



Technische
Universität
Braunschweig

Institute for Numerical Analysis



Ramp Up Mathematics — Numerical Analysis Ramp Up for Data Science

Matthias Bolhöfer, SS 2024

Contents I

■ Basics

- Finite Precision Representation
- Round off Errors
- Error Propagation

■ Error Analysis

- Introduction
- Conditioning
- Forward and Backward Analysis

■ Linear Systems

- Matrix Norms
- Condition when Solving Linear Systems
- The LU Decomposition
- The Cholesky Decomposition
- The CG Method

■ Nonlinear Systems

- Newton's Method



Contents II

- **Linear Least Squares**

- Linear Least Squares and Normal Equations
- The QR Decomposition
- Stable Computation of the QR Decomposition

- **The Singular Value Decomposition**

- Basic Properties
- The Key Property

- **Fourier Transformation**

- Fourier Series and Discrete Fourier Transformation
- Fast Fourier Transformation



Discrete Fourier Transformation

- signal processing, noise reduction often require approximation of functions which allow suppressing high frequency modes
- frequencies are modeled by periodic functions \sin , \cos
- this leads to the Fourier transformation where functions are approximated by Fourier series

Definition 6.1

$$P(x) = c_{-r} e^{-rx} + c_{-r+1} e^{(-r+1)x} + \cdots + c_{-1} e^{-x} + c_0 + c_1 e^{ix} + \cdots + c_s e^{sx},$$

where $j = \sqrt{-1}$, $r, s \in \mathbb{N}$, is called trigonometric polynomial

Example 6.1 (trigonometric polynomial)

$$P(x) = e^{3jx} + e^{-3jx} =$$



Fourier Series

Definition 6.2

Let $f : [0, 2\pi) \rightarrow \mathbb{C}$ such that $\int_0^{2\pi} |f(x)|^2 dx < \infty$. Then we call

$$\hat{f}(x) = \sum_{k=-\infty}^{\infty} c_k e^{ikx},$$

where

$$c_k = \langle f, e^{ikx} \rangle = \frac{1}{2\pi} \int_0^{2\pi} f(x) e^{-ikx} dx$$

the Fourier series associated with f . The coefficients c_k are called Fourier coefficients.

Theorem 6.1 (Representation of a function by its Fourier series)

Let $f : [0, 2\pi) \rightarrow \mathbb{C}$ such that $\int_0^{2\pi} |f(x)|^2 dx < \infty$. Then we have

$$\int_0^{2\pi} |f(x) - \hat{f}(x)|^2 dx = 0$$



Fourier Series

Example 6.2 (Signal)

$$f(x) = \begin{cases} -1 & x \in [0, \pi] \\ 1 & x \in (\pi, 2\pi] \end{cases}$$

$$k \neq 0 : 2\pi c_k = - \int_0^\pi e^{-jkx} dx + \int_\pi^{2\pi} e^{-jkx} dx =$$

=

$$\Rightarrow c_0 = , c_{2l} = , c_{2l+1} = \Rightarrow \hat{f}(x) =$$

real-valued representation: we have $c_{-k} = \overline{c_k}$:

$$\hat{f}(x) =$$

$$= \boxed{\quad}$$



Fourier Polynomials

In order to approximate the Fourier series itself we have to solve two tasks:

1. we have to truncate the series to a *finite Fourier polynomial* $\hat{f}_{-r,s}(x) = \sum_{k=-r}^s c_k e^{ikx}$
2. integrals c_k have to be computed numerically

Task 1.

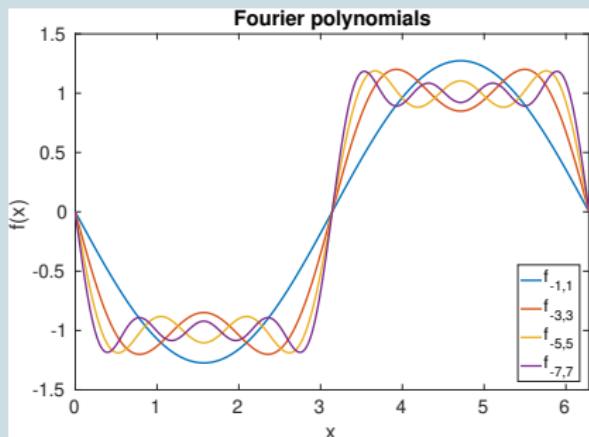
Example 6.3 (Signal approximated via Fourier transformation)

