An introduction to fractional calculus

Fundamental ideas and numerics

Fabio Durastante
Università di Pisa
fabio.durastante@unipi.it
fdurastante.github.io

September, 2022

Solving linear system with Toeplitz-like matrices

In the last lecture we discretized

$$\begin{cases} \frac{\partial W}{\partial t} = \theta \, {}^{RL}D^{\alpha}_{[0,x]} W(x,t) + (1-\theta) {}^{RL}D^{\alpha}_{[x,1]} W(x,t), & \theta \in [0,1], \\ W(0,t) = W(1,t) = 0, & W(x,t) = W_0(x). \end{cases}$$

Solving linear system with Toeplitz-like matrices

In the last lecture we discretized

$$\begin{cases} \frac{\partial W}{\partial t} = \theta \, {}^{GL}D^{\alpha}_{[0,x]} W(x,t) + (1-\theta)\, {}^{GL}D^{\alpha}_{[x,1]} W(x,t), & \theta \in [0,1], \\ W(0,t) = W(1,t) = 0, \qquad W(x,t) = W_0(x). \end{cases}$$

Obtaining

$$\left(I_N - \frac{\Delta t}{h_N^{\alpha}} \left[\theta G_N + (1-\theta) G_N^T \right] \right) \mathbf{w}^{n+1} = \mathbf{w}^n$$

with

$$G_N = \begin{bmatrix} 1 & 0 & \cdots & \cdots & 0 \\ g_2 & g_1 & g_0 & & \\ \vdots & \ddots & \ddots & \ddots & \\ \vdots & & \ddots & \ddots & g_0 \\ g_{N-1} & \cdots & g_3 & g_2 & g_1 \\ 0 & \cdots & \cdots & 0 & 1 \end{bmatrix}.$$

Solving linear system with Toeplitz-like matrices

The matrix

$$A_N = I_N - \frac{\Delta t}{h_N^{\alpha}} \left[\theta G_N + (1 - \theta) G_N^T \right],$$

is a **Toepltiz** matrix plus some rank corrections.

- By rearranging the right-hand side or restricting to solve only for the internal nodes we can avoid the rank corrections.

Solving linear system with Toeplitz-like matrices

The matrix

$$A_N = I_N - \frac{\Delta t}{h_N^{\alpha}} \left[\theta G_N + (1-\theta) G_N^T \right],$$

is a **Toepltiz** matrix plus some rank corrections.

- By rearranging the right-hand side or restricting to solve only for the internal nodes we can avoid the rank corrections.
- How do we solve such systems?

Solving linear system with Toeplitz-like matrices

The matrix

$$A_N = I_N - \frac{\Delta t}{h_N^{\alpha}} \left[\theta G_N + (1-\theta) G_N^T \right],$$

is a **Toepltiz** matrix plus some rank corrections.

- By rearranging the right-hand side or restricting to solve only for the internal nodes we can avoid the rank corrections.
- How do we solve such systems?
  - Direct methods

Solving linear system with Toeplitz-like matrices

The matrix

$$A_N = I_N - \frac{\Delta t}{h_N^{\alpha}} \left[\theta G_N + (1-\theta) G_N^T \right],$$

is a **Toepltiz** matrix plus some rank corrections.

- By rearranging the right-hand side or restricting to solve only for the internal nodes we can avoid the rank corrections.
- How do we solve such systems?
  - Direct methods
  - Iterative methods

Solving linear system with Toeplitz-like matrices

The matrix

$$A_N = I_N - \frac{\Delta t}{h_N^{\alpha}} \left[\theta G_N + (1-\theta) G_N^T \right],$$

is a **Toepltiz** matrix plus some rank corrections.

- By rearranging the right-hand side or restricting to solve only for the internal nodes we can avoid the rank corrections.
- How do we solve such systems?
  - Direct methods $\Rightarrow$ fast and superfast Toeplitz solvers
  - Iterative methods

Solving linear system with Toeplitz-like matrices

The matrix

$$A_N = I_N - \frac{\Delta t}{h_N^{\alpha}} \left[\theta G_N + (1-\theta) G_N^T \right],$$

is a **Toepltiz** matrix plus some rank corrections.

- By rearranging the right-hand side or restricting to solve only for the internal nodes we can avoid the rank corrections.
- How do we solve such systems?
  - Direct methods ⇒ fast and superfast Toeplitz solvers
  - Iterative methods ⇒ preconditioned Krylov methods, multigrid solvers/preconditioners

Direct Toeplitz solvers

Direct Toeplitz solver are *mostly* based on the answer to the following question:

- ❓ is the inverse of a Toeplitz matrix still a Toeplitz matrix?

Direct Toeplitz solvers

Direct Toeplitz solver are *mostly* based on the answer to the following question:

- ❓ is the inverse of a Toeplitz matrix still a Toeplitz matrix?

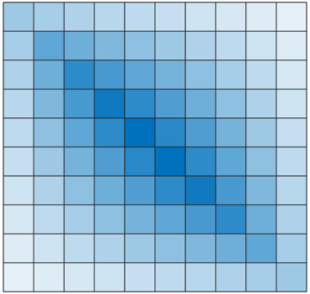
$$T_n = \begin{bmatrix} 2 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 2 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 2 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 2 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 2 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 2 \end{bmatrix}$$

Direct Toeplitz solvers

Direct Toeplitz solver are *mostly* based on the answer to the following question:

- ❓ is the inverse of a Toeplitz matrix still a Toeplitz matrix?

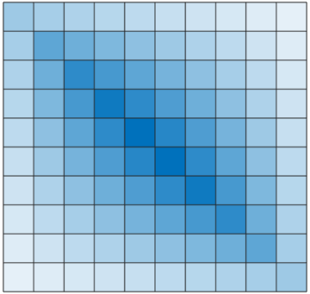
$T_n^{-1} =$



Direct Toeplitz solvers

Direct Toeplitz solver are *mostly* based on the answer to the following question:

- is the inverse of a Toeplitz matrix still a Toeplitz matrix?

$T_n^{-1} =$ 

So the answer is **no**, but... it seems that there is still some structure there, doesn't it?

The Gohberg–Semencul formula

... starting from a **displacement representation** of T_n, i.e.,

$$t_0 T_n = \begin{bmatrix} t_0 & 0 & \cdots & 0 \\ t_1 & t_0 & \ddots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ t_{n-1} & t_{n-2} & \cdots & t_0 \end{bmatrix} \begin{bmatrix} t_0 & t_{-1} & \cdots & t_{1-n} \\ 0 & t_0 & \cdots & t_{2-n} \\ 0 & 0 & & \vdots \\ \vdots & \vdots & & t_{-1} \\ 0 & 0 & \cdots & t_0 \end{bmatrix} - \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ t_1 & 0 & \cdots & 0 & 0 \\ t_2 & t_1 & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & 0 & 0 \\ t_{n-1} & t_{n-2} & \cdots & t_1 & 0 \end{bmatrix} \begin{bmatrix} 0 & t_{-1} & t_{-2} & \cdots & t_{1-n} \\ 0 & 0 & t_{-1} & \cdots & t_{2-n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & & t_{-1} \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}$$

Gohberg and Semencul 1972 obtained a **displacement representation** of the **inverse**

$$z_1 T_n^{-1} = \begin{bmatrix} z_1 & 0 & \cdots & 0 \\ z_2 & z_1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ z_{n-1} & z_{n-2} & \cdots & 0 \\ z_n & z_{n-1} & \cdots & z_1 \end{bmatrix} \begin{bmatrix} v_n & v_{n-1} & \cdots & v_1 \\ 0 & v_n & \cdots & v_2 \\ 0 & 0 & & \vdots \\ \vdots & \vdots & & v_{n-1} \\ 0 & 0 & \cdots & v_n \end{bmatrix} - \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ v_1 & 0 & \cdots & 0 & 0 \\ v_2 & v_1 & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & 0 & 0 \\ v_{n-1} & v_{n-2} & \cdots & v_1 & 0 \end{bmatrix} \begin{bmatrix} 0 & z_n & z_{n-1} & \cdots & z_1 \\ 0 & 0 & z_n & \cdots & z_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & & z_n \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}$$

with $z_1 = v_n$.

Direct Toeplitz solvers

By cleverly computing the vectors $\mathbf{z}$ and $\mathbf{v}$ from the $\{t_n\}_n$ coefficients, one obtains several "fast" and "superfast" algorithms:

Algorithm	Complexity
Levinson 1946	$O(n^2)$
Trench 1964	$O(n^2)$
Zohar 1974	$O(n^2)$
Bitmead and Anderson 1980	$O(n \log^2(n))$
Brent, Gustavson, and Yun 1980	$O(n \log^2(n))$
Hoog 1987	$O(n \log^2(n))$
Ammar and Gragg 1988	$O(n \log^2(n))$
T. F. Chan and Hansen 1992	$O(n^2)$
Bini and Meini 1999	$O(n \log m + m \log^2 m \log {}^{n}/_{m})$

n size of the matrix, m size of the bandwidth.

In our case

To treat our case

$$\left(I_N - \frac{\Delta t}{h_N^{\alpha}} \left[\theta G_N + (1-\theta) G_N^T \right] \right) \mathbf{w}^{n+1} = \mathbf{w}^n$$

we can then apply one of those algorithms (some of them use *symmetry*).

In our case

To treat our case

$$\left(I_N - \frac{\Delta t}{h_N^{\alpha}}\left[\theta G_N + (1-\theta) G_N^T\right]\right)\mathbf{w}^{n+1} = \mathbf{w}^n$$

we can then apply one of those algorithms (some of them use *symmetry*).

❓ What happens if we need to treat the case

$$\left(I_N - \frac{\Delta t}{h_N^{\alpha}}\left[D_n^{(1)} G_N + D_n^{(2)} G_N^T\right]\right)\mathbf{w}^{n+1} = \mathbf{w}^n$$

with $D_n^{(\cdot)}$ diagonal matrices coming from the discretization of **anisotropic space-variant diffusion coefficients**?

In our case

To treat our case

$$\left(I_N - \frac{\Delta t}{h_N^{\alpha}}\left[\theta G_N + (1-\theta) G_N^T\right]\right)\mathbf{w}^{n+1} = \mathbf{w}^n$$

we can then apply one of those algorithms (some of them use *symmetry*).

- What happens if we need to treat the case

$$\left(I_N - \frac{\Delta t}{h_N^{\alpha}}\left[D_n^{(1)} G_N + D_n^{(2)} G_N^T\right]\right)\mathbf{w}^{n+1} = \mathbf{w}^n$$

  with $D_n^{(\cdot)}$ diagonal matrices coming from the discretization of **anisotropic space-variant diffusion coefficients**?
- What happens if we need to treat **multi-dimensional cases**?

Krylov subspace methods

To overcome these challenges, we use an iterative approach based on **Krylov subspaces**.

Krylov subspace

A *Krylov subspace* $\mathcal{K}$ for the matrix A related to a non null vector $\mathbf{v}$ is defined as

$$\mathcal{K}_m(A, \mathbf{v}) = \mathrm{Span}\{\mathbf{v}, A\mathbf{v}, A^2\mathbf{v}, \dots, A^{m-1}\mathbf{v}\}.$$

Krylov subspace methods

To overcome these challenges, we use an iterative approach based on **Krylov subspaces**.

Krylov subspace

A *Krylov subspace* $\mathcal{K}$ for the matrix A related to a non null vector $\mathbf{v}$ is defined as

$$\mathcal{K}_m(A, \mathbf{v}) = \text{Span}\{\mathbf{v}, A\mathbf{v}, A^2\mathbf{v}, \ldots, A^{m-1}\mathbf{v}\}.$$

❗ The fundamental operation is the **matrix-vector** product.

Krylov subspace methods

To overcome these challenges, we use an iterative approach based on **Krylov subspaces**.

Krylov subspace

A *Krylov subspace* $\mathcal{K}$ for the matrix A related to a non null vector $\mathbf{v}$ is defined as

$$\mathcal{K}_m(A, \mathbf{v}) = \mathrm{Span}\{\mathbf{v}, A\mathbf{v}, A^2\mathbf{v}, \ldots, A^{m-1}\mathbf{v}\}.$$

- The fundamental operation is the **matrix-vector** product.
- Their use is *effective* when these **products are cheap**.

Krylov subspace methods

To overcome these challenges, we use an iterative approach based on **Krylov subspaces**.

Krylov subspace

A *Krylov subspace* $\mathcal{K}$ for the matrix A related to a non null vector $\mathbf{v}$ is defined as

$$\mathcal{K}_m(A, \mathbf{v}) = \mathrm{Span}\{\mathbf{v}, A\mathbf{v}, A^2\mathbf{v}, \ldots, A^{m-1}\mathbf{v}\}.$$

- The fundamental operation is the **matrix-vector** product.
- Their use is *effective* when these **products are cheap**.
- We can compute $T_n(f)\mathbf{v}$ in $O(n\log(n))$ operations!

$$C_{2n}\begin{bmatrix}\mathbf{v}\\ \mathbf{0}_n\end{bmatrix} = \underbrace{\begin{bmatrix}T_n(f) & E_n\\ E_n & T_n(f)\end{bmatrix}}_{\text{Circulant}}\begin{bmatrix}\mathbf{v}\\ \mathbf{0}_n\end{bmatrix} = \begin{bmatrix}T_n(f)\mathbf{v}\\ E_n\mathbf{v}\end{bmatrix}, \quad E_n = \begin{bmatrix} 0 & t_{n-1} & \ldots & t_2 & t_1\\ t_{1-n} & 0 & t_{n-1} & \ldots & t_2\\ \vdots & t_{1-n} & 0 & \ddots & \vdots\\ t_{-2} & \ldots & \ddots & \ddots & t_{n-1}\\ t_{-1} & t_{-2} & \ldots & t_{1-n} & 0\end{bmatrix}.$$

The Conjugate Gradient Method

When A is **symmetric positive definite** the method of choice is the **C**onjugate **G**radient.

Theorem.

Let A be SPD and $k_2(A) = \lambda_n/\lambda_1$ be the 2–norm condition number of A. We have:

$$\frac{\|\mathbf{r}^{(m)}\|_2}{\|\mathbf{r}^{(0)}\|_2} \leq \sqrt{k_2(A)} \frac{\|\mathbf{x}^* - \mathbf{x}^{(m)}\|_A}{\|\mathbf{x}^* - \mathbf{x}^{(0)}\|_A}.$$

Corollary.

If A is SPD with eigenvalues $0 < \lambda_1 \leq \ldots \leq \lambda_n$, we have

$$\frac{\|\mathbf{x}^* - \mathbf{x}^{(m)}\|_A}{\|\mathbf{x}^* - \mathbf{x}^{(0)}\|_A} \leq 2 \left(\frac{\sqrt{k_2(A)} - 1}{\sqrt{k_2(A)} + 1} \right)^m .$$

Input: $A \in \mathbb{R}^{n\times n}$ SPD, N_{max}, $\mathbf{x}^{(0)}$
Output: $\tilde{\mathbf{x}}$, candidate approximation.

```
r^(0) ← ||b − A x^(0)||_2, r = r^(0), p ← r;
ρ0 ← ||r^(0)||^2;
for k = 1, ..., N_max do
    if k = 1 then
        p ← r;
    end
    else
        β ← ρ1/ρ0;
        p ← r + β p;
    end
    w ← A p;
    α ← ρ1/p^T w;
    x ← x + α p;
    r ← r − α w;
    ρ1 ← ||r||_2^2;
    if then
        Return: x̃ = x;
    end
end
```

The Conjugate Gradient Method

⚠ The bound in the corollary is **descriptive** of the convergence behavior.

The Conjugate Gradient Method

⚠ The bound in the corollary is **descriptive** of the convergence behavior.

Theorem.

