Open access • Journal Article • DOI:10.1007/S10208-013-9158-8

On the Numerical Stability of Fourier Extensions — Source link

Ben Adcock, Daan Huybrechs, Jesús Martín-Vaquero

Institutions: Purdue University, Katholieke Universiteit Leuven, University of Salamanca

Published on: 01 Aug 2014 - Foundations of Computational Mathematics (Springer US)

Topics: Fourier series, Discrete Fourier series, Fourier inversion theorem, Sine and cosine transforms and Discrete-time Fourier transform

On the numerical stability of Fourier extensions

Ben Adcock, Daan Huybrechs
and Jesús Martín–Vaquero

Report TW 615, June 2012

Katholieke Universiteit Leuven
Department of Computer Science
Celestijnenlaan 200A – B-3001 Heverlee (Belgium)

On the numerical stability of Fourier extensions

Ben Adcock, Daan Huybrechs and Jesús Martín–Vaquero

Report TW 615, June 2012

Department of Computer Science, KU Leuven

Abstract

An effective means to approximate an analytic, nonperiodic function on a bounded interval is by using a Fourier series on a larger domain. When constructed appropriately, this so-called Fourier extension is known to converge exponentially fast in the truncation parameter. However, computing a Fourier extension requires solving an ill-conditioned linear system. The purpose of this paper is to show that, despite such ill-conditioning, Fourier extensions are actually numerically stable when implemented in finite arithmetic. Moreover, the convergence rate of such numerical extensions is at least spectral, and sometimes exponential. Thus, for Fourier extensions at least, ill-conditioning of the linear system does not prohibit a good approximation.

In the second part of this paper we consider the problem of computing Fourier extensions from equispaced data. A result of Platte, Trefethen & Kuijlaars states that no method for this problem can be both numerically stable and exponentially convergent. We explain how Fourier extensions relate to this theoretical barrier, and demonstrate that they are particularly well suited for this problem: namely, they obtain high-order (in particular, always spectral) convergence in a numerically stable manner.

Keywords : Fourier series, Fourier extensions, least squares approximation, ill-conditioned problems
MSC : Primary : 42A10, Secondary : 42C15, 65D15.

On the numerical stability of Fourier extensions

Ben Adcock
Department of Mathematics
Simon Fraser University
Canada

Daan Huybrechs
Department of Computer Science
Katholieke Universiteit Leuven
Belgium

Jesús Martín–Vaquero
Department of Applied Mathematics
E.T.S.I.I. Béjar, University of Salamanca
Spain

June 12, 2012

Abstract

An effective means to approximate an analytic, nonperiodic function on a bounded interval is by using a Fourier series on a larger domain. When constructed appropriately, this so-called Fourier extension is known to converge exponentially fast in the truncation parameter. However, computing a Fourier extension requires solving an ill-conditioned linear system. The purpose of this paper is to show that, despite such ill-conditioning, Fourier extensions are actually numerically stable when implemented in finite arithmetic. Moreover, the convergence rate of such numerical extensions is at least spectral, and sometimes exponential. Thus, for Fourier extensions at least, ill-conditioning of the linear system does not prohibit a good approximation.

In the second part of this paper we consider the problem of computing Fourier extensions from equispaced data. A result of Platte, Trefethen & Kuijlaars states that no method for this problem can be both numerically stable and exponentially convergent. We explain how Fourier extensions relate to this theoretical barrier, and demonstrate that they are particularly well suited for this problem: namely, they obtain high-order (in particular, always spectral) convergence in a numerically stable manner.

1 Introduction

Let $f : [-1, 1] \to \mathbb{R}$ be an analytic function. When periodic, an extremely effective means to approximate f is via its truncated Fourier series. This approximation not only converges exponentially fast in the truncation parameter N, it can also be computed efficiently via the Fast Fourier Transform (FFT). Moreover, Fourier series possess high resolution power. One requires an optimal 2 modes per wavelength to resolve oscillatory behaviour, making Fourier methods well-suited for (most notably) PDE problems with oscillatory solutions [17].

For these reasons, Fourier series are extremely widely used in practice. However, the situation changes completely when f is nonperiodic. In this case, rather than exponential convergence, one witnesses the familiar Gibbs phenomenon near $x = \pm 1$ and only linear pointwise convergence in $(-1, 1)$.

1.1 Fourier extensions

For analytic and nonperiodic functions, one way to retain the good properties of a Fourier series expansion (i.e. exponential convergence and high resolution power) is to seek to approximate f with a Fourier series on an *extended* domain $[-T, T]$. Here $T > 1$ is a user-determined parameter. Thus, we seek an approximation $F_N(f)$ to f from the set

$$\mathcal{G}_N := \operatorname{span}\left\{\phi_n : |n| \leq N\right\}, \qquad \phi_n(x) := \frac{1}{\sqrt{2T}} \mathrm{e}^{\mathrm{i}\frac{n\pi}{T}x}.$$

Although there are many potential ways to define $F_N(f)$, in [7, 10, 18] it was proposed to compute $F_N(f)$ as the best approximation to f on $[-1,1]$ in a least squares sense:

$$F_N(f) := \underset{\phi \in \mathcal{G}_N}{\operatorname{argmin}} \|f - \phi\|. \tag{1.1}$$

Here $\|\cdot\|$ is the standard norm on $\mathrm{L}^2(-1,1)$ – the space of square-integrable functions on $[-1,1]$. Henceforth, we shall refer to $F_N(f)$ as the *continuous* Fourier extension (FE) of f.

In [1, 18] it was shown that the continuous FE $F_N(f)$ converges exponentially fast in N (see also Theorem 2.10), and has a resolution constant (number of degrees of freedom per wavelength required to resolve an oscillatory wave) that ranges between 2 and π depending on the choice of the parameter T, with $T \approx 1$ giving close to the optimal value 2 (see §2.3.1 for a discussion). Thus the continuous FE successfully retains the key properties of rapid convergence and high resolution power of a standard Fourier series for the case of a nonperiodic function.

We remark in passing that, in practice, one does not usually compute the continuous FE [1, 18]. A more convenient approach is to replace (1.1) by the discrete least squares

$$\tilde{F}_N(f) := \underset{\phi \in \mathcal{G}_N}{\operatorname{argmin}} \sum_{|n| \leq N} |f(x_n) - \phi(x_n)|^2, \tag{1.2}$$

for nodes $\{x_n\}_{|n| \leq N} \subseteq [-1,1]$. We refer to $\tilde{F}_N(f)$ as the *discrete* Fourier extension of f. When chosen suitably – in particular, as in (2.10) – such nodes ensure that the difference in approximation properties between the extensions (1.1) and (1.2) is minimal (for details see §2.2).

1.2 Numerical convergence and stability of Fourier extensions

The approximation properties of the continuous and discrete FE's have been analysed previously [1, 18]. Therein it was also observed numerically that the condition numbers of the matrices A and $\tilde{A}$ associated to the least squares (1.1) and (1.2) are exponentially large in N. Thus, if $a = (a_{-N}, \ldots, a_N)^\top$ is the vector of *coefficients* of the continuous or discrete FE (i.e. $F_N(f)$ or $\tilde{F}_N(f)$ is given by $\sum_{|n| \leq N} a_n \phi_n$), one expects small perturbations in f to lead to large errors in a. In other words, the computation of the coefficients of the (continuous or discrete) FE is unstable. In the first result of this paper we prove such exponential growth of the condition numbers of these matrices, and derive the precise rate.

Because of this ill-conditioning, it is tempting to think that FE's will be useless in applications. Indeed, at first sight it is reasonable to expect that the good approximation properties of *exact* FE's (i.e. those obtained in exact arithmetic) will be destroyed when computing *numerical* FE's in finite precision. However, previous numerical studies [1, 7, 10, 18, 20, 21] indicate otherwise. Despite very large condition numbers, one typically obtains an extremely good approximation with a numerical FE, even for poorly behaved functions and in the presence of noise. The aim of this paper is to give a full explanation of this phenomenon.

The explanation we provide can be summarised as follows. In computations, our interest does not lie with the accuracy in computing the coefficients $\{a_n\}_{n=-N}^N$, but rather the accuracy of the numerical FE $\sum_{|n| \leq N} a_n \phi_n$. As we show, although the mapping from a function to its coefficients is ill-conditioned, the mapping from f to its numerical FE is, in fact, well-conditioned. In other words, small singular values of A (or $\tilde{A}$) have a significant effect on a, but little effect on the FE itself.

Whilst this observation explains the apparent stability of numerical FE's, it does not address their approximation properties. In [1, 18] it was shown that the exact continuous and discrete FE's $F_N(f)$ and $\tilde{F}_N(f)$ converge exponentially fast in N. However, the fact that there may be substantial differences between the coefficients of $F_N(f)$, $\tilde{F}_N(f)$ and those of the numerical FE's, which henceforth we denote by $G_N(f)$ and $\tilde{G}_N(f)$, suggests that exponential convergence may not be witnessed in finite arithmetic. As we show later, for a large class of functions, exponential convergence of $F_N(f)$ (or $\tilde{F}_N(f)$) essentially implies exponential growth of the norm $\|a\|$ of the exact (infinite precision) coefficient vector. Hence, whenever N is sufficiently large, there must be a discrepancy between this coefficient vector and its numerically computed counterpart, meaning that the numerical extensions $G_N(f)$ and $\tilde{G}_N(f)$ may not exhibit the same convergence. In the first half of this paper, besides showing stability, we also give a complete analysis and description the convergence of $G_N(f)$ and $\tilde{G}_N(f)$, and discuss how this contrasts with that of $F_N(f)$ and $\tilde{F}_N(f)$.

In summary, the main results of the first half of the paper are as follows:

1. The condition numbers of the matrices A and $\tilde{A}$ of the continuous and discrete FE's are exponentially large in N (Theorems 3.1 and 3.2).

2. The condition number of the numerical continuous FE mapping $f \mapsto G_N(f)$ satisfies

$$\kappa(G_N) \lesssim \frac{1}{\sqrt{\epsilon}}, \quad \forall N \in \mathbb{N},$$

where $\epsilon = \epsilon_{\text{mach}}$ is the machine precision used (Theorem 4.7 – see §4.3 for a definition of κ). Moreover, the error $\|f - G_N(f)\|$ decays exponentially fast in N up to some breakpoint N_0, and spectrally fast once $N > N_0$ down to a particular tolerance (see §4.2.1). This tolerance, the *maximal achievable accuracy* of $G_N(f)$, is of order $\sqrt{\epsilon}$, i.e.

$$\limsup_{N\to\infty} \|f - G_N(f)\| \lesssim \sqrt{\epsilon}.$$

In addition, the breakpoint N_0 is independent of f and depends only on ϵ and T.

3. The numerical discrete FE $f \mapsto \tilde{G}_N(f)$ satisfies $\kappa(\tilde{G}_N) \lesssim 1$, $\forall N \in \mathbb{N}$ (Theorem 4.7). Moreover, the error $\|f - \tilde{G}_N(f)\|$ is exponentially decaying for $N \leq N_1 := 2N_0$, and decreases spectrally once $N > N_1$. The maximal achievable accuracy for $\tilde{G}_N(f)$ is of order ϵ (see §4.2.2).

Remark 1.1 In this paper we refer to three types of convergence of an approximation $f_N \approx f$. We say that f_N converges *algebraically* fast to f at rate k if $\|f - f_N\| = \mathcal{O}\left(N^{-k}\right)$ as $N \to \infty$. Conversely, f_N converges *spectrally* fast if the error $\|f - f_N\|$ decays faster than any algebraic power of N^{-1}. Finally, we say f_N converges exponentially fast to f if there exists a $\rho > 1$ such that $\|f - f_N\| = \mathcal{O}\left(\rho^{-N}\right)$.

As we explain in §4, the reason for the disparity between the exact and numerical FE's can be traced to the fact that the system of functions $\{e^{i\frac{n\pi}{T}\cdot}\}_{n\in\mathbb{Z}}$ forms a *frame* for $L^2(-1,1)$. The inherent redundancy of this frame, i.e. the fact that any function f has infinitely many expansions in this system, leads to both the ill-conditioning in the coefficients of F_N and $\tilde{F}_N$, as well as the numerical stability and differing convergence of G_N and $\tilde{G}_N$.

This comment aside, the above results show that the numerical continuous FE converges exponentially fast in the regime $N < N_0$, and then spectrally fast beyond this point down to a best achievable accuracy of order 10^{-8} (we use standard precision $\epsilon_{\text{mach}} \approx 10^{-16}$ in our experiments). This latter property is undesirable: it means that one cannot obtain more than 7 or 8 digits of accuracy in general. When combined with the fact that the condition number $\kappa(G_N) \approx 10^8$ is rather large, this suggests that the continuous FE may be unreliable in practice. However, the above results also show that the discrete FE is completely stable when implemented numerically. Moreover, it possesses the same qualitative convergence behaviour as the continuous FE, but with two key differences. First, the region of guaranteed exponential convergence is precisely twice as large, $N_1 = 2N_0$, and second the maximal achievable accuracy is on the order of machine precision, as opposed to its square-root. Thus, an important conclusion of the first half of this paper is the following: it is possible to compute numerically stable FE's of analytic functions which converge at least spectrally fast in N (in particular, exponentially fast for all small N), and which attain close to machine accuracy for N sufficiently large.

Remark 1.2 This paper is about the discrepancy between theoretical properties of solutions to (1.1) and (1.2) and their numerical solutions when computed with standard solvers. Throughout we shall consistently use *Mathematica*'s `LeastSquares` routine in our computations, though we would like to stress that *Matlab*'s command \ gives very similar results. Occasionally, to compare theoretical and numerical properties, we shall carry out computations in additional precision to eliminate the effect of round-off error. When done, this will be stated explicitly. Otherwise, it is to be assumed that all computations are carried out as described in standard precision.

1.3 Fourier extensions from equispaced data

In many applications, one is faced with the problem of recovering an analytic function f to high accuracy from its values on an equispaced grid $\left\{f\left(-1+\frac{2m}{M}\right) : m = 0,\dots,M\right\}$. This problem turns out to be quite challenging. For example, recalling the famous Runge phenomenon, one notices that the M^{th} degree polynomial interpolant of this data will diverge exponentially as $M \to \infty$ unless f is analytic in a sufficiently large region.

Numerous approaches have been proposed to address this problem, and thereby 'overcome' the Runge phenomenon (see [8, 24] for a comprehensive list). Indeed, many are quite effective in practice. However, a common feature of many such methods is instability. This was explained recently by Platte, Trefethen & Kuijlaars in [24], wherein it was shown that any exponentially convergent method for recovering analytic functions f from equispaced data must also be exponentially ill-conditioned. As was also proved, the best possible that can be achieved by a stable method is root-exponential convergence in M. This profound result, most likely the first of its kind for this type of problem, places an important theoretical barrier and benchmark against which all such methods must be measured.

As we show in the first half of this paper, the numerical discrete and continuous FE's are well-conditioned and have good convergence properties. However, neither deals with equispaced data. In the second half of this paper we consider the use of FE's for this problem. Specifically, if $x_n = \frac{n}{M}$ for $|n| \leq M$, we study the so-called *equispaced* Fourier extension

$$F_{N,M}(f) := \underset{\phi \in \mathcal{G}_N}{\operatorname{argmin}} \sum_{|n| \leq M} |f(x_n) - \phi(x_n)|^2, \tag{1.3}$$

and its numerically computed counterpart $G_{N,M}(f)$. Note that our primary interest lies with the case where $M = \gamma N$ for some $\gamma \geq 1$, i.e. where the number of equispaced points M scales linearly with N. We refer to γ as the oversampling parameter: observe that (1.3) results in an $(2M+1) \times (2N+1)$ overdetermined least squares problem for the coefficients of $F_{N,M}(f)$. We denote the corresponding matrix by $\bar{A}$.

Our main results concerning the equispaced FE are as follows:

1. The condition number of $\bar{A}$ is exponentially large as $N, M \to \infty$ with $M \geq N$ (see §5.2.2).
2. The approximation $F_{N,\gamma N}(f)$ suffers from a Runge phenomenon for any fixed $\gamma \geq 1$. In particular, if f has a complex singularity sufficiently close to $[-1, 1]$, then the error $\|f - F_{N,\gamma N}(f)\|$ diverges exponentially fast in N (see §5.2).
3. The scaling $M = \mathcal{O}\left(N^2\right)$ is required to avoid a Runge phenomenon in $F_{N,M}(f)$. In this case, $F_{N,M}(f)$ converges at the same rate as the exact continuous FE $F_N(f)$, i.e. exponentially fast in N (see §5.2.1). However, the condition number of $\bar{A}$ remains exponentially large (see §5.2.2).

At first sight, these results appear to indicate that FE's perform suboptimally in view of the barrier of Platte, Trefethen & Kuijlaars. Indeed, convergence can only be ensured with $M = \mathcal{O}\left(N^2\right)$, resulting in only root-exponential convergence in M, and one is still faced with ill-conditioning. However, much like with the continuous and discrete extensions, there is a significant discrepancy between the exact equispaced extension $F_{N,M}(f)$ and its numerical counterpart $G_{N,M}(f)$. Indeed, in §5.3 and 5.4 we establish the following results:

1. The condition number $\kappa(G_{N,\gamma N})$ satisfies

   $$\kappa(G_{N,\gamma N}) \lesssim \epsilon^{-a(\gamma;T)}, \quad \forall N \in \mathbb{N},$$

   where $\epsilon = \epsilon_{\text{mach}}$ is the machine precision used, and $0 < a(\gamma;T) < 1$ is a constant independent of N with $a(\gamma;T) \to 0$ as $\gamma \to \infty$ for fixed T.
2. The error $\|f - G_{N,\gamma N}(f)\|$ behaves as follows:
   (i) If $N < N_2$, where N_2 is a function-independent breakpoint, $\|f - G_{N,\gamma N}(f)\|$ converges/diverges exponentially at the same rate as $\|f - F_{N,\gamma N}(f)\|$.
   (ii) If $N_2 \leq N < N_1$, where N_1 is as introduced previously in §1.2, then $\|f - G_{N,\gamma N}(f)\|$ decays exponentially at the same rate as $\|f - F_N(f)\|$, where $F_N(f)$ is the exact continuous FE.
   (iii) If $N > N_1$ then $\|f - G_{N,\gamma N}(f)\|$ decays spectrally fast in N down to a maximal achievable accuracy of order $\epsilon^{1-a(\gamma;T)}$.

These results show that, after a (function-independent) regime of possible divergence, we witness exponential convergence followed by spectral convergence, down to a best achievable accuracy depending only on the machine precision used. Furthermore, since $a(\gamma;T) \to 0$ as $\gamma \to \infty$, more oversampling leads to an improved maximal achievable accuracy, as well as better stability. As we show via numerical experiment and numerical computation of the relevant constants, double oversampling $\gamma = 2$ gives

perfectly adequate numerical results in practice. Note that a larger γ also yields a less severe rate of exponential divergence for $N < N_2$, with $\gamma = 2$ virtually eliminating such behaviour for all reasonable functions f (see §5.3 for details).

The main conclusion of this analysis is that equispaced FE's obtain overall spectral convergence in a numerically stable manner. Hence, there is no contradiction with the theorem of Platte, Trefethen & Kuijlaars. Moreover, in some senses we get the best possible convergence permitted by this theorem (i.e. spectral convergence), thereby making FE's eminently well suited for this problem. Numerical results confirm both the excellent accuracy and stability of this approach.

1.4 Relation to previous work

One-dimensional FE's for overcoming the Gibbs and Runge phenomena were studied (without analysis) in [7] and [8], and applications to surface parametrizations considered in [10]. Analysis of the convergence of the exact continuous and discrete FE's was presented by the authors in [1, 18], along with a study of resolution power [1]. The content of the first half of this paper, namely stability analysis of exact/numerical FE's, follows on directly from this previous work.

A different approach to FE's, known as the FC–Gram method, was introduced in [22]. This approach forms a central part of an extremely effective method for solving PDE's in complex geometries [2, 9]. For previous work on using FE's for PDE problems (so-called Fourier embeddings) see [4, 23].

FE's from equispaced data were arguably first considered in detail in [7, 10]. In particular, Boyd [7] describes the use of truncated singular value decompositions (SVD's) to compute equispaced FE's, and gives extensive numerical experiments (see also [8]). Most recently Lyon has presented an analysis of equispaced FE's computed using truncated SVD's [20]. In particular, numerical stability and convergence (down to close to machine precision) were shown. In §5.3 we discuss this work in more detail (see, in particular, Remark 5.9), and give further insight into some of the questions raised in [20].

1.5 Outline of the paper

The main results of this paper are as described above. Besides these, a consistent theme of this paper is the question of how to choose the extension parameter T. Specifically, we provide results for the effect of T on exact and numerical FE's, and discuss the cases corresponding to small $T \approx 1$ and large $T \gg 1$. We also consider strategies for varying T with N.

