Numer. Math. 69: 423–440 (1995)

Numerische Mathematik
© Springer-Verlag 1995

Efficient numerical methods in non-uniform sampling theory

Hans G. Feichtinger[1], **Karlheinz Gröchenig**[2] **Thomas Strohmer**[1,⋆]

[1] Department of Mathematics, University of Vienna, Strudlhofgasse 4, A-1090 Wien, Austria
E-mail: FEI@TYCHE.MAT.UNIVIE.AC.AT, STROHMER@TYCHE.MAT.UNIVIE.AC.AT
[2] Department of Mathematics, The University of Connecticut, Storrs, CT. 06269-3009, USA
E-mail: GROCH@MATH.UCONN.EDU

Received December 22, 1993

Summary. We present a new "second generation" reconstruction algorithm for irregular sampling, i.e. for the problem of recovering a band-limited function from its non-uniformly sampled values. The efficient new method is a combination of the adaptive weights method which was developed by the two first named authors and the method of conjugate gradients for the solution of positive definite linear systems. The choice of "adaptive weights" can be seen as a simple but very efficient method of preconditioning. Further substantial acceleration is achieved by utilizing the Toeplitz-type structure of the system matrix. This new algorithm can handle problems of much larger dimension and condition number than have been accessible so far. Furthermore, if some gaps between samples are large, then the algorithm can still be used as a very efficient extrapolation method across the gaps.

Mathematics Subject Classification (1991): 42A15, 65D05, 65D10, 65F10

1. Introduction

One of the principal problems in signal analysis is the reconstruction or approximation of a signal from its discrete samples. In many practical considerations one may safely assume some maximal frequency in the signal. One may thus consider the sampling problem for band-limited functions. In the ideal case of equally spaced samples the reconstruction is routine and can be carried out explicitly by one of the many variations of the Shannon-Whittaker-Kotel'nikov sampling theorem [6, 26, 33].

However, in many applications, for instance in astronomy, seismology, tomography and physics, one is forced to sample signals at nonuniformly spaced points. This problem has received much attention in the past years, see [2, 14–16, 18, 31] for history and references. But despite an abundance of work on the irregular sampling problem — [30] lists about 300 references — its numerical and algorithmic

⋆ The second named author was partially supported by NSF grant DMS-9306430. The first and third named authors have been partially supported through project PH08784 of the Austrian Science foundation FWF
Correspondence to: K. Gröchenig

aspects have been neglected so far. Simple iterative algorithms have been proposed in [2, 13, 18, 21, 22, 32, 35, 40, 41]. These algorithms seem to work decently for well-conditioned problems and for small data sets, but become slow and expensive for more complicated and more realistic problems. A comparison of the performance of the "first generation" of reconstruction algorithms can be found in [12, 18].

In the present paper we introduce a new "superfast" algorithm for the reconstruction of band-limited signals from irregular samples. This algorithm is iterative and cuts the number of iterations to achieve a given accuracy by an order of magnitude when compared with the simple iteration schemes. We will substantiate these claims by both theoretical estimates and by the results of numerical simulations.

The problem

Let f be a band-limited signal of finite energy, i.e.

$$\int_{-\infty}^{\infty} |f(t)|^2 dt < \infty \quad \text{and supp}\hat{f} \subseteq [-\Omega, \Omega] \tag{1}$$

for some $\Omega > 0$, and suppose that the values $f(t_n)$ are known at a biinfinite sampling sequence $\ldots t_{n-1} < t_n < t_{n+1} \ldots$ with $\lim_{n\to\pm\infty} t_n = \infty$. Then the question is whether f is uniquely determined by its samples; if yes, how can it be reconstructed? Is this reconstruction stable and how can error estimates be obtained? This is an infinite-dimensional problem and has been treated in many variations and through many approaches. See [2, 15, 14] for recent contributions and references.

For the problem to be accessible for numerical solution, we first have to create a *finite-dimensional model*. For the discrete theory used below we follow [22]. A *discrete signal* of length N is a finite sequence $(s(0), s(1), \ldots, s(N-1)) \in \mathbb{C}^N$. Its ℓ^2-norm (the square root of the signal energy) is

$$\|s\| = \left(\sum_{n=0}^{N-1} |s(n)|^2 \right)^{1/2} . \tag{2}$$

With the help of the unitary discrete Fourier transform

$$\hat{s}(k) = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} s(n) \mathrm{e}^{-2\pi \mathrm{i} kn/N} \quad k = 0, \ldots, N-1 \tag{3}$$

we can define discrete band-limited signals of bandwidth $M < N/2$ as follows:

$$\mathscr{B}_M = \left\{ s \in \mathbb{C}^N | \hat{s}(k) = 0 \ \text{ for } |k \bmod (N)| > M \right\} \tag{4}$$

Here and throughout the paper we extend finite sequences in $\mathbb{C}^N$ to infinite sequences of period N by $s(n+lN) = s(n)$ and $\hat{s}(k+lN) = \hat{s}(k)$ for $n, k = 0, \ldots, N-1,\ l \in \mathbb{Z}$, so that $s(n)$ and $\hat{s}(k)$ make sense for all $n, k \in \mathbb{Z}$.

The discrete irregular sampling problem can now be formulated in the following way: given a subsequence $0 \le n_1 < \ldots < n_r \le N-1$ and the samples $s(n_i), i = 1, \ldots, r$ of a discrete band-limited signal $s \in \mathscr{B}_M$, when can s be reconstructed? The theoretical answer being trivial (one just has to have enough sampling points), the real question is how can s be reconstructed *fast and efficiently*? Since the number

of samples r is often of the order $r \approx 10^3 - 10^6$, this leads to large systems of linear equations which certainly cannot be solved directly.

The numerical aspects of this problem have not yet been explored satisfactorily. The iterative algorithms that have been proposed in the engineering literature [30, 32, 35, 40, 41] seem to work adequately only for small dimensions, but in realistic problems their performance is by no means conclusive. In large problems most of these are rather slow and they also lack robustness, for instance with respect to irregularities in the sampling set.

Before we proceed to the deduction and presentation of a new "superfast" algorithm, let us give a different and more convenient interpretation of the irregular sampling problem for discrete signals. Since for $s \in \mathscr{B}_M$ we have

$$s(n) = \frac{1}{\sqrt{N}} \sum_{k=-M}^{M} \hat{s}(k) \mathrm{e}^{2\pi \mathrm{i} k n/N}, \tag{5}$$

a discrete signal of length N and bandwidth M can be interpreted as the restriction of a trigonometric polynomial of period 1 and degree M to an arithmetic sequence in $[0, 1)$. More precisely, if $s \in \mathscr{B}_M$ is given by (5) and

$$p(t) = \frac{1}{\sqrt{N}} \sum_{k=-M}^{M} \hat{s}(k) \mathrm{e}^{2\pi \mathrm{i} k t} \tag{6}$$

then $s(n) = p(\frac{n}{N}), n = 0, \ldots, N-1$. The original question now turns into the problem how to reconstruct p from its samples $p(\frac{n_i}{N}), i = 1, \ldots, r$. From this point of view it is completely irrelevant that the sampling sequence $\frac{n_i}{N}$ is a subsequence of an arithmetic progression. For the remainder of the paper we consider therefore the $2M + 1$-dimensional space

$$\mathscr{T}_M = \left\{ p \mid p(t) = \sum_{k=-M}^{M} a_k \mathrm{e}^{2\pi \mathrm{i} k t} \right\} \tag{7}$$

of all trigonometric polynomials of degree M and period 1. The problem discussed in this paper can now be formulated as follows: Given a sampling sequence $0 \leq t_1 < t_2 < \ldots < t_r < 1$, we ask how $p \in \mathscr{T}_M$ can be reconstructed from its sampled values $p(t_i)$.

