ISSN 0084-778X

Department
of

APPLIED MATHEMATICS

On the numerical approximation of derivatives by a modified Fourier collocation method

by

Knut S. Eckhoff and Carl Erik Wasberg

Report no. 99 July 1995

UNIVERSITAS BERGENSIS

Universitetet i Bergen

UNIVERSITY OF BERGEN
BERGEN, NORWAY

Department of Mathematics
University of Bergen
5007 Bergen
Norway

ISSN 0084-778x

On the numerical approximation of derivatives by a modified Fourier collocation method

by

Knut S. Eckhoff and Carl Erik Wasberg

Report No. 99 July 1995

On the numerical approximation of derivatives by a modified Fourier collocation method

Knut S. Eckhoff[*] Carl Erik Wasberg[*]

Abstract

A modified Fourier collocation method is applied to calculate highly accurate approximations to the derivatives of piecewise smooth functions. The theoretical asymptotic error estimates are demonstrated to be obtainable in calculations, and high accuracy is achieved even for small numbers of collocation points. The limitations in accuracy and robustness due to finite numerical precision are discussed, and approximation problems arising from the solution of partial differential equations in complex geometries are solved satisfactorily. Robustness of the method with respect to approximation of more smooth functions is also discussed.

1 Introduction

Spectral methods [3, 11] have proven to be efficient tools for obtaining numerical solutions of partial differential equations when high accuracy is required. However, the obtainable accuracy depends strongly on the degree of smoothness of the solution.

By the modified Fourier collocation method presented in [8], functions that are only piecewise smooth can be represented with high order accuracy by using the discrete Fourier coefficients to approximate jumps in the function and its derivatives at the points where the periodic extension of the function is not smooth. The calculation of approximate derivatives are studied in detail in this paper, bearing in mind applications to the solution of partial differential equations.

The main points concerning the representation of functions and calculation of derivatives presented in [8] are reviewed in section 2, while sections 3 and 4 contain more details on the calculations with respect to accuracy and robustness.

An example of a function with one discontinuity point is studied in section 5 to illustrate the effects of different choices that have to be made in the implementation of the approximation method and to make recommendations for later use. In section 6 an example with more discontinuity points is considered, and section 7 illustrates a situation that may appear in connection with complex geometry applications.

[*] Dept. of Mathematics, University of Bergen, Allégt. 55, N-5007 Bergen, NORWAY.

Section 8 deals with the robustness of the method with respect to functions with a higher degree of smoothness than known a priori, and methods of detecting this smoothness are discussed.

2 The approximation method

The purpose of the method is to calculate accurate approximations to the derivatives of a piecewise smooth function $u(x)$ defined on $[0, 2\pi]$, from given values of u at the set of $N + 1$ points

$$x_i = 2\pi i/N, \qquad i = 0, \dots, N. \tag{1}$$

The M points where $u(x)$ is not smooth are called discontinuity points and are denoted by γ_j, $j = 1, \dots, M$. These points are assumed to be known. As in [8], the point $x = 0$ (but not $x = 2\pi$) is considered to be a discontinuity point if the 2π-periodic extension of $u(x)$ is discontinuous or has jumps in any of its first Q derivatives at $x = 0$.

Following [6, 8] we consider the following decomposition of $u(x)$:

$$u(x) = u^Q(x) + \sum_{j=1}^{M} \sum_{n=0}^{Q} A_j^n \, V_n(x; \gamma_j), \qquad 0 \le x \le 2\pi, \tag{2}$$

where

$$V_n(x; \gamma_j) = U_n(x - \gamma_j), \qquad 0 \le x \le 2\pi, \tag{3}$$

and

$$U_n(x) = -\frac{(2\pi)^n}{(n+1)!} B_{n+1}(x/2\pi), \qquad 0 \le x \le 2\pi,$$
$$U_n(x) = U_n(2\pi + x), \qquad -2\pi \le x < 0,$$

for $n = 0, 1, \dots$, where $B_j(x)$, $j = 1, 2, \dots$, are the Bernoulli polynomials [1].

The 2π-periodic extension of $U_n(x)$ is a known function in the space $C_p^{n-1}(0, 2\pi)$ of 2π-periodic, continuous (if $n > 0$), and $n - 1$ times continuously differentiable functions, and it has a jump-discontinuity of magnitude 1 in the nth derivative at $x = m2\pi$, $m = 0, \pm 1, \pm 2, \dots$. From (3) we see that $V_n(x; \gamma_j)$ is just a translation of $U_n(x)$, with a jump in the nth derivative at $x = \gamma_j$. We will therefore refer to these functions as "jump-functions".

If $u(x)$ is at least Q times continuously differentiable for $x \ne \gamma_j$, $j = 1, 2, \dots, M$, and A_j^n is the jump in the nth derivative of u at $x = \gamma_j$ for $n = 0, 1, \dots, Q$, $j = 1, \dots, M$, then all the discontinuities in $u(x)$ and its Q first derivatives are represented by the double sum in (2). Consequently, the term $u^Q(x)$ in (2) represents a function in $C_p^Q(0, 2\pi)$.

The Fourier coefficients of a function $v(x)$ are defined as

$$\widehat{v}_k = \frac{1}{2\pi} \int_0^{2\pi} v(x) \, e^{-ikx} \, dx, \qquad k = 0, \pm 1, \pm 2, \dots, \tag{4}$$

and the Fourier series associated with $v(x)$ is $\sum_{k=-\infty}^{\infty} \widehat{v}_k e^{ikx}$. It is well known [3, 13] that if a piecewise smooth function $v(x)$ is in $C_p^Q(0, 2\pi)$, then the coefficients $\widehat{v}_k$ decay as $O(|k|^{-Q-2})$

as $|k| \to \infty$. Because of this rapid decay, a truncated Fourier series is well suited for the representation of $u^Q(x)$ when Q is sufficiently large.

The approximation of $u(x)$ is constructed on the form

$$w(x) = \sum_{k=-N/2+1}^{N/2-1} \widehat{w}_k^Q \, e^{ikx} + \sum_{j=1}^{M} \sum_{n=0}^{Q} \bar{A}_j^n \, V_n(x;\gamma_j), \qquad 0 \le x \le 2\pi, \tag{5}$$

where each $\bar{A}_j^n$ is an approximation of the exact jump A_j^n.

For the differentiation of (5) we note that [6, 8]

$$\frac{dU_n}{dx}(x) = U_{n-1}(x), \qquad n \ge 1, \quad 0 < x < 2\pi,$$

and

$$\frac{dU_0}{dx}(x) = -1/2\pi, \qquad 0 < x < 2\pi.$$

Thus when $x \neq \gamma_j$, $j = 1, \ldots, M$, the derivative of (5) can be written

$$\frac{dw}{dx}(x) = \sum_{k=-N/2+1}^{N/2-1} \widehat{w}_k^{Q'} \, e^{ikx} + \sum_{j=1}^{M} \sum_{n=0}^{Q} \bar{B}_j^n \, V_n(x;\gamma_j), \tag{6a}$$

where

$$\widehat{w}_k^{Q'} = ik\,\widehat{w}_k, \quad k = \pm 1, \pm 2, \ldots, \pm(N/2-1), \qquad \widehat{w}_0^{Q'} = -\frac{1}{2\pi} \sum_{j=1}^{M} A_j^0, \tag{6b}$$

$$\bar{B}_j^Q = 0, \quad \bar{B}_j^n = A_j^{n+1}, \qquad j = 1, 2, \ldots, M, \quad n = 0, 1, \ldots, Q-1. \tag{6c}$$

Still following [8], it is assumed that all the jumps A_j^0 in the function $u(x)$ are known. This will for example be the case in an application to partial differential equations with Dirichlet boundary conditions and smooth solutions in the interior [8, 9]. The knowledge of these jumps is used to define a new function that is continuous and has a continuous 2π-periodic extension:

$$u^0(x) = u(x) - \sum_{j=1}^{M} A_j^0 V_0(x;\gamma_j), \tag{7}$$

and the corresponding approximating function becomes

$$w^0(x) = \sum_{k=-N/2+1}^{N/2-1} \widehat{w}_k^Q \, e^{ikx} + \sum_{j=1}^{M} \sum_{n=1}^{Q} \bar{A}_j^n \, V_n(x;\gamma_j). \tag{8}$$

If these jumps were not known, the procedure described below would be based on the original functions $u(x)$ and $w(x)$.

It remains to find values for the approximate jumps $\bar{A}_j^n$ and the coefficients $\widehat{w}_k^Q$ in (8) from known grid point values of u^0. For this purpose, the collocation equations

$$w^0(x_i) = u^0(x_i), \qquad i = 0, \ldots, N, \tag{9}$$

are used at the grid points given by (1), and an N-term discrete Fourier transform of (8) is carried out, producing

$$\widetilde{w}_k^0 = \widehat{w}_k^Q + \sum_{j=1}^{M} \sum_{n=1}^{Q} \bar{A}_j^n \, (\widetilde{V}_n)_k(\gamma_j), \qquad k = 0, \pm 1, \ldots, \pm(N/2 - 1), \tag{10}$$

where the discrete Fourier coefficients of $w^0(x)$ are defined as

$$\widetilde{w}_k^0 = \frac{1}{N} \sum_{j=0}^{N-1} w^0(x_j)\, e^{-ikx_j}, \qquad k = 0, \pm 1, \ldots, \pm(N/2 - 1). \tag{11}$$

To determine the relevant values for $\bar{A}_j^n$ we use the fact that the coefficients $\widehat{w}_k^Q$ of the Q times smooth part will asymptotically decay faster with increasing k than the jump-function coefficients $(\widetilde{V}_n)_k(\gamma_j)$, and can therefore be discarded for large $|k|$. This gives the following system of equations:

$$\sum_{n=1}^{Q} \sum_{j=1}^{M} (\widetilde{V}_n)_k(\gamma_j) \bar{A}_j^n = \widetilde{w}_k^0, \qquad k = k_1, k_2, \ldots, k_K, \tag{12}$$

where the MQ unknowns are $\bar{A}_j^n$, $n = 1, \ldots, Q$, $j = 1, \ldots, M$, and k takes on K different values chosen between $-(N/2-1)$ and $N/2-1$. This linear system of equations can be exactly determined or under- or overdetermined, depending on the choices of K and $k_1, k_2, \ldots, k_K$. The system may become underdetermined (rank-deficient) due to finite numerical precision, even if the number of equations K is equal to or greater than MQ.

The most robust and general way to solve the system (12) is to use a least squares method with singular value decomposition [10], as found e.g. in LAPACK [2], where a minimum norm solution is sought if the problem is rank-deficient. Because the derivation is motivated by asymptotic behaviour of Fourier coefficients, the k's used should have the highest possible absolute values. These choices are discussed in sections 4 and 5.

When approximate jumps have been obtained from (12), the coefficients of the trigonometric part of (8) are calculated as the residuals

$$\widehat{w}_k^Q = \widetilde{w}_k^0 - \sum_{j=1}^{M} \sum_{n=1}^{Q} (\widetilde{V}_n)_k(\gamma_j) \bar{A}_j^n, \qquad k = 0, \pm 1, \ldots, \pm(N/2 - 1). \tag{13}$$

It is shown in [8] that (12) gives $\bar{A}_j^n = A_j^n + O(N^{-(Q+1-n)})$ as $N \to \infty$, and in [7] that the mth derivative of $w(x)$ then approximates $\frac{d^m u(x)}{dx^m}$ with error $O(N^{-(Q+1-m)})$ as $N \to \infty$, for $m = 0, 1, \ldots, Q$.

3 Calculation of discrete Fourier coefficients for the jump-functions

To obtain accurate solutions of the system (12), it is important that the discrete Fourier coefficients $(\widetilde{V}_n)_k(\gamma_j)$ are accurately calculated. The discrete Fourier transform of $V_n(x;\gamma)$ is

$$(\widetilde{V}_n)_k(\gamma) = \frac{1}{N}\sum_{j=0}^{N} V_n(x_j;\gamma)\, e^{-ikx_j}, \qquad k = -N/2, \dots, N/2, \tag{14}$$

where the grid points x_j are given by (1). Because $V_n(x;\gamma)$ is $n-1$ times continuously differentiable, the coefficients $(\widetilde{V}_n)_k(\gamma)$ decrease rapidly when k increases, in particular for large n. These coefficients may therefore often be too small to be calculated accurately by the sum (14).

Using the exact Fourier coefficients of $V_n(x;\gamma)$, defined by (4),

$$(\widehat{V}_n)_k(\gamma) = \frac{e^{-ik\gamma}}{2\pi(ik)^{n+1}}, \qquad n = 0,1,2,\dots, \qquad k = \pm 1, \pm 2, \dots, \tag{15}$$

and the following general relation between the discrete and exact Fourier coefficients [3],

$$(\widetilde{V}_n)_k(\gamma) = (\widehat{V}_n)_k(\gamma) + \sum_{m=1}^{\infty} \left[(\widehat{V}_n)_{k+mN}(\gamma) + (\widehat{V}_n)_{k-mN}(\gamma)\right], \quad |k| < N/2, \quad n = 1,2,\dots, \tag{16}$$

the following analytic formula for the discrete coefficients is derived in [8], assuming that $0 \le \gamma \le 2\pi/N$:

$$(\widetilde{V}_n)_k(\gamma) = \frac{(-1)^n e^{-ik\gamma}}{2(iN)^{n+1} n!}\left[\frac{d^n}{dz^n}\left\{e^{izN\gamma}\cot \pi z\right\}\right]_{z=k/N} - \frac{(-1)^n \gamma^n}{2N\, n!}, \qquad n \ge 1, \quad 0 < |k| \le N/2. \tag{17a}$$

For $2\pi/N < \gamma < 2\pi$, there exists an integer l and a $\gamma' \in [0, 2\pi/N]$ such that $\gamma = \frac{2\pi l}{N} + \gamma'$. It follows from the translation property of the discrete Fourier transform that

$$(\widetilde{V}_n)_k(\gamma) = e^{-i2\pi kl/N}(\widetilde{V}_n)_k(\gamma'), \tag{17b}$$

where $(\widetilde{V}_n)_k(\gamma')$ is found from (17a).

When (17a) is used, the derivatives of $e^{izN\gamma}\cot \pi z$ are calculated exactly (e.g. by Maple [4]), and even though the resulting coefficients are also affected by the precision of the computer, they are much more robust than calculation of the sum (14). (An accuracy problem seems to arise for the smallest $|k|$ when $\gamma \ne 0$, but these coefficients can be calculated with sufficient accuracy by (14).

In the rest of this section we assume a single discontinuity point at $x = 0$, and study $(\widetilde{V}_n)_k(0)$ (or $(\widetilde{U}_n)_k$, by (3)). It is seen from (17a) that the coefficients for $U_{2n+1}(x)$, $n = 0,1,2,\dots$, are real, while the coefficients for $U_{2n}(x)$, $n = 0,1,2,\dots$, are imaginary. In the following we study the accuracy of the real (for odd n) and imaginary (for even n) parts of the coefficients calculated by (14) using a Fast Fourier Transform (FFT), compared with the results from (17a).

k	$N = 32$	$N = 64$	$N = 128$	$N = 256$
$N/2-1$	-1.549e-03	-3.844e-04	-9.593e-05	-2.397e-05
$N/2-2$	-1.595e-03	-3.872e-04	-9.611e-05	-2.398e-05
$N/2-3$	-1.675e-03	-3.919e-04	-9.640e-05	-2.400e-05
$N/2-4$	-1.797e-03	-3.987e-04	-9.680e-05	-2.403e-05

Table 1: Real parts of the discrete Fourier coefficients $(\widetilde{U}_1)_k$, FFT/analytic.