Let $A \in \mathbb{R}^{n\times n}$ be SPD. Let m an integer, $1 < m < n$ and $c > 0$ a constant such that for the eigenvalues of A we have

$$0 < \lambda_1 \leq \lambda_2 \leq \lambda_3 \leq \ldots \leq \lambda_{n-m+1} \leq c < \ldots \leq \lambda_n.$$

Fixed $\varepsilon > 0$ an upper bound in exact arithmetic for the minimum number of iterations k reducing the relative error in energy norm form the approximation $\mathbf{x}^{(k)}$ generated by CG by ε is given by

$$\min\left\{ \left\lceil \frac{1}{2}\sqrt{c/\lambda_1}\log\left(\frac{2}{\varepsilon}\right) + m + 1 \right\rceil, n \right\}$$

The Conjugate Gradient Method

⚠ The bound in the corollary is **descriptive** of the convergence behavior.

Theorem.

Let $A \in \mathbb{R}^{n \times n}$ be SPD. Let m an integer, $1 < m < n$ and $c > 0$ a constant such that for the eigenvalues of A we have

$$0 < \lambda_1 \leq \lambda_2 \leq \lambda_3 \leq \ldots \leq \lambda_{n-m+1} \leq c < \ldots \leq \lambda_n.$$

Fixed $\varepsilon > 0$ an upper bound in exact arithmetic for the minimum number of iterations k reducing the relative error in energy norm form the approximation $\mathbf{x}^{(k)}$ generated by CG by ε is given by

$$\min \left\{ \left\lceil \frac{1}{2} \sqrt{c/\lambda_1} \log\left(\frac{2}{\varepsilon}\right) + m + 1 \right\rceil, n \right\}$$

❓ How can we put ourselves in the hypotheses of the Theorem?

Clustered spectra

A proper cluster

A sequence of matrices $\{A_n\}_{n\geq 0}$, $A_n \in \mathbb{C}^{n\times n}$, has a **proper cluster** of eigenvalues in $p \in \mathbb{C}$ if, $\forall \varepsilon > 0$, if the number of eigenvalues of A_n **not in** $D(p, \varepsilon) = \{z \in \mathbb{C} \mid |z - p| < \varepsilon\}$ is bounded by a constant r that does not depend on n. Eigenvalues not in the *proper cluster* are called **outlier** eigenvalues.

Clustered spectra

A proper cluster

A sequence of matrices $\{A_n\}_{n\geq 0}$, $A_n \in \mathbb{C}^{n\times n}$, has a **proper cluster** of eigenvalues in $p \in \mathbb{C}$ if, $\forall \varepsilon > 0$, if the number of eigenvalues of A_n **not in** $D(p, \varepsilon) = \{z \in \mathbb{C} \mid |z - p| < \varepsilon\}$ is bounded by a constant r that does not depend on n. Eigenvalues not in the *proper cluster* are called **outlier** eigenvalues.

❓ Do the matrices

$$A_N = I_N - \frac{\Delta t}{2h_N^{\alpha}}\left[G_N + G_N^T\right]$$

have a **clustered spectra**?

Clustered spectra

A proper cluster

A sequence of matrices $\{A_n\}_{n\geq 0}$, $A_n \in \mathbb{C}^{n\times n}$, has a **proper cluster** of eigenvalues in $p \in \mathbb{C}$ if, $\forall \varepsilon > 0$, if the number of eigenvalues of A_n **not in** $D(p,\varepsilon) = \{z \in \mathbb{C} \,|\, |z-p| < \varepsilon\}$ is bounded by a constant r that does not depend on n. Eigenvalues not in the *proper cluster* are called **outlier** eigenvalues.

- Do the matrices

$$A_N = I_N - \frac{\Delta t}{2h_N^{\alpha}}\left[G_N + G_N^T\right]$$

have a **clustered spectra**?

- We can investigate this question by looking again at the **spectral distribution** of the sequence $\{A_N\}_N$.

Asymptotic distribution: the symmetric case

$$A_N = I_N - \frac{\Delta t}{2h_N^{\alpha}}\left[G_N + G_N^T\right],$$

the sequence $\{A_N\}_N$ is **not** yet **ready** for the **analysis**, we have the coefficient $\Delta t/2h_N^{\alpha}$ that varies with N.

- For *consistency* reason it makes sense to select $\Delta t \equiv h_N \equiv \nu_N$, then, since $\alpha \in (1,2]$ we have that $\nu^{1-\alpha}$ for $\nu \to 0^+$ goes to $+\infty$.

Asymptotic distribution: the symmetric case

$$A_N = I_N - \frac{\Delta t}{2h_N^\alpha}\left[G_N + G_N^T\right],$$

the sequence $\{A_N\}_N$ is **not** yet **ready** for the **analysis**, we have the coefficient $\Delta t/2h_N^\alpha$ that varies with N.

- ❗ For *consistency* reason it makes sense to select $\Delta t \equiv h_N \equiv \nu_N$, then, since $\alpha \in (1, 2]$ we have that $\nu^{1-\alpha}$ for $\nu \to 0^+$ goes to $+\infty$.
- ⇒ We look instead at the sequence:

$$\{\nu_N^{\alpha-1} A_N\}_N = \{\nu_N^{\alpha-1} I_N - {(G_N + G_N^T)}/{2}\}_N,$$

and is such that $\|\nu^{\alpha-1} I_N\| = \nu^{\alpha-1} < C$ independently of N.

Asymptotic distribution: the symmetric case

$$A_N = I_N - \frac{\Delta t}{2h_N^\alpha}\left[G_N + G_N^T\right],$$

the sequence $\{A_N\}_N$ is **not** yet **ready** for the **analysis**, we have the coefficient ${}^{\Delta t}/_{2h_N^\alpha}$ that varies with N.

- For *consistency* reason it makes sense to select $\Delta t \equiv h_N \equiv \nu_N$, then, since $\alpha \in (1, 2]$ we have that $\nu^{1-\alpha}$ for $\nu \to 0^+$ goes to $+\infty$.
- ⇒ We look instead at the sequence:

$$\{\nu_N^{\alpha-1} A_N\}_N = \{\nu_N^{\alpha-1} I_N - {}^{(G_N+G_N^T)}/_2\}_N,$$

  and is such that $\|\nu^{\alpha-1} I_N\| = \nu^{\alpha-1} < C$ independently of N.
- $\{-{}^{(G_N+G_N^T)}/_2\}_N$ is now a *symmetric* Toeplitz sequence with **known generating function**:

$$p_\alpha(\theta) = f(\theta) + f(-\theta) = -e^{-i\theta}(1 - e^{i\theta})^\alpha - e^{i\theta}(1 - e^{-i\theta})^\alpha.$$

Asymptotic distribution: the symmetric case

$$A_N = I_N - \frac{\Delta t}{2h_N^\alpha}\left[G_N + G_N^T\right],$$

the sequence $\{A_N\}_N$ is **not** yet **ready** for the **analysis**, we have the coefficient $\Delta t/2h_N^\alpha$ that varies with N.

- For *consistency* reason it makes sense to select $\Delta t \equiv h_N \equiv \nu_N$, then, since $\alpha \in (1,2]$ we have that $\nu^{1-\alpha}$ for $\nu \to 0^+$ goes to $+\infty$.

⇒ We look instead at the sequence:

$$\{\nu_N^{\alpha-1}A_N\}_N = \{\nu_N^{\alpha-1}I_N - {}^{(G_N+G_N^T)}/2\}_N,$$

and is such that $\|\nu^{\alpha-1}I_N\| = \nu^{\alpha-1} < C$ independently of N.

- $\{-{}^{(G_N+G_N^T)}/2\}_N$ is now a *symmetric* Toeplitz sequence with **known generating function**:

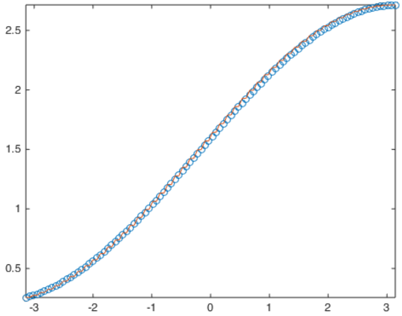
$$p_\alpha(\theta) = f(\theta) + f(-\theta) = -e^{-i\theta}(1-e^{i\theta})^\alpha - e^{i\theta}(1-e^{-i\theta})^\alpha.$$

⇒ We have just discovered that: $\{\nu_N^{\alpha-1}A_N\}_N \sim_\lambda p_\alpha(\theta)$.

Asymptotic distribution: the symmetric case

$$\{\nu_N^{\alpha-1} A_N\} = \left\{\nu_N^{\alpha-1} I_N - \frac{1}{2}\left[G_N + G_N^T\right]\right\}_N \sim_\lambda p_\alpha(\theta) = -e^{-i\theta}(1-e^{i\theta})^\alpha - e^{i\theta}(1-e^{-i\theta})^\alpha,$$

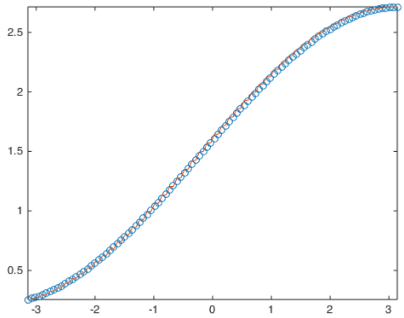
```
N = 100; alpha = 1.3;
hN = 1/(N-1); dt = hN; nu = dt;
G = glmatrix(N,alpha); I = eye(N,N);
A = nu^(alpha-1)*I-0.5*(G+G');
ev = eig(A);
f = @(t)-exp(-1i*t).*(1-exp(1i*t))
↪  .^alpha;
p = @(t)nu^(alpha-1)+0.5*(f(t)
↪  +conj(f(t)));
t = linspace(-pi,pi,N);
plot(t,ev,'o',t,sort(p(t),'ascend'),'-')
```



Asymptotic distribution: the symmetric case

$$\{\nu_N^{\alpha-1} A_N\} = \left\{\nu_N^{\alpha-1} I_N - \frac{1}{2}\left[G_N + G_N^T\right]\right\}_N \sim_\lambda p_\alpha(\theta) = -e^{-i\theta}(1-e^{i\theta})^\alpha - e^{i\theta}(1-e^{-i\theta})^\alpha,$$

```
N = 100; alpha = 1.3;
hN = 1/(N-1); dt = hN; nu = dt;
G = glmatrix(N,alpha); I = eye(N,N);
A = nu^(alpha-1)*I-0.5*(G+G');
ev = eig(A);
f = @(t)-exp(-1i*t).*(1-exp(1i*t))
↪  .^alpha;
p = @(t)nu^(alpha-1)+0.5*(f(t)
↪  +conj(f(t)));
t = linspace(-pi,pi,N);
plot(t,ev,'o',t,sort(p(t),'ascend'),'-')
```



⚠ the spectrum is **not clustered**!

CG with a non clustered spectra

Let us test the CG with different values of α and N.

α	1.8	1.5	1.2
N	Iteration		
100	49	34	16
200	87	42	17
500	155	53	18
1000	209	63	19
5000	398	92	21
10000	523	108	22

- The number if iterations grows with N,
- Smaller values of α seem to be easier.

```
A = nu^(alpha-1)*I-0.5*(G+G'); b = nu^(alpha-1)*ones(N,1);
[x,flag,relres,iter,resvec] = pcg(A,b,1e-6,N)
```

CG with a non clustered spectra

Let us test the CG with different values of α and N.

α	1.8	1.5	1.2
N	Iteration		
100	49	34	16
200	87	42	17
500	155	53	18
1000	209	63	19
5000	398	92	21
10000	523	108	22

- The number if iterations grows with N,
- Smaller values of α seem to be easier.
- We would like **number of iterations** independent on both **size** and value of α. In this case this is called having a method with a **superlinear convergence** and **robust with respect to the parameters**.

```
A = nu^(alpha-1)*I-0.5*(G+G'); b = nu^(alpha-1)*ones(N,1);
[x,flag,relres,iter,resvec] = pcg(A,b,1e-6,N)
```

CG with a non clustered spectra

Let us test the CG with different values of α and N.

α	1.8	1.5	1.2
N	Iteration		
100	49	34	16
200	87	42	17
500	155	53	18
1000	209	63	19
5000	398	92	21
10000	523	108	22

- The number if iterations grows with N,
- Smaller values of α seem to be easier.
- We would like **number of iterations** independent on both **size** and value of α. In this case this is called having a method with a **superlinear convergence** and **robust with respect to the parameters**.
- Can we?

```
A = nu^(alpha-1)*I-0.5*(G+G'); b = nu^(alpha-1)*ones(N,1);
[x,flag,relres,iter,resvec] = pcg(A,b,1e-6,N)
```

Preconditioned CG

- To try and achieve this result we need to *modify the spectrum of the system*, i.e., we need to **precondition**.

Preconditioned CG

- To try and achieve this result we need to *modify the spectrum of the system*, i.e., we need to **precondition**.
- We modify the system

$$A\mathbf{x} = \mathbf{b},$$

into

$$M^{-1}A\mathbf{x} = M^{-1}\mathbf{b},$$

with M SPD and such that $M^{-1}A$ has a **clustered spectra**.

Input: $A \in \mathbb{R}^{n\times n}$ SPD, N_{max}, $\mathbf{x}^{(0)}$, $M \in \mathbb{R}^{n\times n}$ SPD preconditioner

$\mathbf{r}^{(0)} \leftarrow \mathbf{b} - A\mathbf{x}^{(0)}$, $\mathbf{z}^{(0)} \leftarrow M^{-1}\mathbf{r}^{(0)}$, $\mathbf{p}^{(0)} \leftarrow \mathbf{z}^{(0)}$;

for $j = 0, \ldots, N_{max}$ **do**

- $\alpha_j \leftarrow {<\mathbf{r}^{(j)},\mathbf{z}^{(j)}>}/{A\mathbf{p}^{(j)},\mathbf{p}^{(j)}}$;
- $\mathbf{x}^{(j+1)} \leftarrow \mathbf{x}^{(j)} + \alpha_j \mathbf{p}^{(j)}$;
- $\mathbf{r}^{(j+1)} \leftarrow \mathbf{r}^{(j)} - \alpha_j A\mathbf{p}^{(j)}$;
- **if then**
  - **Return:** $\tilde{\mathbf{x}} = \mathbf{x}^{(j+1)}$;
- **end**
- $\mathbf{z}^{(j+1)} \leftarrow M^{-1}\mathbf{r}^{(j+1)}$;
- $\beta_j \leftarrow {<\mathbf{r}^{(j+1)},\mathbf{z}^{(j+1)}>}/{<\mathbf{r}^{(j)},\mathbf{z}^{(j)}>}$;
- $\mathbf{p}^{(j+1)} \leftarrow \mathbf{z}^{(j+1)} + \beta_j \mathbf{p}^{(j)}$;

end

Preconditioned CG

- To try and achieve this result we need to *modify the spectrum of the system*, i.e., we need to **precondition**.
- We modify the system

$$A\mathbf{x} = \mathbf{b},$$

into

$$M^{-1}A\mathbf{x} = M^{-1}\mathbf{b},$$

with M SPD and such that $M^{-1}A$ has a **clustered spectra**.

