ON ONE-SIDED FILTERS FOR SPECTRAL FOURIER APPROXIMATIONS OF DISCONTINUOUS FUNCTIONS

Wei Cai
David Gottlieb
Chi-Wang Shu

Contract No. NAS1-18605
March 1991

Institute for Computer Applications in Science and Engineering
NASA Langley Research Center
Hampton, Virginia 23665-5225

Operated by the Universities Space Research Association
ON ONE-SIDED FILTERS FOR SPECTRAL FOURIER APPROXIMATIONS OF DISCONTINUOUS FUNCTIONS

Wei Cai ¹
Department of Mathematics
University of North Carolina at Charlotte
Charlotte, NC 28223

David Gottlieb ² and Chi-Wang Shu ³
Division of Applied Mathematics
Brown University
Providence, RI 02912

ABSTRACT

In this paper we prove the existence of one-sided filters, for spectral Fourier approximations of discontinuous functions, which can recover spectral accuracy up to the discontinuity from one side. We also use a least square procedure to construct such a filter and test it on several discontinuous functions numerically.

¹Research supported by NSF grant ASC-9005874.
²Research supported by AFOSR grant 90-0093, NSF grant DMS-8810150, and by NASA contract NAS1-18605 while the author was in residence at ICASE, NASA Langley Research Center, Hampton, VA 23665
³Research supported by NSF grant DMS-8810150, NASA Langley grant NAG1-1145, AFOSR grant 90-0093, and by NASA contract NAS1-18605 while the author was in residence at ICASE, NASA Langley Research Center, Hampton, VA 23665
1 Introduction

A spectral Fourier approximation for a $2\pi$ periodic function $u(x)$ is

$$u_N(x) = \sum_{k=-N}^{N} \hat{u}_k e^{ikx} \quad (1.1)$$

where the Fourier coefficients $\hat{u}_k$ are defined by

$$\hat{u}_k = \frac{1}{2\pi} \int_{0}^{2\pi} u(x) e^{-ikx} \, dx \quad (1.2)$$

for Fourier Galerkin, and by

$$\hat{u}_k = \frac{1}{2N + 1} \sum_{j=0}^{2N} u(x_j) e^{-ikx_j}, \quad x_j = \frac{2\pi j}{2N + 1} \quad (1.3)$$

for Fourier collocation. Spectral methods are well-known for their high accuracy in the approximation of smooth functions and in solving partial differential equations with smooth solutions [3], [2]. For a discontinuous function, however, unmodulated spectral approximation produces Gibbs oscillations and yields first order accuracy even in the smooth region [3]. The good news is that, even if the accuracy is poor in the point-wise sense, it is still excellent for the moments [4]:

Lemma 1.1 If $u(x) \in L^2[0,2\pi]$ and $u_N(x)$ is its Fourier Galerkin sum (1.1)-(1.2), then for any $2\pi$ periodic function $v(x) \in C^\infty$, we have

$$\left| \int_{0}^{2\pi} (u(x) - u_N(x))v(x) \, dx \right| \leq \frac{\|u(x)\|_{L^2}}{N^s} \|u\|_{L^2} \quad \forall s \geq 1 \quad (1.4)$$

Proof: By the orthogonality of trigonometric polynomials, we have

$$\left| \int_{0}^{2\pi} (u(x) - u_N(x))v(x) \, dx \right| = \left| \int_{0}^{2\pi} (u(x) - u_N(x))(v(x) - v_N(x)) \, dx \right|$$

$$\leq \|u\|_{L^2} \|v - v_N\|_{L^2}$$

Integration by parts yields
\[ \hat{v}_k = \frac{1}{2\pi} \int_0^{2\pi} v(x) e^{-ikx} dx = \frac{(-1)^s}{2\pi(-i)^s} \int_0^{2\pi} v(x) e^{-ikx} dx \]

We then have

\[ \|v - v_N\|_{L^2} = \sqrt{\sum_{|k| > N} |\hat{v}_k|^2} \leq \frac{1}{N\pi} \sqrt{\sum_{|k| > N} |\hat{v}_k^{(s)}|^2} \leq \frac{\|v(x)\|_{L^2}}{N\pi} \]