The outline of the remainder of this paper is as follows. In §2 we recap properties of the continuous and discrete FE's from [1, 18], including convergence. Ill-conditioning of the coefficient map is proved in §3, and in §4 we consider the stability of the numerical extensions and their convergence. Finally, in §5 we consider the case of equispaced FE's.

Notation. If $I \subseteq \mathbb{R}$ is an interval we write $\mathrm{L}^2(I)$ for the space of square-integrable functions on I, with corresponding inner product $\langle \cdot, \cdot \rangle_I$ and norm $\|\cdot\|_I$. When $I = [-1, 1]$ is the unit interval then we shall merely write $\langle \cdot, \cdot \rangle$ and $\|\cdot\|$. If w is a nonnegative and integrable weight function on I then $\mathrm{L}^2_w(I)$ will denote the space of weighted square-integrable functions on I with respect to w. We write $\langle \cdot, \cdot \rangle_{w,I}$ and $\|\cdot\|_{w,I}$ for the corresponding inner product and norm (as before, we drop the I subscript whenever $I = [-1, 1]$). We shall occasionally write $\|\cdot\|_{\infty,I}$ for the uniform norm on I.

2 Fourier extensions

In this section we discuss FE's, and recap the important details from [1, 18].

2.1 Two interpretations of Fourier extensions

There are two important interpretations of FE's which inform their approximation properties and their stability respectively. These are described in the next two sections.

2.1.1 Fourier extensions as polynomial approximations

The space $\mathcal{G}_N$ can be decomposed as $\mathcal{G}_N = \mathcal{C}_N \oplus \mathcal{S}_N$, where

$$\mathcal{C}_N = \operatorname{span}\left\{\cos \tfrac{n\pi}{T} x : n = 0, \ldots, N\right\}, \quad \mathcal{S}_N = \operatorname{span}\left\{\sin \tfrac{n\pi}{T} x : n = 1, \ldots, N\right\},$$

consist of even and odd functions respectively. Likewise, for f we have

$$f(x) = f_e(x) + f_o(x), \qquad f_e(x) = \tfrac{1}{2}\left[f(x) + f(-x)\right], \quad f_o(x) = \tfrac{1}{2}\left[f(x) - f(-x)\right],$$

and for any FE f_N of f:

$$f_N = f_{e,N} + f_{o,N}, \quad f_{e,N} \in \mathcal{C}_N,\ f_{o,N} \in \mathcal{S}_N. \tag{2.1}$$

Throughout this paper we shall use the notation f_N to denote an arbitrary FE of f when not wishing to specify its construction. From (2.1), it follows that the problem of approximating f via a FE f_N decouples into two problems $f_{e,N} \approx f_e$ and $f_{o,N} \approx f_o$ on the half-interval $[0,1]$.

Let us define the mapping $y = y(x) : [0,1] \to [c(T),1]$ by $y = \cos\frac{\pi}{T}x$, where $c(T) = \cos\frac{\pi}{T}$. The functions $\cos\frac{n\pi}{T}x$ and $\sin\frac{(n+1)\pi}{T}x/\sin\frac{\pi}{T}x$ are algebraic polynomials of degree n in y. Therefore, $\mathcal{C}_N$ and $\mathcal{S}_N$ are (up to multiplication by $\sin\frac{\pi}{T}x$ for the latter) the subspaces $\mathbb{P}_N$ and $\mathbb{P}_{N-1}$ of polynomials of degree N and $N-1$ respectively in the transformed variable y. Letting

$$g_1(y) = f_e(x), \quad g_2(y) = \frac{f_o(x)}{\sin\frac{\pi}{T}x}, \qquad g_{1,N}(y) = f_{e,N}(x), \quad g_{2,N}(y) = \frac{f_{o,N}(x)}{\sin\frac{\pi}{T}x},$$

with $g_{1,N}(y) \in \mathbb{P}_N$ and $g_{2,N}(y) \in \mathbb{P}_{N-1}$, we now conclude that the approximation problem of the FE f_N in the variable x is completely equivalent to two polynomial approximation problems in the transformed variable $y \in [c(T),1]$. This fact is central to the analysis of FE's. Specifically, one can use the rich literature on polynomial approximations to determine explicitly the theoretical behaviour of the exact/discrete Fourier extension (see §2.3).

Remark 2.1 The interpretation of f_N as essentially a polynomial approximation is purely for the purposes of analysis. We always carry out computations in the x-domain, i.e. by using the standard trigonometric basis for $\mathcal{G}_N$ (see §2.2).

The interval $[c(T),1] \subseteq (-1,1]$ is not standard. Therefore, it is convenient to map it affinely to $[-1,1]$. To this end, let

$$u := u(y) = 2\frac{y - c(T)}{1 - c(T)} - 1 \in [-1,1].$$

Observe that $y = y(u) = c(T) + \frac{1-c(T)}{2}(u+1)$. Denote by $m : [0,1] \to [-1,1]$ the mapping $x \mapsto u$, i.e.

$$u = m(x) = 2\frac{\cos\frac{\pi}{T}x - c(T)}{1 - c(T)} - 1. \tag{2.2}$$

Note that $x = m^{-1}(u) = \frac{T}{\pi}\arccos\left[c(T) + \frac{1-c(T)}{2}(u+1)\right]$. If we now let

$$h_i(u) = g_i(y(u)), \quad i = 1,2, \tag{2.3}$$

then the FE f_N is equivalent to the two polynomial approximations

$$h_{1,N}(u) = g_{1,N}(y(u)) = f_{e,N}(m^{-1}(u)), \quad h_{2,N}(u) = g_{2,N}(y(u)) = \frac{f_{o,N}(m^{-1}(u))}{\sin\left(\frac{\pi}{T}m^{-1}(u)\right)}, \tag{2.4}$$

of degree N and $N-1$ respectively in the new variable $u \in [-1,1]$.

2.1.2 Fourier extensions as frame approximations

Definition 2.2. *Let* H *be a Hilbert space with inner product* $\langle\cdot,\cdot\rangle$ *and norm* $\|\cdot\|$. *A set* $\{\phi_n\}_{n=1}^{\infty} \subseteq$ H *is a frame for* H *if (i)* $\operatorname{span}\{\phi_n\}_{n=1}^{\infty}$ *is dense in* H *and (ii) there exist* $c_1, c_2 > 0$ *such that*

$$c_1\|f\|^2 \leq \sum_{n=1}^{\infty} |\langle f, \phi_n\rangle|^2 \leq c_2\|f\|^2, \quad \forall f \in \mathrm{H}. \tag{2.5}$$

If $c_1 = c_2$ *then* $\{\phi_n\}_{n=1}^{\infty}$ *is referred to as a tight frame.*

Introduced by Duffin & Schaeffer [14], frames have become vitally important in signal processing [12]. Note that all orthonormal, indeed Riesz, bases are frames, but a frame need not be a basis. In fact, frames are typically *redundant*: any element $f \in \mathrm{H}$ may well have infinitely many representations of the form $f = \sum_{n=1}^{\infty} \alpha_n \phi_n$ with coefficients $\{\alpha_n\}_{n=1}^{\infty} \in l^2(\mathbb{N})$.

The following lemma is important to subsequent analysis:

Lemma 2.3 ([1]). *The set* $\{\frac{1}{\sqrt{2T}} \mathrm{e}^{\mathrm{i}\frac{n\pi}{T}x}\}_{n\in\mathbb{Z}}$ *is a tight frame for* $\mathrm{L}^2(-1,1)$ *whenever* $T > 1$.

Note that $\{\frac{1}{\sqrt{2T}} \mathrm{e}^{\mathrm{i}\frac{n\pi}{T}x}\}_{n\in\mathbb{Z}}$ is an orthonormal basis for $\mathrm{L}^2(-T,T)$: it is precisely the standard Fourier basis on $[-T,T]$. However, it forms only a frame when considered as a subset of $\mathrm{L}^2(-1,1)$. This fact means that ill-conditioning may well be an issue in numerical algorithms for computing FE's, due to the possibility of redundancies. As it happens, it is trivial to see that the set $\{\frac{1}{\sqrt{2T}} \mathrm{e}^{\mathrm{i}\frac{n\pi}{T}x}\}_{n\in\mathbb{Z}}$ is redundant. Indeed, any function $f \in \mathrm{L}^2(-1,1)$ has infinitely many extensions $\tilde{f} \in \mathrm{L}^2(-T,T)$. For each such $\tilde{f}$, the sum $\sum_{n\in\mathbb{Z}} \alpha_n \phi_n$, where $\alpha_n = \langle \tilde{f}, \phi_n \rangle_{[-T,T]}$ and $\phi_n(x) = \frac{1}{\sqrt{2T}} \mathrm{e}^{\mathrm{i}\frac{n\pi}{T}x}$, coincides with $\tilde{f}$ on $[-T,T]$ (it is precisely the Fourier series of $\tilde{f}$) and therefore f when restricted to $[-1,1]$.

The previous argument is valid for arbitrary $f \in \mathrm{L}^2(-1,1)$. When f has higher regularity, say $f \in \mathrm{H}^k(-1,1)$, where $\mathrm{H}^k(-1,1)$ is the k^{th} standard Sobolev space on $[-1,1]$, it is useful to note that there exist extensions $\tilde{f}$ with the same regularity on the torus $\mathbb{T} = [-T,T)$. This is the content of the next lemma. For convenience, given a domain I, we write $\|\cdot\|_{\mathrm{H}^k(I)}$ for the standard norm on $\mathrm{H}^k(I)$:

Lemma 2.4. *Let* $f \in \mathrm{H}^k(-1,1)$ *for* $k \in \mathbb{N}$. *Then there exists an extension* $\tilde{f} \in \mathrm{H}^k(\mathbb{T})$ *of* f *satisfying* $\|\tilde{f}\|_{\mathrm{H}^k(\mathbb{T})} \leq c_k(T)\|f\|_{\mathrm{H}^k(-1,1)}$, *where* $c_k(T) > 0$ *is independent of* f. *Moreover,* $f = \sum_{n\in\mathbb{Z}} \alpha_n \phi_n$, *where* $\alpha_n = \langle \tilde{f}, \phi_n \rangle_{[-T,T]}$ *satisfies* $\alpha_n = \mathcal{O}\left(n^{-k}\right)$ *as* $|n| \to \infty$.

Proof. The first part of the lemma follows directly from the proof of Theorem 2.1 in [1]. The second follows from integrating by parts k times and the fact that $\tilde{f}$ is periodic. □

This lemma, which shall be used later in studying stability of FE's, states that there exist representations of f in the frame $\{\frac{1}{\sqrt{2T}} \mathrm{e}^{\mathrm{i}\frac{n\pi}{T}x}\}_{n\in\mathbb{Z}}$ that have nice (i.e. rapidly decaying) coefficients and which cannot grow large on the extended region $[-T,T]$.

2.2 The continuous and discrete Fourier extensions

We now describe the two types of FE's we consider in the first part of this paper.

2.2.1 The continuous Fourier extension

The continuous FE of $f \in \mathrm{L}^2(-1,1)$, defined by (1.1), is the orthogonal projection onto $\mathcal{G}_N$. We have the following characterization:

Proposition 2.5 ([1, 18]). *Let* $F_N(f)$ *be the continuous FE (1.1) of a function* f, *and let* $h_{i,N}(u)$ *and* $h_i(u)$ *be given by (2.3) and (2.4) respectively. Then* $h_{1,N}(u)$ *and* $h_{2,N}(u)$ *are the expansions of* $h_1(u)$ *and* $h_2(u)$ *respectively in orthogonal polynomials with respect to the weight functions*

$$w_1(u) = [(1-u)(u-m(T))]^{-\frac{1}{2}}, \quad w_2(u) = [(1-u)(u-m(T))]^{\frac{1}{2}}, \quad u \in [-1,1], \tag{2.6}$$

where $m(T) = 1 - 2\mathrm{cosec}^2\left(\frac{\pi}{2T}\right) < -1$. *In other words,* $h_{i,N}(u)$, $i = 1,2$, *is the orthogonal projection of* $h_i(u)$ *onto* $\mathbb{P}_{N+1-i}$ *with respect to the weighted inner product* $\langle \cdot, \cdot \rangle_{w_i}$ *with weight function* w_i.

Computation of the $F_N(f)$ involves solving a linear system. Let us write $F_N(f) = \sum_{n=-N}^{N} a_n \phi_n$ with unknowns $\{a_n\}_{n=-N}^{N}$. If $a = (a_{-N}, \ldots, a_N)^\top$ and $b = (b_{-N}, \ldots, b_N)^\top$, where

$$b_n = \langle f, \phi_n \rangle = \int_{-1}^{1} f(x)\overline{\phi_n(x)}\,\mathrm{d}x, \quad n = -N, \ldots, N, \tag{2.7}$$

and $A \in \mathbb{C}^{(2N+1)\times(2N+1)}$ is the matrix with $(n,m)^{\text{th}}$ entry

$$A_{n,m} = \langle \phi_m, \phi_n \rangle = \int_{-1}^{1} \phi_m(x)\overline{\phi_n(x)}\,\mathrm{d}x, \quad n,m = -N, \ldots, N, \tag{2.8}$$

then a is the solution of the linear system $Aa = b$. We refer to the values $\{a_n\}_{n=-N}^{N}$ as the *coefficients* of the FE $F_N(f)$. Note that the matrix A is a Hermitian positive-definite, Toeplitz matrix with $A_{n,m} = A_{n-m}$, where $A_0 = \frac{1}{T}$ and $A_n = \frac{\sin \frac{n\pi}{T}}{n\pi}$ otherwise. In fact, A coincides with the so-called *prolate* matrix [27, 29]. We shall discuss this connection further in §4.

2.2.2 The discrete Fourier extension

The discrete FE $\tilde{F}_N(f)$ is defined by (1.2). To use this extension it is first necessary to choose nodes $\{x_n\}_{n=-N}^{N}$. This question was considered in [1], and a solution was obtained by exploiting the characterization of FE's as polynomial approximations in the transformed variable u.

A good system of nodes for polynomial interpolation are the Chebyshev nodes

$$u_n = \cos\left(\frac{(2n+1)\pi}{2N+2}\right), \quad n = 0, \dots, N. \tag{2.9}$$

Mapping these back to the x-variable and symmetrizing about $x = 0$ leads to the so-called *mapped symmetric Chebyshev* nodes

$$x_n = -x_{-n-1} = \frac{T}{\pi} \arccos\left[\frac{1}{2}(1 - c(T)) \cos\left(\frac{(2n+1)\pi}{2N+2}\right) + \frac{1}{2}(1 + c(T))\right], \quad n = 0, \dots, N. \tag{2.10}$$

This gives a set of $2N + 2$ nodes. Therefore, rather than (1.2), we define the discrete FE by

$$\tilde{F}_N(f) := \operatorname*{argmin}_{\phi \in \mathcal{G}'_N} \sum_{n=-N-1}^{N} |f(x_n) - \phi(x_n)|^2, \tag{2.11}$$

from now on, where $\mathcal{G}'_N = \mathcal{C}_N \oplus \mathcal{S}_{N+1}$. Exploiting the relation between FE's and polynomial approximations once more, we obtain the following:

Proposition 2.6. *Let $f_N = \tilde{F}_N(f) \in \mathcal{G}'_N$ be the discrete FE (2.11) based on the nodes (2.10), and let $h_i(u)$ and $h_{i,N}(u) \in \mathbb{P}_N$ be given by (2.3) and (2.4) respectively. Then $h_{i,N}(u)$, $i = 1, 2$ is the N^{th} degree polynomial interpolant of $h_i(u)$ at the Chebyshev nodes (2.9).*

Write $\phi_n(x) = \cos n\pi x$, $\phi_{-n-1}(x) = \sin n\pi x$, $n \in \mathbb{N}$, and let $\tilde{F}_N(f)(x) = \sum_{n=-N-1}^{N} a_n \phi_n(x)$. If $a = (a_{-N-1}, \dots, a_N)^{-T}$ and $\tilde{A} \in \mathbb{R}^{(2N+2)\times(2N+2)}$ has $(n, m)^{\text{th}}$ entry

$$\tilde{A}_{n,m} = \phi_m(x_n), \qquad n, m = -N-1, \dots, N, \tag{2.12}$$

then we have $\tilde{A}a = b$, where $b = (b_{-N-1}, \dots, b_N)^\top$ with $b_n = f(x_n)$.

The following lemma concerning the matrix $\tilde{A}$ will prove useful for what follows:

Lemma 2.7 ([1])**.** *Let $D \in \mathbb{R}^{(2N+2)\times(2N+2)}$ denote the diagonal matrix with entries $\frac{\pi}{N+1}$. Then $A_W = \tilde{A}^\top D \tilde{A}$ has entries*

$$\langle \phi_n, \phi_m \rangle_W = \int_{-1}^{1} \phi_n(x)\phi_m(x)W(x)\,\mathrm{d}x, \quad n, m = -N-1, \dots, N,$$

where W is the positive, integrable weight function given by $W(x) = \frac{\sqrt{2}\pi}{T} \frac{\cos \frac{\pi}{2T} x}{\sqrt{\cos \frac{\pi}{T} x - \cos \frac{\pi}{T}}}$.

Note that this lemma implies that the left-hand side of the normal equations (with constant weighting D) of the discrete FE correspond to the equations of a continuous FE based on the weighted least-squares minimization with weight function W.

2.3 Convergence of Fourier extensions

A detailed analysis of the convergence of the exact continuous FE, which we now recap, was carried out in [1, 18]. We commence with the following theorem:

Theorem 2.8 ([1])**.** *Suppose that $f \in \mathrm{H}^k(-1, 1)$ for some $k \in \mathbb{N}$ and that $T > 1$. If $F_N(f)$ is the continuous FE of f defined by (1.1), then*

$$\|f - F_N(f)\| \leq c_k(T) N^{-k} \|f\|_{\mathrm{H}^k(-1,1)}, \quad \forall n \in \mathbb{N}, \tag{2.13}$$

where $c_k(T) > 0$ is independent of f and N.

This theorem confirms *algebraic* convergence – that is, convergence of order N^{-k} for some fixed k – of the continuous FE $F_N(f)$ whenever the approximated function f has only finite degrees of smoothness. Conversely, if f is smooth, this theorem implies *spectral* convergence of $F_N(f)$, i.e. faster than any algebraic power of N^{-1}.

Suppose now that f is analytic. Although Theorem 2.8 indicates spectral convergence, it transpires that the convergence is actually exponential. This is a direct consequence of the interpretation of the $F_N(f)$ as the sum of two polynomial expansions in the transformed variable u (Proposition 2.5). To state the corresponding theorem, we first require the following definition:

Definition 2.9. *The Bernstein ellipse $B(\rho) \subseteq \mathbb{C}$ of index $\rho \geq 1$ is given by*

$$B(\rho) = \left\{ \tfrac{1}{2} \left(\rho^{-1} \mathrm{e}^{\mathrm{i}\theta} + \rho \mathrm{e}^{-\mathrm{i}\theta} \right) : \theta \in [-\pi, \pi] \right\}.$$

Given a Bernstein ellipse $B(\rho)$, we write $D(\rho) \subseteq \mathbb{C}$ for its image in the complex x-plane under the mapping $x = m^{-1}(u)$, where m is as in (2.2). We now have the following:

Theorem 2.10 ([1], [18])**.** *Suppose that f is analytic in $D(\rho^*)$ and not analytic inside any $D(\rho')$ with $\rho' > \rho^*$. Then, for some constant $c_f > 0$ proportional to $\sup_{z \in D(\rho^*)} |f(z)|$, we have $\|f - F_N(f)\| \leq c_f \rho^{-N}$, where $\rho = \min\{\rho^*, E(T)\}$ and $E(T) = \cot^2\left(\frac{\pi}{4T}\right)$.*

Proof. A full proof was given in [1, Thm 2.3]. The expansion g_N of an analytic function g in a system orthogonal polynomials with respect to some integrable weight function satisfies $\|g - g_N\|_\infty \leq c_g \rho^{-N}$, where c_g is proportional to $\sup_{z \in B(\rho)} |g(z)|$ [26]. Using this and Proposition 2.5, it remains only to determine the maximal parameter ρ of Bernstein ellipse $B(\rho)$ within which $h_1(u)$ and $h_2(u)$ are analytic.

The mapping $u = m(x)$ introduces a square-root type singularity into the functions $h_i(u)$ at the point $u = m(T) < -1$. Hence the maximal possible value of the parameter ρ satisfies

$$\tfrac{1}{2}(\rho + \rho^{-1}) = -m(T). \tag{2.14}$$

Observe that if $F(x) = x + \sqrt{x^2 - 1}$ then

$$F(m(T)) = E(T). \tag{2.15}$$

Thus, since $\rho > 1$, the solution to (2.14) is precisely $\rho = E(T)$. On the other hand, any complex singularity of f introduces a complex singularity of $h_i(u)$, which also limits this value. Hence we obtain stated minimum. □

Remark 2.11 Although Theorems 2.8 and 2.10 apply to $F_N(f)$, they also hold (up to a possible log factor corresponding to the Lebesgue constant of Chebyshev interpolation) for the discrete FE $\tilde{F}_N(f)$. This follows from the interpretation of $\tilde{F}_N(f)$ as a sum of polynomial interpolants (Proposition 2.6).

