This paper is part of the following report:
TITLE: Algorithms For Approximation IV. Proceedings of the 2001 International Symposium
To order the complete compilation report, use: ADA412833

The component part is provided here to allow users access to individually authored sections of proceedings, annals, symposia, etc. However, the component should be considered within the context of the overall compilation report and not as a stand-alone technical report.

The following component part numbers comprise the compilation report:
ADP013708 thru ADP013761
Accurate approximation of functions with discontinuities, using low order Fourier coefficients

R. K. Wright

Department of Mathematics and Statistics, UVM, Burlington, VT, 05445 USA.
wright@emba.uvm.edu

Abstract

In previous work we introduced a method of using polynomial splines with appropriate discontinuities to approximate a piecewise smooth function $f$ with jump discontinuities of $f$ and $f'$. The information used is location of discontinuities, and low order, possibly noisy Fourier coefficients. The number of discontinuities was limited to two at most, and the discontinuities needed to lie at meshpoints in a uniform mesh. We showed that the linear operator corresponding to the method is $L_2$-bounded with a modest bound, and thus that the method is $L_2$-robust in the presence of noise. In the present paper we develop a new method of analysis which enables us to determine operator bounds that are valid for arbitrarily many discontinuities. The new analysis allows discontinuities to be placed arbitrarily. Given a placement, an initially uniform spline mesh of width $h$ must be used such that nearest meshpoints to discontinuities are at least $4h$ apart (discontinuities then replace these meshpoints); the number of available Fourier coefficients must be at least three times the number of mesh intervals in a period. The previous work was restricted to quadratic splines; the present work includes cubic splines. Much of the analysis uses exact computations with a computer algebra system. We give an example to illustrate the accuracy of the method using noisy Fourier coefficients.

1 Introduction

We consider approximating a function $f$ when the information consists of low order, possibly noisy Fourier coefficients, and knowledge that $f$ is smooth except for jumps of $f$ or $f'$ at known locations but unknown magnitudes. We will work with a method, introduced in [10], which amounts to linear least squares fitting of the available coefficients with the coefficients of splines with appropriately placed discontinuities. Since we anticipate applications to ill-posed problems where boundedness of the solution operator is crucial, we develop a method for bounding the norm of this operator. The bounding method depends heavily on exact computations in certain spline spaces. These computations are fundamentally finite dimensional linear algebra with rational integer coefficients. Their goal is to develop upper bounds for the norms of certain projector operators whose norms are naturally expressed in terms of generalized eigenvalues, and to prove by exact computation that the bounds are correct. A computer algebra system is used for the computations. The programming is detailed in [9].
Accurate approximation, discontinuities

In [10] we obtained bounds under much more restrictive conditions than in the present paper. In [10] the splines were quadratic only, while here results also are given for cubic splines. The analysis in [10] required all knots of the approximating splines to be uniformly spaced, and since the discontinuities are at the knots, the location of discontinuities was limited. Further, in [10] the estimation process is linear in the total number of discontinuities, and produces results unacceptably large for cases with more than one discontinuity of \( f \) and two of \( f' \).

Others ([2, 3, 4, 5]) have addressed questions of accurate approximations to functions with discontinuities given Fourier coefficients as information. In [8] we give examples which show that those methods can substantially magnify noise in the coefficients; our main concern here is to prove robustness of our method. We illustrate with an example in Section 5.

2 General linear space-theoretic results

Let \( V \) be a real Hilbert space with inner product \( \langle \cdot , \cdot \rangle \). We will denote the norm associated with \( \langle \cdot , \cdot \rangle \) by \( || \cdot || \). Let \( P \) and \( Q \) be closed subspaces of \( V \); suppose \( P \) is the orthogonal projector on \( P \). Here, as in [10], we deal with the approximation \( f^* \) obtained as the solution to the constrained least squares problem

\[
\min || Pf^* - Pf ||, \quad f^* \in Q.
\]

Assuming that \( P \) is invertible as a mapping on \( Q \), we denote by \( P^+ \) the mapping from \( P(Q) \) to \( Q \) which inverts \( P \). It is not hard to verify that \( f^* = P^+ R P f \) where \( R \) is the orthogonal projector on \( P(Q) \). Let \( A \) denote the operator that takes \( f \) to \( f^* \).

