Class notes

Numerical Analysis

Prof. Annalisa Buffa

2021-2022

Chapter 6

Discrete Fourier transform and data compression

6.1 Introduction

The goal of this chapter is to present some of the most successful applications of Fourier analysis, by illustrating a point of view and a language typical of physicists and applied mathematicians. In particular, we will introduce the reader to the concepts of discrete Fourier and cosine transforms and show how they may be used for audio and image compression, respectively. Note that these applications not only have an interest in themselves, but can also offer a test field for Fourier analysis, and thus contribute to a better understanding of the latter. Moreover, we will present the so-called fast Fourier transform, a breakthrough algorithm that reduced the complexity, for the computation of the discrete Fourier transform, from $\mathcal{O}\left(N^2\right)$ to $\mathcal{O}\left(N\log N\right)$, where N is the size of the data. The chapter is concluded with an application of the discrete Fourier transform to interpolation. We will see how to interpolate periodic functions using partial sums of Fourier series, or trigonometric polynomials. Error estimates will be provided as well.
Let us start with some recalls on fundamental concepts of Fourier analysis that we are going to employ throughout the chapter.

6.2 A few things about Fourier analysis

Definition 6.1. *Let us fix a constant $T > 0$. We say that a function $f : \mathbb{R} \to \mathbb{R}$ is periodic of period T or T-periodic if*

$$f(x+T) = f(x) \qquad \forall\, x \in \mathbb{R}.$$

Remark 6.1.

- *Note that a T-periodic function is trivially kT-periodic for every $k > 1$.*
- *Every periodic continuous non-constant function has a minimal period, called fundamental period. The functions $\sin(x)$, $\cos(x)$, e^{ix} (Definition 6.1 applies to complex functions too) are periodic with fundamental period 2π.*

- *Given $a \in R$ and $T > 0$, any function $f : [a, a+T) \to \mathbb{R}$ can be extended in a unique way to a T-periodic function $\tilde{f} : \mathbb{R} \to \mathbb{R}$. Since every real number can be represented in the form $t + kT$ for $t \in [a, a+T)$ and $k \in \mathbb{Z}$, it suffices to define $\tilde{f}(t+kT) := f(t)$ for every $t \in [a, a+T)$ and every $k \in \mathbb{Z}$.*
- *Let us remark that the periodic functions of a given period form a vector space.*

Let $f : \mathbb{R} \to \mathbb{R}$ be 2π-periodic. We can identify it with its restriction to any interval of length 2π. Let us take, for instance, $[0, 2\pi]$.

Remark 6.2. *When talking about $f : \mathbb{R} \to \mathbb{R}$ 2π-periodic, we may write for the sake of simplicity "$f : [0, 2\pi] \to \mathbb{R}$ 2π-periodic".*

From the Analysis courses you know that, under some special conditions, it is possible to write a given function $f : [0, 2\pi] \to \mathbb{R}$ 2π-periodic through a series of trigonometric polynomials, namely

$$f(x) = \frac{a_0}{2} + \sum_{k=1}^{\infty} a_k \cos(kx) + b_k \sin(kx) \qquad \forall\ x \in \mathbb{R}. \tag{6.1}$$

Expression (6.1) is called Fourier series associated to f. In order to give a precise meaning to the writing (6.1), which for the moment is purely formal, let us assume that $f \in C^0([0, 2\pi])$ 2π-periodic and that the Fourier series associated to f is uniformly convergent. Using Euler formulas

$$e^{ix} = \cos(x) + i \sin(x), \quad \cos(x) = \frac{e^{ix} - e^{-ix}}{2}, \quad \sin(x) = \frac{e^{ix} + e^{-ix}}{2i},$$

and denoting

$$c_0 := \frac{a_0}{2}, \qquad c_k := \frac{a_k - i b_k}{2}, \qquad c_{-k} := \frac{a_k + i b_k}{2} \qquad \forall\ k \in \mathbb{N} \setminus \{0\}, \tag{6.2}$$

let us rewrite (6.1) in the more compact form

$$f(x) = \sum_{k=-\infty}^{+\infty} c_k e^{ikx} \qquad \forall\ x \in \mathbb{R}. \tag{6.3}$$

Remark 6.3. *It is worth noting that while $(c_k)_{k \in \mathbb{Z}} \subset \mathbb{C}$, expression (6.3) is still real!*

Lemma 6.1 (Orthogonality relations)**.** *For every $k, \ell \in \mathbb{Z}$,*

$$\int_0^{2\pi} e^{-i\ell x} e^{ikx} \, \mathrm{d}x = \begin{cases} 0 & \text{if } k \neq \ell, \\ 2\pi & \text{if } k = \ell. \end{cases} \tag{6.4}$$

Proof. See the exercise sessions. □

Remark 6.4. *From Lemma 6.1 and the fact that $\operatorname{span}\left(e^{ikx}\right)_{k \in \mathbb{Z}}$ is dense in $L^2(0, 2\pi)$ (a classical result in Fourier Analysis), it follows that $\left(e^{ikx}\right)_{k \in \mathbb{Z}}$ is an orthogonal basis for $L^2(0, 2\pi)$.*

Let us now recover an explicit formula for the coefficients $(c_k)_{k\in\mathbb{Z}}$ appearing in (6.3). We multiply (6.3) by $e^{-i\ell x}$ and integrate from 0 to 2π.

$$\int_0^{2\pi} f(x)e^{-i\ell x}\,\mathrm{d}x = \sum_{k=-\infty}^{+\infty} c_k \int_0^{2\pi} e^{-i\ell x}e^{ikx}\,\mathrm{d}x.$$

Note that since we assumed the Fourier series of f to be uniformly convergent, we can freely exchange integrals and limit processes. Using Lemma 6.1, we conclude

$$c_k = \frac{1}{2\pi}\int_0^{2\pi} f(x)e^{-ikx}\,\mathrm{d}x, \qquad \forall\ k\in\mathbb{Z}. \tag{6.5}$$

$(c_k)_{k\in\mathbb{Z}}$ are the Fourier coefficients of f. Let us recall the definition of probably the most well-known integral transform.

Definition 6.2. *Given $f\in L^1(0,2\pi)$, we define the Fourier transform of f as*

$$\hat{f}(\xi) := \frac{1}{2\pi}\int_0^{2\pi} f(x)e^{-i\xi x}\,\mathrm{d}x \qquad \forall\ x\in\mathbb{R}. \tag{6.6}$$

Remark 6.5. *Comparing (6.5) and (6.6), we observe that $c_k = \hat{f}(k)$, $\forall\ k\in\mathbb{Z}$.*

The next result will be useful when we study the trigonometric interpolation. From the classical Riemann-Lebesgue Lemma we know that $\hat{f}(k)$ goes to 0 as $|k|\to+\infty$. If some extra regularity on f is assumed, then it is possible to estimate the velocity of the decay of its Fourier coefficients.

Proposition 6.2. *Let $f\in C^m([0,2\pi])$ be 2π-periodic and such that $f^{(m+1)}$ is integrable on $(0,2\pi)$, then*

$$\left|\hat{f}(k)\right| \le C_m |k|^{-m-1} \qquad \forall\ k\in\mathbb{Z},$$

where $C_m := \frac{1}{2\pi}\int_0^{2\pi}\left|f^{(m+1)}(x)\right|\,\mathrm{d}x$.