It should be noted that this problem differs from the trigonometric interpolation which asks to find a trigonometric polynomial of *appropriate* degree which interpolates the given data $(t_i, y_i), i = 1, \ldots, r$, in the sense that $p(t_i) = y_i$. There are several explicit interpolation procedures known, see [44] and also efficient algorithms [4, 34]. However, numerically these methods seem to be fairly unstable.

In practice an upper bound for the degree of p is known as a consequence of the band-limitedness and one avoids the bad conditioning by oversampling, i.e., collecting more data than necessary for uniqueness. If data are noisy the interpolation of p may no longer be possible, but in this case a least square approximation problem can be solved.

In the sequel we will give explicit estimates which indicate how a higher oversampling rate (maximal gap size divided by Nyquist rate) improves the condition number of the reconstruction problem.

2. Uniqueness and a primitive reconstruction algorithm

In the following sections we present the different elements necessary for an efficient reconstruction algorithm. The optimal method then results as a combination of the various considerations.

We first discuss the principal solvability of the reconstruction problem and indicate various methods of reconstruction, cf. also [13].

Lemma 1. *Let* $0 \leq t_1 < \ldots < t_r < 1$ *be* r *arbitrary distinct sampling points in* $[0,1)$ *and* $p \in \mathscr{P}_M$. *Then* p *is uniquely determined by its* r *samples* $p(t_i)$ *if and only if* $r \geq 2M+1$. *In this case there exist two constants* $0 < A \leq B$ *so that*

$$A \int_0^1 |p(t)|^2 \, dt \leq \sum_{i=1}^{r} |p(t_i)|^2 \leq B \int_0^1 |p(t)|^2 \, dt \tag{8}$$

Proof. This is of course well known and relies on the fact that $\mathrm{e}^{\mathrm{i}Mx}p(x)$ is the restriction of a complex polynomial of order $2M$ to the circle $\{z \in \mathbb{C} : |z| = 1\}$. Thus any $2M+1$ distinct points determine p completely. The existence of A and B then follows from the observation that the map $p \in \mathscr{P}_M \to (p(t_i))_{i=1,\ldots,r} \in \mathbb{C}^r$ is one-to-one and thus invertible on its range. □

The equivalence of the norms in (8) can be exploited in several ways to obtain a reconstruction algorithm. Let D_M denote the Dirichlet kernel

$$D_M(t) = \sum_{k=-M}^{M} \mathrm{e}^{2\pi \mathrm{i} k t} = \frac{\sin(M+\frac{1}{2})2\pi t}{\sin \pi t} \tag{9}$$

and observe that

$$p(t) = \int_0^1 p(u) D_M(t-u) \, du = \langle p, D_M(.-t) \rangle \ . \tag{10}$$

Define the *frame operator* S [11, 43] by

$$Sp(t) = \sum_{i=1}^{r} p(t_i) D_M(t-t_i) \ . \tag{11}$$

Following Duffin and Schaeffer [11] we obtain the following simple iterative reconstruction algorithm.

Lemma 2. *Fix* $M \in \mathbb{N}$ *and suppose* $r \geq 2M+1$ *and* $\lambda < \frac{1}{B}$, *where* B *is the constant in (8). Define iteratively* $p_0 = 0$,

$$p_n = p_{n-1} + \lambda S(p - p_{n-1}) \ . \tag{12}$$

Then $\lim_{n\to\infty} p_n = p$ *for* $p \in \mathscr{P}_M$ *and*

$$\|p - p_n\|_2 \leq \gamma^n \|p\|_2 \tag{13}$$

where $\gamma = \max\{|1-\lambda A|, |1-\lambda B|\} < 1$. *Since* Sp *depends only on the samples* $p(t_i)$, *this is indeed a reconstruction from the samples only.*

Proof. [11, 24, 39, 42, 43] Since $\langle Sp, p\rangle = \sum_{i=1}^{r} |p(t_i)|^2$ by (10) and (11), S is a positive operator on $\mathscr{P}_M$ and thus invertible. By (8) one obtains

$$(1-\lambda B)\|p\|_2^2 \le \langle (Id-\lambda S)p, p\rangle \le (1-\lambda A)\|p\|_2^2 \tag{14}$$

and consequently

$$\|p-p_n\|_2 = \|(Id-\lambda S)(p-p_{n-1})\|_2 \le \gamma\|p-p_{n-1}\|_2 \le \ldots \le \gamma^n\|p\|_2 \ . \quad \square$$

In terms of linear algebra this is the Richardson iteration for $Sp = q$ with q given, and it comes with all its advantages and disadvantages [25, 24, 42].

The frame algorithm or Richardson iteration of Lemma 2 yields decent convergence only if

- explicit estimates for the constants A and B can be derived, and furthermore
- the optimal convergence factor $\frac{B-A}{B+A}$ for the Richardson iteration is small. For a convergence analysis see [39, 42].

For the irregular sampling problem these conditions are rarely satisfied. It is a hard problem to derive explicit acceptable estimates of the constants in (8), let alone good estimates. Also, if $r \approx 2M+1$ and the sampling points are distributed very irregularly, then even for low dimensional problems, $r = 25$, say, the condition numbers can be of the order $10^{12} - 10^{15}$. In fact, clusters in the sampling family imply that the upper frame bound B will be very large, whereas a single larger gap implies that the lower frame bound A is going to be very small.

Although it is a general experience of numerical analysts that the Richardson iteration should not be used, the simple algorithm of Lemma 2 is very much in favor in the signal analysis community [30, 32, 40] and even among mathematicians [2, 18]. The reasons why it is still in use are probably its simplicity and the close relation of the algorithm of Lemma 2 to the Shannon-Whittaker-Kotel'nikov sampling theorem. Its discrete version is the reconstruction formula

$$p(t) = \frac{1}{N}\sum_{k=0}^{N-1} p\left(\frac{k}{N}\right) D_M\left(t-\frac{k}{N}\right) \tag{15}$$

for $p \in \mathscr{P}_M$ and any $N > 2M$ [33]. We see that the frame operator (11) mimics the "cardinal series" in (15) literally with the regular samples $\frac{k}{N}$ being replaced by the irregular samples at t_i.