$$f(x) = \begin{cases} -1 & x \in [0, \pi] \\ 1 & x \in (\pi, 2\pi) \end{cases}$$

$$\hat{f}_{-2r-1,2r+1}(x) = -\frac{4}{\pi} \sum_{l=0}^r \frac{\sin((2l+1)x)}{2l+1}$$

$$\hat{f}_{-1,1}(x) =$$

$$\hat{f}_{-3,3}(x) =$$



Fourier Polynomials

Task 2. replace integral by quadrature formula, here: piecewise trapezoidal rule

- set $h = \frac{2\pi}{n}$ and use n grid points

$$x_l = lh, \quad l = 0, \dots, n$$

- replace integral

$$\int_0^{2\pi} f(x) \, dx$$

by piecewise trapezoidal rule for periodic functions

$$T(h) = h \sum_{l=0}^{n-1} f(x_l)$$

Note that for periodic functions we extend f to the real axis via $f(x + 2\pi) \equiv f(x)$, in particular this way we extend $f(2\pi)$ via $f(2\pi) := f(0)$



Fourier Polynomials

- Set $g(x) = e^{j kx}$. We thus replace the complex-valued integral $\langle f, g \rangle$:

$$\langle f, g \rangle = \frac{1}{2\pi} \int_0^{2\pi} f(x) \overline{g(x)} dx \approx \frac{1}{n} \sum_{l=0}^{n-1} f(x_l) \overline{g(x_l)} =: (f, g)$$

The expressions $\langle f, g \rangle$ and (f, g) are *scalar products* and induce norms

$$\|f\| := \sqrt{\frac{1}{2\pi} \int_0^{2\pi} |f(x)|^2 dx} \text{ and } |f| := \sqrt{\frac{1}{n} \sum_{l=0}^{n-1} |f(x_l)|^2}$$

Definition 6.3

Let $f : [0, 2\pi] \rightarrow \mathbb{C}$ piecewise continuous and let $n, r, s \in \mathbb{N}$. Then we call

$$\tilde{f}_{-r,s}(x) = \sum_{k=-r}^s \tilde{c}_k e^{jkx}, \text{ where } \tilde{c}_k = (f, e^{jkx}) = \frac{1}{n} \sum_{l=0}^{n-1} f(x_l) e^{-jkx_l}$$

discrete Fourier transform of f .



Discrete Fourier Transformation

Example 6.4 (Signal approximated by discrete Fourier transform)

$$f(x) = \begin{cases} -1 & x \in (0, \pi) \\ 0 & x = 0, \pi, 2\pi \\ 1 & x \in (\pi, 2\pi) \end{cases}$$

Let n be even (for simplicity) $\Rightarrow \tilde{c}_k =$

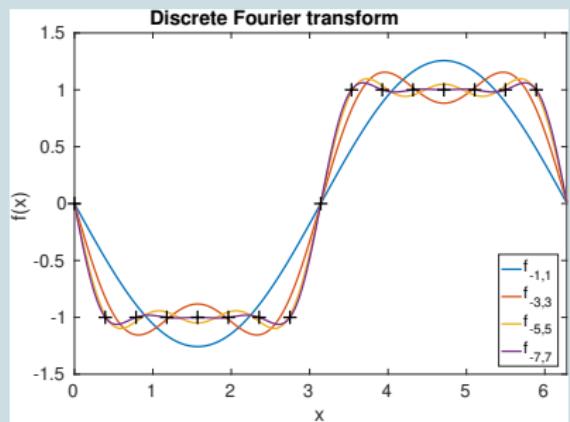
For $n = 16$ we obtain $\tilde{c}_{2l} = 0$,

$$\begin{aligned}\tilde{c}_1 &\approx -1.26, & \tilde{c}_3 &\approx -0.374, \\ \tilde{c}_5 &\approx -0.167, & \tilde{c}_7 &\approx -0.0497.\end{aligned}$$

$$\tilde{f}_{-1,1}(x) =$$

$$\tilde{f}_{-3,3}(x) =$$

$$\tilde{f}_{-5,5}(x) =$$



Discrete Fourier Transformation

- regarding the norms $\|\bullet\|$, $|\bullet|$, Fourier transforms are the optimal approximations to f w.r.t. the space spanned by $\{e^{ikx}\}_{k=-r,\dots,s}$
- $\tilde{f}_{-r,s}$, respectively $\tilde{f}_{r,s}$ are the associated orthogonal projections of f
- the functions $\{e^{ikx}\}_{k=-r,\dots,s}$ form an orthonormal system