Proof. Given $k\in\mathbb{Z}$, let us compute

$$\hat{f}(k) = \frac{1}{2\pi}\int_0^{2\pi} f(x)e^{-ikx}\,\mathrm{d}x = \frac{1}{2\pi ik}\int_0^{2\pi} f'(x)e^{-ikx}\,\mathrm{d}x = \frac{1}{ik}\widehat{f'}(k),$$

where the boundary term coming from the integration by parts vanishes since f and e^{-ikx} are 2π-periodic. Hence, we get

$$\left|\hat{f}(k)\right| \le \frac{1}{2\pi|k|}\int_0^{2\pi}\left|f'(x)\right|\,\mathrm{d}x. \tag{6.7}$$

By reiterating the procedure above by integration by parts, it is possible to show

$$\hat{f}(k) = \frac{1}{(ik)^{m+1}}\widehat{f^{(m+1)}}(k),$$

so that by definition of $\widehat{f^{(m+1)}}$, analogously to (6.7), we have

$$\left|\hat{f}(k)\right| \leq \frac{1}{2\pi\,|k|^{m+1}} \int_0^{2\pi} \left|f^{(m+1)}(x)\right| \,\mathrm{d}x.$$

□

6.3 Discrete Fourier transform (DFT)

Now, we would like to mimic what we have been doing in the previous section, but in the discrete setting. In particular, we are looking for a discrete analogue of formula (6.6). Let us assume, indeed, the function f to be known just at some discretization points of $[0,2\pi]$ and denote

$$x_j = \frac{2\pi j}{N}, \qquad y_j = f(x_j), \qquad \forall\ j = 0,1,\dots,N-1. \tag{6.8}$$

By analogy with (6.3), let us write

$$y_j = \sum_{k=0}^{N-1} z_k e^{ikx_j} = \sum_{k=0}^{N-1} z_k \omega_N^{kj}, \qquad \forall\ j = 0,1,\dots,N-1, \tag{6.9}$$

with $\omega_N := e^{\frac{i2\pi}{N}}$.

Remark 6.6. *Let us recall that $\left(\omega_N^k\right)_{k=0}^{N-1}$ form an abelian group, called the group of N-th roots of unit, since $\left(\omega_N^k\right)^N = 1$ for $k = 0,\dots N-1$.*

When needed, it is possible to prolongate the sequence $\left(y_j\right)_{j=0}^{N-1}$ by defining $\tilde{y}_{j+kN} := y_j$ for every $j = 0,\dots,N-1$ and every $k \in \mathbb{Z}$. With a slight abuse of notation in that case we may write $\left(y_j\right)_{j\in\mathbb{Z}}$ instead of $\left(\tilde{y}_j\right)_{j\in\mathbb{Z}}$.
At this point we want to find $(z_k)_{k=0}^{N-1}$ such that (6.9) holds. Let us mimic the procedure employed in the continuous case in order to recover the Fourier coefficients $(c_k)_{k\in\mathbb{Z}}$ starting from (6.3).

Lemma 6.3 (Discrete orthogonality relations)**.** *For every $\ell, k = 0,1,\dots,N-1$,*

$$\sum_{j=0}^{N-1} \omega_N^{-\ell j}\omega_N^{kj} = \begin{cases} 0 & \text{if} \quad k \neq \ell, \\ N & \text{if} \quad k = \ell. \end{cases} \tag{6.10}$$

Proof. See the exercise sessions. □

Let us multiply (6.9) by $\omega_N^{-\ell j}$ and take the sum over $j = 0,\dots,N-1$.

$$\sum_{j=0}^{N-1} y_j \omega_N^{-\ell j} = \sum_{j=0}^{N-1}\sum_{k=0}^{N-1} z_k \omega_N^{-\ell j}\omega_N^{kj}.$$

By using Lemma 6.3, we obtain

$$z_k = \frac{1}{N}\sum_{j=0}^{N-1} y_j \omega_N^{-kj} \qquad \forall\ k = 0,\dots,N-1, \tag{6.11}$$

the discrete counterpart of (6.5). Because of the analogy with (6.5) and Remark 6.5, we may denote $\hat{f}_N(k) := z_k$, for $k = 0,\dots,N-1$.

Remark 6.7. *Note that formula (6.11) could also be obtained by applying the trapezoidal quadrature rule to approximate (6.5) using as integration nodes $(x_j)_{j=0}^{N-1}$ (see the exercise sessions).*

Since expression (6.11) can be evaluated for any $k \in \mathbb{Z}$, we can write $\left(\hat{f}_N(k)\right)_{k\in\mathbb{Z}}$ when needed.

Definition 6.3 (Discrete Fourier transform)**.** *$\left(\hat{f}_N(k)\right)_{k\in\mathbb{Z}}$ is called discrete Fourier transform (DFT) of f with respect to the discretization $(x_j)_{j=0}^{N-1}$.*

Proposition 6.4. *The DFT $\left(\hat{f}_N(k)\right)_{k\in\mathbb{Z}}$ is an N-periodic sequence.*

Proof. Let us fix $k \in \{0,1,\dots,N-1\}$ and recall that $\omega_N^{Nm} = 1$ for every $m \in \mathbb{Z}$. We have

$$\hat{f}_N(k+N) = \frac{1}{N}\sum_{j=0}^{N-1} y_j \omega_N^{-(k+N)j} = \frac{1}{N}\sum_{j=0}^{N-1} y_j \omega_N^{-kj} = \hat{f}_N(k).$$

□

Remark 6.8. *With Proposition 6.4 in mind, we observe that $\hat{f}_N(k)$ is not an approximation of $\hat{f}$ for every $k \in \mathbb{Z}$: recall that $\hat{f}(k) \to 0$ as $|k| \to +\infty$.*

Since we have made sure that the expression (6.9) makes sense, we can give the following definition.

Definition 6.4 (Inverse Discrete Fourier Transform)**.** *The sequence $(y_j)_{j\in\mathbb{Z}}$ is called inverse discrete Fourier transform (IDFT) of f with respect to the discretization $(x_j)_{j=0}^{N-1}$.*