Theorem 2.10 shows that if f is analytic in a sufficiently large region (for example, if f is entire) then the rate of exponential convergence is precisely $E(T)$. Recall that the parameter T can be chosen by the user. We now discuss the effect of different choices of T.

2.3.1 The choice of T

Note that $E(T) \sim 1 + \pi(T-1)$ as $T \to 1^+$ and $E(T) \sim \frac{16}{\pi^2} T^2$ when $T \to \infty$. Thus, small T leads to a slower rate of exponential convergence, whereas large T gives a faster rate. However, as discussed in [1], a larger value of T leads to a worse resolution power, meaning that more degrees of freedom are required to resolve oscillatory behaviour. On the other hand, setting T sufficiently close to 1 yields a resolution power that is arbitrarily close to optimal.

In [1] a number of fixed values of T were used in numerical experiments. These typically give good results, with small values of T being particularly well suited for oscillatory functions. Another approach for choosing T was also discussed. This involves letting

$$T = T(N; \epsilon_{\mathrm{tol}}) = \frac{\pi}{4} \left(\arctan(\epsilon_{\mathrm{tol}})^{\frac{1}{2N}} \right)^{-1}, \tag{2.16}$$

where $\epsilon_{\mathrm{tol}} \ll 1$ is some tolerance (note that this is very much related to the Kosloff Tal–Ezer map in spectral methods for PDEs [6, 19] – see [1] for a discussion). This choice of T, which now depends

on N, is such that $E(T)^{-N} = \epsilon_{\text{tol}}$. Note that this limits the best achievable accuracy of the FE with this approach to $\mathcal{O}(\epsilon_{\text{tol}})$. However, setting $\epsilon_{\text{tol}} = 10^{-14}$ is normally sufficient in practice. Numerical experiments [1] indicate that this works well, especially for oscillatory functions. In fact, since

$$T(N;\epsilon_{\text{tol}}) \sim 1 - \frac{\log(\epsilon_{\text{tol}})}{\pi N} + \mathcal{O}\left(N^{-2}\right), \quad N \to \infty, \tag{2.17}$$

this approach has formally optimal resolution power.

Remark 2.12 The strategy (2.16) is particularly good for oscillatory problems. However, if this is not a concern, then an optimal choice appears to be $T = 2$. In this case, the FE has a particular symmetry which can actually be exploited to allow the efficient computation of FE's in only $\mathcal{O}\left(N(\log N)^2\right)$ operations [21]. The underlying reason for this is the particular structure of the singular values of the matrix A (see Remark 3.4 for further details).

3 Ill-conditioning of Fourier extension matrices

The redundancy of the frame $\{\frac{1}{\sqrt{2T}}\mathrm{e}^{\mathrm{i}\frac{n\pi}{T}\cdot}\}_{n\in\mathbb{Z}}$ means that the matrices associated with the continuous and discrete FE's are ill-conditioned. We next derive bounds for the condition number of these matrices. The spectrum of A is discussed further in §3.2.

3.1 The condition number of the exact/discrete Fourier extension

Theorem 3.1. *Let A be the matrix (2.8) of the continuous FE. Then the condition number of A is $\mathcal{O}\left(E(T)^{2N}\right)$ for large N. Specifically, the maximal and minimal eigenvalues satisfy*

$$T^{-1} \leq \lambda_{\max}(A) \leq 1, \qquad c_1(T)N^{-3}E(T)^{-2N} \leq \lambda_{\min}(A) \leq c_2(T)N^2E(T)^{-2N}, \tag{3.1}$$

where $c_1(T)$ and $c_2(T)$ are positive constants with $c_1(T), c_2(T) = \mathcal{O}(1)$ as $T \to 1^+$.

Proof. It is a straightforward exercise to verify that

$$\lambda_{\min}(A) = \min_{\phi\in\mathcal{G}_N}\left\{\|\phi\|^2 : \|\phi\|_{[-T,T]} = 1\right\}, \quad \lambda_{\max}(A) = \max_{\phi\in\mathcal{G}_N}\left\{\|\phi\|^2 : \|\phi\|_{[-T,T]} = 1\right\}. \tag{3.2}$$

Using the fact that $\|\phi\| \leq \|\phi\|_{[-T,T]}$ we first notice that $\lambda_{\max}(A) \leq 1$. On the other hand, setting $\phi = \frac{1}{\sqrt{2T}}$, we find that $\lambda_{\max}(A) \geq T^{-1}$, which completes the result for $\lambda_{\max}(A)$.

We now consider $\lambda_{\min}(A)$. Recall that any $\phi \in \mathcal{G}_N$ can be decomposed into even and odd parts ϕ_e and ϕ_o, with each function corresponding to a polynomial in the transformed variable u. Hence,

$$\lambda_{\min}(A) = \min_{\substack{\phi\in\mathcal{G}_N\\ \phi\neq 0}}\left\{\frac{\|\phi\|^2}{\|\phi\|^2_{[-T,T]}}\right\} = \min_{\substack{p_1\in\mathbb{P}_N\\ p_1\neq 0}}\min_{\substack{p_2\in\mathbb{P}_{N-1}\\ p_2\neq 0}}\left\{\frac{\|p_1\|^2_{w_1} + \|p_2\|^2_{w_2}}{\|p_1\|^2_{w_1,[m(T),1]} + \|p_2\|^2_{w_2,[m(T),1]}}\right\}, \tag{3.3}$$

where w_i, $i = 1, 2$, is given by (2.6). Since the weight function w_i is integrable, we have

$$\|p_i\|_{w_i,[m(T),1]} \leq \sqrt{C_i(T)}\|p_i\|_{\infty,[m(T),1]}, \quad i = 1, 2, \tag{3.4}$$

where $C_i(T) = \int_{m(T)}^1 \mathrm{d}w_i$, $i = 1, 2$. Moreover, by Remez's inequality,

$$\|p\|_{\infty,[m(T),1]} \leq \|T_N\|_{\infty,[m(T),1]}\|p\|_\infty, \quad \forall p \in \mathbb{P}_N,$$

where $T_N \in \mathbb{P}_N$ is the N^{th} Chebyshev polynomial. Since T_N is monotonic outside $[-1, 1]$, we have $\|T_N\|_{\infty,[m(T),1]} = |T_N(m(T))|$. Moreover, since

$$T_N(x) = \frac{1}{2}\left[\left(x - \sqrt{x^2-1}\right)^n + \left(x + \sqrt{x^2-1}\right)^n\right],$$

an application of (2.15) gives

$$\|T_N\|_{\infty,[m(T),1]} = \frac{1}{2}\left[\tan^{2N}\left(\tfrac{\pi}{4T}\right) + \cot^{2N}\left(\tfrac{\pi}{4T}\right)\right] \leq E(T)^N. \tag{3.5}$$

Next we note that $w_1(u) \geq D_1(T)$ and $w_2(u) \geq D_2(T)\sqrt{1-u^2}$, $\forall u \in [-1,1]$, for positive constants $D_1(T)$ and $D_2(T)$. Moreover, there exist constants $d_1, d_2 > 0$ independent of T such that

$$\|p\|_\infty \leq d_1 N \|p\|, \quad \|p\|_\infty \leq d_2 N^{\frac{3}{2}} \|p\|_v, \quad p \in \mathbb{P}_N,$$

where $v(u) = \sqrt{1-u^2}$ (this follows from expanding p in orthonormal polynomials $\{p_n\}_{n\in\mathbb{N}}$ on $[-1,1]$ corresponding to the weight function $w(u) = 1$, i.e. Legendre polynomials, or $w(u) = v(u)$, i.e. Chebyshev polynomials of the second kind, and using the known estimate $\|p_n\|_\infty = \mathcal{O}(n^{\frac{1}{2}})$ for the former and $\|p_n\|_\infty = \mathcal{O}(n^{\frac{3}{2}})$ for the latter [3, chpt. X]). Therefore

$$\|p\|_\infty \leq \frac{d_i}{D_i(T)} N^{\frac{1+i}{2}} \|p\|_{w_i}, \quad \forall p \in \mathbb{P}_N, \quad i = 1, 2. \tag{3.6}$$

Substituting (3.4), (3.5) and (3.6) into (3.3) now gives

$$\lambda_{\min}(A) \geq \frac{c}{\max\{C_1(T)/D_1(T), C_2(T)/D_2(T)\}^2} N^{-3} E(T)^{-2N},$$

which gives the lower bound in (3.1).

For the upper bound, we set $p_2 = 1$ and $p_1 = T_N$ in (3.3) to give

$$\lambda_{\min}(A) \leq \frac{\|1\|_{w_1}^2 + \|T_N\|_{w_2}^2}{\|1\|_{w_1,[m(T),1]}^2 + \|T_N\|_{w_2,[m(T),1]}^2} \leq \frac{C_1(T) + C_2(T)}{\|T_N\|_{w_2,[m(T),1]}^2}. \tag{3.7}$$

Using (3.5) we note that $\|T_N\|_{\infty,[m(T),1]} \geq \frac{1}{2} E(T)^N$. Recall also that $\|p\|_\infty \leq d_1 N \|p\|, \forall p \in \mathbb{P}_N$. Scaling this inequality to the interval $[m(T), 1]$ now gives

$$\|p\|_{\infty,[m(T),1]} \leq d_1 \sqrt{\frac{2}{1-m(T)}} N \|p\|_{[m(T),1]} = \sqrt{C_3(T)} N \|p\|_{[m(T),1]}.$$

Note also that $w_1(u) \geq D_3(T)$, $\forall u \in [m(T), 1]$. Therefore,

$$\|T_N\|_{w_1,[m(T),1]} \geq \sqrt{D_3(T)} \|T_N\|_{[m(T),1]} \geq \frac{\sqrt{D_3(T)}}{\sqrt{C_3(T)} N} \|T_N\|_{\infty,[m(T),1]} \geq \frac{\sqrt{D_3(T)}}{2\sqrt{C_3(T)} N} E(T)^N.$$

Substituting this into (3.7) now gives the result. □

We now consider the case of the discrete FE:

Theorem 3.2. *Let $\tilde{A}$ be the matrix (2.12) of the discrete FE. Then the condition number of $\tilde{A}$ is $\mathcal{O}\left(E(T)^N\right)$ for large N. Specifically, the maximal and minimal singular values of $\tilde{A}$ satisfy*

$$c_1(T) \leq \sigma_{\max}(\tilde{A}) \leq c_2(T) N^{\frac{3}{2}}, \qquad d_1(T) N^{-\frac{3}{2}} E(T)^{-N} \leq \sigma_{\min}(\tilde{A}) \leq d_2(T) N^{\frac{5}{2}} E(T)^{-N}, \tag{3.8}$$

where $c_1(T), c_2(T), d_1(T), d_2(T)$ are positive constants that are $\mathcal{O}(1)$ as $T \to 1^+$.

Proof. Using Lemma 2.7, the values $\sigma_{\min}^2(\tilde{A})$ and $\sigma_{\max}^2(\tilde{A})$ may be expressed as in (3.2) (with $\|\cdot\|$ replaced by $\|\cdot\|_W$). Note that $W(0)\|\phi\|^2 \leq \|\phi\|_W^2 \leq \|\phi\|_\infty^2 \int_{-1}^1 \mathrm{d}W$. It is a straightforward exercise (using the bound (3.6) and the fact that ϕ can be expressed as the sum of two polynomials) to show that $\|\phi\|_\infty \leq C_1(T) N^{\frac{3}{2}} \|\phi\|$, where $C_1(T) = \mathcal{O}(1)$ as $T \to 1^+$. Thus we obtain

$$W(0) \frac{\|\phi\|^2}{\|\phi\|_{[-T,T]}^2} \leq \frac{\|\phi\|_W^2}{\|\phi\|_{[-T,T]}^2} \leq \left(C_1(T)^2 \int_{-1}^1 \mathrm{d}W \right) N^3 \frac{\|\phi\|^2}{\|\phi\|_{[-T,T]}^2}.$$

The result now follows immediately from the bounds (3.1). □

Remark 3.3 Theorems 3.1 and 3.2 imply that the matrices of the continuous and discrete FE's are ill-conditioned. Note, however, that the discrete FE is substantially better than its continuous counterpart in this regard: the condition number grows only like $E(T)^N$ as opposed to $E(T)^{2N}$. This can be understood using Lemma 2.7: the normal equations of the discrete FE correspond to a continuous FE with a particular weight function, and therefore $\kappa(\tilde{A}) \approx \sqrt{\kappa(A)}$. As discussed in [1, 18], this translates into improved numerical performance (see also §3.3). With the continuous FE, the best achievable accuracy is typically $\mathcal{O}(\sqrt{\epsilon})$, where ϵ is the machine precision used. On the other hand, for the discrete FE the corresponding value is $\mathcal{O}(\epsilon)$. We give a full analysis of this difference in §4.2.

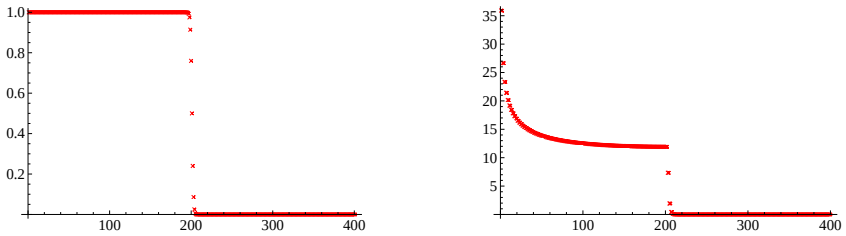


Figure 1: Eigenvalues of the matrices (2.8) (left) and (2.12) (right) for $N = 200$ and $T = 2$.

Since the constants in Theorems 3.1 and 3.2 are bounded as $T \to 1^+$, this allows one to determine the condition number in the case that $T \to 1^+$ as $N \to \infty$ (see §2.3.1). If $T \to 1^+$ sufficiently rapidly, then (up to possibly small algebraic factors in N) $\kappa(A)$ and $\kappa(\tilde{A})$ are $\mathcal{O}(1)$. In particular, when T is given by (2.16), then $\kappa(A)$ and $\kappa(\tilde{A})$ are at worst $\mathcal{O}\left((\epsilon_{\text{tol}})^{-2}\right)$ and $\mathcal{O}\left((\epsilon_{\text{tol}})^{-1}\right)$ respectively.

3.2 The singular value decomposition of A

Although we have now determined the condition number of A, it is actually possible to give a very detailed analysis of its spectrum. This follows from the identification of A with the well-known prolate matrix, which was analysed in detail by Slepian [27, 29]. We now review some of this work.

Using Slepian's notation, let us define the matrix $\rho(N, W) \in \mathbb{C}^{N \times N}$ with entries

$$\rho(N, W)_{m,n} = \begin{cases} \frac{\sin 2\pi W(m-n)}{\pi(m-n)} & m \neq n \\ 2W & m = n, \end{cases} \qquad m, n = 0, \ldots, N-1,$$

where $W < \frac{1}{2}$ is fixed, and write $\lambda_0(N, W) > \ldots > \lambda_{N-1}(N, W) > 0$ for its eigenvalues. Note that

$$\lambda_k(N, \tfrac{1}{2} - W) = 1 - \lambda_{N-1-k}(N, W). \tag{3.9}$$

The following asymptotic results are found in [27]:

(i) For fixed and small k, we have

$$1 - \lambda_k(N, W) \sim \sqrt{\pi}(k!)^{-1} 2^{(14k+9)/4} \alpha^{(2k+1)/4} (2-\alpha)^{-(k+1/2)} N^{k+1/2} \mathrm{e}^{-\gamma N}, \tag{3.10}$$

where $\alpha = 1 - \cos 2\pi W$ and $\gamma = \log\left[\frac{\sqrt{2}+\sqrt{\alpha}}{\sqrt{2}-\sqrt{\alpha}}\right]$.

(ii) For large N and k with $k = \lfloor 2WN(1-\epsilon) \rfloor$ and $0 < \epsilon < 1$, we have $1 - \lambda_k(N, W) \sim \mathrm{e}^{-c_1 - c_2 N}$, for explicitly known constants c_1, c_2 depending only on W and ϵ.

(iii) For large N and k with $k = \lfloor 2WN + (b/\pi) \log N \rfloor$, we have $\lambda_k(N, W) \sim \frac{1}{1+\mathrm{e}^{\pi b}}$

(Slepian also derives similar asymptotic results for the eigenvectors of $\rho(N, W)$ [27]). From these results we conclude that the eigenvalues of the prolate matrix cluster exponentially near 0 and 1 and have a transition region of width $\mathcal{O}(\log N)$ around $k = 2WN$. This is shown in Figure 1.

The matrix A of the continuous FE is precisely $\rho(2N+1, \frac{1}{2T})$. Note that the asymptotic behaviour derived in Theorem 3.1 agrees with that of (3.10): when $W = \frac{1}{2T}$ we have

$$\frac{\sqrt{2}+\sqrt{\alpha}}{\sqrt{2}-\sqrt{\alpha}} = \cot^2\left(\frac{\pi}{4T}\right) = E(T),$$

as expected. Thus, using Slepian's analysis, we see that the eigenvalues of A cluster exponentially at rate $E(T)^2$ near zero and one (note that A corresponds to a prolate matrix of size $2N$), and in particular, the condition number is $\mathcal{O}\left(E(T)^{2N}\right)$, in agreement with Theorem 3.1. We remark, however, that Theorem 3.1 gives explicit bounds for the minimal eigenvalue of A which hold for all N and T (unlike (3.10) which holds only for fixed T and sufficiently large N). In particular, Theorem 3.1 remains valid when T is varied with N, an option which, as discussed in §2.3.1, can be advantageous in practice.

Although the matrix $\tilde{A}$ of the discrete FE is not directly related to A (see Lemma 2.7), we expect a similar structure for the singular values. This is illustrated in Figure 1. Indeed, the only qualitative difference is in the large singular values. The other key features – the narrow transition region and the exponential clustering of singular values near 0 – are much the same.

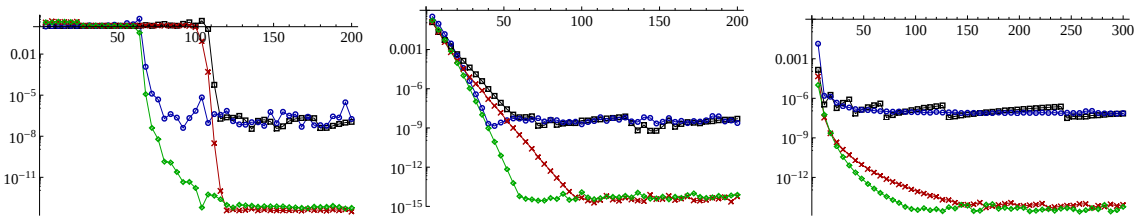


Figure 2: The error $\|f - f_N\|_\infty$, where $f_N = G_N(f)$ (squares and circles) or $f_N = \tilde{G}_N(f)$ (crosses and diamonds) and $T = 2$ (squares/crosses) or $T = T(N; \epsilon_{\text{tol}})$ (circles/diamonds) with $\epsilon_{\text{tol}} = 10^{-14}$.

Remark 3.4 The choice $T = 2$ ($W = \frac{1}{4}$) is special. As shown by (3.9), the eigenvalues $\lambda_k(N, W)$ are symmetric in this case, and the transition region occurs at $k = \frac{1}{2}N$. This is unsurprising. When $T = 2$, the frame $\{\mathrm{e}^{\mathrm{i}\frac{n\pi}{2}x}\}_{n\in\mathbb{Z}}$ decomposes into two orthogonal bases, related to the sine and cosine transforms. Using this decomposition and the associated discrete transforms, M. Lyon has introduced a fast implementation of FE's for equispaced data [21].

3.3 Numerical examples

Having discussed the ill-conditioning of the matrices associated to the continuous and discrete FE's, we now consider several numerical examples. In Figure 2 we plot the error $\|f - f_N\|_\infty$ against N for various choices of f. Here the extension f_N is the the numerically computed continuous or discrete FE – i.e. the result of solving the corresponding linear system in standard precision (recall Remark 1.2). Henceforth, we use the notation $G_N(f)$ and $\tilde{G}_N(f)$ for such extensions, so as to distinguish them from their 'exact' counterparts $F_N(f)$ and $\tilde{F}_N(f)$.

At first sight, Figure 2 appears somewhat surprising: for all three functions we obtain good accuracy, and there is no drift or growth in the error, even in the case where f is nonsmooth or has a complex singularity near $x = 0$. Thus, in practice, the ill-conditioning established in Theorems 3.1 and 3.2 appears to have little effect on the numerical extensions $G_N(f)$ and $\tilde{G}_N(f)$. The purpose of the next section is to explain this apparent contradiction.

