eq.(12).

Furthermore, in agreement with the discussion in Sec. 3.1, we also find that using the filter only modifies the solution locally while away from the sharp gradient, the computed and the exact solution agree very well.

3.3. A Summary

To summarize the above experiments, the key observations are

- The filter improves the accuracy, often dramatically, away from the point of discontinuity.
- The amount of improvement is controlled by the order of the filter, p, and not by the regularity of the function being approximated.

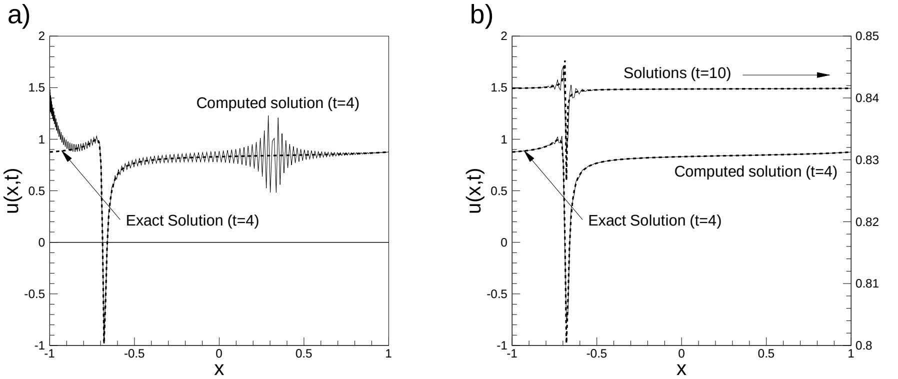

FIG. 7. In a) we show the exact (dashed line) and computed (full line) solution to a variable coefficient linear problem at $t = 4$. The numerical scheme is a Legendre collocation methods with $N = 256$. In b) we show the same problem, however solved with a weak ($p = 16$) filter applied. Also shown is the solution at $t = 10$.

• Increasing the expansion order, N, narrows the region where the discontinuity most severely impacts the convergence rate.

• Low order filters can destroy the expected accuracy even for smooth functions.

• Filtering not only improves on the accuracy but may be essential to maintain the stability.

• The optimal filter and the exponential filter behaves similarly for all practical purposes.

As we shall see shortly, these observations conform well with the insight offered by a more careful analysis.

4. A SECOND LOOK

The purpose of the above has been to illustrate the impact of the filter on a polynomial expansion, both in terms of accuracy of the filtered expansion and through the enhanced stability of a polynomial spectral method used to solve a differential equation.

In the following we shall attempt an analysis to substantiate the main observations and allow us to have confidence in the use of filtering as a more general approach to improve the accuracy of expansions and the stability of spectral methods based on orthogonal polynomials.

4.1. Impact on Accuracy

Let us first seek an understanding of exactly what the filter does and what conditions must be imposed on $\sigma(\eta)$ to achieve this. Without loss of generality, we shall subsequently assume that $u(x)$ is piecewise $H^p(p > 0)$ and that $x = c$ indicates the point of loss of regularity. Note that if $|c| = 1$, the fact that the Legendre polynomials satisfies a singular Sturm-Liouville equations suffices to guarantee convergence controlled by the regularity of $u(x)$. Thus, without loss of generality, we restrict the attention to cases where c is interior, i.e., $|c| < 1$.

Consider the filtered polynomial expansion

$$\mathcal{F}_N u_N(x) = \sum_{n=0}^{N} \sigma\left(\frac{n}{N}\right) \hat{u}_n P_n(x) \quad , \quad \hat{u}_n = \frac{1}{\gamma_n}\int_{-1}^{1} u(x) P_n(x)\, dx \quad , \tag{16}$$

where γ_n is given in Eq.(5). Inserting $\hat{u}_n$ into Eq.(16) yields

$$\mathcal{F}_N u_N(x) = \int_{-1}^{1} u(s) K_N^0(x,s)\, ds \quad , \tag{17}$$

where we have the filtered kernel

$$K_N^0(x,s) = \sum_{n=0}^{N} \frac{1}{\gamma_n} \sigma\left(\frac{n}{N}\right) P_n(s) P_n(x) \quad . \tag{18}$$

Let us define the sequence of functions ($l \geq 1$)

$$K_N^l(x,s) = \sum_{n=1}^{N} \sigma\left(\frac{n}{N}\right) \left(\frac{-1}{\lambda_n}\right)^l \frac{P_n(s)P_n(x)}{\gamma_n} \quad .$$

Note that a special property of this is

$$\mathcal{L}_s K_N^{l+1}(x,s) = K_N^l(x,s) \quad , \quad l = 1, 2, \ldots \quad ,$$

where $\mathcal{L}_s$ signifies the Sturm-Liouville operator, Eq.(2), with respect to s.

4.1.1. Filtering of Smooth Functions

Let us begin by assuming that $u(x) \in H^p[-1,1], p > 1$. In this case we have already seen in Sec. 3.1 that the filter may destroy the expected convergence rate if the order of the filter is too low.

It seems natural to require that filtering of a smooth function does not destroy the rapid convergence rate. Thus, we must understand what conditions to impose on the filter to ensure this. Necessary conditions on the filter is stated in the following theorem.

Theorem 4.1. *Assume that $u(x) \in H^p[-1,1]$ and that the filter, $\sigma(\eta) \in \mathsf{C}^p[0,\infty]$, $p > 1$, obeys*

$$\sigma(\eta) = \begin{cases} \sigma(0) = 1 & \\ \sigma(\eta) = 0 & \eta > 1 \\ \sigma^{(l)}(0) = 0 & l = 1 \ldots p-1 \end{cases} \quad .$$

Then

$$|u(x) - \mathcal{F}_N u_N(x)| \leq N^{1-p} \left\| u^{(p)} \right\|_{L^2[-1,1]} \quad .$$

Proof. Let us write $u(x)$ as

$$u(x) = \int_{-1}^{1} u(s) G^0(x,s)\, ds \;\;,\;\; G^0(x,s) = \sum_{n=0}^{\infty} \frac{1}{\gamma_n} P_n(s) P_n(x) \;\; . \tag{19}$$

Note that $G^0(x,s)$ is nothing but the L^2-projection of the Dirac-function, $\delta(x-s)$, onto the space of polynomials.