Example 6.1 (Sound digitization and data compression)**.** *With this example we try to provide a "very" simplified strategy of compression for audio signals. Figure 6.1[1] shows the digitization (or discretization) of a sound which can be thought as function $f(t)$ of time. The sound considered has been recorded with a sample rate of 22000 Hz (the unit measuring pulses per second) of which 1024 are represented. At this point we can compute the DFT of $(y_j)_{j=0}^{1023}$ using formula (6.11). The scope of data compression is to reduce the size of the DFT without losing the essential information, i.e. in such a way that it is still possible to reconstruct the original audio using the inverse DFT (6.9). In particular, we are interested in the magnitude or amplitude of the DFT $\left(\left|\hat{f}_N(k)\right|\right)_k$. Recall, from Proposition 6.4, that $\left(\hat{f}_N(k)\right)_k$ is an N-periodic sequence, so, since $(y_j)_j \subset \mathbb{R}$, it holds $z_{N-k} = z_{-k} = \overline{z}_k$. Since starting from $k = 170$ the coefficient $\left|\hat{f}_N(k)\right|$ is very small, we plot just the first 169 coefficients in the second picture of Figure 6.1. Note that the theory developed in this section requires the analyzed function or signal to be periodic. However, from Figure 6.1[1], we perceive some*

kind of periodicity, but we can clearly see that the period is not $N = 1024$*. By analyzing the data one realizes that* $y_{N+n} - y_n$ *gets very small for* $N = 997$*, hence we decided that this is the period of our signal. We can restrict the signal to a period of* $N = 997$*, interpolate it with a piecewise linear polynomial and extract other* 1024 *new values of the signal. This time it is much easier to identify the fundamental frequency or harmonic from its spectrum, see Figure 6.1*[3]*. Now, let us get rid of all the discrete Fourier coefficients* $\hat{f}_N(k)$ *such that* $\left|\hat{f}_N(k)\right| < 0.1 \max_{0\le k\le 169} \left|\hat{f}_N(k)\right|$*. In this way we are left with just 16 coefficients, see Figure 6.1*[4]*, starting from 997 real values* $\left(y_j\right)_j$*. Let us decode the audio signal by using formula (6.9): the result is shown in Figure 6.1*[5]*. By comparing the first and the last pictures of Figure 6.1, the reader can observe that the original signal has been faithfully reconstructed. Instead of storing the 997 values of the original signal, it is enough to store a few coefficients of the discrete Fourier transform without losing visible information.*

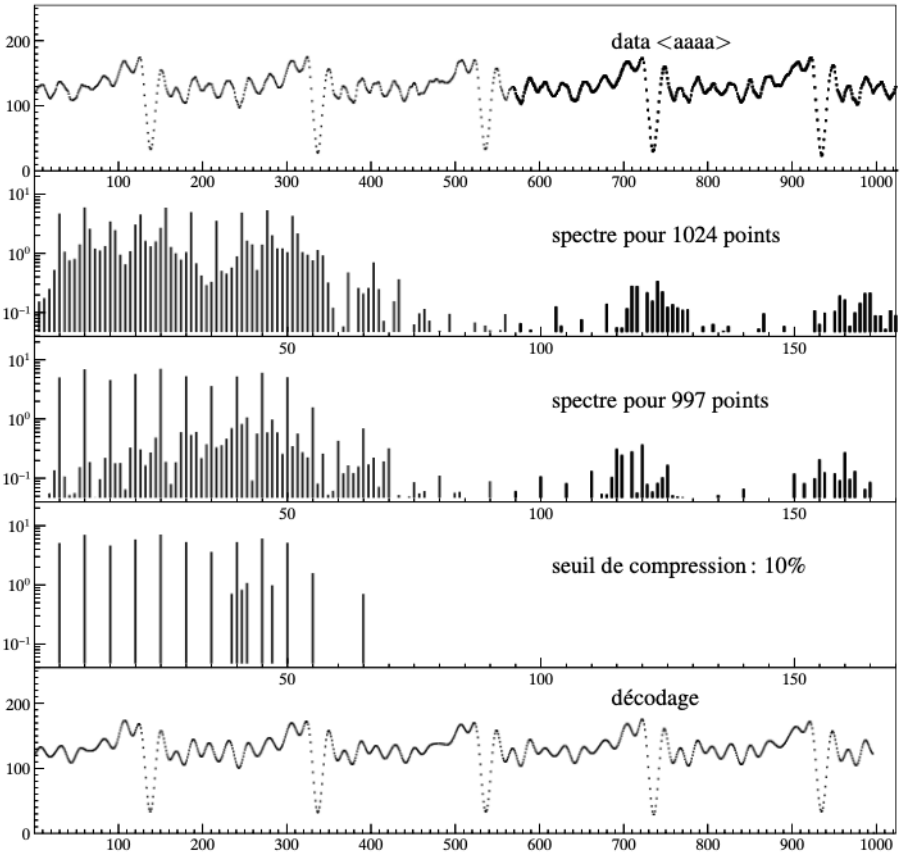


Figure 6.1: Digitization and compression of an audio signal.

6.4 Discrete cosine transform (DCT)

In the previous section we have been focusing on periodic functions. This time the starting point is an arbitrary continuous function $\tilde{f} \in C^0([0,\pi])$. We can extend it to an even function as follows.

Let

$$f_p(x) := \begin{cases} \tilde{f}(x) & \text{if } x \in [0,\pi], \\ \tilde{f}(-x) & \text{if } x \in [-\pi,0). \end{cases}$$

And then to a 2π-periodic function as

$$f(x+2k\pi) := f_p(x) \qquad \forall\ x \in [-\pi,\pi], \forall\ k \in \mathbb{Z}.$$

Since f is even and 2π-periodic, its Fourier series, if convergent, reads as

$$f(x) = \frac{a_0}{2} + \sum_{k=1}^{\infty} a_k \cos(kx) \qquad \forall\ k \in \mathbb{Z}. \tag{6.12}$$

We assume that in (6.12) we have uniform convergence. The following orthogonality relations hold.

Lemma 6.5. *For every $\ell, k \in \mathbb{Z}$*

$$\int_0^{\pi} \cos(\ell x)\cos(kx)\,\mathrm{d}x = \begin{cases} 0 & \text{if} \quad \ell \neq k, \\ \frac{\pi}{2} & \text{if} \quad \ell = k \neq 0, \\ \pi & \text{if} \quad \ell = k = 0. \end{cases}$$

Let us recover the Fourier coefficients $(a_k)_{k\in\mathbb{Z}}$. By multiplying (6.12) by $\cos(\ell x)$, integrating from 0 to π and using Lemma 6.5, we get

$$\int_0^{\pi} f(x)\cos(\ell x)\,\mathrm{d}x = \frac{a_0}{2}\int_0^{\pi} \cos(\ell x)\,\mathrm{d}x + \sum_{k=1}^{\infty} a_k \int_0^{\pi} \cos(kx)\cos(\ell x)\,\mathrm{d}x,$$

thus

$$a_k = \frac{2}{\pi}\int_0^{\pi} f(x)\cos(kx)\,\mathrm{d}x \qquad \forall\ k \in \mathbb{N}. \tag{6.13}$$

We assume the target function f to be known just at some discretization points of $[0,\pi]$, hence we define

$$x_j := \frac{(2j+1)\pi}{2N}, \qquad y_j := f(x_j) \qquad \forall\ j = 0,\dots,N-1. \tag{6.14}$$

The discrete counterpart of (6.12) would be

$$y_j = \frac{z_0}{2} + \sum_{k=1}^{N-1} z_k \cos(kx_j), \qquad \forall\ j = 0,\dots,N-1. \tag{6.15}$$

The following discrete orthogonality relations hold.