**Theorem 2.1** Let \( C \) be a mapping from \( V \) to \( Q \). Let \( \epsilon \) be \( T \)-periodic and in \( L_2(0, T) \). Then

\[
|| A(P f + \epsilon) - f || \leq (|| P^+ || + 1) || C f - f || + || P^+ || || \epsilon ||.
\]

**Proof:** \( A(P f + \epsilon) = A f + A \epsilon \). \( || A f - f || \leq || A f - C f || + || C f - f || = || A(f - C f) || + || C f - f || \leq (|| A || + 1) || f - C f ||. \) \( || A || = || P^+ R P || \leq || P^+ || \) because \( P \) and \( R \) are orthogonal projections. \( \square \)

A main objective of the following work will be to bound \( || P^+ || \). This will be done by establishing upper bounds for \( || I - P || \) as a mapping on \( Q \). From these, bounds can easily be derived for \( || P^+ || \).

**Theorem 2.2** Let \( \eta < 1 \) exist such that \( ||(I - P)q|| \leq \eta || q || \), for all \( q \in Q \). Then \( P \) is injective as a mapping on \( Q \) and for all \( h \in P(Q) \), \( P^+ \), the inverse of the restriction of \( P \) to \( Q \), satisfies

\[
|| P^+ h ||^2 \leq \frac{1}{1 - \eta^2} || h ||^2.
\]
We will obtain bounds for $|I - P|$ by considering the projector perpendicular to a spline space $G$ which is more tractable than $PV$, and on which $I - P$ is small. In the next section, $Q$ is the approximating spline space, $S$ a subspace of maximally continuous splines, and $G$ is a space of maximally continuous splines whose knots are in a mesh refining the mesh for the members of $S$. $S$ and $G$ have orthogonal projectors $S$ and $G$, respectively. The following estimates $|I - P|$ in terms of $|I - G|$.

**Theorem 2.3** Suppose $||(I - P)g|| \leq \eta_0 ||g||$ for all $g \in G$. Suppose $||(I - G)q|| \leq \eta_1 ||q||$ for all $q \in Q$. Then $||(I - P)q|| \leq (\eta_0 + \eta_1) ||q||$ for all for all $q \in Q$.

**Proof:** For $q \in Q$, $||(I - P)q|| \leq ||(I - P)Gq|| + ||(I - P)(I - G)q||$. $||(I - P)Gq|| \leq \eta_0 ||Gq|| \leq \eta_0 ||q||$, and $||(I - P)(I - G)q|| \leq ||(I - G)q|| \leq \eta_1 ||q||$. □

Theorem 2.4 enables us to bound $||I - G||$ on $Q$ by instead bounding projectors orthogonal to small subspaces of $G$, restricted to small subspaces of $Q$.

**Theorem 2.4** Let $G$ and $S$ be closed subspaces of $V$ with $S \subseteq G \cap Q$. Let $Q_1, Q_2, \ldots, Q_r$ be nonzero mutually orthogonal subspaces of $V$. Let $Q_i \subseteq Q \cap V$, $1 \leq i \leq r$ be nonzero closed subspaces such that $Q = S + Q_1 + Q_2 + \cdots + Q_r$. Let $G_i \subseteq G \cap V_i, H_i \subseteq S \cap V_i$, $1 \leq i \leq r$ be nonzero closed subspaces with orthogonal projectors $G_i, H_i$. Let $\nu$ be a constant such that $||(I - G_i)q_i||^2 \leq \nu ||H_i q_i||^2$ for all $q_i \in Q_i, 1 \leq i \leq r$. Then $||(I - G)q||^2 \leq \nu ||q||^2$ for all $q \in Q$.