The calculations are done in double precision, i.e., with 16 significant digits, and the FFT results agree with the analytic coefficients up to at least the 15th decimal place. Taking into account that the largest coefficient ($|k| = 1$) is approximately $1/2\pi$ (exactly so for the exact coefficients (15)), we can assume that FFT can in general give correct answers up to at least the 14th decimal place for a function where the largest coefficient has magnitude 1.

The non-zero parts of the coefficients $(\widetilde{U}_n)_k$, $|k| = N/2-4, \ldots, N/2-1$, calculated by FFT, agree with the results from (17) with at least 4 digits in the following cases:

- $N = 32$: $n \leq 7$.
- $N = 64$: $n \leq 6$.
- $N = 128$: $n \leq 5$.
- $N = 256$: $n \leq 4$.

Results are given in tables 1–7. Because of accumulated round-off errors, $(\widetilde{U}_n)_{N/2-1}$ calculated by FFT are never smaller than 10^{-17} in absolute value, while the correct values can be several orders of magnitude smaller for large n and N.

We also look at the size of the imaginary (for odd n) and real (for even n) parts of the coefficients calculated by FFT. These parts should be zero according to (17a), and are therefore a measure of the errors in the calculations. Sample results are given in tables 8–11, showing that the errors are mostly of order 10^{-17}–10^{-18} without systematic variations.

k	$N = 32$	$N = 64$	$N = 128$	$N = 256$
$N/2 - 1$	-9.763e-08	-1.495e-09	-2.323e-11	-3.626e-13
$N/2 - 2$	-2.120e-07	-3.051e-09	-4.671e-11	-7.260e-13
$N/2 - 3$	-3.648e-07	-4.736e-09	-7.066e-11	-1.091e-12
$N/2 - 4$	-5.904e-07	-6.624e-09	-9.534e-11	-1.460e-12

Table 2: Imaginary parts of the discrete Fourier coefficients $(\widetilde{U}_4)_k$, FFT/analytic.

	$N = 32$	$N = 64$	$N = 128$	$N = 256$	
k	FFT/analytic	FFT/analytic	FFT/analytic	FFT	Analytic
$N/2 - 1$	-2.059e-08	-3.030e-10	-4.663e-12	-7.266e-14	-7.257e-14
$N/2 - 2$	-2.585e-08	-3.218e-10	-4.734e-12	-7.292e-14	-7.285e-14
$N/2 - 3$	-3.639e-08	-3.546e-10	-4.855e-12	-7.340e-14	-7.332e-14
$N/2 - 4$	-5.582e-08	-4.038e-10	-5.028e-12	-7.405e-14	-7.397e-14

Table 3: Real parts of the discrete Fourier coefficients $(\widetilde{U}_5)_k$.

	$N = 32$	$N = 64$	$N = 128$		$N = 256$	
k	FFT/analytic	FFT/analytic	FFT	Analytic	FFT	Analytic
$N/2 - 1$	5.437e-10	2.051e-12	7.851e-15	7.940e-15	-5.190e-17	3.095e-17
$N/2 - 2$	1.250e-09	4.248e-12	1.592e-14	1.602e-14	-6.088e-17	6.204e-17
$N/2 - 3$	2.359e-09	6.754e-12	2.429e-14	2.439e-14	-5.069e-17	9.339e-17
$N/2 - 4$	4.318e-09	9.766e-12	3.314e-14	3.319e-14	9.621e-18	1.252e-16

Table 4: Imaginary parts of the discrete Fourier coefficients $(\widetilde{U}_6)_k$.

	$N = 32$	$N = 64$		$N = 128$		$N = 256$	
k	FFT/analytic	FFT	Analytic	FFT	Analytic	FFT	Analytic
$N/2-1$	8.493e-11	3.000e-13	2.998e-13	1.235e-15	1.141e-15	1.041e-16	4.428e-18
$N/2-2$	1.223e-10	3.319e-13	3.317e-13	1.258e-15	1.171e-15	1.299e-16	4.457e-18
$N/2-3$	2.045e-10	3.888e-13	3.889e-13	1.313e-15	1.222e-15	1.056e-16	4.506e-18
$N/2-4$	3.764e-10	4.776e-13	4.777e-13	1.337e-15	1.296e-15	6.177e-17	4.575e-18

Table 5: Real parts of the discrete Fourier coefficients $(\widetilde{U}_7)_k$.

	$N = 32$		$N = 64$		$N = 128$		$N = 256$	
k	FFT	Analytic	FFT	Analytic	FFT	Analytic	FFT	Analytic
$N/2-1$	-2.798e-12	-2.798e-12	-2.618e-15	-2.590e-15	-7.866e-17	-2.496e-18	2.353e-17	-2.429e-21
$N/2-2$	-6.901e-12	-6.901e-12	-5.529e-15	-5.465e-15	-8.912e-17	-5.059e-18	3.755e-17	-4.875e-21
$N/2-3$	-1.451e-11	-1.452e-11	-9.009e-15	-8.952e-15	-7.086e-17	-7.760e-18	9.492e-17	-7.353e-21
$N/2-4$	-3.053e-11	-3.053e-11	-1.355e-14	-1.348e-14	-2.215e-17	-1.067e-17	1.148e-16	-9.882e-21

Table 6: Imaginary parts of the discrete Fourier coefficients $(\widetilde{U}_8)_k$.

	$N = 32$		$N = 64$		$N = 128$		$N = 256$	
k	FFT	Analytic	FFT	Analytic	FFT	Analytic	FFT	Analytic
$N/2-1$	1.409e-15	1.500e-15	-1.180e-16	2.975e-19	2.082e-17	6.869e-23	7.633e-17	1.653e-26
$N/2-2$	2.853e-15	2.945e-15	-6.221e-17	3.662e-19	1.118e-18	7.263e-23	2.010e-17	1.677e-26
$N/2-3$	6.924e-15	6.903e-15	-1.248e-16	4.969e-19	1.745e-17	7.941e-23	-1.039e-17	1.717e-26
$N/2-4$	1.789e-14	1.789e-14	-7.305e-17	7.190e-19	4.063e-17	8.940e-23	-3.439e-17	1.773e-26

Table 7: Real parts of the discrete Fourier coefficients $(\widetilde{U}_{11})_k$.

k	$\Im((\widetilde{U}_1)_k)$	$\Re((\widetilde{U}_2)_k)$	$\Re((\widetilde{U}_{10})_k)$	$\Im((\widetilde{U}_{11})_k)$
$N/2-1$	-1.529e-17	-4.464e-17	-4.207e-16	5.168e-17
$N/2-2$	2.885e-17	-2.345e-17	-3.829e-16	9.396e-17
$N/2-3$	1.581e-17	-9.041e-17	-3.540e-16	9.520e-17
$N/2-4$	7.457e-18	-5.077e-17	-3.204e-16	1.532e-16

Table 8: Errors in the discrete Fourier coefficients of jump-functions, calculated by FFT with $N = 32$.

k	$\Im((\widetilde{U}_1)_k)$	$\Re((\widetilde{U}_2)_k)$	$\Re((\widetilde{U}_{10})_k)$	$\Im((\widetilde{U}_{11})_k)$
$N/2-1$	-1.055e-17	-5.766e-17	5.881e-17	6.769e-17
$N/2-2$	2.031e-18	-3.044e-17	1.417e-16	4.552e-17
$N/2-3$	-5.035e-18	-1.901e-17	8.006e-17	5.328e-17
$N/2-4$	-2.331e-17	1.507e-17	1.320e-16	1.288e-16

Table 9: Errors in the discrete Fourier coefficients of jump-functions, calculated by FFT with $N = 64$.

k	$\Im((\widetilde{U}_1)_k)$	$\Re((\widetilde{U}_2)_k)$	$\Re((\widetilde{U}_{10})_k)$	$\Im((\widetilde{U}_{11})_k)$
$N/2-1$	-1.859e-17	2.548e-17	5.917e-17	1.289e-17
$N/2-2$	-1.252e-17	3.220e-17	8.056e-17	-6.054e-18
$N/2-3$	-1.116e-18	3.340e-17	1.061e-16	8.198e-18
$N/2-4$	1.064e-18	1.963e-17	1.080e-16	-1.608e-17

Table 10: Errors in the discrete Fourier coefficients of jump-functions, calculated by FFT with $N = 128$.

k	$\Im((\widetilde{U}_1)_k)$	$\Re((\widetilde{U}_2)_k)$	$\Re((\widetilde{U}_{10})_k)$	$\Im((\widetilde{U}_{11})_k)$
$N/2-1$	-1.515e-17	-1.960e-19	3.949e-18	-3.291e-17
$N/2-2$	-1.585e-18	-6.009e-18	-1.085e-17	-7.366e-17
$N/2-3$	-3.027e-18	7.945e-18	-9.406e-18	-5.628e-17
$N/2-4$	-1.381e-18	4.249e-18	3.243e-17	-3.248e-17

Table 11: Errors in the discrete Fourier coefficients of jump-functions, calculated by FFT with $N = 256$.

	$N = 32$		$N = 64$		$N = 128$		$N = 256$	
Q	Largest	Smallest	Largest	Smallest	Largest	Smallest	Largest	Smallest
3	2.78e-03	1.32e-07	6.72e-04	1.88e-09	1.67e-04	2.86e-11	4.15e-05	4.45e-13
4	3.31e-03	2.39e-09	7.81e-04	7.69e-12	1.93e-04	2.86e-14	4.80e-05	1.10e-16
5	3.86e-03	6.06e-11	8.81e-04	4.17e-14	2.16e-04	3.73e-17	5.37e-05	3.56e-20
6	4.45e-03	2.06e-12	9.76e-04	2.86e-16	2.37e-04	6.03e-20	5.88e-05	1.42e-23
7	5.14e-03	9.21e-14	1.07e-03	2.39e-18	2.57e-04	1.17e-22	6.36e-05	6.75e-27

Table 12: Largest and smallest numerically calculated eigenvalues of the jump-function coefficient matrix $\mathbf{A}$ for one discontinuity point ($M = 1$, $\gamma = 0$) when $K = Q$ equations are used to determine the jumps.

4 The jump-function coefficient matrix

The linear system of equations (12) can be written on the form $\mathbf{Ax} = \mathbf{b}$, where $\mathbf{A}$ is a $K \times (MQ)$-matrix consisting of jump-function coefficients $(\widetilde{V}_n)_k(\gamma_j)$, $\mathbf{x}$ is an MQ-vector of the jumps $\bar{A}_j^n$, and $\mathbf{b}$ is a K-vector of the discrete Fourier coefficients $\widetilde{w}_k^0$.

When this system is solved, it is important to take into consideration the fact that the elements of $\mathbf{A}$ are the discrete Fourier coefficients of $C_p^{n-1}(0, 2\pi)$-functions, $n = 1, 2, \ldots, Q$, and consequently decay as $O(|k|^{-n-1})$ as $|k| \to \infty$. These coefficients become very small when k is large, even for moderate values of Q, as shown in tables 1–7.

We take a closer look at the matrix $\mathbf{A}$ in the special case where $M = 1$, i.e., there is only one discontinuity point, which is taken to be $\gamma = 0$. (The subscript to γ is dropped when $M = 1$.) By (3), $(\widetilde{V}_n)_k(0) = (\widetilde{U}_n)$. First, we use

$$k = N/2 - 1, N/2 - 2, \ldots, N/2 - Q \tag{18}$$

in (12), such that $\mathbf{A}$ becomes the following $Q \times Q$-matrix:

$$\mathbf{A} = \begin{pmatrix} (\widetilde{U}_1)_{N/2-1} & (\widetilde{U}_2)_{N/2-1} & \cdots & (\widetilde{U}_Q)_{N/2-1} \\ (\widetilde{U}_1)_{N/2-2} & (\widetilde{U}_2)_{N/2-2} & \cdots & (\widetilde{U}_Q)_{N/2-2} \\ \vdots & \vdots & & \vdots \\ (\widetilde{U}_1)_{N/2-Q} & (\widetilde{U}_2)_{N/2-Q} & \cdots & (\widetilde{U}_Q)_{N/2-Q} \end{pmatrix}. \tag{19}$$

It is interesting to study the eigenvalues of $\mathbf{A}$, because when the span between the largest and smallest eigenvalue becomes large, the matrix is said to be ill-conditioned and the solution of

	$N = 32$		$N = 64$		$N = 128$		$N = 256$	
Q	Q eqs.	$2Q$ eqs.	Q eqs.	$2Q$ eqs.	Q eqs.	$2Q$ eqs.	Q eqs.	$2Q$ eqs.
2	2.03e+02	6.43e+01	8.25e+02	2.61e+02	3.32e+03	1.05e+03	1.33e+04	4.20e+03
3	2.11e+04	3.02e+03	3.58e+05	5.12e+04	5.82e+06	8.31e+05	9.34e+07	1.33e+07
4	1.39e+06	7.02e+04	1.02e+08	5.13e+06	6.74e+09	3.39e+08	4.35e+11	2.19e+10
5	6.37e+07	1.37e+06	2.11e+10	4.62e+08	5.79e+12	1.26e+11	1.51e+15	3.29e+13
6	2.16e+09	1.66e+07	3.41e+12	2.72e+10	3.93e+15	3.12e+13	4.14e+18	3.29e+16
7	5.58e+10	1.70e+08	4.47e+14	1.50e+12	2.20e+18	7.37e+15	9.43e+21	3.16e+19
8	1.12e+12	1.15e+09	4.86e+16	6.02e+13	1.04e+21	1.29e+18	1.83e+25	2.26e+22
9	1.79e+13	6.29e+09	4.47e+18	2.27e+15	4.28e+23	2.20e+20	3.09e+28	1.59e+25
10	2.28e+14	2.29e+10	3.51e+20	6.61e+16	1.54e+26	2.96e+22	4.61e+31	8.84e+27

Table 13: Numerically calculated condition numbers of the jump-function coefficient matrix **A** for one discontinuity point ($M = 1$, $\gamma = 0$) when $K = Q$ and $K = 2Q$ equations are used to determine the jumps.

the system of equations is expected to be unreliable. The effect of this will be demonstrated in section 5.

Eigenvalues of $\mathbf{A}$ for different values of N and Q have been calculated by the routine ZGELSS from LAPACK [2] (using singular value decomposition of $\mathbf{A}$ [10]) and are given in table 12. The corresponding condition numbers in 2-norm [10] (the largest eigenvalue divided by the smallest) are shown in table 13. The smallest eigenvalues can not be expected to be accurately calculated if the span between the largest and smallest eigenvalue exceeds 16 digits. In these cases the numbers given in tables 12 and 13 just indicate that $\mathbf{A}$ is ill-conditioned.