Let us summarize the principle computation of a discrete Fourier transform

- Choose $n \in \mathbb{N}$ and the grid points $(x_0, f_0), \dots, (x_{n-1}, f_{n-1})$, where $x_i = i \frac{2\pi}{n}$.
- Select r, s , such that $r + s \leq n - 1$ and compute the coefficients of the discrete Fourier transform

$$\tilde{c}_k = \frac{1}{n} \sum_{l=0}^{n-1} f_l e^{-ikx_l}, \quad k = -r, \dots, s.$$

- Then our discrete Fourier transform is given by

$$\tilde{f}_{-r,s}(x) = \sum_{k=-r}^s \tilde{c}_k e^{ikx}.$$



Fast Fourier Transformation (FFT)

- efficient computation of the coefficients $\tilde{c}_{-r}, \dots, \tilde{c}_s$. is performed via the *fast Fourier transformation (FFT)*
- note that for $r + s = n - 1$ we have $\tilde{c}_{-k} = \tilde{c}_{n-k}$, i.e., the coefficients are cyclic
- For simplicity we set $r = 0$ and $s = n - 1 \rightarrow \tilde{c}_0, \dots, \tilde{c}_{n-1}$.
- important for the efficient computation: for n only use powers of two

$$n = 2^p,$$

in this case there is a simple recursion formula



Fast Fourier Transformation (FFT)

- our goal is to compute the coefficients

$$\tilde{c}_k = \frac{1}{n} \sum_{l=0}^{n-1} f_l e^{-j \frac{2\pi k l}{n}} = \frac{1}{n} \sum_{l=0}^{n-1} f_l \omega_n^{kl}, \text{ where } \omega_n = e^{-j \frac{2\pi}{n}}$$

efficiently.

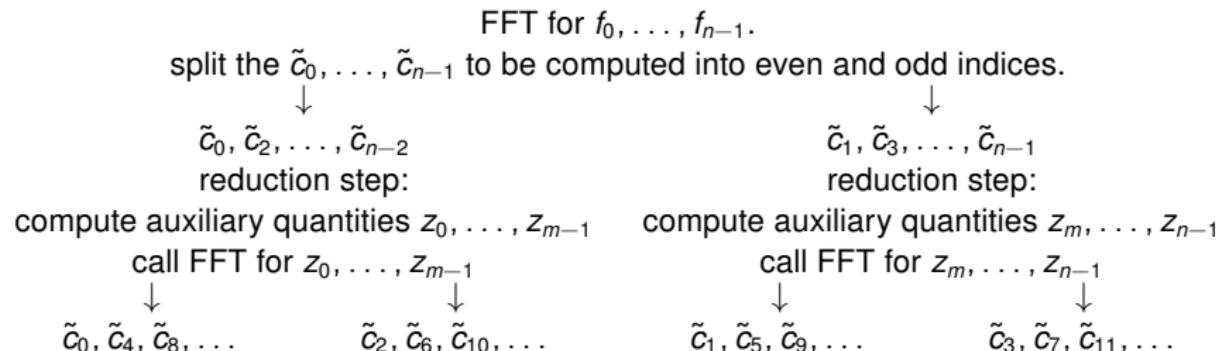
- formally this can be read as a matrix-vector product of type

$$\begin{bmatrix} \tilde{c}_0 \\ \vdots \\ \tilde{c}_{n-1} \end{bmatrix} = \frac{1}{n} W \begin{bmatrix} f_0 \\ \vdots \\ f_{n-1} \end{bmatrix}, \text{ where } W = [\omega_n^{kl}]_{k,l=0,\dots,n-1}.$$

- clue: use the fact that ω_n is n -th unit root, i.e., all powers of ω_n are also located on the unit circle, exponents can be computed easily

Fast Fourier Transformation (FFT)

the basic sketch of the FFT works as follows. Set $m := \frac{n}{2}$.