Input: $A \in \mathbb{R}^{n\times n}$ SPD, N_{max}, $\mathbf{x}^{(0)}$, $M \in \mathbb{R}^{n\times n}$ SPD preconditioner

$\mathbf{r}^{(0)} \leftarrow \mathbf{b} - A\mathbf{x}^{(0)}$, $\mathbf{z}^{(0)} \leftarrow M^{-1}\mathbf{r}^{(0)}$, $\mathbf{p}^{(0)} \leftarrow \mathbf{z}^{(0)}$;

for $j = 0, \ldots, N_{max}$ **do**

- $\alpha_j \leftarrow {}^{<\mathbf{r}^{(j)},\mathbf{z}^{(j)}>}/_{A\mathbf{p}^{(j)},\mathbf{p}^{(j)}}$;
- $\mathbf{x}^{(j+1)} \leftarrow \mathbf{x}^{(j)} + \alpha_j \mathbf{p}^{(j)}$;
- $\mathbf{r}^{(j+1)} \leftarrow \mathbf{r}^{(j)} - \alpha_j A\mathbf{p}^{(j)}$;
- **if then**
  - **Return:** $\tilde{\mathbf{x}} = \mathbf{x}^{(j+1)}$;
- **end**
- $\mathbf{z}^{(j+1)} \leftarrow M^{-1}\mathbf{r}^{(j+1)}$;
- $\beta_j \leftarrow {}^{<\mathbf{r}^{(j+1)},\mathbf{z}^{(j+1)}>}/_{<\mathbf{r}^{(j)},\mathbf{z}^{(j)}>}$;
- $\mathbf{p}^{(j+1)} \leftarrow \mathbf{z}^{(j+1)} + \beta_j \mathbf{p}^{(j)}$;

end

M^{-1} has to be easy to apply, possibly it has to have the *same cost of multiplying by A*.

Circulant preconditioners for Toeplitz matrices

- If M is circulant than applying M^{-1} costs $O(n \log n)$ operations, same as applying A.

Circulant preconditioners for Toeplitz matrices

💡 If M is circulant than applying M^{-1} costs $O(n \log n)$ operations, same as applying A.

👁 Observe that, nevertheless, this **doubles the cost per iteration**, can we do better?

Circulant preconditioners for Toeplitz matrices

💡 If M is circulant than applying M^{-1} costs $O(n \log n)$ operations, same as applying A.

👁 Observe that, nevertheless, this **doubles the cost per iteration**, can we do better?

ω-circulant matrices

Let $\omega = \exp(i\theta)$ for $\theta \in [-\pi, \pi]$. A matrix $W_n^{(\omega)}$ of size n is said to be an ω–**circulant matrix** if it has the spectral decomposition

$$W_n^{(\omega)} = \Omega_n^H F_n^H \Lambda_n F_n \Omega_n,$$

where F_n is the Fourier matrix and $\Omega_n = \operatorname{diag}(1, \omega^{-1/n}, \ldots, \omega^{-(n-1)/n})$ and Λ_n is the diagonal matrix of the eigenvalues. In particular 1–circulant matrices are circulant matrices while $\{-1\}$–circulant matrices are the skew–circulant matrices.

Circulant preconditioners for Toeplitz matrices

If M is circulant than applying M^{-1} costs $O(n \log n)$ operations, same as applying A.
Observe that, nevertheless, this **doubles the cost per iteration**, can we do better?

ω-circulant matrices

Let $\omega = \exp(i\theta)$ for $\theta \in [-\pi, \pi]$. A matrix $W_n^{(\omega)}$ of size n is said to be an ω–**circulant matrix** if it has the spectral decomposition

$$W_n^{(\omega)} = \Omega_n^H F_n^H \Lambda_n F_n \Omega_n,$$

where F_n is the Fourier matrix and $\Omega_n = \operatorname{diag}(1, \omega^{-1/n}, \ldots, \omega^{-(n-1)/n})$ and Λ_n is the diagonal matrix of the eigenvalues. In particular 1–circulant matrices are circulant matrices while $\{-1\}$–circulant matrices are the skew–circulant matrices.

We can use them to reduce the overall cost of the preconditioning step!

Circulant preconditioners for Toeplitz matrices

The key idea is observing that we can decompose any Toeplitz matrix into the **sum of a circulant and of a skew-circulant matrix**

$$T_n = U_n + V_n,\ U_n = F_n^H \Lambda_n^{(1)} F_n,\ V_n = \Omega_n^H F_n^H \Lambda_n^{(2)} F_n \Omega_n$$

where

$$\begin{aligned} \mathbf{e}_1^T U_n &= 1/2 \left[t_0, t_{-1} + t_{n-1}, \ldots, t_{-(n-1)+t_1}\right], \\ W_n \mathbf{e}_1 &= 1/2 \left[t_0, -(t_{n-1} - t_{-1}), \ldots, -(t_{-1} - t_{n-1})\right]^T. \end{aligned}$$

Circulant preconditioners for Toeplitz matrices

The key idea is observing that we can decompose any Toeplitz matrix into the **sum of a circulant and of a skew-circulant matrix**

$$T_n = U_n + V_n,\ U_n = F_n^H \Lambda_n^{(1)} F_n,\ V_n = \Omega_n^H F_n^H \Lambda_n^{(2)} F_n \Omega_n$$

where

$$\begin{aligned} \mathbf{e}_1^T U_n &= {}^1/_2 \left[t_0, t_{-1} + t_{n-1}, \ldots, t_{-(n-1)+t_1} \right], \\ W_n \mathbf{e}_1 &= {}^1/_2 \left[t_0, -(t_{n-1} - t_{-1}), \ldots, -(t_{-1} - t_{n-1}) \right]^T. \end{aligned}$$

Then we can compute the product

$$\begin{aligned} C_n^{-1} T_n = C_n^{-1} (U_n + V_n) &= C_n^{-1} \left(F_n^H \Lambda_n^{(1)} F_n + \Omega_n^H F_n^H \Lambda_n^{(2)} F_n \Omega_n \right) \\ &= F_n^H \Lambda_n^{-1} F_n \left(F_n^H \Lambda_n^{(1)} F_n + \Omega_n^H F_n^H \Lambda_n^{(2)} F_n \Omega_n \right) \\ &= F_n^H \left[\Lambda_n^{-1} \left(\Lambda_n^{(1)} + F_n \Omega_n^H F_n^H \Lambda_n^{(2)} F_n \Omega_n F_n^H \right) \right] F_n. \end{aligned}$$

Circulant preconditioners for Toeplitz matrices

The key idea is observing that we can decompose any Toeplitz matrix into the **sum of a circulant and of a skew-circulant matrix**

$$T_n = U_n + V_n,\ U_n = F_n^H \Lambda_n^{(1)} F_n,\ V_n = \Omega_n^H F_n^H \Lambda_n^{(2)} F_n \Omega_n$$

where

$$\begin{aligned} \mathbf{e}_1^T U_n &= {}^1\!/\!_2 \left[t_0, t_{-1} + t_{n-1}, \ldots, t_{-(n-1)+t_1} \right], \\ W_n \mathbf{e}_1 &= {}^1\!/\!_2 \left[t_0, -(t_{n-1} - t_{-1}), \ldots, -(t_{-1} - t_{n-1}) \right]^T. \end{aligned}$$

Then we can compute the product

$$C_n^{-1} T_n = F_n^H \left[\Lambda_n^{-1} \left(\Lambda_n^{(1)} + F_n \Omega_n^H F_n^H \Lambda_n^{(2)} F_n \Omega_n F_n^H \right) \right] F_n.$$

And solve $C_n^{-1} T_n \mathbf{x} = C_n^{-1} \mathbf{b}$ as

$$\Lambda_n^{-1} \left(\Lambda_n^{(1)} + F_n \Omega_n^H F_n^H \Lambda_n^{(2)} F_n \Omega_n F_n^H \right) \underbrace{F_n \mathbf{x}}_{=\tilde{\mathbf{x}}} = \underbrace{\Lambda_n^{-1} F_n \mathbf{b}}_{=\tilde{\mathbf{b}}}$$

Circulant preconditioners for Toeplitz matrices

The key idea is observing that we can decompose any Toeplitz matrix into the **sum of a circulant and of a skew-circulant matrix**

$$T_n = U_n + V_n, \; U_n = F_n^H \Lambda_n^{(1)} F_n, \; V_n = \Omega_n^H F_n^H \Lambda_n^{(2)} F_n \Omega_n$$

where

$$\mathbf{e}_1^T U_n = ^1\!/_2 \left[t_0, t_{-1} + t_{n-1}, \ldots, t_{-(n-1)+t_1} \right],$$
$$W_n \mathbf{e}_1 = ^1\!/_2 \left[t_0, -(t_{n-1} - t_{-1}), \ldots, -(t_{-1} - t_{n-1}) \right]^T.$$

Then we can compute the product

$$C_n^{-1} T_n = F_n^H \left[\Lambda_n^{-1} \left(\Lambda_n^{(1)} + F_n \Omega_n^H F_n^H \Lambda_n^{(2)} F_n \Omega_n F_n^H \right) \right] F_n.$$

And solve $C_n^{-1} T_n \mathbf{x} = C_n^{-1} \mathbf{b}$ as

$$\Lambda_n^{-1} \left(\Lambda_n^{(1)} + F_n \Omega_n^H F_n^H \Lambda_n^{(2)} F_n \Omega_n F_n^H \right) \underbrace{F_n \mathbf{x}}_{=\tilde{\mathbf{x}}} = \underbrace{\Lambda_n^{-1} F_n \mathbf{b}}_{=\tilde{\mathbf{b}}}$$ 4 FFTs per iteration!

Circulant preconditioners for Toeplitz matrices

❓ This is then *computationally efficient*, can we find the right circulant matrix to have a clustered spectra?

Circulant preconditioners for Toeplitz matrices

❓ This is then *computationally efficient*, can we find the right circulant matrix to have a clustered spectra?

🔧 We need to change the problem into an equivalent one: the aim is **discharging everything on the generating functions**!

Circulant preconditioners for Toeplitz matrices

❓ This is then *computationally efficient*, can we find the right circulant matrix to have a clustered spectra?
🔧 We need to change the problem into an equivalent one: the aim is **discharging everything on the generating functions**!

Continuous convolution

Given two scalar functions f and g in the Schwartz space, i.e., $f, g \in \mathcal{C}^{\infty}(\mathbb{R})$ such that $\exists\, C^{(f)}_{\alpha,\beta}, C^{(g)}_{\alpha',\beta'} \in \mathbb{R}$ with $\|x^{\alpha}\partial_{\beta} f(x)\|_{\infty} \leq C^{\alpha\beta}$ and $\|x^{\alpha'}\partial_{\beta'} g(x)\|_{\infty} \leq C^{\alpha'\beta'}$, α, β, α', β' scalar indices, we define the **convolution operation**, "$*$", as

$$[f * g](t) = \int_{-\infty}^{+\infty} f(\tau) g(t-\tau)\, d\tau = \int_{-\infty}^{+\infty} g(\tau) f(t-\tau)\, d\tau.$$

Circulant preconditioners for Toeplitz matrices

❓ This is then *computationally efficient*, can we find the right circulant matrix to have a clustered spectra?

🔧 We need to change the problem into an equivalent one: the aim is **discharging everything on the generating functions**!

Discrete convolution

For two arbitrary 2π–periodic continuous functions,

$$f(\theta) = \sum_{k=-\infty}^{+\infty} t_k e^{ik\theta} \text{ and } g = \sum_{k=-\infty}^{+\infty} s_k e^{ik\theta}$$

their **convolution product** is given by

$$[f * g](\theta) = \sum_{k=-\infty}^{+\infty} s_k t_k e^{ik\theta}.$$

Circulant preconditioners for Toeplitz matrices

❓ This is then *computationally efficient*, can we find the right circulant matrix to have a clustered spectra?

🔧 We need to change the problem into an equivalent one: the aim is **discharging everything on the generating functions**!

Using a Kernel

Given a kernel $\mathcal{K}_n(\theta)$ defined on $[0, 2\pi]$ and a generating function f for a Toeplitz sequence $T_n(f)$, we consider the circulant matrix C_n with eigenvalues given by

$$\lambda_j(C_n) = [\mathcal{K}_n * f]\left(\frac{2\pi j}{n}\right), 0 \leq j < n,$$

Circulant preconditioners for Toeplitz matrices

❓ This is then *computationally efficient*, can we find the right circulant matrix to have a clustered spectra?

🔧 We need to change the problem into an equivalent one: the aim is **discharging everything on the generating functions**!

Using a Kernel

Given a kernel $\mathcal{K}_n(\theta)$ defined on $[0, 2\pi]$ and a generating function f for a Toeplitz sequence $T_n(f)$, we consider the circulant matrix C_n with eigenvalues given by

$$\lambda_j(C_n) = [\mathcal{K}_n * f]\left(\frac{2\pi j}{n}\right), 0 \leq j < n,$$

- We have rewritten the problem of **finding an appropriate preconditioner** to the problem of **approximating the generating function** of the underlying Toeplitz matrix.

Circulant preconditioners for Toeplitz matrices

Theorem (R. H. Chan and Yeung 1992)

Lef f be a 2π–periodic continuous positive function. Let $\mathcal{K}_n(\theta)$ be a kernel such that $\mathcal{K}_n * f \stackrel{n\to+\infty}{\longrightarrow} f$ uniformly on $[-\pi,\pi]$. If C_n is the sequence of circulant matrices with eigenvalues given by

$$\lambda_j(C_n) = [\mathcal{K}_n * f]\left(\frac{2\pi j}{n}\right), 0 \leq j < n,$$

then the spectra of the sequence $\{C_n^{-1}T_n(f)\}_n$ is clustered around 1.

- Is this the result we need?

Circulant preconditioners for Toeplitz matrices

Theorem (R. H. Chan and Yeung 1992)

Lef f be a 2π–periodic continuous positive function. Let $\mathcal{K}_n(\theta)$ be a kernel such that $\mathcal{K}_n * f \stackrel{n\to+\infty}{\longrightarrow} f$ uniformly on $[-\pi,\pi]$. If C_n is the sequence of circulant matrices with eigenvalues given by

$$\lambda_j(C_n) = [\mathcal{K}_n * f]\left(\frac{2\pi j}{n}\right), 0 \leq j < n,$$

then the spectra of the sequence $\{C_n^{-1}T_n(f)\}_n$ is clustered around 1.

- ❓ Is this the result we need?
- ❗ It requires a *continuous positive function* generating function f! Ours is:

$$p_\alpha(\theta) = -e^{-i\theta}(1-e^{i\theta})^\alpha - e^{i\theta}(1-e^{-i\theta})^\alpha,$$

  and it does seem to have a zero.

Circulant preconditioners: cases with a zero

Order of the zero

Let $f : [a, b] \subset \mathbb{R} \to \mathbb{R}$ be a continuous nonnegative function. We say that f has a zero order $\beta > 0$ at $\theta_0 \in [a, b]$ if there exist two real constants $C_1, C_2 > 0$ such that

$$\liminf_{\theta \to \theta_0} \frac{f(\theta)}{|\theta - \theta_0|^\beta} = C_1, \quad \limsup_{\theta \to \theta_0} \frac{f(\theta)}{|\theta - \theta_0|^\beta} = C_2.$$

Circulant preconditioners: cases with a zero

Order of the zero

Let $f : [a, b] \subset \mathbb{R} \to \mathbb{R}$ be a continuous nonnegative function. We say that f has a zero order $\beta > 0$ at $\theta_0 \in [a, b]$ if there exist two real constants $C_1, C_2 > 0$ such that

$$\liminf_{\theta \to \theta_0} \frac{f(\theta)}{|\theta - \theta_0|^\beta} = C_1, \quad \limsup_{\theta \to \theta_0} \frac{f(\theta)}{|\theta - \theta_0|^\beta} = C_2.$$

Proposition (Donatelli, Mazza, and Serra-Capizzano 2016)

Given $\alpha \in (1, 2)$, then the function $p_\alpha(\theta)$ is nonnegative and has a zero of order α at 0.

Circulant preconditioners: cases with a zero

Order of the zero

Let $f : [a, b] \subset \mathbb{R} \to \mathbb{R}$ be a continuous nonnegative function. We say that f has a zero order $\beta > 0$ at $\theta_0 \in [a, b]$ if there exist two real constants $C_1, C_2 > 0$ such that

$$\liminf_{\theta \to \theta_0} \frac{f(\theta)}{|\theta - \theta_0|^\beta} = C_1, \quad \limsup_{\theta \to \theta_0} \frac{f(\theta)}{|\theta - \theta_0|^\beta} = C_2.$$

Proposition (Donatelli, Mazza, and Serra-Capizzano 2016)

Given $\alpha \in (1, 2)$, then the function $p_\alpha(\theta)$ is nonnegative and has a zero of order α at 0.