As for K_N^{l+1}, we also define the sequence of functions

$$G^l(x,s) = \sum_{n=1}^{\infty} \left(\frac{-1}{\lambda_n}\right)^l \frac{P_n(s)P_n(x)}{\gamma_n} \;\; ,$$

with the same property that

$$\mathcal{L}_s G^{l+1}(x,s) = G^l(x,s) \;\;,\;\; l = 1, 2, \dots \;\; . \tag{20}$$

Assuming $u(x) \in H^{2q}[-1,1]$, $q \geq 1$, repeated integration by parts of Eq. (16) and Eq.(19) yields

$$u(x) - \mathcal{F}_N u_N(x) = \int_{-1}^{1} \mathcal{L}_s^q u(s) \left[G^q(x,s) - K_N^q(x,s)\right] \, ds \;\; .$$

Using the special property, e.g., Eq.(20), of G^q and K_N^q we have

$$|u(x) - \mathcal{F}_N u_N(x)| \leq \left| \int_{-1}^{1} \mathcal{Q}_N(x,s) \mathcal{L}_s^q u(s)\, ds \right| + \left| \int_{-1}^{1} \mathcal{R}_N(x,s) \mathcal{L}^q u(s)\, ds \right| \;\; ,$$

where $\mathcal{Q}_N(x,s)$ and $\mathcal{R}_N(x,s)$ take the form

$$\mathcal{Q}_N(x,s) = \sum_{n=1}^{N} \left[1 - \sigma\left(\frac{n}{N}\right)\right] \left(\frac{-1}{\lambda_n}\right)^q \frac{P_n(x)P_n(s)}{\gamma_n} \;\; ,$$

and

$$\mathcal{R}_N(x,s) = \sum_{n=N+1}^{\infty} \left(\frac{-1}{\lambda_n}\right)^q \frac{P_n(x)P_n(s)}{\gamma_n} \;\; .$$

We can bound the latter as

$$\|\mathcal{R}_N(x,s)\|^2 \leq \sum_{n=N+1}^{\infty} \left(\frac{1}{\lambda_n}\right)^{2q} \frac{1}{\gamma_n} \leq N^{-4q+2} \;\; .$$

The former function, $\mathcal{Q}_N(x,s)$, can be estimated in a similar fashion as

$$\begin{aligned} \|\mathcal{Q}_N(x,s)\|^2 \;&\leq\; \sum_{n=1}^{N} \left[1 - \sigma\left(\frac{n}{N}\right)\right]^2 \left(\frac{1}{\lambda_n}\right)^{2q} \frac{1}{\gamma_n} \\ &\leq\; \frac{N}{\gamma_N \lambda_N^{2q}} \left(\frac{1}{N} \sum_{n=1}^{N} \left[1 - \sigma\left(\frac{n}{N}\right)\right]^2 \left(\frac{\lambda_n}{\lambda_N}\right)^{-2q} \right) \;\; , \end{aligned}$$

since $\gamma_n \geq \gamma_N$. Furthermore, we have

$$\eta^2 \leq \frac{\lambda_n}{\lambda_N} \leq 2\eta^2 \ , \ \eta = \frac{n}{N} \ , \tag{21}$$

which yields the bound

$$\|\mathcal{Q}_N(x,s)\|^2 \leq CN^{-4q+2}\left(\frac{1}{N}\sum_{n=1}^{N}\left[1-\sigma\left(\frac{n}{N}\right)\right]^2\left(\frac{n}{N}\right)^{-4q}\right) \ .$$

We recognize the sum as a Riemann sum, i.e.,

$$\left|\int_0^1 (1-\sigma(\eta))^2\eta^{-4q}\,d\eta\right| < \infty \ \Rightarrow \ \ \|\mathcal{Q}_N(x,s)\|^2 \leq N^{-4q+2} \ .$$

Potential problems with boundedness of the sum arise at $\eta \simeq 0$. However, if we assume that $\sigma(\eta)$ is sufficiently smooth around $\eta \simeq 0$ we can expand it as

$$\sigma(\eta) = \sum_{k=0}^{2q}\frac{1}{k!}\sigma^{(k)}(0)\eta^k + \int_0^{\eta}\frac{1}{(2q+1)!}\sigma^{(2q+1)}(t)t^{2q+1}\,dt \ .$$

Clearly, if $\sigma \in \mathsf{C}^{2q}$ and

$$\begin{cases} \sigma(0) = 1 \\ \sigma^{(k)}(0) = 0 \ \ k = 1..2q-1 \end{cases} ,$$

the integral is bounded.

In combination with the Cauchy-Schwarz inequality and $p = 2q$ this yields the result. ■

This results conforms well with the observations made in Sec. 3.1. In particular, it confirms that the smoothness of $\sigma(\eta)$ around $\eta = 0$ must exceed the global smoothness of $u(x)$ in order to not impact the convergence rate. Examples of this can be found in Figs. 4-6 for $p = 2$ where the 2nd order nature of the filter is the limiting factor.

The importance of the smoothness of $\sigma(\eta)$ around the origin also supports the observed very small differences between the optimal filter, Eq.(12), and the exponential filter, Eq.(13), as they both obey the conditions of Theorem 4.1 up to machine accuracy, i.e., if $\alpha > -\log(\varepsilon_M)$, then $\exp(-\alpha)$ is zero in finite precision.

4.1.2. Filtering of Piecewise Smooth Functions

Let us now return to the more general case where $u(x)$ is piecewise H^p and looses regularity at $x = c$, $|c| < 1$. This is more complicated as we need to understand how the error behaves as a function of the distance $|x - c|$. As we shall see, we are not able to provide such results in completeness but shall base a conjecture on a number of special cases.

Repeated integration by parts of Eq. (16) and Eq.(19), carefully executed not to cross the point of discontinuity, $s = c$, yields

$$
\begin{aligned}
&u(x) - \mathcal{F}_N u_N(x) = \\
&\sum_{l=0}^{q-1} (1-c^2) \left[\mathcal{L}_s^l u(c^-) - \mathcal{L}_s^l u(c^+) \right] \left[\frac{\partial}{\partial s} G^{l+1}(x,s) \Big|_{s=c} - \frac{\partial}{\partial s} K_N^{l+1}(x,s) \Big|_{s=c} \right] \\
&- \sum_{l=0}^{q-1} (1-c^2) \frac{\partial}{\partial s} \left[\mathcal{L}_s^l u(c^-) - \mathcal{L}_s^l u(c^+) \right] \left[G^{l+1}(x,c) - K_N^{l+1}(x,c) \right] \\
&+ \int_{-1}^{1} \mathcal{L}_s^q u(s) \left[G^q(x,s) - K_N^q(x,s) \right] ds \ \ , \qquad (22)
\end{aligned}
$$

provided that $u(x) \in H^{2q}$ $(q \geq 1)$ on $x \in [-1, c^-]$ and $x \in [c^+, 1]$.