**Proof:** $q \in Q$ can be written $q = s + v$ where $s \in S$ and $v = q_1 + q_2 + \cdots + q_r$, $q_i \in Q_i$, $1 \leq i \leq r$. $||(I - G)q|| = ||(I - G)v||$ since $S \subseteq G$. Let $F = G_1 + G_2 + \cdots + G_r$. Since $G_1 + G_2 + \cdots + G_r \subseteq G$, $||(I - G)v||^2 \leq ||(I - F)v||^2 = \sum_{i=1}^r ||(I - G_i)q_i||^2$, the latter equality because of orthogonality of the $G_i$. $||q||^2 \geq ||(I - S)v||^2 \geq \sum_{i=1}^r ||H_i q_i||^2 = \sum_{i=1}^r ||H_i q_i||^2$, since $\sum_{i=1}^r H_i \subseteq S$, and the $H_i$ are orthogonal. If all $H_i q_i = 0$ the hypothesis implies all $(I - G_i)q_i = 0$. The above then implies $(I - G)q = 0$, and the conclusion is true. We proceed assuming $H_i q_i \neq 0$ for some $i$ and let $N_i$ be the set of all those $i$. Then

$$\frac{||(I - G)q||^2}{||q||^2} \leq \sum_{i \in N} \frac{||(I - G_i)q_i||^2}{||H_i q_i||^2}.$$

An elementary argument shows the quotient of sums is $\leq \nu$ since for each $i \in N_i$, $||(I - G_i)q_i||^2 / ||H_i q_i||^2 \leq \nu$. □

### 3 Bounds for restricted projectors

Below, we specialize the spaces of the last section, and get our main results. Let $T > 0$ be a fixed period. We take $V$ to be the space of real-valued $T$-periodic functions which belong $L_2(I)$ for some period interval $I$. On $V$ and its subspaces we define the inner product $(f, g) = \int_I f(t)g(t)dt$, $I$ a period interval. The other realizations are defined in the statements and proofs of the following results. Lemma 3.1 sets up an application of Theorem 2.4; Theorem 3.2 uses this, together with Theorem 2.2, to get our main result.

**Lemma 3.1** Let $X$ be a finite set of points in $[0, T)$. Let $N \geq 4$ be an integer. Let $K = \{iT/N, 0 \leq i \leq N\}$: for each $x \in X$, let $k_x$ be a member of $K$ closest to $x$ where
Accurate approximation, discontinuities

0 is identified with T. Assume N large enough that between any two distinct \( k_x \) are at least three other members of \( K \). Let \( K_X \) result from substituting in \( K \) each \( x \in X \) for its \( k_x \). For \( m = 3, 4 \) let \( Q \) be the space of \( m \)-th order \( T \)-periodic polynomial splines with \( K_X \) as knots and with continuity \( C^{m-2} \) at all knots except the \( x \in X \), where no continuity is required. Let \( G \) be the space of \( m \)-th order periodic splines with knots in \([0, T)\) at the points \( \{iT/(3N), 0 \leq i \leq 3N\} \), and let \( G \) be the orthogonal projector on \( G \). Then \( I - G \) restricted to \( Q \) satisfies \( ||I - G||_2^2 \leq .69 \) if \( m = 3 \), and \( ||I - G||_2^2 \leq .9 \) if \( m = 4 \).

**Proof:** Let \( S \) be the subspace of \( Q \) consisting of those splines which are \( C^\infty \) at the \( k_x \). Clearly \( S \subseteq G \). Let \( h = T/N \). Fix \( x = x_I \in X = \{x_1, x_2, \ldots, x_r\} \) and let \( y_0 = x_I \), \( y_\alpha = k_x - \alpha h, \alpha = -2, -1, 1, 2 \). Take \( V_i \) to be the subspace of \( V \) consisting of those functions with support in \([y_2 - y_2, y_2)\) and its \( T \)-translates.

For \( m = 3 \) let \( j_1 \) and \( j_2 \) be B-splines with knots \( y_{-1}, y_0, y_0, y_1 \) and \( y_0, y_0, y_0, y_1 \); let \( j_3 \) be the difference of the B-splines with knots \( y_{-2}, y_{-1}, y_0, y_1 \) and \( y_{-1}, y_0, y_1, y_2 \) (see [1] for explanation of multiplicity versus degree of continuity). For \( m = 4 \) let \( j_1 \) and \( j_2 \) be B-splines with knots \( y_{-1}, y_0, y_0, y_0 \) and \( y_0, y_0, y_0, y_1 \); let \( j_3 \) be the difference of the B-splines with knots \( y_{-2}, y_{-1}, y_0, y_1 \) and \( y_{-1}, y_0, y_1, y_2 \); and let \( j_4 \) be the B-spline with knots \( y_{-2}, y_{-1}, y_0, y_1, y_2 \). Since \( y_2 - y_2 < T \) we may identify the \( j_\alpha \) with their \( T \)-periodic extensions.