Lemma 6.6. *For all $\ell, k \in \mathbb{Z}$*

$$\sum_{j=0}^{N-1} \cos(\ell x_j)\cos(k x_j) = \begin{cases} 0 & \text{if} \quad \ell \neq k, \\ \frac{N}{2} & \text{if} \quad \ell = k \neq 0, \\ N & \text{if} \quad \ell = k = 0. \end{cases}$$

Proof. See exercise sessions. □

We look for $(z_k)_{k=0}^{N-1}$ such that (6.15) holds. In order to do that, we mimic the procedure above, but in a discrete setting. By multiplying (6.15) with $\cos(\ell x_j)$, summing between $j = 0$ and $j = N-1$ and employing Lemma 6.6, we get

$$\sum_{j=0}^{N-1} y_j \cos(\ell x_j) = \frac{z_0}{2} \sum_{j=0}^{N-1} \cos(\ell x_j) + \sum_{k=1}^{N-1} z_k \sum_{j=0}^{N-1} \cos(k x_j)\cos(\ell x_j),$$

from which we deduce

$$z_k = \frac{2}{N} \sum_{j=0}^{N-1} y_j \cos(k x_j) \qquad \forall\ k = 0, \dots, N-1. \tag{6.16}$$

Note also in this case that since expression (6.16) can be evaluated for every $k \in \mathbb{Z}$, we may write $\left(\hat{f}_N(k)\right)_{k\in\mathbb{Z}}$ when needed.

Definition 6.5 (Discrete Cosine Transform)**.** $\left(\hat{f}_N(k)\right)_{k\in\mathbb{Z}}$ *is called the discrete cosine transform (DCT) of f with respect to the discretization $\left(x_j\right)_{j=0}^{N-1}$ defined in* (6.14).

Remark 6.9. *It is possible to interpret the DCT as the approximation of* (6.13)*, by using the composite midpoint quadrature rule. See exercise session.*

Definition 6.6 (Inverse Discrete Cosine Transform)**.** *The sequence $\left(y_j\right)_{j\in\mathbb{Z}}$ is called inverse discrete cosine transform (IDCT) of f with respect to the discretization $\left(x_j\right)_{j=0}^{N-1}$.*

6.4.1 The JPEG: an image compression standard

The light intensity measured by a camera is generally sampled over rectangular array of picture elements called pixels. Let us consider an image consisting of M^2 pixels, such that each couple (i, j), for $i, j = 0, \dots, M-1$, corresponds to a pixel. For the sake of simplicity of the discussion, let us focus on the case of black and white pictures. A BW picture can be thought as a function $Y : M \times M \to \{0, \dots, 255\}$, $(i, j) \mapsto Y(i, j)$, where $Y(i, j)$ represents the gray level at the pixel (i, j). Thus, $M^2 \cdot 8$ bits (the number 255 is 11111111 in base 2) per pixel are needed in order to store a picture. In principle M may be very large: a typical high resolution color picture for the web contains on the order of one millions pixels. However, state-of-the-art techniques can compress typical images from 1/10 to 1/50 without visibly affecting image quality. One of the most popular procedures is indeed the JPEG (N. Ahmed, T. Natarajan, K. R. Rao, 1974). Let us subdivide the image

into 8×8 blocks. For each block $\left(Y_{i_k,j_\ell}\right)_{k,\ell=0}^{7}$, where $(i_k)_{k=0}^{7}$, $\left(j_\ell\right)_{\ell=0}^{7} \subset \{0,\dots,M\}$ are subsequences of consecutive indices, we can apply the DCT, passing from the spatial domain to the frequency domain. In this way every 8×8 block of source image sample is effectively a discrete signal with 64 entries, which is a function of the two spatial dimensions, denoted for the sake of simplicity of the notation as $\left(Y_{i,j}\right)_{i,j=0}^{N}$, with $N = 7$. By analogy to (6.15), we want to find $\left(Z_{k,\ell}\right)_{k,\ell=0}^{N-1}$ such that

$$Y_{i,j} = \sum_{k=0}^{N-1} \sum_{\ell=0}^{N-1} \tilde{Z}_{k,\ell} \cos(kx_i) \cos(x_j), \qquad i,j = 0,\dots,N-1, \tag{6.17}$$

where, in order to compensate for the factor $1/2$, we employ the notation $\tilde{Z}_{0,0} = Z_{0,0}/4$, $\tilde{Z}_{k,0} = Z_{k,0}/2$, $\tilde{Z}_{0,\ell} = Z_{0,\ell}/2$, $\tilde{Z}_{k,\ell} = Z_{k,\ell}$, $k,\ell \geq 1$. In Figure 6.2 the reader can see the representation of the 64 basis functions $\left(\cos(kx_i)\cos(\ell x_j)\right)_{k,\ell=0}^{N-1}$ on a single 8×8 block. In particular, the columns correspond to the index k and the rows to the index ℓ, $k,\ell = 0,1,\dots,N = 7$. Increasing k, respectively ℓ, corresponds to higher oscillations in the x-direction, respectively y-direction.

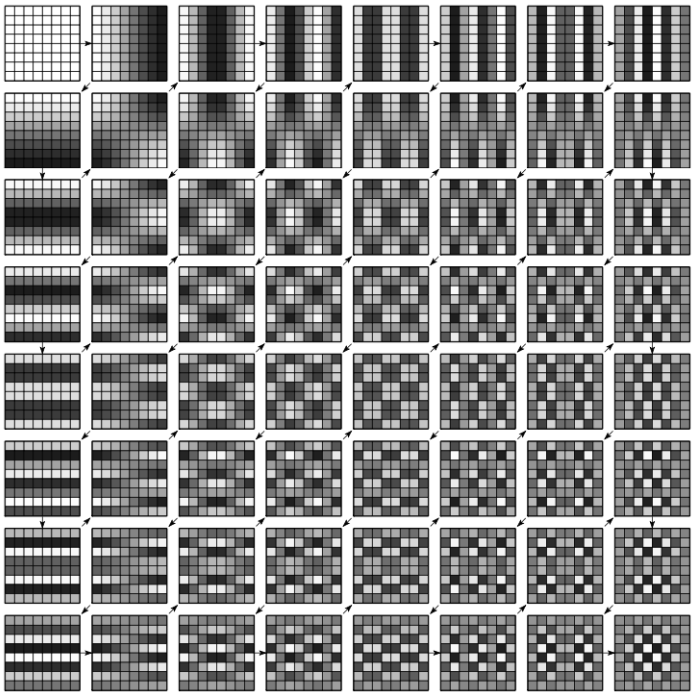


Figure 6.2: Representation of the basis functions $\left(\cos(kx_i)\cos(\ell x_j)\right)_{k,\ell=0}^{N-1}$.