To estimate the impact of the filter, we must thus understand the behavior of the following terms

$$
\begin{aligned}
G^l(x,c) - K_N^l(x,c) = \ & \sum_{n=1}^{N} \left(1 - \sigma\left(\frac{n}{N}\right) \right) \left(\frac{-1}{\lambda_n} \right)^l \frac{P_n(c) P_n(x)}{\gamma_n} \\
& + \sum_{n=N+1}^{\infty} \left(\frac{-1}{\lambda_n} \right)^l \frac{P_n(c) P_n(x)}{\gamma_n} \\
= \ & Q_N^1(x,c) + R_N^1(x,c) \ \ , \qquad (23)
\end{aligned}
$$

and

$$
\begin{aligned}
\frac{\partial}{\partial s} G^l(x,s) \Big|_{s=c} - \frac{\partial}{\partial s} K_N^l(x,s) \Big|_{s=c} = \ & \sum_{n=1}^{N} \left(1 - \sigma\left(\frac{n}{N}\right) \right) \left(\frac{-1}{\lambda_n} \right)^l \frac{P_n'(c) P_n(x)}{\gamma_n} \\
& + \sum_{n=N+1}^{\infty} \left(\frac{-1}{\lambda_n} \right)^l \frac{P_n'(c) P_n(x)}{\gamma_n} \\
= \ & Q_N^2(x,c) + R_N^2(x,c) \ \ . \qquad (24)
\end{aligned}
$$

Before we continue, we need the following result

Lemma 4.1. *Let $m > 0$ be an integer. Then*

$$
\frac{1}{\gamma_{2m}} P_{2m}(0) = (-1)^m \left[\sqrt{\frac{4m}{\pi}} + \sqrt{\frac{1}{16\pi m}} + \mathcal{O}(m^{-3/2}) \right] = (-1)^m \sum_{q=0}^{\infty} \hat{a}_q m^{1/2-q} \ \ ,
$$

where $\hat{a}_0 = 2/\sqrt{\pi}$, $\hat{a}_1 = 1/4\sqrt{\pi}$ etc.

Proof. The result follows directly by combining Eq.(5) and Eq.(3) with Stirling's asymptotic series

$$
\Gamma(1+x) = \sqrt{2\pi x} x^x e^{-x} \left(1 + \frac{1}{12x} + \frac{1}{288x^2} + \mathcal{O}(x^{-3}) \right) \ \ .
$$

■

Let us simplify matters and assume that the point of discontinuity, c, is in the center of the domain, i.e., $c = 0$. We then have

LEMMA 4.2. *Assume that the point of discontinuity, c, is located at the center of the domain, i.e., $c = 0$, and $N \gg 1$. Then filtering does not improve the point wise convergence rate at this point, independent of the choice of filter.*

Proof. It suffices to show that $R^1_N(c,c)$ or $R^2_N(c,c)$ limits the convergence rate. However, the latter is identically zero as $P'_n(0) \cdot P_n(0) = 0$ due to symmetry. Thus, consider

$$\begin{aligned} R^1_N(0,0) &= (-1)^l \sum_{n=N+1}^{\infty} \frac{1}{\lambda^l_n} \frac{P^2_n(0)}{\gamma_n} = \\ &= (-1)^l \sum_{m=M+1}^{\infty} \frac{\gamma_{2m}}{\lambda^l_{2m}} \frac{4m}{\pi} \left(1 + \mathcal{O}(m^{-1})\right) \end{aligned}$$

where the last simplification follows from Lemma 4.1 and using $N = 2M$ since $P_{2m+1}(0) = 0$.

Assume now that N is large to recover

$$|R^1_N(0,0)| \le CN^{-2l+1} + \mathcal{O}(N^{-2l+2}) \ .$$

Thus, for $l = 1$, we can not expect point wise convergence unless $u \in H^p[-1,1]$, $p > 1/2$ due to the jump term, consistent with Eq.(9). Similarly, we see algebraic convergence is obtained at the same rate as predicted in Eq.(9), i.e., the filter has no impact right at the point of discontinuity. ■

Let us now consider the situation away from the point of discontinuity. Before doing so, we recall the following two results, both essentially established in [27].

LEMMA 4.3. *Let n and M be integers. Then for any $g(x) \in \mathsf{C}^{2n}[0,1]$, $n \ge 2$ we have*

$$\begin{aligned} \sum_{m=1}^{M} (-1)^m g\left(\tfrac{m}{M}\right) = \quad & \frac{1}{2}\left[g(1) - g(0)\right] \\ & + \sum_{l=1}^{n-1} M^{-2l+1} \frac{B_{2l}}{(2l)!} (4^l - 1) \left(g^{(2l-1)}(1) - g^{(2l-1)}(0)\right) + \mathcal{O}\left(M^{-2n+1}\right) \end{aligned}$$

Here B_{2l} represents the Bernoulli numbers.

LEMMA 4.4. *Let $n \ge 2$ and M be integers. Then we have*

$$\begin{aligned} \sum_{m=M+1}^{\infty} (-1)^m m^{-\alpha} = \quad & -\frac{1}{2} M^{-\alpha} \\ & + \sum_{l=1}^{n-1} M^{-\alpha-2l+1} \frac{B_{2l}}{(2l)!} (4^l - 1) \frac{\Gamma(\alpha + 2l - 1)}{\Gamma(\alpha)} + \mathcal{O}\left(M^{-\alpha-2n+1}\right) \ , \end{aligned}$$

where $\Gamma(x)$ is the gamma function.

We shall also use the following result:

Lemma 4.5. *Assume $l > 0$. Then ($n \neq a$)*

$$\left(\frac{n}{n-a}\right)^l = \sum_{p=0}^{\infty} \hat{b}_p n^{-p} \quad , \quad \hat{b}_p = \begin{pmatrix} p+l-1 \\ l-1 \end{pmatrix} a^p \quad .$$

Proof. The result follows directly by considering the standard result for Z-transforms as

$$\frac{1}{(z-a)^l} = \sum_{n=0}^{\infty} \hat{b}_n z^{-n} \quad ,$$

where

$$\hat{b}_n = \begin{pmatrix} n-1 \\ l-1 \end{pmatrix} a^{n-l} \quad ,$$

for $n \geq l$ and $\hat{b}_n = 0$ otherwise. ■

We again restrict the attention to a special case, all with $c = 0$, i.e., the point of discontinuity is at the center of the domain. In this case we have

Lemma 4.6. *Assume that the point of discontinuity, c, is located at the center of the domain, i.e., $c = 0$, and $N \gg 1$. Then*

$$\left|Q_N^1(\pm 1, 0) + R_N^1(\pm 1, c)\right| = \mathcal{O}(N^{-p}) \quad ,$$

and

$$\left|Q_N^2(\pm 1, 0) + R_N^2(\pm 1, c)\right| = \mathcal{O}(N^{-p}) \quad ,$$

provided the filter, $\sigma(\eta)$, is order p, as defined in Definition 3.1.