The conditioning of the matrix can be improved by increasing K, the number of equations in (12), a natural choice being to introduce $-k$ corresponding to each k in (18). As $\mathbf{A}$ then becomes a $2Q \times Q$-matrix, the condition number is defined as the largest singular value divided by the smallest [10]. The decrease in condition numbers can be seen from table 13. If this is not sufficient, a possibility is to include smaller values of $|k|$, but this may reduce the accuracy, because the method is based on the asymptotic assumption that the coefficients $\widehat{w}_k^Q$ of the "Q-smooth" part are smaller than the jump-function coefficients. The computational work will of course increase when the number of equations in (12) increases.

If $\mathbf{A}$ is ill-conditioned, it may be necessary to introduce the "numerical rank", i.e. the effective rank of the matrix in a numerical calculation. This will be further discussed in section 5.

5 Approximation of a function with one discontinuity point

We apply the approximation method described in the previous sections to the function

$$u(x) = 1 - \cos 3x/4, \qquad 0 \leq x \leq 2\pi. \tag{20}$$

The 2π-periodic extension of $u(x)$ is discontinuous and has discontinuous derivatives at $x = m2\pi$, $m = 0, \pm 1, \pm 2, \ldots$. As explained in section 2, this is handled by introducing a discontinuity point at $x = 0$.

When the Q values (18) are chosen for k in the system of equations (12) for the jumps A^n, $n = 1, 2, \ldots, Q$, we get the matrix $\mathbf{A}$ given by (19) in the previous section. The discrete Fourier coefficients that are the elements of $\mathbf{A}$ are calculated by the analytic formula (17a). The first derivative of $u(x)$ is calculated by the method in section 2, and the maximum error is plotted as a function of Q for different N in figure 1.

The conditioning of the matrix can be improved as described in the previous section, by using the $2Q$ equations

$$k = \pm N/2 - 1, \pm N/2 - 2, \ldots, \pm N/2 - Q \tag{21}$$

in (12) and solving the system by the least squares method. Approximation results from these calculations are shown in figure 2.

Comparing the results from figures 1 and 2 with the condition numbers in table 13, we find that in most cases the best results are obtained when the condition number of the matrix $\mathbf{A}$ is smaller than 10^{12}. For condition numbers larger than 10^{13}, no further accuracy of the

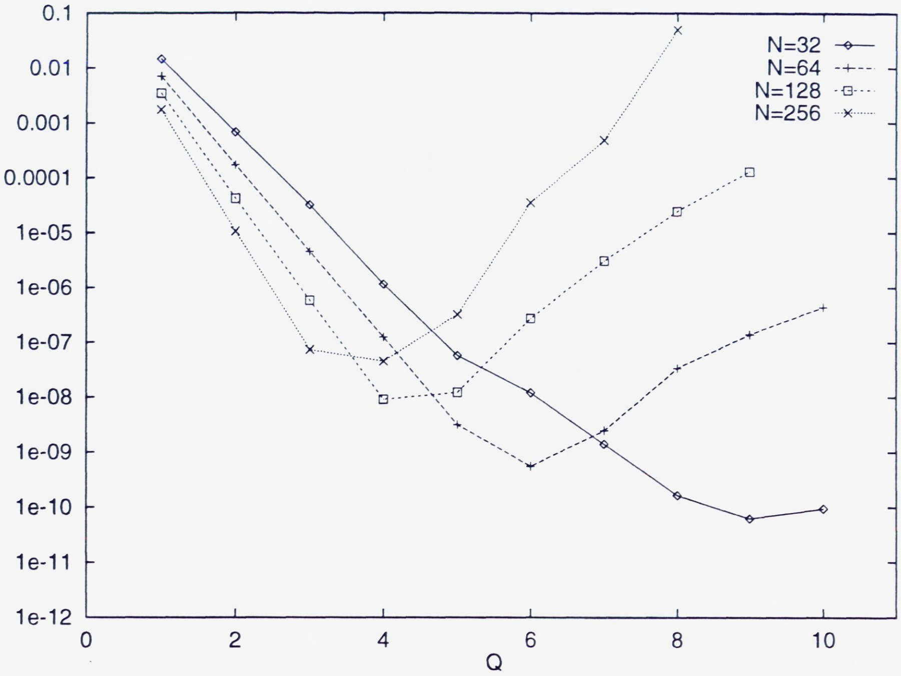

Figure 1: Max-error in the first derivative, $K = Q$ equations, no condition limit.

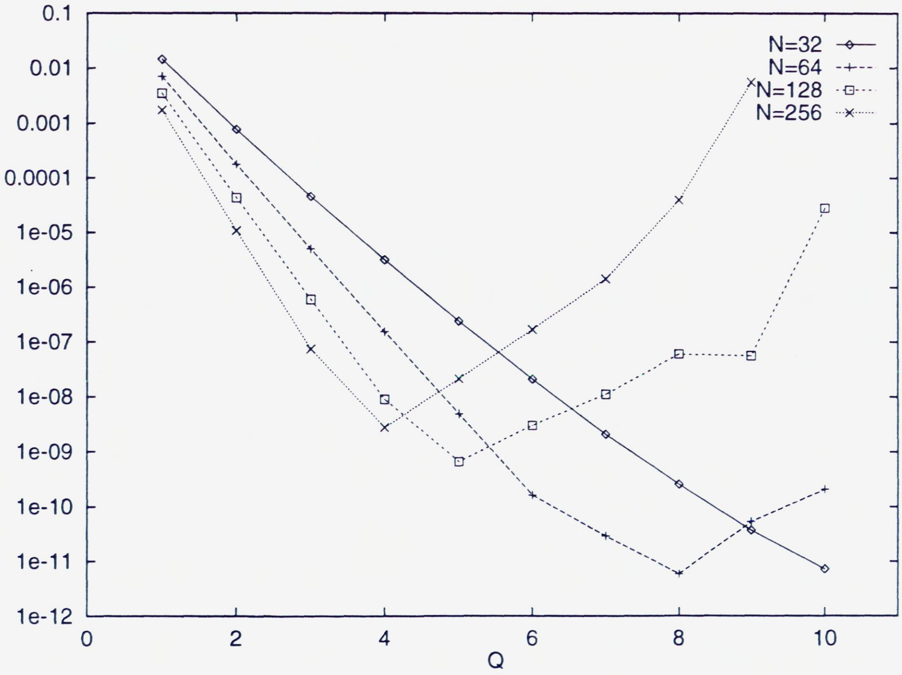

Figure 2: Max-error in the first derivative, $K = 2Q$ equations, no condition limit.

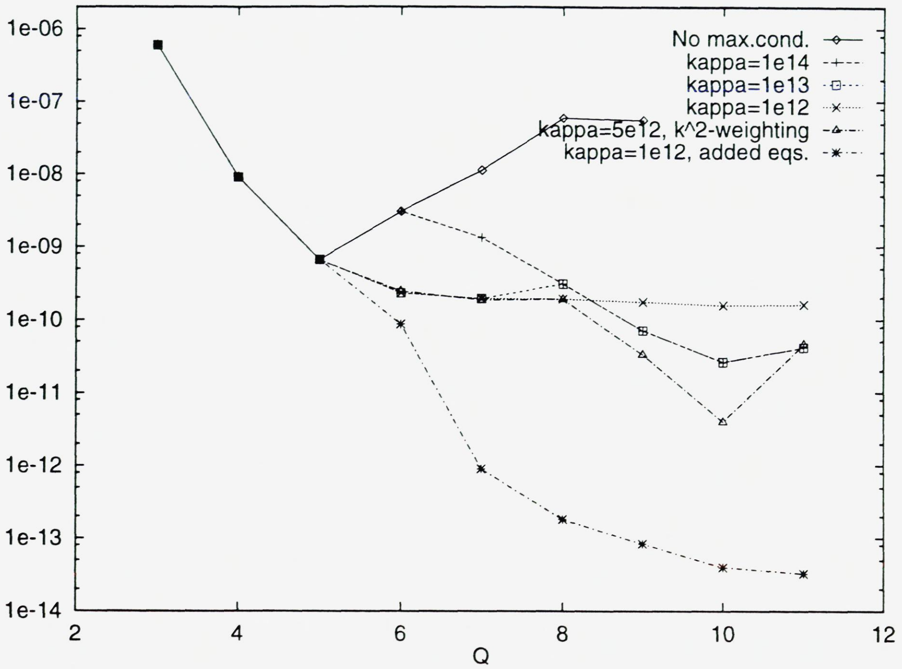

Figure 3: Max-error in the first derivative, $N = 128$, different condition limits and weighted least squares.

solution is obtained, and for the most ill-conditioned matrices the solution becomes highly unreliable.

These observations regarding condition numbers suggest the introduction of "numerical rank". The maximum allowed condition number is set to κ, and if some eigenvalues or singular values are more than a factor 10^{12} smaller than the largest eigenvalue or singular value, they are regarded as zero, and the system is considered rank-deficient. The term "rank" will be used in the following meaning "numerical column rank".

We shall now look more closely at the effects of numerical rank and condition number for the case $N = 128$. Approximation results with $2Q$ equations and different limits on the condition number are shown in figure 3. When the system becomes rank-deficient, the minimum norm least squares solution [10] is used. (See below.) We observe that if κ is too large, the solution may become less reliable, as seen in figure 3 for $\kappa = 10^{14}$. On the other hand, a too small value of κ may not allow us to benefit from the full potential of the equations. A maximum condition number of $\kappa = 5 \cdot 10^{12}$ gives the optimal solution for all Q shown in figure 3. To be on the conservative side, $\kappa = 10^{12}$ could be chosen.

Some comments on the minimum norm least squares solution are appropriate. In the rank-deficient case, the rank r of the matrix $\mathbf{A}$ is smaller than the number of unknowns n, and the solution of the least squares problem is no longer unique. A unique solution can be found by introducing the additional constraint of minimal norm of $\mathbf{x}$ [2, 10]. However, for our system of equations it is more natural to require that the jumps in the $n - r$ highest derivatives are set to zero. Without going into details on singular value decompositions and the solution of least squares problems (see [10]), we just note that the minimum norm solution of a rank-deficient

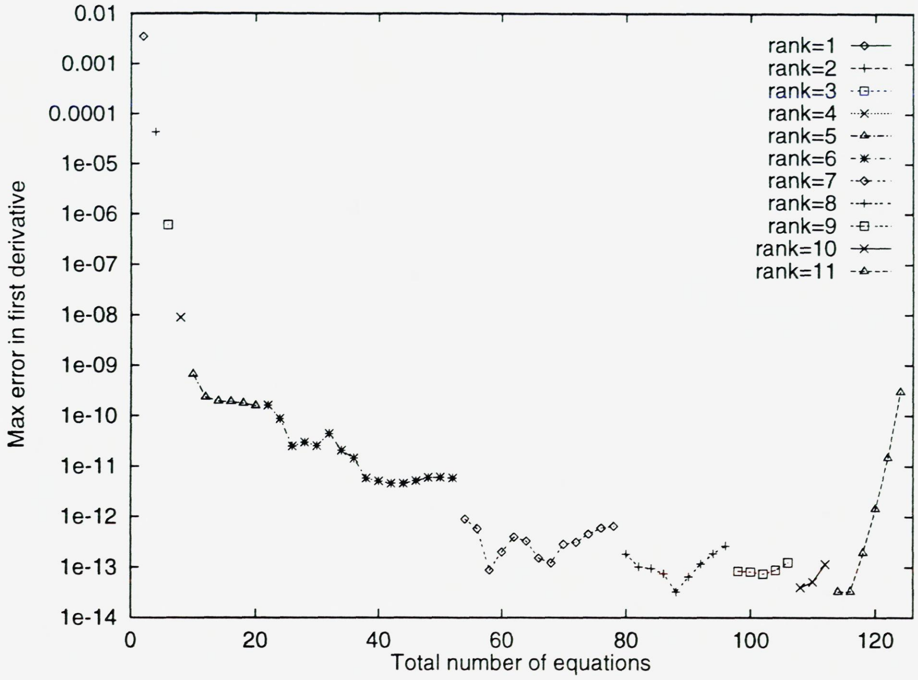

Figure 4: Max-error in the first derivative, $N = 128$, full rank systems, extra equations.

least squares problem does not produce zero for these $n - r$ jumps. The best approach is probably to reduce the number of unknowns to r, and solve a full rank problem. However, the mentioned jumps will usually become very small, and the difference would not be visible in figure 3.

Weighting of the equations in favour of large $|k|$ is suggested for similar problems in [5] and can also be applied here. The weighted results shown in figure 3 were produced by multiplying each equation in (12) by k^2. (Multiplication by $|k|$ had very little effect.) As the results only differ from the unweighted ones for some of the rank-deficient systems, weighting is not elaborated further in this paper.

Another possibility mentioned in the previous section is to use more than $2Q$ equations. More equations improve the conditioning of $\mathbf{A}$, but may at the same time decrease the applicability of the asymptotic assumption the method is based on, and is more time-consuming. Figure 3 includes results ("added eqs.") obtained by adding enough equations for the system to have full rank when the maximum condition number was set to 10^{12}.

The number of equations needed to obtain the different ranks are shown in figure 4. The values of k were

$$k = \pm N/2 - 1, \pm N/2 - 2, \ldots, \pm N/2 - K/2, \tag{22}$$

a total of K equations. Up to $Q = 5$, $K = 2Q$ equations are sufficient to give full rank (as already seen from figure 2), but 22 equations are needed to obtain a rank of 6, and 54 equations for a rank 7 system. For the given function (20), there is no particular structure apart from the jumps at $x = 0$, so the accuracy of the approximation is not decreasing before the very smallest $|k|$ are used in (12), but in general the number of additional equations used would have to be more restricted.

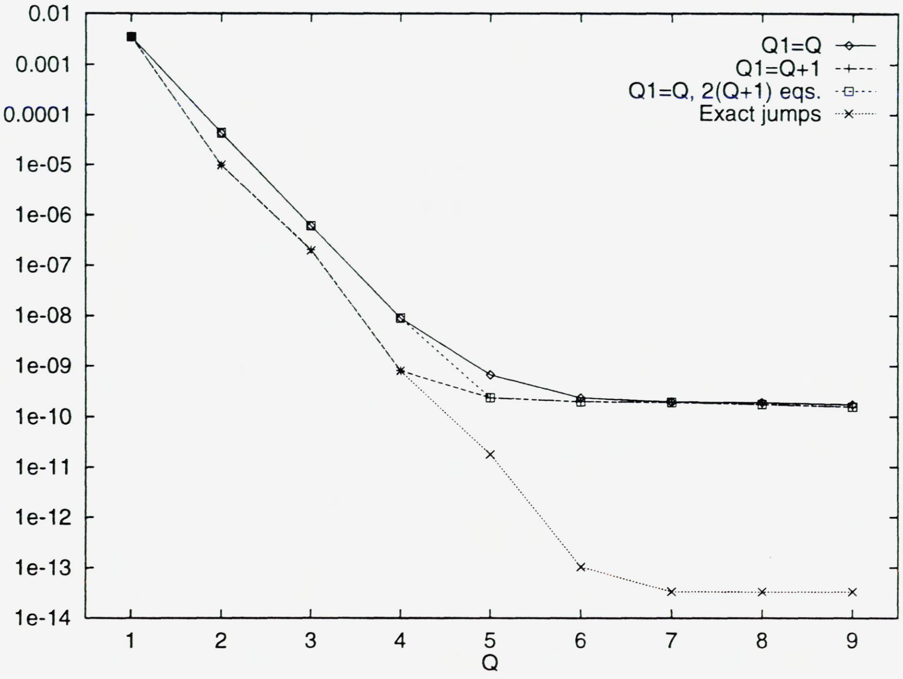

Figure 5: Max-error in the first derivative, $N = 128$, the effect of $Q_1 > Q$.