Fast Fourier Transformation (FFT)

- FFT of size $n = 2^p$ can be traced back to two FFTs of size $m = \frac{n}{2} = 2^{p-1}$
- additional computational effort of complexity $\mathcal{O}(n)$ is required for the reduction step
- in total, the same principle is recursively applied p times, where $p = \log_2(n)$
- instead of a matrix vector multiplication that costs $\mathcal{O}(n^2)$ we obtain a complexity of $\mathcal{O}(n \log_2 n)$
- two FFTs of half size can be computed independently, implementation on a parallel computer is feasible



Fast Fourier Transformation (FFT)

1. discuss even indices $k = 2q$:

- for the reduction step compute the quantities $z_l := f_l + f_{m+l}$ for $l = 0, \dots, m-1$.
- Note that $\omega_n^{2q(m+l)} = \dots = \dots = \dots$
- for the reduction step we sum up the terms with indices l and $m+l$:

$$\Rightarrow n \cdot \tilde{c}_{2q} = \underbrace{\sum_{l=0}^{n-1} f_l \omega_n^{2ql}}_{FFT_n(f_0, \dots, f_{n-1})} =$$

2. analogously proceed for the odd indices $k = 2q+1$:

- compute quantities $z_{m+l} = (f_l - f_{m+l}) \omega_n^l$ for $l = 0, \dots, m-1$
- We have $\omega_n^{(2q+1)(m+l)} = \dots = \dots = \dots = \dots$
- Again, sum up the terms with indices l and $m+l$:

$$\Rightarrow n \cdot \tilde{c}_{2q+1} = \underbrace{\sum_{l=0}^{n-1} f_l \omega_n^{(2q+1)l}}_{FFT_n(f_0, \dots, f_{n-1})} =$$



Fast Fourier Transformation (FFT)

Algorithm 6.1 (FFT Blue Print)

Input: function values $f_0, \dots, f_{n-1} \in \mathbb{C}$

Output: coefficients $n \cdot \tilde{c}_0, \dots, n \cdot \tilde{c}_{n-1}$ of the discrete Fourier transform

- 1: If $n = 1$
- 2: $n\tilde{c}_0 \leftarrow f_0$
- 3: Else
- 4: {reduction step for even indices}
- 5: $z_l \leftarrow f_l + f_{m+l}, l = 0, \dots, m - 1$
- 6: $[n\tilde{c}_0, n\tilde{c}_2, \dots, n\tilde{c}_{n-2}] = FFT(z_0, \dots, z_{m-1})$
- 7: {reduction step for odd indices}
- 8: $z_{m+l} \leftarrow (f_l - f_{m+l})\omega_n^l, l = 0, \dots, m - 1$
- 9: $[n\tilde{c}_1, n\tilde{c}_3, \dots, n\tilde{c}_{n-1}] = FFT(z_m, \dots, z_{n-1})$
- 10: [End] If

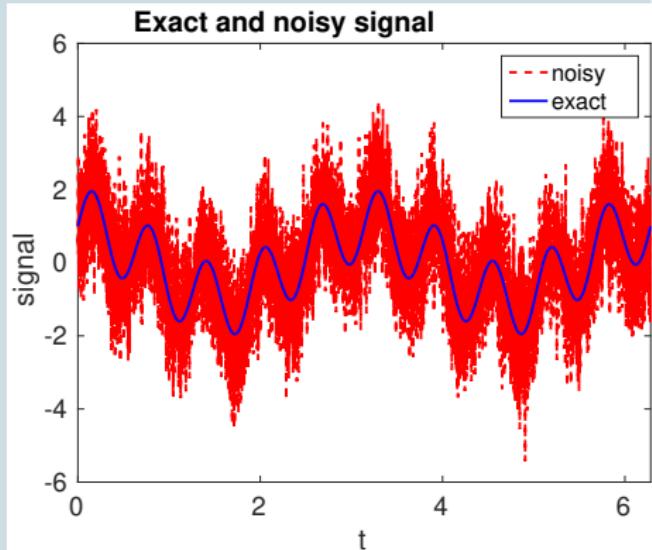
Remark: For the proper Fourier coefficients \tilde{c}_k we have to eventually divide by n .



Fast Fourier Transformation (FFT)

Example 6.5 (Signal Denoising)

- We are seeking to reconstruct the exact signal $g(t) = \sin(10t) + \cos(2t)$ in $[0, 2\pi]$
- we let the signal be superposed by a normally distributed noise $n(t)$,
 $\tilde{g}(t) = g(t) + n(t)$
- we sample the signal with a step size of $\Delta t = 0.001$.
- we compute the Fourier coefficients c_0, \dots, c_{n-1} via FFT
- we only consider those coefficients c_0, \dots, c_{n-1} which are greatest in magnitude and skip the others
- watch the associated Fourier transform $\hat{g}(t)$



Fast Fourier Transformation (FFT)

Example 6.6 (Signal Denoising (ctd.))