In Figure 2 we also compare two choices of T: fixed $T = 2$ and the N-dependent value (2.16) with $\epsilon_{\text{tol}} = 10^{-14}$. Note that the latter typically outperforms the fixed value $T = 2$, especially for oscillatory functions. This is unsurprising in view of the discussion in §2.3.1. Figure 2 also exhibits the disadvantage of the continuous extension described in Remark 3.3: namely, the error levels off at around $\sqrt{\epsilon_{\text{mach}}}$, as opposed to around ϵ_{mach} for the discrete extension. This will be confirmed in the next section by analyzing the numerical method.

4 Numerical stability of Fourier extensions

Figure 2 suggests the following conclusion: although there may be extreme sensitivity in the coefficients of the extensions $G_N(f)$ and $\tilde{G}_N(f)$ (due to the large condition numbers), this has little effect on the extensions themselves. In this section we precisely explain this phenomenon.

4.1 The magnitude of the coefficients

To understand the numerical extensions $G_N(f)$ and $\tilde{G}_N(f)$, it is first necessary to determine the behaviour of the coefficients $\{a_n\}_{n=-N}^N$ of the exact extensions $F_N(f)$ and $\tilde{F}_N(f)$. Specifically, we wish to estimate how large $\|a\|$ can be. We have

Theorem 4.1. *Suppose that f is analytic in $D(\rho^*)$ and not analytic inside any $D(\rho)$ with $\rho > \rho^*$. If $a \in \mathbb{C}^{2N+1}$ is the vector of coefficients of the continuous FE $F_N(f)$, then*

$$\|a\| \leq c_f \begin{cases} \left(\frac{E(T)}{\rho^*}\right)^N & \rho^* < E(T), \\ N & \rho^* \geq E(T), \end{cases} \tag{4.1}$$

where $c_f > 0$ is as in Theorem 2.10. If $f \in \mathrm{L}^2(-1,1)$ is not analytic, then

$$\|a\| \leq c\|f\|E(T)^N. \tag{4.2}$$

Proof. Write $F_N(f) = f_N = f_{e,N} + f_{o,N}$, where $f_{e,N}$ and $f_{o,N}$ are the even and odd parts of f_N respectively. Since the set $\{\phi_n\}_{n\in\mathbb{Z}}$ is orthonormal over $[-T,T]$ we find that

$$\|a\| = \|f_N\|_{[-T,T]} \leq 2\left(\|f_{e,N}\|_{[0,T]} + \|f_{o,N}\|_{[0,T]}\right) \leq 2\sqrt{T}\left(\|f_{e,N}\|_{\infty,[0,T]} + \|f_{o,N}\|_{\infty,[0,T]}\right).$$

Recall from §2.1.1 that $f_{e,N}(x) = h_{1,N}(u)$ and $f_{o,N}(x) = \sin\left(\frac{\pi}{T}m^{-1}(u)\right)h_{2,N}(u)$, where $h_{i,N} \in \mathbb{P}_{N+1-i}$, $i = 1,2$, is defined by (2.4). Thus, $\|a\| \leq c\left(\|h_{1,N}\|_{\infty,[m(T),1]} + \|h_{2,N}\|_{\infty,[m(T),1]}\right)$ for some $c > 0$ independent of N and f. Consider $h_{1,N}(u)$. This is precisely the expansion of h_1 in polynomials $\{p_n\}_{n=0}^{\infty}$ orthogonal with respect to the weight function w_1: i.e. $h_{1,N} = \sum_{n=0}^{N}\langle h_1, p_n\rangle_{w_1}p_n$, where $h_1(u) = f_1(m^{-1}(u))$. Therefore

$$\|h_{1,N}\|_{\infty,[m(T),1]} \leq \sum_{n=0}^{N}|\langle h_1, p_n\rangle_{w_1}|\|p_n\|_{\infty,[m(T),1]}.$$

It is known that $\|p_n\|_{\infty,[m(T),1]} \leq cE(T)^n$ [18]. Also, since h_1 is analytic in $B(\rho')$, where $\rho' = \min\{\rho^*, E(T)\}$, we have that $|\langle h_1, p_n\rangle_{w_1}| \leq c_f(\rho')^{-n}$. Hence

$$\|h_{1,N}\|_{\infty,[m(T),1]} \leq c_f\sum_{n=0}^{N}\left(\frac{E(T)}{\rho'}\right)^n,$$

which gives (4.1). For (4.2) we use the bound $|\langle h_1, p_n\rangle_{w_1}| \leq \|h_1\|_{w_1} \leq c\|f\|$ instead. □

Corollary 4.2. *Let f be as in Theorem 4.1. Then the vector of coefficients $a \in \mathbb{C}^{2N+2}$ of the discrete Fourier extension $\tilde{F}_N(f)$ of f satisfies the same bounds as those given in Theorem 4.1.*

Proof. The functions $h_{i,N}$, $i = 1,2$ are the polynomial interpolants of h_i at the nodes (2.9) (Proposition 2.6). Write $h_{i,N}(u) = \sum_{n=0}^{N}\tilde{u}_nT_n(u)$, where $T_n(u)$ is the n^{th} Chebyshev polynomial, and let $\hat{u}_n = \langle h_i, T_n\rangle_w$ be the exact Chebyshev polynomial coefficient of h_i. Note that $|\hat{u}_n| \leq c_f\rho^{-n}$. Moreover, the aliasing formula gives that $\tilde{u}_n = \hat{u}_n + \sum_{k\neq 0}(\hat{u}_{2kN+n} + \hat{u}_{2kN-n})$ (see [11, Eqn. (2.4.20)]). Therefore,

$$|\tilde{u}_n| \leq c_f\left(\rho^{-n} + \sum_{k=1}^{\infty}\rho^{-2kN-n} + \sum_{k=1}^{\infty}\rho^{-2kN+n}\right) \leq c_f\left(\rho^{-n} + \rho^{n-2N}\right) \leq c_f\rho^{-n}.$$

The result now follows along the same lines as the proof of Theorem 4.1. □

Recall that to compute the continuous or discrete FE we need to solve the linear system $Aa = b$ (respectively $\tilde{A}a = b$). When N is large, the columns of A ($\tilde{A}$) become near-linearly dependent. Indeed, as shown in §3.2, the numerical rank of A is roughly $1/T$ times its dimension for large N. As discussed, this is a direct consequence of the redundancy of the frame.

Now suppose we solve $Aa = b$ with a standard numerical solver. Loosely speaking, the solver will use the extra degrees of freedom to construct approximate solutions a' with small norm. However, the previous theorem and corollary give conditions under which $\|a\|$ is exponentially large in N. Using these, we may now conclude the following: only in the case where f is analytic with $\rho^* \geq E(T)$ can we expect the theoretical coefficient vector a to be produced by the numerical solver for all N. In all other cases, we may well have that $a' \neq a$ for sufficiently large N, and therefore $G_N(f)$ will not coincide with the exact extension $F_N(f)$.

This raises the following question: if the numerical solver does not output the coefficients of $F_N(f)$, then what does it yield? The following proposition confirms the existence of infinitely many approximate solutions of the equations $Aa = b$ with small norm coefficient vectors:

Proposition 4.3. *Suppose that $f \in \mathrm{H}^k(-1,1)$. Then there exist $a^{[N]} \in \mathbb{C}^{2N+1}$, $N \in \mathbb{N}$, satisfying*

$$\|a^{[N]}\| \leq c_k(T)\|f\|_{\mathrm{H}^k(-1,1)}, \tag{4.3}$$

and

$$\|Aa^{[N]} - b\| \leq c_k(T)N^{-k}\|f\|_{\mathrm{H}^k(-1,1)}. \tag{4.4}$$

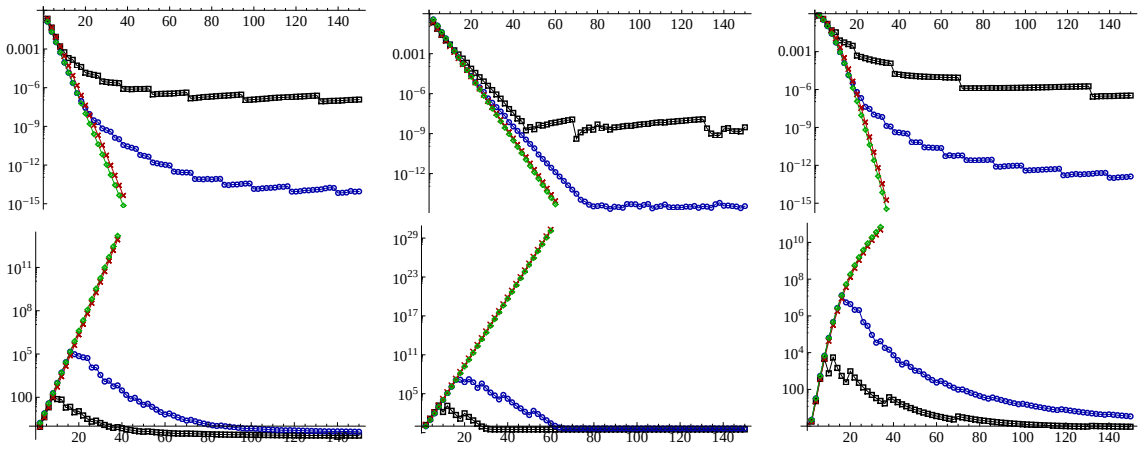


Figure 3: Comparison of the numerical continuous and discrete FE's $G_N(f)$ and $\tilde{G}_N(f)$ (squares and circles) and their exact counterparts $F_N(f)$ and $\tilde{F}_N(f)$ (crosses and diamonds) for $T = 2$. Top row: the uniform error $\|f - f_N\|_\infty$ against N. Bottom row: the norm $\|a\|$ of the coefficient vector.

Moreover, if $g_N = \sum_{|n|\leq N} a_n^{[N]} \phi_n$ *then*

$$\|f - g_N\| \leq c_k(T) N^{-k} \|f\|_{\mathrm{H}^k(-1,1)}. \tag{4.5}$$

Proof. Let $\tilde{f} \in \mathrm{H}^k(\mathbb{T})$ be the extension guaranteed by Lemma 2.4, and write $a^{[N]}$ for the vector of its first $2N + 1$ Fourier coefficients. By Bessel's inequality, $\|a^{[N]}\| \leq \|\tilde{f}\|_{[-T,T]} \leq c_k(T)\|f\|_{\mathrm{H}^k(-1,1)}$ which gives (4.3). For (4.4), we merely note that $(Aa^{[N]} - b)_n = \langle f - g_N, \phi_n \rangle$. Using the frame property (2.5) we obtain $\|Aa^{[N]} - b\| \leq c_2 \|f - g_N\|$. Thus, (4.4) follows directly from (4.5), and the latter is a standard result of Fourier analysis (see [11, eqn. (5.1.10)], for example). □

This proposition states that there exist vectors with norm bounded independently of N which approximately solve the equations $Aa = b$ (up to an error of order N^{-k}). Moreover, these coefficient vectors yield extensions which converge algebraically fast to f at rate k. Whilst it does not imply that these are the coefficient vectors produced by the numerical solver, it does indicate that, in the case where the exact extension $F_N(f)$ has large coefficient norm, exponential convergence of the numerical extension $G_N(f)$ may be sacrificed for spectral convergence so as to retain boundedness of the computed coefficients.

This hypothesis is confirmed in Figure 3 (all computations were carried out in *Mathematica*, with additional precision used to compute the exact FE's and standard precision used otherwise). As we see, exponential convergence of the exact extension is replaced by slower, but still high-order convergence, for sufficiently large N. Note that this 'breakpoint' occurs at roughly the same value of N, regardless of the function. Moreover, the breakpoint occurs at a larger value of N for the discrete extension than the continuous extension.

These observations, as well as the intuitive arguments above, will be confirmed in the next section by an analysis of the numerical extensions $G_N(f)$ and $\tilde{G}_N(f)$. However, let us first make several further comments about the results in Figure 3. First, note that the breakdown of exponential convergence is far less severe for the classical Runge function $f(x) = \frac{1}{1+100x^2}$ than for $f(x) = \frac{1}{8-7x}$ and the entire function $f(x) = 1 + \frac{\cosh 40x}{\cosh 40}$. This can be explained by Proposition 4.3. When $a > 1$ the derivatives of the Runge function $f(x) = \frac{1}{1+a^2x^2}$ are reasonably small, and therefore the approximate coefficient vectors of Proposition 4.3 are also reasonably small in norm for all k. On the other hand, the functions $f(x) = \frac{1}{1+b-bx}$ and $f(x) = 1 + \frac{\cosh bx}{\cosh b}$ have boundary layers near $x = 1$ (also $x = -1$ for the latter). In particular, the k^{th} derivative scales like b^k. Thus, for these functions, approximate solutions corresponding to larger k have much larger norm.

Second, although it is not apparent from Figure 3 that the convergence rate beyond the breakpoint is truly spectral (or merely algebraic of high order), this is in fact the case. Such convergence is confirmed by Figure 4: the slight downward curve in the error indicates spectral convergence.

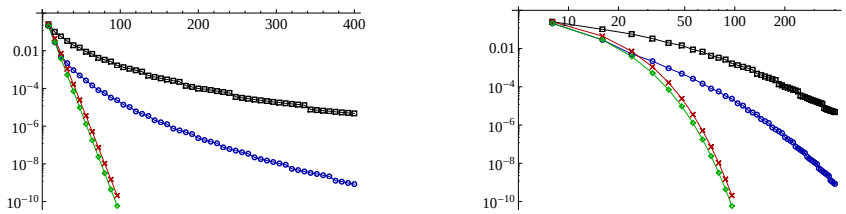


Figure 4: Comparison of the numerical continuous and discrete FE's $G_N(f)$ and $\tilde{G}_N(f)$ (squares and circles) and their exact counterparts $F_N(f)$ and $\tilde{F}_N(f)$ (crosses and diamonds) for $T = 2$ and $f(x) = \frac{1}{101-100x}$. Left: the uniform error in log scale. Right: the error in log-log scale.

4.2 Analysis of the numerical method

We now wish to analyse the numerical extensions $G_N(f)$ and $\tilde{G}_N(f)$ obtained by implementing the continuous and discrete FE's in finite arithmetic. Since the numerical solvers used in environments such as *Matlab* or *Mathematica* are difficult to analyse directly, we now look at the result of solving $Aa = b$ (or $\tilde{A}a = b$) via a truncated singular value decomposition (SVD). This represents an idealization of the numerical solver. Indeed, neither *Matlab*'s \ or *Mathematica*'s `LeastSquares` actually performs a truncated SVD. However, in practice, this simplification appears reasonable: numerical experiments (see later in this section) indicate that these standard solvers give roughly the same results as the truncated SVD with suitably small truncation parameter (typically $\epsilon = 10^{-14}$). We shall also assume throughout that the truncated SVD is computed without error. However, this also seems reasonable: in numerical experiments we observe that the SVD, when computed in finite precision, gives near-identical results to the numerical solver, provided once more that the tolerance is set sufficiently small.

Suppose that A (respectively $\tilde{A}$) has SVD USV^* with $S = \mathrm{diag}(\sigma_0, \ldots, \sigma_{2N})$ being the diagonal matrix of singular values. Given a truncation parameter $\epsilon > 0$, we now consider the solution

$$a_\epsilon = VS^\dagger U^* b, \tag{4.6}$$

where $S^\dagger$ is the diagonal matrix with n^{th} entry $1/\sigma_n$ if $\sigma_n > \epsilon$ and 0 otherwise. We write

$$H_{N,\epsilon}(f) = \sum_{|n|\leq N} (a_\epsilon)_n \phi_n,$$

for the corresponding FE. Let $v_n \in \mathbb{C}^{2N+1}$ be the right singular vector of A with singular value σ_n, and define

$$\Phi_n = \sum_{|m|\leq N} (v_n)_m \phi_m \in \mathcal{G}_N,$$

to be the Fourier series corresponding to v_n. Note that the functions Φ_n are orthonormal with respect to $\langle \cdot, \cdot \rangle_{[-T,T]}$ and span $\mathcal{G}_N$. Also, if we denote $\mathcal{G}_{N,\epsilon} = \mathrm{span}\{\Phi_n : \sigma_n > \epsilon\}$, then we have $H_{N,\epsilon}(f) \in \mathcal{G}_{N,\epsilon}$.

We now consider the cases of the continuous and discrete FE's separately.

4.2.1 The continuous Fourier extension

In this case, since A is Hermitian and positive definite, the singular vector v_n are actually eigenvectors of A, with $Av_n = \sigma_n v_n$. By definition, we have $\langle \Phi_n, \Phi_m \rangle = (v_n)^* A v_m = \sigma_n \delta_{n,m}$, and therefore

$$H_{N,\epsilon}(f) = \sum_{n:\sigma_n > \epsilon} \frac{1}{\sigma_n} \langle f, \Phi_n \rangle \Phi_n. \tag{4.7}$$

Our main result is as follows:

Theorem 4.4. *Let $f \in \mathrm{L}^2(-1,1)$ and suppose that $H_{N,\epsilon}(f)$ is given by (4.7). Then*

$$\|f - H_{N,\epsilon}(f)\| \leq \|f - \phi\| + \sqrt{\epsilon}\|\phi\|_{[-T,T]}, \quad \forall \phi \in \mathcal{G}_N, \tag{4.8}$$

and

$$\|a_\epsilon\| = \|H_{N,\epsilon}(f)\|_{[-T,T]} \leq \frac{1}{\sqrt{\epsilon}}\|f - \phi\| + \|\phi\|_{[-T,T]}, \quad \forall \phi \in \mathcal{G}_N. \tag{4.9}$$

Proof. The function $H_{N,\epsilon}(f)$ is the orthogonal projection of f onto $\mathcal{G}_{N,\epsilon}$ with respect to $\langle\cdot,\cdot\rangle$. Hence for any $\phi \in \mathcal{G}_N$ we have $\|f - H_{N,\epsilon}(f)\| \leq \|f - H_{N,\epsilon}(\phi)\| \leq \|f-\phi\| + \|\phi - H_{N,\epsilon}(\phi)\|$. Consider the latter term. Since $\phi \in \mathcal{G}_N$, the observation that the $\{\Phi_n\}$ are orthonormal on $[-T,T]$ gives

$$\|\phi - H_{N,\epsilon}(\phi)\|^2 = \left\| \sum_{n:\sigma_n<\epsilon} \langle \phi, \Phi_n\rangle_{[-T,T]} \Phi_n \right\|^2 = \sum_{n:\sigma_n<\epsilon} \sigma_n |\langle \phi, \Phi_n\rangle_{[-T,T]}|^2 \leq \epsilon \|\phi\|^2_{[-T,T]}.$$

This yields (4.8). For (4.9) we first write $\|H_{N,\epsilon}(f)\|_{[-T,T]} \leq \|H_{N,\epsilon}(f-\phi)\|_{[-T,T]} + \|H_{N,\epsilon}(\phi)\|_{[-T,T]}$. By orthogonality,

$$\|H_{N,\epsilon}(f-\phi)\|^2_{[-T,T]} = \sum_{n:\sigma_n>\epsilon} \frac{1}{\sigma_n^2} |\langle f-\phi, \Phi_n\rangle|^2 \leq \frac{1}{\epsilon} \sum_{n:\sigma_n>\epsilon} \frac{1}{\sigma_n} |\langle f-\phi, \Phi_n\rangle|^2 = \frac{1}{\epsilon} \|H_{N,\epsilon}(f-\phi)\|^2.$$

Since $H_{N,\epsilon}$ is an orthogonal projection, we conclude that $\|H_{N,\epsilon}(f-\phi)\|^2_{[-T,T]} \leq \frac{1}{\epsilon}\|f-\phi\|^2$, which gives the first term in (4.9). For the second, we notice that

$$\|H_{N,\epsilon}(\phi)\|^2_{[-T,T]} \leq \sum_{n:\sigma_n>\epsilon} |\langle \phi, \Phi_n\rangle|^2 \leq \|\phi\|^2_{[-T,T]},$$

since $\phi \in \mathcal{G}_N$. □

This theorem allows us to explain the behaviour of the numerical FE $G_N(f)$. Suppose that f is analytic in $D(\rho)$, where $\rho < E(T)$ and $D(\rho)$ is as Theorem 2.10. Set $\phi = F_N(f)$ in (4.8), where $F_N(f)$ is the exact continuous FE. Then, using Theorems 2.10 and 4.1, we obtain the bound

$$\|f - H_{N,\epsilon}(f)\| \leq c_f \left(1 + \sqrt{\epsilon} E(T)^N\right) \rho^{-N}.$$

For small N, the first term in the brackets dominates. Thus we expect exponential convergence of $H_{N,\epsilon}(f)$, and therefore also the numerical extension $G_N(f)$, at rate ρ. However, once

$$N > N_0(\epsilon, T) := -\frac{\log \epsilon}{2 \log E(T)}, \tag{4.10}$$

the second term dominates and the bound begins to increase. On the other hand, Proposition 4.3 establishes the existence of functions $\phi \in \mathcal{G}_N$ with bounded coefficients which approximate f to arbitrary orders of accuracy. Substituting such a function ϕ into (4.8) gives

$$\|f - H_{N,\epsilon}(f)\| \leq c_k(T) \left(N^{-k} + \sqrt{\epsilon}\right) \|f\|_{\mathrm{H}^k(-1,1)}, \quad \forall N, k \in \mathbb{N}.$$

Therefore, once $N > N_0(\epsilon, T)$ we expect spectral convergence of $H_{N,\epsilon}(f)$, and consequently $G_N(f)$, down to a maximal achievable accuracy of order $\sqrt{\epsilon}$.