The partition $\left(x_j\right)_{j=0}^{N-1}$ of $[0,\pi]$ is the same as in (6.14). We multiply (6.17) by $\cos(kx_i)\cos(\ell x_j)$, sum over $i,j = 0,\dots,N-1$ and use the discrete orthogonality relations (6.6). In this way we get the $2D$ counterpart of (6.16), that is

$$Z_{k,\ell} = \frac{4}{N^2} \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} Y_{i,j} \cos(kx_i) \cos(\ell x_j), \qquad k,\ell = 0,\dots,N-1. \tag{6.18}$$

The DCT takes the signal representing the block as an input and decomposes it into 64 orthogonal

basis signals, each one of them corresponding to a particular frequency. The value of a frequency reflects the size and speed of a change as you can see from Figure 6.2. The output is the collection of 64 DCT coefficients, representing the amplitudes of these signals. The first coefficient, corresponding to the zero frequency in both spatial dimensions, is often called DC (direct current). The remaining 63 entries are called AC (alternating currents). The high frequencies represent the high contrast areas in the image, i.e. rapid changes in pixel intensity. Note that in a classic image there is a high continuity between pixel values. Hence it turns out that the numerically important AC coefficients can be found in the square 4×4 around the DC coefficient.
Once the DCT coefficients are obtained, we would like to numerically represent them with no greater precision than is necessary to achieve the desired image quality. This step is called quantization of the signal. Each of the 64 DCT coefficients is quantized according to a 8×8 matrix T called, quantization matrix, with integer entries between 1 and 255, which is specified by the user and is conceived to provide greater resolution to more perceptible frequency components on less perceptible ones. In formulas, the quantization step reads as

$$Z_{k,\ell} \mapsto \lfloor \frac{Z_{k,\ell}}{T_{k,\ell}} \rfloor, \qquad k,\ell = 0,\dots,N-1. \tag{6.19}$$

Note that since the entries of T corresponding to the high frequencies are usually high and because we use the function "floor" in (6.19), the resulting high frequency coefficients will be zero. Let $\mathcal{Z}_N$ be the 8×8 matrix of coefficients after quantization. Let us go through its entries by following a zig-zag path (see Figure 6.3) in order to construct a vector of 64 coefficients. We just mention that this trick allows to further reduce the amount of information to be compressed of the image by placing low-frequency coefficients (more likely to be non-zero) before high frequency coefficients.

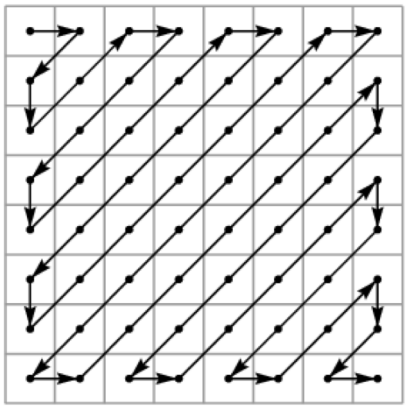
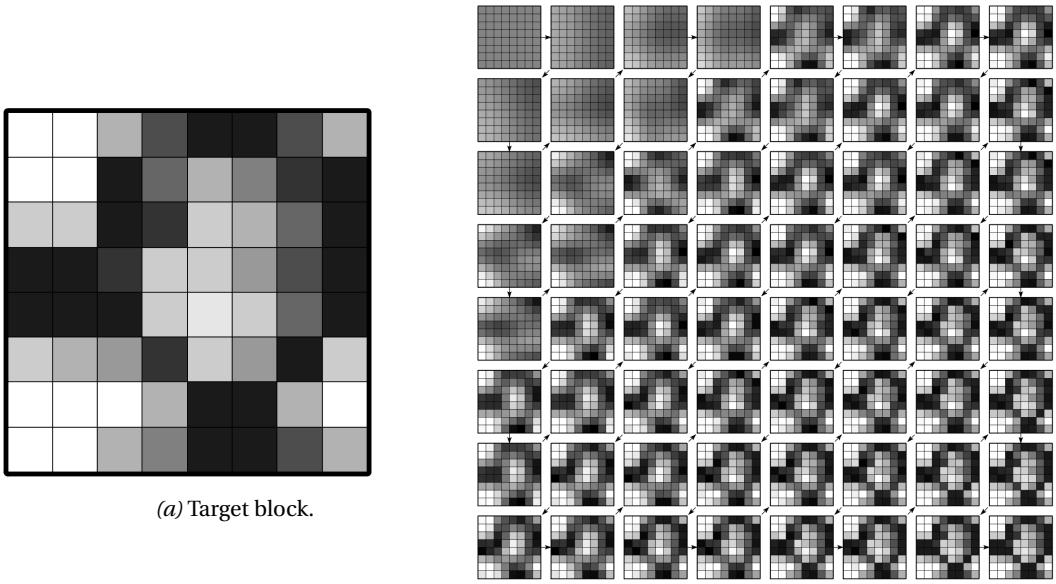


Figure 6.3: Zig-zag path used for the encoding of images in the JPEG.

The procedure described above can be reversed by applying the IDCT which takes the encoded coefficients and reconstructs the image signal by summing the basis signals. However, because of the quantization step, there is an inevitable loss of information. We have indeed introduced a numerical error and we made the whole procedure irreversible. That is why the JPEG is said to be a lossy compression technique. In Figure 6.4a we can see an 8×8 block from an image and in



(a) Target block.

(b) Decoding using the zig-zag procedure.

Figure 6.4: The reconstruction process of a sample 8×8 block from an image.

Figure 6.4b its decoding process that follows the zig-zag path.

6.5 Fast Fourier transform (FFT)

Let $f \in C^0([0,2\pi])$ be 2π-periodic. The starting point is the DFT of f with respect to the discretization (6.8), namely

$$\hat{f}_N(k) = \frac{1}{N} \sum_{j=0}^{N-1} y_j \omega_N^{-kj}, \qquad \forall\ k = 0, \dots, N-1, \tag{6.20}$$

with $\omega_N := e^{\frac{i2\pi}{N}}$. For the sake of simplicity of the notation, we are going to drop the multiplicative factor $\frac{1}{N}$ from (6.20), i.e.

$$\tilde{f}_N(k) = \sum_{j=0}^{N-1} y_j \omega_N^{-kj}, \qquad \forall\ k = 0, \dots, N-1. \tag{6.21}$$

By identifying

$$(\mathbf{z})_i := \tilde{f}_N(i), \quad (\mathbf{y})_i := y_i, \quad (\mathscr{F}_N)_{i,j} := \omega_N^{-ij}, \quad \forall\ i,j = 0, \dots, N-1,$$

so that

$$\mathbf{z}, \mathbf{y} \in \mathbb{R}^N, \qquad \mathscr{F}_N \in \mathbb{R}^{N \times N},$$

we can rewrite (6.21) as a linear system in matrix form, namely

$$\mathbf{z} = \mathscr{F}_N \mathbf{y}. \tag{6.22}$$

N	N^2	$N\log_2 N$	ratio
2^2	16	8	2
2^4	256	64	4
2^8	$\approx 6.5\cdot 10^4$	2048	32
2^{16}	$\approx 4\cdot 10^9$	$\approx 10^6$	4096
2^{32}	$\approx 2\cdot 10^{19}$	$\approx 10^{11}$	$\approx 1.3\cdot 10^8$

Table 6.1: Comparison of number of flops: $\mathcal{O}\left(N^2\right)$ vs $\mathcal{O}\left(N\log_2 N\right)$.