Lemma 1.1 implies that the spectral approximation \( u_N(x) \) does indeed contain the information of \( u(x) \) by preserving accurately the moments of \( u(x) \) against any \( C^\infty \) functions. Various filters are designed in the literature to extract this information. The early work in this direction includes [8] by Majda, McDonough and Osher and [7] by Kreiss and Oliger. The filters, in the phase space, are of the form

\[ u_N^f(x) = \sum_{k=-N}^{N} \sigma_k^N \hat{v}_k e^{ikx} \]  

(1.5)

where \( \sigma_k^N = \sigma_{-k}^N \) are real numbers, which decay smoothly from 1 to 0 when \( |k| \) goes from 0 to \( N \). The filter (1.5) is also equivalent to a convolution in the physical space

\[ u_N^f(x) = u_N \ast K_N(x) = \frac{1}{2\pi} \int_0^{2\pi} u_N(y) K_N(x - y) dy \]  

(1.6)

with a kernel \( K_N(x) \) defined by

\[ K_N(x) = \sum_{k=-N}^{N} \sigma_k^N e^{ikx} \]  

(1.7)

Vandeven [10] studied this type of filters in more detail. He proved the following result, which we will use in Section 2. For simplicity of presentations and without loss of generality, we assume the function \( u(x) \) has only one discontinuity located at \( x = 0 \), i.e., \( u(x) \) is smooth but not periodic over \([0, 2\pi)\).
Lemma 1.2 (Vandeven) If \( u(x) \) is an analytic but not periodic function in \([0, 2\pi)\), then for any \( 0 < \varepsilon < 1 \), the filtered Fourier sum \( u_N^\varepsilon(x) \) defined by (1.5) with

\[
\sigma_k^N = 1 - \frac{(2p - 1)!}{(p - 1)!^2} \int_0^{k/N} [t(1-t)]^{p-1} dt, \quad p = N^\frac{1}{4}
\] (1.8)
satisfies the following error estimates:

\[
\max_{\frac{1}{N^1/4} \leq x \leq 2\pi - \frac{1}{N^1/4}} |u(x) - u_N^\varepsilon(x)| \leq \frac{N^\beta}{(CN^{\frac{1}{4}})^N} (1.9)
\]

where \( C \) and \( \beta \) are constants independent of \( N \).

Similar results were also obtained in Gottlieb and Tadmor [4], by directly constructing the filter kernel \( K_N(x) \) in (1.6).

In this paper, we will use \( C \) or \( \tilde{C} \) for a generic constant independent of \( N \), which may be different in different locations.

Lemma 1.2 establishes the spectral convergence of the filtered Fourier sum (1.5) to the function \( u(x) \), in a distance \( \frac{1}{N^1/4} \) away from the discontinuity \( x = 0 \) \((mod \ 2\pi)\). If \( u(x) \) is not piecewise analytic but piecewise \( C^r \), similar estimates can be obtained for algebraic convergence.

The choice of \( \sigma_k^N \) in (1.8) is not unique. In practice, exponential filter of the form

\[
\sigma_k^N = e^{-\alpha(x/N)^{2p}}
\] (1.10)
is often used due to its simplicity and good numerical results. Here \( 2p \) is the order of the filter and can be chosen depending on \( N \). \( \alpha \) is a constant chosen so that \( e^{-\alpha} \), the filter for the last mode, is machine zero. Other frequently used filters include the raised cosine filter and the Gottlieb-Tadmor filter [4]. We refer the readers to Kopriva [6] for the definition and an extensive comparison of various filters in practical calculations.
All of the filters mentioned above are two-sided filters. That is, a two-sided region around the discontinuity has to be excluded in the error estimates. See for example (1.9). This is expected since one uses a symmetric kernel $K_N(x)$ in (1.6), as a result of using real numbers $\sigma_k^N = \sigma_{-k}^N$ in (1.5). The Gibbs oscillation is still present in the filtered sum $u_N^r(x)$ near the discontinuity, see Figure 1. This is usually unacceptable for solving nonlinear partial differential equations since these oscillations may eventually pollute the smooth regions and/or trigger nonlinear instability. In [1], we proposed a way to overcome this difficulty by introducing non-smooth functions (saw-tooth functions) into the basis of trigonometric functions. It would be nice, however, if one could work within the trigonometric polynomial basis and use filters to completely remove the Gibbs oscillations and to obtain uniform convergence up to the discontinuity. This is the motivation for the one-sided filters discussed in this paper. We refer the readers to Mock and Lax [9] for the early work in this direction.