Theorem 4.1 also explains the behaviour of the coefficient norm $\|a_\epsilon\|$. Observe that breakpoint $N_0(\epsilon, T)$ is (up to a small constant) the largest N for which all singular values of A are included in its truncated SVD (see Theorem 3.1). Thus, when $N < N_0(\epsilon, T)$, we have $H_{N,\epsilon}(f) = F_N(f)$, and Theorem 4.1 gives exponential growth of $\|a_\epsilon\|$. On the other hand, once $N > N_0(\epsilon, T)$, we use (4.9) to obtain

$$\|a_\epsilon\| \leq c_k(T) \left(N^{-k}/\sqrt{\epsilon} + 1\right) \|f\|_{\mathrm{H}^k(-1,1)}, \quad \forall N, k \in \mathbb{N}.$$

In particular, for $N > N_0(\epsilon, T)$, we see decay of $\|a_\epsilon\|$ down from its maximal value at $N = N_0(\epsilon, T)$.

This analysis precisely explains the numerical results of the previous section. Note that the maximal achievable accuracy is $\mathcal{O}(\sqrt{\epsilon})$, where ϵ is the tolerance used. This is also shown in Figure 5. Since $N_0(10^{-6}, 2) \approx 4$, $N_0(10^{-10}, 2) \approx 7$, and $N_0(10^{-14}, 2) \approx 9$, this figure also confirms the expression (4.10) for the breakpoint $N_0(\epsilon, T)$.

Figure 5 also demonstrates that the numerical solver (in this case, *Mathematica*'s `LeastSquares`) exhibits very similar behaviour to a truncated SVD with tolerance $\epsilon = 10^{-14}$, as remarked earlier in this section. In particular, for the numerical continuous FE $G_N(f)$, one cannot expect more than 7 digits of accuracy in general. This explains the comments made in Remark 3.3. As we see next, for the discrete FE the corresponding factor in the error bound is ϵ (as opposed to $\sqrt{\epsilon}$), which confirms the significant advantage of the latter approach.

Observe that the breakpoint $N_0(\epsilon, T)$, although derived by assuming f was analytic in a particular region, is actually independent of f. Interestingly, one still witnesses such a breakdown, even when f

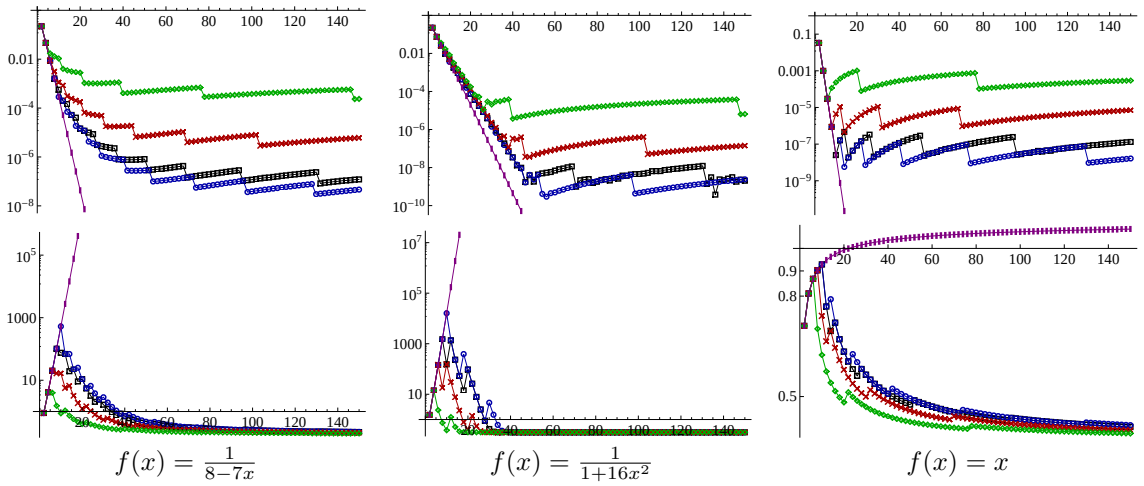


Figure 5: Error (top) and coefficient norm (bottom) against N for the continuous FE with $T = 2$. Squares correspond to the numerical extension $G_N(f)$, circles, crosses and diamonds correspond to the truncated SVD extension $H_{N,\epsilon}(f)$ with $\epsilon = 10^{-14}, 10^{-10}, 10^{-6}$ respectively, and dashes correspond to the exact extension $F_N(f)$. The truncated SVD was computed using additional precision, to avoid the effects of round-off error.

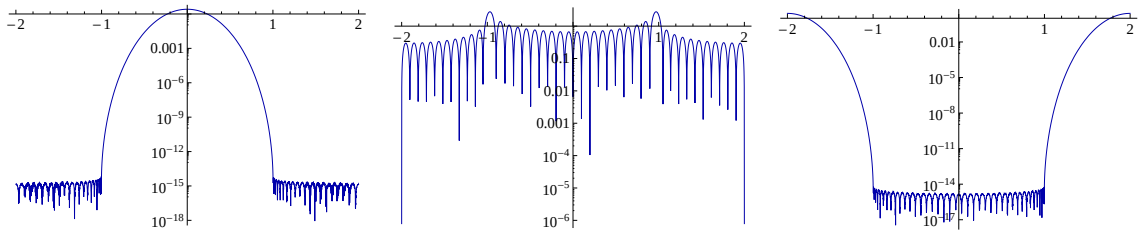


Figure 6: The functions $|\Phi_n(x)|$ for $n = 0$, $n = 20$ and $n = 40$, where $N = 20$ and $T = 2$.

is entire. In Figure 5 we also consider the function $f(x) = x$. The quantity $\|a\|$ initially grows (albeit not exponentially), before decaying slowly once the breakpoint $N_0(T;\epsilon)$ is reached.

This fact is unsurprising. As noted, the breakpoint $N_0(\epsilon, T)$ is the largest N for which $H_{N,\epsilon}(f)$ coincides with $F_N(f)$. Beyond this point, regardless of the function being approximated, we will not obtain $F_N(f)$ from the truncated SVD, and thus we will not in general get further exponential convergence. Having said this, if f is entire, or even analytic in a sufficiently large region, and not possessing large oscillations or derivatives, then once the breakpoint $N_0(\epsilon, T)$ is reached, f will already typically be resolved down to $\mathcal{O}(\sqrt{\epsilon})$. Hence, there is no visible breakdown in exponential convergence.

Remark 4.5 At first sight, it may appear counterintuitive that, by excluding all singular values below a certain tolerance, one can still obtain good accuracy. However, recall that we are not interested in the accuracy of computing a, but rather the accuracy of $F_N(f)$ on the domain $[-1, 1]$. Since the n^{th} singular value σ_n is equal to $\|\Phi_n\|^2$ we see that the functions Φ_n excluded from $H_{N,\epsilon}(f)$ are precisely those for which $\|\Phi_n\|^2 < \epsilon \|\Phi_n\|_{[-T,T]}$. In other words, they have little effect on $F_N(f)$ in $[-1, 1]$.

In Figure 6 we plot the functions Φ_n for several n. As predicted, when n is small, the function Φ_n is large in $[-1, 1]$ and small in $[-T, T]\backslash[-1, 1]$. When n is in the transition region ($n \approx 2N/T$ — see §3.2), the function Φ_n is roughly of equal magnitude in both regions, and for $n \approx 2N$, Φ_n is much smaller in $[-1, 1]$ than on $[-T, T]$. Note also that Φ_n is increasingly oscillatory in $[-1, 1]$ as n increases, and decreasingly oscillatory in $[-T, T]\backslash[-1, 1]$. This follows from the fact that Φ_n has precisely n zeroes in $[-1, 1]$ and $2N - n$ zeroes in $[-T, T]\backslash[-1, 1]$ [27]. Note that this behaviour implies that any 'nice' function will be well approximated by functions Φ_n corresponding to 'nice' eigenvalues, as expected.

4.2.2 The discrete Fourier extension

In this case, we have $(\Phi_n, \Phi_m)_N = \sigma_n^2 \delta_{n,m}$, where

$$(f, g)_N = \frac{\pi}{N+1} \sum_{n=-N-1}^{N} f(x_n)\overline{g(x_n)},$$

is the discrete inner product corresponding to the quadrature nodes $\{x_n\}_{n=-N-1}^{N}$. Therefore

$$\tilde{H}_{N,\epsilon}(f) = \sum_{n:\sigma_n>\epsilon} \frac{1}{\sigma_n^2}(f, \Phi_n)_N \Phi_n \in \mathcal{G}'_{N,\epsilon} := \operatorname{span}\{\Phi_n : \sigma_n > \epsilon\}, \tag{4.11}$$

is the orthogonal projection of f onto $\mathcal{G}'_{N,\epsilon}$ with respect to the discrete inner product $(\cdot,\cdot)_N$.

Theorem 4.6. *Let $f \in \mathrm{L}^\infty(-1,1)$ and $\tilde{H}_{N,\epsilon}(f)$ be given by (4.11). Then*

$$\|f - \tilde{H}_{N,\epsilon}(f)\|_W \leq \|f-\phi\|_W + \sqrt{2\pi S(N;\epsilon)}\|f-\phi\|_\infty + \epsilon\|\phi\|_{[-T,T]}, \quad \forall \phi \in \mathcal{G}_N, \tag{4.12}$$

and

$$\|a_\epsilon\| = \|\tilde{H}_{N,\epsilon}(f)\|_{[-T,T]} \leq \frac{1}{\epsilon}\sqrt{2\pi S(N;\epsilon)}\|f-\phi\|_\infty + \|\phi\|_{[-T,T]}, \quad \forall \phi \in \mathcal{G}_N, \tag{4.13}$$

where $S(N;\epsilon) = |\{n : \sigma_n > \epsilon\}| \leq 2(N+1)$ and W is the weight function of Lemma 2.7.

Proof. By the triangle inequality,

$$\|f - \tilde{H}_{N,\epsilon}(f)\|_W \leq \|f-\phi\|_W + \|\phi - \tilde{H}_{N,\epsilon}(\phi)\|_W + \|\tilde{H}_{N,\epsilon}(f-\phi)\|_W, \quad \forall \phi \in \mathcal{G}'_N.$$

Consider the second term. Since $\phi \in \mathcal{G}'_N$ and the quadrature is exact on $\mathcal{G}'_N$, we have

$$\|\phi - \tilde{H}_{N,\epsilon}(\phi)\|_W^2 = (\phi - \tilde{H}_{N,\epsilon}(\phi), \phi - \tilde{H}_{N,\epsilon}(\phi))_N = \sum_{n:\sigma_n<\epsilon} \sigma_n^2 |\langle \phi, \Phi_n\rangle_{[-T,T]}|^2 \leq \epsilon^2 \|\phi\|_{[-T,T]}^2.$$

For the third term, let g be arbitrary. Then $(\tilde{H}_{N,\epsilon}(g), \tilde{H}_{N,\epsilon}(g))_N = \sum_{n:\sigma_n>\epsilon} \frac{1}{\sigma_n^2}|(g,\Phi_n)_N|^2$. Hence

$$\|\tilde{H}_{N,\epsilon}(g)\|_W^2 = (\tilde{H}_{N,\epsilon}(g), \tilde{H}_{N,\epsilon}(g))_N \leq (g,g)_N \sum_{n:\sigma_n>\epsilon} 1 = (g,g)_N S(N;\epsilon). \tag{4.14}$$

It is straightforward to show that $(g,g)_N \leq 2\pi\|g\|_\infty^2$. Setting $g = f - \phi$ now gives (4.12). For (4.13), we proceed as in the proof of Theorem 4.4. Note that

$$\|\tilde{H}_{N,\epsilon}(g)\|_{[-T,T]}^2 = \sum_{n:\sigma_n>\epsilon} \frac{1}{\sigma_n^4}|(g,\Phi_n)_N|^2 \leq \frac{1}{\epsilon^2}\|\tilde{H}_{N,\epsilon}(g)\|_W^2, \tag{4.15}$$

for any $g \in \mathrm{L}^\infty(-1,1)$. Also,

$$\|\tilde{H}_{N,\epsilon}(\phi)\|_{[-T,T]} \leq \|\phi\|_{[-T,T]}, \quad \phi \in \mathcal{G}_N. \tag{4.16}$$

The result now follows by writing $\|\tilde{H}_{N,\epsilon}(f)\|_{[-T,T]} \leq \|\tilde{H}_{N,\epsilon}(f-\phi)\|_{[-T,T]} + \|\tilde{H}_{N,\epsilon}(\phi)\|_{[-T,T]}$ and using (4.14)–(4.16) with $g = f - \phi$. □

As with the continuous FE, this theorem allows us to analyze the numerical discrete extension $\tilde{G}_N(f)$. Once more we deduce exponential convergence in N up to the function-independent breakpoint $N_1(T;\epsilon) := -\frac{\log\epsilon}{\log E(T)}$, with spectral convergence beyond this point. Note, however, two key differences. First, the bound (4.12) involves ϵ, as opposed to $\sqrt{\epsilon}$, meaning that we expect convergence of $\tilde{G}_N(f)$ down to machine precision (as seen in the experiments of the §4.1). Second, the breakpoint $N_1(T;\epsilon)$ is precisely twice $N_0(T;\epsilon)$. Hence, the regime of exponential convergence of $\tilde{G}_N(f)$ is exactly twice as large as that of the continuous FE $G_N(f)$.

These observations are verified in Figure 7. Note that with the truncated SVD (computed in infinite precision) it is actually possible to get a smaller error than the $\mathcal{O}(\epsilon)$ bound, especially if f does not have large derivatives near the endpoints $x = \pm 1$. However, this effect would obviously be destroyed by roundoff when computing the SVD in finite precision.

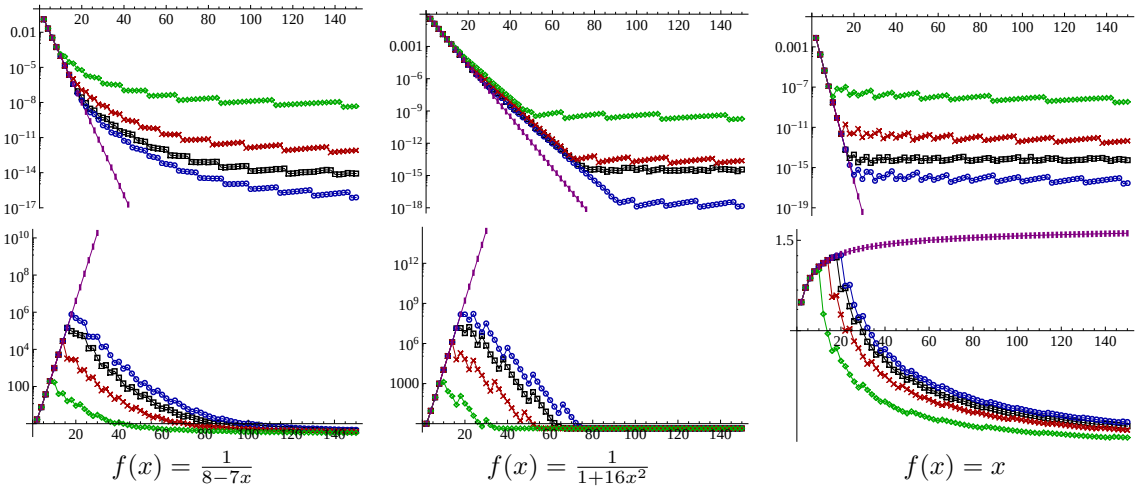


Figure 7: Error (top) and coefficient norm (bottom) against N for the discrete FE with $T = 2$. Squares correspond to the numerical FE $\tilde{G}_N(f)$, circles, crosses and diamonds correspond to the truncated SVD extension $\tilde{H}_{N,\epsilon}(f)$ with $\epsilon = 10^{-14}, 10^{-10}, 10^{-6}$ respectively, and dashes correspond to the exact extension $\tilde{F}_N(f)$.

4.3 The condition number of the numerical method

Having analysed the numerical FE, we next consider its stability with respect to perturbations (e.g. noise). As discussed previously, the condition numbers of the FE matrices (which are exponentially large) are poor predictors for the stability of the FE mappings G_N and $\tilde{G}_N$. Hence we now consider the condition number of the mappings themselves. Proceeding in a standard manner [28], let $F : \mathrm{H} \to \mathrm{H}$ be a mapping taking a vector of inputs $b = b_f$ to an approximation $F(f)$. Here H denotes some function space. We define the condition number $\kappa(F)$ by

$$\kappa(F) = \sup_{f \in \mathrm{H}} \lim_{\delta \to 0} \sup_{\substack{g \in \mathrm{H} \\ 0 < \|b_g\| \leq \delta}} \frac{|||F(f+g) - F(f)|||}{\|b_g\|}, \tag{4.17}$$

where $\|b\|$ is the l^2-norm of the vector b, and $|||\cdot|||$ is a norm on H. Specifically, in the case of the continuous FE, $F = G_N$, $\mathrm{H} = \mathrm{L}^2(-1,1)$, $|||\cdot|||$ is the standard Euclidean norm, and $b = b_f$ has entries given by (2.7). For the discrete extension, $F = \tilde{G}_N$, $\mathrm{H} = \mathrm{L}^2_W(-1,1)$, where the weight function W is as in Lemma 2.7, $|||\cdot||| = \|\cdot\|_W$ is the usual norm on this space, and $b = b_f$ is as defined in §2.2.2.

Note that (4.17) gives the *absolute* condition number of F, as opposed to the more standard *relative* condition number [28]. The key results proved later in this section can easily be reformulated for the latter. However, since the work of [24] is relevant to what follows (§5, in particular), we shall continue to use (4.17), which coincides with the definition used therein. We remark in passing that $\kappa(F)$ measures the absolute sensitivity of the mapping F to perturbations of the inputs b_f. Specifically, one has

$$|||f - F(f+g)||| \leq |||f - F(f)||| + \kappa(F)\|b_g\|.$$

Thus, perturbations are magnified by at most $\kappa(F)$ times their amplitude. This aside, it is also worth noting that, since all the methods considered in this paper are linear, the condition number (4.17) reduces to

$$\kappa(F) = \sup_{\substack{f \in \mathrm{H} \\ b_f \neq 0}} \frac{|||F(f)|||}{\|b_f\|}.$$

As in the previous section, we analyse the condition number for the numerical FE's G_N and $\tilde{G}_N$ by considering the truncated SVD extensions $H_{N,\epsilon}$ and $\tilde{H}_{N,\epsilon}$. Our main result is as follows:

Theorem 4.7. *Let $F = H_{N,\epsilon}$ be the continuous truncated SVD FE given by (4.7). Then*

$$\kappa(H_{N,\epsilon}) = \frac{1}{\min\{\sqrt{\sigma_n} : \sigma_n > \epsilon\}} \leq \min\left\{\frac{1}{\sqrt{\epsilon}}, C(T)N^{\frac{3}{2}}E(T)^N\right\}, \quad N \in \mathbb{N},\ \epsilon > 0.$$

where $C(T)$ is a positive constant independent of N. Conversely, if $F = \tilde{H}_{N,\epsilon}$ is the discrete extension (4.11), then $\kappa(\tilde{H}_{N,\epsilon}) = 1$ for all $N \in \mathbb{N}$ and $\epsilon > 0$.

Proof. Consider the continuous FE first and let g be arbitrary. By definition,

$$\|H_{N,\epsilon}(f)\|^2 = \sum_{n:\sigma_n>\epsilon} \frac{1}{\sigma_n}|\langle g, \Phi_n\rangle|^2 \leq \frac{1}{\min\{\sigma_n : \sigma_n > \epsilon\}} \sum_{n=0}^{2N} |\langle g, \Phi_n\rangle|^2. \tag{4.18}$$

Note that $\langle g, \Phi_n\rangle = \sum_{m=0}^{2N} \langle g, \phi_n\rangle (v_n)_m = \langle b_g, v_n\rangle$. By orthogonality of the singular vectors v_n, we therefore have that $\sum_{n=0}^{2N} |\langle g, \Phi_n\rangle|^2 = \|b_g\|^2$. Recall that the minimal singular value of A is bounded by $c(T)N^{-3}E(T)^{2N}$ (Theorem 3.1). Hence, using (4.18), we deduce that

$$\frac{\|H_{N,\epsilon}(g)\|}{\|b_g\|} \leq \min\left\{\frac{1}{\sqrt{\epsilon}}, c(T)N^{\frac{3}{2}}E(T)^N\right\},$$

as required.