It should be remembered that because the coefficients of the jump-functions $U_n(x)$ decrease rapidly for even moderate n, it is necessary to include equations with low $|k|$ in (12) in order to obtain the necessary rank to get any advantage of these jump-functions.

Another observation from figure 4 is that the error decreases much more slowly and less systematic when extra equations are added. The method is clearly most efficient when the systems are well-conditioned, and this indicates that higher numerical precision would immediately give much better results.

We have also investigated the effect of replacing Q with Q_1 ($> Q$) in the solution of (12), while the original Q is still used in the representation (5) (i.e., only the the jumps in the Q first derivatives are used). As expected from the error estimates referred in section 2, the calculated jumps are closer to the exact ones when $Q_1 > Q$, but the total accuracy of the calculated derivatives versus the amount of computational work must be considered.

Results from using $Q_1 > Q$ are shown in figure 5, and for comparison the approximation error when the jump magnitudes are exactly known are also plotted. We see that already with $Q_1 = Q + 1$, optimal accuracy is achieved for $Q \leq 4$. For larger Q, the systems become rank-deficient and the improvements are smaller. Larger values of Q_1 do not give significant improvements.

Even though $Q_1 = Q + 1$ improves the results in most cases, a larger least squares problem is solved, so the actual computational work is closer to the work for a "$(Q + 1)$-problem" than for a "Q-problem". Considering that the improvement in going from $Q_1 = Q$ to $Q_1 = Q + 1$ is in most cases smaller than the improvement of increasing Q by 1, it may not be favourable after all, at least not from a pure approximation point of view. We also observe that when

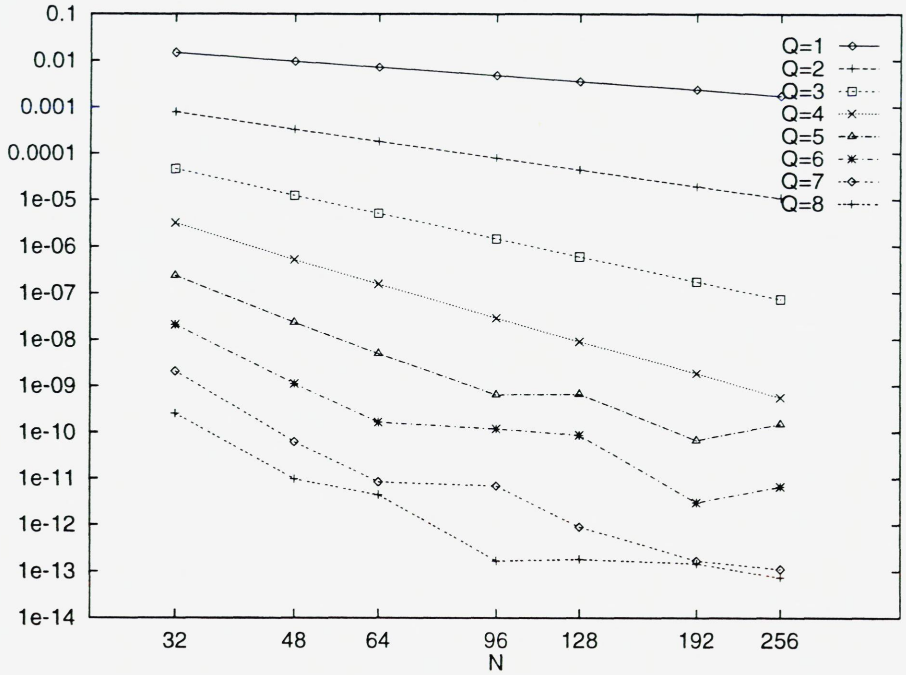

Figure 6: Max-error in the first derivative, illustration of convergence for different Q.

Q	1	2	3	4	5	6
Interval in N	32–256	32–256	32–256	32–256	32–96	32–64
Convergence order	1.0	2.1	3.1	4.2	5.4	7.0

Table 14: Convergence orders for the max error of the first derivative

the system no longer has full rank (for $Q_1 \geq 5$), it is the extra equations in (12) that give the improved accuracy, and not the increased number of unknowns. This is seen from the results where $Q_1 = Q$ and $2(Q+1)$ equations are used.

Examples of the pointwise distribution of errors for calculations with $Q_1 > Q$ are given in section 6. Where nothing else is noted, the same value of Q are used in (12) and (5) in the calculations in the rest of this paper.

In figure 6, we show the approximation results for different N and Q to check the convergence rates. The results are calculated with 10^{12} as maximum condition number of $\mathbf{A}$. Basically, $2Q$ equations are used in (12), but sufficient additional equations are included when necessary to achieve full numerical rank. As shown in figure 4, the results can be improved by adding more equations in (12) than what is necessary to obtain full numerical rank, but a reliable criterion for the optimal number of equations have not been found. It seems to be a good compromise between accuracy, efficiency, and simplicity to use the smallest full rank system.

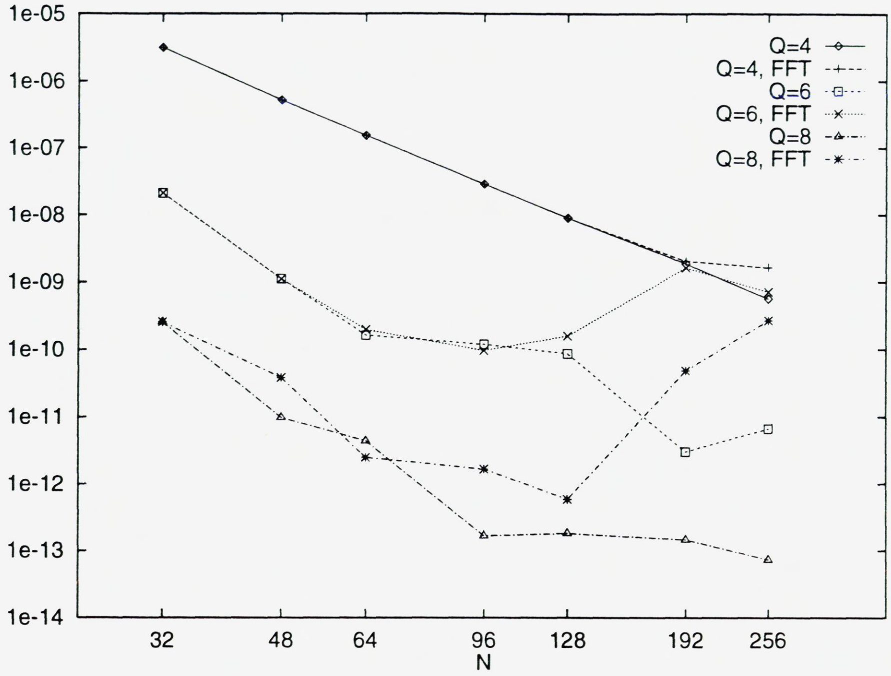

Figure 7: Max-error in the first derivative, FFT vs. analytical calculation of jump-function coefficients.

It it seen from figure 6 that the error for a given Q decays almost linearly in the double-logarithmic plot, as long as Q and N are not too large. This means that the error is proportional to N^{-p}, where the convergence orders p are given in table 14, together with the interval in N they are calculated from. We observe that the orders of convergence are well in accordance with the theoretical estimate of $O(N^{-Q})$ for the error in the first derivative [7].

Figure 6 also confirm the comments made in connection with figure 4, that the method is most efficient when the least squares systems are well-conditioned. Even though the accuracy can be improved for ill-conditioned systems by increasing the number of equations, the convergence is no longer algebraic.

Finally, the effect of using FFT to calculate the jump-function coefficients is shown in figure 7, where the results with analytical coefficients from figure 6 are included for $Q = 4, 6, 8$, for reference. The reduced accuracy from the use of FFT is first seen for $N = 256$ and $Q = 4$ (no differences for smaller Q), and for large Q and N the results are much worse than when the analytic formula (17a) is used. The differences appear for the same combinations of N and Q that give ill-conditioned systems, so additional equations seem to be necessary to take advantage of the improved accuracy of the analytical expressions (17) for the jump-function coefficients.

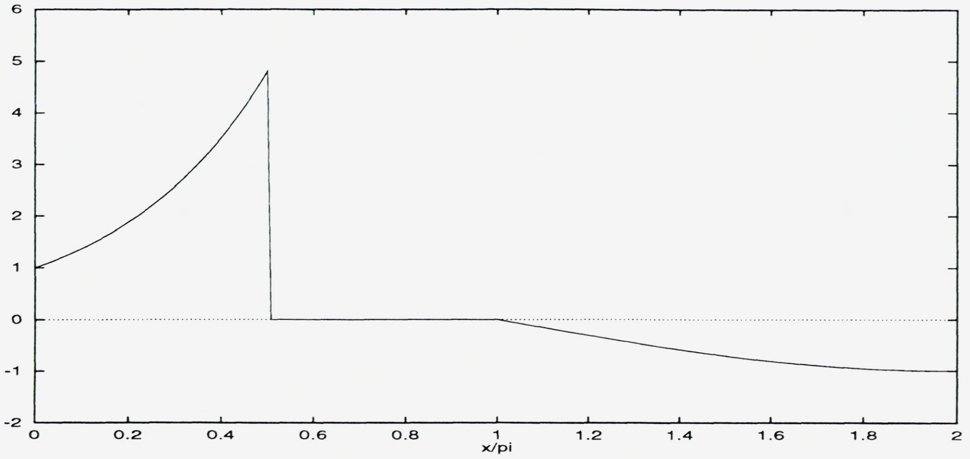

Figure 8: The function (23) with three discontinuity points studied in section 6.

6 Approximation of a function with three discontinuity points

In this section we study a more complicated case, the function defined by (23) and shown in figure (8). It has jumps in the function value and/or the derivatives at $\gamma_1 = 0$, $\gamma_2 = \pi/2$, and $\gamma_3 = \pi$.

$$u(x) = \begin{cases} e^x, & 0 \leq x < \pi/2, \\ 0, & \pi/2 < x < \pi, \\ \cos(x/2), & \pi \leq x \leq 2\pi. \end{cases} \tag{23}$$

The jumps in $u(x)$ are subtracted as in (7), and the coefficients $(\widetilde{V}_n)_k(\gamma_j)$ are calculated by (17) in the following calculations. $K = 2Q$ equations are used in (12) to find approximate jumps, but the maximum condition number is set to 10^{12} and additional equations are added when necessary to obtain full rank systems.

Figures 9–12 show the error in the first and second derivatives at the grid points for different values of the reconstruction degree Q and the number of grid points N. We observe that the dominating errors are located in a relatively small neighbourhood of the discontinuity points, with errors being several orders of magnitude smaller away from these points. This behaviour has been observed also for other methods of approximation of non-smooth functions, see [12] for a similar approximation problem.

When the number of points with dominating errors is $O(1)$, as is clearly the situation here, the RMS-error will be a factor $N^{1/2}$ smaller than the maximum error, so the theoretical results from [7, 8] give $O(N^{-(Q+3/2-m)})$ for the RMS-error of the mth derivative.

The RMS-errors in the first and second derivative for different N and Q are displayed in figures 13 and 14. The results for each Q are approximately straight lines when N is large enough to avoid the equations with too small $|k|$ to be included in (12), and small enough for

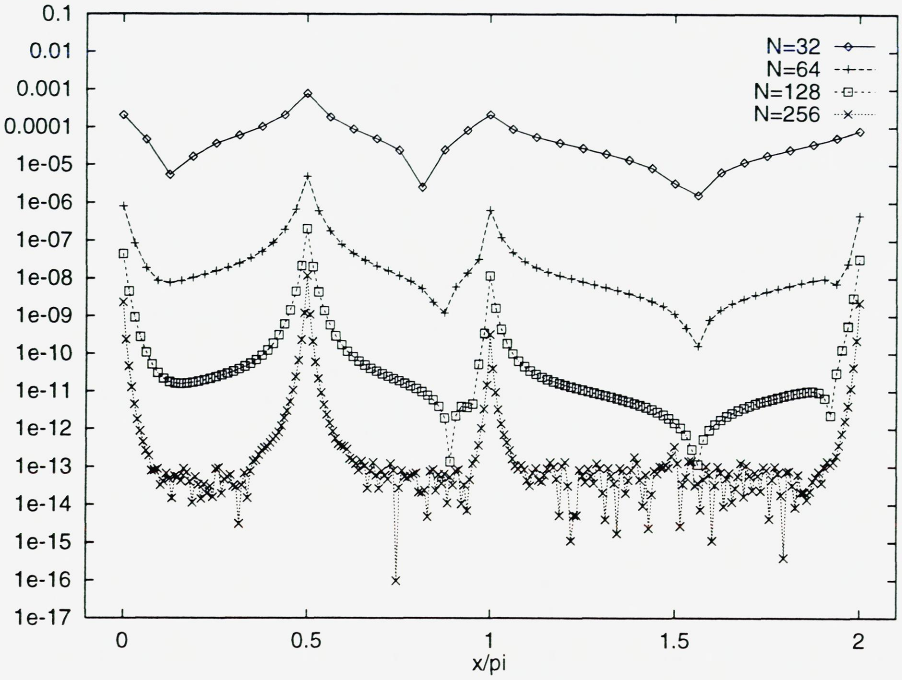

Figure 9: Pointwise error in the first derivative of (23), $Q = 4$.

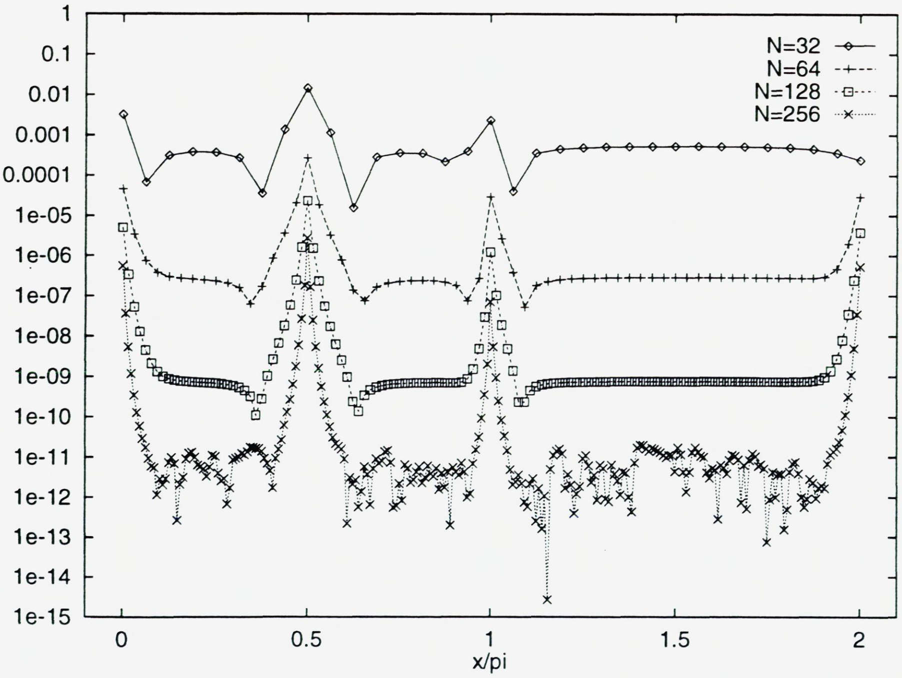

Figure 10: Pointwise error in the second derivative of (23), $Q = 4$.