Proof. Let us first consider the terms in Eq.(23), evaluated at $(x, c) = (\pm 1, 0)$ as

$$\begin{aligned} Q_N^1(\pm 1, 0) &= \sum_{m=1}^{M} \left(1 - \sigma\left(\frac{m}{M}\right)\right) \left(\frac{-1}{\lambda_{2m}}\right)^l \frac{P_{2m}(0)}{\gamma_{2m}} \\ &= \left(\frac{-1}{\lambda_{2M}}\right)^l \sum_{m=1}^{M} \left(1 - \sigma\left(\frac{m}{M}\right)\right) \left(\frac{\lambda_{2m}}{\lambda_{2M}}\right)^{-l} \frac{P_{2m}(0)}{\gamma_{2m}} \end{aligned}$$

and

$$\begin{aligned} R_N^1(\pm 1, 0) &= \sum_{m=M+1}^{\infty} \left(\frac{-1}{\lambda_{2m}}\right)^l \frac{P_{2m}(0)}{\gamma_{2m}} \\ &= \left(\frac{-1}{\lambda_{2M}}\right)^l \sum_{m=M+1}^{\infty} \left(\frac{\lambda_{2m}}{\lambda_{2M}}\right)^{-l} \frac{P_{2m}(0)}{\gamma_{2m}} \end{aligned}$$

where $2M = N$ and we have used the fact that $P_n(0) = 0$ for n being odd.

Assuming that $M \gg 1$, consider now $(m > 0)$

$$\begin{aligned}\left(\frac{\lambda_{2m}}{\lambda_{2M}}\right)^{-l} &= (2M)^{2l}\frac{1}{(2m)^l(2m+1)^l} = \frac{M^{2l}}{m^l}\frac{1}{(m+1/2)^l}\\ &= \left(\frac{m}{M}\right)^{-2l}\sum_{p=0}^{\infty}\hat{b}_p m^{-p} \quad ,\end{aligned}$$

where the coefficients, $\hat{b}_p$, are given in Lemma 4.5.

Combining this with the result of Lemma 4.1 we recover

$$\begin{aligned}\frac{P_{2m}(0)}{\gamma_{2m}}\left(\frac{\lambda_{2m}}{\lambda_{2M}}\right)^{-l} &= \left((-1)^m\sum_{q=0}^{\infty}\hat{a}_q m^{1/2-q}\right)\times\left(\left(\frac{m}{M}\right)^{-2l}\sum_{p=0}^{\infty}\hat{b}_p m^{-p}\right)\\ &= (-1)^m\left(\frac{m}{M}\right)^{-2l}\sum_{r=0}^{\infty}\left(\sum_{p=0}^{r}\hat{a}_{r-p}\hat{b}_p\right)m^{1/2-r}\\ &= (-1)^m\left(\frac{m}{M}\right)^{-2l}\sum_{r=0}^{\infty}\hat{c}_r m^{1/2-r} \quad .\end{aligned}$$

The coefficients can be found directly, e.g.,

$$\hat{c}_0 = \frac{2}{\sqrt{\pi}} \quad , \quad \hat{c}_1 = -\frac{4l-1}{4\sqrt{\pi}} \quad .$$

Utilizing this, we obtain

$$Q_N^1(\pm1,0) = \left(\frac{-1}{4}\right)^l\sum_{r=0}^{\infty}\hat{c}_r\left[M^{-2l-r+1/2}\sum_{m=1}^{M}(-1)^m\left(1+\sigma(\eta)\right)\eta^{-2l-r+1/2}\right] \quad ,$$

with $\eta = m/M$ and

$$R_N^1(\pm1,0) = \left(\frac{-1}{4}\right)^l\sum_{r=0}^{\infty}\hat{c}_r\left[\sum_{m=M+1}^{\infty}(-1)^m m^{-2l-r+1/2}\right] \quad .$$

Combining the two yields

$$\begin{aligned}Q_N^1(\pm1,0)+R_N^1(\pm1,0) = &\left(\frac{-1}{4}\right)^l\sum_{r=0}^{\infty}\hat{c}_r\\ &\times\left[M^{-2l-r+1/2}\sum_{m=1}^{M}(-1)^m g(\eta)+\sum_{m=M+1}^{\infty}(-1)^m m^{-2l-r+1/2}\right] \quad ,\end{aligned}$$

where

$$g(\eta) = \left(1+\sigma(\eta)\right)\eta^{-2l-r+1/2} \quad .$$

Inspection reveals that the two sums in the above can be expressed to arbitrary accuracy using Lemmas 4.3 and 4.4 to obtain

$$
\begin{aligned}
&Q_N^1(\pm 1,0) + R_N^1(\pm 1,0) = \\
&\left(\frac{-1}{4}\right)^l \sum_{r=0}^{\infty} \hat{c}_r \times \left[\frac{1}{2} M^{1/2-2l-r}\left(g(1)-g(0)-1\right)\right. \\
&\sum_{q=1}^{n-1} M^{-2q+3/2-2l-r} \frac{B_{2q}}{(2q)!}(4^q-1)\left[\frac{\Gamma(2l+r-3/2+2q)}{\Gamma(2l+r-1/2)} + g^{(2q-1)}(1) - g^{(2q-1)}(0)\right] \\
&\left.+\mathcal{O}(M^{-2l-r-2n+3/2})\right]
\end{aligned}
$$

Note that since $n, l > 0$, the remainder vanishes for M increasing. We see immediately, that if $2l - 1/2 > p$ we have

$$
2l - 1/2 > p \;:\; \left|Q_N^1(\pm 1,0) + R_N^1(\pm 1,0)\right| = \mathcal{O}(M^{-p}) \;\; .
$$

In this case, the smoothness of the function dominates that of the filter and the latter sets the convergence rate.

In the complementary case, e.g., $2l - 1/2 < p$, we eliminate the first term by requiring that

$$
g(1) - g(0) = 1 \;\; .
$$

However, from Theorem 4.1, $g(0) = 1$, and we must require $g(1) = 0$ as stated in Definition 3.1.