We can write the algorithm of Lemma 2 in terms of matrices. Let A be the $r \times r$ matrix with entries

$$A_{jk} = D_M(t_j - t_k) = \frac{\sin(M+\frac{1}{2})2\pi(t_j-t_k)}{\sin \pi(t_j-t_k)} \tag{16}$$

and let $v^{(n)} \in \mathbb{C}^r$ be the vector determined by $p_n(t) = \sum_{i=1}^{r} v_i^{(n)} D_M(t-t_i)$. Then

$$v^{(n)} = v^{(n-1)} + \lambda v^{(0)} - \lambda A v^{(n-1)}\,, \tag{17}$$

where $v_i^{(0)} = p(t_i)$, $i = 1, \ldots, r$, is just the input. Then by Lemma 2 $v^{(n)}$ converges to $v \in \mathbb{C}^r$ and $p(t) = \sum_{i=1}^{r} v_i D_M(t-t_i)$.

This implementation is rather clumsy because A is a full matrix and because the dimension of the problem grows with the number of sampling points – a quite unnatural feature.

3. Reformulation as a Toeplitz system

To make the dimension independent of the number of samples — after all we are dealing with a problem in the $2M+1$-dimensional space $\mathscr{S}_M$ — we look at the action of S on the coefficients of trigonometric polynomials as in [22]. Let $p(t) = \sum_{k=-M}^{M} a_k \mathrm{e}^{2\pi \mathrm{i} k t} \in \mathscr{S}_M$ with coefficients $\mathbf{a} = (a_k)_{k=-M}^{M} \in \mathbb{C}^{2M+1}$. Then

$$Sp(t) = \sum_{j=1}^{r} p(t_j) D_M(t - t_j) = \sum_{j=1}^{r} \sum_{k=-M}^{M} \sum_{l=-M}^{M} a_k \mathrm{e}^{2\pi \mathrm{i} k t_j} \mathrm{e}^{2\pi \mathrm{i} l t} \mathrm{e}^{-2\pi \mathrm{i} l t_j}$$

$$= \sum_{l=-M}^{M} \left[\sum_{k=-M}^{M} \left(\sum_{j=1}^{r} \mathrm{e}^{-2\pi \mathrm{i}(l-k) t_j} \right) a_k \right] \mathrm{e}^{2\pi \mathrm{i} l t} \ . \tag{18}$$

Thus let T be the $(2M+1) \times (2M+1)$ Toeplitz matrix with entries

$$T_{lk} = T_{l-k} = \sum_{j=1}^{r} \mathrm{e}^{-2\pi \mathrm{i}(l-k) t_j} \quad \text{for } |l|, |k| \leq M \ . \tag{19}$$

Then the coefficients of the trigonometric polynomials $Sp \in \mathscr{S}_M$ are $T\mathbf{a}$. Lemma 2 could now be rewritten in the following form.

Lemma 3. *Let $p(t) = \sum_{k=-M}^{M} a_k \mathrm{e}^{2\pi \mathrm{i} k t} \in \mathscr{S}_M$ with coefficients $\mathbf{a} = (a_k)_{k=-M}^{M} \in \mathbb{C}^{2M+1}$ and let $0 \leq t_1 < \ldots < t_r < 1$ be an arbitrary sampling sequence, with $r \geq 2M+1$. Let $\mathbf{b} \in \mathbb{C}^{2M+1}$ be defined by*

$$b_k = \sum_{j=1}^{r} p(t_j) \mathrm{e}^{-2\pi \mathrm{i} k t_j} \quad \textit{for } |k| \leq M \, . \tag{20}$$

Then for λ small enough, the iteration $\mathbf{a}^{(0)} = 0$,

$$\mathbf{a}^{(k)} = \mathbf{a}^{(k-1)} - \lambda T \mathbf{a}^{(k-1)} + \lambda \mathbf{b} \quad k \geq 1 \tag{21}$$

converges to $\mathbf{a} \in \mathbb{C}^{2M+1}$. As $\mathbf{a}^{(1)} = \lambda \mathbf{b}$ requires only knowledge of the samples of p, this is a reconstruction of p.

This is just a reformulation of Lemma 2 by means of (18) and (19) and

$$Sp(t) = \sum_{k=-M}^{M} \left(\sum_{j=1}^{r} p(t_j) \mathrm{e}^{-2\pi \mathrm{i} k t_j} \right) \mathrm{e}^{2\pi \mathrm{i} k t} \ . \tag{22}$$

We can even obtain a much simpler reconstruction of p.

Lemma 4. *Let $p(t_j), j = 1, \ldots, r$, be the samples of $p \in \mathscr{S}_M$ with $r \geq 2M+1$ and let $\mathbf{b} = (b_k)_{|k| \leq M}$ be defined as in (20). Compute*

$$\mathbf{a} = T^{-1} \mathbf{b} \in \mathbb{C}^{2M+1} \ , \tag{23}$$

then $p(t) = \sum_{k=-M}^{M} a_k \mathrm{e}^{2\pi \mathrm{i} k t} \in \mathscr{S}_M$.

Proof. This is clear from (21), since we obtain $T\mathbf{a} = \mathbf{b}$ as $\mathbf{a}^{(k)}$ converges to $\mathbf{a}$. The invertibility and positivity of T is already part of Lemma 1, because

$$\langle Sp, p\rangle_{L^2(0,1)} = \langle T\mathbf{a}, \mathbf{a}\rangle_{C^{2M+1}} \quad \square \tag{24}$$

Remarks. Although Lemma 3 and 4 are just reformulations of the usual reconstruction algorithm, we have already obtained a much better method.

(a) The dimension of the problem is now $2M + 1$ and it depends *only on the size of the spectrum* and is independent of the number of samples.

(b) We have gained some additional structure. Instead of storing $r^2/2$ entries of A (16) which are non-zero and distinct in general, we have to store only the $2M + 1$ entries

$$T_{0k} = \sum_{j=1}^{r} \mathrm{e}^{2\pi \mathrm{i} k t_j} \quad k = 0, 1, \ldots, 2M \ . \tag{25}$$

(c) The matrix T of equation (19) is a Toeplitz matrix for any spectrum of the form $[M_1, M_2]$.

(d) There is a large repertory of Toeplitz solvers at our disposal [1, 36, 37, 3, 10] which can solve (23) with only $O(M^2)$, $O(M \log^2 M)$ or even $O(M \log M)$ operations. Depending on the size of M and the required precision one may choose either

(I) direct methods for the complete inversion of T, or

(II) iterative methods for an approximation of T^{-1}.

4. The adaptive weights method and estimates of the condition number

Although we would obtain decent reconstructions by means of Lemma 4 and appropriate Toeplitz solvers, the reconstructions can be further improved by the so-called *adaptive weights method*. This method is also more satisfactory in theory because it *improves the condition number* of the irregular sampling problem, *provides explicit estimates* for the rate of convergence depending only on the maximal gap size, and therefore gives useful *stopping criteria*.

The following statement has been derived in [22].