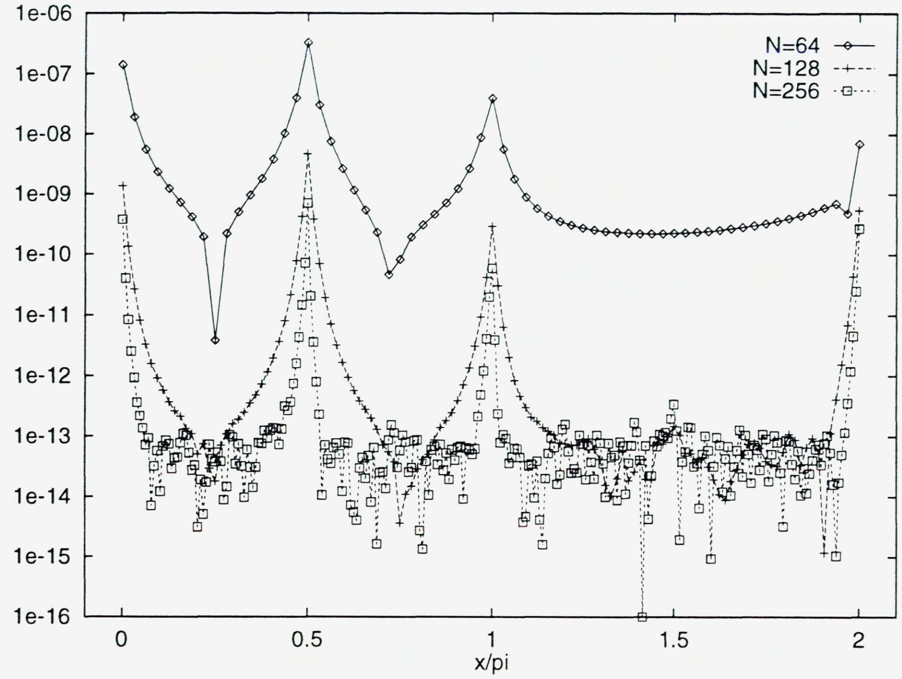

Figure 11: Pointwise error in the first derivative of (23), $Q = 5$.

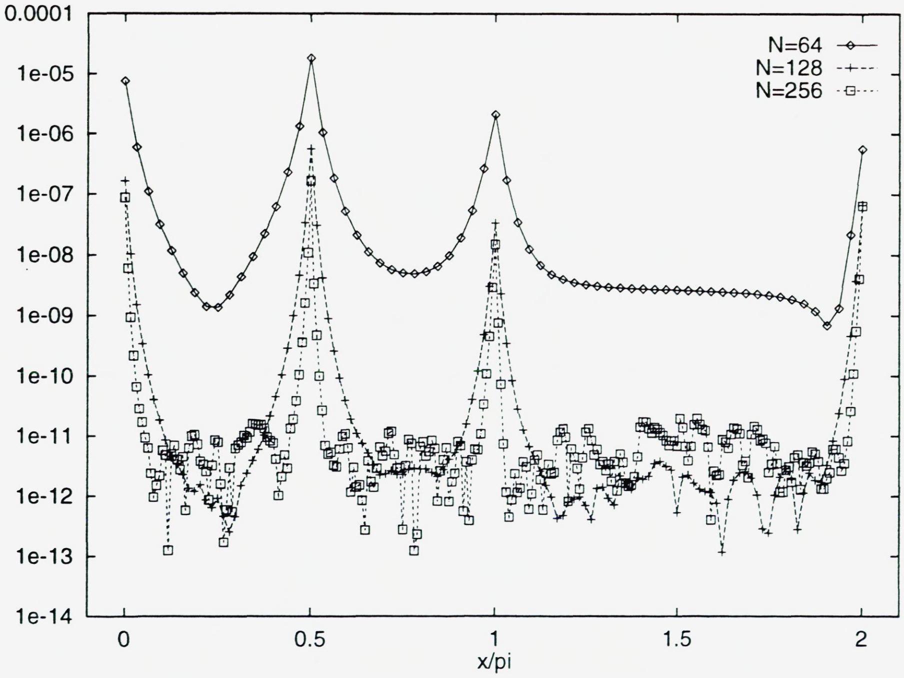

Figure 12: Pointwise error in the second derivative of (23), $Q = 5$.

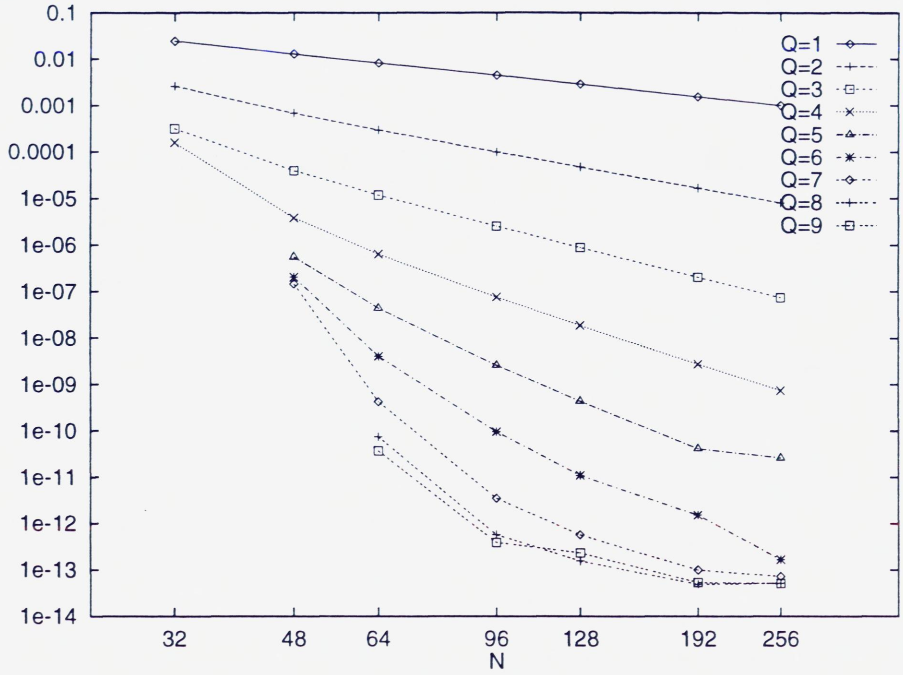

Figure 13: RMS-error in the first derivative for different values of Q and N.

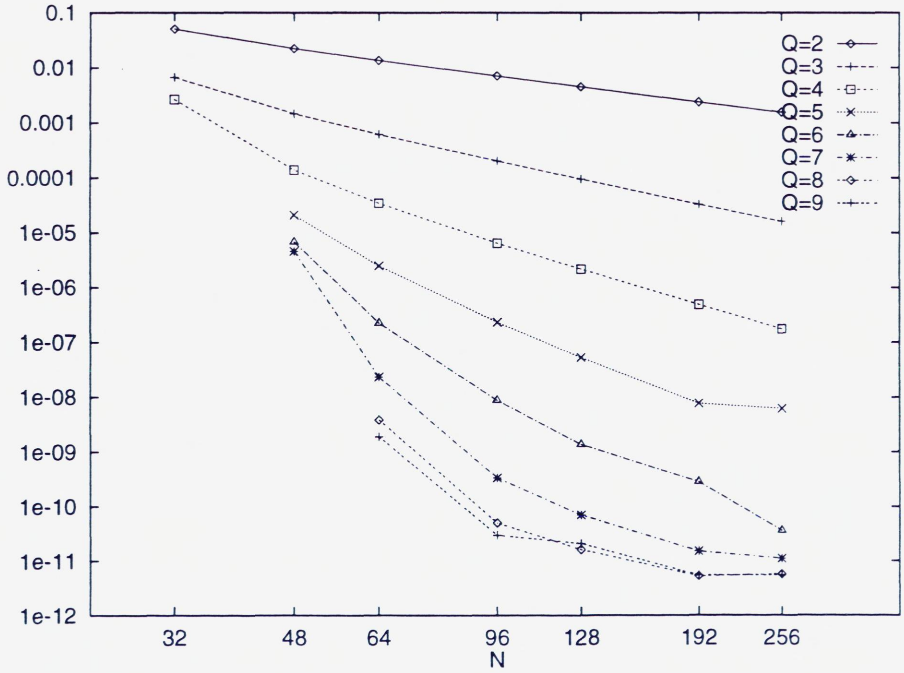

Figure 14: RMS-error in the second derivative for different values of Q and N.

Q	1	2	3	4	5
Interval in N	32–256	64–256	96–256	96–256	96–192
Conv. order, 1st der.	1.5	2.6	3.6	4.7	6.0
Conv. order, 2nd der.	—	1.6	2.6	3.7	4.9

Table 15: Convergence order for the RMS-error of derivatives.

the system to have full rank with only $2Q$ equations.

The convergence rates are calculated for the intervals where the errors exhibit almost algebraic convergence, i.e., straight lines in figures 13 and 14, and are given in table 15. As in the previous section, the results agree with the theoretical estimates.

Finally, the results obtained here are compared with what we get if the exact jumps and the exact Fourier coefficients of the smooth part of the function are known. That would be the optimal representation on the form (5) with given N and Q. In practice we have here used the known jumps of $u(x)$ and its derivatives for the A_j^n's to calculate the Q times smooth part

$$u^Q(x) = u(x) - \sum_{n=0}^{Q} \sum_{j=1}^{M} A_j^n V_n(x;\gamma_j). \tag{24}$$

For $Q > 3$, approximations of the coefficients $\widehat{u}_k^Q$, $|k| < 64$, up to the 15th decimal place can be obtained using a discrete Fourier transform of length 1024.

The pointwise errors in the first derivative are shown for different N and Q in figure 15. In all cases the maximum error is much larger for the approximate jumps, but the distribution of errors throughout the domain is interesting, as it is more uniform when the exact jumps and coefficients are used. For small N and large Q, the approximation has larger error at all points, while for the opposite case (large N and smaller Q), the exact jumps and coefficients give a better solution only at a few grid points closest to the discontinuity points.

To illustrate this further, the absolute values of the real parts of the Fourier coefficients of the smooth part $w^Q(x)$ of the approximating function (5) are displayed in figure 16 for three pairs of N and Q. The exact coefficients decay very regularly, as they are asymptotically $O(|k|^{-(Q+2)})$ as $|k| \to \infty$. The approximate coefficients are much smaller for the highest $3Q$ frequencies, because these coefficients are given by (13) as the residuals from the least square problem for the jumps. The correspondence between the magnitude of the highest coefficients and the minimum pointwise error is seen by comparing figures 15 and 16.

For the case $N = 128$, $Q = 4$, results of calculations with $Q = Q_1 = 5$ in (12) are included in figures 15 and 16. Those results imply that the approximated jumps in the highest derivative (the fourth in this case) play a key part in the pointwise error distribution. As mentioned earlier, it is expected from the theory [8] that the approximate jumps in the highest derivative have the largest error. However, the inclusion of these jumps apparently acts as a "filter"

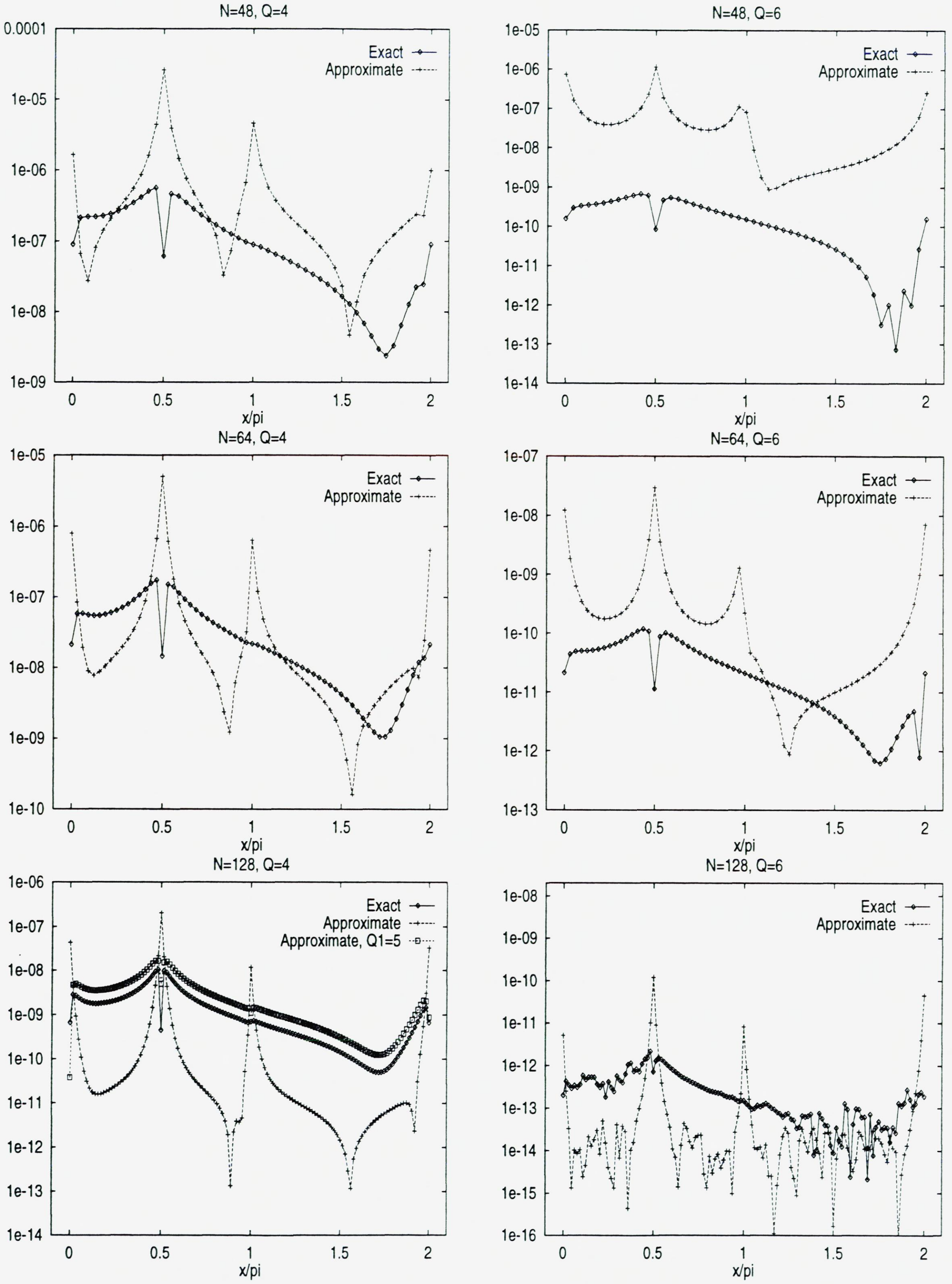

Figure 15: Pointwise errors in the first derivative for approximate and exact jumps and coefficients.