The second term can be controlled by requiring smoothness of the filter, i.e., we can choose n such that

$$
p - 2l + 3/2 < 2n < p - 2l + 5/2 \;\; ,
$$

to control the coefficient, $M^{-2q+3/2-2l-r}$, ensuring that $M^{-2n+3/2-2l-r}$ is bounded by M^{-p}. With this we have

$$
g^{(2q-1)}(1) - g^{(2q-1)}(0) = -\frac{\Gamma(2l+r-3/2+2q)}{\Gamma(2l+r-1/2)} \;\; .
$$

From Theorem 4.1 we recover

$$
\sigma^{(k)}(0) = 0 \;\; , \;\; k = 1..p-1 \;\; ,
$$

which suffices to guarantee that

$$
g^{(2q-1)}(0) = 0 \;\; , \;\; 0 \le q \le n-1 \;\; .
$$

We also note that

$$
\left.\frac{d^{2q-1}\eta^{-2l-r+1/2}}{d\eta^{2q-1}}\right|_1 = -\frac{\Gamma(2l+r-3/2+2q)}{\Gamma(2l+r-1/2)} \;\; .
$$

Thus, provided

$$\left.\frac{d^{2q-1}\sigma(\eta)\eta^{-2l-r+1/2}}{d\eta^{2q-1}}\right|_1 = 0 \quad ,$$

for $0 \leq q \leq n-1$, we recover

$$2l - 1/2 < p \ : \ \left|Q_N^1(\pm 1, 0) + R_N^1(\pm 1, 0)\right| = \mathcal{O}(M^{-p}) \quad .$$

Using Leibniz's rule for differentiation of products one easily finds that since $2n < p$, this is guaranteed if $\sigma^{(k)}(1) = 0$ for $k = 1..p-1$ as required in Definition 3.1. This completes the proof for $Q_N^1(\pm 1, 0) + R_N^1(\pm 1, 0)$.

However, the result for $Q_N^2(\pm 1, 0) + R_N^2(\pm 1, 0)$ can be obtained in a similar way. In particular, we recover that

$$\frac{P'_{2m-1}(0)}{\gamma_{2m-1}} = -\frac{(2m-1)(2m-2)}{2m+1}\frac{P_{2m}(0)}{\gamma_{2m}} \quad ,$$

from which the equivalent first expansion is recovered as

$$\frac{P'_{2m-1}(0)}{\gamma_{2m-1}} = -(-1)^m \frac{(2m-1)(2m-2)}{2m+1}\sum_{q=0}^{\infty} \hat{a}_q m^{1/2-q} = \sum_{q=0}^{\infty} \tilde{a}_q m^{3/2-q} \quad .$$

In a similar way, we recover

$$\left(\frac{\lambda_{2m-1}}{\lambda_{2M}}\right)^{-l} = \left(\frac{m}{M}\right)^{-2l}\left(\frac{m}{m-1/2}\right)^{l} = \left(\frac{m}{M}\right)^{-2l}\sum_{p=0}^{\infty} \tilde{b}_p m^{-p} \quad ,$$

where $\hat{b}_n$ are given in Lemma 4.5 with $a = 1/2$.

With this, the procedure for $Q_N^1(\pm 1, 0) + R_N^1(\pm 1, 0)$ carries through for $M \gg 1$, thus completing the proof.

■

Let us define

$$k_0 = \inf\left\{k \in \mathsf{N} \ : \ \left|\mathcal{L}_s^k u(c^-) - \mathcal{L}_s^k u(c^+)\right| \neq 0\right\} \quad ,$$

i.e., a measure of regularity. Let us also define the broken norm

$$|||u|||_p = \left(\|u\|^2_{H^p[-1,c^-[} + \|u\|^2_{H^p]c^+,1]}\right)^{1/2} \quad .$$

We are now ready to state the following main result

Theorem 4.2. *Assume that the point of discontinuity, c, is located at the center of the domain, i.e., $c = 0$, and the filter $\sigma(\eta)$, is order $p > 1$, as defined in Definition 3.1.*

Then

$$|u(\pm 1) - \mathcal{F}_N u_N(\pm 1)| \leq C N^{1-p} |||u|||_p \quad ,$$

for $N \gg 1$.

Proof. Recall

$$
\begin{aligned}
&u(\pm 1) - \mathcal{F}_N u_N(\pm 1) = \\
&\sum_{l=0}^{q-1} \left[\mathcal{L}_s^l u(0^-) - \mathcal{L}_s^l u(0^+)\right] \left[\frac{\partial}{\partial s} G^{l+1}(\pm 1, s)\bigg|_{s=0} - \frac{\partial}{\partial s} K_N^{l+1}(\pm 1, s)\bigg|_{s=0} \right] \\
&- \sum_{l=0}^{q-1} \frac{\partial}{\partial s} \left[\mathcal{L}_s^l u(0^-) - \mathcal{L}_s^l u(0^+)\right] \left[G^{l+1}(\pm 1, 0) - K_N^{l+1}(\pm 1, c)\right] \\
&+ \int_{-1}^{1} \mathcal{L}_s^q u(s) \left[G^{l+1}(\pm 1, s) - K_N^q(\pm 1, s)\right] \, ds \quad .
\end{aligned}
$$

Clearly, if $k_0 \geq 2q$, the first two terms vanish, and the result follows from Theorem 4.1 by taking $p = 2q$.

On the other hand, if $k_0 < 2q$, we recover

$$
\begin{aligned}
&|u(\pm 1) - \mathcal{F}_N u_N(\pm 1)| \leq \\
&\sum_{l=0}^{q-1} \left|\mathcal{L}_s^l u(0^-) - \mathcal{L}_s^l u(0^+)\right| \left| \frac{\partial}{\partial s} G^{l+1}(\pm 1, s)\bigg|_{s=0} - \frac{\partial}{\partial s} K_N^{l+1}(\pm 1, s)\bigg|_{s=0} \right| \\
&- \sum_{l=0}^{q-1} \frac{\partial}{\partial s} \left|\mathcal{L}_s^l u(0^-) - \mathcal{L}_s^l u(0^+)\right| \left|G^{l+1}(\pm 1, 0) - K_N^{l+1}(\pm 1, c)\right| \\
&+ \int_{-1}^{1} \mathcal{L}_s^q u(s) \left|G^{l+1}(\pm 1, s) - K_N^q(\pm 1, s)\right| \, ds \quad .
\end{aligned}
$$

Borrowing the result from Lemma 4.6, we immediately recover

$$
|u(\pm 1) - \mathcal{F}_N u_N(\pm 1)| \leq C N^{1-2q} |||u|||_{2q} \quad ,
$$

again recovering the desired result with $p = 2q$. ■

Based on this partial result, and the extensive experiments provided previously, we make the following conjecture

Conjecture 1. Assume that the point of discontinuity, c, is located in the interior of the domain and that the filter $\sigma(\eta)$, is order $p > 1$, as defined in Definition 3.1.

Then for all $x \neq c$

$$
|u(x) - \mathcal{F}_N u_N(x)| \leq C N^{1-p} |||u|||_p \quad ,
$$

where C is a constant depending on $|x - c|$. ■

The conjecture is in line with all previous results and similar in spirit to that obtainable for filtered Fourier expansions [27]. In this latter work, a finite set of points of discontinuities is considered and this extension can likewise be pursued in the above also, the key being the proof of Lemma 4.6 for arbitrary separation between x and c. However, the lack of translation invariance of the orthogonal polynomials makes it difficult to see how to accomplish this within the current approach.