In Section 2 we prove the existence of one-sided filters, i.e., we prove the existence of complex numbers $\sigma_k^N$, such that for any analytic but not periodic function $u(x)$ in $[0, 2\pi)$, the filtered Fourier sum (1.5) satisfies a uniform error estimate

$$\max_{0 \leq x \leq x_R} |u(x) - u_N^r(x)| \leq \frac{N^\beta}{(CN^{1/4})^N}$$

(1.11)

where

$$x_R = 2\pi - \frac{2}{N^{1-\frac{1}{4}\varepsilon}}$$

(1.12)

and $C$ and $\beta$ are again constants independent of $N$. This filter is naturally labelled "right-sided" filters since it can recover the spectral accuracy up to the discontinuity $x = 0$ from the right. A symmetry consideration leads to the "left-sided" filters by taking the conjugate of the complex numbers $\sigma_k^N$.

The proof in Section 2 is constructive. However, the filters obtained that way can not be satisfactorily applied in actual computations unless $N$ is extremely large. For a practical range of $N$ between 8 and 32 (16 to 64 grid points for collocation), we try to find a better
one-sided filter through a least-square procedure described in Section 3. Several numerical examples are also provided in that section.

We point out that the one-sided filters discussed in this paper are neither unique nor necessarily optimal in practical calculations. Work is under way to explore the construction of one-sided filters directly in the physical space (1.6). A systematic investigation about one-sided filters in the phase space, similar to the approach used in [10], is also under consideration.

2 Existence of One-Sided Filters

The main result in this section is the following theorem:

**Theorem 2.1** If \( u(z) \) is an analytic (but not periodic) function in \([0, 2\pi)\), then for any \( 0 < \varepsilon < \frac{4}{7} \), the one-sided filter defined by

\[
\sigma_k^N = \sigma_k^{T,N} \sum \limits_{i=1}^{m} \left( \begin{array}{c} m \\ l \end{array} \right) (-1)^{l+1} e^{ik\Delta}
\]

where

\[
\Delta = \frac{1}{N^{1-\varepsilon}}, \quad m = N^4
\]

and \( \sigma_k^{T,N} \) is the Vandeven two-sided filter defined by (1.8), produces a filtered Fourier sum (1.5) satisfying

\[
\max_{0 \leq z \leq x_R} |u(x) - u_N^\varepsilon(x)| \leq \frac{N^\beta}{(CN^{\frac{4}{7}})N^4}
\]

where \( x_R \) is given by (1.12), \( C \) and \( \beta \) are constants independent of \( N \).

**Proof:** We have, by (2.1) and (1.5),

\[
u_N^\varepsilon(x) = \sum_{k=-N}^{N} \sigma_k^{T,N} \sum_{l=1}^{m} \left( \begin{array}{c} m \\ l \end{array} \right) (-1)^{l+1} e^{ik\Delta} \hat{u}_k e^{ikx}
\]
\[
\begin{align*}
E_1 &= \sum_{l=1}^{m} \binom{m}{l} (-1)^{l+1} u_N^{\sigma,T}(x + l\Delta) \\
&= \sum_{l=1}^{m} \binom{m}{l} (-1)^{l+1} u(x + l\Delta) + E_1
\end{align*}
\] (2.4)

where \( E_1 = \sum_{l=1}^{m} \binom{m}{l} (-1)^{l+1} (u_N^{\sigma,T}(x + l\Delta) - u(x + l\Delta)) \), and