Proof. We first prove that is nonnegative by direct computation

$$p_\alpha(\theta) = -\sum_{k=-1}^{+\infty} g_{k+1}^{(\alpha)} (e^{ik\theta} + e^{-ik\theta})$$

Circulant preconditioners: cases with a zero

Order of the zero

Let $f : [a, b] \subset \mathbb{R} \to \mathbb{R}$ be a continuous nonnegative function. We say that f has a zero order $\beta > 0$ at $\theta_0 \in [a, b]$ if there exist two real constants $C_1, C_2 > 0$ such that

$$\liminf_{\theta \to \theta_0} \frac{f(\theta)}{|\theta - \theta_0|^\beta} = C_1, \quad \limsup_{\theta \to \theta_0} \frac{f(\theta)}{|\theta - \theta_0|^\beta} = C_2.$$

Proposition (Donatelli, Mazza, and Serra-Capizzano 2016)

Given $\alpha \in (1, 2)$, then the function $p_\alpha(\theta)$ is nonnegative and has a zero of order α at 0.

Proof. We first prove that is nonnegative by direct computation

$$p_\alpha(\theta) = -\left[2g_1^{(\alpha)} + (g_0^{(\alpha)} + g_2^{(\alpha)})(e^{i\theta} + e^{-i\theta}) + \sum_{k=2}^{+\infty} g_{k+1}^{(\alpha)}(e^{ik\theta} + e^{-ik\theta})\right]$$

Circulant preconditioners: cases with a zero

Order of the zero

Let $f : [a, b] \subset \mathbb{R} \to \mathbb{R}$ be a continuous nonnegative function. We say that f has a zero order $\beta > 0$ at $\theta_0 \in [a, b]$ if there exist two real constants $C_1, C_2 > 0$ such that

$$\liminf_{\theta \to \theta_0} \frac{f(\theta)}{|\theta - \theta_0|^\beta} = C_1, \quad \limsup_{\theta \to \theta_0} \frac{f(\theta)}{|\theta - \theta_0|^\beta} = C_2.$$

Proposition (Donatelli, Mazza, and Serra-Capizzano 2016)

Given $\alpha \in (1, 2)$, then the function $p_\alpha(\theta)$ is nonnegative and has a zero of order α at 0.

Proof. We first prove that is nonnegative by direct computation

$$p_\alpha(\theta) = -\left[2g_1^{(\alpha)} + 2(g_0^{(\alpha)} + g_2^{(\alpha)})\cos\theta + 2\sum_{k=2}^{+\infty} g_{k+1}^{(\alpha)} cos(k\theta)\right]$$

Circulant preconditioners: cases with a zero

Order of the zero

Let $f : [a, b] \subset \mathbb{R} \to \mathbb{R}$ be a continuous nonnegative function. We say that f has a zero order $\beta > 0$ at $\theta_0 \in [a, b]$ if there exist two real constants $C_1, C_2 > 0$ such that

$$\liminf_{\theta \to \theta_0} \frac{f(\theta)}{|\theta - \theta_0|^{\beta}} = C_1, \quad \limsup_{\theta \to \theta_0} \frac{f(\theta)}{|\theta - \theta_0|^{\beta}} = C_2.$$

Proposition (Donatelli, Mazza, and Serra-Capizzano 2016)

Given $\alpha \in (1, 2)$, then the function $p_\alpha(\theta)$ is nonnegative and has a zero of order α at 0.

Proof. We first prove that is nonnegative by direct computation

$$p_\alpha(\theta) = -\left[2g_1^{(\alpha)} + 2(g_0^{(\alpha)} + g_2^{(\alpha)}) \cos\theta + 2\sum_{k=2}^{+\infty} g_{k+1}^{(\alpha)} cos(k\theta)\right] \geq -2\sum_{k=-1}^{+\infty} g_{k+1}^{(\alpha)} = 0.$$

Circulant preconditioners: cases with a zero

Proposition (Donatelli, Mazza, and Serra-Capizzano 2016)

Given $\alpha \in (1, 2)$, then the function $p_\alpha(\theta)$ is nonnegative and has a zero of order α at 0.

Proof. Then we focus on the zero. Let us rewrite

$$1 - e^{i\theta} = \sqrt{2 - 2\cos\theta}\, e^{i\phi}, \quad 1 - e^{-i\theta} = \sqrt{2 - 2\cos\theta}\, e^{i\psi},$$

where

$$\phi = \begin{cases} \arctan\left(\frac{-\sin\theta}{1-\cos\theta}\right), & \theta \neq 0, \\ \lim_{\theta\to 0^+} \arctan\left(\frac{-\sin\theta}{1-\cos\theta}\right) = -\frac{\pi}{2}, & \theta = 0. \end{cases} \quad \psi = -\phi.$$

Circulant preconditioners: cases with a zero

Proposition (Donatelli, Mazza, and Serra-Capizzano 2016)

Given $\alpha \in (1,2)$, then the function $p_\alpha(\theta)$ is nonnegative and has a zero of order α at 0.

Proof. Then we focus on the zero. Let us rewrite

$$1 - e^{i\theta} = \sqrt{2-2\cos\theta}\, e^{i\phi}, \quad 1 - e^{-i\theta} = \sqrt{2-2\cos\theta}\, e^{i\psi},$$

and write

$$\begin{aligned} p_\alpha(\theta) &= -\, e^{-i\theta}(\sqrt{2-2\cos\theta}\, e^{i\phi})^\alpha - e^{i\theta}(\sqrt{2-2\cos\theta}\, e^{-i\phi})^\alpha \\ &= -\,\sqrt{(2-2\cos\theta)^\alpha}\, e^{i(\alpha\phi-\theta)} - \sqrt{(2-2\cos\theta)^\alpha}\, e^{-i(\alpha\phi-\theta)} \\ &= -\,2\sqrt{(2-2\cos\theta)^\alpha}\, r_\alpha(\theta), \qquad r_\alpha(\theta) = \cos(\alpha\phi - \theta). \end{aligned}$$

Circulant preconditioners: cases with a zero

Proposition (Donatelli, Mazza, and Serra-Capizzano 2016)

Given $\alpha \in (1,2)$, then the function $p_\alpha(\theta)$ is nonnegative and has a zero of order α at 0.

Proof. Then we focus on the zero. Let us rewrite

$$1 - e^{i\theta} = \sqrt{2 - 2\cos\theta}\, e^{i\phi}, \quad 1 - e^{-i\theta} = \sqrt{2 - 2\cos\theta}\, e^{i\psi},$$

and write

$$p_\alpha(\theta) = -2\sqrt{(2 - 2\cos\theta)^\alpha}\, r_\alpha(\theta), \qquad r_\alpha(\theta) = \cos(\alpha\phi - \theta).$$

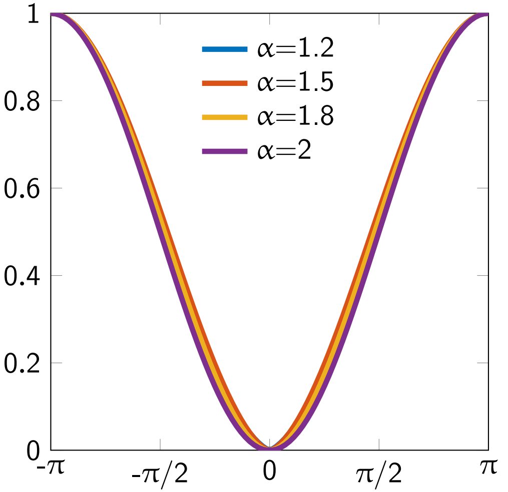
Since $\lim_{\theta\to0^-} r_\alpha(\theta) = \lim_{\theta\to0^+} r_\alpha(\theta) = \cos(\alpha\pi/2)$, we find

$$\lim_{\theta\to0} \frac{p_\alpha(\theta)}{|\theta|^\alpha} = -2\lim_{\theta\to0} \frac{(2 - 2\cos\theta)^{\alpha/2}}{|\theta|^\alpha} r_\alpha(\theta) = -2\cos(\alpha\pi/2) \in (0,2),$$

i.e., p_α has a zero of order α at 0 according to the definition. □

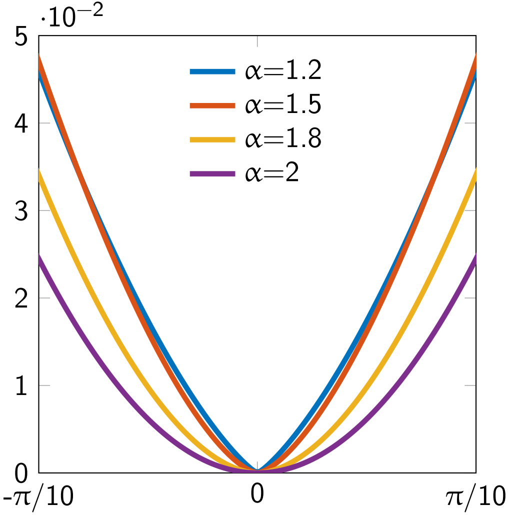
Circulant preconditioners: cases with a zero

```
t = linspace(-pi,pi,100);
f = @(alpha)
 ↪  -exp(-1i*t).*(1-exp(1i*t)).^alpha;
p = @(alpha) f(alpha) +
 ↪  conj(f(alpha));
plot(t,p(1.2)./max(p(1.2)),...
 t,p(1.5)./max(p(1.5)),...
 t,p(1.8)./max(p(1.8)),
 t,p(2)./max(p(2)),...
 'LineWidth',2);
legend({'\alpha=1.2','\alpha=1.5',...
 '\alpha=1.8','\alpha=2'},...
 'Location','north');
```



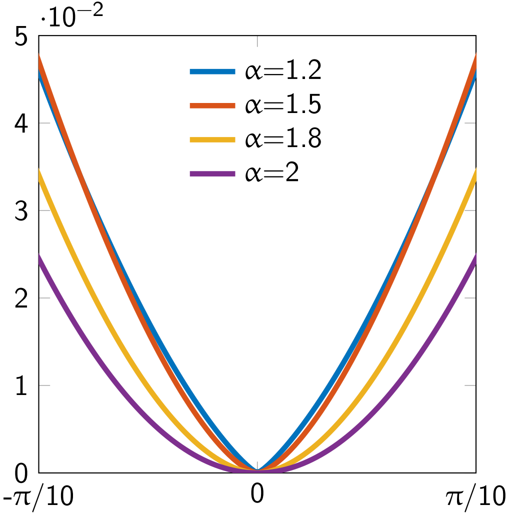
Circulant preconditioners: cases with a zero

- $p_2(\theta) = 2(2 - 2\cos\theta)$, i.e., 2×Laplacian generating function,
- $p_\alpha(\theta)/\|p_\alpha\|_\infty$ approaches the order of the zero of the Laplacian in 0, i.e., it increases up to 2 as α tends to 2.



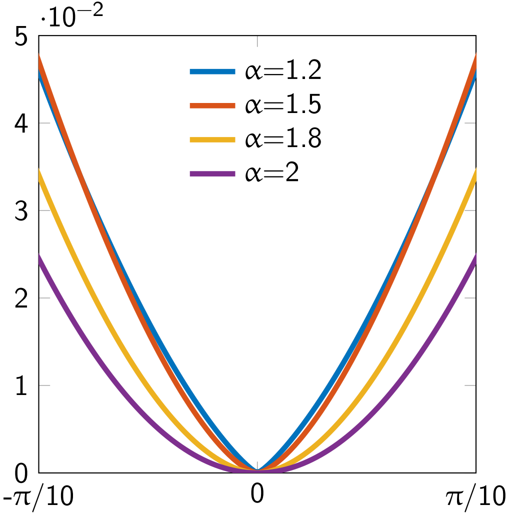
Circulant preconditioners: cases with a zero

- $p_2(\theta) = 2(2 - 2\cos\theta)$, i.e., 2×Laplacian generating function,
- $p_\alpha(\theta)/\|p_\alpha\|_\infty$ approaches the order of the zero of the Laplacian in 0, i.e., it increases up to 2 as α tends to 2.
- What can we do for the case in this case?



Circulant preconditioners: cases with a zero

- $p_2(\theta) = 2(2 - 2\cos\theta)$, i.e., $2\times$Laplacian generating function,
- $p_\alpha(\theta)/\|p_\alpha\|_\infty$ approaches the order of the zero of the Laplacian in 0, i.e., it increases up to 2 as α tends to 2.
- What can we do for the case in this case?
- **matching the zeros** of the generating function, *heuristically*, if the preconditioner have a spectrum that behaves as a function g with zeros of the same order, and in the same place of f, then f/g no loner have the problematic behavior...



Generalized Jackson Kernel

Generalized Jackson Kernel

Given $\theta \in [-\pi, \pi]$, $\mathbb{N} \ni r \geq 1$ and $\mathbb{N} \ni m > 0$ such that $r(m-1) < n \leq rm$, i.e., $m = \lceil n/r \rceil$, the generalized Jackson kernel function is defined as,

$$\mathcal{K}_{m,2r}(\theta) = \frac{k_{m,2r}}{m^{2r-1}} \left(\frac{\sin(m\theta/2)}{\sin(\theta/2)} \right)^{2r}, \; k_{m,2r} \text{ s.t. } \frac{1}{2\pi} \int_{-\pi}^{\pi} \mathcal{K}_{m,2r}(\theta) d\theta = 1.$$

We build a **circulant preconditioner** $J_{n,m,r}$ from its eigenvalues using the Jackson kernel

$$\lambda_j(J_{n,m,r}) = [\mathcal{K}_{m,2r} * f] \left(\frac{2j\pi}{n} \right), \; j = 0, \ldots, n-1.$$

Generalized Jackson Kernel

Theorem (R. H. Chan, Ng, and Yip 2002)

Let f be a nonnegative 2π–periodic continuous function with a zero of order 2ν at θ_0. Let $r > \nu$ and $m = \lceil n/r \rceil$. Then there exists numbers a, b independent from n and such that the spectrum of $J_{n,m,r}^{-1} T_n(f)$ is clustered in $[a, b]$ and at most $2\nu + 1$ eigenvalues are not in $[a, b]$ for n sufficiently large.

We build a **circulant preconditioner** $J_{n,m,r}$ from its eigenvalues using the Jackson kernel

$$\lambda_j(J_{n,m,r}) = [\mathcal{K}_{m,2r} * f]\left(\frac{2j\pi}{n}\right), \quad j = 0, \ldots, n-1.$$

Generalized Jackson Kernel

Theorem (R. H. Chan, Ng, and Yip 2002)

Let f be a nonnegative 2π–periodic continuous function with a zero of order 2ν at θ_0. Let $r > \nu$ and $m = \lceil n/r \rceil$. Then there exists numbers a, b independent from n and such that the spectrum of $J_{n,m,r}^{-1} T_n(f)$ is clustered in $[a, b]$ and at most $2\nu + 1$ eigenvalues are not in $[a, b]$ for n sufficiently large.

We build a **circulant preconditioner** $J_{n,m,r}$ from its eigenvalues using the Jackson kernel

$$\lambda_j(J_{n,m,r}) = [\mathcal{K}_{m,2r} * f]\left(\frac{2j\pi}{n}\right), \quad j = 0, \ldots, n-1.$$

With some work can be **generalized** to the case of **multiple zeros of different order**,

Generalized Jackson Kernel

Theorem (R. H. Chan, Ng, and Yip 2002)

Let f be a nonnegative 2π–periodic continuous function with a zero of order 2ν at θ_0. Let $r > \nu$ and $m = \lceil n/r \rceil$. Then there exists numbers a, b independent from n and such that the spectrum of $J_{n,m,r}^{-1} T_n(f)$ is clustered in $[a, b]$ and at most $2\nu + 1$ eigenvalues are not in $[a, b]$ for n sufficiently large.