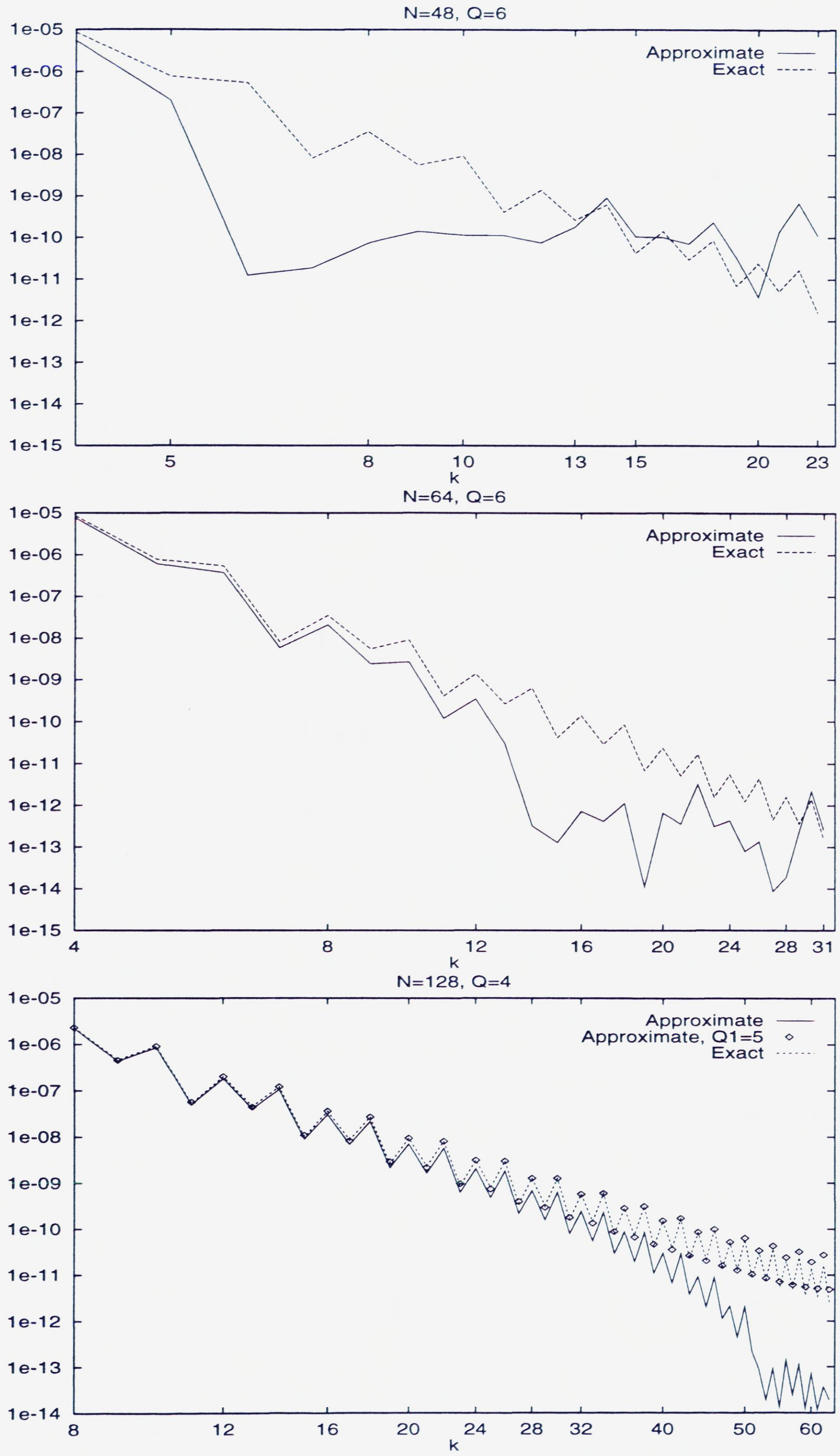

Figure 16: Absolute values of the real parts of the Fourier coefficients for the smooth part $w^Q(x)$ of the approximating function.

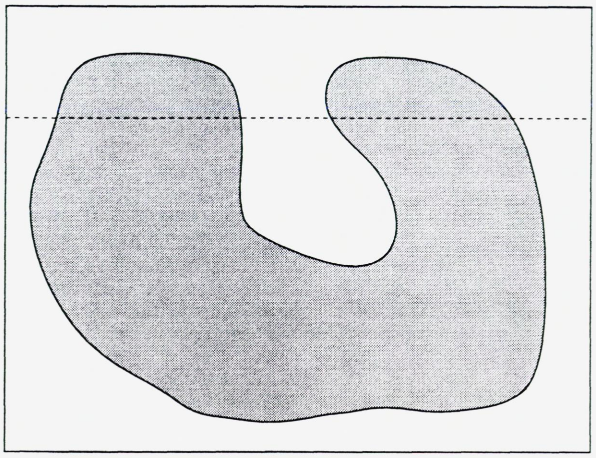

Figure 17: A domain with complex geometry embedded in a rectangular domain, with a horizontal grid line.

on the highest coefficients (as seen in figure 16) and reduces the error away from the jump locations strongly, whereas the approximation close to the discontinuity points suffers from the limited accuracy of the jumps in the highest derivative. When $Q_1 = 5$ (and the results change very little with larger Q_1) is used, the jumps in the first four derivatives become more accurate, but the "filtering" due to the matching of the calculated jumps and the highest coefficients disappear and the resulting approximation is very similar to the one with exact jumps. (The difference between them is mainly due to the calculation of coefficients of the smooth part, i.e., the collocation error, and not to the accuracy of the jumps.)

7 Width of an "exterior" interval

For applications to problems in complex two-dimensional geometries, the approach suggested in [8] is to embed the given domain in a rectangular domain. This is sketched in figure 17, where the gray area is the original domain.

The rectangular computational domain will in this case be divided into an "interior" part which is the original domain, and an "exterior" part added to make the total domain rectangular. The functions involved are taken to be zero in the exterior part of the domain.

A horizontal grid line like the one drawn in figure 17 contains both interior and exterior intervals, and the function (23) can be used as a simplified example of a function along such a grid line. By changing the definition from (23) to

$$u(x) = \begin{cases} e^x, & 0 \le x < \pi/2, \\ 0, & \pi/2 < x < \pi/2 + \delta\,\Delta x, \\ \cos(x/2), & \pi/2 + \delta\,\Delta x \le x \le 2\pi, \end{cases} \tag{25}$$

where $\Delta x = 2\pi/N$ is the grid spacing, we get a function which is zero at an interval of width

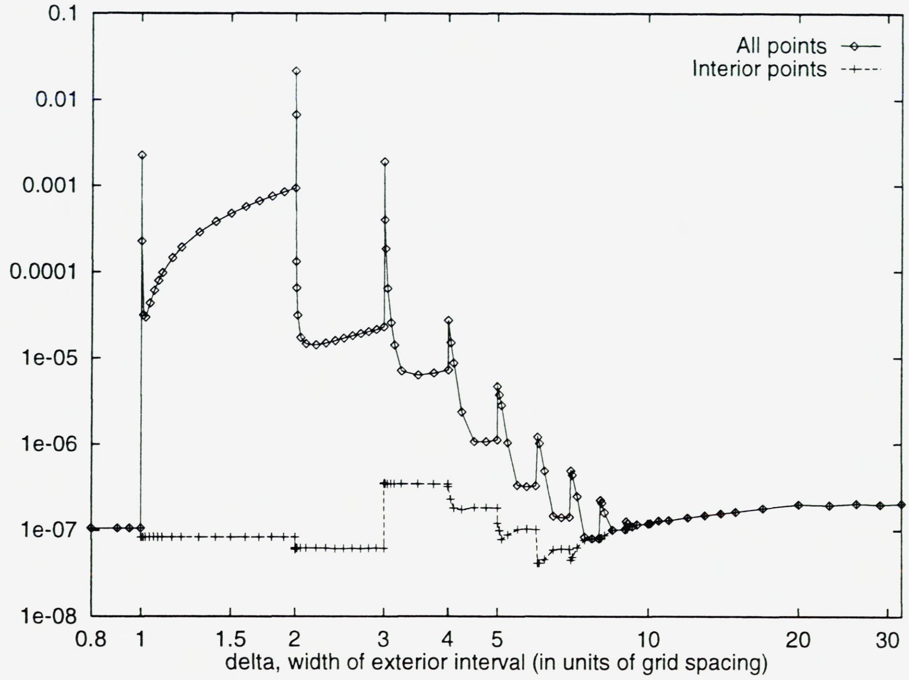

Figure 18: Max-error in the first derivative, N=128, Q=4, for varying width of the exterior interval.

$\delta\,\Delta x$. This interval is called the exterior interval in the following.

The effect of diminishing the exterior interval for $N = 128$ and $Q = 4$ is shown in figure 18. As seen from the previous section, this is a well-conditioned example when $\delta = 32$. When the exterior interval gets smaller, the condition number grows, and the rank of the system (12) becomes 3 (rank-deficient) at $\delta = 3$. For this problem the condition number is not reduced when more equations are added in (12). The rank decreases further at $\delta = 2$ (rank 2) and $\delta = 1$ (rank 1). These results for $N = 128$ and $Q = 4$ have been found to be typical, for better conditioned systems δ may be smaller before rank-deficiency occurs, and vice versa for worse conditioned systems.

Figure 18 shows the maximum error at the collocation points, therefore there are jumps in the error curves at integer values of δ. For δ just larger than an integer, there is a grid point at the right end of the exterior interval, and the error at this point is seen to become quite large for small δ. When δ is equal to or just smaller than an integer, the grid point closest to the discontinuity will be in the right interior domain and the error is smaller. If $0 < \delta \leq 1$, the error is not measured inside the exterior interval, and becomes constant.

In an application, it is not really interesting to see how good the zero solution in the exterior domains is represented, as long as errors does not spread to interior intervals. And we see from figure 18 that when the error is computed only from the interior points, the effect of varying δ is much smaller, which is an encouraging result concerning the methods ability to handle problems in complex geometries.

It is illustrated in figure 19 that the maximum error appears in the exterior domain. Here

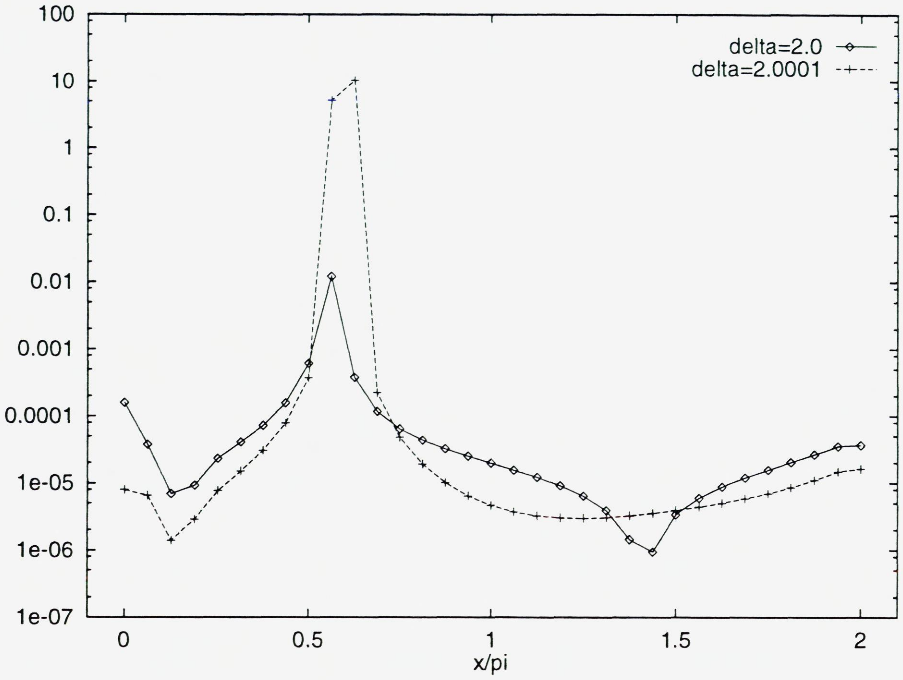

Figure 19: Pointwise error in the first derivative, N=32, Q=4, with 1 and 2 points in the exterior interval.

$N = 32$ is used to give a better view of the pointwise behaviour. When $\delta = 2$, only the one point with the largest error is located in the exterior interval. For $\delta > 2$ there are two exterior points with much larger errors, but at the interior points the errors are still small.

8 Robustness regarding functions without discontinuities

In this section we investigate the robustness of the approximation method for cases where the actual function do not have discontinuities in the relevant derivatives. Such special cases can occur e.g. in time-dependent problems, and must be handled with the same accuracy as seen previously.

The four functions we use as illustrations are shown in figure 20 and given by

$$u_1(x) = \begin{cases} \exp\left(-\frac{5.0(x/\pi-0.8)^2}{0.7^2-(x/\pi-0.8)^2}\right), & |x/\pi - 0.8| < 0.7, \\ 0, & \text{otherwise}, \end{cases} \tag{26a}$$

$$u_2(x) = \left((\xi - 2\pi)\xi\right)^6 / \pi^{12}, \qquad \xi = (2\pi + x - 0.3/\pi) \bmod 2\pi, \quad 0 \le x \le 2\pi, \tag{26b}$$

$$u_3(x) = 4\sin(\xi/2) - 2\xi + \frac{\xi^2}{2\pi}, \qquad \xi = (2\pi + x - 0.3/\pi) \bmod 2\pi, \quad 0 \le x \le 2\pi, \tag{26c}$$

$$u_4(x) = \frac{1}{2}(1 - \cos(3x/4)) - \frac{x}{4\pi}, \qquad 0 \le x \le 2\pi. \tag{26d}$$

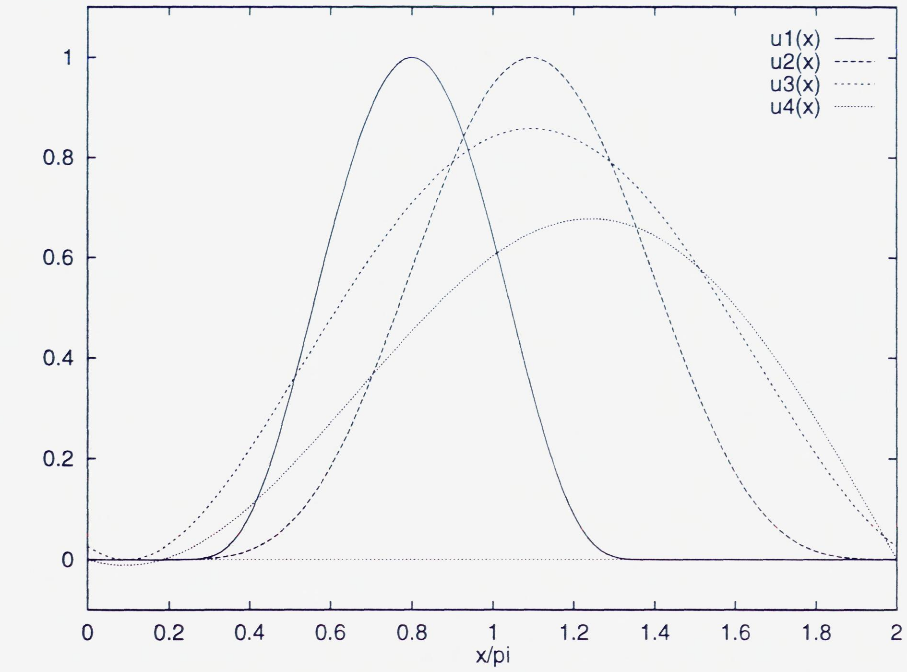

Figure 20: The four functions considered in section 8.