\[
u_N^{\sigma,T}(x) = \sum_{k=-N}^{N} \sigma_k T N_k e^{ikx}
\] (2.5)

is the two-sided filtered Fourier sum. By (2.2) and Lemma 1.2, \( E_1 \) is small whenever \( 0 \leq x \leq x^*_R \), i.e., whenever \( \frac{1}{N^1} \leq x + l\Delta \leq 2\pi - \frac{1}{N^1} \) for \( 1 \leq l \leq m \):

\[
|E_1| \leq \sum_{l=1}^{m} \binom{m}{l} \frac{N^\beta}{(CN^{\frac{1}{2}})N^\ell} = 2^m \frac{N^\beta}{(CN^{\frac{1}{2}})N^\ell} = \frac{N^\beta}{(CN^{\frac{1}{2}})N^\ell}, \quad \hat{C} = \frac{C}{2}
\]

On the other hand, by Newton's formula, the sum on the last line of (2.4) satisfies, for some \( x < \xi < x + m\Delta \),

\[
\sum_{l=1}^{m} \binom{m}{l} (-1)^{l+1} u(x + l\Delta) = u(x) - u^{(m)}(\xi)(-\Delta)^m
\]

Since we assume \( u(x) \) is analytic in \([0, 2\pi]\), we have, for some \( \rho > 0 \),

\[
|u^{(m)}(x)| \leq M(\rho) \frac{m!}{\rho^m}
\]

We then have, by (2.2) and the Stirling's formula \( \lim_{m \to \infty} \frac{m^m \sqrt{2\pi m}}{m!} = 1 \),

\[
|u^{(m)}(\xi)\Delta^m| \leq \frac{Cm!}{(\rho N^{1-\varepsilon})^m} \leq \frac{N^\delta}{(CN^{1-\varepsilon})^\eta N^\ell} \leq \frac{N^\delta}{(CN^{\frac{1}{2}})N^\ell}
\]

where the last inequality is valid because \( \varepsilon < \frac{4}{7} \). This completes the proof.

\[\square\]

**Remark** (1) The filter (2.1) is based on the following simple idea: since the two-sided filtered Fourier sum \( u_N^{\sigma,T}(x) \) in (2.5) is a good approximation to \( u(x) \) in a region \( \Delta \) away from the
discontinuity, we can use extrapolations from points inside this region to approximate points outside of it. The right-sided filtered Fourier sum \( u_N^r(x) \) based on (2.1) is just such an extrapolation using \( u_N^r(x) \). In practice, this simple, equally spaced extrapolation may not be the best. One would expect Chebyshev extrapolations to do a better job.

(2) A left-sided filter is obtained by taking the complex conjugate of \( u_N^l(k) \) in (2.1). The proof follows the same lines.

3 Efficient One-Sided Filters and Numerical Examples

The one-sided filters described in Section 2 are good asymptotically, but a very large \( N \) is needed in order for \( m \) in (2.2) to be of reasonable size. It is our experience that asymptotically correct choices are often not necessarily optimal for small \( N \). We thus use a least square procedure, described below, with the objective of obtaining more efficient one-sided filters for a practical range of \( N \) between 8 and 32 (between 16 and 64 grid points for collocation):

**Least Square Procedure**: We make an ansæze

\[
\sigma_k^N = \begin{cases} 
\sigma_k^{T,N} + w(\frac{k}{N})q(\frac{k}{N}), & k \geq 0 \\
\sigma_k^N, & k < 0 
\end{cases} \quad (3.1)
\]

where the weight function \( w(t) \) is defined by

\[
w(t) = e^{-|t|/N} \quad (3.2)
\]

\( \sigma_k^{T,N} \) is the Van de Ven two-sided filter defined by (1.8) with a parameter \( p \), and \( q(t) \in P_s \), the collection of all polynomials of degree at most \( s \), with complex coefficients, defined on \([0,1]\). The right-sided filter \( \sigma_k^N \) is then defined by taking the following minimum:

\[
\min_{q(t) \in P_s} \sum_{r=1}^{R} \int_0^A [(x^r)^N(x) - x^r]^2 \, dx \quad (3.3)
\]
where $A$ is a parameter between 0 and $2\pi$, and $(x^r)^{\text{f}}(x)$ is the filtered Fourier sum (1.5), using the filter (3.1), of the function $u(x) = x^r$.