We build a **circulant preconditioner** $J_{n,m,r}$ from its eigenvalues using the Jackson kernel

$$\lambda_j(J_{n,m,r}) = [\mathcal{K}_{m,2r} * f]\left(\frac{2j\pi}{n}\right), \quad j = 0, \ldots, n-1.$$

- With some work can be **generalized** to the case of **multiple zeros of different order**,
- One can prove also that a and b are **bounded away from zero**.

Time to do some tests

We consider the following **circulant preconditioners**,

Dirichlet kernel, a.k.a. the Strang circulant preconditioner

$$\mathcal{D}_n(\theta) = \frac{\sin\left((n + 1/2)\theta\right)}{\sin\left(\theta/2\right)} \qquad \begin{cases} t_k, & 0 < k \leq \lfloor n/2 \rfloor, \\ t_{k-n}, & \lfloor n/2 \rfloor < j < n, \\ c_{n+k}, & 0 < -k < n. \end{cases}$$

Modified Dirichlet kernel, a.k.a. the T. Chan circulant preconditioner

$$1/2\left(\mathcal{D}_{n-1}(\theta) + \mathcal{D}_{n-2}(\theta)\right) \qquad \begin{cases} t_1 + 1/2\bar{t}_{n-1}, & k = 1, \\ t_k + t_{n-k}, & 2 \leq k \leq n-2, \\ 1/2 t_{n-1} + \bar{t}_1, & k = n-1. \end{cases}$$

R.H. Chan $\mathcal{D}_{n-1}(\theta)$ $\quad t_k + \bar{t}_{n-k}$, $0 < k \leq n-1$.

Jackson with $r = 2$.

Time to do some tests

We consider the following **circulant preconditioners**,

Dirichlet kernel, a.k.a. the Strang circulant preconditioner

```
c = fft([t(1:n/2);0;conj(t(n/2:-1:2))].')';
```

Modified Dirichlet kernel, a.k.a. the T. Chan circulant preconditioner

```
coef = (1/n:1/n:1-1/n)';
c = fft([t(1);(1-coef).*t(2:n)+coef.*t1]);
```

R.H. Chan `c = fft([t(1);t(2:n)+t1].')';`

Jackson with $r = 2$.

Time to do some tests

We consider the following **circulant preconditioners**,

Dirichlet kernel, a.k.a. the Strang circulant preconditioner

```
c = fft([t(1:n/2);0;conj(t(n/2:-1:2))].')';
```

Modified Dirichlet kernel, a.k.a. the T. Chan circulant preconditioner

```
coef = (1/n:1/n:1-1/n)';
c = fft([t(1);(1-coef).*t(2:n)+coef.*t1]);
```

R.H. Chan
```
c = fft([t(1);t(2:n)+t1].')';
```

Jackson with $r = 2$.

We test both **clustering properties** and **convergence behavior** inside the **P**reconditioned **C**onjugate **G**radient algorithm.

Jackson Kernel Circulant Preconditioner

For $r = 2, 3, 4$ it can be built as

```
n = length(t);
t1 = conj(t(n:-1:2));
if r == 2 || r == 3 || r == 4
 coef = convol(n,r).';
 c = [t(1)*coef(1)
 ↪  (coef(2:n).*t(2:n)...
 +coef(n:-1:2).*t1).'];
 c = fft(c)';
else
 error('r needs to be 2, 3 or 4');
end
c = real(c);
```

```
function [ c ] = jacksonprec( t,r )
m = floor(n/r); a = 1:-1/m:1/m; r0 = 1;
coef = [a(m:-1:2) a];
while r0 < r
 M = (2*r0+3)*m; b1 = zeros(M,1);
 c = zeros(M,1); c(1:m) = a;
 c(M:-1:M-m+2) = a(2:m);
 b1(m:m+2*r0*(m-1)) = coef;
 tp = ifft(fft(b1).*fft(c));
 coef = real(tp(1:2*(r0+1)*(m-1)+1));
 r0 = r0+1;
end
M = r*(m-1)+1;
coef = [coef(M:-1:1)' zeros(1,n-M)]';
coef = coef';
end
```

Back to the example

We try to solve again

$$\begin{cases} \frac{\partial W}{\partial t} = \theta \, {}^{RL}D^{\alpha}_{[0,x]} W(x,t) + (1-\theta)\, {}^{RL}D^{\alpha}_{[x,1]} W(x,t), & \theta \in [0,1], \\ W(0,t) = W(1,t) = 0, \\ W(x,t) = W_0(x). \end{cases}$$

Back to the example

We try to solve again for $\theta = 1/2$

$$T_{N-2}(p_\alpha(\theta))\mathbf{w}^{n+1} \equiv \left(\frac{h_N^\alpha}{\Delta t} I_{N-2} - \frac{1}{2}\left[G_{N-2} + G_{N-2}^T\right]\right)\mathbf{w}^{n+1} = \frac{h_N^\alpha}{\Delta t}\mathbf{w}^n$$

We have removed the *rank corrections* due to the boundary conditions to have a **pure Toeplitz** matrix, i.e., we solve the equation only in the inner nodes.

Back to the example

We try to solve again

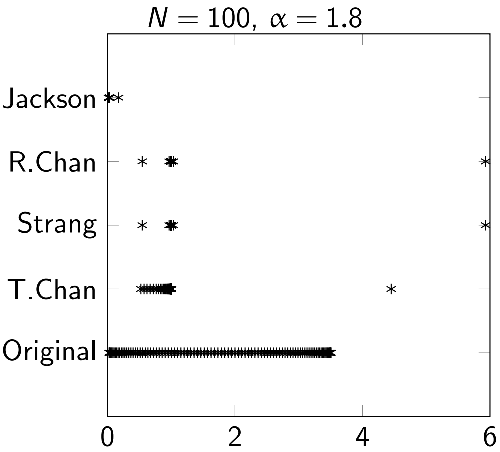
$$T_{N-2}(p_\alpha(\theta))\mathbf{w}^{n+1} \equiv \left(\frac{h_N^\alpha}{\Delta t} I_{N-2} - \frac{1}{2}\left[G_{N-2} + G_{N-2}^T\right]\right)\mathbf{w}^{n+1} = \frac{h_N^\alpha}{\Delta t}\mathbf{w}^n$$

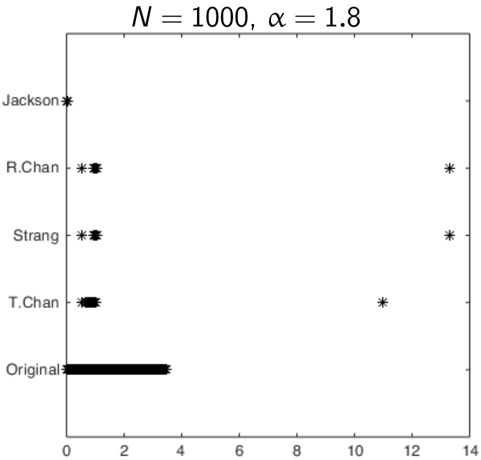
We have removed the *rank corrections* due to the boundary conditions to have a **pure Toeplitz** matrix, i.e., we solve the equation only in the inner nodes.

```
%% Problem data
theta = 0.5;
alpha = 1.8;
w0 = @(x) 5*x.*(1-x);
%% Discretization data
N  = 10;
hN = 1/(N-1); x = 0:hN:1;
dt = hN; t = 0:dt:1;
```

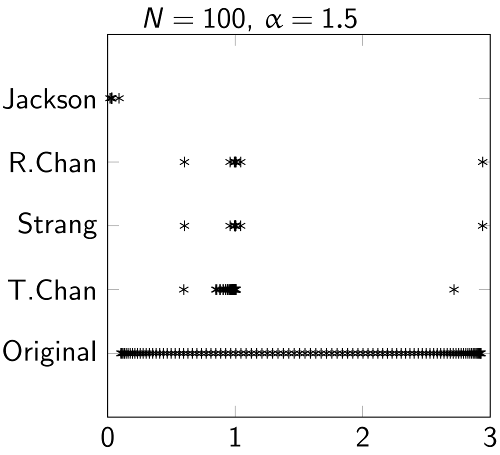
```
%% Discretize
G = glmatrix(N,alpha);
Gr = G(2:N-1,2:N-1); Grt = Gr.';
I = eye(N-2,N-2);
% Left-hand side
nu = hN^alpha/dt;
A = nu*I - (theta*Gr + (1-theta)*Grt);
% Right-hand side
w = w0(x).';
```

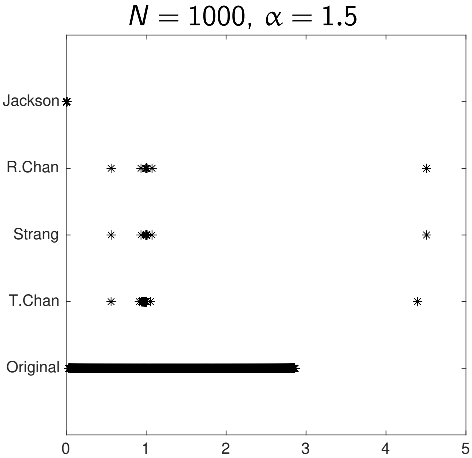
A look at the spectrum



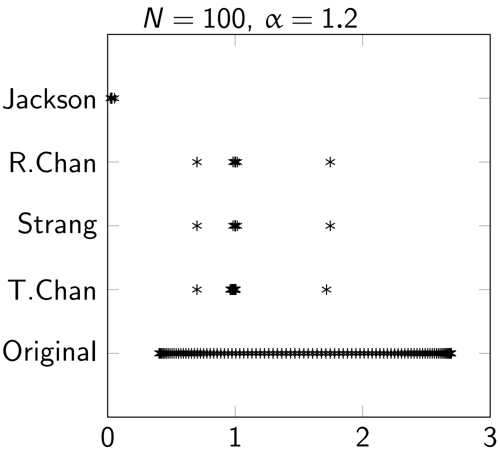


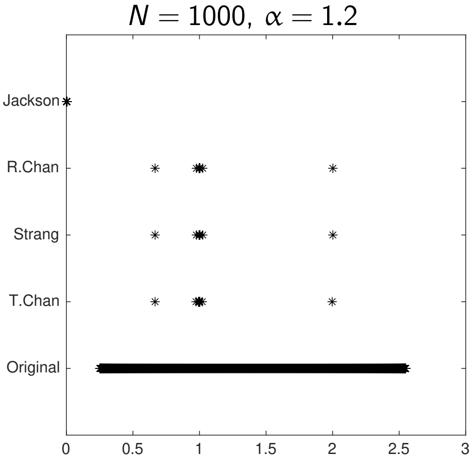
A look at the spectrum



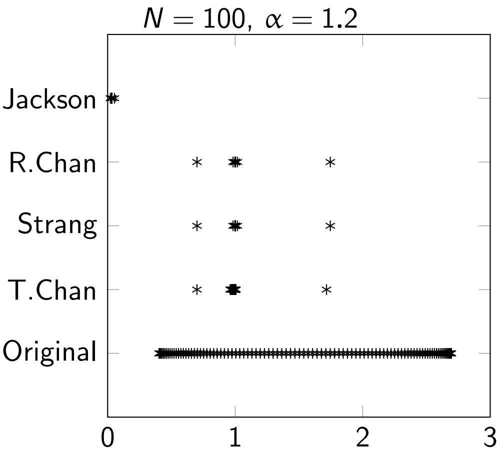


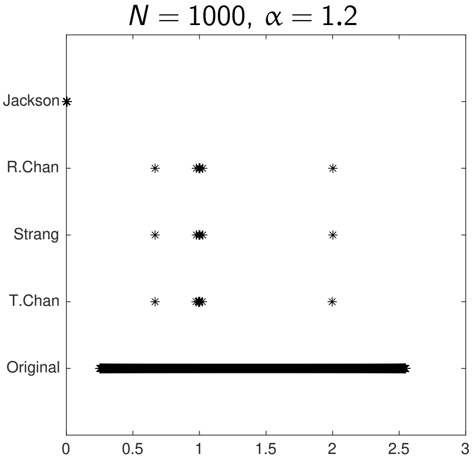
A look at the spectrum





A look at the spectrum

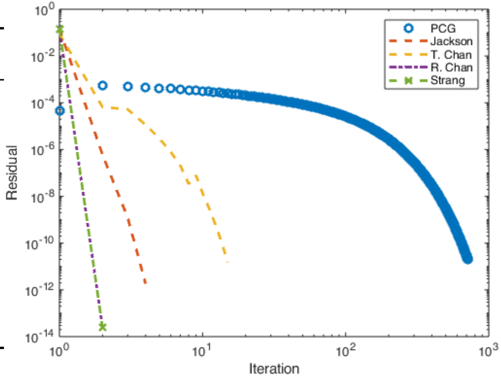




Can you guess what is happening with the Jackson Kernel preconditioner?

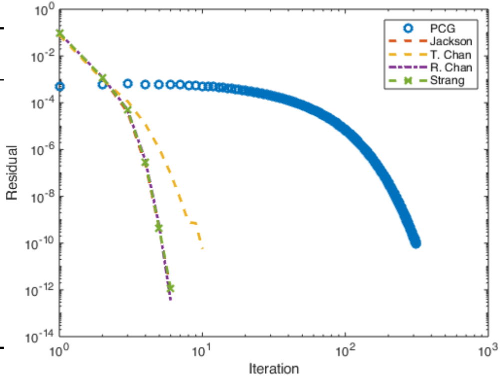
↓ A look at the convergence

α	N	PCG	Jackson	T.Chan	R.Chan	Strang
	2^5	15	6	8	2	2
	2^6	31	6	10	2	2
	2^7	63	6	12	2	2
2.0	2^8	127	5	13	2	2
	2^9	251	5	14	2	2
	2^{10}	464	5	15	2	2
	2^{11}	713	4	15	2	2



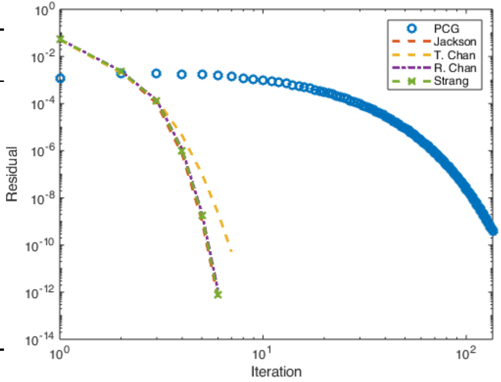
↓ A look at the convergence

α	N	PCG	Jackson	T.Chan	R.Chan	Strang
	2^5	15	6	8	5	5
	2^6	31	6	9	5	5
	2^7	61	6	9	5	5
1.8	2^8	108	6	11	5	5
	2^9	174	6	11	6	5
	2^{10}	234	6	11	6	6
	2^{11}	314	6	10	6	6



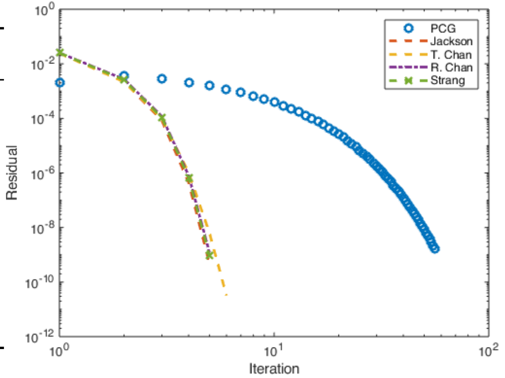
↓ A look at the convergence

α	N	PCG	Jackson	T.Chan	R.Chan	Strang
	2^5	15	6	7	5	5
	2^6	31	6	8	5	5
	2^7	51	6	8	5	5
1.6	2^8	73	5	8	5	5
	2^9	91	5	8	6	5
	2^{10}	111	6	7	6	6
	2^{11}	135	6	7	6	6