Proposition 1. *Suppose that* $0 \le t_1 < \ldots t_r < 1$ *and define the maximal gap* δ *as*

$$\delta := \max(t_{i+1} - t_i) < \frac{1}{2M} \tag{26}$$

where we set $t_0 = t_r - 1$ *and* $t_{r+1} = t_1 + 1$. *Then*

$$(1 - 2\delta M)^2 \ \|p\|_2^2 \le \sum_{i=1}^{r} |p(t_i)|^2 \ \frac{t_{i+1} - t_{i-1}}{2} \le (1 + 2\delta M)^2 \ \|p\|_2^2 \tag{27}$$

holds for all $p \in \mathscr{S}_M$.

Instead of repeating the proof of [22] we comment on the rationale of the weights $w_i = \frac{1}{2}(t_{i+1} - t_{i-1})$. They help to keep the condition number low in the presence of clusters in the sampling set.

To understand the effect assume that there are many sampling points in some subinterval I of $[0, 1)$. Then the upper bound B in (8) increases, roughly proportional

to the number of samples, whereas the lower bound is virtually unaffected – just take a polynomial whose energy is most "concentrated" outside I. Then the convergence factor $\frac{B-A}{A+B} \approx 1 - \frac{2A}{B}$ is almost 1 and the frame algorithm of Lemma 2 converges extremely slow.

On the other hand, using the weighted frame operator

$$S_w p(t) = \sum_{i=1}^{r} p(t_i) w_i D_M(t - t_i) \tag{28}$$

and the modified iteration $p_0 = 0$ and

$$p_n = p_{n-1} + \lambda_{\text{opt}} S_w(p - p_{n-1}) \tag{29}$$

with $\lambda_{\text{opt}} = \frac{1}{1+4\delta^2 M^2}$ leads to the *explicit* error estimates

$$\|p - p_n\|_2 \leq \left(\frac{4\delta M}{1 + 4\delta^2 M^2} \right)^n \|p\|_2 \tag{30}$$

with convergence factor $\gamma = 4\delta M \lambda_{\text{opt}} < 1$.

Thus the rate of convergence is independent of clustering effects of the samples and depends essentially only on the maximal gap between the samples. The use of weights w_i compensates the local variations in the sampling density. If we repeat the computation in (18) with the weighted operator S_w in place of S, we are lead to the Toeplitz matrix T_w with the entries

$$(T_w)_{lk} = (T_w)_{l-k} = \sum_{j=1}^{r} w_j \mathrm{e}^{-2\pi \mathrm{i}(l-k)t_j} \quad \text{for } |l|, |k| \leq M \ . \tag{31}$$

The improved and quantified version of Lemma 4 now reads as follows.

Proposition 2. *Suppose that $\delta < \frac{1}{2M}$. Let $p(t_i)$ be r samples of some $p \in \mathscr{P}_M$. Set $\mathbf{b} \in \mathbb{C}^{2M+1}$ with entries*

$$b_k = \sum_{j=1}^{r} p(t_j) w_j \mathrm{e}^{-2\pi \mathrm{i} k t_j} \quad \textit{for } |k| \leq M \,, \tag{32}$$

and calculate $\mathbf{a} = T_w^{-1}\mathbf{b}$. Then $p(t) = \sum_{k=-M}^{M} a_k \mathrm{e}^{2\pi \mathrm{i} k t}$ is the desired reconstruction of p. Moreover, the condition number of T_w can be estimated by

$$\text{cond } T_w \leq \left(\frac{1 + 2\delta M}{1 - 2\delta M} \right)^2 . \tag{33}$$

Remarks. 1. The estimate (33) shows clearly how oversampling, in other words, *more information, improves the condition number* of the problem. For instance, with twofold oversampling, i.e. $\delta = \frac{1}{4M}$, cond $T_w \leq 9$ holds *uniformly* over all sampling configurations with maximal gap of length $\frac{1}{4M}$.

For sampling sequences that do not satisfy condition (26) on the maximal gap size we do not know any estimates for the condition number. Numerically we have observed that the problem remains well-conditioned if larger gaps are compensated by an increased number of sampling points in other regions. However, if $r \approx 2M + 1$

and if large gaps occur in the sampling set, then the problem is usually ill-conditioned and condition numbers as large as 10^{14} may occur.

2. Once the coefficients **a** of the trigonometric polynomial p are known, its values can be computed at any point t. In particular, p can be evaluated on *any* grid $\{n/N : n = 0, \ldots N-1\}$ by FFT. All our plots of reconstructions have been obtained in this way.

3. By Caratheodory's theorem [20] every positive definite Toeplitz matrix T can be written as

$$T_{lk} = \sum_{j=1}^{r} w_j \, \mathrm{e}^{2\pi \mathrm{i}(l-k)t_j}$$

for some uniquely determined sequence $w_j > 0$ and $t_j \in [0, 1)$. From this point of view the use of the weights w_j serves to minimize the condition number of a Toeplitz matrix with nodes $0 \leq t_1 < \ldots < t_r < 1$. Although we do not have a formal proof that the particular choice of weights in (27) is optimal, we have been able to verify this claim numerically for a large number of examples. See [18] for an illustration of this effect.

5. Acceleration of iterative algorithms

Up to this point we have analyzed only the mathematical structure of the discrete irregular sampling problem without any effort to produce a numerically efficient algorithm. Obviously the simple Richardson iteration of Lemma 2 and 3 suffices only for small and well-conditioned problems. In our experience data sizes $r \approx 50$ and considerable oversampling can be handled well with the simple iteration scheme. However, in this case the direct solution of the Toeplitz system (33) can be easily performed and iterative methods are not needed.

In interesting applications the amount of data r is much larger, typically $r \approx 10^3 - 10^6$, and the sampling takes place close to the critical density, i.e. $r \approx 2M + 1$ or $\delta \approx \frac{1}{2M}$. An improvement of the Richardson iteration is then indispensable. Fortunately, in this problem all operators S and S_w (11), (28), and the Toeplitz matrices T and T_w (19), (31) are positive definite. Thus a large variety of accelerated algorithms is applicable, see [25, 24, 39, 42]. Here we only discuss the variants of conjugate gradient iteration which is the most powerful and most flexible acceleration method. Moreover, it is particularly well suited for the solution of Toeplitz systems [36, 37].

Proposition 3. *[24] Let A be a positive definite $N \times N$ matrix with smallest eigenvalue λ and largest eigenvalue Λ. Let $x_0 \in \mathbb{C}^N$ be arbitrary, $r_0 = q_0 = b - Ax_0$. For $n \geq 1$ set*

$$x_n = x_{n-1} + \frac{\langle r_{n-1}, q_{n-1} \rangle}{\langle Aq_{n-1}, q_{n-1} \rangle} \, q_{n-1} \tag{34}$$

$$r_n = r_{n-1} - \frac{\langle r_{n-1}, q_{n-1} \rangle}{\langle Aq_{n-1}, q_{n-1} \rangle} \, Aq_{n-1} \tag{35}$$

$$q_n = r_n - \frac{\langle r_n, Aq_{n-1} \rangle}{\langle Aq_{n-1}, q_{n-1} \rangle} \, q_{n-1} \tag{36}$$

Then x_n converges in at most N iterations to the exact solution of $Ax = b$. For $n < N$ the error is at most

$$\|x - x_n\|_A \leq 2 \left(\frac{\sqrt{\Lambda} - \sqrt{\lambda}}{\sqrt{\Lambda} + \sqrt{\lambda}} \right)^n \|x - x_0\|_A \tag{37}$$

where $\|x\|_A = \langle x, Ax \rangle^{1/2}$ *is the* A*-norm of* x.