The parameters involved in the Least Square Procedure are $p$ for the order of accuracy of the two-sided filter (1.8), $s$ for the degree of polynomials $q(t)$ in (3.1), $R$ for the number of terms in the summation (3.3), and $A$ in (3.3) for the domain of integration. The implementation of the minimization (3.3) involves the choice of a basis for $P_s$ (we use a Chebyshev basis over $[0,1]$) and the solution of the resulting linear system.

The condition $\sigma_k^{N_k} = \sigma_k^R$ is imposed in (3.1) in order to get a real kernel in (1.7) or, equivalently, to get a real filtered Fourier sum (1.5) for a real function $u(x)$. The exponential weight function (3.2) is used so that the term added to the two-sided filter $\sigma_k^{T_r}$ will not destroy the accuracy (see [10] for details of the relation between properties of $\sigma_k$ and the accuracy of the filters). The minimization (3.3) is based upon the assumption that the right-sided filtered Fourier sums for $u(x) = x^r$, with $1 \leq r \leq R$, should be uniformly accurate in the interval $[0,A]$. We have also tested the procedure by replacing $x^r$ in (3.3) by the $r$-th order Chebyshev polynomial $T_r(x)$, obtaining similar results.

A crucial issue for the success of this approach is the sensitivity of the filters obtained with respect to the parameters $p$, $s$, $R$ and $A$. According to the experience for two-sided filters, $p$ should be chosen related to $N$. One could also try the (very costly) procedure of minimization over some of those parameters in certain ranges. The filter might be expected to work well for polynomials because of the choice $u(x) = x^r$ in the minimization (3.3) but must be extensively tested for other functions.

We perform numerical experiments using different parameters $p$, $s$, $R$, $A$ to get the filters, then testing them on a non-polynomial function

\[ u(x) = \cos(1.5 x) \quad x \in [0,2\pi) \] (3.4)
Since we enforce $2\pi$ periodicity, this function and all its derivatives are discontinuous at $x = 0 \ (mod \ 2\pi)$. We have experimented with $p$ between 3 and 15 depending on the size of $N$ (the choice is based on our experience with two-sided filters). It turns out that the result is most sensitive to the choice of $s$ and the optimal $s$ increases rather rapidly with $N$. This indicates that the choice for the polynomial space $P_s$ or the exponential weight function in (3.2) may not be adequate for large $N$. It also limits the size of $N$ we can use in this approach since the resulting linear system, which we solve by using a routine in IMSL, becomes ill-conditioned for large $s$. However, the result seems not very sensitive to the choice of $R$ (we have tried $4 < R < 15$) and $A$ (bounded away from 0 and $2\pi$).

We now show the results obtained for $N = 8$ with $s = 2$, $p = 3$, $R = 5$ and $A = \pi$; for $N = 16$ with $s = 5$, $p = 5$, $R = 5$ and $A = \pi$; and for $N = 32$ with $s = 18$, $p = 6$, $R = 6$ and $A = \pi$, in Figures 2 to 6.

In Figure 2, we plot the filter $\sigma_k^N$ for its real and imaginary parts. No special pattern can be observed. In Figure 3, we plot the filter $\sigma_k^N$ for its magnitudes and arguments. Clearly it shows a straight line for the arguments immediately after the first few modes. This interesting phenomena can be loosely explained in the following way: A straight line for the arguments corresponds to a pure shifting for the kernel (1.7). It makes the kernel non-symmetric around zero and mainly supported in one-side, allowing for one-sided recovery of accuracies. For the first few modes the arguments have to be close to zero to ensure accuracy of the filter for the low modes.