For the discrete FE, note that $(g, \Phi_n)_N = \langle b_g, \tilde{A}v_n\rangle = \sigma_n\langle b_g, u_n\rangle$, where u_n is the corresponding left singular vector of $\tilde{A}$. Therefore

$$\|b_g\|^2 = \sum_{n=0}^{2N} \frac{1}{\sigma_n^2}|(g, \Phi_n)_N|^2 \geq \sum_{\sigma_n>\epsilon} \frac{1}{\sigma_n^2}|(g, \Phi_n)_N|^2 = \|\tilde{H}_{N,\epsilon}(g)\|_W^2,$$

which gives the result. Finally, we note that the bounds for both the continuous and discrete extension hold with equality since we can take b_g to be the corresponding singular vector. □

This theorem has some interesting consequences. First, the discrete FE is perfectly stable! Second, the continuous FE is, as one might expect, far from stable. Indeed, the condition number grows exponentially fast at rate $E(T)$ until it reaches the level $\frac{1}{\sqrt{\epsilon}}$, where ϵ is the truncation parameter in the SVD. Thus, with the continuous FE we may see perturbations being magnified by a factor of $\frac{1}{\sqrt{\epsilon_{\text{mach}}}} \approx 10^8$ in practice.

We also note another implication of Theorem 4.7: varying T has no substantial effect on stability (for sufficiently large N in the case of the continuous extension). Although the condition number of the FE matrices depends on T (recall Theorems 3.1 and 3.2), the condition numbers of the numerical mappings G_N and $\tilde{G}_N$ are actually independent of this parameter.

Much as in the previous section, it is important to confirm that the results of this theorem on the condition number of the truncated SVD extensions correspond closely to the behaviour of the numerical extensions G_N and $\tilde{G}_N$. Unfortunately, $\kappa(G_N)$ and $\kappa(\tilde{G}_N)$ cannot be computed. However, they can be approximated by taking repeated draws of randomly-chosen input vectors b. We therefore define

$$\kappa_t(F) = \max_{j=1,\dots,t} \frac{\|F(b_j)\|}{\|b_j\|}, \tag{4.19}$$

where each b_j, $j = 1, \dots, t$, is the realization of a vector b whose entries are independent uniformly distributed random variables taking values in $[-1, 1]$ (for the continuous extension we allow complex values of b) and the parameter t is the number of trials.

In Figure 8 we plot $\kappa_t(G_N)$ and $\kappa_t(\tilde{G}_N)$ for various choices of N. As we see, the discrete FE is extremely stable: not only is there no blowup with N, the value of $\kappa_t(\tilde{G}_N)$ is very close in magnitude to 1, indicating that $\kappa(\tilde{G}_N) \approx 1$ in this case. For the continuous extension, $\kappa_t(G_N)$ initially grows exponentially, before levelling off at around $10^8 \approx \sqrt{\epsilon_{\text{mach}}}$ in magnitude. This behaviour is in good agreement with Theorem 4.7.

The difference in stability between the continuous and discrete FE's is highlighted in Figure 9. Here we perturbed the right-hand side b of the function $f(x) = \mathrm{e}^x$ by noise of magnitude δ, and then computed its FE. As is evident, the discrete extension approximates f to an error of magnitude roughly δ, whereas for the continuous extension the error is of magnitude $\approx 10^8\delta$, as predicted by Theorem 4.7.

Remark 4.8 The disparity between the condition number of the coefficients of the FE (i.e. the condition number of FE matrix) and that of the mapping can be explained by intuitive arguments. Perturbations η in the input $b = b_f$ are magnified in the FE coefficients if η has a large component corresponding to small singular vectors v_n. However, since the singular functions Φ_n are small on $[-1, 1]$ (Remark 4.5), any magnification in the FE coefficients is cancelled out in the extension itself.

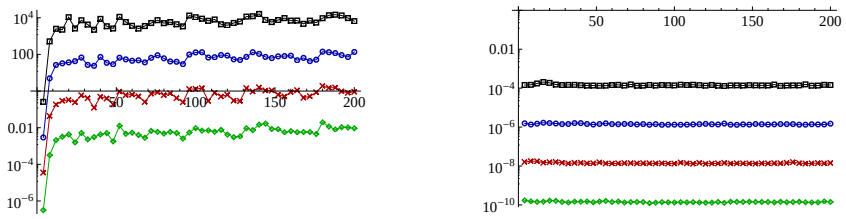


Figure 8: The function $\delta\kappa_t(F)$ for the continuous (left) and discrete (right) FE's G_N and $\tilde{G}_N$ respectively against $N = 1, \ldots, 200$, where $\delta = 10^{-4}, 10^{-6}, 10^{-8}, 10^{-10}$ (squares, circles, crosses and diamonds).

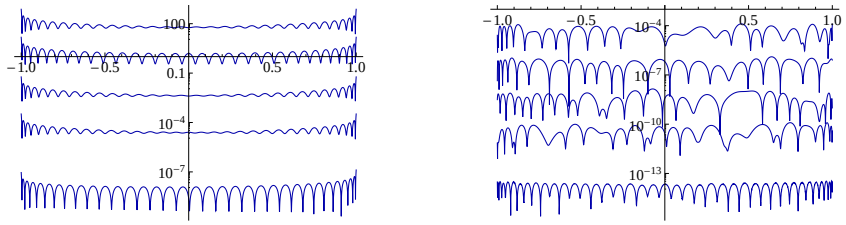


Figure 9: The error $|f(x) - f_N(x)|$ against x, where $f_N = G_N(f)$ (left) or $f_N = \tilde{G}_N(f)$ (right), for $N = 30$, $T = 2$ and $f(x) = \mathrm{e}^x$, with noise at amplitudes $\delta = 10^{-4}, 10^{-6}, 10^{-8}, 10^{-10}, 0$.

5 Fourier extensions from equispaced data

Having analyzed the stability of the continuous and discrete FE's, we now turn our attention to the problem of computing FE's when only equispaced data is prescribed.

As discussed in §1, one feature common to all rapidly convergent methods for overcoming the Runge phenomenon is ill-conditioning. This observation was explained by Platte, Trefethen & Kuijlaars in [24], wherein it was proved that any exponentially convergent method for this problem must also be exponentially ill-conditioned. The presence of such poor conditioning in an, albeit rapidly convergent, numerical method may appear highly disadvantageous. Indeed, at first sight at least, it is reasonable to expect that any such method must be highly susceptible to noise and round-off error, and therefore not useful in practice. However, it was noted in [24] that some methods for this problem do seemingly exhibit both high accuracy and numerical stability. The purpose of the remainder of this paper is to demonstrate that FE's give rise to such a method, the so-called *equispaced Fourier extension*, and explain precisely how this relates the stability barrier of Platte, Trefethen & Kuijlaars.

5.1 The equispaced Fourier extension

Let

$$x_n = \frac{n}{M}, \quad n = -M, \ldots, M, \tag{5.1}$$

be a set of $2M + 1$ equispaced points in $[-1, 1]$, where $M \geq N$. We define the *equispaced Fourier extension* of a function $f \in \mathrm{L}^\infty[-1, 1]$ by

$$F_{N,M}(f) := \operatorname*{argmin}_{\phi \in \mathcal{G}_N} \sum_{|n| \leq M} |f(x_n) - \phi(x_n)|^2. \tag{5.2}$$

Observe that if $F_{N,M}(f) = \sum_{|n| \leq N} a_n \phi_n$, then the vector $a = (a_{-N}, \ldots, a_N)^\top$ is the least squares solution to $\bar{A}a \approx b$, where $\bar{A} \in \mathbb{C}^{(2M+1)\times(2N+1)}$ has $(n, m)^{\text{th}}$ entry $\phi_m(x_n)$ and $b = \{f(x_n)\}_{|n| \leq M}$.

Note that $F_{N,M}(f)$, as defined by (5.2), is (up to minor changes of parameters/notation) identical to the extensions considered in the previous papers [7, 8, 10, 20, 21] on equispaced FE's.

5.2 Theory of the equispaced Fourier extension

Consider first the case $M = N$. Then $F_{N,N}(f)$ is equivalent to polynomial interpolation in u:

Proposition 5.1. *Let $F_{N,N}(f) = f_N = f_{e,N} + f_{o,N} \in \mathcal{G}'_N$ be defined by (5.2) with $N = M$ and let $h_{i,N}(u) \in \mathbb{P}_N$ be given by (2.4). Then $h_{i,N}(u)$, $i = 1, 2$ is the $(N+1-i)^{\text{th}}$ degree polynomial interpolant of $h_i(u)$ at the nodes $\{u_n\}_{n=i-1}^N \subseteq [-1, 1]$, where*

$$u_n = m(x_n) = 2\frac{\cos\left(\frac{n\pi}{NT}\right) - c(T)}{1 - c(T)} - 1, \quad n = 0, \ldots, N. \tag{5.3}$$

This proposition allows us to analyze the theoretical convergence/divergence of $F_{N,N}(f)$ using standard results on polynomial interpolation. Recall that associated to a set of nodes $\{u_n\}_{n=0}^N$ is a *node density function* $\mu(u)$, i.e. a function such that (i) $\int_{-1}^1 \mu(u)\,\mathrm{d}u = 1$ and (ii) each small interval $[u, u+h]$ contains a total of $N\mu(u)h$ nodes for large N [16]. In the case of (5.3) we have

Lemma 5.2. *The nodes (5.3) have node density function $\mu(u) = c\frac{1}{\sqrt{(1-u)(u-m(T))}}$, where $c^{-1} = 2\arctan\left(\frac{\sqrt{2}}{\sqrt{-1-m(T)}}\right)$.*

Proof. Note first that $\int_{-1}^1 \mu(u)\,\mathrm{d}u = 1$. Now let $I = [u, u+h] \subseteq [-1, 1]$ be an interval. Then the node $u_n \in I$ if and only if $m^{-1}(u+h) \leq x_n \leq m^{-1}(u)$. Therefore, as $N \to \infty$, the proportion of nodes lying in I tends to $m^{-1}(u) - m^{-1}(u+h)$. Now suppose that $h \to 0$. Then

$$m^{-1}(u+h) = \frac{T}{\pi}\arccos\left[c(T) + \frac{1-c(T)}{2}(u+h+1)\right] = m^{-1}(u) - \mu(u)h + \mathcal{O}\left(h^2\right).$$

Thus $m^{-1}(u) - m^{-1}(u+h) = \mu(u)h + \mathcal{O}\left(h^2\right)$, as required. □

It is useful to consider the behaviour of $\mu(u)$. Near $u = 1$, $\mu(u) \sim c\frac{1}{\sqrt{1-u}}$. On other hand, μ is continuous at $u = -1$ with $\mu(-1) = c\frac{1}{\sqrt{2(-1-m(T))}}$. Hence the nodes $\{u_n\}_{n=0}^N$ cluster quadratically near $u = 1$ and are linearly distributed near $u = -1$. It is well known that to avoid the Runge phenomenon in a polynomial interpolation scheme, it is essentially necessary that the nodes cluster quadratically near both endpoints (as is the case with Chebyshev nodes) [16]. If this is not the case, one expects the Runge phenomenon: that is, divergence (at exponential rate) in the interpolant at some points $u \in [-1, 1]$ for any function having a singularity in a certain complex region containing $[-1, 1]$ (the *Runge region* for the interpolation scheme). Since the nodes (5.3) do not exhibit the correct clustering, we therefore expect this behaviour in the equispaced FE $F_{N,N}(f)$.

As it transpires, the corresponding Runge region $R = R(T)$ for $F_{N,N}(F)$ can actually be defined in terms of the potential function $\phi(z) = -\int_{-1}^1 \mu(u)\log|z-u|\,\mathrm{d}u + c$ (here c is an arbitrary constant). Standard polynomial interpolation theory [16] then gives that

$$R(T) = \{x \in \mathbb{C} : \phi(m(x)) = \phi(-1)\},$$

(observe that this is a subset of the complex x-plane). We note also that the convergence/divergence of $F_{N,N}(f)$ at a point x will be exponential at a rate $\mathrm{e}^{\phi(m(x_0))-\phi(m(x))}$, where x_0 is the limiting singularity of f (this follows from a general result on polynomial interpolation [16]). In particular, if f has a singularity in $R(T)$, then there will be some points $x \in [-1, 1]$ for which $F_{N,N}(f)$ diverges.

We next discuss two approaches to overcome the Runge phenomenon in $F_{N,N}(f)$.

5.2.1 Overcoming the Runge phenomenon

One way to attempt to overcome (or, at least, mitigate) the Runge phenomenon is to vary the parameter T. We observe the following:

Lemma 5.3. *The Runge region $R(T)$ satisfies $R(T) \to [-1, 1]$ as $T \to 1^+$, and $R(T) \to R$ as $T \to \infty$, where R is the Runge region for equispaced polynomial interpolation.*

This lemma comes as no surprise (we omit the proof for brevity's sake). As $T \to 1^+$, the system $\{\mathrm{e}^{\mathrm{i}\frac{n\pi}{T}\cdot}\}_{|n|\leq N}$ tends to the standard Fourier basis on $[-1, 1]$. The problem of equispaced interpolation with trigonometric polynomials is well-conditioned and convergent. On the other hand, when $T \to \infty$ for fixed N, the subspaces $\mathcal{C}_N$ and $\mathcal{S}_N$ both resemble spaces of algebraic polynomials in x. Thus, in the large T limit, $F_{N,N}(f)$ is an algebraic polynomial interpolant of f at equispaced nodes.

Since the Runge region $R(T)$ can be made arbitrarily small by letting $T \to 1^+$, one way to overcome the Runge phenomenon is to vary T in the way described in §2.3.1 and set $T = T(N;\epsilon)$. Note that, since $T(N;\epsilon) \sim 1$ for large N (see (2.17)), this approach will not suffer from a Runge phenomenon. One could also take $T \approx 1$ fixed. However, this will always lead to a nontrivial Runge region, and therefore there will always be divergence for some functions.

An alternative way to overcome the hypothesized Runge phenomenon is to oversample. In other words, allow $M \geq N$. Oversampling is well known to defeat the Runge phenomenon in equispaced polynomial interpolation [5, 8, 24], and the same is true here (see [7, 10] for previous discussions on oversampling for equispaced FE's). The key theorem is as follows (for brevity we shall omit the proof – a very similar argument is given in [5] for the case of polynomial interpolation):

Theorem 5.4. *Let $F_{N,M}(f)$ be given by (5.2), and suppose that*

$$D(N,M) = \sup\left\{\|\phi\| : \phi \in \mathcal{G}_N,\ \|\phi\|_M = 1\right\}, \tag{5.4}$$

where $\|\phi\|_M^2 = \frac{1}{M+\frac{1}{2}} \sum_{|n|\leq M} |\phi(x_n)|^2$ and $\{x_n\}_{n=0}^{2M+1}$ are the nodes (5.1). Then

$$\|f - F_{N,M}(f)\| \leq (1 + 2D(N,M)) \inf_{\phi\in\mathcal{G}_N} \|f-\phi\|_\infty.$$

This theorem implies that the FE based on equispaced nodes will converge, regardless of the analyticity of f, provided M is chosen such that $D(N,M)$ is bounded. Up to possible small algebraic factors in M and N, the quantity $D(N,M)$ is equivalent to

$$\tilde{D}(N,M) = \sup\left\{\|p\|_\infty : p \in \mathbb{P}_N,\ |p(u_n)| \leq 1,\ n = 0,\dots,M\right\}. \tag{5.5}$$

Note the meaning of this quantity: it informs us how large a polynomial of degree N can be on $[-1,1]$, if that polynomial is bounded at the M points u_n. Unfortunately, numerical evidence suggests that

$$a^{\frac{N^2}{M}} \leq \tilde{D}(N,M) \leq b^{\frac{N^2}{M}}. \tag{5.6}$$

for constants $a, b > 1$. Thus one requires $M = \mathcal{O}\left(N^2\right)$ nodes to ensure boundedness of $D(N,M)$. This is clearly less than ideal: it means that we require many more samples of f to compute its N-term equispaced FE, and, in particular, the FE (5.2) converges only root-exponentially fast in the number of given grid values of f.

Had the nodes $\{u_n\}_{n=0}^M$ clustered quadratically near $u = \pm 1$, then $M = \mathcal{O}(N)$ would be sufficient to ensure boundedness of $\tilde{D}(N,M)$ (note that when $N = M$, $\tilde{D}(N,M)$ is precisely the Lebesgue constant of polynomial interpolation). On the other hand, if $\{u_n\}_{n=0}^M$ were equispaced nodes on $[-1,1]$ then (5.6) would coincide with a well-known result of Coppersmith & Rivlin [13]. The intuition for a bound of the form (5.6) for the nodes (5.3) comes from the fact that these nodes are linearly distributed near $u = -1$. Thus, at least near $u = -1$ they behave like equispaced nodes.

Since the scaling $M = \mathcal{O}\left(N^2\right)$ is undesirable, one can ask what happens when $M = \gamma N$ for some fixed oversampling parameter $\gamma \geq 1$. Using potential theory arguments, one can show that $\tilde{D}_{N,\gamma N}$ grows exponentially in N (with the constant of this growth becoming smaller as γ increases), as predicted by the conjectured bound (5.6). We shall write $c(\gamma;T)$ for such a constant: i.e. $D_{N,\gamma N} = c(\gamma;T)^N$.[1] Incidentally, this also proves that linear oversampling is not sufficient for theoretical convergence.

In view of this behaviour, Theorem 5.4 guarantees convergence of the FE (5.2), provided $\rho \geq c(\gamma;T)$, where ρ is as in Theorem 2.10. In other words, f needs to be analytic in the region $D(c(\gamma;T))$ (recall D from Theorem 2.10) to ensure convergence, and therefore one expects a Runge phenomenon whenever f has a complex singularity lying in the corresponding Runge region $R(\gamma;T) = D(c(\gamma;T))$. Naturally, a larger value of γ leads to a smaller (but still nontrivial) Runge region. However, regardless of the choice of γ, there will always be analytic functions for which one expects divergence of $F_{N,\gamma N}(f)$ (see [5] for a related discussion in the case of equispaced polynomial interpolation).

5.2.2 Numerical examples

In summary, to obtain a convergent FE using equispaced data it appears that one either needs to oversample quadratically (and thereby reduce the convergence rate to only root-exponential), or scale

[1] The constant of growth was obtained in private communication with A. Kuijlaars. A closed expression (up to several integrals involving the potential function ϕ for the nodes u_n) can be found for $c(\gamma;T)$. We omit the full argument as it is rather lengthy, but note that it is based on standard results in potential theory. A general reference is [25].

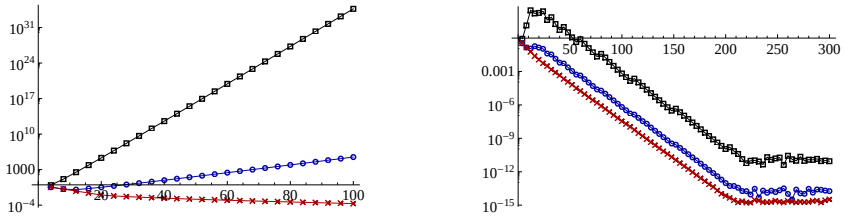


Figure 10: The error $\|f - f_N\|_\infty$ against N for the equispaced FE's $f_N = F_{N,\gamma N}(f)$ (left) and $f_N = G_{N,\gamma N}(f)$ (right) of $f(x) = \frac{1}{1+100x^2}$ with oversampling factor $\gamma = 1, 2, 4$ (squares, circles and crosses) and $T = 2$.

the extension parameter suitably with N or both. However, recall from §4 that an FE obtained from a computation carried out in finite precision may differ quite dramatically from the exact, infinite precision extension. Is it therefore possible that the unpleasant effects described in the previous section may not be witnessed in finite precision? The answer transpires to be yes, and consequently FE's can be used for equispaced data, even in situations where theoretical divergence is expected.

To illustrate, consider the approximation of the function $f(x) = \frac{1}{1+100x^2}$. When $T = 2$, this function has a singularity lying in the Runge region $R(1; 2)$, and so we expect divergence of its equispaced FE. This is shown in Figure 10. Note that double oversampling also gives divergence, whilst with quadratic oversampling the singularity of f no longer lies in $R(\gamma; T)$, and thus we witness exponential convergence (although at a very slow rate).