Consequently we could accelerate each of the simple iterations of Lemma 2, 3, and (4) and (29) [23]. The immediate advantages are:

1. Improvement of the convergence by an order of magnitude.

2. If the error is measured with respect to the operator used in the iteration (S, S_w, T, or T_w), then the *convergence is optimal* in the class of polynomial acceleration algorithms [24].

3. No additional parameter is necessary. In contrast to the Richardson iteration the error estimate (37) does not depend on estimates for the spectrum of A. Bad estimates of the constants in (8) make the simple Richardson iteration notoriously slow and inconvenient. This fact has seriously hampered its use in signal processing. Compare the remark in [32]: "The main problem in choosing a [relaxation parameter] λ is that it cannot be determined theoretically, since A and B are not known. Therefore experimental results will determine the range of λ."

The combination of the adaptive weights method with the conjugate gradient acceleration provides an efficient method for the reconstruction of band-limited signals from non-uniform samples [23]. A detailed discussion of this method can be found in [38].

6. Superfast reconstruction from irregular samples

By combining the reformulation of the original problem as a *Toeplitz system* with the *adaptive weights method* and with the *conjugate gradient acceleration*, we arrive at a fast and efficient reconstruction algorithm. Because of its small storage requirements it is particularly well suited for large data sets. Furthermore it also works well near the critical sampling density and for sampling sets with large gaps.

Since this algorithm is a combination of the adaptive weights method, conjugate gradient acceleration and the use of Toeplitz matrices, we call it ACT algorithm.

Theorem 1 (and Algorithm). *Let* M *be the size of the spectrum and let* $0 \leq t_1 < \ldots < t_r < 1$ *be an* arbitrary *sequence of sampling points with* $r \geq 2M + 1$. *Set* $t_0 = t_r - 1, t_{r+1} = t_1 + 1$ *and* $w_j = \frac{1}{2}(t_{j+1} - t_{j-1})$ *and compute*

$$\gamma_k = \sum_{j=1}^{r} \mathrm{e}^{-2\pi \mathrm{i} k t_j} w_j \quad \textit{for } k = 0, 1, \ldots, 2M \ .$$

The associated Toeplitz matrix has $(T_w)_{lk} = \gamma_{l-k}$ *for* $|l|, |k| \leq M$.

To reconstruct a trigonometric polynomial $p \in \mathscr{P}_M$ *from its samples* $p(t_j)$*, compute first*

$$b_k = \sum_{j=1}^{r} p(t_j) w_j \mathrm{e}^{-2\pi \mathrm{i} k t_j} \quad \textit{for } |k| \leq M \ ,$$

and set $r_0 = q_0 = b \in \mathbb{C}^{2M+1}$, $a_0 = 0$. *Compute iteratively for* $n \geq 1$

$$a_n = a_{n-1} + \frac{\langle r_{n-1}, q_{n-1}\rangle}{\langle T_w q_{n-1}, q_{n-1}\rangle}\, q_{n-1}$$

$$r_n = r_{n-1} - \frac{\langle r_{n-1}, q_{n-1}\rangle}{\langle T_w q_{n-1}, q_{n-1}\rangle}\, T_w q_{n-1}$$

and

$$q_n = r_n - \frac{\langle r_n, T_w q_{n-1}\rangle}{\langle T_w q_{n-1}, q_{n-1}\rangle}\, q_{n-1}\ .$$

Then a_n converges in at most $2M+1$ steps to a vector $\mathbf{a} \in \mathbb{C}^{2M+1}$ solving $T_w\mathbf{a} = \mathbf{b}$. The reconstruction $p \in \mathscr{S}_M$ is then given by $p(t) = \sum_{k=-M}^{M} a_k e^{2\pi i k t}$.

If in addition $\delta < \frac{1}{2M}$ and $p_n(t) = \sum_{k=-M}^{M} a_{n,k} e^{2\pi i k t} \in \mathscr{S}_M$ denotes the approximating polynomial after n iterations, then

$$\left(\sum_{j=1}^{r} |p(t_j) - p_n(t_j)|^2 w_j\right)^{1/2} \le 2(2\delta M)^n \left(\sum_{j=1}^{r} |p(t_j)|^2 w_j\right)^{1/2} \tag{38}$$

Since we have already discussed all steps that lead to this algorithm, we only need to verify (38). Given $p \in \mathscr{S}_M, p(t) = \sum_{k=-M}^{M} a_k e^{2\pi i k t}$ with $\mathbf{a} \in \mathbb{C}^{2M+1}$ and S_w as in (28), we obtain

$$\sum_{j=1}^{r} |p(t_j)|^2 w_j = \langle S_w p, p\rangle = \langle T_w a, a\rangle = \|a\|_{T_w}^2\ .$$

(38) follows from (37), since (27) implies

$$(1 - 2\delta M)^2 \le \lambda \le \Lambda \le (1 + 2\delta M)^2\ .$$

Further discussion

1. The T_w or S_w-norm in (38) is the correct norm to measure the error. In contrast to the previous error estimates in the $\|\cdot\|_2$-norm, (38) uses *only the given data* to compare the quality of the n-th iteration with the initial data. We emphasize once more that without the use of weights no estimate of the form (38) is known.
2. The ACT method can be applied to *any* given sampling set with sufficiently many samples. However, lacking knowledge about the spectrum of T_w, we cannot predict the quality of the approximation. Under the additional condition on the maximal gap length the error estimate (38) yields a *guaranteed* rate of convergence. In general the CG acceleration is significantly faster than in (38), both on account of the fact that $(1 \pm 2\delta M)^2$ are just estimates for the extrema of the spectrum of T_w, and also because the convergence depends on the distribution of the singular values of T_w and not just on the extrema occurring in (37).
3. If we start with data $(t_j, y_j)_{j=1,\dots,r}$, where the y_j's are not necessarily samples of a trigonometric polynomial, then the algorithm – in fact all procedures discussed so far – computes the trigonometric polynomial $p \in \mathscr{S}_M$ which minimizes the expression

$$\sum_{j=1}^{r} |p(t_j) - y_j|^2 w_j$$

The algorithm can therefore also be used for the fitting and least square approximation of data by trigonometric polynomials. Compare also [34].