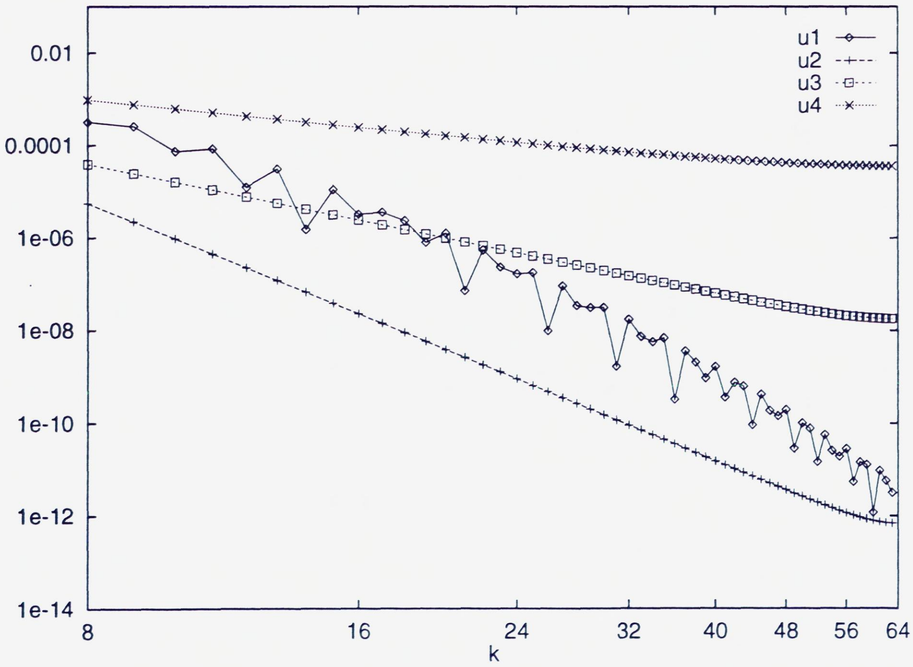

Figure 21: Real parts of the discrete Fourier coefficients (for $N = 128$) of the four functions considered in section 8.

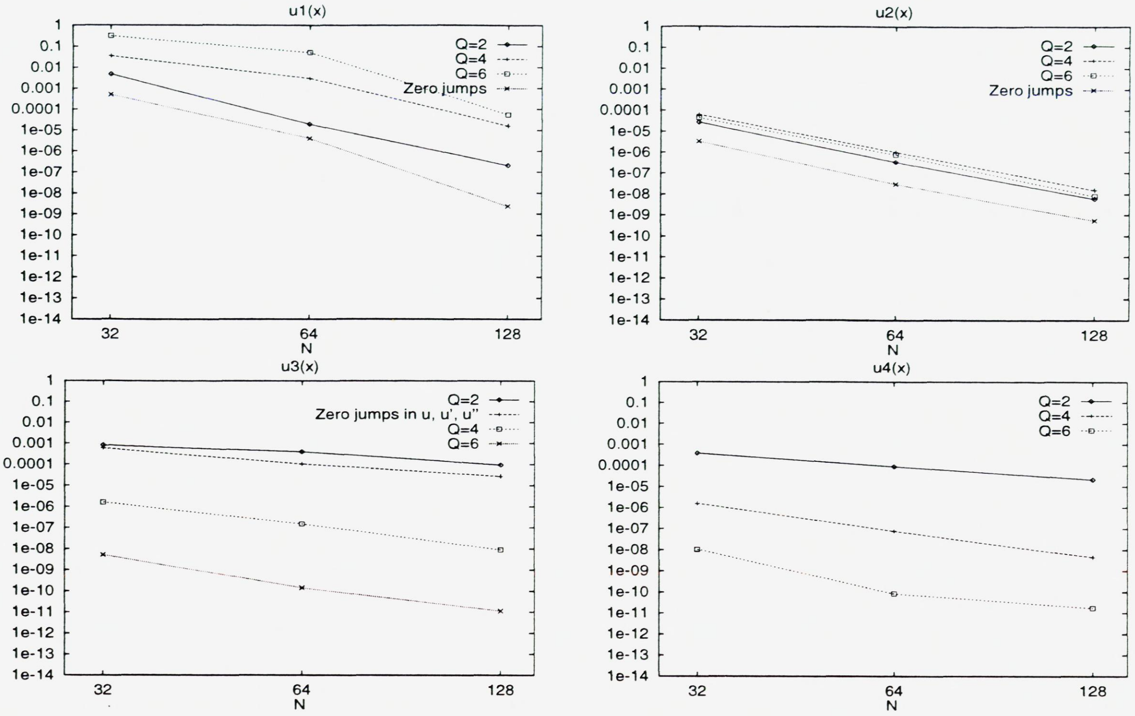

Figure 22: Maximum error in the first derivative from the approximation of the functions (26).

$u_1(x)$ is a smooth function, identically zero in the intervals $[0, 0.1\pi]$ and $[1.5\pi, 2\pi]$, and thus the 2π-periodic extension is also smooth. The 2π-periodic extension of $u_2(x)$ has discontinuities in the 7th, 9th, and 11th derivatives at $x = 0.3/\pi$, while the 2π-periodic extension of u_3 has discontinuities in the third and all the higher odd-numbered derivatives at $x = 0.3/\pi$. $u_4(x)$ is essentially the function we considered in section 5, but the discontinuity at $x = 0$ in the 2π-periodic extension has been removed by subtracting a linear function. The 2π-periodic extension of $u_4(x)$ has discontinuities in all derivatives at $x = 0$. $u_4(x)$ is included in this section to represent the "normal" case, for comparison. The real parts of the discrete Fourier coefficients of the functions (26a) for $N = 128$ are shown in figure 21.

Figure 22 shows how the different degrees of smoothness and periodicity influence the accuracy of the calculation of the first derivative. The curves marked "zero jumps" show the accuracy obtained if the relevant jumps are set to zero. For $u_2(x)$ and $u_3(x)$, the convergence orders are close to the expected orders of 6 and 2, respectively, even when Q is too small to catch the discontinuities. Still the approximations could be improved if it was detected that the functions did not have jumps in the Q first derivatives. This applies in particular for the function $u_1(x)$, and higher values of Q increase the errors drastically. Therefore we search for means to deal with these situations. Three general approaches are discussed below, before we consider a method for a single discontinuity point more in detail.

The first approach ("truncation") is to set a lower limit for the discrete Fourier coefficients used in the system (12), and regard them as zero if they are smaller than this limit, as they will be if the function is sufficiently smooth and the values of $|k|$ are large enough. If the truncation limit is related to the size of the highest coefficient of the jump-function of order

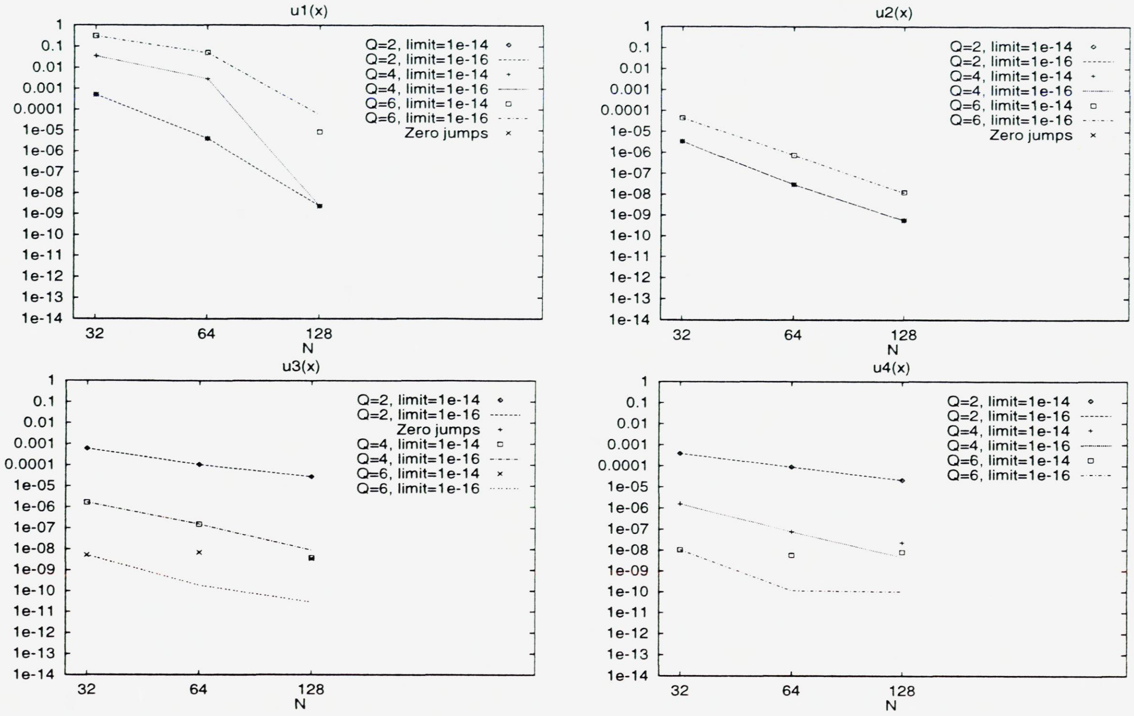

Figure 23: Maximum error in the first derivative from the approximation with truncation of the discrete Fourier coefficients of the functions (26).

Q, i.e., the absolute value of the coefficient $(\widetilde{U}_Q)_{N/2-1}$, the desired effect would be achieved when N is sufficiently large.

If the coefficients are larger than the limit just described, this truncation should not be applied because significant digits of the coefficients would be truncated. Instead, a truncation limit close to the machine epsilon (usually around 10^{-16} for double precision) could be introduced to remove noise generated by the FFT of the input function. This may improve the results for large N.

Results using the truncation strategy are shown in figure 23. It detects the zero jumps in many cases, but not all. Results are shown with the noise removal limit set to 10^{-14} and 10^{-16}, and it appears that this limit should not be larger than 10^{-16}. In fact, when the functions have jumps in some of the first Q derivatives, it is difficult to see that application of the "noise removal" improves the results at all (cf. figure 22 for approximation without any truncation).

Unfortunately, N would in many cases have to be much larger than seen previously in this paper for the truncation to have the desired effect. To illustrate this, consider figure 21, where the coefficients of the smooth, periodic function $u_1(x)$ decreases faster than those of the other functions when $|k|$ grows, but are not smaller in magnitude than the coefficients of $u_2(x)$ until N is greater than 128. So this truncation method only works in calculations with small Q and/or large N.

The second alternative is to consider the residual of the least squares solution of 12, in the case that the number of equations is larger than the number of unknown. A large residual indicates that the calculated jumps are inaccurate. This will happen when the system is solved for jumps that really are zero, but it could also be other reasons (e.g. too small N to be in the asymptotic area), and it is difficult to quantify how small the residual should be for an "accurate" solution.

For a third method, we note that (12) is a system of complex equations. The fact that we are only interested in real jumps as solutions (assuming that the functions to be approximated are real) means that the jumps of two real functions $u_1(x)$ and $u_2(x)$ are usually calculated at once by defining the new complex function $u(x) = u_1(x) + iu_2(x)$.

When pairs of positive frequencies k and $-k$ are used in (12), the real and imaginary parts are decoupled, such that the errors in the jumps for $u_1(x)$ and for $u_2(x)$ will not influence each other. However, if only positive frequencies are used, errors for both functions are distributed in both the real and imaginary parts of the solution. In that case, solving the system for a single real function would produce imaginary parts of the calculated jumps that could indicate the size of the errors in the real parts. Jumps with large relative errors can then be set to zero, assuming that the error is large because there is no jump in the given derivative at the given discontinuity point.

There are some disadvantages of this method too, one is that the system of equations for the jumps becomes more ill-conditioned. The conditioning may be restored by using additional equations with lower $|k|$, but this can decrease the accuracy. Another disadvantage is that jumps for pairs of real functions can no longer be calculated in one complex calculation, with the result that twice as many right-hand sides may be required. Also, as was the case with the residual test mentioned above, it is difficult to prescribe a good quantitative criterion for how small the imaginary part should be before the jump is accepted. But in contrast to the residual test, this approach gives an indication of the accuracy of each jump, not only for the total vector of jumps.

It appears that there is a need for a criterion to decide whether or not the function has jumps in any of its Q first derivatives, for given $Q > 1$. (Still assuming that jumps in the function itself have been subtracted.) We shall discuss another approach for the case of a single jump location, which is closely related to the algorithms given in [5] and [6] for locating discontinuities. The simplest form of those algorithms utilizes the fact that if a 2π-periodic function $u(x)$ is continuous except for a jump discontinuity of magnitude A at $x = \gamma$, its Fourier coefficients asymptotically satisfy

$$\widehat{u}_k = A(\widehat{V}_0)_k(\gamma) + O(|k|^{-2}) = \frac{Ae^{-ik\gamma}}{2\pi ik} + O(|k|^{-2}), \tag{27}$$

as $|k| \to \infty$. Therefore, if we define the complex number z_k as the ratio

$$z_k = \frac{\widehat{u}_k/(\widehat{V}_0)_k(0)}{\widehat{u}_{k-1}/(\widehat{V}_0)_{k-1}(0)} = \frac{2\pi ik\widehat{u}_k}{2\pi i(k-1)\widehat{u}_{k-1}}, \tag{28}$$

we obtain

$$z_k = e^{-i\gamma} + O(|k|^{-1}), \quad \text{as } |k| \to \infty. \tag{29}$$

Thus the location of the discontinuity is given by

$$\gamma = -\arg(z_k) + O(|k|^{-1}), \quad \text{as } |k| \to \infty. \tag{30}$$

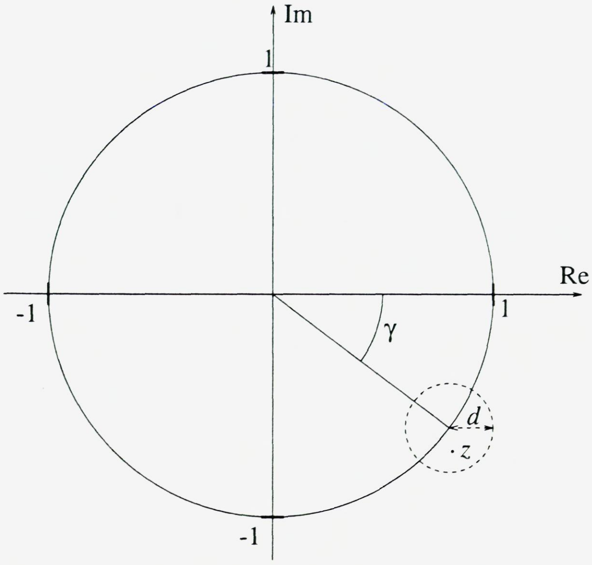

Figure 24: Approximation of discontinuity position using the complex number z given by (28).