Further insight into the working of the filter can be gained by leaving the rigor of the above discussion and consider the effect of filtering a polynomial representation of a general function. Consider

$$u_N = \sum_{n=0}^{N} \hat{u}_n P_n(x) \quad ,$$

and the filtered function

$$\mathcal{F}_N u_N = \sum_{n=0}^{N} \sigma(n) \hat{u}_n P_n(x) \quad .$$

Let us for simplicity assume that

$$\sigma(n) = 1 - \left(\frac{n}{N}\right)^p \quad .$$

We note that this does not strictly adhere to Definition 3.1 but it does contain the minimum smoothness around $n = 0$ to avoid the destruction of the accuracy of the native expansion. Furthermore, the results presented in Sec. 3.1 using the exponential filter, Eq.(13), suggest that the smoothness of the filter around $\eta = 1$, required to complete the proof of Theorem 4.2, may be sufficient but not necessary.

To appreciate the action of the filter on the function, consider

$$\begin{aligned} u_N(x) - \mathcal{F}_N u_N(x) &= \sum_{n=0}^{N} (1 - \sigma(n)) \hat{u}_n P_n(x) \\ &= N^{-p} \sum_{n=0}^{N} n^p \hat{u}_n P_n(x) \quad . \end{aligned}$$

This yields

$$\begin{aligned} |u_N(x) - \mathcal{F}_N u_N(x)| &\leq N^{-p} \left| \sum_{n=0}^{N} \hat{u}_n \lambda_n^{p/2} P_n(x) \right| \\ &= N^{-p} \left| \sum_{n=0}^{N} \hat{u}_n \left(\frac{d}{dx} (1 - x^2) \frac{d}{dx} \right)^{p/2} P_n(x) \right| \\ &= N^{-p} \left| \sum_{n=0}^{N} \hat{u}_n \left((1 - x^2) \frac{d^2}{dx^2} - 2x \frac{d}{dx} \right)^{p/2} P_n(x) \right| \quad . \end{aligned}$$

This highlights the non-uniformity with which the filter modifies the function, u_N. In fact, we have the two extreme cases

$$|x| \simeq 0 \ : \ |u_N(x) - \mathcal{F}_N u_N(x)| \leq N^{-p} \left| \sum_{n=0}^{N} \hat{u}_n \frac{d^p}{dx^p} P_n(x) \right| = N^{-p} \left| \frac{d^p}{dx^p} u_N(x) \right| \quad ,$$

and

$$|x| \simeq 1 \ : \ |u_N(x) - \mathcal{F}_N u_N(x)| \leq \frac{\sqrt{2}^p}{N^p} \left| \sum_{n=0}^{N} \hat{u}_n \frac{d^{p/2}}{dx^{p/2}} P_n(x) \right| = \frac{\sqrt{2}^p}{N^p} \left| \frac{d^{p/2}}{dx^{p/2}} u_N(x) \right| \quad .$$

This highlights the non-uniformity with which the filter modifies a general function, u_N. Clearly, the impact of the filter can be expected to be strongest in the interior of the domain and weakens as one approaches the boundaries. It is worthwhile emphasizing that while the impact of the filter is minimized at the boundaries of the domain, the function is modified everywhere.

4.2. Impact on Stability

With some added understanding of the impact on accuracy of the use of a filter, let us consider the impact a filter may have on stability, as observed in Sec. 3.2.

Consider the linear problem

$$\frac{\partial u}{\partial t} + a(x) \frac{\partial u}{\partial x} = 0 \ \ , x \in [-1, 1] \ \ . \tag{25}$$

For simplicity we assume that $a(\pm 1)$ is positive but that $a(x)$ in general can change sign inside the domain. The data is given as

$$u(-1, t) = g(t) \ \ , \ \ u(x, 0) = f(x) \ \ .$$

As in Sec. 3.2 we seek a polynomial solution, $u_N(x, t)$, that satisfies

$$\left. \frac{du_N}{dt} \right|_{x_i} + a(x_i) \sum_{j=0}^{N} \mathrm{D}_{ij} u_N(x_j) = -l_0(x_i) \frac{N(N+1)}{4} a(-1) \left[u_N(-1, t) - g(t) \right] \tag{26}$$

To first expose the source of potential instabilities, write Eq.(26) as

$$\frac{\partial u_N}{\partial t} + \mathcal{N}_1 u_N + \mathcal{N}_2 u_N + \mathcal{N}_3 u_N = -\mathcal{I}_N(l_0(x)) \frac{N(N+1)}{4} a(-1) \left[u_N(-1, t) \right] \quad ,$$

where $\mathcal{I}_N$ represents the interpolation. We have defined the three operators

$$\mathcal{N}_1 u_N = \frac{1}{2} \frac{\partial}{\partial x} \mathcal{I}_N a(x) u_N + \frac{1}{2} \mathcal{I}_N \left(a(x) \frac{\partial u_N}{\partial x} \right) \quad ,$$

$$\mathcal{N}_2 u_N = \frac{1}{2} \mathcal{I}_N \left(a(x) \frac{\partial u_N}{\partial x} \right) - \frac{1}{2} \mathcal{I}_N \frac{\partial a(x) u_N}{\partial x} \quad ,$$

$$\mathcal{N}_3 u_N = \frac{1}{2}\mathcal{I}_N \frac{\partial a(x) u_N}{\partial x} - \frac{1}{2}\frac{\partial}{\partial x}\mathcal{I}_N a(x) u_N \quad .$$

To understand stability in an energy sense, consider

$$[u_N, \mathcal{N}_1 u_N]_N = \frac{1}{2}\left[u_N, \frac{\partial}{\partial x}\mathcal{I}_N a(x) u_N\right]_N + \frac{1}{2}\left[u_N, \mathcal{I}_N\left(a(x)\frac{\partial u_N}{\partial x}\right)\right]_N \quad .$$

The accuracy of the Gauss-Lobatto quadrature allows integration by parts of the first term to recover

$$[u_N, \mathcal{N}_1 u_N]_N = \frac{1}{2}\left[a(1)u_N^2(1) - a(-1)u_N^2(-1)\right] \quad .$$

By inspection we immediately have

$$[u_N, \mathcal{N}_2 u_N]_N \leq \frac{1}{2}\max_x |a_x| \|u_N\|_N^2 \quad .$$

Finally, consider

$$[u_N, \mathcal{N}_3 u_N]_N \leq C\left(\|u_N\|_N^2 + \|\mathcal{N}_3 u_N\|_N^2\right) \quad .$$

Realizing the $\mathcal{N}_3 u_N$ is simply the commutation error between interpolation and differentiation, we recover

$$\|\mathcal{N}_3 u_N\|_N^2 \leq C N^{2-2q} \|u^{(q)}\|_N^2 \quad ,$$

by the classic result [4, 1] for commutation errors in Legendre-Gauss-Lobatto collocation methods. Note that the constant, C, depends on a and its derivatives but not on N. A similar result can be obtained for a Legendre Galerkin method as the source is to be found in the lack of commutation between differentiation and projection/interpolation and not in the aliasing errors.