4. In many problems arising in practice – such as reconstruction of lost samples in a CD-signal – the sampling sequence is a subsequence of an arithmetic progression. In this case the entries of the Toeplitz matrix T_w can be computed particularly fast, as explained in the following result:

Corollary 1. *Assume that the sampling sequence* $0 \leq t_1 < \ldots < t_r \leq 1$ *is a subsequence of the arithmetic progression* $0, \frac{1}{N}, \frac{2}{N}, \ldots, \frac{N-1}{N}$. *For* $n = 0, \frac{1}{N}, \ldots, \frac{N-1}{N}$ *define* u_w *and* u_b *by*

$$u_w(n) = \begin{cases} w_j & \text{if } n = t_j \\ 0 & \text{else} \end{cases} \tag{39}$$

and

$$u_p(n) = \begin{cases} p(n) \cdot w_j & \text{if } n = t_j \\ 0 & \text{else.} \end{cases} \tag{40}$$

Then the entries $\gamma_0, \gamma_1, \ldots, \gamma_{2M}$ *of the Toeplitz matrix* T_w *and the vector* **b** *can be computed via Fourier transform by*

$$\gamma_k = \hat{u}_w(k) \quad \text{for } k = 0, 1, \ldots, 2M \quad \text{and} \quad b_k = \hat{u}_p(k) \quad \text{for } k \leq |M|. \tag{41}$$

Proof. Since

$$\gamma_k = \sum_{j=1}^{r} w_j \mathrm{e}^{-2\pi \mathrm{i} t_j k} \qquad \text{for } k = 0, 1, \ldots, 2M$$

we can write γ_k in the form

$$\gamma_k = \sum_{n=0}^{N-1} u_w(n) \mathrm{e}^{-2\pi \mathrm{i} n k/N} = \hat{u}_w(k).$$

Part two of assertion (41) for **b** can be obtained in a similar way. □

7. Efficient implementation of the algorithm

An important advantage of Toeplitz matrices is that they can be inverted with considerably less computational effort than matrices without special structure. A number of so-called "superfast" direct inversion methods [1, 10] have been created in the last ten years. However the stability of these fast direct solvers is still a problem [5]. Furthermore, since in many applications a solution is required only with a certain accuracy, but not the exact solution, we prefer to use iterative methods.

In [36] Strang proposed an iterative method for solving positive definite Hermitean Toeplitz systems. In this method the Toeplitz matrix is preconditioned by a *circulant matrix*. The entries of a circulant matrix C of size $n \times n$ satisfy $c_{ij} = c_{i-j} = c_{i-j+n}$. The multiplication of a vector by a circulant C is identical to discrete (cyclic) convolution of that vector by the generating sequence c. In other words the linear system $Ca = b$ is the same as the convolution equation $c * a = b$, where c is the first column of C. Applying the discrete Fourier transform to this equation it becomes $\hat{c}\hat{a} = \hat{b}$. Therefore $\hat{a}$ is given by a component-by-component division, and a is recovered from $\hat{a}$ by inverse Fourier transform [36, 9].

Multiplication of a vector a by a Toeplitz matrix T can also be carried out quickly by means of an appropriate FFT. The $n \times n$ matrix T is extended to a circulant $\tilde{T}$ of

size $(2n-1)\times(2n-1)$, and the vector a is completed to $\tilde{a}$ by $n-1$ zeros. Then Ta appears in the first n components of $\tilde{T}\tilde{a}$. Although $\tilde{T}$ is larger than T, this matrix action can be performed much faster because it is just discrete convolution and is carried out using the FFT.

Since the main calculation in the ACT algorithm of Theorem 1 is a matrix-vector multiplication as described above, each iteration requires roughly $O(2n\log 2n)$ operations. It is often profitable to augment the matrix $\tilde{T}$ somewhat by zeropadding to a length for which the FFT is efficient, i.e. a length with many small prime factors. We demonstrate the augmentation by the first row of $\tilde{T}$:

$$(\gamma_0,\gamma_1,\ldots,\gamma_{n-1},0,\ldots,0,\bar{\gamma}_{n-1},\ldots,\bar{\gamma}_2,\bar{\gamma}_1).$$

Observe that we do not claim that zeropadding is compatible with the FFT. We only make use of the fact that the enlargement of the Toeplitz matrix in the form described above does not influence the first n components of the solution a.

If the signal is sampled near the critical density or if there are many large gaps in the sampling set, it may happen that even the conjugate gradient method converges rather slowly. To remedy this problem one usually turns to the preconditioned system $C^{-1}Ta = C^{-1}b$ [24, 25, 36]. If the preconditioner C is chosen appropriately, the preconditioned conjugate gradient method (referred to as PCG) will converge much faster than the original CG method. Various efficient preconditioners have been developed in the last ten years, see [7, 29, 27, 8] for detailed discussions of preconditioners. In the present paper we use the optimal circulant Frobenius norm approximation C_F introduced by T. Chan [8]. We have mentioned in Sect. 4 that the use of weights serves to minimize the condition number of the Toeplitz matrix, if the sampling set satisfies the Nyquist criterion. In Sect. 8 below we shall discuss the effect of applying T. Chan's preconditioner to the use of weighted Toeplitz matrices and consider a combination of both.

Remark. If many signals of the same bandwidth have to be reconstructed from the same sampling geometry, it is useful to establish the inverse of the Toeplitz matrix in the Gohberg-Semencul formula [19] once, which can be easily done by a slight modification of ACT. The reconstruction of the signals can then be done considerably faster [28, 38, 17].

8. Numerical results

In this section we will discuss some numerical results of the proposed algorithm. We will demonstrate the efficiency of the ACT algorithm and illustrate how appropriate preconditioning may allow to reconstruct signals from sampling sets having many large gaps.

We consider a synthetic signal of length 8192 with bandwidth 500 and randomly generated Fourier coefficients. The sampling set satisfies the Nyquist criterion and consists of about 2300 samples. Since we know the actual signal x, we can measure the error $\|x-x_n\|_2/\|x\|_2$ between x and the approximation x_n of the n-th iteration.

We apply the Marvasti method of Lemma 2, the adaptive weights conjugate gradient method – which can be derived by a combination of Lemma 2, Proposition 1 and Proposition 3 – and the ACT method presented in Theorem 1 to this situation. We measure the error and required time for 25 iterations for each method.

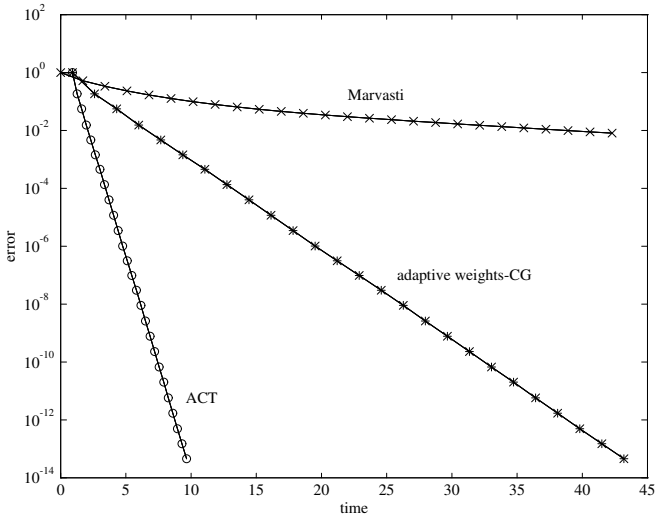

Fig. 1. Comparison of required time for 25 iterations

One can see in Fig. 1 that the Marvasti method with optimal relaxation parameter is very inefficient, both in time and rate of convergence. The adaptive weights-CG method requires only negligibly more time than the Marvasti method, but the rate of convergence is higher by several orders of magnitude. By construction ACT and the adaptive weights-CG method deliver the same error after each iteration if we assume exact arithmetics. However since the bandwidth of the signal is small compared to its length, ACT needs considerably less computation time. The overall improvement of ACT compared to the Marvasti method is significant and very convincing.