Remark 6.10. *From (6.22) we deduce that the DFT can be implemented on a computer and, since it is a matrix-vector multiplication (and $\mathscr{F}_N$ is full) it requires $\mathcal{O}(N^2)$ floating point operations ("flops"). Moreover, let us observe that $\mathscr{F}_N$ is an orthogonal matrix.*

The goal of this section is to present an algorithm which performs (6.22) in $\mathcal{O}(N\log_2 N)$ operations. Note that in practical applications N may be very large, hence being able to reduce the complexity of our algorithm makes a huge difference in terms of computational efforts, see Table 6.1.
The original idea of J.W. Cooley and J. Tukey (1965) was to recursively reduce the initial problem of size N into subproblems of size $\frac{N}{2}$ by dividing the addends of the matrix-vector multiplication (6.21) or (6.22) into two categories: the "even" and the "odd" terms. For the sake of simplicity of the exposition, let us start with $\mathscr{F}_{2N}\mathbf{y}$, the DFT of size $2N$ of a vector $\mathbf{y}\in\mathbb{R}^{2N}$. The goal is to reduce its computation to the ones of two DFTs of size N.
The subsequent result embodies the essence of the idea of Cooley and Tukey.

Lemma 6.7. *Let $\mathbf{u}=(u_0,u_1,\dots,u_{N-1})$, $\mathbf{v}=(v_0,v_1,\dots,v_{N-1})\in\mathbb{R}^N$ and $\mathbf{y}=(u_0,v_0,u_1,v_1,\dots,u_{N-1},v_{N-1})\in\mathbb{R}^{2N}$. Then, for all $k=0,\dots,N-1$, we have*

$$\left(\mathscr{F}_{2N}\mathbf{y}\right)_k=\left(\mathscr{F}_N\mathbf{u}\right)_k+\omega_{2N}^{-k}\left(\mathscr{F}_N\mathbf{v}\right)_k, \tag{6.23}$$

$$\left(\mathscr{F}_{2N}\mathbf{y}\right)_{k+N}=\left(\mathscr{F}_N\mathbf{u}\right)_k-\omega_{2N}^{-k}\left(\mathscr{F}_N\mathbf{v}\right)_k. \tag{6.24}$$

Proof. We focus on (6.23). First of all, notice that periodicity implies

$$\omega_{2N}^2=e^{\frac{i2\pi}{N}}=\omega_N. \tag{6.25}$$

We fix $k \in \{0, \dots, N-1\}$. Let us use (6.21) and separate the addends with even and odd indices

$$\begin{aligned}\left(\mathscr{F}_{2N}\mathbf{y}\right)_k &= \sum_{j=0}^{2N-1} y_j \omega_{2N}^{-kj} = \sum_{j=0}^{N-1} u_j \omega_{2N}^{-2jk} + \sum_{j=0}^{N-1} v_j \omega_{2N}^{-(2j+1)k} \\ &= \sum_{j=0}^{2N-1} y_j \omega_{2N}^{-kj} = \sum_{j=0}^{N-1} u_j \omega_{2N}^{-2jk} + \omega_{2N}^{-k} \sum_{j=0}^{N-1} v_j \omega_{2N}^{-2jk}.\end{aligned}$$

At this point we can exploit the key observation (6.25) so that

$$\begin{aligned}\left(\mathscr{F}_{2N}\mathbf{y}\right)_k &= \sum_{j=0}^{N-1} u_j \omega_N^{-jk} + \omega_{2N}^{-k} \sum_{j=0}^{N-1} v_j \omega_N^{-jk} \\ &= \left(\mathscr{F}_N \mathbf{u}\right)_k + \omega_{2N}^{-k} \left(\mathscr{F}\mathbf{v}\right)_k.\end{aligned}$$

The proof of (6.24) is very similar, hence it is left to the reader. Hint: use $\omega_N^N = 1$ and $\omega_{2N}^M = 1$. □

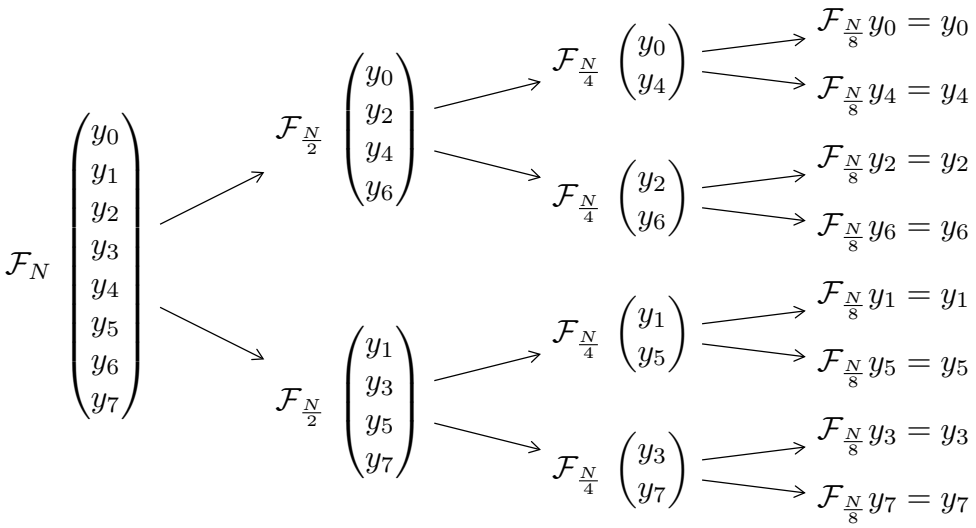


Figure 6.5: Visual illustration of Lemma 6.7 in the case $N = 2^3$.

Remark 6.11. *Lemma 6.7 tells us that we can compute $\mathscr{F}_{2N}\mathbf{y}$, the DFT of size $2N$, from $\mathscr{F}_N\mathbf{u}$ and $\mathscr{F}_N\mathbf{v}$, the DFT of the data with even and odd indices respectively. In particular, it allows to reduce a problem of size $2N$ into two subproblems of size N. Note that if N is even, then we can apply the same procedure to $\mathscr{F}_N\mathbf{u}$ and $\mathscr{F}_N\mathbf{v}$. In the special case $N = 2^m$, we can recursively reiterate the algorithm until we reduce to a problem size 1, i.e. until when the DFT is just the multiplication with a scalar.*

We observe that we will have to apply Lemma 6.7 $\mathcal{O}\left(\log_2 N\right)$ times. By keeping in mind Figure 6.5, we can think there are $\mathcal{O}\left(\log_2 N\right)$ "layers". For each layer there are $\mathcal{O}(N)$ operations to compute: $2N$ additions and N multiplications (see (6.23) and (6.24)).

We can summarize the previous observation in the following result.

Theorem 6.8. *Let $N = 2^m$, $m \in \mathbb{N}$. The computation of $\mathscr{F}_N \mathbf{y}$ can be performed with $\mathcal{O}\left(N \log_2 N\right)$ operations.*

6.6 Trigonometric interpolation

Let us move to the last section of this chapter. The fundamental task of the representation of a general 2π-periodic function using Fourier series is a problem of approximation theory.

Definition 6.7. *Let $n \in \mathbb{N}$, $(a_k)_{k=0}^n$, $(b_k)_{k=1}^n \subset \mathbb{R}$, the function*

$$f(x) = \frac{a_0}{2} + \sum_{k=1}^{n} a_k \cos(kx) + b_k \sin(kx) \tag{6.26}$$

is called trigonometric polynomial of degree n.