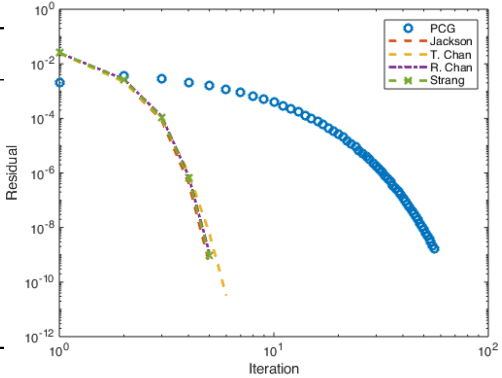
↓ A look at the convergence

α	N	PCG	Jackson	T.Chan	R.Chan	Strang
	2^5	15	5	7	5	5
	2^6	27	5	7	5	5
	2^7	35	5	7	5	5
1.4	2^8	41	5	6	5	5
	2^9	46	5	6	5	5
	2^{10}	51	5	6	5	5
	2^{11}	56	5	6	5	5



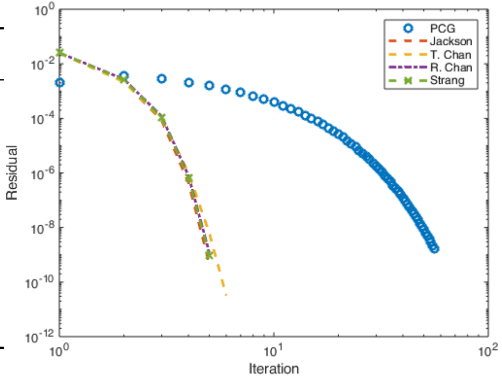
↓ A look at the convergence

α	N	PCG	Jackson	T.Chan	R.Chan	Strang
	2^5	15	5	6	4	4
	2^6	19	5	6	5	5
	2^7	20	5	5	5	5
1.2	2^8	21	5	5	5	5
	2^9	22	5	5	5	5
	2^{10}	22	5	5	5	5
	2^{11}	22	5	5	5	5



↓ A look at the convergence

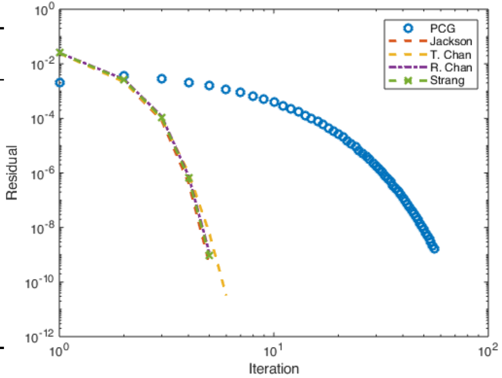
α	N	PCG	Jackson	T.Chan	R.Chan	Strang
	2^5	15	5	6	4	4
	2^6	19	5	6	5	5
	2^7	20	5	5	5	5
1.2	2^8	21	5	5	5	5
	2^9	22	5	5	5	5
	2^{10}	22	5	5	5	5
	2^{11}	22	5	5	5	5



👁 We got **robustness** with respect to both α and N.

A look at the convergence

α	N	PCG	Jackson	T.Chan	R.Chan	Strang
	2^5	15	5	6	4	4
	2^6	19	5	6	5	5
	2^7	20	5	5	5	5
1.2	2^8	21	5	5	5	5
	2^9	22	5	5	5	5
	2^{10}	22	5	5	5	5
	2^{11}	22	5	5	5	5



- We got **robustness** with respect to both α and N.
- What do we do in the non symmetric case, i.e., $\theta \neq 1/2$?

Non symmetric Toeplitz system

If $T_n(f)$ is non symmetric (or more generally, non Hermitian), then f is a complex-valued function then

- we **no longer** have information on the asymptotic **spectral distribution**, but only on the singular values,
- we can **no longer** apply **fast** direct Toeplitz **solvers**,
- we can **no longer** apply the **CG** to $T_n(f)\mathbf{x} = \mathbf{b}$.

❓ What to do?

Non symmetric Toeplitz system

If $T_n(f)$ is non symmetric (or more generally, non Hermitian), then f is a complex-valued function then

- we **no longer** have information on the asymptotic **spectral distribution**, but only on the singular values,
- we can **no longer** apply **fast** direct Toeplitz **solvers**,
- we can **no longer** apply the **CG** to $T_n(f)\mathbf{x} = \mathbf{b}$.

What to do?

Apply the PCG to the normal equations (CGNR):

$$T_n(f)^H T_n(f)\mathbf{x} = T_n(f)^H \mathbf{b},$$

Non symmetric Toeplitz system

If $T_n(f)$ is non symmetric (or more generally, non Hermitian), then f is a complex-valued function then

- we **no longer** have information on the asymptotic **spectral distribution**, but only on the singular values,
- we can **no longer** apply **fast** direct Toeplitz **solvers**,
- we can **no longer** apply the **CG** to $T_n(f)\mathbf{x} = \mathbf{b}$.

What to do?

Apply the PCG to the normal equations (CGNR):

$$T_n(f)^H T_n(f)\mathbf{x} = T_n(f)^H \mathbf{b},$$

Use another Krylov method: GMRES or TFQMR

Non symmetric Toeplitz system

If $T_n(f)$ is non symmetric (or more generally, non Hermitian), then f is a complex-valued function then

- we **no longer** have information on the asymptotic **spectral distribution**, but only on the singular values,
- we can **no longer** apply **fast** direct Toeplitz **solvers**,
- we can **no longer** apply the **CG** to $T_n(f)\mathbf{x} = \mathbf{b}$.

What to do?

Apply the PCG to the normal equations (CGNR):

$$T_n(f)^H T_n(f)\mathbf{x} = T_n(f)^H\mathbf{b},$$

Use another Krylov method: GMRES or TFQMR
do we know how to precondition these methods?

The GMRES method (Saad and Schultz 1986)

The **G**eneralized **M**inimum **Res**idual (GMRES) is a Krylov projection method approximating the solution of linear system

$$A\mathbf{x} = \mathbf{b}$$

on the **affine subspace**

$$\mathbf{x}^{(0)} + \mathcal{K}_m(A, \mathbf{v}_1), \quad \mathbf{r}^{(0)} = \mathbf{b} - A\,\mathbf{x}^{(0)}, \quad \mathbf{v}_1 = {}^{\mathbf{r}^{(0)}}/\|\mathbf{r}^{(0)}\|_2$$

, for $\mathbf{x}^{(0)}$ a *starting guess* for the solution.
By this choice, we enforce the **Arnoldi relation**:

$$A\,V_m = V_m H_m + \mathbf{w}_m \mathbf{e}_m^T = V_{m+1}\overline{H}_m, \quad \text{Span}\,V_m = \text{Span}\{\mathbf{v}_1 \ \cdots \ \mathbf{v}_m\} = \mathcal{K}_m(A, \mathbf{v}_1),$$

and H_m $m \times m$ Hessenberg submatrix extracted from $\overline{H}_m$ by deleting the $(m+1)$th line.

The GMRES method (Saad and Schultz 1986)

Input: $A \in \mathbb{R}^{n\times n}$, $\mathbf{b} \in \mathbb{R}^n$, m, $\mathbf{x}^{(0)}$
$\mathbf{r}^{(0)} \leftarrow \mathbf{b} - A\,\mathbf{x}^{(0)}$, $\beta \leftarrow \|\mathbf{r}^{(0)}\|_2$;
$\mathbf{v}_1 \leftarrow \mathbf{r}^{(0)}/\beta$;
for $j = 1, \ldots, m$ **do**
  $\mathbf{w}_j \leftarrow A\,\mathbf{v}_j$;
  for $i = 1, \ldots, j$ **do**
    $h_{i,j} \leftarrow < \mathbf{w}_j, \mathbf{v}_i >$;
    $\mathbf{w}_j \leftarrow \mathbf{w}_j - h_{i,j}\,\mathbf{v}_i$;
  end
  $h_{j+1,j} \leftarrow \|\mathbf{w}_j\|_2$;
  if $h_{j+1,j} = 0$ *or convergence* **then**
    $m = j$;
    break;
  end
  $\mathbf{v}_{j+1} = \mathbf{w}_j/\|\mathbf{w}_j\|_2$;
end

Compute $\mathbf{y}^{(m)}$ such that $\|\mathbf{r}^{(m)}\|_2 = \|\mathbf{b} - A\,\mathbf{x}^{(m)}\|_2 = \|\beta\mathbf{e}_1 - \underline{H}_m\mathbf{y}\|_2 = \min_{\mathbf{y}\in\mathbb{R}^m}$;
Build candidate approximation $\tilde{\mathbf{x}}$;

The GMRES method (Saad and Schultz 1986)

Input: $A \in \mathbb{R}^{n\times n}$, $\mathbf{b} \in \mathbb{R}^n$, m, $\mathbf{x}^{(0)}$

$\mathbf{r}^{(0)} \leftarrow \mathbf{b} - A\mathbf{x}^{(0)}$, $\beta \leftarrow \|\mathbf{r}^{(0)}\|_2$;

$\mathbf{v}_1 \leftarrow \mathbf{r}^{(0)}/\beta$;

for $j = 1, \ldots, m$ **do**

- $\mathbf{w}_j \leftarrow A\mathbf{v}_j$;
- **for** $i = 1, \ldots, j$ **do**
  - $h_{i,j} \leftarrow <\mathbf{w}_j, \mathbf{v}_i>$;
  - $\mathbf{w}_j \leftarrow \mathbf{w}_j - h_{i,j}\mathbf{v}_i$;
- **end**
- $h_{j+1,j} \leftarrow \|\mathbf{w}_j\|_2$;
- **if** $h_{j+1,j} = 0$ *or convergence* **then**
  - $m = j$;
  - **break**;
- **end**
- $\mathbf{v}_{j+1} = \mathbf{w}_j/\|\mathbf{w}_j\|_2$;

end

Compute $\mathbf{y}^{(m)}$ such that $\|\mathbf{r}^{(m)}\|_2 = \|\mathbf{b} - A\mathbf{x}^{(m)}\|_2 = \|\beta\mathbf{e}_1 - \underline{H}_m\mathbf{y}\|_2 = \min_{\mathbf{y}\in\mathbb{R}^m}$;

Build candidate approximation $\tilde{\mathbf{x}}$;

Minimizing the residual

At step m, the candidate solution $\mathbf{x}^{(m)}$ is the vector minimizing the 2–norm residual:

$$\|\mathbf{r}^{(m)}\|_2 = \|\mathbf{b} - A\mathbf{x}^{(m)}\|_2,$$

with

$$\mathbf{b} - A\mathbf{x}^{(m)} = V_{m+1}(\beta\mathbf{e}_1 - \overline{H}_m\mathbf{y}).$$

The GMRES method (Saad and Schultz 1986)

Input: $A \in \mathbb{R}^{n\times n}$,$\mathbf{b} \in \mathbb{R}^n$, m, $\mathbf{x}^{(0)}$
$\mathbf{r}^{(0)} \leftarrow \mathbf{b} - A\,\mathbf{x}^{(0)}$, $\beta \leftarrow \|\mathbf{r}^{(0)}\|_2$;
$\mathbf{v}_1 \leftarrow \mathbf{r}^{(0)}/\beta$;
for $j = 1, \ldots, m$ **do**
  $\mathbf{w}_j \leftarrow A\,\mathbf{v}_j$;
  for $i = 1, \ldots, j$ **do**
    $h_{i,j} \leftarrow <\mathbf{w}_j, \mathbf{v}_i>$;
    $\mathbf{w}_j \leftarrow \mathbf{w}_j - h_{i,j}\,\mathbf{v}_i$;
  end
  $h_{j+1,j} \leftarrow \|\mathbf{w}_j\|_2$;
  if $h_{j+1,j} = 0$ *or convergence* **then**
    $m = j$;
    break;
  end
  $\mathbf{v}_{j+1} = \mathbf{w}_j/\|\mathbf{w}_j\|_2$;
end

Compute $\mathbf{y}^{(m)}$ such that $\|\mathbf{r}^{(m)}\|_2 = \|\mathbf{b} - A\,\mathbf{x}^{(m)}\|_2 = \|\beta\mathbf{e}_1 - \underline{H}_m\mathbf{y}\|_2 = \min_{\mathbf{y}\in\mathbb{R}^m}$;
Build candidate approximation $\tilde{\mathbf{x}}$;

Minimizing the residual

At step m, the candidate solution $\mathbf{x}^{(m)}$ is the vector minimizing the 2–norm residual:

$$\|\mathbf{r}^{(m)}\|_2 = \|\mathbf{b} - A\,\mathbf{x}^{(m)}\|_2,$$

with

$$\mathbf{b} - A\,\mathbf{x}^{(m)} = V_{m+1}(\beta\mathbf{e}_1 - \overline{H}_m\mathbf{y}).$$

GMRES variants

Variants obtained by different least square problem solutions, and different orthogonalization algorithms.

The GMRES convergence theory (or lack thereof...)

Theorem (Convergence, diagonalizable)

If A can be diagonalized, i.e. if we can find $X \in \mathbb{R}^{n \times n}$ non singular and such that

$$A = X \Lambda X^{-1}, \; \Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n), \; K_2(X) = \|X\|_2 \, \|X^{-1}\|_2,$$

$K_2(X) = \|X\|_2 \, \|X^{-1}\|_2$ condition number of X, then at step m, we have

$$\|r\|_2 \leq K_2(X) \|\mathbf{r}^{(0)}\|_2 \min_{\substack{\mathrm{p}(z) \in \mathbb{P}_m \\ \mathrm{p}(0)=1}} \max_{i=1,\ldots,n} |\mathrm{p}(\lambda_i)|, \tag{DiagGMRES}$$

where $\mathrm{p}(z)$ is the polynomial of degree less or equal to m such that $\mathrm{p}(0) = 1$ and the expression in the right hand side of (DiagGMRES) is minimum.

⚠ The eigenvectors can be arbitrarily *ill-conditioned*, i.e., $K_2(X) \ggg 1$,

❗ being **diagonalizable** can be a **strong assumption**.

The GMRES convergence theory (or lack thereof...)

Theorem (Almostr everything is possible) (Greenbaum, Pták, and Strakoš 1996)

Given a non-increasing positive sequence $\{f_k\}_{k=0,\ldots,n-1}$ with $f_{n-1} > 0$ and a set of non–zero complex numbers $\{\lambda_i\}_{i=1,2,\ldots,n} \subset \mathbb{C}$, there exist a matrix A with eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ and a right-hand side $\mathbf{b}$ with $\|\mathbf{b}\| = f_0$ such that the residual vectors $\mathbf{r}^{(k)}$ at each step of the GMRES algorithm applied to solve $A\mathbf{x} = \mathbf{b}$ with $\mathbf{x}^{(0)} = \mathbf{0}$, satisfy $\|\mathbf{r}^{(k)}\| = f_k$, $\forall\, k = 1, 2, \ldots, n-1$.

"Any non-increasing convergence curve is possible for GMRES".

In the clustered case we can partition $\sigma(A)$ as follows

$$\sigma(A) = \sigma_c(A) \cup \sigma_0(A) \cup \sigma_1(A),$$

where

- $\sigma_c(A)$ denotes the **clustered set** of eigenvalues of A,
- $\sigma_0(A) \cup \sigma_1(A)$ denotes the **set of the outliers**.

The GMRES convergence theory (or lack thereof...)