In Fig. 2 we illustrate the required total number of floating point operations (FLOPS) for complete numerical reconstruction of the signal. We compare the three methods above and the adaptive weights method, which can be derived by combining Lemma 2 and Proposition 1. Although the ordinary adaptive weights method needs about 5 times less FLOPS than the Marvasti method, the computational effort is further reduced by the ACT method by more than 7 times. These plots demonstrate drastically that the methods used previously in engineering are quite inadequate for signals of this size.

To support the claim that the use of weights minimizes the condition number of the Toeplitz matrix, we consider the number of iterations required to obtain complete reconstruction of the signal. We compare the ACT method to the preconditioned version of the ACT method (denoted as "PACT" in Fig. 3) as well as to the method of Lemma 4, accelerated by CG and PCG (labeled as "CG" and "PCG" in the plot). We use the same signal and sampling set as above. Figure 3 shows that ACT and PACT terminate after about half the number of iterations as required by the CG method applied to the unweighted Toeplitz matrix of Lemma 4. Furthermore the rate of convergence of ACT is better than that of solving the Toeplitz system of Lemma 4 by PCG. This result supports the claim that the singular values of the weighted Toeplitz matrix of Proposition 2 are better clustered than those of the unweighted but preconditioned Toeplitz matrix. Observe that all methods terminate within 45 iterations, although the rank of the Toeplitz matrix is 1001.

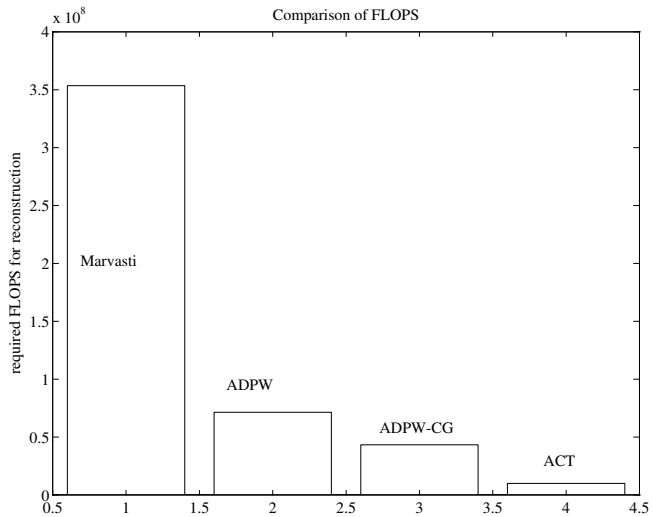

Fig. 2. Comparison of floating point operations

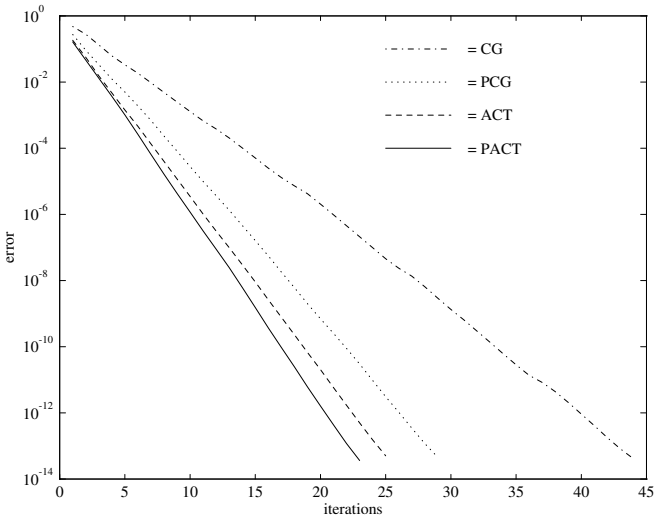

Fig. 3. Comparison of rate of convergence

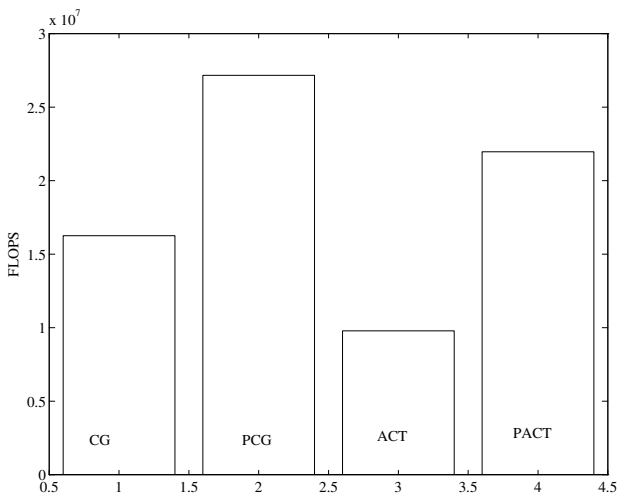

Fig. 4. Comparison of floating point operations

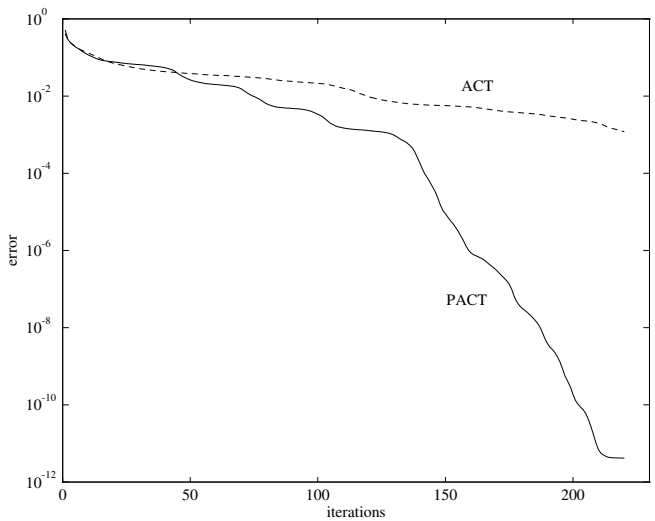

Fig. 5. Convergence rate of ACT and PACT

In Fig. 4 we consider the required floating point operations for the four methods above for complete reconstruction of the signal. Although ACT needs slightly more iterations than PACT (see Fig. 3), it requires significantly less FLOPS.

The last figure illustrates a typical "critical" situation, where the standard methods are by no means able to reconstruct the signal. We consider the same signal as above, but a sampling set consisting of 2210 points. Approximately 1% of the gaps between the sampling points are 2-3 times larger than the Nyquist interval, i.e. 2-3 larger than $\frac{N}{2M+1}$, and about 10% of the gaps are slightly larger than $\frac{N}{2M+1}$.