This yields the result

$$\frac{d}{dt}\|u_N\|_N^2 \leq -a(1)u_N^2(1) - \max_x |a_x| \|u_N\|_N^2 + C\left(\|u_N\|_N^2 + N^{2-2q}\|u^{(q)}\|_N^2\right) \quad .$$

Clearly, the latter term, originating from $\mathcal{N}_3 u_N$ is not controlled and may, thus, drive the scheme unstable. This can be expected to be a particular problem when q is low, i.e., for problems with limited regularity.

As illustrated in the examples in Sec. 3.2, this is a real effect and may well drive the scheme unstable even for smooth problems which develop steep marginally resolved gradients. We also observed, however, that filtering appears to be able to control this effectively.

To obtain an intuitive understanding of how the filter can stabilize this instability, let us approximate the filter as

$$\sigma(\eta) = 1 - \alpha\eta^p \quad .$$

While this may be, but likely is not, insufficient to recover spectral convergence as stated in Conjecture 1 it satisfies the minimum requirements from Theorem 4.1. Furthermore, it is a simple approximation to the exponential filter, Eq.(13), which, for high values of p, modifies the solution significantly less than when using the optimal filter, Eq.(12).

Consider

$$\begin{aligned}\mathcal{F}_N u_N &= \sum_{n=0}^{N} \sigma\left(\frac{n}{N}\right) \hat{u}_n P_n(x) = u_N(x) - \frac{\alpha}{N^p} \sum_{n=0}^{N} n^p \hat{u}_n P_n(x) \\ &\simeq u_N(x) + (-1)^{p/2+1} \frac{\alpha}{N^p} \sum_{n=0}^{N} \hat{u}_n \left(\frac{d}{dx}(1-x^2)\frac{d}{dx}\right)^{p/2} P_n(x) \\ &= u_N(x) + (-1)^{p/2+1} \frac{\alpha}{N^p} \left(\frac{d}{dx}(1-x^2)\frac{d}{dx}\right)^{p/2} u_N(x) \ .\end{aligned}$$

This can be recognized as a forward Euler approximation with time step Δt of

$$\frac{du_N}{dt} = (-1)^{p/2+1} \varepsilon \left(\frac{d}{dx}(1-x^2)\frac{d}{dx}\right)^{p/2} u_N \ \ , \ \ \varepsilon = \frac{\alpha}{\Delta t N^p} \ . \tag{27}$$

Summation over all nodes and repeated integration by parts yields

$$\frac{1}{2}\frac{d}{dt}\|u_N\|_N^2 \leq -\varepsilon \|u_N^{(p/2)}\|_N^2 \leq -\varepsilon N^{-p} \|u_N\|_N^2 \ .$$

We observe immediately that the filtering process is dissipative as one would intuitively expect. Furthermore, filtering corresponds approximately to solving a dissipative equation which is, however, well-posed and stable even in the absence of boundary conditions.

If one assumes, as is most often done, that the filter is applied after each time step, the analogy between the filter and the dissipative problem above allows one to consider the combined problem as that of solving the modified equation

$$\frac{\partial u}{\partial t} + a(x)\frac{\partial u}{\partial x} = (-1)^{p/2+1} \varepsilon \left(\frac{d}{dx}(1-x^2)\frac{d}{dx}\right)^{p/2} u \ \ , \ \ , x \in [-1,1] \ . \tag{28}$$

subject to the same boundary conditions as Eq.(25) due to the special singular nature of the dissipative term. This approximation is valid to an $\mathcal{O}(\Delta t)$ splitting error in time.

Now considering the ordinary exponential filter, Eq.(13), we can repeat the above line of arguments using the representation

$$\sigma_E(\eta) = 1 + \sum_{k=1}^{\infty} \frac{1}{k!}(-\alpha\eta^p)^k \ .$$

Using this in the stability analysis above, we recover a new term of the form

$$
\begin{aligned}
&-\sum_{k=1}^{\infty}(-1)^{kp/2+k}\frac{1}{k!}\frac{\alpha^k}{N^{pk}}\left[u_N,\left(\frac{d}{dx}(1-x^2)\frac{d}{dx}\right)^{pk/2}u_N\right]\\
&\leq \sum_{k=1}^{\infty}\frac{1}{k!}\frac{\alpha^k}{N^{pk}}\|u_N^{(pk/2)}\|_N^2 \leq \|u\|_N^2\sum_{k=1}^{\infty}\frac{1}{k!}\alpha^k = \alpha\exp(\alpha)\|u_N\|_N^2 \quad .
\end{aligned}
$$

Thus, we recover the stability statement for the filtered problem as

$$
\begin{aligned}
\frac{d}{dt}\|u_N\|_N^2 \leq \quad & -a(1)u_N^2(1) - \max_x |a_x|\|u_N\|_N^2\\
& +C\left(\|u_N\|_N^2 + N^2\|u_N\|_N^2\right) - \frac{\alpha\exp(\alpha)}{\Delta t}\|u_N\|_N^2 \quad .
\end{aligned}
$$

Using an explicit time integration scheme, $\Delta t \propto N^{-2}$ [13, 16], implies that

$$
\alpha\exp(\alpha) \geq C \quad ,
$$

suffices to recover stability. Clearly, this can always be done, thus confirming the ability of the exponential filter to fully stabilize the instability observed in Sec. 3.2.

5. CONCLUDING REMARKS

Although filtering in spectral and pseudospectral polynomial methods for solving time-dependent partial differential equations is widely used, a rigorous theory remains largely unknown. With the expected increasing use of high-order and spectral methods in the future, it seems timely to build at least some foundation for such techniques.

In this paper we have initiated this by shedding some light on the impact of filtering in Legendre spectral methods, both in terms of accuracy of the expansion and in terms of the stability and robustness that filtering adds to the method. In the latter case, a simple example and subsequent analysis highlights the stabilization offered by the filter which essentially acts as a high-order dissipative term added to the equation. A central observation is that the dissipative operator is nonuniform and singular at the boundaries, thus not requiring any additional boundary conditions [2].