Theorem (Almostr everything is possible) (Greenbaum, Pták, and Strakoš 1996)

Given a non-increasing positive sequence $\{f_k\}_{k=0,\ldots,n-1}$ with $f_{n-1} > 0$ and a set of non–zero complex numbers $\{\lambda_i\}_{i=1,2,\ldots,n} \subset \mathbb{C}$, there exist a matrix A with eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ and a right-hand side $\mathbf{b}$ with $\|\mathbf{b}\| = f_0$ such that the residual vectors $\mathbf{r}^{(k)}$ at each step of the GMRES algorithm applied to solve $A\mathbf{x} = \mathbf{b}$ with $\mathbf{x}^{(0)} = \mathbf{0}$, satisfy $\|\mathbf{r}^{(k)}\| = f_k$, $\forall\, k = 1, 2, \ldots, n-1$.

- "Any non-increasing convergence curve is possible for GMRES".
- What happens if we have **a clustered spectrum**?
- In the clustered case we can partition $\sigma(A)$ as follows

$$\sigma(A) = \sigma_c(A) \cup \sigma_0(A) \cup \sigma_1(A),$$

where

- $\sigma_c(A)$ denotes the **clustered set** of eigenvalues of A,
- $\sigma_0(A) \cup \sigma_1(A)$ denotes the **set of the outliers**.

GMRES in the clustered and diagonalizable case

$$\sigma(A) = \underbrace{\sigma_c(A)}_{\text{clustered}} \cup \underbrace{\sigma_0(A) \cup \sigma_1(A)}_{\text{outliers}},$$

we assume that

1. the clustered set $\sigma_c(A)$ of eigenvalues is contained in a convex set Ω,
2. and, that denoting two sets of j_0 and j_1 outliers as

$$\sigma_0(A) = \{\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_{j_0}\} \quad \text{and} \quad \sigma_1(A) = \{\tilde{\lambda}_1, \tilde{\lambda}_2, \ldots, \tilde{\lambda}_{j_1}\}$$

where if $\hat{\lambda}_j \in \sigma_0(A)$, we have

$$1 < |1 - z/\hat{\lambda}_j| \leq c_j, \quad \forall z \in \Omega,$$

while, for $\tilde{\lambda}_j \in \sigma_1(A)$,

$$0 < |1 - z/\tilde{\lambda}_j| < 1, \quad \forall z \in \Omega,$$

GMRES in the clustered and diagonalizable case

Theorem

The number of full GMRES iterations j needed to attain a tolerance ε on the relative residual in the 2-norm $\|\mathbf{r}^{(j)}\|_2/\|\mathbf{r}^{(0)}\|_2$ for the linear system $A\mathbf{x} = \mathbf{b}$, where A is diagonalizable, is bounded above by

$$\min\left\{ j_0 + j_1 + \left\lceil \frac{\log(\varepsilon) - \log(\kappa_2(X))}{\log(\rho)} - \sum_{\ell=1}^{j_0} \frac{\log(c_\ell)}{\log(\rho)} \right\rceil, n \right\},$$

where

$$\rho^k = \frac{\left(a/d + \sqrt{(a/d)^2 - 1}\right)^k + \left(a/d + \sqrt{(a/d)^2 - 1}\right)^{-k}}{\left(c/d + \sqrt{(c/d)^2 - 1}\right)^k + \left(c/d + \sqrt{(c/d)^2 - 1}\right)^{-k}},$$

and the set $\Omega \in \mathbb{C}^+$ is the ellipse with center c, focal distance d and major semi axis a.

GMRES the non-diagonalizable case

In this case we have to turn to either the **field of values** or the ε-**pseudospectra** of A.
We need to bound the right-hand side of

$$\|\mathbf{r}_m\|_2 \leq \min_{\substack{p(z)\in\mathbb{P}_m\\ p(0)=1}} \|p(A)\mathbf{r}_0\|, \quad m = 1, 2, \dots$$

or in the worst case scenario

$$\frac{\|\mathbf{r}_m\|_2}{\|\mathbf{r}_0\|} \leq \max_{\substack{\mathbf{v}\in\mathbb{C}^n\\ \|\mathbf{v}\|=1}} \min_{\substack{p(z)\in\mathbb{P}_m\\ p(0)=1}} \|p(A)\mathbf{v}\|, \quad m = 1, 2, \dots$$

- If A is real, and $M = {}^{(A+A^T)}/_2$ is SPD, then (Eisenstat, Elman, and Schultz 1983)

$$\max_{\substack{\mathbf{v}\in\mathbb{R}^n\\ \|\mathbf{v}\|=1}} \min_{\substack{p(z)\in\mathbb{P}_m\\ p(0)=1}} \|p(A)\mathbf{v}\| \leq \left(1 - \frac{\lambda_{\min}(M)^2}{\lambda_{\max}(A^TA)}\right)^{m/2}.$$

GMRES the non-diagonalizable case

$$\|\mathbf{r}_m\|_2 \leq \min_{\substack{\mathrm{p}(z)\in\mathbb{P}_m \\ \mathrm{p}(0)=1}} \|\mathrm{p}(A)\mathbf{r}_0\|, \quad m = 1, 2, \dots$$

we recall that the **field of values** of A is given by

$$W(A) = \{< A\mathbf{v}, \mathbf{v} >: \mathbf{v} \in \mathbb{C}^n, \|\mathbf{v}\| = 1\}, \qquad \nu(A) = \min_{z\in W(A)} |z|,$$

with $\nu(A)$ the distance of $W(A)$ from the origin.

GMRES the non-diagonalizable case

$$\|\mathbf{r}_m\|_2 \leq \min_{\substack{\mathrm{p}(z)\in\mathbb{P}_m\\ \mathrm{p}(0)=1}} \|\mathrm{p}(A)\mathbf{r}_0\|, \quad m = 1, 2, \ldots$$

we recall that the **field of values** of A is given by

$$W(A) = \{< A\mathbf{v}, \mathbf{v} >:\ \mathbf{v} \in \mathbb{C}^n,\ \|\mathbf{v}\| = 1\}, \qquad \nu(A) = \min_{z\in W(A)} |z|,$$

with $\nu(A)$ the distance of $W(A)$ from the origin.

- For a general nonsingular A (Eiermann and Ernst 2001)

$$\max_{\substack{\mathbf{v}\in\mathbb{C}^n\\ \|\mathbf{v}\|=1}} \min_{\substack{\mathrm{p}(z)\in\mathbb{P}_m\\ \mathrm{p}(0)=1}} \|\mathrm{p}(A)\mathbf{v}\| \leq (1 - \nu(A)\nu(A^{-1}))^{m/2}.$$

GMRES the non-diagonalizable case

$$\|\mathbf{r}_m\|_2 \leq \min_{\substack{\mathrm{p}(z)\in\mathbb{P}_m\\ \mathrm{p}(0)=1}} \|\mathrm{p}(A)\mathbf{r}_0\|, \quad m = 1, 2, \ldots$$

we recall that the **field of values** of A is given by

$$W(A) = \{< A\mathbf{v}, \mathbf{v} >: \mathbf{v} \in \mathbb{C}^n, \ \|\mathbf{v}\| = 1\}, \qquad \nu(A) = \min_{z\in W(A)} |z|,$$

with $\nu(A)$ the distance of $W(A)$ from the origin.

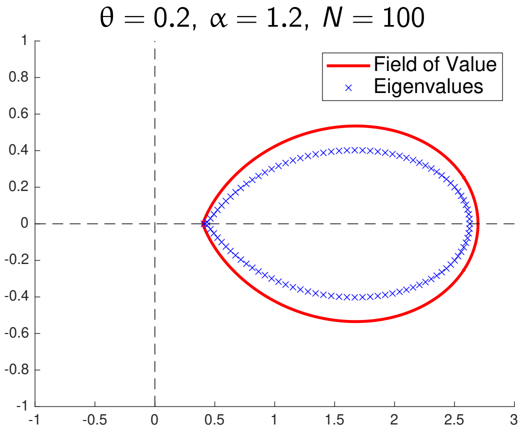
- For a general nonsingular A (Eiermann and Ernst 2001)

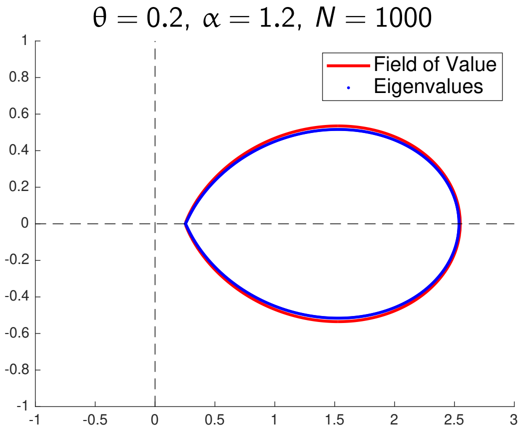
$$\max_{\substack{\mathbf{v}\in\mathbb{C}^n\\ \|\mathbf{v}\|=1}} \min_{\substack{\mathrm{p}(z)\in\mathbb{P}_m\\ \mathrm{p}(0)=1}} \|\mathrm{p}(A)\mathbf{v}\| \leq (1 - \nu(A)\nu(A^{-1}))^{m/2}.$$

- This bound is useful only when $0 \notin W(A)$ and $0 \notin W(A^{-1})$.

Some experimentation with the FOV in our case

$$\nu_N^{\alpha-1} A_N = \nu_N^{\alpha-1} I_N - \theta G_N + (1-\theta) G_N^T,$$

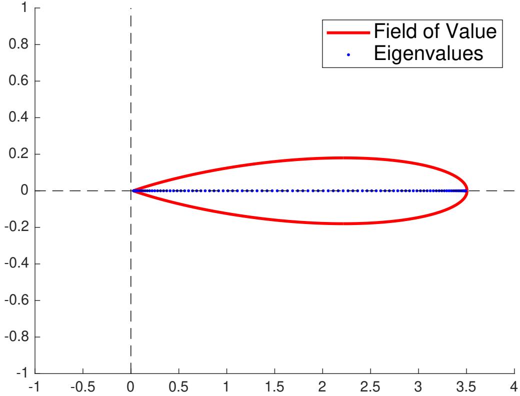




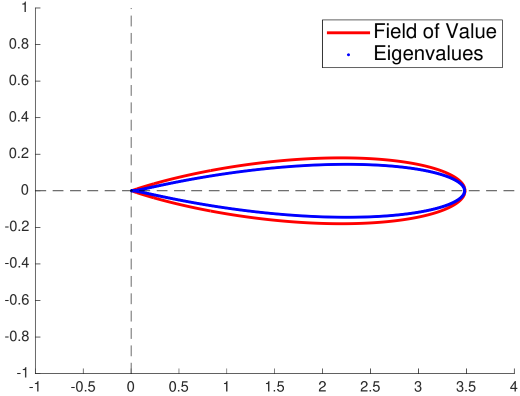
Some experimentation with the FOV in our case

$$\nu_N^{\alpha-1} A_N = \nu_N^{\alpha-1} I_N - \theta G_N + (1-\theta) G_N^T,$$

$\theta = 0.2,\ \alpha = 1.8,\ N = 100$

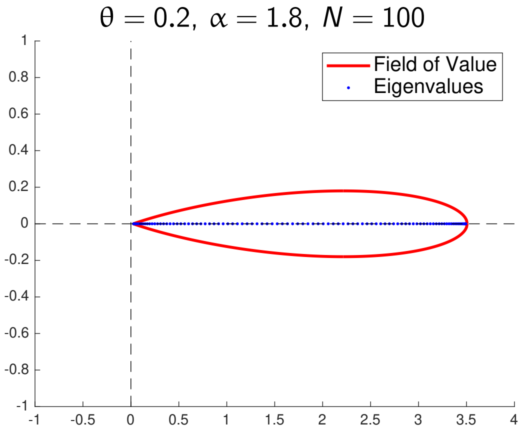


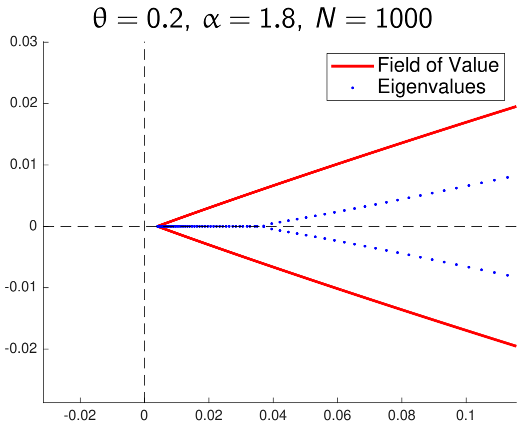
$\theta = 0.2,\ \alpha = 1.8,\ N = 1000$



Some experimentation with the FOV in our case

$$\nu_N^{\alpha-1} A_N = \nu_N^{\alpha-1} I_N - \theta G_N + (1-\theta) G_N^T,$$





Some experimentation with the FOV in our case

Unfortunate truth

In general it is difficult to say something about the Field of Value of preconditioned matrices.

Some experimentation with the FOV in our case

Unfortunate truth

In general it is difficult to say something about the Field of Value of preconditioned matrices.

- What do we do in practice?

  "To speed up the CG-like methods, we can choose a matrix C such that the singular values of the preconditioned matrix $C^{-1}A$ are clustered." – (R. H. Chan and Ng 1996, P. 439)

Some experimentation with the FOV in our case

Unfortunate truth

In general it is difficult to say something about the Field of Value of preconditioned matrices.

- What do we do in practice?

  "To speed up the CG-like methods, we can choose a matrix C such that the singular values of the preconditioned matrix $C^{-1}A$ are clustered." – (R. H. Chan and Ng 1996, P. 439)

- How do we build a **Circulant preconditioner** for a **our non-symmetric Toeplitz** matrix?

Some experimentation with the FOV in our case

Unfortunate truth

In general it is difficult to say something about the Field of Value of preconditioned matrices.

- What do we do in practice?