All standard methods diverge in this situation and are therefore not shown in Fig. 5. Although the ACT method is convergent, the rate of decay is very poor. In such a situation it is useful to apply the preconditioned version of the ACT algorithm (denoted by "PACT" in the plot) to obtain a better rate of convergence. The preconditioned ACT algorithm delivers complete reconstruction after about 200 iterations, while all standard methods fail to reconstruct the signal.

References

1. Ammar, G., Gragg, W. (1988): Superfast solution of real positive definite Toeplitz systems. SIAM J. Matrix Anal. Appl. **9**, 61–76
2. Benedetto, J., Heller, W. (1990): Irregular sampling and the theory of frames, I. Mat. Note **10**, 103–125
3. Bitmead, R.R., Anderson, B.D. (1980): Asymptotically fast solution of Toeplitz and related systems of linear equations. Lin. Alg. Appl. **34**, 103–117
4. Björk, A.A., Pereyra, V. (1970): Solution of Vandermonde systems of equations. Math. Comp. **24**, 893–903
5. Bunch, J.R. (1985): Stability of methods for solving Toeplitz systems of equations. SIAM J. Sci. Statis. Comput. **6**, 349–364
6. Butzer, P.L., Splettstößer, W., Stens, R.L. (1988): The sampling theorem and linear prediction in signal analysis. Jahresbericht der DMV 90, 1–70
7. Chan, R., Nagy, J.G., Plemmons, R.J. (1989): Circulant preconditioned Toeplitz Least Squares Iterations. SIAM J. Matrix Anal. **10**, 542–550
8. Chan, T. (1989): An optimal circulant preconditioner for Toeplitz systems. SIAM J. Sci. Stat. Comput. **9**, 766–771
9. Davis, P.J. (1979): Circulant Matrices. John Wiley
10. De Hoog, F. (1987): A new algorithm for solving Toeplitz systems of equations. Linear Alg. Appl. **88/89**, 349–364
11. Duffin, R., Schaeffer, A. (1952): A class of nonharmonic Fourier series. Trans. Amer. Math. Soc. **72**, 341–366
12. Feichtinger, H.G., Cenker, C., Herrmann, M. (1991): Iterative algorithms in irregular sampling: A first comparison of methods. Conf. ICCCP'91, March 1991, Phoenix, USA, 483–489
13. Feichtinger, H.G., Cenker, C., Steier, H. (1991): Fast iterative and non-iterative reconstruction methods in irregular sampling. Conf. ICASSP'91, May, Toronto, 1773–1776
14. Feichtinger, H.G., Gröchenig, K. (1992): Irregular sampling theorems and series expansions of band-limited functions. J. Math. Anal. Appl **167**, 530–556
15. Feichtinger, H.G., Gröchenig, K. (1992): Iterative reconstruction of multivariate band-limited functions from irregular sampling values. SIAM J. Math. Anal. **231**, 244–261
16. Feichtinger, H.G., Gröchenig, K: (1993) Error analysis in regular and irregular sampling theory. Applicable Analysis **50**, 167–189
17. Feichtinger, H.G., Strohmer, T. (1993): Efficient Band-Limited Signal extrapolation. In preparation
18. Feichtinger, H.G., Gröchenig, K.H. (1993): Theory and Practice of Irregular Sampling, In: Benedetto, J., Frazier M., eds., Wavelets: Mathematics and Applications, pp. 305–363. CRC Press
19. Gohberg, I.C., Semencul, A.A. (1972): On the inversion of finite Toeplitz matrices and their continuous analogs. Mat. Issled. **2**, 201–233
20. Grenander, U., Szegö, G. (1958): Toeplitz forms and their applications. Univ. California Press
21. Gröchenig, K. (1992): Reconstruction algorithms in irregular sampling. Math. Comp. **59**, 181–194

22. Gröchenig, K. (1993): A discrete theory of irregular sampling. Lin. Alg. and Appl. **193**, 129–150
23. Gröchenig, K. (1993): Acceleration of the frame algorithm. IEEE Trans. Signal Proc., Special Issue on Wavelets **41** (12), 3331–3340
24. Hackbusch, W. (1991): Iterative Lösung großer schwachbesetzter Gleichungssysteme. Teubner Studienbücher
25. Hageman, L.A., Young, D.M. (1981): Applied Iterative Methods. Academic Press
26. Higgins, J.R. (1985): Five short stories about the cardinal series. Bull. Am. Math.Soc. **12**, 45–89
27. Huckle, T. (1992): Some Aspects of Circulant Preconditioners. Preprint
28. Jain, J.R. (1979): An efficient algorithm for a large Toeplitz system of equations. IEEE Trans. ASSP **27**, 612–615
29. Ku, T.K., Kuo, C.J. (1992): Design and analysis of Toeplitz preconditioners. IEEE Trans. SSP **40**, 129–141
30. Marks II, R.J. (1993): Advanced Topics in Shannon Sampling and Interpolation Theory. Springer Verlag
31. Marvasti, F.A. (1987): A unified approach to zero-crossing and nonuniform sampling of single and multi-dimensional systems. Nonuniform, P.O.Box 1505, Oak Park, IL 60304
32. Marvasti, F.A., Analoui, M. (1989): Recovery of signals from nonuniform samples using iterative methods. In: Proc. Int. Symp. Circuits Syst., Portland, OR
33. Papoulis, A. (1977): Signal analysis. McGraw-Hill, New York
34. Reichel, L., Ammar, G., Gragg, W. (1991): Discrete Least Squares Approximation by Trigonometric Polynomials. Math. Comp. **57**, 273–289
35. Sauer, K.D., Allebach, J.P. (1987): Iterative reconstruction of band-limited images from nonuniformly spaced samples. IEEE Trans. CAS-34/12
36. Strang, G. (1986): A proposal for Toeplitz matrix calculations. Stud. Appl. Math. **74**, 171–176
37. Strang, G., Chan, R.H. (1989): Toeplitz Equations by Conjugate Gradients with Circulant Preconditioner. SIAM J. Sci. Stat. Comput. **10**, 104–119
38. Strohmer, T. (1993): Efficient Methods for Digital Signal and Image Reconstruction from Nonuniform Samples. PhD thesis, University of Vienna
39. Varga, R.S. (1962): Matrix Iterative Analysis. Prentice-Hall
40. Wiley, R.G. (1978): Recovery of bandlimited signals from unequally spaced samples. IEEE Trans. Com. **26**(1), 135–137
41. Yeh, S., Stark, H. (1990): Iterative and one-step reconstruction from nonuniform samples by convex projections. J. Opt. Soc. Amer. A **7**(3), 491–499
42. Young, D.M. (1971): Iterative Solution of Large Linear Systems. Academic Press, New York
43. Young, R. (1980): An Introduction to Nonharmonic Fourier Series. Academic Press, New York
44. Zygmund, A. (1959): Trigonometric Series, Vol. II. Cambridge Univ. Press

This article was processed by the author using the LaTeX style file *pljour1* from Springer-Verlag.