In Figure 4, we plot the filter kernels (1.7) in the physical space. We can see that they are indeed one-sided approximate $\delta$ functions, i.e., they are supported mainly to the left side of the discontinuity $0 \ (mod \ 2\pi)$.

In Figure 5, we plot the one-sided filtered Fourier sum (1.5), using plus signs, against the exact solution (3.4) (shifted by $\pi$ to show more clearly the discontinuity), with $N = 16$. Each point is either left-side filtered or right-side filtered according to which side it is closer to the discontinuity. We can clearly see the advantage of using one-sided filters in getting
accurate, non-oscillatory results.

Finally, in Figure 6, we plot the errors of the one-sided filtered Fourier sum (1.5) from the exact function (3.4), with \( N = 8, 16, 32 \), on a logarithm scale. Again, each point is either left-side filtered or right-side filtered according to which side it is closer to the discontinuity. We can see the uniform convergence up to the discontinuity and a faster than algebraic convergence for the one-sided filtered Fourier sums.

The same filter is also used to other functions such as \( u(x) = e^{\cos(1.5x)} \), \( u(x) = \frac{1}{1+x^2} \), etc., with similar results. We would also like to point out that even if we have restricted our discussion to the Galerkin method (1.2), the collocation case (1.3) can be analysed in a similar fashion with some assumptions on the location of the discontinuity between the grids (say, in the middle of two grids). We have performed the same numerical experiments with collocation method, and have obtained similar results. Details are omitted here.

The main potential of one-sided filters is in solving hyperbolic partial differential equations with discontinuous solutions. They can either be used during the final stage of post-processing to recover accurate point values of the numerical solution near the discontinuity, or used in each time step to obtain non-oscillatory numerical solutions. In order to perform the latter one might obtain three Fourier sums, namely right-sided filtered \( u_{NR}^a(x) \), two-sided filtered \( u_{NT}^a(x) \) and left-sided filtered \( u_{NL}^a(x) \), at each point (this only involves two additional FFT’s if implemented in the phase space (1.5)), and use some local criterion to decide whether there is a discontinuity nearby, and if yes, to which side, then to decide which filtered solution to use. This is very similar to the ENO idea in finite difference [5], and is currently under investigation.
References


Figure 1: The sawtooth function (background solid line); the Fourier sum (1.1) with $N = 16$ (solid line) and the two-sided filtered Fourier sum using (1.8) with $p = 6$ (dashed line). We can see that the two-sided filter removes oscillations away from the discontinuity but still leaves over- and under-shoots near the discontinuity.

Figure 2: The right-sided filters $\sigma_k^N$: the real parts $\text{Re}(\sigma_k^N)$ (solid lines) and the imaginary parts $\text{Im}(\sigma_k^N)$ (dashed lines), for (a) $N = 8$; (b) $N = 16$ and (c) $N = 32$. 

Figure 2(a)
Figure 2(b)

Figure 2(c)
Figure 3: The right-sided filters $\sigma^N_k$: the magnitudes $|\sigma^N_k|$ (solid lines) and the arguments $\text{Arg}(\sigma^N_k)$ (dashed lines), for (a) $N = 8$; (b) $N = 16$ and (c) $N = 32$.
Figure 3(c)

Figure 4: The kernels $K_N(x)$ in (1.7) for the right-sided filters $\sigma_k^N$: (a) $N = 8$; (b) $N = 16$ and (c) $N = 32$.

Figure 4(a)
Figure 5: The function (3.4) shifted by π (solid line) and the one-sided filtered Fourier sum (the plus signs), with \( N = 16 \).

Figure 6: The errors \( |u(x) - u_N^e(x)| \), of the one-sided filtered Fourier sum to (3.4), in logarithm scales, for \( N = 8, 16, 32 \).
ON ONE-SIDED FILTERS FOR SPECTRAL FOURIER APPROXIMATIONS OF DISCONTINUOUS FUNCTIONS

In this paper we prove the existence of one-sided filters, for spectral Fourier approximations of discontinuous functions, which can recover spectral accuracy up to the discontinuity from one side. We also use a least square procedure to construct such a filter and test it on several discontinuous functions numerically.