  "To speed up the CG-like methods, we can choose a matrix C such that the singular values of the preconditioned matrix $C^{-1}A$ are clustered." – (R. H. Chan and Ng 1996, P. 439)

- How do we build a **Circulant preconditioner** for a **our non-symmetric Toeplitz** matrix?
- We can use a suitably modified Strang preconditioner for our case (Lei and Sun 2013)

A Circulant preconditioner (Lei and Sun 2013)

We can build a circulant preconditioner as

$$P = \frac{h_N^\alpha}{\Delta t} I_N + \theta s(G_N) + (1-\theta) s(G_N^T),$$

where

$$(s(G_N))_{:,1} = - \begin{bmatrix} g_1^{(\alpha)} \\ \vdots \\ g^{\alpha}_{\lfloor (N+1)/2 \rfloor} \\ 0 \\ \vdots \\ 0 \\ g_0^{(\alpha)} \end{bmatrix},$$

```
function [ev,evt] = sunprec(N,alpha)
g = gl(N,alpha);
v = zeros(N,1);
v(1:floor((N+1)/2)) =
↪  g((1:floor((N+1)/2))+1);
v(end) = g(1);
ev = fft(-v);
v = zeros(N,1);
v(1) = g(2);
v(2) = g(1);
v(end:-1:floor((N+1)/2)+2) =
↪  g(3:floor((N+1)/2)+1);
evt = fft(-v);
end
```

A Circulant preconditioner (Lei and Sun 2013)

We can build a circulant preconditioner as

$$P = \frac{h_N^\alpha}{\Delta t} I_N + \theta s(G_N) + (1-\theta) s(G_N^T),$$

where

$$(s(G_N^T))_{:,1} = -\begin{bmatrix} g_1^{(\alpha)} \\ g_0^{(\alpha)} \\ 0 \\ \vdots \\ 0 \\ g^{\alpha}_{\lfloor (N+1)/2 \rfloor} \\ \vdots \\ g_2^{(\alpha)} \end{bmatrix}.$$

```
function [ev,evt] = sunprec(N,alpha)
g = gl(N,alpha);
v = zeros(N,1);
v(1:floor((N+1)/2)) =
 ↪  g((1:floor((N+1)/2))+1);
v(end) = g(1);
ev = fft(-v);
v = zeros(N,1);
v(1) = g(2);
v(2) = g(1);
v(end:-1:floor((N+1)/2)+2) =
 ↪  g(3:floor((N+1)/2)+1);
evt = fft(-v);
end
```

A Circulant preconditioner (Lei and Sun 2013)

We can build a circulant preconditioner as

$$P = \frac{h_N^\alpha}{\Delta t} I_N + \theta s(G_N) + (1-\theta)s(G_N^T),$$

- It uses the **construction of the Strang preconditioner** using only *half o the bandwidth* of the Toeplitz matrices.

```
function [ev,evt] = sunprec(N,alpha)
g = gl(N,alpha);
v = zeros(N,1);
v(1:floor((N+1)/2)) =
 ↪  g((1:floor((N+1)/2))+1);
v(end) = g(1);
ev = fft(-v);
v = zeros(N,1);
v(1) = g(2);
v(2) = g(1);
v(end:-1:floor((N+1)/2)+2) =
 ↪  g(3:floor((N+1)/2)+1);
evt = fft(-v);
end
```

A Circulant preconditioner (Lei and Sun 2013)

We can build a circulant preconditioner as

$$P = \frac{h_N^\alpha}{\Delta t} I_N + \theta s(G_N) + (1-\theta) s(G_N^T),$$

- It uses the **construction of the Strang preconditioner** using only *half o the bandwidth* of the Toeplitz matrices.
- All the eigenvalues of $s(G_N)$ and $s(G_N^T)$ fall inside the open disc $\{z \in \mathbb{C} \,:\, |z - \alpha| < \alpha\}$ by Gershgorin theorem, indeed:

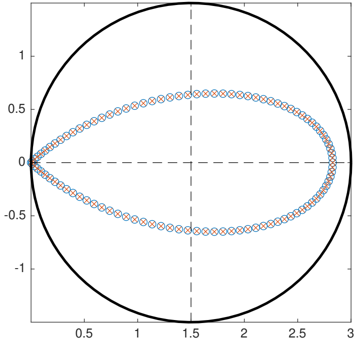
$$r_N = g_0^\alpha + \sum_{k=2}^{\lfloor (N+1)/2 \rfloor} < \sum_{\substack{k=0\\k\neq 1}} g_k^{(\alpha)} = -g_1^{(\alpha)} = \alpha.$$

```
function [ev,evt] = sunprec(N,alpha)
g = gl(N,alpha);
v = zeros(N,1);
v(1:floor((N+1)/2)) =
↪  g((1:floor((N+1)/2))+1);
v(end) = g(1);
ev = fft(-v);
v = zeros(N,1);
v(1) = g(2);
v(2) = g(1);
v(end:-1:floor((N+1)/2)+2) =
↪  g(3:floor((N+1)/2)+1);
evt = fft(-v);
end
```

A Circulant preconditioner (Lei and Sun 2013)

We can build a circulant preconditioner as

$$P = \frac{h_N^{\alpha}}{\Delta t} I_N + \theta s(G_N) + (1 - \theta) s(G_N^T),$$




```
function [ev,evt] = sunprec(N,alpha)
g = gl(N,alpha);
v = zeros(N,1);
v(1:floor((N+1)/2)) =
  ↪  g((1:floor((N+1)/2))+1);
v(end) = g(1);
ev = fft(-v);
v = zeros(N,1);
v(1) = g(2);
v(2) = g(1);
v(end:-1:floor((N+1)/2)+2) =
  ↪  g(3:floor((N+1)/2)+1);
evt = fft(-v);
end
```

A Circulant preconditioner (Lei and Sun 2013)

Will it work?
We can always write:

$$P^{-1}A_N - I_N = P^{-1}(A_N - P) \qquad ,$$

now for the Strang preconditioner of a Toeplitz matrix with with generating function in the Wiener class, it holds that for any $\varepsilon > 0$ exists N' and M' such that

$$A_N - s(A_N) = U_N + V_N, \quad \operatorname{rank}(U_N) \leq M' \text{ and } \|V_N\|_2 < \varepsilon \; \forall N > N'.$$

A Circulant preconditioner (Lei and Sun 2013)

Will it work?

We can always write:

$$P^{-1}A_N - I_N = P^{-1}(A_N - P) = P_N^{-1}U_N - P_N^{-1}V_N,$$

now for the Strang preconditioner of a Toeplitz matrix with with generating function in the Wiener class, it holds that for any $\varepsilon > 0$ exists N' and M' such that

$$A_N - s(A_N) = U_N + V_N, \quad \mathrm{rank}(U_N) \leq M' \text{ and } \|V_N\|_2 < \varepsilon \;\forall N > N'.$$

A Circulant preconditioner (Lei and Sun 2013)

Will it work?
We can always write:

$$P^{-1}A_N - I_N = P^{-1}(A_N - P) = P_N^{-1}U_N - P_N^{-1}V_N,$$

now for the Strang preconditioner of a Toeplitz matrix with with generating function in the Wiener class, it holds that for any $\varepsilon > 0$ exists N' and M' such that

$$A_N - s(A_N) = U_N + V_N, \quad \mathrm{rank}(U_N) \leq M' \text{ and } \|V_N\|_2 < \varepsilon \; \forall N > N'.$$

$\mathrm{rank}(P_N^{-1}U_N) \leq \mathrm{rank}(U_N) \leq M'$,

A Circulant preconditioner (Lei and Sun 2013)

Will it work?

We can always write:

$$P^{-1}A_N - I_N = P^{-1}(A_N - P) = P_N^{-1}U_N - P_N^{-1}V_N,$$

now for the Strang preconditioner of a Toeplitz matrix with with generating function in the Wiener class, it holds that for any $\varepsilon > 0$ exists N' and M' such that

$$A_N - s(A_N) = U_N + V_N, \quad \mathrm{rank}(U_N) \leq M' \text{ and } \|V_N\|_2 < \varepsilon \; \forall N > N'.$$

- $\mathrm{rank}(P_N^{-1}U_N) \leq \mathrm{rank}(U_N) \leq M'$,
- $\forall\, k = 1, 2, \ldots, N$, $|\lambda(P_N)| \geq \Re(\Lambda(P_N)_{k,k}) = {}^{h_N^{\alpha}}/_{\Delta t} + \theta\Re(\Lambda(s(G_N))_{kk}) + (1-\theta)\Re(\Lambda(s(G_N^T))_{kk}) \geq {}^{h_N^{\alpha}}/_{\Delta t} > 0$ and thus $\|P_N^{-1}\|_2 \leq {}^{\Delta t}/_{h_N^{\alpha}}$

A Circulant preconditioner (Lei and Sun 2013)

Will it work?

We can always write:

$$P^{-1}A_N - I_N = P^{-1}(A_N - P) = P_N^{-1}U_N - P_N^{-1}V_N,$$

now for the Strang preconditioner of a Toeplitz matrix with with generating function in the Wiener class, it holds that for any $\varepsilon > 0$ exists N' and M' such that

$$A_N - s(A_N) = U_N + V_N, \quad \operatorname{rank}(U_N) \leq M' \text{ and } \|V_N\|_2 < \varepsilon \; \forall N > N'.$$

- $\operatorname{rank}(P_N^{-1}U_N) \leq \operatorname{rank}(U_N) \leq M'$,
- $\|P_N^{-1}V_N\| \leq \|P_N^{-1}\|_2 \|V_N\|_2 < {}^{\varepsilon\Delta t}/_{h_N^{\alpha}}$.

A Circulant preconditioner (Lei and Sun 2013)

Will it work?
We can always write:

$$P^{-1}A_N - I_N = P^{-1}(A_N - P) = P_N^{-1}U_N - P_N^{-1}V_N \Rightarrow \text{“small rank”} \; + \; \text{“small norm”},$$

now for the Strang preconditioner of a Toeplitz matrix with with generating function in the Wiener class, it holds that for any $\varepsilon > 0$ exists N' and M' such that

$$A_N - s(A_N) = U_N + V_N, \quad \operatorname{rank}(U_N) \leq M' \text{ and } \|V_N\|_2 < \varepsilon \; \forall N > N'.$$

- $\operatorname{rank}(P_N^{-1}U_N) \leq \operatorname{rank}(U_N) \leq M'$,
- $\|P_N^{-1}V_N\| \leq \|P_N^{-1}\|_2 \|V_N\|_2 < {}^{\varepsilon\Delta t}/_{h_N^{\alpha}}$.
- If we select Δt and h_N in such a way that ${}^{h_N^{\alpha}}/_{\Delta t}$ is bounded and bounded away from zero we have the result.

Results with GMRES

$$\left(\frac{h_N^{\alpha}}{\Delta t} I_{N-2} - \left[\theta G_{N-2} + (1-\theta) G_{N-2}^{T} \right] \right) \mathbf{w}^{n+1} = \frac{h_N^{\alpha}}{\Delta t}, \quad \theta = 0.2$$

Results with GMRES

```
[ev,evt] = sunprec(N,alpha);
c = nu + theta*ev + (1-theta)*evt;
P = @(x) cprec(c,x);
[X,FLAGsun,RELRESsun,ITERsun,RESVECsun] = gmres(A,(nu*w),[],1e-9,N,P);
```

α	N	GMRES	P
	2^5	28	6
	2^6	31	6
	2^7	33	6
1.2	2^8	34	6
	2^9	35	6
	2^{10}	36	6
	2^{11}	36	6

α	N	GMRES	P
	2^5	31	6
	2^6	46	6
	2^7	54	6
1.4	2^8	62	7
	2^9	69	7
	2^{10}	78	7
	2^{11}	87	7

α	N	GMRES	P
	2^5	32	6
	2^6	59	6
	2^7	82	7
1.6	2^8	105	7
	2^9	128	7
	2^{10}	156	7
	2^{11}	189	7

α	N	GMRES	P
	2^5	32	6
	2^6	64	6
	2^7	109	6
1.8	2^8	162	7
	2^9	222	7
	2^{10}	287	7
	2^{11}	372	7

Conclusion and summary

- We have discussed the solution of Toeplitz linear systems,
- Studied the usage and convergence of PCG and GMRES method,
- Tested the usage of Circulant preconditioners for Toeplitz linear systems.

Next up

- We need to discuss the next problem in difficulty

$$\begin{cases} \frac{\partial W}{\partial t} = d^{+}(x,t)\, {}^{RL}D^{\alpha}_{[0,x]} W(x,t) + d^{-}(x,t)\, {}^{RL}D^{\alpha}_{[x,1]} W(x,t), & \theta \in [0,1], \\ W(0,t) = W(1,t) = 0, \qquad W(x,t) = W_0(x). \end{cases}$$

- What happens if we go to **more than one spatial dimension**?

Bibliography I

- Ammar, G. S. and W. B. Gragg (1988). "Superfast solution of real positive definite Toeplitz systems". In: *SIAM J. Matrix Anal. Appl.* 9.1, pp. 61–76.
- Bini, D. A. and B. Meini (1999). "Effective methods for solving banded Toeplitz systems". In: *SIAM J. Matrix Anal. Appl.* 20.3, pp. 700–719. ISSN: 0895-4798. DOI: 10.1137/S0895479897324585. URL: https://doi.org/10.1137/S0895479897324585.
- Bitmead, R. R. and B. D. Anderson (1980). "Asymptotically fast solution of Toeplitz and related systems of linear equations". In: *Linear Algebra Appl.* 34, pp. 103–116.
- Brent, R. P., F. G. Gustavson, and D. Y. Yun (1980). "Fast solution of Toeplitz systems of equations and computation of Padé approximants". In: *J. Algorithms* 1.3, pp. 259–295.
- Chan, R. H. and M. K. Ng (1996). "Conjugate gradient methods for Toeplitz systems". In: *SIAM Rev.* 38.3, pp. 427–482. ISSN: 0036-1445. DOI: 10.1137/S0036144594276474. URL: https://doi.org/10.1137/S0036144594276474.

Bibliography II

- Chan, R. H., M. K. Ng, and A. M. Yip (2002). "The best circulant preconditioners for Hermitian Toeplitz systems. II. The multiple-zero case". In: *Numer. Math.* 92.1, pp. 17–40. ISSN: 0029-599X. DOI: 10.1007/s002110100354. URL: https://doi.org/10.1007/s002110100354.
- Chan, R. H. and M.-C. Yeung (1992). "Circulant preconditioners constructed from kernels". In: *SIAM J. Numer. Anal.* 29.4, pp. 1093–1103. ISSN: 0036-1429. DOI: 10.1137/0729066. URL: https://doi.org/10.1137/0729066.
- Chan, T. F. and P. C. Hansen (1992). "A look–ahead Levinson algorithm for general Toeplitz systems". In: *IEEE Transactions on signal processing* 40.5, pp. 1079–1090.
- Donatelli, M., M. Mazza, and S. Serra-Capizzano (2016). "Spectral analysis and structure preserving preconditioners for fractional diffusion equations". In: *J. Comput. Phys.* 307, pp. 262–279. ISSN: 0021-9991. DOI: 10.1016/j.jcp.2015.11.061. URL: https://doi.org/10.1016/j.jcp.2015.11.061.

Bibliography III

- Eiermann, M. and O. G. Ernst (2001). "Geometric aspects of the theory of Krylov subspace methods". In: *Acta Numer.* 10, pp. 251–312. ISSN: 0962-4929. DOI: 10.1017/S0962492901000046. URL: https://doi.org/10.1017/S0962492901000046.
- Eisenstat, S. C., H. C. Elman, and M. H. Schultz (1983). "Variational iterative methods for nonsymmetric systems of linear equations". In: *SIAM J. Numer. Anal.* 20.2, pp. 345–357. ISSN: 0036-1429. DOI: 10.1137/0720023. URL: https://doi.org/10.1137/0720023.
- Gohberg, I. C. and A. A. Semencul (1972). "The inversion of finite Toeplitz matrices and their continual analogues". In: *Mat. Issled.* 7.2(24), pp. 201–223, 290. ISSN: 0542-9994.
- Greenbaum, A., V. Pták, and Z. Strakoš (1996). "Any Nonincreasing Convergence Curve is Possible for GMRES". In: *SIAM J. Matrix Anal. Appl.* 17.3, pp. 465–469. DOI: 10.1137/S0895479894275030. eprint: http://dx.doi.org/10.1137/S0895479894275030. URL: http://dx.doi.org/10.1137/S0895479894275030.
- Hoog, F. de (1987). "A new algorithm for solving Toeplitz systems of equations". In: *Linear Algebra Appl.* 88, pp. 123–138.

Bibliography IV

- Lei, S.-L. and H.-W. Sun (2013). "A circulant preconditioner for fractional diffusion equations". In: *J. Comput. Phys.* 242, pp. 715–725. ISSN: 0021-9991. DOI: 10.1016/j.jcp.2013.02.025. URL: https://doi.org/10.1016/j.jcp.2013.02.025.
- Levinson, N. (1946). "The Wiener (root mean square) error criterion in filter design and prediction". In: *J. Math. Phys.* 25.1, pp. 261–278.
- Saad, Y. and M. H. Schultz (1986). "GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems". In: *SIAM J. Sci. Statist. Comput.* 7.3, pp. 856–869. ISSN: 0196-5204. DOI: 10.1137/0907058. URL: https://doi.org/10.1137/0907058.
- Trench, W. F. (1964). "An algorithm for the inversion of finite Toeplitz matrices". In: *SIAM J. Appl. Math.* 12.3, pp. 515–522.
- Zohar, S. (1974). "The solution of a Toeplitz set of linear equations". In: *J. Assoc. Comput. Mach.* 21.2, pp. 272–276.