Let \( Q_i \) be the space of splines whose generic member is \( q_i = \sum_{\alpha=1}^{m} c_\alpha j_\alpha \) for constants \( c_\alpha \). For each \( i \) nonzero members of \( Q_i \) have continuity from \( C^{m-2} \) through full discontinuity at \( x_I \), while members of \( S \) are \( C^1 \) at \( x_I \). It follows that \( S \cap (Q_1 + Q_2 + \cdots + Q_r) = 0 \) and \( Q = S + Q_1 + \cdots + Q_r \).

Let \( G_i \) be the subspace of \( G \) with basis the \( C^{m-2} \) periodic B-splines whose knots in the period containing \([y_2 - y_2, y_2)\) are length \( m + 1 \) sublists of consecutive knots from the list \((ah/3 + k_x, -6 \leq \alpha \leq 6) \). Let \( H_i \) be the space of those \( m \)-th order periodic splines which in \([-T/2 + k_x, T/2 + k_x] \) have support in \([y_2 - y_2, y_2] \), which have knots at the \( y_i \), \( i \neq 0 \) and at \( x \), are \( C^{m-2} \) at \( y_1 \) and \( y_1 \), which may be fully discontinuous at \( y_2, y_2 \), and which are orthogonal to all members of \( S \). \( ||(I - G_i)q_i||^2 / ||H_i q_i||^2 \) is a ratio of quadratic forms in the \( c_\alpha \). An upper bound \( \nu \) for it can be obtained as an upper bound for the eigenvalues of the pencil \( A - \lambda B \) where \( a_\alpha \beta = \langle (I - G_i)j_\alpha, (I - G_i)j_\beta \rangle, \) \( b_\alpha \beta = \langle H_i j_\alpha, H_i j_\beta \rangle, 1 \leq \alpha, \beta \leq m \).

In [9] explicit bases for the spaces \( G_i \) and \( H_i \) are calculated as \( m \)-th order splines. From their definitions ([1]), B-splines are rational functions of the knots, and thus are also inner products of B-splines. The null-basis and orthogonal projection calculations in [9] use standard methods which involve only rational operations. Thus the \( (I - G_i)j_\alpha \) and \( H_i j_\alpha \) and then the \( a_\alpha \beta \) and \( b_\alpha \beta \) are rational functions of the knots of \( q_i \), so long as \( x \) remains in \([k_x, k_x + h/3]\). When \( x \) crosses into \([k_x + h/3, k_x + h/2]\), thus crossing knots for splines in \( G_i \), the rational functions change, so in general the matrix entries are piecewise rational functions of \( x \).

Let \( \nu \) be a conjectured upper bound for the maximum eigenvalue \( \lambda_{max} \) of \( A - \lambda B \) (in [9] a floating point approximation to \( \lambda_{max} \) is plotted as a function of \( x \); \( \nu \) is determined from inspecting this plot). For computational convenience in [9] we represent \( x \) as \( 2ch/3 + k_x, 0 \leq \epsilon \leq 1/2 \) for \( x \leq k_x + h/3 \), and as \((1 + \epsilon)h/3 + k_x, 0 \leq \epsilon \leq 1/2 \) for \( k_x + h/3 \leq x \leq \)
For further convenience we take \( k_x = 0 \), clearly losing no generality. We have represented only \( x \geq k_x \), but because of symmetry, \( x \leq k_x \) produces the same bounds.