In terms of accuracy, the results are only partial but nevertheless confirms the computational experiments, showing that filtering can restore high-order accuracy away from points of discontinuity. The smoothness of the filter is a critical parameter to ensure that the filter does not adversely affect the accuracy of smooth functions, i.e., a step function as used in classical dealiasing methods [3] may be good for stability but impacts the accuracy adversely.

The filter properties, defined in Definition 3.1, are sufficient but may not be necessary. In fact, one could conjecture, based on computational results, that the conditions in Theorem 4.1 are both necessary and sufficient.

While the analysis provides some foundation for the use of spectral filtering in polynomial methods, it leaves one important question open: how does one choose the order of the filter, p, in a particular applications.

At this point in time, this is largely a question with answers based on experience. However, with the understanding we have developed here, some guidelines can be offered.

In time-dependent computations, the primary concern is often stability, i.e., one must choose p sufficiently low to increase the local dissipation, to ensure a stable computation. Choosing p too low, however, will impact the accuracy significantly throughout the computational domain. Thus, if one finds that $p = 4$ is needed for stability, this indicates severe under resolution and one should explore an increase in the resolution, hopefully with the award that p can also be increased.

A useful guideline is to seek to use as high a value of p as possible without destroying stability. The range of $p = 6..16$ is generally reasonable. Higher values of p has little effect and having to decrease p below 6 most often indicates that one tries to do too much with too little. Using a filter is a matter of striking a balance. However, with a bit of experience and a few tests, one often gains significant advantages in terms of both accuracy and robustness.

ACKNOWLEDGMENT

The work of the first author was partly supported by NSF Career Award DMS-0132967, by NSF International Award INT-0307475, by ARO under contract DAAD19-01-1-0631 and by the Alfred P. Sloan Foundation through a Sloan Research Fellowship.

The second author gratefully acknowledges the partial support of NSF Career Award NSF-CCF0347791 and the computational support and resources provided by the Scientific Computing and Imaging Institute at the University of Utah.

REFERENCES

1. C. Bernardi and Y. Maday, *Polynomial Interpolation Results in Sobolev Spaces*, J. Comput. Appl. Math. **43**(1992), pp. 53-80.
2. J.P. Boyd, *Two Comments on Filtering (Artificial Viscosity) for Chebyshev and Legendre Spectral and Spectral Element Methods: Preserving Boundary Conditions and Interpretation of the Filter as a Diffusion*, J. Comput. Phys. **142**(1998), pp. 283-288.
3. C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang, *Spectral Methods in Fluid Dynamics.* Springer Series in Computational Physics. Springer-Verlag. New York, 1988.
4. C. Canuto and A. Quarteroni, *Approximation Results for Orthogonal Polynomials in Sobolev Spaces*, Math. Comp. **38**(1982), pp. 67-86.
5. W. S. Don, *Numerical Study of Pseudospectral Methods in Shock Wave Applications*, J. Comput. Phys. **110**(1994), pp. 103-111.
6. W. S. Don and D. Gottlieb, *Spectral Simulation of Supersonic Reactive Flows*, SIAM J. Numer. Anal. **35**(1998), pp. 2370-2384.
7. A. Erdélyi (Eds), *Higher Transcendental Functions, Vol II.* Robert E. Krieger Publishing Company, Florida, 1981.
8. M.O. Deville, P.F. Fischer, and E.H. Mund, *High-Order Methods for Incompressible Fluid Flow*, Cambridge University Press, 2002.
9. P.F. Fischer and J.S. Mullen, *Filter-Based Stabilization of Spectral Element Methods*, C. R. Acad. Sci. Paris **332**(2001), pp. 265-270 (2001).
10. D. Gottlieb, S.A. Orszag, and E. Turkel, *Stability of Pseudospectral and Finite-Difference Methods for Variable Coefficient Problems*, Math. Comp. **37**(1981), pp. 293-305.
11. 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.

12. D. Gottlieb and C. W. Shu, *On the Gibbs Phenomenon and its Resolution*, SIAM Review **39**(1997), pp. 644-668.

13. D. Gottlieb and J. S. Hesthaven, *Spectral Methods for Hyperbolic Problems*, J. Comp. Appl. Math. **128**(2001), pp. 83-131.

14. J. S. Hesthaven and D. Gottlieb, *A Stable Penalty Method for the Compressible Navier-Stokes Equations. I. Open Boundary Conditions*, SIAM J. Sci. Comp. **17**(1996), 579-612.

15. J. S. Hesthaven, *Spectral Penalty Methods*, Appl. Numer. Math. **23**(2000), pp. 23-41.

16. J. S. Hesthaven, S. Gottlieb, and D. Gottlieb, *Spectral Methods for Time-Dependent Problems*, Cambridge University Press, Cambridge, UK, 2007.

17. A Kaneveky, M.H. Carpenter, and J.S. Hesthaven, *Idempotent Filtering in Spectral and Spectral Element Methods*, J. Comput. Phys. **220**(2006), pp. 41-58.

18. G.E. Karniadakis and S. J. Sherwin, *Spectral/hp Element Methods for CFD*, Oxford University Press, Oxford, UK, 1999.

19. R.M. Kirby and G. E. Karniadakis, *De-aliasing on non-uniform grids: algorithms and applications*, J. Comput. Phys. **191**(2003) pp. 249-264

20. D.A. Kopriva, *A Practical Assessment of Spectral Accuracy for Hyperbolic Problems with Discontinuities*, J. Sci. Comput. **2**(1987), pp. 249-262.

21. H.O. Kreiss and J. Oliger, *Stability of the Fourier Method*, SIAM J. Numer. Anal. **16**(1979), pp. 421-433.

22. Y. Maday and E. Tadmor, *Analysis of the Spectral Vanishing Viscosity Method for Periodic Conservation Laws*, SIAM J. Numer. Anal. **26**(1989), pp. 854-870.

23. Y. Maday, S. M. Ould Kaper, and E. Tadmor, *Legendre Pseudospectral Viscosity Method for Nonlinear Conservation Laws*, SIAM J. Numer. Anal. **30**(1993), pp. 321-342.

24. R. Pasquetti and C. J. Xu, *Comments on "Filter-Based Stabilization of Spectral Element Methods"*, J. Comput. Phys. **182**(2002), pp. 646-650.

25. E. Tadmor, *Convergence of Spectral Methods for Nonlinear Conservation Laws*, SIAM J. Numer. Anal. **26**(1989), pp. 30-44.

26. E. Tadmor and J. Tanner, *Adaptive mollifiers – High Resolution Recovery of Piecewise Smooth Data from its Spectral Information*, Foundat. Comput. Math. **2**(2002), pp. 155-189.

27. H. Vandeven, *Family of Spectral Filters for Discontinuous Problems*, J. Scient. Comput. **6**(1991), pp. 159-192.