However, the situation changes completely when we carry out computations in finite precision. In all cases, the approximation, which we denote $G_{N,M}(f)$, converges exponentially fast, and there is no drift in the error once the best achievable accuracy is attained. Note that oversampling by a constant factor improves the approximation, but, this aside, in all cases we still witness convergence.

This figure suggests the following conclusion: equispaced FE's are theoretically unstable and divergent, but numerically stable and convergent in finite arithmetic. In the next section we explain this apparent contradiction.

Before doing so, we remark that the condition number $\kappa(\bar{A})$ of the matrix $\bar{A}$ is always exponentially large. Indeed, suppose that we write

$$\sigma_{\min}(\bar{A})^{-1} = \sup\left\{\|\phi\|_{[-T,T]} : \phi \in \mathcal{G}_N, \|\phi\|_M = 1\right\} = B(N, M). \tag{5.7}$$

Then, up to small algebraic factors in M and N, the quantity $B(N, M)$ is equivalent to

$$\tilde{B}(N, M) = \sup\left\{\|p\|_{\infty,[m(T),1]} : p \in \mathbb{P}_N, |p(u_n)| \leq 1, n = 0, \ldots, M\right\}, \tag{5.8}$$

(this is analogous to $D(N, M)$ and $\tilde{D}(N, M)$ – see §5.2.1). From this, it is a straightforward exercise to show that $E(T)^N \leq B(N, M) \leq \tilde{D}(N, M)E(T)^N$, and therefore

$$E(T)^N \lesssim \kappa(\bar{A}) \lesssim D(N, M)E(T)^N,$$

which implies that $\kappa(\bar{A})$ is exponentially large in N, regardless of M. Consequently, one expects extreme sensitivity in the coefficients of the equispaced FE. However, in the next section we show that this does not lead to significant instability in the equispaced FE itself, much as in the case of the continuous and discrete FE's.

5.3 Analysis of the numerical method

Proceeding as in §4.2 we now analyze the truncated SVD approximation, which we denote $H_{N,M,\epsilon}(f)$. Note that a similar analysis has also recently been presented in [20] – see Remark 5.9 for further details.

Let $\Phi_n \in \mathcal{G}_N$ be the function corresponding to singular value σ_n of the matrix $\bar{A}$. Write $\mathcal{G}_{N,M,\epsilon} = \text{span}\{\Phi_n : \sigma_n > \epsilon\}$ and $\mathcal{G}^{\perp}_{N,M,\epsilon} = \text{span}\{\Phi_n : \sigma_n < \epsilon\}$, and let $(\cdot,\cdot)_M$ be the discrete bilinear form $(g, h)_M = \frac{1}{M+1}\sum_{n=0}^{2M+1} g(x_n)\overline{h(x_n)}$, with corresponding discrete semi-norm $\|\cdot\|_M$. Observe that $H_{N,M,\epsilon}$ is the orthogonal projection onto $\mathcal{G}_{N,M,\epsilon}$ with respect to $(\cdot,\cdot)_M$. Since $(\Phi_n, \Phi_m)_M = \sigma_n^2\delta_{n,m}$, we have

$$H_{N,M,\epsilon}(f) = \sum_{n:\sigma_n>\epsilon} \frac{1}{\sigma_n^2}(f, \Phi_n)_M\Phi_n. \tag{5.9}$$

Our main result is as follows:

Theorem 5.5. *Let $f \in \mathrm{L}^\infty(-1,1)$ and $H_{N,M,\epsilon}(f)$ be given by (5.9). Then*

$$\|f - H_{N,M,\epsilon}(f)\| \leq \|f-\phi\| + \sqrt{2}C_1(N,M;T,\epsilon)\|f-\phi\|_\infty + C_2(N,M;T,\epsilon)\|\phi\|_{[-T,T]}, \quad \forall \phi \in \mathcal{G}_N, \tag{5.10}$$

and

$$\|a_\epsilon\| = \|H_{N,M,\epsilon}(f)\|_{[-T,T]} \leq \frac{\sqrt{2}}{\epsilon}\|f-\phi\|_\infty + \|\phi\|_{[-T,T]}, \quad \forall \phi \in \mathcal{G}_N, \tag{5.11}$$

where

$$C_1(N,M;T,\epsilon) = \sup_{\substack{\phi \in \mathcal{G}_{N,M,\epsilon} \\ \phi \neq 0}} \left\{ \frac{\|\phi\|}{\|\phi\|_M} \right\}, \quad C_2(N,M;T,\epsilon) = \sup_{\substack{\phi \in \mathcal{G}_{N,M,\epsilon}^\perp \\ \phi \neq 0}} \left\{ \frac{\|\phi\|}{\|\phi\|_{[-T,T]}} \right\}. \tag{5.12}$$

Proof. Let $\phi \in \mathcal{G}_N$. Then

$$\|f - H_{N,M,\epsilon}(f)\| \leq \|f-\phi\| + \|H_{N,M,\epsilon}(f-\phi)\| + \|\phi - H_{N,M,\epsilon}(\phi)\|. \tag{5.13}$$

Consider the second term. By definition of $C_1(N,M;T,\epsilon)$,

$$\|H_{N,M,\epsilon}(f-\phi)\| \leq C_1(N,M,\epsilon)\|H_{N,M,\epsilon}(f-\phi)\|_M \leq C_1(N,M,\epsilon)\|f-\phi\|_M,$$

where the second inequality follows from the fact that $H_{N,M,\epsilon}$ is an orthogonal projection with respect to $(\cdot,\cdot)_M$. Noting that $\|g\|_M \leq \sqrt{2}\|g\|_\infty$ for any function g now gives the corresponding term in (5.10). The bound for the third term of (5.13) follows immediately from the definition of $C_2(N,M;T,\epsilon)$ and the inequality $\|\phi - H_{N,M,\epsilon}(\phi)\|_{[-T,T]} \leq \|\phi\|_{[-T,T]}$.

For (5.11), we first write $\|H_{N,M,\epsilon}(f)\|_{[-T,T]} \leq \|H_{N,M,\epsilon}(f-\phi)\|_{[-T,T]} + \|H_{N,M,\epsilon}(\phi)\|_{[-T,T]}$. Observe that, for any g, we have

$$\|H_{N,M,\epsilon}(g)\|_{[-T,T]}^2 = \sum_{n:\sigma_n > \epsilon} \frac{1}{\sigma_n^4}|(g,\Phi_n)_M|^2 \leq \frac{1}{\epsilon^2}\|H_{N,M,\epsilon}(g)\|_M^2 \leq \frac{1}{\epsilon^2}\|g\|_M^2 \leq \frac{2}{\epsilon^2}\|g\|_\infty^2.$$

Also, $\|H_{N,M,\epsilon}(\phi)\|_{[-T,T]} \leq \|\phi\|_{[-T,T]}$ for $\phi \in \mathcal{G}_N$. Combining these two bounds, and setting $g = f - \phi$, now gives (5.11). □

An immediate corollary of this theorem shows that the truncated SVD equispaced Fourier extension $H_{N,M,\epsilon}$ does not suffer from a Runge phenomenon:

Corollary 5.6. *If $f \in \mathrm{L}^\infty(-1,1)$ then*

$$\|H_{N,M,\epsilon}(f)\| \leq \frac{\sqrt{2}}{\epsilon}\|f\|_\infty, \quad \forall N \in \mathbb{N},\ M \geq N.$$

Moreover, if $f \in \mathrm{H}^1(-1,1)$, then

$$\limsup_{\substack{N,M\to\infty \\ M \geq N}} \|H_{N,M,\epsilon}(f)\| \leq \inf\left\{ \|\tilde{f}\|_{[-T,T]} : \tilde{f} \in \mathrm{H}^1(\mathbb{T}),\ \tilde{f}|_{[-1,1]} = f \right\} \leq c_1(T)\|f\|_{\mathrm{H}^1(-1,1)},$$

where $\mathbb{T} = [-T,T)$ is the torus and $c_1(T) > 0$ is as in Theorem 2.8.

Proof. By (5.11), we have

$$\|H_{N,M,\epsilon}(f)\| \leq \|H_{N,M,\epsilon}(f)\|_{[-T,T]} \leq \frac{\sqrt{2}}{\epsilon}\|f-\phi\|_\infty + \|\phi\|_{[-T,T]}, \quad \forall \phi \in \mathcal{G}_N. \tag{5.14}$$

Setting $\phi = 0$ gives the first result. For the second, we let ϕ be the N-term Fourier series of $\tilde{f}$, so that $\|f-\phi\|_\infty \to 0$ as $N \to \infty$. The final inequality follows from the fact that there exists an extension $\tilde{f}$ of f with $\|\tilde{f}\|_{\mathrm{H}^1(\mathbb{T})} \leq c_1(T)\|f\|_{\mathrm{H}^1(-1,1)}$ (Lemma 2.4). □

This corollary shows that the Runge phenomenon, i.e. divergence of $H_{N,M,\epsilon}(f)$, cannot occur in finite precision. This should come as no surprise. Divergence of $H_{N,M,\epsilon}(f)$ would imply unboundedness of the coefficients a_ϵ, a behaviour which is prohibited by truncating the singular values of $\bar{A}$ at level ϵ. Note that this corollary actually shows a much stronger result, namely that $H_{N,M,\epsilon}(f)$ is bounded on the extended domain $[-T,T]$, not just on $[-1,1]$.

Although this corollary demonstrates lack of divergence of $H_{N,M,\epsilon}(f)$, it says littles about its convergence besides the observation that $\|H_{N,M,\epsilon}(f)\|$ is asymptotically bounded by $\|f\|_{\mathrm{H}^1(-1,1)}$. To study convergence we shall use (5.10). For this we first need to understand the constants $C_i(N,M;T,\epsilon)$.

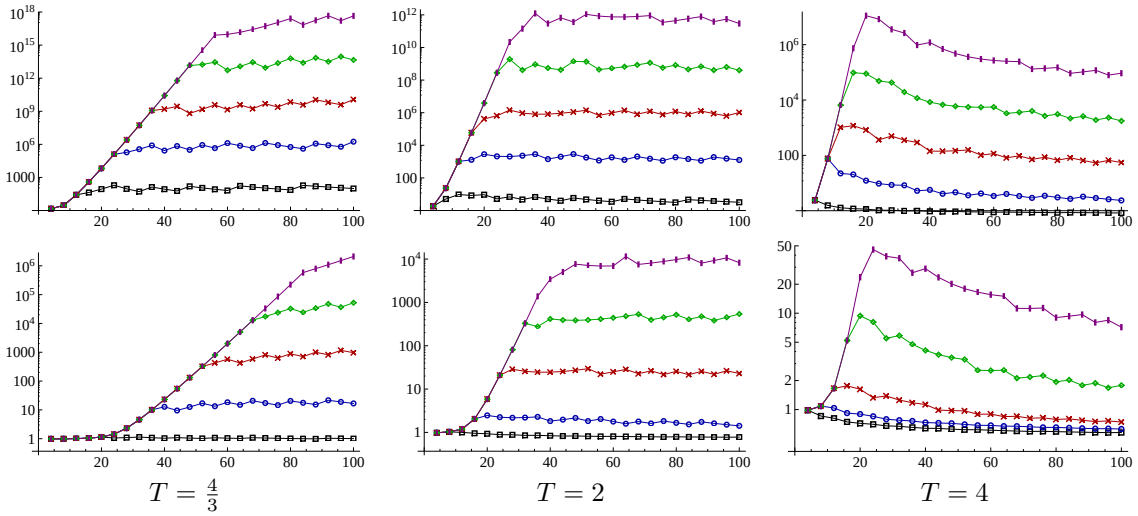


Figure 11: The quantity $C_1(N, \gamma N; T, \epsilon)$ against N for $\gamma = 1$ (top row) or $\gamma = 2$ (bottom row) and $\epsilon = 10^{-6}, 10^{-12}, 10^{-18}, 10^{-24}, 10^{-30}$ (squares, circles, crosses, diamonds and dashes respectively).

5.3.1 Behaviour of $C_i(N, M; T, \epsilon)$

Although Theorem 5.5 holds for arbitrary $M \geq N$, we now focus on the case of linear oversampling, i.e. $M = \gamma N$ for some $\gamma \geq 1$.

Let $N_2(\gamma, T, \epsilon)$ be the largest N such that all the singular values of the matrix $\bar{A}$ are at least ϵ in magnitude. For $N \leq N_2(\gamma, T, \epsilon)$ we have $\mathcal{G}_{N,\gamma N,\epsilon} = \mathcal{G}_N$ and therefore $C_1(N, \gamma N; T, \epsilon) = D(N, \gamma N)$, where $D(N, M)$ is given by (5.4). Thus we witness exponential divergence of $C_1(N, \gamma N; T, \epsilon)$ at rate $c(\gamma; T)$, where $c(\gamma; T)$ is the fixed constant introduced in §5.2.1. This is shown in the Figure 11.

However, once $N > N_2(\gamma, T, \epsilon)$ the numerical results in Figure 11 indicate a completely different behaviour: namely, $C_1(N, \gamma N; T, \epsilon)$ appears to be bounded. Although we have no proof of this fact, extensive numerical experiments indicate that

$$C_1(N, \gamma N; T, \epsilon) \lesssim C_1(N_2, \gamma N_2; T, \epsilon) \sim c(\gamma; T)^{N_2}, \quad \forall N > N_2. \tag{5.15}$$

Thus, $C_1(N, \gamma N; T, \epsilon)$ achieves its maximal value at $N \approx N_2$, and is approximately bounded by this value for all $N > N_2$.

Recall the expression (5.7) for the minimal singular value of $\bar{A}$. Much like before, potential theory arguments can be used once more to obtain the behaviour of the quantity $\tilde{B}(N, \gamma N)$. In particular, one can show that $B(N, \gamma N) \sim d(\gamma; T)^N$ as $N \to \infty$, for some constant $d(\gamma; T) > 1$. With this in hand, it is now easy to verify that

$$N_2(\gamma, T, \epsilon) \sim -\frac{\log \epsilon}{\log d(\gamma; T)}. \tag{5.16}$$

Thus, substituting this into the conjectured bound (5.15) gives

$$C_1(N, \gamma N; T, \epsilon) \lesssim \min\left\{c(\gamma; T)^N, \epsilon^{-\frac{\log c(\gamma;T)}{\log d(\gamma;T)}}\right\}, \quad \forall N \in \mathbb{N}. \tag{5.17}$$

In particular, $C_1(N, \gamma N; T, \epsilon)$ is bounded for all N by some power of ϵ^{-1}. Importantly, this power cannot be too large. Note that $c(\gamma; T) \leq d(\gamma; T)$, $\forall T > 1$, since the maximum of a polynomial on $[m(T), 1]$ is bigger than its maximum on the smaller interval $[-1, 1]$ – compare (5.8) to (5.5). Therefore the ratio $\frac{\log c(\gamma;T)}{\log d(\gamma;T)}$ is at most one. Moreover, for T not too close to 1, we have $c(\gamma; T) \ll d(\gamma; T)$.

The quantity $C_2(N, M; T, \epsilon)$ is harder to analyze, although clearly we have $C_2(N, M; T, \epsilon) = 0$ when $N < N_2$. Figure 12 demonstrates that $C_2(N, \gamma N, \epsilon)$ is bounded in N. Moreover, closer comparison with Figure 11 indicates the existence of a bound of the form

$$C_2(N, \gamma N; T, \epsilon) \lesssim \epsilon C_1(N, \gamma N; T, \epsilon). \tag{5.18}$$

Once more, we have no proof of this observation (see §6 for a discussion).

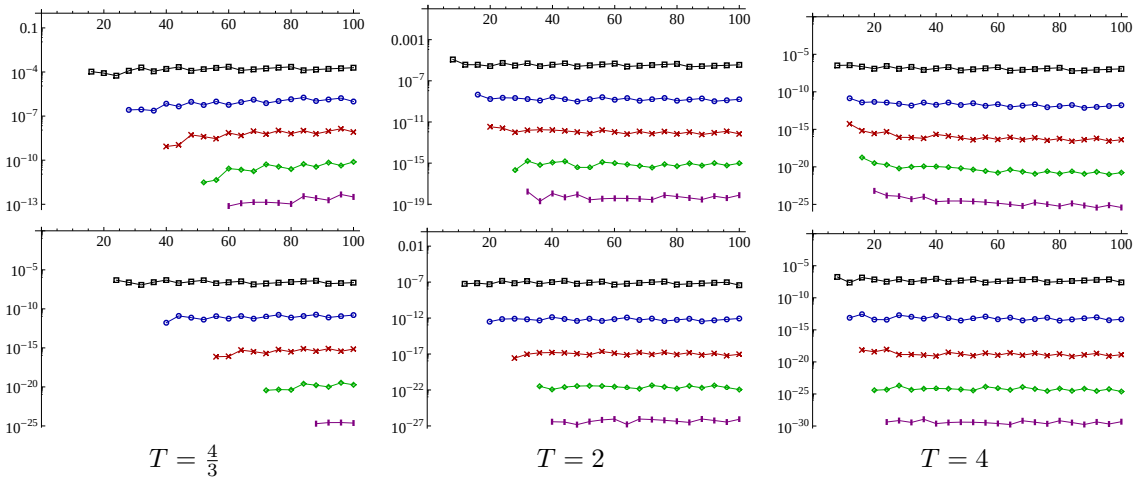


Figure 12: The quantity $C_2(N, \gamma N; T, \epsilon)$ against N for $\gamma = 1$ (top row) or $\gamma = 2$ (bottom row) and $\epsilon = 10^{-6}, 10^{-12}, 10^{-18}, 10^{-24}, 10^{-30}$ (squares, circles, crosses, diamonds and dashes respectively).

Remark 5.7 Note that $C_1(N, M; T, \epsilon)$ and $C_2(N, M; T, \epsilon)$ can each be identified with the norm of a certain matrix, meaning that they can be computed explicitly. These expressions were used to obtain the numerical results in Figures 11 and 12 (computations were carried out with additional precision to avoid effects due to round-off).

With the above results in hand, we can now explain the effect of oversampling on the constants $C_i(N, \gamma N; T, \epsilon)$. Observe that, for fixed T,

$$c(\gamma; T) \to 1, \quad d(\gamma; T) \to E(T), \qquad \gamma \to \infty. \tag{5.19}$$

Thus, increasing the oversampling factor γ leads to a smaller bound in (5.17). As can be seen in Figure 11 and 12, this effect is quite dramatic. In particular, $C_1(N, \gamma N, ; 2, 10^{-12}) \approx 10^4$ for $\gamma = 1$, whereas by setting $\gamma = 2$ we reduce this to the significantly smaller value $C_1(N, \gamma N; 2, 10^{-12}) \approx 10$.

Remark 5.8 The main conclusion of this section is that one requires a lower asymptotic scaling of M with N for the numerical equispaced FE than the exact equispaced FE. Indeed, since $\mathcal{G}_{N,M,\epsilon}$ is a subset of $\mathcal{G}_N$, we clearly have $C_1(N, M; T, \epsilon) \leq D(N, M)$, where $D(N, M)$ is given by (5.4). Hence the discussion in §5.2.1 implies that quadratic scaling $M = \mathcal{O}\left(N^2\right)$ is sufficient to ensure boundedness of $C_1(N, M; T, \epsilon)$ (one can make a similar argument for $C_2(N, M; T, \epsilon)$). However, Figures 11 and 12 indicate that this condition is not necessary, and that one can get away with $M = \mathcal{O}(N)$ in practice.

This difference can be understood intuitively in terms of the singular values of $\bar{A}$. Recall that small singular values of $\bar{A}$ correspond to functions $\phi \in \mathcal{G}_N$ with $\|\phi\|_{[-T,T]} \gg \|\phi\|_M$. Now consider an arbitrary $\phi \in \mathcal{G}_N$. If the ratio $\|\phi\|/\|\phi\|_M$ is large, then this suggests that ϕ must lie approximately in the space $\mathcal{G}^{\perp}_{N,M,\epsilon}$ corresponding to small singular values. Hence, $\|\phi\|/\|\phi\|_M$ cannot be too large over $\phi \in \mathcal{G}_{N,M,\epsilon}$, and thus we see boundedness of $C_1(N, M, \epsilon)$, even when $D(N, M)$ – the supremum of this ratio over the whole of $\mathcal{G}_N$ – is unbounded.