Since \( h \) is a linear factor in all knots in the calculation, we see that \( a_{\alpha \beta} \) and \( b_{\alpha \beta} \) can be written as \( h \) multiplying piecewise rational functions of \( \epsilon \) (with integer rational coefficients). The determinant of \( A - \nu B \) is thus \( h^m \) times a piecewise rational function of \( \epsilon \). The MAXRAT algorithm ([9]) proves that its reciprocal is bounded as a function of \( \epsilon \) in the appropriate ranges, so the determinant itself is bounded away from 0. In [9], \( \epsilon \) is then set equal to 0 in \( A - \tau B \), and the determinant of that matrix is then shown to have \( m \) sign changes as \( \tau \) decreases from \( \nu \). Thus the conjectured value \( \nu \) bounds all eigenvalues of \( A - \lambda B \) for all values of \( x \). The upper bounds thus obtained are \( \nu = .69 \) for \( m = 3 \) and \( \nu = .9 \) for \( m = 4 \). We emphasize that the B-splines, matrix entries, and determinants all are calculated exactly, using the Maple ([6, 7]) computer algebra system, so the bounding property of \( \nu \) is rigorously proven. Since the bounds we obtain apply to the spaces \( G_i \) and \( B_i \) associated with any one of the \( x_i \), they satisfy the hypotheses of Theorem 2.4 which now provides our conclusions.

Our main result now follows.

**Theorem 3.2** Let the hypotheses be those of Lemma 3.1. In addition, let \( P \) be the orthogonal projector onto the space of \( n \)-th order real-valued \( T \)-periodic trigonometric polynomials, where \( n \geq 3N \). If \( m = 3 \), we have \( \|P^+\|_2 \leq 2.4 \), while if \( m = 4 \), we have \( \|P^+\|_2 \leq 4.5 \).

**Proof:** The space \( G \) in Lemma 3.1 consists of periodic splines with uniformly spaced knots. Theorem 3.1 of [10] implies that \( \|I - P\|_2 \leq (\alpha/(1 + \alpha))^{1/2} \) where

\[
\alpha = 4 \sum_{r=1}^{\infty} (1/(1 + 2r))^{2m}.
\]

In [9] we use this formula to get upper bounds of \( .076 \) when \( m = 3 \) and \( .025 \) when \( m = 4 \). Taking these bounds as \( \eta_0 \) in Theorem 2.3 and taking the bounds from Lemma 3.1 as \( \eta_1 \) in Theorem 2.3, we obtain from that theorem bounds for \( \|I - P\|_2 \) of \( .907 \) for \( m = 3 \) and \( .974 \) for \( m = 4 \). Theorem 2.2 now applies to produce the present results.

Above, we required \( n \geq 3N \); under this condition we can get our simplest and most comprehensive results. Since we contemplate applying our results where the number \( n \) of useful coefficients may be limited, we have tried to get versions of Theorem 3.2 where \( n \) is smaller compared with \( N \). We have no useful versions for \( n < 3N \) and \( m = 4 \) (cubic splines). The following result for quadratic splines may be useful. To formulate it, let \( \epsilon_1 = \max\{|x - k_x|N/T\} \). In the previous results, the separation of the values \( x \) from their nearest uniform mesh points \( k_x \) was unrestricted, which corresponds to \( \epsilon_1 = 1/2 \). Here, we can get results for quadratic splines, and \( n \geq 2N \), provided the \( x \) are more restricted; our methods of analysis "blow up" for \( n \geq 2N \) as \( \epsilon_1 \) approaches a number slightly larger than \( .25 \).
Theorem 3.3  Let \( m = 3 \) (quadratic splines); let \( n \geq 2N \). Otherwise, let the hypotheses be those of Theorem 3.2. Corresponding to the list 0, .1, .2, .25 for values of \( \varepsilon_1 \), we have the list of values 1.7, 2.1, 3.9, 16 as bounds for \( \|P^+\| \).