Definition 6.8.

$$\mathbb{T}_n := \operatorname{span}\left\{\frac{1}{2}, \cos(kx), \sin(kx) : \ k = 1, \dots, n\right\}$$

denotes the space of trigonometric polynomials up to degree n.

Remark 6.12. *Using Euler's formulas, it is possible to rewrite (6.26) in the complex framework as*

$$f(x) = \sum_{k=-n}^{n} c_k e^{ikx}, \tag{6.27}$$

where $(c_k)_{|k| \le n} \subset \mathbb{C}$ have been defined in (6.2). Let us remark that even though its coefficients are complex, the expression (6.27) still denotes a real valued function.

Proposition 6.9. *$\mathbb{T}_n$, the set of trigonometric polynomials of degree n, is a vector space of dimension $2n+1$ over $\mathbb{C}$.*

Proof. It is constructed via the Fourier basis e^{ikx}, $k = -n, \dots, n$. □

In this Section we analyse the approximation properties of the space $\mathbb{T}_N$, in a spirit that is similar to the one used in Chapter 3 for algebraic polynomials.
Let us consider a 2π-periodic function f, which is known at some discretization points of $[0, 2\pi]$,

$$x_j := \frac{2\pi j}{N}, \qquad y_j := f(x_j) \qquad \forall \ j = 0, \dots, N-1. \tag{6.28}$$

We look for $p_N \in \mathbb{T}_{\frac{N}{2}}$ interpolating f at the discretization points $\left(x_j\right)_{j=0}^{N-1}$. Let us firstly introduce some useful notation.

Notation 6.1. *We denote:*

$$\sum_{|k|\le n}^{\circ} x_k := \frac{1}{2}(x_{-n}+x_n) + \sum_{|k|<n} x_k,$$
$$\sum_{|k|\ge n}^{\circ} x_k := \frac{1}{2}(x_{-n}+x_n) + \sum_{|k|>n} x_k.$$

Theorem 6.10. *Let $\left(\hat{f}_N(k)\right)_{k=0}^{N-1}$ be the DFT of f with respect to the discretization (6.28). Then the trigonometric polynomial*

$$p_N(x) = \sum_{|k|\le N/2}^{\circ} \hat{f}_N(k) e^{ikx} \tag{6.29}$$

satisfies $p_N(x_j) = y_j$ for every $j = 0,\dots,N-1$.

Proof. Let us fix $j = 0,\dots,N-1$ and observe that the sequence $\left(\hat{f}_N(k)e^{ikx_j}\right)_{k\in\mathbb{Z}}$ is N-periodic. This follows from the fact that the DFT is N-periodic and we can easily check $e^{ikx_j} = e^{i(k+N)x_j}$ for every $k\in\mathbb{Z}$. This means that we can write

$$p_N(x_j) = \sum_{|k|\le N/2}^{\circ} \hat{f}_N(k)e^{ikx_j} = \sum_{k=0}^{N-1} \hat{f}_N(k)e^{ikx_j} = y_j,$$

where in the last equality we just used (6.9). □

6.6.1 Error analysis for trigonometric polynomial interpolation

Let us assume for the rest of the chapter that $f\in C^0([0,2\pi])$ 2π-periodic and

$$f(x) = \sum_{k=-\infty}^{+\infty} \hat{f}(k)e^{ikx} \qquad \forall\ k\in\mathbb{Z} \tag{6.30}$$

is absolutely convergent.

The goal is to study the error $f(x)-p_N(x)$, where p_N is the trigonometric polynomial interpolating f at the discretization points $(x_j)_{j=0}^{N-1}$, constructed in Theorem 6.10. By keeping in mind expressions (6.29) and (6.30), we observe that this is related to how the DFT $\left(\hat{f}_N(k)\right)_{|k|\le N}$ approximates the Fourier coefficients $\left(\hat{f}(k)\right)_{k\in\mathbb{Z}}$.

Since f is assumed to be at least continuous, the natural norm in order to study the behavior of the error $f(x)-p_N(x)$ is $\|\cdot\|_\infty$, where $\|u\|_\infty := \max_{x\in[0,2\pi]}|u(x)|$.

Let us recall, without proof, a simple version of the so-called Dirichlet theorem giving a sufficient condition for the pointwise convergence of p_N to f.

Theorem 6.11 (Dirichlet)**.** *Let $f\in C^0([0,2\pi])$ be 2π-periodic and such that for a $x\in\mathbb{R}$ there exist $f'_+(x)\in\mathbb{R}$ and $f'_-(x)\in\mathbb{R}$. Then*

$$\lim_{N\to+\infty} p_N(x) = f(x).$$

Note that from the numerical point of view the Dirichlet theorem is not satisfying. Indeed it does not provide any measure about the way $\|f-p_N\|_\infty$ decreases with respect to N. However, we

will see that if we require extra regularity on f, then it is possible to derive such estimates, see Theorem 6.14 below. We first analyze the relation between $\hat{f}_N(k)$ and $\hat{f}(k)$.

Lemma 6.12 (Aliasing formula)**.** *For every $k \in \mathbb{Z}$, it holds*

$$\hat{f}_N(k) - \hat{f}(k) = \sum_{\ell \neq 0} \hat{f}(k + \ell N). \tag{6.31}$$

Proof. Using (6.30), we have that $f(x_j) = \sum_{k=-\infty}^{+\infty} \hat{f}(k) e^{ikx_j} = \sum_{k=-\infty}^{+\infty} \hat{f}(k) \omega_N^{kj}$, hence we can rewrite (6.11), the DFT of f with respect to $(x_j)_{j=0}^{N-1}$, as follows

$$\hat{f}_N(k) = \frac{1}{N} \sum_{j=0}^{N-1} f(x_j) \omega_N^{-kj} = \frac{1}{N} \sum_{j=0}^{N-1} \sum_{n=-\infty}^{+\infty} \hat{f}(n) \omega_N^{nj} \omega_N^{-kj} \qquad k = 0, \dots, N-1.$$

From the discrete orthogonality relations stated in Lemma 6.3, it holds

$$\hat{f}_N(k) = \sum_{n=-\infty}^{+\infty} \hat{f}(n) \frac{1}{N} \sum_{j=0}^{N-1} \omega_N^{nj} \omega_N^{-kj} = \sum_{\substack{\ell=-\infty \\ \ell \neq 0}}^{+\infty} \hat{f}(k + \ell N) \qquad k = 0, \dots, N-1,$$

hence (6.31) follows. □

We have seen that for a sufficiently smooth function the Fourier coefficients rapidly converge to zero. The following result states that, for a smooth function, $\hat{f}_N(k)$ is a good approximation of $\hat{f}(k)$ if $|k| \leq N/2$, but it is bad if $|k| > N/2$.