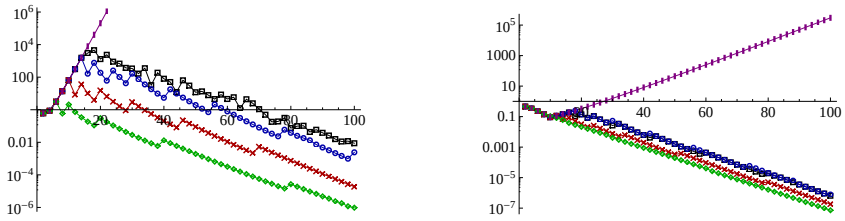
5.3.2 Behaviour of the truncated SVD solution

Combining the analysis of the previous section with Theorem 5.5, we now conjecture the bound

$$\|f - H_{N,\gamma N,\epsilon}(f)\| \leq C(\gamma, T, \epsilon) \left(\|f - \phi\|_\infty + \epsilon \|\phi\|_{[-T,T]} \right), \quad \forall \phi \in \mathcal{G}_N, \tag{5.20}$$

where $C(\gamma, T, \epsilon)$ is proportional to $\epsilon^{-\frac{\log c(\gamma;T)}{\log d(\gamma;T)}}$. In particular, numerical results (Figures 11 and 12) indicate that using $T = 2$ and $\gamma = 2$ gives a bound of a little over 1 in magnitude for $\epsilon = 10^{-14}$.

This estimate allows us to understand the behaviour of the equispaced FE in finite precision. Clearly, when $N < N_2$ the numerically computed extension will diverge exponentially in N whenever f has a singularity in the Runge region $R(\gamma; T)$ (see §5.2.1). However, once N exceeds N_2, one witnesses convergence. Indeed, substituting the exact continuous FE into (5.20), we expect exponential convergence



$\gamma = 1$ $\gamma = 2$

Figure 13: Error for the equispaced FE of $f(x) = \frac{1}{1+100x^2}$. Squares correspond to the numerical solution, circles, crosses and diamonds correspond to truncated SVDs with cutoff $10^{-14}, 10^{-10}, 10^{-6}$ respectively. The dashes correspond to the exact equispaced FE.

up to the breakpoint N_1, followed by spectral convergence. Note that the maximal achievable accuracy is of order $C(\gamma, T, \epsilon)\epsilon \sim \epsilon^{1-\frac{\log c(\gamma;T)}{\log d(\gamma;T)}}$.

Figure 13 confirms these observations for the function $f(x) = \frac{1}{1+100x^2}$. For $\gamma = 1$ the initial exponential divergence is quite noticeable, however this effect can almost be completely removed by doubling γ. Notice that a large cutoff ϵ actually gives a smaller error (since there is a smaller regime of divergence). However, this will also limit the maximal achievable accuracy accordingly.

To summarize, we have now identified three regimes that distinguish the behaviour of $H_{N,\gamma N,\epsilon}(f)$:

(i) $N < N_2(\gamma, T, \epsilon) \approx -\frac{\log \epsilon}{\log d(\gamma;T)}$. Exponential divergence/convergence of $H_{N,\gamma N,\epsilon}(f)$ at rate $c(\gamma;T)/\rho$, where ρ is as in Theorem 2.10.

(ii) $N_2(\gamma, T, \epsilon) \leq N < N_1(T, \epsilon) \approx -\frac{\log \epsilon}{\log E(T)}$. Exponential convergence of $H_{N,\gamma N,\epsilon}(f)$ at rate ρ.

(iii) $N \geq N_1(\gamma, T)$. Spectral convergence of $H_{N,\gamma N,\epsilon}(f)$ down to a maximal achievable accuracy proportional to $\epsilon^{1-\frac{\log c(\gamma;T)}{\log d(\gamma;T)}}$.

Let us make several remarks. First, in practice the regime $N < N_1$ is typically very small (recall that N_1 is around 20 for $T = 2$ – see §4.2.2), and therefore one usually does not witness all three types of behaviour in numerical examples. Second, as $\gamma \to \infty$, we have $N_2 \to N_1$ (recall (5.19)). Thus, with a sufficient amount of oversampling, the regime (ii) of exponential convergence will be arbitrarily small. On the other hand, oversampling decreases $c(\gamma; T)$, and therefore the rate of divergence in the regime (i) is also lessened by taking $\gamma > 1$. Indeed, the numerical experiments in this section indicate that oversampling by a factor of 2 is typically sufficient in practice to mitigate the effects of divergence for all reasonable functions.

Remark 5.9 A similar analysis of the equispaced FE, also based on truncated SVD's, was recently presented by M. Lyon in [20]. In particular, our expressions (5.10) and (5.20) are similar to equations (30) and (31) of [20] (albeit slightly sharper). Lyon also provides extensive numerical results for his analogues of the quantities $C_1(N, M; T, \epsilon)$ and $C_2(N, M; T, \epsilon)$, and describes a bound which is somewhat easier to use in computations. The main contributions of our analysis are the scaling of the constant $C(\gamma, T, \epsilon)$ in terms of ϵ, γ and T, the description and analysis of the breakpoints N_2 and N_1, and the differing convergence/divergence in the corresponding regions.

5.4 The condition number of the numerical method

Having analysed the convergence of the numerical equispaced FE, we now wish to consider its condition number

$$\kappa(G_{N,M}) = \sup_{\substack{f \in \mathrm{L}^\infty(-1,1) \\ b_f \neq 0}} \frac{\|G_{N,M}(f)\|}{\|b_f\|}.$$

Here $b_f = \{f(x_n)\}_{|n| \leq M}$ (note that this is an instance of the condition number of Platte et al [24]). We also define $\kappa_t(G_{N,M})$ analogously to (4.19).

In Figure 14 we plot the $\kappa_t(G_{N,\gamma N})$ against N. The results indicate numerical stability, and, as we expect, improved stability with more oversampling. To illustrate this further, in Figure 15 we

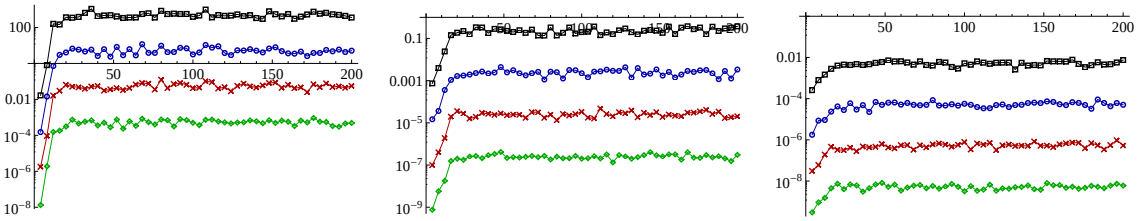


Figure 14: The function $\delta\kappa_t(G_{N,\gamma N})$ against $N = 1, \ldots, 200$, where $T = 2$, $\delta = 10^{-4}, 10^{-6}, 10^{-8}, 10^{-10}$ (squares, circles, crosses and diamonds) and $t = 10$.

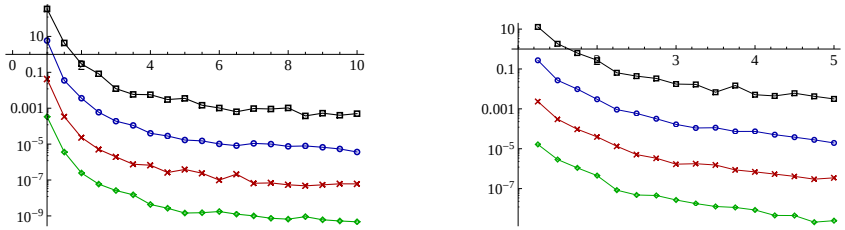


Figure 15: The function $\delta\kappa_t(G_{N,\gamma N})$ for $t = 10$, $\delta = 10^{-4}, 10^{-6}, 10^{-8}, 10^{-10}$ and $N = 200$. Left: $\delta\kappa_t(G_{N,\gamma N})$ against $1 < \gamma \leq 10$ for $T = 2$. Right: $\delta\kappa_t(G_{N,\gamma N})$ against $1 < T \leq 5$ for $\gamma = 2$.

plot $\kappa_t(G_{N,\gamma N})$ as a function of the oversampling parameter γ. As is evident, stability improves with increasing γ. The rate of improvement also declines as γ increases, therefore, in most circumstances, setting $\gamma = 2$ is typically sufficient.

The disadvantage of increasing γ is that one requires more evaluations of f. As an alternative, we may consider varying T. In Figure 15 we also plot the stability function against T. Evidently, larger T leads to better stability. However, the downside of this approach is worse resolution power. In practice, unless given some additional information on the problem, a neutral choice appears to be $T = 2$ and $\gamma = 2$ (recall also Remark 3.4).

The fact that larger T actually leads to better stability may at first sight be surprising. The analysis of §3 suggests that increasing T worsens the condition number of $\bar{A}$. However, numerical stability actually appears to improve. This discrepancy can be explained using Proposition 4.3. When T is small, the constant $c_k(T)$ is large. In other words, extending a function periodically from $[-1, 1]$ to $[-T, T]$ with k orders of smoothness involves large derivatives for $T \approx 1$. On the other hand, for large T, such a construction need not involve such large derivatives, and hence stability improves.

Such arguments can be made precise using the analysis of §5.3 (the proof is similar to that of Theorem 4.7 and hence omitted):

Theorem 5.10. *The condition number $\kappa(H_{N,M,\epsilon})$ for the truncated SVD equispaced FE $H_{N,M,\epsilon}$ satisfies $\kappa(H_{N,M,\epsilon}) \leq C_1(N, M; T, \epsilon)$, where $C_1(N, M; T, \epsilon)$ is given by (5.12).*

From the analysis of §5.3.1 we conclude that $\kappa(H_{N,\gamma N,\epsilon}) \lesssim \epsilon^{-\frac{\log c(\gamma;T)}{\log d(\gamma;T)}}$. In particular, (5.19) implies that $\kappa(H_{N,\gamma N,\epsilon}) \lesssim 1$ as $\gamma \to \infty$: in other words, by oversampling sufficiently we can render the equispaced FE completely stable. On the other hand, $d(\gamma; T) \to \infty$ as $T \to \infty$ (for fixed γ), whereas $c(\gamma; T)$ is bounded. Hence $\kappa(H_{N,\gamma N,\epsilon})$ can again be made arbitrarily close to 1 by taking T sufficiently large. This confirms the arguments given above.

5.5 Numerical examples and relation to Platte, Trefethen & Kuijlaars

In Figure 16 we consider the equispaced FE for three test functions. In all cases we use $\gamma = 2$, with T either being fixed, or given by (2.16). As is evident, all choices of T give good, stable numerical results, with the best achievable accuracy being at least 10^{-12}. Moreover, we observe that larger T leads to a slightly better best achievable accuracy, exactly as predicted in §5.4. On the other hand, larger T means worse resolution power, as can be seen in the oscillatory example.

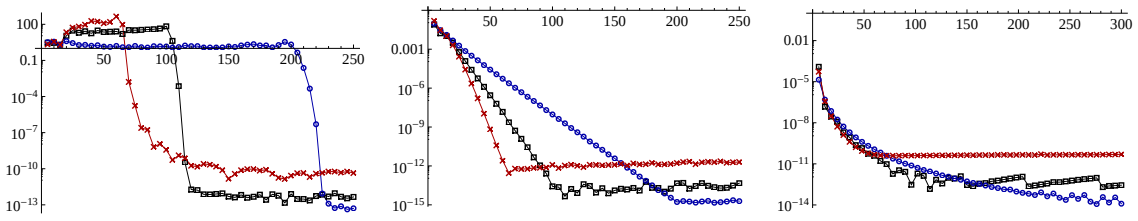


Figure 16: The error $\|f - G_{N,\gamma N}(f)\|_\infty$, where $\gamma = 2$ and $T = 2$ (squares), $T = 4$ (circles) or $T = T(N; \epsilon_{\mathrm{tol}})$ (crosses), $\epsilon_{\mathrm{tol}} = 10^{-14}$.

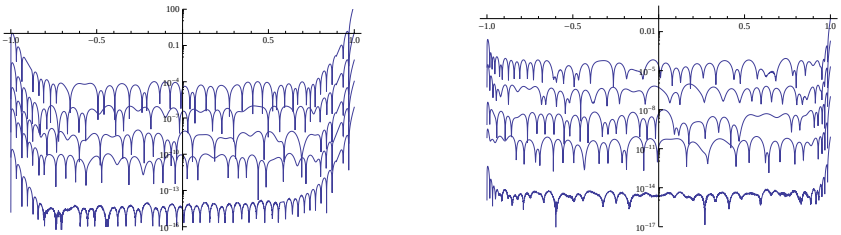


Figure 17: The error $|f(x) - G_{N,\gamma N}(f)(x)|$ against x, where $\gamma = 1$ (left) or $\gamma = 2$ (right), for $N = 30$, $T = 2$ and $f(x) = \mathrm{e}^x$, with noise at amplitudes $\delta = 10^{-4}, 10^{-6}, 10^{-8}, 10^{-10}, 0$.

Robustness in the presence of noise is shown in Figure 17. Observe that when $\gamma = 1$, noise at amplitude δ is magnified by around 10^5, consistently with Theorem 5.10 (note that the constant $C_1(N, N; T, \epsilon)$ is roughly 10^5 in magnitude for $\epsilon = 10^{-14}$ – see Figure 11). Conversely, with double oversampling, this factor drops to around 10^2, again in agreement with Theorem 5.10.

Finally, we are now in a position to explain how the equispaced FE relates to the theorem of Platte, Trefethen & Kuijlaars. First, it has bounded condition number. Second, after small regimes (whose sizes are function independent) of possible exponential divergence and convergence, it is spectrally convergent down to a maximal achievable accuracy on the order of machine precision.

6 Conclusions and challenges

We conclude by making the following remark. Extensive numerical results [7, 8, 10, 20, 21] have shown the effectiveness of FE's in approximating even badly behaved functions to high accuracy in a stable fashion. The purpose of this paper has been to present an explanation and analysis as to why this is the case. For the three types of extensions considered, we have shown that, whilst the computation of the coefficients of the FE is exponentially ill-conditioned, the computation of the extension itself is, in fact, numerically stable. Moreover, even when there is significant error in the computed coefficients, one still obtains a spectrally convergent approximation. This is due to the facts that the FE is a frame approximation and that, for all functions f, even those with oscillations or r:scalinglarge derivatives, there eventually exist coefficient vectors with small norm which approximate f to high accuracy.

The main outstanding challenge is to understand the constants $C_i(N, M; T, \epsilon)$ of the equispaced FE. In particular, we wish to show that linear scaling $M = \gamma N$ is sufficient to ensure boundedness of these constants in N, with a larger γ corresponding to a smaller bound. Note that the analysis of §5.2.1 implies the suboptimal result that $M = \mathcal{O}\left(N^2\right)$ is sufficient (Remark 5.8). It is also a relatively straightforward exercise to show that if $M = cN/\epsilon$ for some $c > 0$, then $C_i(N, M; T, \epsilon)$ is bounded (this is based on making rigorous the arguments given in Remark 5.8 – we do not report it here for brevity's sake). Unfortunately, although this estimate give the correct scaling $M = \mathcal{O}(N)$, it is wildly pessimistic: it implies that M should scale like $\approx 10^{16} N$, whereas the numerics in §5.3.1 indicate that $M = \gamma N$ is sufficient for *any* $\gamma \geq 1$.

One approach to establish a more satisfactory result is to perform a closer analysis of the singular values of the matrix $\bar{A}$. Some preliminary insight into this problem was given in [15], wherein it was proved that (whenever $M = N$ and $2T \in \mathbb{N}$) the singular values cluster near zero and one, and the

transition region is $\mathcal{O}(\log N)$ in width (much like for the prolate matrix A). Unfortunately, little is known outside of this result: there is no existing analysis for $\bar{A}$ akin to that of Slepian's for the prolate matrix – see [15] for a discussion. Note, however, that the normal form $B = \bar{A}^*\bar{A}$, with entries $B_{n,m} = \frac{\sin\frac{(n-m)\pi}{T}}{MT\sin\frac{(n-m)\pi}{MT}}$, can be viewed as a discretized version of the prolate matrix A. Indeed, $B \to A$ as $M \to \infty$ for fixed N. Given the similarities between the two matrices, there is potential for Slepian's analysis to be extended to this case. However, this remains an open problem.

Acknowledgements

The authors would like to thank Doug Cochran, Laurent Demanet, Anne Gelb, Anders Hansen, Arieh Iserles, Arno Kuijlaars, Mark Lyon, Nilima Nigam, Sheehan Olver, Rodrigo Platte and Nick Trefethen for useful discussions and comments.

References

[1] B. Adcock and D. Huybrechs. On the resolution power of Fourier extensions for oscillatory functions. Technical Report TW597, Dept. Computer Science, K.U. Leuven., 2011.

[2] N. Albin and O. P. Bruno. A spectral FC solver for the compressible Navier–Stokes equations in general domains I: Explicit time-stepping. *J. Comput. Phys.*, 230(16):6248–6270, 2011.

[3] H. Bateman. *Higher Transcendental Functions*. Vol. 2, McGraw–Hill, New York, 1953.

[4] J. Boyd. Fourier embedded domain methods: extending a function defined on an irregular region to a rectangle so that the extension is spatially periodic and C^∞. *Appl. Math. Comput.*, 161(2):591–597, 2005.

[5] J. Boyd and F. Xu. Divergence (Runge phenomenon) for least-squares polynomial approximation on an equispaced grid and mock-Chebyshev subset interpolation. *Appl. Math. Comput.*, 210(1):158–168, 2009.

[6] J. P. Boyd. *Chebyshev and Fourier Spectral Methods*. Springer–Verlag, 1989.

[7] J. P. Boyd. A comparison of numerical algorithms for Fourier Extension of the first, second, and third kinds. *J. Comput. Phys.*, 178:118–160, 2002.

[8] J. P. Boyd and J. R. Ong. Exponentially-convergent strategies for defeating the Runge phenomenon for the approximation of non-periodic functions. I. Single-interval schemes. *Commun. Comput. Phys.*, 5(2–4):484–497, 2009.

[9] O. Bruno and M. Lyon. High-order unconditionally stable FC-AD solvers for general smooth domains I. Basic elements. *J. Comput. Phys.*, 229(6):2009–2033, 2010.

[10] O. P. Bruno, Y. Han, and M. M. Pohlman. Accurate, high-order representation of complex three-dimensional surfaces via Fourier continuation analysis. *J. Comput. Phys.*, 227(2):1094–1125, 2007.

[11] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang. *Spectral methods: Fundamentals in Single Domains*. Springer, 2006.

[12] O. Christensen. *An Introduction to Frames and Riesz Bases*. Birkhauser, 2003.

[13] D. Coppersmith and T. Rivlin. The growth of polynomials bounded at equally spaced points. *SIAM J. Math. Anal.*, 23:970–983, 1992.

[14] R. J. Duffin and A. C. Schaeffer. A class of nonharmonic Fourier series. *Trans. Amer. Math. Soc.*, 72:341–366, 1952.

[15] A. Edelman, P. McCorquodale, and S. Toledo. The future Fast Fourier Transform? *SIAM J. Sci. Comput.*, 20(3):1094–1114, 1999.

[16] B. Fornberg. *A Practical Guide to Pseudospectral Methods*. Cambridge University Press, 1996.

[17] D. Gottlieb and S. A. Orszag. *Numerical Analysis of Spectral Methods: Theory and Applications*. Society for Industrial and Applied Mathematics, 1st edition, 1977.

[18] D. Huybrechs. On the Fourier extension of non-periodic functions. *SIAM J. Numer. Anal.*, 47(6):4326–4355, 2010.

[19] D. Kosloff and H. Tal-Ezer. A modified Chebyshev pseudospectral method with an $\mathcal{O}(N^{-1})$ time step restriction. *J. Comput. Phys.*, 104:457–469, 1993.

[20] M. Lyon. Approximation error in regularized SVD-based Fourier continuations. *Preprint*, 2011.

[21] M. Lyon. A fast algorithm for Fourier continuation. *SIAM J. Sci. Comput.*, 33(6):3241–3260, 2012.

[22] M. Lyon and O. Bruno. High-order unconditionally stable FC-AD solvers for general smooth domains II. Elliptic, parabolic and hyperbolic PDEs; theoretical considerations. *J. Comput. Phys.*, 229(9):3358–3381, 2010.

[23] R. Pasquetti and M. Elghaoui. A spectral embedding method applied to the advection–diffusion equation. *J. Comput. Phys.*, 125:464–476, 1996.

[24] R. Platte, L. N. Trefethen, and A. Kuijlaars. Impossibility of fast stable approximation of analytic functions from equispaced samples. *SIAM Rev.*, 53(2):308–318, 2011.

[25] T. Ransford. *Potential theory in the complex plane*. Cambridge Univ. Press, Cambridge, UK, 1995.

[26] T. J. Rivlin. *Chebyshev Polynomials: from Approximation Theory to Algebra and Number Theory*. Wiley New York, 1990.

[27] D. Slepian. Prolate spheriodal wave functions. Fourier analysis, and uncertainty V: The discrete case. *Bell System Tech J.*, 57:1371–1430, 1978.

[28] L. N. Trefethen and D. Bau. *Numerical Linear Algebra*. SIAM, 1997.

[29] J. Varah. The prolate matrix. *Linear Algebra Appl.*, 187(1):269–278, 1993.