Proof: For each of the cases for \( \varepsilon \), an argument similar to the proof of Lemma 3.1 applies to produce a bound \( \eta_1 \) for \( \|I - G\|_2 \) where \( G \) now is defined using the uniform knot spacing \( 1/(2N) \) rather than \( 1/(3N) \). The only difference in the argument is that here, a discontinuity location \( x \) always stays in the interval \( [k_x, k_x + \varepsilon_1 h] \) where \( h = T/N \), so the matrix entries and determinants can be treated as functions of \( \varepsilon \) in \( [0, \varepsilon_1] \). Each bound \( \eta_1 \) now is used just as in the proof of Theorem 3.2, to get the present bounds for \( \|P^+\|_2 \). \( \square \)

4  Uniform norm bounds

Using representers of point evaluation, as in [8], we can get uniform norm bounds for \( P^+ \), and thus for \( A \). The arguments are similar to those in [8]. The main difference is that there the mesh is uniform and the order \( m \) is 3. The constructions of representers extend fairly easily to the present case: here the norms of representers are functions both of the evaluation point and the location of the discontinuity nearest to the evaluation point. One can show that for each point \( t \in [0, T) \), a spline \( r_t \) exists in a space \( \mathcal{U} \) containing \( \mathcal{Q} \), such that \( \langle r_t, q \rangle = q(t) \) for each \( q \in \mathcal{Q} \), and such that \( \|r_t\|_2 \leq k/\sqrt{h} \) where \( k = 5, m = 3 \) and \( k = 7, m = 4; h = T/N \) as before. The computations for the construction and bound calculations are in [9]. Noting that \( \sqrt{T}/\sqrt{h} = \sqrt{N} \), we have

\[
\|Af\|_\infty \leq \max_t \|r_t\|_2 \|Af\|_2 \leq (k/\sqrt{h})\|P^+\|_2 \sqrt{T}\|f\|_\infty \leq k\sqrt{N}\|P^+\|_2 \|f\|_\infty.
\]

When \( N \leq 100 \) and the hypotheses are those of Lemma 3.1, this gives \( \|Af\|_\infty \leq 120\|f\|_\infty \) for \( m = 3 \), and \( \|Af\|_\infty \leq 315\|f\|_\infty \) for \( m = 4 \).

5  Example

FIG. 1.
We illustrate the method using an example where the function $f$ is $2\pi$-periodic and on $[0, 2\pi)$ consists of the function $e^{-x/6}$ with a piecewise quadratic added, so as to produce discontinuities at 0, 1.5, 2.5, and 4. $f$ is a modification of an example in [2]; for convenience we have shifted that example left by 1 unit, and we have added the exponential term because our method can represent a piecewise quadratic exactly in the absence of noise. Exact (up to 17-decimal digit floating point error) Fourier coefficients are derived from $f$ by exact integration using the Maple ([6, 7]) system. Noisy approximate coefficients are also derived by sampling $f$ at 1024 equidistant values in $[0, 2\pi]$, adding uniformly distributed pseudo-random noise to the samples, and taking the discrete Fourier transform of the samples. In effect, we work with $f + \epsilon$ where $\epsilon$ is a perturbing function. The level of the noise is set so that the discrete $L_2$-norm of the noise vector is 1% of the discrete $L_2$-norm of the vector of samples of $f$. $N = 45$ and thus $n = 135$ are the smallest values of $n$ and $N$ for which the hypotheses of the previous section are satisfied. Using these values, we proceed with $m = 4$ (cubic splines) for each of these cases for Fourier coefficients. Plots of $f$ and of the error for the two cases appear in the figure. The ratio $\|f - A(f + \epsilon)\|_2/\|f\|_2$ is about .005 for the case of 1% noise. In [9] we develop a probabilistic estimate of .0037 for the ratio of $\|\epsilon\|_2/\|f\|_2$. This estimate indicates an $L_2$-norm noise magnification of about 1.35-fold, compared with the upper bound of 4.5 given in Theorem 3.2. The uniform error, for noise-free coefficients, is about $10^{-9}$; computational experiments show this is dominated by truncation error in approximating the exponential term. In [9] we do the corresponding calculations for $m = 3$, and find similar results for 1% noise, with larger, but still small, error for noise-free coefficients.

In [9], we implement Eckhoff's method as described in [3], used on the above data. For noiseless data, the results are comparable to those reported by Eckhoff for similar examples. The uniform norm error seems to be about .06, with errors at jumps somewhat smaller. For 1% noise, the results of Eckhoff's method are about 750-fold in error.

Bibliography