Theorem 6.13 (Sampling theorem)**.** *For every $x \in \mathbb{R}$ it holds*

$$\left| f(x) - p_N(x) \right| \leq 2 \sum_{|k| \geq N/2}^{\circ} \left| \hat{f}(k) \right|.$$

Proof. Let us fix $x \in \mathbb{R}$. Let us subtract expression (6.29) from (6.30):

$$\begin{aligned} f(x) - p_N(x) = \sum_{k=-\infty}^{+\infty} \hat{f}(k) e^{ikx} - \sum_{|k| \leq N/2}^{\circ} z_k e^{ikx} = \left(\hat{f}(N/2) - \frac{1}{2} z_{N/2} \right) e^{i\frac{N}{2}x} \\ + \left(\hat{f}(-N/2) - \frac{1}{2} z_{-N/2} \right) e^{-i\frac{N}{2}x} + \sum_{|k| < N/2} \left(\hat{f}(k) - z_k \right) e^{ikx} + \sum_{|k| > N/2} \hat{f}(k) e^{ikx}. \end{aligned}$$

Now, we rearrange the terms as follows

$$f(x) - p_N(x) = \sum_{|k| \leq N/2}^{\circ} \left(\hat{f}(k) - z_k \right) e^{ikx} + \sum_{|k| \geq N/2}^{\circ} \hat{f}(k) e^{ikx}.$$

Lemma 6.12 comes to help, hence

$$f(x) - p_N(x) = \sum_{|k| \leq N/2}^{\circ} \sum_{\ell \neq 0} \hat{f}(k + \ell N) e^{ikx} + \sum_{|k| \geq N/2}^{\circ} \hat{f}(k) e^{ikx}.$$

The reader can check with a change of variables that

$$\mathring{\sum_{|k|\le N/2}} \sum_{\ell\neq 0} \hat{f}(k+\ell N) = \mathring{\sum_{|k|\ge N/2}} \hat{f}(k).$$

Finally we conclude using the triangle inequality and the identity $\left|e^{ikx}\right| = 1$ for every $k \in \mathbb{R}$. □

Remark 6.13. *Let $f : [0,2\pi] \to \mathbb{R}$ be 2π-periodic and with maximal frequency M, namely $\hat{f}(k) = 0$ for every $k \in \mathbb{Z}$ such that $|k| > M$. Then, the Sampling Theorem 6.13 tells us that the trigonometric polynomial $p_N \in \mathbb{T}_{N/2}$ interpolating f at the discretization points (6.28) is exact, i.e. it satisfies*

$$p_N(x) = f(x) \qquad \forall\, x \in \mathbb{R},$$

if we take $N > 2M$.

Theorem 6.14. *Let $f \in C^m([0,2\pi])$ be 2π-periodic and such that $f^{(m+1)}$ is integrable on $[0,2\pi]$. Then, it holds*

$$\left\|f - p_N f\right\|_\infty \le \tilde{C}_m (N/2)^{-m},$$

where $\tilde{C}_m = 2\left(2+\frac{1}{m}\right)\frac{1}{2\pi}\int_0^{2\pi}\left|f^{(m+1)}(x)\right|\,\mathrm{d}x$.

Proof. Using Proposition 6.2 and Theorem 6.13, we have

$$\begin{aligned}\left\|f - p_N\right\|_\infty \le& 2 \mathring{\sum_{|k|\ge N/2}} \left|\hat{f}(k)\right| \le \hat{f}(-N/2) + \hat{f}(N/2) + 2\sum_{|k|>N/2} \hat{f}(k) \\ \le& 2C_m (N/2)^{-m-1} + 2C_m \sum_{k>N/2} |k|^{-m-1}.\end{aligned}$$

The following estimate, whose proof is left as an exercise, comes to help.

$$\sum_{k>N/2} |k|^{-m-1} \le \int_{N/2}^{+\infty} x^{-m-1}\,\mathrm{d}x = /\frac{1}{m}(N/2)^{-m}. \tag{6.32}$$

Using (6.32), we get

$$\begin{aligned}\left\|f - p_N\right\|_\infty \le& 2C_m (N/2)^{-m-1} + 2C_m \frac{1}{m}(N/2)^{-m} \\ =& 2C_m (N/2)^{-m}\left(2/N + \frac{1}{m}\right) \le 2C_m\left(2+\frac{1}{m}\right)(N/2)^{-m}.\end{aligned}$$

□

6.7 Discrete Fourier transform and Chebyshev polynomials

The goal of this section is to give another interpretation of Chebyshev interpolation. The starting point is a function $f \in C^0([-1,1])$, not necessarily periodic, which we aim to interpolate. Let us pass to the variable θ such that $x = \cos\theta$ through the diffeomorphism $[-1,1] \to [0,\pi]$,

$x \mapsto \theta := \arccos x$. We define $g(\theta) := f(\cos\theta)$ for every $\theta \in [0,\pi]$ which can readily be extended to a 2π-periodic and even function as in Section 6.16. Hence we can consider the trigonometric polynomial $q \in \mathbb{T}_{\frac{N}{2}}$ interpolating g at

$$\theta_j := \frac{(2j+1)\pi}{2N}, \qquad j = 0,\dots,N-1, \tag{6.33}$$

which, by even symmetry, reads

$$q(\theta) = \frac{\hat{g}_N(0)}{2} + \sum_{k=1}^{N-1} \hat{g}_N(k)\cos(k\theta), \tag{6.34}$$

$\big(\hat{g}_N(k)\big)_{k=0}$ being the DCT of g with respect to the equispaced covering of $[0,\pi]$ in (6.33). The reader can easily check, using formula (6.15), that $q(\theta_j) = g(\theta_j)$ for $j = 0,\dots,N-1$. Performing a change of variable back to $[-1,1]$, we get the sample points

$$x_j = \cos\theta_j = \cos\left(\frac{(2j+1)\pi}{2j}\right), \qquad j = 0,\dots,N-1. \tag{6.35}$$

We have just recovered the Chebyshev points introduced in Chapter 3!

Remark 6.14. *Notice that the Chebyshev points are not equispaced anymore and that, in particular, they cluster near the edges of* $[-1,1]$. *They are indeed the projection on the x-axis of the equidistant subdivision* (6.33) *of the upper unit circle.*

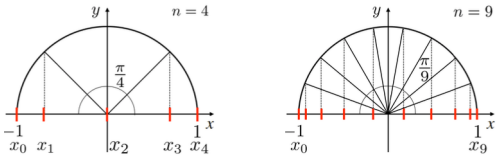


Figure 6.6: Chebyshev points as projection on the x-axis from equidistant subdivision on the circle.

Recalling the definition of Chebyshev polynomials T_n, Definition 3.4, the interpolant can be expressed, in terms of the variable x, as

$$p(x) = q(\arccos x) = \frac{\hat{g}_N(0)}{2} + \sum_{k=1}^{N-1} \hat{g}_N(k)\cos(k\arccos x) \tag{6.36}$$

$$= \frac{\hat{g}_N(0)}{2} + \sum_{k=1}^{N-1} \hat{g}_N(x)T_k(x) = \sum_{k=0}^{N-1} c_k T_k(x), \tag{6.37}$$

with

$$c_k = \begin{cases} \frac{\hat{g}_N(0)}{2} & \text{if } k = 0, \\ \hat{g}_N(k) & \text{if } k \neq 0. \end{cases}$$

This means in particular that the error analysis for the interpolant defined in terms of the Chebyshev polynomials is inherited from trigonometric interpolation.