(It is shown in [6] that if $u(x)$ is continuously differentiable for all $x \neq \gamma$, the last term in (29) and (30) will actually be $O(|k|^{-2})$.)

Figure 24 shows the circle of radius d around $e^{-i\gamma}$ where the calculated z_k is expected to be. The radius d is at most $O(|k|^{-1})$ as $|k| \to \infty$.

In [5] and [6], an approximate discontinuity position was calculated as $\tilde{\gamma} = -\arg(|z_k|/z_k)$, but in our context the position is assumed to be known. Instead, we shall describe how z_k can be used to check whether there is a jump at a given position.

Because it is also assumed in this paper that the jump in the function is known and subtracted using (7), we rewrite (27), (28) for use with a continuous 2π-periodic function with $n-1$ continuous derivatives and a discontinuity of magnitude A^n in the nth derivative at $x=\gamma$ for $n \geq 1$. The asymptotic behaviour of the Fourier coefficients becomes

$$\widehat{u}_k = A^n (\widehat{V}_n)_k(\gamma) + O(|k|^{-(n+2)}) = \frac{A^n e^{-ik\gamma}}{2\pi(ik)^{n+1}} + O(|k|^{-(n+2)}), \quad \text{as } |k| \to \infty, \tag{31}$$

and we define z_k^n by

$$z_k^n = \frac{\widehat{u}_k/(\widehat{V}_n)_k(0)}{\widehat{u}_{k-1}/(\widehat{V}_n)_{k-1}(0)} = \frac{2\pi(ik)^{n+1}\widehat{u}_k}{2\pi\left(i(k-1)\right)^{n+1}\widehat{u}_{k-1}}. \tag{32}$$

We expect (29) to hold with z^k replaced by z_k^n (and with $O(|k|^{-2})$ as the last term if $u^{(n+1)}(x)$ is continuous for all $x \neq \gamma$). Therefore the quantity

$$d_k^n = |z_k^n - e^{-i\gamma}| \tag{33}$$

is expected to be small if $u(x)$ satisfies the regularity assumptions given prior to eq. (31) and if $|k|$ is large enough for the asymptotic assumption to be valid. If (31) does not hold, d_k^n will generally be larger.

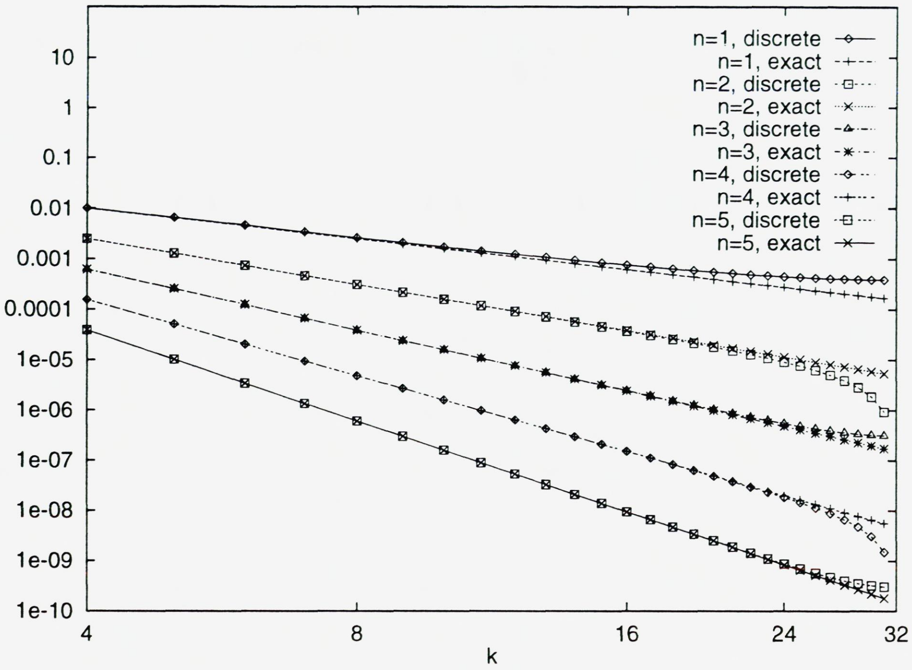

Figure 25: Absolute values of the discrete (for $N = 64$) and exact Fourier coefficients of the jump-functions $V_n(x;0)$.

However, because a collocation method is used in this paper, only the *discrete* Fourier coefficients of $u(x)$ are available, not the exact ones. The difference between discrete and exact Fourier coefficients are given in general by (16) and illustrated for the jump-functions $V_n(x;0)$, $n = 1,2,3,4,5$, in figure 25.

Because of this difference, we can not in general expect (31) to hold with the same accuracy unless we decrease $|k|$ to avoid the highest discrete coefficients, and figure 25 shows that the decrease in $|k|$ has to be substantial for the small values of n. This is clearly unfortunate, as we wish to use the highest coefficients because of the asymptotic assumptions. We shall therefore study the difference between discrete and exact coefficients more closely, using (15) and (16).

When the first terms of the sum in (16) are written explicitly and (15) is used, we obtain for $k = 1,2,\ldots,N/2-1$, and $n = 1,2,\ldots$:

$$\begin{aligned}(\widetilde{V}_n)_k(\gamma) &= \frac{e^{-ik\gamma}}{2\pi(ik)^{n+1}} + \sum_{m=1}^{\infty}\left(\frac{e^{-i(k-mN)\gamma}}{2\pi i^{n+1}(k-mN)^{n+1}} + \frac{e^{-i(k+mN)\gamma}}{2\pi i^{n+1}(k+mN)^{n+1}}\right)\\ &= \frac{e^{-ik\gamma}}{2\pi i^{n+1}}\left(\frac{1}{k^{n+1}} + \frac{e^{iN\gamma}}{(k-N)^{n+1}} + \frac{e^{-iN\gamma}}{(k+N)^{n+1}}\right.\\ &\quad \left. + \frac{e^{i2N\gamma}}{(k-2N)^{n+1}} + \frac{e^{-i2N\gamma}}{(k+2N)^{n+1}} + \ldots\right). \qquad (34)\end{aligned}$$

The first term in the sum on the right-hand side gives the exact coefficient. The rest of the terms have decreasing absolute values, and for $\gamma = 0$ it is easy to see that they are all positive

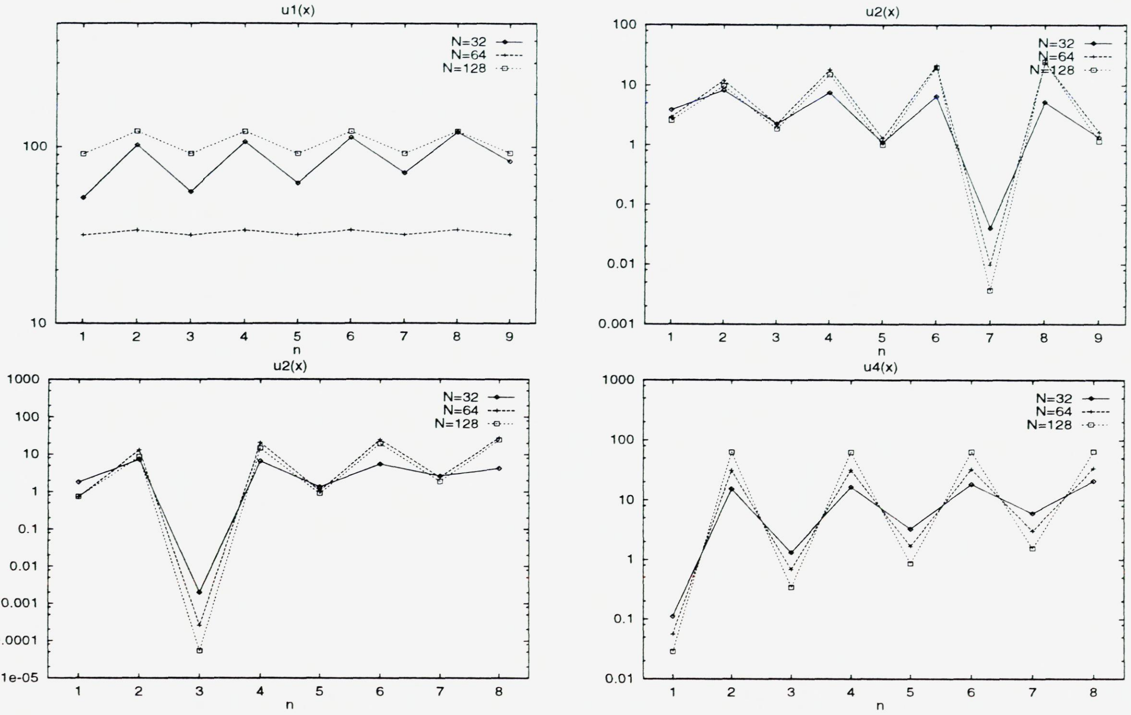

Figure 26: Size of the quantity $k\delta_k^n$, indicating whether the leading discontinuity of $u(x)$ is in the nth derivative and at the given position. The functions $u_1(x)$–$u_4(x)$ are given by (26).

for odd n, while they form an alternating series if n is even. This explains the differences between the discrete and exact coefficients in figure 25. The same behaviour was seen in the discrete coefficients of $u_2(x)$, $u_3(x)$, and $u_4(x)$ in figure 21.

Unless $\gamma = j2\pi/N$ for an integer j (i.e. γ is a grid-point), the sum over m in (34) will depend on γ. The use of (32) with the discrete Fourier coefficients of $u(x)$ and $V_n(x;0)$ instead of the exact ones, will therefore generally not lead to a good approximation to γ. However, because γ is assumed to be known, we can define

$$\zeta_k^n = \frac{\widetilde{u}_k/(\widetilde{V}_n)_k(\gamma)}{\widetilde{u}_{k-1}/(\widetilde{V}_n)_{k-1}(\gamma)}. \tag{35}$$

With the regularity assumptions given for (31), the asymptotic behaviour of the discrete Fourier coefficients is

$$\widetilde{u}_k = A(\widetilde{V}_n)_k(\gamma) + O(|k|^{-(n+2)}), \quad \text{as } |k| \to \infty, \tag{36}$$

and we obtain

$$\zeta_k^n = 1 + O(|k|^{-1}), \quad \text{as } |k| \to \infty. \tag{37}$$

The quantity

$$\delta_k^n = |\zeta_k^n - 1| = ((\Re(\zeta_k^n - 1)^2 + (\Im(\zeta_k^n))^2)^{1/2} \tag{38}$$

is expected to be small if $u(x)$ satisfies the regularity assumptions given prior to eq. (31) and $|k|$ is large enough for the asymptotic assumption to be valid.

Figure 26 shows $k\delta_k^n$ for $k = N/2 - 1$ and different values of n for the four functions (26). When the leading discontinuity of the function is in the nth derivative, $k\delta_k^n$ has a minimum value significantly smaller than 1 for this n. Using this method, the "optimal" accuracy (zero jumps) shown in figure 22 is obtained for the cases where the first Q derivatives are continuous, otherwise the results are not affected.

9 Conclusions

We have studied the accuracy and robustness of calculation of approximate derivatives by the modified Fourier collocation method [8], which uses piecewise polynomials to represent discontinuities in the approximated function and its Q first derivatives. The magnitudes of these jump discontinuities are determined from the discrete Fourier coefficients of the function. It has been demonstrated through numerical examples that the theoretical, asymptotic orders of convergence [7, 8] can be achieved with relatively few collocation points.

The numerical precision limits the accuracy of the calculations in different ways. The smallest discrete Fourier coefficients for the piecewise polynomials involved become very small for large values of Q, but this can be overcome by using analytic formulae for these coefficients as discussed in section 3. However, for these values of Q and N, the system of equations (12) for the jumps becomes numerically singular (or rank-deficient if the system is overdetermined), so the accuracy is nevertheless reduced. With the high orders of accuracy obtainable by this method, increased numerical precision would immediately be rewarded by substantial improvement of the results.

For 16-digit (double) precision, an upper limit of 10^{12} is recommended for the condition number of the matrix used in the system for the jumps. For a problem with M discontinuity points, a system including the MQ highest positive and negative discrete Fourier coefficients gives a good balance between accuracy and robustness. This overdetermined system can be solved by the least squares method.

The pointwise errors displayed in section 6 show that the errors are essentially concentrated near the discontinuity points, as is common in approximation of piecewise smooth functions. High orders of accuracy are obtained everywhere, but the results are in many cases several orders of magnitude better away from the discontinuity points.

An example of a situation occurring in applications to problems with complex geometries was studied in section 7, namely when a part of the computational domain is regarded as "exterior", and not really interesting for the results. The sensitivity of collocation point positions relative to the discontinuity points were studied, with encouraging results. This topic is discussed further in [9].

Finally we considered the situation where the functions to be approximated were smoother than expected, in order not to introduce artificially large errors in the approximation of such functions. Several aspects of the topic were discussed, but the choice of strategy and the degree of sensitivity to these properties of the functions are probably problem-dependent and should be reconsidered in the context of specific applications.

Stability of the differentiation operators in applications to partial differential equations are not discussed in this paper. The implementation of boundary conditions has a major influence on the stability properties, in particular for spectral methods, so the stability should be discussed in connection with particular applications. Stability for the heat equation with Dirichlet boundary conditions is discussed in [9].

References

[1] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions.* Dover, New York, 1970.

[2] E. Anderson, Z. Bai, C. Bischof, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, S. Ostrouchov, and D. Sorensen. *LAPACK Users' Guide.* Society for Industrial and Applied Mathematics, Philadelphia, 1992.

[3] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang. *Spectral methods in fluid dynamics.* Springer-Verlag, New York, 1988.

[4] B. W. Char, K. O. Geddes, G. H. Gonnet, B. L. Leong, M. B. Monagan, and S. M. Watt. *First Leaves: A Tutorial Introduction to Maple V.* Springer-Verlag, New York, 1992.

[5] K. S. Eckhoff. Accurate and efficient reconstruction of discontinuous functions from truncated series expansions. *Math. Comp.*, 61:745–763, 1993.

[6] K. S. Eckhoff. Accurate reconstructions of functions of finite regularity from truncated series expansions. *Math. Comp.*, 64:671–690, 1995.

[7] K. S. Eckhoff. On a high order numerical method for functions with singularities. In preparation, 1995.

[8] K. S. Eckhoff. On a high order numerical method for solving partial differential equations in complex geometries. To appear in *J. Sci. Comput.*, 1995.

[9] K. S. Eckhoff and C. E. Wasberg. Solution of parabolic partial differential equations in complex geometries by a modified Fourier collocation method. To appear in the Proceedings from ICOSAHOM'95, 1995.

[10] G. H. Golub and C. F. Van Loan. *Matrix Computations.* John Hopkins University, Baltimore, MD, 1983.

[11] D. Gottlieb and S. A. Orszag. *Numerical analysis of spectral methods: Theory and applications.* SIAM, Philadelphia, 1977.

[12] D. Gottlieb, C.-W. Shu, A. Solomonoff, and H. Vandeven. On the Gibbs phenomenon I: recovering exponential accuracy form the Fourier partial sum of a nonperiodic analytic function. *J. Comp. Appl. Math.*, 43:81–98, 1992.

[13] A. Zygmund. *Trigonometric Series.* Cambridge University Press, Cambridge, 1968.